Learning without Interaction Requires Separation

Amit Daniely
Hebrew University and Google

Vitaly Feldman
Google Brain

Abstract

One of the key resources in large-scale learning systems is the number of rounds of communication between the server and the clients holding the data points. We study this resource for systems with two types of constraints on the communication from each of the clients: local differential privacy and limited number of bits communicated. For both models the number of rounds of communications is captured by the number of rounds of interaction when solving the learning problem in the statistical query (SQ) model. For many learning problems known efficient algorithms require many rounds of interaction. Yet little is known on whether this is actually necessary. In the context of classification in the PAC learning model, Kasiviswanathan et al. [KLNRS11] constructed an artificial class of functions that is PAC learnable with respect to a fixed distribution but cannot be learned by an efficient non-interactive (or one-round) SQ algorithm. Here we show that a similar separation holds for learning linear separators and decision lists without assumptions on the distribution. To prove this separation we show that non-interactive SQ algorithms can only learn function classes of low margin complexity, that is classes of functions that can be represented as large-margin linear separators.

1 Overview

We consider learning in distributed systems where each client i (or user) holds a data point $z_i \in Z$ drawn i.i.d. from some unknown distribution P and the goal of the server is to solve some statistical learning problem using the data stored at the clients. In addition, the communication from the client to the server is constrained. The first constraint we consider is that of local differential privacy (LDP) [KLNRS11]. In this model each user i applies a differentially-private algorithm to their point z_i and then sends the result to the server. The specific algorithm applied by each user is determined by the server. In the general version of the model the server can determine which LDP algorithm the user should apply on the basis of all the previous communications the server has received. In practice, however waiting for the client's response often takes a relatively large amount of time. Therefore in such systems it is necessary to limit the number of rounds of interaction. That is, the queries of the server need to be split into a small number of batches such that the LDP algorithms used in each batch depend only on responses to queries in previous batches (a query specifies the LDP algorithm to apply). Indeed, currently deployed systems that use local differential privacy use very few rounds (usually just one) [EPK14; App17; DKY17].

The second constraint we consider is the small number of bits communicated from each client. Namely, each client applies a function with range $\{0,1\}^k$ to their input and sends the result to the server (for some $k \ll \log |Z|$). As in the case of LDP, the specific function used is chosen by the server. One motivation for this model is collection of data from remote sensors where the cost of communication is highly asymmetric. In the context of learning this model was introduced by Ben-David and Dichterman [BD98] and generalized by Steinhardt et al. [SVW16]. Identical and closely related models are often studied in the context of distributed statistical estimation with communication constraints (*e.g.* [Luo05; RWV06; RG06; ZDJW13;

SD15; SYMK16]). As in the setting of LDP, the number of rounds of interaction that the server uses to solve a learning problem in this model is a critical resource.

To understand the round complexity of solving a learning problem in these models we will use the fact that both of these models are known to be closely related to learning in the statistical query model of Kearns [Kea98]. In this model an algorithm has access to a statistical query oracle for P in place of i.i.d. samples from P. The most commonly studied SQ oracle give an estimate of the mean of any bounded function with fixed tolerance.

Definition 1.1. *Let P be a distribution over a domain Z and $\tau > 0$. A statistical query oracle STAT$_P(\tau)$ is an oracle that given as input any function $\phi : Z \to [-1, 1]$, returns some value v such that $|v - \mathbf{E}_{z\sim P}[\phi(z)]| \leq \tau$.*

Tolerance τ of statistical queries roughly corresponds to the number of random samples in the traditional setting. Namely, the Chernoff-Hoeffding bounds imply that n i.i.d. samples allow estimation of $\mathbf{E}_P[\phi]$ with tolerance $\tau = \Theta(1/\sqrt{n})$ (with high probability). A special case of statistical queries are counting or linear queries in which the distribution P is uniform over the elements of a given database $S \in Z^n$. In other words the goal is to estimate the empirical mean of ϕ on the given set of data points. This setting is studied extensively in the literature on differential privacy (see [DR14] for an overview) and our discussion applies to this setting as well.

An algorithm $\mathcal{A}$ in the SQ model is said to have r rounds of interaction if the queries asked by $\mathcal{A}$ can be split into r batches in such a way that the queries in batch t only depend on answers to queries in the previous batches. We also say that the algorithm is non-interactive (also referred as non-adaptive) if it requires only one round of interaction. Reductions between learning in the SQ model and the two constrained communication models above were given by Kasiviswanathan et al. [KLNRS11] and Steinhardt et al. [SVW16]. For our purposes it is important to note that all these four reductions preserve the number of rounds of interaction of a learning algorithm. In particular, this implies that it is sufficient for our purposes to study the round complexity of solving the learning problem in the SQ model.

In this paper we will focus on the standard PAC learning of a class of Boolean functions C over some domain X. In this setting the input distribution P is over labeled examples $(x, y) \in X \times \{-1, 1\}$ where x is drawn from some distribution D and $y = f(x)$ for some unknown $f \in C$ (referred to as the target function). The goal of the learning algorithm is to output a function h such that the error $\mathbf{Pr}_{x\sim D}[f(x) \neq h(x)]$ is small. In the distribution-independent setting D is not known to the learning algorithm while in the distribution-specific setting the learning algorithm only needs to succeed for some specific D.

For many of the important classes of functions all known learning algorithms require many rounds of interaction. Yet there are almost no known lower bounds on the round complexity of SQ algorithms. The only example that we are aware of is the result of Kasiviswanathan et al. [KLNRS11] who were also motivated by the local differential privacy model. They constructed a class of functions C over $\{0, 1\}^d$ that can be PAC learned relative to the uniform distribution over $\{0, 1\}^d$ in the SQ model. At the same time C cannot be learned by an efficient non-interactive SQ algorithm (the complexity is exponential in d). The class C is highly unnatural. It splits the domain into two parts. Target function learned on the first half gives the key to the learning problem on the second half of the domain. That problem is exponentially hard to solve without the key. This approach does not extend to distribution-independent learning setting (intuitively, the learning algorithm will not be able to obtain the key if the distribution does not place any probability on the first half of the domain).

1.1 Our Result

We give a separation between the power of interactive and non-interactive algorithms for distribution-independent PAC learning of two natural classes of Boolean functions. Specifically, we prove that only classes that have polynomially small *margin complexity* can be efficiently PAC learned by a non-interactive SQ algorithm. The margin complexity of a class of Boolean functions C, denoted by $\mathsf{MC}(C)$, measures (the inverse of) the largest margin of separation achievable by an embedding of X in $\mathbb{R}^d$ that makes the positive and negative examples of each function in C linearly separable (see Definition 2.9). It is a well-studied measure of complexity of classes of functions and corresponding sign matrices in learning theory and communication complexity (*e.g.* [Nov62; ABR64; BGV92; FSS01; She08; LS09; KS11]).

Theorem 1.2. *Let C be a class of Boolean functions closed under negation. Assume that for some m there exists a non-interactive possibly randomized SQ algorithm $\mathcal{A}$ that, with success probability at least $2/3$, PAC learns C with error less than $1/2$ using at most m queries to STAT$(1/m)$. Then* $\mathsf{MC}(C) \leq \frac{1}{6}m^{3/2}$.

The class of decision lists and the class of linear separators (or halfspaces) over $\{0,1\}^d$ are known to have exponentially large margin complexity [GHR92; BVW07; She08] (and are also negation closed). In contrast, these classes are known to be learnable efficiently by SQ algorithms [Kea98; DV04]. Combining these results with Theorem 1.2 gives the claimed separation for the SQ model. Using the reductions between the SQ model and the two models of distributed learning we obtain the separation in those models.

1.2 Related work

Smith et al. [STU17] address the question of the power of non-interactive LDP algorithms in the closely related setting of stochastic convex optimization. They derive new non-interactive LDP algorithms for the problem albeit requiring an exponential dependence in the dimension number of queries. They also give a strong lower bound for non-interactive algorithms that are further restricted to obtain only local information about the optimized function. Upper and lower bounds on the number of queries to the gradient oracle for algorithms with few rounds of interaction have been recently studied by Duchi et al. [DRY18]. In the context of discrete optimization from queries for the value of the optimized function the round complexity has been recently investigated by Balkanski et al. [BRS17] and Balkanski and Singer [BS18]. To the best of our knowledge, the techniques used in these works are unrelated to ours.

The round complexity of PAC learning a class of functions C by an SQ algorithm has been shown to determine the number of generations necessary to evolve C in a variant of Valiant's model of evolvability [Val09; Kan11]. The number of data samples necessary to answer statistical queries chosen interactively has recently been studied in a line of work on adaptive data analysis [DFHPRR14; HU14; BNSSSU16; SU15].

2 Preliminaries

For integer $n \geq 1$ let $[n] \doteq \{1, \ldots, n\}$.

2.1 Local Differential Privacy

In the local differential privacy (LDP) model [War65; KLNRS11] it is assumed that each data sample obtained by the server is randomized in a differentially private way. This is modeled by assuming that the server running the learning algorithm accesses the dataset via an oracle defined below.

Definition 2.1. *An ϵ-local randomizer $R : Z \to W$ is a randomized algorithm that satisfies $\forall z_1, z_2 \in Z$ and $w \in W$, $\mathbf{Pr}[R(z_1) = w] \leq e^{\epsilon}\,\mathbf{Pr}[R(z_2) = w]$. For a dataset $S \in Z^n$, an LR_S oracle takes as an input an index i and a local randomizer R and outputs a random value w obtained by applying $R(z_i)$. An algorithm is ϵ-LDP if it accesses S only via the LR_S oracle with the following restriction: for all $i \in [n]$, if $\mathrm{LR}_S(i, R_1), \ldots, \mathrm{LR}_S(i, R_k)$ are the algorithm's invocations of LR_S on index i where each R_j is an ϵ_j-randomizer then $\sum_{j \in [k]} \epsilon_j \leq \epsilon$.*

This model can be contrasted with the standard, or central, model of differential privacy where the entire dataset is held by the learning algorithm whose output needs to satisfy differential privacy [DMNS06]. This is a stronger model and an ϵ-LPD algorithm also satisfies ϵ-differential privacy.

2.2 Bounded communication

In the bounded communication model [BD98; SVW16] it is assumed that the total number of bits learned by the server about each data sample is bounded by ℓ for some $\ell \ll \log|Z|$. As in the case of LDP this is modeled by using an appropriate oracle for accessing the dataset.

Definition 2.2. *We say that an algorithm $R\colon Z \to \{0,1\}^{\ell}$ extracts ℓ bits. For a dataset $S \in Z^n$, an COMM_S oracle takes as an input an index i and an algorithm R and outputs a random value w obtained by applying $R(z_i)$. An algorithm is ℓ-bit communication bounded if it accesses S only via the COMM_S oracle with the following restriction: for all $i \in [n]$, if $\mathrm{COMM}_S(i, R_1), \ldots, \mathrm{COMM}_S(i, R_k)$ are the algorithm's invocations of COMM_S on index i where each R_j extracts ℓ_j bits then $\sum_{j \in [k]} \ell_j \leq \ell$.*

2.3 Loss minimization algorithm

Let $W \subset \mathbb{R}^d$ and let $F : W \times X \to \mathbb{R}$ a function such that for any $x \in X$, $\mathbf{w} \mapsto F(\mathbf{w}, x)$ is differentiable and L_1-Lipschitz. Let $\ell : \mathbb{R} \times Y$ be a differentiable and L_2-Lipschitz loss function. For a distribution $\mathcal{D}$ on $X \times Y$ and $f : X \to \mathbb{R}$ we denote $\ell_{\mathcal{D}}(f) = \mathbf{E}_{(x,y)\sim\mathcal{D}}\,\ell(f(x), y)$.

A *gradient oracle* $\mathrm{GRAD}_{\mathcal{D},F,\ell}(\tau)$ gets a vector $\mathbf{w} \in W$ and return a vector $g \in W$ such that $\|g - \nabla\ell_{\mathcal{D}}(\mathbf{w})\| \leq L_1 L_2 \tau$ [[[TODO: Motivate multiplication by τ]]]

A *loss minimization algorithm* $\mathcal{A}$ is defined by a sequence of update rules $U_t : W^t \to W$ a tolerance parameter τ, step limit T, and initial vector $\mathbf{w}_1 \in W$. At it step $t = 1, 2, \ldots, T$, it call $\mathrm{GRAD}_{\mathcal{D},F,\ell}(\tau)$ with $\mathbf{w}_t$ and get an estimated gradient $\mathbf{g}_t$. If $\|\mathbf{g}_t\| \leq \tau$ or $t = T$, it halts end return $x \mapsto F(\mathbf{w}_t, x)$. Else, it updates $\mathbf{w}_{t+1} = \mathbf{w}_t + U_t(\mathbf{g}_1, \ldots, \mathbf{g}_t)$. The *complexity* of $\mathcal{A}$ is the maximal number of steps that it runs, multiplied by $\frac{1}{\tau}$. [[[TODO: put the maximal number of iteration in the definition of the algorithm?]]]

[[[TODO: motivate the definition. Give examples – Neural networks and SVMs]]]

Theorem 2.3. *Fix an initial vector $\mathbf{w}_1 \in W$ Let $\mathcal{H} \subset \{\pm 1\}^{\{\pm 1\}^n}$. There are are two $\mathcal{H}$-realizable [[[TODO: define]]] distributions $\mathcal{D}_1$, $\mathcal{D}_2$ such that*

1. *[[[TODO: $\mathcal{D}_1$ and $\mathcal{D}_2$ are different]]]*

2. *[[[$\nabla\ell_{\mathcal{D}_1}(\mathbf{w}_1) \approx \nabla\ell_{\mathcal{D}_1}(\mathbf{w}_1)$]]]*

[[[TODO: Two different lemmas – for biased and un biased initialization?]]]
[[[TODO: corollary for loss minimization algorithms]]]
[[[TODO: open qeustions: (i) Non-zero initializations (ii) Random $\mathbf{w}_1$]]]

2.4 Equivalence to statistical queries

The third model we consider is the statistical query model of Kearns [Kea98] that is defined by having access to $\mathrm{STAT}_P(\tau)$ oracle, where P is the unknown data distribution (see Def. 1.1). To solve a learning problem in this model an algorithm needs to succeed for any valid (that is satisfying the guarantees on the tolerance) oracle's responses. In other words, the guarantees of the algorithm should hold in the worst case over the responses of the oracle. A randomized learning algorithm needs to succeed for any SQ oracle whose responses may depend on the all queries asked so far but not on the internal randomness of the learning algorithm.

For an algorithm in any of these oracle models we say that the algorithm is *non-interactive* (or *non-adaptive* if all its queries are determined before observing any of the oracle's responses.

Kasiviswanathan et al. [KLNRS11] show that one can simulate $\mathrm{STAT}_P(\tau)$ oracle with success probability $1-\delta$ by an ϵ-LDP algorithm using LR_S oracle for S containing $n = O(\log(1/\delta)/(\epsilon\tau)^2)$ i.i.d. samples from P. This has the following implication for simulating SQ algorithms.

Theorem 2.4 ([KLNRS11])**.** *Let $\mathcal{A}_{SQ}$ be an algorithm that makes at most t queries to STAT$_P(\tau)$. Then for every $\epsilon > 0$ and $\delta > 0$ there is an ϵ-local algorithm $\mathcal{A}$ that uses LR_S oracle for S containing $n \geq n_0 = O(t\log(t/\delta)/(\epsilon\tau)^2)$ i.i.d. samples from P and produces the same output as $\mathcal{A}_{SQ}$ (for some valid answers of STAT$_P(\tau)$) with probability at least $1-\delta$. Further, if $\mathcal{A}_{SQ}$ is non-interactive then $\mathcal{A}$ is non-interactive.*

Kasiviswanathan et al. [KLNRS11] also prove a converse of this theorem.

Theorem 2.5 ([KLNRS11])**.** *Let $\mathcal{A}$ be an ϵ-LPD algorithm that makes at most t queries to LR_S for S drawn i.i.d. from P^n. Then for every $\delta > 0$ there is an SQ algorithm $\mathcal{A}_{SQ}$ that in expectation makes $O(t \cdot e^\epsilon)$ queries to STAT$_P(\tau)$ for $\tau = \Theta(\delta/(e^{2\epsilon}t)$ and produces the same output as $\mathcal{A}$ with probability at least $1-\delta$. Further, if $\mathcal{A}$ is non-interactive then $\mathcal{A}_{SQ}$ is non-interactive.*

As first observed by Ben-David and Dichterman [BD98], it is easy to simulate a single query to $\mathrm{STAT}_P(\tau)$ by extracting a single bit from each of the $O(1/\tau^2)$ samples. This gives the following simulation.

Theorem 2.6 ([BD98])**.** *Let $\mathcal{A}_{SQ}$ be an algorithm that makes at most t queries to STAT$_P(\tau)$. Then for every $\delta > 0$ there is an ϵ-local algorithm $\mathcal{A}$ that uses COMM_S oracle for S containing $n \geq n_0 = O(t\log(t/\delta)/\tau^2)$ i.i.d. samples from P and produces the same output as $\mathcal{A}_{SQ}$ (for some valid answers of STAT$_P(\tau)$) with probability at least $1-\delta$. Further, if $\mathcal{A}_{SQ}$ is non-interactive then $\mathcal{A}$ is non-interactive.*

The converse of this theorem for the simpler COMM oracle that accesses each sample once was given in [BD98; FGRVX12]. For the stronger oracle in Definition 2.2, the converse was given by Steinhardt et al. [SVW16].

Theorem 2.7 ([SVW16])**.** *Let $\mathcal{A}$ be an ℓ-bit communication bounded algorithm that makes queries to COMM_S for S drawn i.i.d. from P^n. Then for every $\delta > 0$, there is an SQ algorithm $\mathcal{A}_{SQ}$ that makes $2n\ell$ queries to STAT$_P\left(\delta/(2^{\ell+1}n)\right)$ and produces the same output as $\mathcal{A}$ with probability at least $1-\delta$. Further, if $\mathcal{A}$ is non-interactive then $\mathcal{A}_{SQ}$ is non-interactive.*

Note that in this simulation we do not need to assume a separate bound on the number of queries since at most ℓn queries can be asked.

2.5 PAC Learning and Margin complexity

Our results are for the standard PAC model of learning [Val84].

Definition 2.8. *Let X be a domain and C be a class of Boolean functions over X. An algorithm $\mathcal{A}$ is said to PAC learn C with error α if for every distribution D over X and $f \in C$, given access (via oracle or samples) to the input distribution over examples $(x, f(x))$ for $x \sim D$, the algorithm outputs a function h such that $\mathbf{Pr}_D[f(x) \neq h(x)] \leq \alpha$.*

We say that the learning algorithm is efficient if its running time is polynomial in $\log|X|$, $\log|C|$ and $1/\epsilon$.

We say that a class of Boolean ($\{-1,1\}$-valued) functions C is closed under negation if for every $f \in C$, $-f \in C$. For dimension d, we denote by $\mathcal{B}^d(1)$ the unit ball in ℓ_2 norm in $\mathbb{R}^d$.

Definition 2.9. *Let X be a domain and C be a class of Boolean functions over X. The* margin complexity *of C, denoted $\mathsf{MC}(C)$, is the minimal number $M \geq 0$ such that for some d, there is an embedding $\Psi : X \to \mathcal{B}^d(1)$ for which the following holds: for every $f \in C$ there is $w \in \mathcal{B}^d(1)$ such that*

$$\min_{x \in X}\{f(x) \cdot \langle w, \Psi(x)\rangle\} \geq \frac{1}{M}.$$

We remark that margin complexity is closely related to the smallest dimension d for which for which there exists a mapping of X to $\{0,1\}^d$ such that every $f \in C$ becomes expressible as a majority function over some subset $T \subset [d]$ of variables (a majority function is equal to 1 if and only if the number of variables with indices in T that are equal to 1 is larger than the number of those in T set to 0).

As pointed out in [Fel08], margin complexity is equivalent (up to a polynomial) to the existence of a (possibly randomized) algorithm that outputs a small set of functions such that with significant probability one of those functions is correlated with the target function. The upper bound in [Fel08] was sharpened by Kallweit and Simon [KS11] although they proved it only for determistic algorithms (which corresponds to a single fixed set of functions). It is however easy to see that their sharper bound extends to randomized algorithms and we give the resulting statement below:

Lemma 2.10 ([Fel08; KS11])**.** *Let X be a domain and C be a class of Boolean functions over X. Assume that there exists a (possibly randomized) algorithm $\mathcal{A}$ that generates a set of functions $h_1, \ldots, h_m$ satisfying: for every $f \in C$ and distribution D over X with probability at least $\beta > 0$ (over the randomness of $\mathcal{A}$) there exists $i \in [m]$ such that $|\mathbf{E}_{x\sim D}[f(x)h_i(x)]| \geq 1/m$. Then*

$$\mathsf{MC}(C) \leq \frac{\beta}{2} m^{3/2}.$$

The conditions in Lemma 2.10 are also known to be necessary for low margin complexity.

Lemma 2.11 ([Fel08; KS11])**.** *Let X be a domain, C be a class of Boolean functions over X and $d = \mathsf{MC}(C)$. Then for $m = O(\ln(|C||X|)d^2)$, there exists a set of functions $h_1, \ldots, h_m$ satisfying: for every $f \in C$ and distribution D over X there exists $i \in [m]$ such that $|\mathbf{E}_{x\sim D}[f(x)h_i(x)]| \geq 1/m$.*

3 Lower Bounds for Non-Interactive Algorithms

We start by proving our main result.

Proof of Theorem 1.2. We first recall a simple observation from [BF02] that allows to decompose each statistical query into a correlational and target-independent parts. Namely, for a function $\phi\colon X \times \{-1,1\} \to [-1,1]$,

$$\phi(x,y) = \frac{1-y}{2}\phi(x,-1) + \frac{1+y}{2}\phi(x,1) = \frac{\phi(x,-1)+\phi(x,1)}{2} + y \cdot \frac{\phi(x,1)-\phi(x,-1)}{2}.$$

For a query ϕ, we will use h and g to denote the parts of the decomposition $\phi(x,y) = g(x) + yh(x)$:

$$h(x) \doteq \frac{\phi(x,1)-\phi(x,-1)}{2}$$

and

$$g(x) \doteq \frac{\phi(x,1)+\phi(x,-1)}{2}.$$

For every input distribution D and target functions f, we define the following SQ oracle. Given a query ϕ, if $|\mathbf{E}_D[f(x)g(x)]| \geq 1/m$ then the oracle provides the exact expectation $\mathbf{E}_D[\phi(x,f(x)]$ as the response. Otherwise it answers with $\mathbf{E}_D[g(x)]$. Note that, by the properties of the decomposition, this is a valid implementation of the SQ oracle.

Let $\mathcal{A}(r)$ denote $\mathcal{A}$ with its random bits set to r, where r is drawn from some distribution R. Let $\phi_1^r, \ldots, \phi_m^r\colon X \times \{-1,1\} \to [-1,1]$ be the non-interactive statistical queries asked by $\mathcal{A}(r)$. Let g_i^r and h_i^r denote the decomposition of these queries into correlational and target-independent parts. Let $h_{f,D}^r$ denote the hypothesis output by $\mathcal{A}(r)$ when used with the SQ oracle defined above.

We claim that if $\mathcal{A}$ achieves error $< 1/2$ with probability at least $2/3$, then for every $f \in C$ and distribution D, with probability at least $1/3$, there exists $i \in [m]$ such that $|\,\mathbf{E}_D[f(x)h_i^r(x)]| \geq 1/m$ (satisfying the conditions of Lemma 2.10 with $\beta = 1/3$). To see this, assume for the sake of contradiction that for some distribution D and function $f \in C$,

$$\mathop{\mathbf{Pr}}_{r\sim R}[r \in T(f,D)] > 2/3,$$

where $T(f,D)$ is the set of all random strings r such that for all $i \in [m]$, $|\,\mathbf{E}_D[f(x)h_i^r(x)]| < 1/m$. Let $S(f,D)$ denote the set of random strings r for which $\mathcal{A}$ succeeds (with the given SQ oracle), that is $\mathbf{Pr}_D[f(x) \neq h_{f,D}^r(x)] < 1/2$.

By our assumption, $\mathbf{Pr}_{r\sim R}[r \in S(f,D)] \geq 2/3$ and therefore

$$\mathop{\mathbf{Pr}}_{r\sim R}[r \in T(f,D) \cap S(f,D)] > 1/3. \tag{1}$$

Now, observe that $T(-f,D) = T(f,D)$ and, in particular, the answers of our SQ oracle to $\mathcal{A}(r)$'s queries are identical for f and $-f$ whenever $r \in T(f,D)$. Further, if $\mathbf{Pr}_D[f(x) \neq h_{f,D}^r(x)] < 1/2$ then $\mathbf{Pr}_D[-f(x) \neq h_{f,D}^r(x)] > 1/2$. This means that for every $r \in T(f,D) \cap S(f,D)$, $\mathcal{A}(r)$ fails for the target function is $-f$ and the distribution D (by definition, $-f \in C$). By eq. (1) we obtain that $\mathcal{A}$ fails with probability $> 1/3$ for $-f$ and D. This contradicts our assumption and therefore we obtain that

$$\mathop{\mathbf{Pr}}_{r\sim R}[r \notin T(f,D)] \geq 1/3.$$

By Lemma 2.10, we obtain the claim. □

We will now apply this result to obtain the claimed separations. We start with the class of halfspaces over $\{0,1\}^d$ which we denote by C_{HS}. We will use the following rather involved result about the margin complexity of halfspaces.

Theorem 3.1 ([GHR92; She08])**.** $\mathsf{MC}(C_{HS}) = 2^{\Omega(d)}$.

Combining this result with Theorem 1.2 we obtain that:

Corollary 3.2. *Any non-interactive SQ algorithm that PAC learns C_{HS} over $\{0,1\}^d$ with error less than $1/2$ and success probability $\geq 2/3$ using at most m queries to STAT$(1/m)$ must have $m = 2^{\Omega(d)}$.*

On the other hand, Dunagan and Vempala [DV04] give an efficient algorithm for PAC learning halfspaces (their description is not in the SQ model but it is known that their algorithm can be easily converted to the SQ model [BF15]).

Theorem 3.3 ([DV04; BF15])**.** *There exists an SQ learning algorithm that for every $\alpha > 0$, PAC learns C_{HS} with error α using $\mathrm{poly}(d/\alpha)$ queries to STAT$(1/\mathrm{poly}(d/\alpha))$.*

A similar separation also holds for the more restricted class of decision lists over $\{0,1\}^d$ that we denote by C_{DL} (see [KV94] for a definition).

Theorem 3.4 ([BVW07])**.** $\mathsf{MC}(C_{DL}) = 2^{\Omega(d^{1/3})}$.

Corollary 3.5. *Any non-interactive SQ algorithm that PAC learns C_{DL} over $\{0,1\}^d$ with error less than $1/2$ and success probability $\geq 2/3$ using at most m queries to STAT$(1/m)$ must have $m = 2^{\Omega(d^{1/3})}$.*

On the other hand a simple algorithm of Kearns [Kea98] shows that decision list are efficiently PAC learnable in the SQ model.

Theorem 3.6 ([Kea98])**.** *There exists an SQ learning algorithm that for every $\alpha > 0$, learns C_{DL} over $\{0,1\}^d$ with error α using $O(d)$ queries to STAT$(\alpha/(4d))$.*

It is known that by using the SQ algorithm for learning halfspaces (or decision lists) and simulations of SQ algorithms by ϵ-LDP algorithms (Theorem 2.4) and 1-bit communication bounded algorithms (Theorem 2.6) one obtains efficient PAC learning algorithm for halfspaces in these models. Hence we only state the lower bounds for non-interactive algorithms here that follow from Theorems 2.5 and 2.7.

Corollary 3.7. *Any non-interactive ϵ-LPD algorithm that PAC learns C_{HS} over $\{0,1\}^d$ with error less than $1/2$ and success probability at least $3/4$ using at most m queries to LR_S for S drawn i.i.d. from P^n must have $me^{\epsilon} = 2^{\Omega(d)}$.*

Corollary 3.8. *Any non-interactive ℓ-communication bounded algorithm that PAC learns C_{HS} over $\{0,1\}^d$ with error less than $1/2$ and success probability at least $3/4$ using queries to COMM_S for S drawn i.i.d. from P^n must have $n2^{\ell} = 2^{\Omega(d)}$.*

4 Discussion

Our work shows that polynomial margin complexity is a necessary condition for learning a class of binary classifiers by a non-interactive SQ/LDP/limited-communication algorithm. We first point out a sense in which our condition is also sufficient. Lemma 2.11 implies that for every class C of margin complexity d, there exists a PAC learning algorithm that uses m non-interactive SQ queries to STAT$(1/m)$ and learns C with error of at most $1/2 - 1/m$ for $m = O(d^2 \ln(|C||X|))$ (since the correlated function or its negation will have error of at most $1/2 - 1/m$). Therefore our result implies that margin complexity characterizes (up to a polynomial) the complexity of *weak* PAC learning by non-interactive algorithms in the three models we consider. On the other hand, for PAC learning with small constant error (say $1/4$) all known algorithms for learning large-margin halfspaces require many rounds of interaction [BFKV97; FGV15]. Proving that this is necessary is a natural open problem.

A significant limitation of our result is that it does not rule out even a 2-round algorithm for learning halfspaces (or decision lists). This is, again, in contrast to the fact that learning algorithms for these classes require at least d rounds of interaction. We believe that extending our lower bounds to multiple-round algorithms and quantifying the tradeoff between the number of rounds and the complexity of learning is an important direction for future work.

Acknowledgements

We thank Kobbi Nissim, Adam Smith and Justin Thaler for insightful discussions of this problem.

References

[ABR64] M. A. Aizerman, E. A. Braverman, and L. Rozonoer. “Theoretical foundations of the potential function method in pattern recognition learning.” In: *Automation and Remote Control,* Automation and Remote Control, 25. 1964, pp. 821–837.

[BD98] S. Ben-David and E. Dichterman. “Learning with Restricted Focus of Attention”. In: *J. Comput. Syst. Sci.* 56.3 (1998), pp. 277–298.

[BF02] N. Bshouty and V. Feldman. “On using extended statistical queries to avoid membership queries”. In: *Journal of Machine Learning Research* 2 (2002), pp. 359–395.

[BF15] M. Balcan and V. Feldman. “Statistical Active Learning Algorithms for Noise Tolerance and Differential Privacy”. In: *Algorithmica* 72.1 (2015), pp. 282–315.

[BFKV97] A. Blum, A. Frieze, R. Kannan, and S. Vempala. “A polynomial time algorithm for learning noisy linear threshold functions”. In: *Algorithmica* 22.1/2 (1997), pp. 35–52.

[BGV92] B. E. Boser, I. Guyon, and V. Vapnik. “A Training Algorithm for Optimal Margin Classifiers”. In: *COLT*. ACM, 1992, pp. 144–152.

[BNSSSU16] R. Bassily, K. Nissim, A. D. Smith, T. Steinke, U. Stemmer, and J. Ullman. “Algorithmic stability for adaptive data analysis”. In: *STOC*. 2016, pp. 1046–1059.

[BRS17] E. Balkanski, A. Rubinstein, and Y. Singer. “The limitations of optimization from samples”. In: *STOC*. 2017.

[BS18] E. Balkanski and Y. Singer. “The adaptive complexity of maximizing a submodular function”. In: *STOC*. 2018, pp. 1138–1151.

[BVW07] H. Buhrman, N. Vereshchagin, and R. de Wolf. “On Computation and Communication with Small Bias”. In: *IEEE Conference on Computational Complexity*. 2007, pp. 24–32.

[DFHPRR14] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. “Preserving Statistical Validity in Adaptive Data Analysis”. In: *CoRR* abs/1411.2664 (2014). Extended abstract in STOC 2015.

[DKY17] B. Ding, J. Kulkarni, and S. Yekhanin. “Collecting Telemetry Data Privately”. In: *31st Conference on Neural Information Processing Systems (NIPS)*. 2017, pp. 3574–3583.

[DMNS06] C. Dwork, F. McSherry, K. Nissim, and A. Smith. “Calibrating noise to sensitivity in private data analysis”. In: *TCC*. 2006, pp. 265–284.

[DR14] C. Dwork and A. Roth. *The Algorithmic Foundations of Differential Privacy*. Vol. 9. 3-4. 2014, pp. 211–407.

[DRY18] J. C. Duchi, F. Ruan, and C. Yun. “Minimax Bounds on Stochastic Batched Convex Optimization”. In: *COLT*. 2018, pp. 3065–3162.

[DV04] J. Dunagan and S. Vempala. “A simple polynomial-time rescaling algorithm for solving linear programs”. In: *STOC*. 2004, pp. 315–320.

[EPK14] Ú. Erlingsson, V. Pihur, and A. Korolova. “RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response”. In: *ACM SIGSAC Conference on Computer and Communications Security*. 2014, pp. 1054–1067.

[Fel08] V. Feldman. “Evolvability from Learning Algorithms”. In: *STOC*. 2008, pp. 619–628.

[FGRVX12] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. “Statistical Algorithms and a Lower Bound for Detecting Planted Cliques”. In: *arXiv, CoRR* abs/1201.1214 (2012). Extended abstract in STOC 2013.

[FGV15] V. Feldman, C. Guzman, and S. Vempala. “Statistical Query Algorithms for Mean Vector Estimation and Stochastic Convex Optimization”. In: *CoRR* abs/1512.09170 (2015). Extended abstract in SODA 2017.

[FSS01] J. Forster, N. Schmitt, and H. U. Simon. “Estimating the Optimal Margins of Embeddings in Euclidean Half Spaces”. In: *Proceedings of COLT 2001 and EuroCOLT 2001*. 2001, pp. 402–415.

[GHR92] M. Goldmann, J. Håstad, and A. Razborov. “Majority gates vs. general weighted threshold gates”. In: *Computational Complexity* 2 (1992), pp. 277–300.

[HU14] M. Hardt and J. Ullman. “Preventing False Discovery in Interactive Data Analysis Is Hard”. In: *FOCS*. 2014, pp. 454–463.

[Kan11] V. Kanade. “Evolution with Recombination”. In: *FOCS*. 2011, pp. 837–846.

[Kea98] M. Kearns. “Efficient noise-tolerant Learning from statistical queries”. In: *Journal of the ACM* 45.6 (1998), pp. 983–1006.

[KLNRS11] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. “What Can We Learn Privately?” In: *SIAM J. Comput.* 40.3 (June 2011), pp. 793–826.

[KS11] M. Kallweit and H. Simon. “A Close Look to Margin Complexity and Related Parameters”. In: *COLT*. 2011, pp. 437–456.

[KV94] M. Kearns and U. Vazirani. *An introduction to computational learning theory*. Cambridge, MA: MIT Press, 1994.

[LS09] N. Linial and A. Shraibman. “Learning Complexity vs Communication Complexity”. In: *Comb. Probab. Comput.* 18.1-2 (Mar. 2009), pp. 227–245. ISSN: 0963-5483.

[Luo05] Z.-Q. Luo. “Universal decentralized estimation in a bandwidth constrained sensor network”. In: *IEEE Transactions on information theory* 51.6 (2005), pp. 2210–2219.

[Nov62] A. Novikoff. “On convergence proofs on perceptrons”. In: *Proceedings of the Symposium on Mathematical Theory of Automata*. Vol. XII. 1962, pp. 615–622.

[RG06] A. Ribeiro and G. B. Giannakis. “Bandwidth-constrained distributed estimation for wireless sensor networks-part I: Gaussian case”. In: *IEEE transactions on signal processing* 54.3 (2006), pp. 1131–1143.

[RWV06] R. Rajagopal, M. J. Wainwright, and P. Varaiya. “Universal quantile estimation with feedback in the communication-constrained setting”. In: *Information Theory, 2006 IEEE International Symposium on*. IEEE. 2006, pp. 836–840.

[SD15] J. Steinhardt and J. C. Duchi. “Minimax rates for memory-bounded sparse linear regression”. In: *COLT*. 2015, pp. 1564–1587.

[She08] A. A. Sherstov. “Halfspace Matrices”. In: *Computational Complexity* 17.2 (2008), pp. 149–178.

[STU17] A. D. Smith, A. Thakurta, and J. Upadhyay. “Is Interaction Necessary for Distributed Private Learning?” In: *2017 IEEE Symposium on Security and Privacy, SP 2017*. 2017, pp. 58–77.

[SU15] T. Steinke and J. Ullman. “Interactive Fingerprinting Codes and the Hardness of Preventing False Discovery”. In: *COLT*. 2015, pp. 1588–1628.

[SVW16] J. Steinhardt, G. Valiant, and S. Wager. “Memory, Communication, and Statistical Queries”. In: *COLT*. 2016, pp. 1490–1516.

[SYMK16] A. T. Suresh, F. X. Yu, H. B. McMahan, and S. Kumar. “Distributed mean estimation with limited communication”. In: *arXiv preprint arXiv:1611.00429* (2016).

[Val09] L. G. Valiant. “Evolvability”. In: *Journal of the ACM* 56.1 (2009). Earlier version in ECCC, 2006., pp. 3.1–3.21.

[Val84] L. G. Valiant. “A theory of the learnable”. In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142.

[War65] S. L. Warner. “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias”. In: *J. of the American Statistical Association* 60.309 (1965), pp. 63–69.

[ZDJW13] Y. Zhang, J. C. Duchi, M. I. Jordan, and M. J. Wainwright. “Information-theoretic lower bounds for distributed statistical estimation with communication constraints”. In: *NIPS*. 2013, pp. 2328–2336.

[App17] Apple’s Differential Privacy Team. “Learning with Privacy at Scale”. In: *Apple Machine Learning Journal* 1.9 (Dec. 2017).