An Õptimal Differentially Private PAC Learner for Concept Classes with VC Dimension 1

Chao Yan[*]

May 13, 2025

Abstract

We present the first nearly optimal differentially private PAC learner for any concept class with VC dimension 1 and Littlestone dimension d. Our algorithm achieves the sample complexity of $\tilde{O}_{\varepsilon,\delta,\alpha,\delta}(\log^* d)$, nearly matching the lower bound of $\Omega(\log^* d)$ proved by Alon et al. [1] [STOC19]. Prior to our work, the best known upper bound is $\tilde{O}(VC \cdot d^5)$ for general VC classes, as shown by Ghazi et al. [13] [STOC21].

1 Introduction

Differentially private learning, introduced by Kasiviswanathan et al. [17], studies the task of learning a hypothesis from data while preserving the privacy of individual entries in the dataset. Formally, the goal is to construct a learner that satisfies the requirements of PAC learning and, simultaneously, differential privacy.

Definition 1 (PAC learning [21]). *Given a set of n points $S = (\mathcal{X} \times \{0,1\})^n$ drawn i.i.d. from an unknown distribution $\mathcal{D}$ and labels that are given by an unknown concept $c \in \mathcal{C}$, we say a learner L (possibly randomized) (α, β)-PAC learns $\mathcal{C}$ if $h = L(S)$ and*

$$\Pr[error_{\mathcal{D}}(c, h) \leq \alpha] \geq 1 - \beta.$$

Definition 2 (differential privacy [11]). *A randomized algorithm M is called (ε, δ)-differentially private if for any two dataset S and S' that differ on one entry and any event E, it holds that*

$$\Pr[M(S) \in E] \leq e^{\varepsilon} \cdot \Pr[M(S') \in E] + \delta.$$

Specifically, when $\delta = 0$, we call it pure-differential privacy. When $\delta > 0$, we call it approximate-differential privacy

Definition 3 (differentially private learning [17]). *We say a learner L $(\alpha, \beta, \varepsilon, \delta)$-differentially privately PAC learns the concept class $\mathcal{C}$ if*

1. *L (α, β)-PAC learns $\mathcal{C}$.*

2. *L is (ε, δ)-differentially private.*

We call the dataset size of the learning task the *sample complexity*, which is a fundamental question in learning theory. For non-private PAC learning, it is well-known that the sample complexity is linear to the VC dimension of the concept class [21]. However, the sample complexity is less well understood in the differentially private setting. Kasiviswanathan et al. [17] define the private learning and give a general upper

bound $O(\log|\mathcal{C}|)$, which works for pure differential privacy ($\delta = 0$). This bound is tight for several natural concept classes [2, 12].

In the approximate privacy regime ($\delta > 0$), several works [4, 8, 3, 16, 15, 9, 19] show that the sample complexity can be significantly lower than that in the pure setting. Alon et al. [1] and Bun et al. [7] find that the sample complexity of approximately differentially private learning is related to the Littlestone dimension [18] of the concept class $\mathcal{C}$. In detail, for a concept class $\mathcal{C}$ with Littlestone dimension d, Alon et al. [1] prove a lower bound $\Omega(\log^* d)$ and Bun et al. [7] provide an upper bound $\tilde{O}(2^{2^d})$. Subsequently, Ghazi et al. [13] improve the upper bound to $\tilde{O}(VC \cdot d^5)$.

This leaves a large gap between the lower bound $\Omega(VC + \log^* d)$ and the upper bound $\tilde{O}(VC \cdot d^5)$, even when the VC dimension is as small as 1. The main question is: what is the correct dependence on d? One important example is the halfspaces class, which can be privately learned with $poly(VC, \log^* d)$ examples by the work of Nissim et al. [19]. So it's natural to ask:

> Could we privately learn any concept class with sample size $poly(VC, \log^* d)$?

Unfortunately, the technique of Nissim et al. [19] cannot extends to the general VC class because it depends on the data structure of halfspaces. However, the work by Nissim et al. [19] also shows an upper bound of $\tilde{O}(\log^*|\mathcal{X}|)$ for any VC 1 class. Although $|\mathcal{X}|$ can be infinitely larger than d for the general concept class, and a significant gap remained between $\log^* d$ and $\min\{d^5, \log^*|\mathcal{X}|\}$, it shows that in some cases (say when $\log|\mathcal{X}| = d$), the sample complexity of of private learning depends on $\log^* d$.

To the target of fully understand the sample complexity of differentially private learning, the first step is to ask:

> Could we privately learn VC 1 class with sample size $poly(\log^* d)$?

In this work, we give a positive answer to this question and give a nearly tight bound $\tilde{\theta}_{\alpha,\beta,\varepsilon,\delta}(\log^* d)$.

1.1 Our result

Theorem 1. *For any concept class $\mathcal{C}$ with VC dimension 1 and Littlestone dimension d, there is an (ε, δ)-differentially private algorithm that (α, β)-PAC learns $\mathcal{C}$ if the given labeled dataset has size*

$$N \geq O\left(\frac{\log^* d \cdot \log^2(\frac{\log^* d}{\varepsilon\beta\delta})}{\varepsilon} \cdot \frac{48}{\alpha}\left(10\log(\frac{48e}{\alpha}) + \log(\frac{5}{\beta})\right)\right) = \tilde{O}_{\beta,\delta}\left(\frac{\log^* d}{\alpha\varepsilon}\right)$$

1.2 Overview of Technique

The key observation is that any concept with VC dimension 1 has a tree structure, which is observed by Ben-David [5]. In the tree structure, each node is a point from the domain $\mathcal{X}$, and each hypothesis is a path from one node to the tree's root. We show that the tree's height is upper bounded by the threshold dimension of the concept class, which has an upper bound $O(2^d)$ [20, 14, 1].

We use the partition and aggregate method to construct the private learner. The labeled dataset S is randomly partitioned into subsets $S_1, \ldots, S_t$, where $t = O(\log^* d)$. Each subset has a set of "deterministic points", whose labels are fixed to be 1 by the given labeled points. We show that the deterministic points can be used to construct an accurate hypothesis, and all the sets of deterministic points of $S_1, \ldots, S_t$ are on the same path. Since the length of the path is at most $O(2^d)$, we use the private median algorithm [9] to select a "good length" with sample size $\tilde{O}(\log^* d)$. Then we can use the choosing mechanism [4] to select a "good path" with the "good length" with a sample size $O(1)$ because the deterministic points of $S_1, \ldots, S_t$ are on the same path. Finally, we show that the selected "good path" is an accurate hypothesis.

2

Figure 1: Example of tree structure

1.2.1 Example

Here we give an example. Let $\mathcal{X} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$. Let $I(f) = \{x : f(x) = 1\}$. Then $\mathcal{H} = \{h_1, h_2, h_3, h_4, h_5, h_6, h_7, h_8\}$, where $I(h_1) = \{x_1\}$, $I(h_2) = \{x_2\}$, $I(h_3) = \{x_3\}$, $I(h_4) = \{x_1, x_4\}$, $I(h_5) = \{x_1, x_5\}$, $I(h_6) = \{x_1, x_5, x_6\}$, $I(h_7) = \{x_1, x_5, x_7\}$, $I(h_8) = \emptyset$. The tree structure is shown in Figure 1. In the tree structure, each concept can be represented by a path from one node to the root (For example, h_5 is the path $x_5 \to x_1 \to \emptyset$. That means h_5 gives x_1, x_5 label 1 and gives other points label 0.). The tree structure has four layers. The first layer has $\emptyset$, the second layer has x_1, x_2, x_3, the third layer has x_4, x_5, the fourth layer has x_6, x_7. We will show that the layer number cannot be "too large" if the Littlestone dimension is bounded.

Let the underlying concept be h_7, for all subsets of the given dataset, the deterministic points set can only be $\emptyset$ or $\{x_1\}$ or $\{x_1, x_5\}$ or $\{x_1, x_5, x_7\}$. For instance, if we have S_1, S_2 and their deterministic points are $\{x_1\}$ and $\{x_1, x_5, x_7\}$. We record the maximum number of layers of deterministic points, which are 2 and 4. We can privately output the median of the layer numbers, say it is 3. Then we only consider the points on the third layer, which are x_4 and x_5. Notice that x_4 will never be the deterministic point, and approximately half of the subsets will make x_5 the deterministic point (because layer 3 is the median layer). Then we can use the choosing mechanism to select x_5. We consider the corresponding path to the root and its hypothesis h_5. Notice that if S_1 and S_2 contain enough points, we can show that h_1 and h_7 have high accuracy by VC theory. Then we can show h_5 also has high accuracy because it is in the middle of h_1 and h_7 (see Figure 2).

Figure 2: Example of output hypothesis

3

2 Notations

In learning theory, we call $\mathcal{X}$ a domain. The element $x \in \mathcal{X}$ is the point. The concept c is a function $c : \mathcal{X} \to \{0, 1\}$. The concept class $\mathcal{C}$ is a set of concepts. We call a learned function h a hypothesis. For any function f, we denote $I(f) = \{x : f(x) = 1\}$

3 Preliminary

3.1 Learning theory

Definition 4 (Error). *Let $\mathcal{D}$ be a distribution, c be a concept and h be a hypothesis. The error of h over $\mathcal{D}$ is defined as*

$$error_{\mathcal{D}}(c, h) = \Pr_{x \sim \mathcal{D}}[c(x) \neq h(x)].$$

For a finite set S, the error of h over S is defined as

$$error_S(c, h) = \frac{|\{x \in S : c(x) \neq h(x)\}|}{|S|}.$$

We write it as $error_S(h)$ because $c(x)$ is given as the label of x in S.

Definition 5 (PAC Learning [21]). *Given a set of n points $S = (\mathcal{X} \times \{0, 1\})^n$ sampled from distribution $\mathcal{D}$ and labels that are given by an underlying concept c, we say a learner L (L could be randomized) (α, β)-PAC learns $\mathcal{C}$ if $h = L(S)$ and*

$$\Pr[error_{\mathcal{D}}(c, h) \leq \alpha] \geq 1 - \beta.$$

Definition 6 (VC Dimension[22]). *For a domain $\mathcal{X}$ and a concept class $\mathcal{C}$, We say $x_1, \ldots, x_k$ is* shattered *if for any subset $S \subseteq \{x_1, \ldots, x_k\}$, there exists a concept $c \in \mathcal{C}$, such that $c(x) = 1$ for any $x \in S$ and $c(x) = 0$ for any $x \in \{x_1, \ldots, x_k\} \backslash S$. The maximum size of the shattered set is called* VC dimension.

Theorem 2 ([6]). *Let $\mathcal{C}$ be a concept class, and let $\mathcal{D}$ be a distribution over the domain $\mathcal{C}$. Let $\alpha, \beta > 0$, and $m \geq \frac{48}{\alpha}\left(10VC(\mathcal{X}, \mathcal{C})\log(\frac{48e}{\alpha}) + \log(\frac{5}{\beta})\right)$. Let S be a sample of m points drawn i.i.d. from $\mathcal{D}$. Then*

$$\Pr[\exists c, h \in \mathcal{C} \ s.t. \ error_S(c, h) \leq \alpha/10 \ and \ error_{\mathcal{D}}(c, h) \geq \alpha] \leq \beta.$$

Definition 7 (Thresholds Dimension). *For $x_1, \ldots, x_k \in \mathcal{X}$ and $c_1, \ldots, c_k \in \mathcal{C}$, if for any $i \in [k]$, we have $c_i(x_j) = 1$ for all $j \geq i$ and $c_i(x_j) = 0$ for all $j < i$, then we call $((x_1, \ldots, x_k), (c_1, \ldots, c_k))$ as a class of thresholds. The* thresholds dimension *$TD(\mathcal{X}, \mathcal{C}) = argmax_k\{\exists x_1, \ldots, x_k \in \mathcal{X}, c_1, \ldots, c_k \in \mathcal{C}, ((x_1, \ldots, x_k), (c_1, \ldots, c_k)) \ is \ a \ thresholds \ class\}$, i.e. the length of the longest thresholds in $(\mathcal{X}, \mathcal{C})$.*

Definition 8 (Online Learning [18]). *In the ith turn of the online learning setting, the learner receives a data point x_i and predicts the label of x_i. Then the learner receives the true label of x_i.*

Definition 9 (Littlestone Dimension [18]). *In online learning, we say the learner makes a mistake if the learner's prediction is different from the true label in one turn. We say a learner is optimal if the learner can make the minimum number of mistakes when the learner outputs the true concept. The maximum number of mistakes the optimal learner makes is called the Littlestone dimension of $(\mathcal{X}, \mathcal{C})$. We denote it as $d_L(\mathcal{X}, \mathcal{C})$.*

Theorem 3. *[20, 14, 1] $\lfloor \log d_L(\mathcal{X}, \mathcal{C}) \rfloor \leq TD(\mathcal{X}, \mathcal{C}) \leq 2^{d_L(\mathcal{X}, \mathcal{C})+1}$.*

Corollary 1. $O(\log^*(TD(\mathcal{X}, \mathcal{C}))) = O(\log^*(d_L(\mathcal{X}, \mathcal{C})))$

3.2 Differential privacy

Definition 10 (Differential Privacy [11]). *A mechanism M is called (ε, δ)-differentially private if for any two dataset S and S' that differ on one entry and any event E, it holds that*

$$\Pr[M(S) \in E] \leq e^\varepsilon \cdot \Pr[M(S') \in E] + \delta.$$

Definition 11 (Differentially Private PAC Learning [17]). *Given a set of n points $S = (\mathcal{X} \times \{0, 1\})^n$ sampled from distribution $\mathcal{D}$ and labels that are given by an underlying concept c, we say a learner L $(\alpha, \beta, \varepsilon, \delta)$-differentially privately PAC learns the concept class $\mathcal{C}$ if*

1. *L (α, β)-PAC learns $\mathcal{C}$.*

2. *L is (ε, δ)-differentially private.*

Theorem 4 (Post-processing [11]). *For any (ε, δ)-differentially private mechanism M and any function A (A could be randomized), the mechanism $A \circ M$ is (ε, δ)-differentially private.*

Theorem 5 (Composition [10]). *For an $(\varepsilon_1, \delta_1)$-differentially private mechanism M_1 and an $(\varepsilon_2, \delta_2)$-differentially private mechanism M_2, the composed mechanism $M(X) = (M_1(X), M_2(X))$ is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-differentially private.*

Definition 12 (α-median). *For a set of number $S = \{x_1, \ldots, x_n\}$, we say a number $\hat{x}$ is an α-median of S if $\min\{|\{x : x \leq \hat{x}, x \in S\}|, |\{x : x \geq \hat{x}, x \in S\}|\} \geq (1/2 - \alpha) \cdot |S|$*

Fact 1. *For any set of number $S = \{x_1, \ldots, x_n\}$, there exists a $1/2$-median.*

Theorem 6 ([9]). *Let $\mathcal{X}$ be a finite ordered domain. There exists an (ε, δ)-differentially private algorithm PrivateMedian[1] that on input $S \in \mathcal{X}^n$ outputs an α-median point with probability $1 - \beta$ provided that $n > n_{PM}(|\mathcal{X}|, \alpha, \beta, \varepsilon, \delta)$ for $n_{PM}(|\mathcal{X}|, \alpha, \beta, \varepsilon, \delta) \in O\left(\frac{\log^* |\mathcal{X}| \cdot \log^2(\frac{\log^* |\mathcal{X}|}{\beta\delta})}{\alpha\varepsilon}\right)$.*

Definition 13 (k-bounded function). *We call a quality function $q : X^* \times Z \to \mathbb{R}$ is k-bounded if adding a new element to the data set can only increase the score of at most k solutions, Specifically, it holds that*

1. *$q(\emptyset, z) = 0$ for every $z \in Z$.*

2. *If $D' = D \cup \{x\}$, then $q(D', z) \in \{q(D, z), q(D, z) + 1\}$ for every $z \in Z$, and*

3. *There are at most k solutions z such that $q(D', z) = q(D, z) + 1$*

Lemma 1 (Choosing Mechanism [4]). *Let $\varepsilon \in (0, 2)$ and $\delta > 0$. Let $q : X^* \times Z \to \mathbb{R}$ be a k-bounded quality function. There is an (ε, δ)-DP algorithm $\mathcal{A}$, such that given a dataset $D \in X^n$, $\mathcal{A}$ outputs a solution z and*

$$\Pr[q(D, z) \geq \max_{z \in Z}\{q(D, z)\} - \frac{16}{\varepsilon} \log(\frac{4kn}{\beta\varepsilon\delta})] \geq 1 - \beta$$

4 Structure of Classes with VC Dimension 1

Without loss of generality, we have the following two assumptions. So that every point in the domain $\mathcal{X}$ is different and every concept in $\mathcal{C}$ is different.

Assumption 1. *Assume for any two different points $x_1, x_2 \in \mathcal{X}$, there exists a concept $c \in \mathcal{C}$ makes $c(x_1) \neq c(x_2)$. Otherwise, we can replace all x_2 by x_1 in the given dataset and remove x_2 from $\mathcal{X}$.*

---

[1]In [9], they provide an algorithm that can privately select interior point. That is given $x_1, \ldots, x_k$, the algorithm privately outputs a number $\hat{x}$ satisfying $\min\{x_i\} \leq \hat{x} \leq \max\{x_i\}$. It can be extended to the median by removing the smallest and largest $(1/2 - \alpha/2)$ fraction of numbers. This reduction is found by Bun et al. [8].

Assumption 2. *Assume for any point x, there exists two different $c_1, c_2 \in \mathcal{C}$ make $c_1(x) \neq c_2(x)$. Otherwise, we can remove c_2 from $\mathcal{C}$.*

Definition 14 (Partial Order). *Given $\mathcal{X}, \mathcal{C}$ and $x_1, x_2 \in \mathcal{X}$, we say $x_1 \preceq x_2$ under $\mathcal{C}$ if for all $c \in \mathcal{C}$, $(c(x_1) = 1) \Rightarrow (c(x_2) = 1)$*

Example 1. *For thresholds $((x_1, \ldots, x_k), (c_1, \ldots, c_k))$, we have $x_1 \preceq \cdots \preceq x_k$.*

We say x_1 and x_2 are comparable under $\mathcal{C}$ if $x_1 \preceq x_2$ or $x_2 \preceq x_1$ under $\mathcal{C}$.

Definition 15 (f-representation). *For function $c, f : \mathcal{X} \to \{0, 1\}$, the f-representation of c is*

$$c_f(x) = \begin{cases} 1 & f(x) \neq c(x) \\ 0 & f(x) = c(x) \end{cases}$$

For the class of function $\mathcal{C}$, the f-representation of $\mathcal{C}$ is $\mathcal{C}_f = \{c_f : c \in \mathcal{C}\}$.

Given f and c_f, we can transform c_f to f:

$$c(x) = \begin{cases} 1 & f(x) \neq c_f(x) \\ 0 & f(x) = c_f(x) \end{cases}$$

Given the pair of the point and label $(x, c(x))$ that is labeled by c, we can transform it to a corresponding pair labeled by c_f: let label be 1 if $c(x) \neq f(x)$ and be 0 if $c(x) = f(x)$.

Lemma 2. $VC(\mathcal{X}, \mathcal{C}) = VC(\mathcal{X}, \mathcal{C}_f)$ *and* $d_L(\mathcal{X}, \mathcal{C}) = d_L(\mathcal{X}, \mathcal{C}_f)$ *for any f.*

Proof. Since $\mathcal{C}$ is the f-representation of $\mathcal{C}_f$, we only consider one direction.

1. **VC dimension**. Let $x_1, \ldots, x_{VC(\mathcal{X}, \mathcal{C})}$ be a set of points shattered by $\mathcal{C}$. For any set of dichotomy $(c(x_1), \ldots, c(x_{VC(\mathcal{X}, \mathcal{C})}))$, there exists a concept $c' \in \mathcal{C}$ makes $(c'(x_1), \ldots, c'(x_{VC(\mathcal{X}, \mathcal{C})})) = ((c_f(x_1), \ldots, c_f(x_{VC(\mathcal{X}, \mathcal{C})})))$ because $x_1, \ldots, x_{VC(\mathcal{X}, \mathcal{C})}$ are shattered. Thus, the corresponding $c'_f \in \mathcal{C}_f$ makes $\left(c'_f(x_1), \ldots, c'_f(x_{VC(\mathcal{X}, \mathcal{C})})\right) = ((c(x_1), \ldots, c(x_{VC(\mathcal{X}, \mathcal{C})})))$. So that all $2^{VC(\mathcal{X}, \mathcal{C})}$ dichotomies can be labeled by concepts of $\mathcal{C}_f$, which implies $x_1, \ldots, x_{VC(\mathcal{X}, \mathcal{C})}$ can be shattered by $\mathcal{C}_f$. Therefore $VC(\mathcal{X}, \mathcal{C}) \leq VC(\mathcal{X}, \mathcal{C}_f)$.

2. **Littlestone dimension**. Assume for $\mathcal{C}$, there is an optimal function $\mathcal{O} : (\mathcal{X} \times \{0, 1\})^* \times \mathcal{X} \to \{0, 1\}^2$ that receives pairs of points and labels and one new point and outputs a prediction label of the new point. Then, we can construct the corresponding $\mathcal{O}_f$:

   (a) for any pair of $(x, c_f(x))$, if $c_f(x) = 1$, set the label to $1 - f(x)$, if $c_f(x) = 0$, set the label to $f(x)$.
   (b) feed all pair of points and labels and the new point x_{new} to $\mathcal{O}$, receive a label y.
   (c) output 1 if $y \neq f(x_{new})$, otherwise output 0.

   So that for the concept class $\mathcal{C}_f$, the number of mistakes made by $\mathcal{O}_f$ is at most $d_L(\mathcal{X}, \mathcal{C})$, which implies $d_L(\mathcal{X}, \mathcal{C}_f) \leq d_L(\mathcal{X}, \mathcal{C})$.

   $\square$

In the remaining part of this paper, we only consider $f \in \mathcal{C}$.

Observation 1. *When $f \in \mathcal{C}$, there exists a function $c \in \mathcal{H}_f$, such that $c(x) = 0$ for all $x \in \mathcal{C}$*

Proof. f_f is such a function. $\square$

Lemma 3. *[5] When $f \in \mathcal{C}$, if x_1 and x_2 are incomparable under $\mathcal{C}_f$, then there is no $c \in \mathcal{C}_f$, such that $c(x_1) = 1$ and $c(x_2) = 1$.*

Proof. If x_1 and x_2 are incomparable, then there exists c_1 makes $c_1(x_1) = 1$ and $c_1(x_2) = 0$ (and c_2 makes $c_2(x_1) = 0$ and $c_2(x_2) = 1$, respectively). If there is $c \in \mathcal{C}_f$, such that $c(x_1) = 1$ and $c(x_2) = 1$, then x_1, x_2 are shattered because $f_f \in \mathcal{C}_f$. It makes the VC dimension at least 2. $\square$

---

[2]Littlestone [18] provides a general method to achieve the optimal number of mistakes called standard optimal algorithm (SOA).

4.1 Tree Structure

Then, we can build the tree structure of $\mathcal{C}_f$ according to the partial order relationship. Specifically, for any point x, if there is no x' makes $x \prec x'$, we define $x \prec \emptyset$.

Algorithm 1: MakeTree

/*In this algorithm, we call a node a leaf if the node does not have children.*/
Inputs: a concept class $\mathcal{C}$ with VC dimension 1

1. Select a function $f \in \mathcal{C}$, construct $\mathcal{C}_f$ according to Definition 15. In this algorithm, all the partial order relationships are under $\mathcal{C}_f$.

2. add $\emptyset$ to the tree T.

3. If there is $x \in \mathcal{X}$ and $x \notin T$:

   (a) Let L be the set of leaves of the tree. For $\ell \in L$:

      i. select all the points x' satisfying $x' \prec \ell$ and there is no x'' makes $x' \prec x'' \prec x$. Then make x' to be the child of ℓ.

4. Output the tree T.

Definition 16 (Deterministic Point). *For a point x and a labeled dataset S, we say x is deterministic by S in $\mathcal{C}$ if, for all $h \in \{c \in \mathcal{C} : error_S(c) = 0\}$ (that is all concepts that agrees with S), we have $h(x) = 1$.*

Definition 17 (Distance). *For an ordered sequence $x \prec x_1 \prec \cdots \prec x_k \prec \emptyset$ under $\mathcal{C}_f$. We define $d_{\mathcal{C}_f}(x) = \max k + 1$ as the distance of a point x in $\mathcal{C}_f$. Specifically, $d(\emptyset) = 0$.*

Lemma 4. $\max_{x \in \mathcal{X}} d_{\mathcal{C}_f}(x) \leq TD(\mathcal{X}, \mathcal{C}_f)$

Proof. It is equivalent to show for any $x_0 \prec x_1 \cdots \prec x_k \prec \emptyset$, there exist corresponding $c_0, \ldots, c_k \in \mathcal{C}_f$ make $c_i(x_j) = 1$ if and only if $i \leq j$.

We first select c_0. There must exist a $c_0 \in \mathcal{C}_f$ makes $c_0(x_0) = 1$, otherwise all concepts $c_f \in \mathcal{C}_f$ make $c_f(x_0) = 0$, which makes corresponding $c(x_0) = f(x_0)$ for all $c \in \mathcal{C}$. It contradicts Assumption 2. By Definition 14, $c_0(x_i) = 1$ for all $i \geq 0$.

Assume we already find $c_0, \ldots, c_{k'}$, by Assumtion 1, there exist a concept c' makes $c'(x_{k'}) \neq c'(x_{k'+1})$. Since $c(x_{k'}) = 1 \Rightarrow c(x_{k'+1}) = 1$, it must be that $c'(x_{k'}) = 0$ and $c'(x_{k'+1}) = 1$. By Definition 14, $c'(x_i) = 1$ for all $i \geq k' + 1$ and we can set c' to be $c_{k'+1}$.

At the end, we have $f_f(x_i) = 0$ for all i. $\square$

5 Private learner

Theorem 7. *OPTPrivateLearner is $(2\varepsilon, 2\delta)$-differentially private.*

Proof. For each different entry of S, there is at most one different element in $y_i, \ldots, y_t$ and $q_1, \ldots, q_t$. Thus by Theorem 6, z is (ε, δ)-differentially private in Step 4. Notice that $q_1, \ldots, q_m$ is 1-bouned function (Definition 13). By Lemma 1, x_{good} is (ε, δ)-differentially private in Step 7. By the composition (Theorem 5) and post-processing (Theorem 4), $\hat{h}_{good}$ is $(2\varepsilon, 2\delta)$-differentially private. $\square$

Lemma 5. *[5] For every $c \in \mathcal{C}_f$ and the corresponding set $I(c) = \{x_1, \ldots, x_r\}$, the following two statements are true:*

1. *There is an order $\pi(1), \pi(2), \ldots, \pi(r)$ to make $x_{\pi(1)} \prec \cdots \prec x_{\pi(r)}$.*

Algorithm 2: OPTPrivateLearner

Parameter: Confidence parameter $\beta > 0$, privacy parameter $\varepsilon, \delta > 0$,
$n_{PM}(d, 1/3, \beta, \varepsilon, \delta) = O\left(\frac{\log^* d \cdot \log^2(\frac{\log^* d}{\beta \delta})}{\varepsilon}\right)$ and number of subsets

$t = \max\left\{n_{PM}(d, 1/3, \beta, \varepsilon, \delta), O\left(\frac{1}{\varepsilon}\log(\frac{4n_{PM}(d, 1/3, \beta, \varepsilon, \delta)}{\beta \varepsilon \delta})\right)\right\} = O\left(\frac{\log^* d \cdot \log^2(\frac{\log^* d}{\varepsilon \beta \delta})}{\varepsilon}\right)$, where
$d = TD(\mathcal{X}, \mathcal{C}_f) + 1$.

Inputs: Labeled dataset $S \in (\mathcal{X} \times \{0, 1\})^N$, where $N = t \cdot \frac{48}{\alpha}\left(10\log(\frac{48e}{\alpha}) + \log(\frac{5}{\beta})\right)$

Operation:

1. construct tree according to Algorithm 1

2. randomly partition S into $S_1, \ldots, S_t$

3. For $i \in [t]$

   (a) Let B_i be the set of points deterministic by S_i in $\mathcal{C}_f$. Let $y_i = \max_{x \in B_i} d(x)$, that is the largest distance of a point in B_i.

4. compute the $1/3$-median $z = PrivateMedian(y_1, \ldots, y_t)$ with parameter $\varepsilon, \delta, \beta$.

5. let $P = \{x : d_{\mathcal{C}_f}(x) = z\}$, i.e. the set of points with distance z. Define $P = \{x_1, \ldots, x_m\}$ and $q_1 = q_2 = \cdots = q_m = 0$

6. For $i \in [t]$:

   (a) if $y_i \geq z$:
       i. for $j \in [m]$, if $x_j \in B_i$, make $q_j = q_j + 1$.
   (b) if $y_i < z$, do nothing.

7. run choosing mechanism on $(q_1, q_2, \ldots, q_m)$ with parameter $\varepsilon, \delta, \beta$, select p_{good} and the corresponding point x_{good}.

8. let $I_{good} = \{x | x_{good} \preceq x\}$. Construct $\hat{I}_{good} = \{x : (x \in I_{good} \wedge f(x) = 0) \vee (x \notin I_{good} \wedge f(x) = 1)\}$.

9. construct and output
$$\hat{h}_{good}(x) = \begin{cases} 1 & x \in \hat{I}_{good} \\ 0 & x \notin \hat{I}_{good} \end{cases}$$

---

2. *There is no $\hat{x} \in \mathcal{X}$ to make $x_{\pi(r)} \prec \hat{x}$.*

Proof. By Lemma 3, every $x, x' \in I(c)$ are comparable. Sort all points in $I(c)$ by their distances and it is the order required.

There is no $\hat{x} \in \mathcal{X}$ to make $x_{\pi(r)} \prec \hat{x}$ because if there is a $x_{\pi(r)} \prec \hat{x}$, by Definition 14, we have $\hat{x} \in I(c)$. Then $x_{\pi(r)}$ is not the last point in this order sequence. $\square$

Lemma 6. *For every deterministic point x, we have $x \in I(c_f^*)$, where c^* is the underlying true concept and c_f^* is the f-representation of c^*.*

Proof. For a dataset S, a point x is deterministic if $c(x) = 1$ for any c with $error_S(c, c_f^*) = 0$. The lemma can be concluded by substituting c with c_f^*. $\square$

Lemma 7. *Let h_i be the hypothesis with $I(h_i) = B_i$, then with probability $1 - \beta t$, all h_i are α-good hypothesis.*

Proof. Let c_f^* be the f-representation of the underlying true concept. If $h_i = c_f^*$, then we are done. Otherwise, note that y_i is the point with the largest distance in B_i. It means for any $x \prec y_i$ with $c_f^*(x) = 1$, there exists one $h' \neq c_f^*$ makes $h'(x) = 0$ and $error_{S_i}(h', c_f^*) = 0$ (otherwise x is also deterministic, but $d_{\mathcal{C}_f}(x) > d_{\mathcal{C}_f}(y_i)$, contradicting to y_i is the point with largest distance). By Theorem 2, with probability $1 - \beta$, h_i is an α-good hypothesis.

Notice that $B_i \subseteq I(h')$. Consider the set $I(h') \backslash B_i$. For all $x \in (h_1 \backslash B_i)$, we have $c_f^*(x) = 0$ because $x \notin I(c_f^*)$. Therefore $error_{\mathcal{D}}(h_i, c_f^*) \leq error_{\mathcal{D}}(h', c_f^*) \leq \alpha$ (because for all points that h' and h_i make different predictions, h_i gives the correct label).

Finally, the lemma can be concluded by union bound. $\square$

Lemma 8. *For all $y_1, \ldots, y_t$, there is an order $\pi(1), \pi(2), \ldots, \pi(t)$ to make $y_{\pi(1)} \preceq \cdots \preceq y_{\pi(t)}$.*

Proof. By Lemma 6, all $y_i \in I(c_f^*)$. By Lemma 5, there is an order $\pi(1), \pi(2), \ldots, \pi(t)$ to make $y_{\pi(1)} \preceq \cdots \preceq y_{\pi(t)}$ (here it is possible to have $y_i = y_j$ for $i \neq j$). $\square$

Lemma 9. *Let h_i be the hypothesis with $I(h_i) = B_i$ and h_{good} be the hypothesis with $I(h_{good}) = I_{good}$, when all h_i are α-good hypothesis, with probability $1 - 2\beta$, we have $error_{\mathcal{D}}(h_{good}, c_f^*) \leq \alpha$.*

Proof. Let $\pi(1), \pi(2), \ldots, \pi(t)$ be the order in Lemma 8. By Theorem 6, with probability $1 - \beta$, there are at least $t/6$ y_i's make $y_i \preceq x_{good}$. It means $\sum_i^m q_i \geq t/6$.

We claim that every different point in S makes at most one q_i different. Otherwise, there exist different x, x' with distance z and one B_i to make $x, x' \in B_i$. By Lemma 3, x and x' are comparable. Assume $x \prec x'$, it makes $d_{\mathcal{C}_f}(x) > d_{\mathcal{C}_f}(x')$.

So that we can apply the choosing mechanism. For any $x \neq x_{good}$, they will get a 0 score. For x_{good}, it will get a score of at least $t/6 \geq \frac{16}{\varepsilon} \log(\frac{4n_{PM}(d, 1/3, \beta, \varepsilon, \delta)}{\beta \varepsilon \delta})$. By Lemma 1, with probability at least $1 - \beta$, choosing mechanism outputs x_{good}.

Since there is at least one y_i make $x_{good} \preceq y_i$, we have $B_i \subseteq I_{good}$. Thus $error_{\mathcal{D}}(h_{good}, c_f^*) \leq error_{\mathcal{D}}(h_i, c_f^*) \leq \alpha$ because for all points that h_{good} and h_i make different predictions, h_{good} gives the correct label $\square$

Corollary 2. *With probability $1 - (t+2)\beta$, OPTPrivateLearner outputs $\hat{h}_{good}$ satisfying $error_{\mathcal{D}}(\hat{h}_{good}, c^*) \leq \alpha$.*

Proof. The accuracy is because $error_{\mathcal{D}}(\hat{h}_{good}, c^*) = error_{\mathcal{D}}(h_{good}, c_f^*)$. The confidence is by the union bound. $\square$

Substitute ε by $\varepsilon/2$, δ by $\delta/2$, and β by $\beta/(t+2)$, and considering Corollary 1, we have the main result.

Theorem 8. *For any concept class $\mathcal{C}$ with VC dimension 1 and Littlestone dimension d, and given labeled dataset with size*

$$N \geq O\left(\frac{\log^* d \cdot \log^2(\frac{\log^* d}{\varepsilon \beta \delta})}{\varepsilon} \cdot \frac{48}{\alpha}\left(10 \log(\frac{48e}{\alpha}) + \log(\frac{5}{\beta})\right)\right) = \tilde{O}_{\beta, \delta}\left(\frac{\log^* d}{\alpha \varepsilon}\right)$$

there is an (ε, δ)-differentially private algorithm that (α, β)-PAC learns $\mathcal{C}$.

References

[1] Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite littlestone dimension. In Moses Charikar and Edith Cohen, editors, *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 852–860. ACM, 2019.

[2] 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.

[3] Amos Beimel, Shay Moran, Kobbi Nissim, and Uri Stemmer. Private center points and learning of halfspaces. In *Conference on Learning Theory*, pages 269–282. PMLR, 2019.

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

[5] Shai Ben-David. 2 notes on classes with vapnik-chervonenkis dimension 1. *ArXiv*, abs/1507.05307, 2015.

[6] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, October 1989.

[7] Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 389–402. IEEE, 2020.

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

[9] Edith Cohen, Xin Lyu, Jelani Nelson, Tamás Sarlós, and Uri Stemmer. Õptimal differentially private learning of thresholds and quasi-concave optimization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023*, pages 472–482, 2023.

[10] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *STOC*, pages 371–380. ACM, May 31–June 2 2009.

[11] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

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

[13] Badih Ghazi, Noah Golowich, Ravi Kumar, and Pasin Manurangsi. Sample-efficient proper pac learning with approximate differential privacy. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 183–196, 2021.

[14] Wilfrid Hodges. *A shorter model theory*. Cambridge university press, 1997.

[15] Haim Kaplan, Katrina Ligett, Yishay Mansour, Moni Naor, and Uri Stemmer. Privately learning thresholds: Closing the exponential gap. In *COLT*, pages 2263–2285, 2020.

[16] Haim Kaplan, Yishay Mansour, Uri Stemmer, and Eliad Tsfadia. Private learning of halfspaces: Simplifying the construction and reducing the sample complexity. *Advances in Neural Information Processing Systems*, 33:13976–13985, 2020.

[17] 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, 2011.

[18] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. In *28th Annual Symposium on Foundations of Computer Science (sfcs 1987)*, pages 68–77, 1987.

[19] Kobbi Nissim, Eliad Tsfadia, and Chao Yan. Differentially private quasi-concave optimization: Bypassing the lower bound and application to geometric problems. *arXiv preprint arXiv:2504.19001*, 2025.

[20] Saharon Shelah. *Classification theory: and the number of non-isomorphic models*. Elsevier, 1990.

[21] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, November 1984.

[22] Vladimir N. Vapnik and Alexey Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.