Private PAC learning implies finite Littlestone dimension

Noga Alon* Roi Livni† Maryanthe Malliaris‡ Shay Moran§

March 8, 2019

Abstract

We show that every approximately differentially private learning algorithm (possibly improper) for a class H with Littlestone dimension d requires $\Omega\big(\log^*(d)\big)$ examples. As a corollary it follows that the class of thresholds over $\mathbb{N}$ can not be learned in a private manner; this resolves open questions due to [Bun et al., 2015, Feldman and Xiao, 2015]. We leave as an open question whether every class with a finite Littlestone dimension can be learned by an approximately differentially private algorithm.

1 Introduction

Private learning concerns the design of learning algorithms for problems in which the input sample contains sensitive data that needs to be protected. Such problems arise in various contexts, including those involving social network data, financial records, medical records, etc. The notion of differential privacy [Dwork et al., 2006b,a], which is a standard mathematical formalism of privacy, enables a systematic study of algorithmic privacy in machine learning. The question

"Which problems can be learned by a private learning algorithm?"

has attracted considerable attention [Rubinstein et al., 2009, Kasiviswanathan et al., 2011, Chaudhuri et al., 2011, Beimel et al., 2013, 2014, Chaudhuri et al., 2014, Balcan and Feldman,

*Department of Mathematics, Princeton University, Princeton, New Jersey, USA and Schools of Mathematics and Computer Science, Tel Aviv University, Tel Aviv, Israel. Noga Alon's research is supported in part by ISF grant No. 281/17, GIF grant No. G-1347-304.6/2016 and the Simons Foundation.

†Department of Computer Science, Princeton University, Princeton, New Jersey, USA. Research supported in part by the Eric and Wendy Schmidt foundation for strategic innovation.

‡Department of Mathematics, University of Chicago, Chicago, Illinois, USA. Research partially supported by NSF 1553653 and a Minerva research foundation membership at IAS.

§Department of Computer Science, Princeton University, Princeton, New Jersey, USA. Part of this research was done while the author was a member at the Institute for Advanced Study, Princeton and was supported by the National Science Foundation under agreement No. CCF-1412958.

2015, Bun et al., 2015, Beimel et al., 2015, Feldman and Xiao, 2015, Wang et al., 2016, Cummings et al., 2016, Beimel et al., 2016, Bun et al., 2016, Bassily et al., 2016, Ligett et al., 2017, Bassily et al., 2018, Dwork and Feldman, 2018].

Learning thresholds is one of the most basic problems in machine learning. This problem consists of an unknown threshold function $c : \mathbb{R} \to \{\pm 1\}$, an unknown distribution D over $\mathbb{R}$, and the goal is to output an hypothesis $h : \mathbb{R} \to \{\pm 1\}$ that is close to c, given access to a limited number of input examples $(x_1, c(x_1)), \ldots, (x_m, c(x_m))$, where the x_i's are drawn independently from D.

The importance of thresholds stems from that it appears as a subclass of many other well-studied classes. For example, it is the one dimensional version of the class of *Euclidean Half-Spaces* which underlies popular learning algorithms such as kernel machines and neural networks (see e.g. Shalev-Shwartz and Ben-David [2014]).

Standard PAC learning of thresholds without privacy constraints is known to be easy and can be done using a constant number of examples. In contrast, whether thresholds can be learned privately turned out to be more challenging to decide, and there has been an extensive amount of work that addressed this task: the work by Kasiviswanathan et al. [2011] implies a *pure differentially private* algorithm (see the next section for formal definitions) that learns thresholds over a finite $X \subseteq \mathbb{R}$ of size n with $O(\log n)$ examples. Feldman and Xiao [2015] showed a matching lower bound for any pure differentially private algorithm. Beimel et al. [2016] showed that by relaxing the privacy constraint to *approximate differential privacy*, one can significantly improve the upper bound to some $2^{O(\log^*(n))}$. Bun et al. [2015] further improved the upper bound from [Beimel et al., 2016] by polynomial factors and gave a lower bound of $\Omega(\log^*(n))$ that applies for any proper learning algorithm. They also explicitly asked whether the dependence on n can be removed in the improper case. Feldman and Xiao [2015] asked more generally whether any class can be learned privately with sample complexity depending only on its VC dimension (ignoring standard dependencies on the privacy and accuracy parameters). Our main result (Theorem 1) answers these questions by showing that a similar lower bound applies for any (possibly improper) learning algorithm.

Despite the impossibility of privately learning thresholds, there are other natural learning problems that can be learned privately, In fact, even for the class of Half-spaces, private learning is possible if the target half-space satisfies a large *margin*[1] assumption [Blum et al., 2005, Chaudhuri et al., 2011].

Therefore, it will be interesting to find a natural invariant that characterizes which classes can be learned privately (like the way the VC dimension characterizes PAC learning [Blumer et al., 1989, Vapnik and Chervonenkis, 1971]). Such parameters exist in the case of pure differentially private learning; these include the *one-way communication complexity* characterization by Feldman and Xiao [2015] and the *representation dimension* by Beimel et al. [2013]. However, no such parameter is known for approximate differentially private learning. We next suggest a candidate invariant that rises naturally from this work.

[1]The margin is a geometric measurement for the distance between the separating hyperplane and typical points that are drawn from the target distribution.

1.1 Littlestone dimension vs. approximate private learning

The Littlestone dimension [Littlestone, 1987] is a combinatorial parameter that characterizes learnability of binary-labelled classes within *Online Learning* both in the realizable case [Littlestone, 1987] and in the agnostic case [Ben-David et al., 2009].

It turns out that there is an intimate relationship between thresholds and the Littlestone dimension: a class H has a finite Littlestone dimension if and only if it does not embed thresholds as a subclass (for a formal statement, see Theorem 3); this follows from a seminal result in model theory by Shelah [1978]. As explained below, Shelah's theorem is usually stated in terms of orders and ranks. Chase and Freitag [2018] noticed[2] that the Littlestone dimension is the same as the model theoretic rank. Meanwhile, order translates naturally to thresholds. To make Theorem 3 more accessible for readers with less background in model theory, we provide a combinatorial proof in the appendix.

While it still remains open whether finite Littlestone dimension is indeed equivalent to private learnability, our main result (Theorem 1) combined with the above connection between Littlestone dimension and thresholds (Theorem 3) imply an implication in one direction: At least $\Omega(\log^* d)$ examples are required for privately learning any class with Littlestone dimension d (see Corollary 2).

It is worth noting that Feldman and Xiao [2015] studied the Littlestone dimension in the context of pure differentially private learning: (i) they showed that $\Omega(d)$ examples are required for learning a class with Littlestone dimension d in a pure differentially private manner, (ii) they exhibited classes with Littlestone dimension 2 that can not be learned by pure differentially private algorithms, and (iii) they showed that these classes can be learned by approximate differential private algorithms.

Organization The rest of this manuscript is organized as follows: Section 2 presents the two main results, Section 3 contains definitions and technical background from machine learning and differential privacy, and Section 4 and Section 5 contain the proofs.

2 Main Results

We next state the two main results of this paper. The statements use technical terms from differential privacy and machine learning whose definitions appear in Section 3.

We begin by the following statement that resolves an open problem in Feldman and Xiao [2015] and Bun et al. [2015]:

Theorem 1 (Thresholds are not privately learnable)**.** *Let $X \subseteq \mathbb{R}$ of size $|X| = n$ and let $\mathcal{A}$ be a $(\frac{1}{16}, \frac{1}{16})$-accurate learning algorithm for the class of thresholds over X with sample complexity m which satisfies (ϵ, δ)-differential privacy with $\epsilon = 0.1$ and $\delta = O(\frac{1}{m^2 \log m})$. Then,*

$$m \geq \Omega(\log^* n).$$

In particular, the class of thresholds over an infinite X can not be learned privately.

[2] Interestingly, though the Littlestone dimension is a basic parameter in Machine Learning (ML), this result has not appeared in the ML literature.

Theorem 1 and Theorem 3 (which is stated in the next section) imply that any privately learnable class has a finite Littlestone dimension. As elaborated in the introduction, this extends a result by Feldman and Xiao [2015].

Corollary 2 (Private learning implies finite Littlestone dimension)**.** *Let H be an hypothesis class with Littlestone dimension $d \in \mathbb{N} \cup \{\infty\}$ and let $\mathcal{A}$ be a $(\frac{1}{16}, \frac{1}{16})$-accurate learning algorithm for H with sample complexity m which satisfies (ϵ, δ)-differential private with $\epsilon = 0.1$ and $\delta = O(\frac{1}{m^2 \log m})$. Then,*

$$m \geq \Omega(\log^* d).$$

In particular any class that is privately learnable has a finite Littlestone dimension.

3 Preliminaries

3.1 PAC learning

We use standard notation from statistical learning, see e.g. [Shalev-Shwartz and Ben-David, 2014]. Let X be a set and let $Y = \{\pm 1\}$. An *hypothesis* is an $X \to Y$ function. An *example* is a pair in $X \times Y$. A *sample* S is a finite sequence of examples. The *loss of h with respect to S* is defined by

$$L_S(h) = \frac{1}{|S|} \sum_{(x_i, y_i) \in S} 1[h(x_i) \neq y_i].$$

The *loss of h with respect to a distribution D* over $X \times Y$ is defined by

$$L_D(h) = \Pr_{(x,y) \sim D}[h(x) \neq y].$$

Let $\mathcal{H} \subseteq Y^X$ be an *hypothesis class*. S is said to be *realizable by $\mathcal{H}$* if there is $h \in H$ such that $L_S(h) = 0$. D is said to be *realizable by $\mathcal{H}$* if there is $h \in H$ such that $L_D(h) = 0$. A *learning algorithm A* is a (possibly randomized) mapping taking input samples to output hypotheses. We denote by $A(S)$ the distribution over hypotheses induced by the algorithm when the input sample is S. We say that A *learns*[3] *a class $\mathcal{H}$* with *α-error*, $(1-\beta)$*-confidence*, and *sample-complexity m* if for every realizable distribution D:

$$\Pr_{S \sim D^m,\, h \sim A(S)}[L_D(h) > \alpha] \leq \beta,$$

For brevity if A is a learning algorithm with α-error and $(1-\beta)$-confidence we will say that A is an (α, β)*-accurate learner*.

Littlestone Dimension The Littlestone dimension is a combinatorial parameter that characterizes regret bounds in Online Learning [Littlestone, 1987, Ben-David et al., 2009]. The definition of this parameter uses the notion of *mistake-trees*: these are binary decision trees whose internal nodes are labelled by elements of X. Any root-to-leaf path in a mistake tree can be described as a sequence of examples $(x_1, y_1), ..., (x_d, y_d)$, where x_i is the label of the

[3] We focus on the realizable case.

i'th internal node in the path, and $y_i = +1$ if the $(i+1)$'th node in the path is the right child of the i'th node, and otherwise $y_i = -1$. We say that a tree T is *shattered* by $\mathcal{H}$ if for any root-to-leaf path $(x_1, y_1), ..., (x_d, y_d)$ in T there is $h \in \mathcal{H}$ such that $h(x_i) = y_i$, for all $i \le d$. The Littlestone dimension of $\mathcal{H}$, denoted by $\mathrm{Ldim}(\mathcal{H})$, is the depth of largest complete tree that is shattered by $\mathcal{H}$.

Recently, Chase and Freitag [2018] noticed that the Littlestone dimension coincides with a model-theoretic measure of complexity, Shelah's 2-rank.

A classical theorem of Shelah connects bounds on 2-rank (Littlestone dimension) to bounds on the so-called order property in model theory. The order property corresponds naturally to the concept of *thresholds*. Let $\mathcal{H} \subseteq \{\pm 1\}^X$ be an hypothesis class. We say that $\mathcal{H}$ *contains* k *thresholds* if there are $x_1, \ldots, x_k \in X$ and $h_1, \ldots, h_k \in \mathcal{H}$ such that $h_i(x_j) = 1$ if and only if $i \le j$ for all $i, j \le k$.

Shelah's result (part of the so-called Unstable Formula Theorem[4]) [Shelah, 1978, Hodges, 1997], which we use in the following translated form, provides a simple and elegant connection between Littlestone dimension and thresholds.

Theorem 3. *(Littlestone dimension and thresholds [Shelah, 1978, Hodges, 1997])*
Let $\mathcal{H}$ be an hypothesis class, then:

1. *If the* $\mathrm{Ldim}(\mathcal{H}) \ge d$ *then $\mathcal{H}$ contains* $\lfloor \log d \rfloor$ *thresholds*
2. *If $\mathcal{H}$ contains* d *thresholds then its* $\mathrm{Ldim}(\mathcal{H}) \ge \lfloor \log d \rfloor$.

For completeness, we provide a combinatorial proof of Theorem 3 in Appendix B.

In the context of model theory, Theorem 3 is used to establish an equivalence between finite Littlestone dimension and *stable theories*. It is interesting to note that an analogous connection between theories that are called *NIP theories* and VC dimension has also been previously observed and was pointed out by Laskowski [1992]; this in turn led to results in Learning theory: in particular within the context of compression schemes [Livni and Simon, 2013] but also some of the first polynomial bounds for the VC dimension for sigmoidal neural networks [Karpinski and Macintyre, 1997].

3.2 Privacy

We use standard notation from differential privacy. For more background see e.g. the surveys [Dwork and Roth, 2014, Vadhan, 2017]. For $s, t \in \mathbb{R}$ let $a =_{\epsilon,\delta} b$ denote the statement

$$a \le e^{\epsilon} b + \delta \text{ and } b \le e^{\epsilon} a + \delta.$$

We say that two distributions p, q are (ϵ, δ)-*indistinguishable* if $p(E) =_{\epsilon,\delta} q(E)$ for every event E. Note that when $\epsilon = 0$ this specializes to the total variation metric.

Definition 4 (Private Learning Algorithm)**.** *A randomized learning algorithm*

$$A : (X \times \{\pm 1\})^m \to \{\pm 1\}^X$$

is (ϵ, δ)-differentially private if for every two samples $S, S' \in (X \times \{\pm 1\})^m$ that disagree on a single example, the output distributions $A(S)$ and $A(S')$ are (ϵ, δ)-indistinguishable.

[4]Shelah [1978] provides a qualitative statement, a quantitative one that is more similar to Theorem 3 can be found at Hodges [1997]

The parameters ϵ, δ are usually treated as follows: ϵ is a small constant (say 0.1), and δ is negligible, $\delta = m^{-\omega(1)}$, where m is the input sample size. The case of $\delta = 0$ is also referred to as *pure differential privacy*. A common interpretation of a negligible $\delta > 0$ is that there is a tiny chance of a catastrophic event (in which perhaps all the input data is leaked) but otherwise the algorithm satisfies pure differential privacy. Thus, a class $\mathcal{H}$ is privately learnable if it is PAC learnable by an algorithm A that is $(\epsilon(m), \delta(m))$-differentially private with $\epsilon(m) \leq o(1)$, and $\delta(m) \leq m^{-\omega(1)}$.

We will use the following corollary of the *Basic Composition Theorem* from differential privacy (see, e.g. Theorem 3.16 in [Dwork and Roth, 2014]).

Lemma 5. *[Dwork et al., 2006a, Dwork and Lei, 2009] If p, q are (ϵ, δ)-indistinguishable then for all $k \in \mathbb{N}$, p^k and q^k are $(k\epsilon, k\delta)$-indistinguishable, where p^k, q^k are the k-fold products of p, q (i.e. corresponding to k independent samples).*

For completeness, a proof of this statement appears in Appendix A.

Private Empirical Learners It will be convenient to consider the following task of minimizing the empirical loss.

Definition 6 (Empirical Learner). *Algorithm A is (α, β)-accurate empirical learner for a hypothesis class $\mathcal{H}$ with sample complexity m if for every $h \in \mathcal{H}$ and for every sample $S = ((x_1, h(x_1), \ldots, (x_m, h(x_m))) \in (X \times \{0,1\})^m$ the algorithm A outputs a function f satisfying*

$$\Pr_{f \sim A(S)}\big(L_S(f) \leq \alpha\big) \geq 1 - \beta$$

This task is simpler to handle than PAC learning, which is a distributional loss minimization task. Replacing PAC learning by this task does not lose generality; this is implied by the following result by Bun et al. [2015].

Lemma 7. *[Bun et al. [2015], Lemma 5.9] Suppose $\epsilon < 1$ and A is an (ϵ, δ)-differentially private (α, β)–accurate learning algorithm for a hypothesis class $\mathcal{H}$ with sample complexity m. Then there exists an (ϵ, δ)–differentially private (α, β)–accurate empirical learner for $\mathcal{H}$ with sample complexity $9m$.*

3.3 Additional notations

A sample S of an even length is called *balanced* if half of its labels are $+1$'s and half are -1's.

For a sample S, let S_X denote the underlying set of unlabeled examples: $S_X = \big\{x | (\exists y) : (x, y) \in S\big\}$. Let A be a randomized learning algorithm. It will be convenient to associate with A and S the function $A_S : X \to [0, 1]$ defined by

$$A_S(x) = \Pr_{h \sim A(S)}\big[h(x) = 1\big].$$

Intuitively, this function represents the average hypothesis outputted by A when the input sample is S.

For the next definitions assume that the domain X is linearly ordered. Let $S = ((x_i, y_i))_{i=1}^m$ be a sample. We say that S is *increasing* if $x_1 < x_2 < \ldots < x_m$. For $x \in X$ define $\mathrm{ord}_S(x)$

by $|\{i|x_i \leq x\}|$. Note that the set of points $x \in X$ with the same $\mathrm{ord}_S(x)$ form an interval whose endpoints are two consecutive examples in S (consecutive with respect to the order on X, i.e. there is no example x_i between them).

The *tower function* $\mathsf{twr}_k(x)$ is defined by the recursion

$$\mathsf{twr}^{(i)}x = \begin{cases} x & i = 1, \\ 2^{\mathsf{twr}(i-1)(x)} & i > 1. \end{cases}$$

The iterated logarithm, $\log^{(k)}(x)$ is defined by the recursion

$$\log^{(i)} x = \begin{cases} \log x & i = 0, \\ 1 + \log^{(i-1)} \log x & i > 0. \end{cases}$$

The function $\log^* x$ equals the number of times the iterated logarithm must be applied before the result is less than or equal to 1. It is defined by the recursion

$$\log^* x = \begin{cases} 0 & x \leq 1, \\ 1 + \log^* \log x & x > 1. \end{cases}$$

4 A lower bound for privately learning thresholds

In this section we prove Theorem 1.

4.1 Proof overview

We begin by considering an arbitrary differentially private algorithm A that learns the class of thresholds over an ordered domain X of size n. Our goal is to show a lower bound of $\Omega(\log^* n)$ on the sample complexity of A. A central challenge in the proof follows because A may be improper and output arbitrary hypotheses (this is in contrast with proving impossibility results for proper algorithms where the structure of the learned class can be exploited).

The proof consists of two parts: (i) the first part handles the above challenge by showing that for any algorithm (in fact, for any mapping that takes input samples to output hypotheses) there is a large subset of the domain that is *homogeneous with respect to the algorithm*. This notion of homogeneity places useful restrictions on the algorithm when restricting it to the homogeneous set. (ii) The second part of the argument utilizes such a large homogeneous set $X' \subseteq X$ to derive a lower bound on the sample complexity of the algorithm in terms of $|X'|$.

We note that the Ramsey argument in the first part is quite general: it does not use the definition of differential privacy and could perhaps be useful in other sample complexity lower bounds. Also, a similar argument was used by Bun [2016] in a weaker lower bound for privately learning thresholds in the proper case. However, the second and more technical part of the proof is tailored specifically to the definition of differential privacy. We next outline each of these two parts.

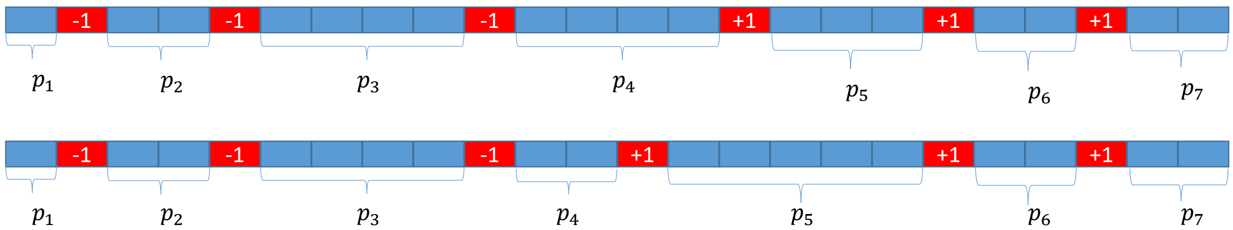

Figure 1: Depiction of two possible outputs of an algorithm over an homogeneous set, given two input samples from the set (marked in red). The number p_i denote, for a given point x, the probability that $h(x) = 1$, where $h \sim A(S)$ is the hypothesis h outputted by the algorithm on input sample S. These probabilities depends (up to a small additive error) only on the interval that x belongs to. In the figure above we changed in the input the fourth example – this only affects the interval and not the values of the p_i's (again, up to a small additive error).

Reduction to an algorithm over an homogeneous set As discussed above, the first step in the proof is about identifying a large homogeneous subset of the input domain X on which we can control the output of A: a subset $X' \subseteq X$ is called *homogeneous with respect to* A if there is a list of numbers $p_0, p_1, \dots, p_m$ such that for every increasing balanced sample S of points from X' and for every x' from X' with $\text{ord}_S(x') = i$:

$$|A_S(x') - p_i| \leq \gamma,$$

where γ is sufficiently small. For simplicity, in this proof-overview we will assume that $\gamma = 0$ (in the formal proof γ is some $O(1/m)$ - see Definition 8). So, for example, if A is deterministic then $h = A(S)$ is constant over each of the intervals defined by consecutive examples from S. See Figure 1 for an illustration of homogeneity.

The derivation of a large homogeneous set follows by a standard application of Ramsey Theorem for hyper-graphs using an appropriate coloring (Lemma 9).

Lower bound for an algorithm defined on large homogeneous sets We next assume that $X' = \{1, \dots, k\}$ is a large homogeneous set with respect to A (with $\gamma = 0$). We will obtain a lower bound on the sample complexity of A, denoted by m, by constructing a family $\mathcal{P}$ of distributions such that: (i) on the one hand $|\mathcal{P}| \leq 2^{\tilde{O}(m^2)}$, and (ii) on the other hand $|\mathcal{P}| \geq \Omega(k)$. Combining these inequalities yields a lower bound on m and concludes the proof.

The construction of $\mathcal{P}$ proceeds as follows and is depicted in Figure 2: let S be an increasing balanced sample of points from X'. Using the fact that A learns thresholds it is shown that for some $i_1 < i_2$ we have that $p_{i_1} \leq 1/3$ and $p_{i_2} \geq 2/3$. Thus, by a simple averaging argument there is some $i_1 \leq i \leq i_2$ such that $p_i - p_{i-1} \geq \Omega(1/m)$.

The last step in the construction is done by picking an increasing sample S such that the interval (x_{i-1}, x_{i+1}) has size $n = \Omega(k)$. For $x \in (x_{i-1}, x_{i+1})$, let S_x denote the sample obtained by replacing x_i with x in S. Each output distribution $A(S_x)$ can be seen as a distribution over the cube $\{\pm 1\}^n$ (by restricting the output hypothesis to the interval (x_{i-1}, x_{i+1}), which is of size n). This is the family of distributions $\mathcal{P} = \{P_j : j \le n\}$. Since A is private, and by choice of the interval (x_i, x_{i+1}) we obtain that $\mathcal{P}$ has the following two properties:

- $P_{j'}, P_{j''}$ are (ϵ, δ)-indistinguishable for all j', j'', and
- Put $r = \frac{p_{i-1}+p_i}{2}$, then for all P_j

$$(\forall x \le n) : \Pr_{v \sim P_j}\left[v(x) = 1\right] = \begin{cases} r - \Omega(1/m) & x < j, \\ r + \Omega(1/m) & x > j. \end{cases}$$

It remains to show that $\Omega(k) \le |\mathcal{P}| \le 2^{\tilde{O}(m^2)}$. The lower bound follows directly from the definition of $\mathcal{P}$. The upper bound requires a more subtle argument: it exploits the assumption that δ is small and Lemma 5 via a binary-search argument and concentration bounds. This argument appears in Lemma 13, whose proof is self-contained.

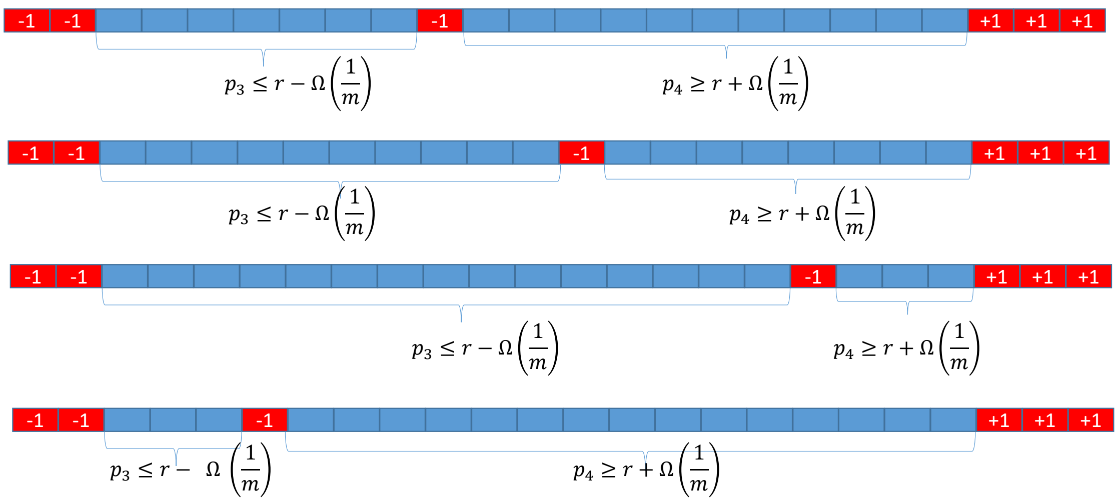

Figure 2: An illustration of the definition of the family P. Given an homogeneous set and two consecutive intervals where there is a gap of at least $\Omega(1/m)$ between p_i and p_{i-1} (here $i = 4$). The distributions in P correspond to the different positions of the i'th example, which separates between the $(i-1)$'th and the i'th intervals.

4.2 Proof of Theorem 1

The proof uses the following definition of homogeneous sets. Recall the definitions of balanced sample and of an increasing sample. In particular that a sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ of an even size is realizable (by thresholds), balanced, and increasing if and only if $x_1 < x_2 < \ldots < x_m$ and the first half of the y_i's are -1 and the second half are $+1$.

Definition 8 (m-homogeneous set). *A set $X' \subseteq X$ is m-homogeneous with respect to a learning algorithm A if there are numbers $p_i \in [0, 1]$, for $0 \le i \le m$ such that for every increasing balanced realizable sample $S \in \left(X' \times \{\pm 1\}\right)^m$ and for every $x \in X' \setminus S_X$:*

$$\left|A_S(x) - p_i\right| \le \frac{1}{10^2 m},$$

where $i = ord_S(x)$. The list $(p_i)_{i=0}^m$ is called the probabilities-list of X' with respect to A.

Proof of Theorem 1. Let A be a $(1/16, 1/16)$-accurate learning algorithm that learns the class of thresholds over X with m examples and is (ϵ, δ)-differential private with $\epsilon = 0.1, \delta = \frac{1}{10^3 m^2 \log m}$. By Lemma 7 we may assume without loss of generality that A is an empirical learner with the same privacy and accuracy parameters and sample size that is at most 9 times larger.

Theorem 1 follows from the following two lemmas:

Lemma 9 (Every algorithm has large homogeneous sets)**.** *Let A be a (possibly randomized) algorithm that is defined over input samples of size m over a domain $X \subseteq R$ with $|X| = n$. Then, there is a set $X' \subseteq X$ that is m-homogeneous with respect to A of size*

$$|X'| \geq \frac{\log^{(m)}(n)}{2^{O(m \log m)}}.$$

Lemma 9 allows us to focus on a large homogeneous set with respect to A. The next Lemma implies a lower bound in terms of the size of a homogeneous set. For simplicity and without loss of generality assume that the homogeneous set is $\{1, \ldots, k\}$.

Lemma 10 (Large homogeneous sets imply lower bounds for private learning)**.** *Let A be an $(0.1, \delta)$-differentially private algorithm with sample complexity m and $\delta \leq \frac{1}{10^3 m^2 \log m}$. Let $X = \{1, \ldots, k\}$ be m-homogeneous with respect to A. Then, if A empirically learns the class of thresholds over X with $(1/16, 1/16)$-accuracy, then*

$$k \leq 2^{O(m^2 \log^2 m)}$$

(i.e. $m \geq \Omega\left(\frac{\sqrt{\log k}}{\log \log k}\right)$).

We prove Lemma 9 and Lemma 10 in the following two subsections.

With these lemmas in hand, Theorem 1 follows by a short calculation: indeed, Lemma 9 implies the existence of an homogeneous set X' with respect to A of size $k \geq \log^{(m)}(n)/2^{O(m \log m)}$. We then restrict A to input samples from the set X', and by relabeling the elements of X' assume that $X' = \{1, \ldots, k\}$. Lemma 10 then implies that $k = 2^{O(m^2 \log^2 m)}$. Together we obtain that

$$\log^{(m)}(n) \leq 2^{c \cdot m^2 \log m}$$

for some constant $c > 0$. Applying the iterated logarithm $t = \log^*(2^{c \cdot m^2 \log m}) = \log^*(m) + O(1)$ times on the inequality yields that

$$\log^{(m+t)}(n) = \log^{(m + \log^*(m) + O(1))}(n) \leq 1,$$

and therefore $\log^*(n) \leq \log^*(m) + m + O(1)$, which implies that $m \geq \Omega(\log^* n)$ as required. □

4.3 Proof of Lemma 9

We next prove that every learning algorithm has a large homogeneous set. We will use the following quantitative version of Ramsey Theorem due to Erdős and Rado [1952] (see also the book [Graham et al., 1990], or Theorem 10.1 in the survey by Mubayi and Suk [2017]):

Theorem 11. *[Erdős and Rado, 1952] Let $s > t \geq 2$ and q be integers, and let*

$$N \geq \mathsf{twr}_t(3sq\log q).$$

Then for every coloring of the subsets of size t of a universe of size N using q colors there is a homogeneous subset[5] of size s.

Proof of Lemma 9. Define a coloring on the $(m+1)$-subsets of X as follows. Let $D = \{x_1 < x_2 < \ldots < x_{m+1}\}$ be an $(m+1)$-subset of X. For each $i \leq m+1$ let $D^{-i} = D \setminus \{x_i\}$, and let S^{-i} denote the balanced increasing sample on D^{-i}. Set p_i to be the fraction of the form $\frac{t}{10^2 m}$ that is closest to $A_{S^{-i}}(x_i)$ (in case of ties pick the smallest such fraction). The coloring assigned to A is the list $(p_1, p_2, \ldots, p_{m+1})$.

Thus, the total number of colors is $(10^2m+1)^{(m+1)}$. By applying Theorem 11 with $t := m+1, q := (10^2m+1)^{(m+1)}$, and $N := n$ there is a set $X' \subseteq X$ of size

$$|X'| \geq \frac{\log^{(m)}(n)}{3(10^2m+1)^{m+1}(m+1)\log(10^2m+1)} = \frac{\log^{(m)}(N)}{2^{O(m\log m)}}$$

such that all $m+1$-subsets of X' have the same color. One can verify that X' is indeed m-homogeneous with respect to A. ☐

4.4 Proof of Lemma 10

The lower bound is proven by using the algorithm A to construct a family of distributions $\mathcal{P}$ with certain properties, and use these properties to derive that $\Omega(k) \leq \mathcal{P} \leq 2^{O(m^2\log^2 m)}$, which implies the desired lower bound.

Lemma 12. *Let A, X', m, k as in Lemma 10, and set $n = k - m$. Then there exists a family $\mathcal{P} = \{P_i : i \leq n\}$ of distributions over $\{\pm 1\}^n$ with the following properties:*

1. *Every $P_i, P_j \in \mathcal{P}$ are $(0.1, \delta)$-indistinguishable.*
2. *There exists $r \in [0,1]$ such that for all $i, j \leq n$:*

$$\Pr_{v\sim P_i}\big[v(j) = 1\big] = \begin{cases} \leq r - \frac{1}{10m} & j < i, \\ \geq r + \frac{1}{10m} & j > i. \end{cases}$$

Lemma 13. *Let $\mathcal{P}, n, m, r$ as in Lemma 12. Then $n \leq 2^{10^3m^2\log^2 m}$.*

By the above lemmas, $k - m = |\mathcal{P}| \leq 2^{10^3m^2\log^2 m}$, which implies that $k = 2^{O(m^2\log^2 m)}$ as required. Thus, it remains to prove these lemmas, which we do next.

[5] A subset of the universe is homogeneous if all of its t-subsets have the same color.

4.4.1 Proof of Lemma 12

For the proof of lemma 12 we will need the following claim:

Claim 14. *Let $(p_i)_{i=0}^m$ denote the probabilities-list of X' with respect to A. Then for some $0 < i \leq m$:*

$$p_i - p_{i-1} \geq \frac{1}{4m}$$

Proof. The proof of this claim uses the assumption that A empirically learns thresholds. Let S be a balanced increasing realizable sample such that $S_X = \{x_1 < \ldots < x_m\} \subseteq X'$ are evenly spaced points on K (so, $S = (x_i, y_i)_{i=1}^m$, where $y_i = -1$ for $i \leq m/2$ and $y_i = +1$ for $i > m_2$).

A is an $(\alpha = 1/16, \beta = 1/16)$-empirical learner and therefore its expected empirical loss on S is at most $(1-\beta)\cdot\alpha + \beta \cdot 1 \leq \alpha + \beta = 1/8$, and so:

$$\begin{aligned}\frac{7}{8} &\leq \mathop{\mathbb{E}}_{h\sim A(S)}(1 - L_S(h)) \\ &= \frac{1}{m}\sum_{i=1}^{m/2}[1 - A_S(x_i)] + \frac{1}{m}\sum_{i=m/2+1}^{m}[A_S(x_i)]\,. \qquad \text{(since } S \text{ is balanced)}\end{aligned}$$

This implies that there is $m/2 \leq m_1 \leq m$ such that $A_S(x_{m_1}) \geq 3/4$. Next, by privacy if we consider S' the sample where we replace x_{m_1} by $x_{m_1} + 1$ (with the same label), we have that

$$A_{S'}(x_{m_1}) \geq \Big(\frac{3}{4} - \delta\Big)e^{-0.1} \geq \frac{2}{3}.$$

Note that $\mathrm{ord}_{S'}(x_{m_1}) = m_1 - 1$, hence by homogeneity: $p_{m_1-1} \geq \frac{2}{3} - \frac{1}{10^2 m}$. Similarly we can show that for some $1 \leq m_2 \leq \frac{m}{2}$ we have $p_{m_2-1} \leq \frac{1}{3} + \frac{1}{10^2 m}$. This implies that for some $m_2 - 1 \leq i \leq m_1 - 1$:

$$p_i - p_{i-1} \geq \frac{1/3}{m} - \frac{1}{50m^2} \geq \frac{1}{4m},$$

as required. □

Proof of Lemma 12. Let i be the index guaranteed by Claim 14 such that $p_i - p_{i-1} \geq 1/4m$. Pick an increasing realizable sample $S \in \big(X' \times \{\pm 1\}\big)^m$ so that the interval $J \subseteq X'$ between x_{i-1} and x_{i+1},

$$J = \big\{x \in \{1, \ldots, k\} : x_{i-1} < x < x_{i+1}\big\},$$

is of size $k - m$. For every $x \in J$ let S_x be the neighboring sample of S that is obtained by replacing x with x_i. This yields family of neighboring samples $\big\{S_x : x \in (x_{i-1}, x_{i+1})\big\}$ such that

- every two output-distributions $A(S_{x'})$, $A(S_{x''})$ are (ϵ, δ)-indistinguishable (because A satisfies (ϵ, δ) differential privacy).
- Set $r = \frac{p_{i+1}+p_i}{2}$. Then for all $x, x' \in J$:

$$\Pr_{h\sim A(S_x)}\big[h(x') = 1\big] = \begin{cases} \leq r - \frac{1}{10m} & x' < x, \\ \geq r + \frac{1}{10m} & x' > x. \end{cases}$$

The proof is concluded by restricting the output of A to J, and identifying J with $[n]$ and each output-distributions $A(S_x)$ with a distribution over $\{\pm 1\}^n$.

□

4.4.2 Proof of Lemma 13

Proof. Set $T = 10^3 m^2 \log^2 m - 1$, and $D = 10^2 m^2 \log T$. We want to show that $n \leq 2^{T+1}$. Assume towards contradiction that $n > 2^{T+1}$. Consider the family of distributions $Q_i = P_i^D$ for $i = 1, \ldots, n$. By Lemma 5, each Q_i, Q_j are $(0.1D, \delta D)$-indistinguishable.

We next define a set of mutually disjoint events E_i for $i \leq 2^T$ that are measurable with respect to each of the Q_i's. For a sequence of vectors $\mathbf{v} = (v_1, \ldots, v_D)$ in $\{\pm 1\}^n$ we let $\bar{\mathbf{v}} \in \{\pm 1\}^n$ be the threshold vector defined by

$$\bar{\mathbf{v}}(j) = \begin{cases} -1 & \frac{1}{D}\sum_{i=1}^{D} v_i(j) \leq r, \\ +1 & \frac{1}{D}\sum_{i=1}^{D} v_i(j) \geq r. \end{cases}$$

Given a point in the support of any of the Q_i's, namely a sequence $\mathbf{v} = (v_1, \ldots, v_D)$ of D vectors in $\{\pm 1\}^n$ define a mapping B according to the outcome of T steps of binary search on $\bar{\mathbf{v}}$ as follows: probe the $\frac{n}{2}$'th entry of $\bar{\mathbf{v}}$; if it is $+1$ then continue recursively with the first half of $\bar{\mathbf{v}}$. Else, continue recursively with the second half of $\bar{\mathbf{v}}$. Define the mapping $B = B(\mathbf{v})$ to be the entry that was probed at the T'th step. The events E_j correspond to the 2^T different outcomes of B. These events are mutually disjoint by the assumption that $n > 2^{T+1}$.

Notice that for any possible i in the image of B, applying the binary search on a sufficiently large i.i.d sample $\mathbf{v}$ from P_i would yield $B(\mathbf{v}) = i$ with high probability. Quantitatively, a standard application of Chernoff inequality and a union bound imply that the event $E_i = \{\mathbf{v} : B(\bar{\mathbf{v}}) = i\}$ for $\mathbf{v} \sim Q_i$, has probability at least

$$1 - T\exp\Big(-2\frac{1}{10^2 m^2}D\Big) = 1 - T\exp(-2\log T) \geq \frac{2}{3}.$$

We claim that for all $j \leq n$, and i in the image of B:

$$Q_j(E_i) \geq \frac{1}{2}\exp(-0.1D). \tag{1}$$

This will finish the proof since the 2^T events are mutually disjoint, and therefore

$$\begin{aligned} 1 &\geq Q_j(\cup_i E_i) \\ &= \sum_i Q_j(E_i) \\ &\geq 2^T \cdot \frac{1}{2} e^{-0.1D} \\ &= 2^{T-1} e^{-0.1D}, \end{aligned}$$

however, $2^{T-1}e^{-0.1D} > 1$ by the choice of T, D, which is a contradiction.

Thus it remains to prove Equation (1). This follows since Q_i, Q_j are $(0.1D, D\delta)$-indistinguishable:

$$\frac{2}{3} \leq Q_i(E_i) \leq \exp(0.1D)Q_j(E_i) + D\delta,$$

and by the choice of δ, which implies that $\frac{2}{3} - D\delta \geq \frac{1}{2}$. □

5 Privately learnable classes have finite Littlestone dimension

We conclude the paper by deriving Corollary 2 that gives a lower bound of $\Omega(\log^* d)$ on the sample complexity of privately learning a class with Littlestone dimension d.

Proof of Corollary 2. The proof is a direct corollary of Theorem 3 and Theorem 1. Indeed, let H be a class with Littlestone dimension d, and let $c = \lfloor \log d \rfloor$. By Item 1 of Theorem 3, there are $x_1, \ldots, x_c$ and $h_1, \ldots, h_c \in H$ such that $h_i(x_j) = +1$ if and only if $j \geq i$. Theorem 1 implies a lower bound of $m \geq \Omega(\log^* c) = \Omega(\log^* d)$ for any algorithm that learns $\{h_i : i \leq c\}$ with accuracy $(1/16, 1/16)$ and privacy $(0.1, O(1/m^2 \log m))$. □

6 Conclusion

The main result of this paper is a lower bound on the sample complexity of private learning in terms of the Littlestone dimension.

We conclude with an open problem. There are many mathematically interesting classes with finite Littlestone dimension, see e.g. [Chase and Freitag, 2018]. It is natural to ask whether the converse to our main result holds, i.e. whether every class with finite Littlestone dimension may be learned privately.

Acknowledgements

We thank Raef Bassily and Mark Bun for many insightful discussions and suggestions regarding an earlier draft of this manuscript. We also thank the anonymous reviewers for helping improving the presentation of this paper.

References

Maria-Florina Balcan and Vitaly Feldman. Statistical active learning algorithms for noise tolerance and differential privacy. *Algorithmica*, 72(1):282–315, 2015.

Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *STOC*, pages 1046–1059. ACM, 2016.

Raef Bassily, Om Thakkar, and Abhradeep Thakurta. Model-agnostic private learning via stability. *CoRR*, abs/1803.05101, 2018.

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

Amos Beimel, Hai Brenner, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. *Machine Learning*, 94(3): 401–437, 2014.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Learning privately with labeled and unlabeled examples. In *SODA*, pages 461–477. SIAM, 2015.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. *Theory of Computing*, 12(1):1–61, 2016.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, 2009.

Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In *PODS*, pages 128–138. ACM, 2005.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989. ISSN 0004-5411. doi: 10.1145/76359.76371.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil P. Vadhan. Differentially private release and learning of threshold functions. In *FOCS*, pages 634–649. IEEE Computer Society, 2015.

Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. In *ITCS*, pages 369–380. ACM, 2016.

Mark Mar Bun. *New Separations in the Complexity of Differential Privacy*. PhD thesis, Harvard University, Graduate School of Arts & Sciences, 2016.

Hunter Chase and James Freitag. Model theory and machine learning. *arXiv preprint arXiv:1801.06566*, 2018.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

Kamalika Chaudhuri, Daniel J. Hsu, and Shuang Song. The large margin mechanism for differentially private maximization. In *NIPS*, pages 1287–1295, 2014.

Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive learning with robust generalization guarantees. In *COLT*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 772–814. JMLR.org, 2016.

Cynthia Dwork and Vitaly Feldman. Privacy-preserving prediction. *CoRR*, abs/1803.10266, 2018.

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

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, volume 4004 of *Lecture Notes in Computer Science*, pages 486–503. Springer, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006b.

P. Erdős and R. Rado. Combinatorial theorems on classifications of subsets of a given set. *Proceedings of the London Mathematical Society*, s3-2(1):417–439, 1952. doi: 10.1112/plms/s3-2.1.417. URL `https://londmathsoc.onlinelibrary.wiley.com/doi/abs/10.1112/plms/s3-2.1.417`.

Vitaly Feldman and David Xiao. Sample complexity bounds on differentially private learning via communication complexity. *SIAM J. Comput.*, 44(6):1740–1764, 2015.

R.L. Graham, R.L. Graham, B.L. Rothschild, and J.H. Spencer. *Ramsey Theory*. A Wiley-Interscience publication. Wiley, 1990. ISBN 9780471500469. URL `https://books.google.com/books?id=55oXT60dC54C`.

Wilfrid Hodges. *A Shorter Model Theory*. Cambridge University Press, New York, NY, USA, 1997. ISBN 0-521-58713-1.

Marek Karpinski and Angus Macintyre. Polynomial bounds for vc dimension of sigmoidal and general pfaffian neural networks. *Journal of Computer and System Sciences*, 54(1): 169–176, 1997.

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

Michael C Laskowski. Vapnik-chervonenkis classes of definable sets. *Journal of the London Mathematical Society*, 2(2):377–384, 1992.

Katrina Ligett, Seth Neel, Aaron Roth, Bo Waggoner, and Steven Z. Wu. Accuracy first: Selecting a differential privacy level for accuracy constrained ERM. In *NIPS*, pages 2563–2573, 2017.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1987.

Roi Livni and Pierre Simon. Honest compressions and their application to compression schemes. In *Conference on Learning Theory*, pages 77–92, 2013.

D. Mubayi and A. Suk. A survey of quantitative bounds for hypergraph Ramsey problems. *ArXiv e-prints*, July 2017.

Benjamin I. P. Rubinstein, Peter L. Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. *CoRR*, abs/0911.5708, 2009.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014. ISBN 1107057132, 9781107057135.

Saharon. Shelah. *Classification theory and the number of non-isomorphic models*. North-Holland Pub. Co. ; sole distributors for the U.S.A. and Canada, Elsevier/North-Holland Amsterdam ; New York : New York, 1978. ISBN 0720407575.

Salil P. Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer International Publishing, 2017.

V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971. ISSN 0040-585X; 1095-7219/e. doi: 10.1137/1116025.

Yu-Xiang Wang, Jing Lei, and Stephen E. Fienberg. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of ERM principle. *Journal of Machine Learning Research*, 17:183:1–183:40, 2016. URL `http://jmlr.org/papers/v17/15-313.html`.

A Proof of Lemma 5

The theorem follows by induction from the following lemma.

Lemma 15. *Let p_1, q_1 be distributions over a countable domain X_1 and p_2, q_2 be distributions over a countable domain X_2. Assume that p_i, q_i are (ϵ_i, δ_i)-indistinguishable for $i = 1, 2$. Then $p_1 \times p_2, q_1 \times q_2$ are $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-indistinguishable.*

Proof. Let $a \wedge b$ denote $\min\{a, b\}$. For $x \in X_i$ let

$$\Delta_i(x) = \begin{cases} p_i(x) - e^{\epsilon} q_i(x) & p_i(x) - e^{\epsilon} q_i(x) \geq 0, \\ 0 & p_i(x) - e^{\epsilon} q_i(x) < 0. \end{cases}$$

Extend Δ_1, Δ_2 to be a measure on X in the obvious way. Note that

- $\Delta_i(X) \leq \delta_i$, and that
- $p_i(x) \leq e^{\epsilon} q_i(x) + \Delta_i(x)$ for all $x \in X_i$.

Let $S \subseteq X_1 \times X_2$. We show that $(p_1 \times p_2)\big(S\big) \leq e^{\epsilon_1 + \epsilon_2}\big(q_1 \times q_2\big)(S) + \delta_1 + \delta_2$, the other direction can be derived similarly.

For $a \in X_1$ let $S_a \subseteq X_2$ denote the set $\{b : (a, b) \in S\}$

$$
\begin{aligned}
(p_1 \times p_2)(S) &= \sum_{a \in X_1} p_1(a) p_2(S_a) \\
&\leq \sum_{a \in X_1} p_1(a) \Big(\big(1 \wedge e^{\epsilon_2} q_2(S_a)\big) + \delta_2 \Big) \\
&\leq \delta_2 + \sum_{a \in X_1} p_1(a) \big(1 \wedge e^{\epsilon_2} q_2(S_a)\big) \\
&\leq \delta_2 + \sum_{a \in X_1} \big(e^{\epsilon_1} q_1(a) + \Delta_1(a)\big) \big(1 \wedge e^{\epsilon_2} q_2(S_a)\big) \\
&\leq \delta_2 + \sum_{a \in X_1} e^{\epsilon_1} q_1(a) e^{\epsilon_2} q_2(S_a) + \Delta_1(a) \\
&= \delta_2 + \Delta_1(X) + e^{\epsilon_1 + \epsilon_2} \big(q_1 \times q_2\big)(S) \\
&\leq e^{\epsilon_1 + \epsilon_2} \big(q_1 \times q_2\big)(S) + \delta_1 + \delta_2.
\end{aligned}
$$

□

B Proof of Theorem 3

In this appendix we prove Theorem 3. Throughout the proof a labeled binary tree means a full binary tree whose internal vertices are labeled by instances.

The second part of the theorem is easy. If $\mathcal{H}$ contains 2^t thresholds then there are $h_i \in \mathcal{H}$ for $0 \leq i < 2^t$ and there are x_j for $0 \leq j < 2^t - 1$ such that $h_i(x_j) = 0$ for $j < i$ and $h_i(x_j) = 1$ for $j \geq i$. Define a labeled binary tree of height t corresponding to the binary search process. That is, the root is labeled by $x_{2^{t-1}-1}$, its left child by $x_{2^{t-1}+2^{t-2}-1}$ and its right child by $x_{2^{t-1}-2^{t-2}-1}$ and so on. If the label of an internal vertex of distance q from the root, where $0 \leq q \leq t-1$, is x_p, then the label of its left child is $x_{p+2^{t-q-1}}$ and the label of its right child is $x_{p-2^{t-q-1}}$. It is easy to check that the root-to-leaf path corresponding to each of the functions h_i leads to leaf number i from the right among the leaves of the tree (counting from 0 to $2^t - 1$).

To prove the first part of the theorem we first define the notion of a subtree T' of depth h of a labeled binary tree T by induction on h. Any leaf of T is a subtree of height 0. For $h \geq 1$ a subtree of height h is obtained from an internal vertex of T together with a subtree of height $h-1$ of the tree rooted at its left child and a subtree of height $h-1$ of the tree rooted at its right child. Note that if T is a labeled tree and it is shattered by the class $\mathcal{H}$, then any subtree T' of it with the same labeling of its internal vertices is shattered by the class $\mathcal{H}$. With this definition we prove the following simple lemma.

Lemma 16. *Let p, q be positive integers and let T be a labeled binary tree of height $p+q-1$ whose internal vertices are colored by two colors, red and blue. Then T contains either a subtree of height p in which all internal vertices are red (a red subtree), or a subtree of height q in which all vertices are blue (a blue subtree).*

Proof: We apply induction on $p+q$. The result is trivial for $p = q = 1$ as the root of T is either red or blue. Assuming the assertion holds for $p' + q' < p + q$, let T be of height

$p+q-1$. Without loss of generality assume the root of t is red. If $p=1$ we are done, as the root together with a leaf in the subtree of its left child and one in the subtree of its right child form a red subtree of height p. If $p>1$ then, by the induction hypothesis, the tree rooted at the left child of the root of T contains either a red subtree of height $p-1$ or a blue subtree of height q, and the same applies to the tree rooted at the right child of the root. If at least one of them contains a blue subtree as above we are done, otherwise, the two red subtrees together with the root provide the required red subtree. $\square$

We can now prove the first part of the theorem, showing that if the Littlestone dimension of $\mathcal{H}$ is at least $2^{t+1}-1$ then $\mathcal{H}$ contains $t+2$ thresholds. We apply induction on t. If $t=0$ we have a tree of height 1 shattered by $\mathcal{H}$. Its root is labeled by some variable x_0 and as it is shattered there are two functions $h_0, h_1 \in \mathcal{H}$ so that $h_0(x_0)=1, h_1(x_0)=0$, meaning that $\mathcal{H}$ contains two thresholds, as needed. Assuming the desired result holds for $t-1$ we prove it for $t, t \geq 1$. Let T be a labeled binary tree of height $2^{t+1}-1$ shattered by $\mathcal{H}$. Let h be an arbitrary member of $\mathcal{H}$ and define a two coloring of the internal vertices of T as follows. If an internal vertex is labeled by x and $h(x)=1$ color it red, else color it blue. Since $2^{t+1}-1=2\cdot 2^t-1$, Lemma 16 with $p=q=2^t$ implies that T contains either a red or a blue subtree T' of height 2^t. In the first case define $h_0=h$ and let X be the set of all variables x so that $h(x)=1$. Let x_0 be the root of T' and let T'' be the subtree of T' rooted at the left child of T'. Let $\mathcal{H}'$ be the set of all $h' \in \mathcal{H}$ so that $h'(x_0)=0$. Note that $\mathcal{H}'$ shatters the tree T'', and that the depth of T'' is 2^t-1. We can thus apply the induction hypothesis and get a set of $t+1$ thresholds $h_1, h_2, \ldots, h_{t+1} \in \mathcal{H}'$ and variables $x_1, x_2, \ldots, x_t \in X$ so that $h_i(x_j)=1$ iff $j \geq i$. Adding h_0 and x_0 to these we get the desired $t+2$ thresholds.

Similarly, if T contains a blue subtree T', define $h_{t+1}=h$ and let X be the set of all variables x so that $h(x)=0$. In this case denote the root of T' by x_t and let T'' be the subtree of T' rooted at the right child of T'. Let $\mathcal{H}'$ be the set of all $h' \in \mathcal{H}$ so that $h'(x_t)=1$. As before, $\mathcal{H}'$ shatters the tree T'' whose depth is 2^t-1. By the induction hypothesis we get $t+1$ thresholds $h_0, h_1, \ldots, h_t$ and variables $x_0, x_1, \ldots, x_{t-1} \in X$ so that $h_i(x_j)=1$ iff $j \geq i$, and the desired result follows by appending to them h_{t+1} and x_t. This completes the proof. $\square$