Spectral Norm in Learning Theory: Some Selected Topics *

Hans Ulrich Simon

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

Abstract. In this paper, we review some known results that relate the statistical query complexity of a concept class to the spectral norm of its correlation matrix. Since spectral norms are widely used in various other areas, we are then able to put statistical query complexity in a broader context. We briefly describe some non-trivial connections to (seemingly) different topics in learning theory, complexity theory, and cryptography. A connection to the so-called Hidden Number Problem, which plays an important role for proving bit-security of cryptographic functions, will be discussed in somewhat more detail.

1 Introduction

Kearns' Statistical Query (SQ) model [8] is an elegant abstraction from Valiant's PAC learning model [15].[1] 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 function classes known to be efficiently learnable in the PAC learning model. This had raised the question of whether any concept class that is efficiently learnable by a noise-tolerant learner in the PAC learning model might already be efficiently learnable in the SQ model. This question was (at least partially) answered to the negative by Blum, Kalai, and Wasserman [4] who presented a concept class that has an efficient noise-tolerant[2] PAC learner but (provably) has no efficient SQ learner. However, classes that distinguish between the model of noise-tolerant PAC learning and SQ learning

* 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.

[1] The model of "Learning by Distances" [1] seems to be equivalent to the SQ model. For our purpose, the notation within the SQ model is more convenient.

[2] for noise rate bounded away from 1/2

seem to be extremely rare. Thus, the non-existence of an efficient SQ learner often indicates that finding an efficient and noise-tolerant PAC learner might also be hard to achieve.

Information-theoretic lower bounds in the PAC model are stated in terms of the VC-dimension. Information-theoretic lower bounds in the in the SQ model have a much different (more algebraic) flavor. Blum et al. [3] have shown that, roughly speaking, the number of statistical queries being necessary and sufficient 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 [18] presented an alternative (stronger but polynomially equivalent) lower bound on the number of statistical queries. It is stated in terms of the eigenvalues of the correlation matrix C associated with the concept class. A simplified variant of this lower bound is stated in terms of the spectral norm (= largest eigenvalue) of C. There exist classes which are weakly learnable in the SQ model, but not strongly. Ke Yang [17, 18] invented a technique that allows to prove statements of this type (at least occasionally for specific classes). A variant of this technique [19] yields lower bounds in the so-called model of SQ Sampling (related to quantum computing). In this model, the goal of the learner is to find a positive example of the unknown target concept (again by means of statistical queries).

In the first part of this paper (up to section 5), we partially review the work about SQ learning in [3, 17–19]. In order to present the results in the technically simplest fashion, we make use of the equivalence between the SQ and the so-called Correlation Query (CQ) model that replaces statistical queries by correlation queries. Correlation queries deal with predicates on instances (whereas statistical queries deal with predicates on *labeled* instances). Upon a correlation query, the CQ oracle will return an estimate for the correlation between this predicate and the target concept. The equivalence between the two models was proven explicitly by Bshouty and Feldman [2] (but has been used implicitly in predecessor papers already). For sake of completeness, we repeat the short proof in section 3. The CQ model will be much more convenient for our purposes.

In the second part of the paper (beginning in section 6), we briefly describe some non-trivial connections to (seemingly) different topics in learning theory, complexity theory, and cryptography. A connection to the so-called Hidden Number Problem, which plays an important role for proving bit-security of cryptographic functions, will be discussed in somewhat more detail.

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 $h : X \to \{\pm 1\}$ (not necessarily from $\mathcal{F}$) is called a *hypothesis*. We say that h is *ε-accurate for f* if $\Pr_D[h(x) \neq f(x)] \leq \varepsilon$. Informally, the goal of a learner for concept class $\mathcal{F}$ is to infer an ε-accurate hypothesis for an unknown *target concept* $f \in \mathcal{F}$ from "partial information". There must be a uniform learning algorithm for all choices of $f \in \mathcal{F}$ and $\varepsilon > 0$ (and sometimes, but not in this paper, also for all choices of D and other relevant parameters). The formal notion of "partial information" depends on the learning model in consideration. We assume the reader to be familiar with the PAC learning model [15] where the information given to the learner consists of random examples for the unknown target concept. We will be mainly concerned with the SQ model that is outlined in the next paragraph.

Statistical Query Learning: In this model, the learner has access to random examples only indirectly through statistical queries that are answered by an oracle. A statistical query is of the form $\mathrm{SQ}(\hbar, \tau)$ where $\hbar : X \times \{\pm 1\} \to \{\pm 1\}$ denotes a binary predicate on labeled instances and $\tau > 0$ is a *tolerance parameter*. Upon such a query, the oracle returns a *τ-approximation for* $\mathbb{E}_D[\hbar(x, f(x)]$, i.e., it returns a number d that satisfies

$$\mathbb{E}_D[\hbar(x, f(x)] - \tau \leq d \leq \mathbb{E}_D[\hbar(x, f(x)] + \tau \enspace .$$

We briefly note that

$$\mathbb{E}_D[\hbar(x, f(x)] = \Pr_D[\hbar(x, f(x)) = +1] - \Pr_D[\hbar(x, f(x)) = -1] \enspace .$$

We will measure the "efficiency" of the learner by the number of statistical queries, say q, that it passes to the oracle (in the worst-case), by the smallest tolerance parameter τ that is ever used during learning, and by the accuracy ε of the final hypothesis. It was shown by Kearns [8] that an SQ learner can be simulated by a noise-tolerant PAC learner that has access to $\mathrm{poly}(q, 1/\tau, 1/\varepsilon)$ random examples.

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

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

With a concept class $\mathcal{F}$, we associate the *incidence matrix* $M \in \{\pm 1\}^{\mathcal{F} \times X}$ given by $M[f, x] := f(x)$ and the *correlation matrix* $C \in [-1, 1]^{\mathcal{F} \times \mathcal{F}}$ given by $C[f_1, f_2] := \langle f_1, f_2 \rangle_D$. Clearly, C is positive semidefinite. Let $\mathrm{diag}(D) \in [0, 1]^{X \times X}$ denote the matrix with values $D(x)$ on the main diagonal and zeroes elsewhere. Then matrices C and M satisfy the relation $C = M \cdot \mathrm{diag}(D) \cdot M^\top$. In particular, we have $C = \frac{1}{|X|} \cdot M \cdot M^\top$ if D is the uniform distribution.

Warning: We will use the analogous notations (and analogous remarks are valid) for the more general case, where $\mathcal{F}$ contains real-valued functions.

Correlation Query Learning: The definition of Correlation Query (CQ) learning is completely analogous to the definition of SQ learning with the only difference that we replace the SQ oracle by a CQ oracle. A correlation query is of the form $\mathrm{CQ}(h, \tau)$ for some function $h : X \to \{-1, 0, +1\}$ and $\tau > 0$. The oracle will return a τ-approximation of $\langle h, f \rangle_D$, where (as always) $f \in \mathcal{F}$ denotes the unknown target concept.

Warning: We will use this model (for technical reasons) also in the more general case where $\mathcal{F}$ contains real-valued functions.

Lower Bounds and Adversaries: The emphasis of this paper is on lower bounds. A well-known adversary argument for proving lower bounds is as follows. An adversary of the learner runs the learning algorithm, waits for queries, answers them in a malicious fashion and keeps track of the so-called *version space*. The latter, by definition, consists of all target concepts being consistent with all answers that have been returned so far to the learner. Intuitively, the adversary tries to keep the version space as "rich" as possible in order to slow down the progress made by the learner.
In order to make the lower bounds as strong as possible, we do not impose unnecessary restrictions on the learner. In particular, our lower bounds will be valid even in the following setting:

- The learner may output arbitrary functions $h : X \to \{-1, 0, +1\}$ as hypotheses.
- The learning algorithm need not be uniform w.r.t. domain distributions (since there is a fixed distribution D).
- The learner is already considered successful when its final hypothesis has a significant correlation, say γ, with the target concept. Note that a hypothesis whose correlation with f is γ is $(1/2 - \gamma/2)$-accurate for f, i.e., it has advantage $\gamma/2$ over random guessing.

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 $\|u\|$. For a matrix M, $\|M\|$ denotes its spectral norm:

$$\|M\| = \sup_{u:\|u\|\leq 1} \|Mu\|$$

It is well-known that $\|M\|$ coincides with the largest singular value of M. If M is symmetric and positive semidefinite, then the largest singular value coincides with the largest eigenvalue. A matrix of the form $M \cdot M^\top$ (symmetric and positive semidefinite !) satisfies $\|M \cdot M^\top\| = \|M\|^2$. In particular, if M is an incidence matrix for a concept class $\mathcal{F}$ and $C = \frac{1}{|X|} M \cdot M^\top$ is the corresponding correlation matrix under the uniform distribution, we get the relation $\|C\| = \frac{1}{|X|}\|M\|^2$.

3 Statistical Queries versus Correlation Queries

As explained in the introduction (and proven in [2]) already, statistical queries and correlation queries are equivalent from an information-theoretic point of view: for any query of one kind, there is a query of the other kind that reveals precisely the same amount of information. This equivalence can be seen as follows. Consider a function $\hbar : X \times \{\pm 1\} \rightarrow \{\pm 1\}$. It induces the following split of the instance space:

$$X_0(\hbar) = \{x \in X \mid \hbar(x, +1) = \hbar(x, -1)\}$$
$$X_1(\hbar) = \{x \in X \mid \hbar(x, +1) \neq \hbar(x, -1)\}$$

Loosely speaking, $X_1(\hbar)$ contains the instances on which predicate $\hbar$ is "label-sensitive" whereas $\hbar$ is insensitive to labels of instances from $X_0(\hbar)$. With $\hbar$ we associate the following function $h : X \rightarrow \{-1, 0, +1\}$:

$$h(x) = \begin{cases} 0 & \text{if } x \in X_0(\hbar) \\ \hbar(x,1) & \text{if } x \in X_1(\hbar) \end{cases}$$

Note that, for every $x \in X_1$ and every $b = \pm 1$,

$$\hbar(x, b) = b \cdot \hbar(x, 1) = b \cdot h(x) \ .$$

Now $\mathbb{E}_D[\hbar(x, f(x))]$ can be written in terms of $\langle h, f \rangle_D$ as follows:

$$\begin{aligned} \mathbb{E}_D[\hbar(x, f(x))] &= \sum_{x \in X_0(\hbar)} D(x)\hbar(x, f(x)) + \sum_{x \in X_1(\hbar)} D(x)\hbar(x, f(x)) \\ &= \underbrace{\sum_{x \in X_0(\hbar)} D(x)\hbar(x, 1))}_{=:k(\hbar)} + \sum_{x \in X_1(\hbar)} D(x)f(x)h(x) \\ &= k(\hbar) + \langle h, f \rangle_D \ , \end{aligned}$$

where $k(\hbar)$ depends on $\hbar$ only (and not on the target concept f).

Notice that the mapping $\hbar \mapsto h$ is surjective since any mapping $h : X \rightarrow \{-1, 0, +1\}$ has a pre-image, for instance the mapping $\hbar : X \times \{\pm 1\} \rightarrow \{\pm 1\}$ given by

$$\hbar(x, b) = \begin{cases} 1 & \text{if } h(x) = 0 \\ bh(x) & \text{otherwise} \end{cases} \ .$$

We conclude from these considerations, and in particular from the relation $\mathbb{E}_D[\hbar(x, f(x))] = k(\hbar) + \langle h, f \rangle_D$ that there are mutual simulations between an SQ-oracle and a CQ-oracle such that answers to $\mathrm{SQ}(\hbar, \tau)$ and to $\mathrm{CQ}(h, \tau)$ provide the same amount of information. Thus, we arrive at the following result:

Theorem 1. *There exists an algorithm that finds a ε-accurate hypothesis for every target concept $f \in \mathcal{F}$ by means of q statistical queries whose tolerance parameters are lower-bounded by τ respectively if and only if there exists an algorithm that finds a ε-accurate hypothesis for every target concept $f \in \mathcal{F}$ by means of q correlation queries whose tolerance parameters are lower-bounded by τ respectively.*

4 A Lower Bound in Terms of the Spectral Norm

In the sequel, we say "function class" (as opposed to "concept class") when we allow real function values different from ± 1.

Theorem 2 ([18]). *Consider a finite function class $\mathcal{F}$ (with correlation matrix C) and a learning algorithm for $\mathcal{F}$ in the CQ model with the following features:*

- *It outputs a hypothesis h satisfying $\langle h, f\rangle_D > \gamma$, where f denotes the target concept.*
- *It makes at most q statistical queries none of which uses a tolerance parameter smaller than τ.*

Let $\lambda_1(C) \geq \cdots \geq \lambda_{|\mathcal{F}|}(C)$ be the eigenvalues of C. Then,

$$\sum_{i=1}^{q+1} \lambda_i(C) \geq |\mathcal{F}| \cdot \min\{\gamma^2, \tau^2\} \ . \tag{1}$$

Proof. We basically present the proof of Ke Yang [18] (modulo some slight simplifications resulting from the more convenient CQ model).[3] Consider an adversary that returns 0 upon correlation queries as long as this does not lead to an empty version space. Choose $q' \leq q$ maximal such that the first q' queries of the learner, say $\mathrm{CQ}(h_1, \tau_1), \mathrm{CQ}(h_2, \tau_2), \ldots, \mathrm{CQ}(h_{q'}, \tau_{q'})$, are answered 0. Let $h_{q'+1}$ denote the query function of the next correlation query if $q > q'$, and let $h_{q'+1}$ denote the final hypothesis of the learner if $q = q'$. Let $\mathcal{V} \subseteq \mathcal{F}$ denote the version space resulting after "queries" with query functions $h_1, \ldots, h_{q'}, h_{q'+1}$. By definition of q', $\mathcal{V}$ is empty if $q > q'$. Let $\mathcal{Q}$ denote the (at most) $q'+1$-dimensional vector space spanned by $h_1, \ldots, h_{q'}, h_{q'+1}$. For every function $f \in \mathcal{F}$, the following holds:

- If $f \in \mathcal{V}$, then $q = q'$. Since $h_{q'+1}$ is the final hypothesis of the learner, it follows that $\langle h_{q'+1}, f\rangle_D \geq \gamma$.
- If $f \notin \mathcal{V}$, then f was eliminated after one of the first $q'+1$ "queries". Thus there exists some $i \in \{1, \ldots, q'+1\}$ such that $\langle h_i, f\rangle_D \geq \tau$.

We use $f^{\mathcal{Q}}$ to denote the projection of f into subspace $\mathcal{Q}$ and conclude that

$$\sum_{f \in \mathcal{F}} \|f^{\mathcal{Q}}\|^2 \geq |\mathcal{F}| \cdot \min\{\gamma^2, \tau^2\} \ . \tag{2}$$

On the other hand, it is a well-known fact from matrix theory that

$$\sum_{f \in \mathcal{F}} \|f^{\mathcal{Q}}\|^2 \leq \sum_{i=1}^{q'+1} \lambda_i \leq \sum_{i=1}^{q+1} \lambda_i \ . \tag{3}$$

Solving (2) and (3) for $\sum_{i=1}^{q+1} \lambda_i$ yields the result. □

[3] He considered only classes $\mathcal{F}$ with ± 1-valued functions, but his argument holds in the more general case of real-valued functions. The adversary technique applied in the proof goes back to [3].

The following result is immediate from Theorem 2 and the fact that $\mathcal{F}$ is not easier to learn than a subclass $\mathcal{F}' \subseteq \mathcal{F}$ (with correlation matrix C').

Corollary 1. *The number q of queries needed to learn $\mathcal{F}$ in the sense of Theorem 2 satisfies the following condition:*

$$\forall \mathcal{F}' \subseteq \mathcal{F} : \sum_{i=1}^{q+1} \lambda_i(C') \geq |\mathcal{F}'| \cdot \min\{\gamma^2, \tau^2\} \ .$$

Since the spectral norm of C, $\|C\|$, coincides with the largest eigenvalue, λ_1, (1) implies the following inequality:

$$q \geq \frac{|\mathcal{F}| \cdot \min\{\gamma^2, \tau^2\}}{\|C\|} - 1 \ . \tag{4}$$

Define

$$L(\mathcal{F}) := \sup_{\mathcal{F}' \subseteq \mathcal{F}} \frac{|\mathcal{F}'|}{\|\mathrm{C}'\|} \ , \tag{5}$$

where C' denotes the correlation matrix associated with $\mathcal{F}'$. If D is the uniform distribution on X and M' denotes the incidence matrix of $\mathcal{F}'$, we have $\|C'\| = \frac{1}{|X|}\|M'\|^2$ and can rewrite (5) as follows:

$$L(\mathcal{F}) = \sup_{\mathcal{F}' \subseteq \mathcal{F}} \frac{|\mathcal{F}'| \cdot |X|}{\|M'\|^2} \ . \tag{6}$$

Analogously to Corollary 1, we obtain

Corollary 2. $q \geq L(\mathcal{F}) \cdot \min\{\gamma^2, \tau^2\} - 1.$ [4]

5 Related Lower Bounds

The lower bounds from section 4 are tailored to weak learning in the SQ model. These bounds get trivial when a class is efficiently weakly, but not strongly, learnable. Ke Yang [17, 18] presented a technique that allows to reduce the problem of (better than weakly) learning a class consisting of "(almost) uniformly correlated" concepts to the problem of weakly learning a corresponding class of "(almost) uncorrelated" concepts. To the latter class, the bounds from section 4 *do* apply. In this section, we present another (and even older) technique that is slightly simpler and fits the same purpose.[5] Furthermore, both techniques (Yang's and the technique from this section) lead to lower bounds in the so-called SQ Sampling model.

Throughout this section, D is the uniform distribution on X (and we omit index D in the notation of probabilities, expectations, and inner products). Consider a concept class $\mathcal{F}$ with the following properties:

[4] Even a slightly stronger bound of the form $q \geq L(\mathcal{F})\tau^2 - (\tau^2/\gamma^2)$ is valid (and proven in the full paper).

[5] A nice feature of Yang's technique is that it applies even when the SQ oracle is assumed as "honest". We conjecture that the alternative technique (described in this section) applies to honest oracles as well.

Property 1 There exists a constant $0 < \rho < 1/2$ such that, for every $f \in \mathcal{F}$, $\Pr[f(x) = 1] = \rho$.

Property 2 There exists a constant s such that the matrix M_s, given by

$$M_s[f, x] = \begin{cases} s & \text{if } f(x) = 1 \\ -1 & \text{if } f(x) = -1 \end{cases} ,$$

has pairwise orthogonal rows.

It is easy to see that s and ρ must satisfy the relation $s \geq (1 - 2\rho)/\rho$. It is furthermore easy to infer from property 1 that the constant function $\mathbf{-1}$ is positively correlated with every concept from $\mathcal{F}$:

$$\forall f \in \mathcal{F} : \langle f, \mathbf{-1} \rangle = (1 - \rho) - \rho = 1 - 2\rho \ .$$

We will pursue the question how many queries it takes to find a hypothesis whose correlation with the target concept is significantly greater than $1 - 2\rho$.

The first important observation is that we can consider M_s as the incidence matrix of the function class $\mathcal{F}_s = \{f_s | \ f \in \mathcal{F}\}$ where f_s is given by

$$f_s(x) = \begin{cases} s & \text{if } f(x) = 1 \\ -1 & \text{if } f(x) = -1 \end{cases} .$$

Since M_s has pairwise orthogonal rows and every row vector has squared Euclidean length $\rho|X|s^2 + (1 - \rho)|X|$, the correlation matrix $C_s = \frac{1}{|X|} \cdot M_s M_s^\top$ satisfies

$$C_s = \mathrm{diag}(\rho s^2 + 1 - \rho, \ldots, \rho s^2 + 1 - \rho) \ .$$

Thus $\|C_s\| = \rho s^2 + 1 - \rho$. Note that Corollary 2 applies to $\mathcal{F}_s$ and leads to a lower bound of the form

$$\frac{|\mathcal{F}|}{\rho s^2 + 1 - \rho} \min\{\gamma^2, \tau^2\} - 1 \tag{7}$$

with the usual meaning of γ and τ.

The second important observation is that the problems of learning $\mathcal{F}$ and $\mathcal{F}_s$ by correlation queries exhibit a close relationship. For a query function g, consider the following calculation:

$$\begin{aligned} \mathbb{E}[f(x)g(x)] &= \mathbb{E}[f(x)g(x)|f(x) = 1] \cdot \rho + \mathbb{E}[f(x)g(x)|f(x) = -1] \cdot (1 - \rho) \\ &= \mathbb{E}[g(x)|f(x) = 1] \cdot \rho - \mathbb{E}[g(x)|f(x) = -1] \cdot (1 - \rho) \\ \mathbb{E}[f_s(x)g(x)] &= \mathbb{E}[g(x)|f(x) = 1] \cdot \rho s - \mathbb{E}[g(x)|f(x) = -1] \cdot (1 - \rho) \\ \mathbb{E}[g(x)] &= \mathbb{E}[g(x)|f(x) = 1] \cdot \rho + \mathbb{E}[g(x)|f(x) = -1] \cdot (1 - \rho) \end{aligned}$$

Now an easy calculation shows that

$$\mathbb{E}[f(x)g(x)] - \frac{s-1}{s+1}\mathbb{E}[-g] = \frac{2}{s+1}\mathbb{E}[f_s(x)g(x)] \ . \tag{8}$$

According to (8), the following holds:

Lemma 1. *g has a correlation of at least $\alpha + \frac{s-1}{s+1}\mathbb{E}[-g]$ with f iff g has a correlation of at least $\frac{s+1}{2}\alpha$ with f_s.*

Since $\mathbb{E}[-g] \leq 1$ (with equality for $g = \mathbf{-1}$), we obtain

Corollary 3. *If g has a correlation of at least $\alpha + \frac{s-1}{s+1}$ with f, then g has a correlation of at least $\frac{s+1}{2}\alpha$ with f_s.*

Since the latter two results are valid for tolerance parameter τ in the role of α and for final correlation γ in the role of α, we get

Corollary 4. *The number of correlation queries (with smallest tolerance τ) needed to achieve a correlation of at least $\frac{s-1}{s+1} + \gamma$ with an unknown target concept from $\mathcal{F}$ is not smaller than the number of correlation queries (with smallest tolerance $\frac{s+1}{2}\tau$) needed to achieve a correlation of at least $\frac{s+1}{2}\gamma$ with an unknown target function from $\mathcal{F}_s$.*

An application of the lower bound in (7) with $\frac{s+1}{2}\tau$ in the role of τ and $\frac{s+1}{2}\gamma$ in the role of γ finally leads to

Corollary 5. *The number of correlation queries (with smallest tolerance τ) needed to achieve a correlation of at least $\frac{s-1}{s+1}+\gamma$ with an unknown target concept from $\mathcal{F}$ is at least*

$$\frac{|\mathcal{F}|}{\rho s^2 + 1 - \rho} \frac{(s+1)^2}{4} \min\{\gamma^2, \tau^2\} \ .$$

Note $\frac{(s+1)^2}{4} \min\{\gamma^2, \tau^2\} \leq 1$ since $\frac{s-1}{s+1} + \gamma \leq 1$.

Here is a concrete example[6] to which Corollary 5 applies. Remember that the elements (projective points) of the $(n-1)$-dimensional projective space over $\mathbb{Z}_p$ are the 1-dimensional linear subspaces of $\mathbb{Z}_p^n$. We will represent projective points by elements in $\mathbb{Z}_p^n$. We say that a projective point Q is *orthogonal* to a projective point Q', denoted as $Q \perp Q'$, if $\langle Q, Q' \rangle = 0$. We view the matrix M such that

$$M[Q, Q'] = \begin{cases} 1 & \text{if } Q \perp Q' \\ -1 & \text{otherwise} \end{cases}$$

as the incidence matrix of a concept class $\mathrm{ORT}(p, n)$ (over domain $X = \mathrm{ORT}(p, n)$). According to results in [12], $\mathrm{ORT}(p, n)$ has properties 1 and 2, where

$$\rho = \frac{p^{n-1} - 1}{p^n - 1} \approx \frac{1}{p} \text{ and } s = \frac{(p-1)p^{n/2-1}}{1 + p^{n/2-1}} \approx p \ .$$

Combining this with $|\mathrm{ORT}(p, n)| = \frac{p^n-1}{p-1}$ and with Corollary 5, we get

Corollary 6. *The number of correlation queries (with smallest tolerance τ) needed to achieve a correlation of at least $1 - 2\frac{p^{n/2-1}}{p^{n/2}+1} + \gamma$ with an unknown target concept from $\mathcal{F}$ is asymptotically at least $p^{n-2}(p^2/4)\min\{\gamma^2, \tau^2\}$.*

Note that $\gamma \leq 2\frac{p^{n/2-1}}{p^{n/2}+1} \approx 2/p$ such that $(p^2/4)\min\{\gamma^2, \tau^2\}$ is asymptotically at most 1.

[6] taken from [12] and used in connection with half-space embeddings in [7]

A Note on the SQ Sampling Model: A query in the SQ Sampling model has the same form as a query in the CQ model but is answered by a τ-approximation for $\mathbb{E}[g(x)|f(x) = 1]$. In the SQ sampling model, the learner pursues the goal to find a positive example for the unknown target concept. Blum and Yang [19] showed that the technique of Yang from [17, 18] leads to lower bounds in the SQ sampling model (when properly applied). The same remark is valid for the alternative technique that we have used in this section. It can be shown that classes with properties 1 and 2 are "hard" in the SQ Sampling model. For example, the retrieval of a positive example for an unknown concept from $\mathrm{ORT}(p, n)$ requires exponentially many queries. The corresponding results and proofs are found in the full paper. Here, we give only the equation that plays the the same key role for the SQ Sampling model as equation (8) for the SQ model:

$$\mathbb{E}[g(x)|f(x) = 1] - \frac{1}{\rho(s+1)}\mathbb{E}[g] = \frac{1}{\rho(s+1)}\mathbb{E}[f_s(x)g(x)] \ .$$

6 Characterizations of Statistical Query Learnability

As for this section, we pass to the parameterized scenario, where we consider ensembles of domains, distributions, and concept classes, respectively:

$$X = (X_n)_{n\geq 1},\ D = (D_n)_{n\geq 1},\ \text{and}\ \mathcal{F} = (\mathcal{F}_n)_{n\geq 1} \ .$$

We furthermore focus on "weak polynomial" learners, i.e., the number of queries, q, the inverse of required correlation with the target concept, $1/\gamma$, and the inverse of the lower bound on the tolerance parameters, $1/\tau$, should be bounded by a polynomial in n, respectively.

Blum et al. [3] have shown that $\mathcal{F}$ has a weak polynomial learner in the SQ model if and only if $\mathrm{SQDim}(\mathcal{F}_n)$ is bounded by a polynomial in n, where $\mathrm{SQDim}(\mathcal{F}_n)$ denotes the so-called SQ dimension. In the full paper, we show that $L(\mathcal{F}_n)$ is polynomially related to $\mathrm{SQDim}(\mathcal{F}_n)$. Thus:

Corollary 7. *The following statements are equivalent:*

1. $\mathcal{F}$ *admits a weak polynomial learner in the SQ model.*
2. $\mathrm{SQDim}(\mathcal{F}_n)$ *is polynomially bounded in* n*.*
3. $L(\mathcal{F}_n)$ *is polynomially bounded in* n*.*

We close this section by listing some rather surprising connections between SQ learning and (seemingly) different questions in learning and complexity theory, respectively:

Corollary 8. *There is a weak polynomial SQ learner for* $\mathcal{F} = (\mathcal{F}_n)_{n\geq 1}$ *under the uniform distribution if at least one of the following conditions is satisfied:*

- *There exists a* $\mathrm{poly}(n)$*-dimensional half-space embedding for* $\mathcal{F}_n$*.*
- *There exists a half-space embedding for* $\mathcal{F}_n$ *that achieves a margin whose inverse is polynomially bounded in* n*.*

– *The probabilistic communication complexity of the evaluation problem for $\mathcal{F}_n$ in the unbounded error model is logarithmically bounded in n.*
– *The evaluation problem for $\mathcal{F}_n$ can be solved by a depth-2 threshold circuit whose size and whose weights associated with nodes on the hidden layer are polynomially bounded in n (but the weights may be arbitrary at the top gate).*

Proof. It suffices to show that all conditions mentioned in the corollary will be violated if there is no weak polynomial learner for $\mathcal{F}$. According to Corollary 7, the non-existence of a weak polynomial learner implies that $L(\mathcal{F}_n)$ is super-polynomial in n. Since we assume a uniform distribution, we may apply (6) and write $L(\mathcal{F}_n)$ in the form $\frac{|\mathcal{F}_n|\cdot|X_n|}{\|\mathrm{M}_n\|^2}$. The proof can now be completed by calling the following facts into mind:

– It was shown by Forster [6] that $\frac{\sqrt{|\mathcal{F}_n|\cdot|X_n|}}{\|\mathrm{M}_n\|}$ is a lower bound on the smallest dimension and on the inverse of largest margin that can be achieved by a half-space embedding for $\mathcal{F}_n$.
– It is well-known [14] that the probabilistic communication complexity in the unbounded error model coincides (up to rounding) with the logarithm of the smallest Euclidean dimension that allows for a half-space embedding.
– It is furthermore well-known [7] that the smallest size of a threshold circuit (of the type described above) grows linearly with the smallest dimension that allows for a half-space embedding. □

7 Hardness of learning a "Hidden Number"

In the final section, we outline a relation between learning and the concept of bit-security in cryptography.

7.1 Hidden Number Problem and Bit Security

Let p be an n-bit prime, $(\mathbb{Z}_p, +, \cdot)$ be the field of integers modulo p, $(\mathbb{Z}_p^*, \cdot)$ be the (cyclic) group of prime residuals modulo p, and g be a generator of $\mathbb{Z}_p^*$. We will identify $\mathbb{Z}_p$ with $\{0, 1, \ldots, p-1\}$, the smallest residuals modulo p, and $\mathbb{Z}_p^*$ with $\{1, \ldots, p-1\}$.

Loosely speaking, the Hidden Number Problem (HNP) is the problem of inferring a hidden number $u \in \mathbb{Z}_p^*$ by means of an oracle that provides us with information about $u \cdot z$ for (known) random instances $z \in \mathbb{Z}_p^*$.[7]
In order to cast HNP as a learning problem, consider a binary predicate $B : \mathbb{Z}_p^* \rightarrow \{\pm 1\}$. Now we view the hidden number $u \in \mathbb{Z}_p^*$ as an unknown target concept that assigns label $B(u \cdot z)$ to instance z. We denote the resulting concept class $\mathrm{HNP}[B] = (\mathrm{HNP}[B]_n)_{n \geq 1}$ where n always denotes the bit-length of prime p.
The Hidden Number Problem is closely related to the cryptographic concept of bit security, a connection that we now outline briefly for motivational purposes.

[7] Terms involving elements from $\mathbb{Z}_p$ are always understood modulo p.

The Diffie-Hellman function (w.r.t. the cyclic group $\mathbb{Z}_p^*$) is given by

$$\mathrm{DH}(g^a, g^b) = g^{ab} \ .$$

It is widely believed that the evaluation of DH is computationally intractable. Even if DH cannot be efficiently evaluated, it might still be possible to compute some bits of information about g^{ab}. In fact, it is known that the Legendre symbol of g^{ab} is easy to compute from p, g, p^a, p^b. This bit of information is therefore "insecure". In order to show that a particular bit of a cryptographic function G is "secure", one typically proves that a general efficient procedure, which reliably predicts this bit, can be converted into an efficient procedure which evaluates G. Everybody who believes that the evaluation problem for G is intractable is therefore forced to believe that no such efficient procedure for the particular bit is conceivable. The analogous remarks are valid for the security of a particular collection of bits.[8]

The Hidden Number Problem was introduced by Boneh and Venkatesan [5] as a tool for showing that the collection of $\sqrt{n}$ unbiased most significant bits of the Diffie-Hellman function is secure. An error in the proof from [5] was spotted and corrected by Vasco and Shparlinski [16]. A technically simpler correction was outlined by Nguyen and Stern [13], but they had to assume that the evaluation problem for the Diffie-Hellman function remains intractable when the input p, g, g^a, g^b is augmented by the prime factorization of $p - 1$.

Here comes the central relation [16] that helps to understand the connection between the Hidden Number Problem and bit-security:

$$g^{(b+r)a} \cdot g^{(b+r)x} = g^{ab+ar+xb+xr} = g^{(a+x)(b+r)} = \mathrm{DH}\left(g^{a+x}, g^{b+r}\right) \qquad (9)$$

Note that $g^{(b+r)a}$ is as hard to compute as g^{ab} whereas g^{a+x}, g^{b+r} are easy to compute from input parameters g^a and g^b (assuming r and x are known). Because of (9), a reliable bit predictor for $B \circ \mathrm{DH}$ provides us with information about the "hidden number" $g^{(b+r)a}$. If we were able to efficiently infer the hidden number from this information, we would end up with a conversion of a bit-predictor for DH into an efficient algorithm that computes the whole function (thereby proving security for bit B of the Diffie-Hellman function).

As emphasized by [9], the Hidden Number Problem plays a central role for the bit-security of a variety of cryptosystems (not just systems employing the Diffie-Hellman function).

7.2 Hidden Number Problem and Learning

Consider a family $B = (B_p)$ of binary predicates $B_p : \mathbb{Z}_p^* \to \{\pm 1\}$. We say that B *distinguishes hidden numbers* if there exists a polynomial $P(n)$ such that

[8] Note that the security of a collection of bits is easier to show than the security of an individual bit because a predictor for a collection of bits is a more powerful tool for solving the evaluation problem for G than the predictor for only one individual bit of the collection.

the correlation between two different concepts u_1, u_2 from $\mathbb{Z}_p^*$ under the uniform distribution is at most $1 - 1/P(n)$:

$$\Pr[B(u_1 \cdot z) = B(u_2 \cdot z)] - \Pr[B(u_1 \cdot z) \neq B(u_2 \cdot z)] \leq 1 - \frac{1}{P(n)}$$

Here, z is drawn uniformly at random from $\mathbb{Z}_p^*$ and n denotes the bit-length of prime p.

The proof of the following theorem strongly builds on problem reductions performed in [16] and [13]. New is only the "compilation" in a learning-theoretic framework.

Theorem 3. *For every binary predicate B that distinguishes hidden numbers, the following holds. If* $\mathrm{HNP}[B]$ *is properly PAC learnable under the uniform distribution, then bit B of the Diffie-Hellman function is secure.*

Proof. Consider a fixed n-bit prime p and a generator g of $\mathbb{Z}_p^*$. We will show how a PAC learner and a reliable predictor for $B \circ \mathrm{DH}$ can be used to compute g^{ab} from p, g, g^a, g^b and the prime factorization of $p-1$:

– Select repeatedly a random element r from $\{0, \ldots, p-2\}$ (independent drawings according to the uniform distribution) until g^{b+r} is a generator.[9]
– Simulate the PAC learner, say A, for target concept $g^{a(b+r)}$ and for "sufficiently small" parameters ε, δ. Whenever A asks for an example, temporarily suspend the simulation and do the following:
  • Select a random element x from $\{0, \ldots, p-2\}$ and compute $g^{a+x} = g^a g^x$ and $g^{b+r} = g^b g^r$.
  • Ask the reliable bit predictor for label $b = B(\mathrm{DH}(g^{a+x}, g^{b+r}))$ and return example (x, b) to the PAC learner.

  Receive finally a hypothesis, say $u \in \mathbb{Z}_p^*$, from the PAC learner.

Note that b correctly labels x w.r.t. target concept $g^{a(b+r)}$ because of the general equation (9). Thus, with probability at least $1-\delta$, hypothesis u will be ε-accurate for concept $g^{a(b+r)}$. Let P be the polynomial such that either $u = g^{a(b+r)}$ or the correlation of u and $g^{a(b+r)}$ is bounded above by $1 - 1/P(n)$. In the latter case, the labels $B(uz)$ and $B(g^{a(b+r)}z)$ assigned to a random instance $z \in \mathbb{Z}_p^*$ are different with probability at least $1/(2P(n))$. Choosing $\varepsilon = 1/(3P(n))$, we can force the PAC learner to be probably *exactly* correct. Thus, with probability at least $1 - \delta$, $u = g^{a(b+r)}$. In this case, we can retrieve g^{ab} by making use of the equation $g^{ab} = g^{a(b+r)}(g^a)^{-r}$. □

The *unbiased most significant bit* is the following predicate on $\mathbb{Z}_p^*$:

$$\mathrm{MSB}(z) = \begin{cases} 1 \text{ if } \frac{p+1}{2} \leq z \leq p-1 \\ 0 \text{ otherwise} \end{cases}$$

[9] Although we do not know b, we can easily test for the generator-property thanks to the factorization of $p-1$.

Lemma 2 ([10]). *The unbiased most significant bit distinguishes hidden numbers in the following strong sense:*

$$\forall u_1, u_2 \in \mathbb{Z}_p^* : u_1 \neq u_2 \Rightarrow \Pr[\mathrm{MSB}(u_1 z) = \mathrm{MSB}(u_2 z)] \leq \frac{2}{3} \ .$$

The corresponding statement for the least significant bit was proven by Kiltz and Simon [10]. Because of $\mathrm{MSB}(x) = \mathrm{LSB}(2x)$, the result carries over to the unbiased most significant bit.

Corollary 9. *If* HNP[MSB] *is properly PAC learnable under the uniform distribution, then the unbiased most significant bit of the Diffie-Hellman function is secure.*

To the best of our knowledge, it is not known whether HNP[MSB] is properly PAC learnable under the uniform distribution. However, from our general lower bound on the number of statistical queries, we can infer the following result:

Lemma 3. *For every n-bit prime,* $L(\mathrm{HNP}[\mathrm{MSB}]_n) = p^{1-o(1)}$.

Proof. Consider the multiplication table $M \in (\mathbb{Z}_p^*)^{(p-1)\times(p-1)}$. i.e, $M[u_1, u_2] = u_1 \cdot u_2 \in \mathbb{Z}_p^*$ for all $u_1, u_2 \in \mathbb{Z}_p^*$. For any binary predicate $B : \mathbb{Z}_p^* \to \{\pm 1\}$, let $B \circ M$ denote the matrix given by $(B \circ M)[u_1, u_2] = B(u_1 \cdot u_2)$. It is shown by Kiltz and Simon (see Corollary 5.1 in [11]) that

$$\|B \circ M\| = p^{1/2+o(1)}$$

provided that predicate B is "balanced". For the purpose of this paper, we do not need to know the definition of "balanced" because it is also shown in the appendix of [11] that every predicate that can be represented by union of $o(\log n)$ intervals actually is balanced. Since MSB is represented by one interval, namely $\{(p+1)/2, \ldots, p-1\}$, it follows that

$$\|\mathrm{MSB} \circ M\| = p^{1/2+o(1)} \ .$$

Note that $\mathrm{MSB} \circ M$ is the incidence matrix for concept class $\mathrm{HNP}[\mathrm{MSB}]_n$. From (6), with the full concept class in the role of $\mathcal{F}'$, we conclude that

$$L(\mathrm{HNP}[\mathrm{MSB}]_n) \geq \frac{(p-1)\cdot(p-1)}{(p^{1/2+o(1)})^2} = p^{1-o(1)} \ .$$

□

Since $p \approx 2^n$, we get the

Corollary 10. $\mathrm{HNP}[MSB]$ *is not weakly polynomially learnable in the SQ model.*

This result indicates that the same concept class might be hard to learn even in the PAC learning model.

References

1. Shai Ben-David, Alon Itai, and Eyal Kushilevitz. Learning by distances. *Information and Computation*, 117(2):240–250, 1995.
2. 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.
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. Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the Association on Computing Machinery*, 50(4):506–519, 2003.
5. Dan Boneh and Ramarathnam Venkatesan. Hardness of computing the most significant bits of secret keys in Diffie-Hellman and related schemes. In *Proceedings of the Conference on Advances in Cryptology — CRYPTO '96*, pages 129–142, 1996.
6. Jürgen Forster. A linear lower bound on the unbounded error communication complexity. *Journal of Computer and System Sciences*, 65(4):612–625, 2002.
7. Jürgen Forster, Matthias Krause, Satyanarayana V. Lokam, Rustam Mubarakzjanov, Niels Schmitt, and Hans Ulrich Simon. Relations between communication complexity, linear arrangements, and computational complexity. In *Proceedings of the 21'st Annual Conference on the Foundations of Software Technology and Theoretical Computer Science*, pages 171–182, 2001.
8. Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the Association on Computing Machinery*, 45(6):983–1006, 1998.
9. Eike Kiltz. A useful primitive to prove security of every bit and about hard core predicates and universal hash functions. In *Proceedings of the 14th International Symposium on Fundamentals of Computation Theory*, pages 388–392, 2001.
10. Eike Kiltz and Hans Ulrich Simon. Unpublished Manuscript about the Hidden Number Problem.
11. Eike Kiltz and Hans Ulrich Simon. Threshold circuit lower bounds on cryptographic functions. *Journal of Computer and System Sciences*, 71(2):185–212, 2005.
12. Matthias Krause and Stephan Waack. Variation ranks of communication matrices and lower bounds for depth two circuits having symmetric gates with unbounded fan-in. *Mathematical System Theory*, 28(6):553–564, 1995.
13. Phong Q. Nguyen and Jacques Stern. The two faces of lattices in cryptology. In *Proceedings of the International Conference on Cryptography and Lattices*, pages 146–180, 2001.
14. Ramamohan Paturi and Janos Simon. Probabilistic communication complexity. *Journal of Computer and System Sciences*, 33(1):106–123, 1986.
15. Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
16. Maria Isabel Gonzáles Vasco and Igor E. Shparlinski. On the security of Diffie–Hellman bits. In *Proceedings of the Workshop on Cryptography and Computational Number Theory*, pages 331–342, 2000.
17. Ke Yang. On learning correlated boolean functions using statistical query. In *Proceedings of the 12th International Conference on Algorithmic Learning Theory*, pages 59–76, 2001.

18. Ke Yang. New lower bounds for statistical query learning. In *Proceedings of the 15th Annual Conferene on Computational Learning Theory*, pages 229–243, 2002.
19. Ke Yang and Avrim Blum. On statistical query sampling and nmr quantum computing. In *Proceedings of the 18th Annual Conference on Computational Complexity*, pages 194–208, 2003.