A Characterization of Strong Learnability in the Statistical Query Model *

Hans Ulrich Simon

Fakultät für Mathematik, Ruhr-Universität Bochum, 44780 Bochum, Germany
simon@lmi.rub.de

Abstract. In this paper, we consider Kearns' [5] Statistical Query Model of learning. It is well known [3] that the number of statistical queries, needed for "weakly learning" an unknown target concept (i.e. for gaining significant advantage over random guessing) is polynomially related to the so-called Statistical Query dimension of the concept class. In this paper, we provide a similar characterization for "strong learning" where the learners final hypothesis is required to approximate the unknown target concept up to a small rate of misclassification. The quantity that characterizes strong learnability in the Statistical Query model is a surprisingly close relative of (though not identical to) the Statistical Query dimension. For the purpose of proving the main result, we provide other characterizations of strong learnability which are given in terms of covering numbers and related notions. These results might find some interest in their own right. All characterizations are purely information-theoretical and ignore computational issues.

1 Introduction

Kearns' Statistical Query (SQ) model [5] is an elegant abstraction from Valiant's PAC learning model [7]. In this model, instead of having direct access to random examples (as in the PAC learning model) the learner obtains information about random examples via an oracle that provides estimates of various statistics about the unknown concept. Kearns showed that any learning algorithm that is successful in the SQ model can be converted, without much loss of efficiency, into a learning algorithm that is successful in the PAC learning model despite noise uniformly applied to the class labels of the examples. In the same paper where Kearns showed that SQ learnability implies noise-tolerant PAC learnability, he developed SQ algorithms for almost all concept classes known to be efficiently learnable in the PAC learning model. This is why the SQ model attracted a lot of attention in the Computational Learning community.

* This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views. Part of this work was done during a visit of the Helsinki Institute of Information Technology.

Blum et al. [3] have shown that the number of statistical queries, needed for weakly learning a concept class, is polynomially related to the largest number of pairwise "almost orthogonal" concepts from this class (the so-called SQ dimension). Ke Yang [8] presented an alternative (stronger but polynomially related) lower bound on the number of statistical queries. It is stated in terms of the eigenvalues of the correlation matrix associated with the concept class. The problem of characterizing strong learnability in the SQ model was left open in these papers.

Köbler and Lindner [6] characterized strong learnability in the SQ model in terms of the so-called "general dimension" (a variant of the abstract combinatorial dimension from [1]). The general dimension can be viewed as the number of queries needed when the SQ oracle behaves like a uniform non-adaptive adversary that returns answers according to a fixed scheme (the *same* scheme for all learners). Köbler and Lindner show that this number is polynomially related to the number of queries needed in the worst-case where the SQ oracle behaves like an adaptive and non-uniform adversary (the strongest adversary of the particular learner). Loosely speaking, the result by Köbler and Lindner tells that the handicap of being non-adaptive and uniform does not make the adversary substantially weaker. The general dimension has no algebraic flavor. In contrast, the results in this paper make extensive use of algebraic properties of the correlation matrix associated with the concept class. We conjecture that our results can also serve as a tool for the computation of the general dimension.

Models related to the SQ model: We would like to mention briefly that the SQ model has a predecessor: the model of "Learning by Distances" [2]. The former model is a special case of the latter (but the inventors of the latter model did not reveal the relation to noise-tolerant learning).
The SQ model is furthermore equivalent to a conceptually simpler model where statistical queries (as they were originally defined by Kearns) are replaced by "correlation queries". For the purpose of our theoretical analysis, the simpler model is more convenient. For this reason, we do not bother the reader with the definition of the original model. Instead we define the SQ model in section 2 directly within the Correlation Query framework. The reader interested in details about the equivalence between the two models is referred to [4].
In the original model by Kearns, "concepts" and "query functions" are ± 1-valued. In this paper, it will be more convenient to deal with real-valued functions instead. The traditional setting of concept learning (with binary functions) will however be subsumed as a special case.

Structure of the paper: In section 2, we provide the reader with rigorous definitions and notations (including the notions of weak and strong learning in the SQ model). Section 3 briefly reviews the known bounds on the number of statistical queries and the known characterization of weak learnability in terms of the (Weak) Statistical Query dimension (SQdim). In section 4, we define the Strong Statistical Query dimension (SQDim*) and present a characterization of strong learnability in terms of this dimension. It should be stressed that, although SQDim* is a close relative of SQdim, the mathematical analysis of strong learnability in the SQ model is much more involved than the preceding analysis of weak learnability.

2 Definitions and Notations

Concept Learning: A *concept* is a function f of the form $f : X \rightarrow \{\pm 1\}$. A concept class, denoted as $\mathcal{F}$, is a set of concepts. Throughout this paper, X is a finite set called the *domain*. An element $x \in X$ is called an *instance*. A labeled instance $(x, b) \in X \times \{\pm 1\}$ is called an *example for* f if $b = f(x)$. $D : X \rightarrow [0, 1]$ denotes a mapping that assigns probabilities to instances such that $D(x) \geq 0$ for all $x \in X$ and $\sum_{x \in X} D(x) = 1$. Notations $\Pr_D[\cdot]$ and $\mathbb{E}_D[\cdot]$ refer to the probability of an event and to the expectation of a random variable, respectively. A function $h : X \rightarrow \{\pm 1\}$ (not necessarily from $\mathcal{F}$) is called a *(binary) hypothesis*. Informally, the goal of a learner for concept class $\mathcal{F}$ is to infer from "partial information" about an unknown *target concept* $f_* \in \mathcal{F}$ a hypothesis h_* whose probability of misclassification, $\Pr_D[h(x) \neq f(x)]$, is small. A notion that helps to keep track of the progress made by the learner is the so-called "version space". Formally, the *version space* is a subclass $\mathcal{F}' \subseteq \mathcal{F}$ of the concept class (initially $\mathcal{F}' = \mathcal{F}$) that consists of all concepts being consistent with the partial information that is currently available to the learner. The formal notion of "partial information" depends on the learning model in consideration. The reader might perhaps be familiar with the PAC learning model [7] where the information given to the learner consists of random examples for the unknown target concept. In this paper, we will be mainly concerned with the SQ model that is outlined in the remainder of this section.

Correlation Matrix: The real-valued functions on domain X form an $|X|$-dimensional vector space that can be equipped with the following inner product:

$$\langle h_1, h_2 \rangle_D := \sum_{x \in X} D(x) h_1(x) h_2(x) = \mathbb{E}_D[h_1(x) h_2(x)]$$

This inner product induces the so-called D-*norm*: $\|h\|_D := \langle h, h \rangle_D$. Note that $\|h\|_D = 1$ if h is a ± 1-valued function, and $\|h\|_D \leq 1$ if h is a function with values in $[-1, 1]$. In the sequel, $\mathcal{F}$ denotes a class of real-valued functions, and $C_\mathcal{F} \in [-1, 1]^{\mathcal{F} \times \mathcal{F}}$ such that $C_\mathcal{F}[f_1, f_2] := \langle f_1, f_2 \rangle_D$ denotes the *correlation matrix* induced by $\mathcal{F}$. Note that $C_\mathcal{F}$ is positive semidefinite.

Ingredients of the SQ Model: Notions $\mathcal{F}, X, D, f_*$ are understood as in the general concept learning framework except that $\mathcal{F}$ is a function class (with concept classes as a special case). The learner has access to random examples only indirectly through queries that are answered by an oracle. To this end, let $\mathcal{Q}$ denote a function class from which the learner picks its so-called query functions and its final hypothesis. Throughout the paper, we assume that $\mathcal{F} \subseteq \mathcal{Q}$ and that $1 \leq B < \infty$ is an upper bound on the D-norm of functions from $\mathcal{Q}$. Moreover, we assume that $h \in \mathcal{Q}$ implies that $-h \in \mathcal{Q}$. A *query* is of the form (h, τ) where $h \in \mathcal{Q}$ and $\tau > 0$. τ is called the *tolerance parameter*. Upon such a query, the oracle returns a τ-*approximation* for $\langle h, f_* \rangle_D$, i.e. it returns a number c that satisfies

$$\langle h, f_* \rangle_D - \tau \leq c \leq \langle h, f_* \rangle_D + \tau \ . \tag{1}$$

Like in the general framework of concept learning, the learner will stop at some point and return a final hypothesis $h_* \in \mathcal{Q}$. We will measure the "efficiency" of the learner by the number of queries that it passes to the oracle (in the worst-case) and by the smallest tolerance parameter τ that is ever used during learning. We will measure the "success" of the learner by the correlation between the target function f_* and the final hypothesis h_*.

Statistical Learning Complexity: We say that $q = q(\tau, \gamma)$ queries are sufficient to learn $\mathcal{F}$ with query functions and final hypothesis chosen from $\mathcal{Q}$ if there exists a learning procedure L (with unbounded computational resources) that achieves the following. For every possible target function $f_* \in \mathcal{F}$ and for every possible policy for the oracle (subject to (1)), L asks at most $q-1$ queries (with tolerance parameter set to τ or a larger value) until it comes up with a final hypothesis $h_* \in \mathcal{Q}$ (a kind of q-th query function) whose correlation with f_* is at least γ. $\mathrm{SLC}_{\mathcal{F},\mathcal{Q}}(\tau, \gamma)$ denotes the smallest number of queries that is sufficient for that purpose. If $\mathcal{Q}$ consists of all functions of the form $h : X \to \mathbb{R}$ such that $\|h\|_D \leq 1$, we simply write $\mathrm{SLC}_{\mathcal{F}}(\tau, \gamma)$.[1]

Parameterized Classes, Weak and Strong Learnability: A parameterized function class is of the form $\mathcal{F} = \cup_{n \geq 1} \mathcal{F}_n$. Similarly, $\mathcal{Q} = \cup_{n \geq 1} \mathcal{Q}_n$. We say that $\mathcal{F}$ is *weakly learnable* in the SQ model with query functions and final hypothesis chosen from $\mathcal{Q}$ if there exist functions $\gamma(n) > 0$ and $\tau(n) > 0$ such that $1/\gamma(n)$, $1/\tau(n)$, and $\mathrm{SLC}_{\mathcal{F}_n,\mathcal{Q}_n}(\tau(n), \gamma(n))$ are polynomially bounded in n. We say that $\mathcal{F}$ is *strongly learnable* in the SQ model with query functions and final hypothesis chosen from $\mathcal{Q}$ if there exists a function $\tau(n, \varepsilon) > 0$ such that $1/\tau(n, \varepsilon)$ and $\mathrm{SLC}_{\mathcal{F}_n,\mathcal{Q}_n}(\tau(n), 1 - \varepsilon)$ are polynomially bounded in n and $1/\varepsilon$.

Notations and Facts from Matrix Theory: Although we assume some familiarity with basic concepts from matrix theory, we provide the reader with a refreshment of his or her memory and fix some notation. The Euclidean norm of a vector $u \in \mathbb{R}^d$ is denoted as $\|u\|$. The real eigenvalues of a symmetric matrix A are denoted as $\lambda_1(A) \geq \lambda_2(A) \geq \lambda_3(A) \geq \cdots$. We finally would like to mention the following fact (known as Geršgorin's Disc Theorem): for any $(d \times d)$-matrix M with complex entries, the union of discs,

$$\bigcup_{i=1}^{d} \left\{ z \,\middle|\, |z - M_{i,i}| \leq \sum_{j:1 \leq j \leq d, j \neq i} |M_{i,j}| \right\} \ ,$$

[1] An analogous convention is made in the sequel for other notions depending on $\mathcal{Q}$.

covers all (complex) eigenvalues of M. Here, $|\cdot|$ denotes the absolute value of a complex number. The theorem holds in particular for symmetric real-valued matrices where all eigenvalues are reals.

3 Known Bounds on the Number of Statistical Queries

The following beautiful result by Ke Yang presents a lower bound on statistical learning complexity in terms of the eigenvalues of the correlation matrix $C_{\mathcal{F}}$.

Theorem 1 ([8]). *Assume that all query and target functions have D-norm at most* 1. *Then,* $\sum_{i=1}^{\mathrm{SLC}_{\mathcal{F}}(\tau,\gamma)} \lambda_i(C_{\mathcal{F}}) \geq |\mathcal{F}| \cdot \min\{\gamma^2, \tau^2\}$.

Definition 1 (Covering Number). *A set of functions $h_1, \ldots, h_k \in \mathcal{Q}$ is called a τ-*cover *of $\mathcal{F}$ if, for every function $f \in \mathcal{F}$, there exists an index $i \in \{1, \ldots, k\}$ such that $|\langle f, h_i\rangle_D| \geq \tau$. We denote the size of the smallest τ-cover by* $\mathrm{CNum}_{\mathcal{F},\mathcal{Q}}(\tau)$.

The following result is easy to infer from Ke Yang's proof of Theorem 1:

Corollary 1. *Assume that $\tau \leq \gamma$. Then,* $\mathrm{SLC}_{\mathcal{F}}(\tau,\gamma) \geq \mathrm{CNum}_{\mathcal{F}}(\tau)$ *and*

$$\sum_{i=1}^{\mathrm{CNum}_{\mathcal{F}}(\tau)} \lambda_i(C_{\mathcal{F}}) \geq |\mathcal{F}| \cdot \tau^2 \ . \tag{2}$$

Definition 2 (Weak SQ Dimension). *Let* $\mathrm{SQDim}_{\mathcal{F}}$ *denote the largest number d such that $\mathcal{F}$ contains functions $f_1, \ldots, f_d$ with the following property:*[2]

$$\forall 1 \leq i < j \leq d : \ |\langle f_i, f_j\rangle_D| < \frac{1}{d-1} \tag{3}$$

Let $d := \mathrm{SQDim}_{\mathcal{F}}$. Consider a maximal sequence $f_1, \ldots, f_s \in \mathcal{F}$ with the property that $|\langle f_i, f_j\rangle_D| < 1/d$ for all $1 \leq i < j \leq s$. These functions form a $1/d$-cover (because, otherwise, the sequence could be extended). Thus, $s \geq \mathrm{CNum}_{\mathcal{F},\mathcal{F}}(1/d)$. On the other hand, $s \leq d$ (because, otherwise, the sequence witnesses that $\mathrm{SQDim}_{\mathcal{F}} \geq d+1$). We conclude that the following holds (provided that $\mathcal{F} \subseteq \mathcal{Q}$):

$$\mathrm{SQDim}_{\mathcal{F}} \geq \mathrm{CNum}_{\mathcal{F},\mathcal{F}}\left(\frac{1}{\mathrm{SQDim}_{\mathcal{F}}}\right) \geq \mathrm{CNum}_{\mathcal{F},\mathcal{Q}}\left(\frac{1}{\mathrm{SQDim}_{\mathcal{F}}}\right) \ . \tag{4}$$

Ke Yang [8] inferred from Theorem 1 that for $d = \mathrm{SQDim}_{\mathcal{F}}$ the following (slight improvement on an older result from [3]) holds:

$$\mathrm{SLC}_{\mathcal{F}}(d^{-1/3}, d^{-1/3}) \geq \frac{1}{2} \cdot d^{1/3}$$

From Corollary 1, one can infer a slightly more general result:

[2] If we replace $1/(d-1)$ on the right hand side of (3) by $1/d^3$, we obtain the original definition (from [3]) of the SQ dimension of $\mathcal{F}$.

Corollary 2. *Assume $\tau \leq \gamma$. Then,*

$$\mathrm{SLC}_{\mathcal{F}}(\tau, \gamma) \geq \mathrm{CNum}_{\mathcal{F}}(\tau) \geq \frac{1}{2} \cdot \tau^2 \cdot \mathrm{SQDim}_{\mathcal{F}} \ .$$

Weak learnability in the SQ model can be perfectly characterized in terms of the weak SQ dimension:

Corollary 3 ([3]). *A parameterized function class $\mathcal{F} = \cup_{n \geq 1} \mathcal{F}_n$ is weakly learnable in the SQ-model if and only if the growth rate of $\mathrm{SQDim}_{\mathcal{F}_n}$ is polynomially bounded in n.*

Proof. Direction "only if" is obvious from Corollary 2. Direction "if" easily follows from (4). □

4 Strong Learnability in the SQ Model

It is well-known that there exist concept classes that are weakly, but not strongly, learnable. These classes will have a reasonably small SQ dimension. This rules out the possibility of characterizing strong learnability in the SQ model by the weak SQ dimension. As for strong learnability, we need, on one hand, stronger lower bounds and, on the other hand, (reasonably well) matching upper bounds. It turns out that satisfactory (lower and upper) bounds are obtained in terms of the SQ dimension of classes of the form

$$\mathcal{F} - g := \{f - g : \ f \in \mathcal{F}\}$$

provided that the set of "admissible shift functions" g is carefully chosen. A serious technical obstacle is the fact that possibly $\mathrm{SLC}_{\mathcal{F}-g,\mathcal{Q}}(\tau, \gamma) > \mathrm{SLC}_{\mathcal{F},\mathcal{Q}}(\tau, \gamma)$ such that lower bounds on $\mathrm{SLC}_{\mathcal{F}-g,\mathcal{Q}}(\tau, \gamma)$ are not necessarily lower bounds on $\mathrm{SLC}_{\mathcal{F},\mathcal{Q}}(\tau, \gamma)$. We shall proceed as follows:

- As for the upper bound, we will pursue a "halving strategy" where the learner proceeds in stages and, at any fixed stage, tries to "halve" the current version space. This will lead to the notion of "statistical halving complexity" $\mathrm{SHC}_{\mathcal{F},\mathcal{Q}}(\tau)$.
- As for the lower bound, we will exploit the fact that the statistical halving complexity is invariant under shifts such that lower bounds on $\mathrm{SHC}_{\mathcal{F}-g,\mathcal{Q}}(\tau)$ *are* lower bounds on $\mathrm{SHC}_{\mathcal{F},\mathcal{Q}}(\tau)$.
- Shift functions g that lead to "unbiased" classes $\mathcal{F} - g$ will play a crucial role.

The remainder of this section is devoted to the statement and proof of the main result.

4.1 Central Definitions and Main Result

The following definition is tailored such that a learner with a γ-nontrivial version space is actually *forced* to "halve" it in order to achieve a correlation of at least γ with the target function. This prepares the ground for relating statistical learning and statistical halving complexity to each other.

Definition 3 (Trivial Subclasses). *We say that $\mathcal{F}' \subseteq \mathcal{F}$ is a $(\gamma, \mathcal{Q})$-*trivial *subclass of $\mathcal{F}$ if there exists a function $h \in \mathcal{Q}$ that has a 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 $(\gamma, \mathcal{Q})$-*nontrivial.

Definition 4 (Unbiased Function Classes and Admissible Functions). *The average of the functions in $\mathcal{F}$ is given by*

$$B_{\mathcal{F}} := \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} f \ .$$

$\mathcal{F}$ is called unbiased *if $B_{\mathcal{F}}(x) = 0$ for every $x \in X$. A function $g : X \to \mathbb{R}$ is called $(\gamma, \mathcal{Q})$-*admissible *for $\mathcal{F}$ if it has the form $g = B_{\mathcal{F}'}$ for a $(\gamma, \mathcal{Q})$-nontrivial subclass $\mathcal{F}'$ of $\mathcal{F}$.*

The following (obvious) result sheds some light on the significance of the preceding definitions:

Lemma 1. *1. $\mathcal{F} - B_{\mathcal{F}}$ is unbiased.*
2. For every function $h : X \to \mathbb{R}$, the correlation between h and the functions from an unbiased class (like, for example, $\mathcal{F} - B_{\mathcal{F}}$) is zero on the average.

Here comes the central new notion that allows for a characterization of strong learnability in the SQ model.

Definition 5 (Strong SQ Dimension). *The* strong SQ dimension *associated with $\mathcal{F}$ and $\mathcal{Q}$ is the function*

$$\mathrm{SQDim}^*_{\mathcal{F},\mathcal{Q}}(\gamma) := \sup_{\mathcal{F}'} \ \mathrm{SQDim}_{\mathcal{F}' - B_{\mathcal{F}'}} \ ,$$

where $\mathcal{F}'$ ranges over all $(\gamma, \mathcal{Q})$-nontrivial subclasses of $\mathcal{F}$. In addition, we define

$$\mathrm{SQDim}^{**}_{\mathcal{F},\mathcal{Q}}(\gamma) := \sup_{g} \ \mathrm{SQDim}_{\mathcal{F} - g} \geq \mathrm{SQDim}^*_{\mathcal{F},\mathcal{Q}}(\gamma) \ ,$$

where g ranges over all functions that are $(\gamma, \mathcal{Q})$-admissible for $\mathcal{F}$.[3]

With these definitions, our main result reads as follows:

Theorem 2. *Assume that all query and target functions have D-norm at most 1. A parameterized function class $\mathcal{F} = \cup_{n \geq 1} \mathcal{F}_n$ is strongly learnable in the SQ-model with query functions and final hypothesis chosen from $\mathcal{Q} = \cup_{n \geq 1} \mathcal{Q}_n$ if and only if the growth rate of $\mathrm{SQDim}^*_{\mathcal{F}_n,\mathcal{Q}_n}(1 - \varepsilon)$ is polynomially bounded in n and $1/\varepsilon$.*

[3] Throughout the paper, the default value for a supremum ranging over the empty set is 1.

4.2 A Detour on Halving Complexity and Half-Covering Numbers

As explained in the beginning of section 4, statistical halving complexity will be the right notion for our purpose because of its invariance to shifts. Here comes the formal definition:

Definition 6 (Statistical Halving Complexity). $\mathrm{SHC}_{\mathcal{F},\mathcal{Q}}(\tau)$ *denotes the smallest number of queries that is sufficient to "halve" function class $\mathcal{F}$ in the sense that at most half of the functions from $\mathcal{F}$ are consistent with the answers that are received from the oracle. We furthermore define*

$$\mathrm{SHC}^*_{\mathcal{F},\mathcal{Q}}(\tau,\gamma) := \sup_{\mathcal{F}'} \mathrm{SHC}_{\mathcal{F}',\mathcal{Q}}(\tau)$$

where $\mathcal{F}'$ ranges over all $(\gamma,\mathcal{Q})$-nontrivial subclasses of $\mathcal{F}$.

Note that a subclass might be harder to halve than the class itself. It is however obvious that, for every function $g : X \to \mathbb{R}$,

$$\mathrm{SHC}_{\mathcal{F},\mathcal{Q}}(\tau) = \mathrm{SHC}_{\mathcal{F}-g,\mathcal{Q}}(\tau) \ . \tag{5}$$

The following notion relates to statistical halving complexity pretty much as the covering number relates to statistical learning complexity:

Definition 7 (Half-Covering Number). *A set of functions $h_1,\dots,h_k \in \mathcal{Q}$ is called a* τ-half-cover *of $\mathcal{F}$ if, for at least half of the functions $f \in \mathcal{F}$, there exists an index $i \in \{1,\dots,k\}$ such that $|\langle f,h_i\rangle_D| \geq \tau$. We denote the size of the smallest τ-half-cover by* $\mathrm{HCNum}_{\mathcal{F},\mathcal{Q}}(\tau)$.

Recall that B denotes an upper bound on the D-norm of query functions. The following result is the analogue to Corollary 1:

Theorem 3. $\mathrm{SHC}_{\mathcal{F},\mathcal{Q}}(\tau) \geq \mathrm{HCNum}_{\mathcal{F},\mathcal{Q}}(\tau)$ *and*

$$\sum_{i=1}^{\mathrm{HCNum}_{\mathcal{F},\mathcal{Q}}(\tau)} \lambda_i(C_{\mathcal{F}}) \geq \frac{|\mathcal{F}|}{2}\cdot\frac{\tau^2}{B^2} \ .$$

The proof is a slight modification of Ke Yang's proof of Theorem 1. We briefly note that

$$\mathrm{SHC}_{\mathcal{F},\mathcal{Q}}(\tau) \overset{(5)}{=} \mathrm{SHC}_{\mathcal{F}-g,\mathcal{Q}}(\tau) \overset{Th.3}{\geq} \mathrm{HCNum}_{\mathcal{F}-g,\mathcal{Q}}(\tau) \ . \tag{6}$$

We will see in Lemmas 2 and 4 that SLC is closely related to SHC^*. Furthermore, SHC^* can be shown to be closely related to the following variants of half-covering numbers:

$$\mathrm{HCNum}^*_{\mathcal{F},\mathcal{Q}}(\tau,\gamma) := \sup_{\mathcal{F}'} \mathrm{HCNum}_{\mathcal{F}'-B_{\mathcal{F}'},\mathcal{Q}}(\tau)$$
$$\mathrm{HCNum}^{**}_{\mathcal{F},\mathcal{Q}}(\tau,\gamma) := \sup_{\mathcal{F}'} \sup_{g:X\to\mathbb{R}} \mathrm{HCNum}_{\mathcal{F}'-g,\mathcal{Q}}(\tau)$$

In both definitions, $\mathcal{F}'$ ranges over all $(\gamma,\mathcal{Q})$-nontrivial subclasses of $\mathcal{F}$.

Lemma 2. $\mathrm{SLC}_{\mathcal{F},\mathcal{Q}}(\tau,\gamma) \geq \mathrm{SHC}^*_{\mathcal{F},\mathcal{Q}}(\tau,\gamma) \geq \mathrm{HCNum}^{**}_{\mathcal{F},\mathcal{Q}}(\tau,\gamma)$.

Proof. The second inequality is obtained by applying (6) to every $(\gamma,\mathcal{Q})$-nontrivial subclass $\mathcal{F}' \subseteq \mathcal{F}$.
The first inequality can be seen as follows. If there exists no $(\gamma,\mathcal{Q})$-nontrivial subclass of $\mathcal{F}$, it holds because $\mathrm{SLC}_{\mathcal{F},\mathcal{Q}}(\tau,\gamma) \geq 1$. Otherwise, consider a $(\gamma,\mathcal{Q})$-nontrivial subclass $\mathcal{F}' \subseteq \mathcal{F}$ such that $\mathrm{SHC}_{\mathcal{F}',\mathcal{Q}}(\tau) = \mathrm{SHC}^*_{\mathcal{F},\mathcal{Q}}(\tau,\gamma)$. It is easily seen that

$$\mathrm{SLC}_{\mathcal{F},\mathcal{Q}}(\tau,\gamma) \geq \mathrm{SLC}_{\mathcal{F}',\mathcal{Q}}(\tau,\gamma) \geq \mathrm{SHC}_{\mathcal{F}',\mathcal{Q}}(\tau)\ ,$$

where the second inequality holds thanks to the $(\gamma,\mathcal{Q})$-non-triviality of $\mathcal{F}'$. □

In order to establish an upper bound on SHC^* in terms of HCNum^*, we have to work harder. Imagine a learner who received the information that the target function, f_*, has a correlation of at least τ with query function h. If the vast majority of concepts shared this property with f_*, the learner would make only little progress. If, however, the correlation between h and the functions of the current version space is zero on the average, then, as compensation for the functions with correlation at least τ, there will be a certain fraction of concepts that correlate negatively with h and that fall out of the version space. This explains why the fact that $\mathcal{F} - B_{\mathcal{F}}$ is unbiased will play a central role in the proof of the following result:

Lemma 3. $\mathrm{SHC}^*_{\mathcal{F},\mathcal{Q}}(\tau/4,\gamma) < 2/\tau + \mathrm{HCNum}^*_{\mathcal{F},\mathcal{Q}}(2\tau,\gamma)$.

Proof. Consider a $(\gamma,\mathcal{Q})$-nontrivial subclass $\mathcal{F}' \subseteq \mathcal{F}$ such that $\mathrm{SHC}^*_{\mathcal{F},\mathcal{Q}}(\tau/4,\gamma) = \mathrm{SHC}_{\mathcal{F}',\mathcal{Q}}(\tau/4)$. (The case that no such subclass exists is trivial.) Recall that $\mathrm{SHC}_{\mathcal{F}',\mathcal{Q}}(\tau/4) = \mathrm{SHC}_{\mathcal{F}'-B_{\mathcal{F}'},\mathcal{Q}}(\tau/4)$ such that we can focus on the halving problem for $\mathcal{F}' - B_{\mathcal{F}'}$. We proceed by case analysis.

Case 1: $\exists h \in \mathcal{Q} : |\{f \in \mathcal{F}' :\ |\langle h, f - B_{\mathcal{F}'}\rangle_D| \geq 3\tau/2\}| \geq |\mathcal{F}'|/2$.
Pass $(h,\tau/4)$ to the oracle. If it returns a number whose absolute value is smaller than than $5\tau/4$ such that the absolute value of the true correlation with the target function is smaller than $3\tau/2$, then, by our case assumption, we have halved $\mathcal{F}'$ at the expense of one query only. So let us assume that the oracle returns a number whose absolute value is at least $5\tau/4$ such that the absolute value of the true correlation c with the target function from $\mathcal{F}' - B_{\mathcal{F}'}$ must be at least τ. We may assume $c \geq \tau$ because the case $c \leq -\tau$ is symmetric. Let α denote the fraction of functions from $\mathcal{F}' - B_{\mathcal{F}'}$ whose correlation with h is at least τ such that the version space shrinks by factor α. According to Lemma 1, the average correlation between h and functions from $\mathcal{F}' - B_{\mathcal{F}'}$ is zero. It follows that $\alpha\tau + (1-\alpha)(-1) \leq 0$, or equivalently, $\alpha \leq 1/(1+\tau)$.

Case 2: $\forall h \in \mathcal{Q} : |\{f \in \mathcal{F}' :\ |\langle h, f - B_{\mathcal{F}'}\rangle_D| \geq 3\tau/2\}| < |\mathcal{F}'|/2$.
Consider the smallest (2τ)-half-cover, say $h_1,\ldots,h_k \in \mathcal{Q}$, for $\mathcal{F}' - B_{\mathcal{F}'}$. Clearly, $k \leq \mathrm{HCNum}^*_{\mathcal{F},\mathcal{Q}}(2\tau,\gamma)$. Pass $(h_1,\tau/4),\ldots,(h_k,\tau/4)$ to the oracle. If it ever returns a number whose absolute value is at least $7\tau/4$, then the sbaolute value of the true correlation with the target function is at least $3\tau/2$

and, according to our case assumption, the version space shrinks immediately by factor 1/2. If the oracle returns only numbers whose absolute values are smaller than $7\tau/4$, then the absolute values of the true correlations are smaller than 2τ and, by the definition of a (2τ)-half-cover, $\mathcal{F}' - B_{\mathcal{F}'}$ is halved after the k-th query.

The punch-line of this discussion is that, either we can shrink $\mathcal{F}' - B_{\mathcal{F}'}$ by factor $1/(1+\tau)$ at the expense of a single query, or we can halve $\mathcal{F}' - B_{\mathcal{F}'}$ at the expense of at most $\mathrm{HCNum}^*_{\mathcal{F},\mathcal{Q}}(2\tau,\gamma)$ queries. In the former case, we are not quite done with halving $\mathcal{F}' - B_{\mathcal{F}'}$ but we can iterate the procedure. Since $1/(\tau+1)^{2/\tau} < 1/2$, the theorem follows. □

Putting Lemmas 2 and 3 together, it follows that SHC* can be "sandwiched" by HCN*:

$$\mathrm{HCNum}^*_{\mathcal{F},\mathcal{Q}}(\tau,\gamma) \leq \mathrm{HCNum}^{**}_{\mathcal{F},\mathcal{Q}}(\tau,\gamma) \leq \mathrm{SHC}^*_{\mathcal{F},\mathcal{Q}}(\tau,\gamma) < \frac{1}{2\tau} + \mathrm{HCNum}^*_{\mathcal{F},\mathcal{Q}}(8\tau,\gamma) \tag{7}$$

For the purpose of sandwiching SLC by SHC*, we state the following result:

Lemma 4. $\mathrm{SLC}_{\mathcal{F},\mathcal{Q}}(\tau,\gamma-2\tau) \leq \mathrm{SHC}^*_{\mathcal{F},\mathcal{Q}}(\tau,\gamma) \cdot \lfloor \log|\mathcal{F}| \rfloor$.

Proof. As long as the current version space $\mathcal{F}'$ is not $(\gamma,\mathcal{Q})$-trivial, we can use $\mathrm{SHC}_{\mathcal{F}',\mathcal{Q}}(\tau,\gamma) \leq \mathrm{SHC}^*_{\mathcal{F},\mathcal{Q}}(\tau,\gamma)$ queries to halve it. If the current version space $\mathcal{F}'$ is $(\gamma,\mathcal{Q})$-trivial, we can use the function $h \in \mathcal{Q}$ with correlation at least γ to at least half of the functions in $\mathcal{F}'$ as query function. The value returned by the oracle will either prove that h has correlation at least $\gamma - 2\tau$ with the target function or it will halve the version space. In the former case, we can output h as the final hypothesis. In the latter case, we enter the next stage. We conclude that either we identify the target function exactly after $\lfloor \log|\mathcal{F}| \rfloor$ halving-stages or we have found a function with correlation at least $\gamma - 2\tau$ on the way. □

Combining Lemma 2 and Lemma 4, we obtain

$$\mathrm{SHC}^*_{\mathcal{F},\mathcal{Q}}(\tau,\gamma) \leq \mathrm{SLC}_{\mathcal{F},\mathcal{Q}}(\tau,\gamma) \leq \mathrm{SHC}^*_{\mathcal{F},\mathcal{Q}}(\tau,\gamma+2\tau) \cdot \lfloor \log|\mathcal{F}| \rfloor \ . \tag{8}$$

4.3 Proof of the Main Result

Theorem 2 is a direct consequence of the three lemmas below and of the previous "sandwich-results" (inequalities (7) and (8)). Lemma 5 implies that parameterized classes with a super-polynomial strong SQ dimension are not strongly learnable. Lemmas 6 and 7 imply that parameterized classes with a polynomially bounded strong SQ dimension are strongly learnable.

Lemma 5. *Let* $d := \mathrm{SQDim}^{**}_{\mathcal{F},\mathcal{Q}}(1-\varepsilon)$*. If* $\varepsilon^2 \cdot d \geq 640$*, then the following holds:*

$$\mathrm{HCNum}^{**}_{\mathcal{F},\mathcal{Q}}\left(\tau, 1 - \frac{\varepsilon}{4}\right) \geq \frac{1}{40} \cdot \tau^2 \cdot d$$

Proof. Choose a $(1-\varepsilon, \mathcal{Q})$-nontrivial subclass $\mathcal{F}' \subseteq \mathcal{F}$ and set $g := B_{\mathcal{F}'}$ such that $\mathrm{SQDim}^{**}_{\mathcal{F},\mathcal{Q}}(1-\varepsilon) = \mathrm{SQDim}_{\mathcal{F}-g}$. Furthermore choose $f_1, \ldots, f_d \in \mathcal{F}$ such that the the absolute values of the pairwise correlations of $f_1 - g, \ldots, f_d - g$ are smaller than $1/(d-1)$. The following three claims are proven in the full paper:

Claim 1: The correlation between g and any function from $\mathcal{Q}$ is bounded by $1 - \varepsilon/2$.
Claim 2: Let $\mathcal{F}'' := \{f_1, \ldots, f_d\}$. Then, $\lambda_1(C_{\mathcal{F}''-g}) < 5$.
Claim 3: $\mathcal{F}''$ is $(1-\varepsilon/4, \mathcal{Q})$-nontrivial.

We are now in the position to complete the proof of the lemma. We can conclude from the $(1-\varepsilon/4, \mathcal{Q})$-non-triviality of $\mathcal{F}''$ that $\mathrm{HCNum}^{**}_{\mathcal{F},\mathcal{Q}}(\tau, 1-\varepsilon/4) \geq \mathrm{HCNum}_{\mathcal{F}''-g,\mathcal{Q}}(\tau)$. We can furthermore conclude from Theorem 3 (with $\mathcal{F}'' - g$ in the role of $\mathcal{F}$, and $B = 2$) and Claim 2 that

$$5 \cdot \mathrm{HCNum}_{\mathcal{F}''-g,\mathcal{Q}}(\tau) > \lambda_1(C_{\mathcal{F}''-g}) \cdot \mathrm{HCNum}_{\mathcal{F}''-g,\mathcal{Q}}(\tau) \geq \frac{d}{8} \cdot \tau^2 \ .$$

Division by 5 yields the lemma. □

Lemma 6. *Let $d := \mathrm{SQDim}^{*}_{\mathcal{F},\mathcal{Q}}(1-\varepsilon)$ and assume that*[4]

$$\tilde{\mathcal{Q}} := \{f - B_{\mathcal{F}'} :\ f \in \mathcal{F}, \mathcal{F}' \subseteq \mathcal{F}\} \subseteq \mathcal{Q} \ . \tag{9}$$

Then, the following holds:

$$\mathrm{HCNum}^{*}_{\mathcal{F},\mathcal{Q}}\left(\frac{1}{d}, 1-\varepsilon\right) \leq d$$

Proof. Choose a $(1-\varepsilon, \mathcal{Q})$-nontrivial subclass $\mathcal{F}' \subseteq \mathcal{F}$ and set $g := B_{\mathcal{F}'}$ such that

$$\mathrm{HCNum}^{*}_{\mathcal{F},\mathcal{Q}}\left(\frac{1}{d}, 1-\varepsilon\right) = \mathrm{HCNum}_{\mathcal{F}'-g,\mathcal{Q}}\left(\frac{1}{d}\right) \ .$$

Let $d' := \mathrm{SQDim}_{\mathcal{F}'-g}$. Since $\mathcal{F}'$ is $(1-\varepsilon, \mathcal{Q})$-nontrivial, we can conclude that $d' \leq d$. Furthermore, by assumption (9), $\mathcal{F}' - g \subseteq \tilde{\mathcal{Q}} \subseteq \mathcal{Q}$. The proof is now completed as follows:

$$\mathrm{HCNum}_{\mathcal{F}'-g,\mathcal{Q}}\left(\frac{1}{d}\right) \leq \mathrm{CNum}_{\mathcal{F}'-g,\mathcal{Q}}\left(\frac{1}{d}\right) \leq \mathrm{CNum}_{\mathcal{F}'-g,\mathcal{Q}}\left(\frac{1}{d'}\right) \overset{(4)}{\leq} d' \leq d$$

□

Assumption (9) is not essential for the following reason: if $\mathcal{F}$ is strongly learnable in the SQ model with query functions and final hypothesis chosen from $\mathcal{Q} \cup \tilde{\mathcal{Q}}$, then $\mathcal{F}$ is already strongly learnable in the SQ model with query functions and final hypothesis chosen from $\mathcal{Q}$. This is an immediate consequence of the following

[4] Assumption (9) looks strange. See however Lemma 7 below for a justification.

Lemma 7. *A query of the form* $(\tilde{h}, \tau)$ *such that* $\tilde{h} \in \tilde{\mathcal{Q}}$ *can be simulated by means of* $O(1/\tau^2)$ *queries of the form* $(h, \tau/3)$ *such that* $h \in \mathcal{F}$.

Proof. According to the definition of $\tilde{\mathcal{Q}}$, $\tilde{h} = f - B_{\mathcal{F}'}$ for some function $f \in \mathcal{F}$ and a subclass $\mathcal{F}' \subseteq \mathcal{F}$. Let f_* denote the unknown target function. Query $(f, \tau/3)$ provides us with a $\tau/3$-approximation $\hat{c}_1$ of $\langle f, f_* \rangle_D$. $\langle B_{\mathcal{F}'}, f_* \rangle_D$ is the average correlation between the target function and a function from $\mathcal{F}'$. Hoeffding bounds imply that there exist $O(1/\tau^2)$ functions in $\mathcal{F}'$ such that their average correlation with f_* equals $\langle B_{\mathcal{F}'}, f_* \rangle_D$ up to an additive term $\tau/3$. If we get to know these $O(1/\tau^2)$ correlations only up to another additive term $\tau/3$ (the tolerance parameter), the resulting estimation, say $\hat{c}_2$, for $\langle B_{\mathcal{F}'}, f_* \rangle_D$ will be only a $2\tau/3$-approximation. Clearly, $\hat{c}_1 - \hat{c}_2$ is a τ-approximation of $\langle \tilde{h}, f_* \rangle_D$. This shows that $O(1/\tau^2)$ queries with tolerance $\tau/3$ and query functions from $\mathcal{F}$ are sufficient to simulate the query $(\tilde{h}, \tau)$. $\square$

Our results imply the following lower and upper bound on the statistical halving complexity in terms of the strong SQ dimension $d := \mathrm{SQDim}^*_{\mathcal{F},\mathcal{Q}}(1-\varepsilon)$:

$$\begin{array}{rcccc} \mathrm{SHC}^*_{\mathcal{F},\mathcal{Q}}\left(\tau, 1-\frac{\varepsilon}{4}\right) & \overset{L.2}{\geq} & \mathrm{HCNum}^{**}_{\mathcal{F},\mathcal{Q}}\left(\tau, 1-\frac{\varepsilon}{4}\right) & \overset{L.5}{\geq} & \frac{1}{40} \cdot \tau^2 \cdot d \\ \mathrm{SHC}^*_{\mathcal{F},\mathcal{Q}}\left(\frac{1}{8d}, 1-\varepsilon\right) & \overset{L.3}{<} & 4d + \mathrm{HCNum}^*_{\mathcal{F},\mathcal{Q}}\left(\frac{1}{d}, 1-\varepsilon\right) & \overset{L.6}{\leq} & 5 \cdot d \end{array}$$

The inequality marked "L.6" is only valid under assumption (9). Without this assumption, we have to replace $5 \cdot d$ by $O(d/\tau^2)$. The gap between the lower and the upper bound is not tremendous but it would be interesting to know whether it can be narrowed.

Acknowledgments: Thanks to Heikki Mannila, Esko Ukkonen, and Aristides Gionis for helpful discussions. Thanks to the anonymous referees for their comments and suggestions, for pointing my attention to the paper by Köbler and Lindner about the general dimension (and for fixing a flaw in the first version of the paper).

References

1. José L. Balcázar, Jorge Castro, and David Guijarro. A new abstract combinatorial dimension for exact learning via queries. *Journal of Computer and System Sciences*, 64(1):2–21, 2002.
2. Shai Ben-David, Alon Itai, and Eyal Kushilevitz. Learning by distances. *Information and Computation*, 117(2):240–250, 1995.
3. Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishai Mansour, and Steven Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Proceedings of the 26th Annual Symposium on Theory of Computing*, pages 253–263, 1994.
4. Nader H. Bshouty and Vitaly Feldman. On using extended statistical queries to avoid membership queries. *Journal of the Machine Learning Research*, 2:359–395, 2002.

5. Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the Association on Computing Machinery*, 45(6):983–1006, 1998.
6. Johannes Köbler and Wolfgang Lindner. A general dimension for approxmately learning boolean functions. In *Proceedings of the 13th International Conference on Algorithmic Learning Theory*, pages 139–148, 2002.
7. Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
8. Ke Yang. New lower bounds for statistical query learning. *Journal of Computer and System Sciences*, 70(4):485–509, 2005.