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

Chao Yan[*]
Technion

Abstract

We present the first nearly optimal differentially private PAC learner for any concept class with VC dimension 1 and Littlestone dimension d. Our algorithm achieves the sample complexity of $\tilde{O}_{\varepsilon,\delta,\alpha,\beta}(\log^* d)$, nearly matching the lower bound of $\Omega(\log^* d)$ proved by Alon et al. [2] [STOC19]. Prior to our work, the best known upper bound is $\tilde{O}(d^5)$ for general classes, as shown by Lyu [21] [STOC26]. The main idea is to combine the tree structure of VC-dimension-1 classes with the private median primitive for interior-point selection. We first privately locate a good depth in the tree, then privately identify a good node on that layer. For proper learning, we show how to descend from the improper output to a proper leaf while controlling the additional false positives.

1 Introduction

Machine-learning models trained on sensitive data may inadvertently reveal information about individual training examples. Differential privacy (DP) [13] provides a principled way to limit such leakage, and Kasiviswanathan et al. [19] initiated the study of *differentially private learning*, where one seeks a learner that is simultaneously probably approximately correct (PAC) [25] and differentially private.

Definition 1 (Differential Privacy [13])**.** *A randomized algorithm M is (ε, δ)-differentially private if for any two datasets S, S' that differ in one entry, and any event E, it holds that*

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

When $\delta = 0$ we say M satisfies pure *differential privacy; when $\delta > 0$ we say it satisfies* approximate *differential privacy.*

In the PAC learning setting, the learner receives an i.i.d. sample

$$S = \{(x_i, y_i)\}_{i=1}^{n}, \qquad x_i \sim \mathcal{D}, \quad y_i = c(x_i),$$

where $\mathcal{D}$ is an unknown distribution over the domain $\mathcal{X}$ and $c \in \mathcal{C}$ is an unknown target concept from a concept class $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$. The goal is to output a hypothesis $h : \mathcal{X} \to \{0,1\}$ with

$$\Pr_{x \sim \mathcal{D}}[h(x) \neq c(x)] \leq \alpha$$

with probability at least $1 - \beta$. The minimum required sample size is the *sample complexity*. In the non-private setting, sample complexity is characterized (up to constants) by the VC dimension [25]. In contrast, the optimal sample complexity under differential privacy is less well understood. For

[*]Department of Mathematics, Technion. cy399@georgetown.edu. Funded by the European Union (ERC, GENERALIZATION, 101039692). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. Research is supported in part by a gift from Georgetown University.

Conference on Neural Information Processing Systems (NeurIPS).

pure DP ($\delta = 0$), Kasiviswanathan et al. [19] give a generic bound whose dependence scales with $\log |\mathcal{C}|$.

In the approximate privacy regime ($\delta > 0$), a sequence of works [5, 10, 4, 18, 17, 11, 23] shows that substantially fewer samples can suffice. Moreover, Alon et al. [2] and Bun et al. [9] relate the sample complexity of approximately differentially private learning to the Littlestone dimension [20], the standard complexity measure for online learning. Concretely, for a class $\mathcal{C}$ with Littlestone dimension d, Alon et al. [2] prove a lower bound of $\Omega(\log^* d)$[2], while Bun et al. [9] give an upper bound of $O(2^{O(2^d)})$. Subsequently, Ghazi et al. [15] improved the general upper bound to $\tilde{O}(\cdot d^6)$, and Lyu [21] improved it to $\tilde{O}(\cdot d^5)$.

This still leaves a large gap between $\Omega(\log^* d)$ and $\tilde{O}(\mathrm{VC} \cdot d^5)$, even for classes with $\mathrm{VC} = 1$. An instructive example is the class of halfspaces: Nissim et al. [23] obtain private learning with $poly(\mathrm{VC}, \log^* d)$ samples. This suggests that $\log^* d$ may be the correct dependence on d more broadly.

More generally, Nissim et al. [23] also show that any VC-dimension-1 class can be privately learned with $\tilde{O}(\log^* |\mathcal{X}|)$ samples (where $|\mathcal{X}|$ is the domain size). However, $|\mathcal{X}|$ can be arbitrarily larger than d, leaving a significant gap between $\log^* d$ and $\min\{d^5, \log^* |\mathcal{X}|\}$. This motivates the following question:

For VC-dimension-1 classes, can we achieve sample complexity $O(\log^* d)$?

In this work, we reduce privately learning VC-dimension-1 classes to the private interior point problem. This yields a sample complexity $\tilde{O}(\log^* d)$.

1.1 Our Results

Theorem 1 (Improper learning)**.** *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a concept class with $\mathrm{VC}(\mathcal{C}) = 1$ and Littlestone dimension d. Then there exists an (ε, δ)-differentially private (possibly improper) algorithm that (α, β)-PAC learns $\mathcal{C}$, provided it is given a labeled sample of size N satisfying*

$$N = \widetilde{O}_{\beta,\delta}\left(\frac{\log^* d}{\alpha\varepsilon}\right).$$

Theorem 2 (Proper learning)**.** *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a concept class with $\mathrm{VC}(\mathcal{C}) = 1$ and Littlestone dimension d. Then there exists an (ε, δ)-differentially private algorithm that* properly *(α, β)-PAC learns $\mathcal{C}$, provided it is given a labeled sample of size N satisfying*

$$N = \widetilde{O}_{\beta,\delta}\left(\frac{\log^* d}{\alpha\varepsilon} + \frac{1}{\alpha^{2.5}\varepsilon}\right).$$

1.2 Other Related Works

Two well-studied VC-dimension-1 classes. A first canonical example is the class of *point functions* (also called *indicator* or *singleton* functions). For each $t \in \mathcal{X}$, define $f_t(x) = 1$ if $x = t$ and $f_t(x) = 0$ otherwise. The point-functions class has Littlestone dimension 1, and Beimel et al. [5] showed it can be privately learned with $O(1)$ samples using the choosing mechanism (Lemma 1).

A second canonical example is the class of *thresholds* over a totally ordered domain. For each $t \in \mathcal{X}$, define $f_t(x) = 1$ if $x \geq t$ and $f_t(x) = 0$ if $x < t$. Over a finite ordered domain, the thresholds class has Littlestone dimension $\Theta(\log |\mathcal{X}|)$. A sequence of works [5, 10, 17, 11] culminated in algorithms that privately learn thresholds with $\tilde{O}(\log^* |\mathcal{X}|)$ samples. A key observation, due to Bun et al. [10], is the equivalence between privately learning thresholds and solving the *private interior point* problem.

Proper vs. improper private learning. We call a learner *proper* if it outputs a hypothesis in the concept class, and *improper* otherwise. Under pure differential privacy, Beimel et al. [3] proved that

[2] $\log^* d$ denotes the iterated logarithm of d.

proper learning of point functions requires $\Omega(\log |\mathcal{X}|)$ samples, while improper learning is possible with $O(1)$ samples. This exhibits a sharp separation between proper and improper learning in the pure-DP setting. In contrast, our result shows that for any VC-dimension-1 class under approximate differential privacy, the gap between proper and improper learning is at most polynomial in $1/\alpha$ (suppressing the dependence on the other parameters).

1.3 Overview of Technique

1.3.1 Tree Structure

Our starting point is the *tree/partial-order structure* of VC-dimension-1 classes, observed by Ben-David [7].[3] In this representation, each node corresponds to a domain point in $\mathcal{X}$ (together with a distinguished root), and each hypothesis corresponds to a path from some node to the root. Intuitively, the points on the path are labeled 1, and points off the path are labeled 0.

A second key idea is to work with the *threshold dimension* (rather than the Littlestone dimension directly). We show that the height of the above tree is upper bounded by the threshold dimension of $\mathcal{C}$, which in turn is at most $2^{O(d)}$ when $\mathrm{Ldim}(\mathcal{C}) = d$ [24, 16, 2]. This allows us to apply the private median primitive (Theorem 9) and obtain a $\tilde{O}(\log^* d)$ dependence.

1.3.2 Improper Learning

Using the tree structure, we show that every VC-dimension-1 class can be privately learned via two private subroutines.

Step 1: privately find a "good depth". We use a partition-and-aggregate approach. The sample S is randomly partitioned into t blocks $S_1, \ldots, S_t$, where $t = \tilde{O}(\log^* d)$. For each block S_i, consider the set of *deterministic points* (with respect to S_i): these are points $x \in \mathcal{X}$ whose label is forced to be 1 by consistency with S_i (equivalently, every concept consistent with S_i labels x as 1). In the tree representation, the deterministic points of each block form an initial segment of a single root-to-leaf path; moreover, the deterministic sets across blocks lie on the *same* underlying path.

We summarize each block S_i by a single number: the depth (layer index) of its deepest deterministic point. Since the tree height is at most $2^{O(d)}$, we can privately aggregate these depths using the private median algorithm (Theorem 9 [11]) to obtain a robust "good depth" z, using $\tilde{O}(\log^* d)$ samples (up to the dependence on $(\alpha, \beta, \varepsilon, \delta)$).

Step 2: privately find a "good node" at that depth. Having fixed depth z, we restrict attention to nodes on layer z. For each candidate node v at depth z, define a score equal to the number of blocks S_i for which v is deterministic. This score is 1-sensitive, hence can be handled by standard private selection primitives. We then apply a choosing mechanism (Lemma 1) to select a high-scoring node v^* at depth z. The selected node determines a rootward path, which we project back to an (improper) hypothesis; we show this hypothesis is accurate.

1.3.3 Example

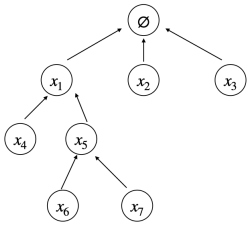

Figure 1: Example of the tree representation.

[3] This tree representation is often discussed via the one-inclusion graph; many works focus on *maximum* classes, whereas Ben-David's observation applies to general (not necessarily maximum) VC-dimension-1 classes.

We illustrate the construction on a small domain $\mathcal{X} = \{x_1, \ldots, x_7\}$. For a hypothesis h, write $I(h) = \{x \in \mathcal{X} : h(x) = 1\}$. Consider $\mathcal{H} = \{h_1, \ldots, h_8\}$ with

$$I(h_1) = \{x_1\}, \quad I(h_2) = \{x_2\}, \quad I(h_3) = \{x_3\},$$
$$I(h_4) = \{x_1, x_4\}, \quad I(h_5) = \{x_1, x_5\},$$
$$I(h_6) = \{x_1, x_5, x_6\}, \quad I(h_7) = \{x_1, x_5, x_7\}, \quad I(h_8) = \emptyset.$$

The corresponding tree is shown in Figure 1. Each hypothesis corresponds to a path to the root; for instance, h_5 corresponds to the path $x_5 \to x_1 \to \emptyset$, meaning that h_5 labels x_1 and x_5 as 1 and all other points as 0. In this example the tree has four layers: layer 0 contains the root $\emptyset$, layer 1 contains x_1, x_2, x_3, layer 2 contains x_4, x_5, and layer 3 contains x_6, x_7.

Now suppose the target concept is h_7. For any block S_i, the deterministic set must be one of

$$\emptyset, \quad \{x_1\}, \quad \{x_1, x_5\}, \quad \{x_1, x_5, x_7\},$$

corresponding to depths $0, 1, 2, 3$. If (say) two blocks yield depths 1 and 3, then the private median of the multiset of depths may output $z = 2$. We then restrict to layer 2, which contains x_4 and x_5. Here x_4 is never deterministic under h_7, whereas x_5 is deterministic for roughly half of the blocks (those whose deepest deterministic point reaches layer 3). Applying the choosing mechanism selects x_5, yielding the hypothesis corresponding to the path $x_5 \to x_1 \to \emptyset$, i.e., h_5.

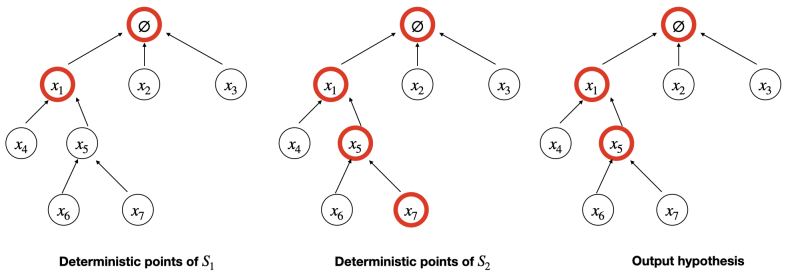

Figure 2: Illustration of the selected hypothesis in the example.

Finally, we argue that if each block contains enough samples, then (by standard VC generalization) the hypotheses induced by "early" and "late" deterministic depths both have small error, and therefore the intermediate hypothesis h_5 is also accurate; see Figure 2.

1.3.4 Proper Learning

The key observation is that the improper learner (Algorithm 2) outputs a node x_{good} that lies on the path corresponding to the true target concept (in the tree representation). Let $x_{c_f^*}$ denote the node corresponding to the target concept. Then $x_{c_f^*}$ is a descendant of x_{good}, and every node on the chain between them induces a hypothesis that is highly accurate (informally, moving down the chain only adds additional points labeled 1).

We call a node *proper* if it can be projected to a concept in $\mathcal{C}$. We therefore restrict attention to the compressed subtree T_{good} rooted at x_{good} (see Figure 3) obtained by enforcing:

1. The root of T_{good} is x_{good}.
2. Every leaf of T_{good} is a proper node.
3. Every internal node of T_{good} is *not* proper.
4. Every internal node has at least two children (if a node has a single child, we contract the edge and merge the two nodes, see Figure 4).

By construction, there exists a proper leaf on the true concept path below x_{good}, hence it induces a proper hypothesis with high accuracy. The remaining task is to find such a leaf *privately*. The analysis of the improper learner implies that for every node in the subtree below x_{good}, the induced

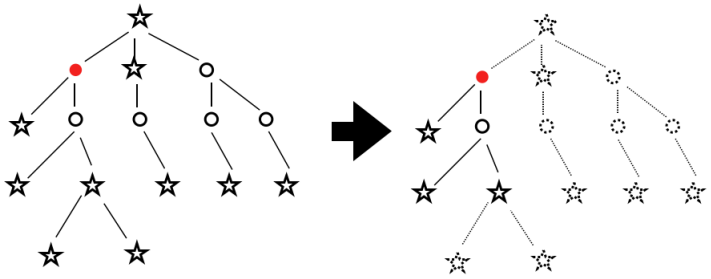

Figure 3: Internal nodes (circles) are improper; leaves (stars) are proper. The red node is x_{good}, and we consider the subtree rooted at it.

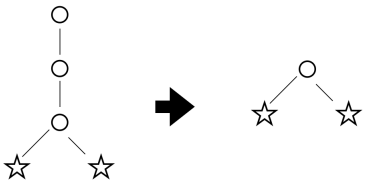

Figure 4: We repeatedly merge any node with a single child into its child, ensuring that every non-leaf node has at least two children.

hypothesis already classifies most *label*-1 points correctly. Thus, to obtain a proper hypothesis it suffices to control the error contributed by *label*-0 points.

We use an additional labeled sample of size N for this proper-selection step. For a node x below x_{good}, let

$$v(x) \;:=\; \#\{\text{label-0 sample points lying on the chain from } x \text{ up to } x_{\text{good}}\},$$

and let

$$w(x) \;:=\; \#\{\text{label-0 sample points that lie in the (original) subtree of } x\}.$$

Intuitively, $v(x)$ counts the number of additional false positives incurred when moving from the hypothesis induced by x_{good} to the hypothesis induced by x.

Fact 1. *There exists a leaf ℓ of T_{good} such that $v(\ell) = 0$.*

Proof. Let ℓ be the (proper) leaf of T_{good} that lies on the true concept path below x_{good}. Every node on the chain from ℓ to x_{good} is labeled 1 by the target concept, so no label-0 sample point can lie on this chain. Hence $v(\ell) = 0$. □

Fact 2. *If x' is a descendant of x (i.e., x' lies strictly below x), then*

$$v(x') \le v(x) + w(x).$$

Proof. The chain from x' to x_{good} is the concatenation of the chain from x' to x and the chain from x to x_{good}. The label-0 points on the first segment lie inside the subtree rooted at x, whose total number of label-0 points is $w(x)$. Thus $v(x') - v(x) \le w(x)$. □

Private search for a good leaf. We maintain a pointer x_{flag} initialized to x_{good}. Let X_{children} be the set of children of x_{flag} in T_{good}. We distinguish two cases.

1. **Case 1 (a light child exists).** If there exists a child $x \in X_{\text{children}}$ with

   $$w(x) \;\le\; O(\alpha N),$$

   then every leaf ℓ in the subtree rooted at x satisfies

   $$v(\ell) \;\le\; v(x) + w(x) \;\le\; v(x) + O(\alpha N),$$

by the inequality above. In this case we privately select a child with small $w(\cdot)$ using the exponential mechanism (even if $|X_{\text{children}}|$ is large), and then output an arbitrary leaf in its subtree (all leaves are proper by construction).

2. **Case 2 (all children are heavy).** Otherwise, every child $x \in X_{\text{children}}$ satisfies $w(x) > \Omega(\alpha N)$. Since the subtrees of distinct children are disjoint and $\sum_{x \in X_{\text{children}}} w(x) \leq N$, we get

$$|X_{\text{children}}| \;\leq\; O(1/\alpha).$$

We can therefore apply the exponential mechanism over this set of size $O(1/\alpha)$ to choose a child whose subtree contains a near-optimal leaf (i.e., a leaf ℓ with small $v(\ell)$). The privacy/utility loss per iteration scales like $O(\log(1/\alpha)/\varepsilon)$ due to the bounded candidate set.

In Case 2, moving x_{flag} to a heavy child removes at least $\Omega(\alpha N)$ label-0 sample points from consideration (those contained in the discarded siblings' subtrees). Hence Case 2 can occur at most $O(1/\alpha)$ times before reaching a leaf. Overall, we obtain a proper leaf ℓ satisfying

$$v(\ell) \;\leq\; \min_{\ell' \in \text{Leaves}(T_{\text{good}})} v(\ell') \;+\; O\left(\frac{\log(1/\alpha)}{\alpha\varepsilon}\right),$$

and since $\min_{\ell'} v(\ell') = 0$ (up to the $O(\alpha N)$ slack already present from the improper-learning analysis), this yields

$$v(\ell) \;\leq\; O\left(\alpha N + \frac{\log(1/\alpha)}{\alpha\varepsilon}\right).$$

Applying advanced composition over the $O(1/\alpha)$ iterations, we obtain the stated proper-learning bound: the sample complexity equals that of the improper learner plus an additional $\widetilde{O}\big(\frac{1}{\alpha^{2.5}\varepsilon}\big)$ examples.

Finally, we can think of $v(\ell)$ as the 'distance' from the improper output: it is the number of sample points on which the hypothesis at ℓ disagrees with the improper hypothesis. Picking ℓ with small $v(\ell)$ preserves accuracy while making the hypothesis proper.

2 Discussion for Agnostic Learning

Our algorithm is presented for the realizable setting, where the labeled sample is consistent with some concept in the class. A realizable-to-agnostic transformation for private learning is known, see, e.g., Beimel et al. [6] and the explicit discussion in Alon et al. [1].

References

[1] Noga Alon, Amos Beimel, Shay Moran, and Uri Stemmer. Closure properties for private classification and online prediction. In *Conference on Learning Theory*, pages 119–152. PMLR, 2020.

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

[3] Amos Beimel, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. In *TCC*, volume 5978 of *LNCS*, pages 437–454. Springer, 2010.

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

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

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

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

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

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

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

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

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

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

[14] Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *FOCS*, pages 51–60, 2010.

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

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

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

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

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

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

[21] Xin Lyu. Private learning of littlestone classes, revisited. In *Proceedings of the 58th Annual ACM Symposium on Theory of Computing*, pages 2218–2229, 2026.

[22] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103. IEEE, Oct 20–23 2007.

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

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

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

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

A Notation

We write $\mathcal{X}$ for the *domain* (or instance space), and $x \in \mathcal{X}$ for a *point* (instance). A *concept* is a function $c : \mathcal{X} \to \{0,1\}$, and a *concept class* is a set of concepts $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$. A *hypothesis* is any function $h : \mathcal{X} \to \{0,1\}$ (not necessarily in $\mathcal{C}$).

For any function $f : \mathcal{X} \to \{0,1\}$, we denote its set of positive points by

$$I(f) \;:=\; \{x \in \mathcal{X} : f(x) = 1\}.$$

Given a distribution $\mathcal{D}$ over $\mathcal{X}$, the (0–1) error of h w.r.t. c is $\mathrm{err}_{\mathcal{D}}(c,h) = \Pr_{x\sim\mathcal{D}}[h(x) \neq c(x)]$.

B Preliminaries

B.1 Learning Theory

Definition 2 (Error)**.** *Let $\mathcal{D}$ be a distribution over $\mathcal{X}$, let $c : \mathcal{X} \to \{0,1\}$ be a concept, and let $h : \mathcal{X} \to \{0,1\}$ be a hypothesis. The error of h with respect to c under $\mathcal{D}$ is*

$$\mathrm{err}_{\mathcal{D}}(c,h) \;:=\; \Pr_{x\sim\mathcal{D}}\big[c(x) \neq h(x)\big].$$

For a labeled sample $S = \{(x_i, y_i)\}_{i=1}^n$ with $y_i = c(x_i)$, the empirical error is

$$\mathrm{err}_S(c,h) \;:=\; \frac{1}{|S|}\,\big|\{i \in [n] : c(x_i) \neq h(x_i)\}\big| \;=\; \frac{1}{|S|}\,\big|\{i \in [n] : y_i \neq h(x_i)\}\big|\,.$$

When the labels in S are understood, we also write $\mathrm{err}_S(h)$.

Definition 3 (PAC learning [25])**.** *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a concept class. A (possibly randomized) learner L (α,β)-*PAC learns *$\mathcal{C}$ if for every distribution $\mathcal{D}$ over $\mathcal{X}$ and every target concept $c \in \mathcal{C}$, given an i.i.d. labeled sample*

$$S = \{(x_i, c(x_i))\}_{i=1}^n, \qquad x_i \sim \mathcal{D},$$

the output hypothesis $h = L(S)$ satisfies

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

where the probability is over the draw of S and the internal randomness of L.

Definition 4 (Proper vs. improper learning)**.** *A learner is* proper *if it always outputs a hypothesis $h \in \mathcal{C}$. It is* improper *if it may output $h \notin \mathcal{C}$.*

Definition 5 (VC dimension [26])**.** *A set of points $\{x_1,\dots,x_k\} \subseteq \mathcal{X}$ is* shattered *by $\mathcal{C}$ if for every subset $T \subseteq \{x_1,\dots,x_k\}$ there exists a concept $c \in \mathcal{C}$ such that $c(x) = 1$ for all $x \in T$ and $c(x) = 0$ for all $x \in \{x_1,\dots,x_k\} \setminus T$. The VC dimension $\mathrm{VC}(\mathcal{C})$ is the maximum k such that some k-point set is shattered by $\mathcal{C}$.*

Theorem 3 ([8])**.** *Let $\mathcal{C}$ be a concept class over domain $\mathcal{X}$ and let $\mathcal{D}$ be a distribution over $\mathcal{X}$. Fix $\alpha,\beta > 0$, and let*

$$m \;\geq\; \frac{48}{\alpha}\left(10\,\mathrm{VC}(\mathcal{C})\log\Big(\frac{48e}{\alpha}\Big) + \log\Big(\frac{5}{\beta}\Big)\right).$$

If S is an i.i.d. sample of size m from $\mathcal{D}$ (labeled by some $c \in \mathcal{C}$), then

$$\Pr[\exists\, c,h \in \mathcal{C} \text{ s.t. } \mathrm{err}_S(c,h) \leq \alpha/10 \text{ and } \mathrm{err}_{\mathcal{D}}(c,h) \geq \alpha] \;\leq\; \beta.$$

Definition 6 (Threshold dimension)**.** *We say that points $x_1,\dots,x_k \in \mathcal{X}$ and concepts $c_1,\dots,c_k \in \mathcal{C}$ form a* threshold sequence of length k *if for every $i \in [k]$,*

$$c_i(x_j) = \begin{cases} 0 & \text{if } j < i, \\ 1 & \text{if } j \geq i. \end{cases}$$

The threshold dimension $\mathrm{TD}(\mathcal{X},\mathcal{C})$ *is the maximum k for which such a threshold sequence exists.*

Definition 7 (Online learning [20])**.** *In round t, the learner receives a point $x_t \in \mathcal{X}$, predicts a label $\hat{y}_t \in \{0,1\}$, and then observes the true label $y_t \in \{0,1\}$. The learner makes a* mistake *in round t if $\hat{y}_t \neq y_t$.*

Definition 8 (Littlestone dimension [20]). *A* mistake tree *of depth d for $\mathcal{C}$ is a complete binary tree of depth d whose internal nodes are labeled by points in $\mathcal{X}$, such that for every root-to-leaf path (interpreted as a sequence of left/right choices, i.e., labels $0/1$), there exists a concept $c \in \mathcal{C}$ consistent with all labels along the path. The* Littlestone dimension $\mathrm{Ldim}(\mathcal{X}, \mathcal{C})$ *is the largest d for which such a mistake tree exists. Equivalently,* $\mathrm{Ldim}(\mathcal{X}, \mathcal{C})$ *is the optimal worst-case mistake bound in realizable online learning.*

Theorem 4 ([24, 16, 2]). *For any concept class $\mathcal{C} \subseteq 2^{\mathcal{X}}$,*

$$\left\lfloor \log \mathrm{Ldim}(\mathcal{X}, \mathcal{C}) \right\rfloor \;\le\; \mathrm{TD}(\mathcal{X}, \mathcal{C}) \;\le\; 2^{\mathrm{Ldim}(\mathcal{X}, \mathcal{C})+1}.$$

Corollary 1.

$$\log^*\big(\mathrm{TD}(\mathcal{X}, \mathcal{C})\big) = \Theta\big(\log^* \mathrm{Ldim}(\mathcal{X}, \mathcal{C})\big).$$

B.2 Differential Privacy

Definition 9 (Differential privacy [13]). *A randomized algorithm (mechanism) M is (ε, δ)-differentially private if for any two datasets S, S' that differ in one entry and any event E,*

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

Theorem 5 (Post-processing [13]). *If M is (ε, δ)-differentially private and A is any (possibly randomized) algorithm, then $A \circ M$ is also (ε, δ)-differentially private.*

Theorem 6 (Basic composition [12]). *If M_1 is $(\varepsilon_1, \delta_1)$-differentially private and M_2 is $(\varepsilon_2, \delta_2)$-differentially private, then the mechanism $M(S) = (M_1(S), M_2(S))$ is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-differentially private.*

Theorem 7 (Advanced composition [14]). *Let $M_1, \ldots, M_k : \mathcal{X}^n \to \mathcal{Y}$ be (ε, δ)-differentially private mechanisms. Then the algorithm that on input $S \in \mathcal{X}^n$ outputs $(M_1(S), \ldots, M_k(S))$ is $(\varepsilon', k\delta + \delta')$-differentially private, where $\varepsilon' = \sqrt{2k \ln(1/\delta')} \cdot \varepsilon + k\varepsilon \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}$ for every $\delta' > 0$.*

Specifically, when $\varepsilon < \frac{1}{\sqrt{k}}$, we have $\varepsilon' = \theta(\sqrt{k \ln(1/\delta')} \cdot \varepsilon)$.

Theorem 8 (Exponential mechanism [22]). *Let $q(D, r)$ be a score function where D is a dataset and $r \in R$ is a candidate output. Let Δ be the sensitivity of q, i.e.,*

$$\Delta \;:=\; \max_{r \in R} \max_{\text{neighbors } D \sim D'} |q(D, r) - q(D', r)|.$$

The exponential mechanism M_E outputs $r \in R$ with probability proportional to

$$\exp\left(\frac{\varepsilon\, q(D, r)}{2\Delta}\right).$$

It has the following two properties:

1. ***(Privacy)** It is ε-differentially private.*
2. ***(Accuracy)** Let $OPT(D) = \max_{r \in R} q(D, r)$ be the score attained by the best object r with respect to the dataset D. Then*

$$\Pr\left[q(M_E(D)) \le OPT(D) - \frac{2\Delta}{\varepsilon}\left(\ln |R| + \ln(\frac{1}{\beta})\right)\right] \le \beta.$$

Definition 10 (Differentially private PAC learning [19]). *A learner L $(\alpha, \beta, \varepsilon, \delta)$-*DP PAC learns *$\mathcal{C}$ if: (i) L (α, β)-PAC learns $\mathcal{C}$, and (ii) L is (ε, δ)-differentially private.*

Definition 11 (α-median). *Let $S = \{x_1, \ldots, x_n\}$ be a multiset of real numbers (or totally ordered elements). We say $\hat{x}$ is an* α-median *of S if at least a $(\frac{1}{2} - \alpha)$ fraction of the elements of S are $\le \hat{x}$ and at least a $(\frac{1}{2} - \alpha)$ fraction are $\ge \hat{x}$, i.e.,*

$$\left|\{i : x_i \le \hat{x}\}\right| \;\ge\; (\tfrac{1}{2} - \alpha)n \qquad \textit{and} \qquad \left|\{i : x_i \ge \hat{x}\}\right| \;\ge\; (\tfrac{1}{2} - \alpha)n.$$

Fact 3. *For every finite multiset S, there exists a 0-median (i.e., a median) of S.*

Theorem 9 ([11]). *Let $\mathcal{X}$ be a finite totally ordered domain. There exists an (ε, δ)-differentially private algorithm* PrivateMedian *that, on input $S \in \mathcal{X}^n$, outputs an α-median with probability at least $1 - \beta$, provided that*

$$n \;>\; n_{\mathrm{PM}}(|\mathcal{X}|, \alpha, \beta, \varepsilon, \delta),$$

where

$$n_{\mathrm{PM}}(|\mathcal{X}|, \alpha, \beta, \varepsilon, \delta) \in O\left(\frac{\log^* |\mathcal{X}| \cdot \log^2\!\left(\frac{\log^* |\mathcal{X}|}{\beta\delta}\right)}{\alpha\varepsilon}\right).$$

Definition 12 (k-bounded quality function). *Let $q : \mathcal{X}^* \times Z \to \mathbb{R}$ be a quality function. We say q is* k-bounded *if for every dataset D and every element $x \in \mathcal{X}$, letting $D' = D \cup \{x\}$, the following hold:*

1. *$q(\emptyset, z) = 0$ for all $z \in Z$;*
2. *for all $z \in Z$, $q(D', z) \in \{q(D, z),\ q(D, z) + 1\}$; and*
3. *there are at most k values $z \in Z$ such that $q(D', z) = q(D, z) + 1$.*

Lemma 1 (Choosing mechanism [5]). *Let $\varepsilon \in (0, 2)$ and $\delta > 0$. Let $q : \mathcal{X}^* \times Z \to \mathbb{R}$ be a k-bounded quality function. There exists an (ε, δ)-DP algorithm $\mathcal{A}$ such that for every dataset $D \in \mathcal{X}^n$, with probability at least $1 - \beta$,*

$$q(D, \mathcal{A}(D)) \;\geq\; \max_{z \in Z} q(D, z) \;-\; \frac{16}{\varepsilon} \log\left(\frac{4kn}{\beta\varepsilon\delta}\right).$$

C Structure of Classes with VC Dimension 1

Throughout, let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a concept class with $\mathrm{VC}(\mathcal{C}) = 1$. We make two standard "without loss of generality" simplifications to remove redundant points/concepts.

Assumption 1 (Distinct points). *For any two distinct points $x_1, x_2 \in \mathcal{X}$, there exists a concept $c \in \mathcal{C}$ such that $c(x_1) \neq c(x_2)$. Otherwise, x_1 and x_2 are indistinguishable by $\mathcal{C}$ and can be identified (merged) without affecting the learning problem.*

Assumption 2 (Non-constant points). *For every point $x \in \mathcal{X}$, there exist concepts $c_0, c_1 \in \mathcal{C}$ such that $c_0(x) = 0$ and $c_1(x) = 1$. Otherwise, if all concepts agree on x, then x carries no information and can be removed from the domain (its label is fixed under every concept).*

C.1 A Partial Order on the Domain

Definition 13 (Partial order). *Given $\mathcal{X}, \mathcal{C}$ and points $x_1, x_2 \in \mathcal{X}$, we write $x_1 \preceq x_2$ (with respect to $\mathcal{C}$) if for every $c \in \mathcal{C}$,*

$$c(x_1) = 1 \implies c(x_2) = 1.$$

We write $x_1 \prec x_2$ if $x_1 \preceq x_2$ and $x_1 \neq x_2$.

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

Example 1. *For the thresholds class over a totally ordered domain, the order $\preceq$ coincides with the underlying total order, so $x_1 \preceq x_2 \preceq \cdots \preceq x_k$.*

C.2 f-representation

Definition 14 (f-representation). *Fix functions $c, f : \mathcal{X} \to \{0,1\}$. The f-*representation *of c is the function $c_f : \mathcal{X} \to \{0,1\}$ defined by*

$$c_f(x) \;:=\; \mathbf{1}[c(x) \neq f(x)] \;=\; c(x) \oplus f(x),$$

where $\oplus$ denotes XOR. For a class $\mathcal{C}$, define its f-representation as

$$\mathcal{C}_f \;:=\; \{c_f : c \in \mathcal{C}\}.$$

[3] In [11], the authors give an algorithm for privately selecting an *interior point*, i.e., given $x_1, \ldots, x_n$, output $\hat{x}$ with $\min_i x_i \leq \hat{x} \leq \max_i x_i$. One can obtain a private median by discarding the smallest and largest $(\frac{1}{2} - \alpha/2)$ fraction of the data and then applying the interior-point algorithm; see Bun et al. [10].

Algorithm 1: MakeTree

Input: a concept class $\mathcal{C}$ with $\mathrm{VC}(\mathcal{C}) = 1$.
Output: a rooted tree T over $\mathcal{X} \cup \{\emptyset\}$.

1. Choose any $f \in \mathcal{C}$ and form $\mathcal{C}_f$ (Definition 14). All order relations below are with respect to $\mathcal{C}_f$.
2. Create a root node $\emptyset$.
3. For each $x \in \mathcal{X}$, connect x to its *parent* defined as follows:
   - If there is no $y \in \mathcal{X}$ with $x \prec y$, set $\mathrm{parent}(x) = \emptyset$.
   - Otherwise, let $\mathrm{parent}(x)$ be any element y such that $x \prec y$ and there is no z with $x \prec z \prec y$ (i.e., an immediate successor of x in the Hasse diagram).
4. Output the resulting rooted graph T.

The map $c \mapsto c_f$ is a bijection with inverse $c(x) = f(x) \oplus c_f(x)$. Accordingly, a labeled example $(x, c(x))$ can be converted to the corresponding label under c_f by replacing the label with $c(x) \oplus f(x)$.

Lemma 2. *For every $f : \mathcal{X} \to \{0, 1\}$,*

$$\mathrm{VC}(\mathcal{C}) = \mathrm{VC}(\mathcal{C}_f) \qquad \textit{and} \qquad \mathrm{Ldim}(\mathcal{C}) = \mathrm{Ldim}(\mathcal{C}_f).$$

Proof. For VC dimension, $c \mapsto c_f$ is a bijection on labelings: for any set of points, the family of label patterns realized by $\mathcal{C}$ is in one-to-one correspondence with those realized by $\mathcal{C}_f$ (obtained by XOR with the fixed pattern f restricted to the set). Hence shattering is preserved.

For Littlestone dimension, a mistake tree for $\mathcal{C}$ can be converted to a mistake tree for $\mathcal{C}_f$ by flipping the interpretation of the labels at node x according to $f(x)$ (equivalently, XOR every label along every path by $f(x)$ at that node). This preserves realizability of every root-to-leaf labeling and therefore preserves the maximum shattered depth. □

In the remainder, we fix an arbitrary $f \in \mathcal{C}$ and work with $\mathcal{C}_f$.

Observation 1. *If $f \in \mathcal{C}$, then $\mathcal{C}_f$ contains the all-zero function.*

Proof. The concept f_f satisfies $f_f(x) = f(x) \oplus f(x) = 0$ for all $x \in \mathcal{X}$. □

Lemma 3 ([7])**.** *Assume $f \in \mathcal{C}$. If x_1 and x_2 are incomparable under $\mathcal{C}_f$, then there is no $c \in \mathcal{C}_f$ such that $c(x_1) = c(x_2) = 1$.*

Proof. If x_1 and x_2 are incomparable, then there exist $c_1, c_2 \in \mathcal{C}_f$ with $c_1(x_1) = 1, c_1(x_2) = 0$ and $c_2(x_1) = 0, c_2(x_2) = 1$. If there were also $c \in \mathcal{C}_f$ with $c(x_1) = c(x_2) = 1$, then together with the all-zero concept $f_f \in \mathcal{C}_f$ (Observation 1), the set $\{x_1, x_2\}$ would be shattered, contradicting $\mathrm{VC}(\mathcal{C}_f) = 1$. □

C.3 Tree Structure

Ben-Davide [7] shows that every concept class with VC dimension 1 has a tree structure. We build a rooted tree from the poset $(\mathcal{X}, \preceq)$ induced by $\mathcal{C}_f$ (see Algorithm 1). Introduce a formal root $\emptyset$ and declare that every $\preceq$-maximal point is a child of $\emptyset$.

(For VC-dimension-1 classes, the order structure ensures this construction yields a tree; we use this tree representation throughout.)

Definition 15 (Deterministic point)**.** *Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a labeled sample. Define the* version space

$$V(S) \;:=\; \{c \in \mathcal{C} : \mathrm{err}_S(c) = 0\}.$$

A point $x \in \mathcal{X}$ is deterministic (with respect to S) *if every concept consistent with S labels it as 1, i.e.,*

$$\forall c \in V(S), \quad c(x) = 1.$$

Definition 16 (Depth / distance)**.** *Let T be the rooted tree built from $\mathcal{C}_f$ with root $\emptyset$. For $x \in \mathcal{X}$, define $d_{\mathcal{C}_f}(x)$ to be the length of the longest chain from x to the root:*

$$d_{\mathcal{C}_f}(x) \;:=\; \max\{k : \exists\, x = x_0 \prec x_1 \prec \cdots \prec x_{k-1} \prec \emptyset\}\,.$$

We set $d_{\mathcal{C}_f}(\emptyset) = 0$.

Lemma 4.

$$\max_{x \in \mathcal{X}} d_{\mathcal{C}_f}(x) \;\le\; \mathrm{TD}(\mathcal{X}, \mathcal{C}_f).$$

Proof. Fix any chain

$$x_0 \prec x_1 \prec \cdots \prec x_k \prec \emptyset$$

in the partial order induced by $\mathcal{C}_f$. It suffices to construct concepts $c_0, \ldots, c_k \in \mathcal{C}_f$ such that for all $i, j \in \{0, \ldots, k\}$,

$$c_i(x_j) = 1 \quad \Longleftrightarrow \quad j \ge i,$$

i.e., $(x_0, \ldots, x_k)$ together with $(c_0, \ldots, c_k)$ form a threshold sequence of length $k+1$.

Step 1: choose c_0. Since $f \in \mathcal{C}$ and by Assumption 2 (non-constant at each point), for the point x_0 there exists some $c \in \mathcal{C}$ with $c(x_0) \neq f(x_0)$, hence its f-representation satisfies $c_f(x_0) = 1$. Let $c_0 := c_f \in \mathcal{C}_f$ with $c_0(x_0) = 1$. Because $x_0 \preceq x_j$ for every $j \ge 0$, Definition 13 implies $c_0(x_j) = 1$ for all j.

Step 2: inductively construct c_i for $i \ge 1$. Assume we have constructed $c_0, \ldots, c_{i-1}$. By Assumption 1 (applied to the class $\mathcal{C}_f$), there exists a concept $c' \in \mathcal{C}_f$ such that

$$c'(x_{i-1}) \neq c'(x_i).$$

Since $x_{i-1} \preceq x_i$, the pattern $c'(x_{i-1}) = 1$ and $c'(x_i) = 0$ is impossible (it would contradict the definition of $\preceq$). Therefore we must have

$$c'(x_{i-1}) = 0 \qquad \text{and} \qquad c'(x_i) = 1.$$

Now we claim that this c' already has the required threshold behavior on the chain:

- For any $j < i$, we have $x_j \preceq x_{i-1}$. If $c'(x_j) = 1$, then by Definition 13 we would get $c'(x_{i-1}) = 1$, a contradiction. Hence $c'(x_j) = 0$ for all $j < i$.
- For any $j \ge i$, we have $x_i \preceq x_j$, and since $c'(x_i) = 1$, Definition 13 implies $c'(x_j) = 1$ for all $j \ge i$.

Thus c' satisfies $c'(x_j) = \mathbf{1}[j \ge i]$ on the chain. Set $c_i := c'$.

We have constructed $c_0, \ldots, c_k$ with the threshold property on $(x_0, \ldots, x_k)$, so $\mathrm{TD}(\mathcal{X}, \mathcal{C}_f) \ge k+1$. Since the chain was arbitrary, the maximum possible chain length is at most $\mathrm{TD}(\mathcal{X}, \mathcal{C}_f)$. □

D Improper Learner

We present the improper learner in Algorithm 2.

Theorem 10. *OPTPrivateLearner is $(2\varepsilon, 2\delta)$-differentially private.*

Proof. A change to one entry of S affects only one block S_i, and therefore can only affect the corresponding quantities (B_i, b_i, y_i) and the increments contributed by that block to the vector $(q_1, \ldots, q_m)$.

Step 4 applies PrivateMedian once with parameters (ε, δ), so the released depth z is (ε, δ)-DP. Step 7 applies the choosing mechanism once with parameters (ε, δ) to select x_{good}. All remaining steps are post-processing. By basic composition (Theorem 6) and post-processing (Theorem 5), the overall algorithm is $(2\varepsilon, 2\delta)$-differentially private. □

Lemma 5 ([7])**.** *For every $c \in \mathcal{C}_f$, the set $I(c) = \{x : c(x) = 1\}$ is totally ordered by $\preceq$. Moreover, if x is the maximum element of $I(c)$ under $\preceq$, then there is no $\hat{x} \in \mathcal{X}$ such that $x \prec \hat{x}$.*

Algorithm 2: OPTPrivateLearner

Parameters: confidence $\beta > 0$, accuracy $\alpha > 0$, privacy (ε, δ). Let $d := \mathrm{TD}(\mathcal{X}, \mathcal{C}_f) + 1$ (so depths lie in $[0, d]$). Let

$$n_{\mathrm{PM}}(d, 1/3, \beta, \varepsilon, \delta) \in O\left(\frac{\log^* d \cdot \log^2\left(\frac{\log^* d}{\beta\delta}\right)}{\varepsilon}\right),$$

and set

$$t = \max\left\{n_{\mathrm{PM}}(d, 1/3, \beta, \varepsilon, \delta),\ O\left(\frac{1}{\varepsilon}\log\left(\frac{4n_{\mathrm{PM}}(d, 1/3, \beta, \varepsilon, \delta)}{\beta\varepsilon\delta}\right)\right)\right\} = O\left(\frac{\log^* d \cdot \log^2\left(\frac{\log^* d}{\varepsilon\beta\delta}\right)}{\varepsilon}\right).$$

Input: labeled sample $S \in (\mathcal{X} \times \{0,1\})^N$ where

$$N = t \cdot \frac{48}{\alpha}\left(10\log\left(\frac{48e}{\alpha}\right) + \log\left(\frac{5}{\beta}\right)\right).$$

Output: a hypothesis $\hat{h}_{\mathrm{good}} : \mathcal{X} \to \{0,1\}$.

Procedure:

1. Build the tree T via Algorithm 1. Let $f \in \mathcal{C}$ be the selected concept, and convert labels to the f-representation (Definition 14), so the sample is labeled by c_f^*.
2. Randomly partition S into t disjoint blocks $S_1, \ldots, S_t$ of equal size.
3. For each $i \in [t]$:
   (a) Let B_i be the set of points that are deterministic with respect to S_i in $\mathcal{C}_f$ (Definition 15).
   (b) Let $b_i \in B_i$ be a deepest deterministic point:
   $$b_i \in \arg\max_{x \in B_i} d_{\mathcal{C}_f}(x),$$
   and let $y_i := d_{\mathcal{C}_f}(b_i)$ denote its depth.
   Specifically, if $B_i = \emptyset$, let $y_i = 0$
4. Compute a $(1/3)$-median depth
   $$z \leftarrow \mathsf{PrivateMedian}(y_1, \ldots, y_t)$$
   with parameters $(\varepsilon, \delta, \beta)$.
5. Let $P = \{x \in \mathcal{X} : d_{\mathcal{C}_f}(x) = z\} = \{x_1, \ldots, x_m\}$ and initialize counters $q_1 = \cdots = q_m = 0$.
6. For each $i \in [t]$:
   (a) If $y_i \geq z$, then for each $j \in [m]$ with $x_j \in B_i$, increment $q_j \leftarrow q_j + 1$.
   (b) If $y_i < z$, do nothing.
7. Run the choosing mechanism (Lemma 1) with score vector $(q_1, \ldots, q_m)$ and parameters $(\varepsilon, \delta, \beta)$, and obtain an index j^* and the corresponding point $x_{\mathrm{good}} := x_{j^*}$.
8. Define the upset of x_{good} in the partial order:
   $$I_{\mathrm{good}} := \{x \in \mathcal{X} : x_{\mathrm{good}} \preceq x\}.$$
   Convert back from the f-representation by outputting the hypothesis
   $$\hat{h}_{\mathrm{good}}(x) \ := \ f(x) \oplus \mathbf{1}[x \in I_{\mathrm{good}}].$$

Proof. By Lemma 3, any two points in $I(c)$ are comparable, hence $I(c)$ is totally ordered. If there were $\hat{x}$ with $x \prec \hat{x}$, then by Definition 13 we would have $c(\hat{x}) = 1$, contradicting maximality of x in $I(c)$. □

Lemma 6. *Let S be labeled by the target c_f^*. If a point x is deterministic with respect to S in $\mathcal{C}_f$, then $c_f^*(x) = 1$.*

Proof. If x is deterministic, then every concept in the version space $V(S) = \{c \in \mathcal{C}_f : \mathrm{err}_S(c) = 0\}$ labels x as 1. In particular, the true concept c_f^* belongs to $V(S)$, so $c_f^*(x) = 1$. □

Lemma 7. *Let $S_1, \ldots, S_t$ be the blocks in Algorithm 2. For each $i \in [t]$, let B_i be the set of points deterministic with respect to S_i in $\mathcal{C}_f$, and let h_i be the hypothesis defined by $I(h_i) = B_i$ (i.e., $h_i(x) = 1$ iff $x \in B_i$). Then, with probability at least $1 - \beta t$, we have*

$$\forall i \in [t], \qquad \mathrm{err}_{\mathcal{D}}(h_i, c_f^*) \le \alpha,$$

i.e., all h_i are α-good.

Proof. Fix an index $i \in [t]$ and let c_f^* be the f-representation of the target concept. First we show that h_i is perfectly consistent with S_i:

- If $(x, 1) \in S_i$, then every $c \in \mathcal{C}_f$ with $\mathrm{err}_{S_i}(c) = 0$ must satisfy $c(x) = 1$, hence $x \in B_i$ by definition, so $h_i(x) = 1$.
- If $(x, 0) \in S_i$, then no $c \in \mathcal{C}_f$ with $\mathrm{err}_{S_i}(c) = 0$ can satisfy $c(x) = 1$, so $x \notin B_i$, hence $h_i(x) = 0$.

Therefore $\mathrm{err}_{S_i}(h_i) = 0$.

Now consider two cases.

Case 1: $h_i \in \mathcal{C}_f$. Since both h_i and c_f^* belong to $\mathcal{C}_f$ and they agree on S_i (indeed, $\mathrm{err}_{S_i}(c_f^*, h_i) = 0$), Theorem 3 implies that with probability at least $1 - \beta$,

$$\mathrm{err}_{\mathcal{D}}(h_i, c_f^*) \le \alpha.$$

Case 2: $h_i \notin \mathcal{C}_f$. Let $b_i \in B_i$ be a deepest deterministic point (i.e., a point in B_i of maximum depth in T). Because h_i is not a concept in $\mathcal{C}_f$, the version space restricted below b_i must branch: there exist two distinct children x_1, x_2 of b_i and concepts $c_1, c_2 \in \mathcal{C}_f$ such that

$$\mathrm{err}_{S_i}(c_1) = \mathrm{err}_{S_i}(c_2) = 0, \qquad c_1(x_1) = 1, \qquad c_2(x_2) = 1.$$

(Otherwise, if only one child could ever be labeled 1 by a consistent concept, that child would be forced and hence would lie in B_i, contradicting the maximality of b_i.)

Since x_1 and x_2 lie on different branches, one of them is not on the true path of c_f^*. Without loss of generality assume $c_f^*(x_1) = 0$ and keep the corresponding consistent concept c_1.

We claim that

$$\mathrm{err}_{\mathcal{D}}(h_i, c_f^*) \;\le\; \mathrm{err}_{\mathcal{D}}(c_1, c_f^*). \tag{1}$$

Indeed, every $x \in B_i$ is labeled 1 by every consistent concept, hence in particular $c_1(x) = 1$; thus $B_i \subseteq I(c_1)$. So the only points where h_i and c_1 disagree are those in $I(c_1) \setminus B_i$, on which h_i outputs 0 and c_1 outputs 1. Such points lie strictly below b_i along the c_1-branch, hence are descendants of x_1. Since $c_f^*(x_1) = 0$, every descendant $u \preceq x_1$ must satisfy $c_f^*(u) = 0$ (otherwise $c_f^*(u) = 1$ would imply $c_f^*(x_1) = 1$ by Definition 13). Therefore, on every disagreement point between h_i and c_1, the label under c_f^* is 0, so h_i agrees with c_f^* there. This proves (1).

Finally, since $c_1 \in \mathcal{C}_f$ and $\mathrm{err}_{S_i}(c_1, c_f^*) = 0$, Theorem 3 gives $\mathrm{err}_{\mathcal{D}}(c_1, c_f^*) \le \alpha$ with probability at least $1 - \beta$, and then (1) implies $\mathrm{err}_{\mathcal{D}}(h_i, c_f^*) \le \alpha$ as well.

In both cases, for each fixed i we have $\Pr[\mathrm{err}_{\mathcal{D}}(h_i, c_f^*) > \alpha] \le \beta$. A union bound over $i \in [t]$ yields that with probability at least $1 - \beta t$, all h_i are α-good. □

Lemma 8. *The deepest deterministic points $b_1, \ldots, b_t$ can be ordered so that*

$$b_{\pi(1)} \preceq b_{\pi(2)} \preceq \cdots \preceq b_{\pi(t)}.$$

Proof. By Lemma 6, each $b_i \in I(c_f^*)$. Lemma 5 implies $I(c_f^*)$ is totally ordered by $\preceq$, hence the b_i's admit such an ordering (allowing ties). □

Lemma 9. *Let h_i be the hypothesis with $I(h_i) = B_i$, and let h_{good} be the hypothesis with $I(h_{\mathrm{good}}) = I_{\mathrm{good}} = \{x : x_{\mathrm{good}} \preceq x\}$. Conditioned on the event that all h_i are α-good, with probability at least $1 - 2\beta$ (over the randomness of PrivateMedian and the choosing mechanism) we have*

$$\mathrm{err}_{\mathcal{D}}(h_{\mathrm{good}}, c_f^*) \leq \alpha.$$

Proof. For each i, let $b_i \in B_i$ be a deepest deterministic point and let $y_i = d_{\mathcal{C}_f}(b_i)$. Let z be the output of PrivateMedian on $(y_1, \ldots, y_t)$ (Step 4). By Theorem 9, with probability at least $1 - \beta$, the value z is a $(1/3)$-median of $\{y_1, \ldots, y_t\}$; in particular,

$$|\{i : y_i \geq z\}| \;\geq\; t/6 \qquad \text{and} \qquad |\{i : y_i \leq z\}| \;\geq\; t/6. \tag{2}$$

Let $P = \{x \in \mathcal{X} : d_{\mathcal{C}_f}(x) = z\} = \{x_1, \ldots, x_m\}$, and let q_j be the counters constructed in Step 6. By Lemma 6, every deterministic point lies on the true path, i.e. $B_i \subseteq I(c_f^*)$ for all i. Moreover, by Lemma 5, the set $I(c_f^*)$ is a chain and therefore contains a *unique* node at depth z; denote this node by x^*.

Now fix any index i with $y_i \geq z$. Since B_i is upward closed along the chain (if $x \in B_i$ and $x \preceq x'$ then $x' \in B_i$), and $b_i \in B_i$ has depth at least z, it follows that the unique depth-z node on the true path, namely x^*, belongs to B_i. Consequently, every such i contributes $+1$ to the counter corresponding to x^*, and no B_i contributes to any other node in P (because $B_i \subseteq I(c_f^*)$ and $I(c_f^*) \cap P = \{x^*\}$). Hence

$$q(x^*) \;\geq\; |\{i : y_i \geq z\}| \;\geq\; t/6, \qquad \text{and} \qquad q(x) = 0 \;\text{ for all } x \in P \setminus \{x^*\}.$$

Next, observe that the score function is 1-bounded: changing one example in S affects only one block S_i, and for a fixed i the set $B_i \cap P$ has size at most 1, so at most one counter q_j can increase by 1. Therefore we may apply the choosing mechanism (Lemma 1) to select a point in P. Since $q(x^*) \geq t/6$ while all other scores are 0, and t is chosen so that

$$t/6 \;\geq\; \frac{16}{\varepsilon} \log\!\left(\frac{4t}{\beta\varepsilon\delta}\right),$$

Lemma 1 implies that with probability at least $1 - \beta$ the algorithm outputs $x_{\mathrm{good}} = x^*$. Combining with the median event (2), this holds with probability at least $1 - 2\beta$.

It remains to bound the error of h_{good}. On the event (2), choose an index i_0 with $y_{i_0} \leq z$. Then b_{i_0} is at depth at most z, and since both b_{i_0} and $x_{\mathrm{good}} = x^*$ lie on the same chain, we have $x_{\mathrm{good}} \preceq b_{i_0}$. Because B_{i_0} is upward closed, this implies

$$B_{i_0} \;\subseteq\; I_{\mathrm{good}} = \{x : x_{\mathrm{good}} \preceq x\}.$$

Moreover, since $x_{\mathrm{good}} = x^* \in I(c_f^*)$, we have $I_{\mathrm{good}} \subseteq I(c_f^*)$, i.e., every point in I_{good} has label 1 under c_f^*.

Therefore, on every point where h_{good} and h_{i_0} disagree (namely points in $I_{\mathrm{good}} \setminus B_{i_0}$), we have $c_f^*(x) = 1$ and $h_{\mathrm{good}}(x) = 1$, so h_{good} is correct on all such disagreements. Hence

$$\mathrm{err}_{\mathcal{D}}(h_{\mathrm{good}}, c_f^*) \;\leq\; \mathrm{err}_{\mathcal{D}}(h_{i_0}, c_f^*).$$

Conditioned on all h_i being α-good, in particular $\mathrm{err}_{\mathcal{D}}(h_{i_0}, c_f^*) \leq \alpha$, and thus $\mathrm{err}_{\mathcal{D}}(h_{\mathrm{good}}, c_f^*) \leq \alpha$ as claimed. □

Corollary 2. *With probability at least $1-(t+2)\beta$, OPTPrivateLearner outputs a hypothesis $\hat{h}_{\mathrm{good}}$ satisfying*

$$\mathrm{err}_{\mathcal{D}}(\hat{h}_{\mathrm{good}}, c^*) \leq \alpha.$$

Proof. By construction, $\hat{h}_{\mathrm{good}}$ is obtained from h_{good} by undoing the f-representation, hence

$$\mathrm{err}_{\mathcal{D}}(\hat{h}_{\mathrm{good}}, c^*) = \mathrm{err}_{\mathcal{D}}(h_{\mathrm{good}}, c_f^*).$$

Let $\mathcal{E}_1$ be the event that all h_i are α-good; by Lemma 7, $\Pr[\mathcal{E}_1] \geq 1 - t\beta$. Let $\mathcal{E}_2$ be the event that the median step succeeds, and $\mathcal{E}_3$ the event that the choosing step succeeds; each holds with probability at least $1 - \beta$. Conditioned on $\mathcal{E}_1$, Lemma 9 implies that on $\mathcal{E}_2 \cap \mathcal{E}_3$ we have $\mathrm{err}_{\mathcal{D}}(h_{\mathrm{good}}, c_f^*) \leq \alpha$. Therefore, by a union bound,

$$\Pr\big[\mathrm{err}_{\mathcal{D}}(\hat{h}_{\mathrm{good}}, c^*) \leq \alpha\big] \;\geq\; 1 - (t\beta + \beta + \beta) \;=\; 1 - (t+2)\beta.$$

□

Finally, set $\varepsilon \leftarrow \varepsilon/2$, $\delta \leftarrow \delta/2$, and $\beta \leftarrow \beta/(t+2)$, and apply Corollary 1. This yields the stated main result.

Theorem 11. *For any concept class $\mathcal{C}$ with $\mathrm{VC}(\mathcal{C}) = 1$ and $\mathrm{Ldim}(\mathcal{C}) = d$, there exists an (ε, δ)-differentially private algorithm that (α, β)-PAC learns $\mathcal{C}$ given*

$$N \;\geq\; O\left(\frac{\log^* d \cdot \log^2\!\left(\frac{\log^* d}{\varepsilon\beta\delta}\right)}{\varepsilon} \cdot \frac{1}{\alpha}\left(\log\!\left(\frac{1}{\alpha}\right) + \log\!\left(\frac{1}{\beta}\right)\right)\right) = \widetilde{O}_{\beta,\delta}\left(\frac{\log^* d}{\alpha\varepsilon}\right).$$

E Proper Learner

When $\mathcal{C}$ is not a maximum class, the hypothesis $\hat{h}_{\mathrm{good}}$ produced by the improper learner may not belong to $\mathcal{C}$. From the analysis of the improper learner, with high probability $\hat{h}_{\mathrm{good}}$ misclassifies at most an $O(\alpha)$ fraction (under the underlying distribution) of the points labeled 1 in the tree representation. Therefore, in the proper-learning step it suffices to control the additional error contributed by points labeled 0.

E.1 Construction of the Proper Learner

We continue working with the tree T constructed from $(\mathcal{X}, \mathcal{C}_f)$. Unlike the improper learner, we restrict attention to a subtree T_{good} rooted at x_{good} such that *only leaves* correspond to concepts in $\mathcal{C}$ (equivalently, to concepts in $\mathcal{C}_f$).

E.1.1 Additional Definitions and Notation

Definition 17 (Proper node)**.** *A node x is* proper *if there exists $c_f \in \mathcal{C}_f$ such that*

$$I(c_f) = \{y \in \mathcal{X} : x \preceq y\}.$$

Equivalently, x induces a hypothesis whose set of 1-points is exactly the upset of x.

Definition 18 (Weight)**.** *Let T be the tree from Algorithm 1 and let S be a labeled dataset (in the f-representation). For a node $x \in T$, define its* weight

$$w(x) \;:=\; \big|\{(x', 0) \in S : x' \preceq x\}\big|,$$

i.e., the number of label-0 sample points that lie in the subtree rooted at x (with respect to $\preceq$).

Definition 19 (Value)**.** *Let T_{good} be a subtree rooted at x_{good} and let S be a labeled dataset (in the f-representation). For a node $x \in T_{\mathrm{good}}$, define its* value

$$v(x) \;:=\; \big|\{(x', 0) \in S : x \preceq x' \prec x_{\mathrm{good}}\}\big|,$$

i.e., the number of label-0 sample points lying on the (unique) path from x up to x_{good} (excluding x_{good}).

Definition 20 (Subtree T_x)**.** *For a node $x \in T_{\mathrm{good}}$, let T_x denote the subtree of T_{good} rooted at x.*

E.1.2 Subtree Construction

We first build T_{good} so that proper nodes appear exactly at the leaves (Algorithm 3).

We additionally contract degree-1 internal nodes so that every internal node has at least two children (Algorithm 4).

Algorithm 3: MakeSubTree

Input: a tree T from Algorithm 1 and a node $x_{\mathrm{good}} \in \mathcal{X}$ from Algorithm 2.
Output: a subtree T_{good} rooted at x_{good}.

1. If x_{good} is proper (Definition 17), output the single-node tree containing only x_{good}.
2. Delete every node x that is either incomparable with x_{good}, or lies strictly above x_{good} (i.e., $x_{\mathrm{good}} \prec x$). (This makes x_{good} the root of the remaining tree.)
3. Delete every node x for which there exists a proper node x' satisfying $x \prec x' \prec x_{\mathrm{good}}$. (This removes everything strictly below the first proper node on each branch, ensuring proper nodes can only appear as leaves.)
4. Output the resulting tree as T_{good}.

Algorithm 4: ContractUnaryNodes

Input: a tree T_{good}.
Output: the tree obtained by repeatedly contracting any node with exactly one child.

1. While there exists a node x with a single child y, contract the edge (x, y) (remove x and connect its parent directly to y).

E.1.3 The Proper Learner

The proper learner is presented in Algorithm 5.

E.2 Analysis

E.2.1 Privacy

When the differing datum lies in S_1, privacy follows from Theorem 11. Hence we focus on the case where the differing datum lies in S_2.

Theorem 12. *For any $\delta' > 0$, Algorithm 5 is*

$$\big(O(\sqrt{T \ln(1/\delta')} \cdot \varepsilon),\ \delta'\big)\text{-differentially private},$$

where $T = \lceil 2/\alpha \rceil$.

Proof. The only privacy loss from S_2 occurs inside the loop, where each iteration uses (i) one Laplace release and (ii) one exponential-mechanism selection, each with privacy parameter ε. Thus each iteration is $(2\varepsilon, 0)$-DP with respect to S_2. The claim follows from advanced composition (Theorem 7) and post-processing. □

E.2.2 Accuracy

We say a node x *represents an α-accurate hypothesis* if the hypothesis induced by x,

$$h_x(\cdot) := \mathbf{1}[x \preceq \cdot],$$

satisfies $\Pr_{u \sim \mathcal{D}}[h_x(u) \neq c_f^*(u)] \leq \alpha$.

Claim 1. *Conditioned on the event that x_{good} represents an α-accurate hypothesis (with respect to c_f^* and $\mathcal{D}$), let S_2 be the (re-labeled) sample used in the proper-learning stage, with $|S_2| = N_2$. Then*

$$\Pr\Big[\big|\{(x,1) \in S_2 : x \prec x_{\mathrm{good}}\}\big| > 2\alpha N_2\Big] \ \leq\ \exp\left(-\frac{\alpha N_2}{3}\right).$$

In particular, if $N_2 \geq \frac{3}{\alpha}\ln(1/\beta)$, then with probability at least $1-\beta$, at most $2\alpha N_2$ sample points in S_2 with label 1 lie below x_{good}.

Algorithm 5: ProperLearner

Parameters: $\alpha, \beta > 0$ and privacy (ε, δ).
Input: sample $S \in (\mathcal{X} \times \{0,1\})^{N_1+N_2}$, where N_1 is as in Theorem 11, and

$$N_2 = O\left(\frac{\log(1/\alpha) + \log(1/\beta)}{\alpha^2\,\varepsilon}\right).$$

Output: a *proper* hypothesis in $\mathcal{C}$.

1. Randomly split S into S_1 and S_2 with $|S_1| = N_1$ and $|S_2| = N_2$.
2. Run Algorithm 2 on S_1 to obtain x_{good} and the corresponding hypothesis $\hat{h}_{\mathrm{good}}$. (Record the associated $f \in \mathcal{C}$ chosen by MakeTree. Convert labels in S_2 to the same f-representation.)
3. If x_{good} is a proper node, output $\hat{h}_{\mathrm{good}}$ (it is then proper by Definition 17).
4. Compute weights $w(\cdot)$ on T using only the label-0 points in S_2 (Definition 18).
5. Construct T_{good} rooted at x_{good} using Algorithm 3, contract unary nodes via Algorithm 4, and compute values $v(\cdot)$ on T_{good} using S_2 (Definition 19).
6. Initialize $x_{\mathrm{flag}} \leftarrow x_{\mathrm{good}}$ and set $T := \lceil 2/\alpha \rceil$.
7. For $i = 1, 2, \ldots, T$:
   (a) Let X_{children} be the children of x_{flag} in T_{good}.
   (b) Let $w_{\min} = \min_{x \in X_{\mathrm{children}}} w(x)$ and release $\tilde{w}_{\min} = w_{\min} + \mathrm{Lap}(1/\varepsilon)$.
   (c) If $\tilde{w}_{\min} \le \alpha N_2$ ("non-uniform" case): select x_{new} from X_{children} using the exponential mechanism with quality $q(x) = -w(x)$ and privacy ε, set $x_{\mathrm{flag}} \leftarrow x_{\mathrm{new}}$, and **break**.
   (d) If $\tilde{w}_{\min} > \alpha N_2$ ("uniform" case): for each child x, let $L(x)$ be the leaves of T_x and define $q(x) = -\min_{\ell \in L(x)} v(\ell)$. Use the exponential mechanism (privacy ε) to select $x_{\mathrm{new}} \in X_{\mathrm{children}}$, and set $x_{\mathrm{flag}} \leftarrow x_{\mathrm{new}}$.
8. Select any leaf ℓ of $T_{x_{\mathrm{flag}}}$. Output the proper hypothesis corresponding to ℓ by taking

$$I_\ell := \{x \in \mathcal{X} : \ell \preceq x\}, \qquad \hat{h}(x) := f(x) \oplus \mathbf{1}[x \in I_\ell].$$

Proof. Let $h_{\mathrm{good}}(x) = \mathbf{1}[x_{\mathrm{good}} \preceq x]$ be the hypothesis represented by x_{good}. Points satisfying $x \prec x_{\mathrm{good}}$ are exactly those on which $h_{\mathrm{good}}(x) = 0$. Since x_{good} is α-accurate, we have

$$p \;:=\; \Pr_{x \sim \mathcal{D}}\big[c_f^*(x) = 1 \;\wedge\; x \prec x_{\mathrm{good}}\big] \;\le\; \mathrm{err}_{\mathcal{D}}(h_{\mathrm{good}}, c_f^*) \;\le\; \alpha.$$

Let

$$Z \;:=\; \big|\{(x,1) \in S_2 : x \prec x_{\mathrm{good}}\}\big|.$$

Then $Z \sim \mathrm{Bin}(N_2, p)$. For a fixed threshold, $\Pr[Z > 2\alpha N_2]$ is maximized at the largest allowed p, so it suffices to upper bound the case $p = \alpha$. Let $\mu = \mathbb{E}[Z] = \alpha N_2$. By the multiplicative Chernoff bound,

$$\Pr[Z > 2\mu] \;\le\; \exp(-\mu/3) \;=\; \exp\left(-\frac{\alpha N_2}{3}\right).$$

The final "in particular" statement follows by requiring $\exp(-\alpha N_2/3) \le \beta$. □

Claim 2. *There exists a leaf ℓ_{good} of T_{good} such that $v(\ell_{\mathrm{good}}) = 0$.*

Proof. Let x^* denote the node corresponding to the true target concept c_f^* (so that $I(c_f^*) = \{x : x^* \preceq x\}$). By the analysis of the improper learner, x_{good} lies on the true path, i.e., $x^* \preceq x_{\mathrm{good}}$.

Consider the unique branch of T_{good} that contains x^*. Since x^* is a proper node, and by construction of T_{good} all proper nodes appear only as leaves, this branch contains a (unique) leaf; let ℓ_{good} be that leaf. In particular,

$$x^* \preceq \ell_{\mathrm{good}} \prec x_{\mathrm{good}}.$$

Now take any node u on the path from ℓ_{good} up to x_{good} (excluding x_{good}). Because $\ell_{\text{good}} \preceq u$ and $x^* \preceq \ell_{\text{good}}$, we have $x^* \preceq u$, hence $c_f^*(u) = 1$. Under the realizability assumption (labels are generated by c_f^*), no sample point at such a node can have label 0. Therefore there is no pair $(u, 0)$ in the sample along this path, and by Definition of $v(\cdot)$ we obtain $v(\ell_{\text{good}}) = 0$. □

Claim 3. *With probability at least $1 - T\beta$, Algorithm 5 enters Step 7c (the "non-uniform" case) in some iteration of the loop Step 7.*

Proof. (Throughout this proof, $N_2 := |S_2|$ is the sample size used in the proper-learning loop.)

In iteration $r \in [T]$, let $\eta_r \sim \text{Lap}(1/\varepsilon)$ be the noise added to $w_{\min}$, so that $\tilde{w}_{\min} = w_{\min} + \eta_r$. Define the event

$$\mathcal{E} := \bigcap_{r=1}^{T} \left\{ |\eta_r| \le \frac{1}{\varepsilon} \ln\left(\frac{1}{2\beta}\right) \right\}.$$

Since $\Pr[|\eta_r| > u] = e^{-\varepsilon u}$ for $\eta_r \sim \text{Lap}(1/\varepsilon)$, we have $\Pr[\mathcal{E}] \ge 1 - T\beta$ by a union bound.

Condition on $\mathcal{E}$. Suppose the algorithm takes the "uniform" branch (Step 7d) in some iteration. Then $\tilde{w}_{\min} > \alpha N_2$. Under $\mathcal{E}$ and assuming $\frac{1}{\varepsilon} \ln(\frac{1}{2\beta}) \le \alpha N_2/2$, we get

$$w_{\min} \;\ge\; \tilde{w}_{\min} - |\eta_r| \;>\; \alpha N_2 - \frac{\alpha N_2}{2} \;=\; \frac{\alpha N_2}{2}.$$

When the algorithm moves from the current node x_{flag} to the chosen child x_{new}, it discards all other children subtrees. After contracting unary nodes, x_{flag} has at least two children; hence the total discarded weight is at least $w_{\min}$, and therefore the remaining total weight decreases by at least $\alpha N_2/2$ in every iteration that takes the uniform branch.

Initially, the total weight below x_{good} is at most N_2 (since $w(\cdot)$ counts label-0 points from S_2). Thus after $T - 1 = \lceil 2/\alpha \rceil - 1$ uniform iterations, the remaining total weight is

$$\le\; N_2 - (T-1) \cdot \frac{\alpha N_2}{2} \;\le\; \frac{\alpha N_2}{2}.$$

At the next iteration, since there are at least two children, we must have $w_{\min} \le \frac{1}{2} \cdot \frac{\alpha N_2}{2} = \frac{\alpha N_2}{4}$. Under $\mathcal{E}$,

$$\tilde{w}_{\min} \;\le\; w_{\min} + |\eta_r| \;\le\; \frac{\alpha N_2}{4} + \frac{\alpha N_2}{2} \;<\; \alpha N_2,$$

so the algorithm must enter Step 7c, contradicting the assumption that it stayed in the uniform case.

Therefore, conditioned on $\mathcal{E}$, the algorithm reaches Step 7c within T iterations. Since $\Pr[\mathcal{E}] \ge 1 - T\beta$, the claim follows. □

Claim 4. *Assume Claim 3 holds, i.e., Algorithm 5 reaches Step 7c within T iterations. Then, with probability at least $1 - T\beta$, in every iteration in which the algorithm executes Step 7d, the selected child x_{new} satisfies*

$$q(x_{\text{new}}) \;\ge\; -2\alpha N_2 \;-\; T \cdot \frac{2\ln(\frac{2}{\alpha\beta})}{\varepsilon},$$

where $N_2 = |S_2|$ and the quality function in Step 7d is $q(x) = -\min_{\ell \in L(x)} v(\ell)$.

Proof. We analyze only iterations in which the algorithm runs Step 7d. In such an iteration, the algorithm chooses $x_{\text{new}} \in X_{\text{children}}$ using the exponential mechanism with quality $q(x) = -\min_{\ell \in L(x)} v(\ell)$.

Bounding $|X_{\text{children}}|$. Whenever Step 7d is executed we have (by definition of the branch) $\tilde{w}_{\min} > \alpha N_2$. On the high-probability event from Claim 3 (noise is at most $\alpha N_2/2$ in magnitude), this implies $w_{\min} \ge \alpha N_2/2$. Since $\sum_{x \in X_{\text{children}}} w(x) \le N_2$, we obtain

$$|X_{\text{children}}| \;\le\; \frac{N_2}{w_{\min}} \;\le\; \frac{2}{\alpha}.$$

One-step guarantee. The score function q has sensitivity 1, so by the standard guarantee of the exponential mechanism (Theorem 8), in any fixed uniform-case iteration, with probability at least $1-\beta$ the selected x_{new} satisfies

$$q(x_{\text{new}}) \;\geq\; \max_{x\in X_{\text{children}}} q(x) \;-\; \frac{2}{\varepsilon}\ln\!\left(\frac{|X_{\text{children}}|}{\beta}\right) \;\geq\; \max_{x\in X_{\text{children}}} q(x) \;-\; \frac{2}{\varepsilon}\ln\!\left(\frac{2}{\alpha\beta}\right).$$

Induction over iterations. Let r index the uniform-case iterations (there are at most T of them). For the base case $r=1$, we use that (by the existence of a good leaf and the preceding accuracy bounds) $\max_{x\in X_{\text{children}}} q(x) \geq -2\alpha N_2$, and thus, with probability at least $1-\beta$,

$$q(x_{\text{new}}) \;\geq\; -2\alpha N_2 \;-\; \frac{2}{\varepsilon}\ln\!\left(\frac{2}{\alpha\beta}\right).$$

For the inductive step, assume that before iteration r we have maintained

$$q(x_{\text{flag}}) \;\geq\; -2\alpha N_2 \;-\; (r-1)\cdot\frac{2}{\varepsilon}\ln\!\left(\frac{2}{\alpha\beta}\right).$$

Since $\min_{\ell\in L(x_{\text{flag}})} v(\ell)$ is achieved in some child subtree, we have $\max_{x\in X_{\text{children}}} q(x) \geq q(x_{\text{flag}})$. Applying the one-step guarantee gives, with probability at least $1-\beta$,

$$q(x_{\text{new}}) \;\geq\; q(x_{\text{flag}}) \;-\; \frac{2}{\varepsilon}\ln\!\left(\frac{2}{\alpha\beta}\right) \;\geq\; -2\alpha N_2 \;-\; r\cdot\frac{2}{\varepsilon}\ln\!\left(\frac{2}{\alpha\beta}\right).$$

Finally, taking a union bound over at most T uniform-case iterations yields that all these bounds hold simultaneously with probability at least $1-T\beta$. □

Claim 5. *Assume we are in an iteration where Algorithm 5 executes Step 7c. Suppose further that*

$$q(x_{\text{flag}}) \;\geq\; -2\alpha N_2 \;-\; T\cdot\frac{2\ln\!\big(2/(\alpha\beta)\big)}{\varepsilon},$$

where $q(x) = -\min_{\ell\in L(x)} v(\ell)$ (as in Step 7d) and $N_2 := |S_2|$.

Then, with probability at least $1-\beta/\alpha$, the exponential mechanism in Step 7c (run on X_{children} with score $q'(x) = -w(x)$) selects a child $x_{\text{new}} \in X_{\text{children}}$ satisfying

$$q(x_{\text{new}}) \;\geq\; -4\alpha N_2 \;-\; T\cdot\frac{2\ln\!\big(2/(\alpha\beta)\big)}{\varepsilon}.$$

Equivalently, $\min_{\ell\in L(x_{\text{new}})} v(\ell) \leq 4\alpha N_2 + T\cdot\frac{2\ln(2/(\alpha\beta))}{\varepsilon}$.

Proof. **Step 1: the exponential mechanism outputs a light child.** In Step 7c we have $\tilde{w}_{\min} \leq \alpha N_2$. On the same high-probability event used in Claim 3, the Laplace noise satisfies $|\eta| \leq \alpha N_2/2$, hence

$$w_{\min} \;\leq\; \tilde{w}_{\min} + |\eta| \;\leq\; \frac{3\alpha N_2}{2}.$$

In particular, there exists at least one child with weight at most $3\alpha N/2$. Define

$$X_{\text{good}} := \{x \in X_{\text{children}} : w(x) \leq 3\alpha N_2/2\}, \qquad X_{\text{verybad}} := \{x \in X_{\text{children}} : w(x) > 2\alpha N_2\}.$$

Then $X_{\text{good}} \neq \emptyset$.

The exponential mechanism in Step 7c selects $x \in X_{\text{children}}$ with probability proportional to $\exp\!\big(-\varepsilon w(x)/2\big)$. Therefore,

$$\Pr[x_{\text{new}} \in X_{\text{verybad}}] \;\leq\; \frac{\sum_{x\in X_{\text{verybad}}} e^{-\varepsilon w(x)/2}}{\sum_{x\in X_{\text{good}}} e^{-\varepsilon w(x)/2}} \;\leq\; \frac{|X_{\text{children}}|\cdot e^{-\varepsilon\cdot(2\alpha N_2)/2}}{e^{-\varepsilon\cdot(3\alpha N_2/2)/2}} \;=\; |X_{\text{children}}|\cdot e^{-\varepsilon\alpha N_2/4}.$$

In Step 7d (and hence also when we reach Step 7c), we have $w_{\min} \geq \tilde{w}_{min} - |\eta| \geq \alpha N_2/2$, which implies $|X_{\text{children}}| \leq 2/\alpha$ since $\sum_{x\in X_{\text{children}}} w(x) \leq N_2$. Thus

$$\Pr[x_{\text{new}} \in X_{\text{verybad}}] \;\leq\; \frac{2}{\alpha}\cdot e^{-\varepsilon\alpha N_2/4}.$$

By choosing N large enough so that $2e^{-\varepsilon\alpha N_2/4} \le \beta$ (equivalently, $N_2 \ge \frac{4}{\varepsilon\alpha}\ln(2/\beta)$), we obtain

$$\Pr[x_{\text{new}} \in X_{\text{verybad}}] \;\le\; \beta/\alpha.$$

Hence, with probability at least $1-\beta/\alpha$, we have $w(x_{\text{new}}) \le 2\alpha N_2$.

Step 2: translating $w(x_{\text{new}})$ into a lower bound on $q(x_{\text{new}})$. Recall $q(x) = -\min_{\ell\in L(x)} v(\ell)$. Since v is nondecreasing as we move down the tree, for every leaf $\ell \in L(x_{\text{flag}})$ we have $v(x_{\text{flag}}) \le v(\ell)$, and therefore

$$v(x_{\text{flag}}) \;\le\; \min_{\ell\in L(x_{\text{flag}})} v(\ell) \;=\; -q(x_{\text{flag}}) \;\le\; 2\alpha N_2 + T\cdot\frac{2\ln\bigl(2/(\alpha\beta)\bigr)}{\varepsilon}.$$

Now consider any leaf $\ell \in L(x_{\text{new}})$. The additional label-0 points on the path from x_{new} down to ℓ lie inside the subtree rooted at x_{new}, and are therefore counted by $w(x_{\text{new}})$. Hence

$$v(\ell) \;\le\; v(x_{\text{flag}}) + w(x_{\text{new}}) \;\le\; 2\alpha N_2 + T\cdot\frac{2\ln\bigl(2/(\alpha\beta)\bigr)}{\varepsilon} + 2\alpha N_2.$$

Taking the minimum over $\ell \in L(x_{\text{new}})$ and negating yields

$$q(x_{\text{new}}) \;=\; -\min_{\ell\in L(x_{\text{new}})} v(\ell) \;\ge\; -4\alpha N_2 \;-\; T\cdot\frac{2\ln\bigl(2/(\alpha\beta)\bigr)}{\varepsilon},$$

as claimed. □

Claim 6. *Conditioned on $N_2 \ge \frac{4\lg(2/(\alpha\beta))}{\alpha^2\varepsilon}$, Algorithm 5 is an $(O(\alpha),\, O(\beta/\alpha))$-PAC learner.*

Proof. Let S_2 be the second subsample used by the proper learner, with $|S_2| = N_2$ (after converting labels to the f-representation). Notice that we select a leaf ℓ satisfying $v(\ell) \le 5\alpha N_2$ by Claim 5, which means the output hypothesis disagree at most $5\alpha N_2$ examples with label 0 in S_2. From the preceding analysis (Claims 1, 2, 3, and 5), with probability at least $1 - O(\beta/\alpha)$ the algorithm (of improper learner) outputs a hypothesis $\hat{h}_{\text{good}}$ whose empirical error on S_2 satisfies

$$\text{err}_{S_2}(\hat{h}_{\text{good}}) \;\le\; c\,\alpha$$

for some absolute constant $c > 0$.

Set $\alpha' := 10c\,\alpha$ and $\beta' := \Theta(\beta/\alpha)$. Then the event $\text{err}_{S_2}(\hat{h}_{\text{good}}) \le c\alpha$ implies $\text{err}_{S_2}(\hat{h}_{\text{good}}) \le \alpha'/10$. By Theorem 3 (with $\text{VC}(\mathcal{C}) = 1$), provided N_2 is large enough as specified there (with parameters α', β'), we have

$$\Pr\bigl[\text{err}_{\mathcal{D}}(\hat{h}_{\text{good}}) > \alpha' \text{ and } \text{err}_{S_2}(\hat{h}_{\text{good}}) \le \alpha'/10\bigr] \;\le\; \beta'.$$

Combining this with the probability $1 - O(\beta/\alpha)$ of the empirical-error event above (and absorbing constants into $O(\cdot)$) yields that, with probability at least $1 - O(\beta/\alpha)$,

$$\text{err}_{\mathcal{D}}(\hat{h}_{\text{good}}) \;\le\; \alpha' \;=\; O(\alpha).$$

Therefore Algorithm 5 is an $(O(\alpha), O(\beta/\alpha))$-PAC learner. □

Therefore, setting

$$\varepsilon \leftarrow \frac{\varepsilon}{\sqrt{2T\ln(2/\delta)}} \quad (\text{with } T = \Theta(1/\alpha)), \qquad \delta \leftarrow \frac{\alpha\delta}{2}, \qquad \delta' \leftarrow \frac{\delta}{2}, \qquad \beta \leftarrow \alpha\beta,$$

and applying the advanced composition bound over the $T = O(1/\alpha)$ iterations, we obtain the stated sample complexity for privately proper learning.

Theorem 13. *For any concept class $\mathcal{C}$ with $\text{VC}(\mathcal{C}) = 1$ and $\text{Ldim}(\mathcal{X},\mathcal{C}) = d$, there exists an (ε,δ)-differentially private algorithm that* properly *(α,β)-PAC learns $\mathcal{C}$ given a labeled sample of size*

$$N \;\ge\; O\left(\frac{\log^* d\cdot\log^2\!\left(\frac{\log^* d}{\varepsilon\beta\delta}\right)}{\varepsilon}\cdot\frac{1}{\alpha}\left(\log\!\left(\frac{1}{\alpha}\right)+\log\!\left(\frac{1}{\beta}\right)\right)+\frac{(\log(1/\alpha)+\log(1/\alpha\beta))\sqrt{\log(1/\delta)}}{\alpha^{2.5}\varepsilon}\right)$$
$$= \widetilde{O}_{\beta,\delta}\left(\frac{\log^* d}{\alpha\varepsilon}+\frac{1}{\alpha^{2.5}\varepsilon}\right).$$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: This paper researches the sample complexity of private learning on concept classes with VC dimension 1. All the theorems in this paper only work for concept classes with VC dimension 1.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification:

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: This is theoretical research, and there is no experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: This is theoretical research, and there is no experiment or code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: This is theoretical research, and there is no experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: This is theoretical research, and there is no experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: This is theoretical research, and there is no experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.