Privately Answering Classification Queries in the Agnostic PAC Model

Anupama Nandi * NANDI.10@OSU.EDU
Department of Computer Science & Engineering
The Ohio State University
Columbus,OH

Raef Bassily * BASSILY.1@OSU.EDU
Department of Computer Science & Engineering
The Ohio State University
Columbus,OH

Editors: Aryeh Kontorovich and Gergely Neu

Abstract

We revisit the problem of differentially private release of classification queries. In this problem, the goal is to design an algorithm that can accurately answer a sequence of classification queries based on a private training set while ensuring differential privacy. We formally study this problem in the agnostic PAC model and derive a new upper bound on the private sample complexity. Our results improve over those obtained in a recent work (Bassily et al., 2018) for the agnostic PAC setting. In particular, we give an improved construction that yields a tighter upper bound on the sample complexity. Moreover, unlike (Bassily et al., 2018), our accuracy guarantee does not involve any blow-up in the approximation error associated with the given hypothesis class.

Given any hypothesis class with VC-dimension d, we show that our construction can privately answer up to m classification queries with average excess error α using a private sample of size $\approx \frac{d}{\alpha^2}\,\max\left(1, \sqrt{m}\,\alpha^{3/2}\right)$ (assuming the privacy parameter $\epsilon = \Theta(1)$). Using recent results on private learning with auxiliary public data, we extend our construction to show that one can privately answer *any* number of classification queries with average excess error α using a private sample of size $\approx \frac{d}{\alpha^2}\,\max\left(1, \sqrt{d}\,\alpha\right)$. When $\alpha = O\left(\frac{1}{\sqrt{d}}\right)$ and the privacy parameter $\epsilon = \Theta(1)$, our private sample complexity bound is essentially optimal.

Keywords: Differential privacy, agnostic PAC model, classification queries.

1. Introduction

In this paper, we revisit the problem of answering a sequence of classification queries in the *agnostic* PAC model under the constraint of (ϵ, δ)-differential privacy. An algorithm for this problem is given a *private* training dataset $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of n i.i.d. binary-labeled examples drawn from some unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ denotes an arbitrary data domain (space of feature-vectors) and $\mathcal{Y}$ denotes a set of binary labels (e.g., $\{0, 1\}$). The algorithm is also given as input some hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ of binary functions mapping $\mathcal{X}$ to $\mathcal{Y}$. The algorithm accepts a sequence of classification queries given by a sequence of i.i.d. feature-vectors $\mathcal{Q} = (\tilde{x}_1, \tilde{x}_2, \ldots)$, drawn from the marginal distribution of $\mathcal{D}$ over $\mathcal{X}$, denoted as $\mathcal{D}_{\mathcal{X}}$. Here, the feature-vectors defining the set of queries $\mathcal{Q}$ do not involve any privacy constraint. The queries are also assumed to arrive one at a time, and the algorithm is required to answer the current query $\tilde{x}_j$ by predicting a label $\hat{y}_j$ for it before seeing the next query (*online setting*). The goal is to answer up to a given number m of queries (which is a parameter of the problem) such that, (i) the entire process of answering the m queries is (ϵ, δ)-differentially private, and (ii) the average excess error in the predicted labels does not

* Part of this work was done while the authors were visiting Simons Institute for the Theory of Computing. Research supported by NSF Awards AF-1908281, SHF-1907715, Google Faculty Research Award, and OSU faculty start-up support.

exceed some desired level $\alpha \in (0,1)$; specifically, $\frac{1}{m}\sum_{j=1}^{m} \mathbf{1}\left(\hat{y}_j \neq \tilde{y}_j\right) \leq \alpha + \min_{h\in\mathcal{H}} \mathsf{err}\left(h;\mathcal{D}\right)$, where $\tilde{y}$ is the corresponding (hidden) true label, and $\min_{h\in\mathcal{H}} \mathsf{err}(h;\mathcal{D})$ is the approximation error associated with $\mathcal{H}$, i.e., the least possible true (population) error that can be attained by a hypothesis in $\mathcal{H}$ (see Section 2 for formal definitions).

One could argue that a more direct approach for differentially private classification would be to design a differentially private learner that, given a private training set as input, outputs a classifier that is safe to publish and then can be used to answer any number of classification queries. However, there are several pessimistic results that either limit or eliminate the possibility of differentially private learning even for elementary problems such as one-dimensional thresholds (Bun et al., 2015; Alon et al., 2018). Therefore, it is natural to study the problem of classification-query release under differential privacy as an alternative approach.

A recent formal investigation of this problem was carried out in (Bassily et al., 2018). This recent work gives an algorithm based on a combination of two useful techniques from the literature on differential privacy, namely, the *sub-sample-and-aggregate* technique (Nissim et al., 2007; Smith and Thakurta, 2013) and the *sparse-vector* technique (Dwork and Roth, 2014). The algorithm by Bassily et al. (2018), hereafter denoted as $\mathcal{A}_{\mathsf{SubSamp}}$, assumes oracle access to a generic, non-private (agnostic) PAC learner $\mathcal{B}$ for $\mathcal{H}$. In this work, we give non-trivial improvements over the results of (Bassily et al., 2018) in the agnostic PAC setting. More details on the comparison with (Bassily et al., 2018) are given in the "Related work" section below. Our improvements are in terms of the attainable accuracy guarantees and the associated private sample complexity bounds in the agnostic setting. These improvements are achieved via importing new ideas and techniques from literature (particularly, the elegant agnostic-to-realizable reduction technique of (Beimel et al., 2015)) to provide an improved construction for the one that appeared in (Bassily et al., 2018).

Main results

In this work, we formally study algorithms for classification queries release under differential privacy in the agnostic PAC model. We focus on the sample complexity of such algorithms as a function of the privacy and accuracy parameters as well as the number of queries to be answered. For simplicity, in the expressions given below for our upper bounds, we will assume that $\epsilon = \Theta(1)$.

- We give an algorithm for this problem that is well-suited for the agnostic setting. Our algorithm is a two-stage construction that is based on a careful combination of the relabeling technique of (Beimel et al., 2015) and the private classification algorithm $\mathcal{A}_{\mathsf{SubSamp}}$ by Bassily et al. (2018) (see "Techniques" section below).

- We show that our construction provides significant improvements over the results of Bassily et al. (2018) for the agnostic setting:

  - The error guarantees in (Bassily et al., 2018) involves a constant blow-up (a multiplicative factor ≈ 3) in the approximation error $\min_{h\in\mathcal{H}} \mathsf{err}(h;\mathcal{D})$ associated with the given hypothesis class $\mathcal{H}$. Using our construction, we give a standard excess error guarantee that does not involve such a blow-up.
  - We show that our construction can answer up to m queries with average excess error α using a private sample whose size $\approx \mathsf{VC}(\mathcal{H})/\alpha^2 \cdot \max\left(1, \sqrt{m}\,\alpha^{3/2}\right)$ (assuming ϵ is a constant, e.g. 0.1), where $\mathsf{VC}(\mathcal{H})$ is the VC-dimension of $\mathcal{H}$. Note that this implies that we can answer up to $\approx 1/\alpha^3$ queries with private sample size that is essentially the same as the standard non-private sample complexity in the agnostic PAC model. i.e., that many queries can be answered with essentially no additional cost due to privacy.
  - Using a recent result of Alon et al. (2019) on the sample complexity of *semi-private learners* (introduced by Beimel et al. (2013)), we show that our construction immediately leads to a *universal* private classification algorithm that can answer *any* number of classification queries using a private sample of size

$\approx \frac{\mathsf{VC}(\mathcal{H})}{\alpha^2} \cdot \max\left(1, \sqrt{\mathsf{VC}(\mathcal{H})}\,\alpha\right)$, which is independent of the number of queries. We note that when $\alpha = O\left(1/\sqrt{\mathsf{VC}(\mathcal{H})}\right)$ and assuming the privacy parameter $\epsilon = \Theta(1)$, our sample bound nearly matches the standard non-private sample complexity in agnostic PAC model. This implies that in this regime, we attain a nearly optimal sample complexity bound for privately answering *any* number of classification queries. Equivalently, our bound is nearly optimal for any class $\mathcal{H}$ with $\mathsf{VC}(\mathcal{H}) = O(1/\alpha^2)$. We note that the setting studied by Alon et al. (2019) is tantamount to the setting of offline (batch) classification where the whole set of unlabeled data (the set of queries in our case) is available and given to the algorithm beforehand. Whereas, as described earlier, in this work we study the online setting (which was also studied by Bassily et al. (2018)). Hence, the upper bound on the private sample complexity obtained by Alon et al. (2019) is *not* valid in our setting.

Techniques: Our algorithm is a two-stage construction. In the first stage, the input training set is pre-processed once and for all via a relabeling procedure due to Beimel et al. (2015) in which the labels are replaced with the labels generated by an appropriately chosen hypothesis in the given hypothesis class $\mathcal{H}$. This step allows us to reduce the agnostic setting to a realizable one. In the second stage, we first sample a new training set from the empirical distribution of the relabeled set in the first stage, then feed it to $\mathcal{A}_{\mathsf{SubSamp}}$ of Bassily et al. (2018) together with other appropriately chosen input parameters. To formally prove the accuracy guarantee of our construction, in our analysis we use some tools from learning theory (e.g., the uniform-convergence argument of Claim 10). As mentioned earlier, we also use the framework of semi-private learning (Beimel et al., 2013; Alon et al., 2019) to transform our algorithm into a universal private classification algorithm.

Related work

Our results are most closely related to the results given by Bassily et al. (2018). They provide formal accuracy guarantees for their algorithm in both the realizable and agnostic settings of the PAC model. However, the accuracy guarantees Bassily et al. (2018) provide for the agnostic setting is far from optimal. In particular, their guarantees involves a constant blow-up in the approximation error $\min_{h\in\mathcal{H}} \mathsf{err}(h;\mathcal{D})$, which would limit the utility of their construction in scenarios where the approximation error is not negligible. In fact, in most typical scenarios in practice, the approximation error associated with the hypothesis (model) class is a non-negligible constant, (e.g., the test error attained by some state-of-the-art neural networks on benchmark datasets can be as large as 5%, or 10%). Our improved construction avoids this blow-up in the approximation error.

The construction by Bassily et al. (2018) can answer up to m queries with average excess error $\alpha + O(\gamma)$ (where $\gamma = \min_{h\in\mathcal{H}} \mathsf{err}(h;\mathcal{D})$ is the approximation error) using a private sample of size $\approx \frac{\mathsf{VC}(\mathcal{H})}{\alpha^2} \cdot \max\left(1, \sqrt{m\,\alpha}\right)$ (follows from Theorem 3.5 Bassily et al., 2018). Given our results discussed in the "Main results" section above, it follows that our sample complexity bound is tighter than that of Bassily et al. (2018) by roughly a factor of $\max\left(1, \min\left(\sqrt{m\,\alpha},\ 1/\alpha\right)\right)$. In particular, our bound is tighter by roughly a factor of $\sqrt{m\alpha}$ for $\frac{1}{\alpha} \leq m < \frac{1}{\alpha^3}$, and it is tighter by roughly a factor of $\frac{1}{\alpha}$ for $m \geq \frac{1}{\alpha^3}$. Equivalently, for the same private sample size, our construction can answer roughly a factor of $1/\alpha^2$ more queries than that of Bassily et al. (2018).

Bassily et al. (2018) also extend their construction to provide a semi-private learner that can finally produce a classifier. This is done by answering a sufficiently large number of queries then applying the *knowledge transfer* technique using the new training set formed by the set of answered queries. The output classifier can then be used to answer any subsequent queries, and hence, their extended construction provides a universal private classification algorithm. Their private sample complexity bound for this task is $\approx \mathsf{VC}(\mathcal{H})^{3/2}/\alpha^{5/2}$ (see Theorem 4.3 Bassily et al., 2018). On the other hand, our universal private classification algorithm yields a private sample complexity bound $\approx \frac{\mathsf{VC}(\mathcal{H})}{\alpha^2} \cdot \max\left(1, \sqrt{\mathsf{VC}(\mathcal{H})}\,\alpha\right)$, which is tighter than that of Bassily

et al. (2018) by roughly a factor of $\min\left(\sqrt{\frac{\mathsf{VC}(\mathcal{H})}{\alpha}}, \frac{1}{\alpha^{3/2}}\right)$. Moreover, our bound is nearly optimal when $\alpha = O\left(1/\sqrt{\mathsf{VC}(\mathcal{H})}\right)$.

Other related works: Dwork and Feldman (2018) consider the same problem, but focus on the sample complexity of a single query. In the agnostic PAC setting, their accuracy guarantee does not suffer from the constant blow-up in the approximation error as in the results of Bassily et al. (2018). However, their sample complexity upper bound scales as $\approx \mathsf{VC}(\mathcal{H})/\epsilon\,\alpha^3$, which is sub-optimal. Assuming $\epsilon = \Theta(1)$, our bound in the single-query setting (i.e., $m = 1$) is essentially optimal as it nearly matches the standard non-private sample complexity in the agnostic PAC model. In an independent work, Dagan and Feldman (2019) further study the connections between uniform stability and differential privacy in the context of PAC learning, and give a new algorithm that yields a sample complexity bound of $\approx \mathsf{VC}(\mathcal{H})/\alpha^2 + \mathsf{VC}(\mathcal{H})^2/\epsilon\,\alpha$ in the single-query setting. Their bound exhibits the optimal dependence on ϵ, and when $\mathsf{VC}(\mathcal{H}) \leq \epsilon/\alpha$, it is tighter than our bound by a factor of $1/\epsilon$. They also give a new, simpler algorithm based on connections to uniform stability that yields the same bound as ours in the single-query setting.

Prior to the work of Bassily et al. (2018); Dwork and Feldman (2018), there have been several works that considered similar problem settings, (e.g. Hamm et al., 2016; Papernot et al., 2017, 2018). The last two references gave different algorithms and offered extensive empirical evaluation, however, they do not provide any formal accuracy guarantees.

2. Preliminaries

Notation: For classification tasks we denote the space of feature vectors by $\mathcal{X}$, the set of labels by $\mathcal{Y}$, and the data universe by $U = \mathcal{X} \times \mathcal{Y}$. A function $h : \mathcal{X} \to \mathcal{Y}$ is called a hypothesis and it labels data points in the feature space $\mathcal{X}$ by either 0 or 1 i.e. $\mathcal{Y} = \{0, 1\}$. A set of hypotheses $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ is called a hypothesis class. The VC dimension of $\mathcal{H}$ is denoted by $\mathsf{VC}(\mathcal{H})$. We use $\mathcal{D}$ to denote a distribution defined over $U = \mathcal{X} \times \mathcal{Y}$, and $\mathcal{D}_{\mathcal{X}}$ to denote the marginal distribution over $\mathcal{X}$. A sample dataset of n i.i.d. draws from $\mathcal{D}$ is denoted by $S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$.

Expected error: The expected error of a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ with respect to a distribution $\mathcal{D}$ over U is defined by $\mathsf{err}(h; \mathcal{D}) \triangleq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\mathbf{1}(h(x) \neq y)]$. The excess expected error is defined as $\mathsf{err}(h; \mathcal{D}) - \min_{h\in\mathcal{H}} \mathsf{err}(h; \mathcal{D})$.

Empirical error: The empirical error of a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ with respect to a labeled set S is denoted by $\widehat{\mathsf{err}}(h; S) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(h(x_i) \neq y_i)$.

The problem of minimizing the empirical error on a dataset is known as Empirical Risk Minimization (ERM). We use h_S^{ERM} to denote the hypothesis that minimizes the empirical error with respect to a dataset S, $h_S^{\mathsf{ERM}} \triangleq \arg\min_{h\in\mathcal{H}} \widehat{\mathsf{err}}(h; S)$.

Expected disagreement: The expected disagreement between a pair of hypotheses h_1 and h_2 with respect to a distribution $\mathcal{D}_{\mathcal{X}}$ is defined as $\mathsf{dis}(h_1, h_2; \mathcal{D}_{\mathcal{X}}) \triangleq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathbf{1}(h_1(x)) \neq h_2(x))]$.

Empirical disagreement: The empirical disagreement between a pair of hypotheses h_1 and h_2 w.r.t. an unlabeled dataset $S_u = \{x_1, \ldots, x_n\}$ is defined as $\widehat{\mathsf{dis}}(h_1, h_2; S_u) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(h_1(x_i)) \neq h_2(x_i))$.

Realizable setting: In the realizable setting of the PAC model, there exists a $h^* \in \mathcal{H}$ such that $\mathsf{err}(h^*; \mathcal{D}) = 0$ i.e., the true labeling function is assumed to be in $\mathcal{H}$. In this case, the distribution $\mathcal{D}$ can be described by $\mathcal{D}_{\mathcal{X}}$ and the hypothesis $h^* \in \mathcal{H}$. Such a distribution $\mathcal{D}$ is called *realizable* by $\mathcal{H}$. Hence, for realizable distributions, the expected error of a hypothesis h will be denoted as $\mathsf{err}(h; (\mathcal{D}_{\mathcal{X}}, h^*)) \triangleq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathbf{1}(h(x) \neq h^*(x))]$.

Definition 1 (Differential Privacy (Dwork et al., 2006a,b)) *Let $\epsilon, \delta > 0$. A (randomized) algorithm $M : U^n \to \mathcal{R}$ is (ϵ, δ)-differentially private if for all pairs of datasets $S, S' \in U^n$ that differs in exactly one data point, and every measurable $\mathcal{O} \subseteq \mathcal{R}$, with probability at least $1 - \delta$ over the coin flips of M, we have:*

$$\Pr\left(M(S) \in \mathcal{O}\right) \leq e^{\epsilon} \cdot \Pr\left(M(S') \in \mathcal{O}\right) + \delta.$$

We study private classification algorithms that take as input a private labeled dataset $S \sim \mathcal{D}^n$, and a sequence of classification queries $\mathcal{Q} = (\tilde{x}_1, \ldots, \tilde{x}_m) \sim \mathcal{D}_{\mathcal{X}}^m$, defined by m unlabeled feature-vectors from $\mathcal{X}$, (where m is an input parameter), and output a corresponding sequence of predictions, i.e., labels, $(\hat{y}_1, \ldots, \hat{y}_m)$. Here, we assume that the classification queries come one at a time and the algorithm is required to generate a label for the current query before seeing and responding to the next query. The goal is: i) after answering m queries the algorithm should satisfy (ϵ, δ)-differential privacy, and ii) the labels generated should be (α, β)-accurate with respect to a hypothesis class $\mathcal{H}$: a notion of accuracy which we formally define shortly. We give a generic description of the above classification paradigm in Algorithm 1 below (denoted as $\mathcal{A}_{\mathsf{PrivClass}}$).

Algorithm 1 $\mathcal{A}_{\mathsf{PrivClass}}$: Private Classification-Query Release Algorithm

Input: Private dataset: $S \in (\mathcal{X} \times \mathcal{Y})^n$, upper bound on the number of queries: m, online sequence of classification queries: $\mathcal{Q} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_m)$, hypothesis class: $\mathcal{H}$, privacy parameters $\epsilon, \delta > 0$, accuracy: α, and failure probability: β

1: **for** $j = 1, \ldots, m$ **do**

2: $\quad \hat{y}_j \leftarrow \mathsf{PrivLabel}(S, \mathcal{H}, \{(\tilde{x}_i, \hat{y}_i)\}_{i=1}^{j-1}, \tilde{x}_j)$ {Generic procedure that, given $S, \mathcal{H}$, the history $\{(\tilde{x}_i, \hat{y}_i)\}_{i=1}^{j-1}$, and the current query $\tilde{x}_j$, generates a label $\hat{y}_j$}

3: $\quad$ Output $\hat{y}_j$

The algorithm $\mathcal{A}_{\mathsf{PrivClass}}$ invokes a procedure PrivLabel, which is a generic classification procedure that given the input private training set S, the knowledge of hypothesis class $\mathcal{H}$, and the previous queries and outputs, it generates a label for an input query (feature-vector) $\tilde{x} \in \mathcal{X}$.

Definition 2 ($(\epsilon, \delta, \alpha, \beta, n, m)$-Private Classification-Query Release Algorithm) *Let $\mathcal{H}$ be a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$. Let $\epsilon, \delta, \alpha, \beta \in (0,1)$. A randomized algorithm $\mathcal{A}$ (whose generic format is described in Algorithm 1) is said to be an $(\epsilon, \delta, \alpha, \beta, n, m)$-PCQR (private classification-query release) algorithm for $\mathcal{H}$, if the following conditions hold:*

1. *For any sequence $\mathcal{Q} \in \mathcal{X}^m$, $\mathcal{A}$ is (ϵ, δ)-differentially private with respect to its input dataset.*
2. *For every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, given a dataset $S \sim \mathcal{D}^n$ and a sequence $V \triangleq ((\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_m, \tilde{y}_m)) \sim \mathcal{D}^m$ (where $\tilde{x}_i$'s are the queried feature-vectors in $\mathcal{Q}$ and $\tilde{y}_i$'s are their true hidden labels), $\mathcal{A}$ is (α, β)-accurate with respect to $\mathcal{H}$, where our notion of (α, β)-accuracy is defined as follows: With probability at least $1 - \beta$ over the choice of S, V, and the internal randomness in* PrivLabel *(Step 2 in Algorithm 1), we have*

$$\frac{1}{m} \sum_{j=1}^{m} \mathbf{1}(\hat{y}_j \neq \tilde{y}_j) \leq \alpha + \gamma,$$

where $\gamma \triangleq \min_{h \in \mathcal{H}} \mathsf{err}(h; \mathcal{D})$.

In the realizable setting, we have an analogous definition where $\gamma = 0$. In this case, we say that the algorithm is a PCQR *algorithm for $\mathcal{H}$ in the realizable setting.*

2.1. Previous work on private classification-query release (Bassily et al., 2018)

Bassily et al. (2018) give a construction for a PCQR algorithm (referred to as $\mathcal{A}_{\mathsf{SubSamp}}$), which combines the sub-sample-aggregate framework (Nissim et al., 2007; Smith and Thakurta, 2013) with the sparse-vector technique (Dwork and Roth, 2014). They provide formal privacy and accuracy guarantees with sample complexity bounds for $\mathcal{A}_{\mathsf{SubSamp}}$ in both the realizable and agnostic settings of the PAC model. As in the sparse-vector technique, one important input parameter to $\mathcal{A}_{\mathsf{SubSamp}}$ is cut-off paramter T, which gives bound on the number of the so-called "unstable queries" that $\mathcal{A}_{\mathsf{SubSamp}}$ can answer before the privacy budget is consumed. We formally describe $\mathcal{A}_{\mathsf{SubSamp}}$ and the notion of "unstable queries" in Appendix B for completeness. Here, we restate the results by Bassily et al. (2018) for the realizable and agnostic settings.

Lemma 3 (Realizable Setting: follows from Theorems 3.2 & 3.4, (Bassily et al., 2018)) *Let $\epsilon, \delta > 0$ and, $\alpha, \beta \in (0,1)$. Let $\mathcal{H}$ be a hypothesis class with $\mathsf{VC}(\mathcal{H}) = d$. Suppose that $\mathcal{B}$ in $\mathcal{A}_{\mathsf{SubSamp}}$ is a PAC learner for $\mathcal{H}$. Let $\mathcal{D}$ be any distribution over U that is realizable by $\mathcal{H}$. There is a setting for the cut-off parameter $T = \max\left(1,\ \tilde{O}\left(m\,\alpha\right)\right)$ such that $\mathcal{A}_{\mathsf{SubSamp}}$ is an $(\epsilon, \delta, \alpha, \beta, n, m)$-PCQR algorithm for $\mathcal{H}$ in the realizable setting where the private sample size is $n = \tilde{O}\left(\frac{d}{\epsilon\,\alpha} \cdot \max\left(1, \sqrt{m\,\alpha}\right)\right)$.*

In the agnostic setting, the accuracy guarantee of (Bassily et al., 2018) is not compatible with Definition 2; the accuracy guarantee therein has a sub-optimal dependency on the approximation error, γ (where $\gamma \triangleq \min_{h \in \mathcal{H}} \mathsf{err}(h; \mathcal{D})$). In particular, their result entails a blow-up in γ by a constant factor (≈ 3). This significantly limit the applicability of this result in scenarios where $\gamma \gg \alpha$.

Lemma 4 (Agnostic Setting: follows from Theorems 3.2 & 3.5, (Bassily et al., 2018)) *Let $\epsilon, \delta, \alpha, \beta \in (0,1)$. Let $\mathcal{H}$ be a hypothesis class with $\mathsf{VC}(\mathcal{H}) = d$. Suppose $\mathcal{B}$ in $\mathcal{A}_{\mathsf{SubSamp}}$ is an agnostic PAC learner for $\mathcal{H}$. Let $\mathcal{D}$ be any distribution over U, and let $\gamma \triangleq \min_{h \in \mathcal{H}} \mathsf{err}(h; \mathcal{D})$. Let $S \sim \mathcal{D}^n$ denote the input private sample to $\mathcal{A}_{\mathsf{SubSamp}}$. Let $V \triangleq ((\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_m, \tilde{y}_m)) \sim \mathcal{D}^m$, where $\tilde{x}_i$'s are the queried feature-vectors in $\mathcal{Q}$ and $\tilde{y}_i$'s are their true (hidden) labels. Let $(y_1^{\mathsf{priv}}, \ldots, y_m^{\mathsf{priv}})$ denote the output labels of $\mathcal{A}_{\mathsf{SubSamp}}$. There is a setting for the cut-off parameter $T = \max\left(1,\ \tilde{O}\left(m\,(\alpha + \gamma)\right)\right)$ such that: 1) $\mathcal{A}_{\mathsf{SubSamp}}$ is (ϵ, δ)-differentially private with respect to the input training set; 2) when the private sample is of size $n = \tilde{O}\left(\frac{d}{\epsilon\,\alpha^2} \cdot \max\left(1, \sqrt{m\,\alpha}\right)\right)$, then with probability at least $1 - \beta$ over S, V and the randomness in $\mathcal{A}_{\mathsf{SubSamp}}$, we have:*

$$\frac{1}{m} \sum_{j=1}^{m} \mathbf{1}(y_j^{\mathsf{priv}} \neq \tilde{y}_j) \leq \alpha + 3\gamma.$$

3. Private Release of Classification Queries in the Agnostic PAC Setting

In this section we give an improved construction for the private classification-query release algorithm of Bassily et al. (2018) in the agnostic setting. Our construction can privately answer up to m queries with excess classification error α, and input sample size $\tilde{O}\left(\frac{\mathsf{VC}(\mathcal{H})}{\epsilon\,\alpha^2} \cdot \max\left(1, \sqrt{m}\,\alpha^{3/2}\right)\right)$, (where $\tilde{O}$ hides $\log$ factors of $m, \frac{1}{\alpha}, \frac{1}{\delta}, \frac{1}{\beta}$). Comparing to the result by Bassily et al. (2018) for the agnostic setting, where the private sample size is $\approx \frac{\mathsf{VC}(\mathcal{H})}{\epsilon\,\alpha^2} \cdot \max(1, \sqrt{m\alpha})$ (Lemma 4), our sample complexity bound is tighter by a factor of $\approx \sqrt{m\alpha}$ when $\frac{1}{\alpha} \leq m < \frac{1}{\alpha^3}$, and it is tighter by a factor of $\approx \frac{1}{\alpha}$ when $m \geq \frac{1}{\alpha^3}$.

Overview

Our construction is made up of two phases. The first phase is a pre-processing phase in which a subset S', of the input private sample S, is *relabeled* using a "good" hypothesis $\hat{h} \in \mathcal{H}$ to obtain a new sample S''. This

phase is a reenactment of the elegant technique due to Beimel et al. (2015), which was called *LabelBoost Procedure* therein. By construction $\hat{h}$ is chosen such that its empirical error is close to that of the ERM hypothesis. Hence, we can formally show that when input sample size is sufficiently large, $\hat{h}$ attains low excess error. Moving forward, one may view $\hat{h}$ as if it is the true labeling hypothesis, and hence the agnostic setting can be reduced to the realizable setting. In Section 3.1, we describe this pre-processing phase and state its guarantees.

Now as we reduced the problem to the realizable setting, in the next phase we invoke the techniques of (Bassily et al., 2018). In the second phase, the relabeled training set S'' is used to provide input training examples to $\mathcal{A}_{\mathsf{SubSamp}}$ (described in Section 2.1). Note that S'' is no longer i.i.d. from the original distribution. We form a new dataset $\widehat{S}$ by sampling data points uniformly with replacement from S'' and then feed $\widehat{S}$ to $\mathcal{A}_{\mathsf{SubSamp}}$ as input. This new training set $\widehat{S}$ is now i.i.d. from the empirical distribution of S''. Via a uniform-convergence argument (see Claim 10), we can show that that this re-sampling step does not impact our desired accuracy guarantees. We also need to carefully calibrate the privacy parameters of $\mathcal{A}_{\mathsf{SubSamp}}$ to take into account the fact that $\widehat{S}$ may contain repetitions of the elements in S''. Algorithm $\mathcal{A}_{\mathsf{SubSamp}}$ uses $\widehat{S}$ to privately generate labels for an online sequence of classification queries. We formally show that for any setting of the target parameters (accuracy, privacy, and total number of queries), there is a sufficient size for the original input sample S such that our construction attains the desired accuracy and privacy guarantees w.r.t. the entire sequence of queries. We formally describe our construction and provide formal analysis for its privacy and accuracy guarantees in Section 3.2.

3.1. From the agnostic to the realizable setting: A generic reduction

In this section, we describe the pre-processing procedure, denoted as $\mathcal{A}_{\mathsf{Relabel}}$ (given by Algorithm 2 below), which follows from the relabeling technique devised by Beimel et al. (2015).

The algorithm $\mathcal{A}_{\mathsf{Relabel}}$ operates on a private labeled dataset $S' \sim \mathcal{D}^{n'}$ and on a hypothesis class $\mathcal{H}$. Let S'_u denote the unlabeled version of S', i.e., $S'_u = \{x_1, \ldots, x_{n'}\}$, and $\prod_{\mathcal{H}}(S'_u)$ denote the set of all possible dichotomies that can be generated by $\mathcal{H}$ on the set S'_u. First the algorithm chooses a finite subset $\widetilde{H}$ of $\mathcal{H}$ such that each dichotomy in $\prod_{\mathcal{H}}(S'_u)$ is represented by one of the hypotheses in $\widetilde{H}$. Note that by Sauer's lemma (see Sauer, 1972), the size of $\widetilde{H}$ is $O\left((n'/d)^d\right)$, where $d = \mathsf{VC}(\mathcal{H})$. Next, $\mathcal{A}_{\mathsf{Relabel}}$ chooses a hypothesis $\widehat{h}$ using the exponential mechanism with privacy parameter $\tilde{\epsilon} = 1$ and a score function $q(S', h) = -\widehat{\mathsf{err}}(h; S')$. Finally, $\mathcal{A}_{\mathsf{Relabel}}$ uses $\widehat{h}$ to rebalel S'_u, and outputs this labeled set S''.

Algorithm 2 $\mathcal{A}_{\mathsf{Relabel}}$: Relabel Procedure

Input: Private dataset: $S' \in (\mathcal{X} \times \mathcal{Y})^{n'}$, a hypothesis class: $\mathcal{H}$

1: $\widetilde{H} \leftarrow \emptyset$
2: Let $S'_u = \{x_1, \ldots, x_{n'}\}$ be the unlabeled version of S'.
3: For every $(y_1, \ldots, y_{n'}) \in \prod_{\mathcal{H}}(S'_u) = \{(h(x_1), \ldots, h(x_{n'})) : h \in \mathcal{H}\}$, add to $\widetilde{H}$ any arbitrary hypothesis $h \in \mathcal{H}$ s.t. $h(x_i) = y_i, \forall i \in [n']$.
4: Use the exponential mechanism with inputs S', $\widetilde{H}$, privacy parameter $\tilde{\epsilon} = 1$, and a score function $q(S', h) \triangleq -\widehat{\mathsf{err}}(h; S')$ to select $\widehat{h}$ from $\widetilde{H}$.
5: Relabel S'_u using $\widehat{h}$, and denote this relabeled dataset as S''.
6: Output S''.

The following lemmas give the privacy and accuracy guarantees of $\mathcal{A}_{\mathsf{Relabel}}$. Lemma 5 follows directly from (Beimel et al., 2015). We prove Lemma 6 in Appendix A.

Lemma 5 (Lemma 4.1 in (Beimel et al., 2015) restated) *Let $\mathcal{A}$ be an $(1,\delta)$- differentially private algorithm. Let $\mathcal{B}$ be an algorithm that on input dataset S' invokes $\mathcal{A}$ on the output of $\mathcal{A}_{\mathsf{Relabel}}(S',\mathcal{H})$. Then, $\mathcal{B}$ is $(4,4e\delta)$- differentially private.*

Lemma 6 *Let $\mathcal{H}$ be a hypothesis class with $\mathsf{VC}(\mathcal{H}) = d$. Let $\alpha,\beta \in (0,1)$. Let $\mathcal{D}$ be an arbitrary distribution over $\mathcal{X}\times\mathcal{Y}$, and $S' \sim \mathcal{D}^{n'}$ be an input dataset to $\mathcal{A}_{\mathsf{Relabel}}$, where $n' \geq 256\frac{(d+\log(3/\beta))}{\alpha^2}$. With probability at least $1-\beta$, hypothesis $\widehat{h}$ (generated in Step 4 of $\mathcal{A}_{\mathsf{Relabel}}$) satisfies the following:*

$$\mathsf{err}\left(\widehat{h};\mathcal{D}\right) - \mathsf{err}\left(h^{\mathsf{ERM}}_{S'};\mathcal{D}\right) \leq \alpha,$$

where $h^{\mathsf{ERM}}_{S'}$ is the ERM hypothesis w.r.t. the input sample S'.

3.2. A Private Classification-Query Release Algorithm

In this section, we describe our PCQR algorithm $\mathcal{A}_{\mathsf{AgPrivCl}}$ (Algorithm 3 below) that combines the two techniques given by $\mathcal{A}_{\mathsf{Relabel}}$, and $\mathcal{A}_{\mathsf{SubSamp}}$. As a PCQR algorithm, $\mathcal{A}_{\mathsf{AgPrivCl}}$ takes as input: a private dataset $S \sim \mathcal{D}^n$, the number of queries m, an online sequence of classification queries $\mathcal{Q} = (\tilde{x}_1,\ldots,\tilde{x}_m) \sim \mathcal{D}^m_{\mathcal{X}}$, a hypothesis class $\mathcal{H}$, as well as the desired privacy and accuracy parameters. Together with these, $\mathcal{A}_{\mathsf{AgPrivCl}}$ also has oracle access to a PAC learner $\mathcal{B}_{\mathsf{PAC}}$ for $\mathcal{H}$. First, we randomly sample a subset S' of S of size n', where $n' \approx \epsilon n$, and invoke $\mathcal{A}_{\mathsf{Relabel}}$ on S' and $\mathcal{H}$. This sampling step is used to boost the privacy guarantee of $\mathcal{A}_{\mathsf{AgPrivCl}}$. Note that, dataset S'' (output by $\mathcal{A}_{\mathsf{Relabel}}$) is relabeled using hypothesis $\widehat{h} \in \mathcal{H}$. In order to ensure that our input to the next stage is i.i.d., we sample n' points uniformly with replacement from S'' to form a new dataset $\widehat{S}$ (i.e., $\widehat{S}$ is made up of n' i.i.d. draws from the empirical distribution of S''). Next, we invoke $\mathcal{A}_{\mathsf{SubSamp}}$ in the realizable setting on the dataset $\widehat{S}$, m, $\mathcal{Q}$, and $\mathcal{B}_{\mathsf{PAC}}$ as inputs. We set the cut-off parameter of $\mathcal{A}_{\mathsf{SubSamp}}$ as $T = \max\left(1,\ \tilde{O}(m\alpha)\right)$, where α is the accuracy parameter of $\mathcal{B}_{\mathsf{PAC}}$. The privacy parameters to $\mathcal{A}_{\mathsf{SubSamp}}$ are set to $(\hat{\epsilon},\hat{\delta})$ defined in Step 1 of $\mathcal{A}_{\mathsf{AgPrivCl}}$. This is needed to ensure (ϵ,δ)-differential privacy for the entire construction. Finally, we output the sequence of private labels $\{y^{\mathsf{priv}}_1,\ldots,y^{\mathsf{priv}}_m\}$ generated by $\mathcal{A}_{\mathsf{SubSamp}}$ for the input sequence of queries.

Algorithm 3 $\mathcal{A}_{\mathsf{AgPrivCl}}$: Private Agnostic-PAC Classification-Query Release Algorithm

Input: Private dataset: $S \in (\mathcal{X}\times\mathcal{Y})^n$, upper bound on the number of queries: m, online sequence of classification queries: $\mathcal{Q} = (\tilde{x}_1,\tilde{x}_2,\ldots,\tilde{x}_m)$, a hypothesis class: $\mathcal{H}$, oracle access to non-private learner: $\mathcal{B}_{\mathsf{PAC}}$ for $\mathcal{H}$, privacy parameters: $\epsilon,\delta > 0$, accuracy parameter: α, and, failure probability: β

1: $n' \leftarrow \frac{\epsilon}{56}n$, $\quad T \leftarrow \max\left(1,\ \frac{1}{8}\,m\alpha + \frac{1}{4}\sqrt{3\,m\alpha\log\left(\frac{m}{\beta}\right)}\right)$,
$\epsilon' \leftarrow \alpha\max\left(1,\sqrt{m\alpha}\right)$, $\quad \hat{\epsilon} \leftarrow \frac{1}{\log(2/\delta)}\min\left(1,\ \epsilon'\right)$, $\quad \hat{\delta} \leftarrow \frac{\delta}{2\ e^{\min(1,\ \epsilon')}\ \log(2/\delta)}$

2: Uniformly sample without replacement a subset S' of n' data points from S

3: $S'' \leftarrow \mathcal{A}_{\mathsf{Relabel}}(S',\mathcal{H})$.

4: $\widehat{S} \leftarrow$ Uniformly sample n' points from S'' with replacement.

5: Output $(y^{\mathsf{priv}}_1,\ldots,y^{\mathsf{priv}}_m) \leftarrow \mathcal{A}_{\mathsf{SubSamp}}(\widehat{S},m,\mathcal{Q},\mathcal{B}_{\mathsf{PAC}},T,\hat{\epsilon},\hat{\delta},\beta)$.

We formally state the main result of this section in the following theorem.

Theorem 7 *Let $\mathcal{H}$ be a hypothesis class with $\mathsf{VC}(\mathcal{H}) = d$. For any $\epsilon,\delta,\alpha,\beta \in (0,1)$, $\mathcal{A}_{\mathsf{AgPrivCl}}$ (Algorithm 3) is an $(\epsilon,\delta,\alpha,\beta,n,m)$-PCQR algorithm for $\mathcal{H}$, where private sample size*

$$n = O\left(\frac{\left(d\,\log\left(\frac{1}{\alpha}\right) + \log\left(\frac{m}{\beta}\right)\right)\,\log^{3/2}\left(\frac{2}{\delta}\right)\,\log\left(\frac{m\alpha}{\min(\delta,\beta/2)}\right)}{\epsilon\,\alpha^2}\cdot\max\left(1,\sqrt{m}\,\alpha^{3/2}\right)\right),$$

and number of queries $m = \Omega\left(\frac{\log(1/\alpha\beta)}{\alpha}\right)$.

We will prove the theorem via the following lemmas that establish the privacy and accuracy guarantees of $\mathcal{A}_{\mathsf{AgPrivCl}}$.

Lemma 8 (Privacy Guarantee of $\mathcal{A}_{\mathsf{AgPrivCl}}$) *$\mathcal{A}_{\mathsf{AgPrivCl}}$ is (ϵ, δ)-differentially private (with respect to its input dataset).*

Proof Fix the randomness in dataset S' due to sampling in Step 2 of $\mathcal{A}_{\mathsf{AgPrivCl}}$. Let $\mathcal{R}(\cdot)$ denote the uniform sampling procedure in Step 4 in $\mathcal{A}_{\mathsf{AgPrivCl}}$; that is, Step 4 can be written as $\widehat{S} \leftarrow \mathcal{R}(S'')$. Note that Steps 4-5 in $\mathcal{A}_{\mathsf{AgPrivCl}}$ can now be expressed as a composition $\mathcal{R} \circ \mathcal{A}_{\mathsf{SubSamp}}$, where $\mathcal{R} \circ \mathcal{A}_{\mathsf{SubSamp}}(\cdot) \triangleq \mathcal{A}_{\mathsf{SubSamp}}(\mathcal{R}(\cdot))$.

Let $\epsilon^* = \min(1,\ \epsilon')$, where $\epsilon' = \alpha \max(1, \sqrt{m\alpha})$ (as defined in Step 1 of $\mathcal{A}_{\mathsf{AgPrivCl}}$). Note that the input to $\mathcal{R} \circ \mathcal{A}_{\mathsf{SubSamp}}$ is the dataset S'', which is the output of $\mathcal{A}_{\mathsf{Relabel}}$. Note that if we can show that $\mathcal{R} \circ \mathcal{A}_{\mathsf{SubSamp}}$ is (ϵ^*, δ)-differentially private, then, from Lemma 5, it follows that $\mathcal{A}_{\mathsf{Relabel}} \circ \mathcal{R} \circ \mathcal{A}_{\mathsf{SubSamp}}$ is $(\epsilon^* + 3, 4e\delta)$-differentially private. Next, by taking into account account the randomness due to sampling in Step 2, then by privacy amplification via sampling (Kasiviswanathan et al., 2008; Li et al., 2012), it follows that $\mathcal{A}_{\mathsf{AgPrivCl}}$ is (ϵ, δ)-differentially private. Hence, it remains to show that $\mathcal{R} \circ \mathcal{A}_{\mathsf{SubSamp}}$ is (ϵ^*, δ)-differentially private with respect to S''.

Let S''_1 and S''_2 be neighboring datasets. W.l.o.g., assume that S''_1 and S''_2 differ in index $j \in [n']$. Let r be the number of times the j-th index is sampled by $\mathcal{R}$. By the definition of $\mathcal{R}$, and Chernoff bound, w.p. $\geq 1 - \delta/2$, we have $r \leq \log(2/\delta)$.

Using the result in (Theorem 3.1 Bassily et al., 2018), $\mathcal{A}_{\mathsf{SubSamp}}$ is $(\hat{\epsilon}, \hat{\delta})$-differentially private with respect to $\widehat{S}$. Conditioned on $r \leq \log(\frac{2}{\delta})$ and by the notion of group privacy we have, $\mathcal{R} \circ \mathcal{A}_{\mathsf{SubSamp}}$ is $(r\hat{\epsilon}\ ,\ re^{r\hat{\epsilon}}\ \hat{\delta})$-differentially private. Hence, by the above high probability bound on the event $r \leq \log(\frac{2}{\delta})$, we conclude that $\mathcal{R} \circ \mathcal{A}_{\mathsf{SubSamp}}$ is $(\min(1,\ \epsilon')\,, \delta)$-differentially private.

■

Lemma 9 (Accuracy Guarantee of $\mathcal{A}_{\mathsf{AgPrivCl}}$) *Let $\mathcal{H}$ be a hypothesis class with $\mathsf{VC}(\mathcal{H}) = d$. Let $\mathcal{B}_{\mathsf{PAC}}$ (invoked by $\mathcal{A}_{\mathsf{SubSamp}}$) be a PAC learner for $\mathcal{H}$ (in the realizable setting). Let $\mathcal{D}$ be any distribution over $\mathcal{X} \times \mathcal{Y}$, and let $\gamma \triangleq \min_{h \in \mathcal{H}} \mathsf{err}(h; \mathcal{D})$. Let $S \sim \mathcal{D}^n$ denote the input private sample to $\mathcal{A}_{\mathsf{AgPrivCl}}$, where*

$$n = O\left(\frac{\left(d\,\log\left(\frac{1}{\alpha}\right) + \log\left(\frac{m}{\beta}\right)\right)\,\log^{3/2}\left(\frac{2}{\delta}\right)\,\log\left(\frac{m\alpha}{\min(\delta,\beta/2)}\right)\max\left(1, \sqrt{m}\,\alpha^{3/2}\right)}{\epsilon\,\alpha^2}\right),$$

and $m \geq 8\,\frac{\log(1/\alpha\beta)}{\alpha}$. Let $(\tilde{y}_1, \ldots, \tilde{y}_m)$ denote the corresponding true (hidden) labels for $\mathcal{Q}$. Then, w.p. at least $1 - \beta$ (over the choice of S, $\mathcal{Q}$, and the randomness in $\mathcal{A}_{\mathsf{AgPrivCl}}$), we have:

$$\frac{1}{m}\sum_{j=1}^{m} \mathbf{1}(y_j^{\mathsf{priv}} \neq \tilde{y}_j) \leq \alpha + \gamma.$$

In the proof of Lemma 9 we will use the following claim. We defer its proof after the proof of the lemma.

Claim 10 *Let $\mathcal{H}$ be a hypothesis class with $\mathsf{VC}(\mathcal{H}) = d$. Let S_u be an an unlabeled training set of size n_o, where $n_o \geq 50\frac{d\,\log(1/\alpha)+\log(1/\beta')}{\alpha^2}$. Then, with probability at least $1 - \beta'$ for any $h_1, h_2 \in \mathcal{H}$, we have $\left|\mathsf{dis}(h_1, h_2; \mathcal{D}_{\mathcal{X}}) - \widehat{\mathsf{dis}}(h_1, h_2; S_u)\right| \leq \alpha$. (Recall that $\mathsf{dis}(h_1, h_2; \mathcal{D}_{\mathcal{X}})$ and $\widehat{\mathsf{dis}}(h_1, h_2; S_u)$ are the expected and empirical disagreement rates, respectively, as defined in Section 2.)*

Proof of Lemma 9 Consider the description of $\mathcal{A}_{\mathsf{Relabel}}$ in Algorithm 2. Note that, hypothesis $\widehat{h} \in \mathcal{H}$ selected in Step 4 of $\mathcal{A}_{\mathsf{Relabel}}$ is used to generate labels of S'' (output dataset of $\mathcal{A}_{\mathsf{Relabel}}$). Note that by choosing n to be sufficiently large, we can ensure that the size of S'' is given by

$$n' = 8000 \frac{\left(d \log\left(\frac{1}{\alpha}\right) + \log\left(\frac{m}{\beta}\right)\right) \text{polylog}\left(m, \frac{1}{\delta}, \frac{1}{\beta}\right)}{\alpha^2} \cdot \max\left(1, \sqrt{m}\,\alpha^{3/2}\right).$$

Let $\mathcal{D}_{S''}$ denote the empirical distribution induced by S''. Note that $\mathsf{err}(\widehat{h}; \mathcal{D}_{S''}) = 0$. In $\mathcal{A}_{\mathsf{AgPrivCl}}$, dataset $\widehat{S}$ (input to $\mathcal{A}_{\mathsf{SubSamp}}$) is created by n' i.i.d. draws from $\mathcal{D}_{S''}$.

From the description of $\mathcal{A}_{\mathsf{SubSamp}}$ (Algorithm 5), $\mathcal{A}_{\mathsf{SubSamp}}$ splits $\widehat{S}$ into k equal-sized sub-samples, where $k = \tilde{O}\left(\frac{\sqrt{T}}{\hat{\epsilon}}\right)$. Here T is the input cut-off parameter of $\mathcal{A}_{\mathsf{SubSamp}}$ whose setting is given in Step 1 of $\mathcal{A}_{\mathsf{AgPrivCl}}$. Note that since $m = \Omega\left(\frac{\log(1/\alpha\beta)}{\alpha}\right)$, we have $T = O(m\alpha)$. Each sub-sample is then fed separately as an input to $\mathcal{B}_{\mathsf{PAC}}$. For each input sub-sample, $\mathcal{B}_{\mathsf{PAC}}$ outputs a classifier $h_j, j \in [k]$. Hence we end up with an ensemble of classifiers $h_1, \cdots, h_k$. Note that the size of the input sub-sample to $\mathcal{B}_{\mathsf{PAC}}$ is $\frac{n'}{k}$. Observe that

$$n' = 8000 \frac{\left(d \log\left(\frac{1}{\alpha}\right) + \log\left(\frac{m}{\beta}\right)\right) \log^{3/2}\left(\frac{2}{\delta}\right) \log\left(\frac{m\alpha}{\min(\delta, \beta/2)}\right)}{\alpha^2} \cdot \max\left(1, \sqrt{m}\,\alpha^{3/2}\right),$$

and the number of sub-samples k is set in Step 1 of $\mathcal{A}_{\mathsf{SubSamp}}$ as follows

$$k = O\left(\frac{\sqrt{m\alpha \log(\frac{2}{\delta})} \cdot \log\left(\frac{m\alpha}{\min(\delta, \beta/2)}\right)}{\hat{\epsilon}}\right)$$

Hence, using the setting of $\hat{\epsilon}$ in Step 1 of $\mathcal{A}_{\mathsf{AgPrivCl}}$, we have

$$\begin{aligned}\frac{n'}{k} &= \Omega\left(\frac{\left(d \log\left(\frac{1}{\alpha}\right) + \log\left(\frac{m}{\beta}\right)\right)}{\sqrt{m}\,\alpha^{5/2}} \cdot \min\left(1, \sqrt{m}\,\alpha^{3/2}\right) \cdot \max\left(1, \sqrt{m}\,\alpha^{3/2}\right)\right) \\ &= \Omega\left(\frac{\left(d \log\left(\frac{1}{\alpha}\right) + \log\left(\frac{m}{\beta}\right)\right)}{\alpha}\right) = \Omega\left(\frac{\left(d \log\left(\frac{1}{\alpha}\right) + \log\left(\frac{16k}{\beta}\right)\right)}{\alpha}\right).\end{aligned}$$

By standard results in learning theory, it is easy to see that the size of the input sub-sample to $\mathcal{B}_{\mathsf{PAC}}$ is sufficient for $\mathcal{B}_{\mathsf{PAC}}$ to PAC-learn $\mathcal{H}$ with respect to $\mathcal{D}_{S''}$ with accuracy $\frac{\alpha}{24}$ and confidence $\frac{\beta}{16k}$.

Fix any $j \in [k]$. Using the above fact about $\mathcal{B}_{\mathsf{PAC}}$, w.p. at least $1 - \frac{\beta}{16k}$, $\mathsf{err}(h_j; \mathcal{D}_{S''}) \leq \frac{\alpha}{24}$. Since $\mathcal{D}_{S''}$ is the empirical distribution of S'', equivalently, we have $\widehat{\mathsf{dis}}(h_j, \widehat{h}; S''_u) \leq \frac{\alpha}{24}$, where S''_u is the unlabeled version of S''. Note that the size of S'' is $n' \geq 7200 \frac{\left(d \log\left(\frac{1}{\alpha}\right) + \log\left(\frac{8k}{\beta}\right)\right)}{\alpha^2}$. Hence, by Claim 10, it follows that w.p. $\geq 1 - \frac{\beta}{8k}$, $\mathsf{dis}(h_j, \widehat{h}; \mathcal{D}_{\mathcal{X}}) \leq \frac{\alpha}{12}$. Equivalently, w.p. $\geq 1 - \frac{\beta}{8k}$, $\mathsf{err}\left(h_j; (\mathcal{D}_{\mathcal{X}}, \widehat{h})\right) \leq \frac{\alpha}{12}$.

From the above and the fact that the queries in $\mathcal{Q}$ are i.i.d. from $\mathcal{D}_{\mathcal{X}}$, we invoke the same counting argument in the proof of (Theorem 3.2 Bassily et al., 2018) to show that w.p. $\geq 1 - \frac{\beta}{4}$, the output labels of $\mathcal{A}_{\mathsf{SubSamp}}$ satisfy:

$$\frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\left(y_i^{\mathsf{priv}} \neq \widehat{h}(\tilde{x}_i)\right) \leq \frac{\alpha}{4}. \tag{1}$$

Let $h_{S'}^{\mathsf{ERM}}$ denote the ERM hypothesis with respect to the dataset S' constructed in Step 2 of $\mathcal{A}_{\mathsf{AgPrivCl}}$. Note that Lemma 6 implies that w.p. $\geq 1-\beta/4$, $\mathsf{err}(\widehat{h};\mathcal{D}) - \mathsf{err}\left(h_{S'}^{\mathsf{ERM}};\mathcal{D}\right) \leq \alpha/4$.

Since the queries and their true labels $\left((\tilde{x}_1,\tilde{y}_1),\ldots,(\tilde{x}_m,\tilde{y}_m)\right)$ are drawn i.i.d. from $\mathcal{D}$, then by Chernoff's bound and the fact that $m \geq 8\frac{\log(1/\beta)}{\alpha}$, we get that w.p. $\geq 1-\frac{\beta}{2}$,

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\left(\widehat{h}(\tilde{x}_i)\neq\tilde{y}_i\right) - \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\left(h_{S'}^{\mathsf{ERM}}(\tilde{x}_i)\neq\tilde{y}_i\right) \leq \frac{\alpha}{2}. \tag{2}$$

Moreover, from the bound on n' and using a basic fact from learning theory, w.p. $\geq 1-\beta/8$, the ERM hypothesis $h_{S'}^{\mathsf{ERM}}$ satisfies: $\mathsf{err}\left(h_{S'}^{\mathsf{ERM}};\mathcal{D}\right) \leq \alpha/8+\gamma$, where $\gamma = \min_{h\in\mathcal{H}}\mathsf{err}(h;\mathcal{D})$. Again, since $\left((\tilde{x}_1,\tilde{y}_1),\ldots,(\tilde{x}_m,\tilde{y}_m)\right)$ are i.i.d. from $\mathcal{D}$, then by Chernoff's bound and the fact that $m \geq 8\frac{\log(1/\beta)}{\alpha}$, w.p. $\geq 1-\beta/4$, we have

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\left(h_{S'}^{\mathsf{ERM}}(\tilde{x}_i)\neq\tilde{y}_i\right) \leq \frac{\alpha}{4}+\gamma. \tag{3}$$

Now, using (1), (2), and (3) together with a simple application of the triangle inequality and the union bound, we conclude that w.p. $\geq 1-\beta$, $\frac{1}{m}\sum_{j=1}^{m}\mathbf{1}\left(y_j^{\mathsf{priv}}\neq\tilde{y}_j\right) \leq \alpha+\gamma$. ■

Proof of Claim 10 For $S_u \sim \mathcal{D}_{\mathcal{X}}^{n_o}$, define the event

$$\mathsf{Bad} = \{\exists h_1,h_2\in\mathcal{H} : |\mathsf{dis}(h_1,h_2;\mathcal{D}_{\mathcal{X}}) - \widehat{\mathsf{dis}}(h_1,h_2;S_u)| > \alpha\}$$

We will show that $\mathbb{P}_{S_u\sim\mathcal{D}_{\mathcal{X}}^{n_o}}[\mathsf{Bad}] \leq 2\left(\frac{en_o}{d}\right)^{2d}\exp\left(-n_o\alpha^2/8\right)$. Hence, by using a standard manipulation, one can easily show that the right-hand side is bounded by β' when n_o is as given in the statement of the claim. Let $\mathcal{H}_\Delta$ be a hypothesis class defined as $\mathcal{H}_\Delta \triangleq \{h_1\Delta h_2 : h_1,h_2\in\mathcal{H}\}$, where $h_1\Delta h_2 : \mathcal{X}\to\{0,1\}$ is defined as: $\forall x\in\mathcal{X},\ h_1\Delta h_2(x) \triangleq \mathbf{1}(h_1(x)\neq h_2(x))$.

Let $\mathcal{G}_{\mathcal{H}_\Delta}$ denote the growth function of $\mathcal{H}_\Delta$; i.e. for any number t, $\mathcal{G}_{\mathcal{H}_\Delta}(t) \triangleq \max_{V:|V|=t}\left|\prod_{\mathcal{H}_\Delta}(V)\right|$, where $\prod_{\mathcal{H}_\Delta}(V)$ is the set of all dichotomies that can be generated by $\mathcal{H}_\Delta$ on a set V of size t. Now for any set V of size t, every dichotomy in $\prod_{\mathcal{H}_\Delta}(V)$ is determined by a pair of dichotomies in $\prod_{\mathcal{H}}(V)$, and thus we get $|\prod_{\mathcal{H}_\Delta}(V)| \leq |\prod_{\mathcal{H}}(V)|^2$. Hence, by Sauer's Lemma $\mathcal{G}_{\mathcal{H}_\Delta}(t) \leq \mathcal{G}_{\mathcal{H}}(t) \leq \left(\frac{et}{d}\right)^{2d}$. Let h_0 be the all-zero hypothesis. Note that $h_0\in\mathcal{H}_\Delta$. Now, using a standard VC-based uniform convergence argument we have,

$$\begin{aligned}
&\mathbb{P}_{S_u\sim\mathcal{D}_{\mathcal{X}}^{n_o}}\left[\exists h_1,h_2\in\mathcal{H} : |\mathsf{dis}(h_1,h_2;\mathcal{D}_{\mathcal{X}}) - \widehat{\mathsf{dis}}(h_1,h_2;S_u)| > \alpha\right] \\
&\leq \mathbb{P}_{S_u\sim\mathcal{D}_{\mathcal{X}}^{n_o}}\left[\exists h\in\mathcal{H}_\Delta : |\mathsf{dis}(h,h_0;\mathcal{D}_{\mathcal{X}}) - \widehat{\mathsf{dis}}(h,h_0);S_u)| > \alpha\right] \\
&\leq 2\mathcal{G}_{\mathcal{H}_\Delta}\exp\left(-n_o\alpha^2/8\right) \leq 2\left(\frac{en_o}{d}\right)^{2d}\exp\left(-n_o\alpha^2/8\right)
\end{aligned}$$

Note that the first inequality in the third line is non-trivial, and is used unanimously in VC-based uniform convergence bounds (see e.g. Shalev-Shwartz and Ben-David, 2014). ■

4. Privately Answering Any Number of Classification Queries

In this section, we describe an universal PCQR algorithm that can answer *any* number of queries with private sample size that is independent of the number of queries. The main idea is that after answering a number

of queries $\approx \frac{\mathsf{VC}(\mathcal{H})}{\alpha}$, we can use the feature-vectors defining those queries as an auxiliary "public" dataset. Recall that as defined earlier in our problem statement, the set of queries themselves do not entail any privacy constraints. We can then invoke the framework of *semi-private learning* introduced by Beimel et al. (2013), where such auxiliary public dataset can be exploited to finally generate a classifier that is safe to publish. In particular, a semi-private learner takes as input two types of datasets: a private labeled dataset, and another auxiliary public dataset. The algorithm needs to satisfy differential privacy only with respect to the private dataset. Alon et al. (2019) describe a construction of a semi-private learner (referred to as $\mathcal{A}_{\mathsf{SSPP}}$), and show that it suffices to have a public unlabeled dataset of size $\approx \frac{\mathsf{VC}(\mathcal{H})}{\alpha}$ to privately learn any hypothesis class $\mathcal{H}$ with excess error α in the *agnostic* setting. In particular, for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, given a set of feature-vectors of size $\approx \frac{\mathsf{VC}(\mathcal{H})}{\alpha}$ drawn i.i.d. from $\mathcal{D}_{\mathcal{X}}$, and a private labeled training set of $\approx \frac{\mathsf{VC}(\mathcal{H})}{\epsilon\ \alpha^2}$ drawn i.i.d. from $\mathcal{D}$, $\mathcal{A}_{\mathsf{SSPP}}$ outputs a classifier $h_{\mathsf{priv}} \in \mathcal{H}$ such that $\mathsf{err}(h_{\mathsf{priv}}; \mathcal{D}) - \min_{h \in \mathcal{H}} \mathsf{err}(h; \mathcal{D}) \leq \alpha$ w.r.t $\mathcal{D}$. Hence, $\mathcal{A}_{\mathsf{SSPP}}$ outputs a classifier that can be used to answer any number of subsequent classification queries. For the sake of completeness, we give a formal description of $\mathcal{A}_{\mathsf{SSPP}}$ in Appendix B.

Using this result, we can extend our construction in Section 3.2 to allow for privately answering any number of classification queries using a private training set whose size is independent of the number of queries. In Algorithm 4 below (denoted as $\mathcal{A}_{\mathsf{UnvPrivCl}}$), we describe our universal PCQR algorithm.

Algorithm 4 $\mathcal{A}_{\mathsf{UnvPrivCl}}$: Universal Private Classification-Query Release Algorithm

Input: Private dataset: $S \in U^n$, number of queries: m, online sequence of classification queries: $\mathcal{Q} = (\tilde{x}_1, \ldots, \tilde{x}_m)$, hypothesis class: $\mathcal{H}$, oracle access to a non-private PAC learner for $\mathcal{H}$: $\mathcal{B}_{\mathsf{PAC}}$, privacy parameters $\epsilon, \delta > 0$, accuracy parameter α, and failure probability β.

1: $m_o \leftarrow 32\, \frac{d \log(1/\alpha) + \log(1/\beta)}{\alpha}$, $m' \leftarrow \min(m_o, m)$

2: Output $(y_1^{\mathsf{priv}}, \ldots, y_{m'}^{\mathsf{priv}}) \leftarrow \mathcal{A}_{\mathsf{AgPrivCl}}\Big(S,\ m',\ (\tilde{x}_1, \ldots, \tilde{x}_{m'}),\ \mathcal{H},\ \mathcal{B}_{\mathsf{PAC}},\ \epsilon,\ \delta,\ \alpha,\ \beta\Big)$

3: **if** $m' = m_o$ **then**

4: $\quad T_{\mathsf{pub}} \leftarrow (\tilde{x}_1, \ldots, \tilde{x}_{m_o})$

5: $\quad h_{\mathsf{priv}} \leftarrow \mathcal{A}_{\mathsf{SSPP}}(S, T_{\mathsf{pub}}, \mathcal{H}, \epsilon)$

6: $\quad$ **for** $j = m_o + 1, \ldots, m$ **do**

7: $\quad\quad$ Output $y_j^{\mathsf{priv}} \leftarrow h_{\mathsf{priv}}(\tilde{x}_j)$

We finally formalize this observation in the following theorem.

Theorem 11 *Let $\mathcal{H}$ be any hypothesis class with $\mathsf{VC}(\mathcal{H}) = d$. For any $\epsilon, \delta, \alpha, \beta \in (0, 1)$ and any $m < \infty$, $\mathcal{A}_{\mathsf{UnvPrivCl}}$ is an $(\epsilon, \delta, \alpha, \beta, n, m)$-PCQR algorithm for $\mathcal{H}$ with private sample size*

$$n = O\left(\frac{\left(d\, \log\left(\frac{1}{\alpha}\right) + \log\left(\frac{m_o}{\beta}\right)\right) \log^{3/2}\left(\frac{2}{\delta}\right) \log\left(\frac{m_o \alpha}{\min(\delta, \beta/2)}\right)}{\epsilon\ \alpha^2} \cdot \max\left(1, \sqrt{d}\, \alpha \log^{1/2}\left(\frac{1}{\alpha}\right)\right)\right),$$

where $m_o = O\left(\frac{d \log(1/\alpha) + \log(1/\beta)}{\alpha}\right)$ (as set in Step 1). In particular, when $\alpha \leq \frac{1}{\sqrt{d}}$, it would suffice to have a private sample of size $n = \tilde{O}\left(\frac{d}{\epsilon\ \alpha^2}\right)$.

Near-optimality of our sample complexity bound: Note that without any privacy constraints, the sample complexity of this problem in the agnostic PAC setting is $\Theta\left(\left(\frac{\mathsf{VC}(\mathcal{H}) + \log(1/\beta)}{\alpha^2}\right)\right)$. Note that this follows from the standard agnostic PAC learning bound and the fact that access to unlabeled data (the set of queries) does not improve the sample complexity (Ben-David et al., 2008, Theorem 15), unless one makes assumptions about the conditional distribution of the true label given the unlabeled domain point. Now, when

$\alpha = O\left(\frac{1}{\sqrt{\mathsf{VC}(\mathcal{H})}}\right)$, and assuming $\epsilon = \Theta(1)$, our private sample complexity bound in Theorem 11 nearly matches the non-private sample complexity. This shows that our bound is optimal (up to log factors) in that parameters regime[1].

Remark 12 *It is worth mentioning that applying the same technique (the semi-private learner of Alon et al. (2019)) to the construction of (Bassily et al., 2018) also yields a universal PCQR algorithm but with a worse sample complexity bound than ours. In particular, it is not hard to see that the resulting sample complexity bound based on the construction by Bassily et al. (2018) is* $\tilde{O}\left(\frac{\mathsf{VC}(\mathcal{H})}{\alpha^2} \cdot \max\left(1, \sqrt{\mathsf{VC}(\mathcal{H})}\right)\right)$*, where* $\tilde{O}$ *hides* $\operatorname{polylog}$ *factors in* $(1/\alpha, 1/\beta, 1/\delta)$*. Our bound is tighter by a factor of* $\approx \alpha$ *when* $\mathsf{VC}(\mathcal{H}) = \omega(1)$.

Acknowledgments

The authors would like to thank Uri Stemmer, Amos Beimel, and Kobbi Nissim for pointing us to their elegant LabelBoost procedure in Beimel et al. (2015) which is the crux of the pre-processing step of our algorithm. We are also grateful to them for the several insightful discussions we had about this line of research. We also thank Vitaly Feldman for his helpful comments on the manuscript, and the anonymous reviewer for pointing out the fact that we can save a factor of $1/\epsilon$ (in an earlier version of the bounds) via a small tweak in the argument.

References

Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private pac learning implies finite littlestone dimension. *arXiv preprint arXiv:1806.00949 (STOC 2019, in Press)*, 2018.

Noga Alon, Raef Bassily, and Shay Moran. Limits of private learning with access to public data. *To appear in NeuRIPS 2019, also available at arXiv:1910.11519 [cs.LG]*, 2019.

Raef Bassily, Abhradeep Thakurta, and Om Thakkar. Model-agnostic private learning. In *Advances in Neural Information Processing Systems 31*, pages 7102–7112. Curran Associates, Inc., 2018.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 363–378. Springer, 2013.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Learning privately with labeled and unlabeled examples. *CoRR*, abs/1407.2662 (appeared at SODA 2015), 2015.

Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *COLT*, pages 33–44, 2008.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 634–649. IEEE, 2015.

Yuval Dagan and Vitaly Feldman. Pac learning with stable and private predictions. *arXiv:1911.10541 [cs.LG]*, 2019.

[1] Note that the accuracy of $\mathcal{A}_{\mathsf{UnvPrivCl}}$ is defined in terms of the average misclassification rate over the given set of queries rather the expected classification error, however, since the queries are i.i.d., it is easy to see that the two accuracy definitions are essentially equivalent (by applying a standard concentration argument).

Cynthia Dwork and Vitaly Feldman. Privacy-preserving prediction. *arXiv preprint arXiv:1803.10266*, 2018.

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

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006b.

Jihun Hamm, Yingjun Cao, and Mikhail Belkin. Learning privately from multiparty data. In *International Conference on Machine Learning*, pages 555–563, 2016.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? In *FOCS*, pages 531–540. IEEE Computer Society, 2008.

Ninghui Li, Wahbeh Qardaji, and Dong Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, pages 32–33. ACM, 2012.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, 2007.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, 2007.

Nicolas Papernot, Martın Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *stat*, 1050, 2017.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

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

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Adam Smith and Abhradeep Thakurta. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *COLT*, 2013.

Appendix A. Proof of Lemma 6

Proof of Lemma 6

Note that the score function for the exponential mechanism is $-\widehat{\mathsf{err}}(h; S')$ whose global sensitivity is $1/n'$. Now, by using the standard accuracy guarantees of exponential mechanism of (McSherry and Talwar, 2007) (and the fact that its instantiated here with privacy parameter $= 1$), w.p. $\geq 1 - \beta/3$ we have

$$\widehat{\mathsf{err}}\left(\widehat{h}; S'\right) - \widehat{\mathsf{err}}\left(h_{S'}^{\mathsf{ERM}}; S'\right) \leq \frac{2}{n'}\left(\log\left(|\widetilde{H}|\right) + \log(\frac{3}{\beta})\right).$$

Using the value of n' given in the lemma statement, together with Sauer's Lemma (Sauer, 1972) to bound the size of $\tilde{H}$, it follows that:

$$\begin{aligned}\widehat{\mathsf{err}}\left(\widehat{h}; S'\right) - \widehat{\mathsf{err}}\left(h_{S'}^{\mathsf{ERM}}; S'\right) &\leq \frac{2}{n'}\left(d\log\left(\frac{en'}{d}\right) + \log\left(\frac{3}{\beta}\right)\right) \\ &\leq \frac{80\alpha^2\ \left(d\log(1/\alpha) + \log(3/\beta)\right)}{256(\ d + \log(3/\beta))} \\ &\leq \alpha/3. \end{aligned} \tag{4}$$

Given the bound on n' and the fact that $S' \sim \mathcal{D}^{n'}$, by a standard uniform convergence argument from learning theory (Shalev-Shwartz and Ben-David, 2014), we have the following generalization error bounds. With probability $\geq 1 - 2\beta/3$, we have:

$$|\mathsf{err}(\widehat{h}; \mathcal{D}) - \widehat{\mathsf{err}}(\widehat{h}; S')| \leq \alpha/3, \tag{5}$$

$$|\mathsf{err}\left(h_{S'}^{\mathsf{ERM}}; \mathcal{D}\right) - \widehat{\mathsf{err}}\left(h_{S'}^{\mathsf{ERM}}; S'\right)| \leq \alpha/3 \tag{6}$$

Putting (4)-(6) together, we conclude that w.p. $\geq 1 - \beta$, we have $\mathsf{err}\left(\widehat{h}; \mathcal{D}\right) - \mathsf{err}\left(h_{S'}^{\mathsf{ERM}}; \mathcal{D}\right) \leq \alpha$. This completes the proof. ■

Appendix B. Constructions from previous works

B.1. Description of Algorithm $\mathcal{A}_{\mathsf{SubSamp}}$

For completeness, here we briefly describe the algorithm $\mathcal{A}_{\mathsf{SubSamp}}$ (Algorithm 5 below) due to Bassily et al. (2018). The input to $\mathcal{A}_{\mathsf{SubSamp}}$ is a private labeled dataset $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, an online sequence of classification queries $Q = (\tilde{x}_1, \ldots, \tilde{x}_m)$, and a generic non-private PAC learner $\mathcal{B}$ for a hypothesis class $\mathcal{H}$. The algorithm outputs a sequence of private labels $(y_1^{\mathsf{priv}}, \ldots, y_m^{\mathsf{priv}})$. The key idea in $\mathcal{A}_{\mathsf{SubSamp}}$ is as follows: first, it arbitrarily splits S into k equal-sized sub-samples $S_1, \ldots, S_k$ for appropriately chosen k. Each of those sub-samples is used to train $\mathcal{B}$. Hence, we obtain an ensemble of k classifiers $h_{S_1}, \cdots, h_{S_k}$. Next for each input query $\tilde{x}_i \in Q$, the votes $(h_{S_1}(\tilde{x}_i), \ldots, h_{S_k}(\tilde{x}_i))$ are computed. It then applies the distance-to-instability test (Smith and Thakurta, 2013) on the difference between the largest count of votes and the second largest count. If the majority vote is sufficiently stable, $\mathcal{A}_{\mathsf{SubSamp}}$ returns the majority vote as the predicted label for $\tilde{x}_i$; otherwise, it returns a random label. The sparse-vector framework is employed to efficiently manage the privacy budget over the m queries. In particular, by employing the sparse-vector technique, the privacy budget of $\mathcal{A}_{\mathsf{SubSamp}}$ is only consumed by those queries where the majority vote is not stable. Algorithm $\mathcal{A}_{\mathsf{SubSamp}}$ takes an input cut-off parameter T, which represents a bound on the total number of "unstable queries" the algorithm can answer before it halts in order to ensure (ϵ, δ)-differential privacy.

Algorithm 5 $\mathcal{A}_{\mathsf{SubSamp}}$ Bassily et al. (2018): Private Classification via subsample-aggregate and sparse-vector

Input: Private dataset: S, upper bound on the number of queries: m, online sequence of classification queries: $\mathcal{Q} = \{\tilde{x}_1, \ldots, \tilde{x}_m\}$, hypothesis class $\mathcal{H}$, oracle access to a PAC learner of $\mathcal{H}$: $\mathcal{B}_{\mathsf{PAC}}$, unstable query cutoff: T, privacy parameters: $\epsilon, \delta > 0$, failure probability: β.

$c \leftarrow 0$, $\lambda \leftarrow \frac{\sqrt{32T\log(2/\delta)}}{\epsilon}$ and $k \leftarrow 34\sqrt{2}\lambda \cdot \log\left(4mT/\min\left(\delta, \beta/2\right)\right)$

$w \leftarrow 2\lambda \cdot \log(2m/\delta)$, $\hat{w} \leftarrow w + \mathsf{Lap}(\lambda)$ $\{\mathsf{Lap(b)}$ denotes the Laplace distribution with scale $b\}$

Split S into k non-overlapping sub-samples $S_1, \cdots, S_k$.

for $j \in [k]$ **do**

  $h_{S_j} \leftarrow \mathcal{B}_{\mathsf{PAC}}(S_j)$

for $i \in [m]$ and $c \leq T$ **do**

  $\mathcal{F}_i \leftarrow \{h_{S_1}(x_i), \cdots, h_{S_k}(x_i)\}$ {For every $y \in \{0,1\}$, let $\mathsf{ct}(y) = \#$ times y appears in $\mathcal{F}_i$.}

  $\widehat{q}_{x_i} \leftarrow \arg\max_{y \in \{0,1\}} [\mathsf{ct}(y)]$, $\mathsf{dist}_{\widehat{q}_{x_i}} \leftarrow$ largest $\mathsf{ct}(y)$ - second largest $\mathsf{ct}(y)$

  $y_i^{\mathsf{priv}} \leftarrow \mathcal{A}_{\mathsf{stab}}(S, \widehat{q}_{x_i}, \mathsf{dist}_{\widehat{q}_{x_i}}, \hat{w}, \frac{1}{2\lambda})$ {Stability test for $\widehat{q}_{x_i}$, given by Algorithm 6 below.}

  if $y_i^{\mathsf{priv}} = \perp$, **then** $c \leftarrow c + 1$, $\hat{w} \leftarrow w + \mathsf{Lap}(\lambda)$

  Output y_i^{priv}

Algorithm 6 $\mathcal{A}_{\mathsf{stab}}$ Smith and Thakurta (2013): Private estimator for f via distance to instability

Input: Dataset: S, function: $f : U^n \to \mathcal{R}$, distance to instability: $\mathsf{dist}_f : U^n \to \mathbb{R}$, threshold: Γ, privacy parameter: $\epsilon > 0$

$\widehat{\mathsf{dist}} \leftarrow \mathsf{dist}_f(S) + \mathsf{Lap}\left(1/\epsilon\right)$

if $\widehat{\mathsf{dist}} > \Gamma$, **then** output $f(S)$, **else** output $\perp$

B.2. Description of Algorithm $\mathcal{A}_{\mathsf{SSPP}}$

In Section 4, we use a semi-supervised semi-private learner construction from (Alon et al., 2019) (referred to as $\mathcal{A}_{\mathsf{SSPP}}$) to give a construction for a universal PCQR algorithm that can answer any number of classification queries (Algorithm 4). For completeness, we describe the construction of this semi-private learner $\mathcal{A}_{\mathsf{SSPP}}$ in Algorithm 7 below[2]. Algorithm 7 takes as input two datasets: a private dataset S of size n, and an unlabeled public dataset T_{pub} of size m_o, and outputs a hypothesis $h_{\mathsf{priv}} : \mathcal{X} \to \{0,1\}$. The main idea of the construction in (Alon et al., 2019) is that the public unlabeled dataset can be used to create a finite α-cover for $\mathcal{H}$ (see Definition 13 below), and hence, reducing the task of privately learning $\mathcal{H}$ to the task of learning a finite sub-class of $\mathcal{H}$ (the α-cover).

Definition 13 (α-cover for a hypothesis class) *A family of hypotheses $\widetilde{\mathcal{H}}$ is said to form an alpha-cover for a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ with respect to distribution $\mathcal{D}_{\mathcal{X}}$ if for every $h \in \mathcal{H}$ there exists a $\tilde{h} \in \widetilde{\mathcal{H}}$ such that* $\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}\left[\mathbf{1}(h(x) \neq \tilde{h}(x)\right] \leq \alpha$.

[2] A similar construction of the semi-private learner $\mathcal{A}_{\mathsf{SSPP}}$ has also appeared in the earlier work by Beimel et al. (2013).

Algorithm 7 $\mathcal{A}_{\mathsf{SSPP}}$ Alon et al. (2019): Semi-Supervised Semi-Private Agnostic Learner

Input: Private labeled dataset: $S \in U^n$, a public unlabeled dataset: $T_{\mathsf{pub}} = (\tilde{x}_1, \cdots, \tilde{x}_{m_o}) \in \mathcal{X}^{m_o}$, a hypothesis class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$, and a privacy parameter $\epsilon > 0$.

Let $\tilde{T} = \{\hat{x}_1, \ldots, \hat{x}_{\hat{m}}\}$ be the set of points $x \in \mathcal{X}$ appearing at least once in T_{pub}.

Let $\Pi_{\mathcal{H}}(\tilde{T}) = \{(h(\hat{x}_1), \ldots, h(\hat{x}_{\hat{m}})) : \ h \in \mathcal{H}\}$.

Initialize $\tilde{\mathcal{H}}_{T_{\mathsf{pub}}} = \emptyset$.

for each $\mathbf{c} = (c_1, \ldots, c_{\hat{m}}) \in \Pi_{\mathcal{H}}(\tilde{T})$: **do**

    Add to $\tilde{\mathcal{H}}_{T_{\mathsf{pub}}}$ arbitrary $h \in \mathcal{H}$ that satisfies $h(\hat{x}_j) = c_j$ for every $j = 1, \ldots, \hat{m}$.

Use the exponential mechanism with inputs S, $\tilde{\mathcal{H}}_{T_{\mathsf{pub}}}, \epsilon$, and score function $q(S,h) \triangleq -\widehat{\mathsf{err}}(h;S)$ to select $h_{\mathsf{priv}} \in \tilde{\mathcal{H}}_{T_{\mathsf{pub}}}$.

return h_{priv}.