A General Dimension for Approximately Learning Boolean Functions*

Johannes Köbler and Wolfgang Lindner

Institut für Informatik, Humboldt-Universität zu Berlin, 10099 Berlin, Germany

Abstract. We extend the notion of general dimension, a combinatorial characterization of learning complexity for arbitrary query protocols, to encompass approximate learning. This immediately yields a characterization of the learning complexity in the statistical query model. As a further application, we consider approximate learning of DNF formulas and we derive close upper and lower bounds on the number of statistical queries needed. In particular, we show that with respect to the uniform distribution, and for any constant error parameter $\varepsilon < 1/2$, the number of statistical queries needed to approximately learn DNF formulas (over n variables and s terms) with tolerance $\tau = \Theta(1/s)$ is $n^{\Theta(\log s)}$.

1 Introduction

A main issue in computational learning theory is to decide whether a concept class can be learned with a polynomial number of interactions with a teacher, regardless of the computational resources required by the learning algorithm. One way of addressing this problem is to apply combinatorial characterizations such as the VC dimension for the PAC learning model [8], or the approximate fingerprint property for the exact learning model with equivalence queries [2]. Further combinatorial characterizations for various query-types in the exact learning model can be found in, e.g., [11, 12, 5].

In [4], Balcázar et al. introduce the notion of a general dimension as a uniform way to characterize, up to a logarithmic factor, the learning complexity, i.e., the number of queries both necessary and sufficient for exact learning, with any set of queries.

Here we address the question whether the general dimension can be extended to capture approximate learning, where the learning algorithm has to identify the target only with respect to some error parameter ε. We show that the general dimension of [4] can in fact be adapted so that the resulting notion characterizes the learning complexity for approximate learning, provided that the learning algorithm is able to decide (possibly with the help of the teacher) whether a hypothesis is already a sufficiently close approximation of the target. This immediately yields characterizations of the learning complexity in the statistical query model. In contrast to the SQ-dimension of [7], our general dimension for

* Work supported by the DFG under project KO 1053/1-1

statistical queries works for any error parameter ε and any tolerance parameter $\tau \leq \varepsilon$.

As an application, we consider approximate learning of DNF formulas (over n variables and s terms) with statistical queries and we show that with respect to the uniform distribution, and for any constant error parameter $\varepsilon < 1/2$ it is possible to choose the tolerance parameter $\tau = \Theta(1/s)$ such that the number of statistical queries needed to approximately learn DNF formulas with tolerance τ is $n^{\Theta(\log s)}$.

2 Approximate Learning for Arbitrary Query Protocols

In this paper we only consider concept learning of boolean functions of some arity n. We let B_n denote the set of all boolean functions $f : \{0,1\}^n \to \{0,1\}$ and $[n]$ denotes the set $\{1, \ldots, n\}$.

The learning model presented in [4] is a generalization of the query learning model of Angluin [1]. In this model, learning of a concept class $C \subseteq B_n$ can be viewed as a game between a learning algorithm A and a teacher T with respect to some target $f \in C$ that is only known to T. In each round, A asks a query q from a set Q of queries and T responds with a subset $\Lambda \subseteq B_n$ that contains f. Thereby, T provides some partial knowledge about the target f in form of a property Λ shared by f. The communication between A and T is guided by some *protocol* $P \subseteq Q \times 2^{B_n}$, i.e., the teacher is only allowed to respond with an answer Λ such that $\langle q, \Lambda \rangle \in P$. Further, P is required to be *complete* in the sense that for all $f \in B_n$ and for all queries $q \in Q$ there exists an answer $\Lambda \in P_q^f$, where $P_q^f = \{\Lambda \mid \langle q, \Lambda \rangle \in P \wedge f \in \Lambda\}$.

In the approximate learning setting, the goal of the learning algorithm is to collect enough knowledge about f such that every function in C that shares all properties exposed by T is a close approximation to f. As in [6] we formally express the success condition in terms of an arbitrary *pseudo-metric* ϱ on B_n, i.e., ϱ is a function from B_n^2 to non-negative reals satisfying $\varrho(f,f) = 0$, $\varrho(f,g) = \varrho(g,f)$, and $\varrho(f,h) \leq \varrho(f,g) + \varrho(g,h)$, for all functions f, g and h in B_n. We will frequently use pseudo-metrics of the form $\varrho(f,g) = \Pr_D[f(x) \neq g(x)]$, for some distribution D on $\{0,1\}^n$.

In the following we will further use the following notations. The *diameter* of a set of boolean functions $F \subseteq B_n$ is the maximal distance between two members of F, $\varrho(F) = \max_{f,g \in F} \varrho(f,g)$, and for $f \in B_n$ and $\rho \geq 0$, we let $B_\varrho(f, \rho) = \{g \in B_n \mid \varrho(f,g) \leq \rho\}$ denote the *ball* of radius ρ with center f.

Let $C \subseteq B_n$ be a concept class, let $P \subseteq Q \times 2^{B_n}$ be a protocol, and let ϱ be a pseudo-metric on B_n. A teacher T *answers with respect to f using* P if for each query $q \in Q$, T responds with an answer $\Lambda \in P_q^f$. The *current version space* at any stage in the run of a learning algorithm A is the set of all functions in C that are contained in all answers received by A so far. The concept class C is *ε-learnable with at most d queries under* P, if there exists a learning algorithm A such that for all $f \in C$, and for any teacher T which answers with respect to f using P, after at most d queries the current version space is contained in the ball $B_\varrho(f,\varepsilon)$. We define the *learning complexity* of C under P, denoted by $\mathrm{LC}_\varrho(C,P,\varepsilon)$, as the smallest integer $d \geq 0$ such that C is ε-learnable with at most d queries under P. If no such integer d exists, then $\mathrm{LC}_\varrho(C,P,\varepsilon) = \infty$.

Before we proceed, let us consider some examples. The protocol $\mathrm{Eq}(C)$ for equivalence queries with hypotheses from a concept class $C \subseteq B_n$ is the set of all pairs $\langle h, \{f \in \mathcal{B}_n \mid f(x) \neq h(x)\}\rangle$, for $h \in C$ and $x \in \{0,1\}^n$, together with all pairs $\langle h, \{h\}\rangle$ with $h \in C$. The first set of pairs corresponds to the case that a hypothesis h is answered with a counterexample x, and the second set of pairs corresponds to the answer "Yes". Note that, for $\varepsilon = 0$, the learning complexity of C under the protocol $\mathrm{Eq}(C)$ describes precisely the number of queries both necessary and sufficient for exact learning C with proper equivalence queries.

In the model of learning from statistical queries [14], queries are of the form $\chi : \{0,1\}^n \times \{0,1\} \to \{0,1\}$ (i.e., $\chi \in B_{n+1}$) for which we receive an estimate σ of the probability $\Pr_D[\chi(x,f(x)) = 1]$, for a target f a randomly chosen $x \in_D \{0,1\}^n$ and with respect to some distribution D on $\{0,1\}^n$. The estimate has to be accurate within an additive error $\tau \geq 0$, which we refer to as the tolerance. Thus, a protocol for statistical queries can be defined as

$$\mathrm{STAT}_D(\tau) = \{\langle \chi, \Lambda_{\chi,\sigma}\rangle \mid \chi \in B_{n+1}, \sigma \in \mathbb{R}^+\},$$

where $\Lambda_{\chi,\sigma}$ is the set of all functions $f \in B_n$ such that

$$|\Pr_D[\chi(x,f(x)) = 1] - \sigma| \leq \tau.$$

In the less known model of learning by distance [6], a query is a hypothesis $h \in B_n$ for which we get an approximation of the distance between h and the target f with respect to some given pseudo metric ϱ. Thus, the corresponing protocol can be defined as

$$\mathrm{DIST}_\varrho(\tau) = \{\langle h, S_\varrho(h,\rho,\tau)\rangle \mid h \in B_n, \rho \in \mathbb{R}^+\},$$

where $S_\varrho(h,\rho,\tau)$ denotes the *sphere* around h of radius ρ and width τ, $S_\varrho(h,\rho,\tau) = \{f \in B_n \mid |\varrho(f,h) - \rho| \leq \tau\}$.

In order to compare the strength of different protocols, we use the following notion of reducibility between protocols.

Definition 1. *A protocol $P' \subseteq Q' \times 2^{B_n}$* simulates a protocol $P \subseteq Q \times 2^{B_n}$ with at most k queries *(in which case we also write $P \leq_k P'$), if for every query $q \in Q$ there exist k queries $q'_1, \ldots, q'_k \in Q'$, such that for all answers $\Lambda'_i \in P'_{q_i}$, $i = 1, \ldots, k$, there exists an answer $\Lambda \in P_{q'}$ such that $\bigcap_{i=1}^k \Lambda'_i \subseteq \Lambda$.*

The following basic properties of the reducibility $\leq_k$ are easily verified.

Proposition 1. *1. $P \leq_k P$, $k \geq 1$,*
2. $P \leq_k P'$ and $P' \leq_l P''$ implies $P \leq_{kl} P''$,
3. $P \leq_k P'$ implies both $\mathrm{LC}_\varrho(C,P',\varepsilon) \leq k\mathrm{LC}_\varrho(C,P,\varepsilon)$ and $\mathrm{Gdim}(C,P',\varepsilon) \leq k\mathrm{Gdim}(C,P,\varepsilon)$.

For a statistical query $\chi(x,y) = y \oplus h(x)$, where $h \in B_n$, it holds that $\Pr_D[\chi(x, f(x)) = 1] = \Pr_D[h(x) \oplus f(x) = 1] = \Pr_D[h \neq f] = \varrho(f,h)$. Hence, any answer to such a query χ is just some sphere $S_\varrho(h,\sigma,\tau)$ around h of radius σ and width τ with respect to the metric $\varrho(f,g) = \Pr_D[f \neq g]$. This means that $\mathrm{DIST}_\varrho(\tau) \leq_1 \mathrm{STAT}_D(\tau)$. On the other hand, arguments from [7] can be used to show that every statistical query can be simulated by two distance queries [9].

Proposition 2. *For any distribution D on $\{0,1\}^n$, and for the metric $\varrho(f,g) = \Pr_D[f \neq g]$, it holds that $\mathrm{DIST}_\varrho(\tau) \leq_1 \mathrm{STAT}_D(\tau) \leq_2 \mathrm{DIST}_\varrho(\tau)$.*

3 General Dimension for Approximate Learning

Now we extend the general dimension of [4] to encompass approximate learning. For a query $q \in Q$ and a subset $T \subseteq P$, T_q denotes the set of all answers Λ such that $\langle q, \Lambda\rangle \in T$, whereas $T_Q = \bigcup_{q\in Q} T_q$. The definition is based on the notion of an *answering scheme* $T \subseteq P$, i.e., T has the property that for all queries $q \in Q$ there exists an answer $\Lambda \in T_q$. Note that in contrast to P, T need not be complete since there might exist a query $q \in Q$ and a function $f \in B_n$ such that no answer $\Lambda \in T_q$ contains f.

Definition 2. *The* general dimension $\mathrm{Gdim}_\varrho(C,P,\varepsilon)$ *of C under protocol P and pseudo-metric ϱ is the smallest integer $d \geq 0$ such that for all answering schemes $T \subseteq P$ there exists a set $S \subseteq T_Q$ of cardinality at most d such that the diameter of the set*

$$\bigcap S \cap C = \{h \in C \mid (\forall \Lambda \in S)[h \in \Lambda]\}$$

is at most ε. If no such integer d exists, then $\mathrm{Gdim}_\varrho(C,P,\varepsilon) = \infty$.

Lemma 1. $\mathrm{Gdim}_\varrho(C,P,\varepsilon) \leq \mathrm{LC}_\varrho(C,P,\varepsilon)$

Proof. Suppose $\mathrm{Gdim}_\varrho(C,P,\varepsilon) > d$. Then there exists an answering scheme $T \subseteq P$ such that for all $S \subseteq T_Q$ of cardinality at most d it holds that $\varrho(\bigcap S \cap C) > \varepsilon$. Fix some learning algorithm A and consider a run of A where the i-th query q_i of A is answered by some $\Lambda_i \in T_{q_i}$. Since T is an answering scheme, these answers always exist. By the assumption, the version space after d queries contains two functions f and g with $\varrho(f,g) > \varepsilon$. It follows that there is some teacher which answers with respect to f using P such that after d queries, the current version space is not contained in $B_\varrho(f,\varepsilon)$, i.e., $\mathrm{LC}_\varrho(C,P,\varepsilon) > d$. □

In the next lemma we use the following notation. For any set $V \subseteq B_n$ and $f \in B_n$, let $r(f,V,\varepsilon) = \|B_\varrho(f,\varepsilon) \cap V\|$ denote the number of functions $g \in V$ with $\varrho(f,g) \leq \varepsilon$. Further, let $r_{\max}(V,\varepsilon) = \max_{f\in V} r(f,V,\varepsilon)$.

Lemma 2. *Suppose* $\mathrm{Gdim}_\varrho(C,P,\varepsilon) = d \geq 1$*, and let V be a non-empty subset of C. Then there exists a query $q \in Q$ such that for all $\Lambda \in P_q$ it holds that*

$$\|V \setminus \Lambda\| \geq \frac{\|V\| - r_{\max}(V,\varepsilon)}{d}.$$

Proof. Assume that for any query $q \in Q$ there exists an answer $\Lambda_q \in P_q$ for which

$$\|V \setminus \Lambda_q\| < \frac{\|V\| - r_{\max}(V, \varepsilon)}{d}.$$

Consider the answering scheme $T = \{\langle q, \Lambda_q \rangle \mid q \in Q\}$. For all $S \subseteq T_Q$ with $\|S\| \leq d$ it follows that

$$\|V \setminus \bigcap S\| = \|\bigcup_{\Lambda \in S} (V \setminus \Lambda)\| \leq \sum_{\Lambda \in S} \|V \setminus \Lambda\| < \|V\| - r_{\max}(V, \varepsilon).$$

Thus, the set $\bigcap S \cap V$ is larger than $r_{\max}(V, \varepsilon)$, which means that there is a function $f \in \bigcap S \cap V$ such that $\bigcap S \cap V \not\subseteq B_\varrho(f, \varepsilon)$. This, however, implies that $\varrho(\bigcap S \cap C) \geq \varrho(\bigcap S \cap V) > \varepsilon$, contradicting the assumption $\mathrm{Gdim}_\varrho(C, P, \varepsilon) = d$. □

In order to apply Lemma 2, we have to assume that the protocol P allows the learning algorithm to (approximately) check if a hypothesis h is already sufficiently close to the target with respect to some tolerance parameter τ. Formally, we express this assumption by requiring that P is able to simulate the protocol

$$\mathrm{SUCCESS}_\varrho(\varepsilon, \tau) = \{\langle h, B_n \setminus B_\varrho(h, \varepsilon - \tau)\rangle \mid h \in B_n\} \cup \{\langle h, B_\varrho(h, \varepsilon + \tau)\rangle \mid h \in B_n\}.$$

Intuitively, by asking a query h under the protocol $\mathrm{SUCCESS}_\varrho(\varepsilon, \tau)$, the learning algorithm A can determine whether the target is already sufficiently close to h (in case he gets the answer $B_\varrho(h, \varepsilon + \tau)$) or not (in case he gets the answer $B_n \setminus B_\varrho(h, \varepsilon - \tau)$). If $\mathrm{SUCCESS}_\varrho(\varepsilon, \tau) \leq_k P$ holds, then A can also get this information from the teacher by asking k queries under P.

Based on Lemma 2 we can bound the learning complexity by the general dimension as follows.

Lemma 3. *For any protocol P with $\mathrm{SUCCESS}_\varrho(\varepsilon, \tau) \leq_k P$ it holds that* $\mathrm{LC}_\varrho(C, P, 2(\varepsilon + \tau)) \leq 2k\mathrm{Gdim}_\varrho(C, P, \varepsilon - \tau)\lceil \ln \|C\| \rceil$.

Proof. Suppose that $\mathrm{Gdim}_\varrho(C, P, \varepsilon - \tau) = d$. We have to describe a learning algorithm such that the version space after at most $2kd\lceil \ln \|C\| \rceil$ queries is contained in the ball $B_\varrho(f, 2(\varepsilon + \tau))$ around the target $f \in C$.

Given the current version space V, we first compute $r_{\max}(V, \varepsilon - \tau)$.

- If $r_{\max}(V, \varepsilon - \tau) \leq \|V\|/2$, then we use the query of Lemma 2 to discard at least
$$\|V \setminus \Lambda\| \geq \frac{\|V\| - r_{\max}(V, \varepsilon - \tau)}{d} \geq \frac{\|V\|}{2d}$$
hypotheses from V.
- If $r_{\max}(V, \varepsilon - \tau) > \|V\|/2$, then there exists a function $h \in V$ with $r(h, V, \varepsilon - \tau) = \|B_\varrho(h, \varepsilon - \tau) \cap V\| > \|V\|/2$. Since $\mathrm{SUCCESS}_\varrho(\varepsilon, \tau) \leq_k P$, we can find k queries $q_1, \ldots, q_k$ such that the intersection of the corresponding answers $\Lambda_1, \ldots, \Lambda_k$ is either contained in $B_n - B_\varrho(h, \varepsilon - \tau)$ or in $B_\varrho(h, \varepsilon + \tau)$.

In case the intersection $\bigcap_{i=1}^{k} \Lambda_i$ is contained in $B_n - B_\varrho(h, \varepsilon - \tau)$, then the answers $\Lambda_1, \ldots, \Lambda_k$ discard more than half of V, since $\|B_\varrho(h, \varepsilon - \tau) \cap V\| > \|V\|/2$.
In case the intersection $\bigcap_{i=1}^{k} \Lambda_i$ is contained in $B_\varrho(h, \varepsilon + \tau)$, it follows that the version space after the queries $q_1, \ldots, q_k$ is contained in the ball $B_\varrho(f, 2(\varepsilon + \tau))$.

This means that by asking a set of at most k queries we can either get V within the ball $B_\varrho(f, 2(\varepsilon + \tau))$ or reduce the size of V by a fraction of at least $(1 - 1/2d)$. Thus, after at most km queries, where m satisfies the inequality

$$(1 - 1/2d)^m \|C\| \leq 1,$$

V must be contained in the ball $B_\varrho(f, 2(\varepsilon + \tau))$. A simple calculation shows that any $m \geq 2d\lceil \ln \|C\| \rceil$ fulfills the inequality. □

Combining Lemmas 1 and 3 we get the following theorem.

Theorem 1. *For any protocol P with* $\mathrm{SUCCESS}_\varrho(\varepsilon, \tau) \leq_k P$ *it holds that* $\mathrm{Gdim}_\varrho(C, P, 2(\varepsilon + \tau)) \leq \mathrm{LC}_\varrho(C, P, 2(\varepsilon + \tau)) \leq 2k\mathrm{Gdim}_\varrho(C, P, \varepsilon - \tau)\lceil \ln \|C\| \rceil$.

Next we apply Theorem 1 to the protocols for statistical and distance queries. In order to see how the protocol $\mathrm{SUCCESS}_\varrho(\varepsilon, \tau)$ can be simulated by the protocol $\mathrm{DIST}_\varrho(\tau)$, consider a distance query $h \in B_n$ and an answer $S_\varrho(h, \rho, \tau)$ to h with respect to $\mathrm{DIST}_\varrho(\tau)$. If $\rho \leq \varepsilon$, then the sphere $S_\varrho(h, \rho, \tau)$ is contained in the ball $B_\varrho(h, \varepsilon + \tau)$, and if $\rho > \varepsilon$, then the sphere $S_\varrho(h, \rho, \tau)$ is disjoint from $B_\varrho(h, \varepsilon - \tau)$, and hence contained in $B_n \setminus B_\varrho(h, \varepsilon - \tau)$. This means that $\mathrm{SUCCESS}_\varrho(\varepsilon, \tau) \leq_1 \mathrm{DIST}_\varrho(\tau)$, for any $\varepsilon \geq 0$.

Further, since $\mathrm{DIST}_\varrho(\tau) \leq_1 \mathrm{STAT}_D(\tau)$ for any metric of the form $\varrho(f, g) = \Pr_D[f \neq g]$, it also follows that $\mathrm{SUCCESS}_\varrho(\varepsilon, \tau) \leq_1 \mathrm{STAT}_D(\tau)$ for any $\varepsilon \geq 0$. Thus, as an application of Theorem 1 we get the following characterizations of the learning complexity for statistical and distance queries.

Corollary 1. *For* $P \in \{\mathrm{DIST}_\varrho(\tau), \mathrm{STAT}_D(\tau)\}$ *and all* $\varepsilon \geq 0$ *it holds that* $\mathrm{Gdim}_\varrho(C, P, 2(\varepsilon + \tau)) \leq \mathrm{LC}_\varrho(C, P, 2(\varepsilon + \tau)) \leq 2\mathrm{Gdim}_\varrho(C, P, \varepsilon - \tau) \ln\lceil \|C\| \rceil$.

4 The Learning complexity of DNF formulas

In this section we give upper und lower bounds for the learning complexity of DNF formulas with respect to distance queries and the metric $\varrho(f, g) = \Pr_U[f \neq g]$, where U denotes the uniform distribution on $\{0, 1\}^n$. The DNF-size of a boolean function f is the minimal number of terms of a DNF formula for f. We use the Fourier transform of boolean functions and basic ideas from [7].

In order to describe the Fourier transform of a boolean function f of arity n, it is convenient to think of f as a mapping from $\{0, 1\}^n$ to $\{-1, +1\}$. Then, for $A \subseteq [n]$, the parity function $\chi_A : \{0, 1\}^n \to \{-1, +1\}$ is defined as

$$\chi_A(x) = (-1)^{\sum_{i \in A} x_i}.$$

The 2^n parity functions $\{\chi_A \mid A \subseteq [n]\}$ form an orthonormal basis for the vector space of all real-valued functions on $\{0,1\}^n$, provided that the inner product is defined as $\langle f, g\rangle = \mathrm{E}[f \cdot g]$ and the norm of f is $\|f\| = \sqrt{\mathrm{E}[f^2]}$, where the expectation is taken with respect to the uniform distribution on $\{0,1\}^n$. Hence, every function $f : \{0,1\}^n \to \mathbb{R}$ can be uniquely expressed as a linear combination $f(x) = \sum_{A\subseteq[n]} \hat{f}(A)\chi_A(x)$, where $\hat{f}(A) = \langle \chi_A, f\rangle$.

A particularly useful fact is *Parseval's identity* which states that for every $f : \{0,1\}^n \to \mathbb{R}$ it holds that $\mathrm{E}[f^2] = \sum_{A\subseteq[n]} \hat{f}(A)^2$. For a boolean function $f : \{0,1\}^n \to \{-1,+1\}$, we thus get $\sum_{A\subseteq[n]} \hat{f}(A)^2 = 1$, which in particular gives us a bound on the number of coefficients exceeding a threshold $\theta > 0$,

$$\|\{A \subseteq [n] \mid \hat{f}(A) > \theta\}\| < 1/\theta^2.$$

This is useful in proving the following lemma which is inspired by a similar result in [9].

Lemma 4. *Let $C \subseteq B_n$ and suppose that $k = \|\{A \subseteq [n] \mid \chi_A \in C\}\| \geq 2$. Then, for $\varrho(f,g) = \mathrm{Pr}_U[f \neq g]$, and for every $\varepsilon < 1/2$ and $\tau \geq 0$ it holds that $\mathrm{Gdim}_\varrho(C, \mathrm{DIST}_\varrho(\tau), \varepsilon) > k\tau^2$.*

Proof. Suppose that $\mathrm{Gdim}_\varrho(C, \mathrm{DIST}_\varrho(\tau), \varepsilon) = d$, let $K = \{A \subseteq [n] \mid \chi_A \in C\}$, and consider the answering scheme $T = \{\langle h, \Lambda_h\rangle \mid h \in B_n\}$, where $\Lambda_h = \{f \mid |\varrho(f,h) - 1/2| \leq \tau\}$. Note that $\mathrm{E}[fg] = 2\Pr[f = g] - 1$ and therefore $\varrho(f,g) - 1/2 = -\langle f, g\rangle/2$. By Parseval's identity, we can bound the number of parity functions that can be eliminated from K by an answer Λ_h by

$$\begin{aligned}
\|\{A \in K \mid \chi_A \notin \Lambda_h\}\| &= \|\{A \in K \mid |\varrho(h,\chi_A) - 1/2| > \tau\}\| \\
&= \|\{A \in K \mid |\langle h, \chi_A\rangle| > 2\tau\}\| \\
&\leq \|\{A \subseteq [n] \mid |\hat{h}(A)| > 2\tau\}\| \\
&< 1/4\tau^2.
\end{aligned}$$

It follows that for any set $S \subseteq T_Q$ of cardinality at most $d \leq k\tau^2$,

$$\begin{aligned}
\|\{A \in K \mid (\forall \Lambda_h \in S)[\chi_A \in \Lambda_h]\}\| &= k - \|\{A \in K \mid (\exists \Lambda_h \in S)[\chi_A \notin \Lambda_h]\}\| \\
&> k - d/4\tau^2 \\
&\geq k - k/4 \\
&> 1.
\end{aligned}$$

Since $\varepsilon < 1/2$, and since the distance between two parity functions is precisely $1/2$, it follows that $\mathrm{Gdim}_\varrho(C, P, \varepsilon) > k\tau^2$. □

Every parity function χ_A with $\|A\| \leq \log s$ can be represented by a DNF formula with at most s terms. Together with Lemma 4 we get that for every $\varepsilon < 1/2$,

$$\mathrm{LC}_\varrho(DNF_{s,n}, \mathrm{DIST}_\varrho(\tau)) > \binom{n}{\log s}\tau^2 = n^{\Omega(\log s)}\tau^2,$$

where $DNF_{s,n}$ is the class of n-ary boolean functions that can be represented by a DNF formula with at most s terms. Recall that for a real-valued function f over $\{0,1\}^n$, the norm $L_\infty(f)$ is defined as $\max_{x\in\{0,1\}^n} |f(x)|$.

Theorem 2. *For $\varrho(f,g) = \Pr_U[f \neq g]$, for all $\varepsilon > 0$, s and n, and for $\tau = \Omega(\varepsilon^2/s)$,*

$$\mathrm{LC}_\varrho(DNF_{s,n}, \mathrm{DIST}_\varrho(\tau)) = n^{O(\log(s/\varepsilon^2))}.$$

Proof. The proof is based on the fact that every boolean function f of DNF-size s has a large coefficient. In particular, for every Boolean function f of DNF-size s, and for every distribution D on $\{0,1\}^n$ there exists a set $A \subseteq [n]$ of cardinality at most $\log((2s+1)L_\infty(2^n D))$ such that $|\mathrm{E}_D[f\chi_A]| \geq 1/(2s+1)$ [7, 13, 9].

It follows that for any distribution D on $\{0,1\}^n$, and for every target f of DNF-size s, a simple learning algorithm under the protocol $\mathrm{DIST}_\varrho(\tau)$ with respect to $\varrho(f,g) = \Pr_D[f \neq g] = (1 - \mathrm{E}_D[fg])/2$ can do an exhaustive search to produce a hypothesis $h \in \{\chi_A, -\chi_A\}$, for some A of cardinality at most $\log((2s+1)L_\infty(2^n D))$, such that h satisfies $\mathrm{E}_D[fh] \geq 1/(2s+1) - 4\tau$. Hence, $\Pr_D[f = h] \geq 1/2 + 1/(4s+2) - 2\tau$, and the number of queries required is

$$\binom{n}{\log((2s+1)L_\infty(2^n D))}.$$

Now we can invoke known boosting techniques due to Freund [10] which, as shown by Aslam and Decatur [3], also can be applied to statistical queries as follows. Suppose that there is a learning algorithm WL, such that for every distribution D, and for all $f \in C$, WL uses N_0 queries under the protocol $\mathrm{STAT}_D(\tau_0)$ with respect to $\varrho(f,g) = \Pr_D[f \neq g]$ to output a hypothesis h such that $\Pr_D[h = f] \geq 1/2 + \Omega(1/s)$. Then there is an algorithm B such that for every distribution D, every $f \in C$, and for every $\varepsilon > 0$, B under the protocol $\mathrm{STAT}_D(\tau)$ uses $N = O(N_0\tau^{-4}\log^2(\varepsilon^{-1}))$ queries to output a hypothesis h such that $\Pr_D[h = f] \geq 1-\varepsilon$. The required tolerance is $\tau = \Omega(\tau_0\varepsilon^2)$, and B simulates WL with respect to distributions D_i with $L_\infty(D_i) \leq O(1/\varepsilon^2)L_\infty(D)$.

Since $\mathrm{DIST}_\varrho(\tau) \leq_1 \mathrm{STAT}_D(\tau) \leq_2 \mathrm{DIST}_\varrho(\tau)$, for any distribution D and metric $\varrho(f,g) = \Pr_D[f \neq g]$, this construction can also be applied to the above exhaustive-search algorithm for learning DNF formulas with distance queries. This yields for the uniform distribution a learning algorithm B that for all f of DNF-size s and every $\varepsilon > 0$ uses

$$\binom{n}{O(\log(s/\varepsilon^2))} s^4 \log^2(1/\varepsilon) = n^{O(\log(s/\varepsilon^2))}$$

many queries under the protocol $\mathrm{DIST}_\varrho(\tau)$, where $\tau = \Omega(\varepsilon^2/s)$, and outputs a hypothesis h such that $\Pr_U[h \neq f] \leq \varepsilon$.

Finally, we have to take into account that in our definition of learning complexity, we require that the final version space (consisting of all functions in C not yet discarded by the answers received so far) is contained in the ball $B_\varrho(f,\varepsilon)$. It is clear that the version space of a learning algorithm A that for every $f \in C$

produces a hypothesis h with distance at most ε from f, must be contained in the ball $B_\varrho(f, 2\varepsilon)$ when A outputs the hypothesis (otherwise, there would be a function g left in the final version space whose distance from f is more than 2ε; but since there is no (boolean) function whose distance to both f and g is at most ε, such a function g could be used to fool the algorithm). Thus, the theorem follows by replacing ε by $\varepsilon/2$. □

By combining the upper bound of 2 with the lower bound that can be derived from Lemma 4 and from Corollary 1, we get the following bound on the number of queries needed to approximately learn DNF formulas in the statistical query model as well as in the model of learning by distance.

Corollary 2. *Fix some constant* $0 < \varepsilon_0 < 1/2$. *Then, for all* s *and* n, *there is some* $\tau = \Theta(1/s)$ *such that for* $\varrho(f, g) = \Pr_U[f \neq g]$ *and* $P \in \{\mathrm{DIST}_\varrho(\tau), \mathrm{STAT}_U(\tau)\}$,

$$LC(DNF_{s,n}, P, \varepsilon_0) = n^{\Theta(\log s)}.$$

5 Acknowledgements

For helpful conversations and suggestions on this work we are very grateful to José Balcázar, Jorge Castro, and Richard Gavaldà.

References

[1] D. Angluin. Queries and concept learning. *Machine Learning*, 2:319–342, 1988.
[2] D. Angluin. Negative results for equivalence queries. *Machine Learning*, 5(2):121–150, 1990.
[3] J. Aslam and S. Decatur. General bounds on statistical query learning and PAC learning with noise via hypothesis boosting. *Information and Computation*, 141(2):85–118, 1998.
[4] J. Balcázar, J. Castro, and D. Guijarro. A general dimension for exact learning. In *Proc. 14th ACM Conference on Computational Learning Theory*, volume 2111 of *Lecture Notes in Artificial Intelligence*, pages 354–367. Springer-Verlag, Berlin Heidelberg New York, 2001.
[5] J. L. Balcázar, J. Castro, D. Guijarro, and H. U. Simon. The consistency dimension and distribution-dependent learning from queries. In *Proc. 10th International Workshop on Algorithmic Learning Theory (ALT'99)*, Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin Heidelberg New York, 1999.
[6] S. Ben-David, A. Itai, and E. Kushilevitz. Learning by distances. *Information and Computation*, pages 240–250, 1995.
[7] A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, and S. Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In ACM, editor, *Proc. 26thACM Symposium on Theory of Computing*, pages 253–262. ACM Press, 1994.
[8] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[9] N. H. Bshouty and V. Feldman. On using extended statistical queries to avoid membership queries. *Journal of Machine Learning Research*, 2:359–395, 2002.

[10] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285, Sept. 1995.

[11] T. Hegedüs. Generalized teaching dimensions and the query complexity of learning. In *Proceedings of the 8th Annual Conference on Computational Learning Theory (COLT'95)*, pages 108–117. ACM Press, July 1995.

[12] L. Hellerstein, K. Pillaipakkamnatt, V. Raghavan, and D. Wilkins. How many queries are needed to learn? *Journal of the ACM*, 43(5):840–862, 1996.

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

[14] M. J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.