Average-case statistical query algorithms

Enric Boix-Adserà (MIT)
Report, MDS Lab Internship, EPFL

1 Introduction

1.1 PAC-learning and the SQ model

The statistical query (SQ) model was first introduced by Kearns in [6] as a restriction of the probably approximately correct (PAC) model of Valiant in [8]. In both models, an algorithm seeks to learn an unknown function $F : \mathcal{X} \to \mathcal{Y}$ that is a member of a concept class $\mathcal{F} \ni F$.

PAC-learning model In the PAC-learning model, the learning algorithm A is given n samples $(X_1, F(X_1)), \ldots, (X_t, F(X_t))$ where the X_i's are drawn i.i.d. from the alphabet $\mathcal{X}$ according to some distribution P_X. The algorithm is allowed to manipulate these samples however it chooses, and at the end it must output a hypothesis $\hat{F} : \mathcal{X} \to \mathcal{Y}$. The algorithm A is said to (ϵ, δ)-PAC learn $(\mathcal{F}, P_X)$ if for any $F \in \mathcal{F}$,

$$\mathbb{P}_{A, X_1, \ldots, X_n}[\mathbb{P}_{X \sim P_X}[\hat{F}(X) \neq F(X)] \leq \epsilon] \geq 1 - \delta.$$

In other words, in order to PAC-learn $(\mathcal{F}, P_X)$ the algorithm must output an approximately-correct hypothesis $\hat{F}$ with high probability over the samples $X_1, \ldots, X_n$ and its internal randomness.

It is of particular interest to understand which PAC-learning problems $(\mathcal{F}, P_X)$ are learnable by an efficient algorithm. Typically, a PAC-learning algorithm is said to be efficient if it uses only $n = \text{poly}(1/\epsilon, 1/\delta)$ samples and also runs in $\text{poly}(1/\epsilon, 1/\delta)$ time.

SQ model Kearns [6] introduced the statistical query (SQ) model in order to capture the subset of PAC-learning algorithms that only access the distribution by taking overall statistics of the distribution. Under the SQ model a learning algorithm does not have direct access to the samples $(X_i, F(X_i))$ but rather interacts with F through queries $\phi : \mathcal{X} \times \mathcal{Y} \to [-1, 1]$ to an adversarial oracle $\mathcal{O}^{adv,\tau}$. Given a query ϕ, the oracle $\mathcal{O}^{adv,\tau}$ adversarially returns a value

$$\hat{\mu} \in \mathbb{E}_{X \sim P_X}[\phi(X, F(X))] + [-\tau, \tau],$$

where $\tau > 0$ is an additive error tolerance parameter for the the oracle $\mathcal{O}^{adv,\tau}$. The SQ algorithm can adaptively choose its queries as a function of its previous queries and the oracle's responses to those queries. At a certain point, once $F \in \mathcal{F}$ has been uniquely identified by the oracle's responses, the SQ algorithm returns F.

Note that every SQ algorithm A can be simulated by a corresponding PAC algorithm B. The algorithm B works identically to A, except that for each query $\phi : \mathcal{X} \times \mathcal{Y} \to [-1, 1]$ made by A, instead B computes $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i, F(X_i))$, which by Hoeffding bounds will be within τ of the true mean $\mu = \mathbb{E}_{X \sim P_X}[\phi(X, F(X))]$ with high probability as long as B has enough samples relative to the number q of statistical queries made by A (i.e., B $(0, \delta)$-PAC-learns $\mathcal{F}$ if $n = \Omega(\log(q/\delta)/\tau^2)$).

As with PAC algorithms, it is of interest to understand whether a learning problem $(\mathcal{F}, P_X)$ has an efficient SQ algorithm. In this case, the complexity of an SQ algorithm is generally measured in terms of the number q of queries that it makes, and the SQ complexity of $(\mathcal{F}, P_X)$ is the complexity of the most efficient SQ algorithm learning it. Remarkably, Blum et al. [2] showed that the SQ complexity of a statistical learning problem $(\mathcal{F}, P_X)$ can be characterized via Fourier analysis. Therefore it is often a simple exercise to prove lower bounds on the SQ complexity of a class $(\mathcal{F}, P_X)$. These SQ complexity lower bounds can then be taken as evidence that a learning problem is hard even for general PAC learning algorithms (e.g.,[4], [5]) – although there are known concept classes (such as learning parities) with high SQ complexity that are also efficiently PAC-learnable [6].

1.2 Average-case SQ

Notice that in many respects the SQ model is more a worst-case model than an average-case model. For instance, (1) the complexity of a SQ algorithm is measured as the algorithm's number of queries on a worst-case, rather than average-case, input $F \in \mathcal{F}$, (2) $\mathcal{O}^{adv,\tau}$ returns an answer with adversarial, rather than statistical, noise, and (3) the SQ algorithm is expected to learn $F \in \mathcal{F}$ exactly rather than output a hypothesis that is simply correct on average for inputs $X \sim P_X$. We address these points in some detail:

1. **Worst-case input F vs. average-case F** Instead of a worst-case function, one may be asked to learn a function F drawn from a distribution $F \sim P_F$ on $\mathcal{F}$. Given such an average-case problem (P_F, P_X), we define the average-case statistical query (ASQ) complexity to be the expected query complexity of a Las Vegas (zero probability of error) statistical query algorithm for learning F exactly, when given access to F via the oracle $\mathcal{O}^{adv,\tau}$.

   In Section 2 we provide a simple characterization of the ASQ complexity, and we also show how to lower-bound the ASQ complexity in terms of the "cross-predictability" of (P_F, P_X).

2. **Adversarial noise vs. statistical noise** Instead of an oracle $\mathcal{O}^{adv,\tau}$ that answers queries with adversarial noise, suppose that one has an oracle that answers queries

with statistical noise. How does the SQ complexity change?

This question was studied by Yang [9] and Feldman et al. [5]. In their 1-STAT query model, given a query $\phi : \mathcal{X} \times \mathcal{Y} \to \{-1, 1\}$ the oracle $\mathcal{O}^{1STAT}$ draws a fresh sample $X \sim P_X$ and returns $\phi(X, F(X))$.[1] Notice that by taking $O((\log(q/\delta))/\tau^2)$ 1-STAT samples for each query to $\mathcal{O}^{adv,\tau}$, and computing the average of these samples, with probability of success $\geq 1 - \delta$ one can simulate q queries of $\mathcal{O}^{adv,\tau}$ using $O(q(\log(q/\delta))/\tau^2)$ queries $\mathcal{O}^{1STAT}$. In fact, [9] and [5] prove that this is roughly the best that one can do with a $\mathcal{O}^{1STAT}$ oracle, and so adversarial and statistical noise are roughly comparable.

In this work, whenever we consider statistical noise we will work instead with a $\mathcal{O}^{stat,\sigma}$ oracle that adds i.i.d. $\mathcal{N}(0, \sigma^2)$ noise to the true expectation of each query. A coupling argument, also noted by [7], allows us to show that $\mathcal{O}^{adv,\tau}$ and $\mathcal{O}^{stat,\sigma}$ are closely related. We will consider statistical noise in the context of weak learning (see below).

3. **Exact learning vs. weak learning** Instead of requiring the SQ algorithm to learn F exactly, we can relax the requirement to learning F approximately: outputting $\hat{F}$ such that $\hat{F}$ is positively correlated with F better than a random guess.

   Given an average-case SQ problem (P_F, P_X) as defined above, if (P_F, P_X) is balanced and binary-valued[2] then weakly-learning (P_F, P_X) amounts to returning $\hat{F}$ such that

   $$\mathbb{P}_{F \sim P_F, X \sim P_X}[\hat{F} = F(X)] \geq 1/2 + \epsilon.$$

   We show lower bounds on the number of statistical-noise queries needed to weakly-learn a balanced-binary problem (P_F, P_X) in Section 3.

1.3 Overview

In Section 2 we characterize the average-case statistical query (ASQ) complexity of a problem (P_F, P_X) when the noise is adversarial and we wish to exactly learn F. In Section 2.3 we show how to lower-bound the ASQ complexity by upper-bounding the "cross-predictability" of (P_F, P_X). In Section 3 we lower-bound the number of queries to $\mathcal{O}^{stat,\sigma}$ that are necessary to weakly-learn (P_F, P_X) when the functions $F \in \mathcal{F}$ are binary-valued and balanced with respect to P_X. In Section 4 we apply the lower bounds from Section 3 in order to prove limitations on neural networks for learning parities using gradient descent. This latter result is comparable to and strictly weaker than the result of [1], but the proof technique is different.

[1] Yang [9] calls this model the "Honest Statistical Query" model.

[2] That is, each function $F : \mathcal{X} \to cY$ has output alphabet $\mathcal{Y} = \{-1, 1\}$ and also $\mathbb{P}_{F \sim P_F, X \sim P_X}[F(X) = 1] = \frac{1}{2} + o(1)$

2 Average-case statistical query complexity

2.1 Definition

Definition 2.1 (Average-case exact-learning SQ)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite alphabets, and let $\mathcal{F}$ be a class of functions $F : \mathcal{X} \to \mathcal{Y}$. Given a distribution P_F over $\mathcal{F}$, a distribution P_X over $\mathcal{X}$, and $\tau > 0$ a noise threshold, we may define the average-case exact-learning statistical query problem (P_F, P_X, τ) with adversarial noise as follows.*

An algorithm A is said to solve (P_F, P_X, τ) if it learns a hidden function $F \sim P_F$, given access only to an oracle $\mathcal{O}^{adv,\tau}$ that on query $\phi : \mathcal{X} \times \mathcal{Y} \to [-1, 1]$ returns an adversarial value in

$$\mathbb{E}_{X \sim P_X}[\phi(X, F(X))] + [-\tau, \tau].$$

The complexity $C(A)$ of the algorithm A is defined to be the expected number of queries that it performs before outputting F. That is, letting $Q(A, F)$ be a random variable denoting the number of queries of A,

$$C(A) = \mathbb{E}_{F \sim P_F, A}[Q(A, F)],$$

where the expectation is taken over the hidden function $F \sim P_F$ and the internal randomness of A.

However, instead of studying the exact-learning problem directly, we will study the many-to-one distribution testing problem, which is closely related.

Definition 2.2 (Average-case many-to-one SQ)**.** *Let $\mathcal{D}$ be a set of distributions over an alphabet $\mathcal{Z}$. Given a tuple (P_D, D_0, τ) consisting of a distribution P_D supported in $\mathcal{D}$, a "reference" distribution $D_0 \notin \mathcal{D}$, and a noise tolerance $\tau > 0$, the average-case many-to-one SQ testing problem is to distinguish between D_0 and a random distribution $D \sim P_D$ using only queries $\phi : \mathcal{Z} \to [-1, 1]$ to an oracle $\mathcal{O}^{adv,\tau}$ that returns an adversarial response in $D[\phi] + [-\tau, \tau]$, where $D[\phi] := \mathbb{E}_{z \sim D}[\phi(z)]$.*

The complexity $C(A)$ of an algorithm A solving (P_D, D_0, τ) is the expected number of queries that it makes before returning. The complexity $C(P_D, D_0, \tau)$ of the the problem is the minimum complexity $C(A)$ over algorithms A solving (P_D, D_0, τ).

Setting $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, we may construct a distribution-testing problem given an ASQ exact-learning problem (P_F, P_X, τ). First, for each function $F \in \mathcal{F}$, we define the distribution D_F over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with probability mass function

$$D_F(x, y) = P_X(x) \times 1(F(x) = y).$$

Then we create a distribution P_D over the set of distributions $\mathcal{D}$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ which has the probability mass

$$P_D(D_F) = P_F(F).$$

Given any reference distribution D_0, a lower-bound on the statistical query complexity of the many-to-one distribution testing problem (P_D, D_0, τ) will directly translate into a lower-bound on the statistical query complexity of the exact-learning problem (P_F, P_X, τ). Conversely, Feldman et al. [3] show that for any exact-learning statistical query problem[3] in the worst-case setting, the statistical query complexity is roughly approximated by the statistical query complexity of a corresponding many-to-one decision problem. Here we do not show this converse in the average-case setting.

2.2 Characterization via average-case statistical dimension

Given a many-to-one distribution testing problem (P_D, D_0, τ), we define the average-case statistical dimension of (P_D, D_0, τ) in a manner analogous to the definition of the statistical dimension of $(\mathcal{D}, D_0, \tau)$ found in [3]. In order to define this, let $\mathcal{D}'[\phi, \tau, D_0] \subset \mathcal{D}' \subset \mathcal{D}$ denote the subset of distributions distinguished from D_0 by query ϕ :

$$\mathcal{D}'[\phi, \tau, D_0] = \{D' \in \mathcal{D} : |D[\phi] - D_0[\phi]| > \tau\}.$$

Definition 2.3. *The average-case statistical dimension (ASD) is given by*

$$ASD(P_D, D_0, \tau) := \max_{\mathcal{D}' \subset \mathcal{D}} \cdot \left(\min_{\phi : \mathcal{Z} \to [-1,1]} \frac{P_D[\mathcal{D}']^2}{P_D\left[\mathcal{D}'[\phi, \tau, D_0]\right]} \right).$$

Theorem 2.4. $ASD(P_D, D_0, \tau)/2 \leq C(P_D, D_0, \tau) \leq ASD(P_D, D_0, 2\tau) \cdot O(\log(|\mathcal{D}|))$.

Proof. $ASD(P_D, D_0, \tau)/4 \leq C(P_D, D_0, \tau)$: let A be an algorithm that sequentially queries $\phi_1, \phi_2, \ldots$, and returns as soon as the oracle $\mathcal{O}^{adv,\tau}$ returns an answer that disagrees with D_0. Suppose that $\mathcal{O}^{adv,\tau}$ returns $D_0[\phi]$ to query ϕ, as long as $D \notin \mathcal{D}[\phi, \tau, D_0]$. Let $\mathcal{D}' \subset \mathcal{D}$ be the subset of distributions that attains the maximum for $ASD(P_D, D_0, \tau)$. Let $\mathcal{D}'_i \subset \mathcal{D}'$ be the set of distributions in $\mathcal{D}'$ that have not been disqualified by the answers to queries $\phi_1, \ldots, \phi_{i-1}$. Then since $P_D[\mathcal{D}'_i] \geq P_D[\mathcal{D}'] \cdot (1 - P_D[\mathcal{D}'](i-1)/ASD(P_D, D_0, \tau))$,

$$\begin{aligned} C(P_D, D_0, \tau) &\geq \sum_{i=1}^{\infty} P_D[\mathcal{D}'_i] \geq \sum_{i=1}^{ASD(P_D, D_0, \tau)/P_D[\mathcal{D}']} P_D[\mathcal{D}'] \cdot (1 - \frac{P_D[\mathcal{D}'](i-1)}{ASD(P_D, D_0, \tau)}) \\ &\geq (P_D[\mathcal{D}'])/2 \cdot (ASD(P_D, D_0, \tau)/2P_D[D']) \geq ASD(P_D, D_0, \tau)/4. \end{aligned}$$

$C(P_D, D_0, \tau/2) \leq ASD(P_D, D_0, \tau)$: Consider the following greedy statistical query algorithm. On iteration i, the algorithm queries ϕ_i. Let $\mathcal{D}_i \subseteq \mathcal{D}$ be the set of hypotheses

[3] and, more generally, for any search problem

that are still possible after having queried $\phi_1, \ldots, \phi_{i-1}$ and received answers from the oracle. Choose ϕ_i so as to minimize $P_D[\mathcal{D}_i]/P_D[\mathcal{D}_i[\phi_i, \tau, D_0]]$, so

$$P_D[\mathcal{D}_i[\phi_i, \tau, D_0]] \geq \frac{(P_D[\mathcal{D}_i])^2}{ASD(P_D, D_0, \tau)}.$$

Note that $\mathcal{D}_{i+1} \subseteq \mathcal{D}_i \setminus \mathcal{D}_i[\phi_i, \tau, D_0]$.Therefore,

$$P_D[\mathcal{D}_{i+1}] \leq \left(1 - \frac{P_D[\mathcal{D}_i]}{ASD(P_D, D_0, \tau)}\right) P_D[\mathcal{D}_i]. \tag{1}$$

In particular,

$$C(P_D, D_0, \tau/2) \leq \sum_{i=1}^{\infty} P_D[\mathcal{D}_i].$$

Letting $i^* = \max\{i : P_D[\mathcal{D}_i] > 1/|\mathcal{D}|\}$, since the procedure terminates in at most $|\mathcal{D}|$ steps we have

$$C(P_D, D_0, \tau/2) \leq \sum_{i=1}^{i^*} P_D[\mathcal{D}_i] + \sum_{i=i^*+1}^{|\mathcal{D}|} P_D[\mathcal{D}_i] \leq \sum_{i=1}^{i^*} P_D[\mathcal{D}_i] + \sum_{i=i^*+1}^{|\mathcal{D}|} \frac{1}{|\mathcal{D}|} \leq 1 + \sum_{i=1}^{i^*} P_D[\mathcal{D}_i].$$

Now for each j let $a_j = \min\{i \geq 1 : P_D[\mathcal{D}_i] \leq 2^{-j}\}$. Grouping terms in the above equation,

$$C(P_D, D_0, \tau/2) \leq \sum_{j=0}^{\lceil \log_2(|\mathcal{D}|)\rceil - 1} \sum_{i \in [a_j, a_{j+1})} P_D[\mathcal{D}_i] \leq \sum_{j=0}^{\lceil \log_2(|\mathcal{D}|)\rceil - 1} \sum_{i \in [a_j, a_{j+1})} 2^{-j}.$$

We note that for any $j, k \geq 0$, by (1),

$$P_D[D_{a_j+k}] \leq (1 - \frac{2^{-j}}{ASD(P_D, D_0, \tau)})^k.$$

As a result, $|a_{j+1} - a_j| \leq 2^j \cdot ASD(P_D, D_0, \tau) \cdot \ln(2)$. So

$$C(P_D, D_0, \tau/2) \leq \sum_{j=0}^{\log_2(|\mathcal{D}|)} 2^{-j} \cdot |a_{j+1} - a_j|$$

$$\leq \sum_{j=0}^{\log_2(|\mathcal{D}|)} ASD(P_D, D_0, \tau) = ASD(P_D, D_0, \tau) \cdot O(\log(|\mathcal{D}|)).$$

□

2.3 Low cross-predictability implies high ASD

We will now introduce the cross-predictability, which can be used to lower-bound the average-case statistical query dimension. First we introduce some notation: for any distribution D on an alphabet $\mathcal{Z}$ and functions $f, g : \mathcal{Z} \to [-1, 1]$, define the inner product

$$\langle f, g \rangle_D = \mathbb{E}_{z \sim D} f(z) g(z).$$

Let D_0 be a reference distribution, and define

$$\hat{D} = \frac{D}{D_0} - 1.$$

The cross-predictability of P_D with respect to D_0 is defined to be

Definition 2.5. $\Pi_2(P_D, D_0) := \Pi(P_D, D_0) := \mathbb{E}_{D_1, D_2 \overset{i.i.d}{\sim} P_D}[\langle \hat{D}_1, \hat{D}_2 \rangle^2_{D_0}]$.

Notice that in the case in which P_F is over a set of functions $\mathcal{F} : \mathcal{X} \to \{-1, 1\}$ and the reference distribution $D_0 = \mathbb{E}_{F_0} D_{F_0}$ for F_0 a random function (i.e., $D_0(x, y) = P_X(x)/2$ for all $x \in \mathcal{X}, y \in \{-1, 1\}$), then in fact

$$\Pi(P_D, D_0) = \mathbb{E}_{F_1, F_2 \overset{i.i.d.}{\sim} P_F} \left(\mathbb{E}_{x \sim P_X} F_1(x) F_2(x) \right)^2.$$

We also define an L_1-norm version of the cross-predictability:

Definition 2.6. $\Pi_1(P_D, D_0) := \mathbb{E}_{D_1, D_2 \overset{i.i.d}{\sim} P_D}[|\langle \hat{D}_1, \hat{D}_2 \rangle_{D_0}|]$.

Theorem 2.7 (Low cross-predictability implies high ASD)**.**

$$ASD(P_D, D_0, \tau) \geq \min_{\phi : \mathcal{Z} \to [-1,1]} \frac{1}{P_D[\mathcal{D}[\phi, \tau, D_0]]} \geq \frac{\tau^2}{\Pi(P_D, D_0)^{1/2}}$$

Proof. We lower-bound $\min_{\phi : \mathcal{Z} \to [-1,1]} \frac{P_D[\mathcal{D}]}{P_D[\mathcal{D}[\phi, \tau, D_0]]}$, thereby lower-bounding the ASD. Let $\phi : \mathcal{Z} \to [-1, 1]$ be a statistical query. Define

$$\Psi = \langle \phi, \sum_{D \in \mathcal{D}[\phi, \tau, D_0]} P_D[D] \operatorname{sgn}(\langle \phi, \hat{D} \rangle) \cdot \hat{D} \rangle_{D_0}.$$

Then

$$\begin{aligned}
\Psi^2 &\le \langle \phi, \phi \rangle_{D_0} \langle \sum_{D \in \mathcal{D}[\phi,\tau,D_0]} P_D[D] \operatorname{sgn}(\langle \phi, \hat{D} \rangle) \cdot \hat{D}, \sum_{D \in \mathcal{D}[\phi,\tau,D_0]} P_D[D] \operatorname{sgn}(\langle \phi, \hat{D} \rangle) \cdot \hat{D} \rangle_{D_0} \\
&\qquad\qquad \text{(by Cauchy-Schwarz)} \\
&\le \langle \sum_{D \in \mathcal{D}[\phi,\tau,D_0]} P_D[D] \operatorname{sgn}(\langle \phi, \hat{D} \rangle) \cdot \hat{D}, \sum_{D \in \mathcal{D}[\phi,\tau,D_0]} P_D[D] \operatorname{sgn}(\langle \phi, \hat{D} \rangle) \cdot \hat{D} \rangle_{D_0} \\
&= \sum_{D_1, D_2 \in \mathcal{D}[\phi,\tau,D_0]} P_D[D_1] P_D[D_2] \operatorname{sgn}(\langle \phi, \hat{D_1} \rangle) \operatorname{sgn}(\langle \phi, \hat{D_2} \rangle) \cdot \langle \hat{D_1}, \hat{D_2} \rangle_{D_0} \\
&\le \sum_{D_1, D_2 \in \mathcal{D}[\phi,\tau,D_0]} P_D[D_1] P_D[D_2] \cdot |\langle \hat{D_1}, \hat{D_2} \rangle_{D_0}| \\
&= P_D[\mathcal{D}[\phi,\tau,D_0]]^2 \mathbb{E}_{D_1, D_2 \sim P_D[\cdot | \mathcal{D}[\phi,\tau,D_0]]} |\langle \hat{D_1}, \hat{D_2} \rangle_{D_0}| \\
&\le P_D[\mathcal{D}[\phi,\tau,D_0]]^2 \sqrt{\mathbb{E}_{D_1, D_2 \sim P_D[\cdot | \mathcal{D}[\phi,\tau,D_0]]} \left(\langle \hat{D_1}, \hat{D_2} \rangle_{D_0}^2 \right)} \\
&\le P_D[\mathcal{D}[\phi,\tau,D_0]] \sqrt{\mathbb{E}_{D_1, D_2 \sim P_D} \left(\langle \hat{D_1}, \hat{D_2} \rangle_{D_0}^2 \right)} \\
&= P_D[\mathcal{D}[\phi,\tau,D_0]] \cdot \Pi(P_D, D_0)^{1/2}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\Psi^2 &= \left(\sum_{D \in \mathcal{D}[\phi,\tau,D_0]} P_D[D] \cdot |\langle \phi, \hat{D} \rangle_{D_0}| \right)^2 \\
&= \left(\sum_{D \in \mathcal{D}[\phi,\tau,D_0]} P_D[D] \cdot |\mathbb{E}_D \phi - \mathbb{E}_{D_0} \phi| \right)^2 \\
&\ge \left(\sum_{D \in \mathcal{D}[\phi,\tau,D_0]} P_D[D] \cdot \tau \right)^2 \\
&= P_D[\mathcal{D}[\phi,\tau,D_0]]^2 \cdot \tau^2.
\end{aligned}$$

So overall

$$P_D[\mathcal{D}[\phi,\tau,D_0]] \le \frac{\Pi(P_D, D_0)^{1/2}}{\tau^2}.$$

□

Theorem 2.8. *Similarly,*

$$ASD(P_D, D_0, \tau) \ge \min_{\phi : \mathcal{Z} \to [-1,1]} \frac{1}{P_D[\mathcal{D}[\phi,\tau,D_0]]} \ge \frac{\tau}{\Pi_1(P_D, D_0)^{1/2}}$$

Proof. Define Ψ as in the proof of Theorem 2.7. Similarly to the previous proof

$$\begin{aligned}\Psi^2 &\le P_D[\mathcal{D}[\phi,\tau,D_0]]^2 \mathbb{E}_{D_1,D_2\sim P_D[\cdot|\mathcal{D}[\phi,\tau,D_0]]}[|\langle \hat{D}_1, \hat{D}_2\rangle_{D_0}|] \\ &\le \mathbb{E}_{D_1,D_2\sim P_D}[|\langle \hat{D}_1, \hat{D}_2\rangle|] = \Pi_1(P_D, D_0).\end{aligned}$$

And also $\Psi^2 \ge P_D[\mathcal{D}[\phi,\tau,D_0]]^2 \cdot \tau^2$, so

$$P_D[\mathcal{D}[\phi,\tau,D_0]] \le \frac{\Pi_1(P_D,D_0)^{1/2}}{\tau},$$

proving the theorem. □

2.4 Cross-predictability does not characterize ASD

We construct an example with high cross-predictability, but also high ASD. Let $\mathcal{X} = \{+1,-1\}^n$ be the input alphabet and let $\mathcal{Y} = \{+1,-1\}$ be the output alphabet. Let $\mathcal{F}_1$ be a set of 2^n standard basis functions. That is,

$$\mathcal{F}_1 = \{F_h : h \in \{-1,+1\}^n\},$$

where for each $h \in \{-1,+1\}^n$,

$$F_h(x) = \begin{cases} -1, & x \ne h \\ 1, & x = h \end{cases}.$$

Let $\mathcal{F}_2$ be the set of $2^n - 1$ nontrivial parity functions:

$$\mathcal{F}_2 = \{\chi_S : S \subseteq [n], S \ne \emptyset\},$$

where for each $S \subseteq [n]$,

$$\chi_S(x) = \prod_{i\in S} x_i.$$

Let $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$, and consider the uniform distribution P_F on $\mathcal{F}$ and the corresponding distribution P_D on $\mathcal{D}$, the set of distributions supported on $\mathcal{Z} = \{-1,+1\}^n \times \{-1,+1\}$. Let the reference distribution D_0 be the uniform distribution on $\mathcal{Z}$.

Then for any $F \ne F' \in \mathcal{F}$, if (i) $F, F' \in \mathcal{F}_1$, we have $\langle D_F, D_{F'}\rangle_{D_0} = 1 - o(1)$, if (ii) $F, F' \in \mathcal{F}_2$, we have $\langle D_F, D_{F'}\rangle_{D_0} = 0$, and if (iii) $F \in \mathcal{F}_1$ and $F' \in \mathcal{F}_2$ we have $\langle D_F, D_{F'}\rangle_{D_0} = o(1)$. So overall the cross-predictability is high:

$$\Pi_1(P_D, D_0) \approx \Pi(P_D, D_0) \approx \frac{1}{4}.$$

Nevertheless, the average statistical dimension of (P_D, D_0) is also high, because $\mathcal{D}$ contains a large subset $\mathcal{D}_2 = D_{\mathcal{F}_2}$ with high statistical dimension, since $P_D[\mathcal{D}_2] \ge 1/2$ and $\Pi(P_D[\cdot|\cdot \in \mathcal{D}_2], D_0) = o(1)$. Formally:

Lemma 2.9. *Suppose $\mathcal{D}' \subseteq \mathcal{D}$ and define $P_{D'} = P_D[\cdot|\cdot \in \mathcal{D}']$. Then for any D_0 and τ,*

$$ASD(P_D, D_0, \tau) \geq ASD(P_{D'}, D_0, \tau) \cdot P_D[\mathcal{D}'].$$

Proof. This is due to the fact that

$$\begin{aligned} ASD(P_D, D_0, \tau) &= \max_{\mathcal{D}'' \subseteq \mathcal{D}} \min_{\phi} \frac{P_D[\mathcal{D}'']^2}{P_D[\mathcal{D}''[\phi, \tau, D_0]]} \\ &\geq \max_{\mathcal{D}'' \subseteq \mathcal{D}'} \min_{\phi} \frac{P_D[\mathcal{D}'']^2}{P_D[\mathcal{D}''[\phi, \tau, D_0]]} \\ &= P_D[\mathcal{D}'] \cdot \max_{\mathcal{D}'' \subseteq \mathcal{D}'} \min_{\phi} \frac{P_{D'}[\mathcal{D}'']^2}{P_{D'}[\mathcal{D}''[\phi, \tau, D_0]]} \\ &= P_D[\mathcal{D}'] \cdot ASD(P_{D'}, D_0, \tau). \end{aligned}$$

□

Therefore in our example above

$$\begin{aligned} ASD(P_D, D_0, \tau) &\geq \frac{1}{2} \cdot ASD(P_D[\cdot|\cdot \in \mathcal{D}_2], D_0, \tau) \\ &\geq \frac{\tau^2}{2 \cdot \Pi(P_D[\cdot|\cdot \in \mathcal{D}_2], D_0)^{1/2}} = \omega(\tau^2), \end{aligned}$$

which is high, even though the cross-predictability is also high.

Motivated by this construction, we can modify the definition of the cross-predictability in order to strengthen Theorems 2.7 and 2.8.

Definition 2.10 (Modified cross-predictability)**.**

$$\Pi^M(P_D, D_0) = \min_{\mathcal{D}' \subseteq \mathcal{D}} \frac{\mathbb{E}_{D_1, D_2 \overset{i.i.d.}{\sim} P_D[\cdot|\cdot \in \mathcal{D}']}[\langle \hat{D}_1, \hat{D}_2 \rangle_{D_0}^2]}{P_D[\mathcal{D}']^2}.$$

Definition 2.11 (Modified L_1 cross-predictability)**.**

$$\Pi_1^M(P_D, D_0) = \min_{\mathcal{D}' \subseteq \mathcal{D}} \frac{\mathbb{E}_{D_1, D_2 \overset{i.i.d.}{\sim} P_D[\cdot|\cdot \in \mathcal{D}']}[|\langle \hat{D}_1, \hat{D}_2 \rangle_{D_0}|]}{P_D[\mathcal{D}']^2}.$$

And the modifications of Theorems 2.7 and 2.8 are:

Theorem 2.12 (Modification of Theorem 2.7)**.**

$$ASD(P_D, D_0, \tau) \geq \frac{\tau^2}{\Pi^M(P_D, D_0)^{1/2}}$$

Theorem 2.13 (Modification of Theorem 2.8)**.**

$$ASD(P_D, D_0, \tau) \geq \frac{\tau}{\Pi_1^M(P_D, D_0)^{1/2}}$$

The proofs are straightforward from Lemma 2.9 and Theorems 2.7 and 2.8.

3 Average-case statistical query complexity of weak learning

3.1 Average-case statistical query weak-learning with adversarial noise

Theorems 2.12 and 2.13 lower-bound the expected number of statistical queries needed to exactly learn F. In this section, we show how to lower-bound the expected number of statistical queries needed to weakly-learn F.

First we define weak learning, specifically for the case of balanced binary-valued functions:

Definition 3.1. *We say that (P_F, P_X) is a balanced distribution of binary functions if the image of the functions in P_F is binary ($\mathcal{Y} = \{-1, 1\}$), and they are balanced (i.e., $\mathbb{P}[F(X) = 0] = \mathbb{P}[F(X) = 1] + o_n(1)$ for $(X, F) \sim P_X \times P_F$).*

Definition 3.2. *Let A be a statistical query algorithm for a balanced-binary problem (P_F, P_X) that, given a statistical query oracle on $D = D_F$ for $F \sim P_F$ outputs a function $\hat{F} : \mathcal{X} \to \mathcal{Y}$ such that*

$$\mathbb{P}_{X,F,A}[\hat{F}(X) = F(X)] \geq \frac{1}{2} + \epsilon,$$

where the probability is over the internal randomness of A, and also $F \sim P_F$ and $X \sim P_X$.

We say that A weakly-learns, or ϵ-learns, (P_F, P_X).

A first observation is that low cross-predictability implies that one cannot weakly-learn $F \sim P_F$ without any queries:

Lemma 3.3. *Let $G : \mathcal{X} \to \{-1, +1\}$. Then*

$$\mathbb{P}_{X \sim P_X, F \sim P_F}[G(X) = F(X)] \leq \frac{1}{2} + \Pi_1(P_D, D_0)^{1/2} \leq \frac{1}{2} + \Pi_2(P_D, D_0)^{1/4}.$$

Proof. We have

$$\mathbb{P}[G(X) = F(X)] = \frac{1}{2} + \frac{\mathbb{E}[G(X)F(X)]}{2}.$$

Let D_0 be the uniform distribution on functions from $\mathcal{X}$ to $\mathcal{Y} = \{-1, 1\}$. We have

$$\begin{aligned}
\mathbb{E}[G(X)F(X)] &= \mathbb{E}_X[G(X)\mathbb{E}_F F(X)] \\
&\leq \left(\mathbb{E}_X[G(x)^2]\mathbb{E}_X[\mathbb{E}_F\mathbb{E}'_F[F(X)F'(X)]]\right)^{1/2} \qquad \text{(by Cauchy-Schwarz)} \\
&= \left(\mathbb{E}_{F,F'}(\mathbb{E}_X F(X)F'(X))\right)^{1/2} \\
&\leq \left(\mathbb{E}_{F,F'}|\mathbb{E}_X F(X)F'(X)|\right)^{1/2} \\
&= \Pi_1(P_D, D_0)^{1/2} \leq \Pi_2(P_D, D_0)^{1/4}
\end{aligned}$$

□

We show a lower bound on the number of queries necessary to weakly-learn P_F:

Theorem 3.4. *Suppose A makes at most n queries to an oracle $\mathcal{O}^{adv,\tau}$ with adversarial additive error $\leq \tau$, and suppose that A outputs an estimator $\hat{F}$ of F. If (P_X, P_F) is a balanced distribution of binary functions, we have*

$$\mathbb{P}_{X,F,A}[\hat{F}(X) = F(X)] \leq \frac{1}{2} + (\frac{n}{\tau} + 1)\Pi_1^{1/2}$$

and

$$\mathbb{P}_{X,F,A}[\hat{F}(X) = F(X)] \leq \frac{1}{2} + \frac{n\Pi^{1/2}}{\tau^2} + \Pi^{1/4}.$$

Proof. Let $F \sim P_F$ be the hidden function, and let D_F be the distribution corresponding to F. Let the reference distribution D_0 be the uniform distribution.

Consider an adversarial oracle that upon query ϕ returns $D_0[\phi] = 0$ if $|D_0[\phi] - D_F[\phi]| \leq \tau$ and returns $D_F[\phi]$ otherwise. Consider the first n statistical queries $\phi_1, \ldots, \phi_n$ made by A. Let $\mathcal{E}$ be the event that there exists i such that $|D_0[\phi_i] - D_F[\phi_i]| > \tau$. By a union bound,

$$\mathbb{P}[\mathcal{E}] \leq n \cdot \max_{\phi:\mathcal{Z}\to[-1,1]} P_D[\mathcal{D}[\phi_i, \tau, \phi]] \leq \frac{n\Pi(P_D, D_0)^{1/2}}{\tau^2}.$$

Note that if the event $\neg\mathcal{E}$ holds, then $\phi_1, \ldots, \phi_n$ are fixed (since the queries always return 0 in this case), so we can define $\mathcal{D}' \subset \mathcal{D}$ as $\mathcal{D}' = \{D : |D[\phi_i] - D_0[\phi_i]| \leq \tau \ \forall i\}$. Since $\mathsf{P}_D(\mathcal{D}') = 1 - \mathbb{P}[\mathcal{E}]$, then $\Pi_1(P_{D'}, D_0) \leq \frac{\Pi_1(P_D, D_0)}{P_D[\mathcal{D}']^2}$ and $\Pi_2(P_{D'}, D_0) \leq \frac{\Pi_2(P_D, D_0)}{P_D[\mathcal{D}']^2}$.

Let $\hat{F} : \mathcal{X} \to \{-1, 1\}$ is the estimator outputted by A. Notice that $\mathbb{P}[\hat{F}(X) = F(X)|\mathcal{E}] \leq \Pi_1(P_{D'}, D_0)^{1/2} \leq \frac{1}{2} + \Pi_2(P_{D'}, D_0)^{1/4}$,. Thus since

$$\mathbb{P}[\hat{F}(X) = F(X)] \leq \mathbb{P}[\mathcal{E}] + \mathbb{P}[\hat{F}(X) = F(X), \neg\mathcal{E}],$$

we have

$$\mathbb{P}[\hat{F}(X) = F(X)] \leq \frac{1}{2} + \frac{n\Pi_2^{1/2}}{\tau^2} + \Pi_2^{1/4}.$$

And

$$\mathbb{P}[\hat{F}(X) = F(X)] \leq \frac{1}{2} + n\Pi_1^{1/2}/\tau + \Pi_1^{1/2},$$

as well.

□

3.2 Average-case statistical query weak-learning with statistical noise

We extend the bound of the previous section to work with statistical $N(0, s^2)$ noise on each oracle query. The main argument in this section was also pointed out in [7]. The idea is that adding $N(0, s^2)$ noise on top of any adversarial magnitude-$O(cs/n)$ additive noise,

one obtains obtain noise that is c/n-close in total variation to $N(0, s^2)$. In particular, if one peforms n adversarial queries and adds $N(0, s^2)$ i.i.d. noise to each one, the resulting n values are at a total $\leq c$ total variation distance from n purely statistical queries with noise $N(0, s^2)$. Thus, as a corollary of Theorem 3.4:

Theorem 3.5. *In the balanced binary P_F setting of Theorem 3.4 let algorithm A make n statistical queries to the oracle $\mathcal{O}^{stat,\sigma}$ with statistical noise distributed as $N(0, s^2)$ (where $s \leq 1$) and output estimator $\hat{F}$ of F. Then*

$$\mathbb{P}[\hat{F}(X) = F(X)] \leq .\frac{1}{2} + O(n\Pi_1^{1/4}/\sqrt{s})$$

and

$$\mathbb{P}[\hat{F}(X) = F(X)] \leq \frac{1}{2} + O(n\Pi^{1/12}/s^{2/3}).$$

Proof. (a) Π_1 bound: By the above discussion, for any $c > 0$ we have

$$\mathbb{P}[\hat{F}(X) = F(X)] \leq \frac{1}{2} + O(n^2/(cs) + 1)\Pi_1^{1/2}) + c.$$

To minimize the right-hand side set $c = (n^2/(cs))\Pi_1^{1/2}$ so that $c = n\Pi_1^{1/4}/\sqrt{s}$. (b) Π bound: Again, for all $c > 0$ we have

$$\mathbb{P}[\hat{F}(X) = F(X)] \leq \frac{1}{2} + O(n^3\Pi^{1/4}/(c^2s^2)) + \Pi^{1/4} + c.$$

To minimize the right-hand side set $c = (n^3\Pi^{1/4})/(c^2s^2)$ so that $c = \Pi^{1/12}n/s^{2/3}$. □

4 Applications to neural networks

In the neural network model of Abbe and Sandon [1], a neural network is trained to learn a function $F \sim P_F$ drawn from a balanced distribution of binary functions (P_F, P_X) as in definition 3.1. The neural network has size $|E|$, any differentiable non-linearity and any initialization of the weights $W^{(0)}$. It is trained with gradient descent with learning rate γ on any differentiable loss function for S steps as follows:

$$W^{(t)} = W^{(t-1)} - \gamma \left[\nabla \mathbb{E}_{x \sim P_X} L(W^{(t-1)}(X), F(X))\right]_A + Z^{(t)}, \quad t = 1, \ldots, S.$$

The gradients of the network have an overflow range given by A and on each step Gaussian noise of variance σ^2 is added.

Abbe and Sandon [1] obtain the following error bound on the success probability of the neural network:

$$\mathbb{P}\left\{W^{(S)}(X) = F(X)\right\} \leq 1/2 + \text{grapes}, \tag{2}$$

where

$$\text{grapes} = \gamma \frac{1}{\sigma} A \Pi^{1/4} |E|^{1/2} S. \tag{3}$$

Therefore, in order to **exactly learn** F one needs $\gamma \frac{1}{\sigma} A \Pi^{1/4} |E|^{1/2} S = \Omega(1)$.

Another way to obtain a lower bound on the number of steps necessary to train the neural network is to view gradient descent as a statistical query algorithm: each time step $t \in \{1, \ldots, S\}$ involves $|E|$ different edge update steps. Each edge update step needs to make only one statistical query to the gradient of the loss. The query returns a value in the range $[-\gamma A, \gamma A]$, and the added noise is distributed as $N(0, \sigma^2)$. In other words training a neural network involves $n = |E|S$ queries in the range $[-1, 1]$ with noise distributed as $N(0, s^2) = N(0, (\frac{\sigma}{\gamma A})^2)$. How do the lower bounds from the statistical query method compare to the lower bounds from [1]?

SQ adversarial-noise exact-learning If we knew that the neural nets algorithm worked with **adversarial** additive noise of magnitude s, then in order to **exactly learn** the parity function, by Theorems 2.7 and 2.8 we would need in expectation $n \geq s^2/\Pi^{1/2}$ and $n \geq s/\Pi_1^{1/2}$. This corresponds to the bounds

$$\gamma^2 \frac{1}{\sigma^2} A^2 \Pi^{1/2} |E| (\mathbb{E}S) \geq \Omega(1), \tag{4}$$

and

$$\gamma \frac{1}{\sigma} A \Pi_1^{1/2} |E| (\mathbb{E}S) \geq \Omega(1), \tag{5}$$

respectively. $\mathbb{E}S$ denotes the number of steps that the neural network must take in expectation in order to return the correct parity function with no error probability.

The adversarial SQ bounds (4) and (5) are incomparable to [1], and they can be stronger than the bound from [1] if Π or Π_1 is very small, for example. However, the bound of [1] is more refined in that (a) handles statistical noise instead of adversarial noise, and (b) lower-bounds the error probability of the algorithm – instead of just handling the zero-error case.

SQ statistical-noise weak-learning The case in which the algorithm has statistical $N(0, s^2) = N(0, (\sigma/(\gamma A))^2)$ noise and in which we care about weak learning is directly comparable to [1]. For this case, the bounds that we get from Theorem 3.5 are the following:

$$\mathbb{P}\{W^{(S)}(X) = F(X)\} \leq \frac{1}{2} + O(\gamma^{1/2} \frac{1}{\sigma^{1/2}} A^{1/2} \Pi_1^{1/4} |E| S) \tag{6}$$

and using the Π cross-predictability instead of the Π_1 cross-predictability:

$$\mathbb{P}\{W^{(S)}(X) = F(X)\} \leq \frac{1}{2} + O(\gamma^{2/3} \frac{1}{\sigma^{2/3}} A^{2/3} \Pi^{1/12} |E| S) \tag{7}$$

Assuming $s = \sigma/(\gamma A) \leq 1$, the bound of [1] is stronger than both of these.[4]

Can the SQ statistical-noise weak-learning bound be improved? If we assume additive adversarial additive $s = \sigma/(\gamma A)$ noise instead of statistical $N(0, s^2)$ noise then we obtain the following bounds from Theorem 3.4:

$$\mathbb{P}\{W^{(S)}(X) = F(X)\} \leq \frac{1}{2} + O(\gamma \frac{1}{\sigma} A \Pi_1^{1/2} |E|S) \tag{8}$$

and using the Π cross-predictability instead of the Π_1 cross-predictability:

$$\mathbb{P}\{W^{(S)}(X) = F(X)\} \leq \frac{1}{2} + \gamma^2 \frac{1}{\sigma^2} A^2 \Pi^{1/2} |E|S + \Pi^{1/4} \tag{9}$$

These bounds could be stronger than [1] if Π_1 (respectively Π) were very small. However, they apply to **adversarial** noise – not statistical noise. Note also that in the case of very small Π_1 such as for parities, [1] obtains a tighter bound than the one used above.

References

[1] Emmanuel Abbe and Colin Sandon. "Provable limitations of deep learning". In: *arXiv preprint arXiv:1812.06369* (2018).

[2] Avrim Blum et al. "Weakly learning DNF and characterizing statistical query learning using Fourier analysis". In:

[3] Vitaly Feldman. "A general characterization of the statistical query complexity". In: *arXiv preprint arXiv:1608.02198* (2016).

[4] Vitaly Feldman, Will Perkins, and Santosh Vempala. "On the complexity of random satisfiability problems with planted solutions". In: *SIAM Journal on Computing* 47.4 (2018), pp. 1294–1338.

[5] Vitaly Feldman et al. "Statistical algorithms and a lower bound for detecting planted cliques". In: *Journal of the ACM (JACM)* 64.2 (2017), p. 8.

[6] Michael Kearns. "Efficient noise-tolerant learning from statistical queries". In: *Journal of the ACM (JACM)* 45.6 (1998), pp. 983–1006.

[7] Oded Regev. *Personal communication*. 2019.

[8] Leslie G Valiant. "A theory of the learnable". In: *Proceedings of the sixteenth annual ACM symposium on Theory of computing*. ACM. 1984, pp. 436–445.

[9] Ke Yang. "New lower bounds for statistical query learning". In: *Journal of Computer and System Sciences* 70.4 (2005), pp. 485–509.

[4] Perhaps better bounds than (6) and (7) can be proved for the statistical noise case using the statistical query approach (e.g., by leveraging techniques of [9] or [5] for the 1-STAT model).