Locally Private Learning without Interaction Requires Separation

Amit Daniely
Hebrew University and Google

Vitaly Feldman*
Google Brain

Abstract

We consider learning under the constraint of local differential privacy (LDP). For many learning problems known efficient algorithms in this model require many rounds of communication between the server and the clients holding the data points. Yet multi-round protocols are prohibitively slow in practice due to network latency and, as a result, currently deployed large-scale systems are limited to a single round. Despite significant research interest, very little is known about which learning problems can be solved by such non-interactive systems. The only lower bound we are aware of is for PAC learning an artificial class of functions with respect to a uniform distribution [Kas+11].

We show that the margin complexity of a class of Boolean functions is a lower bound on the complexity of any non-interactive LDP algorithm for distribution-independent PAC learning of the class. In particular, the classes of linear separators and decision lists require exponential number of samples to learn non-interactively even though they can be learned in polynomial time by an interactive LDP algorithm. This gives the first example of a natural problem that is significantly harder to solve without interaction and also resolves an open problem of Kasiviswanathan et al. [Kas+11]. We complement this lower bound with a new efficient learning algorithm whose complexity is polynomial in the margin complexity of the class. Our algorithm is non-interactive on labeled samples but still needs interactive access to unlabeled samples.

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 primary model we consider is that of local differential privacy (LDP) [Kas+11]. 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].

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

---

*Part of this work was done while the author was visiting the Simons Institute for the Theory of Computing.

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 LDP learning algorithms require many rounds of interaction. Yet there are no results that rule out solving these problems without interaction. This problem was first addressed by Kasiviswanathan et al. [Kas+11] who demonstrated existence of an artificial class of Boolean functions C over $\{0, 1\}^d$ with the following property. C can be PAC learned efficiently relative to the uniform distribution over $\{0, 1\}^d$ by an interactive LDP protocol but requires $2^{\Omega(d)}$ samples to learn by any non-interactive learning algorithm. 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).

Deriving a technique that applies to distribution independent learning is posed as a natural open problem in this area [Kas+11]. Even beyond PAC learning, there are no examples of natural problems that provably require exponentially more samples to solve non-interactively.

1.1 Our results

We give a new technique for proving lower bounds on the power of non-interactive LDP algorithms for distribution-independent PAC learning. Our technique is based on a connection between the power of interaction and margin complexity of Boolean function classes that we establish. The margin complexity of a class of Boolean functions C, denoted by $\mathsf{MC}(C)$, is 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 2.5). 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; BES02; She08; LS09; KS11]).

We prove that only classes that have polynomially small *margin complexity* can be efficiently PAC learned by a non-interactive LDP algorithm. Our lower bound implies that two natural and well-studied classes of functions: linear separators and decision lists require an exponential number of samples to learn non-interactively. Importantly, it is known that these classes can be learned efficiently by interactive LDP algorithms (this follows from the results for the statistical query model that we discuss later). Thus our result gives an exponential separation between the power of interactive and non-interactive protocols. To the best of our knowledge this is only known such separation for a natural statistical problem.

Our result follows from a stronger lower bound that also holds against algorithms for which only the queries that depend on the label of the point are non-interactive (also referred to as *non-adaptive* in related contexts). We will refer to such algorithms *label-non-adaptive* LDP algorithms. Formally, our lower bounds for such algorithms is as follows. We say that a class of Boolean ($\{-1, 1\}$-valued) functions C is closed under negation if for every $f \in C$, $-f \in C$.

Theorem 1.1. *Let C be a class of Boolean functions closed under negation. Assume that there exists a label-non-adaptive ϵ-LDP algorithm $\mathcal{A}$ that, with success probability at least $2/3$, PAC learns C distribution-independently with error less than $1/2$ using at most n examples. Then $n = \Omega(\mathsf{MC}(C)^{2/3}/e^{\epsilon})$.*

Our second contribution is an algorithm for learning large-margin linear separators that matches (up to polynomial factors) our lower bound.

Theorem 1.2. *Let C be a class of Boolean functions over X. For any $\alpha, \epsilon > 0$ and $n = \operatorname{poly}\left(\mathsf{MC}(C)/(\alpha\epsilon)\right)$ there is a label-non-adaptive ϵ-LDP algorithm that PAC learns C distribution-independently with accuracy $1 - \alpha$ using at most n examples.*

Learning of large-margin classifiers is a classical learning problem and various algorithms for the problem are widely used in practice. Our learning algorithm is computationally efficient as long as an embedding of C into a $d = \operatorname{poly}\left(\mathsf{MC}(C)\log|X|\right)$-dimensional space can be computed efficiently (such an embedding is known to exists by the Johnson-Lindenstrauss random projection argument [AV99]). Together these results show an equivalence (up to polynomials) between margin complexity and PAC learning with this limited form of interaction in the LDP model.

Another implication of Theorem 1.2 is that if the distribution over X is fixed (and known to the learning algorithm) then the learning algorithm becomes non-interactive.

Corollary 1.3. *Let C be a class of Boolean functions over X and D be an arbitrary distribution over X. For any $\alpha, \epsilon > 0$ and $n = \operatorname{poly}\left(\mathsf{MC}(C)/(\alpha\epsilon)\right)$ there is a non-interactive ϵ-LDP algorithm that PAC learns C relative to D with accuracy $1 - \alpha$ using at most n examples.*

Techniques: Following the approach of Kasiviswanathan et al. [Kas+11], we use the characterization of LDP protocols using the *statistical query* (SQ) 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 1.4. *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 $\phi\colon Z \to [-1, 1]$, 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. Non-adaptive (or non-interactive) SQ algorithms are defined analogously to LDP protocols. The reductions between learning in the SQ model and learning in the LDP model given by Kasiviswanathan et al. [Kas+11] preserve the number of rounds of interaction of a learning algorithm.

The key technical tool we apply to prove our lower bound is a result of Feldman [Fel08] relating margin complexity and a certain notion of complexity for statistical queries. The result shows that the existence of a (possibly randomized) algorithm that outputs a set T of m functions such that for every $f \in C$ and distribution D, with significant probability one of the functions in T is at least $1/m$-correlated with f relative to D implies that $\mathsf{MC}(C) \leq \operatorname{poly}(m)$. We then show that such a set of functions can be easily extracted from the queries of any label-non-adaptive SQ algorithm for learning C.

Our label-non-adaptive LDP learning algorithm for large-margin halfspaces relies on a new formulation of halfspace learning as a stochastic convex optimization problem. The crucial property of this program is that (approximately) computing sub-gradients can be done by using a fixed set of non-adaptive queries (that measure the correlation of each of the attributes with the label) and (adaptive) but label-independent queries. We can then use an arbitrary gradient-descent-based LDP algorithm for stochastic convex optimization. Such algorithms were first described by Duchi et al. [DJW13]. For simplicity, we appeal to the fact that such algorithms can also be implemented in the statistical query model [FGV15].

Corollaries: The class of decision lists (see [Riv87; KV94] for a definition) 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] and thus also by LDP algorithms. Formally, we obtain the following lower bounds:

Corollary 1.5. *Any label-non-adaptive ϵ-LPD algorithm that PAC learns the class of linear separators over $\{0,1\}^d$ with error less than $1/2$ and success probability at least $3/4$ must use $n = 2^{\Omega(d)}/e^\epsilon$ i.i.d. examples. For learning the class of decision lists under the same conditions the algorithm must use $n = 2^{\Omega(d^{1/3})}/e^\epsilon$ i.i.d. examples.*

Our use of the statistical query model to prove the results implies that we can derive the analogues of our results in other models that have connections to the SQ model. One of such models is the distributed model in which only a small number of bits is 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; Zha+13; SD15; Sur+16]). 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. Using the equivalence between this model and SQ learning that preserves the number of rounds of interaction we immediately obtain analogous results for this model. We are not aware of any prior results on the power of interaction in the context of this model.

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. Subsequently, upper and lower bounds on the number of queries to the gradient/second-order oracles for algorithms with few rounds of interaction have been studied by several groups [DRY18; Woo+18; BS18a; DG18]. In the context of discrete optimization from queries for the value of the optimized function the round complexity has been recently investigated in [BRS17; BS18b; BRS19]. To the best of our knowledge, the techniques used in these works are unrelated to ours. Also in all these works the lower bounds rely heavily on the fact that the oracle provides only local information about the optimized function. In contrast, statistical queries allow getting global information about the optimized function.

Acharya et al. [ACT18] implicitly give a separation between interactive and non-interactive protocols for the problem of identity testing for a discrete distribution over k elements, albeit a relatively weak one ($O(k)$ vs $\Omega(k^{3/2})$ samples). A very recent work of Joseph et al. [Jos+19] explores a different aspect of interactivity in LDP. Specifically, they distinguish between two types of interactive protocols: fully-interactive and sequentially-interactive ones. Fully-interactive protocols place no restrictions on interaction whereas sequentially-interactive ones only allows asking one query per user. They give a separation showing that sequentially-interactive protocols may require more samples than fully interactive ones (although only by a polynomial factor). This separation is orthogonal to ours since our lower bounds are against completely non-interactive protocols and we separate them from sequentially interactive protocols. There are also a number of lower bounds on the sample complexity of LDP algorithms that demonstrate that LDP is less efficient than the central model of differential privacy (*e.g.* [DWJ13; DR19]). The techniques in this literature are completely unrelated to ours.

Margin complexity has also played a key role in separation of the power of correlational statistical query algorithms (CSQ) from general SQ algorithms. A statistical query ϕ is *correlational* if $\phi(x, \ell) = \ell \cdot \psi(x)$ for

some function $\psi \colon X \to [-1, 1]$. Such queries allow measuring the correlation between the target function and an arbitrary predictor. CSQ algorithm are known to capture the power of learning algorithm in Valiant's [Val09] model of evolvability [Fel08]. In this context it was shown in [Fel08] that the complexity of *weak* and distribution-independent learning in this model is exactly characterized by margin complexity. Our work relies on some of the properties of margin complexity given in that work. At the same time these are incomparable restrictions on the power of algorithms and our lower bound is incomparable to the lower bound for CSQ algorithms. For example, Boolean conjunctions are (strongly) PAC learnable distribution independently and efficiently by a non-interactive SQ algorithm [Kea98] but not by CSQ algorithms [Fel11]. On the other hand, the function class in [Kas+11] is PAC learnable by a CSQ algorithm relative to the uniform distribution but not by a non-interactive SQ algorithm.

The number of data samples necessary to answer statistical queries chosen adaptively has recently been studied in a line of work on adaptive data analysis [Dwo+14; HU14; Bas+16; SU15]. Our work provably demonstrates that the use of such adaptive queries is important for solving basic learning problems.

2 Preliminaries

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

Local differential privacy: In the local differential privacy (LDP) model [War65; EGS03; Kas+11] 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 ([Kas+11]). *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 $\sum_{j \in [k]} \epsilon_j \leq \epsilon$.*

For a non-interactive LDP algorithm one can assume without loss of generality that each sample is queried only once since the application of k fixed local randomizers with $\sum_{j \in [k]} \epsilon_j \leq \epsilon$ can be seen as an execution of a single ϵ-randomizer. 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 [Dwo+06]. This is a stronger model and an ϵ-LPD algorithm also satisfies ϵ-differential privacy. We note that an alternative way to measure the privacy guarantee of an LDP protocol is by examining the privacy of the entire transcript of the protocol. However this distinction is only relevant for fully-interactive LDP protocols [Jos+19] whereas we will only consider non-interactive or sequentially-interactive ones.

Equivalence to statistical queries: The statistical query model of Kearns [Kea98] is defined by having access to $\mathrm{STAT}_P(\tau)$ oracle, where P is the unknown data distribution. 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.

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.

For an algorithm in LDP and SQ 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. Similarly, we say that the algorithm is *label-non-adaptive* if all the queries that depend on oracle's response are label-independent (the query function depends only on the point).

Kasiviswanathan et al. [Kas+11] 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 2.2 ([Kas+11]). *Let $\mathcal{A}_{SQ}$ be an algorithm that makes at most t queries to $\mathrm{STAT}_P(\tau)$. 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 $\mathrm{STAT}_P(\tau)$) with probability at least $1 - \delta$. Further, if $\mathcal{A}_{SQ}$ is non-interactive then $\mathcal{A}$ is non-interactive.*

Kasiviswanathan et al. [Kas+11] also prove a converse of this theorem.

Theorem 2.3 ([Kas+11]). *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 $\mathrm{STAT}_P(\tau)$ 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.*

PAC learning and margin complexity: Our results are for the standard PAC model of learning [Val84].

Definition 2.4. *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 $\mathbf{Pr}_D[f(x) \neq h(x)] \leq \alpha$ with probability at least $2/3$.*

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

For dimension d, we denote by $\mathcal{B}^d(1)$ the unit ball in ℓ_2 norm in $\mathbb{R}^d$.

Definition 2.5. *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 deterministic algorithms (which corresponds to a single fixed set of functions). It is however easy to see that their sharper bound extends to randomized algorithms with an appropriate adjustment of the bound and we give the resulting statement below:

Lemma 2.6 ([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*

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

The conditions in Lemma 2.6 are also known to be necessary for low margin complexity.

Lemma 2.7 ([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 label-non-adaptive algorithms

We prove the SQ version of our lower bound. Theorem 1.1 then follows immediately by applying the simulation result from Lemma 2.3.

Theorem 3.1. *Let C be a class of Boolean functions closed under negation. Assume that for some m there exists a label-non-adaptive possibly randomized SQ algorithm $\mathcal{A}$ that, with success probability at least $2/3$, PAC learns C distribution-independently with error less than $1/2$ using at most m queries to STAT$(1/m)$. Then $\mathsf{MC}(C) \leq 6m^{3/2}$.*

Proof. We first recall a simple observation from [BF02] that allows to decompose each statistical query into a correlational and label-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 $|\mathbf{E}_D[f(x)h(x)]| \geq 1/m$ 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-adaptive statistical queries asked by $\mathcal{A}(r)$ (where $m' \leq m$).

Let g_i^r and h_i^r denote the decomposition of these queries into correlational and label-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 2.6 with $\beta = 1/3$). To see this, assume for the sake of contradiction that for some distribution D and function $f \in C$,

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

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

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

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

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

By Lemma 2.6, we obtain the claim. □

3.1 Applications

We will now spell out several easy corollaries of our lower bound, simulation results and existing SQ algorithms. Together they imply the claimed separations for halfspaces and decision lists. We start with the class of halfspaces over $\{0, 1\}^d$ which we denote by C_{HS}. The lower bound on the margin complexity of halfspaces is implied by a celebrated work of Goldmann et al. [GHR92] on the complexity of linear threshold circuits (the connection of this result to margin complexity is due to Sherstov [She08]):

Theorem 3.2 ([GHR92; She08]). $\mathsf{MC}(C_{HS}) = 2^{\Omega(d)}$.

We denote the class of decision lists over $\{0, 1\}^d$ by C_{DL} (see [KV94] for a standard definition). A lower bound on the margin complexity decision lists was derived by Buhrman et al. [BVW07] in the context of communication complexity.

Theorem 3.3 ([BVW07]). $\mathsf{MC}(C_{DL}) = 2^{\Omega(d^{1/3})}$.

Combining these results with Theorem 1.1 we obtain the lower bound on complexity of LDP algorithms for learning linear classifiers and decision lists given in Corollary 1.5.

Learnability of decision list using statistical queries is a classical result of Kearns [Kea98]. Applying the simulation in Theorem 2.2 we obtain polynomial time learnability of this class by (interactive) LDP algorithms.

Theorem 3.4 ([Kea98]). *For every $\epsilon, \alpha > 0$, there exists an ϵ-LDP learning algorithm that PAC learns C_{DL} with error α using $\mathrm{poly}(d/(\epsilon\alpha))$ i.i.d. examples (with one query per example).*

In the case of halfspaces, Dunagan and Vempala [DV04] give the first 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]). Applying Theorem 2.2 we obtain learnability of this class by (interactive) LDP algorithms.

Theorem 3.5 ([DV04; BF15]). *For every $\epsilon, \alpha > 0$, there exists an ϵ-LDP learning algorithm that PAC learns C_{HS} with error α using $\mathrm{poly}(d/(\epsilon\alpha))$ i.i.d. examples (with one query per example).*

4 Label-non-adaptive learning algorithm for halfspaces

Our algorithm for learning large-margin halfspaces relies on the formulation of the problem of learning a halfspace as the following convex optimization problem.

Lemma 4.1. *Let P be a distribution on $\mathcal{B}^d(1) \times \{-1, 1\}$. Suppose that there is a vector $w^* \in \mathcal{B}^d(1)$ such that $\mathbf{Pr}_{(x,\ell)\sim P}[\langle w^*, \ell x \rangle \geq \gamma] = 1$. Let $(e_1, \ldots, e_d)$ denote the standard basis of $\mathbb{R}^d$ and let w be a unit vector such that for $\alpha, \beta \in (0, 1)$*

$$F(w) \doteq \mathop{\mathbf{E}}_{(x,\ell)\sim P}\left[\sum_{i=1}^{d} |\langle w + \gamma e_i, x \rangle| - \langle w + \gamma e_i, \ell x \rangle + \sum_{i=1}^{d} |\langle w - \gamma e_i, x \rangle| - \langle w - \gamma e_i, \ell x \rangle\right] \leq \alpha\beta. \quad (2)$$

Then, $F(w^*) = 0$ and

$$\mathop{\mathbf{Pr}}_{P}\left[\langle w, \ell x \rangle \geq -\frac{\beta}{2} + \frac{\gamma^2}{\sqrt{d}}\right] \geq 1 - \alpha.$$

In particular, if $\beta < \frac{2\gamma^2}{\sqrt{d}}$ then $\mathbf{Pr}_P[\langle w, \ell x \rangle > 0] \geq 1 - \alpha$.

Proof. To see the first part, note that with probability 1 over $(x, \ell) \sim P$,

$$\langle w^*, \ell x \rangle \geq \gamma \geq |\langle \gamma e_i, \ell x \rangle|.$$

Therefore $\langle w^* + \gamma e_i, \ell x \rangle \geq 0$ and $\langle w^* - \gamma e_i, \ell x \rangle \geq 0$ and thus $\langle w^* + \gamma e_i, \ell x \rangle = |\langle w^* + \gamma e_i, x \rangle|$ and $\langle w^* - \gamma e_i, \ell x \rangle = |\langle w^* + \gamma e_i, x \rangle|$ making $F(w^*) = 0$.

By Markov's inequality, with probability at least $1 - \alpha$, the sum is less that β. In this event, for all $i \in [d]$,

$$|\langle w + \gamma e_i, x \rangle| - \langle w + \gamma e_i, \ell x \rangle \leq \beta$$

and

$$|\langle w - \gamma e_i, x \rangle| - \langle w - \gamma e_i, \ell x \rangle \leq \beta,$$

which implies that $\langle w + \gamma e_i, \ell x \rangle \geq -\beta/2$ and $\langle w - \gamma e_i, \ell x \rangle \geq -\beta/2$. Furthermore, in this event, there exists i such that $|\langle x, e_i \rangle| \geq \frac{\gamma}{\sqrt{d}}$ (This is true with probability 1 since $\|x\| \geq \langle w^*, \ell x \rangle \geq \gamma$). For this i one of $\langle w + \gamma e_i, \ell x \rangle, \langle w - \gamma e_i, \ell x \rangle$ must be $\geq -\frac{\beta}{2} + \frac{2\gamma^2}{\sqrt{d}}$. Hence,

$$\langle w, \ell x \rangle = \frac{\langle w + \gamma e_i, \ell x \rangle + \langle w - \gamma e_i, \ell x \rangle}{2} \geq -\frac{\beta}{2} + \frac{\gamma^2}{\sqrt{d}}.$$

$\square$

We now describe how to solve the convex optimization problem given in Lemma 4.1. Both the running time and accuracy of queries of our solution depend on the ambient dimension d. This dimension is not necessarily upper-bounded by a polynomial in $\mathsf{MC}(C)$. However, the well-known random projection argument shows that the dimension can be reduced to $O(\log(1/\delta)/\mathsf{MC}(C)^2)$ at the expense of small multiplicative decrease in the margin and probability of at most δ of failure (for every individual point over the randomness of the random projection) [AV99; BES02]. This fact together with Markov's inequality implies the following standard lemma:

Lemma 4.2. *Let d be an arbitrary dimension. For every δ and γ, there exists a distribution Ψ over mappings $\psi\colon \mathcal{B}^d(1) \to \mathcal{B}^{d'}(1)$, where $d' = O(\log(1/\delta)/\gamma^2)$ such that: For every distribution D and function f over $\mathcal{B}^d(1)$, if there exists $w \in \mathcal{B}^d(1)$ such that $\mathbf{Pr}_{x \sim D}[f(x) \cdot \langle w, x \rangle \geq \gamma] = 1$ then*

$$\mathop{\mathbf{Pr}}_{\psi \sim \Psi}\left[\exists w', \ \mathop{\mathbf{Pr}}_{x \sim D}\left[f(x) \cdot \langle w', \psi(x) \rangle \geq \frac{\gamma}{2}\right] \geq 1 - \delta\right] \geq 1 - \delta.$$

Lemma 4.2 ensures that at most a tiny fraction δ of the points (according to D) does not satisfy the margin condition. This is not an issue as we will be implementing our algorithm in the SQ model that, by definition allows any of the answers to its queries to be imprecise. The lemma also allows for tiny probability that the mapping will fail altogether (making the overall algorithm randomized).

Therefore the only ingredient missing for establishing Theorem 1.2 is a label-non-adaptive SQ algorithm for solving the convex optimization algorithm in dimension d using a polynomial in d, $1/\gamma$ and $1/\alpha$ number of queries and (the inverse of) tolerance:

Lemma 4.3. *Let P be a distribution on $\mathcal{B}^d(1) \times \{-1, 1\}$. Suppose that there is a vector $w^* \in \mathcal{B}^d(1)$ such that $\mathbf{Pr}_{(x,\ell) \sim P}[\langle w^*, \ell x \rangle \geq \gamma] = 1$. There is a label-non-adaptive SQ algorithm that for every $\alpha \in (0, 1)$ uses $O(d^4/(\gamma^4 \alpha^2))$ queries to $\mathrm{STAT}_P(\Omega(\gamma^4 \alpha^2/d^3))$, and finds a vector w such that $\mathbf{Pr}_P[\langle w, \ell x \rangle > 0] \geq 1 - \alpha$.*

Proof. By Lemma 4.1, it is enough to find a vector w with $F(w) \leq \alpha\beta$ for $\beta = \frac{2\gamma^2}{\sqrt{d}}$. We next explain how to find such a vector. To this end, consider the stochastic convex program of minimizing $F(w)$ subject to the constraint that $\|w\| \leq 1$. Since $F(w)$ is $4d$-Lipschitz over $\mathcal{B}^d(1)$, and $F(w^*) = 0$, a solution with $F(w) \leq \alpha\beta$ can be found using projected (sub-)gradient descent using $O(d^3/(\alpha\beta)^2)$ queries to $\mathrm{STAT}_P(\Omega((\alpha\beta/d)^2))$ [FGV15]. It remains to verify that the sub-gradient computations for this algorithm can be done using label-non-adaptive statistical queries. We decompose $F(w)$ into two parts $F(w) = F_1(w) + F_2(w)$ where,

$$F_1(w) \doteq \mathop{\mathbf{E}}_{(x,\ell) \sim P}\left[\sum_{i=1}^{d}|\langle w + \gamma e_i, x \rangle| + \sum_{i=1}^{d}|\langle w - \gamma e_i, x \rangle|\right]$$

and

$$F_2(w) \doteq -\mathop{\mathbf{E}}_{(x,\ell) \sim P}\left[\sum_{i=1}^{d}\langle w + \gamma e_i, \ell x \rangle + \sum_{i=1}^{d}\langle w - \gamma e_i, \ell x \rangle\right].$$

Now the sub-gradient of F_1 is

$$\nabla F_1(w) = \mathop{\mathbf{E}}_{(x,\ell) \sim P}\left[\sum_{i=1}^{d}(\texttt{sign}(\langle w + \gamma e_i, x \rangle) + \texttt{sign}(\langle w - \gamma e_i, x \rangle))x\right],$$

and is just a function of x. Hence, computing an estimate requires of sub-gradient of F_1 requires only label-independent SQs.

The gradient of (the linear) $F_2(w)$ is

$$\nabla F_2(w) = -2d \mathop{\mathbf{E}}_{(x,\ell)\sim P}[\ell x].$$

Crucially, it does not depend on w and hence can be computed using d non-adaptive statistical queries once. $\qquad\square$

5   Implications for distributed learning with communication constraints

In this section we briefly define the model of bounded communication per sample, state the known equivalence results to the SQ model and spell out the immediate corollary of our lower bound.

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 5.1. *We say that an algorithm $R: Z \rightarrow \{0,1\}^{\ell}$ 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 $\sum_{j\in[k]} \ell_j \leq \ell$.*

We use (non-)adaptive in the same sense as we do for LDP.

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 5.2 ([BD98]). *Let $\mathcal{A}_{SQ}$ be an algorithm that makes at most t queries to $\mathrm{STAT}_P(\tau)$. 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 $\mathrm{STAT}_P(\tau)$) 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; Fel+12]. For the stronger oracle in Definition 5.1, the converse was given by Steinhardt et al. [SVW16].

Theorem 5.3 ([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 $\mathrm{STAT}_P\left(\delta/(2^{\ell+1}n)\right)$ 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.

A direct corollary of Theorems 3.1 and 5.3 and is the following lower bound:

Corollary 5.4. *Let C be a class of Boolean functions closed under negation. Any label-non-adaptive ℓ-communication bounded algorithm that PAC learns C 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 $n = \mathsf{MC}(C)^{2/3}/2^{\ell}$.*

Our other results can be extended analogously.

6 Discussion

Our work shows that polynomial margin complexity is a necessary and sufficient condition for efficient distribution-independent PAC learning of a class of binary classifiers by a label-non-adaptive SQ/LDP/limited-communication algorithm. A natural open problem that is left open is whether there exists an efficient and fully non-interactive algorithm for any class of polynomial margin complexity. We conjecture that the answer is "no" and in this case the question is how to characterize the problems that are learnable by non-interactive algorithms. We remark that for weak learning the complexity is again characterized by the margin complexity: our lower bound rules out weak learning and Lemma 2.7 immediately gives a weak non-interactive SQ learner [Fel08].

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.

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.

[ACT18]   J. Acharya, C. L. Canonne, and H. Tyagi. "Inference under Information Constraints I: Lower Bounds from Chi-Square Contraction". In: *arXiv preprint arXiv:1812.11476* (2018).

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

[AV99]   R. Arriaga and S. Vempala. "An Algorithmic Theory of Learning: Robust Concepts and Random Projection". In: *Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS)*. New York, NY, 1999, pp. 616–623.

[Bas+16]   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.

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

[BES02]   S. Ben-David, N. Eiron, and H. U. Simon. "Limitations of Learning Via Embeddings in Euclidean Half Spaces". In: *Journal of Machine Learning Research* 3 (2002), pp. 441–461.

[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.

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

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

[BRS19]   E. Balkanski, A. Rubinstein, and Y. Singer. "An Exponential Speedup in Parallel Running Time for Submodular Maximization without Loss in Approximation". In: *SODA*. 2019, pp. 283–302.

[BS18a]   E. Balkanski and Y. Singer. "Parallelization does not Accelerate Convex Optimization: Adaptivity Lower Bounds for Non-smooth Convex Minimization". In: *CoRR* abs/1808.03880 (2018). arXiv: `1808.03880`.

[BS18b]   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.

[DG18]   J. Diakonikolas and C. Guzmán. "Lower Bounds for Parallel and Randomized Convex Optimization". In: *CoRR* abs/1811.01903 (2018). arXiv: `1811.01903`.

[DJW13]   J. C. Duchi, M. I. Jordan, and M. J. Wainwright. "Local Privacy and Statistical Minimax Rates". In: *FOCS*. 2013, pp. 429–438.

[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.

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

[DR19]   J. Duchi and R. Rogers. "Lower Bounds for Locally Private Estimation via Communication Complexity". In: *arXiv preprint arXiv:1902.00582* (2019).

[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.

[DWJ13]   J. C. Duchi, M. J. Wainwright, and M. I. Jordan. "Local Privacy and Minimax Bounds: Sharp Rates for Probability Estimation". In: *NIPS*. 2013, pp. 1529–1537.

[Dwo+06]   C. Dwork, F. McSherry, K. Nissim, and A. Smith. "Calibrating noise to sensitivity in private data analysis". In: *TCC*. 2006, pp. 265–284.

[Dwo+14]   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.

[EGS03]   A. V. Evfimievski, J. Gehrke, and R. Srikant. "Limiting privacy breaches in privacy preserving data mining". In: *PODS*. 2003, pp. 211–222.

[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.

[Fel11]   V. Feldman. "Distribution-Independent Evolvability of Linear Threshold Functions". In: *CoRR* abs/1103.4904 (2011).

[Fel+12]   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.

[Jos+19]   M. Joseph, J. Mao, S. Neel, and A. Roth. "The Role of Interactivity in Local Differential Privacy". In: *CoRR* abs/1904.03564 (2019). arXiv: `1904.03564`.

[Kas+11]   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.

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

[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.

[Riv87]   R. Rivest. "Learning decision lists". In: *Machine Learning* 2.3 (1987), pp. 229–246.

[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.

[Sur+16]   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).

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

[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.

[Woo+18]   B. E. Woodworth, J. Wang, A. D. Smith, B. McMahan, and N. Srebro. "Graph Oracle Models, Lower Bounds, and Gaps for Parallel Stochastic Optimization". In: *NeurIPS*. 2018, pp. 8505–8515.

[Zha+13]   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.