Characterizing Statistical Query Learning: Simplified Notions and Proofs*

Balázs Szörényi

[1] Fakultät für Mathematik, Ruhr-Universität Bochum, D-44780 Bochum, Germany
[2] Hungarian Academy of Sciences and University of Szeged, Research Group on Artificial Intelligence, H-6720 Szeged
szorenyi@inf.u-szeged.hu

Abstract. The Statistical Query model was introduced in [6] to handle noise in the well-known PAC model. In this model the learner gains information about the target concept by asking for various statistics about it. Characterizing the number of queries required by learning a given concept class under *fixed distribution* was already considered in [3] for weak learning; then in [8] strong learnability was also characterized. However, the proofs for these results in [3, 10, 8] (and for strong learnability even the characterization itself) are rather complex; our main goal is to present a simple approach that works for both problems. Additionally, we strengthen the result on strong learnability by showing that a class is learnable with polynomially many queries iff *all* consistent algorithms use polynomially many queries, and by showing that proper and improper learning are basically equivalent. As an example, we apply our results on conjunctions under the uniform distribution.

1 Introduction

The *Statistical Query model* (called SQ model for short) was introduced by Kearns [6] as an approach to handle noise in the well-known PAC model. The general idea is that—instead of using random examples as in the PAC model—the learner gains information about the unknown function by asking various statistics (called *queries*) over the distribution of labeled examples. As it was shown by Kearns [6], any learning algorithm in the SQ model can be transformed to a PAC algorithm without much loss in efficiency. It is even more interesting that the resulting algorithm is robust to noise. He has also shown that many efficient PAC algorithms can also be converted to an efficient SQ algorithm.

Despite the power of the model that is apparent from the above results, it is still weaker than the PAC model. Indeed, already in [6] it was shown that the parities, which is a PAC-learnable class, cannot be efficiently learned in the SQ model under the uniform distribution. The proof used an information theoretic

* This work was supported in part by the Deutsche Forschungsgemeinschaft Grant SI 498/8-1, and the NKTH grant of the National Technology Programme 2008 (project codename AALAMSRK NTP OM-00192/2008) of the Hungarian government..

argument, which was generalized later by Blum et al. in [3] to characterize *weak learnability* of a concept class (where the goal is to do slightly better than random guessing) in the SQ model for the *distribution dependent* case (i.e., when the underlying distribution is fixed in advance and is known by the learner). The characterization is based on the so called *SQ dimension* of the class which is, roughly, the maximal size of an "almost orthogonal" system in the class. However, the proof in [3] is rather long and complex. Subsequently Yang gave an alternative, elegant proof for basically the same result [10]. In this paper we present yet another, but much shorter proof, thereby significantly simplifying on both existing proofs.

Strong learnability (i.e., when the goal is to approximate the target concept with arbitrary accuracy) of a concept class in the distribution dependent case was first characterized by Köbler and Lindner [7] in terms of a general framework for protocols, called the *general dimension*. Independently Simon in [8] gave another characterization for strong learnability that was based on the SQ dimension (more precisely it was based on the SQ dimension of the class after some translation), and is more of an algebraic flavor. However, both the characterization and the proof are rather complex in [8]; as we shall show in this paper, our simple approach that is successful in characterizing weak learnability, can be also applied for strong learnability, thereby giving an alternative, simple characterization for this problem as well, which might also have the potential to be easier to apply and calculate for concrete concept classes. Recently Feldman has also obtained a simple characterization of strong SQ learnability of a similar flavor [5], however the two papers focus on different perspectives: Feldman is interested in applications to agnostic learning and evolvability, meanwhile our main interest is to find a really simple proof and a unified view of weak and strong learnability. Additionally our approach also reveals that in the distribution dependent case query-efficient learnability is possible if and only if *all* consistent learning algorithms learn the given concept class query-efficiently.[3] As far as we know, this was not known before. We also show that in the distribution dependent case proper learning (i.e., when the queries of the learner are restricted to use functions from the given concept class) is as strong as improper learning, but we would like to point out that this can be easily deduced already from the characterization result of Simon (see Observation 11).

Finally we show that in the *distribution independent* case (i.e., when the learner doesn't know anything about the underlying distribution) proper and improper learning can differ significantly, and we contrast this with the above mentioned result on their equivalence in the distribution dependent case.

Equivalent models. Ben-David et al. have introduced an equivalent model, called *learning by distances* [2], and have also given upper and lower bound on the minimal number of queries required for learning. However their upper bound

[3] Query-efficiency means that the number of queries used by the learner is bounded by some polynomial of the various parameters. When query-efficiency is in focus, then usually no restrictions are set on the running time.

is exponential in their lower bound (see also our discussion on the topic in Sect. 6) and the paper does not reveal the relation of the model to noise-tolerant PAC learning (which gave the importance of the SQ model).

In [11] Yang has introduced the model of *honest SQ model* using stronger queries and less adversarial settings than the ones used in the SQ model. In [9] it is shown how to apply the results and methods of this paper to prove a somewhat surprising result: the equivalence of the honest and the "pure" SQ model.

Organization of the paper. Section 2 contains the formal introduction of the SQ model and also some basic definitions. In Sect. 3 we present our alternative proof for characterizing weak learnability with the SQ dimension, in Sect. 4 we discuss the relation of strong and weak learnability, and then in Sect. 5 we characterize strong learnability. In Sect. 6 we analyze the relation of our strong SQ dimension to the ones of Simon and Feldman. In Sect. 7 as an example, we compute our dimension notion for conjunctions under the uniform distribution. Finally, in Sec. 8 we contrast the result on the equivalence of proper and improper learning in the distribution dependent case with the fact that they occasionally significantly differ in the distribution independent case.

2 Preliminaries

A *concept* is a mapping from some domain to $\{-1, 1\}$. A *concept class* is a set of concepts with the same domain. A *Boolean concept* over n variables is a concept of the form $\{-1, 1\}^n \to \{-1, 1\}$. A *family of concept classes* is an infinite set $\{\mathcal{F}_n\}_{n=1}^{\infty}$, such that each $\mathcal{F}_n$ is a concept class. The class of all concepts over some domain X is denoted $\mathcal{C}(X)$.

The *correlation* of two functions $f, g : X \to \mathbb{R}$ under some distribution $\mathcal{D}$ over X is defined as $\langle f, g \rangle_{\mathcal{D}} = \mathbb{E}[f(\xi)g(\xi)]$, where ξ is a random variable with distribution $\mathcal{D}$. The *norm* of f under $\mathcal{D}$ is $\|f\|_{\mathcal{D}} := \sqrt{\langle f, f \rangle_{\mathcal{D}}}$. f is said to be a γ-*approximation* of g, if $\langle f, g \rangle_{\mathcal{D}} \geq \gamma$.

In the *Statistical Query model* a learner (or learning algorithm) L can make *queries* of the form (h, τ), where τ is a positive constant called *tolerance*, and h is chosen from some concept class $\mathcal{H}$ called the *query class*. Each such query is answered with some c satisfying $|c - \langle f^*, h \rangle_{\mathcal{D}}| \leq \tau$, where f^* is some fixed concept, called the *target concept* that is unknown for the learner, and where $\mathcal{D}$ is some distribution over the input domain of f^*. (Here the learner is supposed to be familiar with $\mathcal{D}$.) The learner *succeeds* when he finds some function $f \in \mathcal{H}$ having correlation at least γ with f^* for some constant $\gamma > 0$ fixed ahead of the learning process. Parameter γ is called *accuracy*. Let $q_{\mathcal{F},\mathcal{H}}^{\mathcal{D},L}(\tau, \gamma)$ denote the smallest integer q such that L always succeeds in the above setting using at most q queries when the target concept belongs to $\mathcal{F}$. Finally, $\mathrm{SLC}_{\mathcal{F},\mathcal{H}}^{\mathcal{D}}(\tau, \gamma)$ (or the *statistical learning complexity*) is defined to be the minimum value of $q_{\mathcal{F},\mathcal{H}}^{\mathcal{D},L}(\tau, \gamma)$ over all possible learning algorithms L. We would like to emphasize that in this

paper we are interested only in the number of queries during the learning process (i.e., the *information complexity* of learning), and do not consider the running time.

Note that originally in [6] the SQ model allowed much more general queries, but in [4] Bshouty and Feldman have shown that the two models are equivalent.[4]

We also consider the following variants of the above described learning model. The learning is called *proper* when $\mathcal{F} = \mathcal{H}$, and is called *improper* when $\mathcal{F} \subsetneq \mathcal{H}$. Also, in general, a query (h, τ) is *proper* if $h \in \mathcal{F}$, otherwise it is *improper*. The learner is a *consistent learner*, if $|\langle h_i, h_j \rangle_{\mathcal{D}} - c_i| \leq \tau_i$ for $i < j$, where (h_i, τ_i) is the i-th query of the learner and c_i is the answer for it. Finally, note that in the above definition the learner is supposed to be familiar with the underlying distribution, but the model can also be defined for the case when this is not true. We are mainly interested in the former case (except for Sect. 8), but when we want to explicitly refer to one case or the other, we shall call the former the *distribution dependent* and the latter the *distribution independent* case.

For simplicity, when it causes no confusion, we omit $\mathcal{D}$ from notations like $\mathrm{SLC}^{\mathcal{D}}_{\mathcal{F},H}(\tau, \gamma)$ and $\langle f, g \rangle_{\mathcal{D}}$, and simply use $\mathrm{SLC}_{\mathcal{F},H}(\tau, \gamma)$ and $\langle f, g \rangle$ instead.

Definition 1. *We say that a family $\{\mathcal{F}_n\}_{n=1}^{\infty}$ of concept classes is* weakly learnable *in the SQ model with a family $\{\mathcal{H}_n\}_{n=1}^{\infty}$ of query classes if there exists some $\gamma(n) > 0$ and $\tau(n) > 0$ such that $1/\gamma(n)$, $1/\tau(n)$ and $\mathrm{SLC}_{\mathcal{F}_n,\mathcal{H}_n}(\tau(n), \gamma(n))$ are polynomially bounded in n. $\{\mathcal{F}_n\}_{n=1}^{\infty}$ is* strongly learnable *in the SQ model with queries from $\{\mathcal{H}_n\}_{n=1}^{\infty}$ if there exists some $\tau(n, \epsilon) > 0$ such that $1/\tau(n, \epsilon)$ and $\mathrm{SLC}_{\mathcal{F}_n,\mathcal{H}_n}(\tau(n), 1 - \epsilon)$ are polynomially bounded in n and $1/\epsilon$.*

The following Observation, which we shall apply several times later, is basically the reason for the equivalence of the proper and improper learning in the distribution dependent model.

Observation 2. Let f, g and h be arbitrary concepts. If $\langle f, h \rangle \geq 1 - \epsilon$ and $\langle g, h \rangle \geq 1 - \epsilon$, then $\langle f, g \rangle = (1/2)\langle f + g, f + g \rangle - 1 \geq \langle f + g, h \rangle - 1 \geq 1 - 2\epsilon$.

Although this paper mainly considers concepts and concept classes, we would like to point out that all the results remain valid for classes of functions with norm bounded by 1 (which might be tempting to use for example in query classes)—albeit in some cases, when the proof applies Observation 2, the constants get slightly worse.[5] The reason for this is the following lemma which is the generalization of Observation 2 for these functions.

[4] Actually they have shown how to simulate an arbitrary statistical query using two statistical queries that are independent of the target function and two correlation queries. However, when running time is not considered and the underlying distribution is known, one can omit the two former queries and just compute them directly.

[5] The choice of 1 as an upper bound for the query function is arbitrary, one can use any other constant instead. (But note that smaller constants would exclude all concepts.) However, unbounded queries should not be allowed, because they make the learning problem trivial. Indeed, for example when the target concept is Boolean over n variables, and one uses a query with tolerance $1/2$ and with the function that

Proposition 3. *When $f, g, h : \{-1,1\}^n \to \{-1,1\}$ have norm at most 1, and $\langle f,h\rangle \geq 1-\epsilon$ and $\langle g,h\rangle \geq 1-\epsilon$, then $\langle f,g\rangle \geq 1-6\epsilon$.*

Proof. First of all, by Cauchy-Schwarz, $\|f\| \geq \langle f,h\rangle \geq 1-\epsilon$, and similarly $\|g\| \geq 1-\epsilon$. Using this

$$\begin{aligned} 2(1-2\epsilon) - 2\langle f,g\rangle &\leq \|f\|^2 + \|g\|^2 - 2\langle f,g\rangle \\ &= \|f-g\|^2 \\ &\leq (\|f-h\| + \|g-h\|)^2 \\ &\leq \left(\sqrt{2-2\langle f,h\rangle} + \sqrt{2-2\langle g,h\rangle}\right)^2 \\ &\leq 8\epsilon \ , \end{aligned}$$

implying $1-6\epsilon \leq \langle f,g\rangle$. □

Finally for integer d let $[d]$ denote the set $\{1,\ldots,d\}$.

3 Characterizing Weak Learnability

According to the definition, weak learnability is possible if and only if there exists some polynomial $p(n)$ such that $\mathrm{SLC}_{\mathcal{F}_n,\mathcal{H}_n}(1/p(n), 3/p(n)) \leq p(n)$ (simply define $p(n)$ to be a polynomial that upper bounds $1/\tau(n)$, $3/\gamma(n)$ and $\mathrm{SLC}_{\mathcal{F}_n,\mathcal{H}_n}(\tau(n),\gamma(n))$). This way the task of weak learning is basically to find functions $h_{n,1},\ldots,h_{n,p(n)} \in \mathcal{H}_n$ such that all $f \in \mathcal{F}_n$ has correlation at least $3/p(n)$ with at least one of $h_{n,1},\ldots,h_{n,p(n)}$. Thus $p(n)$ (and this way SLC itself) can be considered as a kind of covering number. Bshouty and Feldman in [4] make this property explicit in their characterization of weak learnability.

On the other hand, the notion of SQ dimension introduced by Blum et al. [3] is rather a packing number in nature:

Definition 4. *The* SQ dimension *(or* weak SQ dimension*) of a class of real valued functions $\mathcal{F}$ over some domain X and under distribution $\mathcal{D}$ over X, denoted* $\mathrm{SQDim}^{\mathcal{D}}_{\mathcal{F}}$*, is the biggest integer d such that $\mathcal{F}$ contains some distinct functions $f_1,\ldots f_d$ with pairwise correlations between $-1/d$ and $1/d$.*

(Note that SQDim is defined not only for concept classes but also for more general classes; Definition 10 will really make use of this generality.) For simplicity, as mentioned, we use $\mathrm{SQDim}_{\mathcal{F}}$ instead of $\mathrm{SQDim}^{\mathcal{D}}_{\mathcal{F}}$ when this leads to no confusion.

The nice feature of the characterization result in [3] is that it binds the two different type of notions. One direction, namely that $\mathrm{SQDim}_{\mathcal{F}}$ queries are

evaluates $x \in \{-1,1\}^n$ to $\sum_{i=1}^n (1/\epsilon)\cdot 2^n \cdot 2^{i(x_i+1)/2}$, then the value of the target concept on inputs with probability at least $\epsilon/2^n$ can be reconstructed from the answer to this query, meanwhile the sum of the probabilities of the rest of the inputs is less than ϵ.

enough for weakly learning concept class $\mathcal{F}$ (properly!) is easy: if $\{f_1, \ldots, f_d\}$ is a maximal subset of $\mathcal{F}$ fulfilling $|\langle f_i, f_j \rangle| \leq 1/d$ for $1 \leq i < j \leq d$, then (due to the maximality) it obviously holds that at least one of them has correlation at least $1/d$ with the target concept, thus the learner simply needs to query $f_1, \ldots, f_d$ with tolerance $1/(3d)$ in order to find an $1/(3d)$ approximation of it. However the proof in [3] for the other direction was rather long and complex. Subsequently Yang in [10] gave another, elegant proof for this direction, based on the eigenvalues of the correlation matrix of the concept class.[6]

Here we show that basically the same result can be derived using a very simple argument, thus significantly simplifying on both of the above mentioned proofs. The proof in some sense follows the same line of thought they use, but lacks the machineries applied in them.

Theorem 5. *Let $\mathcal{F}$ be a concept class and let $d := \mathrm{SQDim}_{\mathcal{F}}$. Then any learning algorithm that uses tolerance parameter lower bounded by $\tau > 0$ requires in the worst case at least $(d\tau^2 - 1)/2$ queries for learning $\mathcal{F}$ with accuracy at least τ. In particular, when $\tau = 1/\sqrt[3]{d}$, this means $(\sqrt[3]{d} - 1)/2$ queries.*

Proof. Assume that $f_1, \ldots, f_d \in \mathcal{F}$ fulfill $|\langle f_i, f_j \rangle| \leq 1/d$ for $i, j \in [d]$ distinct. We show an (adversary) answering strategy that ensures to eliminate only a small number of these functions after each query. Let h be an arbitrary query function used by the learner (having thus norm at most 1) and let $A := \{i \in [d] : \langle f_i, h \rangle \geq \tau\}$. Then, by the Cauchy-Schwarz inequality

$$\left\langle h, \sum_{i \in A} f_i \right\rangle^2 \leq \left\| \sum_{i \in A} f_i \right\|^2 = \sum_{i,j \in A} \langle f_i, f_j \rangle \leq \sum_{i \in A} \left(1 + \frac{|A| - 1}{d} \right) \leq |A| + \frac{|A|^2}{d} \ ,$$

meanwhile, by the choice of A it holds that $\left\langle h, \sum_{i \in A} f_i \right\rangle \geq |A|\tau$, and the two together implies that $1/|A| + 1/d \geq \tau^2$ or equivalently, that $|A| \leq d/(d\tau^2 - 1)$. Similar argument shows that at most $d/(d\tau^2 - 1)$ of the f_i functions have correlation at most $-\tau$ with h. Thus at most $2d/(d\tau^2 - 1)$ of the functions will be inconsistent with the answer if the adversary returns 0 to this query. This, in turn, implies the desired lower bound $(d\tau^2 - 1)/2$ on the learning complexity. □

It is also worth mentioning that this result is quite tight in the improper case, when the learner can use arbitrary functions of norm 1 in the queries. Indeed, if the concept class itself is $\{f_1, \ldots, f_d\}$, then defining $g_i := \sum_{j=i \cdot d^{2/3}+1}^{(i+1) \cdot d^{2/3}} f_j$ for $i = 0, 1, \ldots, d^{1/3} - 1$ (assuming for simplicity that $\sqrt[3]{d}$ is integer), at least one $h_i = g_i / \|g_i\|$, $i = 0, 1, \ldots, d^{1/3} - 1$ will have correlation at least

$$\left(1 - d^{2/3} \frac{1}{d} \right) \frac{1}{\sqrt{d^{2/3} + d^{2/3} d^{2/3} (1/d)}} \tag{1}$$

with the target function. Note that (1) asymptotically equals to $1/\sqrt[3]{d}$.

[6] The correlation matrix of the concept class $\mathcal{F} = \{f_1, \ldots, f_s\}$ is the $s \times s$ matrix C such that $C_{i,j} = \langle f_i, f_j \rangle$.

4 Weak and Strong Learning

Aslam and Decatur [1] apply the boosting techniques from the PAC model to SQ learning and show how to use (efficiently) a weak learning algorithm to achieve strong learnability. Their primary concern is the distribution independent case, but their result (combined with results for weak learning) also has the following consequence in the distribution dependent case:

$$\max_{\mathcal{D}} \mathrm{SLC}^{\mathcal{D}}_{\mathcal{F},\mathcal{H}}\left(\frac{\epsilon}{3d}\log\frac{1}{\epsilon}\ ,\ 1-\epsilon\right) = O\left(d^5\log^2\frac{1}{\epsilon}\right)\ ,$$

when $\mathcal{H} \supseteq \mathcal{F}$, and where $d = \max_{\mathcal{D}} \mathrm{SQDim}^{\mathcal{D}}_{\mathcal{F}}$. However, this does not imply any result on fixed distributions in general. Indeed, when the support of a distribution consists of only a single input, then one query is enough both in the weak and in the strong setting—for *any* concept class. Thus the gap between the upper bound in the above equation and the number of queries required for strong learning under some given (known) distribution can be as big as possible: exponential versus constant. What is more, we cannot expect to bound the strong SQ dimension under some distribution $\mathcal{D}$ using the weak SQ dimension under the same distribution. Indeed, consider for example the uniform distribution and the concept class $\mathcal{F}_n$ consisting of all the functions of the form $v_1 \vee f$, where f is any parity function over variables $v_2, \ldots, v_n$. Then $|\mathcal{F}_n| = 2^{n-1}$, and any two distinct elements $(v_1 \vee f), (v_1 \vee f') \in \mathcal{F}_n$ have correlation $1/2$:

$$\langle v_1 \vee f, v_1 \vee f' \rangle = 2\,\mathbb{P}\left[(v_1 \vee f) = (v_1 \vee f')\right] - 1 = 2\left(\frac{1}{2} + \frac{1}{2}\,\mathbb{P}[f = f']\right) - 1 = \frac{1}{2}$$

(as the parity functions are uncorrelated under the uniform distribution), and so by Theorem 8 strong learning of $\mathcal{F}_n$ requires superpolynomial number of queries, meanwhile weak learning requires none.[7]

5 Characterizing Strong Learnability

In this section we give a complete characterization of strong learnability. More precisely we define a dimension notion that is a generalization of the weak SQ dimension SQDim from Sect. 3, and show that it is closely related to the learning complexity.

Definition 6. *For a concept class $\mathcal{F}$ let $d_0(\mathcal{F}, \gamma)$ denote the largest d such that some $f_1, \ldots, f_d \in \mathcal{F}$ fulfill*

- $|\langle f_i, f_j \rangle| \leq \gamma$ *for* $1 \leq i < j \leq d$, *and*
- $|\langle f_i, f_j \rangle - \langle f_k, f_\ell \rangle| \leq 1/d$ *for all* $1 \leq i < j \leq d$ *and* $1 \leq k < \ell \leq d$.

[7] Yang [11] has also shown a similar result for another concept class, but the argument there is more complicated.

Actually, this dimension notion is a kind of combination of the strong SQ dimension of Simon [8] (see also Sect. 6) and Yang [10].

Theorem 7. *Let $\mathcal{F}$ be a concept class and let $d := d_0(\mathcal{F}, 1-\epsilon/2)$. Then any consistent algorithm that uses tolerance $\tau \leq \min\{1/(4d+4), \epsilon/4\}$ requires at most d/τ queries to learn $\mathcal{F}$ with accuracy $1-\epsilon$. Specifically, setting $\tau = \min\{1/(4d+4), \epsilon/4\}$, the algorithm finds an $(1-\epsilon)$-approximation of the target concept after $4d \cdot \max\{d+1, 1/\epsilon\}$ queries, implying $\mathrm{SLC}_{\mathcal{F},\mathcal{F}}(\tau, 1-\epsilon) \leq 4d \cdot \max\{d+1, 1/\epsilon\}$.*

Proof. Assume that some consistent algorithm used tolerance as above, queried $h_1, \ldots, h_q$ in this order, and got the answers $c_1, \ldots, c_q$ in this order. Suppose that for some $1 \leq i_1 < i_2 < \cdots < i_\ell \leq q$ and some $c \in [-1, 1]$ it holds that $c_{i_j} \in [c-\tau, c+\tau]$ for $j = 1, \ldots, \ell$. The algorithm is consistent, thus $\langle h_{i_j}, h_{i_k} \rangle \in [c_{i_j} - \tau, c_{i_j} + \tau]$ for $1 \leq j < k \leq \ell$, consequently $\langle h_{i_j}, h_{i_k} \rangle \in [c-2\tau, c+2\tau] \subseteq [c - 1/(2d+2), c + 1/(2d+2)]$ for $1 \leq j < k \leq \ell$. Also note that $|\langle h_i, h_j \rangle| \leq |c_i| + \tau \leq 1 - \epsilon/2$ for $1 \leq i < j \leq q$, since $c_1, \ldots, c_q$ have absolute value less then $1 - 3\epsilon/4$ (as otherwise the algorithm would have successfully terminated). The two together imply however that $\ell \leq d_0(\mathcal{F}, 1-\epsilon/2)$. As this was true for any c, it follows that $q \leq d_0(\mathcal{F}, 1-\epsilon/2)(2/(2\tau))$. □

The proof for the other direction has the same structure as the proof for Theorem 5, with some necessary modifications.

Theorem 8. *Let $\mathcal{F} \subseteq \mathcal{C}(X)$ be any concept class for some domain X, and assume $d := d_0(\mathcal{F}, 1-2\epsilon) \geq 3$. Then if the tolerance τ is bigger than $\sqrt{3/(2\lfloor d/2 \rfloor)}$, then $\mathrm{SLC}_{\mathcal{F},\mathcal{C}(X)}(\tau, 1-\epsilon) \geq d\tau^2/3$. In particular $\mathrm{SLC}_{\mathcal{F},\mathcal{C}(X)}(1/\sqrt[3]{d}, 1-\epsilon) \geq \sqrt[3]{d}/3$.*

Proof. Assume $3/(2\tau^2) \leq \lfloor d/2 \rfloor$ and let $d' := \lceil 3/(2\tau^2) \rceil$. By the choice of d there exist $f_1, \ldots, f_d \in \mathcal{F}$ and $c \in (-1+2\epsilon, 1-2\epsilon)$ satisfying $|\langle f_i, f_j \rangle - c| \leq 1/(2d)$ for all $1 \leq i < j \leq d$. We show an (adversary) answering strategy that ensures to eliminate only a small number of the f_i functions after each query. Let $h \in \mathcal{C}(X)$ be an arbitrary query function used by the learner, and assume for simplicity that $\langle f_1, h \rangle \geq \langle f_2, h \rangle \geq \cdots \geq \langle f_d, h \rangle$. Define $\alpha := \langle f_{d'}, h \rangle$, $\beta := \langle f_{d-d'+1}, h \rangle$, $A := [d']$ and $B := \{d-d'+1, d-d'+2, \ldots, d\}$. Then $1 - \epsilon \geq \alpha \geq \beta \geq -1 + \epsilon$ whenever $d \geq 3$ (recall Observation 2 and that $d' \leq d/2$ by our assumption on τ), furthermore A and B are disjoint sets of cardinality d'. First note that

$$
\begin{aligned}
&\left\| \frac{1}{d'} \sum_{i \in A} f_i - \frac{1}{d'} \sum_{i \in B} f_i \right\|^2 \\
&= \frac{1}{(d')^2} \left(\sum_{i \in A} \|f_i\|^2 + \sum_{i \in B} \|f_i\|^2 + \sum_{i,j \in A : i \neq j} \langle f_i, f_j \rangle \right. \\
&\qquad \left. + \sum_{i,j \in B : i \neq j} \langle f_i, f_j \rangle - 2 \sum_{i \in A} \sum_{j \in B} \langle f_i, f_j \rangle \right) \\
&\leq \frac{1}{(d')^2} \left(2d' + d'(d'-1) \left(c + \frac{1}{2d} \right) + d'(d'-1) \left(c + \frac{1}{2d} \right) - 2(d')^2 \left(c - \frac{1}{2d} \right) \right)
\end{aligned}
$$

$$\leq \frac{4}{d'} + \frac{2}{d} \ ,$$

and so, by the Cauchy-Schwarz inequality

$$\left\langle h,\ \frac{1}{d'}\sum_{i\in A} f_i - \frac{1}{d'}\sum_{i\in B} f_i \right\rangle \leq \left\| \frac{1}{d'}\sum_{i\in A} f_i - \frac{1}{d'}\sum_{i\in B} f_i \right\| \leq \sqrt{\frac{4}{d'} + \frac{2}{d}} \leq \sqrt{\frac{6}{d'}} \ .$$

On the other hand, by the definition of A and B it also holds that

$$\left\langle h,\ \frac{1}{d'}\sum_{i\in A} f_i - \frac{1}{d'}\sum_{i\in B} f_i \right\rangle = \frac{1}{d'}\sum_{i\in A} \langle h, f_i\rangle - \frac{1}{d'}\sum_{j\in B} \langle h, f_j\rangle \geq \alpha - \beta \ ,$$

and so $\alpha - \beta \leq \sqrt{6/d'} \leq 2\tau$. Thus, answering the learner's query with $(\alpha+\beta)/2$, all but at most $2d'-2$ functions will be consistent with the answer. This, in turn, implies the desired lower bound $d/(2d'-2) \geq d\tau^2/3$ on the learning complexity. □

The main result of this section is the following corollary of the two theorems above:

Corollary 9. *The following statements are equivalent for any family $\{\mathcal{F}_n\}_{n=1}^{\infty}$ of concept classes under arbitrary (fixed) distribution:*

- *$d_0(\mathcal{F}_n, 1-\epsilon)$ is polynomially bounded in n and $1/\epsilon$,*
- *$\{\mathcal{F}_n\}_{n=1}^{\infty}$ is strongly learnable by some (possibly improper) algorithm,*
- *$\{\mathcal{F}_n\}_{n=1}^{\infty}$ is strongly learnable by* all *consistent learning algorithms.*

6 The Relation of d_0 to Other Learnability Related Notions

In this section we consider the relation of d_0 and the strong SQ dimensions of Simon [8] and Feldman [5]. We also discuss the relation of d_0 to the notion introduced in [2] to analyze learnability.

6.1 SQDim*

Let us first deal with SQDim* from [8].

Definition 10 ([8]). *Given some concept class $\mathcal{F}$, a subclass $\mathcal{F}'$ of it is* $(\gamma, \mathcal{H})$-trivial *for some query class $\mathcal{H}$ and constant $0 < \gamma < 1$, if some function $h \in \mathcal{H}$ has correlation of at least γ with at least half of the functions in $\mathcal{F}'$. The remaining subclasses of $\mathcal{F}$ are said to be* $(\gamma, \mathcal{H})$-nontrivial. *The* strong SQ dimension *associated with concept class $\mathcal{F}$ and query class $\mathcal{H}$ is the function* $\mathrm{SQDim}^*_{\mathcal{F},\mathcal{H}}(\gamma) := \sup_{\mathcal{F}'} \mathrm{SQDim}_{\mathcal{F}'-B_{\mathcal{F}'}}$, *where $\mathcal{F}'$ ranges over all $(\gamma, \mathcal{H})$-nontrivial subclasses of $\mathcal{F}$, and where $B_{\mathcal{F}'} = (1/|\mathcal{F}'|)\sum_{f\in\mathcal{F}'} f$.*

As it turns out below, it doesn't really matter, which query class is used, as long as it contains the concept class itself.

Observation 11. When $\mathcal{F} \subseteq \mathcal{H}$, then any $(1-\epsilon, \mathcal{F})$-trivial subset of $\mathcal{F}$ is also $(1-\epsilon, \mathcal{H})$-trivial, meanwhile, by Observation 2, it also holds that any $(1-\epsilon/2, \mathcal{H})$-trivial subset of $\mathcal{F}$ is also $(1-\epsilon, \mathcal{F})$-trivial. Thus

$$\mathrm{SQDim}^*_{\mathcal{F},\mathcal{H}}(1-\epsilon) \leq \mathrm{SQDim}^*_{\mathcal{F},\mathcal{F}}(1-\epsilon) \leq \mathrm{SQDim}^*_{\mathcal{F},\mathcal{H}}\left(1-\frac{\epsilon}{2}\right) \ .$$

.

The following equation we shall need later.

$$\langle f, g\rangle = \langle f-B, g-B\rangle + \langle f, B\rangle + \langle g, B\rangle - \|B\|^2 \ . \tag{2}$$

Theorem 12. *For any concept classes $\mathcal{F}$ and $\mathcal{H}$ satisfying $\mathcal{F} \subseteq \mathcal{H}$ it holds that* $\max\{32/\epsilon^2, 9d_0^2(\mathcal{F}, 1-\epsilon^2/32)\} \geq \mathrm{SQDim}^*_{\mathcal{F},\mathcal{H}}(1-\epsilon)$.

Proof. According to Observation 11, it is enough to show that the statement of the theorem holds for $\mathcal{H} = \mathcal{F}$.

Let $\mathcal{F}'$ be a $(1-\epsilon, \mathcal{F})$-nontrivial subset of $\mathcal{F}$, and let $\mathcal{F}'_0$ be a subset of $\mathcal{F}'$ such that $\mathrm{SQDim}_{\mathcal{F}'_0 - B_{\mathcal{F}'}} = |\mathcal{F}'_0|$. Assume furthermore that $d := |\mathcal{F}'_0| \geq 32/\epsilon^2$.

Consider the correlation of $B_{\mathcal{F}'}$ with all the functions in $\mathcal{F}'_0$. Obviously there exist some $c \in [-1,1]$ and some $d' \geq \sqrt{d}/3$ such that for some distinct functions $f_1, \ldots, f_{d'} \in \mathcal{F}'_0$ it holds that $\langle f_i, B_{\mathcal{F}'}\rangle \in [c-1/\sqrt{9d},\ c+1/\sqrt{9d}]$ for $j = 1, \ldots, d'$. Then for arbitrary indices $i, j, k, \ell \in [d']$ fulfilling $i \neq j$ and $k \neq \ell$ it holds (using (2)) that

$$\begin{aligned}
|\langle f_i, f_j\rangle - \langle f_k, f_\ell\rangle| =& |(\langle f_i - B_{\mathcal{F}'}, f_j - B_{\mathcal{F}'}\rangle - \langle f_k - B_{\mathcal{F}'}, f_\ell - B_{\mathcal{F}'}\rangle) \\
&+ (\langle f_i, B_{\mathcal{F}'}\rangle - \langle f_k, B_{\mathcal{F}'}\rangle) + (\langle f_j, B_{\mathcal{F}'}\rangle - \langle f_\ell, B_{\mathcal{F}'}\rangle)| \\
\leq& \frac{2}{d} + \frac{2\cdot 2}{\sqrt{9d}} \\
\leq& \frac{3}{\sqrt{d}}
\end{aligned} \tag{3}$$

using that $d \geq 32$. To prove the theorem it thus suffices to show that the correlation of any two distinct elements of $\mathcal{F}'$ has absolute value at most $1-\epsilon^2/32$.[8]

To upper bound $\langle f_i, f_j\rangle$ for some $1 \leq i < j \leq d'$ first note that using (2) with $f = f_i$, $g = f_j$ and $B = B_{\mathcal{F}'}$, and then applying the Cauchy-Schwarz inequality

$$\langle f_i, f_j\rangle \leq \frac{1}{d} + \|B_{\mathcal{F}'}\|\,(2 - \|B_{\mathcal{F}'}\|) \ . \tag{4}$$

[8] Note that we cannot apply Observation 2 (or Proposition 3) directly to bound $\langle f_i, f_j\rangle$, because nontriviality only guarantees that none of the f_i functions have high correlation with at least half of $\mathcal{F}'$, which doesn't prevent them from having really high correlation with some smaller portion of $\mathcal{F}'$. It thus has to be shown that no such set contains another f_i.

Also note that the $(1-\epsilon,\mathcal{F})$-nontriviality of $\mathcal{F}'$ implies that

$$\|B_{\mathcal{F}'}\|^2 = \frac{1}{|\mathcal{F}'|^2}\sum_{g,f\in\mathcal{F}'}\langle g,f\rangle \le \frac{1}{|\mathcal{F}'|}\sum_{g\in\mathcal{F}'}\frac{1}{|\mathcal{F}'|}\left(\frac{|\mathcal{F}'|}{2}(1-\epsilon)+\frac{|\mathcal{F}'|}{2}\right) = 1-\frac{\epsilon}{2}\ ,$$

and therefore $\|B_{\mathcal{F}'}\| \le \sqrt{1-\epsilon/2} \le 1-\epsilon/4$. Combining this with (4), and noting that $x(2-x)$ is monotone increasing on $(0,1)$ we get that

$$\langle f_i, f_j\rangle \le \frac{1}{d} + \left(1-\frac{\epsilon}{4}\right)\left(1+\frac{\epsilon}{4}\right) = 1+\frac{1}{d}-\frac{\epsilon^2}{16}\ .$$

Thus, since $d \ge 32/\epsilon^2$, we have $\langle f_i, f_j\rangle \le 1-\epsilon^2/32$.

Finally, let us give a lower bound for the pairwise correlation. If one pair had correlation less than $-1+1/32$, then, according to (3) all other pairs would have correlation at most $-1+1/32+3/\sqrt{d}$, implying

$$\begin{aligned}
0 &\le \left\|\sum_{i=1}^{d'} f_i\right\|^2 \\
&= \sum_{i=1}^{d'}\|f_i\|^2 + 2\sum_{1\le i<j\le d'}\langle f_i, f_j\rangle \\
&\le d' + d'(d'-1)\left(-1+\frac{1}{32}+\frac{3}{\sqrt{d}}\right)\ ,
\end{aligned}$$

which would lead to a contradiction, as $d \ge 32$. Consequently $\langle f_i, f_j\rangle \ge -1+\epsilon^2/32$ for $1 \le i < j \le d'$. □

Theorem 13. *Let $\mathcal{F}$ and $\mathcal{H}$ be concept classes satisfying $\mathcal{F} \subseteq \mathcal{H}$. Then*

$$d_0(\mathcal{F}, 1-\epsilon) \le \max\{2, 2\cdot \mathrm{SQDim}^*_{\mathcal{F},\mathcal{F}}(1-\epsilon/2)\} \le \max\{2, 2\cdot \mathrm{SQDim}^*_{\mathcal{F},\mathcal{H}}(1-\epsilon/4)\}\ .$$

Proof. The second inequality follows from Observation 11.

To prove the first inequality, let $\mathcal{F}' := \{f_1,\ldots,f_d\} \subseteq \mathcal{F}$ be such that $|\langle f_i, f_j\rangle| < 1-\epsilon$ and $|\langle f_i, f_j\rangle - \langle f_k, f_\ell\rangle| < 1/d$ for $1 \le i < j \le d$ and $1 \le k < \ell \le d$. Then

$$\begin{aligned}
&|\langle f_i - B_{\mathcal{F}'}, f_j - B_{\mathcal{F}'}\rangle| \\
&= \left|\langle f_i, f_j\rangle + \frac{1}{d^2}\sum_{k,\ell=1}^{d}\langle f_k, f_\ell\rangle - \frac{1}{d}\sum_{k=1}^{d}(\langle f_i, f_k\rangle + \langle f_j, f_k\rangle)\right| \\
&\le \left|\frac{1}{d}\sum_{k=1}^{d}(\langle f_i, f_j\rangle - \langle f_i, f_k\rangle)\right| + \left|\frac{1}{d^2}\sum_{k,\ell=1}^{d}(\langle f_k, f_\ell\rangle - \langle f_j, f_k\rangle)\right| \\
&\le \frac{2}{d}\ .
\end{aligned}$$

Furthermore, by Observation 2, $\mathcal{F}'$ is $(1-\epsilon/2,\mathcal{F})$-nontrivial. □

6.2 SSQ-DIM

The dimension notion introduced in [5] is a kind of simplified version of SQDim^*:

Definition 14 ([5]). *For concept class $\mathcal{F}$ over domain X let* $\text{SSQ-DIM}(\mathcal{F},\epsilon) := \max_h \mathrm{SQDim}_{\{f'\in(\mathcal{F}-h):\|f'\|^2\geq\epsilon\}}$, *where h ranges over all mappings from X to $[-1,1]$.*

Furthermore the proof of Theorem 12 and Theorem 13 can be easily modified to show some similar results about the relation of d_0 and SSQ-DIM as follows.

Theorem 15. *For any concept class $\mathcal{F}$ it holds that* $\max\{32, 2/\epsilon, 9d_0^2(\mathcal{F}, 1-\epsilon/2)\} \geq \text{SSQ-DIM}(\mathcal{F},\epsilon)$.

Proof. Let $d := \text{SSQ-DIM}(\mathcal{F},\epsilon)$, let $h : X \to [-1,1]$ (where X is the domain of $\mathcal{F}$), and let $f_1,\ldots,f_d \in \mathcal{F}$ fulfill $|\langle f_i - h, f_j - h\rangle| \leq 1/d$ and $\|f_i - h\|^2 \geq \epsilon$ for $i,j \in [d]$ distinct. Let $d' := \lceil\sqrt{d}/3\rceil$. Then, again, we can assume without loss of generality that for some $c \in [-1,1]$ we have $|\langle f_i, h\rangle - c| \leq 1/d'$ for $i \in [d']$. Consequently (as in the proof of Theorem 12) $|\langle f_i, f_j\rangle - \langle f_k, f_\ell\rangle| \leq 3/\sqrt{d}$ for $1 \leq i < j \leq d'$ and $1 \leq k < \ell \leq d'$.

Now all that is left is to bound the pairwise correlations of functions f_i, $i \in [d]$. For this note that

$$\epsilon \leq \|f_i - h\|^2 = \|f_i\|^2 + \|h\|^2 - 2\langle f_i, h\rangle \ ,$$

and so

$$\langle f_i, h\rangle \leq \frac{1}{2}\left(1 - \epsilon + \|h\|^2\right) \tag{5}$$

for $i \in [d']$. Thus

$$\langle f_i, f_j\rangle \overset{(2)}{=} \langle f_i - h, f_j - h\rangle + \langle f_i, h\rangle + \langle f_j, h\rangle - \|h\|^2 \overset{(5)}{\leq} \frac{1}{d} + 1 - \epsilon \ ,$$

which is upper bounded by $1 - \epsilon/2$ when $d \geq 2/\epsilon$. Finally, the lower bound for the pairwise correlations (for $d \geq 32$) can be proved just as in the proof of Theorem 12.

Theorem 16. *For any concept class $\mathcal{F}$ it holds that* $d_0(\mathcal{F}, 1-\epsilon) \leq \max\{2, 2 \cdot \text{SSQ-DIM}(\mathcal{F}, \epsilon^2/16)\}$.

Proof. Let $\mathcal{F}' = \{f_1,\ldots,f_d\} \subseteq \mathcal{F}$ be such that $|\langle f_i, f_j\rangle - \langle f_k, f_\ell\rangle| \leq 1/d$ and $|\langle f_i, f_j\rangle| \leq 1 - \epsilon$ for $1 \leq i < j \leq d'$ and $1 \leq k < \ell \leq d'$. We choose $h := B_{\mathcal{F}'}$. Then, as in the proof of Theorem 13, $|\langle f_i - B_{\mathcal{F}'}, f_j - B_{\mathcal{F}'}\rangle| \leq 2/d$.

Now we show that the f_i functions are "not too close" to $B_{\mathcal{F}'}$. For this note that by Cauchy-Schwarz

$$\|f_i - B_{\mathcal{F}'}\|^2 = 1 + \|B_{\mathcal{F}'}\|^2 - 2\langle f_i, B_{\mathcal{F}'}\rangle \geq 1 + \|B_{\mathcal{F}'}\|\left(\|B_{\mathcal{F}'}\| - 2\right) \ ,$$

and so, since $x(x-2)$ is monotone decreasing on $(0,1)$, and because

$$\|B_{\mathcal{F}'}\|^2 = \frac{1}{d^2}\sum_{i=1}^{d}\|f_i\|^2 + \frac{2}{d^2}\sum_{1\leq i<j\leq d}\langle f_i, f_j\rangle \leq \frac{1}{d} + \frac{d-1}{d}(1-\epsilon) \leq 1 - \frac{\epsilon}{2}$$

for $d \geq 2$, which implies $\|B_{\mathcal{F}'}\| \leq 1 - \epsilon/4$, we conclude that $\|f_i - B_{\mathcal{F}'}\|^2 \geq \epsilon^2/16$.

6.3 Capacity

In [2] Ben-David et al. related the learning complexity of a class $\mathcal{F}$ to its *capacity* $\mathrm{CAP}(\mathcal{F}, \epsilon) := \min\{|G| : \forall f \in \mathcal{F}\ \exists g \in G \text{ s.t. } \langle f, g\rangle \geq 1 - \epsilon\}$. For the concept class $\mathcal{F}_n = \{v_1, \ldots v_n\}$ consisting of the monotone conjunctions of length 1 this is polynomial in the learning complexity (in specific $\mathrm{CAP}(\mathcal{F}_n, \epsilon) = n$) under uniform distribution, but for the monotone conjunctions this is superpolynomial (choose $s = \log\log n$ in Example 19). The two notions are thus not polynomially related.

7 d_0 for Conjunctions Under the Uniform Distribution

In this section, as an example, we compute the exact value of d_0 for the class of conjunctions under the uniform distribution, up to a constant factor. (Note however that this class is efficiently learnable in the Statistical Query model even distribution independently [6], so d_0 is obviously polynomial in n and in $1/\epsilon$.)

First of all let us compute the correlation of two conjunctions t and t' that have length ℓ and ℓ' respectively, and share exactly s literals (as usual, -1 is interpreted as "true" and 1 as "false"):

$$\begin{aligned}\langle t, t'\rangle &= \mathbb{E}[t \cdot t'] \\ &= 1 - 2\,\mathbb{P}[t \neq t'] \\ &= 1 - 2(\mathbb{P}[t = -1] + \mathbb{P}[t' = -1] - 2\,\mathbb{P}[t = t' = -1]) \\ &= \begin{cases} 1 - 2/2^{\ell} - 2/2^{\ell'} & \text{if } t \text{ and } t' \text{ conflict,} \\ 1 - 2/2^{\ell} - 2/2^{\ell'} + 4/2^{\ell+\ell'-s} & \text{otherwise.} \end{cases}\end{aligned} \tag{6}$$

Next we prove a technical lemma we shall need later. Here we apply the convention that for some $x \in \{0,1\}^n$ the number of 1s in x is denoted $|x|$, and that for $x, y \in \{0,1\}^n$ $x \vee y$ (resp. $x \wedge y$) is the vector of length n with 1 on those components that are 1 in at least one of x and y (resp. in both x and y), and is 0 everywhere else. For conjunctions we use similar notations, that is, $|t|$ denotes the number of literals appearing in term t, and $t \wedge t'$ denotes the term obtained by joining the literals appearing in terms t and t'.

Lemma 17. *If for some $H \subseteq \{0,1\}^n$ and for some integer c it holds that $|x \vee y| = c$ for arbitrary distinct $x, y \in H$, then $|H| \leq n + 1$.*

Proof. For $x \in H$ let x^{c} denote the vector obtained by flipping the bits in x. Then by De Morgan $x^{\mathsf{c}} \wedge y^{\mathsf{c}} = (x \vee y)^{\mathsf{c}}$, and thus $|x^{\mathsf{c}} \wedge y^{\mathsf{c}}| = n - c$ for arbitrary $x, y \in H$. Construct the $n \times |H|$ matrix X such that its columns are the vectors from H in an arbitrary order, and let C be the $|H| \times |H|$ matrix having $n - c$ in each entry. First of all note that $X^{\top}X - C$ is a diagonal matrix. If it contains some zero element in the diagonal, then $|x^{\mathsf{c}}| = n - c$ for some $x \in H$, implying that for all other $y \in H$ y^{c} has 1 everywhere where x does and that each such y^{c}

must have 1 at some unique position where the others have 0. This immediately implies $|H| \leq n+1$. Otherwise, when $X^\top X - C$ is a nonsingular diagonal matrix,

$$|H| = \operatorname{rank}\left(X^\top X - C\right) \leq \operatorname{rank}\left(X^\top X\right) + 1 = \operatorname{rank}(X) + 1 \leq \min\{n, |H|\} + 1$$

implying the statement of the claim. □

Proposition 18. *Let $\mathcal{F}_n$ be the set of conjunctions over variables $v_1, \ldots, v_n$. Then under the uniform distribution $d_0(\mathcal{F}_n, 1-\epsilon) \leq 1 + \max\{2n+2, 8/\epsilon^2\}$.*

Proof. Let $t_1, \ldots, t_d$ be terms satisfying $|\langle t_i, t_j \rangle| \leq 1-\epsilon$ and $|\langle t_i, t_j \rangle - \langle t_k, t_\ell \rangle| \leq 1/d$ for $i, j, k, \ell \in [d]$ fulfilling $i \neq j$ and $k \neq \ell$. Assume for simplicity that t_d is the longest term among them. Then by (6) it holds that $1 - \epsilon \geq \langle t_i, t_d \rangle \geq 1 - 4\,\mathbb{P}[t_i = -1]$, implying

$$\mathbb{P}[t_i = -1] = 2^{-|t_i|} \geq \frac{\epsilon}{4} \ , \tag{7}$$

and thus

$$\mathbb{P}[t_i = t_j = -1] = \begin{cases} 0 & \text{if } t_i \text{ and } t_j \text{ conflict} \\ 2^{-|t_i \wedge t_j|} \geq (\epsilon/4)^2 & \text{otherwise} \end{cases} \tag{8}$$

for distinct $i, j \in [d-1]$.

Let us assume that $1/d < \epsilon^2/8$.

If for some $I \subseteq [d-1]$ it holds that all t_i, $i \in I$, has the same length, then for any indices $i, j, k, \ell \in I$ fulfilling $i \neq j$ and $k \neq \ell$

$$\frac{\epsilon^2}{32} > \frac{1}{4d} \geq \frac{1}{4}|\langle t_i, t_j \rangle - \langle t_k, t_\ell \rangle| \overset{(6)}{=} |\mathbb{P}[t_i = t_j = -1] - \mathbb{P}[t_k = t_\ell = -1]| \, .$$

Note that it cannot happen that t_i and t_j conflict with each other, but t_k and t_ℓ do not—or vice versa—, since by (8) that would mean that the right hand side is at least $\epsilon^2/16$, resulting in a contradiction. So either all t_i with $i \in I$ conflict each other, or there is no conflicting pair among the terms with index in I. The former case implies that $\{t_i = -1\}_{i \in I}$ are all contradicting events, and so $1 \geq \sum_{i \in I} \mathbb{P}[t_i = -1] \overset{(7)}{\geq} |I| \cdot (\epsilon/4)$, giving the bound $|I| \leq 4/\epsilon$. In the latter case, since by (8) both $2^{-|t_i \vee t_j|}$ and $2^{-|t_k \vee t_\ell|}$ are at least $\epsilon^2/16$, we have that $2^{-|t_i \vee t_j|} > (1/2)2^{-|t_k \vee t_\ell|}$ and $2^{-|t_k \vee t_\ell|} > (1/2)2^{-|t_i \vee t_j|}$. This, however implies that $|t_i \vee t_j| = |t_k \vee t_\ell|$, and so, by Lemma 17 (applied for $H \subseteq \{0,1\}^n$ consisting of the vectors that represent some t_i with $i \in I$ by having 1 on position j iff t_i contains variable v_j), I has cardinality at most $n+1$.

We have just seen that the sum of the number of terms of minimal length and the number of terms of length one more is at most $\max\{2n+2, 8/\epsilon\}$. However, there cannot be distinct indices $i, j, k \in [d-1]$ fulfilling $|t_i| + 2 \leq |t_j|, |t_k|$, as otherwise

$$\frac{\epsilon^2}{8} > \frac{1}{d}$$

$$
\begin{aligned}
&\geq |\langle t_i, t_j\rangle - \langle t_j, t_k\rangle| \\
&= |2\,\mathbb{P}[t_i = -1] - 4\,\mathbb{P}[t_i = t_j = -1] - 2\,\mathbb{P}[t_k = -1] + 4\,\mathbb{P}[t_k = t_j = -1]| \\
&\geq \frac{1}{2} \cdot \mathbb{P}[t_i = -1] \\
&\overset{(7)}{\geq} \frac{\epsilon}{8}\ ,
\end{aligned}
$$

a contradiction. □

Note that this bound is sharp up to a constant factor according to the example below and that the terms consisting of one unnegated variable form an orthogonal system of cardinality n. It also immediately follows that these results remain tight even if we restrict $\mathcal{F}_n$ to be the set of *monotone* conjunctions over $v_1, \ldots, v_n$.

Example 19. Let $\mathcal{F}_n$ be the set of all monotone conjunctions over variables $v_1, \ldots, v_n$ and let $\mathcal{F}_n(\ell)$ consist of all $t \in \mathcal{F}_n$ of length ℓ. Set $\epsilon := 2^{-\ell}$ and note that if $t_1, t_2 \in \mathcal{F}_n(\ell)$ share $s < \ell$ variables, then under the uniform distribution $|\langle t_1, t_2\rangle| \overset{(6)}{=} 1 - 4/2^{\ell} + 4/2^{2\ell - s} \leq 1 - 2\epsilon$. If additionally $t_3, t_4 \in \mathcal{F}_n(\ell)$ share $s' < \ell$ variables, then $|\langle t_1, t_2\rangle - \langle t_3, t_4\rangle| \overset{(6)}{=} \left|4/2^{2\ell-s} - 4/2^{2\ell-s'}\right| = \epsilon^2 4\left|2^s - 2^{s'}\right|$. Now we choose $\ell = \ell(n) := c \log n$ for some $c > 1$ (and thus $\epsilon = \epsilon(n) = 1/n^c$) and $s = s(n) := \log\log n$, and prove that $d_0(\mathcal{F}_n, 1-\epsilon) = \Omega(\epsilon^2) = \Omega(n^{2c})$ by showing that one can find an $I \subseteq \mathcal{F}_n(\ell)$ of cardinality $\Omega(n^{2c})$ that contains no two distinct conjunctions sharing more than s variables. Such an I can simply be obtained using the greedy method, since when $n - \ell \geq 2(\ell - s)$ then for any $t \in \mathcal{F}_n(\ell)$ there are exactly $\sum_{i=0}^{\ell-s} \binom{\ell}{i}\binom{n-\ell}{i} \leq \ell 2^{\ell}\binom{n-\ell}{\ell-s}$ conjunctions in $\mathcal{F}_n(\ell)$ that share at least s variables with t, thus (noting that $|\mathcal{F}_n(\ell)| = \binom{n}{\ell}$) I can always be expanded by some term when it has cardinality less than

$$
\begin{aligned}
\frac{\binom{n}{\ell}}{\ell 2^{\ell}\binom{n-\ell}{\ell-s}} \sim& \frac{1}{\ell 2^{\ell}} \cdot \sqrt{\frac{n(\ell-s)(n-2\ell+s)}{\ell(n-\ell)(n-\ell)}} \cdot \frac{n^n(\ell-s)^{\ell-s}(n-2\ell+s)^{n-2\ell+s}}{\ell^{\ell}(n-\ell)^{n-\ell}(n-\ell)^{n-\ell}} \\
\sim& \frac{1}{\ell 2^{\ell}} \cdot 1 \cdot \left(\frac{n^n \ell^{\ell-s} n^{n-2\ell+s}}{\ell^{\ell} n^{n-\ell} n^{n-\ell}} \right. \\
& \left. \cdot \frac{(1-s/\ell)^{(\ell/s-1)s}(1-(2\ell-s)/n)^{(n/(2\ell-s)-1)(2\ell-s)}}{(1-\ell/n)^{(n/\ell-1)\ell}(1-\ell/n)^{(n/\ell-1)\ell}} \right) \\
\sim& \frac{1}{\ell 2^{\ell}} \cdot 1 \cdot \left(\frac{n^s}{\ell^s} \cdot \frac{e^{-s}e^{-(2\ell-s)}}{e^{-\ell}e^{-\ell}} \right) \\
=& \frac{1}{\ell 2^{\ell}} \frac{n^s}{\ell^s}
\end{aligned}
$$

(using Stirling's formula).

8 Proper vs. Improper Learning in the Distribution Independent Case

In the distribution dependent case (i.e., when the learner knows the underlying distribution) proper and improper learning are basically the same (recall Corollary 9). In this section we contrast this result showing that in the distribution independent case proper and improper learning can differ significantly. Consider for example the class of singletons: $\mathcal{F}_n := \{f_x : x \in \{-1,1\}^n\}$, where f_x evaluates to -1 on x, and evaluates to 1 on every other input. Since $\mathcal{F}_n$ is a subset of conjunctions, which was shown by Kearns in [6] to be efficiently learnable in the Statistical Query model, $\mathcal{F}_n$ can be learned using polynomially many *improper* queries.

Let us now define for each $x, y \in \{-1,1\}^n$ a distribution $\mathcal{D}_{x,y}$, which assigns probability $1/2$ to both x and y, and assigns probability 0 to every other input. The key observation is that in case of proper learning each query must be one of the f_x functions. But then, as long as there are at least two of them that are not yet queried, the adversary can just return 0 as the answer. Finally, when only two singletons—say f_x and f_y—are unqueried, the adversary chooses one of them as the target concept, and says that the underlying distribution is $\mathcal{D}_{x,y}$. This way the answers of the adversary remain consistent (no matter how small the tolerance parameter of the learner was), and, at the same time, force the learner to ask at least $2^n - 1$ queries—even for weakly learning the class.[9]

It might also worth mentioning that for the singletons $\mathrm{SQDim}^{\mathcal{D}}_{\mathcal{F}_n} \leq 5$ under any distribution $\mathcal{D}$, because, denoting by p_x the probability assigned to input $x \in \{-1,1\}^n$, $1/6 \geq \langle f_x, f_y \rangle_{\mathcal{D}} = 1 - 2p_x - 2p_y$ implies that at least one of p_x and p_y is $5/24$ or greater, and thus if six functions from $\mathcal{F}_n$ had pairwise correlation at most $1/6$, then at least five distinct inputs would have probability $5/24$ or greater—a contradiction. This result shows that the number of *proper* queries required for weakly learning some concept class can differ significantly in the distribution dependent and in the distribution independent case: in some cases it is constant versus exponential.

9 Open Problems

The approach used in the proof of Theorem 7 raises the problem, how to find some hypothesis h that is consistent with the answers so far. When we do consider the running time of the learning algorithm, this becomes a crucial question regarding the applicability of this approach. It is also interesting whether this problem is always efficiently solvable when the class is efficiently learnable in the SQ model. (Note that for example the learning algorithm for conjunctions in [6] is *not* consistent.)

Recall also that the characterization results in the paper are all for the distribution dependent characterization. It would be nice to have characterization

[9] This doesn't contradict the result of Aslam and Decatur [1] mentioned in Section 4, since their boosting algorithm uses improper queries.

results for the distribution independent case as well. However, the overall goal would be to characterize learnability in the PAC model in the presence of random noise.

Acknowledgements. I would like to thank Hans Ulrich Simon for suggesting me to work on this topic. I am also thankful to him, Thorsten Doliwa and Michael Kallweit for the motivating discussions on the problem.

References

1. Aslam, J.A., Decatur, S.E.: General bounds on statistical query learning and PAC learning with noise via hypothesis boosting. Inf. Comput. **141**(2) (1998) 85–118
2. Ben-David, S., Itai, A., Kushilevitz, E.: Learning by distances. Inform. Comput. **117**(2) (1995) 240–250
3. Blum, A., Furst, M., Jackson, J., Kearns, M., Mansour, Y., Rudich, S.: Weakly learning DNF and characterizing statistical query learning using fourier analysis. In: Proc. of 26th ACM Symposium on Theory of Computing. (1994)
4. Bshouty, N.H., Feldman, V.: On using extended statistical queries to avoid membership queries. Journal of Machine Learning Research **2** (2002) 359–395
5. Feldman, V.: A complete characterization of statistical query learning with applications to evolvability. FOCS 2009 (to appear)
6. Kearns, M.: Efficient noise-tolerant learning from statistical queries. J. ACM **45**(6) (1998) 983–1006
7. Köbler, J., Lindner, W.: The complexity of learning concept classes with polynomial general dimension. Theor. Comput. Sci. **350**(1) (2006) 49–62
8. Simon, H.U.: A characterization of strong learnability in the statistical query model. In: STACS. (2007) 393–404
9. Szörényi, B.: Honest queries do not help in the statistical query model. Manuscript
10. Yang, K.: New lower bounds for statistical query learning. J. Comput. Syst. Sci. **70**(4) (2005) 485–509 Special issue on COLT 2002.
11. Yang, K.: On learning correlated boolean functions using statistical queries. In: Proc. of the 12th Internat. Conf. on Algorithmic Learning Theory. (2001) 59–76