Proceedings of Machine Learning Research vol 125:1–34, 2020 33rd Annual Conference on Learning Theory

Closure Properties for Private Classification and Online Prediction

Noga Alon[*] NALON@MATH.PRINCETON.EDU

Amos Beimel[†] AMOS.BEIMEL@GMAIL.COM

Shay Moran[‡] SHAYMORAN1@GMAIL.COM

Uri Stemmer[§] U@URI.CO.IL

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

Let $\mathcal{H}$ be a class of boolean functions and consider a *composed class* $\mathcal{H}'$ that is derived from $\mathcal{H}$ using some arbitrary aggregation rule (for example, $\mathcal{H}'$ may be the class of all 3-wise majority-votes of functions in $\mathcal{H}$). We upper bound the Littlestone dimension of $\mathcal{H}'$ in terms of that of $\mathcal{H}$. As a corollary, we derive closure properties for online learning and private PAC learning.

The derived bounds on the Littlestone dimension exhibit an undesirable exponential dependence. For private learning, we prove close to optimal bounds that circumvents this suboptimal dependency. The improved bounds on the sample complexity of private learning are derived algorithmically via transforming a private learner for the original class $\mathcal{H}$ to a private learner for the composed class $\mathcal{H}'$. Using the same ideas we show that any (*proper or improper*) private algorithm that learns a class of functions $\mathcal{H}$ in the realizable case (i.e., when the examples are labeled by some function in the class) can be transformed to a private algorithm that learns the class $\mathcal{H}$ in the agnostic case.

1. Introduction

We study closure properties for learnability of binary-labeled hypothesis classes in two related settings: online learning and differentially private PAC learning.

Closure Properties for Online Learning. Let $\mathcal{H}$ be a class of experts that can be online learned with vanishing regret. That is, there exists an algorithm $\mathcal{A}$ such that given any sequence of T prediction tasks, the number of false predictions made by $\mathcal{A}$ is larger by at most $R(T) = o(T)$ than the number of false predictions made by the best expert in $\mathcal{H}$.

Consider a scenario where the sequence of tasks is such that every single expert in $\mathcal{H}$ predicts poorly on it, however there is a small unknown set of experts $h_1, \ldots, h_k \in \mathcal{H}$ that can predict

[*] Department of Mathematics, Princeton University, Princeton, New Jersey, USA and Schools of Mathematics and Computer Science, Tel Aviv University, Tel Aviv, Israel. Research supported in part by NSF grant DMS-1855464, ISF grant 281/17, BSF grant 2018267, and the Simons Foundation.

[†] Department of Computer Science, Ben-Gurion University, Beer-Sheva, Israel. This work was done while visiting Georgetown University, supported by NSF grant no. 1565387, TWC: Large: Collaborative: Computing Over Distributed Sensitive Data and also supported by ISF grant 152/17, by ERC grant 742754 (project NTSC), and by a grant from the Cyber Security Research Center at Ben-Gurion University of the Negev.

[‡] Google AI, Princeton.

[§] Department of Computer Science, Ben-Gurion University, Beer-Sheva, Israel, and Google Research. Partially supported by ISF grant 1871/19.

well by collaborating. More formally, there is an aggregation rule $G : \{0,1\}^k \to \{0,1\}$ such that the combined expert $G(h_1, \ldots, h_k)$ exhibits accurate predictions on a significant majority of the tasks. For example, a possible aggregation rule G could be the majority-vote of the k experts. Since we assume that the identities of the k experts are not known, it is natural to consider the class $\mathcal{H}' = \{G(h_1, \ldots, h_k) : h_i \in \mathcal{H}\}$, which consists of all possible G-aggregations of k experts from $\mathcal{H}$. We study the following question:

Question 1 *Can the optimal regret with respect to $\mathcal{H}'$ be bounded in terms of that of $\mathcal{H}$?*

The *Littlestone dimension* is a combinatorial parameter that determines online learnability (Littlestone, 1987; Ben-David et al., 2009). In particular, $\mathcal{H}$ is online learnable if and only if it has a finite Littlestone dimension $d < \infty$, and the best possible regret $R(T)$ for online learning $\mathcal{H}$ satisfies

$$\Omega(\sqrt{dT}) \le R(T) \le O(\sqrt{dT \log T}). \tag{1}$$

Furthermore, if it is known that if one of the experts never errs (a.k.a the realizable setting), then the optimal regret is exactly d.[1] (The regret is called mistake-bound in this context.)

Thus, the above question boils down to asking whether the Littlestone dimension of $\mathcal{H}'$ is bounded by a function of the Littlestone dimension of $\mathcal{H}$. One of the two main results in this work provides an affirmative answer to this question (Theorem 3).

We next discuss a variant of this question in the setting of Differentially Private (DP) learning. The two settings of online and DP-learning are intimately related (see, e.g., Bun et al. (2020); Abernethy et al. (2017); Joseph et al. (2019); Gonen et al. (2019)). In particular, both online learning and DP-learning are characterized by the finiteness of the Littlestone dimension (Littlestone, 1987; Ben-David et al., 2009; Bun et al., 2015; Alon et al., 2019; Bun et al., 2020).

Closure Properties for Differentially Private Learning. Imagine the following medical scenario: consider a family $\mathcal{H}$ of viruses for which there is an algorithm $\mathcal{A}$ that can learn to diagnose any specific virus $h \in \mathcal{H}$ given enough labeled medical data. Further assume that $\mathcal{A}$ has the desired property of being a differentially private learning algorithm as defined by (Kasiviswanathan et al., 2011); that is, it is a PAC learning algorithm in which the privacy of every patient whose data is used during training is guarded in the formal sense of differential privacy (Dwork et al., 2006b).

Assume that an outbreak of a deadly disease h' has occurred in several locations all over the world and that it is known that h' is caused by some relatively small, yet unknown group of viruses from $\mathcal{H}$. That is, our prior information is that there are unknown viruses $h_1, \ldots, h_k \in \mathcal{H}$ for a relatively small k such that $h' = G(h_1, \ldots, h_k)$ for some rule G. For example, G could be the OR function in which case h' occurs if and only if the patient is infected with at least one of the viruses $h_1, \ldots, h_k$.

It would be highly beneficial if one could use the algorithm $\mathcal{A}$ to diagnose h' in an automated fashion. Moreover, doing it in a private manner could encourage health institutions in the different locations to contribute their patients' data. This inspires the following question:

Question 2 *Can one use the algorithm $\mathcal{A}$ to* privately *learn to diagnose h'? How does the sample complexity of this learning task scale as a function of G?*

[1] More precisely, there is a deterministic algorithm that makes no more than d mistakes, and for every deterministic algorithm there is a (realizable) input sequence on which it makes at least d mistakes. For randomized algorithms a slightly weaker lower bound of $d/2$ holds with respect to the expected number of mistakes.

Differential Privacy, Online Learning, and the Littlestone Dimension. Question 2 and Question 1 are equivalent in the sense that both online learning and DP-learning are characterized by the finiteness of the Littlestone dimension (Littlestone, 1987; Ben-David et al., 2009; Bun et al., 2015; Alon et al., 2019; Bun et al., 2020).

Note however that unlike the bounds relating the Littlestone dimension to online learning, which are tight up to logarithmic factors (see (1)), the bounds relating the Littlestone dimension and DP-learning are *very far from each other*; specifically, if d denotes the Littlestone dimension of $\mathcal{H}$ then the lower bound on the sample complexity of privately learning $\mathcal{H}$ scales with $\log^* d$ (Bun et al., 2015; Alon et al., 2019), while the best known[2] upper bound scales with $\exp(d)$ (Bun et al., 2020).

Thus, while our solution to Question 1 yields an affirmative answer to Question 2, the implied quantitative bounds are far from being realistically satisfying. Specifically, every finite $\mathcal{H}$ is learnable with privacy using $O(\log|\mathcal{H}|)$ samples (Kasiviswanathan et al., 2011), and so if $\mathcal{H}$ is finite and not too large, the bounds implied by the Littlestone dimension are not meaningful. We therefore focus on deriving effective bounds for private learning, which is the content of Theorem 5 (see Theorem 29 for a precise statement).

Littlestone Classes. It is natural to ask which natural hypothesis classes have bounded Littlestone dimension. First, it holds that $\mathrm{Ldim}(\mathcal{H}) \leq \log|\mathcal{H}|$ for every $\mathcal{H}$, so for finite classes the Littlestone dimension scales rather gracefully with their size.

There are also natural infinite Littlestone classes: for example, let the domain $X = \mathbb{F}^n$ be an n-dimensional vector space over some field $\mathbb{F}$ and let $\mathcal{H} \subseteq \{0,1\}^X$ consist of all affine subspaces of V of dimension $\leq d$. It can be shown here that $\mathrm{Ldim}(\mathcal{H}) = d$. (For example, the class of all lines in $\mathbb{R}^{100}$ has Littlestone dimension 1.) A bit more generally, any class of hypotheses that can be described by polynomial *equalities* of a bounded degree has bounded Littlestone dimension. (Observe that if one replaces "equalities" with "inequalities" then the Littlestone dimension may become unbounded, however the VC dimension remains bounded (e.g. halfspaces).) This can be further generalized to classes that are definable in *stable theories*, which is a deep and well-explored notion in model theory. We refer the reader to Chase and Freitag (2019), Section 5.1 for such examples.

1.1. Main Results

Let $G : \{0,1\}^k \to \{0,1\}$ be a boolean function and let $\mathcal{H}_1, \ldots, \mathcal{H}_k \subseteq \{0,1\}^X$ be hypothesis classes. Denote by $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ the following class $G(\mathcal{H}_1, \ldots, \mathcal{H}_k) = \{G(h_1, \ldots, h_k) : h_i \in \mathcal{H}_i\}$. For example, if $G(x_1, x_2) = x_1 \wedge x_2$ then $G(\mathcal{H}_1, \mathcal{H}_2) = \mathcal{H}_1 \wedge \mathcal{H}_2 = \{h_1 \wedge h_2 : h_i \in \mathcal{H}_i\}$ is the class of all pairwise intersections/conjunctions of a function from $\mathcal{H}_1$ and a function from $\mathcal{H}_2$.

Theorem 3 (A Closure Theorem for the Littlestone Dimension) *Let $G : \{0,1\}^k \to \{0,1\}$ be a boolean function, let $\mathcal{H}_1, \ldots, \mathcal{H}_k \subseteq \{0,1\}^X$ be classes, and let $d \in \mathbb{N}$ such that $\mathrm{Ldim}(\mathcal{H}_i) \leq d$ for every $i \leq k$. Then,*

$$\mathrm{Ldim}(G(\mathcal{H}_1, \ldots, \mathcal{H}_k)) \leq \tilde{O}(2^{2k} k^2 d),$$

where $\tilde{O}$ conceals polynomial factors in $\log k$ and $\log d$.

[2] The lower bound is tight up to polynomial factors (Kaplan et al., 2020a), however the upper bound is not known to be tight: for example, as far as we know, it is possible that the sample complexity of private learning scales linearly with $\mathrm{VC}(\mathcal{H}) + \log^*(\mathrm{Ldim}(\mathcal{H}))$.

In particular, if $\mathrm{Ldim}(\mathcal{H}_i) < \infty$ for all $i \leq d$ then $\mathrm{Ldim}(G(\mathcal{H}_1, \ldots, \mathcal{H}_k)) < \infty$. Consequently, if each of the $\mathcal{H}_i$'s is online learnable then $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ is online learnable. We comment that if the aggregating function G is simple then one can obtain better bounds. For example, if G is a majority-vote, a k-wise OR, or a k-wise AND function then a bound of $\tilde{O}(k^2 \cdot d)$ holds. (See Appendix A.2.1.)

Another combinatorial parameter that arises in the relationship between online and DP learning is the *threshold dimension*: a sequence $x_1, \ldots, x_k \in X$ is *threshold-shattered* by $\mathcal{H}$ if there are $h_1, \ldots, h_k \in \mathcal{H}$ such that $h_i(x_j) = 1$ if and only if $i \leq j$ for all $i, j \leq k$. The *threshold dimension*, $T(\mathcal{H})$, is the maximum size of a sequence that is threshold-shattered by $\mathcal{H}$. The threshold dimension plays a key role in showing that DP learnable classes have a finite Littlestone dimension (Alon et al., 2019). A classical theorem by Shelah (1978) in model theory shows that the Littlestone and the threshold dimensions are exponentially related.[3] In particular $\mathrm{Ldim}(\mathcal{H}) < \infty$ if and only if $T(\mathcal{H}) < \infty$. (See Theorem 8 in the preliminaries section.) We prove the following closure theorem in terms of the threshold dimension.

Theorem 4 (A Closure Theorem for the Threshold Dimension) *Let $G : \{0,1\}^k \to \{0,1\}$ be a boolean function, let $\mathcal{H}_1, \ldots, \mathcal{H}_k \subseteq \{0,1\}^X$ be classes, and let $t \in \mathbb{N}$ such that $T(\mathcal{H}_i) < t$ for every $i \leq k$. Then,*

$$T\big(G(\mathcal{H}_1, \ldots, \mathcal{H}_k)\big) < 2^{4k4^k \cdot t}.$$

Moreover, an exponential dependence in t is necessary: for every $t \geq 6$ there exists a class $\mathcal{H}$ such that $T(\mathcal{H}) \leq t$ and

$$T\Big(\{h_1 \vee h_2 : h_1, h_2 \in \mathcal{H}\}\Big) \geq 2^{\lfloor t/5 \rfloor}.$$

Note that the bounds in Theorem 3 and Theorem 4 escalate rapidly with k (the arity of G) and with t. It will be interesting to determine tight bounds.

Closure of Private Learning. By Alon et al. (2019); Bun et al. (2020), Theorem 3 also implies closure properties for DP-learnable classes. However, the quantitative bounds are even worse: not only do the bounds on the Littlestone dimension of $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ escalate rapidly with k, also the quantitative relationship between the Littlestone dimension and DP-learning sample complexity is very loose, and the best bounds exhibit a tower-like gap between the upper and lower bounds. For example, if the class of functions $\mathcal{H}$ is finite and its Littlestone dimension is $\omega(\log \log |\mathcal{H}|)$, then the bound of Theorem 3 is most likely to be much worse than the generic application of the exponential mechanism, whose sample complexity is the logarithm of the size of the class. We therefore explore the closure properties of differentially-private learning algorithms directly and derive the following bound.

Theorem 5 (A Closure Theorem for Private Learning (informal)) *Let $G : \{0,1\}^k \to \{0,1\}$ be a boolean function. Let $\mathcal{H}_1, \ldots, \mathcal{H}_k \subseteq \{0,1\}^X$ be classes that are (ε, δ)-differentially private and (α, β)-accurate learnable in the PAC model with sample complexity m_i respectively. Then,*

[3] The threshold dimension may be interpreted as a combinatorial abstraction of the geometric notion of *margin*. Under this interpretation, Shelah's result may be seen as an extension of the classical Perceptron's mistake-bound analysis by Rosenblatt (1958).

$G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ is (ε, δ)-private and (α, β)-accurate learnable with sample complexity

$$\tilde{O}\left(\sum_{i=1}^{k} m_i\right) \cdot \mathrm{poly}(k, 1/\varepsilon, 1/\alpha, \log(1/\beta)).$$

The exact quantitative statement of the results appears in Theorem 29. We remark that closure properties for *pure* differentially-private learning algorithms (i.e., when $\delta = 0$) are implied by the characterization of (Beimel et al., 2019b). Similarly, closure properties for *non-private* PAC learning are implied by the characterization of their sample complexity in terms of the VC dimension and by the Sauer-Shelah-Perles Lemma (Sauer, 1972). However, since there is no tight characterization of the sample complexity of approximate differentially-private learning algorithms (i.e., when $\delta > 0$), we prove Theorem 5 algorithmically by constructing a (non-efficient) learning algorithm for $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ from private learning algorithms for $\mathcal{H}_1, \ldots, \mathcal{H}_k$. For approximate differentially-private learning, closure properties were not known even for simple compositions. For example, Beimel et al. (2019a) and Kaplan et al. (2020b) showed how to privately learn halfspaces in a discrete space X^d with sample complexity $\mathrm{poly}(d, 2^{\log^* |X|})$, however, privately learning intersection of two halfspaces with sample complexity $\mathrm{poly}(d, 2^{\log^* |X|})$ was not known prior to our results.

Agnostic Learning. Beimel et al. (2015) proved that any *proper* private learning algorithm in the realizable case[4] can be transformed into an agnostic[5] private learning algorithm, with only a mild increase in the sample complexity. We show that the same result holds even for *improper* private learning (i.e., when the private learning algorithm can return an arbitrary hypothesis).

Theorem 6 (Private Learning Implies Agnostic Private Learning) *For every $0 < \alpha, \beta, \delta < 1$, every $m \in \mathbb{N}$, and every concept class $\mathcal{H}$, if there exists a $(1, \delta)$-differentially private (α, β)-accurate PAC learner for the hypothesis class $\mathcal{H}$ with sample complexity m, then there exists an $(O(1), O(\delta))$-differentially private $(O(\alpha), O(\beta + \delta n))$-accurate* agnostic *learner for $\mathcal{H}$ with sample complexity*

$$n = O\left(m + \frac{1}{\alpha^2}\left(\mathrm{VC}(\mathcal{H}) + \log\frac{1}{\beta}\right)\right).$$

Furthermore, if the original learner is proper, then the agnostic learner is proper.[6]

We obtain this result by showing that a variant of the transformation of (Beimel et al., 2015) also works for the improper case; we do not know if the original transformation of (Beimel et al., 2015) also works for the improper case. Our analysis of the transformation for the improper case is more involved than the analysis for the proper case.

[4] That is, when the examples are labeled by some $h \in \mathcal{H}$.

[5] That is, when the examples are labeled arbitrarily and the goal is to find a hypothesis whose error is close to the smallest error of a hypothesis in $\mathcal{H}$.

[6] By standard subsampling techniques (Kasiviswanathan et al., 2011; Beimel et al., 2014), for every $\varepsilon < 1$ we can get an (ε, δ)-differentially agnostic private PAC learner by multiplying the sample complexity by a factor of $O(1/\varepsilon)$.

1.2. Technical Overview

1.2.1. CLOSURE FOR LITTLESTONE DIMENSION

Our proof of Theorem 3 exploits tools from online learning. It may be instructive to compare Theorem 3 with an analogous result for VC classes: a classical result by Dudley (1978) upper bounds the VC dimension of $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ by $\tilde{O}(d_1 + \cdots + d_k)$, where d_i is the VC dimension of $\mathcal{H}_i$. The argument uses the Sauer-Shelah-Perles Lemma (Sauer, 1972) to bound the growth-rate (a.k.a. shatter function) of $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ by some $n^{d_1+\cdots+d_k}$: indeed, if we let $n = \mathrm{VC}(G(\mathcal{H}_1, \ldots, \mathcal{H}_k))$, then by the definition of the shatter function, $2^n \leq n^{d_1+\cdots+d_k}$, which implies that $n = \tilde{O}(d_1 + \cdots + d_k)$ as stated. It is worth noting that a notion of growth-rate as well as a corresponding variant of the Sauer-Shelah-Perles Lemma also exist for Littlestone classes (Bhaskar, 2017; Chase and Freitag, 2018). However we are not aware of a way of using it to prove Theorem 3.

We take a different approach. We first focus on the case where G is a majority-vote. That is, the class $\mathcal{H} = G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ consists of all k-wise majority-votes of experts $h_i \in \mathcal{H}_i$. We bound the Littlestone dimension of $\mathcal{H}$ by exhibiting an online learning algorithm A that learns $\mathcal{H}$ in the mistake-bound model with at most $\tilde{O}(k^2 \cdot d)$ mistakes. The derivation of A exploits fundamental tools from online learning such as the *Weighted Majority Algorithm* by Littlestone and Warmuth (1989) and *Online Boosting* (Chen et al., 2012; Beygelzimer et al., 2015; Brukhim et al., 2020).

Then, the bound for a general $G : \{0,1\}^k \to \{0,1\}$ is obtained by expressing G as a formula that only uses majority-votes and negations gates. The exponential dependence in k in the final bound is a consequence of the formula-size that can be exponential in k. We do not know whether this exponential dependence is necessary.

1.2.2. CLOSURE FOR THRESHOLD DIMENSION

Our proof of Theorem 4 is combinatorial. First, note that an inferior bound follows from Theorem 3, using the fact that the Littlestone and threshold dimensions are exponentially related (see Theorem 8). However this approach yields a super-exponential bound on $T(G(\mathcal{H}_1, \ldots, \mathcal{H}_k))$.

The bound in Theorem 4 follows by arguing contra-positively that if $T(G(\mathcal{H}_1, \ldots, \mathcal{H}_k))$ is large then $T(\mathcal{H}_i)$ is also “largish” for some $i \leq k$. Specifically, if $T(G(\mathcal{H}_1, \ldots, \mathcal{H}_k)) \geq \exp(t \exp(k))$ then $T(\mathcal{H}_i) \geq t$ for some $i \leq k$. This is shown using a Ramsey argument that asserts that any large enough sequence $x_1, \ldots, x_n$ that is threshold-shattered by $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ must contain a relatively large subsequence that is threshold-shattered by one of the $\mathcal{H}_i$'s. Quantitatively, if $n \geq \exp(t \exp(k))$ then there must be a subsequence $x_{j_1}, \ldots, x_{j_t}$ that is threshold-shattered by one of the $\mathcal{H}_i$'s.

This upper bounds $T(G(\mathcal{H}_1, \ldots, \mathcal{H}_k))$ by some $\exp(t \exp(k))$, where $t = \max_i T(\mathcal{H}_i)$. It is worth noting that, in contrast with Theorem 3, an exponential dependence here is inevitable: we prove in Theorem 4 that for any t there exists a class $\mathcal{H}$ with $T(\mathcal{H}) \leq t$ such that $T(\{h_1 \vee h_2 : h_1, h_2 \in \mathcal{H}\}) \geq \exp(t)$. This lower bound is achieved by a randomized construction.

1.2.3. PRIVATE LEARNING IMPLIES AGNOSTIC PRIVATE LEARNING

We start by describing the transformation of (Beimel et al., 2015) from a proper private learning algorithm of a class $\mathcal{H}$ to an agnostic proper private learning algorithm for $\mathcal{H}$. Assume that there is a private learning algorithm $\mathcal{A}$ for $\mathcal{H}$ with sample complexity m. The transformation takes a sample S of size $O(m)$ and constructs all possible behaviors, denoted by H, of functions in $\mathcal{H}$

on the points of the sample. By the Sauer-Shelah-Perles Lemma, the number of such behaviors is at most $\left(\frac{e|S|}{\mathrm{VC}(\mathcal{H})}\right)^{\mathrm{VC}(\mathcal{H})}$. Then, it finds using the exponential mechanism a behavior $h' \in H$ that minimizes the empirical error on the sample. (The exponential mechanism is a differentially-private algorithm that is guaranteed to identify a hypothesis with small empirical error when the number of possible hypotheses is relatively small as in our case.) Finally, the transformation relabels the sample S using h' and applies $\mathcal{A}$ on the relabeled sample. If $\mathcal{A}$ is a proper learning algorithm then, by standard VC arguments, the resulting algorithm is an agnostic algorithm for $\mathcal{H}$. The privacy guarantees of the resulting algorithm are more delicate, and it is only $O(1)$-differentially private, even if $\mathcal{A}$ is ε-differentially private for a small ε. (The difficulty in the privacy analysis is that the set of behaviors H is *data-dependent*. Therefore, the privacy guarantees of the resulting algorithm *are not* directly implied by those of the exponential mechanism, which assume that the set of possible outcomes is fixed and data-independent.)

When $\mathcal{A}$ is improper, we cannot use VC arguments to argue that the resulting algorithm is an agnostic learner. We rather use the generalization properties of differential privacy (proved in (Dwork et al., 2015; Bassily et al., 2016; Rogers et al., 2016; Feldman and Steinke, 2017; Nissim and Stemmer, 2017; Jung et al., 2020)): if a differentially private algorithm has a small empirical error on a sample chosen i.i.d. from some distribution, then it also has a small generalization error on the underlying distribution (even if the labeling hypothesis is chosen after seeing the sample). There are technical issues in applying these results in our case that require some modifications in the transformation.

1.2.4. CLOSURE FOR DIFFERENTIALLY PRIVATE LEARNING

We prove Theorem 5 by constructing a private algorithm $\mathcal{A}_{\mathrm{ClosureLearn}}$ for the class $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ using private learning algorithms for the classes $\mathcal{H}_1, \ldots, \mathcal{H}_k$. Algorithm $\mathcal{A}_{\mathrm{ClosureLearn}}$ uses the relabeling procedure (the one that we use to transform a private PAC learner into a private agnostic learner) in a new setting.

The input to $\mathcal{A}_{\mathrm{ClosureLearn}}$ is a sample labeled by some function in $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$. The algorithm finds hypotheses $h_1, \ldots, h_k$ in steps, where in the i'th step, the algorithm finds a hypothesis h_i such that $h_1, \ldots, h_i$ have a completion $c_{i+1}, \ldots, c_k$ to a hypothesis $G(h_1, \ldots, h_i, c_{i+1}, \ldots, c_k)$ with small error (assuming that $h_1, \ldots, h_{i-1}$ have a good completion).

Each step of $\mathcal{A}_{\mathrm{ClosureLearn}}$ is similar to the algorithm for agnostic learning described above. That is, in the i'th step, $\mathcal{A}_{\mathrm{ClosureLearn}}$ first relabels the input sample S using some $h \in \mathcal{H}_i$ in a way that guarantees completion to a hypothesis with small empirical error. The relabeling h is chosen using the exponential mechanism with an appropriate score function. The relabeled sample is then fed to the private algorithm for the class $\mathcal{H}_i$ to produce a hypothesis h_i and then the algorithm proceeds to the next step $i + 1$. As in the algorithm for agnostic learning, the proof that the hypothesis $G(h_1, \ldots, h_k)$ returned by the algorithm has a small generalization error is easier when the private algorithms for $\mathcal{H}_1, \ldots, \mathcal{H}_k$ are proper, and it is more involved if they are improper.

Organization. Definitions and background results are provided in Section 2. The complete proofs appear in the appendix in two parts. Closure properties for Littlestone classes is proved in Appendix A. The effective bounds for private learning (i.e., agnostic learning and closure) are given in Appendices B to D. We note that each of these parts can be read independently of the other.

2. Preliminaries

This section is organized as follows: Section 2.1 contains basic definitions and tools related to the Littlestone dimension and Section 2.2 contains basic definitions and tools related to private learning.

2.1. Preliminaries on the 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 labeled by elements of X. Any root-to-leaf path in a mistake tree can be described as a sequence of examples $(x_1, y_1), \dots, (x_d, y_d)$, where x_i is the label of the 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 $y_i = 0$. We say that a tree T is *shattered* by $\mathcal{H}$ if for any root-to-leaf path $(x_1, y_1), \dots, (x_d, y_d)$ 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 the largest complete tree that is shattered by $\mathcal{H}$.

Definition 7 (Subtree) *Let T be labeled binary tree. We will use the following notion of a subtree T' of depth h of T by induction on h:*

1. *Any leaf of T is a subtree of height 0.*
2. *For $h \ge 1$ a subtree of height h is obtained from an internal vertex v of T together with a subtree of height $h-1$ of the tree rooted at v's left child and a subtree of height $h-1$ of the tree rooted at v's 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 $\mathcal{H}$.

Threshold Dimension. A classical theorem of Shelah in model-theory connects bounds on 2-rank (Littlestone dimension) to the concept of *thresholds*: let $\mathcal{H} \subseteq \{0,1\}^X$ be a hypothesis class. We say that a sequence $x_1, \dots, x_k \in X$ is *threshold-shattered* by $\mathcal{H}$ if there are $h_1, \dots, h_k \in \mathcal{H}$ such that $h_i(x_j) = 1$ if and only if $i \le j$ for all $i, j \le k$. Define the *threshold dimension*, $T(\mathcal{H})$, as the maximum size of a sequence that is threshold-shattered by $\mathcal{H}$.

Theorem 8 (Littlestone Dimension versus Threshold Dimension (Shelah, 1978; Hodges, 1997)) *Let $\mathcal{H}$ be a hypothesis class, then:*

$$T(\mathcal{H}) \ge \lfloor \log \mathrm{Ldim}(\mathcal{H}) \rfloor \quad \textit{and} \quad \mathrm{Ldim}(\mathcal{H}) \ge \lfloor \log T(\mathcal{H}) \rfloor.$$

2.2. Preliminaries on Private Learning

Differential Privacy. Consider a database where each record contains information of an individual. An algorithm is said to preserve differential privacy if a change of a single record of the database (i.e., information of an individual) does not significantly change the output distribution of the algorithm. Intuitively, this means that the information inferred about an individual from the output of a differentially-private algorithm is similar to the information that would be inferred had the individual's record been arbitrarily modified or removed. Formally:

Definition 9 (Differential privacy (Dwork et al., 2006b,a)) *A randomized algorithm $\mathcal{A}$ is (ε, δ)-differentially private if for all neighboring databases $S_1, S_2 \in X^m$ (i.e., databases differing in one entry), and for all sets $\mathcal{F}$ of outputs,*

$$\Pr[\mathcal{A}(S_1) \in \mathcal{F}] \leq \exp(\varepsilon) \cdot \Pr[\mathcal{A}(S_2) \in \mathcal{F}] + \delta, \tag{2}$$

where the probability is taken over the random coins of $\mathcal{A}$. When $\delta = 0$ we omit it and say that $\mathcal{A}$ preserves pure *ε-differential privacy. When $\delta > 0$, we use the term* approximate *differential privacy, in which case δ is typically a negligible function of the database size m.*

PAC Learning. We next recall the probably approximately correct (PAC) model of Valiant (1984). A hypothesis $c : X \to \{0, 1\}$ is a predicate that labels *examples* taken from the domain X by either 0 or 1. We sometime refer to a hypothesis as a concept. A *hypothesis class* $\mathcal{H}$ over X is a set of hypotheses (predicates) mapping X to $\{0, 1\}$. A learning algorithm is given examples sampled according to an unknown probability distribution $\mathcal{P}$ over X, and labeled according to an unknown *target* concept $c \in \mathcal{H}$. The learning algorithm is successful when it outputs a hypothesis h that approximates the target concept over samples from $\mathcal{P}$. More formally:

Definition 10 *The* generalization error *of a hypothesis $h : X \to \{0, 1\}$ with respect to a concept c and a distribution $\mathcal{P}$ over X is defined as $\mathrm{error}_{\mathcal{P}}(c, h) = \Pr_{x \sim \mathcal{P}}[h(x) \neq c(x)]$. If $\mathrm{error}_{\mathcal{P}}(c, h) \leq \alpha$ we say that h is α-*good *for c and $\mathcal{P}$.*

Definition 11 (PAC Learning (Valiant, 1984)) *An algorithm $\mathcal{A}$ is an (α, β)-*accurate PAC learner *for a hypothesis class $\mathcal{H}$ over X if for all concepts $c \in \mathcal{H}$, all distributions $\mathcal{P}$ on X, given an input of m samples $S = (z_1, \ldots, z_m)$, where $z_i = (x_i, c(x_i))$ and each x_i is drawn i.i.d. from $\mathcal{P}$, algorithm $\mathcal{A}$ outputs a hypothesis h satisfying*

$$\Pr[\mathrm{error}_{\mathcal{P}}(c, h) \leq \alpha] \geq 1 - \beta,$$

where the probability is taken over the random choice of the i.i.d. examples in S according to $\mathcal{P}$ and the random coins of the learner $\mathcal{A}$. If the output hypothesis h always satisfies $h \in \mathcal{H}$ then $\mathcal{A}$ is called a proper *PAC learner; otherwise, it is called an* improper *PAC learner.*

Definition 12 *For an unlabeled sample $S = (x_i)_{i=1}^m$, the* empirical error *between two concepts c, h is $\mathrm{error}_S(c, h) = \frac{1}{m}|\{i : c(x_i) \neq h(x_i)\}|$. For a labeled sample $S = (x_i, y_i)_{i=1}^m$, the* empirical error *of h is $\mathrm{error}_S(h) = \frac{1}{m}|\{i : h(x_i) \neq y_i\}|$.*

The previous definition of PAC learning captures the realizable case, that is, the examples are drawn from some distribution and labeled according to some concept $c \in \mathcal{H}$. We next define agnostic learning, i.e., where there is a distribution over labeled examples and the goal is to find a hypothesis whose error is close to the error of the best hypothesis in $\mathcal{H}$ with respect to the distribution. Formally, for a distribution μ on $X \times \{0, 1\}$ and a function $f : X \to \{0, 1\}$ we define $\mathrm{error}_{\mu}(f) = \Pr_{(x,a) \sim \mu}[f(x) \neq a]$.

Definition 13 (Agnostic PAC Learning) *Algorithm $\mathcal{A}$ is an (α, β)-accurate* agnostic PAC learner *for a hypothesis class $\mathcal{H}$ with sample complexity m if for all distributions μ on $X \times \{0, 1\}$, given an* *input of m labeled samples $S = (z_1, \ldots, z_m)$, where each labeled example $z_i = (x_i, a_i)$ is drawn i.i.d. from μ, algorithm $\mathcal{A}$ outputs a hypothesis $h \in \mathcal{H}$ satisfying*

$$\Pr\left[\left|\text{error}_\mu(h) - \min_{c\in\mathcal{H}}\{\text{error}_\mu(c)\}\right| \le \alpha\right] \ge 1-\beta,$$

where the probability is taken over the random choice of the examples in S according to μ and the random coins of the learner $\mathcal{A}$. If the output hypothesis h always satisfies $h \in \mathcal{H}$ then $\mathcal{A}$ is called a proper *agnostic PAC learner; otherwise, it is called an* improper *agnostic PAC learner.*

The following bound is due to (Vapnik and Chervonenkis, 1971; Blumer et al., 1989).

Theorem 14 (VC-Dimension Generalization Bound) *Let $\mathcal{H}$ and $\mathcal{P}$ be a concept class and a distribution over a domain X. Let $\alpha, \beta > 0$, and*

$$m \ge \frac{80}{\alpha}\left(\text{VC}(\mathcal{H})\ln\left(\frac{16}{\alpha}\right) + \ln\left(\frac{2}{\beta}\right)\right).$$

Suppose that we draw an unlabeled sample $S = (x_i)_{i=1}^m$, where x_i are drawn i.i.d. from $\mathcal{P}$. Then,

$$\Pr[\exists c, h \in \mathcal{H} \text{ s.t. } \text{error}_\mathcal{P}(h,c) > \alpha \ \wedge\ \text{error}_S(h) < \alpha/2] \le \beta.$$

The next theorem, due to (Vapnik and Chervonenkis, 1971; Anthony and Bartlett, 2009; Anthony and Shawe-Taylor, 1993), handles (in particular) the agnostic case.

Theorem 15 (VC-Dimension Agnostic Generalization Bound) *There exists a constant γ such that for every domain X, every concept class $\mathcal{H}$ over the domain X, and every distribution μ over the domain $X \times \{0,1\}$: For a sample $S = (x_i, y_i)_{i=1}^m$ where*

$$m \ge \gamma\frac{\text{VC}(\mathcal{H}) + \ln(\frac{1}{\beta})}{\alpha^2}$$

and $\{(x_i, y_i)\}$ are drawn i.i.d. from μ, it holds that

$$\Pr\left[\exists\, h \in \mathcal{H} \text{ s.t. } \left|\text{error}_\mu(h) - \text{error}_S(h)\right| \ge \alpha\right] \le \beta.$$

Notice that in Theorem 15 the sample complexity is proportional to $\frac{1}{\alpha^2}$, as opposed to $\frac{1}{\alpha}$ in Theorem 14.

Private Learning. Consider an algorithm $\mathcal{A}$ in the probably approximately correct (PAC) model of Valiant (1984). We say that $\mathcal{A}$ is a *private* learner if it also satisfies differential privacy w.r.t. its training data.

Definition 16 (Private PAC Learning (Kasiviswanathan et al., 2011)) *Let $\mathcal{A}$ be an algorithm that gets an input $S = (z_1, \ldots, z_m)$, where each z_i is a labeled example. Algorithm $\mathcal{A}$ is an (ε, δ)-*differentially private (α, β)-accurate PAC learner with sample complexity *m for a class $\mathcal{H}$ over X if*

PRIVACY. *Algorithm $\mathcal{A}$ is (ε, δ)-differentially private (as in Definition 9);*

Utility. *Algorithm $\mathcal{A}$ is an (α, β)-*accurate PAC learner *for $\mathcal{H}$ with sample complexity m (as in Definition 11).*

Note that the utility requirement in the above definition is an average-case requirement, as the learner is only required to do well on typical samples. In contrast, the privacy requirement is a worst-case requirement that must hold for every pair of neighboring databases (no matter how they were generated).

The following definition and lemma are taken from Bun et al. (2015).

Definition 17 (Empirical Learner) *Algorithm $\mathcal{A}$ is an (α, β)-accurate empirical learner for a class $\mathcal{H}$ over X with sample complexity m if for every $c \in \mathcal{H}$ and for every sample S of size m that is labeled by c, algorithm $\mathcal{A}$ outputs a hypothesis $h \in H$ satisfying*

$$\Pr[\mathrm{error}_S(c, h) \le \alpha] \ge 1 - \beta.$$

Lemma 18 (Bun et al. (2015)) *Suppose $\mathcal{A}$ is an (ε, δ)-differentially private (α, β)-accurate PAC learner for a concept class $\mathcal{H}$ with sample complexity m. Let $\mathcal{A}'$ be an algorithm, whose input sample S contains $9m$ randomly labeled examples. Further assume that $\mathcal{A}'$ samples with repetitions m labeled examples from S and returns the output of $\mathcal{A}$ on these examples. Then, $\mathcal{A}'$ is an (ε, δ)-differentially private (α, β)-accurate empirical learner for $\mathcal{H}$ with sample complexity $9m$. Clearly, if $\mathcal{A}$ is proper, then so is $\mathcal{A}'$.*

The Exponential Mechanism. We next describe the exponential mechanism of McSherry and Talwar (2007). Let X be a domain and H a set of solutions. Given a score function $q : X^* \times H \to \mathbb{N}$, and a database $S \in X^*$, the goal is to chooses a solution $h \in H$ approximately minimizing $q(S, h)$. The mechanism chooses a solution probabilistically, where the probability mass that is assigned to each solution h decreases exponentially with its score $q(S, h)$:

Algorithm 1 $\mathcal{A}_{\text{ExponentialMechanism}}$

Input: parameter ε, finite solution set H, database $S \in X^m$, and a sensitivity 1 score function q (i.e., $|q(D) - q(D')| \le 1$ for every neighboring $D, D' \in X^m$).

```
1. Randomly choose $h \in H$ with probability $\frac{\exp(-\varepsilon \cdot q(S,h)/2)}{\sum_{f \in H} \exp(-\varepsilon \cdot q(S,f)/2)}$.
2. Output $h$.
```

Proposition 19 (Properties of the Exponential Mechanism) *(i) The exponential mechanism is ε-differentially private. (ii) Let $\hat{e} \triangleq \min_{f \in H}\{q(S, f)\}$ and $\Delta > 0$. The exponential mechanism outputs a solution h such that $q(S, h) \ge (\hat{e} + \Delta m)$ with probability at most $|H| \cdot \exp(-\varepsilon \Delta m/2)$.*

Kasiviswanathan et al. (2011) showed that the exponential mechanism can be used as a generic private learner for a class $\mathcal{H}$. Specifically, when used with the score function $q(S, h) = |\{i : h(x_i) \neq y_i\}| = m \cdot \mathrm{error}_S(h)$, the exponential mechanism outputs a "bad" hypothesis $h \in \mathcal{H}$ such that $\mathrm{error}_S(h) > \min_{f \in \mathcal{H}}\{\mathrm{error}_S(f)\} + \Delta$ with probability at most $|\mathcal{H}| \cdot \exp(-\varepsilon \Delta m/2)$. This results in a generic private proper-learner for every finite concept class $\mathcal{H}$, with sample complexity $O_{\alpha,\beta,\varepsilon}(\log |\mathcal{H}|)$.

Generalization Properties of Differentially Private Algorithms. In this paper we use the fact that differential privacy implies generalization (Dwork et al., 2015; Bassily et al., 2016; Rogers et al., 2016; Feldman and Steinke, 2017; Nissim and Stemmer, 2017; Jung et al., 2020): differentially private learning algorithms satisfy that their empirical loss is typically close to their population loss. We use the following variant of this result, which is a multiplicative version that applies also to the case that $\varepsilon > 1$ (as needed in this paper).

Theorem 20 (DP Generalization – Multiplicative version (Dwork et al., 2015; Bassily et al., 2016; Feldman and Steinke, 2017; Nissim and Stemmer, 2017)) *Let $\mathcal{A}$ be an (ε, δ)-differentially private algorithm that operates on a database of $S \in X^n$ and outputs a predicate $\mathrm{test} : X \to \{0, 1\}$. Let $\mathcal{P}$ be a distribution over X and S be a database containing n i.i.d. elements from $\mathcal{P}$. Then,*

$$\Pr_{\substack{S \in_{\mathcal{P}} X^n,\\ \mathrm{test} \leftarrow \mathcal{A}(S)}} \left[\mathop{\mathbb{E}}_{x \in_{\mathcal{P}} X}[\mathrm{test}(x)] > e^{2\varepsilon} \left(\frac{\sum_{x \in S} \mathrm{test}(x)}{n} + \frac{10}{\varepsilon n} \log\left(\frac{1}{\varepsilon \delta n}\right)\right)\right] < O\left(\frac{\varepsilon \delta n}{\log(\frac{1}{\varepsilon \delta n})}\right).$$

Organization

We refer the reader to the appendix for the technical sections of this paper.

Acknowledgments

We thank Adam Klivans, Roi Livni, and Kobbi Nissim for insightful discussions.

References

Jacob D. Abernethy, Chansoo Lee, Audra McMillan, and Ambuj Tewari. Online learning via differential privacy. *CoRR*, abs/1711.10019, 2017. URL http://arxiv.org/abs/1711.10019.

Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite Littlestone dimension. In *Proceedings of the 51st Annual ACM Symposium on the Theory of Computing*, STOC '19, New York, NY, USA, 2019. ACM.

Martin Anthony and John Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47(3):207–217, 1993.

Matin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 2009. ISBN 9780521118620. URL http://books.google.co.il/books?id=UH6XRoEQ4h8C.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM Symposium on the Theory of Computing*, STOC '16, pages 1046–1059, New York, NY, USA, 2016. ACM.

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 *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 461–477, Philadelphia, PA, USA, 2015. SIAM.

Amos Beimel, Shay Moran, Kobbi Nissim, and Uri Stemmer. Private center points and learning of halfspaces. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 269–282. PMLR, 2019a. URL http://proceedings.mlr.press/v99/beimel19a.html.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of pure private learners. *Journal of Machine Learning Research*, 20(146):1–33, 2019b. URL http://jmlr.org/papers/v20/18-269.html.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT 2009 – The 22nd Conference on Learning Theory*, 2009. URL http://www.cs.mcgill.ca/%7Ecolt2009/papers/032.pdf#page=1.

Alina Beygelzimer, Satyen Kale, and Haipeng Luo. Optimal and adaptive algorithms for online boosting. In *International Conference on Machine Learning*, pages 2323–2331, 2015.

Siddharth Bhaskar. Thicket density. Technical Report arXiv:1702.03956, ArXiV, 2017.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

Nataly Brukhim, Xinyi Chen, Elad Hazan, and Shay Moran. Online agnostic boosting via regret minimization. *CoRR*, abs/2003.01150, 2020.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 634–649, Washington, DC, USA, 2015. IEEE Computer Society.

Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. *CoRR*, abs/2003.00563, 2020. URL https://arxiv.org/abs/2003.00563.

Hunter Chase and James Freitag. Model theory and combinatorics of banned sequences, 2018.

Hunter Chase and James Freitag. Model theory and machine learning. *The Bulletin of Symbolic Logic*, 25(03):319–332, Feb 2019. ISSN 1943-5894. doi: 10.1017/bsl.2018.71. URL http://dx.doi.org/10.1017/bsl.2018.71.

Shang-Tse Chen, Hsuan-Tien Lin, and Chi-Jen Lu. An online boosting algorithm with theoretical justifications. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1873–1880, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

R. M. Dudley. Central limit theorems for empirical measures. *Ann. Probab.*, 6(6):899–929, 12 1978. doi: 10.1214/aop/1176995384. URL https://doi.org/10.1214/aop/1176995384.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Proceedings of the 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, EUROCRYPT '06, pages 486–503, Berlin, Heidelberg, 2006a. Springer.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006b. Springer.

Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '10, pages 51–60, Washington, DC, USA, 2010. IEEE Computer Society.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

Vitaly Feldman and Thomas Steinke. Generalization for adaptively-chosen estimators via stable median. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 728–757. PMLR, 2017. URL http://proceedings.mlr.press/v65/feldman17a.html.

Alon Gonen, Elad Hazan, and Shay Moran. Private learning implies online learning: An efficient reduction. *NeurIPS*, 2019.

R. E. Greenwood and A. M. Gleason. Combinatorial relations and chromatic graphs. *Canadian Journal of Mathematics*, 7:1–7, 1955. doi: 10.4153/CJM-1955-001-4.

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

Matthew Joseph, Jieming Mao, Seth Neel, and Aaron Roth. The role of interactivity in local differential privacy. In *FOCS*, 2019.

Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. In Thomas Vidick, editor, *11th Innovations in Theoretical Computer Science Conference, ITCS 2020, January 12-14, 2020, Seattle, Washington, USA*, volume 151 of *LIPIcs*, pages 31:1–31:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020. doi: 10.4230/LIPIcs.ITCS.2020.31. URL https://doi.org/10.4230/LIPIcs.ITCS.2020.31.

Haim Kaplan, Katrina Ligett, Yishay Mansour, Moni Naor, and Uri Stemmer. Privately learning thresholds: Closing the exponential gap. In *Conference on Learning Theory, COLT 2020*, 2020a.

Haim Kaplan, Yishay Mansour, Uri Stemmer, and Eliad Tsfadia. Private learning of halfspaces: Simplifying the construction and reducing the sample complexity. *CoRR*, abs/2004.07839, 2020b. URL https://arxiv.org/abs/2004.07839.

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

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. In *Proceedings of the 30th Annual Symposium on Foundations of Computer Science*, SFCS '89, page 256–261, USA, 1989. IEEE Computer Society. ISBN 0818619821. doi: 10.1109/SFCS.1989.63487. URL https://doi.org/10.1109/SFCS.1989.63487.

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

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 94–103, Washington, DC, USA, 2007. IEEE Computer Society.

Kobbi Nissim and Uri Stemmer. Personal communication, 2017.

F. P. Ramsey. On a problem of formal logic. *Proceedings of the London Mathematical Society*, s2-30(1):264–286, 1930. doi: 10.1112/plms/s2-30.1.264. URL https://londmathsoc.onlinelibrary.wiley.com/doi/abs/10.1112/plms/s2-30.1.264.

Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 487–494, Washington, DC, USA, 2016. IEEE Computer Society.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. ISSN 0033-295X. doi: 10.1037/h0042519. URL http://dx.doi.org/10.1037/h0042519.

N. Sauer. On the density of families of sets. *J. Comb. Theory, Ser. A*, 13:145–147, 1972. ISSN 0097-3165. doi: 10.1016/0097-3165(72)90019-2.

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.

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

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.

Appendix A. Closure of Littlestone Classes

In this section we study closure properties for Littlestone classes. We begin in Appendix A.1 with a rather simple (and tight) analysis of the behavior of the Littlestone and Threshold dimension under unions. Then, in Appendix A.2 we prove our main results in this part (Theorems 3 and 4) that bound the variability of the Littlestone and Thresholds dimension under arbitrary compositions.

A.1. Closure Under Unions

We begin with two basic bounds on the variability of the Littlestone/Threshold dimension under union. Note that here $\mathcal{H}_1 \cup \mathcal{H}_2$ denotes the usual union: $\mathcal{H}_1 \cup \mathcal{H}_2 = \{h : h \in \mathcal{H}_1 \text{ or } h \in \mathcal{H}_2\}$. These bounds are useful as they allows us to reduce a bound on the dimension of $G(\mathcal{H}_1, \mathcal{H}_2)$ for arbitrary $\mathcal{H}_1, \mathcal{H}_2$ to the case where $\mathcal{H}_1 = \mathcal{H}_2$ (because $G(\mathcal{H}_1, \mathcal{H}_2) \subseteq G(\mathcal{H}, \mathcal{H})$ for $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$).

Observation 21 (Threshold Dimension Under Union) *Let $\mathcal{H}_1, \mathcal{H}_2 \subseteq \{0,1\}^X$ be hypothesis classes with $T(\mathcal{H}_i) = t_i$. Then,*

$$T(\mathcal{H}_1 \cup \mathcal{H}_2) \leq t_1 + t_2.$$

Moreover, this bound is tight: for every t_1, t_2, there are classes $\mathcal{H}_1, \mathcal{H}_2$ with Threshold dimension t_1, t_2 respectively such that $T(\mathcal{H}_1 \cup \mathcal{H}_2) = t_1 + t_2$.

Proof For the upper bound, observe that if $h_1, \ldots, h_m \in \mathcal{H}_1 \cup \mathcal{H}_2$ threshold-shatters the sequence $x_1, \ldots, x_m$ then $\{h_i : h_i \in \mathcal{H}_j\}$ threshold-shatters $\{x_i : h_i \in \mathcal{H}_j\}$ for $j \in \{1,2\}$. For the lower bound, set $X = [t_1 + t_2]$, $\mathcal{H}_1 = \{h_i : i \leq t_1\}$, and $\mathcal{H}_2 = \{h_i : t_1 < i \leq t_1 + t_2\}$, where $h_i(j) = 1$ if and only if $i \leq j$. ■

Proposition 22 (Littlestone Dimension Under Union) *Let $\mathcal{H}_1, \mathcal{H}_2 \subseteq \{0,1\}^X$ be hypothesis classes with $\mathrm{Ldim}(\mathcal{H}_i) = d_i$. Then,*

$$\mathrm{Ldim}(\mathcal{H}_1 \cup \mathcal{H}_2) \leq d_1 + d_2 + 1.$$

Moreover, this bound is tight: for every d_1, d_2, there are classes $\mathcal{H}_1, \mathcal{H}_2$ with Littlestone dimension d_1, d_2 respectively such that $\mathrm{Ldim}(\mathcal{H}_1 \cup \mathcal{H}_2) = d_1 + d_2 + 1$.

Proof There are several ways to prove this statement. One possibility is to use the realizable online mistake-bound setting (Littlestone, 1987) and argue that $\mathcal{H}_1 \cup \mathcal{H}_2$ can be learned with at most $d_1 + d_2 + 1$ mistakes in this setting. We will present here an alternative inductive argument, which may be of independent interest. Towards this end, it is convenient to define the depth of the empty tree as -1, and that of a tree consisting of one vertex (leaf) as 0.

Consider a shattered tree T of depth $d = \mathrm{Ldim}(\mathcal{H}_1 \cup \mathcal{H}_2)$ with leaves labelled $\mathcal{H}_1$ and $\mathcal{H}_2$ in the obvious way. Recall the notion of a subtree in Theorem 7, and let $x \leq \mathrm{Ldim}(\mathcal{H}_1)$ be the maximum depth of a complete binary subtree all whose leaves are $\mathcal{H}_1$ leaves, and $y \leq \mathrm{Ldim}(\mathcal{H}_2)$ the maximum depth of a subtree all whose leaves are $\mathcal{H}_2$-leaves. Similarly, let x_L, y_L denote the maximum depth of a $\mathcal{H}_1$-subtree and a $\mathcal{H}_2$-subtree in the tree rooted at the left child of the root of T, and let x_R, y_R be the same for the tree rooted at the right child.

It suffices to show that $x + y \geq d - 1$: clearly $x \geq \max(x_L, x_R)$ and also $x \geq \min(x_L, x_R) + 1$ thus $x \geq (x_L + x_R)/2 + 1/2$. Similarly $y \geq (y_L + y_R)/2 + 1/2$, hence

$$x + y \geq \frac{x_L + y_L}{2} + \frac{x_R + y_R}{2} + 1$$

and this gives by induction on d (starting with $d = 0$ or 1) that $x + y \geq d - 1$ as required.

To see that this bound is tight, pick $n \geq d_1 + d_2 + 1$ and set

$$\mathcal{H}_1 = \Big\{h : [n] \to \{0,1\} \ : \sum_i h_i \leq d_1\Big\} \quad \text{and} \quad \mathcal{H}_1 = \Big\{h : [n] \to \{0,1\} \ : \sum_i h_i \geq n - d_2\Big\}.$$

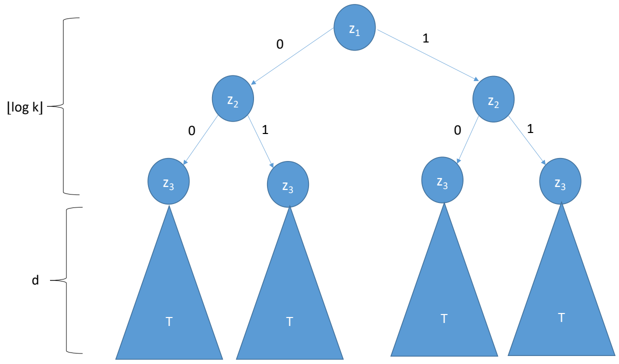

Figure 1: An illustration of the tree shattered by $\mathcal{H}'$ in the construction in Theorem 23. In this illustration $\lfloor \log k \rfloor$ equals 3.

One can verify that $\mathrm{Ldim}(\mathcal{H}_i) = d_i$, for $i = 1, 2$ and that $\mathrm{Ldim}(\mathcal{H}_1 \cup \mathcal{H}_2) = d_1 + d_2 + 1$, as required (in fact, even the VC dimension of $\mathcal{H}_1 \cup \mathcal{H}_2$ is $d_1 + d_2 + 1$). ■

Theorem 22 implies that $\mathrm{Ldim}(\cup_{i=1}^{k} \mathcal{H}_i) = O(k \cdot d)$ provided that $\mathrm{Ldim}(\mathcal{H}_i) \leq d$ for al i, and that this inequality can be tight when $k = 2$. Theorem 23 shows that for a larger k this bound can be significantly improved to $3d + 3 \log k$. This proposition demonstrates a difference with the threshold dimension. Indeed, while the bound above scales logarithmically with k, in the case of the threshold dimension a linear dependence in k is necessary: set $X = [k \cdot t]$, $\mathcal{H}_i = \{h_j : (i-1) \cdot t < ji \leq i \cdot t\}$, where $h_i(j) = 1$ if and only if $i \leq j$. Thus, $\mathrm{Ldim}(\mathcal{H}_i) = t$ for all i and $\mathrm{Ldim}(\cup_{i=1}^{k} \mathcal{H}_i) = k \cdot t >> t + \log k$.

Proposition 23 (Littlestone Dimension Under Multiple Unions) *Let $\mathcal{H}_1, \ldots, \mathcal{H}_k \subseteq \{0,1\}^X$ be hypothesis classes with $\mathrm{Ldim}(\mathcal{H}_i) \leq d$. Then,*

$$\mathrm{Ldim}(\bigcup_{i=1}^{k} \mathcal{H}_i) \leq 3d + 3 \log k.$$

Moreover, this bound is tight up to a constant factor: for every k, there are classes $\mathcal{H}_1, \ldots, \mathcal{H}_k$ with $\mathrm{Ldim}(\mathcal{H}_i) \leq d$ such that $\mathrm{Ldim}(\cup_i \mathcal{H}_i) \geq d + \lfloor \log k \rfloor$.

Proof We begin with the lower bound: pick any class $\mathcal{H} \subseteq \{0,1\}^X$ with Littlestone dimension d, and let T be a tree of depth d that is shattered by $\mathcal{H}$. Pick $\lfloor \log k \rfloor$ new points $z_1, \ldots, z_{\lfloor \log k \rfloor} \notin X$, and extend the domain X to $X' = X \cup \{z_1, \ldots, z_{\lfloor \log k \rfloor}\}$. Define $\mathcal{H}' \subseteq \{0,1\}^{X'}$ by extending each $h \in \mathcal{H}$ to the z_i's in each of the $k' = 2^{\lfloor \log k \rfloor}$ possible ways. (So, each $h \in \mathcal{H}$ has k' copies in $\mathcal{H}'$, one for each possible pattern on the z_i's.) Thus, $\mathcal{H}'$ is a union of k' copies of $\mathcal{H}$, one copy for each boolean pattern on the z_i's. In particular, $\mathcal{H}'$ is the union of k' classes with Littlestone dimension d. Also note that $\mathrm{Ldim}(\mathcal{H}') \geq \lfloor \log k \rfloor + d$, as witnessed by the tree that is illustrated in Figure 1.

The upper bound is based on a multiplicative-weights argument. Recall that the Littlestone dimension equals the optimal number of mistakes performed by a deterministic online learner in the mistake-bound model (i.e. online learning when the sequence of input examples is labelled by some $h \in \mathcal{H}$). Thus, it suffices to demonstrate an online learner for $\cup_{i=1}^{k}\mathcal{H}_i$ that makes at most $3d + 3\log k$ mistakes. Pick for every $\mathcal{H}_i$ an online learner $\mathcal{A}_i$ that makes at most d mistakes on input sequences consistent with $\mathcal{H}_i$. We set the online learning algorithm $\mathcal{A}$ for $\mathcal{H} = \cup_{i=1}^{k}\mathcal{H}_i$ to be *the Weighted Majority Algorithm* of Littlestone and Warmuth (1989) with the k experts being the algorithms $\mathcal{A}_1, \dots, \mathcal{A}_k$. Now, consider an input sequence $S = (x_1, y_1), \dots, (x_T, y_T)$ consistent with $\mathcal{H}$. Thus, S is consistent with $\mathcal{H}_i$ for some $i \le k$ and therefore $\mathcal{A}_i$ makes at most d mistakes on it. Thus, by the multiplicative weights analysis (see e.g. Corollary 2.1 in Littlestone and Warmuth (1989)), the number of mistakes $\mathcal{A}$ makes on S is at most

$$\frac{\log k + d\log\frac{1}{\beta}}{\log\frac{2}{1+\beta}},$$

where $0 \le \beta < 1$ is multiplicative factor that discounts the weights of wrong experts. The upper bound follows by setting $\beta = 1/2$. ■

A.2. Closure Under Composition

A.2.1. Littlestone Dimension

Proof of Theorem 3. We will first show that for an odd k, the majority-vote $G = \mathsf{MAJ}_k$ satisfies

$$\mathrm{Ldim}(\mathsf{MAJ}_k(\mathcal{H}_1, \dots, \mathcal{H}_k)) \le \tilde{O}(k^2 \cdot d). \tag{3}$$

(Recall that $d = \max_i\{\mathrm{Ldim}(\mathcal{H}_i)\}$.) Then, we use this to argue that for any G,

$$\mathrm{Ldim}(G(\mathcal{H}_1, \dots, \mathcal{H}_k)) \le \tilde{O}(2^{2k}k^2 d). \tag{4}$$

We start with proving (3). Let $\mathcal{H} = \cup_{i=1}^{k}\mathcal{H}_i$ and $\mathcal{H}_{\mathrm{maj}} = \mathsf{MAJ}_k(\mathcal{H}, \dots, \mathcal{H})$. Since $\mathsf{MAJ}_k(\mathcal{H}_1, \dots, \mathcal{H}_k) \subseteq \mathcal{H}_{\mathrm{maj}}$, it suffices to show that $\mathrm{Ldim}(\mathcal{H}_{\mathrm{maj}}) \le \tilde{O}(k^2 d)$. We use online boosting towards this end. Online boosting (in the realizable setting) is an algorithmic framework that allows to transform a weak online learner for $\mathcal{H}$ with a non-trivial mistake-bound of $(1/2 - \gamma)T + R(T)$, where $R(T) = o(T)$ is a sublinear regret function, to a strong online learner with a vanishing mistake-bound of $O(R(T)/\gamma)$. Online boosting has been studied by several works (e.g. Chen et al. (2012); Beygelzimer et al. (2015); Brukhim et al. (2020)). We use here the variant given by Brukhim et al. (2020) (see Theorem 2 there)[7].

We want to construct a weak learner for $\mathcal{H}_{\mathrm{maj}}$. Which weak learner to use? Recall that by Ben-David et al. (2009) (see Equation (1)) there exists an *agnostic* online learning algorithm W for $\mathcal{H}$ whose (expected) regret bound is

$$R(T) = O(\sqrt{\mathrm{Ldim}(\mathcal{H})T\log T}).$$

We claim that W is a weak learner for $\mathcal{H}_{\mathrm{maj}}$ with mistake-bound

$$(1/2 - 1/k) \cdot T + R(T). \tag{5}$$

[7] The bound in Theorem 2 in (Brukhim et al., 2020) contains an additional term that depends on N, the number of copies of the weak learner that are used by the boosting algorithm. Since here we are only concerned with the number of mistakes, we can eliminate this term by letting $N \to \infty$.

To prove this, it suffices to show that for every sequence of examples $(x_1, y_1), \ldots, (x_T, y_T)$ that is consistent with $\mathcal{H}_{\mathrm{maj}}$ there exists $h \in \mathcal{H}$ that makes at most $(1/2 - 1/k) \cdot T$ mistakes on it. Indeed, let $h_1, \ldots, h_k$ such that $y_t = \mathsf{MAJ}_k(h_1(x_t), \ldots, h_k(x_t))$ for $t \leq T$. Thus, on every example (x_t, y_t) at most $1/2 - 1/k$ fraction of the h_i's make a mistake on it. By averaging, this implies that one of the h_i makes at most $(1/2 - 1/k)T$ mistakes in total, and (5) follows.

Now, by applying online boosting with W as a weak learner, we obtain an algorithm with a mistake-bound of at most

$$O\Big(\frac{R(T)}{1/k}\Big) = O\Big(k\sqrt{\mathrm{Ldim}(\mathcal{H})T\log T}\Big).$$

Thus, since the Littlestone dimension characterizes the optimal mistake-bound, letting $D = \mathrm{Ldim}(\mathcal{H}_{\mathrm{maj}})$, we get that

$$(\forall T \geq D) : D \leq O\Big(k\sqrt{\mathrm{Ldim}(\mathcal{H})T\log T}\Big),$$

and in particular $D \leq O\Big(k\sqrt{\mathrm{Ldim}(\mathcal{H})D\log D}\Big)$, which implies that

$$\begin{aligned} D &\leq \tilde{O}(k^2\mathrm{Ldim}(\mathcal{H})) \\ &\leq \tilde{O}(k^2 d + k^2\log k) && \text{(Theorem 23)} \\ &= \tilde{O}(k^2 d), \end{aligned}$$

and finishes the proof of (3).

We next set out to prove (4). The idea is to represent an arbitrary G using a formula that only uses majority-votes and negations. Let $G : \{0,1\}^k \to \{0,1\}$ be an arbitrary boolean function. It is a basic fact that G can be represented by a Disjunctive Normal Form (DNF) as follows:

$$G = \bigvee_{i=1}^{m}\Big(\bigwedge_{j=1}^{k} z_{i,j}\Big),$$

where each $z_{i,j} \in \{x_j, \neg x_j\}$, and $m \leq 2^k$. Now, note that

$$\bigwedge_{j=1}^{k} z_{i,j} = \mathsf{MAJ}_{2k-1}(z_{i,1}, \ldots, z_{i,k}, \underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{k-1}),$$

and similarly

$$\bigvee_{i=1}^{m}\Big(\bigwedge_{j=1}^{k} z_{i,j}\Big) = \mathsf{MAJ}_{2m-1}\Big(\bigwedge_{j=1}^{k} z_{1,j}, \ldots, \bigwedge_{j=1}^{k} z_{m,j}, \underbrace{\mathbf{1}, \ldots, \mathbf{1}}_{m-1}\Big).$$

Thus, $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ can be written as $\mathsf{MAJ}_{2m-1}(\mathcal{H}'_1, \ldots, \mathcal{H}'_{2m-1})$, where for $i > m$, $\mathcal{H}'_i = \{h_0\}$ is the class that contains the all-zero function h_0, and for $i \leq m$,

$$\mathcal{H}'_i = \mathsf{MAJ}_{2k-1}(\mathcal{H}''_{i,1}, \ldots, \mathcal{H}''_{i,2k-1})$$

such that each class $\mathcal{H}''_{i,j}$ is either $\mathcal{H}_t$ or its negation $\neg\mathcal{H}_t$ for some $t \leq k$, or $\mathcal{H}''_{i,j}$ is the class $\{h_1\}$ that contains the all-one function. We now apply (3) to conclude that $\mathrm{Ldim}(\mathcal{H}'_i) = \tilde{O}(k^2 d)$ for all $i \leq m$, and that

$$\mathrm{Ldim}(G(\mathcal{H}_1, \ldots, \mathcal{H}_k)) = \tilde{O}\big(m^2(k^2 d)\big) = \tilde{O}(2^{2k}k^2 d)$$

as required. ■

A.2.2. THRESHOLD DIMENSION

Proof of Theorem 3. We begin with the upper bound. Let $T(G(\mathcal{H}_1,\ldots,\mathcal{H}_k))=n$. It suffices to show that if $n\geq 2^{4k4^k\cdot t}$ then there is $i\leq k$ such that $T(\mathcal{H}_i)\geq t$. By assumption, there are $x_1,x_2,\ldots,x_n\in X$ and functions $h_{i,j}\in\mathcal{H}_j$, for $1\leq i\leq n, 1\leq j\leq k$ such that

$$(\forall i,j\leq n): G(h_{i,1},h_{i,2},\ldots,h_{i,k})(x_j)=1 \iff i\leq j.$$

Construct a coloring of the edges of the complete graph on $[n]$ by 4^k colors as follows: for each $1\leq p<q\leq n$, the color of the edge $\{p,q\}$ is given by the following ordered sequence of $2k$ bits:

$$\begin{aligned}\bigl(&h_{p,1}(x_q),h_{p,2}(x_q),\ldots,h_{p,k}(x_q),\\ &h_{q,1}(x_p),h_{q,2}(x_p),\ldots,h_{q,k}(x_p)\bigr).\end{aligned}$$

By Ramsey Theorem (Ramsey, 1930), if $n\geq(4^k)^{2t\cdot 4^k}=2^{4k4^k\cdot t}$ then there is a monochromatic set $A\subseteq[n]$ of size $|A|=2t$.[8] Denote the elements of A by

$$A=\{i_1<j_1<i_2<j_2<\cdots<i_t<j_t\},$$

and let $u=(u_1,\ldots,u_k)$, $v=(v_1,\ldots,v_k)$ such that the color of every pair in A is

$$\begin{aligned}(&v_1,v_2,\ldots,v_k,\\ &u_1,u_2,\ldots,u_k).\end{aligned}$$

Thus, for every pair $p,q\leq d$ and every $r\leq k$:

$$h_{i_p,r}(x_{j_q})=\begin{cases}v_r & p\leq q\\ u_r & p>q.\end{cases}$$

We claim that $u\neq v$: indeed, $x_{j_1},x_{j_2},x_{j_3},\ldots,x_{j_t}$ is threshold-shattered by the functions

$$G(h_{i_1,1},h_{i_1,2},\ldots,h_{i_1,k}),G(h_{i_2,1},h_{i_2,2},\ldots,h_{i_2,k}),\cdots,G(h_{i_d,1},h_{i_d,2},\ldots,h_{i_t,k}).$$

Thus,

$$\begin{aligned}p\leq q &\implies G(v)=G(h_{i_p,1},\ldots,h_{i_p,k})(x_q)=1,\\ p>q &\implies G(u)=G(h_{i_p,1},\ldots,h_{i_p,k})(x_q)=0.\end{aligned}$$

Therefore, $v\in G^{-1}(0)$ and $u\in G^{-1}(1)$ and in particular $u\neq v$. Pick an index r so that $u_r\neq v_r$. Therefore, for every $p,q\leq t$:

$$h_{i_p,r}(x_{j_q})=\begin{cases}v_r & p\leq q\\ u_r & p>q,\end{cases}\text{ and }\quad v_r\neq u_r.$$

This shows that either $x_1,\ldots,x_t$ is threshold shattered by $\mathcal{H}_r$ (if $v_r=1,u_r=0$), or $x_t,\ldots,x_1$ is thresholds shattered by $\mathcal{H}_r$ (if $v_r=0,u_r=1$); in either way, the threshold dimension of $\mathcal{H}_r$ is at least t. This completes the proof of the upper bound.

[8] We use here the following basic bound: if $n\geq c^{r\cdot c}$, then for every coloring of the edges of the complete graph K_n in c colors there exists a monochromatic set of size r. This follows, e.g. from Corollary 3 in Greenwood and Gleason (1955).

Lower Bound. We next prove the lower bound. Let $m = 2^{\lfloor t/5 \rfloor}$, and construct $\mathcal{H} \subseteq \{0,1\}^m$ randomly as follows: $\mathcal{H}$ consists of $2m$ random functions

$$\mathcal{H} = \{f_1, \ldots, f_m, g_1, \ldots, g_m\},$$

where for each i set $f_i(j) = g_j(j) = 0$ for $j > i$, and for $j \leq i$, pick uniformly at random one of f_i, g_i, set it to be 1 in position j and set the other to be 0 in position j. All of the above $\binom{m-1}{2}$ random choices are done independently. By construction, $\{h_1 \vee h_2 : h_1, h_2 \in \mathcal{H}\}$ threshold-shatters the sequence $1, 2, \ldots, m$ with probability 1 and hence has threshold dimension at least m. It suffices to show that with a positive probability it holds that

$$T(\mathcal{H}) \leq 2k, \tag{6}$$

where $k = (2 + \frac{1}{\log m}) \log m = 2\lfloor t/5 \rfloor + 1$. Indeed, $2k = 4\lfloor t/5 \rfloor + 2 \leq t$ whenever $t \geq 6$.

We set out to prove (6). Consider the following event:

$\mathcal{E}$:= *There exist no* $x_1, \ldots, x_k \in [m]$, $h_1, \ldots, h_k \in \mathcal{H}$ *such that* $h_i(x_j) = 1$ *for all* $i, j \leq k$.

Note that $\mathcal{E}$ implies that $T(\mathcal{H}) \leq 2k$ and therefore it suffices to show that $\Pr[E] > 0$. Towards this end we use a union bound: we define a family of "bad" events whose total sum of probabilities is less than one with the property that if none of the bad events occurs then $\mathcal{E}$ occurs. The bad events are defined as follows: for any pair of subsets $A, B \subseteq [m]$ of size $|A| = |B| = k$, let $\mathcal{B}_{A,B}$ denote the event

$\mathcal{B}_{A,B}$:= *"For every* $i \in A$ *there exists* $r_i \in \{f_i, g_i\}$ *such that* $r_i(j) = 1$ *for all* $j \in B$."

Note that indeed $\neg\mathcal{E}$ implies $\mathcal{B}_{A,B}$ for some A, B and thus it suffices to show that with a positive probability none of the $\mathcal{B}_{A.B}$ occurs. We claim that

$$\Pr[\mathcal{B}_{A,B}] \leq 2^{-k(k-1)}.$$

Indeed, for a fixed $i \in A$, the probability that one of f_i, g_i equals to 1 on all $j \in B$ is at most $2^{-(k-1)}$. By independence, the probability that the latter simultaneously holds for every $i \in A$ is at most $2^{-k(k-1)}$. Thus, the probability that $\mathcal{B}_{A,B}$ occurs for at least one pair A, B is at most

$$\binom{m}{k}^2 2^{-k(k-1)} < 2^{2k \log m - k(k-1)} \leq 1,$$

where the last inequality holds because $k = (2 + \frac{1}{\log m}) \log m$. ■

Appendix B. The Private Relabeling Procedure

In this section we describe a private relabeling procedure $\mathcal{A}_{\text{Relabel}}$ (discussed in Section 1.1). This procedure is used both for constructing a private agnostic learner and for proving closure of private learning. We start by explaining the difficulties in designing the relabeling procedure and how we overcome them. We then provide a formal description of the relabeling procedure in $\mathcal{A}_{\text{Relabel}}$ and prove that it can be used to construct a private algorithm that produces a hypothesis that has good generalization properties; this is done by presenting an algorithm $\mathcal{A}_{\text{RelabelAndLearn}}$.

Let $\mathcal{H}$ be a hypothesis class, and suppose that we have a differentially private learning algorithm $\mathcal{A}$ for $\mathcal{H}$ for the realizable setting. That is, $\mathcal{A}$ is guaranteed to succeed in its learning task whenever it is given a labeled database that is consistent with some hypothesis in $\mathcal{H}$. Now suppose that we are given a database S sampled from some distribution $\mathcal{P}$ on X and labeled by some concept c^* (not necessarily in $\mathcal{H}$). So, S might *not* be consistent with any hypothesis in $\mathcal{H}$, and we cannot directly apply $\mathcal{A}$ on S. Heuristically, one might first *relabel* the database S using some function from $\mathcal{H}$, and then apply $\mathcal{A}$ on the relabeled database. Can we argue that such a paradigm would satisfy differential privacy, or is it the case that the relabeling process "vaporises" the privacy guarantees of algorithm $\mathcal{A}$?

Building on a result of Beimel et al. (2015), we show that it *is* possible to relabel the database before applying algorithm $\mathcal{A}$ while maintaining differential privacy. As we mentioned in the introduction, the relabeling procedure of Beimel et al. (2015) instantiates the exponential mechanism in order to choose a hypothesis h that is (almost) as close as possible to the original labels in S, uses this hypothesis to relabel the database, and applies the given differentially private algorithm $\mathcal{A}$ on the relabeled database to obtain an outcome f.

Now we want to argue that f has low generalization error. We known (by the guarantees of the exponential mechanism) that the hypothesis h with which we relabeled S has a relatively small empirical error on S (close to the lowest possible error). Via standard VC arguments, we also know that h has a relatively small generalization error. Therefore, in order to show that the returned hypothesis f has low generalization error, it suffices to show that $\mathrm{error}_{\mathcal{P}}(f, h)$ is small. This might seem trivial at first sight: Since as $\mathcal{A}$ is a PAC learner, and since it is applied on a database S labeled by the hypothesis $h \in \mathcal{H}$, it must (w.h.p.) return a hypothesis f with small error w.r.t. h. Is that really the case?

The difficulty with formalizing this argument is that $\mathcal{A}$ is only guaranteed to succeed in identifying a good hypothesis when it is applied on an *i.i.d.* sample from some underlying distribution. This is *not* true in our case. Specifically, we first sampled the database S from the underlying distribution, then *based on* S, we identified the hypothesis h and relabeled S using h. For all we know, $\mathcal{A}$ might completely fail when executed on such a database (not sampled in an i.i.d. manner).[9] Therefore, before applying $\mathcal{A}$ on the relabeled database, we subsample i.i.d. elements from it, and apply $\mathcal{A}$ on this newly sampled database. Now we know that $\mathcal{A}$ is applied on an i.i.d. sampled database, and so, by the utility guarantees of $\mathcal{A}$, the hypotheses f and h are close w.r.t. the underlying distribution. However, this subsampling step changes the distribution from which the inputs of $\mathcal{A}$ are coming from. This distribution is no longer $\mathcal{P}$ (the original distribution from which S was sampled), rather it is the uniform distribution on the empirical sample S. This means that what we get from the utility guarantees of $\mathcal{A}$ is that $\mathrm{error}_S(f, h)$ is small. We need to show that $\mathrm{error}_{\mathcal{P}}(f, h)$ is small.

If $\mathcal{A}$ is a *proper* learner, then f is itself in $\mathcal{H}$, and hence, using standard VC arguments, the fact that $\mathrm{error}_S(f, h)$ is small implies that $\mathrm{error}_{\mathcal{P}}(f, h)$ is small. However, if $\mathcal{A}$ is an improper learner, then this argument breaks because f might come from a different hypothesis class with a much larger VC dimension.

[9] To illustrate this issue, suppose that the learner $\mathcal{A}$ first checks to see if *exactly* half of the elements in its input sample are labeled by 1 and *exactly* half of them are labeled by 0. If that it the case, then $\mathcal{A}$ fails. Otherwise, $\mathcal{A}$ identifies a hypothesis f with small empirical error. When the input elements are correctly sampled (i.i.d. from some underlying distribution) the probability that exactly half of them will be labeled as 0 is low enough such that $\mathcal{A}$ remains a valid learning algorithm. However, if we first sample the elements, and then choose a hypothesis that evaluates to 1 on exactly half of them, then this breaks the utility guarantees of $\mathcal{A}$ completely.

To overcome this difficulty, we will instead relate $\mathrm{error}_S(f,h)$ and $\mathrm{error}_{\mathcal{P}}(f,h)$ using the generalization properties of differential privacy. These generalization properties state that if a predicate t was identified using a differentially private algorithm, then (w.h.p.) the empirical average of this predicate and its expectation over the underlying distribution are close. More specifically, we would like to consider the predicate $(h{\oplus}f)(x) = h(x) \oplus f(x)$, which would complete our mission because the empirical average of that predicate on S is $\mathrm{error}_S(f,h)$, and its expectation over $\mathcal{P}$ is $\mathrm{error}_{\mathcal{P}}(f,h)$. However, while f is indeed the outcome of a differentially private computation, h *is not*, and we cannot directly apply the generalization properties of differential privacy to this predicate. Specifically, our relabeling procedure does not reveal the chosen hypothesis h.

We overcome this issue by introducing the following conceptual modification to the relabeling procedure. Let us think about the input database S as *two* databases $S = D{\circ}T$. In the relabeling procedure we still relabel all of S using h. We show that (a small variant of) this relabeling procedure still satisfies differential privacy w.r.t. D *even if the algorithm publicly releases the relabeled database* T. This works in our favour because given the relabeled database T we can identify a hypothesis $h' \in \mathcal{H}$ that agrees with it, and by standard VC arguments we know that $\mathrm{error}_{\mathcal{P}}(h,h')$ is small (since both h,h' come from $\mathcal{H}$). In addition, h' is computed by post-processing the relabeled database T which we can view as the result of a private computation w.r.t. D. Therefore, we can now use the generalization properties of differential privacy to argue that $\mathrm{error}_D(f,h) \approx \mathrm{error}_{\mathcal{P}}(f,h)$, which would allow us to complete the proof. We remark that the conceptual modification of treating S as two databases $S = D{\circ}T$ is crucial for our analysis. We do not know if the original relabeling procedure of Beimel et al. (2015) can be applied when $\mathcal{A}$ is an improper learner.

In Algorithm 2 we formally describe $\mathcal{A}_{\mathrm{Relabel}}$. We next provide an informal description of the algorithm. Let $\mathcal{H}$ be a hypothesis class, and let q be a score function. Algorithm $\mathcal{A}_{\mathrm{Relabel}}$ takes two input databases $D, T \in (X \times \{0,1\})^*$, where the labels in D and T are arbitrary. The algorithm *relabels* D and T using a hypothesis $h \in \mathcal{H}$ with near optimal score $q(D,h)$. The output of this algorithm is the two relabeled databases $\tilde{D}$ and $\tilde{T}$. Observe that algorithm $\mathcal{A}_{\mathrm{Relabel}}$ is clearly *not* differentially private, since it outputs its input database (with different labels). Before formally presenting algorithm $\mathcal{A}_{\mathrm{Relabel}}$, we introduce the following definition.

Definition 24 *Let X be a domain and let $\mathcal{H}$ be a class of functions over X. A function $q : (X \times \{0,1\})^* \times \mathcal{H} \to \mathbb{R}$ has* matched-sensitivity *k if for every $S \in (X \times \{0,1\})^*$, every $(x,y),(x'y') \in X \times \{0,1\}$, and every $h,h' \in \mathcal{H}$ that agree on every element of S we have that*

$$\left|q\left(S \cup \{(x,y)\}, h\right) - q\left(S \cup \{(x',y')\}, h'\right)\right| \le k.$$

In words, a score function q has low *matched-sensitivity* if given "similar" databases it assigns "similar" scores to "similar" solutions. Note that if a function q has matched-sensitivity 1, then in particular, it has (standard) sensitivity (at most) 1.

Example 1 *Let $\mathcal{H}$ be a concept class over X. Then, the score function $q(S,h)$ that takes a labeled database $S \in (X \times \{0,1\})^*$ and a concept $h \in H$ and returns the number of errors h makes on S has matched-sensitivity at most 1.*

We next present an algorithm $\mathcal{A}_{\mathrm{RelabelAndLearn}}$ and analyze its properties. This algorithm abstracts parts of the techniques we apply in the following sections, and is used for unifying the proofs of privacy and correctness of in these two sections. We start with an informal description of algorithm $\mathcal{A}_{\mathrm{RelabelAndLearn}}$. The algorithm first applies the relabeling algorithm $\mathcal{A}_{\mathrm{Relabel}}$

```
Algorithm 2 $\mathcal{A}_{\text{Relabel}}$
Global parameters:
  • $\mathcal{H}$ is a concept class over a domain $X$,
  • $q : (X \times \{0,1\})^* \times \mathcal{H} \to \mathbb{R}$ is a score function with matched-sensitivity at most 1 (see Definition 24), which given a labeled database assigns scores to concepts from $\mathcal{H}$,
Inputs: Labeled databases $D, T \in (X \times \{0,1\})^*$. We denote $S = D \circ T$.
  1. Let $P = \{p_1, \dots, p_\ell\}$ be the set of all unlabeled points appearing at least once in $S$.
  2. Let $H = \Pi_{\mathcal{H}}(P) = \{h|_P : h \in \mathcal{H}\}$, where $h|_P$ denotes the restriction of $h$ to $P$ (i.e., $H$ contains all patterns of $\mathcal{H}$ when restricted to $P$).
  3. Choose $h \in H$ using the exponential mechanism with privacy parameter $\varepsilon$=1, score function $q$, solution set $H$, and the database $D$.
  4. Relabel $S$ using $h$, and denote the relabeled databases as $S^h = D^h \circ T^h$. That is, if $D = (x_i, y_i)_{i=1}^d$ then $D^h = (x_i, h(x_i))_{i=1}^d$, and similarly with $T^h$.
  5. Output $D^h, T^h$.
```

and then applies a private algorithm to the relabeled database. For the analysis of our algorithms in the sequence, $\mathcal{A}_{\text{RelabelAndLearn}}$ also publishes part of the relabeled database. We prove that $\mathcal{A}_{\text{RelabelAndLearn}}$ guarantees differential privacy w.r.t. to the part of the database that it did not publish.

```
Algorithm 3 $\mathcal{A}_{\text{RelabelAndLearn}}$
Global parameters:
  • $\mathcal{H}$ is a concept class over a domain $X$,
  • $q : (X \times \{0,1\})^* \times \mathcal{H} \to \mathbb{R}$ is a score function with matched-sensitivity at most 1, which given a labeled database assigns scores to concepts from $\mathcal{H}$,
  • $\mathcal{A}$ is a an $(\varepsilon, \delta)$-differentially private algorithm.
Inputs: Labeled databases $D, V, W \in (X \times \{0,1\})^*$. We denote $S = D \circ V \circ W$.
  1. Execute $\mathcal{A}_{\text{Relabel}}(D, V \circ W)$ with score function $q$ and hypothesis class $\mathcal{H}$ to obtain relabeled databases $\tilde{D}, \tilde{V}, \tilde{W}$.
  2. Let $\overline{h}$ be a hypothesis in $\mathcal{H}$ that is consistent with $\tilde{V}$.
  3. Output $\mathcal{A}(S), \tilde{V}, \overline{h}$.
```

In Lemma 25, we analyze the privacy properties of algorithm $\mathcal{A}_{\text{RelabelAndLearn}}$.

Lemma 25 *Let $\mathcal{A}$ be an (ε, δ)-differentially private algorithm and q be a score function with matched-sensitivity 1. Then, for every V, algorithm*

$$\mathcal{A}^V_{\text{RelabelAndLearn}}(D, W) = \mathcal{A}_{\text{RelabelAndLearn}}(D, V, W)$$

is $(\varepsilon+3, 4e\delta)$-differentially private w.r.t. $D{\circ}W$. In particular, $\mathcal{A}(\mathcal{A}_{\text{Relabel}}(D,T))$ is $(\varepsilon+3, 4e\delta)$-differentially private.

Proof Fix a database V, and let $D_1{\circ}W_1$ and $D_2{\circ}W_2$ be two neighboring databases. We assume that $D_1{\circ}W_1$ and $D_2{\circ}W_2$ differ on their D portion, so that $W_1 = W_2 = W$ and $D_1 = D \cup \{(\mathsf{p}_1, y_1)\}$ and $D_2 = D \cup \{(\mathsf{p}_2, y_2)\}$. The analysis for the other case is essentially identical. Consider the executions of $\mathcal{A}_{\text{Relabel}}$ on $S_1 = D_1{\circ}V{\circ}W$ and on $S_2 = D_2{\circ}V{\circ}W$, and denote by H_1, P_1 and by H_2, P_2 the elements H, P as they are in the executions of algorithm $\mathcal{A}_{\text{Relabel}}$ on S_1 and on S_2.

Since S_1 and S_2 are neighbors, it follows that $|P_1 \setminus P_2| \leq 1$ and $|P_2 \setminus P_1| \leq 1$. Let $K = P_1 \cap P_2$. Since every pattern in $\Pi_{\mathcal{H}}(K)$ has at most two extensions in $\Pi_{\mathcal{H}}(H_t)$, we get that for every $t \in \{1,2\}$.

$$|\Pi_{\mathcal{H}}(K)| \leq |\Pi_{\mathcal{H}}(P_t)| \leq 2|\Pi_{\mathcal{H}}(K)|.$$

Thus, $|H_1| \leq 2|H_2|$ and similarly $|H_2| \leq 2|H_1|$.

More specifically, for every $t \in \{1,2\}$ and every pattern $h \in \Pi_C(K)$ there are either one or two (but not more) patterns in H_t that agree with h on K. We denote these one or two patterns by $h_t^{(0)}$ and $h_t^{(1)}$, which may be identical if only one unique pattern exists. By the fact that q has matched-sensitivity at most 1, for every $t_1, t_2 \in \{1,2\}$ and every $b_1, b_2 \in \{0,1\}$ we have that

$$|q(D_1, h_{t_1}^{(b_1)}) - q(D_2, h_{t_2}^{(b_1)})| = |q(D \cup \{(\mathsf{p}_1, y_1)\}, h_1) - q(D \cup \{(\mathsf{p}_2, y_2)\}, h_2)| \leq 1,$$

where the last inequality is because $h_{t_1}^{(b_1)}$ and $h_{t_2}^{(b_2)}$ agree on every point in D and because q has matched-sensitivity at most 1.

For every $h \in \Pi_{\mathcal{H}}(K)$ and $t \in \{1,2\}$, let $w_{t,h}$ be the probability that the exponential mechanism chooses either $h_t^{(0)}$ or $h_t^{(1)}$ in Step (3) of the execution of $\mathcal{A}_{\text{Relabel}}$ on S_i. We get that for every $h \in \Pi_C(K)$,

$$\begin{aligned}
w_{1,h} &\leq \frac{\exp(\frac{1}{2}\cdot q(D_1, h_1^{(0)})) + \exp(\frac{1}{2}\cdot q(D_1, h_1^{(1)}))}{\sum_{f\in\Pi_{\mathcal{H}}(P_1)} \exp(\frac{1}{2}\cdot q(D_1, f))} \\
&\leq \frac{\exp(\frac{1}{2}\cdot q(D_1, h_1^{(0)})) + \exp(\frac{1}{2}\cdot q(D_1, h_1^{(1)}))}{\sum_{f\in\Pi_{\mathcal{H}}(K)} \exp(\frac{1}{2}\cdot q(D_1, f_1^{(0)}))} \\
&\leq \frac{\exp(\frac{1}{2}\cdot [q(D_2, h_2^{(0)})+1]) + \exp(\frac{1}{2}\cdot [q(D_2, h_2^{(1)})+1])}{\frac{1}{2}\sum\limits_{f\in\Pi_{\mathcal{H}}(K)} \Big(\exp(\frac{1}{2}[q(D_2, h_2^{(0)})-1]) + \exp(\frac{1}{2}[q(D_2, h_2^{(1)})-1])\Big)} \\
&\leq 2e\cdot\frac{\exp(\frac{1}{2}\cdot q(D_2, h_2^{(0)})) + \exp(\frac{1}{2}\cdot q(D_2, h_2^{(1)}))}{\sum_{f\in\Pi_{\mathcal{H}}(P_2)} \exp(\frac{1}{2}\cdot q(D_2, f))} \\
&\leq 4e\cdot w_{2,h}.
\end{aligned}$$

We are now ready to conclude the proof. For every $h \in \Pi_{\mathcal{H}}(K)$, let I_t be the event that the exponential mechanism chooses in Step (3) of the execution on S_t either $h_t^{(0)}$ or $h_t^{(1)}$ and h_t be the random variable denoting the pattern that the exponential mechanism chooses in Step (3) of the execution on S_t conditioned on the event I_t. Observe that S^{h_0} and S^{h_1} are distributions on neighboring databases; thus, applying the differentially private $\mathcal{A}$ on them satisfies differential privacy, i.e., for every possible sets of outputs F of $\mathcal{A}$:

$$\Pr\left[\mathcal{A}\left(S_1^{h_1}\right) \in F\right] \le e^{\varepsilon} \Pr\left[\mathcal{A}\left(S_2^{h_2}\right) \in F\right] + \delta.$$

Recall that algorithm $\mathcal{A}_{\mathrm{RelabelAndLearn}}$ returns *three* outcomes: the relabeled database V^h, hypothesis $\overline{h}$ that is consistent with V^h, and the output of algorithm $\mathcal{A}$. As $\overline{h}$ is computed from V^h, we can consider it as post-processing and ignore it, and assume for the the privacy analysis that $\mathcal{A}_{\mathrm{RelabelAndLearn}}$ only has two outputs: V^h and the output of algorithm $\mathcal{A}$. Also recall that the database V is fixed, and observe that once the hypothesis h is fixed (in Step (3) of algorithm $\mathcal{A}_{\mathrm{Relabel}}$), the relabeled database V^h is also fixed. Furthermore, for every $h \in \Pi_{\mathcal{H}}(K)$ we have that $V^{h_t^{(0)}} = V^{h_t^{(1)}}$, since $h_t^{(0)}$ and $h_t^{(1)}$ agree on all of V.

Let $F \subseteq (X \times \{0,1\})^* \times R$ be a set of possible outcomes for algorithm $\mathcal{A}_{\mathrm{RelabelAndLearn}}$, where R is the range of algorithm $\mathcal{A}$. For every h we denote

$$F_h = \left\{ r \in R : (V^h, r) \in F \right\}.$$

Observe that for every $h \in \Pi_C(K)$ we have that

$$F_{h_1^{(0)}} = F_{h_1^{(1)}} = F_{h_2^{(1)}} = F_{h_2^{(2)}} = F_h,$$

because $h_1^{(0)}, h_1^{(1)}, h_2^{(0)}, h_2^{(1)}$ agree on all points in V. We calculate,

$$\begin{aligned}
\Pr[\mathcal{A}_{\mathrm{RelabelAndLearn}}(S_1) \in F] &= \sum_{h \in \Pi_{\mathcal{H}}(K)} w_{1,h} \cdot \Pr\left[\mathcal{A}_{\mathrm{RelabelAndLearn}}(S_1) \in F \middle| I_t\right] \\
&= \sum_{h \in \Pi_{\mathcal{H}}(K)} w_{1,h} \cdot \Pr\left[\mathcal{A}\left(S_1^{h_1}\right) \in F_{h_1}\right] \\
&\le \sum_{h \in \Pi_{\mathcal{H}}(K)} 4e\, w_{2,h} \cdot \left(e^{\varepsilon} \Pr\left[\mathcal{A}\left(S_2^{h_2}\right) \in F_{h_2}\right] + \delta\right) \\
&\le e^{\varepsilon+3} \cdot \Pr[\mathcal{B}(S_2) \in F] + 4e\delta.
\end{aligned}$$

■

The next claim proves that $\mathcal{A}_{\mathrm{Relabel}}$ returns a hypothesis whose score is close to the hypothesis with smallest score in the class $\mathcal{H}$.

Claim 26 *Fix α and β, and let $S = D \circ T \in (X \times \{0,1\})^*$ be a labeled database such that*

$$|D| \ge \frac{2}{\alpha} \ln\left(\frac{1}{\beta}\right) + \frac{2\,\mathrm{VC}(\mathcal{H})}{\alpha} \ln\left(\frac{e|S|}{\mathrm{VC}(\mathcal{H})}\right).$$

Consider the execution of $\mathcal{A}_{\mathrm{Relabel}}$ on S, and let h denote the hypothesis chosen on Step (3). With probability at least $(1-\beta)$ we have that $q(D,h) \le \min_{c \in \mathcal{H}}\{q(D,c)\} + \alpha|D|$. In particular, assuming that $|D| \ge |S|/2$, it suffices that

$$|D| \ge \frac{4}{\alpha} \ln\left(\frac{1}{\beta}\right) + \frac{10\,\mathrm{VC}(\mathcal{H})}{\alpha} \ln\left(\frac{20e}{\alpha}\right).$$

Proof Note that by Sauer's lemma,

$$|H| \;=\; |\Pi_{\mathcal{H}}(P)| \leq \left(\frac{e|P|}{\mathrm{VC}(\mathcal{H})}\right)^{\mathrm{VC}(\mathcal{H})} \leq \left(\frac{e|S|}{\mathrm{VC}(\mathcal{H})}\right)^{\mathrm{VC}(\mathcal{H})}.$$

As H contains all patterns of $\mathcal{H}$ restricted to S, the set H contains a pattern f^* s.t. $q(D,f^*) = \min_{c\in\mathcal{H}}\{q(D,c)\}$. Hence, Proposition 19 (properties of the exponential mechanism) ensures that the probability of the exponential mechanism choosing an h s.t. $q(D,h) > \min_{c\in\mathcal{H}}\{q(D,c)\} + \alpha$ is at most

$$|H| \cdot \exp(-\frac{\alpha|D|}{2}) \leq \left(\frac{e|S|}{\mathrm{VC}(\mathcal{H})}\right)^{\mathrm{VC}(\mathcal{H})} \cdot \exp(-\frac{\alpha|D|}{2}),$$

which is at most β whenever $|D| \geq \frac{2}{\alpha}\ln(\frac{1}{\beta}) + \frac{2\,\mathrm{VC}(\mathcal{H})}{\alpha}\ln\left(\frac{e|S|}{\mathrm{VC}(\mathcal{H})}\right)$. ■

Let f denote the hypothesis returned by $\mathcal{A}$ and let h be a hypothesis consistent with the pattern chosen on Step (3) of $\mathcal{A}_{\mathrm{Relabel}}$. The next lemma relates the generalization error $\mathrm{error}_{\mathcal{P}}(f,h)$ to the empirical error $\mathrm{error}_D(f,h)$.

Lemma 27 *Fix α and β, and let μ be a distribution on $X \times \{0,1\}$ and $\mathcal{P}$ be the marginal distribution on unlabeled examples from X. Furthermore, let $S = D{\circ}V{\circ}W \in (X \times \{0,1\})^*$ be database sampled i.i.d. from μ such that*

$$|V| \geq O\left(\frac{\mathrm{VC}(\mathcal{H})\ln\left(\frac{1}{\alpha}\right) + \ln\left(\frac{1}{\beta}\right)}{\alpha}\right),$$

and

$$|D| \geq O\left(\frac{\mathrm{VC}(\mathcal{H}) + \ln\left(\frac{1}{\beta}\right)}{\alpha^2}\right).$$

Consider the execution of $\mathcal{A}_{\mathrm{RelabelAndLearn}}$ on S, let $h \in \mathcal{H}$ be a hypothesis consistent with the pattern chosen on Step (3) of $\mathcal{A}_{\mathrm{Relabel}}$ and assume that $\mathcal{A}$ outputs some hypothesis f. With probability at least $1 - O(\beta + \delta|D|)$ we have that

$$\mathrm{error}_{\mathcal{P}}(f,h) \leq O(\mathrm{error}_D(f,h) + \alpha).$$

Proof Let $\overline{h}$ be the third output of $\mathcal{A}_{\mathrm{RelabelAndLearn}}$, i.e., a hypothesis from $\mathcal{H}$ that is consistent with V^h. Since $\overline{h}$ and h agree on V and $|V|$ is big enough, by Theorem 14, with probability at least $1-\beta$ (over sampling V),

$$\mathrm{error}_{\mathcal{P}}(\overline{h},h) \leq \alpha. \tag{7}$$

Since $|D|$ is big enough, by Theorem 15 (applied to $\mathcal{H} \oplus \mathcal{H}$ and the distribution μ that samples x according to $\mathcal{P}$ and labels it with 0), with probability at least $1-\beta$,

$$\mathrm{error}_D(\overline{h},h) \leq \mathrm{error}_{\mathcal{P}}(\overline{h},h) + \alpha \leq 2\alpha. \tag{8}$$

We will now use the generalization properties of differential privacy to argue that $\mathrm{error}_{\mathcal{P}}(f,h)$ is small. By Lemma 25, algorithm $\mathcal{A}_{\mathrm{RelabelAndLearn}}$ is $(O(1),O(\delta))$-differentially private w.r.t. the database D. In addition, by post-processing the outcomes of $\mathcal{A}_{\mathrm{RelabelAndLearn}}$ (the hypotheses f and $\overline{h}$) we can define the following predicate $\mathrm{test}: X\times\{0,1\}\to\{0,1\}$ where $\mathrm{test}(x,y)=1$ if $\overline{h}(x)\neq f(x)$, and $\mathrm{test}(x,y)=0$ otherwise. Now observe that

$$\mathrm{error}_{\mathcal{P}}(f,\overline{h})=\Pr_{x\sim\mathcal{P}}[\overline{h}(x)\neq f(x)]=\mathop{\mathbb{E}}_{x\sim\mathcal{P}}[\mathbb{1}\{\overline{h}(x)\neq f(x)\}]=\mathop{\mathbb{E}}_{(x,y)\sim\mu}[\mathrm{test}(x,y)]. \tag{9}$$

Similarly,

$$\mathrm{error}_D\left(f,\overline{h}\right)=\frac{1}{|D|}\sum_{(x,y)\in D}\mathbb{1}\left\{\overline{h}(x)\neq f(x)\right\}=\frac{1}{|D|}\sum_{(x,y)\in D}\mathrm{test}(x,y). \tag{10}$$

Recall that test is the result of a private computation on the database D (obtained as a post-processing of the outcomes of $\mathcal{A}_{\mathrm{RelabelAndLearn}}$). Also observe that since $\mathcal{A}_{\mathrm{RelabelAndLearn}}$ is $(O(1),O(\delta))$-differentially private, it is in particular, $\left(O(1),O\left(\delta+\frac{\beta}{|D|}\right)\right)$-differentially private for every choice of β and $|D|$. Hence, assuming $|D|\geq O\left(\frac{1}{\alpha}\log\frac{1}{\beta}\right)$, Theorem 20 (the generalization properties of differential privacy) states that with probability at least $1-O(\delta|D|+\beta)$,

$$\begin{aligned}\mathop{\mathbb{E}}_{(x,y)\sim\mu}[\mathrm{test}(x,y)]&\leq O\left(\frac{1}{|D|}\sum_{(x,y)\in D}\mathrm{test}(x,y)+\frac{1}{|D|}\log\left(\frac{1}{\delta|D|+\beta}\right)\right)\\&\leq O\left(\frac{1}{|D|}\sum_{(x,y)\in D}\mathrm{test}(x,y)+\frac{1}{|D|}\log\left(\frac{1}{\beta}\right)\right)\\&\leq O\left(\frac{1}{|D|}\sum_{(x,y)\in D}\mathrm{test}(x,y)+\alpha\right).\end{aligned} \tag{11}$$

So, by (9), (10), and (11), with probability at least $1-O(\beta+\delta|D|)$

$$\mathrm{error}_{\mathcal{P}}(f,\overline{h})\leq O(\mathrm{error}_D(f,\overline{h})+\alpha). \tag{12}$$

Thus, the next inequality, which concludes the proof, holds with probability $1-O(\beta+\delta|D|)$.

$$\begin{aligned}\mathrm{error}_{\mathcal{P}}(f,h)&=\mathrm{error}_{\mathcal{P}}(f,\overline{h})+\mathrm{error}_{\mathcal{P}}(\overline{h},h)\\&\leq O(\mathrm{error}_D(f,\overline{h})+\alpha) && \text{(by (7) and (12))}\\&\leq O(\mathrm{error}_D(f,h)+\mathrm{error}_D(h,\overline{h})+\alpha)\\&\leq O(\mathrm{error}_D(f,h)+\alpha) && \text{(by (8)).}\end{aligned}$$

■

Appendix C. Private PAC Implies Private Agnostic PAC

In this section we show that private learning implies private agnostic learning (with essentially the same sample complexity) even for improper learning algorithms. Algorithm $\mathcal{A}_{\text{PrivateAgnostic}}$, the agnostic algorithm for a class $\mathcal{H}$, first applies algorithm $\mathcal{A}_{\text{Relabel}}$ on the data and relabels the sample using a hypothesis in $\mathcal{H}$ that has close to minimal empirical error, and then uses the private learning algorithm (after sub-sampling) to learn the relabeled database.

```
Algorithm 4 $\mathcal{A}_{\text{PrivateAgnostic}}$
Inputs: A labeled sample $S \in (X \times \{0,1\})^m$.
Auxiliary algorithm: A private learner $\mathcal{A}$ for the concept class $\mathcal{H}$.
  1. Partition $S$ into $S = D \circ T$, where $|D| = |T| = |S|/2$.
  2. Execute $\mathcal{A}_{\text{Relabel}}$ with input $D, T$ and score function $q(D,h) = |D| \cdot \text{error}_D(h)$ to obtain
     relabeled databases $\tilde{D}, \tilde{T}$.
  3. Execute a private empirical learner on $\tilde{D}$: Choose $|D|/9$ samples with replacements from
     $\tilde{D}$. Denote the resulting database by $Q$ and let $f \leftarrow \mathcal{A}(Q)$.
  4. Return $f$.
```

Theorem 28 (Theorem 6 Restated) *Let $0 < \alpha, \beta, \delta < 1$, $m \in \mathbb{N}$, and $\mathcal{A}$ be a $(1, \delta)$-differentially private (α, β)-accurate PAC learner for $\mathcal{H}$ with sample complexity m. Then, $\mathcal{A}_{\text{PrivateAgnostic}}$ is an $(O(1), O(\delta))$-differentially private $(O(\alpha), O(\beta + \delta n))$-accurate* agnostic *learner for $\mathcal{H}$ with sample complexity*

$$n = O\left(m + \frac{1}{\alpha^2}\left(\text{VC}(\mathcal{H}) + \log\frac{1}{\beta}\right)\right).$$

Proof The privacy properties of the algorithm are straightforward. Specifically, by Lemma 18, Step (3) the algorithm (applying $\mathcal{A}$ on a subsample from $\tilde{D}$) satisfies $(O(1), O(\delta))$-differential privacy. Algorithm $\mathcal{A}_{\text{PrivateAgnostic}}$ is, therefore, $O(1), O(\delta))$-differentially private by Lemma 25. In particular, if $\mathcal{A}$ is $(1, 0)$-differentially private then $\mathcal{A}_{\text{PrivateAgnostic}}$ is $(O(1), 0)$-differentially private.

As for the utility analysis, fix a target distribution μ over $X \times \{0,1\}$, and denote

$$\Delta = \min_{c \in \mathcal{H}} \{\text{error}_\mu(c)\}.$$

Also let $\mathcal{P}$ denote the marginal distribution on unlabeled examples from X. Let S be a sample containing n i.i.d. samples from μ, and denote $S = D \circ T$ where $|D| = |T| = |S|/2$. By Theorem 15 (the agnostic VC generalization bound), assuming that $|S| \geq O\left(\frac{1}{\alpha^2}\left(\text{VC}(\mathcal{H}) + \ln\frac{1}{\beta}\right)\right)$, with probability at least $1 - \beta$ (over sampling S), the following event occur.

Event E_1 : $\forall c \in \mathcal{H}$ we have $|\text{error}_\mu(c) - \text{error}_D(c)| \leq \alpha$.

We continue with the analysis assuming that this event occurs, and show that (w.h.p.) the hypothesis f returned by the algorithm has low generalization error. Consider the execution of $\mathcal{A}_{\text{PrivateAgnostic}}$ on S. In Step (2) we apply algorithm $\mathcal{A}_{\text{Relabel}}$ to obtain the relabeled databases $\tilde{D}, \tilde{T}$. Let $h \in \mathcal{H}$

be a hypothesis extending the pattern used by algorithm $\mathcal{A}_{\text{Relabel}}$ to relabel these databases. By Claim 26, assuming that $|D|$ is big enough, with probability at least $1-\beta$ it holds that

$$\text{error}_D(h) \leq \min_{c\in\mathcal{H}}\{\text{error}_D(c)\} + \alpha. \tag{13}$$

In this case, by Event E_1 we have that

$$\text{error}_\mu(h) \leq \text{error}_D(h) + \alpha \leq \min_{c\in\mathcal{H}}\{\text{error}_D(c)\} + 2\alpha \leq \min_{c\in\mathcal{H}}\{\text{error}_\mu(c)\} + 3\alpha = \Delta + 3\alpha. \tag{14}$$

Recall that $\mathcal{A}$ is executed on the database Q containing $|\tilde{D}|/9$ i.i.d. samples from $\tilde{D}$. By Lemma 18, with probability at least $1-\beta$, the hypothesis f chosen in Step (3) satisfies

$$\text{error}_D(f,h) = \text{error}_{\tilde{D}}(f) \leq \alpha. \tag{15}$$

By Lemma 27 and (15) with probability at least $1 - O(\beta + |D|\delta)$

$$\text{error}_{\mathcal{P}}(f,h) \leq O(\text{error}_D(f,h) + \alpha) \leq O(\alpha). \tag{16}$$

Finally, by (14) and (16)

$$\text{error}_\mu(f) \leq \text{error}_{\mathcal{P}}(f,h) + \text{error}_\mu(h) \leq \Delta + O(\alpha).$$

■

Appendix D. Closure of Private Learning

In this section we prove Theorem 29 – if $\mathcal{H}_1,\ldots,\mathcal{H}_k$ are privately learnable, then $G(\mathcal{H}_1,\ldots,\mathcal{H}_k)$ is privately learnable.

Theorem 29 (Closure Theorem for Private Learning) *Let $G : \{0,1\}^k \to \{0,1\}$ be a boolean function and $\mathcal{H}_1,\ldots,\mathcal{H}_k \subseteq \{0,1\}^X$ be classes that are (ε,δ)-differentially private and (α,β)-accurate learnable by a possibly* improper *learning algorithms with sample complexity $m_i(\alpha,\beta,\varepsilon,\delta)$ respectively. Then, $G(\mathcal{H}_1,\ldots,\mathcal{H}_k)$ is $(O(1),O(\delta))$-private and $(O(\alpha),O(\beta+\delta m))$-accurate learnable with sample complexity*

$$m = O\left(\frac{k^3\,\text{VC}(G(\mathcal{H}_1,\ldots,\mathcal{H}_k)) + k^2\ln\left(\frac{k}{\beta}\right)}{\alpha^2} + \sum_{i=1}^{k} m_i\left(\frac{\alpha}{k},\frac{\beta}{k},1,\delta\right)\right).$$

To prove Theorem 29, we present $\mathcal{A}_{\text{ClosureLearn}}$ – a generic transformation of private learning algorithms $\mathcal{A}_1,\ldots,\mathcal{A}_k$ for the classes $\mathcal{H}_1,\ldots,\mathcal{H}_k$ respectively to a private learner for $G(\mathcal{H}_1,\ldots,\mathcal{H}_k)$. This transformation could be applied to proper as well as improper learners, and to a learners that preserves pure or approximate privacy. Given a labeled sample S of size N, algorithm $\mathcal{A}_{\text{ClosureLearn}}$ finds hypotheses $h_1,\ldots,h_k$ in steps, where in the i'th step, the algorithm finds a hypothesis h_i such that $h_1,\ldots,h_i$ have a completion $c_{i+1},\ldots,c_k$ to a hypothesis $G(h_1,\ldots,h_i,c_{i+1},\ \ldots,c_k)$ with small error (assuming that $h_1,\ldots,h_{i-1}$ have a good completion). In the i'th step, $\mathcal{A}_{\text{ClosureLearn}}$ relabels the input sample S so that the relabeled sample is realizable by $\mathcal{H}_i$. The relabeling h is chosen using $\mathcal{A}_{\text{Relabel}}$ in a way that guarantees completion to a hypothesis with small empirical error. That is, using an appropriate score-function in $\mathcal{A}_{\text{Relabel}}$ (i.e., in the exponential mechanism), it is guaranteed that for the hypotheses $h_1, \dots, h_{i-1}$ computed in the previous steps there are some $c_{i+1} \in \mathcal{H}_i, \dots, c_k \in \mathcal{H}_k$ such that the function $G(h_1, \dots, h_{i-1}, h, c_{i+1}, \dots, c_k)$ has a small loss with respect to the original sample S. The relabeled sample is fed (after subsampling) to the private algorithm $\mathcal{A}_i$ to produce a hypothesis h_i and then the algorithm proceeds to the next step $i + 1$.

Algorithm 5 $\mathcal{A}_{\text{ClosureLearn}}$

```
Input: A labeled sample $S \in (X \times \{0,1\})^N$, where $N$ will be fixed later.
Auxiliary Algorithms: Private learners $\mathcal{A}_1, \dots, \mathcal{A}_k$ for the class $\mathcal{H}_1, \dots, \mathcal{H}_k$ respectively.
  1. Partition $S$ into $k$ parts $S = S_1 \circ S_2 \circ \cdots \circ S_k$ – the size of the $S_i$ will be determined later.
  2. For every $i \in [k]$:
      (a) Partition $S_i$ into $S_i = D_i \circ T_i$, where $|D_i| = |T_i| = |S_i|/2$.
      (b) Execute $\mathcal{A}_{\text{Relabel}}$ with input $D, T$, hypothesis class $\mathcal{H}_i$, and score function
              $q(S_i, z) = |S_i| \cdot \min_{c_{i+1} \in \mathcal{H}_{i+1}, \dots, c_k \in \mathcal{H}_k} \text{error}_{S_i}(G(h_1, \dots, h_{i-1}, h, c_{i+1}, \dots, c_k)),$   (17)
          to obtain relabeled databases $\tilde{D}_i, \tilde{T}_i$.
      (c) Execute a private empirical learner on $\tilde{D}_i$:
          i. Choose $|D_i|/9$ samples with replacements from $\tilde{D}_i$. Denote the resulting database by $Q$.
          ii. Execute the private learner $\mathcal{A}_i$ on the sample $Q$ with accuracy parameters $(\alpha/k, \beta/k)$ and privacy parameters $(\varepsilon = 1, \delta)$. Let $h_i$ be its output.
  3. Output $c = G(h_1, \dots, h_k)$.
```

In Lemma 30, we analyze the privacy guarantees of $\mathcal{A}_{\text{ClosureLearn}}$.

Lemma 30 *Let $\varepsilon < 1$ and assume the algorithms $\mathcal{A}_1, \dots, \mathcal{A}_k$ are $(1, \delta)$-private. Then, $\mathcal{A}_{\text{ClosureLearn}}$ is $(\varepsilon, O(\delta))$-differentially private.*

Proof Fix $i \in [k]$ and consider the i'th step of the algorithm. By Lemma 18, Step (2c) of algorithm $\mathcal{A}_{\text{ClosureLearn}}$ (i.e., sub-sampling with replacement and executing a $(1, \delta)$-private algorithm) is $(1, \delta)$-differentially private. Thus, by Lemma 25, Steps (2b)–(2c) of algorithm $\mathcal{A}_{\text{ClosureLearn}}$ are $(O(1), O(\delta))$-differentially private. Since each step is executed on a disjoint set of examples, $\mathcal{A}_{\text{ClosureLearn}}$ is $(O(1), O(\delta))$-differentially private. ■

In the next lemma we prove that $\mathcal{A}_{\text{ClosureLearn}}$ is an accurate learner for the class $G(\mathcal{H}_1, \dots, \mathcal{H}_k)$.

Lemma 31 *Assume that $\mathcal{A}_1, \dots, \mathcal{A}_t$ are $(1, \delta)$-differentially private $(\alpha/k, \beta/k)$-accurate (possibly improper) learning algorithms for $\mathcal{H}_1, \dots, \mathcal{H}_k$ with sample complexity $m_i(\alpha/k, \beta/k, 1, \delta)$. If at each iteration i*

$$|S_i| \geq O\left(\frac{k^2\,\mathrm{VC}(G(\mathcal{H}_1, \dots, \mathcal{H}_k)) + k \ln\left(\frac{k}{\beta}\right)}{\alpha^2} + m_i\left(\frac{\alpha}{k}, \frac{\beta}{k}, 1, \delta\right)\right),$$

then with probability at least $1 - O(\beta + k\delta|S_i|)$ *we have that* $\mathrm{error}_{\mathcal{P}}(c) \leq O(\alpha)$*, where* c *is the hypothesis returned by* $\mathcal{A}_{\mathrm{ClosureLearn}}$ *on* S*.*

Proof Let $h_1, \ldots, h_k$ be the hypotheses that $\mathcal{A}_{\mathrm{ClosureLearn}}$ computes in Step (2c). We prove by induction that for every $i \in [k]$ with probability at least $1 - \frac{O(i)\cdot\beta}{k} + O(i \cdot \delta|S_i|)$ there exist $c_{i+1} \in \mathcal{H}_{i+1}, \ldots, c_k \in \mathcal{H}_k$ such that

$$\mathrm{error}_{\mathcal{P}}(G(h_1, \ldots, h_i, c_{i+1}, \ldots, c_k) \leq \frac{O(i)\cdot\alpha}{k}. \tag{18}$$

The induction basis for $i = 0$ is implied by the fact that the examples are labeled by some $G(c_1, \ldots, c_k)$ from $G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$. For the induction step, assume that there are $c_i \in \mathcal{H}_i, \ldots, c_k \in \mathcal{H}_k$ such that

$$\mathrm{error}_{\mathcal{P}}\Big(G(h_1, \ldots, h_{i-1}, c_i, c_{i+1}, \ldots, c_k)\Big) \leq \frac{O(i-1)\cdot\alpha}{k}.$$

We need to prove that with probability at least $1 - \frac{O(1)\cdot\beta}{k} - O(\delta|S_i|)$ there are $c'_{i+1} \in \mathcal{H}_{i+1}, \ldots, c'_k \in \mathcal{H}_k$ such that

$$\mathrm{error}_{\mathcal{P}}(G(h_1, \ldots, h_{i-1}, h_i, c'_{i+1}, \ldots, c'_k)) \leq \frac{O(i)\cdot\alpha}{k}.$$

Recall that each example in S, and hence in S_i, is chosen i.i.d. from the distribution in $\mathcal{P}$. Since

$$|S_i| \geq O\left(\frac{k^2\,\mathrm{VC}(G(\mathcal{H}_1, \ldots, \mathcal{H}_k)) + \ln\left(\frac{k}{\beta}\right)}{\alpha^2}\right), \tag{19}$$

by Theorem 15 applied to $G(\mathcal{H}_1, \ldots, \mathcal{H}_k) \oplus G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$, with probability at least $1 - \frac{\beta}{k}$ (over the sampling of S_i) the following event occurs:

Event E_1 : $\forall c \in G(\mathcal{H}_1, \ldots, \mathcal{H}_k)$ we have $|\mathrm{error}_{\mathcal{P}}(c) - \mathrm{error}_{S_i}(c)| \leq \frac{\alpha}{k}$.

We continue proving the induction step assuming that E_1 occurs. The proof of the induction step is as follows:

Since E_1 occurs:

$$\begin{aligned}
&\mathrm{error}_{S_i}(G(h_1, \ldots, h_{i-1}, c_i, c_{i+1}, \ldots, c_k)) \\
&\qquad \leq \mathrm{error}_{\mathcal{P}}(G(h_1, \ldots, h_{i-1}, c_i, c_{i+1}, \ldots, c_k)) + \frac{\alpha}{k} \\
&\qquad \leq \frac{(O(i-1)+1)\alpha}{k}.
\end{aligned} \tag{20}$$

By the definition of H, there is $h = h_{\mathrm{opt}} \in H$ that agrees with c_i on S_i, and therefore

$$q(S_i, h_{\mathrm{opt}}) \leq |S_i| \frac{(O(i-1)+1)\alpha}{k}.$$

By Claim 26, if

$$|S_i| \geq O\left(\frac{k}{\alpha}\ln\left(\frac{k}{\beta}\right) + \frac{k\,\mathrm{VC}(|\mathcal{H}_i|)}{\alpha}\ln\left(\frac{k}{\alpha}\right)\right), \tag{21}$$

then with probability at least $1-\frac{\beta}{k}$, the exponential mechanism returns $h \in H$ such that

$$q(S_i, h) \le q(S_i, h_{\text{opt}}) + |S_i|\frac{\alpha}{k} \le |S_i|\frac{(O(i-1)+2)\alpha}{k}.$$

We assume that the above event occurs, thus, the latter implies that there are $c'_{i+1}, \dots, c'_k$ such that

$$\text{error}_{S_i}(G(h_1, \dots, h_{i-1}, h, c'_{i+1}, \dots, c'_k)) \le \frac{(O(i-1)+2)\alpha}{k}. \tag{22}$$

Since E_1 occurs, by (22),

$$\begin{aligned}\text{error}_{\mathcal{P}}(G(h_1, \dots, h_{i-1}, h, c'_{i+1}, \dots, c'_k)) &\le \frac{\alpha}{k} + \text{error}_{S_i}(G(h_1, \dots, h_{i-1}, h, c'_{i+1}, \dots, c'_k)) \\ &\le \frac{(O(i-1)+3)\alpha}{k}. \end{aligned} \tag{23}$$

Since

$$|D_i^h| \ge 9m_i\Big(\frac{\alpha}{k}, \frac{\beta}{k}, 1, \delta\Big), \tag{24}$$

Lemma 18 implies that Step (2c) of $\mathcal{A}_{\text{ClosureLearn}}$ is an $(\frac{\alpha}{k}, \frac{\beta}{k})$ empirical learner and, therefore, with probability at least $1-\frac{\beta}{k}$

$$\text{error}_{D_i}(h, h_i) = \text{error}_{D_i^h}(h_i) \le \frac{\alpha}{k}. \tag{25}$$

Again, we assume in the rest of the proof that the above event occurs. By Lemma 27, since

$$|D_i| = \frac{|S_i|}{2} \ge O\left(\frac{k^2(\text{VC}(\mathcal{H}_i) + \ln\left(\frac{k}{\beta}\right))}{\alpha^2}\right) \tag{26}$$

with probability at least $1-\frac{O(\beta)}{k} - O(\delta|D_i|)$

$$\text{error}_{\mathcal{P}}(h, h_i) \le O\left(\text{error}_{D_i}(h, h_i) + \frac{\alpha}{k}\right).$$

Thus, by (25), with probability at least $1-\frac{O(\beta)}{k}$

$$\text{error}_{\mathcal{P}}(h, h_i) \le O\left(\frac{(O(i-1)+O(1))\alpha}{k}\right). \tag{27}$$

The latter, combined with (23), implies the induction step: with probability at least $1-\frac{O(\beta)}{k} - O(\delta|D_i|)$

$$\begin{aligned}&\text{error}_{\mathcal{P}}(G(h_1, \dots, h_{i-1}, h_i, c'_{i+1}, \dots, c'_k)) \\ &\quad\le \text{error}_{\mathcal{P}}(G(h_1, \dots, h_{i-1}, h, c'_{i+1}, \dots, c'_k)) + \text{error}_{\mathcal{P}}(h, h_i) \\ &\quad\le \frac{(O(i-1)+O(1))\alpha}{k} = \frac{O(i)\cdot\alpha}{k}.\end{aligned}$$

By (19), (21), (24), and (26), the sample complexity $|S_i|$ the i'th step is

$$O\left(\frac{k^2\,\mathrm{VC}(G(\mathcal{H}_1,\ldots,\mathcal{H}_k))+\ln\left(\frac{k}{\beta}\right)}{\alpha^2}+\frac{k}{\alpha}\ln\left(\frac{k}{\beta}\right)+m_i\left(\frac{\alpha}{k},\frac{\beta}{k},1,\delta\right)+\frac{k^2(\mathrm{VC}(\mathcal{H}_i)+\ln\left(\frac{k}{\beta}\right))}{\alpha^2}\right)$$
$$=O\left(\frac{k^2\,\mathrm{VC}(G(\mathcal{H}_1,\ldots,\mathcal{H}_k))+k\ln\left(\frac{k}{\beta}\right)}{\alpha^2}+m_i\left(\frac{\alpha}{k},\frac{\beta}{k},1,\delta\right)\right)$$

To conclude, by a union bound, $\mathcal{A}_{\text{ClosureLearn}}$ returns, with probability at least $1-O(\beta+\delta\sum_{i=1}^{k}|S_i|)$, a hypothesis $G(h_1,\ldots,h_k)$ with error less than $O(\alpha)$ with respect to the distribution $\mathcal{P}$. ∎

Proof of Theorem 29. **Proof** Theorem 29 follows from Lemmas 30 and 31. Specifically, by Lemma 31, to prove that $\mathcal{A}_{\text{ClosureLearn}}$ is $(O(\alpha),O(\beta+\delta m))$-accurate it suffices that

$$\sum_{i=1}^{k}|S_i|\geq\sum_{i=1}^{k}O\left(\frac{k^2\,\mathrm{VC}(G(\mathcal{H}_1,\ldots,\mathcal{H}_k))+k\log(\frac{k}{\beta})}{\alpha^2}+m_i\left(\frac{\alpha}{k},\frac{\beta}{k},1,\delta\right)\right).$$

By Lemma 30, $\mathcal{A}_{\text{ClosureLearn}}$ is $(O(1),O(\delta))$-differentially private. ∎

Remark 32 *Since each $\mathcal{A}_i$ is an (α,β)-accurate learning algorithm for the class $\mathcal{H}_1$,*

$$m_i\left(\frac{\alpha}{k},\frac{\beta}{k},1,\delta\right)=\Omega\left(\frac{k\,\mathrm{VC}(\mathcal{H}_i)}{\alpha}\right).$$

Furthermore, by the Sauer-Shelah-Perles Lemma, $\mathrm{VC}(G(\mathcal{H}_1,\ldots,\mathcal{H}_k)=\tilde{O}(\sum_{i=1}^{k}\mathrm{VC}(\mathcal{H}_i))$. Thus, the sample complexity of $\mathcal{A}_{\text{ClosureLearn}}$ is

$$\tilde{O}\left(\sum_{i=1}^{k}m_i\left(\frac{\alpha}{k},\frac{\beta}{k},1,\delta\right)\right)\cdot\mathrm{poly}(k,1/\alpha,\log(1/\beta)).$$

For constant k,α,β this is nearly tight. By using sub-sampling (see e.g., Kasiviswanathan et al. (2011); Beimel et al. (2014)), we can achieve $(\varepsilon,O(\delta))$-differential privacy by increasing the sample complexity by a factor of $O(1/\varepsilon)$. Furthermore, by using private boosting Dwork et al. (2010), one can start with a private algorithm that is, for example, $(1/4,\beta)$ accurate and get a private algorithm that is (α,β) by increasing the sample complexity by a factor of $O(1/\alpha)$, and by simple technique, one can boost β by increasing the sample complexity by a factor of $O(\log(1/\beta))$. Thus, we get an $(\varepsilon,O(\delta))$-differentially private (α,β)-accurate learner for $G(\mathcal{H}_1,\ldots,\mathcal{H}_k)$ whose sample complexity is

$$\frac{\tilde{O}\left(\sum_{i=1}^{k}m_i\left(1/4,1/2,1,\delta\right)\right)}{\varepsilon}\cdot\mathrm{poly}(k,1/\alpha,\log(1/\beta)).$$