JMLR: Workshop and Conference Proceedings vol 35:1–20, 2014

Sample Complexity Bounds on Differentially Private Learning via Communication Complexity

Vitaly Feldman VITALY@POST.HARVARD.EDU
IBM Research - Almaden. *

David Xiao DXIAO@LIAFA.UNIV-PARIS-DIDEROT.FR
CNRS, Université Paris 7 †

Abstract

In this work we analyze the sample complexity of classification by differentially private algorithms. Differential privacy is a strong and well-studied notion of privacy introduced by Dwork et al. (2006) that ensures that the output of an algorithm leaks little information about the data point provided by any of the participating individuals. Sample complexity of private PAC and agnostic learning was studied in a number of prior works starting with (Kasiviswanathan et al., 2011) but a number of basic questions still remain open (Beimel et al., 2010; Chaudhuri and Hsu, 2011; Beimel et al., 2013a,b).

Our main contribution is an equivalence between the sample complexity of differentially-private learning of a concept class C (or $\mathrm{SCDP}(C)$) and the randomized one-way communication complexity of the evaluation problem for concepts from C. Using this equivalence we prove the following bounds:

- $\mathrm{SCDP}(C) = \Omega(\mathrm{LDim}(C))$, where $\mathrm{LDim}(C)$ is the Littlestone's dimension characterizing the number of mistakes in the online-mistake-bound learning model (Littlestone, 1987). This result implies that $\mathrm{SCDP}(C)$ is different from the VC-dimension of C, resolving one of the main open questions from prior work.
- For any t, there exists a class C such that $\mathrm{LDim}(C) = 2$ but $\mathrm{SCDP}(C) \geq t$.
- For any t, there exists a class C such that the sample complexity of (pure) α-differentially private PAC learning is $\Omega(t/\alpha)$ but the sample complexity of the relaxed (α, β)-differentially private PAC learning is $O(\log(1/\beta)/\alpha)$. This resolves an open problem from (Beimel et al., 2013b).

We also obtain simpler proofs for a number of known related results. Our equivalence builds on a characterization of sample complexity by Beimel et al. (2013a) and our bounds rely on a number of known results from communication complexity.

1. Introduction

In learning tasks, the training data often consists of information collected from individuals. This data can be highly sensitive, for example in the case of medical or financial information, and therefore privacy-preserving data analysis is becoming an increasingly important area of study in machine learning, data mining and statistics (Dwork and Smith, 2009; Sarwate and Chaudhuri, 2013; Dwork and Roth, 2014).

† Part of this work done while visiting LIAFA, Université Paris 7.
† Part of this work done while visiting Harvard's Center for Research on Computation and Society (CRCS).

In this work we consider learning in PAC (Valiant, 1984) and agnostic (Haussler, 1992; Kearns et al., 1994) learning models by differentially-private algorithms. Differential privacy gives a formal semantic guarantee of privacy, saying intuitively that no single individual's data has too large of an effect on the output of the algorithm, and therefore observing the output of the algorithm does not leak much information about an individual's private data (Dwork et al., 2006) (see Section 2 for the formal definitions). The downside of this desirable guarantee is that for some problems achieving it has an additional cost: both in terms of the amount of data, or sample complexity, and computation.

The cost of differential privacy in PAC and agnostic learning was first studied by Kasiviswanathan et al. (2011). They showed that the sample complexity[1] of differentially privately learning a concept class C over domain X, denoted by $\mathrm{SCDP}(C)$, is $O(\log|C|)$ and left open the natural question of whether $\mathrm{SCDP}(C) = O(\mathrm{VC}(C))$. Note that the gap between these two measures can be as large as (and no more than) $\log(|X|)$.

Subsequently, Beimel et al. (2010) showed that there exists a large concept class, specifically single points, for which the sample complexity is a constant. They also show that differentially-private *proper* learning (the output hypothesis has to be from C) of single points Point_b and threshold functions Thr_b on the set $I_b = \{0, 1, \dots, 2^b - 1\}$ requires $\Omega(b)$ samples. These results demonstrate that the sample complexity can be lower than $O(\log(|C|))$ and also that lower bounds on the sample complexity of proper learning do not necessarily apply to non-proper learning that we consider here. A similar lower bound on proper learning of thresholds on an interval was given by Chaudhuri and Hsu (2011) in a continuous setting where the sample complexity becomes infinite. They also showed that the sample complexity can be reduced to essentially $\mathrm{VC}(C)$ by either adding distributional assumptions or by requiring only the privacy of the labels.

The upper bound of Beimel et al. (2010) is based on an observation from (Kasiviswanathan et al., 2011) that if there exists a class of functions H such that for every $f \in C$ and every distribution $\mathcal{D}$ over the domain, there exists $h \in H$ such that $\Pr_{x\sim\mathcal{D}}[f(x) \neq h(x)] \leq \varepsilon$ then the sample complexity of differentially-private PAC learning with error 2ε can be reduced to $O(\log(|H|)/\varepsilon)$. They refer to such H as an ε-representation of C, and define the (deterministic) ε-representation dimension of C, denoted as $\mathrm{DRDim}_\varepsilon(C)$, as $\log(|H|)$ for the smallest H that ε-represents C. We note that this natural notion can be seen as a distribution-independent version of the usual ε-covering of C in which the distribution over the domain is fixed (*e.g.* Benedek and Itai, 1991).

Subsequently, Beimel et al. (2013a) defined a probabilistic relaxation of ε-representation defined as follows. A distribution $\mathcal{H}$ over sets of boolean functions on X is said to (ε, δ)-probabilistically represent C if for every $f \in C$ and distribution $\mathcal{D}$ over X, with probability $1 - \delta$ over the choice of $H \overset{R}{\leftarrow} \mathcal{H}$, there exists $h \in H$ such that $\Pr_{x\sim\mathcal{D}}[h(x) \neq f(x)] \leq \varepsilon$. The (ε, δ)-probabilistic representation dimension $\mathrm{PRDim}_{\varepsilon,\delta}(C)$ is the minimal $\max_{H\in\mathrm{supp}(\mathcal{H})} \log|H|$, where the minimum is over all $\mathcal{H}$ that (ε, δ)-probabilistically represent C. Rather surprisingly[2], Beimel et al. (2013a) proved that $\mathrm{PRDim}_{\varepsilon,\delta}(C)$ characterizes the sample complexity of differentially-private PAC learning. In addition, they show that PRDim can be upper-bounded by the simpler DRDim as $\mathrm{PRDim}(C) = O(\mathrm{DRDim}(C) + \log\log(|X|))$, where we omit ε and δ when they are equal to $1/4$.

Beimel et al. (2013b) consider PAC learning with a more relaxed (α, β)-differential privacy where the privacy guarantee holds with probability $1 - \beta$. They show that Thr_b can be PAC learned

1. For now we ignore the dependence on other parameters and consider them to be small constants.
2. While many other sample complexity bounds in learning theory rely on covering numbers their lower bound does not involve the standard step of constructing a large packing implied by covering. It is unclear to us if a packing implies a covering of the same size in this distribution-independent setting (as it does in the case of metric covering).

using $O(16^{\log^*(b)} \cdot \log(1/\beta))$ samples (α is a constant as before). Their algorithm is proper so this separates the sample complexity of (pure) differentially-private proper PAC learning from the relaxed version. This work leaves open the question of whether such a separation can be proved for non-proper PAC learning.

1.1. Our results

In this paper we resolve the open problems described above. In the process we also establish a new relation between SCDP and Littlestone's dimension, a well-studied measure of complexity of online learning (Littlestone, 1987) (see Section 2.4 for the definition). The main ingredient of our work is a characterization of DRDim and PRDim in terms of randomized one-way communication complexity of associated evaluation problems (Kremer et al., 1999). In such a problem Alice is given as input a function $f \in C$ and Bob is given an input $x \in X$. Alice sends a single message to Bob, and Bob's goal is to compute $f(x)$. The question is how many bits Alice must communicate to Bob in order for Bob to be able to compute $f(x)$ correctly, with probability at least $2/3$ over the randomness used by Alice and Bob.

In the standard or "private-coin" version of this model, Alice and Bob each have their own source of random coins. The minimal number of bits needed to solve the problem for all $f \in C$ and $x \in X$ is denoted by $\mathrm{R}^{\rightarrow}(C)$. In the stronger "public coin" version of the model, Alice and Bob share the access to the same source of random coins. The minimal number of bits needed to evaluate C (with probability at least $2/3$) in this setting is denoted by $\mathrm{R}^{\rightarrow,\mathsf{pub}}(C)$. See Section 2.3 for formal definitions.

We show that these communication problems are equivalent to deterministic and probabilistic representation dimensions of C and, in particular, $\mathrm{SCDP}(C) = \theta(\mathrm{R}^{\rightarrow,\mathsf{pub}}(C))$ (for clarity we omit the accuracy and confidence parameters, see Theorem 7 and Theorem 8 for details).

Theorem 1 $\mathrm{DRDim}(C) = \Theta(\mathrm{R}^{\rightarrow}(C))$ *and* $\mathrm{PRDim}(C) = \Theta(\mathrm{R}^{\rightarrow,\mathsf{pub}}(C))$.

The evaluation of threshold functions on a (discretized) interval I_b corresponds to the well-studied "greater than" function in communication complexity denoted as GT. $\mathsf{GT}_b(x, y) = 1$ if and only if $x > y$, where $x, y \in \{0,1\}^b$ are viewed as binary representations of integers. It is known that $\mathrm{R}^{\rightarrow,\mathsf{pub}}(\mathsf{GT}_b) = \Omega(b)$ (Miltersen et al., 1998). By combining this lower bound with Theorem 1 we obtain a class whose VC dimension is 1 yet it requires at least $\Omega(b)$ samples to PAC learn differentially-privately.

This equivalence also shows that some of the known results in (Beimel et al., 2010, 2013a) are implied by well-known results from communication complexity, sometimes also giving simpler proofs. For example (1) the constant upper bound on the sample complexity of single points follows from the communication complexity of the equality function and (2) the bound $\mathrm{PRDim}(C) = O(\mathrm{DRDim}(C) + \log\log(|X|))$ follows from the classical result of Newman (1991) on the relationship between the public and private coin models. See Section 3.1 for more details and additional examples.

Our second contribution is a relationship of $\mathrm{SCDP}(C)$ (via the equivalences with $\mathrm{R}^{\rightarrow,\mathsf{pub}}(C)$) to Littlestone's (1987) dimension of C. Specifically, we prove

Theorem 2

1. $\mathrm{R}^{\rightarrow,\mathsf{pub}}(C) = \Omega(\mathrm{LDim}(C))$.

2. *For any t, there exists a class C such that* $\mathrm{LDim}(C) = 2$ *but* $\mathrm{R}^{\rightarrow,\mathsf{pub}}(C) \geq t$.

The first result follows from a natural reduction to the augmented index problem, which is well-studied in communication complexity (Bar-Yossef et al., 2004). While new in our context, the relationship of Littlestone's dimension to quantum communication complexity was shown by Zhang (2011). Together with numerous known bounds on LDim (*e.g.* Littlestone, 1987; Maass and Turán, 1994), our result immediately yields a number of new lower bounds on SCDP. In particular, results of Maass and Turán (1994) imply that linear threshold functions over I_b^d require $\Omega(d^2 \cdot b)$ samples to learn differentially privately. This implies that differentially private learners need to pay an additional dimension d factor as well as a bit complexity of point representation b factor over non-private learners. To the best of our knowledge such strong separation was not known before for problems defined over i.i.d. samples from a distribution (as opposed to worst case inputs). Note that this lower bound is also almost tight since $\log |\mathsf{HS}_b^d| = O(d^2(\log d + b))$ (*e.g.* Muroga, 1971).

In the second result of Thm. 2 we use the class Line_p of lines in $\mathbb{Z}_p^2$ (a plane over a finite field $\mathbb{Z}_p$). A lower bound on the one-way quantum communication complexity of this class was first given by Aaronson (2004) using his trace distance based method.

Finally, we consider PAC learning with (α, β)-differential privacy. Our lower bound of $\Omega(b)$ on SCDP of thresholds together with the upper bound of $O(16^{\log^*(b)} \cdot \log(1/\beta))$ from (Beimel et al., 2013b) immediately imply a separation between the sample complexities of pure and approximate differential privacy. We show a stronger separation for the concept class Line_p:

Theorem 3 *The sample complexity of (α, β)-differentially-privately learning Line_p is $O(\frac{1}{\alpha}\log(1/\beta))$.*

Our upper bound is also substantially simpler. See Section 5 for details.

Some of the proofs and related discussions are omitted in this version due to space constraints. The reader is referred to the full version for a more detailed presentation (Feldman and Xiao, 2014).

1.2. Related work

There is now an extensive amount of literature on differential privacy in machine learning and related areas which we cannot hope to cover here. The reader is referred to the excellent surveys in (Sarwate and Chaudhuri, 2013; Dwork and Roth, 2014).

Blum et al. (2005) showed that algorithms that can be implemented in the statistical query (SQ) framework of Kearns (1998) can also be easily converted to differentially-private algorithms. This result implies polynomial upper bounds on the sample (and computational) complexity of all learning problems that can be solved using statistical queries (which includes the vast majority of problems known to be solvable efficiently). Formal treatment of differentially-private PAC and agnostic learning was initiated in the seminal work of Kasiviswanathan et al. (2011). Aside from the results we already mentioned, they separated SQ learning from differentially private learning. Further, they showed that SQ learning is (up to polynomial factors) equivalent to *local* differential privacy a more stringent model in which each data point is privatized before reaching the learning algorithm.

The results of this paper are for the distribution-independent learning, where the learner does not know the distribution over the domain. Another commonly-considered setting is distribution-specific learning in which the learner only needs to succeed with respect to a single fixed distribution $\mathcal{D}$ known to the learner. Differentially-private learning in this setting and its relaxation in which the learner only knows a distribution close to $\mathcal{D}$ were studied by Chaudhuri and Hsu (2011).

$\mathrm{DRDim}_\varepsilon(C)$ restricted to a fixed distribution $\mathcal{D}$ is denoted by $\mathrm{DRDim}^{\mathcal{D}}_\varepsilon(C)$ and equals to the logarithm of the smallest ε-cover of C with respect to the disagreement metric given by $\mathcal{D}$ (also referred to as the *metric entropy*). The standard duality between packing and covering numbers also implies that $\mathrm{PRDim}^{\mathcal{D}}_{\frac{\varepsilon}{2},\delta}(C) \geq \mathrm{DRDim}^{\mathcal{D}}_\varepsilon(C) - \log(\frac{1}{1-\delta})$, and therefore these notions are essentially identical. It also follows from the prior work (Kasiviswanathan et al., 2011; Chaudhuri and Hsu, 2011), that $\mathrm{DRDim}^{\mathcal{D}}_\varepsilon(C)$ characterizes the complexity of differentially-private PAC and agnostic learning up to the dependence on the error parameter ε in the same way as it does for (non-private) learning (Benedek and Itai, 1991). Namely, $\Omega(\mathrm{DRDim}^{\mathcal{D}}_{2\varepsilon}(C)/\alpha)$ samples are necessary to learn α-differentially-privately with error ε and $O(\mathrm{DRDim}^{\mathcal{D}}_{\varepsilon/2}(C)/(\varepsilon\alpha))$ samples suffice for α-differentially private PAC learning (and even if only weaker label differentially-privacy is desired (Chaudhuri and Hsu, 2011)). This implies that in this setting there are no dimension or bit-complexity costs incurred by differentially-private learners. Chaudhuri and Hsu (2011) also show that doubling dimension at an appropriate scale can be used to give upper and lower bounds on sample complexity of distribution-specific private PAC learning that match up to logarithmic factors.

In a related problem of sanitization of queries from the concept class C the input is a database D of points in X and the goal is to output differentially-privately a "synthetic" database $\hat{D}$ such that for every $f \in C$, $\left|\frac{1}{|D|}\sum_{x\in D} f(x) - \frac{1}{|\hat{D}|}\sum_{x\in\hat{D}} f(x)\right| \leq \varepsilon$. This problem was first considered by Blum et al. (2013) who showed an upper bound of $O(\mathrm{VC}(C)\cdot\log(|X|)$ on the size of the database sufficient for this problem and also showed a lower bound of $\Omega(b)$ on the number of samples required for solving this problem when $X = I_b$ for $C = \mathsf{Thr}_b$. It is easy to see that from the point of view of sample complexity this problem is at least as hard as (differentially-private) *proper* agnostic learning of C (*e.g.* Gupta et al., 2011). Therefore lower bounds on proper learning such as those in (Beimel et al., 2010) and (Chaudhuri and Hsu, 2011) apply to this problem and can be much larger than SCDP that we study. That said, to the best of our knowledge, the lower bound for linear threshold functions that we give was not known even for this harder problem. Aside from sample complexity this problem is also computationally intractable for many interesting classes C (see (Ullman, 2013) and references therein for recent progress).

Sample complexity of more general problems in statistics was investigated in several works starting with Dwork and Lei (2009) (measured alternatively via convergence rates of statistical estimators) (Smith, 2011; Chaudhuri and Hsu, 2012; Duchi et al., 2013a,b). A recent work of Duchi et al. (2013a) shows a number of d-dimensional problems where differentially-private algorithms must incur an additional factor d/α^2 cost in sample complexity. However their lower bounds apply only to a substantially more stringent local and non-interactive model of differential privacy.

Differentially-private communication protocols were studied by McGregor et al. (2010) who showed that differential-privacy can be exploited to obtain a low-communication protocol and vice versa. Conceptually this is similar to the equivalence we establish but our contribution is mostly orthogonal to (McGregor et al., 2010) since the main step in our work is going from a learning setting to a communication-protocol.

2. Preliminaries

We defer some standard definitions and preliminaries to Section A.

2.1. Differentially Private Learning

Two sample sets $S = \{(x_i, \ell_i)\}_{i \in [n]}, S' = \{(x'_i, \ell'_i)\}_{i \in [n]}$ are said to be *neighboring* if there exists $i \in [n]$ such that $(x_i, \ell_i) \neq (x'_i, \ell'_i)$, and for all $j \neq i$ it holds that $(x_j, \ell_j) = (x'_j, \ell'_j)$. For $\alpha, \beta > 0$, an algorithm A is (α, β)-differentially private if for all neighboring $S, S' \in (X \times \{0,1\})^n$ and for all $T \subseteq \mathsf{Range}(A)$:

$$\Pr[A(S) \in T] \leq e^{\alpha} \Pr[A(S') \in T] + \beta,$$

where the probability is over the randomness of A (Dwork et al., 2006). When A is $(\alpha, 0)$-differentially private we say that it satisfies *pure* differential privacy, which we also write as α-differential privacy.

Intuitively, each sample (x_i, ℓ_i) used by a learning algorithm is the record of one individual, and the privacy definition guarantees that by changing one record the output distribution of the learner does not change by much. We remark that, in contrast to the accuracy of learning requirement, the differential privacy requirement holds *in the worst case* for all neighboring sets of examples S, S', not just those sampled i.i.d. from some distribution. We refer the reader to the literature for a further justification of this notion of privacy (Dwork et al., 2006; Dwork, 2006).

The *sample complexity* $\mathrm{SCDP}_{\alpha,\varepsilon,\delta}(C)$ is the minimal n such that it is information-theoretically possible to (ε, δ)-accurately and α-differentially-privately PAC learn C with n examples. SCDP without subscripts refers to $\mathrm{SCDP}_{\frac{1}{4},\frac{1}{4},\frac{1}{4}}$.

2.2. Representation Dimension

Definition 4 (Beimel et al., 2010) *The deterministic representation dimension of C, denoted as $\mathrm{DRDim}_{\varepsilon}(C)$ equals $\log(|H|)$ for the smallest H that ε-represents C. We also let $\mathrm{DRDim}(C) = \mathrm{DRDim}_{\frac{1}{4}}(C)$.*

Definition 5 (Beimel et al., 2013a) *The (ε, δ)-probabilistic representation dimension $\mathrm{PRDim}_{\varepsilon,\delta}(C)$ equals the minimal value of $\max_{H \in \mathsf{supp}(\mathcal{H})} \log |H|$, where the minimum is over all $\mathcal{H}$ that (ε, δ)-probabilistically represent C. We also let $\mathrm{PRDim}(C) = \mathrm{PRDim}_{\frac{1}{4},\frac{1}{4}}(C)$.*

Beimel et al. (2013a) proved the following characterization of SCDP by PRDim.

Theorem 6 (Kasiviswanathan et al., 2011; Beimel et al., 2013a)

$$\mathrm{SCDP}_{\alpha,\varepsilon,\delta}(C) = O\left(\frac{1}{\alpha\varepsilon}\left(\log(1/\varepsilon) \cdot \left(\mathrm{PRDim}_{\frac{1}{4},\frac{1}{4}}(C) + \log\log\frac{1}{\varepsilon\delta}\right) + \log\frac{1}{\delta}\right)\right)$$

$$\mathrm{SCDP}_{\alpha,\varepsilon,\delta}(C) = \Omega\left(\frac{1}{\alpha\varepsilon}\mathrm{PRDim}_{1/4,1/4}(C)\right)$$

For agnostic learning we have that sample complexity is at most

$$O\left(\left(\frac{1}{\alpha\varepsilon} + \frac{1}{\varepsilon^2}\right)\left(\log(1/\varepsilon) \cdot \left(\mathrm{PRDim}_{\frac{1}{4},\frac{1}{4}}(C) + \log\log\frac{1}{\varepsilon\delta}\right) + \log\frac{1}{\delta}\right)\right).$$

This form of upper bounds combines accuracy and confidence boosting from (Beimel et al., 2013a) to first obtain (ε, δ)-probabilistic representation and then the use of exponential mechanism as in (Kasiviswanathan et al., 2011). The results in (Kasiviswanathan et al., 2011) also show the extension of this bound to agnostic learning. Note that the characterization for PAC learning is tight up to logarithmic factors.

2.3. Communication Complexity

Let X and Y be some sets. A private-coin one-way protocol $\pi(x, y)$ from Alice who holds $x \in X$ to Bob who holds $y \in Y$ is given by Alice's randomized algorithm producing a communication σ and Bob's randomized algorithm which outputs a boolean value. We describe Alice's algorithm by a function $\pi_A(x; r_A)$ of the input x and random bits and Bob's algorithm $\pi_B(\sigma, y; r_B)$ by a function of input y, communication σ and random bits. (These algorithms need not be efficient.) The (randomized) output of the protocol on input (x, y) is the value of $\pi(x, y; r_A, r_B) \triangleq \pi_B(\pi_A(x; r_A), y; r_B)$ on a randomly and uniformly chosen r_A and r_B. The cost of the protocol $\mathrm{CC}(\pi)$ is given by the maximum $|\sigma|$ over all $x \in X, y \in Y$ and all possible random coins.

A public-coin one-way protocol $\pi(x, y)$ is given by a randomized Alice's algorithm described by a function $\pi_A(x; r)$ and a randomized Bob's algorithm described by a function $\pi_B(\sigma, x; r)$. The (randomized) output of the protocol on input (x, y) is the value of $\pi(x, y; r) \triangleq \pi_B(\pi_A(x; r), y; r)$ on a randomly and uniformly chosen r. The cost of the protocol $\mathrm{CC}(\pi)$ is defined as in the private-coin case.

Let $\Pi_\varepsilon^{\rightarrow}(g)$ denote the class of all private-coin one-way protocols π computing g with error ε, namely private-coin one-way protocols π satisfying for all $x \in X, y \in Y$

$$\Pr_{r_A, r_B}[\pi(x, y; r_A, r_B) = g(x, y)] \geq 1 - \varepsilon$$

Define $\Pi_\varepsilon^{\rightarrow, \mathsf{pub}}(g)$ similarly as the class of all public-coin one-way protocols π computing g and $\mathrm{R}_\varepsilon^{\rightarrow}(g) = \min_{\pi \in \Pi_\varepsilon^{\rightarrow}(g)} \mathrm{CC}(\pi)$ and $\mathrm{R}_\varepsilon^{\rightarrow, \mathsf{pub}}(g)$.

A deterministic one-way protocol π and its cost are defined as above but without dependence on random bits. We will also require distributional notions of complexity, where there is a fixed input distribution from which x, y are drawn. We define $\Pi_\varepsilon^{\rightarrow}(g; \mu)$ to be all deterministic one-way protocols π such that

$$\Pr_{(x,y) \xleftarrow{R} \mu}[\pi(x, y) = g(x, y)] \geq 1 - \varepsilon$$

Define $\mathrm{D}_\varepsilon^{\rightarrow}(g; \mu) = \min_{\pi \in \Pi_\varepsilon^{\rightarrow}(g; \mu)} \mathrm{CC}(\pi)$. A standard averaging argument shows that the quantity $\mathrm{D}_\varepsilon^{\rightarrow}(g; \mu)$ remains unchanged even if we took the minimum over randomized (either public or private coin) protocols computing g with error $\leq \varepsilon$ (*i.e.* since there must exist a fixing of the private coins that achieves as good error as the average error).

Yao's minimax principle (Yao, 1977) tells that for all functions g:

$$\mathrm{R}_\varepsilon^{\rightarrow, \mathsf{pub}}(g) = \max_\mu \mathrm{D}_\varepsilon^{\rightarrow}(g; \mu) \tag{2.1}$$

2.4. Littlestone's Dimension

Let C be a concept class over domain X. A *mistake tree* T over X and C is a binary tree in which each internal node v is labelled by a point $x_v \in X$, each leaf ℓ is labelled by a concept $c_\ell \in C$ and for every node v and leaf ℓ: if ℓ is in the right subtree of v then $c_\ell(x_v) = 1$, otherwise $c_\ell(x_v) = 0$. We remark that a mistake tree over X and C does not necessarily include all concepts from C in its leaves. Such a tree is *complete* if all its leaves are at the same depth. Littlestone's (1987) dimension $\mathrm{LDim}(C)$ is defined as the depth of the deepest complete mistake tree T over X and C. Littlestone's dimension precisely characterizes the smallest number of mistakes that a learning algorithm for C will make (in the worst case) in the online mistake-bound learning model. It is also known to characterize the number of (general) equivalence queries required to learn C in Angluin's (1988) exact model of learning (Littlestone, 1987).

3. Equivalence between representation dimension and communication complexity

We relate communication complexity to private learning by considering the communication problem associated with evaluating a function f from a concept class C on an input $x \in X$. Formally, for a Boolean concept class C over domain X, define $\mathsf{Eval}_C : C \times X \to \{0,1\}$ to be the function defined as $\mathsf{Eval}_C(f, x) = f(x)$.

Our main result is the following two bounds.

Theorem 7 *For any $\varepsilon \in [0, 1/2]$ and $\delta \in [0, 1]$, and any concept class C, it holds that:*

- $\mathrm{PRDim}_{\varepsilon,\delta}(C) \leq \mathrm{R}^{\rightarrow,\mathsf{pub}}_{\varepsilon\delta}(\mathsf{Eval}_C)$.
- $\mathrm{PRDim}_{\varepsilon,\delta}(C) \geq \mathrm{R}^{\rightarrow,\mathsf{pub}}_{\varepsilon+\delta}(\mathsf{Eval}_C)$

Proof ($\leq$): let π be the public-coin one-way protocol that achieves the optimal communication complexity c. For each choice of the public random coins r, let H_r denote the set of functions $h_\sigma(x) = \pi_B(\sigma, x; r)$ over all possible σ. Thus, each H_r has size at most 2^c. Let the distribution $\mathcal{H}$ be to choose uniformly random r and then output H_r.

We show that this family (ε, δ)-probabilistically represents C. We know from the fact that π computes Eval_C with error $\varepsilon\delta$ that it must hold for all $f \in C$ and $x \in X$ that:

$$\Pr_r[\pi_B(\pi_A(f; r), x; r) \neq f(x)] \leq \varepsilon\delta$$

In particular, it must hold for any distribution $\mathcal{D}$ over X that:

$$\Pr_{\mathcal{D},r}[\pi_B(\pi_A(f; r), x; r) \neq f(x)] \leq \varepsilon\delta$$

Therefore, it must hold that

$$\Pr_r\left[\Pr_{\mathcal{D}}[\pi_B(\pi_A(f; r), x; r) \neq f(x)] > \varepsilon\right] < \delta$$

Note that $\pi_B(\pi_A(f; r), x; r) \equiv h_{\pi_A(f;r)}(x) \in H_r$ and therefore, with probability $\geq 1 - \delta$ over the choice of $H_r \overset{R}{\leftarrow} \mathcal{H}$, there exists $h \in H_r$ such that $\Pr_{\mathcal{D}}[h(x) \neq f(x)] \leq \varepsilon$.

($\geq$): let $\mathcal{H}$ be the distribution over sets of boolean functions that achieves $\mathrm{PRDim}_{\varepsilon,\delta}(C)$. We will show that for each distribution μ over inputs (f, x), we can construct a $(\varepsilon + \delta)$-correct protocol for Eval_C over μ that has communication bounded by $\mathrm{PRDim}_{\varepsilon,\delta}(C)$. Namely, we will prove that

$$\max_\mu \mathrm{D}^{\rightarrow}_{\varepsilon+\delta}(g; \mu) \leq \mathrm{PRDim}_{\varepsilon,\delta}(C) \tag{3.1}$$

By Yao's minimax principle (Equation (2.1)) (Yao, 1977) this implies that

$$\mathrm{R}^{\rightarrow,\mathsf{pub}}_{\varepsilon+\delta}(g) \leq \mathrm{PRDim}_{\varepsilon,\delta}(C)$$

Fix μ. This induces a marginal distribution $\mathcal{F}$ over functions $f \in C$ and for every $f \in C$ a distribution $\mathcal{D}_f$ which is μ conditioned on the function being f (note that μ is equivalent to drawing

f from $\mathcal{F}$ and then x from $\mathcal{D}_f$). The protocol π is defined as follows: use public coins to sample $H \stackrel{R}{\leftarrow} \mathcal{H}$. Alice knows f and so knows the distribution $\mathcal{D}_f$. Alice sends the index of $h \in H$ such that $\Pr_{\mathcal{D}_f}[h(x) \neq f(x)] \leq \varepsilon$ if such h exists or an arbitrary $h \in H$ otherwise. Bob returns $h(x)$.

The error of this protocol can be analyzed as follows. Fix f and let G_f denote the event that $H \stackrel{R}{\leftarrow} \mathcal{H}$ contains h such that $\Pr_{\mathcal{D}_f}[h(x) \neq f(x)] \leq \varepsilon$. Observe that G_f is independent of $\mathcal{D}_f$ so that even conditioned on G_f x remains distributed according to $\mathcal{D}_f$. Also, since $\mathcal{H}$ (ε, δ)-probabilistically represents C, we know that for every f, $\Pr_r[G_f] \geq 1 - \delta$. Therefore we can then deduce that:

$$\begin{aligned}\Pr_{r,(f,x)\stackrel{R}{\leftarrow}\mu}[\pi(f,x;r) = f(x)] &= \Pr_{r,(f,x)\stackrel{R}{\leftarrow}\mu}[\pi(f,x;r) = f(x) \wedge G_f] + \Pr_{r,(f,x)\stackrel{R}{\leftarrow}\mu}[\pi(f,x;r) = f(x) \wedge \neg G_f] \\ &\geq \Pr_{r,f\stackrel{R}{\leftarrow}\mathcal{F}}[G_f] \Pr_{r,\mathcal{D}_f}[\pi(f,x;r) = f(x) \mid G_f] \\ &\geq (1-\delta)(1-\varepsilon) > 1 - \delta - \varepsilon\end{aligned}$$

Thus π computes C with error at most $\varepsilon + \delta$ and it has communication bounded by $\mathrm{PRDim}_{\varepsilon,\delta}(C)$. ■

We also establish an analogous equivalence for DRDim and private-coin protocols.

Theorem 8 *For any $\varepsilon \in [0, 1/2]$, it holds that:*

- $\mathrm{DRDim}_\varepsilon(C) \leq \mathrm{R}^{\rightarrow}_{\varepsilon/2}(\mathsf{Eval}_C)$
- $\mathrm{DRDim}_\varepsilon(C) \geq \mathrm{R}^{\rightarrow}_{\varepsilon}(\mathsf{Eval}_C)$

The proof of this theorem is similar to that of Thm. 7 and appears in the full version (Feldman and Xiao, 2014).

3.1. Applications

Our equivalence theorems allow us to import many results from communication complexity into the context of private PAC learning, both proving new facts and simplifying proofs of previously known results in the process.

Separating SCDP **and** VC **dimension.** Define Thr_b as the family of functions $t_x : I_b \to \{0, 1\}$ for $x \in I_b$ where $t_x(y) = 1$ if and only if $y \geq x$. The lower bound follows from an observation that $\mathsf{Eval}_{\mathsf{Thr}_b}$ is equivalent to the "greater-than" function $\mathsf{GT}_b(x, y) = 1$ if and only if $x > y$, where $x, y \in \{0, 1\}^b$ are viewed as binary representations of integers in I_b. Note $\mathsf{Eval}_{\mathsf{Thr}_b}(t_x, y) = 1 - \mathsf{GT}_b(x, y)$ and therefore these functions are the same up to the negation. GT_b is a well studied function in communication complexity and it is known that $\mathrm{R}^{\rightarrow,\mathsf{pub}}_{1/3}(\mathsf{GT}_b) = \Omega(b)$ (Miltersen et al., 1998). By combining this lower bound with Theorem 7 we obtain that $\mathrm{VC}(\mathsf{Thr}_b) = 1$ yet $\mathrm{PRDim}(\mathsf{Thr}_b) = \Omega(b)$. From Theorem 6 it follows that $\mathrm{SCDP}(\mathsf{Thr}_b) = \Omega(b)$.

We note that it is known that VC dimension corresponds to the maximal distributional one-way communication complexity over all *product* input distributions. Hence this separation is analogous to separation of distributional one-way complexity over product distributions and the maximal distributional complexity over all distributions achieved using the greater-than function (Kremer et al., 1999).

We also give more such separations using lower bounds on PRDim based on Littlestone's dimension. These are discussed in Section 4.

Probabilistic vs. deterministic representation dimension. It was shown by Newman (1991) that public and private coin complexity are the same up to additive logarithmic terms. In our setting (and with a specific choice of error bounds to simplify presentation), Newman's theorem says that

$$\mathrm{R}^{\rightarrow}_{1/3}(\mathsf{Eval}_C) \leq \mathrm{R}^{\rightarrow,\mathsf{pub}}_{1/9}(\mathsf{Eval}_C) + O(\log\log(|C||X|)) \tag{3.2}$$

We know by Sauer's lemma that $\log|C| \leq O(\mathrm{VC}(C) \cdot \log|X|)$, therefore we deduce that:

$$\mathrm{R}^{\rightarrow}_{1/3}(\mathsf{Eval}_C) \leq \mathrm{R}^{\rightarrow,\mathsf{pub}}_{1/9}(\mathsf{Eval}_C) + O(\log\log\mathrm{VC}(C) + \log\log|X|)$$

By our equivalence theorems, this implies that

$$\mathrm{DRDim}_{1/3}(C) \leq \mathrm{PRDim}_{1/3,1/3}(C) + O(\log\log\mathrm{VC}(C) + \log\log|X|)$$

A (slightly weaker) version of this was first proved in (Beimel et al., 2013a), whose proof is similar in spirit to the proof of Newman's theorem. We also remark that the fact that $\mathrm{DRDim}_{1/3}(\mathsf{Point}_b) = \Omega(\log b)$ while $\mathrm{PRDim}_{1/3}(\mathsf{Point}_b) = O(1)$ (Beimel et al., 2010, 2013a) corresponds to the fact that the private-coin complexity of the equality function is $\Omega(\log b)$, while the public-coin complexity is $O(1)$. Here Point_b is the family of point functions, *i.e.* functions that are zero everywhere except on a single point.

Simpler learning algorithms. Using our equivalence theorems, we can "import" results from communication complexity to give simple private PAC learners. For example, the well-known constant communication equality protocol using hashing can be converted to a probabilistic representation using Theorem 7, which can then be used to learn point functions. While the resulting learner resembles the constant sample complexity learner for point functions described in (Beimel et al., 2010), we believe that this view provides useful intuition.

Furthermore, in some cases, this connection even leads to *efficient* private PAC learning algorithms. Namely, if there is a communication protocol for Eval_C where both Alice's and Bob's algorithms are polynomial-time, and in addition the resulting probabilistic representation is of polynomial size, then one can run the exponential mechanism efficiently to differentially privately learn C. For example, this is the case with point functions, where the probabilistic representation has constant size.

Another way in which our equivalence theorems simplify the study of private PAC learning is by giving an alternative way to reduce error, notably *without* explicitly using sequential boosting as was done in (Beimel et al., 2013a). Given a private PAC learner with constant error, say $(\varepsilon, \delta) = (1/8, 1/8)$, one can first convert the learner to a communication protocol with error $1/4$, use $O(\log \frac{1}{\varepsilon'\delta'})$ simple independent repetitions to reduce the error to $\varepsilon'\delta'$, and then convert the protocol back into a (ε', δ')-probabilistic representation.[3]

4. Lower Bounds via Littlestone's Dimension

In this section, we show that Littlestone's dimension lower bounds the sample complexity of differentially-private learning. Let C be a concept class over X of LDim d. Our proof is based on a reduction from the communication complexity of Eval_C to the communication complexity of Augmented Index problem on d bits. $\mathsf{AugIndex}$ is the promise problem where Alice gets a string

3. The "magic" here happens when we convert between the communication complexity and probabilistic representation using min-max type arguments. This is the same tool that can be used to prove (computationally inefficient) boosting theorems.

$x_1, \ldots, x_d \in \{0,1\}^d$ and Bob gets $i \in [d]$ and $x_1, \ldots, x_{i-1}$, and $\mathsf{AugIndex}(x, (i, x_{[i-1]})) = x_i$ where $x_{[i-1]} = (x_1, \ldots, x_{i-1})$. A variant of this problem in which the length of the prefix is not necessarily i but some additional parameter m was first explicitly defined by Bar-Yossef et al. (2004) who proved that it has randomized one-way communication complexity of $\Omega(d-m)$. The version defined above is from (Ba et al., 2010) where it is also shown that a lower bound for AugIndex follows from an earlier work of (Miltersen et al., 1998). We use the following lower bound for AugIndex.

Lemma 9 $\mathrm{R}_\varepsilon^{\rightarrow}(\mathsf{AugIndex}) \geq (1 - \mathrm{H}(\varepsilon))d$, *where* $\mathrm{H}(\varepsilon) = \varepsilon \log(1/\varepsilon) + (1-\varepsilon)\log(1/(1-\varepsilon))$ *is the binary entropy function.*

A proof of this lower bound can be easily derived by adapting the proof in (Bar-Yossef et al., 2004) and we include it in the full version (Feldman and Xiao, 2014).

We now show that if $\mathrm{LDim}(C) = d$ then one can reduce AugIndex on d bit inputs to Eval_C.

Lemma 10 *Let C be a concept class over X and $d = \mathrm{LDim}(C)$. There exist two mappings $m_C : \{0,1\}^d \rightarrow C$ and $m_X : \bigcup_{i\in[d]}\{0,1\}^i \rightarrow X$ such that for every x and $i \in [d]$, the value of $m_C(x)$ on point $m_X(x_{[i-1]})$ is equal to* $\mathsf{AugIndex}(x, (i, x_{[i-1]})) = x_i$.

Proof By the definition of LDim there exists a complete mistake tree T over X and C of depth d. For $x \in \{0,1\}^d$ consider a path from the root of the tree such that at step $j \in [d]$ we go to left subtree if $x_j = 0$ and right subtree if $x_j = 1$. Such path will end in a leaf which we denote by ℓ_x and the concept that labels it by c_x. Let $v_{x_{[i-1]}}$ denote the internal node at depth i on this path (with $v_\emptyset$ being the root) and let $z_{x_{[i-1]}}$ denote the point in X that labels $v_{x_{[i-1]}}$. Note that $z_{x_{[i-1]}}$ is uniquely determined by $x_{[i-1]}$. We define the mapping m_C as $m_C(x) = c_x$ for all $x \in \{0,1\}^d$ and the mapping m_X as $m_X(y) = z_y$ for all $y \in \bigcup_{i\in[d]}\{0,1\}^i$. To prove that the mappings correctly reduce AugIndex to Eval_C it suffices to note that by definition of a mistake tree over X and C, ℓ_x is in the subtree of $v_{x_{[i-1]}}$ and the value of c_x on $z_{x_{[i-1]}}$ is determined by whether ℓ_x is in the right (1) or left (0) subtree of $v_{x_{[i-1]}}$. By the definition of ℓ_x this is exactly x_i. ■

An immediate corollary of Lem. 10 and 9 is the following lower bound.

Corollary 11 *Let C be a concept class over X and $d = \mathrm{LDim}(C)$.* $\mathrm{R}_\varepsilon^{\rightarrow}(\mathsf{Eval}_C) \geq (1 - \mathrm{H}(\varepsilon))d$.

A stronger form of this lower bound was proved by Zhang (2011) who showed that the power of Partition Tree lower bound technique for one-way *quantum* communication complexity of Nayak (1999) can be expressed in terms of LDim of the concept class associated with the communication problem.

4.1. Applications

We can now use numerous known lower bounds for Littlestone's dimension of C to obtain lower bounds on sample complexity of private PAC learning. Here we list several examples of known results where $\mathrm{LDim}(C)$ is (asymptotically) larger than the VC dimension of C.

1. $\mathrm{LDim}(\mathsf{Thr}_b) = b$ (Littlestone, 1987). $\mathrm{VC}(\mathsf{Thr}_b) = 1$.

2. Let BOX_b^d denote the class of all axis-parallel rectangles over $[2^b]^d$, namely all concepts $r_{s,t}$ for $s,t \in [2^b]^d$ defined as $r_{s,t}(x) = 1$ if and only if for all $i \in [d]$, $s_i \le x_i \le t_i$. $\mathrm{LDim}(\mathsf{BOX}_b^d) \ge b \cdot d$ (Littlestone, 1987). $\mathrm{VC}(\mathsf{BOX}_b^d) = d+1$.

3. Let HS_b^d denote class of all linear threshold functions over $[2^b]^d$. $\mathrm{LDim}(\mathsf{HS}_b^d) = b \cdot d(d-1)/2$. This lower bound is stated in (Maass and Turan, 1994). We are not aware of a published proof and therefore a proof based on counting arguments in (Muroga, 1971) appears in the full version (Feldman and Xiao, 2014). $\mathrm{VC}(\mathsf{HS}_b^d) = d+1$.

4. Let BALL_b^d denote class of all balls over $[2^b]^d$, that is all functions obtained by restricting a Euclidean ball in $\mathbb{R}^d$ to $[2^b]^d$. Then $\mathrm{LDim}(\mathsf{BALL}_b^d) = \Omega(b \cdot d^2)$ (Maass and Turán, 1994). $\mathrm{VC}(\mathsf{BALL}_b^d) = d+1$.

4.2. Separation from PRDim

While it is natural to ask whether PRDim is equal to LDim, in fact the communication complexity literature (Zhang, 2011) already contains the following counter-example separating PRDim and LDim. Define:

$$\mathsf{Line}_p = \{f : \mathbb{Z}_p^2 \to \{0,1\} : \exists a, b \in \mathbb{Z}_p^2 \; s.t. \; f(x,y) = 1 \text{ iff } ax + b = y\}$$

It is easy to see that $\mathrm{LDim}(\mathsf{Line}_p) = 2$. It was also shown (Aaronson, 2004) that the quantum one-way communication complexity of $\mathsf{Eval}_{\mathsf{Line}_p}$ is $\Theta(\log p)$. This already implies a separation between LDim and PRDim using Theorem 7 and the fact that quantum one-way communication lower-bounds randomized public-coin communication.

We give a new and simpler proof of Aaronson's result for randomized public-coin communication in the full version (Feldman and Xiao, 2014).

5. Separating pure and (α, β)-differential privacy

We prove that it is possible to learn Line_p with (α, β)-differential privacy and (ε, δ) accuracy using $O(\frac{1}{\varepsilon\alpha} \log \frac{1}{\beta} \log \frac{1}{\delta})$ samples. This gives further evidence that it is possible to obtain much better sample complexity with (α, β)-differential privacy than pure differential privacy. Our separation is somewhat stronger than that implied by our lower bound for Thr_b and the upper bound of $O(16^{\log^*(b)})$ in (Beimel et al., 2013b) since for Line_p we are able to *match* the non-private sample complexity (when the privacy and accuracy parameters are constant[4]), even though, as mentioned in the previous section, randomized one-way communication complexity and therefore the SCDP of Line_p is asymptotically $\Theta(\log p)$. We note that our learner is not proper since in addition to lines it may output point functions and the all zero function.

Theorem 12 *For any prime p, any $\varepsilon, \delta, \alpha, \beta \in (0, 1/2)$, one can (ε, δ)-accurately learn Line_p with (α, β)-differential privacy using $O(\frac{1}{\varepsilon\alpha} \log \frac{1}{\beta} \log \frac{1}{\delta})$ samples.*

We sketch the idea here and defer the full proof to Section B. The key observation is that for Line_p, any two positively labeled points uniquely define the hidden concept. By sampling enough points, intuitively we will fall into one of three cases:

4. Formally a bound for constant β is uninformative since weak $1/\beta$ dependence is achievable by naive subsampling. In our case the dependence on $1/\beta$ is logarithmic and we can ignore this issue.

1. We see two positively labeled points and can recover the hidden concept.
2. We see only one positively labeled point, in which case we can safely output just a point function that is positive on this point.
3. We see no positively labeled points, in which case we can safely output the all zero function.

We define a "basic learner" that takes $O(1)$ samples and outputs a concept according to the above rule.

If indeed we are in one of the above cases, we can then use the "propose-test-release" paradigm (Dwork and Lei, 2009) to release the hidden concept: we run the learner many times and hope that in almost every execution it will output the *exact same hypothesis*. If this is the case we can release this unique hypothesis as follows: compute the number of samples that must be modified in order to change the majority hypothesis, add noise to make this number differentially private, and if it exceeds some appropriate threshold output the hypothesis, otherwise output the constant zero hypothesis.

There is a technical detail to overcome: it may be the case that the input distribution does not fall into any of the above cases, but lands "between" two of them, in which case the basic learner will oscillate between, say, outputting a line or outputting a point function. To handle this case, we randomize the *number* of samples we feed to the basic learner, and show that with high probability we pick a number such that we land firmly in one of the three good cases.

Finally, randomizing the number of samples leads to constant sample complexity but the dependence on confidence δ ends up being bad. We boost the confidence by running the poor-sample-complexity learner many times and sampling a single output using the exponential mechanism of McSherry and Talwar (2007).

Acknowledgements

We are grateful to Kobbi Nissim for first drawing our attention to the intriguing problem of understanding the relationship between probabilistic representation dimension and VC dimension, and for valuable discussions regarding the sample complexity of privately learning threshold functions. We thank Nina Balcan and Avrim Blum who brought up the relationship of our bounds for intervals in Section 3.1 to those based on Littlestone's dimension. Their insightful comments and questions have lead to our result in Theorem 2. We also thank Sasha Rakhlin and Sasha Sherstov for useful suggestions and references.

D.X. was supported in part by the French ANR Blanc program under contract ANR-12-BS02-005 (RDAM project), by NSF grant CNS-1237235, a gift from Google, Inc., and a Simons Investigator grant to Salil Vadhan.

References

Scott Aaronson. Limitations of quantum advice and one-way communication. In *IEEE Conference on Computational Complexity*, pages 320–332, 2004.

D. Angluin. Queries and concept learning. *Machine Learning*, 2:319–342, 1988.

Khanh Do Ba, Piotr Indyk, Eric Price, and David P. Woodruff. Lower bounds for sparse recovery. In *SODA*, pages 1190–1197, 2010.

Ziv Bar-Yossef, T. S. Jayram, Robert Krauthgamer, and Ravi Kumar. The sketching complexity of pattern matching. In *APPROX-RANDOM*, pages 261–272, 2004.

Amos Beimel, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. In *TCC*, pages 437–454, 2010.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of private learners. In *ITCS*, pages 97–110, 2013a.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *APPROX-RANDOM*, pages 363–378, 2013b.

Gyora M. Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377 – 389, 1991.

A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In *Proceedings of PODS*, pages 128–138, 2005.

Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to noninteractive database privacy. *J. ACM*, 60(2):12, 2013.

Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *COLT*, pages 155–186, 2011.

Kamalika Chaudhuri and Daniel Hsu. Convergence rates for differentially private statistical estimation. In *ICML*, 2012.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *FOCS*, pages 429–438, 2013a.

John C. Duchi, Martin J. Wainwright, and Michael I. Jordan. Local privacy and minimax bounds: Sharp rates for probability estimation. In *NIPS*, pages 1529–1537, 2013b.

C. Dwork and A. Roth. *The Algorithmic Foundations of Differential Privacy (preprint)*. 2014.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

Cynthia Dwork. Differential privacy. In *ICALP (2)*, pages 1–12, 2006.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *STOC*, pages 371–380, 2009.

Cynthia Dwork and Adam Smith. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2):135–154, 2009.

Vitaly Feldman and David Xiao. Sample complexity bounds on differentially private learning via communication complexity. *CoRR*, abs/1402.6278, 2014.

A. Gupta, M. Hardt, A. Roth, and J. Ullman. Privately releasing conjunctions and the statistical query barrier. In *STOC*, pages 803–812, 2011.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992. ISSN 0890-5401.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, June 2011.

M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6): 983–1006, 1998.

M. Kearns, R. Schapire, and L. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17 (2-3):115–141, 1994.

Ilan Kremer, Noam Nisan, and Dana Ron. On randomized one-round communication complexity. *Computational Complexity*, 8(1):21–49, 1999.

N. Littlestone. Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1987.

W. Maass and G. Turan. *How fast can a threshold gate learn?*, pages 381–414. MIT Press, 1994.

Wolfgang Maass and György Turán. Algorithms and lower bounds for on-line learning of geometrical concepts. *Machine Learning*, 14(1):251–269, 1994.

Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil P. Vadhan. The limits of two-party differential privacy. In *FOCS*, pages 81–90, 2010.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103, 2007.

Peter Bro Miltersen, Noam Nisan, Shmuel Safra, and Avi Wigderson. On data structures and asymmetric communication complexity. *J. Comput. Syst. Sci.*, 57(1):37–49, 1998.

S. Muroga. *Threshold logic and its applications*. Wiley-Interscience, New York, 1971.

Ashwin Nayak. Optimal lower bounds for quantum automata and random access codes. In *FOCS*, pages 369–377, 1999.

Ilan Newman. Private vs. common random bits in communication complexity. *Inf. Process. Lett.*, 39(2):67–71, 1991.

Anand D. Sarwate and Kamalika Chaudhuri. Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *IEEE Signal Process. Mag.*, 30 (5):86–94, 2013.

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *STOC*, pages 813–822, 2011.

Jonathan Ullman. Answering n2+O(1) counting queries with differential privacy is hard. In *STOC*, pages 361–370, 2013.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Andrew Yao. Probabilistic computations: Toward a unified measure of complexity. In *Proceedings of the 18th Annual Symposium on Foundations of Computer Science*, pages 222–227, 1977.

Shengyu Zhang. On the power of lower bound methods for one-way quantum communication complexity. In *ICALP (1)*, pages 49–60, 2011.

Appendix A. Additional preliminaries

A.1. Learning models

Definition 13 *An algorithm A PAC learns a concept class C from n examples if for every $\epsilon > 0, \delta > 0$, $f \in C$ and distribution $\mathcal{D}$ over X, A given access to $S = \{(x_i, \ell_i)\}_{i \in [n]}$ where each x_i is drawn randomly from $\mathcal{D}$ and $\ell_i = f(x_i)$, outputs, with probability at least $1 - \delta$ over the choice of S and the randomness of A, a hypothesis h such that $\mathbf{Pr}_{x \sim \mathcal{D}}[f(x) \neq h(x)] \leq \varepsilon$.*

Agnostic learning: The *agnostic* learning model was introduced by Haussler (1992) and Kearns et al. (1994) in order to model situations in which the assumption that examples are labeled by some $f \in C$ does not hold. In its least restricted version the examples are generated from some unknown distribution P over $X \times \{0, 1\}$. The goal of an agnostic learning algorithm for a concept class C is to produce a hypothesis whose error on examples generated from P is close to the best possible by a concept from C. For a Boolean function h and a distribution P over $X \times \{0, 1\}$ let $\Delta(P, h) = \mathbf{Pr}_{(x,\ell) \sim P}[h(x) \neq \ell]$. Define $\Delta(P, C) = \inf_{h \in C}\{\Delta(P, h)\}$. Kearns et al. (1994) define agnostic learning as follows.

Definition 14 *An algorithm A* agnostically *learns a concept class C if for every $\epsilon > 0, \delta > 0$, distribution P over $X \times \{0, 1\}$, A given access to $S = \{(x_i, \ell_i)\}_{i \in [n]}$ where each (x_i, ℓ_i) is drawn randomly from P, outputs, with probability at least $1 - \delta$ over the choice of S and the randomness of A, a hypothesis h such that $\Delta(P, h) \leq \Delta(P, C) + \varepsilon$.*

In both PAC and agnostic learning model an algorithm that outputs a hypothesis in C is referred to as *proper*.

Appendix B. Separation between PRDim and impure differential privacy

We restate the learner satisfying impure differential privacy.

Theorem 15 (Restatement of Theorem 12) *For any prime p, any $\varepsilon, \delta, \alpha, \beta \in (0, 1/2)$, one can (ε, δ)-accurately learn Line_p with (α, β)-differential privacy using $O(\frac{1}{\varepsilon\alpha} \log \frac{1}{\beta} \log \frac{1}{\delta})$ samples.*

We prove this theorem in two steps: first we construct a learner with poor dependence on δ and then amplify using the exponential mechanism to obtain a learner with good dependence on δ.

B.1. A learner with poor dependence on δ

Lemma 16 *For any prime p, any $\varepsilon, \delta, \alpha, \beta \in (0, 1/2)$, it suffices to take $O(\frac{1}{\varepsilon} 2^{6/\delta} \cdot \frac{1}{\alpha} \log \frac{1}{\beta\delta})$ samples in order to (ε, δ)-learn Line_p with (α, β)-differential privacy.*

Proof At a high level, we run the basic (non-private) learner based on VC-dimension $O(\frac{1}{\alpha}\log\frac{1}{\beta})$ times. We use the fact that Line_p is *stable* in that after a constant number of samples, with high probability there is a *unique* hypothesis that classifies the samples correctly. (This is simply because any two distinct points on a line define the line.) Therefore, in each of the executions of the non-private learner, we are likely to recover the same hypothesis. We can then release this hypothesis (α,β)-privately using the "Propose-Test-Release" framework.

The main challenge in implementing this intuition is to eliminate corner cases, where with roughly probability $1/2$ the sample set may contain two distinct positively labeled points and with probability $1/2$ only a single positively labeled point, as this would lead to unstable outputs. We do this by randomizing the *number* of samples we take.

Let t be a number of samples, to be chosen later. Given t samples $(x_1,y_1),\ldots,(x_t,y_t)$, our *basic learner* will do the following:

1. See if there exist two distinct samples $(x_i,y_i)\neq(x_j,y_j)$ that are both classified positively. If so, output the unique line defined by these points.
2. Otherwise, see if there exists any sample (x_i,y_i) classified positively. Output the point function that outputs 1 on (x_i,y_i) and zero elsewhere.
3. Otherwise, output the constant 0 hypothesis.

Our *overall learner* uses the basic learner as follows: first sample an integer k uniformly from the interval $[\log(\ln(3/2)/\varepsilon),\log(\ln(3/2)/\varepsilon)+6/\delta]$ and set $t=2^k$. Set $\ell=\max\{\frac{12}{\alpha}\ln\frac{1}{\beta\delta}+13,72\ln\frac{4}{\delta}\}$. Set $n=t\ell$.

1. Take n samples and cut them into ℓ subsamples of size t, and run the basic learner on each of these.
2. Let the returned hypotheses be $h_1,\ldots,h_\ell$. Define $\mathsf{freq}(h_1,\ldots,h_\ell)=\mathrm{argmax}_h\,|\{h_i=h\mid i\in[\ell]\}|$, *i.e.* the most frequently occurring hypothesis, breaking ties using lexicographical order. We define $\overline{h}=\mathsf{freq}(h_1,\ldots,h_\ell)$. Compute c to be the smallest number of h_i that must be changed in order to change the most frequently occuring hypothesis, *i.e.*
$$c=\min\left\{c\mid\exists h'_1,\ldots,h'_\ell,\mathsf{freq}(h'_1,\ldots,h'_\ell)\neq\overline{h},c=|\{i\mid h_i\neq h'_i\}|\right\}$$
3. If $c+\Lambda(1/\alpha)>\frac{1}{\alpha}\ln\frac{1}{2\beta}+1$ then output $\overline{h}$, otherwise output the constant 0 hypothesis.

Here, $\Lambda(1/\alpha)$ denotes the Laplace distribution, whose density function at point x equals $\alpha e^{-\alpha|x|}$. It is easy to check that adding $\Lambda(1/\alpha)$ to a sum of Boolean values renders that sum α-differentially private (Dwork et al., 2006).

We analyze the overall learner. Observe that once t is fixed, the basic learner is deterministic algorithm.

Privacy: we prove that the overall learner is (α,β)-differentially private. Consider any two neighboring inputs $x,x'\in(\mathbb{Z}_p^2)^n$. There are two cases:

- The most frequent hypothesis $\overline{h}$ returned by running the basic learner on the ℓ subsamples of x,x' is the same. In this case, there are two possible outputs of the mechanism, either $\overline{h}$ or the

0 hypothesis. Due to the fact that we decide between them using a count with Laplace noise and the count has sensitivity 1, the probability assigned to either output changes by at most a multiplicative $e^{-\alpha}$ factor between x, x'.

- The most frequent hypotheses are different. In this case $c = 1$ for both x, x'. The probability of *not* outputting 0 in either case is given by

$$\Pr[\Lambda(1/\alpha) > \frac{1}{\alpha} \ln \frac{1}{2\beta}] = \beta$$

  Otherwise, in both cases they output 0.

Accuracy: we now show that the overall learner (ε, δ)-PAC learns. We claim that:

Claim 17 *Fix any hidden line f and any input distribution $\mathcal{D}$. With probability $1 - \delta/2$ over the choice of t, there is a* unique *hypothesis with error $\leq \varepsilon$ that the basic learner will output with probability at least $2/3$ when given t independent samples from $\mathcal{D}$.*

Let us first assume this claim is true. Then it is easy to show that the overall learner (ε, δ)-learns: suppose we are in the $1 - \delta/2$ probability case where there is a unique hypothesis with error $\leq \varepsilon$ output by the basic learner. Then, by Chernoff, since $\ell \geq 72 \ln \frac{4}{\delta}$ it holds that with probability $1-\delta/4$ at least $7/12$ fraction of the basic learner outputs will be this unique hypothesis. This means that the number of samples that must be modified to change the most frequent hypothesis is $c \geq \frac{\ell}{12}$. Therefore since $\ell \geq \frac{12}{\alpha} \ln \frac{1}{\beta\delta} + 13$, in this case the probability that the overall learner does not output this unique hypothesis is bounded by:

$$\Pr[c + \Lambda(\tfrac{1}{\alpha}) \leq \tfrac{1}{\alpha} \ln \tfrac{1}{2\beta} + 1] \leq \Pr[\Lambda(\tfrac{1}{\alpha}) < -\tfrac{1}{\alpha} \ln \tfrac{1}{\delta}] = \tfrac{\delta}{4}$$

Thus the overall probability of not returning an ε-good hypothesis is at most δ.

Proof of Claim 17 Fix a concept f defined by a line given by $(a, b) \in \mathbb{Z}_p^2$ and any input distribution $\mathcal{D}$ over $\mathbb{Z}_p^2$.

Define the following events $\mathsf{None}_t, \mathsf{One}_t, \mathsf{Two}_t$ parameterized by an integer $t > 0$ and defined over the probability space of drawing $(x_1, y_1), \ldots, (x_t, y_t)$ independently from $\mathcal{D}$:

- None_t is the event that all of the (x_i, y_i) are not on the line (a, b).
- One_t is the event that there exists some (x_i, y_i) on the line (a, b), and furthermore for every other (x_j, y_j) on the line (a, b) is in fact equal to (x_i, y_i).
- Two_t is the event that there exists distinct $(x_i, y_i) \neq (x_j, y_j)$ that are both on the line (a, b).

Next we will show that with probability $1 - \delta/2$ over the choice of t, one of these three events has probability at least $2/3$, and then we show that this suffices to imply the claim.

Let $r = \Pr_{(x,y)\sim\mathcal{D}}[f(x, y) = 1]$, let $q_{x,y} = \Pr_{(x',y')\sim\mathcal{D}}[(x', y') = (x, y)]$, and let $q = \max_{(x,y)\in f^{-1}(1)} q_{x,y}$. We can characterize the probabilities of $\mathsf{None}_t, \mathsf{One}_t, \mathsf{Two}_t$ in terms of r, q, t as follows:

$$\begin{aligned}
\Pr[\mathsf{None}_t] &= (1 - r)^t \\
\Pr[\mathsf{One}_t] &= \sum_{(x,y)\in f^{-1}(1)} ((1 - r + q_{x,y})^t - (1 - r)^t) \\
\Pr[\mathsf{Two}_t] &= 1 - \Pr[\mathsf{None}_t] - \Pr[\mathsf{One}_t]
\end{aligned}$$

The characterizations for $\mathsf{None}_t, \mathsf{Two}_t$ are obvious. The characterization of One_t is exactly the probability over all $(x,y) \in f^{-1}(1)$ that all samples are either labeled 0 or equal (x,y), excluding the event that they are all labeled 0.

From the above and by considering the (x,y) maximizing $q_{x,y}$, we have the following bounds:

$$\Pr[\mathsf{None}_t] \geq 1 - rt \tag{2.1}$$
$$\Pr[\mathsf{One}_t] \geq (1-r+q)^t - (1-r)^t \geq 1 - (r-q)t - e^{-rt} \tag{2.2}$$
$$\Pr[\mathsf{Two}_t] \geq (1-e^{-rt/2})(1-e^{-(r-q)t/2}) \tag{2.3}$$

The first two follow directly from the fact that for all $x \in \mathbb{R}$ it holds that $1 - x \leq e^x$ and also for all $x \in [0,1]$ and $y \geq 1$ it holds that $(1-x)^y \geq 1 - xy$. Equation (2.3) follows from the following argument. Two_t contains the sub-event where there is at least one positive example in the first $t/2$ samples and a different positive example in the second $t/2$ samples. The probability of this sub-event is lower-bounded by $(1-(1-r)^{t/2})(1-(1-r+q)^{t/2}) \geq (1-e^{-rt/2})(1-e^{-(r-q)t/2})$.

t is good with high probability. Let us say that t is good for None_t if $t \leq \frac{1}{3r}$. We say t is good for One_t if $t \in [\frac{\ln 6}{r}, \frac{1}{6(r-q)}]$. We say t is good for Two_t if $t \geq \frac{2\ln 6}{r-q}$. (It is possible that some of these events may be empty, but this does not affect our argument.) Using Equation (2.1), Equation (2.2) and Equation (2.3), it is clear that if t is good for some event, then the probability of that event is at least $2/3$.

Let us say t is good if it is good for any one of $\mathsf{None}_t, \mathsf{One}_t, \mathsf{Two}_t$. t is good means the following when viewed on the logarithmic scale:

$$\log t \in [0, \log\tfrac{1}{r} - \log 3] \ \cup\ [\log\tfrac{1}{r} + \log\ln 6, \log\tfrac{1}{r-q} - \log 6] \ \cup\ [\log\tfrac{1}{r-q} + \log(2\ln 6), \infty)$$

But this means that t is bad on the logarithmic scale is equivalent to:

$$\log t \in (\log\tfrac{1}{r} - \log 3, \log\tfrac{1}{r} + \log\ln 6) \ \cup\ (\log\tfrac{1}{r-q} - \log 6, \log\tfrac{1}{r-q} + \log(2\ln 6)) \tag{2.4}$$

Thus, for any r, there are at most 3 integer values of $\log t$ that are bad. But recall that $t = 2^k$ where k is uniformly chosen from $\{\log(\ln(3/2)/\varepsilon), \ldots, \log(\ln(3/2)/\varepsilon) + 6/\delta\}$. Therefore the probability that $k = \log t$ is one of the bad values defined in Equation (2.4) is at most $\delta/2$.

When t is good, basic learner outputs unique accurate hypothesis. To conclude, we argue that when t is good then the basic learner will output a unique hypothesis with error $\leq \varepsilon$ with probability $\geq 2/3$. This is obvious when t is good for Two_t, since whenever the basic learner sees two points on the line, it recovers the exact line. It is also easy to see that when t is good for None_t, the basic learner outputs the 0 hypothesis with probability $2/3$, and this has error at most ε since

$$2/3 \leq \Pr[\mathsf{None}_t] \leq (1-r)^t \leq e^{-rt} \Rightarrow r \leq \ln(3/2)/t \leq \varepsilon$$

It remains to argue that the basic learner outputs a unique hypothesis with error at most ε when t is good for One_t. Observe that we have actually set the parameters so that when t is good for One_t, it holds that:

$$\Pr[\mathsf{One}_t \wedge \text{unique positive point is } (x_{\max}, y_{\max})] \geq 2/3 \tag{2.5}$$

where $(x_{\max}, y_{\max}) = \operatorname{argmax}_{(x,y)\in f^{-1}(1)} q_{x,y}$. Therefore, for such t, the basic learner will output the point function that is positive on exactly $(x_{\max}, y_{\max})$ with probability at least $2/3$.

To show that this point function has error at most ε, it suffices to prove that

$$\Pr[f(x,y) = 1 \wedge (x,y) \neq (x_{\max}, y_{\max})] = r - q \leq \varepsilon$$

From Equation (2.5), we deduce that:

$$2/3 \leq (1-r+q)^t - (1-r)^t \leq e^{-(r-q)t} \Rightarrow r - q \leq \ln(3/2)/t \leq \varepsilon$$

This concludes the proof. ■

Improving dependence on δ: We now improve the exponential dependence on $1/\delta$ in Lemma 16 to prove Theorem 12. We will use the algorithm of Lemma 16 with $\delta = 1/2$ and accuracy $\varepsilon/2$ $k = O(\log(1/\delta))$ times independently in order to construct a set H of k hypotheses. We then draws a fresh sample S of $O(\log(1/\delta)/(\varepsilon\alpha))$ examples and select one of the hypotheses based on their error on S using the exponential mechanism of (McSherry and Talwar, 2007). This mechanism chooses a hypothesis from H with probability proportional to $e^{-\alpha \cdot \mathsf{err}_S(h)/2}$, where $\mathsf{err}_S(h)$ is $\mathsf{err}_S(h) = |\{(x,\ell) \in S \mid h(x) \neq \ell\}|$. Simple analysis (*e.g.* Kasiviswanathan et al., 2011; Beimel et al., 2013a) then shows that the selection mechanism is α-differentially private and outputs a hypothesis that has error of at most ε on $\mathcal{D}$ with probability at least $1 - \delta$. Note that each of the k copies of the low-confidence algorithm and the exponential mechanism are run on disjoint sample sets and therefore there is no privacy loss from such composition. Hence the resulting algorithm is also (α, β)-differentially private. We include formal details in the full version (Feldman and Xiao, 2014).