Spectral Norm in Learning Theory: Some Selected Topics

Hans U. Simon (RUB)

Email: simon@lmi.rub.de

Homepage: http://www.ruhr-uni-bochum.de/lmi

The Star of the Talk

$$\lambda_1$$

Spectral Norm of a Matrix

For matrix $A \in \mathbb{R}^{m \times n}$, consider the singular values

$$\sigma_1(A) \geq \sigma_2(A) \geq \cdots \ .$$

If A is symmetric and positive semidefinite, they coincide with the eigenvalues

$$\lambda_1(A) \geq \lambda_2(A) \geq \cdots \ .$$

Definition of Spectral Norm:

$$\|A\| := \sigma_1(A) = \sqrt{\lambda_1(A^\top \cdot A)} \ .$$

Structure of the Talk

I Statistical Query Learning and Correlation Query Learning

II Ke Yang's Lower Bound on the Number of Statistical Queries

III Other Bounds in Terms of the Spectral Norm

IV Hidden Number Problem from the Perspective of Learning Theory

Part I

$$\mathbf{SQ} = \mathbf{CQ}$$

Concept Learning

- X, a domain
- $x \in X$, an instance
- D, a probability distribution on X
- $(x, b) \in X \times \{-1, +1\}$, a labeled instance
- $f : X \to \{-1, +1\}$, a concept
- $\mathcal{F}$, a class of concepts

Goal: Find a good approximation (hypothesis) $h_* : X \to \{-1, +1\}$ for an unknown target concept $f_* \in \mathcal{F}$.

Statistical Query Learning (fixed distribution D)

$h' : X \times \{-1,+1\} \rightarrow \{-1,+1\}$, a query function such that

$$\mathbb{E}_D[h'(x, f_*(x))] = \Pr[h'(x, f_*(x)) = +1] - \Pr[h'(x, f_*(x)) = -1]$$

$\tau > 0$, a tolerance parameter

Information Gathering Mechanism:

- LEARNER $\xrightarrow{h',\tau}$ ORACLE
- ORACLE $\xrightarrow{d}$ LEARNER s.t.

$$\mathbb{E}_D[h'(x, f_*(x))] - \tau \leq d \leq \mathbb{E}_D[h'(x, f_*(x))] + \tau$$

After gathering enough information, the learner outputs a final hypothesis h_*.

Statistical Query Learning (continued)

Efficiency of the learner measured by:

- q, the number of specified query functions (including the final hypothesis)
- τ, smallest tolerance parameter ever used during learning

Success of the learner measured by:

- ε, probability of misclassification of the final hypothesis

Kearns' Result: An SQ learner can be simulated by a noise-tolerant PAC learner that has access to $\text{poly}(q, 1/\tau, 1/\varepsilon)$ random examples.

Evaluation and Correlation Matrix

Equip $\mathbb{R}^X$ with the following inner product:

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

Associate with a concept class $\mathcal{F}$, the evaluation matrix

$$E_{\mathcal{F}}[x, f] := f(x)$$

and the (positive semidefinite) correlation matrix

$$C_{\mathcal{F}}[f_1, f_2] := \langle f_1, f_2 \rangle_D \ .$$

Column f of matrix $E_{\mathcal{F}}$ contains the "function table" for f.

Correlation Query Learning (fixed distribution D)

$h : X \rightarrow \{-1, 0, +1\}$, a query function

$\tau > 0$, a tolerance parameter

$\gamma > 0$, desired correlation between h_* and f_*

$$\text{correlation } \gamma \Leftrightarrow \text{misclassification rate } \frac{1}{2} - \frac{\gamma}{2}$$

Information Gathering Mechanism:

- LEARNER $\xrightarrow{h,\tau}$ ORACLE
- ORACLE $\xrightarrow{c}$ LEARNER s.t.

$$\langle h, f_* \rangle_D - \tau \leq c \leq \langle h, f_* \rangle_D + \tau$$

After gathering enough information, the learner outputs a final hypothesis h_*.

Equivalence of the SQ and the CQ Model

Proof Idea: Mutual conversion

$$h' \leftrightarrow h$$

of query functions

$$h' : X \times \{-1, +1\} \to \{-1, +1\} \text{ and } h : X \to \{-1, 0, +1\}$$

such that queries

$$(h', \tau) \text{ and } (h, \tau)$$

yield precisely the same amount of information.

First Conversion

Given $h' : X \times \{\pm 1\} \rightarrow \{\pm 1\}$, consider

$$X_0(h') \quad := \quad \{x \in X :\ h'(x,+1) = h'(x,-1)\}$$
$$X_1(h') \quad := \quad \{x \in X :\ h'(x,+1) \neq h'(x,-1)\}$$

and the following function $h : X \rightarrow \{-1, 0, +1\}$:

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

Note that:

$$\forall x \in X_1(h'), \forall b = \pm 1 :\ h'(x,b) = b \cdot h'(x,1) = b \cdot h(x)$$

First Conversion (continued)

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

Reverse Direction:

Exploit the same relation between $\mathbb{E}_D[h'(x, f_*(x))]$ and $\langle h, f_* \rangle_D$ again !

Part II

Ke Yang's Lower Bound

Yang's Bound

Theorem (Ke Yang, COLT 2002, JCSS2005): The smallest possible number $q(\mathcal{F})$ of statistical queries is lower-bounded as follows:

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

Corollary: Because of $\|C_{\mathcal{F}}\| = \lambda_1(C_{\mathcal{F}})$:

$$q(\mathcal{F}) \geq \frac{|\mathcal{F}|}{\|C_{\mathcal{F}}\|} \cdot \min\{\gamma^2, \tau^2\}$$

Because $\mathcal{F}$ is not easier to learn than $\mathcal{F}' \subseteq \mathcal{F}$:

$$q(\mathcal{F}) \geq \left(\sup_{\mathcal{F}' \subseteq \mathcal{F}} \frac{|\mathcal{F}'|}{\|C_{\mathcal{F}'}\|} \right) \cdot \min\{\gamma^2, \tau^2\}$$

Sketch of Proof (Adversary Argument)

- CQ-oracle returns answer "0" as long as possible.
- Say $(h_1, \tau_1), \ldots, (h_{q'-1}, \tau_{q'-1})$ are answered "0".
- Let $h_{q'}$ be the next query function (final hypothesis if $q' = q(\mathcal{F})$).
- Let $\mathcal{Q} := \langle h_1, \ldots, h_{q'} \rangle$.
- Let $f^{\mathcal{Q}}$ denote the projection of f onto subspace $\mathcal{Q}$.

With these notations:

$$\sum_{i=1}^{q(\mathcal{F})} \lambda_i(C_{\mathcal{F}}) \geq \sum_{i=1}^{q'} \lambda_i(C_{\mathcal{F}}) \geq \sum_{f \in \mathcal{F}} \|f^{\mathcal{Q}}\|^2 \geq \|\mathcal{F}\| \cdot \min\{\gamma^2, \tau^2\}$$

Characterization of Weak SQ-Learnability

The following statements are equivalent:

1. $(\mathcal{F}_n)_{n\geq 1}$ admits a weak polynomial learner in the SQ model.
2. The statistical query dimension of $(\mathcal{F}_n)_{n\geq 1}$ is polynomially bounded.
3. $\sup_{\mathcal{F}'_n \subseteq \mathcal{F}_n} \frac{|\mathcal{F}'_n|}{\|C_{\mathcal{F}'_n}\|}$ is polynomially bounded.

1.⇔2.: Blum, Furst, Jackson. Kearns, Mansour, Rudich (STOC 1994)

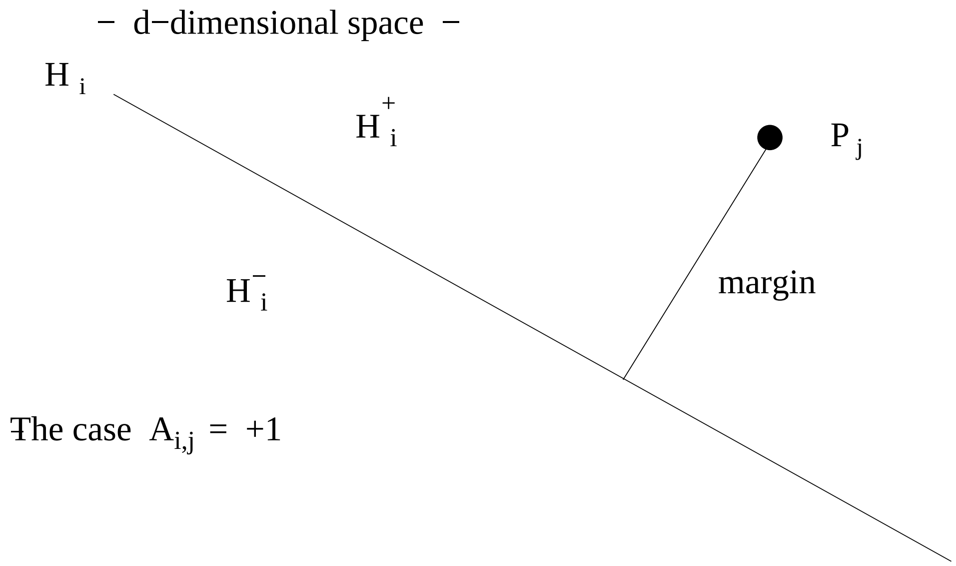
2.⇔3.: Statistical query dimension polynomially related to $\sup_{\mathcal{F}' \subseteq \mathcal{F}} \frac{|\mathcal{F}'_n|}{\|C_{\mathcal{F}'_n}\|}$

Part III

The many faces of λ_1

Spectral Norm and Half-space Embeddings

A half-space embedding for a matrix $A \in \{-1,+1\}^{m\times n}$:



$$
\begin{aligned}
d(A) &:= \text{smallest possible dimension} \\
&:= \text{smallest rank of a sign-equivalent matrix} \\
\mu(A) &:= \text{largest possible "guaranteed" margin}
\end{aligned}
$$

Results by Forster (CCC 2001, JCSS 2002)

For every $A \in \{-1,+1\}^{m \times n}$:

$$d(A) \geq \frac{\sqrt{mn}}{\|A\|}$$

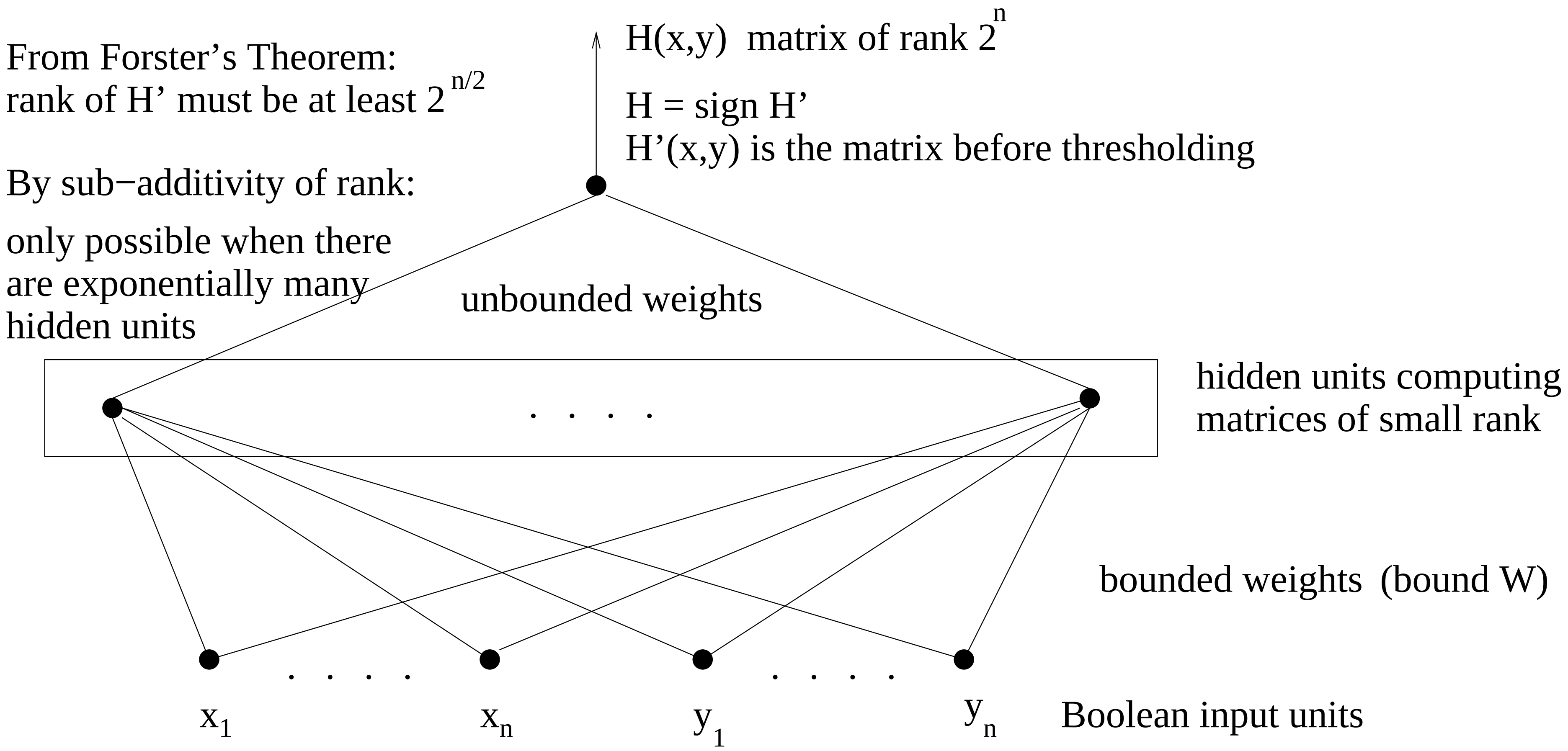
$$\mu(A) \leq \frac{\|A\|}{\sqrt{mn}}$$

In particular for the Hadamard matrix $H_n \in \{-1,+1\}^{2^n \times 2^n}$:

$$d(H_n) \geq 2^{n/2} \text{ } (\Rightarrow \text{ probabilistic communication complexity } n/2)$$

$$\mu(H_n) \leq 2^{-n/2}$$

Application: Threshold Circuits of Exponential Size



Result by Forster, Krause, Lokam, Mubarakzjanov, Schmitt, H.U.S.: there must be at least $\frac{2^{n/2}}{nW+1}$ units in the hidden layer (FSTTCS 2001).

Sufficient Conditions for Weak SQ Learnability

- $d(E_{\mathcal{F}_n})$ is poly(n)-bounded.
- $\mu(E_{\mathcal{F}_n})^{-1}$ is poly(n)-bounded.
- The probabilistic communication complexity for $E_{\mathcal{F}_n}$ in the unbounded error model is $O(\log(n))$-bounded.
- Functions from $\mathcal{F}_n$ can be evaluated by a depth-2 threshold circuit (unbounded weights at top unit, poly(n)-bounded weights at hidden units) of poly(n) size.

Part IV

Can we learn "hidden numbers" ??

Prerequisites for the Hidden Number Problem

- p, an n-bit prime
- $\mathbb{Z}_p = \{0, \ldots, p-1\}$, smallest residuals modulo p
- $(\mathbb{Z}_p, +, \cdot)$, the corresponding prime field
- $\mathbb{Z}_p^* = \{1, \ldots, p-1\}$
- $(\mathbb{Z}_p^*, \cdot)$, the cyclic group of prime residuals modulo p
- g, a generator of $\mathbb{Z}_p^*$
- $B : \mathbb{Z}_p^* \rightarrow \{-1, +1\}$, a binary predicate for prime residuals

`--- All arithmetic is understood modulo` p `---`

The Hidden Number Problem (HNP[B])

Goal: Infer a hidden number $u \in \mathbb{Z}_p^*$ from "observations".

Information-gathering Mechanism:

$$\xrightarrow{z} \boxed{u} \xrightarrow{B(u \cdot z)}$$

Elements $z \in \mathbb{Z}_p^*$ independently drawn at random (uniform distribution)

Motivation: Bit Security

Diffie-Hellman Function: $\mathrm{DH}(g^a, g^b) = g^{ab}$.

Central Relation: (Boneh, Venkatesan 1996; Vasco, Shparlinski 2000)

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

Known: n, p, g, g^a, g^b, r, x

Unknown: $g^{ab}, g^{(b+r)a}$

Hidden Number: $g^{(b+r)a}$

Random Instance: $g^{(b+r)x}$

Note: g^{b+r} should be a generator !!

Cast HNP[B] as Learning Problem

Put problem in the concept learning framework:

- View u as target concept.
- View z as random instance.
- View $B(u \cdot z)$ as the correct classification label.

Note: The correlation between u_1, u_2 can be expressed as

$$\Pr[B(u_1 \cdot z) = B(u_2 \cdot z)] - \Pr[B(u_1 \cdot z) \neq B(u_2 \cdot z)] \ .$$

Assumption: Predicate B must "distinguish" different hidden numbers:

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

Example for a "distinguishing" predicate

Consider the unbiased most significant bit:

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

Kiltz, H.U.S. (unpublished manuscript): The unbiased most significant bit distinguishes hidden numbers in the following strong sense:

$$\Pr[\mathrm{MSB}(u_1 \cdot z) = \mathrm{MSB}(u_2 \cdot z)] - \Pr[\mathrm{MSB}(u_1 \cdot z) \neq \mathrm{MSB}(u_2 \cdot z)] \le \frac{2}{3}$$

Proper PAC Learners imply Bit Security

Theorem:

If HNP[B] is properly and polynomially PAC-learnable under the uniform distribution, then bit B of the Diffie-Hellman function is secure.

The proof translates similar proofs by (Nguyen and Stern, 2001 ; Vasco and Shparlinski, 2000) into a learning-theoretic framework.

Conjecture: Such PAC-learners do not exist.

Hidden Number Problem and SQ Learning

- Let M denote the multiplication table for $\mathbb{Z}_p^*$.
- Then $\mathrm{MSB} \circ M$ is the evaluation matrix for HNP[MSB].

Lemma (Kiltz, H.U.S., JCSS 2005): $\|\mathrm{MSB} \circ M\| = p^{1/2+o(1)}$.

Because of the general relations (assuming uniform distribution D)

$$C_{\mathcal{F}} = \frac{1}{|X|} \cdot E_{\mathcal{F}}^\top \cdot E_{\mathcal{F}} \text{ and } \|C_{\mathcal{F}}\| = \frac{1}{|X|}\|E_{\mathcal{F}}\|^2 \ ,$$

and because of Ke Yang's lower bound:

Corollary: 1. $q(\mathrm{HNP[MSB]}) \geq \frac{(p-1)^2}{p^{1+o(1)}} = p^{1-o(1)}$.

2. HNP[MSB] is not weakly learnable in the SQ model.

Conclusions

We have put the framework of SQ Learning into a broader context, relating it to various (seemingly different) problems in learning theory, complexity theory, and cryptography. The key notion has been the spectral norm.

--- The End ---