Available online at www.sciencedirect.com
ScienceDirect
Journal of Computer and System Sciences 73 (2007) 924–940
JOURNAL OF COMPUTER AND SYSTEM SCIENCES
www.elsevier.com/locate/jcss

A general dimension for query learning ☆

José L. Balcázar [a,*], Jorge Castro [a], David Guijarro [b], Johannes Köbler [c], Wolfgang Lindner [d]

[a] *Departament de LSI, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain*
[b] *Mannes Technology Consulting, Pl. Tirant lo Blanc 7, 08005 Barcelona, Spain*
[c] *Institut für Informatik, Humboldt-Universität zu Berlin, 10099 Berlin, Germany*
[d] *Fakultät für Informatik, Universität Ulm, D-89069 Ulm, Germany*

Received 7 February 2006; received in revised form 28 February 2007

Available online 12 March 2007

Abstract

We introduce a combinatorial dimension that characterizes the number of queries needed to exactly (or approximately) learn concept classes in various models. Our general dimension provides tight upper and lower bounds on the query complexity for all sorts of queries, not only for example-based queries as in previous works.

As an application we show that for learning DNF formulas, unspecified attribute value membership and equivalence queries are not more powerful than standard membership and equivalence queries. Further, in the approximate learning setting, we use the general dimension to characterize the query complexity in the statistical query as well as the learning by distances model. Moreover, we derive close bounds on the number of statistical queries needed to approximately learn DNF formulas.

Keywords: Query learning; UAV queries; Learning by distances; Statistical queries; Learning DNF formulas

1. Introduction

Starting with Angluin's seminal paper [1] a variety of different query types has been investigated in learning theory. The *query complexity* of a concept class C is the minimum number of queries needed to learn C and provides significant information on the difficulty of learning C in a specific learning model. Therefore, determining (or approximating) the query complexity is an important task. To this end, combinatorial notions have been introduced for many query types, as e.g., certificates [19] and consistency dimension [5] for membership and equivalence queries,

☆ Results from Sections 3–5 were presented at COLT/EUROCOLT 2001 [J.L. Balcázar, J. Castro, D. Guijarro, A general dimension for exact learning, in: Proceedings of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory, in: Lecture Notes in Artificial Intelligence, vol. 2111, Springer, 2001, pp. 354–367]; results from Sections 6 and 7 were presented at ALT 2002 [J. Köbler, W. Lindner, A general dimension for approximately learning Boolean functions, in: Algorithmic Learning Theory, Proceedings of the 13th International Conference, ALT 2002, Lübeck, Germany, November 2002, in: Lecture Notes in Artificial Intelligence, vol. 2533, Springer, 2002, pp. 139–148].

* Corresponding author.
E-mail addresses: balqui@lsi.upc.edu (J.L. Balcázar), castro@lsi.upc.edu (J. Castro), david@mannes-tech.com (D. Guijarro), koebler@informatik.hu-berlin.de (J. Köbler), lindner@informatik.uni-ulm.de (W. Lindner).

0022-0000/$ – see front matter
doi:10.1016/j.jcss.2007.03.003

approximate fingerprints [2,15] and strong consistency dimension [5] for equivalence queries and extended teaching dimension [18] for membership queries.

Furthermore, there is a unifying concept, the abstract identification dimension AIdim [4], that gives a single characterization of the query complexity for all types of *example-based* queries. An answer to a query of this type fully specifies the target function f on a subset X of its domain. Hence, it can be described by a set of examples $S = \{(x, f(x)) \mid x \in X\}$.

All query types mentioned above are example-based. On the other hand, several interesting query types are not example-based as, e.g., the restricted equivalence queries [1] and unspecified attribute value (UAV) queries (see [7, 10,17]). A restricted equivalence query is answered by YES or NO, i.e., there are no counterexamples. UAV queries involve examples in which some of the attributes may remain unspecified. The most popular queries in UAV learning are UAV membership and UAV equivalence queries that are natural extensions of their ordinary counterparts. For some of these query types, AIdim turns out to be unable to provide a useful approximation for the query complexity (see Examples 6 and 10).

In this paper we introduce a new combinatorial notion Gdim and show that it characterizes the query complexity for *any* type of queries. In fact, it turns out that the general dimension essentially coincides with AIdim for example-based queries, whereas in general they can be far apart. We note in passing that although the general dimension only characterizes the query complexity and hence provides only a lower bound for the time complexity, it has been shown in [22] that a polynomially bounded value of the general dimension implies polynomial-time learnability with additional access to an oracle in a low level of the polynomial time hierarchy.

As a main application of the general dimension in the exact learning setting we show that for learning DNF formulas, UAV memberships and equivalences are not superior to their ordinary counterparts. To prove this result we introduce UAV versions of the consistency dimension and the extended teaching dimension and use the general dimension to show that these dimensions tightly approximate the query complexity in the corresponding learning models.

Furthermore, we extend the general dimension to approximate learning and show that also in this setting it characterizes the query complexity. As an application, we get a characterization of the query complexity in the statistical query and the learning by distances model. In contrast to the SQ-dimension introduced by Blum et al. [8], the general dimension works for any reasonable choice of the error parameter ε and the tolerance parameter τ. Finally, we consider the problem of learning DNF formulas (over n variables and m terms) with statistical queries and show that for any constant error $\varepsilon < 1/2$ and a suitable choice of the tolerance $\tau = \Theta(1/m)$, the query complexity of this class with respect to the uniform distribution is $n^{\Theta(\log m)}$.

The paper is organized as follows. In Section 2 we provide the framework for query learning that we use in this paper. In Section 3 we introduce our new dimension and show that it is appropriate to characterize the query complexity in the exact learning setting. In Section 4 we apply the general dimension to UAV learning and in Section 5 we investigate the problem of learning DNF formulas in this model. Then, in Section 6, we extend the general dimension to the approximate learning setting and in Section 7 we consider the problem of learning DNF formulas with statistical queries.

2. A framework for query learning

The cardinality of a finite set X is denoted by $\|X\|$. We use $\log x$ and $\ln x$ to denote the logarithm to base 2 and e, respectively. The Boolean constants *false* and *true* are identified with 0 and 1, and B_n denotes the set of all Boolean functions $f : \{0,1\}^n \to \{0,1\}$. We denote the constant zero function by $\bar{0}$. Elements x of $\{0,1\}^n$ are called *assignments* and any pair (x, b) with $b \in \{0,1\}$ is called an *example*. In case $b = f(x)$ we call (x, b) an *example of* f. A *sample* S (of f) is just a set of examples (of f). For a class $C \subseteq B_n$ and a sample $S = \{(x_1, b_1), \ldots, (x_k, b_k)\}$ we use

$$C(S) = C(x_1, b_1, \ldots, x_k, b_k) = \left\{ f \in C \;\middle|\; f(x_i) = b_i \text{ for } i = 1, \ldots, k \right\}$$

to denote the class of all functions in C that are *consistent* with S. In order to formally describe the answers of a teacher we use subsets of B_n, where the answer $\Lambda \subseteq B_n$ provides the information that the target concept f belongs to Λ. For example, the answer YES to a membership query $x \in \{0,1\}^n$ corresponds to the set $B_n(x, 1) = \{f \in B_n \mid f(x) = 1\}$. Further, an equivalence query $h \in B_n$ is either answered by a counterexample $y \in \{0,1\}^n$ which is described by the set $B_n(y, 1 - h(y)) = \{f \in B_n \mid f(y) \neq h(y)\}$ or by the answer YES which corresponds to the singleton set $\{h\}$.

We say that a set of answers $A = \{\Lambda_1, \ldots, \Lambda_k\}$ is *consistent* with a function $f \in B_n$ (or f *satisfies* A), if f is contained in all answers from A. A is called *consistent* (or *satisfiable*) if A is consistent with some function $f \in B_n$. Further, for a class $C \subseteq B_n$, we use

$$C(A) = C(\Lambda_1, \ldots, \Lambda_k) = \{f \in C \mid \forall \Lambda \in A\colon\ f \in \Lambda\}$$

to denote the set of all concepts $f \in C$ that satisfy all answers in A.

Now we are ready to define the abstract notion of a learning protocol. A protocol P consists of two components: a set Q of queries and a specification of which answers are possible for each query $q \in Q$. Formally, a *protocol with query set* Q is a relation

$$P \subseteq Q \times \mathcal{P}(B_n),$$

where $\mathcal{P}(B_n)$ denotes the power set of B_n. For a function $f \in B_n$,

$$P^f = \big\{(q, \Lambda) \in P \;\big|\; f \in \Lambda\big\}$$

denotes the protocol providing only answers that are consistent with f. Further, for a query set $Q' \subseteq Q$, $P_{Q'}$ denotes the answer set

$$P_{Q'} = \bigcup_{q \in Q'} P_q, \quad \text{where } P_q = \big\{\Lambda \subseteq B_n \;\big|\; (q, \Lambda) \in P\big\}$$

contains all possible answers of P for the query q. P is called *complete*, if for any function f, the protocol P^f provides at least one answer for any query q,

$$\forall f \in B_n\ \forall q \in Q\colon\quad P_q^f \neq \emptyset,$$

where $P_q^f = P^f \cap P_q$. In this paper we will only consider learning models that can be described by complete protocols. For instance, Angluin's [1] model of learning by membership queries from $\{0,1\}^n$ is formalized by the protocol

$$MQ = \big\{\big(y, B_n(y, b)\big) \;\big|\; y \in \{0,1\}^n,\ b \in \{0,1\}\big\}$$

and the protocol for learning by equivalence queries with hypotheses from a subset $H \subseteq B_n$ is

$$EQ^H = \big\{\big(h, B_n\big(y, 1 - h(y)\big)\big) \;\big|\; h \in H,\ y \in \{0,1\}^n\big\} \cup \big\{\big(h, \{h\}\big) \;\big|\; h \in H\big\}.$$

We can combine these two protocols by taking the union

$$MEQ^H = MQ \cup EQ^H$$

which describes learning by membership and equivalence queries. Note that in general we may need to rename the queries upon taking the union of protocols in order to get disjoint query sets. Angluin also introduced a *restricted* version of equivalence queries that can only be answered by YES or NO. This model corresponds to the protocol

$$EQ_r^H = \big\{\big(h, \{h\}\big) \;\big|\; h \in H\big\} \cup \big\{\big(h, B_n - \{h\}\big) \;\big|\; h \in H\big\}.$$

Henceforth we omit the superscript H in case $H = B_n$ and simply write EQ, MEQ and EQ_r. A protocol is called *example-based* if it only admits answers of the form $B_n(S)$ for a sample S [16]. For example, all combinations of membership, equivalence, subset, and superset queries are example-based. On the other hand, some popular query types are not example-based as, e.g., restricted equivalence or UAV membership and equivalence queries.

A teacher T *answers according to a protocol* P *and a target* $f \in B_n$, if for each query $q \in Q$, T provides an answer $\Lambda \in P_q^f$. A class $C \subseteq B_n$ is *learnable with* d *queries under* P if there is an algorithm L such that for any target $f \in C$ and for any teacher T that answers according to P^f, L asks at most d queries and f is the only function in C that satisfies all answers of T. Note that there is no restriction on the *computational* complexity of L. Further, L may choose the queries adaptively, based on the answers to previous queries.

For a class $C \subseteq B_n$ and a protocol P we define the *query complexity*, $\mathrm{QC}(C, P)$, as the smallest integer $d \geqslant 0$ such that C is learnable with d queries under P. If no such integer exists then $\mathrm{QC}(C, P) = \infty$.

Since L must be successful with respect to any teacher, T can be seen as an adversary who tries to force the learner L to ask as many queries as possible. From this point of view, the choice of a target $f \in C$ restricts the teacher

to select his answers according to the protocol $P^f \subseteq P$. More generally, we can interpret any subset $T \subseteq P$ as a (more or less specific) adversary strategy.

Formally, we call a set $T \subseteq P$ an *answering scheme for* P, if T_q contains for each query $q \in Q$ exactly one answer. We call T *satisfiable* if the answer set T_Q is satisfiable. We denote the set of all answering schemes for P by $\mathcal{T}(P)$ and the set of all satisfiable answering schemes by $\mathcal{S}(P)$. Note that $\mathcal{S}(P) = \bigcup_{f \in B_n} \mathcal{T}(P^f)$. As explained above answering schemes play the role of adversary strategies in our arguments below.

In order to keep our presentation concise, we only consider concept learning of classes $C \subseteq B_n$ for some fixed arity n. However we highlight that for many query learning models our results can be extended to the case where the concept domain contains words of variable length. Castro [11] treats in detail the relationships between query learning concept classes $C \subseteq B_n$ and standard models using $\{0, 1\}^\star$ as the concept domain (see also Gavaldà [14]).

3. The general dimension for exact learning

In this section we introduce the general dimension and show that it characterizes the query complexity for *any* type of queries. A universal combinatorial parameter that exactly identifies the number of queries needed to learn can be easily defined by using a chain of alternating quantifiers of queries and answers, where the number of alternations corresponds to the query complexity. However, we would rather prefer a "flat" characterization defined by using only a small number of quantifier alternations, provided that it gives a good approximation for the query complexity. In fact, Balcázar et al. [4] introduced the abstract identification dimension AIdim which uses only one quantifier alternation, but nevertheless works well for all example-based learning models.

Definition 1. (See [4].) Let P be a protocol on a query set Q and let T be an answering scheme for P. We say that a set $A \subseteq T_Q$ of answers from T *succeeds* on a concept class $C \subseteq B_n$ ($A \in Succ(C, T)$ for short), if $\|C(A)\| \leqslant 1$.

The *abstract identification dimension* of C under P, $\mathrm{AIdim}(C, P)$, is the minimum integer k (if it exists) such that any satisfiable answering scheme T provides a set A of k answers that succeeds on C,

$$\mathrm{AIdim}(C, P) = \max_{T \in \mathcal{S}(P)} \min_{A \in Succ(C,T)} \|A\|.$$

(Here and in the following we adopt the convention that $\min_{A \in \emptyset} \|A\| = \infty$.)

The notion of AIdim corresponds to the scenario where the adversary T has to reveal his answer for any potential query before the learner actually selects an appropriate query set. But note that also T has an advantage since unlike a teacher, T only needs to be consistent with some function (not necessarily with some $f \in C$). We illustrate the notion by some examples (see also Fig. 1).

Example 2. We call $f \in B_n$ a singleton function ($f \in \mathrm{SING}_n$ for short) if $f(x) = 1$ holds for exactly one $x \in \{0, 1\}^n$. Clearly, if an answering scheme T for MQ answers some query by YES, this single answer succeeds on SING_n. On the other hand, if T gives only NO answers (i.e., $T = MQ^{\bar{0}}$), then any answer set $A \in Succ(\mathrm{SING}_n, T)$ must contain at least $2^n - 1$ answers. Hence, $\mathrm{AIdim}(\mathrm{SING}_n, MQ) = 2^n - 1$ (which in fact coincides with $\mathrm{QC}(\mathrm{SING}_n, MQ)$).

Next we consider the problem of learning singletons by equivalence queries. Since any answer to the hypothesis $h = \bar{0}$ succeeds on SING_n, it follows for $P \in \{EQ^H, MEQ^H \mid \bar{0} \in H\}$ that $\mathrm{AIdim}(\mathrm{SING}_n, P) = \mathrm{QC}(\mathrm{SING}_n, P) = 1$. On the other hand, if $\bar{0} \notin H$, then we again get $\mathrm{AIdim}(\mathrm{SING}_n, P) = \mathrm{QC}(\mathrm{SING}_n, P) = 2^n - 1$ for these protocols, provided that $\mathrm{SING}_n \subseteq H$ in case $P = EQ^H$.

As shown in [4], AIdim is a crucial notion in the sense that it unifies all known learning dimensions of example-based query models. Moreover, the application of this concept to specific protocols such as equivalence and/or membership queries exactly yields the combinatorial notions that are known to characterize the query complexity in these models, such as strong consistency dimension [5], extended teaching dimension [18] and consistency dimension (or certificate size) [5,19]. AIdim thus fully unifies all these characterizations.

Now we are ready to introduce the general dimension where in contrast to AIdim, the adversary T does not need to be consistent with any function. Surprisingly, this harder requirement for the learner turns out to be an adequate compensation for the learner's advantage of knowing all answers in advance, even in non-example-based models.

Example	Class C	Protocol P	AIdim(C, P)	Gdim(C, P)	QC(C, P)
2, 4	SING$_n$	$EQ^H, MEQ^H, \bar{0} \in H$	1	1	1
2, 4	SING$_n$	$MQ, MEQ^H, \bar{0} \notin H$	$2^n - 1$	$2^n - 1$	$2^n - 1$
4	B_n	EQ	1	2	2^n
6	$C \subseteq B_n$	EQ_r	1	$\|C\| - 1$	$\|C\| - 1$
9	SING$_n$	MQ_{uav}, MEQ^H_{uav}	1	$\min(n, 3)$	n
10	SING$_n$	EQ_{uav}	1	$2^n - 1$	$2^n - 1$

Fig. 1. Values of the dimensions AIdim, Gdim and the query complexity for some concept classes C and protocols P considered in this article.

Definition 3. The *general dimension* of a concept class C under a protocol P, Gdim(C, P), is the minimum integer k (if it exists) such that any answering scheme T provides a set A of k answers that succeeds on C,

$$\text{Gdim}(C, P) = \max_{T \in \mathcal{T}(P)} \min_{A \in Succ(C,T)} \|A\|.$$

Although Gdim appears to be harder to estimate than other dimensions (as the maximum operator ranges also over inconsistent adversary strategies), often simpler incarnations of Gdim can be derived for specific protocols. Section 4 provides two such examples for non-example-based protocols. Further, Theorem 5 below shows that AIdim can be interpreted as a simplified version of Gdim for example-based query models. We first consider some examples.

Example 4. It is easy to see that Gdim coincides with AIdim on the class SING$_n$ for the protocols MQ, EQ^H and MEQ^H. Next we consider the problem of learning the class B_n with equivalence queries. Clearly, if an answering scheme T for EQ is consistent with some function $h \in B_n$, then the YES answer to the query h succeeds on B_n, implying that AIdim$(B_n, EQ) = 1$. Otherwise $B_n(T_Q) = \emptyset$ and since EQ is example-based, each answer $\Lambda \in T_Q$ is of the form $B_n(S)$ for some sample S. Thus T_Q must contain two contradictory answers, implying that Gdim$(B_n, EQ) = 2$. Recall that by applying the halving algorithm [1,24] it is easy to see that QC$(B_n, EQ) = 2^n$.

Since in an example-based query model, any inconsistent adversary strategy can be unmasked by just two queries, AIdim and Gdim essentially coincide for these models (see Theorem 5). In contrast, exponentially many (in $\log \|C\|$) non-example-based queries may be necessary for this task (see Example 6 below). Hence, it is not possible to define a useful dimension for such models by considering only satisfiable answering schemes (as AIdim).

Theorem 5. *For any example-based protocol P and any class $C \subseteq B_n$ it holds that*

$$\text{AIdim}(C, P) \leqslant \text{Gdim}(C, P) \leqslant \max\big(2, \text{AIdim}(C, P)\big).$$

Proof. The first inequality immediately follows by definition. For the second inequality assume that AIdim$(C, P) = k$ and let T be any answering scheme for P. If T is satisfiable, then the assumption AIdim$(C, P) = k$ guarantees that there is an appropriate answer set. Otherwise $B_n(T_Q) = \emptyset$ and since P is example-based, T_Q must contain two contradictory answers. □

Theorem 5 implies that on example-based protocols P, the general and the abstract identification dimension can only differ when AIdim$(C, P) = 1$ and Gdim$(C, P) = 2$. As shown in Example 4, this happens for the protocol EQ and the concept class B_n. The next example shows that AIdim is unable to provide a good estimation for the number of restricted equivalence queries.

Example 6. Let T be any answering scheme for the protocol EQ_r. First observe that for any class $C \subseteq B_n$ with $\|C\| \geqslant 2$, AIdim$(C, EQ_r) = 1$, since T_Q must contain the answer $\{h\}$ in case T is satisfied by some function $h \in B_n$. On the other hand, it is easy to see that Gdim$(C, EQ_r) =$ QC$(C, EQ_r) = \|C\| - 1$, since each negative answer eliminates just one concept from C.

Our next goal is to show that in contrast to AIdim, Gdim provides a useful approximation of the query complexity for all query types. For the upper bound we prove in the following lemma that at any point in the learning process,

there is a smart query $q \in Q$. More precisely, q has the property that any answer in P_q considerably shrinks the actual set D of target candidates, where the shrinking factor depends on the value of the general dimension. A similar result has been shown in [4, Lemma 4] for the abstract identification dimension.

Lemma 7. *Let $C \subseteq B_n$ and let P be a protocol with* $\mathrm{Gdim}(C, P) = k \geqslant 1$. *Then for any class $D \subseteq C$ there is a query $q \in Q$ such that for all answers $\Lambda \in P_q$, at least $(\|D\| - 1)/k$ functions from D are inconsistent with Λ.*

Proof. Let $\|D\| = d$. The result is trivial when $d \leqslant 1$. Otherwise, suppose that for each $q \in Q$ there is some $\Lambda_q \in P_q$ such that fewer than $(d-1)/k$ functions are inconsistent with Λ_q. Consider the answering scheme $T = \{(q, \Lambda_q) \mid q \in Q\}$. Then for any $A \subseteq T_Q$ with $\|A\| \leqslant k$, D contains fewer than $d-1$ functions that are inconsistent with A, implying that $D(A)$ (a subset of $C(A)$) contains at least two functions. But this contradicts $\mathrm{Gdim}(C, P) = k$. □

We note that Dasgupta et al. [12] have independently shown a similar result focusing on some type of membership queries. Now we are ready to prove the main result of this section.

Theorem 8. *For any protocol P and any class C it holds that*

$$\mathrm{Gdim}(C, P) \leqslant \mathrm{QC}(C, P) \leqslant \lceil \ln \|C\| \rceil \, \mathrm{Gdim}(C, P).$$

Proof. We first show that if $\mathrm{Gdim}(C, P) > k$ then any learning algorithm must ask more than k queries. If $\mathrm{Gdim}(C, P) > k$ then there is an answering scheme T with the property that $\|C(A)\| > 1$ for any set $A \subseteq T_Q$ that contains at most k answers. Now we can use T to answer the queries of an arbitrary algorithm L in such a way that after k interactions, at least two functions in C are consistent with all answers given to L. This implies that L cannot learn C with k queries.

To show the upper bound let $\mathrm{Gdim}(C, P) = k$. If $k \leqslant 1$ it is easy to see (by using Lemma 7 when $k = 1$) that $\mathrm{QC}(C, P) = k$. Otherwise, consider the learning algorithm L that starting with the empty answer set $A = \emptyset$, in round i asks the query q_i provided by Lemma 7 for the class $C(A)$ and includes the answer Λ_i into A. L stops as soon as $c_i = \|C(\Lambda_1, \ldots, \Lambda_i)\|$ becomes smaller than 2. Now it follows that $c_0 = \|C\|$ and $c_{i+1} \leqslant c_i(1 - 1/k) + 1/k$ which in turn implies that

$$c_i \leqslant c_0(1 - 1/k)^i + (1/k) \sum_{j=0}^{i-1} (1 - 1/k)^j.$$

Since the second term evaluates to $1 - (1 - 1/k)^i < 1$, we can use the inequality $1 - x \leqslant e^{-x}$ to conclude that $c_i < 2$ for $i = \lceil k \ln \|C\| \rceil$. □

In Example 4 we have seen that $\mathrm{Gdim}(\mathrm{SING}_n, MQ) = \mathrm{QC}(\mathrm{SING}_n, MQ) = 2^n - 1$, implying that the lower bound for the query complexity provided by Theorem 8 in terms of the general dimension is sharp. Further, since $\mathrm{QC}(B_n, EQ) = 2^n$ but $\mathrm{Gdim}(B_n, EQ) = 2$ (see Example 4), it follows that $\mathrm{QC}(B_n, EQ) = \alpha \, \mathrm{Gdim}(B_n, EQ)$ for $\alpha = (\log \|B_n\|)/2$. This shows that also the upper bound of Theorem 8 is essentially optimal.

4. Unspecified attribute value queries

In this section we apply the general dimension to different types of unspecified attribute value queries (UAV queries for short). We first introduce some additional notation. A *partial assignment* α is a word from $\{0, 1, \star\}^n$. An assignment $x \in \{0, 1\}^n$ *satisfies* α if α and x coincide in all non-star positions of α. We use t_α to denote the Boolean function with $t_\alpha(x) = 1$ if and only if x satisfies α. Functions of this kind are called *terms*. We use Term_n to denote the class of all terms in B_n and X_α to denote the hypercube $\{x \in \{0, 1\}^n \mid t_\alpha(x) = 1\}$.

Following Goldman et al. [17], we extend each function $h \in B_n$ to a ternary function $\hat{h} : \{0, 1, \star\}^n \to \{0, 1, ?\}$ as follows: If for some $b \in \{0, 1\}$, $h(x) = b$ for all $x \in X_\alpha$, then $\hat{h}(\alpha) = b$. Otherwise, $\hat{h}(\alpha) = ?$. We say that h is *single-valued on* α, if $\hat{h}(\alpha) \in \{0, 1\}$. Pairs $(\alpha, a) \in \{0, 1, \star\}^n \times \{0, 1, ?\}$ are called *UAV examples*. In case $a \in \{0, 1\}$ we call

(α, a) *single-valued* and if $\hat{h}(\alpha) = a$ we call (α, a) a *UAV example of* h. We use SV^h to denote the set of all single-valued UAV examples of h. Further, for a class $C \subseteq B_n$ and a *sample* (i.e., set) $S = \{(\alpha_1, a_1), \ldots, (\alpha_k, a_k)\}$ of UAV examples we use

$$C(S) = C(\alpha_1, a_1, \ldots, \alpha_k, a_k) = \left\{h \in C \mid \hat{h}(\alpha_i) = a_i \text{ for } i = 1, \ldots, k\right\}$$

to denote the class of all functions in C that are *consistent* with S.

For a target $f \in B_n$, the UAV membership query $\alpha \in \{0, 1, \star\}^n$ returns YES if $\hat{f}(\alpha) = 1$, NO if $\hat{f}(\alpha) = 0$, and ? otherwise. This leads to the protocol

$$MQ_{uav} = \left\{\left(\alpha, B_n(\alpha, a)\right) \mid \alpha \in \{0, 1, \star\}^n,\ a \in \{0, 1, ?\}\right\}.$$

A UAV equivalence query $h \in B_n$ is answered either by a UAV example (α, a) (meaning that $\hat{h}(\alpha) \neq \hat{f}(\alpha) = a$) or by YES (meaning that $h = f$). For a hypothesis space $H \subseteq B_n$, this leads to the protocol

$$EQ^H_{uav} = \left\{\left(h, \{h\}\right), \left(h, B_n(\alpha, a)\right) \mid h \in H,\ \alpha \in \{0, 1, \star\}^n,\ a \neq \hat{h}(\alpha)\right\}.$$

Further, MEQ^H_{uav} denotes the protocol $MQ_{uav} \cup EQ^H_{uav}$ where we omit H in case $H = B_n$. Note that UAV membership and UAV equivalence queries admit answers of the form $B_n(\alpha, ?)$ which are not example-based unless α is star-free. In the following example we consider the problem of learning singletons with UAV membership queries.

Example 9. Let T be any answering scheme for MQ_{uav}. If T is consistent with some function $h \in B_n - \{\bar{0}\}$, then the answer to any query $x \in \{0, 1\}^n$ with $h(x) = 1$ succeeds on SING_n. Further, if T is consistent with the null function $\bar{0}$, then the answer NO to the query $\alpha = \star^n$ discards all singletons. This shows that $\mathrm{AIdim}(\mathrm{SING}_n, MQ_{uav}) = 1$.

Next we argue that $\mathrm{Gdim}(\mathrm{SING}_n, MQ_{uav}) \leqslant 3$. Clearly, if T answers some query $\alpha \in \{0, 1\}^n$ with $B_n(\alpha, a)$, $a \neq 0$, or if T answers the query $\alpha = \star^n$ with NO, then a single answer of T succeeds on SING_n. Otherwise there is some query $\alpha \in \{0, 1, \star\}^n - \{0, 1\}^n$ which is answered by $a \neq 0$ but all queries β for which X_β is properly contained in X_α get the answer NO. But then we have $\mathrm{SING}_n(\alpha, a, \alpha_0, 0, \alpha_1, 0) = \emptyset$, where α_b, $b \in \{0, 1\}$, is obtained from α by replacing the first star in α by b. This shows that $\mathrm{Gdim}(\mathrm{SING}_n, MQ_{uav}) \leqslant 3$. On the other hand, the answering scheme

$$T = \left\{(\alpha, 0) \mid \alpha \in \{0, 1\}^n\right\} \cup \left\{(\alpha, ?) \mid \alpha \in \{0, 1, \star\}^n - \{0, 1\}^n\right\}$$

witnesses $\mathrm{Gdim}(\mathrm{SING}_n, MQ_{uav}) \geqslant \min(n, 2)$. (This lower bound is sharp for $n \leqslant 2$. By using a similar answering scheme for the case $n \geqslant 3$ it can be shown that $\mathrm{Gdim}(\mathrm{SING}_n, MQ_{uav}) = \min(n, 3)$.) Further, it is not hard to see that the query complexity of SING_n in the MQ_{uav} model (as well as in the MEQ_{uav} model) is n, implying an exponential gap compared to the MQ model (see Example 2).

The next example shows that AIdim fails to give a useful characterization of the query complexity in the EQ_{uav} model.

Example 10. First observe that $\mathrm{AIdim}(\mathrm{SING}_n, EQ_{uav}) = 1$, since any consistent answering scheme T for EQ_{uav} provides a YES answer $\{h\}$. However, each answer from the inconsistent answering scheme

$$T = \left\{\left(h, B_n(x, 0)\right) \mid h \in B_n - \{\bar{0}\},\ h(x) = 1\right\} \cup \left\{\left(\bar{0}, B_n\left(\star^n, ?\right)\right)\right\}$$

discards at most one function in SING_n. Hence, $\mathrm{Gdim}(\mathrm{SING}_n, EQ_{uav}) = 2^n - 1$ (which coincides with $\mathrm{QC}(\mathrm{SING}_n, EQ_{uav})$).

Several relationships between UAV and ordinary query types are known [7,17]. Although UAV memberships are more powerful than standard memberships (see Example 9), UAV equivalences are strictly weaker than standard ones (see Examples 2 and 10) since, intuitively speaking, in the UAV setting the teacher has more freedom in the choice of counterexamples.

On the other hand, the combination of UAV memberships and UAV equivalences turns out to be again more powerful than its ordinary counterpart [7,17]. In fact, any $MQ_{uav} \cup EQ^H$ learning algorithm using UAV memberships and ordinary equivalences can be simulated by a MEQ^H_{uav} learning algorithm which asks at most n times as many queries. Indeed, in [17] it is explained how a counterexample $(\alpha, ?)$ provided by a UAV teacher can be transformed

into an ordinary counterexample by asking at most n appropriately chosen UAV membership queries. This shows that the protocol MEQ^H_{uav} is at least as strong as MEQ^H. In addition, observe that the class SING_n is learnable with n UAV membership queries (see Example 9), whereas learning SING_n in the MEQ^H model requires an exponential number of queries (provided that the null function is not contained in H, see Example 2). Hence, the MEQ^H_{uav} model is properly stronger than the MEQ^H model.

The observation in the last paragraph raises the question whether also for other interesting concept classes, UAV memberships and equivalences are superior to their ordinary counterparts. In the next section, we give a negative answer for the case of DNF formulas. The proof of this result makes use of a UAV version of the consistency dimension which we introduce later in this section. We first consider the MQ_{uav} model and derive from Gdim a UAV counterpart of the extended teaching dimension for ordinary membership queries [18]. The idea is just to use single-valued UAV examples instead of ordinary examples.

Definition 11. The *UAV extended teaching dimension of* C, $\mathrm{ETdim}_{uav}(C)$, is the smallest integer $k \geqslant 0$ such that for any function g there are k single-valued UAV examples of g which are satisfied by at most one function in C,

$$\forall g \in B_n\ \exists S \subseteq SV^g\colon \quad \|S\| \leqslant k \wedge \big\|C(S)\big\| \leqslant 1.$$

Note that if we require in Definition 11 that for any g there are k examples of g which are satisfied by at most one function in C, then we get a formal definition of the extended teaching dimension $\mathrm{ETdim}(C)$ of C.

Example 12. Since SV^g contains the UAV example $(\star^n, 0)$ in case g is the null function, and a UAV example of the form $(x, 1)$ otherwise, it follows that $\mathrm{ETdim}_{uav}(\mathrm{SING}_n) = 1$.

The following theorem shows that the UAV extended teaching dimension is very close to the general dimension and hence can serve as a dimension for UAV membership learning (see Corollary 14).

Theorem 13. *For any class $C \subseteq B_n$ it holds that*

$$\mathrm{ETdim}_{uav}(C)/2 \leqslant \mathrm{Gdim}(C, MQ_{uav}) \leqslant \max\big(3, \mathrm{ETdim}_{uav}(C)\big).$$

Proof. Let $\mathrm{ETdim}_{uav}(C) = k$. We first show that $\mathrm{Gdim}(C, MQ_{uav}) \leqslant \max(3, k)$. Let T be any answering scheme for MQ_{uav} and let $Q = \{0, 1, \star\}^n$ be the query set of the protocol MQ_{uav}. We have to show that there is a set $A \subseteq T_Q$ of at most $\max(3, k)$ answers with $\|C(A)\| \leqslant 1$. If T_Q is satisfied by some function g, then we can use the fact that $\mathrm{ETdim}_{uav}(C) = k$ to get a sample $S \subseteq SV^g$ of at most k UAV examples with $\|C(S)\| \leqslant 1$, implying that $A = \{B_n(\alpha, a) \mid (\alpha, a) \in S\}$ has the required properties.

Next we show that any unsatisfiable answering scheme T provides an unsatisfiable set of at most three answers. This is certainly true if T replies to some query $x \in \{0, 1\}^n$ with the inconsistent answer $B_n(x, ?)$. Hence we can assume that T's answers to all queries $x \in \{0, 1\}$ are consistent with some function $h \in B_n$. Now, since T is inconsistent, there is a query α which gets an answer $a \neq \hat{h}(\alpha)$ but all queries β for which X_β is properly contained in X_α get the answer $\hat{h}(\beta)$. The following case analysis shows that T gives two or three contradictory answers.

- If $a \in \{0, 1\}$, then we can find an $x \in X_\alpha$ with $h(x) \neq a$ implying that the two answers of T for α and x are unsatisfiable.
- If $a = ?$, then the three answers of T for α, α_0 and α_1 are unsatisfiable, where α_b is obtained from α by replacing the first star in α by b.

It remains to show that $\mathrm{ETdim}_{uav}(C) \leqslant 2k$ for $k = \mathrm{Gdim}(C, MQ_{uav})$. Let $g \in B_n$. In order to find a sample $S \subseteq SV^g$ of at most $2k$ single-valued UAV examples with $\|C(S)\| \leqslant 1$, consider the answering scheme $T = MQ^g_{uav}$ which answers each query α with $\hat{g}(\alpha)$. By the fact that $\mathrm{Gdim}(C, MQ_{uav}) = k$ there is a set $A \subseteq T_Q$ of at most k answers with $\|C(A)\| \leqslant 1$, implying that $\|C(S')\| \leqslant 1$ for the UAV sample $S' = \{(\alpha, a) \mid B_n(\alpha, a) \in A\}$. Since A is consistent with g, S' is a UAV sample of g. Now for each UAV example $(\alpha, ?)$ in S' we can find two examples (x, a) and (y, b) in SV^g with $x, y \in X_\alpha$ and $a \neq b$, implying that $C(x, a, y, b) \subseteq C(\alpha, ?)$. Hence, we can obtain S by replacing each UAV example $(\alpha, ?)$ in S' by two (single-valued) examples. □

By combining Theorems 8 and 13 it now easily follows that the UAV extended dimension is useful for estimating the number of membership queries in the UAV setting.

Corollary 14. *For any class $C \subseteq B_n$,*

$$\mathrm{ETdim}_{uav}(C)/2 \leqslant \mathrm{QC}(C, MQ_{uav}) \leqslant \lceil \ln\|C\| \rceil \big(\mathrm{ETdim}_{uav}(C) + 2\big).$$

Now we introduce the UAV counterpart of the consistency dimension which we will use in the next section to show that learning DNF formulas in the UAV setting requires approximately the same number of membership and equivalence queries as in the ordinary setting.

Definition 15. Let $H \subseteq B_n$ be a hypothesis class and let $C \subseteq H$ be a concept class. The *UAV consistency dimension* of C and H, $\mathrm{Cdim}_{uav}(C, H)$, is the smallest integer $k \geqslant 0$ such that for any function $g \notin H$ there are k single-valued UAV examples of g which are inconsistent with any function in C,

$$\forall g \in B_n - H\ \exists S \subseteq SV^g\colon \quad \|S\| \leqslant k \wedge C(S) = \emptyset.$$

Again note that if we require in Definition 15 that for any $g \notin H$ there are k examples of g which are inconsistent with any function in C, then we get a formal definition of the consistency dimension $\mathrm{Cdim}(C, H)$ of C and H. The next theorem shows that the UAV consistency dimension provides a useful measure in the MEQ^H_{uav} model (see Corollary 17).

Theorem 16. *For all classes $C \subseteq H \subseteq B_n$ it holds that*

$$\big(\mathrm{Cdim}_{uav}(C, H) - 1\big)/2 \leqslant \mathrm{Gdim}\big(C, MEQ^H_{uav}\big) \leqslant \max\big(3, \mathrm{Cdim}_{uav}(C, H)\big).$$

Proof. Let $k = \mathrm{Cdim}_{uav}(C, H)$. We first show that $\mathrm{Gdim}(C, MEQ^H_{uav}) \leqslant \max(3, k)$. Let T be any answering scheme for MEQ^H_{uav}. We have to show that there is a set $A \subseteq T_Q$ of at most $\max(3, k)$ answers with $\|C(A)\| \leqslant 1$. If T is satisfied by some function $g \in H$, then the YES answer $\{g\}$ succeeds on C. Also, if T is satisfied by some function $g \notin H$, then we get an answer set A from a suitable UAV sample $S \subseteq SV^g$ as in the proof of Theorem 13.

Next we show that if T is unsatisfiable, then T_Q contains an unsatisfiable subset of at most three answers. If the set $M = \bigcup_{q \in \{0,1,\star\}^n} T_q$ of T's answers to all membership queries is already inconsistent we argue exactly as in the proof of Theorem 13. Otherwise, M is consistent with some function $g \in B_n$. As T is unsatisfiable, some equivalence query $h \in H$ gets an answer Λ that is inconsistent with g. In case Λ provides a counterexample (α, a), Λ is inconsistent with T's answer $\hat{g}(\alpha)$ for the membership query α. Similarly, if the answer to the equivalence query h is YES, then h must be different from g, implying that Λ is again inconsistent with T's answer $g(x)$ for some membership query x.

Now let $k = \mathrm{Gdim}(C, MEQ^H_{uav})$. It remains to show that $\mathrm{Cdim}_{uav}(C) \leqslant 2k + 1$. Let $g \in B_n - H$. In order to find a sample $S' \subseteq SV^g$ of at most $2k + 1$ single-valued UAV examples with $C(S') = \emptyset$, let T be any answering scheme for MEQ^H_{uav} consistent with g. Since $\mathrm{Gdim}(C, MEQ^H_{uav}) = k$, T_Q contains an answer set A of size k that succeeds on C. Since $g \notin H$, all answers in T_Q are of the form $B_n(\alpha, a)$ and hence, exactly as in the proof of Theorem 13 we obtain a sample $S \subseteq SV^g$ of size at most $2k$ with $\|C(S)\| \leqslant 1$. Since $g \notin C$, it suffices to add one more SV^g example to S to get the desired sample S'. □

Theorems 8 and 16 together imply that the UAV consistency dimension provides a good estimation of the number of membership and equivalence queries needed to learn a concept class in the UAV setting.

Corollary 17. *For all classes $C \subseteq H \subseteq B_n$ it holds that*

$$\big(\mathrm{Cdim}_{uav}(C, H) - 1\big)/2 \leqslant \mathrm{QC}\big(C, MEQ^H_{uav}\big) \leqslant \lceil \ln\|C\| \rceil \big(\mathrm{Cdim}_{uav}(C, H) + 3\big).$$

5. Learning DNF formulas with UAV queries

The learnability of DNF formulas with polynomially many membership and equivalence queries is an important open problem for a variety of hypothesis classes. As explained in the last section it is conceivable that this class is

easier to learn in the UAV setting. However, in Theorem 19 below we show that the consistency and UAV consistency dimensions for this class are close, implying that up to a polynomial factor, the query complexity of DNF formulas is the same in both models (see Corollary 20). Of course, this does not exclude the possibility that UAVs are still more efficient in terms of *computational* complexity measures.

We call $h \in B_n$ an m-term DNF, if h can be expressed as a disjunction $t_1 \vee \cdots \vee t_m$ of m terms. The class of all m-term DNFs is denoted by $m\text{-DNF}_n$. A function $h \in B_n$ *ε-satisfies* a sequence $\sigma = (x_1, \ldots, x_s)$ of s (not necessarily pairwise distinct) assignments $x_i \in \{0,1\}^n$, if $\|\{i \mid h(x_i) = 1\}\| \geqslant \varepsilon s$, i.e., $h(x_i) = 1$ holds for at least an ε-fraction of all strings in the sequence σ. The following lemma is useful for proving Theorem 19.

Lemma 18. *For any assignment $x_0 \in \{0,1\}^n$ and any partial assignment $\alpha \in \{0,1,\star\}^n$ there is an assignment $x_0^\alpha \in X_\alpha$ such that $\text{Term}_n(x_0, 1, x_0^\alpha, 0) \subseteq \text{Term}_n(\alpha, 0)$, i.e., any term t with $t(x_0) = 1$ and $t(x_0^\alpha) = 0$ fulfills $\hat{t}(\alpha) = 0$.*

Proof. For $u, v \in \{0,1\}^n$, let $u \oplus v$ denote the bitwise XOR of u and v, and for a subset $X \subseteq \{0,1\}^n$ let $X \oplus v$ denote the set $\{u \oplus v \mid u \in X\}$. Now let y_0 be the assignment in the hypercube $X_\alpha \oplus x_0 \oplus 1^n$ with the maximum number of ones and define x_0^α as $y_0 \oplus x_0 \oplus 1^n$. Note that x_0^α belongs to X_α.

We have to show that any term t with $t(x_0) = 1$ and $t(x_0^\alpha) = 0$ fulfills $\hat{t}(\alpha) = 0$. In order to derive a contradiction, assume that $t(u) = 1$ for some $u \in X_\alpha$ and consider the term t' defined by $t'(x) = t(x \oplus x_0 \oplus 1^n)$. Since $t'(1^n) = t(x_0) = 1$, t' is monotone. Further, since $t'(u \oplus x_0 \oplus 1^n) = t(u) = 1$ and since y_0 is the maximum assignment in the hypercube $X_\alpha \oplus x_0 \oplus 1^n$, it follows that also $t'(y_0) = 1$, implying that $t(x_0^\alpha) = t(y_0 \oplus x_0 \oplus 1^n) = t'(y_0) = 1$. □

Now we are ready to show that the class of DNF formulas has close consistency and UAV consistency dimensions (see Definition 15 and the subsequent paragraph for formal definitions of these notions).

Theorem 19. *Let $4 \leqslant l < \sqrt{m}/n$. Then it holds for any hypothesis class H with $m\text{-DNF}_n \subseteq H \subseteq B_n$ that*

$$\text{Cdim}(l\text{-DNF}_n, H) \leqslant \left(l^2 n/3\right) \text{Cdim}_{uav}(m\text{-DNF}_n, H).$$

Proof. Let $\text{Cdim}_{uav}(m\text{-DNF}_n, H) = k$ and let $g \in B_n - H$. Then there is a sample $S \subseteq SV^g$ of at most k single-valued UAV examples (α, a) of g with $m\text{-DNF}_n(S) = \emptyset$. Letting $S^0 = S \cap (\{0,1,\star\}^n \times \{0\})$, $Y = \bigcup_{(\alpha,1)\in S} X_\alpha$ and $N = \bigcup_{(\alpha,0)\in S} X_\alpha$ we get a sample $S' = (Y \times \{1\}) \cup (N \times \{0\})$ of g with $m\text{-DNF}_n(S') = \emptyset$. To prove the theorem we show that there is a sample $S'' \subseteq S'$ of size at most $kl^2 n/3$ with $l\text{-DNF}_n(S'') = \emptyset$.

Claim 1. There is a sequence σ of assignments from Y such that no term $t \in \text{Term}_n(S^0)$ n/m-satisfies σ.

Proof of Claim 1. Let $\varepsilon = n/m$ and in order to derive a contradiction assume that any sequence σ of assignments from Y is ε-satisfied by some term $t_\sigma \in \text{Term}_n(S^0)$. Starting with the sequence σ_0 of all assignments from Y and $k = 0$ let $t_k = t_{\sigma_k}$ and let σ_{k+1} be the subsequence of σ_k containing all assignments x with $t_k(x) = 0$. Then the functions $t_0 \vee \cdots \vee t_k$, $k \geqslant 0$, ε_k-satisfy σ_0, where $\varepsilon_k = \varepsilon \sum_{j=0}^{k} (1-\varepsilon)^j$. Since $\varepsilon_{m-1} = 1 - (1-\varepsilon)^m > 1 - 2^{-n}$, it follows that the m-term DNF $t_0 \vee \cdots \vee t_{m-1}$ is consistent with the sample $S' = (Y \times \{1\}) \cup (N \times \{0\})$. This completes the proof of Claim 1. □

Since by assumption, $n/m < 1/(l^2 n)$, it follows that no term $t \in \text{Term}_n(S^0)$ $1/(l^2 n)$-satisfies σ.

Claim 2. There is a sequence $\tau = (x_1, \ldots, x_s)$ of $s = \lfloor l^2 n/3 \rfloor$ assignments $x_i \in Y$ such that no term $t \in \text{Term}_n(S^0)$ $1/l$-satisfies τ.

Proof of Claim 2. For any term $t \in \text{Term}_n(S^0)$ let p_t be the probability that t $1/l$-satisfies a sequence τ of s randomly chosen assignments from σ. Since t does not $1/(l^2 n)$-satisfy σ, it follows that

$$p_t \leqslant \binom{s}{\lceil s/l \rceil} \left(1/l^2 n\right)^{\lceil s/l \rceil} \leqslant s^{\lceil s/l \rceil} (1/3s)^{\lceil s/l \rceil} < 1/3^n.$$

As $\|\text{Term}_n\| = 3^n$ it follows that the probability that some term $t \in \text{Term}_n(S^0)$ $1/l$-satisfies τ is smaller than 1. This completes the proof of Claim 2. □

Now Lemma 18 guarantees that for each assignment x_i from τ and each UAV example $(\alpha, 0)$ from S^0 there is an assignment $x_i^\alpha \in X_\alpha$ such that $\mathrm{Term}_n(x_i, 1, x_i^\alpha, 0) \subseteq \mathrm{Term}_n(\alpha, 0)$. We argue that the sample

$$S'' = \{(x_1, 1), \ldots, (x_s, 1)\} \cup \{(x_1^\alpha, 0), \ldots, (x_s^\alpha, 0) \mid (\alpha, 0) \in S^0\}$$

has the desired properties. In fact, assume that some l-term DNF $h = t_1 \vee \cdots \vee t_l$ satisfies S''. Then some term t_j $1/l$-satisfies τ. Hence, $t_j(x_i) = 1$ holds for some assignment x_i from τ, and since $t_j(x_i^\alpha) = 0$ holds for all $(\alpha, 0) \in S^0$, Lemma 18 implies that t_j is consistent with S^0. But this contradicts Claim 2. □

Theorem 19 has the following consequence for the query complexity of DNF formulas.

Corollary 20. *Let $4 \leqslant l < \sqrt{m}/n$. Then it holds for any hypothesis class H with m-$\mathrm{DNF}_n \subseteq H \subseteq B_n$ that*

$$\mathrm{QC}(l\text{-}\mathrm{DNF}_n, MEQ^H) = O(l^3 n^2)\,\mathrm{QC}(m\text{-}\mathrm{DNF}_n, MEQ^H_{uav}).$$

Proof. The result trivially holds for $H = B_n$. Otherwise, the values of both $\mathrm{Cdim}(l\text{-}\mathrm{DNF}_n, H)$ and $\mathrm{Cdim}_{uav}(m\text{-}\mathrm{DNF}_n, H)$ are non-zero. Hence, since $\mathrm{QC}(l\text{-}\mathrm{DNF}_n, MEQ^H) \leqslant \mathrm{Cdim}(l\text{-}\mathrm{DNF}_n, H) \log \|l\text{-}\mathrm{DNF}_n\|$ (see, e.g., [5]), the result follows by applying Theorem 19 and Corollary 17. □

Corollary 20 implies that a positive result for the query complexity of DNF formulas in the MEQ^H_{uav} model, say polynomial in the number of variables and terms, would imply that DNF formulas are also learnable with polynomially many ordinary membership and equivalence queries.

6. A general dimension for approximate learning

In this section we extend the general dimension to the approximate learning setting, where the goal of the learning algorithm consists in finding a close approximation to the target f. We use a generalization of the learning by distances (*LBD*) model of Ben-David et al. [6]. In this model, concepts are considered as points in a metric space (M, d), where $d : M \times M \to \mathbb{R}$ is an arbitrary *pseudo-metric* on M. This means that for all $f, g, h \in M$, $d(f, g) \geqslant 0$, with equality if $f = g$, $d(f, g) = d(g, f)$, and $d(f, h) \leqslant d(f, g) + d(g, h)$. In this paper we restrict M to be the set B_n of Boolean functions. Note that we do not require that $d(f, g) = 0$ implies $f = g$. Hence, any function of the form $d(f, g) = \Pr_D[f(x) \neq g(x)]$, where x is chosen randomly according to some arbitrary distribution D on $\{0, 1\}^n$, defines a pseudo-metric on B_n. In the following, δ denotes the metric induced by the uniform distribution U.

Let $\varepsilon \geqslant 0$. A subset $B \subseteq B_n$ is called an *ε-ball*, if there is a function $h \in B_n$ (called the *center* of B) such that $B = B_{\varepsilon,d}(h)$, where $B_{\varepsilon,d}(h)$ consists of all concepts $f \in B_n$ with $d(f, h) \leqslant \varepsilon$. The query complexity of a class $C \subseteq B_n$ under a protocol P is extended to the approximate learning setting as follows.

C is *ε-learnable* with k queries under P and d, if there is an algorithm L such that for any $f \in C$ and for any teacher that answers according to P^f, the set of functions in C that satisfy all answers received after at most k interactions is contained in some ε-ball. The *query complexity* of C under P, d and ε, denoted by $\mathrm{QC}_{\varepsilon,d}(C, P)$, is the smallest integer $k \geqslant 0$ such that C is ε-learnable with k queries under P and d. If no such integer k exists, then $\mathrm{QC}_{\varepsilon,d}(C, P) = \infty$.

Let L be an algorithm that ε-learns a class C and let A be the set of answers given by the teacher during some run of L with respect to some target $f \in C$. Since f belongs to the set $C(A)$ which is covered by some ε-ball $B_{\varepsilon,d}(h)$, L "knows" some concept $h \in B_n$ with error $d(f, h) \leqslant \varepsilon$. On the other hand, if a learning algorithm outputs for every target $f \in C$ some $h \in B_n$ with error $d(f, h) \leqslant \varepsilon$, then the set $C(A)$ of target candidates has to be contained in $B_{\varepsilon,d}(h)$, since otherwise, L would not succeed on any target $g \in C(A)$ with $d(g, h) > \varepsilon$. Thus, a class C is ε-learnable with k queries if and only if there is an algorithm that outputs after at most k queries a hypothesis $h \in B_n$ with error $d(f, h) \leqslant \varepsilon$. Further note that if d is a metric, i.e., if $d(f, g) = 0$ implies $f = g$, then $\mathrm{QC}(C, P)$ coincides with $\mathrm{QC}_{\varepsilon,d}(C, P)$ for $\varepsilon = 0$. Now we extend the general dimension to the approximate learning setting.

Definition 21. Let T be an answering scheme for a protocol P on a query set Q and let d be a pseudo-metric on B_n. We say that a set $A \subseteq T_Q$ of answers from T *ε-succeeds* on a concept class C ($A \in Succ_{\varepsilon,d}(C, T)$ for short), if $C(A)$

is covered by some ε-ball. The *general ε-dimension* of C under P and d, $\mathrm{Gdim}_{\varepsilon,d}(C, P)$, is the smallest integer $k \geqslant 0$ such that any answering scheme T for P provides a set A of at most k answers that ε-succeeds on C,

$$\mathrm{Gdim}_{\varepsilon,d}(C, P) = \max_{T \in \mathcal{T}(P)} \min_{A \in Succ_{\varepsilon,d}(C,T)} \|A\|.$$

We illustrate the notion with some easy examples. We first consider the problem of ε-learning the class C_X with membership queries under the uniform distribution, where X is a subset of $\{0,1\}^n$ and C_X contains all functions whose support is contained in X.

Example 22. Let $C_X \subseteq B_n$ be the class of all functions f with $f(x) = 0$ for $x \notin X$. First note that an ε-ball $B_{\varepsilon,\delta}(h)$ contains all functions g with $\|\{x \in \{0,1\}^n \mid h(x) \neq g(x)\}\| \leqslant \varepsilon 2^n$. Hence, if $\|X\| \leqslant \varepsilon 2^n$, then the whole class C_X is covered by the ε-ball $B_{\varepsilon,\delta}(\bar{0})$ around the null function $\bar{0}$, implying that the empty answer set *ε-succeeds* on C_X.

In case $\|X\| > \varepsilon 2^n$ consider any answering scheme T for MQ. Since T gives unique answers, T is consistent with some function $f \in B_n$. If f does not belong to C_X, then T provides an answer of the form $B_n(x, 1)$ with $x \notin X$, implying that $C_X(x, 1) = \emptyset$. Otherwise observe that the class $C_X(A)$ can be covered by some ε-ball if and only if A contains the answers of at least $k = \lceil \|X\| - \varepsilon 2^n \rceil$ membership queries from X. This shows that $\mathrm{Gdim}_{\varepsilon,\delta}(C_X, MQ) = \max(0, \|X\| - \lfloor \varepsilon 2^n \rfloor)$ (which coincides with the query complexity).

Next we consider the problem of learning singletons in the LBD model.

Example 23. In the LBD model, queries are arbitrary hypotheses $h \in B_n$ which are answered by SUCCESS, if the distance $d(f, h)$ between the target f and h is at most ε. Otherwise, the teacher returns for some tolerance parameter τ an estimate s for $d(f, h)$ with $|d(f, h) - s| \leqslant \tau$. This leads to the protocol

$$LBD_{\varepsilon,\tau,d} = \big\{\big(h, B_{\varepsilon,d}(h)\big) \bigm| h \in B_n\big\} \cup \big\{\big(h, \Lambda_{\varepsilon,\tau,d}(s, h)\big) \bigm| h \in B_n,\ s \in \mathbb{R}^+\big\},$$

where $\Lambda_{\varepsilon,\tau,d}(s, h) = \{g \in B_n \mid d(g, h) > \varepsilon \text{ and } s - \tau \leqslant d(g, h) \leqslant s + \tau\}$.

As a specific example we consider the problem of learning singletons under the uniform distribution in this model. Clearly, if $\varepsilon \geqslant 2^{-n}$, then any singleton is ε-close to the null function $\bar{0}$, implying that $\mathrm{Gdim}_{\varepsilon,\delta}(\mathrm{SING}_n, LBD_{\varepsilon,\tau,\delta}) = \mathrm{QC}_{\varepsilon,\delta}(\mathrm{SING}_n, LBD_{\varepsilon,\tau,\delta}) = 0$. Otherwise, ε-learning is equivalent to exact learning, and we have the following two subcases.

If $\varepsilon < 2^{-n} \leqslant \tau$, then letting $s = \delta(\bar{0}, h)$, the answer $\Lambda_{\varepsilon,\tau,\delta}(s, h)$ only discards h from SING_n (assuming that h actually is a singleton). Hence, using the same argument as in Example 6 for the EQ_r model, it follows that $\mathrm{Gdim}_{\varepsilon,\delta}(\mathrm{SING}_n, LBD_{\varepsilon,\tau,\delta}) = \mathrm{QC}_{\varepsilon,\delta}(\mathrm{SING}_n, LBD_{\varepsilon,\tau,\delta}) = \|\mathrm{SING}_n\| - 1 = 2^n - 1$.

If $\varepsilon, \tau < 2^{-n}$, then each answer claims exactly one value for the distance $\delta(f, h)$ between the query h and the target f. Hence, for any answering scheme we can find two answers to queries of the form h_k, h_{k+1}, where $h_k(x) = 1$ if and only if $(x)_2 \leqslant k$, such that at most one singleton function is consistent with these answers (here we use $(x)_2$ to denote the value of x as a binary number). This shows that $\mathrm{Gdim}_{\varepsilon,\delta}(\mathrm{SING}_n, LBD_{\varepsilon,\tau,\delta}) = \min(n, 2)$ and similarly it follows that $\mathrm{QC}_{\varepsilon,\delta}(\mathrm{SING}_n, LBD_{\varepsilon,\tau,\delta}) = n$.

Now we consider the statistical queries model introduced by Kearns [21] who has shown that any class learnable under this model is in fact learnable with classification noise in Valiant's PAC model. Here we only consider the distribution-specific variant with respect to some fixed distribution D.

Example 24. A statistical query is of the form $g : \{0,1\}^n \times \{0,1\} \to \{0,1\}$, i.e., $g \in B_{n+1}$, and the answer provides an estimate s of the probability $\Pr_D[g(x, f(x)) = 1]$, where f is the target, x is chosen randomly according to D and s has to be accurate within an additive error $\tau \geqslant 0$, referred to as the *tolerance*. The goal is to achieve a good approximation of the target with respect to the pseudo-metric $d(f, h) = \Pr_D[f(x) \neq h(x)]$. Formally, the model can be described by the protocol

$$STAT_{\tau,D} = \big\{\big(g, \Lambda_{\tau,D}(s, g)\big) \bigm| g \in B_{n+1},\ s \in \mathbb{R}^+\big\},$$

where $\Lambda_{\tau,D}(s, g) = \{h \in B_n \mid s - \tau \leqslant \Pr_D[g(x, h(x)) = 1] \leqslant s + \tau\}$.

The LBD and statistical queries models are essentially equivalent [6]. More precisely, let $LBD^*_{\tau,d}$ be the variant of the LBD model in which the teacher also returns an estimate s for the distance $d(f,h)$ when $d(f,h) \leqslant \varepsilon$, i.e.,

$$LBD^*_{\tau,d} = \left\{ \left(h, \Lambda^*_{\tau,d}(s,h)\right) \mid h \in B_n,\ s \in \mathbb{R}^+ \right\},$$

where $\Lambda^*_{\tau,d}(s,h) = \{g \in B_n \mid s - \tau \leqslant d(g,h) \leqslant s + \tau\}$. Since the answer $\Lambda_{\varepsilon,\tau,d}(s,h)$ does not provide less information than the answer $\Lambda^*_{\tau,d}(s,h)$, it follows that $\mathrm{QC}_{\varepsilon,d}(C, LBD_{\varepsilon,\tau,d}) \leqslant \mathrm{QC}_{\varepsilon,d}(C, LBD^*_{\tau,d})$. Further, it is not hard to verify that $\mathrm{QC}_{\varepsilon+2\tau,d}(C, LBD^*_{\tau,d}) \leqslant \mathrm{QC}_{\varepsilon,d}(C, LBD_{\varepsilon,\tau,d})$. Now assume that d is a pseudo-metric induced by some distribution D. Since every LBD^* query can be simulated by one statistical query which in turn can be simulated by two LBD^* queries [8,9], all queries having the same tolerance, it follows that $\mathrm{QC}_{\varepsilon,d}(C, STAT_{\tau,D}) \leqslant \mathrm{QC}_{\varepsilon,d}(C, LBD^*_{\tau,d}) \leqslant 2\,\mathrm{QC}_{\varepsilon,d}(C, STAT_{\tau,D})$. Hence we can state the following proposition.

Proposition 25. *Let d be a pseudo-metric induced by some distribution D. Then for any class $C \subseteq B_n$ it holds that*

$$\mathrm{QC}_{\varepsilon+2\tau,d}(C, STAT_{\tau,D}) \leqslant \mathrm{QC}_{\varepsilon,d}(C, LBD_{\varepsilon,\tau,d}) \leqslant 2\,\mathrm{QC}_{\varepsilon,d}(C, STAT_{\tau,D}).$$

In Theorem 29 below we show that Gdim is appropriate for approximate learning. As in the exact learning setting (see Lemma 7) a small value of Gdim guarantees the existence of a smart query that considerably shrinks the current set D of target candidates. But now the reduction factor does not only depend on the dimension but also on the maximum population size $r_{\varepsilon,d}(D) = \max_{h \in B_n} \|B_{\varepsilon,d}(h) \cap D\|$ of concepts from D inside a single ε-ball.

Lemma 26. *Let $\mathrm{Gdim}_{\varepsilon,d}(C,P) = k \geqslant 1$. Then for any class $D \subseteq C$ there is a query $q \in Q$ such that for all answers $\Lambda \in P_q$, at least $(\|D\| - r_{\varepsilon,d}(D))/k$ functions from D are inconsistent with Λ.*

Proof. Let $\mathrm{Gdim}_{\varepsilon,d}(C,P) = k \geqslant 1$ and suppose that for any query q there is an answer $\Lambda_q \in P_q$ such that $\|D - D(\Lambda_q)\| < (\|D\| - r_{\varepsilon,d}(D))/k$. Consider the answering scheme $T = \{(q, \Lambda_q) \mid q \in Q\}$. Then for any $A \subseteq T_Q$ with $\|A\| \leqslant k$, D contains fewer than $\|D\| - r_{\varepsilon,d}(D)$ functions that are inconsistent with A, implying that $\|D(A)\|$ is larger than $r_{\varepsilon,d}(D)$. Hence, $C(A)$ (which is a superset of $D(A)$) cannot be contained in an ε-ball. But this contradicts the assumption $\mathrm{Gdim}_{\varepsilon,d}(C,P) = k$. □

In order to get a learning algorithm L out of Lemma 26, the protocol P should enable L to check whether a hypothesis h is already sufficiently close to the target f. More precisely, we assume that for any $h \in B_n$ there is some query q such that any answer in P_q provides the information whether $d(h,f) \leqslant \varepsilon$ (implying success) or $d(h,f) > \varepsilon - \gamma$ or both. Here we use an additional tolerance parameter γ, since in some models like learning with statistical queries the teacher returns only an estimate of the distance between h and f.

Definition 27. Let $\varepsilon, \gamma \geqslant 0$. A protocol P is called (ε, γ, d)-*affirmative*, if for each hypothesis $h \in B_n$ there is some query $q \in Q$ such that each answer $\Lambda \in P_q$ either fulfills $\Lambda \subseteq B_{\varepsilon,d}(h)$ or $\Lambda \subseteq B_n - B_{\varepsilon-\gamma,d}(h)$.

Example 28. Clearly, the protocol $LBD_{\varepsilon,\tau,d}$ is (ε,γ,d)-affirmative for any γ. Further, since we can express $d(f,h)$ as the probability $\Pr_D[g_h(x, f(x)) = 1]$, where g_h is defined as $g_h(x,b) = h(x) \oplus b$, it follows that the answer $\Lambda_{\tau,D}(s, g_h)$ is contained in the ball $B_{\varepsilon,d}(h)$ if $s \leqslant \varepsilon - \tau$, and in the complement of the ball $B_{\varepsilon-2\tau,d}(h)$, otherwise. This shows that the protocol $STAT_{\tau,D}$ is $(\varepsilon, 2\tau, d)$-affirmative for $0 \leqslant \tau \leqslant \varepsilon/2$.

Theorem 29. *Let d be a pseudo-metric and let P be an (ε,γ,d)-affirmative protocol with $\gamma \leqslant \varepsilon$. Then for any class $C \subseteq B_n$ it holds that*

$$\mathrm{Gdim}_{\varepsilon,d}(C,P) \leqslant \mathrm{QC}_{\varepsilon,d}(C,P) \leqslant 2\left\lceil \ln\|C\| \right\rceil \mathrm{Gdim}_{\varepsilon-\gamma,d}(C,P).$$

Proof. We prove the first inequality by showing that $\mathrm{Gdim}_{\varepsilon,d}(C,P) > k$ implies $\mathrm{QC}_{\varepsilon,d}(C,P) > k$. Assume that $T \subseteq P$ is an answering scheme with the property that for all sets $A \subseteq T_Q$ of at most k answers, $C(A)$ is not contained in any ε-ball. Then no learning algorithm L is able to ε-learn C with k queries, since L may receive for each query q an answer $\Lambda \in T_q$. In this case, L receives a set A of at most k answers $\Lambda \in T_Q$ and hence, the set $C(A)$ of target candidates is not contained in any ε-ball.

To see the second inequality, assume that $\mathrm{Gdim}_{\varepsilon-\gamma,d}(C, P) = k \geqslant 1$. We describe an algorithm L that ε-learns C under P with $m = 2k\lceil \ln\|C\|\rceil$ queries. Starting with the empty answer set $A = \emptyset$, in each round, L first determines the value $r(A) = r_{\varepsilon-\gamma,d}(C(A))$. If $r(A) \leqslant \|C(A)\|/2$, L asks the query q provided by Lemma 26 for the class $D = C(A)$ and includes the answer Λ into A. This answer discards at least

$$\big\|C(A) - C\big(A \cup \{\Lambda\}\big)\big\| \geqslant (1/k)\big(\big\|C(A)\big\| - r(A)\big) \geqslant \big\|C(A)\big\|/2k$$

hypotheses from $C(A)$. If $r(A) > \|C(A)\|/2$, then there is a function $h \in B_n$ such that $\|B_{\varepsilon-\gamma,d}(h) \cap C(A)\| > \|C(A)\|/2$. Since P is (ε, γ, d)-affirmative, L can ask a query q whose answer Λ is either contained in $B_{\varepsilon,d}(h)$ or in the complement of $B_{\varepsilon-\gamma,d}(h)$. In the first case, $C(A \cup \{\Lambda\})$ is covered by some ε-ball, and in the second case, the cardinality of $C(A)$ shrinks at least by $1/2$. Hence, any answer Λ which does not lead to the success of L discards at least a $1/2k$ fraction from $C(A)$. A simple calculation shows that $(1 - 1/2k)^m\|C\| \leqslant 1$. Hence L succeeds after at most m queries. □

Since the protocol $STAT_{\tau,D}$ is $(\varepsilon, 2\tau, d)$-affirmative for $0 \leqslant \tau \leqslant \varepsilon/2$, Theorem 29 provides the following bounds for the statistical query model.

Corollary 30. *Let P be the protocol $STAT_{\tau,D}$ and let d be the pseudo-metric induced by D. Then for any class $C \subseteq B_n$ and $0 \leqslant \tau \leqslant \varepsilon/2$ it holds that*

$$\mathrm{Gdim}_{\varepsilon,d}(C, P) \leqslant \mathrm{QC}_{\varepsilon,d}(C, P) \leqslant 2\big\lceil \ln\|C\|\big\rceil \mathrm{Gdim}_{\varepsilon-2\tau,d}(C, P).$$

Blum et al. [8] defined the statistical query dimension $\mathrm{SQdim}(C, D)$ of a class C under a distribution D to be the largest number k such that C contains k functions $f_1, \ldots, f_k$ satisfying $|d(f_i, f_j) - \frac{1}{2}| \leqslant 1/2k^3$ for all $i \neq j$, where d is the pseudo-metric induced by D. Let $\mathrm{SQdim}(C, D) = k$. Blum et al. showed that $\mathrm{QC}_{\varepsilon,d}(C, STAT_{\tau,D}) \geqslant k^{1/3}/2$ if $\varepsilon < 1/2 - 1/k^3$ and $\tau \geqslant 1/k^{1/3}$, and $\mathrm{QC}_{\varepsilon,d}(C, STAT_{\tau,D}) \leqslant k$ if $\varepsilon \geqslant 1/2 - 1/3k^3$ and $\tau \leqslant 1/3k^3$. Note that the upper bound holds only for values of ε which correspond to weak learning. In contrast, the bound given by Corollary 30 is valid for all $0 \leqslant \tau \leqslant \varepsilon/2$.

Since the protocol $LBD_{\varepsilon,\tau,d}$ is (ε, γ, d)-affirmative for $\gamma = 0$, Theorem 29 immediately provides the following bounds.

Corollary 31. *Let d be a pseudo-metric and let P be the protocol $LBD_{\varepsilon,\tau,d}$. Then for any class $C \subseteq B_n$ and $\tau, \varepsilon \geqslant 0$ it holds that*

$$\mathrm{Gdim}_{\varepsilon,d}(C, P) \leqslant \mathrm{QC}_{\varepsilon,d}(C, P) \leqslant 2\big\lceil \ln\|C\|\big\rceil \mathrm{Gdim}_{\varepsilon,d}(C, P).$$

In [6] the query complexity of a class C in the LBD model is compared to the capacity $\mathrm{Cap}_{\varepsilon,d}(C)$ of C which is defined as the smallest integer k such that k ε-balls are sufficient to cover C. There it is shown that $\log_q \mathrm{Cap}_{\varepsilon,d}(C) \leqslant \mathrm{QC}_{\varepsilon,d}(C, LBD_{\varepsilon,\tau,d}) \leqslant \mathrm{Cap}_{\varepsilon,d}(C)$, where $q = D/\tau$ and $D = \max_{f,h\in C} d(f, h)$ denotes the diameter of C. In fact, when the capacity is used to bound the query complexity in the LBD model, the exponential gap between the lower and the upper bound is in general unavoidable. In contrast, Corollary 31 shows that the general dimension yields lower and upper bounds that only differ by a factor which is logarithmic in the cardinality of C.

Next we show that in the specific setting where the pseudo-metric d is induced by some distribution D and the error bound ε is sufficiently close to $1/2$, also the capacity yields tight lower and upper bounds. To this end we first derive the following relationship between the capacity and the general dimension.

Lemma 32. *Let d be a pseudo-metric induced by some distribution. Then for any class $C \subseteq B_n$ and $0 \leqslant 1/2 - \tau \leqslant \varepsilon \leqslant 1/2$ it holds that*

$$\mathrm{Cap}_{\varepsilon,d}(C) \leqslant 2\,\mathrm{Gdim}_{\varepsilon,d}(C, LBD_{\varepsilon,\tau,d}) + 1.$$

Proof. Let $k = \mathrm{Gdim}_{\varepsilon,d}(C, LBD_{\varepsilon,\tau,d})$ and consider the answering scheme $T = \{(h, \Lambda(h)) \mid h \in B_n\}$ where $\Lambda(h) = \Lambda_{\varepsilon,\tau,d}(1/2, h)$. Then there is a set A of at most k answers from T such that $C(A)$ is contained in some ε-ball B. Note that this is equivalent to saying that the complements of the answers $\Lambda(h) \in A$ together with B cover C. Any function

$g \notin \Lambda(h)$ either fulfills $d(g,h) \leqslant \varepsilon$ or $d(g,h) < 1/2 - \tau$ or $d(g,h) > 1/2 + \tau$. Note that in the latter case we have $d(g,\neg h) = 1 - d(g,h) < 1/2 - \tau$. By the assumption that $1/2 - \tau \leqslant \varepsilon$ it follows that the complement of $\Lambda(h)$ is covered by the two ε-balls $B_{\varepsilon,d}(h)$ and $B_{\varepsilon,d}(\neg h)$. Thus we can conclude that $\mathrm{Cap}_{\varepsilon,d}(C) \leqslant 2k+1$. □

It is easy to see that the upper bound $\mathrm{QC}_{\varepsilon,d}(C, LBD_{\varepsilon,\tau,d}) \leqslant \mathrm{Cap}_{\varepsilon,d}(C)$ holds for any ε. Thus, Lemma 32 and Theorem 29 together imply the following tight relationship between the capacity and the learning complexity in the LBD model when the error bound ε is close to $1/2$ (see also [9]).

Corollary 33. *Let d be a pseudo-metric induced by some distribution. Then for any τ, ε with $0 \leqslant 1/2 - \tau \leqslant \varepsilon \leqslant 1/2$ and any $C \subseteq B_n$ it holds that*

$$\big(\mathrm{Cap}_{\varepsilon,d}(C) - 1\big)/2 \leqslant \mathrm{QC}_{\varepsilon,d}(C, LBD_{\varepsilon,\tau,d}) \leqslant \mathrm{Cap}_{\varepsilon,d}(C).$$

7. Learning DNF formulas with statistical queries

In this section we derive tight upper and lower bounds for the query complexity of DNF formulas in the statistical queries model under the uniform distribution. First we recall the Fourier transform of a real valued function f on $\{0,1\}^n$. For every subset $A \subseteq [n]$, where $[n]$ denotes the set $\{1,\ldots,n\}$, let the parity function χ_A be defined as $\chi_A(x) = (-1)^{\sum_{i\in A} x_i}$, where the summation is over $GF(2)$. Then every real valued function f on $\{0,1\}^n$ can be uniquely expressed as a linear combination $f(x) = \sum_{A\subseteq[n]} \hat{f}(A)\chi_A(x)$, where each coefficient $\hat{f}(A)$ is given by the expectation $\mathrm{E}_U[f(x)\chi_A(x)]$. These coefficients constitute the Fourier transform of f. When applying the Fourier transform to a Boolean function f, it is convenient to think of f as a mapping from $\{0,1\}^n$ to $\{-1,1\}$ and, consequently, of a statistical query as a mapping from $\{0,1\}^n \times \{-1,1\}$ to $\{-1,1\}$. Then the Fourier coefficients $\hat{f}(A)$ of f can be expressed as $\mathrm{E}_U[f(x)\chi_A(x)] = \Pr_U[\chi_A(x) = f(x)] - \Pr_U[\chi_A(x) \neq f(x)] = 1 - 2\delta(\chi_A, f)$, where δ denotes the metric induced by the uniform distribution U.

For the upper bound we use the fact that every DNF formula has a large coefficient in its Fourier transform [8]. This holds with respect to arbitrary distributions D on $\{0,1\}^n$ [20]. Recall that the norm $L_\infty(D)$ is defined as $\max_{x\in\{0,1\}^n} D(x)$.

Lemma 34. *(See [9].) For every m-term DNF formula f and every distribution D on $\{0,1\}^n$, there is a set $A \subseteq [n]$ of size at most $\log(2^n(2m+1)L_\infty(D))$ such that $|\mathrm{E}_D[f(x)\chi_A(x)]| \geqslant 1/(2m+1)$.*

By Lemma 34 we can use a simple weak learning algorithm (cf. [20]) to weakly predict DNF formulas. We then apply Freund's well-known boosting algorithm *F1* [13] which, as shown by Aslam and Decatur [3], also works in the statistical queries model.

Lemma 35. *(Cf. [3, Theorem 2].) Let $C \subseteq B_n$ and $\varepsilon > 0$. Then any weak learning algorithm which produces under any distribution D with $L_\infty(D) \leqslant 3/65\varepsilon^2 2^n$ for any target $f \in C$ after at most N_0 statistical queries of tolerance $\tau_0 \geqslant 0$ a hypothesis h with error $\Pr_D[f(x) \neq h(x)] \leqslant 1/2 - \gamma$ can be transformed into an algorithm for learning C which uses at most $\gamma^{-4}\ln^2(\varepsilon^{-1})N_0$ statistical queries of tolerance $\varepsilon^2\tau_0/65$ to produce under the uniform distribution a hypothesis h with error $\delta(f,h) \leqslant \varepsilon$.*

Theorem 36. *There is a constant $c > 0$ such that for all $\varepsilon < 1/2$ and $\tau \leqslant c\varepsilon^2/m$ it holds that*

$$\mathrm{QC}_{\varepsilon,\delta}(m\text{-DNF}_n, STAT_{\tau,U}) \leqslant n^{O(\log(m/\varepsilon))}.$$

Proof. Let $\tau_0 = 1/(16m+8)$. In order to apply Lemma 35 we first describe a weak learning algorithm WL for m-DNF$_n$ that under any distribution D with $L_\infty(D) \leqslant 3/65\varepsilon^2 2^n$ uses $N_0 = n^{O(\log(m/\varepsilon))}$ statistical queries of tolerance τ_0 to produce a hypothesis h with error at most $1/2 - \tau_0$.

WL asks for all sets $A \subseteq [n]$ of size at most $\log((6m+3)/65\varepsilon^2)$ for estimates a_A of the expectations $\mathrm{E}_D[f(x)\chi_A(x)]$ within additive error $2\tau_0$ until it receives an answer a_A satisfying $|a_A| \geqslant 1/(2m+1) - 2\tau_0$. Then *WL* outputs the hypothesis χ_A if a_A is positive, and $-\chi_A$ otherwise.

Since $\mathrm{E}_D[f(x)h(x)] = 1 - 2\Pr_D[f(x) \neq h(x)]$, each estimate a_A can be obtained by a statistical query of tolerance τ_0 with respect to the unknown distribution D and the target f. Hence, *WL* only needs to ask $N_0 = n^{O(\log(m/\varepsilon))}$ many queries. Further, for any distribution D with $L_\infty(D) \leqslant 3/65\varepsilon^2 2^n$, Lemma 34 guarantees that *WL* receives an answer a_A with $|a_A| \geqslant 1/(2m+1) - 2\tau_0$, implying that $|\mathrm{E}_D[f(x)\chi_A(x)]| \geqslant 1/(2m+1) - 4\tau_0 = 2\tau_0$. It follows that *WL* indeed produces a hypothesis h with error $\Pr_D[f(x) \neq h(x)] \leqslant 1/2 - \tau_0$.

Now, if we choose $c = 1/1560$, then the assumption $\tau \leqslant c\varepsilon^2/m$ implies that $\tau \leqslant \varepsilon^2/65(16m+8) = \varepsilon^2\tau_0/65$. Hence we can apply Lemma 35 with $C = m\text{-DNF}_n$ and $\gamma = \tau_0$ to get an algorithm which uses

$$\gamma^{-4}\ln^2\bigl(\varepsilon^{-1}\bigr)N_0 = n^{O(\log(m/\varepsilon))}$$

many queries under the protocol $STAT_{\tau,U}$ to produce for any target $f \in C$ a hypothesis h with error $\delta(f,h) \leqslant \varepsilon$. □

We note that by applying more powerful boosting strategies, it can be shown [23] that m-term DNFs are ε-learnable from $n^{O(\log(m/\varepsilon))}$ statistical queries with tolerance $\Omega(\varepsilon/m)$ instead of $\Omega(\varepsilon^2/m)$ as in Theorem 36.

Next we consider the lower bound. Since every parity function χ_A, $A \neq \emptyset$, can be represented by a $2^{\|A\|-1}$-term DNF, the number of parities representable as a DNF with at most $m \leqslant 2^{\sqrt{n}}$ terms is at least

$$\sum_{i \leqslant \lceil \log m \rceil} \binom{n}{i} \geqslant (n/\log m)^{\log m} \geqslant n^{(\log m)/2}.$$

Further recall that $\delta(\chi_A, \chi_B) = 1/2$ for $A \neq B$. Hence, using the bound provided by the SQ dimension [8] (see the paragraph after Corollary 30), it follows that under the uniform distribution at least $n^{(\log m)/6}/2$ statistical queries of tolerance $\tau = n^{-(\log m)/6}$ are needed to ε-learn m-term DNFs, provided that $m \leqslant 2^{\sqrt{n}}$ and $\varepsilon < 1/2 - n^{-2(\log m)/3}$.

Theorem 37. *(Cf. [8].) For all $m \leqslant 2^{\sqrt{n}}$, $\tau \geqslant n^{-(\log m)/6}$ and $\varepsilon < 1/2 - n^{-2(\log m)/3}$ it holds that*

$$\mathrm{QC}_{\varepsilon,\delta}(m\text{-DNF}_n, STAT_{\tau,U}) \geqslant n^{(\log m)/6}/2.$$

Alternately we can use the capacity to bound $\mathrm{QC}_{\varepsilon,\delta}(m\text{-DNF}_n, STAT_{\tau,U})$ (see Corollary 33). In fact, using Parseval's identity and the fact that the coefficient $\hat{f}(A)$ for any parity function χ_A in an ε-ball $B_{\varepsilon,\delta}(f)$ is $1 - 2\delta(\chi_A, f) \geqslant 1 - 2\varepsilon$ it is easy to see that a single ε-ball contains at most $1/(1-2\varepsilon)^2$ parity functions, implying that $\mathrm{Cap}_{\varepsilon,\delta}(m\text{-DNF}_n) \geqslant (1-2\varepsilon)^2 n^{(\log m)/2}$. In order to combine the upper and lower bounds of Theorems 36 and 37, we consider a fixed error bound $\varepsilon_0 < \frac{1}{2}$ (say $\varepsilon_0 = \frac{1}{3}$) and use $\mathrm{QC}_{\mathrm{const},\delta}(C, P)$ to denote the corresponding query complexity.

Corollary 38. *There is a constant $c > 0$, such that for all $m \leqslant 2^{\sqrt{n}}$ and all τ with $n^{-(\log m)/6} \leqslant \tau \leqslant c/m$ it holds that*

$$\mathrm{QC}_{\mathrm{const},\delta}(m\text{-DNF}_n, STAT_{\tau,U}) = n^{\Theta(\log m)}.$$

Acknowledgments

Section 5 of the paper strongly benefits from conversations with Vijay Raghavan during his sabbatical in Barcelona. Shai Ben-David discussed with one of the authors the learning by distance model and its relationship with the statistical query model. Also, we thank the referees for their useful comments.

References

[1] D. Angluin, Queries and concept learning, Mach. Learn. 2 (1988) 319–342.
[2] D. Angluin, Negative results for equivalence queries, Mach. Learn. 5 (1990) 121–150.
[3] J. Aslam, S. Decatur, General bounds on statistical query learning and PAC learning with noise via hypothesis boosting, Inform. and Comput. 141 (2) (1998) 85–118.
[4] J.L. Balcázar, J. Castro, D. Guijarro, A new abstract combinatorial dimension for exact learning via queries, J. Comput. System Sci. 64 (2002) 2–21.
[5] J.L. Balcázar, J. Castro, D. Guijarro, H.-U. Simon, The consistency dimension and distribution-dependent learning from queries, Theoret. Comput. Sci. 288 (2) (2002) 197–215.
[6] S. Ben-David, A. Itai, E. Kushilevitz, Learning by distances, Inform. and Comput. 117 (2) (1995) 240–250.

[7] A. Birkendorf, N. Klasner, C. Kuhlmann, H.U. Simon, Structural results about exact learning with unspecified attribute values, J. Comput. System Sci. 60 (2) (2000) 258–277.
[8] A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, S. Rudich, Weakly learning DNF and characterizing statistical query learning using Fourier analysis, in: Proceedings of the Twenty-Sixth Annual ACM Symposium on the Theory of Computing, Montréal, Québec, Canada, 1994, pp. 253–262.
[9] N.H. Bshouty, V. Feldman, On using extended statistical queries to avoid membership queries, J. Mach. Learn. Res. 2 (2002) 359–395.
[10] N.H. Bshouty, D.K. Wilson, On learning in the presence of unspecified attribute values, in: Proceedings of the 12th Annual Conference on Computational Learning Theory, ACM Press, 1999, pp. 81–87.
[11] J. Castro, A note on bounded query learning, Technical report, Universitat Politècnica de Catalunya. LSI-05-16-R, available at http://www.lsi.upc.edu/~castro/bounded.ps, 2005.
[12] S. Dasgupta, W. Lee, P. Long, A theoretical analysis of query selection for collaborative filtering, Mach. Learn. 51 (2003) 283–298.
[13] Y. Freund, Boosting a weak learning algorithm by majority, Inform. and Comput. 121 (2) (1995) 256–285.
[14] R. Gavaldà, The complexity of learning with queries, in: Proceedings of the 9th Annual Conference on Structure in Complexity Theory, Los Alamitos, CA, USA, IEEE Computer Society Press, 1994, pp. 324–337.
[15] R. Gavaldà, On the power of equivalence queries, in: Proceedings of the 1st European Conference on Computational Learning Theory: EuroCOLT '93, Oxford University Press, 1994, pp. 193–203.
[16] S. Goldman, M. Kearns, On the complexity of teaching, J. Comput. System Sci. 50 (1995) 20–31.
[17] S.A. Goldman, S.S. Kwek, S.D. Scott, Learning from examples with unspecified attribute values, Inform. and Comput. 180 (2) (2003) 82–100.
[18] T. Hegedüs, Generalized teaching dimension and the query complexity of learning, in: Proceedings of the 8th Annual Conference on Computational Learning Theory, ACM Press, 1995, pp. 108–117.
[19] L. Hellerstein, K. Pillaipakkamnatt, V. Raghavan, D. Wilkins, How many queries are needed to learn?, J. Assoc. Comput. Mach. 43 (5) (1996) 840–862.
[20] J.C. Jackson, An efficient membership-query algorithm for learning DNF with respect to the uniform distribution, J. Comput. System Sci. 55 (1997) 414–440.
[21] M. Kearns, Efficient noise-tolerant learning from statistical queries, J. ACM 45 (6) (1998) 983–1006.
[22] J. Köbler, W. Lindner, The complexity of learning concept classes with polynomial general dimension, Theoret. Comput. Sci. 350 (1) (2006) 49–62.
[23] W. Lindner, Learning DNF by statistical and proper distance queries under the uniform distribution, in: Proc. 16th International Workshop on Algorithmic Learning Theory, 2005, pp. 198–210.
[24] N. Littlestone, Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm, Mach. Learn. 2 (1988) 285–318.