Improved Bounds for Pure Private Agnostic Learning: Item-Level and User-Level Privacy

Bo Li [1] **Wei Wang** [1] **Peng Ye** [1]

Abstract

Machine Learning has made remarkable progress in a wide range of fields. In many scenarios, learning is performed on datasets involving sensitive information, in which privacy protection is essential for learning algorithms. In this work, we study pure private learning in the agnostic model – a framework reflecting the learning process in practice. We examine the number of users required under item-level (where each user contributes one example) and user-level (where each user contributes multiple examples) privacy and derive several improved upper bounds. For item-level privacy, our algorithm achieves a near optimal bound for general concept classes. We extend this to the user-level setting, rendering a tighter upper bound than the one proved by Ghazi et al. (2023). Lastly, we consider the problem of learning thresholds under user-level privacy and present an algorithm with a nearly tight user complexity.

1. Introduction

Differential Privacy (DP) (Dwork et al., 2006b;a) is a mathematical definition for measuring the privacy of algorithms. An algorithm is considered private if the presence or absence of a single user does not significantly affect the output. Due to its soundness and quantifiability, DP has become the gold standard for ensuring data privacy and has been employed by the industry (Apple Differential Privacy Team, 2017) and the governments (Abowd, 2018).

Machine learning models are usually trained on datasets that contain sensitive data (e.g., in medical or financial applications). Thus, it is necessary to design privacy-preserving machine learning algorithms. Kasiviswanathan et al. (2011) initiated the study of private learning and defined private PAC learning as a combination of probably approximately correct (PAC) learning (Valiant, 1984) and differential privacy. The most important measure for the utility of a learner is the minimum amount of data required to find a hypothesis that achieves some target accuracy, which is called the sample complexity. Subsequent work (e.g., (Feldman & Xiao, 2014; Alon et al., 2019)) showed that the privacy constraint makes the sample complexity much higher than the non-private setting.

The PAC model assumes that all the samples are labeled by a concept in the given concept class. This is called the realizable assumption. However, in many situations, such an assumption could be unrealistic. For example, the collected data may be noisy, or one may choose an inappropriate concept class that cannot classify the samples perfectly. The agnostic learning model, introduced by Haussler (1992) and Kearns et al. (1994), addresses this issue by requiring the learner to output a hypothesis within some additive error compared to the best one in the concept class. Motivated by this, Kasiviswanathan et al. (2011) also defined private agnostic learning with the agnostic model replacing the PAC model in private PAC learning.

For pure private realizable learning, tight sample complexity bounds were shown by Beimel et al. (2019). However, for agnostic learning, there were no such tight results. Though several algorithms were proposed (Kasiviswanathan et al., 2011; Beimel et al., 2015; Alon et al., 2020), none of them achieve a sample complexity that matches the trivial lower bound, i.e., the one combines lower bounds of private realizable learning and non-private agnostic learning. On the other hand, there have been no non-trivial lower bounds obtained in the literature. Thus, a gap exists, leading to the following question:

> *What is the sample complexity of pure private agnostic learning?*

What we have discussed so far only considers the situation in that each user contributes one example, which we refer to as *item-level* DP. In practice, however, one user may have many items (e.g., in federated learning (Kairouz et al., 2021)). Here, the goal becomes protecting all the examples contributed by a single user. In this *user-level* DP setting,

[1]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China. Correspondence to: Peng Ye <pyeac@connect.ust.hk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

we are more interested in the user complexity of learning algorithms, i.e., the number of users needed to achieve the target accuracy.

User-level pure private PAC learning was first studied by Ghazi et al. (2021b), who showed that compared to the item-level setting, one could learn with fewer users if each user contributes sufficiently many examples. The work of Ghazi et al. (2023) further tightly determines the user complexity as a function of the number of examples per user. They also proved upper and lower bounds for the agnostic model. However, like the item-level setting, there is a gap between the upper and lower bounds. Thus, it is natural to ask:

Can we obtain tighter bounds on the user complexity of pure private agnostic learning?

1.1. Our Results

In this work, we study the problem of agnostic learning with pure differential privacy. We list our results as follows. The formal definitions of our notions can be found in Section 2.

Item-Level DP. We show that every concept class $\mathcal{C}$ can be learned up to excess error α by an ε-differentially private algorithm using $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{\alpha^2}\right)$ samples. Compared with private realizable learning (Beimel et al., 2019), our result only incurs an additive term of $\widetilde{O}\left(\frac{\mathrm{VC}(\mathcal{C})}{\alpha^2}\right)$, which is known to be tight up to polylogarithmic factors even for non-private agnostic learning. In contrast, previous results (Kasiviswanathan et al., 2011; Beimel et al., 2019; Alon et al., 2020) incur higher sample complexity than ours.

User-Level DP. We provide a generic learner that agnostically learns any $\mathcal{C}$ using $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\varepsilon} + \frac{\mathrm{RepD}(\mathcal{C})}{\sqrt{m}\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{m\alpha^2}\right)$ users, where each user holds m examples. This improves the previous result of Ghazi et al. (2023), wherein the third term is $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{m\alpha^2}\right)$. As indicated by their lower bound, the third term in our upper bound is optimal while a gap still exists regarding the second term.

Thresholds with User-Level DP. For the case that $\mathcal{C}$ is the class of thresholds over some domain $\mathcal{X}$, we prove that the user complexity upper bound can be further improved to $\widetilde{O}\left(\frac{\log|\mathcal{X}|}{\varepsilon} + \frac{\log|\mathcal{X}|}{m\alpha\varepsilon} + \frac{1}{\sqrt{m}\alpha\varepsilon} + \frac{1}{m\alpha^2}\right)$, matching the lower bound proved by Ghazi et al. (2023)[1].

[1] In their work, they claimed that their algorithm can achieve this complexity for thresholds. We have confirmed with them that the claim made a mistake – their algorithm still has a user complexity of $\widetilde{O}\left(\frac{\log|\mathcal{X}|}{\varepsilon} + \frac{\log|\mathcal{X}|}{\sqrt{m}\alpha\varepsilon} + \frac{1}{m\alpha^2}\right)$, which is not optimal.

The improvement made by our results is significant in the high-accuracy regime. To see this, fix m, ε and let $\alpha \to 0$. Our upper bound is dominated by $\widetilde{O}\left(\frac{\mathrm{VC}(\mathcal{C})}{m\alpha^2}\right)$, which even matches the cost of non-private agnostic learning (including the item-level setting, where $m = 1$). In other words, privacy is free – the privacy constraint can be satisfied with the same number of users. In comparison, the number of users required by the algorithm of Ghazi et al. (2023) is $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{m\alpha^2}\right)$, which can be much larger than ours as suggested by the results of Feldman & Xiao (2014).

We summarize the previous and new bounds in Table 1.

1.2. Technical Overview

We now provide a rough overview of our proofs.

Item-Level DP. We start by describing the algorithm proposed by Beimel et al. (2019). Their algorithm first samples a hypothesis class $\mathcal{H}$ with $\log|\mathcal{H}| = \widetilde{O}(\mathrm{RepD}(\mathcal{C}))$. Then, it applies the exponential mechanism, with the score function being the empirical error. By the utility property of the exponential mechanism, it returns some hypothesis with a small empirical error. To ensure that the generalization error of the output hypothesis is also small, we have to prove agnostic generalization for $\mathcal{H}$. This requires $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha^2}\right)$ samples, which is higher than our goal.

To reduce the sample complexity, we construct a surrogate error as the score function. Consider a hypothesis $h \in \mathcal{H}$. Note that by triangle inequality, for any concept c, the error of h is no more than the error of c plus the disagreement between c and h. Thus, by taking a minimum over $c \in \mathcal{C}$, we get an upper bound on the error of h. This quantity is the surrogate error of h. We then use the empirical surrogate error as the score function of h.

The exponential mechanism finds a hypothesis h_0 with a low empirical surrogate error. By our construction, we can prove that there exists some c_0 s.t. both the empirical error of c_0 and the empirical disagreement between c_0 and h_0 are small. To ensure they are also small on the distribution, we need to show agnostic generalization for $\mathcal{C}$ and realizable generalization for $\mathcal{C} \times \mathcal{H}$. The former needs $\widetilde{O}\left(\frac{\mathrm{VC}(\mathcal{C})}{\alpha^2}\right)$ samples, reducing the $\mathrm{RepD}(\mathcal{C})$ factor to $\mathrm{VC}(\mathcal{C})$. The latter can be satisfied by $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha}\right)$ samples, saving a $1/\alpha$ multiplicaive factor. We thus achieve our goal.

User-Level DP. We extend our item-level algorithm to the user-level. The high-level idea relies on a key observation from Adell & Jodrá (2006); Liu et al. (2020), which states that the total variation distance between two binomial distributions $\mathrm{Bin}(m, p)$ and $\mathrm{Bin}(m, q)$ scales as $\sqrt{m}|p - q|$. We define a notion called user-level error and use the observa-

Table 1. Summary of our results and previous results. For user-level DP, we assume $m \leq 1/\alpha^2$ for simplicity, since as shown by Ghazi et al. (2023), increasing m won't make the user complexity decrease when $m > 1/\alpha^2$.

Concept Class	DP	Lower Bound	Previous Upper Bound	Our Upper Bound
Arbitrary $\mathcal{C}$	Item	$\widetilde{\Omega}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{\alpha^2}\right)$ (Simon, 1996) and (Beimel et al., 2019)	$\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha\varepsilon} + \frac{\mathrm{RepD}(\mathcal{C})}{\alpha^2}\right)$ (Beimel et al., 2019) or $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{\alpha^2\varepsilon}\right)$ (Alon et al., 2020)	$\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{\alpha^2}\right)$ (Theorem 3.4)
Arbitrary $\mathcal{C}$	User	$\widetilde{\Omega}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\min(1,m\alpha)\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{\sqrt{m}\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{m\alpha^2}\right)$ (Ghazi et al., 2023)	$\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\sqrt{m}\alpha\varepsilon} + \frac{\mathrm{RepD}(\mathcal{C})}{m\alpha^2}\right)$ (Ghazi et al., 2023)	$\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\sqrt{m}\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{m\alpha^2}\right)$ (Theorem 4.6)
Thresholds over $\mathcal{X}$	User	$\widetilde{\Omega}\left(\frac{\log\lvert\mathcal{X}\rvert}{\min(1,m\alpha)\varepsilon} + \frac{1}{\sqrt{m}\alpha\varepsilon} + \frac{1}{m\alpha^2}\right)$ (Ghazi et al., 2023)	$\widetilde{O}\left(\frac{\log\lvert\mathcal{X}\rvert}{\sqrt{m}\alpha\varepsilon} + \frac{1}{m\alpha^2}\right)$ (Ghazi et al., 2023)	$\widetilde{O}\left(\frac{\log\lvert\mathcal{X}\rvert}{\min(1,m\alpha)\varepsilon} + \frac{1}{\sqrt{m}\alpha\varepsilon} + \frac{1}{m\alpha^2}\right)$ (Theorem 5.2)

tion to show that the user-level error amplifies the item-level error by $\sqrt{m}$ when there are m examples per user. That is, suppose hypothesis h has an additive error of α worse than the best concept in $\mathcal{C}$, its user-level error must be larger than the best by $\Omega(\sqrt{m}\alpha)$.

Thus, we can then apply our item-level algorithm with the item-level error replaced by the user-level error. For a target error α, we only have to find a hypothesis with a user-level error of $O(\sqrt{m}\alpha)$. This explains why fewer users are required in the user-level setting.

There are other technical issues in directly applying the item-level algorithm, such as user-level generalization lemmata and estimation of the minimum error. We develop some techniques to handle these. The details of these are presented in Section 4.

Thresholds with User-Level DP. Our algorithm for learning thresholds employs binary search. We start with the entire domain. In each iteration, we select the median w.r.t. the distribution of points in the current interval to split it into two parts. We add Laplace noise to the minimum user-level error of thresholds in each part and choose the one with a smaller error to continue. After $O(\log(1/\alpha))$ iterations, we will reach an interval with density at most $O(\alpha)$. By our user-level error results and the utility guarantee of the Laplace mechanism, $\widetilde{O}\left(\frac{1}{\sqrt{m}\alpha\varepsilon}\right)$ users are sufficient for any threshold in the final interval to have an additive error of α.

To select the median, we exploit an observation that the user-level setting amplifies the density by m. We use this to design an algorithm that approximates the median with a user complexity of $\widetilde{O}\left(\frac{\log\lvert\mathcal{X}\rvert}{m\alpha\varepsilon}\right)$, which matches our target.

1.3. Related Work

As the initial work of private learning, Kasiviswanathan et al. (2011) defined the notion of private learning and gave algorithms for finite concept classes. Beimel et al. (2014) showed that properly learning point functions under pure DP requires much more samples than non-private learning, indicating that the structure of learning becomes very different under the privacy restrictions. They also proposed the idea of representation to devise algorithms for pure private improper learning. This idea was further developed by the work of Beimel et al. (2019), which defined the notion of probabilistic representation dimension and showed that it characterizes the sample complexity of (improper) pure private learners. Feldman & Xiao (2014) showed an equivalence between the representation dimension and the communication complexity of the evaluation problem and used this relation to separate the sample complexity of pure private learning from non-private learning. For agnostic learning, Beimel et al. (2015) proposed a realizable-to-agnostic transformation that works for proper learners. Based on this result, Alon et al. (2020) proposed a more general one, which works for improper learners.

Though this work only considers pure DP, it is worth mentioning that there were also a great number of results on learning under approximate DP. It is a relaxed notion of pure DP that allows privacy to be violated with a negligible probability of δ. The most remarkable result is the equivalence between approximate private learning and (non-private) online learning (Alon et al., 2019; Bun et al., 2020; Ghazi et al., 2021a; Alon et al., 2022), which demonstrates a strong connection between these two tasks.

The study of user-level private learning PAC learnable classes was initiated by Ghazi et al. (2021b). They showed that for both pure and approximate DP, if each user contributes sufficiently many examples, the learning task can be done with much fewer users than in the item-level DP setting. Ghazi et al. (2023) improved this result by representing the user complexity as a function of the number of samples per user. For pure DP, they showed the exact

user complexity of private PAC learning and established non-trivial upper and lower bounds for agnostic learning. For approximate DP, they proposed a general transformation that can convert any item-level DP algorithm for any statistical tasks (including but not limited to learning) to a user-level one and yield a multiplicative saving of $\sqrt{m}$ on the number of users.

2. Preliminaries

Notation. We use $\widetilde{O}, \widetilde{\Omega}$ and $\widetilde{\Theta}$ to hide $\mathrm{polylog}(1/\alpha, 1/\beta)$ factors. Throughout the paper, we use m to denote the number of samples per user. The item-level setting refers to the case that $m = 1$.

We first recall the notion of DP. A dataset can be represented as $\boldsymbol{z} = (z_{1,1}, \ldots, z_{1,m}, z_{2,1}, \ldots, z_{n,m}) \in \mathcal{Z}^{nm}$, where $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,m})$ is the set of samples contributed by the i-th user. Two datasets are neighboring if one can be obtained from the other by adding or removing a single user.

Definition 2.1 (Differential Privacy (Dwork et al., 2006b;a))**.** A randomized algorithm A is ε-differentially private if for any neighboring datasets $\boldsymbol{z}$ and $\boldsymbol{z}'$, and any subset $\mathcal{O}$ of outputs, it holds that $\Pr[\mathsf{A}(\boldsymbol{z}) \in \mathcal{O}] \leq e^{\varepsilon} \Pr[\mathsf{A}(\boldsymbol{z}') \in \mathcal{O}]$.

2.1. Learning

Let $\mathcal{X}$ be an arbitrary domain[2] and $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$ be the domain of data samples. A concept $c : \mathcal{X} \to \{0, 1\}$ is a function that labels each (unlabeled) sample taken from $\mathcal{X}$ by either 0 or 1. A concept class $\mathcal{C}$ is a set of concepts over $\mathcal{X}$. The VC dimension of $\mathcal{C}$ is denoted by $\mathrm{VC}(\mathcal{C})$. We use $\mathcal{D}$ to denote a distribution over $\mathcal{Z}$, and $\mathcal{D}_{\mathcal{X}}$ to denote its marginal distribution over $\mathcal{X}$.

In this work, we focus on the agnostic learning model (Haussler, 1992; Kearns et al., 1994), where the data distribution can be arbitrary, and the goal of the learning algorithm is to produce a hypothesis whose generalization error is close to the best possible (by concepts in $\mathcal{C}$). The generalization error of a hypothesis $h : \mathcal{X} \to \{0, 1\}$ with respect to a distribution $\mathcal{D}$ over $\mathcal{Z}$ is defined as $\mathrm{err}_{\mathcal{D}}(h) = \Pr_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$.

Definition 2.2 (Agnostic Learning)**.** We say a learning algorithm A is an (α, β)-agnostic learner for concept class $\mathcal{C}$ if for any distribution $\mathcal{D}$ on $\mathcal{X} \times \{0, 1\}$, it takes in a dataset $\boldsymbol{z}$, where each $z_{i,j}$ is drawn i.i.d. from $\mathcal{D}$, and outputs a hypothesis h such that

$$\mathrm{err}_{\mathcal{D}}(h) \leq \inf_{c \in \mathcal{C}} \mathrm{err}_{\mathcal{D}}(c) + \alpha$$

with probability at least $1 - \beta$. The probability is over the random choice of the samples and the coin tosses of A.

[2]We assume $\mathcal{X}$ is countable to avoid making technical measurability assumptions.

In the standard PAC learning model (Valiant, 1984), it is assumed that there is some $c \in \mathcal{C}$ s.t. $\mathrm{err}_{\mathcal{D}}(c) = 0$. We call this realizable learning. Another way to describe the goal of realizable learning is finding a hypothesis h with small generalization disagreement between h and c over $\mathcal{D}_{\mathcal{X}}$, which is defined as $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(c, h) = \Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[c(x) \neq h(x)]$. The notion of disagreement plays a crucial role in our proofs.

2.2. Probabilistic Representation Dimension

The probabilistic representation dimension is a combinatorial parameter introduced by Beimel et al. (2019) to characterize the sample complexity of pure private realizable learner. Let $\mathcal{P}$ be a distribution on hypothesis classes. Define the size of $\mathcal{P}$ to be $\mathrm{size}(\mathcal{P}) = \max_{\mathcal{H}\in\mathrm{supp}(\mathcal{P})} \ln |\mathcal{H}|$.

Definition 2.3 (Probabilistic Representation Dimension)**.** A distribution $\mathcal{P}$ on hypothesis classes is said to be an (α, β)-probabilistic representation of a concept class $\mathcal{C}$ if for any $c \in \mathcal{C}$ and any distribution $\mathcal{D}_{\mathcal{X}}$ on $\mathcal{X}$, with probability $1 - \beta$ over $\mathcal{H} \sim \mathcal{P}$, there exists $h \in \mathcal{H}$ such that $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(c, h) \leq \alpha$. The (α, β)-probabilistic representation dimension of a concept class $\mathcal{C}$ is defined as

$$\mathrm{RepD}_{\alpha,\beta}(\mathcal{C}) = \min_{\mathcal{P} \text{ is an } (\alpha,\beta)\text{-probabilistic representation of } \mathcal{C}} \mathrm{size}(\mathcal{P}).$$

Moreover, the probabilistic representation dimension of $\mathcal{C}$ is defined by taking $\alpha = \beta = 1/4$:

$$\mathrm{RepD}(\mathcal{C}) = \mathrm{RepD}_{1/4,1/4}(\mathcal{C}).$$

The following lemma (Beimel et al., 2019) shows that a probabilistic representation with arbitrary α and β can be constructed from one with $\alpha = \beta = 1/4$.

Lemma 2.4 (Boosting Probabilistic Representation)**.** *For any concept class $\mathcal{C}$, we have* $\mathrm{RepD}_{\alpha,\beta}(\mathcal{C}) = O(\log(1/\alpha) \cdot (\mathrm{RepD}(\mathcal{C}) + \log\log\log(1/\alpha) + \log\log(1/\beta)))$ *for* $0 < \alpha, \beta < 1$.

2.3. Tools from Differential Privacy

We introduce some useful tools for achieving differential privacy. We say a function $f : \mathcal{Z}^* \to \mathbb{R}$ has sensitivity Δ if $|f(\boldsymbol{z}) - f(\boldsymbol{z}')| \leq \Delta$ for all neighboring datasets $\boldsymbol{z}$ and $\boldsymbol{z}'$. Let $\mathrm{Lap}(b)$ denote the Laplace distribution with mean 0 and scale b. The Laplace mechanism is an algorithm that outputs $f(\boldsymbol{z}) + r$ where $r \sim \mathrm{Lap}(\Delta/\varepsilon)$.

Lemma 2.5 (The Laplace Mechanism)**.** *The Laplace mechanism is ε-differentially private. Moreover, it holds that $|r| \leq \ln(1/\beta)\Delta/\varepsilon$ with probability at least $1 - \beta$, where $r \sim \mathrm{Lap}(\Delta/\varepsilon)$*

We next describe the exponential mechanism (McSherry & Talwar, 2007), which has been widely used in designing differentially private learning algorithms (e.g., (Kasiviswanathan et al., 2011; Beimel et al., 2019; Alon et al.,

2020)). Let $\mathcal{H}$ be a finite set and $q : \mathcal{Z}^* \times \mathcal{H} \to \mathbb{R}$ be a score function. We say q has sensitivity Δ if $\max_{h\in\mathcal{H}} |q(\boldsymbol{z}, h) - q(\boldsymbol{z}', h)| \le \Delta$ for any neighboring datasets $\boldsymbol{z}$ and $\boldsymbol{z}'$. The exponential mechanism outputs an $h \in \mathcal{H}$ with probability

$$\frac{\exp(-\varepsilon \cdot q(\boldsymbol{z}, h)/2\Delta)}{\sum_{f\in\mathcal{H}} \exp(-\varepsilon \cdot q(\boldsymbol{z}, f)/2\Delta)}.$$

Lemma 2.6 (The Exponential Mechanism)**.** *The exponential mechanism is ε-differentially private. Moreover, it outputs an h s.t.*

$$q(\boldsymbol{z}, h) \le \min_{f\in\mathcal{H}} q(\boldsymbol{z}, f) + \frac{2\Delta}{\varepsilon} \cdot \ln(|\mathcal{H}|/\beta)$$

with probability at least $1 - \beta$.

3. Item-Level Privacy

In this section, we give a nearly tight characterization of the sample complexity of agnostic learning under item-level privacy.

The work of Beimel et al. (2019) has shown that the sample complexity of pure private realizable learning is $\widetilde{\Theta}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha\varepsilon}\right)$. Combining with the well-known $\widetilde{\Omega}\left(\frac{\mathrm{VC}(\mathcal{C})}{\alpha^2}\right)$ lower bound on non-private agnostic learning (Vapnik & Chervonenkis, 1974; Simon, 1996), we get a lower bound of $\widetilde{\Omega}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{\alpha^2}\right)$ on the sample complexity of pure private agnostic learning.

For the upper bound, the first result was given by Kasiviswanathan et al. (2011). In their work, they proposed an algorithm using $\widetilde{O}\left(\frac{|\mathcal{C}|}{\alpha\varepsilon} + \frac{|\mathcal{C}|}{\alpha^2}\right)$ samples for finite concept class $\mathcal{C}$ and $\widetilde{O}\left(\frac{\mathrm{VC}(\mathcal{C})\log|\mathcal{X}|}{\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})\log|\mathcal{X}|}{\alpha^2}\right)$ samples for finite domain $\mathcal{X}$. Their analysis can be applied to the realizable learner of Beimel et al. (2019) by proving the generalization property for the hypothesis class $\mathcal{H}$ sampled from representation $\mathcal{P}$. This leads to a sample complexity of $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha\varepsilon} + \frac{\mathrm{RepD}(\mathcal{C})}{\alpha^2}\right)$, which exceeds the lower bound since the representation dimension can be much larger than the VC dimension (Feldman & Xiao, 2014).

Another existing approach is to use the realizable-to-agnostic transformation (Beimel et al., 2015; Alon et al., 2020), which states that every private realizable learner can be transformed into a private agnostic learner. Applying it to the algorithm of Beimel et al. (2019) gives a sample complexity of $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha} + \frac{\mathrm{VC}(\mathcal{C})}{\alpha^2}\right)$ for constant privacy parameter. For an arbitrary ε, one has to use the amplification-by-subsampling trick (Kasiviswanathan et al., 2011), resulting in a sample complexity of $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{\alpha^2\varepsilon}\right)$. Though here the second term depends on the VC dimension, it involves a $1/\varepsilon$ multiplicaive factor, which does not appear in the lower bound. Thus, this method is still suboptimal.

To obtain a tighter result, we propose a new algorithm whose sample complexity matches the lower bound. Before presenting our algorithm, we first introduce some definitions. Let $\boldsymbol{z} = (z_1, \ldots, z_n)$ be the input dataset, where $z_i = (x_i, y_i) \in \mathcal{X} \times \{0, 1\}$. We use $\boldsymbol{x} = (x_1, \ldots, x_n)$ to denote the corresponding unlabeled dataset. For a hypothesis h, define

$$\mathrm{err}_{\boldsymbol{z}}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[h(x_i) \ne y_i]$$

to be the empirical error of h on $\boldsymbol{z}$. For two hypotheses c and h, define their empirical disagreement on $\boldsymbol{x}$ as

$$\mathrm{dis}_{\boldsymbol{x}}(c, h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[c(x_i) \ne h(x_i)].$$

Our algorithm follows the steps of the learner proposed by Beimel et al. (2019): first samples an $\mathcal{H}$ from some probabilistic representation $\mathcal{P}$, then runs the exponential mechanism on $\mathcal{H}$. But unlike their algorithm, which uses $\mathrm{err}_{\boldsymbol{z}}(h)$ as the score function, we adopt the following:

$$q(\boldsymbol{z}, h) = \min_{c\in\mathcal{C}} \mathrm{err}_{\boldsymbol{z}}(c) + \mathrm{dis}_{\boldsymbol{x}}(c, h).$$

The sensitivity of the above score function is $2/n$ because each term may change by at least $1/n$ when moving to an adjacent dataset. Therefore, we can still apply the exponential mechanism to ensure privacy. The benefit of adopting such a score function is that it reduces the number of samples needed so that every hypothesis in $\mathcal{H}$ with a low score must have a small generalization error. Proving this for the algorithm of Beimel et al. (2019) requires agnostic generalization for all $h \in \mathcal{H}$, resulting in a sample complexity of $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha^2}\right)$. When applying our score function, we instead only need agnostic generalization for all $c \in \mathcal{C}$ and realizable generalization for the disagreement between all $c \in \mathcal{C}$ and $h \in \mathcal{H}$. The former can be ensured by the following agnostic generalization result (Talagrand, 1994; Anthony & Bartlett, 1999), which requires $\widetilde{O}\left(\frac{\mathrm{VC}(\mathcal{C})}{\alpha^2}\right)$ samples only.

Lemma 3.1 (VC Agnostic Generalization Bound)**.** *Let $\mathcal{C}$ be a concept class over $\mathcal{X}$, $\mathcal{D}$ be a distribution over $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$, and*

$$n \ge \frac{576}{\alpha^2}\left(4\mathrm{VC}(\mathcal{C}) + \ln(8/\beta)\right).$$

Suppose $\boldsymbol{z} \in \mathcal{Z}^n$ is a dataset with each z_i drawn i.i.d from $\mathcal{D}$, then

$$\Pr\left[\exists c \in \mathcal{C}\ s.t.\ |\mathrm{err}_{\mathcal{D}}(c) - \mathrm{err}_{\boldsymbol{z}}(c)| > \alpha\right] \le \beta.$$

For the latter one, we can apply the following realizable generalization bound (Vapnik & Chervonenkis, 1971; Blumer et al., 1989) to $\mathcal{C} \cup \mathcal{H}$.

Lemma 3.2 (VC Realizable Generalization Bound)**.** *Let $\mathcal{C}$ be a concept class and $\mathcal{D}_{\mathcal{X}}$ be a distribution over domain $\mathcal{X}$. Let*

$$n \geq \frac{96}{\alpha}\left(2\mathrm{VC}(\mathcal{C})\ln(384/\alpha) + \ln(4/\beta)\right).$$

Suppose $\boldsymbol{x} \in \mathcal{X}^n$ is an unlabeled dataset with each x_i drawn i.i.d. from $\mathcal{D}_{\mathcal{X}}$, then

$$\Pr\left[\begin{array}{c}\exists c, h \in \mathcal{C}\ s.t.\ \mathrm{err}_{\mathcal{D}_{\mathcal{X}}}(c,h) > \alpha \\ \textit{and}\ \mathrm{err}_{\boldsymbol{x}}(c,h) \leq \alpha/2\end{array}\right] \leq \beta.$$

Note that realizable generalization saves a $1/\alpha$ factor compared to agnostic generalization. This saving is crucial for our tighter upper bound. To identify the VC dimension of $\mathcal{C} \cup \mathcal{H}$, we need the following bound on the VC dimension of the union.

Lemma 3.3 ((Shalev-Shwartz & Ben-David, 2014))**.** *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two hypothesis class. Then $\mathrm{VC}(\mathcal{H}_1 \cup \mathcal{H}_2) = O(\mathrm{VC}(\mathcal{H}_1) + \mathrm{VC}(\mathcal{H}_2))$.*

This lemma suggests that the VC dimension of $\mathcal{C} \cup \mathcal{H}$ is $\widetilde{O}(\mathrm{RepD}(\mathcal{C}))$. Thus, the realizable generalization property requires $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha}\right)$ samples to hold. Combining with the $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha\varepsilon}\right)$ cost incurred by the exponential mechanism, we achieve the desired upper bound.

Theorem 3.4. *Let $\mathcal{C}$ be a concept class. In the item-level setting, $\widetilde{\Theta}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{\alpha^2}\right)$ samples are necessary and sufficient to (α, β)-agnostically learn $\mathcal{C}$ with ε-differential privacy.*

4. User-Level Privacy

In this section, we study the user complexity of pure private agnostic learning under user-level privacy. The work of Ghazi et al. (2023) gives an upper bound of $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\varepsilon} + \frac{\mathrm{RepD}(\mathcal{C})}{\sqrt{m}\alpha\varepsilon} + \frac{\mathrm{RepD}(\mathcal{C})}{m\alpha^2}\right)$ and a lower bound of $\widetilde{\Omega}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{\sqrt{m}\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{m\alpha^2}\right)$. We present an algorithm that improves the last term in their upper bound to $\widetilde{O}\left(\frac{\mathrm{VC}(\mathcal{C})}{m\alpha^2}\right)$, matching that in their lower bound.

Our algorithm works in the same way as the item-level one (Section 3), but with the empirical error and disagreement replaced by their user-level analogies. For a concept c and a dataset $\boldsymbol{z} \in \mathcal{Z}^{nm}$, define the user-level empirical error

$$\mathrm{err}^t_{\boldsymbol{z}}(c) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[m \cdot \mathrm{err}_{\boldsymbol{z}_i}(c) > t]$$

for some $t \in \{0, \ldots, m\}$. That is, $\mathrm{err}^t_{\boldsymbol{z}}(c)$ is the fraction of users on whose examples c makes more than t mistakes. Also, we can define the user-level generalization error of c on distribution $\mathcal{D}^m$:

$$\mathrm{err}^t_{\mathcal{D}^m}(c) = \Pr_{\boldsymbol{z}_0 \sim \mathcal{D}^m}[m \cdot \mathrm{err}_{\boldsymbol{z}_0}(c) > t].$$

Similarly, let $\boldsymbol{x}_i$ be the corresponding unlabeled dataset held by user i. For two hypotheses c and h, define their user-level empirical disagreement on unlabeled dataset $\boldsymbol{x} \in \mathcal{X}^{nm}$ to be

$$\mathrm{dis}^s_{\boldsymbol{x}}(c,h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[m \cdot \mathrm{dis}_{\boldsymbol{x}_i}(c,h) > s]$$

and their user-level generalization disagreement on distribution $\mathcal{D}^m_{\mathcal{X}}$ to be

$$\mathrm{dis}^s_{\mathcal{D}^m_{\mathcal{X}}}(c,h) = \Pr_{\boldsymbol{x}_0 \sim \mathcal{D}^m_{\mathcal{X}}}[m \cdot \mathrm{dis}_{\boldsymbol{x}_0}(c,h) > s].$$

The advantage of adopting user-level error is that it amplifies the additive error by a multiplicaive factor of $\sqrt{m}$. This is due to the following property of binomial distribution (Adell & Jodrá, 2006; Liu et al., 2020).

Lemma 4.1. *Given $m \in \mathbb{N}$ and $p, q \in [0,1]$. Let*

$$K = \min\left(m|p-q|, \frac{\sqrt{m}|p-q|}{\sqrt{p(1-p)}}, 1\right).$$

Then we have

$$\frac{1}{700}K \leq d_{TV}(\mathrm{Bin}(m,p), \mathrm{Bin}(m,q)) \leq K,$$

where $\mathrm{Bin}(m,p)$ is the binomial distribution with m trials and succeed probability p, and d_{TV} is the total variation distance. Moreover, there exists an ℓ such that

$$|\Pr[\mathrm{Bin}(m,p) > \ell] - \Pr[\mathrm{Bin}(m,q) > \ell]| \geq \frac{1}{700}K.$$

Let η be the minimum achievable item-level error. Then the minimum achievable user-level error can be represented by $\psi = \Pr[\mathrm{Bin}(m,\eta) > t]$. Consider a hypothesis h whose item-level error is greater than $\eta + \alpha$. Lemma 4.1 suggests that the user-level error of h is at least $\psi + \Omega(\sqrt{m}\alpha)$ for some appropriate t. Thus, if we wish the error of our output hypothesis to be at most $\eta + \alpha$, it is sufficient to find a hypothesis whose user-level error is less than $\psi + O(\sqrt{m}\alpha)$.

The above discussion focuses on user-level generalization error, i.e., user-level error on the underlying distribution $\mathcal{D}$. Since our algorithm only has access to the dataset, we have to show generalization properties for user-level error and disagreement as in the item-level setting. We will make use of the following relative uniform convergence lemma (Anthony & Bartlett, 1999).

Lemma 4.2. *Let $\mathcal{C}$ be a concept class over $\mathcal{X}$ and $\mathcal{D}$ be a distribution over $\mathcal{Z} = \mathcal{X} \times \{0,1\}$. Suppose $\boldsymbol{z} \in \mathcal{Z}^n$ is a dataset with each z_i drawn i.i.d. from $\mathcal{D}$, then for $\gamma \in (0,1)$ and $\xi > 0$, we have*

$$\Pr\left[\exists c \in \mathcal{C}\ s.t.\ \mathrm{err}_{\mathcal{D}}(c) > (1+\gamma)\mathrm{err}_{\boldsymbol{z}}(c) + \xi\right] \leq 4\Pi_{\mathcal{C}}(2n)\exp\left(\frac{-\gamma\xi n}{4(\gamma+1)}\right).$$

The $\Pi_{\mathcal{C}}(2n)$ term in the above lemma is the growth function, which represents the maximum number of labelings of $2n$ samples by some concept in $\mathcal{C}$ (see Appendix A for the exact definition). Note that this lemma naturally extends to the user-level setting: for agnostic generalization, one could pack each user's samples as a whole, which can be regarded as a single data drawn from the distribution $\mathcal{D}^m$ over domain $\mathcal{Z}^m$. Thus, Lemma 4.2 holds with the $\Pi_{\mathcal{C}}(2n)$ term increases to $\Pi_{\mathcal{C}}(2nm)$ since there are nm samples in total. Sauer's Lemma (Sauer, 1972) suggests that this only raises the number of users required by roughly $\mathrm{VC}(\mathcal{C})\ln(m)/\alpha^2$. Standard argument (see, e.g., the book of Anthony & Bartlett (1999)) gives the following generalization bound on user-level error.

Lemma 4.3 (User-Level Agnostic Generalization)**.** *Let $\mathcal{C}$ be a concept class over $\mathcal{X}$, $\mathcal{D}$ be a distribution over $\mathcal{Z} = \mathcal{X} \times \{0,1\}$, and m be the number of samples per user. Let*

$$n \geq \frac{64}{\alpha^2}\left(\mathrm{VC}(\mathcal{C})\ln(128m/\alpha^2) + \ln(8/\beta)\right).$$

Suppose $\boldsymbol{z} \in \mathcal{Z}^{nm}$ is a dataset with each $z_{i,j}$ drawn i.i.d. from $\mathcal{D}$, then

$$\Pr\left[\exists c \in \mathcal{C}\ s.t.\ |\mathrm{err}^t_{\mathcal{D}^m}(c) - \mathrm{err}^t_{\boldsymbol{z}}(c)| > \alpha\right] \leq \beta.$$

For the realizable case, we can similarly prove the following bound by setting γ and ξ appropriately. Like the item-level setting, the bound is proportional to $1/\alpha$, as opposed to $1/\alpha^2$ in the agnostic case.

Lemma 4.4 (User-Level Realizable Generalization)**.** *Let $\mathcal{C}$ be a concept class, $\mathcal{D}_{\mathcal{X}}$ be a distribution over domain $\mathcal{X}$, and m be the number of samples per user. Let*

$$n \geq \frac{96}{\alpha}\left(2\mathrm{VC}(\mathcal{C})\ln(384m/\alpha) + \ln(4/\beta)\right).$$

Suppose $\boldsymbol{x} \in \mathcal{X}^{nm}$ is an unlabeled dataset with each $x_{i,j}$ drawn i.i.d. from $\mathcal{D}_{\mathcal{X}}$, then

$$\Pr\left[\begin{array}{c}\exists c, h \in \mathcal{C}\ s.t.\ \mathrm{dis}^s_{\mathcal{D}^m_{\mathcal{X}}}(c,h) > \alpha \\ \textit{and}\ \mathrm{dis}^s_{\boldsymbol{x}}(c,h) \leq \alpha/2\end{array}\right] \leq \beta.$$

Now, we still have one missing piece: what are the values of t and s? It is easy to choose the value of s since we only have to separate $\mathrm{Bin}(m,0)$ and $\mathrm{Bin}(m, C\alpha)$ for predetermined constant C. Thus, s is a predetermined constant (indeed, $s = 0$). However, for t we have to select a value that separates $\mathrm{Bin}(m,\eta)$ and $\mathrm{Bin}(m, \eta + O(\alpha))$. Therefore, the choice of t depends on the minimum achievable error η, which is unknown since $\mathcal{D}$ is unknown.

We resolve this issue by using an approximate value of η to decide the value of t. It can be shown that an estimation within error $O(\alpha)$ suffices. We design an algorithm that estimates η privately. Our idea is based on binary search. Suppose in each iteration, we can compare η to the midpoint of the current interval. Then we can obtain an estimation of η within error $O(\alpha)$ after $O(\log(1/\alpha))$ iterations.

In each iteration, we have to compare η to some guessed value $\widetilde{\eta}$ (the midpoint). Due to agnostic generalization, the comparison can be done by calculating the minimum empirical user-level error with some $\widetilde{t}$, where the parameter $\widetilde{t}$ can be derived from $\widetilde{\eta}$. To ensure privacy, we have to add Laplace noise to the error.

As illustrated before, the user-level error provides a $\sqrt{m}$ amplification. Thus, the binary search yields a reduced user complexity, as stated in the following lemma.

Lemma 4.5. *Let $\mathcal{C}$ be a concept class and $\mathcal{D}$ be a distribution over $\mathcal{Z}$. Suppose each user holds m samples. Let*

$$n \geq \widetilde{O}\left(\frac{1}{\varepsilon} + \mathrm{VC}(\mathcal{C}) + \frac{\mathrm{VC}(\mathcal{C})}{m\alpha^2} + \frac{1}{\sqrt{m}\alpha\varepsilon}\right)$$

and $\boldsymbol{z} \in \mathcal{Z}^{nm}$ be a dataset with each $\boldsymbol{z}_{i,j}$ drawn i.i.d from $\mathcal{D}$. Then there exists an ε-differentially private algorithm that returns some $\hat{\eta}$ s.t. $|\hat{\eta} - \eta| \leq \alpha$ with probability $1-\beta$, where $\eta = \inf_{c\in\mathcal{C}} \mathrm{err}_{\mathcal{D}}(c)$.

Now, we are ready to prove our main result. We first use Lemma 4.5 to compute $\hat{\eta}$, which is an estimation of η within error $O(\alpha)$. Then, we construct an appropriate parameter t from $\hat{\eta}$. Finally, we run the exponential mechanism as in the item-level (Theorem 4.6), but with a score function constructed from the user-level empirical error and disagreement. We state the result in the following theorem.

Theorem 4.6. *Let $\mathcal{C}$ be a concept class and m be the number of samples per user. There exists an ε-differentially private algorithm that (α,β)-agnostically learns $\mathcal{C}$ using $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\varepsilon} + \frac{\mathrm{RepD}(\mathcal{C})}{\sqrt{m}\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{m\alpha^2}\right)$ users.*

5. Learning Thresholds with User-level Privacy

In this section, we focus on learning thresholds with user-level DP and give a nearly tight bound on the user complexity. In the problem of learning thresholds, the domain is $\mathcal{X} = \{1,\ldots,|\mathcal{X}|\}$, and the concept class $\mathcal{C}$ is the collection of all thresholds on $\mathcal{X}$. More formally, a threshold function

f_u is specified by an element $u \in \{0\} \cup \mathcal{X}$ so that $f_u(x) = 1$ if $x > u$ and $f_u(x) = 0$ if $x \leq u$ for $x \in \mathcal{X}$. The concept class $\mathcal{C} = \{f_u \mid u \in \{0\} \cup \mathcal{X}\}$.

The class of thresholds on $\mathcal{X}$ has VC dimension 1 (see e.g., (Shalev-Shwartz & Ben-David, 2014)) and representation dimension $\Theta(\log |X|)$ (Feldman & Xiao, 2014). Thus, directly applying Theorem 4.6 gives a user complexity of $\widetilde{O}\left(\frac{\log |X|}{\varepsilon} + \frac{\log |X|}{\sqrt{m}\alpha\varepsilon} + \frac{1}{m\alpha^2}\right)$. As discussed in Section 4, the second term is still larger than the lower bound proved by Ghazi et al. (2023). We will show how to reduce this term to $\widetilde{O}\left(\frac{1}{\sqrt{m}\alpha\varepsilon}\right)$.

Our method is based on binary search. We start with the entire range $[0, |\mathcal{X}|]$. In each iteration, we pick some point to split the current interval into two parts. We then examine the minimum achievable error of the thresholds in each part and choose the smaller one to continue. This can be done by injecting Laplace noise to the minimum empirical user-level error, which saves the user complexity by a $\sqrt{m}$ factor as demonstrated in Section 4.

Suppose we naïvely select the middle point of the current interval in each iteration, then the binary search has to run for $O(\log |X|)$ rounds. This incurs a user complexity of $\widetilde{O}\left(\frac{\log |X|}{\sqrt{m}\alpha\varepsilon}\right)$ by the basic composition, which is not satisfactory. Thus, we need to reduce the number of iterations.

Our key insight is that, instead of simply picking the middle point, we choose a point in the middle of the distribution (i.e., the median) in each iteration. Though we cannot hope to find the exact median since $\mathcal{D}$ is unknown, a constant approximation is sufficient. Suppose for a given interval $[l, r]$, we can choose some mid such that

$$\max\left(\Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \in [l, mid-1]], \Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \in [mid+1, r]]\right) \leq \theta \cdot \Pr_{x\in\mathcal{D}_\mathcal{X}}[x \in [l, r]]$$

for some constant $\theta < 1$. Then we only have to run the binary search for $O(\log(1/\alpha))$ rounds. After that, the disagreement between any two thresholds in the final interval is $O(\alpha)$. Thus, we can output any one in the final interval.

Note that mid should not be included in any part. Otherwise, the above property may be impossible to hold since the distribution can concentrate at a single point. However, we cannot simply ignore f_{mid} since it may be the only threshold that achieves our target error. Thus, in each iteration, we also calculate the error of f_{mid}. Our algorithm will terminate and output f_{mid} if it has the smallest error among all thresholds in the current interval. We describe the steps in Algorithm 1.

Our method for finding the meadian requires an observation from Lemma 4.1 by letting $p = c_1\alpha$ and $q = c_2\alpha$ for some constant c_1 and c_2:

$$d_{TV}(\mathrm{Bin}(m, c_1\alpha), \mathrm{Bin}(m, c_2\alpha)) = \Theta(m\alpha)$$

when $m = O(1/\alpha)$. Thus, it provides an amplification factor of m when p and q are small. Such an observation was also utilized by Liu et al. (2020) to derive upper bounds on the problem of learning discrete distributions and by Ghazi et al. (2023) to characterize the user complexity of user-level pure private realizable learning. We did not elaborate on this in Section 4 since the $\sqrt{m}$ amplification is sufficient there. However, for median selection, the $\sqrt{m}$ amplification only produces an undesirable user complexity of $\widetilde{O}\left(\frac{\log |\mathcal{X}|}{\sqrt{m}\alpha\varepsilon}\right)$.

The above observation implies that, for the probability density in $[l, r]$ to be less than α, we only need the probability that one user has more than ℓ points in $[l, r]$ to be $O(m\alpha)$. By the realizable generalization property, we can use the empirical counterpart as an alternative and apply the exponential mechanism to select an approximate median privately.

Algorithm 1 PrivateThreshold

Input: dataset $\boldsymbol{z} \in \mathcal{Z}^{nm}$, privacy parameter ε, number of iterations T, user-level error parameter t, algorithm PrivateMedian, decaying factor θ.
$l \leftarrow 0, r \leftarrow |\mathcal{X}|$
$\varepsilon' \leftarrow \varepsilon/(4T)$
for $k \leftarrow 1$ **to** T **do**
  if $l = r$ **then**
    break
  end if
  $mid \leftarrow \mathsf{PrivateMedian}(\boldsymbol{z}, \varepsilon', l, r, \theta^{k-1})$
  $v_{mid} \leftarrow \mathrm{err}^t_{\boldsymbol{z}}(f_{mid}) + \mathrm{Lap}(1/n\varepsilon')$
  $v_l \leftarrow \min_{u\in\{l,\ldots,mid-1\}} \mathrm{err}^t_{\boldsymbol{z}}(f_u) + \mathrm{Lap}(1/n\varepsilon')$
  $v_r \leftarrow \min_{u\in\{mid+1,\ldots,r\}} \mathrm{err}^t_{\boldsymbol{z}}(f_u) + \mathrm{Lap}(1/n\varepsilon')$
  % The min operator returns $+\infty$ if the range is empty.
  if $v_{mid} < \min(v_l, v_r)$ **then**
    return f_{mid}
  else if $v_l < v_r$ **then**
    $r \leftarrow mid - 1$
  else
    $l \leftarrow mid + 1$
  end if
end for
return f_l

Lemma 5.1. *Let $\boldsymbol{z} \in \mathcal{Z}^{nm}$ be a dataset, where each $\boldsymbol{z}_{i,j}$ is drawn i.i.d. from some distribution $\mathcal{D}$, and m be the number of samples per user. Suppose $\Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \in [l, r]] \leq \alpha$ and*

$$n \geq \widetilde{O}\left(\frac{\log |\mathcal{X}|}{\varepsilon} + \frac{\log |\mathcal{X}|}{m\alpha\varepsilon}\right).$$

Then there exists an ε-differentially private algorithm that

takes $\boldsymbol{z}, \varepsilon, l, r, \alpha$ *as input and output some* u_0 *such that*

$$\max\left(\Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \in [l, u_0 - 1]], \Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \in [u_0 + 1, r]]\right) \leq \frac{2}{3}\alpha$$

with probability at least $1 - \beta$.

Now, let us summarize the entire algorithm. We first use Lemma 4.5 to determine the value of t. Then we run Algorithm 1 with PrivateMedian being the algorithm in Lemma 5.1. Note that though the $\frac{\log|\mathcal{X}|}{m\alpha\varepsilon} = \frac{\mathrm{RepD}(\mathcal{C})}{m\alpha\varepsilon}$ term does not explicitly appear in the lower bound presented by Ghazi et al. (2023), they also show that this term is necessary even for pure private realizable learning. Thus, we obtain the following nearly tight user complexity bound.

Theorem 5.2. *Let* $\mathcal{C}$ *be the concept class of thresholds over* $\mathcal{X}$ *and* m *be the number of samples per user. Then*

$$\widetilde{\Theta}\left(\frac{\log|\mathcal{X}|}{\varepsilon} + \frac{\log|\mathcal{X}|}{m\alpha\varepsilon} + \frac{1}{\sqrt{m}\alpha\varepsilon} + \frac{1}{m\alpha^2}\right)$$

users are necessary and sufficient to (α, β)*-agnostically learn* $\mathcal{C}$ *with* ε*-differential privacy in the user-level setting.*

6. Conclusion

This work investigates private agnostic learning under item-level and user-level pure DP. In the item-level setting, we devise an algorithm that achieves the optimal sample complexity up to polylogarithmic factors. In the user-level setting, we propose a generic learner for arbitrary concept classes, with an enhanced user complexity than the best-known result in (Ghazi et al., 2023). For the specific task of learning thresholds, we develop a near-optimal upper bound and conjecture that such improvement can also be made in the generic case. We leave the problem of closing the gap for general concept classes as future work. Another interesting direction for future work is to investigate the scenario where different users may hold different amounts of data.

Acknowledgements

The research was supported in part by a RGC RIF grant under the contract R6021-20, RGC CRF grants under the contracts C7004-22G and C1029-22G, and RGC GRF grants under the contracts 16209120, 16200221, 16207922, and 16211123. We would like to thank the anonymous reviewers for their helpful comments.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abowd, J. M. The U.S. Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2867. Association for Computing Machinery, 2018.

Adell, J. A. and Jodrá, P. Exact Kolmogorov and total variation distances between some familiar discrete distributions. *Journal of Inequalities and Applications*, 2006: 1–8, 2006.

Alon, N., Livni, R., Malliaris, M., and Moran, S. Private PAC learning implies finite Littlestone dimension. In *Proceedings of the 51st Annual ACM Symposium on Theory of Computing*, pp. 852–860. Association for Computing Machinery, 2019.

Alon, N., Beimel, A., Moran, S., and Stemmer, U. Closure properties for private classification and online prediction. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125, pp. 119–152. PMLR, 2020.

Alon, N., Bun, M., Livni, R., Malliaris, M., and Moran, S. Private and online learnability are equivalent. *Journal of the ACM*, 69(4):1–34, 2022.

Anthony, M. and Bartlett, P. L. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Apple Differential Privacy Team. Learning with privacy at scale. *Apple Machine Learning Journal*, 2017.

Beimel, A., Brenner, H., Kasiviswanathan, S. P., and Nissim, K. Bounds on the sample complexity for private learning and private data release. *Machine learning*, 94:401–437, 2014.

Beimel, A., Nissim, K., and Stemmer, U. Learning privately with labeled and unlabeled examples. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 461–477. SIAM, 2015.

Beimel, A., Nissim, K., and Stemmer, U. Characterizing the sample complexity of pure private learners. *Journal of Machine Learning Research*, 20:1–33, 2019.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

Bun, M., Livni, R., and Moran, S. An equivalence between private classification and online prediction. In *Proceedings of the 61st Annual IEEE Symposium on Foundations of Computer Science*, pp. 389–402, 2020.

Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pp. 493–507, 1952.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *Proceedings of the 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503, Berlin, Heidelberg, 2006a. Springer.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, pp. 265–284, Berlin, Heidelberg, 2006b. Springer.

Feldman, V. and Xiao, D. Sample complexity bounds on differentially private learning via communication complexity. In *Proceedings of the 27th Conference on Learning Theory*, volume 35, pp. 1000–1019. PMLR, 2014.

Ghazi, B., Golowich, N., Kumar, R., and Manurangsi, P. Sample-efficient proper PAC learning with approximate differential privacy. In *Proceedings of the 53rd Annual ACM Symposium on Theory of Computing*, pp. 183–196. Association for Computing Machinery, 2021a.

Ghazi, B., Kumar, R., and Manurangsi, P. User-level differentially private learning via correlated sampling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20172–20184. Curran Associates, Inc., 2021b.

Ghazi, B., Kamath, P., Kumar, R., Manurangsi, P., Meka, R., and Zhang, C. User-level differential privacy with few examples per user. In *Advances in Neural Information Processing Systems*, 2023.

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

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Kearns, M. J., Schapire, R. E., and Sellie, L. M. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.

Liu, Y., Suresh, A. T., Yu, F. X. X., Kumar, S., and Riley, M. Learning discrete distributions: user vs item-level privacy. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20965–20976. Curran Associates, Inc., 2020.

McSherry, F. and Talwar, K. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, pp. 94–103. IEEE Computer Society, 2007.

Sauer, N. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

Simon, H. U. General bounds on the number of examples needed for learning probabilistic concepts. *Journal of Computer and System Sciences*, 52(2):239–254, 1996.

Talagrand, M. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 22:28–76, 1994.

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

Vapnik, V. and Chervonenkis, A. *Theory of pattern recognition*. Nauka, Moscow, 1974.

Vapnik, V. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2): 264–280, 1971.

A. Additional Preliminaries

A.1. The Vapnik-Chervonenkis Dimension

The Vapnik-Chervonenkis dimension (VC dimension) is a combinatorial measure of concept classes, which characterizes the sample complexity of (non-private) PAC learning and agnostic learning. Consider a concept class $\mathcal{C}$ over domain $\mathcal{X}$. Let $\boldsymbol{x} \in \mathcal{X}^n$ be an unlabeled dataset of size n. The set of all dichotomies on $\boldsymbol{x}$ that are realized by $\mathcal{C}$ is denoted by $\Pi_{\mathcal{C}}(\boldsymbol{x}) = \{(c(\boldsymbol{x}_1), \ldots, c(\boldsymbol{x}_n)) \mid c \in \mathcal{C}\}$. The growth function of $\mathcal{C}$ is defined as

$$\Pi_{\mathcal{C}}(n) = \max_{x \in \mathcal{X}^n} |\Pi_{\mathcal{C}}(\boldsymbol{x})|.$$

Definition A.1. The VC dimension of a concept class $\mathcal{C}$ is defined as the largest number d such that $\Pi_{\mathcal{C}}(d) = 2^d$ (or infinity, if the maximum does not exist).

Though the number of binary vectors of length n is 2^n, Sauer's Lemma states that the growth function is polynomially bounded.

Lemma A.2 (Sauer's Lemma)**.** *Let $\mathcal{C}$ be a concept class with VC dimension d. Then for $n \geq d$ we have $\Pi_{\mathcal{C}}(n) \leq \left(\frac{en}{d}\right)^d$.*

The following technical inequality from (Anthony & Bartlett, 1999) is useful in deriving sample/user complexity bounds:

$$\ln a \leq ab + \ln(1/b) - 1 \text{ for all } a, b > 0. \tag{1}$$

Applying Sauer's lemma usually gives an inequality of the form $A + B \ln n \leq n$. To find an n that satisfies this inequality, let $a = n$ and $b = 1/2B$ in inequality (1), we get

$$A + B \ln n \leq A + B(n/2B + \ln(2B) - 1) = A + n/2 + B \ln(2B/e).$$

Thus, it suffices to find an n s.t. $A + B \ln(2B/e) \leq n/2$ or, equivalently, $2A + 2B \ln(2B/e) \leq n$.

A.2. Concentration Bounds

Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with probability p of being 1. The expected value of their sum is $\mathbb{E}[X_1 + \ldots X_n] = np$. The following inequalities bound the probability that the summation deviates from the its expected value:

$$\Pr\left[\sum_{i=1}^{n} X_i \geq (1+\delta)np\right] \leq \exp(-\delta^2 np/(2+\delta)) \text{ for } \delta > 0,$$
$$\Pr\left[\sum_{i=1}^{n} X_i \leq (1-\delta)np\right] \leq \exp(-\delta^2 np/2) \text{ for } 0 < \delta < 1,$$
$$\Pr\left[\left|\sum_{i=1}^{n} X_i - np\right| \geq \delta\right] \leq 2\exp(-2\delta^2/n) \text{ for } \delta \geq 0.$$

The first two inequalities are knowns as the Chernoff bounds (Chernoff, 1952) and the last one is kowns as Hoeffding's inequality (Hoeffding, 1963).

B. Proof of Theorem 3.4

Proof. The lower bound can be shown by combining the $\widetilde{\Omega}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha\varepsilon}\right)$ lower bound on private realizable learning (Beimel et al., 2019) and the $\widetilde{\Omega}\left(\frac{\mathrm{VC}(\mathcal{C})}{\alpha^2}\right)$ lower bound on agnostic learning (Vapnik & Chervonenkis, 1974; Simon, 1996).

Now, let us focus on the upper bound. Our algorithm A works as follows: first samples $\mathcal{H} \sim \mathcal{P}$, where $\mathcal{P}$ is a $(\alpha/18, \beta/5)$-probabilistic representation of $\mathcal{C}$ with $\mathrm{size}(\mathcal{P}) = \mathrm{RepD}_{\alpha/18,\beta/5}(\mathcal{C})$, then runs the exponential mechanism on $\mathcal{H}$ with privacy parameter ε, sensitivity parameter $\Delta = 2/n$, and score function $q(\boldsymbol{z}, h)$ defined as

$$q(\boldsymbol{z}, h) = \min_{c \in \mathcal{C}} \mathrm{err}_{\boldsymbol{z}}(c) + \mathrm{dis}_{\boldsymbol{x}}(c, h).$$

Note that each term in the minimization will change by at most $1/n$ when moving to a neighboring dataset. This implies that the sensitivity of q is at most $2/n$. Thus, the privacy guarantee of the exponential mechanism ensures that A is ε-differentially private.

Now it remains to show that A is an (α, β)-agnostic learner. Let $\eta = \inf_{c \in \mathcal{C}} \mathrm{err}_{\mathcal{D}}(c)$, then there exists some $c' \in \mathcal{C}$ s.t. $\mathrm{err}_{\mathcal{D}}(c') \leq \eta + \alpha/18$. By Lemma 3.1, for $n \geq \widetilde{O}\left(\frac{\mathrm{VC}(\mathcal{C})}{\alpha^2}\right)$, we have $|\mathrm{err}_{\mathcal{D}}(c) - \mathrm{err}_{\boldsymbol{z}}(c)| \leq \alpha/18$ for all $c \in \mathcal{C}$ with probability $1 - \beta/5$. Thus,

$$\mathrm{err}_{\boldsymbol{z}}(c') \leq \mathrm{err}_{\mathcal{D}}(c') + \alpha/18 \leq \eta + \alpha/9.$$

Since $\mathcal{P}$ is an $(\alpha/18, \beta/5)$-probabilistic representation, we know that with probability $1 - \beta/5$, there exists some h' s.t. $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(c', h') \leq \alpha/18$. Then by the Chernoff bound, for $n \geq \widetilde{O}(1/\alpha)$, with probability $1 - \beta/5$, we have

$$\mathrm{dis}_{\boldsymbol{x}}(c', h') \leq 2\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(c', h') \leq \alpha/9.$$

Therefore, by Lemma 2.6, for $n \geq \widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\varepsilon\alpha}\right)$, the exponential mechanism chooses some h_0 s.t.

$$\begin{aligned} q(\boldsymbol{z}, h_0) &\leq \min_{f \in \mathcal{H}} q(\boldsymbol{z}, f) + \alpha/18 \\ &\leq q(\boldsymbol{z}, h') + \alpha/18 \\ &\leq \mathrm{err}_{\boldsymbol{z}}(c') + \mathrm{dis}_{\boldsymbol{x}}(c', h') + \alpha/18 \\ &\leq \eta + 5\alpha/18 \end{aligned}$$

with probability at least $1 - \beta/5$.

Suppose $q(\boldsymbol{z}, h_0) = \mathrm{err}_{\boldsymbol{z}}(c_0) + \mathrm{dis}_{\boldsymbol{x}}(c_0, h_0)$, we have

$$\mathrm{err}_{\boldsymbol{z}}(c_0) \leq q(\boldsymbol{z}, h_0) \leq \eta + 5\alpha/18.$$

Then by agnostic generalization, we can bound the generalization error of c_0:

$$\mathrm{err}_{\mathcal{D}}(c_0) \leq \mathrm{err}_{\boldsymbol{z}}(c_0) + \alpha/18 \leq \eta + \alpha/3.$$

Since every concept in $\mathcal{C}$ has error at least η on $\mathcal{D}$, agnostic generalization also implies that

$$\mathrm{err}_{\boldsymbol{z}}(c_0) \geq \mathrm{err}_{\mathcal{D}}(c_0) - \alpha/18 \geq \eta - \alpha/18.$$

Thus, we can bound the empirical disagreement between c_0 and h_0:

$$\begin{aligned} \mathrm{dis}_{\boldsymbol{z}}(c_0, h_0) &= q(\boldsymbol{z}, h_0) - \mathrm{err}_{\boldsymbol{z}}(c_0) \\ &\leq \eta + 5\alpha/18 - (\eta - \alpha/18) \\ &= \alpha/3. \end{aligned}$$

We now show the realizable generalization property. By Lemma 2.4 and 3.3, we have

$$\begin{aligned} \mathrm{VC}(\mathcal{C} \cup \mathcal{H}) &= O(\mathrm{VC}(\mathcal{C}) + \mathrm{VC}(\mathcal{H})) \\ &= O(\mathrm{VC}(\mathcal{C}) + \log_2 |\mathcal{H}|) \\ &= O(\mathrm{VC}(\mathcal{C}) + \mathrm{RepD}_{\alpha/18, \beta/5}(\mathcal{C})) \\ &= \widetilde{O}(\mathrm{VC}(\mathcal{C}) + \mathrm{RepD}(\mathcal{C})) \\ &= \widetilde{O}(\mathrm{RepD}(\mathcal{C})). \end{aligned}$$

Then applying Lemma 3.2, for $n \geq \widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha}\right)$, with probability $1 - \beta/5$ it holds that $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(c, h) \leq 2\alpha/3$ for all $(c, h) \in \mathcal{C} \times \mathcal{H}$ s.t. $\mathrm{dis}_{\boldsymbol{x}}(c, h) \leq \alpha/3$. Thus, $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(c_0, h_0) \leq 2\alpha/3$.

By the union bound, we can conclude that

$$\begin{aligned}\mathrm{err}_{\mathcal{D}}(h_0) &\le \mathrm{err}_{\mathcal{D}}(c_0) + \mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(c_0, h_0)\\ &\le \eta + \alpha/3 + 2\alpha/3\\ &= \eta + \alpha\end{aligned}$$

with probability $1-\beta$ for some $n = \widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{\alpha^2}\right)$.

□

C. Proofs from Section 4

C.1. Proof of Lemma 4.3

Proof. For a concept $c \in \mathcal{C}$, define a concept $\varphi_c : \mathcal{Z}^m \to \{0,1\}$ such that $\varphi_c(\boldsymbol{z}_0) = \mathbf{1}[m \cdot \mathrm{err}_{\boldsymbol{z}_0}(c) > t]$ for $\boldsymbol{z}_0 \in \mathcal{Z}^m$. Let $\mathcal{C}_\varphi = \{\varphi_c \mid c \in \mathcal{C}\}$, $\mathcal{D}'$ be a distribution on $\mathcal{Z}^m \times \{0,1\}$ s.t.

$$\Pr_{\mathcal{D}'}[(\boldsymbol{z}_0, y_0)] = \begin{cases}\Pr_{\mathcal{D}^m}[\boldsymbol{z}_0], & y_0 = 0,\\ 0, & y_0 = 1,\end{cases}$$

and $\boldsymbol{z}' = (z'_1, \ldots, z'_n) \in (\mathcal{Z}^m \times \{0,1\})^n$ be a dataset with $z'_i = (\boldsymbol{z}_i, 0)$, where $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ is a dataset sampled from $\mathcal{D}^{nm}$. We have

$$\mathrm{err}_{\mathcal{D}'}(\varphi_c) = \Pr_{(\boldsymbol{z}_0, y_0)\sim\mathcal{D}'}[\varphi_c(\boldsymbol{z}_0) \ne y_0] = \Pr_{\boldsymbol{z}_0\sim\mathcal{D}^m}[m \cdot \mathrm{err}_{\boldsymbol{z}_0}(c) > t] = \mathrm{err}^t_{\mathcal{D}^m}(c)$$

and

$$\mathrm{err}_{\boldsymbol{z}'}(\varphi_c) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}[\varphi_c(\boldsymbol{z}_i) \ne 0] = \frac{1}{n}\sum_{i=1}^n \mathbf{1}[m \cdot \mathrm{err}_{\boldsymbol{z}_i}(c) > t] = \mathrm{err}^t_{\boldsymbol{z}}(c).$$

Moreover, each z'_i can be seen as i.i.d. drwan from D'. Applying Lemma 4.2 with concept class $\mathcal{C}_\varphi$, distribution $\mathcal{D}'$, dataset $\boldsymbol{z}'$, and $\gamma = \xi = \alpha/2$, we have

$$\begin{aligned}&\Pr[\exists c \in \mathcal{C}\ s.t.\ \mathrm{err}^t_{\mathcal{D}^m}(c) > \mathrm{err}^t_{\boldsymbol{z}}(c) + \alpha]\\ =&\Pr[\exists \varphi_c \in \mathcal{C}_\varphi\ s.t.\ \mathrm{err}_{\mathcal{D}'}(\varphi_c) > \mathrm{err}_{\boldsymbol{z}'}(\varphi_c) + \alpha]\\ \le&\Pr[\exists \varphi_c \in \mathcal{C}_\varphi\ s.t.\ \mathrm{err}_{\mathcal{D}'}(\varphi_c) > (1+\alpha/2)\mathrm{err}_{\boldsymbol{z}'}(\varphi_c) + \alpha/2]\\ \le&4\Pi_{\mathcal{C}_\varphi}(2n)\exp\left(\frac{-\alpha^2 n/4}{4(1+\alpha/2)}\right)\\ \le&4\Pi_{\mathcal{C}}(2nm)\exp\left(-\alpha^2 n/32\right),\end{aligned}$$

where the last inequality is due to the fact that

$$\begin{aligned}\Pi_{\mathcal{C}_\varphi}(2n) &= \max_{\boldsymbol{z}\in\mathcal{Z}^{2nm}} |\Pi_{\mathcal{C}_\varphi}(\boldsymbol{z})|\\ &= \max_{\boldsymbol{z}\in\mathcal{Z}^{2nm}} \left|\left\{\left(\mathbf{1}\left[\sum_{i=1}^m \mathbf{1}\left[c(x_{1,i}) \ne y_{1,i}\right] > t\right], \ldots, \mathbf{1}\left[\sum_{i=1}^m \mathbf{1}\left[c(x_{2n,i}) \ne y_{2n,i}\right] > t\right]\right) \,\middle|\, c \in \mathcal{C}\right\}\right|\\ &\le \max_{\boldsymbol{z}\in\mathcal{Z}^{2nm}} |\{(\mathbf{1}[c(x_{1,1}) \ne y_{1,1}], \ldots, \mathbf{1}[c(x_{2n,m}) \ne y_{2n,m}]) \mid c \in \mathcal{C}\}|\\ &= \max_{\boldsymbol{x}\in\mathcal{X}^{2nm}} |\Pi_{\mathcal{C}}(\boldsymbol{x})|\\ &= \Pi_{\mathcal{C}}(2nm).\end{aligned}$$

By Sauer's Lemma, we have $\Pi_{\mathcal{C}}(2nm) \le \left(\frac{2enm}{d}\right)^d$ if $2nm \ge d$, where $d = \mathrm{VC}(\mathcal{C})$. We want the above quantity to be at most $\beta/2$. Thus, it suffices to show

$$n \ge \frac{32}{\alpha^2}\left(d\ln n + d\ln(2em/d) + \ln(8/\beta)\right).$$

By inequality (1), we have

$$\frac{32d}{\alpha^2}\ln n \le \frac{n}{2} + \frac{32d}{\alpha^2}\ln\left(\frac{64d}{e\alpha^2}\right).$$

Therefore,

$$n \ge \frac{64}{\alpha^2}\left(d\ln(128m/\alpha^2) + \ln(8/\beta)\right) \tag{2}$$

suffices. We still need to prove the other side. For a concept c, define $c^-(x) = 1 - c(x)$. Let $\mathcal{C}^- = \{c^- \mid c \in \mathcal{C}\}$ and $t^- = m - t - 1$. It is easy to verify that $\mathrm{VC}(\mathcal{C}^-) = \mathrm{VC}(\mathcal{C})$, $\mathrm{err}^{t^-}_{\boldsymbol{z}}(c^-) = 1 - \mathrm{err}^t_{\boldsymbol{z}}(c)$, and $\mathrm{err}^{t^-}_{\mathcal{D}^m}(c^-) = 1 - \mathrm{err}^t_{\mathcal{D}^m}(c)$. Applying the above conclusion to $\mathcal{C}^-$ and t^- gives

$$\begin{aligned}
&\Pr[\exists c \in \mathcal{C}\ s.t.\ \mathrm{err}^t_{\mathcal{D}^m}(c) + \alpha < \mathrm{err}^t_{\boldsymbol{z}}(c)]\\
=&\Pr[\exists c^- \in \mathcal{C}^-\ s.t.\ 1 - \mathrm{err}^{t^-}_{\mathcal{D}^m}(c^-) + \alpha < 1 - \mathrm{err}^{t^-}_{\boldsymbol{z}}(c^-)]\\
=&\Pr[\exists c^- \in \mathcal{C}^-\ s.t.\ \mathrm{err}^{t^-}_{\mathcal{D}^m}(c^-) > \mathrm{err}^{t^-}_{\boldsymbol{z}}(c^-) + \alpha]\\
\le&\beta/2
\end{aligned}$$

if n satisfies inequality (2). Applying the union bound completes the proof. □

C.2. Proof of Lemma 4.4

Proof. For any $c, h \in \mathcal{C}$, define a concept $\varphi_{c,h} : \mathcal{X}^m \to \{0,1\}$ such that $\varphi_{c,h}(\boldsymbol{x}_0) = \mathbf{1}[m \cdot \mathrm{dis}_{\boldsymbol{x}_0}(c,h) > s]$ for $\boldsymbol{x}_0 \in \mathcal{X}^m$. Let $\mathcal{C}_\varphi = \{\varphi_{c,h} \mid c, h \in \mathcal{C}\}$, $\mathcal{D}'$ be a distribution on $\mathcal{X}^m \times \{0,1\}$ s.t.

$$\Pr_{\mathcal{D}'}[(\boldsymbol{x}_0, y_0)] = \begin{cases} \Pr_{\mathcal{D}^m_{\mathcal{X}}}[\boldsymbol{x}_0], & y_0 = 0,\\ 0, & y_0 = 1, \end{cases}$$

and $\boldsymbol{z}' \in (\mathcal{X}^m \times \{0,1\})^n$ be a dataset with $z'_i = (\boldsymbol{x}_i, 0)$. Similar to the proof of Lemma 4.3, we have $\mathrm{err}_{\mathcal{D}'}(\varphi_{c,h}) = \mathrm{dis}^s_{\mathcal{D}^m_{\mathcal{X}}}(c,h)$ and $\mathrm{err}_{\boldsymbol{z}'}(\varphi_{c,h}) = \mathrm{dis}^s_{\boldsymbol{x}}(c,h)$. By Lemma 4.2, let $\gamma = 1/2$ and $\xi = \alpha/4$, we have

$$\begin{aligned}
&\Pr\left[\exists c, h \in \mathcal{C}\ s.t.\ \mathrm{dis}^s_{\mathcal{D}^m_{\mathcal{X}}}(c,h) > \alpha\ and\ \mathrm{dis}^s_{\boldsymbol{x}}(c,h) \le \alpha/2\right]\\
\le&\Pr\left[\exists c, h \in \mathcal{C}\ s.t.\ \mathrm{dis}^s_{\mathcal{D}^m_{\mathcal{X}}}(c,h) > 3/2\mathrm{dis}^s_{\boldsymbol{x}}(c,h) + \alpha/4\right]\\
=&\Pr[\exists \varphi_{c,h} \in \mathcal{C}_\varphi\ s.t.\ \mathrm{err}_{\mathcal{D}'}(\varphi_{c,h}) > 3/2\mathrm{err}_{\boldsymbol{z}'}(\varphi_{c,h}) + \alpha/4]\\
\le&4\Pi_{\mathcal{C}_\varphi}(2n)\exp\left(-\frac{-1/2\cdot\alpha n/4}{4\cdot(1+1/2)}\right)\\
\le&4\left(\Pi_{\mathcal{C}}(2nm)\right)^2\exp(-\alpha n/48),
\end{aligned}$$

where the last inequality comes from the fact that

$$\begin{aligned}
\Pi_{\mathcal{C}_\varphi}(2n) =& \max_{\boldsymbol{x}\in\mathcal{X}^{2nm}} |\Pi_{\mathcal{C}_\varphi}(\boldsymbol{x})|\\
=& \max_{\boldsymbol{x}\in\mathcal{X}^{2nm}} \left|\left\{\left(\mathbf{1}\left[\sum_{i=1}^m \mathbf{1}[c(x_{1,i}) \ne h(x_{1,i})] > s\right], \ldots, \mathbf{1}\left[\sum_{i=1}^m \mathbf{1}[c(x_{2n,i}) \ne h(x_{2n,i})] > s\right]\right) \,\middle|\, c, h \in \mathcal{C}\right\}\right|\\
\le& \max_{\boldsymbol{x}\in\mathcal{X}^{2nm}} |\{(\mathbf{1}[c(x_{1,1}) \ne h(x_{1,1})], \ldots, \mathbf{1}[c(x_{2n,m}) \ne h(x_{2n,m})]) \mid c, h \in \mathcal{C}\}|\\
\le& \max_{\boldsymbol{x}\in\mathcal{X}^{2nm}} |\Pi_{\mathcal{C}}(\boldsymbol{x})| \cdot |\Pi_{\mathcal{C}}(\boldsymbol{x})|\\
=& (\Pi_{\mathcal{C}}(2nm))^2.
\end{aligned}$$

Let $d = \mathrm{VC}(\mathcal{C})$, Sauer's Lemma ensures that the above quantity is at most β if $2nm \ge d$ and

$$n \ge \frac{48}{\alpha}\left(2d\ln n + 2d\ln(2em/d) + \ln(4/\beta)\right).$$

By inequality (1), we have

$$\frac{96d}{\alpha}\ln n \le \frac{n}{2} + \frac{96d}{\alpha}\ln\left(\frac{192d}{e\alpha}\right)$$

Thus,

$$n \ge \frac{96}{\alpha}\left(2d\ln(384m/\alpha) + \ln(4/\beta)\right)$$

suffices. □

C.3. Proof of Lemma 4.5

The key insight behind Lemma 4.5 is that we can decide whether a given value $\widetilde{\eta}$ is greater or lees than the minimum error η up to an error of $O(\alpha)$, as demonstrated in the following lemma.

Lemma C.1. *Let $\mathcal{C}$ be a concept class, $\mathcal{D}$ be a distribution over $\mathcal{Z}$, m be the number of samples per user, and $\eta = \inf_{c\in\mathcal{C}} \mathrm{err}_{\mathcal{D}}(c)$. Suppose*

$$n \ge \widetilde{O}\left(\frac{1}{\varepsilon} + \mathrm{VC}(\mathcal{C}) + \frac{\mathrm{VC}(\mathcal{C})}{m\alpha^2} + \frac{1}{\sqrt{m}\alpha\varepsilon}\right)$$

and $\boldsymbol{z} \in \mathcal{Z}^{nm}$ be a dataset with each $\boldsymbol{z}_{i,j}$ drawn i.i.d. from $\mathcal{D}$. Then there exists an ε-differentially private algorithm that takes $\boldsymbol{z}, \varepsilon, \widetilde{\eta}, \alpha$ as input and outputs some $\sigma \in \{0,1\}$, such that with probability $1-\beta$:

1. *If $\sigma = 0$, then $\widetilde{\eta} \ge \eta - \alpha/2$.*
2. *If $\sigma = 1$, then $\widetilde{\eta} \le \eta + \alpha/2$.*

Proof. We will assume $m \le 1/\alpha^2$ and prove an upper bound of $\widetilde{O}\left(\frac{\mathrm{VC}(\mathcal{C})}{m\alpha^2} + \frac{1}{\sqrt{m}\alpha\varepsilon}\right)$. When $m > 1/\alpha^2$, the result simply holds by discarding the extra samples.

By Lemma 4.1, there exists some t s.t.

$$\Pr[\mathrm{Bin}(m,\widetilde{\eta}) > t] + \sqrt{m}\alpha/1400 \le \Pr[\mathrm{Bin}(m,\widetilde{\eta}+\alpha/2) > t].$$

Let A be an algorithm that returns 0 if $\min_{c\in\mathcal{C}} \mathrm{err}^t_{\boldsymbol{z}}(c) + r \le \Pr[\mathrm{Bin}(m,\widetilde{\eta}) > t] + \sqrt{m}\alpha/2800$, where $r \sim \mathrm{Lap}(1/\varepsilon n)$, and otherwise returns 1. By Lemma 2.5, A is ε-differentially private and $|r| \le \sqrt{m}\alpha/5600$ with probability $1-\beta/2$ if $n \ge \widetilde{O}(\frac{1}{\sqrt{m}\alpha\varepsilon})$.

The agnostic generalization bound (Lemma 4.3) shows that if $n \ge \widetilde{O}\left(\frac{\mathrm{VC}(\mathcal{C})}{\sqrt{m}\alpha^2}\right)$, then with probability $1-\beta/2$ it holds that $|\mathrm{err}^t_{\boldsymbol{z}}(c) - \mathrm{err}^t_{\mathcal{D}^m}(c)| \le \sqrt{m}\alpha/5600$ for all $c \in \mathcal{C}$. For the remainder of the proof, we condition on the event that $|r| \le \sqrt{m}\alpha/5600$ and $|\mathrm{err}^t_{\boldsymbol{z}}(c) - \mathrm{err}^t_{\mathcal{D}^m}(c)| \le \sqrt{m}\alpha/5600$ for all $c \in \mathcal{C}$. By the union bound, this happens with probability $1-\beta$.

First consider the case that $\eta \le \widetilde{\eta}$. We have

$$\begin{aligned}
\min_{c\in\mathcal{C}} \mathrm{err}^t_{\boldsymbol{z}}(c) + r &\le \Pr[\mathrm{Bin}(m,\eta) > t] + \sqrt{m}\alpha/5600 + r \\
&\le \Pr[\mathrm{Bin}(m,\widetilde{\eta}) > t] + \sqrt{m}\alpha/5600 + \sqrt{m}\alpha/5600 \\
&\le \Pr[\mathrm{Bin}(m,\widetilde{\eta}) > t] + \sqrt{m}\alpha/2800.
\end{aligned}$$

Thus, A must return 0 in this case. Equivalently, it guarantees that $\eta > \widetilde{\eta} \ge \widetilde{\eta} - \alpha/2$ if A returns 1.

Now suppose $\eta > \widetilde{\eta}$. Note that if $\eta > \widetilde{\eta} + \alpha/2$, it holds that

$$\Pr[\mathrm{Bin}(m,\widetilde{\eta}) > t] + \sqrt{m}/1400 < \Pr[\mathrm{Bin}(m,\eta) > t] = \inf_{c\in\mathcal{C}} \mathrm{err}^t_{\mathcal{D}^m}(c).$$

Thus, if $\inf_{c\in\mathcal{C}} \mathrm{err}^t_{\mathcal{D}^m}(c) \le \Pr[\mathrm{Bin}(m,\widetilde{\eta}) > t] + \sqrt{m}/1400$ we must have $\eta \le \widetilde{\eta} + \alpha/2$.

Algorithm 2 PrivateMinError

Input: dataset $\boldsymbol{z} \in \mathcal{Z}^{nm}$, privacy parameter ε, accruacy parameter α, confidence parameter β, algorithm PrivateCompare.
$l \leftarrow 0, r \leftarrow 1$
$T \leftarrow \lceil \log_2(2/\alpha) \rceil$
$\varepsilon' \leftarrow \varepsilon/T, \beta' \leftarrow \beta/T$
for $k \leftarrow 1$ **to** T **do**
  $mid \leftarrow (l+r)/2$
  if PrivateCompare$(\boldsymbol{z}, \varepsilon', mid, \alpha, \beta') = 1$ **then**
    $r \leftarrow mid$
  else
    $l \leftarrow mid$
  end if
end for
return l

Suppose A returns 0. Let $c_0 \in \mathcal{C}$ s.t. $\mathrm{err}^t_{\boldsymbol{z}}(c_0) = \min_{c \in \mathcal{C}} \mathrm{err}^t_{\boldsymbol{z}}(c)$. Then we have

$$\begin{aligned}\mathrm{err}^t_{\mathcal{D}^m}(c_0) &\le \mathrm{err}^t_{\boldsymbol{z}}(c_0) + \sqrt{m}\alpha/5600 \\ &\le \Pr[\mathrm{Bin}(m, \widetilde{\eta}) > t] + \sqrt{m}\alpha/2800 - r + \sqrt{m}\alpha/5600 \\ &\le \Pr[\mathrm{Bin}(m, \widetilde{\eta}) > t] + \sqrt{m}\alpha/1400.\end{aligned}$$

Therefore, $\inf_{c \in \mathcal{C}} \mathrm{err}^t_{\mathcal{D}^m}(c) \le \Pr[\mathrm{Bin}(m, \widetilde{\eta}) > t] + \sqrt{m}\alpha/1400$, which implies that $\eta \le \widetilde{\eta} + \alpha/2$.

In summary, with probability $1-\beta$, we have $\widetilde{\eta} \ge \eta - \alpha/2$ if A returns 0, and $\widetilde{\eta} \le \eta + \alpha/2$ if A return 1. □

Now we can accurately estimate the value of η by binary research. The details are described in Algorithm 2.

Proof of Lemma 4.5. We execute Algorithm 2 with PrivateCompare being the algorithm in Lemma C.1.

Let l_k, r_k be the endpoints after the k-th iteration. Initially, we have $l_0 = 0$ and $r_0 = 1$. Then after T iterations, the length of $[l_T, r_T]$ is $r_T - l_T \le 1/2^T \le \alpha/2$.

We now show that if every call of PrivateCompare succeeds, then $[l_k, r_k] \cap [\eta - \alpha/2, \eta + \alpha/2] \ne \emptyset$ for all k. We prove this by induction. At the beginning, we have $[l_0, r_0] \cap [\eta - \alpha/2, \eta + \alpha/2] \ne \emptyset$ since $\eta \in [0,1] = [l_0, r_0]$. Suppose $[l_{k-1}, r_{k-1}] \cap [\eta - \alpha/2, \eta + \alpha/2] \ne \emptyset$. In the k-th iteration, we set $mid = (l_{k-1} + r_{k-1})/2$. Let σ be the return value of PrivateCompare. By Lemma C.1, with probability $1 - \beta' = 1 - \beta/T$, we have:

1. If $\sigma = 0$, then $mid \ge \eta - \alpha/2$. Thus, $[l_{k-1}, mid] \cap [\eta - \alpha/2, \eta + \alpha/2] \ne \emptyset$ since $\eta + \alpha/2 \ge l_{k-1}$.
2. If $\sigma = 1$, then $mid \le \eta + \alpha/2$. Thus, $[mid, r_{k-1}] \cap [\eta - \alpha/2, \eta + \alpha/2] \ne \emptyset$ since $\eta - \alpha/2 \ge r_{k-1}$.

For both cases, we have $[l_k, r_k] \cap [\eta - \alpha/2, \eta + \alpha/2] \ne \emptyset$.

By the union bound, it holds with probability $1 - \beta$ that $|\eta - x| \le \alpha/2$ for some $x \in [l_T, r_T]$. This implies $|l_T - \eta| \le |l_T - x| + |x - \eta| \le \alpha$, which completes the proof. □

C.4. Proof of Theorem 4.6

Proof. We will assume that $m \le 1/\alpha^2$ and show an upper bound of $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\sqrt{m}\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{m\alpha^2}\right)$. When $m > 1/\alpha^2$, the $\widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\varepsilon}\right)$ term becomes the dominated term and the user complexity can be achieved by discarding the extra samples.

Let $\eta = \inf_{c \in \mathcal{C}} \mathrm{err}_{\mathcal{D}}(c)$. We first run the algorithm in Lemma 4.5 with privacy parameter $\varepsilon/2$. For $n \ge \widetilde{O}\left(\frac{\mathrm{VC}(\mathcal{C})}{m\alpha^2} + \frac{1}{\sqrt{m}\alpha\varepsilon}\right)$, it returns some $\hat{\eta}$ s.t. $|\hat{\eta} - \eta| \le \alpha/6$ with probability $1 - \beta/7$.

By Lemma 4.1, there exists some t s.t.

$$\Pr[\mathrm{Bin}(m, \hat{\eta} + \alpha/6) > t] + \sqrt{m}\alpha/4200 \leq \Pr[\mathrm{Bin}(m, \hat{\eta} + \alpha/3) > t].$$

Let $\psi = \Pr[\mathrm{Bin}(m, \eta) > t]$ and $\hat{\psi} = \Pr[\mathrm{Bin}(m, \hat{\eta} + \alpha/6) > t]$. Then we have $\psi \leq \hat{\psi}$ since $\eta \leq \hat{\eta} + \alpha/6$.

For any concept c with $\mathrm{err}_{\mathcal{D}}(c) > \eta + \alpha/2$, we have $\mathrm{err}_{\mathcal{D}}(c) > \hat{\eta} + \alpha/3$ by the triangle inequality. Thus,

$$\mathrm{err}^t_{\mathcal{D}^m}(c) > \Pr[\mathrm{Bin}(m, \hat{\eta} + \alpha/3) > t] \geq \hat{\psi} + \sqrt{m}\alpha/4200.$$

Thus, for any concept c with $\mathrm{err}^t_{\mathcal{D}^m}(c) \leq \hat{\psi} + \sqrt{m}\alpha/4200$ it must hold that $\mathrm{err}_{\mathcal{D}}(c) \leq \eta + \alpha/2$.

Again by Lemma 4.1, there exists some s (actually, $s = 0$) s.t.

$$\Pr[\mathrm{Bin}(m, \alpha/3) > s] \geq \Pr[\mathrm{Bin}(m, 0) > s] + \sqrt{m}\alpha/2100 = \sqrt{m}\alpha/2100.$$

Thus, for any hypotheses c and h with $\mathrm{dis}^s_{\mathcal{D}^m_{\mathcal{X}}}(c, h) \leq \sqrt{m}\alpha/2100$, it holds that $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(c, h) \leq \alpha/3$.

Our algorithm then works as follows: first samples $\mathcal{H} \sim \mathcal{P}$, where $\mathcal{P}$ is a $(\alpha/134400\sqrt{m}, \beta/7)$-probabilistic representation of $\mathcal{C}$ with $\mathrm{size}(\mathcal{P}) = \mathrm{RepD}_{\alpha/134400\sqrt{m},\beta/7}(\mathcal{C})$, then run the exponential mechanism on $\mathcal{H}$ with parameter $\varepsilon/2$, sensitivity parameter $\Delta = 2/n$, and score function

$$q(\boldsymbol{z}, h) = \min_{c \in \mathcal{C}} \mathrm{err}^t_{\boldsymbol{z}}(c) + \mathrm{dis}^s_{\boldsymbol{x}}(c, h).$$

The privacy is satisfied by the composition property of DP. The proof of the utility guarantee is similar to that of Theorem 3.4. Firstly, there exists some $c' \in \mathcal{C}$ s.t. $\mathrm{err}^t_{\mathcal{D}^m}(c') \leq \psi + \sqrt{m}\alpha/134400$. By Hoeffding's inequality, for $n \geq \widetilde{O}\left(\frac{1}{m\alpha^2}\right)$, we have $\mathrm{err}^t_{\boldsymbol{z}}(c') \leq \mathrm{err}^t_{\mathcal{D}^m}(c') + \sqrt{m}\alpha/134400 \leq \psi + \sqrt{m}\alpha/67200$ with probability $1 - \beta/7$.

Since $\mathcal{P}$ is an $(\alpha/134400\sqrt{m}, \beta/7)$-probabilistic representation, with probability $1 - \beta/7$ there exists some $h' \in \mathcal{H}$ s.t. $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(c', h') \leq \alpha/134400\sqrt{m}$. By the upper bound in Lemma 4.1, $\mathrm{dis}^s_{\mathcal{D}^m_{\mathcal{X}}}(c', h') \leq \sqrt{m}\alpha/134400$. Then by the Chernoff bound, for $n \geq \widetilde{O}\left(\frac{1}{\sqrt{m}\alpha}\right)$ it holds that $\mathrm{dis}^s_{\boldsymbol{x}}(c', h') \leq 2\mathrm{dis}^s_{\mathcal{D}^m_{\mathcal{X}}}(c', h') \leq \sqrt{m}\alpha/67200$ with probability at least $1 - \beta/7$.

So, $q(\boldsymbol{z}, h') \leq \mathrm{err}^t_{\boldsymbol{z}}(c') + \mathrm{dis}^s_{\boldsymbol{x}}(c', h') \leq \psi + \sqrt{m}\alpha/33600$. By Lemma 2.6, for $n \geq \widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\sqrt{m}\alpha\varepsilon}\right)$, the exponential mechanism chooses some h_0 s.t.

$$q(\boldsymbol{z}, h_0) \leq \min_{f \in \mathcal{H}} q(\boldsymbol{z}, f) + \sqrt{m}\alpha/16800 \leq q(\boldsymbol{z}, h') + \sqrt{m}\alpha/16800 \leq \psi + \sqrt{m}\alpha/8400$$

with probability $1 - \beta/7$.

Let $q(\boldsymbol{z}, h_0) = \mathrm{err}^t_{\boldsymbol{z}}(c_0) + \mathrm{dis}^s_{\boldsymbol{x}}(c_0, h_0)$. Then we have $\mathrm{err}^t_{\boldsymbol{z}}(c_0) \leq q(\boldsymbol{z}, h_0) \leq \psi + \sqrt{m}\alpha/8400$. By Lemma 4.3, for $n \geq \widetilde{O}\left(\frac{\mathrm{VC}(\mathcal{C})}{m\alpha^2}\right)$ we have $|\mathrm{err}^t_{\mathcal{D}^m}(c) - \mathrm{err}^t_{\boldsymbol{z}}(c)| \leq \sqrt{m}\alpha/8400$ for all $c \in \mathcal{C}$ with probability $1 - \beta/7$. Thus,

$$\mathrm{err}^t_{\mathcal{D}^m}(c_0) \leq \mathrm{err}^t_{\boldsymbol{z}}(c_0) + \sqrt{m}\alpha/8400 \leq \psi + \sqrt{m}\alpha/4200 \leq \hat{\psi} + \sqrt{m}\alpha/4200$$

and $\mathrm{err}^t_{\boldsymbol{z}}(c_0) \geq \psi - \sqrt{m}\alpha/8400$. The former implies $\mathrm{err}_{\mathcal{D}}(c_0) \leq \eta + \alpha/2$ and the latter implies

$$\mathrm{dis}^s_{\boldsymbol{x}}(c_0, h_0) = q(\boldsymbol{z}, h_0) - \mathrm{err}_{\boldsymbol{z}}(c_0) \leq \psi + \sqrt{m}\alpha/8400 - (\psi - \sqrt{m}\alpha/8400) \leq \sqrt{m}\alpha/4200.$$

As shown in the proof of Theorem 3.4, the VC dimension of $\mathcal{C} \cup \mathcal{H}$ is $\widetilde{O}(\mathrm{RepD}(\mathcal{C}))$. Then by Lemma 4.4, for $n \geq \widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\sqrt{m}\alpha}\right)$, it holds with probability $1 - \beta/7$ that $\mathrm{dis}^s_{\mathcal{D}^m_{\mathcal{X}}}(c, h) \leq \sqrt{m}\alpha/2100$ for all $c \in \mathcal{C}$ and $h \in \mathcal{H}$ with $\mathrm{dis}^s_{\boldsymbol{x}}(c, h) \leq \sqrt{m}\alpha/4200$. Thus, we have $\mathrm{dis}^s_{\mathcal{D}^m_{\mathcal{X}}}(c_0, h_0) \leq \sqrt{m}\alpha/2100$, which further indicates $\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(c_0, h_0) \leq \alpha/3$.

By the union bound, we have

$$\mathrm{err}_{\mathcal{D}}(h_0) \leq \mathrm{err}_{\mathcal{D}}(c_0) + \mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(c_0, h_0) \leq \eta + \alpha/2 + \alpha/3 < \eta + \alpha$$

with probability $1 - \beta$ for $n \geq \widetilde{O}\left(\frac{\mathrm{RepD}(\mathcal{C})}{\sqrt{m}\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{m\alpha^2}\right)$. □

D. Proofs from Section 5

D.1. Proof of Lemma 5.1

We first prove an corollary of Lemma 4.1 by taking $p = \alpha/2$ and $q = 2\alpha/3$, which states that the total variation distance between $\mathrm{Bin}(m, p)$ and $\mathrm{Bin}(m, q)$ scales linearly with m.

Corollary D.1. *Suppose $0 < \alpha \leq 1$ and $m \leq 1/\alpha$. Then there exists an ℓ such that*

$$\Pr[\mathrm{Bin}(m, \alpha/2) > \ell] + m\alpha/4200 \leq \Pr[\mathrm{Bin}(m, 2\alpha/3) > \ell].$$

Proof. Let $p = \alpha/2$ and $q = 2\alpha/3$. Then by Lemma 4.1, there exists an ℓ s.t.

$$\Pr[\mathrm{Bin}(m, \alpha/2) > \ell] + K/700 \leq \Pr[\mathrm{Bin}(m, 2\alpha/3) > \ell],$$

where

$$K = \min\left(m|p-q|, \frac{\sqrt{m}|p-q|}{\sqrt{p(1-p)}}, 1\right) = \min(m\alpha/6, \sqrt{m/(p(1-p))} \cdot \alpha/6) = m\alpha/6.$$

since $m/(p(1-p)) > m/p = 2m/\alpha \geq 2m^2$. □

Proof of Lemma 5.1. We will assume $m \leq 1/\alpha$ and prove an upper bound of $\widetilde{O}\left(\frac{\log|\mathcal{X}|}{m\alpha\varepsilon}\right)$. When $m > 1/\alpha$, the conclusion follows by discarding the extra samples.

Let $\boldsymbol{x}$ be the corresponding unlabeled dataset of $\boldsymbol{z}$. By Corollary D.1, there exists some s .s.t.

$$\Pr[\mathrm{Bin}(m, \alpha/2) > s] + m\alpha/4200 \leq \Pr[\mathrm{Bin}(m, 2\alpha/3) > s].$$

Let $\mathrm{dis}^s_{\boldsymbol{x}}(f_{a-1}, f_b)$ be the fraction of users that have more than ms samples lie in $[a, b]$. Define score function

$$q(\boldsymbol{z}, u) = \max\left(\mathrm{dis}^s_{\boldsymbol{x}}(f_{l-1}, f_{u-1}), \mathrm{dis}^s_{\boldsymbol{x}}(f_u, f_r)\right).$$

The sensitivity of $q(\boldsymbol{z}, u)$ is $1/n$. Thus, we can run the exponential mechanism on $\{l, \ldots, r\}$ with privacy parameter ε and $\Delta = 1/n$. Lemma 2.6 proves the privacy guarantee. To analyze the accuracy, we start with the claim that there exists some $u' \in \{l, \ldots, r\}$ such that

$$\max\left(\Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in [l, u'-1]], \Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in [u'+1, r]]\right) \leq \alpha/2.$$

To show this, let u' be the greatest number in $\{l, \ldots, r\}$ s.t. $\Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in [l, u'-1]] \leq \alpha/2$. We will prove that $\Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in [u'+1, r]] \leq \alpha/2$. Suppose this does not hold, namely, $\Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in [u'+1, r]] > \alpha/2$. Then by the definition of u' we have $\Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in [l, u']] > \alpha/2$. Thus,

$$\alpha \geq \Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in [l, r]] = \Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in [l, u']] + \Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in [u'+1, r]] > \alpha,$$

a contradiction.

Thus, we have

$$\max\left(\Pr_{\boldsymbol{x}_0\sim\mathcal{D}^m_{\mathcal{X}}}[\mathrm{dis}_{\boldsymbol{x}_0}(f_{l-1}, f_{u'-1}) > s], \Pr_{\boldsymbol{x}_0\sim\mathcal{D}^m_{\mathcal{X}}}[\mathrm{dis}_{\boldsymbol{x}_0}(f_{u'}, f_r) > s]\right) \leq \Pr[\mathrm{Bin}(m, \alpha/2) > s].$$

Moreover, by the upper bound in Lemma 4.1, the right-hand side of the above is at most $m\alpha/2$. Now apply the Chernoff bound, for $n \geq \widetilde{O}\left(\frac{1}{m\alpha}\right)$, with probability $1 - \beta/3$, we have

$$\begin{aligned} q(\boldsymbol{z}, u') &= \max\left(\Pr[\mathrm{dis}^s_{\boldsymbol{x}}(f_{l-1}, f_{u'-1})], \Pr[\mathrm{dis}^s_{\boldsymbol{x}}(f_{u'}, f_r)]\right) \\ &\leq \Pr[\mathrm{Bin}(m, \alpha/2) > s](1 + 1/6300) \\ &\leq \Pr[\mathrm{Bin}(m, \alpha/2) > s] + m\alpha/12600. \end{aligned}$$

By Lemma 2.6, for $n \geq \widetilde{O}\left(\frac{\log|\mathcal{X}|}{m\alpha\varepsilon}\right)$, the exponential mechanism returns a mid s.t.

$$q(\boldsymbol{z}, mid) \leq \min_{u\in\{l,\ldots,r\}} q(\boldsymbol{z}, u) + m\alpha/12600 \leq q(\boldsymbol{z}, u') + m\alpha/12600 \leq \Pr[\mathrm{Bin}(m, \alpha/2) > s] + m\alpha/6300$$

with probability at least $1-\beta/3$. By the Chernoff bound, when $n \geq \widetilde{O}\left(\frac{\log|X|}{m\alpha}\right)$, it holds with probability $1-\beta/3$ that

$$\mathrm{dis}^s_{\mathcal{D}^m_{\mathcal{X}}}(f_u, f_v) \leq (\Pr[\mathrm{Bin}(m, \alpha/2) > s] + m\alpha/6300) \cdot \frac{6303}{6302}$$

for all $u, v \in \{0\} \cup \mathcal{X}$ with $\mathrm{dis}^s_{\boldsymbol{x}}(f_u, f_v) \leq \Pr[\mathrm{Bin}(m, \alpha/2) > s] + m\alpha/6300$. Therefore, we have

$$\begin{aligned}
\max\left(\mathrm{dis}^s_{\mathcal{D}^m_{\mathcal{X}}}(f_{l-1}, f_{mid-1}), \mathrm{dis}^s_{\mathcal{D}^m_{\mathcal{X}}}(f_{mid+1}, r)\right) &\leq (\Pr[\mathrm{Bin}(m, \alpha/2) > s] + m\alpha/6300) \cdot \frac{6303}{6302} \\
&\leq \Pr[\mathrm{Bin}(m, \alpha/2) > s] + m\alpha/6300 + (m\alpha/2 + m\alpha/6300) \cdot \frac{1}{6302} \\
&= \Pr[\mathrm{Bin}(m, \alpha/2) > s] + m\alpha/4200 \\
&\leq \Pr[\mathrm{Bin}(m, 2\alpha/3) > s].
\end{aligned}$$

Thus, we got

$$\max\left(\Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in [l, mid-1]], \Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in [mid+1, r]]\right) = \max\left(\mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(f_{l-1}, f_{mid-1}), \mathrm{dis}_{\mathcal{D}_{\mathcal{X}}}(f_{mid+1}, r)\right) \leq 2\alpha/3.$$

By the union bound, the above happens with probability at least $1-\beta$, which completes the proof. □

D.2. Proof of Theorem 5.2

Proof. As proved by Ghazi et al. (2023), private realizable learning requires $\Omega\left(\frac{\mathrm{RepD}(\mathcal{C})}{\varepsilon} + \frac{\mathrm{RepD}(\mathcal{C})}{m\alpha\varepsilon}\right)$ users and private agnostic learning requires $\Omega\left(\frac{\mathrm{VC}(\mathcal{C})}{\sqrt{m}\alpha\varepsilon} + \frac{\mathrm{VC}(\mathcal{C})}{m\alpha^2}\right)$. By the fact that the class of thresholds has VC dimension 1 (see, e.g., (Shalev-Shwartz & Ben-David, 2014)) and representation dimension $\Theta(\log|\mathcal{X}|)$ (Feldman & Xiao, 2014), the lower bound is proved.

For the upper bound, we assume that $m \leq 1/\alpha^2$. The bound for $m > 1/\alpha^2$ follows by discarding the extra samples.

Let $\eta = \inf_{c\in\mathcal{C}} \mathrm{err}_{\mathcal{D}}(c)$ and $T = \log_{3/2}\left(\frac{2}{\alpha}\right) = O(\log(1/\alpha))$.We first run the algorithm in Lemma 4.5 with parameter $\varepsilon/2$. For $n \geq \widetilde{O}\left(\frac{1}{m\alpha^2} + \frac{1}{\sqrt{m}\alpha\varepsilon}\right)$, it returns some $\hat{\eta}$ s.t. $|\hat{\eta} - \eta| \leq \alpha/6$ with probability $1-\beta/4$. By Lemma 4.1, there exists some t s.t.

$$\Pr[\mathrm{Bin}(m, \hat{\eta} + \alpha/6) > t] + \sqrt{m}\alpha/4200 \leq \Pr[\mathrm{Bin}(m, \hat{\eta} + \alpha/3) > t]$$

Let $\psi = \Pr[\mathrm{Bin}(m, \eta) > t]$ and $\hat{\psi} = \Pr[\mathrm{Bin}(m, \hat{\eta} + \alpha/6)]$. Since $\eta \leq \hat{\eta} + \alpha/6$, we have $\psi \leq \hat{\psi}$. Thus, for any u s.t. $\mathrm{err}^t_{\mathcal{D}^m}(f_u) \leq \psi + \sqrt{m}\alpha/4200$, it holds that $\mathrm{err}_D(f_u) \leq \hat{\eta} + \alpha/3 \leq \eta + \alpha/2$.

We then run Algorithm 1 with dataset $\boldsymbol{z}$, privacy parameter $\varepsilon/2$, T and t as declared before, algorithm PrivateCompare as the one in Lemma 5.1, and $\theta = 2/3$. The privacy guarantee of this part direcly follows from the property of Laplace mechanism (Lemma 2.5) and the basic composition. Thus the privacy of the entire algorithm is ensured by applying the basic composition again.

We now consider the utility. Let l_k, r_k be the endpoints after the k-th iteration and T' be the number of finished iterations during the execution. Then there are three possible cases:

- $T' < T$ and the algorithm returns f_{mid} during the $(T'+1)$-th iteration
- $T' < T$ and the algorithm returns $f_{l_{T'}} = f_{r_{T'}}$
- $T' = T$ and the algorithm returns f_{l_T}.

We first prove that

$$\min_{u\in\{l_k,\ldots,r_k\}} \mathrm{err}^t_{\boldsymbol{z}}(u) \le \psi + \sqrt{m}\alpha/12600 + k\sqrt{m}\alpha/(12600T) \tag{3}$$

for $0 \le k \le T'$ by induction.

Let $n \ge \widetilde{O}\left(\frac{1}{\sqrt{m}\alpha\varepsilon}\right)$, then by Lemma 2.5 and the union bound, with probability $1-\beta/4$, in each iteration the absolute value of the generated Laplace noise is no larger than $\alpha\sqrt{m}/(25200T)$. At the beginning, we have $l_0 = 0$ and $r_0 = |\mathcal{X}|$. By agnostic generalization (Lemma 4.3), for $n \ge \widetilde{O}\left(\frac{1}{m\alpha^2}\right)$, we have $|\mathrm{err}^t_{\boldsymbol{z}}(f_u) - \mathrm{err}^t_{\mathcal{D}^m}(f_u)| \le \sqrt{m}\alpha/12600$ for all $u \in \{0,\ldots,|\mathcal{X}|\}$ with probability $1-\beta/4$. This implies $\min_{u\in\{l_0,\ldots,r_0\}} \mathrm{err}^t_{\boldsymbol{z}}(f_u) \le \psi + \sqrt{m}\alpha/12600$.

Now suppose (3) holds for some $k < T'$. Since the algorithm does not return during the $(k+1)$-th iteration, we thus have

$$\begin{aligned}
\min_{u\in\{l_{k+1},\ldots,r_{k+1}\}} \mathrm{err}^t_{\boldsymbol{z}}(f_u) &\le \min(v_l, v_r, v_{mid}) + \sqrt{m}\alpha/(25200T) \\
&\le \min_{u\in\{l_k,\ldots,r_k\}} \mathrm{err}^t_{\boldsymbol{z}}(f_u) + \sqrt{m}\alpha/(25200T) + \sqrt{m}\alpha/(25200T) \\
&\le \psi + \sqrt{m}\alpha/12600 + (k+1)\sqrt{m}\alpha/(12600T).
\end{aligned}$$

Thus, we prove (3). Now, let's consider the three cases one by one. If the algorithm returns f_{mid} during the $(T'+1)$-th iteraiton, then we have

$$\begin{aligned}
\mathrm{err}^t_{\boldsymbol{z}}(f_{mid}) &\le \min(v_l, v_r, v_{mid}) + \sqrt{m}\alpha/(25200T) \\
&\le \min_{u\in\{l_{T'},\ldots,r_{T'}\}} \mathrm{err}^t_{\boldsymbol{z}}(f_u) + \sqrt{m}\alpha/(25200T) + \sqrt{m}\alpha/(25200T) \\
&\le \psi + \sqrt{m}\alpha/12600 + T'\cdot\sqrt{m}\alpha/(12600T) + \sqrt{m}\alpha/(12600T) \\
&\le \psi + \sqrt{m}\alpha/6300.
\end{aligned}$$

Otherwise if the algorithm returns $f_{l_{T'}}$ due to $l_{T'} = r_{T'}$ after T' iterations, we also have

$$\mathrm{err}^t_{\boldsymbol{z}}(f_{l_{T'}}) \le \psi + \sqrt{m}\alpha/12600 + T'\cdot\sqrt{m}\alpha/(12600T) \le \psi + \sqrt{m}\alpha/6300.$$

Thus by agnostic generalization, for both of these two cases, the algorithm returns some f_{u_0} s.t.

$$\mathrm{err}_{\mathcal{D}^m}(f_{u_0}) \le \psi + \sqrt{m}\alpha/6300 + \sqrt{m}\alpha/12600 \le \psi + \sqrt{m}\alpha/4200.$$

This implies $\mathrm{err}_{\mathcal{D}}(f_{u_0}) \le \eta + \alpha/2 < \eta + \alpha$.

Now consider the remaining case that the algorithm returns f_{l_T} after T iterations. The same argument shows that there exists some $l_T \le u' \le r_T$ s.t. $\mathrm{err}_{\mathcal{D}}(f_{u'}) \le \eta + \alpha/2$. By Lemma 5.1 and the union bound, we know that for $n \ge \widetilde{O}\left(\frac{\log|\mathcal{X}|}{\varepsilon} + \frac{\log|\mathcal{X}|}{m\alpha\varepsilon}\right)$, it holds with probability $1-\beta/4$ that $\Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in [l_T, r_T]] \le \left(\frac{2}{3}\right)^T \le \alpha/2$. Thus,

$$\mathrm{err}_{\mathcal{D}}(f_{l_T}) \le \mathrm{err}_{\mathcal{D}}(f_{u'}) + \Pr_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in [l_T, r_T]] \le \eta + \alpha/2 + \alpha/2 \le \eta + \alpha.$$

All in all, when $n \ge \widetilde{O}\left(\frac{\log|\mathcal{X}|}{\varepsilon} + \frac{\log|\mathcal{X}|}{m\alpha\varepsilon} + \frac{1}{\sqrt{m}\alpha\varepsilon} + \frac{1}{m\alpha^2}\right)$, the algorithm is an (α,β)-agnostic learner with ε-differential privacy. □