Public-data Assisted Private Stochastic Optimization: Power and Limitations

Enayat Ullah
Meta *
enayat@meta.com

Michael Menart
Department of Computer Science & Engineering
The Ohio State University †
Department of Computer Science, University of Toronto
Vector Institute
menart.2@osu.edu

Raef Bassily
Department of Computer Science & Engineering
Translational Data Analytics Institute (TDAI)
The Ohio State University
bassily.1@osu.edu

Cristóbal Guzmán
Inst. for Mathematical and Comput. Eng.
Fac. de Matemáticas and Esc. de Ingeniería
Pontificia Universidad Católica de Chile
crguzmanp@uc.cl

Raman Arora
Department of Computer Science
Johns Hopkins University
arora@cs.jhu.edu

Abstract

We study the limits and capability of public-data assisted differentially private (PA-DP) algorithms. Specifically, we focus on the problem of stochastic convex optimization (SCO) with either labeled or unlabeled public data. For complete/labeled public data, we show that any (ϵ, δ)-PA-DP has excess risk $\tilde{\Omega}\big(\min\big\{\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\epsilon}\big\}\big)$, where d is the dimension, n_{pub} is the number of public samples, n_{priv} is the number of private samples, and $n = n_{\text{pub}} + n_{\text{priv}}$. These lower bounds are established via our new lower bounds for PA-DP mean estimation, which are of a similar form. Up to constant factors, these lower bounds show that the simple strategy of either treating all data as private or discarding the private data, is optimal. We also study PA-DP supervised learning with *unlabeled* public samples. In contrast to our previous result, we here show novel methods for leveraging public data in private supervised learning. For generalized linear models (GLM) with unlabeled public data, we show an efficient algorithm which, given $\tilde{O}(n_{\text{priv}}\epsilon)$ unlabeled public samples, achieves the dimension independent rate $\tilde{O}\big(\frac{1}{\sqrt{n_{\text{priv}}}} + \frac{1}{\sqrt{n_{\text{priv}}\epsilon}}\big)$. We develop new lower bounds for this setting which shows that this rate cannot be improved with more public samples, and any fewer public samples leads to a worse rate. Finally, we provide extensions of this result to general hypothesis classes with finite *fat-shattering dimension* with applications to neural networks and non-Euclidean geometries.

*Work done while the author was at the Johns Hopkins University.
†This work was done while M. Menart was at The Ohio State University.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1 Introduction

The framework of differential privacy has become the primary standard for protecting individual privacy in data analysis and machine learning. Unfortunately, this rigorous framework has also been shown to lead to worse performance on such tasks both empirically and in theory [BST14, PVX+23]. However, it is often the case that, in addition to a collection of privacy-sensitive data points, analysts have access to a pool of public data, for which guaranteeing privacy protections is not required. This can happen, for example, when consumers deem their own data non-sensitive and opt-in to sell this data to a company. This has motivated a long line of work analyzing how public data can be leveraged in tandem with private data to provide better utility [BNS13, ABM19, BCM+20, ZWB21, BKS22, AGM+22, NMT+23]. In machine learning. for example, two commonly proposed strategies are public pretraining and using public data to identify gradient subspaces [ZWB21, KDRT21]. Public pretraining, in particular, has proven effective in practice [YNB+22a, BWZK22], and prior work has even identified a *specific* problem instance where public and private data used in tandem leads to better rates than is possible using only the public or private datasets in isolation [GHN+23]. Despite this surge of work, theory has struggled to show that public data leads to fundamental rate improvements more generally. Recent work has even shown that, for the problem of pure PA-DP stochastic convex optimization, a small amount of public data, $n_{\text{pub}} \leq n\epsilon/d$, leads to no rate improvement, where $n = n_{\text{pub}} + n_{\text{priv}}$ and n_{priv} is the number of private samples [LLHR23].

One particularly important version of this problem is in supervised learning when the public data is unlabeled. This setting has found importance in medical domains and deep learning more generally [LW19, SCZ+20, PAE+17]. Notably, unlabeled data is much less time intensive to collect than labeled data. Due to this fact, and the fact that the unlabeled public data does not contain the same kind of information contained in the private data, the regime $n_{\text{pub}} = \Omega(n_{\text{priv}})$ is meaningful both in theory and in practice. We also note this setting is a stronger (in terms of privacy) version of the label-private setting, where only the labels of the dataset are considered private [CH11, BNS13].

Motivated by the importance of these settings and the lack of existing theory for them in stochastic optimization, we study fundamental limitations and applications of public data in (ϵ, δ)-PA-DP stochastic optimization. In the case where the public data is complete/labeled, we show that the application of public data is fundamentally limited. We then contrast this result with new results in the unlabeled public data setting. In this setting, we provide new results for GLMs, and extend these results to more general hypothesis classes, with finite fat-shattering dimension, and non-Euclidean geometries.

1.1 Our Contributions

We outline our primary contributions in the following.

Limits of Private Stochastic Convex Optimization with Public Data. First, we show a tight lower bound for the problem of differentially-private stochastic convex optimization (DP-SCO) assisted with complete public data, that is, the public data and private data have the same number of features (and labels when applicable). Specifically, we show a lower bound of $\Omega\Big(\min\big\{\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\epsilon}\big\}\Big)$ on the excess population risk for this problem. When $d \geq n\epsilon$ and $n_{\text{pub}} \leq \frac{n}{\log(1/\delta)}$, we further improve this lower bound to $\Omega\Big(\min\big\{\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\big\}\Big)$. This lower bound is matched by the simple upper bound strategy which either discards the private data entirely and outputs the public mean or simply treats all data-points as private. Barring constant factors, this shows more sophisticated attempts at leveraging public data will yield no benefit. These results also hold even for generalized linear models. Our results are based on new results we establish for DP mean estimation with public data, and a reduction of mean estimation to SCO. We note that previous work [LLHR23], on this problem either focused on the pure PA-DP case when $n_{\text{pub}} \leq n\epsilon/d$, or, in the approximate PA-DP case, did not obtain the dimension dependence. Our mean estimation lower bound uses a novel analysis of fingerprinting codes [BUV14], and our SCO reduction further builds on ideas from [BST14, CWZ21]. We also show that, when $d \geq n\epsilon$, our lower bounds for approximate PA-DP SCO directly imply a tight lower bound for *pure* PA-DP.

Private Supervised Learning with Unlabeled Public Data. While the previously discussed results show there is no hope for leveraging public data in "interesting" ways, even for GLMs, they do not preclude settings where the public data is less informative. In particular, in the setting where the

public data is *unlabeled*, it makes sense to even consider $n_{\text{pub}} \geq n_{\text{priv}}$. In this setting, we provide the following results.

- For (Euclidean) GLMs we develop an efficient algorithm which, given $\tilde{O}(n_{\text{priv}}\epsilon)$ unlabeled public data points, achieves the dimension independent rate $\tilde{O}\big(\frac{1}{\sqrt{n_{\text{priv}}}} + \frac{1}{\sqrt{n_{\text{priv}}\epsilon}}\big)$. We obtain this result via a dimensionality reduction procedure of the private feature vectors using the public data, and then running an efficient private algorithm in the lower dimensional space. The key idea is that public data can be used to identify a low dimensional subspace, which under the appropriate metric acts as a cover for the higher dimensional space. We elucidate the tightness of our upper bound by proving two new lower bounds which show that access to a greater number of unlabeled public samples cannot improve this rate, and that any fewer public samples lead to a worse rate. While dimension independent rates for the GLMs have previously been developed in the *unconstrained* setting [SSTT21, ABG+22], in the constrained setting which we study, dependence on dimension is known to be unavoidable even for GLMs if no public data is available [BST14]. Our result thus allows us to bypass these limitations.
- By observing that the key requirement in our GLM result is the construction of an appropriate cover, we extend this result to general hypothesis classes with bounded *fat-shattering dimension*. In the non-private setting, it is known that finiteness of fat-shattering dimension characterizes learnability of real-valued predictors with *scale-sensitive* losses [BLW94, ABDCBH97]. In the private setting, such a result is not known, and is in fact impossible in the *proper learning* setting. This follows from the fact that norm bounded linear predictors, regardless of their (ambient) dimension d, have the same fat-shattering dimension [SST10]. However, it is known that they are not learnable privately in high dimensions $d \geq (n\epsilon)^2$ [BST14]. In contrast, in the PA-DP setting, we show that it is possible to properly learn such classes with a rate of roughly $O\left(\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H}) + \inf_{\alpha>0}\left(\frac{\text{fat}_\alpha(\mathcal{H})}{n_{\text{priv}}\epsilon} + \alpha\right)\right)$, where $\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H})$ denotes the Rademacher complexity of $\mathcal{H}$ and $\text{fat}_\alpha(\mathcal{H})$ denotes its fat-shattering dimension at scale α (see Section 2 for preliminaries).
- As applications of our result for hypothesis classes with bounded fat-shattering dimension, we obtain guarantees for learning feed-forward neural networks and non-Euclidean GLMs. In particular, for depth M feed-forward neural networks with weights bounded as $\|W_j\|_F \leq R_j$ and 1-Lipschitz positive homogeneous activation, we achieve an excess risk bound of essentially $\tilde{O}\left(\frac{\sqrt{M}\prod_{j=1}^M R_j}{\sqrt{n_{\text{priv}}}} + \left(\frac{M\left(\prod_{j=1}^M R_j\right)^2}{n_{\text{priv}}\epsilon}\right)^{1/3}\right)$. For non-Euclidean GLMs, our guarantees are dimension-independent which is not known to be achievable, as of yet, even in the unconstrained setting with no public data (unlike Euclidean GLMs).

1.2 Related Work

With regards to labeled public data, the most directly related work to ours is the recent work of [LLHR23]. This work proves a lower bound of $\Omega\left(\min\left\{\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{1}{\sqrt{n}} + \frac{1}{n\epsilon}\right\}\right)$ for approximate PA-DP mean-estimation/SCO. We note that our results for approximate PA-DP crucially obtain a dependence on d that is the key "price" paid for privacy in this setting. [LLHR23] also show a lower bound of $\Omega\left(\min\left\{\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{1}{\sqrt{n}} + \frac{d}{n\epsilon}\right\}\right)$ on a *pure* PA-DP mean estimation/SCO, but this result only holds when $d \leq \frac{n\epsilon}{n_{\text{pub}}}$. As such, their result is orthogonal to our result in the pure PA-DP setting, which operates in the regime $d \geq n\epsilon$. In both cases, our proof technique is fundamentally different than theirs.[3] Tangentially, [BKS22] showed a small amount of public data is useful in pure-DP mean estimation when the range parameters on the data are unknown.

An important setting where public data is shown to be useful is *PAC learning*. Non-privately, it is known that the finiteness of *VC dimension* characterizes learnability [VC71, BEHW89]. However, under DP, it is impossible to PAC learn even the class of *thresholds*, which has VC dimension of one [BNS13]. The works of [BNS13, BTGT18, ABM19] showed that given access to a small unlabelled public data, it is possible to go beyond this limitation and privately learn VC classes, essentially by reducing a hypothesis class with finite VC dimension to a finite hypothesis class

[3] We note that concurrently and independently, version 2 of [LLHR23], [LLHR24], obtained a lower bound of $\Omega\left(\min\left\{\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\epsilon}\right\}\right)$, but their lower bound is limited to *symmetric* procedures.

A number of works have studied the impact of public data in applied settings as well. A common technique is to use public data to reduce the problem dimension in some way [ZWB21, YZCL21, PHYS24]. The work of [GHN+23] identified a specific problem instance which supports the method of public pretraining commonly used in practice.

With regards to unlabelled public data, there are several existing works. Transfer learning is a common approach in this setting. Besides the benefits in PAC learning, this setting also has applications in deep learning, where (empirically) unlabeled public data has been used to obtain performance improvements [PAE+17, PSM+18]. Unlabeled public data has also yielded impressive results used for pre-training large language models [LTLH22, YNB+22b]. We also remark that, in practice, it is reasonable to expect the private and public datasets to come from slightly different distributions. Accounting for this distribution shift has also been the study of several recent works [BKS22, BDBC+23]. However, in this work we focus on first characterizing the more fundamental problem where the public and private datasets are drawn i.i.d. from the same distribution.

2 Preliminaries

Here, we describe the concepts and assumptions used in the rest of this paper. In this work, $\|\cdot\|$ always denotes the ℓ_2 norm unless stated otherwise.

Public-Data Assisted Differential Privacy. We first present the traditional notion of differential privacy (DP). Let $n, d \in \mathbb{N}$ and $\mathcal{X}$ be some data domain. When no public data is present, we say that an algorithm $\mathcal{A}$ satisfies (ϵ, δ)-differential privacy (DP) if for all datasets S and S' differing in one data point and all events $\mathcal{E}$ in the range of $\mathcal{A}$, $\mathbb{P}\left[\mathcal{A}(S) \in \mathcal{E}\right] \leq e^{\epsilon}\mathbb{P}\left[\mathcal{A}(S') \in \mathcal{E}\right] + \delta$ [DMNS06].

In our work, we denote the number of public samples in the dataset, $S = (S_{\text{pub}}, S_{\text{priv}}) \in \mathcal{X}^n$, as n_{pub} and the number of private samples as n_{priv}, such that $n = n_{\text{pub}} + n_{\text{priv}}$. In keeping with previous work [BNS13, BCM+20], we define public data assisted differentially private algorithms in the following way [4].

Definition 1 (PA-DP)**.** *An algorithm $\mathcal{A}$ is (ϵ, δ) public-data assisted differentially private (PA-DP) algorithm with public sample size n_{pub} and private sample size n_{priv} if for any public dataset $S_{pub} \in \mathcal{X}^{n_{pub}}$, and any pair of private datasets $S_{priv}, S'_{priv} \in \mathcal{X}^{n_{priv}}$ differing in at most one entry, it holds for any event $\mathcal{E}$ that $\mathbb{P}\left[\mathcal{A}(S_{pub}, S_{priv}) \in \mathcal{E}\right] \leq e^{\epsilon}\mathbb{P}[\mathcal{A}(S_{pub}, S'_{priv}) \in \mathcal{E}] + \delta$. When $\delta = 0$, we refer to this notion as pure PA-DP, denoted as ϵ-PA-DP.*

Stochastic Convex Optimization Let $\mathcal{D}$ be a distribution supported on $\mathcal{X}$. Given some constraint set $\mathcal{W} \subseteq \mathbb{R}^d$ of diameter at most D, and a G Lipschitz convex loss $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, we are interested in minimizing the population loss, $L(w; \mathcal{D}) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\ell(w; x)\right]$. Denote the minimizer as $w^* = \min_{w\in\mathcal{W}}\{L(w; \mathcal{D})\}$. We evaluate the quality of the approximate solution, w, via the excess risk, $L(w; \mathcal{D}) - L(w^*; \mathcal{D})$. Specifically, we are interested in PA-DP algorithms which minimizes this quantity when given $S_{\text{pub}}, S_{\text{priv}} \overset{i.i.d.}{\sim} \mathcal{D}$. For a datset S we also define the empirical loss $\widehat{L}(w; S) = \frac{1}{|S|}\sum_{x\in S}\ell(w; x)$.

Supervised Learning and Generalized Linear Models (GLMs) In the supervised learning setting, in addition to the feature space $\mathcal{X}$, we define the label space $\mathcal{Y}$. We here let $\mathcal{D}$ be a joint probability distribution over $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{D}_{\mathcal{X}}$ and $\mathcal{D}_{\mathcal{Y}}$ denote the respective marginal distributions. Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ be a hypothesis class of real-valued predictors, and let $\text{fat}_{\alpha}(\mathcal{H})$ denote its fat shattering dimension at scale α. Consider the loss function $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, such that $\ell(h; x, y) = \phi_y(h(x))$ for some function ϕ_y. We assume that the map $\phi_y : \mathbb{R} \to \mathbb{R}$ is G-Lipschitz for all $y \in \mathcal{Y}$ and is B-bounded. Further, we assume that $\sup_{x\in\mathcal{X}}|h(x)| \leq R$ and define $\sup_{x\in\mathcal{X}}\|x\| = \|\mathcal{X}\|$.

GLMs are a special case of supervised learning setting where the hypothesis class is that of linear predictors, $\mathcal{H} = \mathcal{W} \subseteq \mathbb{R}^d$, over $\mathcal{X} \subseteq \mathbb{R}^d$, and $h(x) = w^\top x$. We refer to the public dataset of unlabeled feature vectors as X_{pub}.

Covering numbers, fat-shattering and Rademacher Complexity Given $X = (x_1, x_2, \cdots, x_m)$ the ℓ_p distance between two hypothesis $h_1, h_2 \in \mathcal{H}$ with respect to the empirical measure over X, is defined as, $\|h_1 - h_2\|_{p,X} = \left(\frac{1}{m}\sum_{x\in X}|h_1(x) - h_2(x)|^p\right)^{1/p}$. Similarly, the distance with respect

[4] The term semi-DP algorithm has also been used in some works.

to the population, is given by $\|h_1 - h_2\|_{p,\mathcal{D}_\mathcal{X}} = \left(\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}} |h_1(x) - h_2(x)|^p\right)^{1/p}$. The covering number of $\mathcal{H}$ at scale $\alpha > 0$ and given dataset X, denoted as $\mathcal{N}_p(\mathcal{H}, \alpha, X)$ is the size of the minimal set of hypothesis, $\tilde{\mathcal{H}}$, such that for any $h \in \mathcal{H}$ there exists $\tilde{h}$ with $\|h - \tilde{h}\|_{p,X} \leq \alpha$. We define $\mathcal{N}_p(\mathcal{H}, \alpha, m) = \sup_{X:|X|=m} \mathcal{N}_p(\mathcal{H}, \alpha, X)$, the covering number with respect to all datasets of size m. We define fat-shattering dimension below.

Definition 2. *[BLW94] Let $\mathcal{H} \subseteq \mathbb{R}^\mathcal{X}$ and $\alpha > 0$. We say that $\mathcal{H}$ α-shatters $X = \{x_1, x_2, \ldots, x_m\}$ if $\sup_{r\in\mathbb{R}^m} \min_{y\in\{-1,1\}^m} \sup_{h\in\mathcal{H}} \min_{i\in[m]} y_i(h(x_i) - r_i) \geq \alpha$. The fat-shattering dimension, $\text{fat}_\alpha(\mathcal{H})$, is the size of the largest α-shattered set.*

We define $\mathfrak{R}_m(\mathcal{H})$, the worst-case Rademacher complexity of $\mathcal{H}$ with respect to m data points, as $\mathfrak{R}_m(\mathcal{H}) = \sup_{X:|X|=m} \mathbb{E}_{\sigma_i} \sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i)$. An important example is that of norm-bounded linear predictors $\mathcal{H} = \{w : x \mapsto \langle w, x\rangle : \|w\| \leq D\}$ over $\mathcal{X} = \{x : \|x\| \leq \|\mathcal{X}\|\}$. Herein, $\text{fat}_\alpha(\mathcal{H}) = \Theta\left(\frac{D^2\|\mathcal{X}\|^2}{\alpha^2}\right)$ and $\mathfrak{R}_m(\mathcal{H}) = \Theta\left(\frac{D\|\mathcal{X}\|}{\sqrt{m}}\right)$ [KST08, SST10].

3 Private Stochastic Convex Optimization with Labeled Public Data

In this section, we present our lower bounds for private stochastic convex optimization with public data. When interpreting the following results, it is helpful to note that in the nontrivial regime, $n_{\text{pub}} = \Theta(n)$ and $n_{\text{pub}} = o(n)$, although our results hold regardless. Further, recall that an upper bound for this problem of $O\left(R\min\left\{\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right\}\right)$ can be obtained by simply either applying an optimal SCO algorithm to only the public data (and discarding the private data) or applying an optimal DP-SCO algorithm and treating the entire dataset as private [BFTGT19]. As we will see, this strategy is essentially optimal.

3.1 Lower Bound for Stochastic Convex Optimization

We start by stating our lower bound for public-data assisted differentially private SCO.

Theorem 1. *Let $\delta \leq \frac{1}{16nd}$, $\epsilon \leq 1$, and d be larger than some universal constant. For any (ϵ, δ)-PA-DP algorithm, there exists a distribution $\mathcal{D}$, and a G-Lipschitz loss such that $\mathbb{E}\left[L(\mathcal{A}(S_{pub}, S_{priv}); \mathcal{D}) - \min_{w:\|w\|\leq D}\{L(w;\mathcal{D})\}\right] = \Omega\left(GD\cdot\Psi(n_{pub}, n, d, \epsilon, \delta)\right)$, where for some universal constant c,*

$$\Psi(n_{pub}, n, d, \epsilon, \delta) = \begin{cases} \min\left\{\frac{1}{\sqrt{n_{pub}}}, \frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right\}, & d \geq cn\epsilon,\ n_{pub} \leq \frac{n\epsilon}{c\log\left(1/[\sqrt{nd}\delta]\right)} \\ \min\left\{\frac{1}{\sqrt{n_{pub}}}, \frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{n\epsilon}\right\}, & \text{else} \end{cases}$$

The function Ψ is defined to avoid repetitive notation in the rest of this section. Barring the mild restriction on n_{pub}, even though the $\sqrt{\log(1/\delta)}$ term is only obtained when $d \geq n\epsilon$, the "aggregate" lower bound is tight for all $d \notin [\frac{n\epsilon^2}{\log(1/\delta)}, n\epsilon]$ since when $d \leq \frac{n\epsilon^2}{\log(1/\delta)}$ the non-private $\frac{1}{\sqrt{n}}$ lower bound dominates. It is also pertinent to our results in Section 4 that the problem construction used to achieve this lower bound is a convex GLM, and as a result this lower bound holds even for GLMs.

Finally, similar statements can be made about strongly convex optimization. We again provide just one such statement here.

Theorem 2. *Let $\delta \leq \frac{1}{16nd}$, $\epsilon \leq 1$. For any (ϵ, δ)-PA-DP algorithm there exists a distribution $\mathcal{D}$, λ-strongly convex and G-Lipschitz loss such that*

$$\mathbb{E}\left[L(\mathcal{A}(S_{pub}, S_{priv}); \mathcal{D}) - \min_{w:\|w\|\leq D}\{L(w;\mathcal{D})\}\right] = \Omega\left(\frac{G^2}{\lambda}\Psi^2(n_{pub}, n, \epsilon, \delta)\right).$$

The crux of the proofs for both the above results lies in establishing new mean estimation lower bounds for PA-DP mean estimation, which we give in Appendix B.1. These mean estimation lower bound use a novel application of a construction known as fingerprinting codes. In particular, the introduction of public data introduces significant challenges in the traditional analysis of fingerprinting codes. As these challenges are more technical in nature, we defer their discussion to Appendix B.2.

After establishing the mean estimation lower bounds, we can adapt the reductions first used in [BST14] that show mean estimation lower bounds can be used to provide lower bounds for risk minimization without public data. Full proofs for the above claims, and in particular details for the above reductions, are found in Appendix B.3.

Lower Bound for Pure DP Case. While not the primary focus of this work, the previous lower bound directly leads to a lower bound for pure PA-DP SCO. Since any ϵ-DP algorithm is (ϵ, δ)-DP for any $\delta > 0$, we can use the above theorem to obtain a non-trivial lower bound for the pure DP case by setting δ small. Specifically, by setting δ such that $\log(1/\delta) = \frac{n\epsilon}{120^2 n_{\text{pub}}}$, one immediately obtains a lower bound of $\Omega\left(\min\{\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{\sqrt{d}}{\sqrt{n_{\text{pub}} \cdot n\epsilon}}\}\right)$ for d large enough. Simplifying this expression yields the following.

Corollary 1. *Let $d \geq cn\epsilon$ for a constant c, and $\mathcal{A}$ be an ϵ-PA-DP algorithm. There exist a distribution $\mathcal{D}$ and a G-Lipschitz loss such that $\mathbb{E}\big[L(\mathcal{A}(S_{pub}, S_{priv}); \mathcal{D}) - \min_{w:\|w\|\leq D} \{L(w; \mathcal{D})\}\big] = \Omega\left(\frac{GD}{\sqrt{n_{pub}}}\right)$.*

The known $O\left(GD \min\left\{\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{1}{\sqrt{n}} + \frac{d}{n\epsilon}\right\}\right)$ upper bound for this problem shows that this bound is tight (in the regime in which it holds). Essentially, this bound states that when $d \geq n\epsilon$, the public dataset is not useful (at least asymptotically). Previously [LLHR23, Theorem 31] established that when $d \leq n\epsilon/n_{\text{pub}}$, a tight lower bound of $\Omega\left(GD\left(\frac{d}{n\epsilon} + \frac{1}{\sqrt{n}}\right)\right)$ holds, effectively showing that in this regime the public dataset is not useful[5]. We leave the remaining regime where $d \in (\frac{n\epsilon}{n_{\text{pub}}}, n\epsilon)$ as an interesting open problem for future work. Finally, we note that similar statements can be made about strongly convex losses using Theorem 2.

4 Private Supervised Learning with Unlabeled Public Data

In this section, we consider supervised learning with real-valued predictors given labeled private data and unlabeled public data. Our results show that, in this setting, it is possible to go beyond the limitations established in the prior section.

4.1 Efficient PA-DP learning of Convex Generalized Linear Models

We start with learning linear predictors with convex loses a.k.a. convex generalized learning models. We propose Algorithm 1, which uses the public unlabeled data to perform dimensionality reduction of the private labeled feature vectors. In the following, we use span to denote the span of a set of vectors and dim to denote the dimension of a subspace. The dataset of public unlabeled feature vectors is denote as X_{pub}. Our algorithm projects the private feature vectors onto the subspace spanning $\mathcal{W} \cap \text{span}(X_{\text{pub}})$ to get $\dim(\text{span}(X_{\text{pub}}) \cap \mathcal{W})$-dimensional representation of the private feature vectors. It then reparametrizes the loss function so that its domain is $\dim(\text{span}(S_{\text{pub}}) \cap \mathcal{W})$-dimensional and applies a private subroutine in the lower dimensional space. The output of the subroutine is then embedded back in $\mathbb{R}^d$. Algorithms similar to Algorithm 1 have appeared in the literature (e.g. [PHYS24]). We emphasize that our key contribution is the formal analysis of this technique and the fact that we provide tight upper and lower bounds while simultaneously avoiding many of the strong assumptions seen in previous work, such as large margin assumptions.

Algorithm 1 Efficient PA-DP learning of GLMs with unlabeled public data

Input: Private labeled dataset S_{priv}, public unlabeled dataset X_{pub}, privacy parameters $\epsilon, \delta > 0$.

1: Let $U \in \mathbb{R}^{d \times \dim(\mathcal{W} \cap \text{span}(X_{\text{pub}}))}$ denote the orthogonal projection onto $\text{span}(X_{\text{pub}}) \cap \mathcal{W}$.
2: Define $\tilde{S}_{\text{priv}} = \left\{(U^\top x_i, y_i)\right\}_{i=1}^{n_{\text{priv}}}$ and let $\tilde{\mathcal{W}} = \left\{U^\top w : w \in \mathcal{W}\right\}$.
3: Apply (ϵ, δ)-DP subroutine, $\tilde{\mathcal{A}}$, on loss function $w \mapsto \phi_y(\langle w, x\rangle)$ with dataset $\tilde{S}_{\text{priv}}$ over the constraint set $\tilde{\mathcal{W}}$, to get $\tilde{w} \in \mathbb{R}^{\dim(\mathcal{W} \cap \text{span}(S_{\text{pub}}))}$.

Output: $\widehat{w} = U\tilde{w}$.

[5] This claim is based on a simplification of their theorem statement. Specifically, because $n_{\text{pub}} \leq n\epsilon/d$, which also implies $d \leq n\epsilon$, their lower bound $\Omega\left(R \min\left\{\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{d}{n\epsilon} + \frac{1}{\sqrt{n}}\right\}\right)$ simplifies.

Our main result for convex Lipschitz losses is the following.

Theorem 3. *Let $\epsilon > 0, \delta > 0$ and $\epsilon \leq \log(1/\delta)$. For a G-Lipschitz, B-bounded convex loss function, Algorithm 1 satisfies (ϵ, δ)-PA-DP. If the private subroutine $\tilde{\mathcal{A}}$ guarantees the following, with probability at least $1-\beta$,*

$$\widehat{L}(\tilde{\mathcal{A}}(\tilde{S}_{priv}); \tilde{S}_{priv}) - \min_{w \in \tilde{\mathcal{W}}} \widehat{L}(w; \tilde{S}_{priv}) = O\left(GD\|\mathcal{X}\|\left(\frac{\sqrt{n_{pub}\log(1/\delta)} + \sqrt{\log(1/\beta)}}{n_{priv}\epsilon}\right)\right) \quad (1)$$

then with $n_{pub} = \tilde{O}\left(\frac{n_{priv}\epsilon}{(\log(2/\beta)+\log(1/\delta))^{1/2}}\right)$, with probability at least $1-\beta$, $L(\widehat{w}; \mathcal{D}) - L(w^*; \mathcal{D})$ is

$$O\left(GD\|\mathcal{X}\|\left(\frac{\sqrt{\log(4/\beta)}}{\sqrt{n_{priv}}} + \frac{(\log(2/\beta)+\log(1/\delta))^{1/4}}{\sqrt{n_{priv}\epsilon}}\right) + \frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{priv}}}\right).$$

We note that DP algorithms such as projected noisy SGD [BST14] and the regularized exponential mechanism [GLL22], both of which can be implemented efficiently, are can be used to achieve (1), since the projected problem is at most n_{pub} dimensional.

The above result shows that in the usual regime of $\epsilon = \Theta(1)$, there is no *price* of privacy, thereby obtaining the non-private rate of $O\left(\frac{1}{\sqrt{n_{\text{priv}}}}\right)$. We contrast this with the rate of $O\left(\frac{1}{\sqrt{n_{\text{priv}}}} + \frac{\sqrt{d}}{n_{\text{priv}}\epsilon}\right)$, achievable without public data. Our result is better when $d \geq n_{\text{priv}}\epsilon$, which is the interesting regime since herein the private error dominates the non-private error. Further, our lower bound (Theorem 4 below) shows that this is the non-trivial regime (for any $\epsilon = O(1)$), since otherwise, even with unlimited public data, the optimal rate is achieved without using any of it. We also note that the above rate is achievable without public data, but in the unconstrained setting where the output $\widehat{w}$ can have very large norm and so may lie outside $\mathcal{W}$ [ABG+22].

The proof of the result primarily follows from the more general result with fat-shattering hypothesis classes (Theorem 7). We provide the key ideas as well as some details pertaining to linear predictors in Section 4.2 after Theorem 7. The full proof of this result is deferred to Appendix C.

Lower Bounds. The above rate as well as the number of public samples used are nearly-optimal. The first claim is due to the following result, which gives a lower bound on excess risk of DP algorithms under full knowledge of the marginal distribution, for Lipschitz GLMs. As unlabeled public data can only reveal information about the marginal distribution, this shows that further unlabeled public samples cannot hope to improve the rate we give in Theorem 3.

Theorem 4. *Let $\epsilon \leq 1, \delta \leq \epsilon$ and $\mathcal{A}$ be an (ϵ, δ)-DP algorithm. There exists a G-Lipschitz convex GLM loss function, and joint distribution $\mathcal{D}$ such that given a dataset S comprising n i.i.d. samples from $\mathcal{D}$ and full knowledge of the marginal distribution $\mathcal{D}_{\mathcal{X}}$, we have the following:* $\mathbb{E}_{\mathcal{A},S}\left[L(\mathcal{A}(S); \mathcal{D}) - \min_{w:\|w\| \leq D} L(w; \mathcal{D})\right] = \Omega\left(GD\|\mathcal{X}\|\left(\frac{1}{\sqrt{n}} + \min\left\{\frac{1}{\sqrt{n\epsilon}}, \frac{\sqrt{d}}{n\epsilon}\right\}\right)\right)$.

We note that the bound with $\frac{\sqrt{d}}{n_{\text{priv}}\epsilon}$ can be achieved without using any public data via standard results [BFTGT19, ABG+22]. This result is largely a corollary of [ABG+22, Theorem 6]. We provide full details in Appendix C.3.1.

To establish optimality of public sample complexity, we give the following lower bound which shows that $\tilde{\Omega}(n_{\text{priv}}\epsilon)$ samples are necessary to achieve the above rate. See Appendix C.3.2 for proof.

Theorem 5. *Let $n_{priv}, n_{pub}, d \in \mathbb{N}, \epsilon \leq 1, \delta < \frac{1}{16dn}$ and $d = \omega(n_{priv}\epsilon)$. If there exists an (ϵ, δ)-PA-DP algorithm $\mathcal{A}$, which, for any G-Lispschitz convex GLM, achieves excess risk $\mathbb{E}\left[L(\mathcal{A}(X_{pub}, S_{priv}); \mathcal{D}) - \min_{w:\|w\| \leq D} L(w; \mathcal{D})\right] = O\big(GD\|\mathcal{X}\|\left(\frac{1}{\sqrt{n_{priv}}} + \frac{\sqrt{\log(1/\delta)}}{\sqrt{n_{priv}\epsilon}}\right)\big)$, for $S_{priv} \sim \mathcal{D}^{n_{priv}}$ and $X_{pub} \sim \mathcal{D}_{\mathcal{X}}^{n_{pub}}$, then $n_{pub} = \Omega(\frac{n_{priv}\epsilon}{\log(1/\delta)})$.*

Optimistic rates. We now consider additional assumptions that the loss function is non-negative and H-smooth, such as in the case of linear regression where $\phi_y(a) = (a-y)^2$. This is a well-studied setting [SST10] especially since it allows for obtaining optimistic rates: those that interpolate between a slow worst-case rate and a faster rate under (near) realizability or interpolation conditions. The main result is the following.

Theorem 6. *Let $\epsilon > 0, \delta > 0$ and $\epsilon \leq \log(1/\delta)$. For a G-Lipschitz, B-bounded non-negative H-smooth loss function, Algorithm 1 satisfies (ϵ, δ)-PA-DP. If the private subroutine $\tilde{\mathcal{A}}$ guarantees Equation (1) with probability at least $1-\beta$, then with $n_{pub} = \tilde{O}\Big(\frac{(HD\|\mathcal{X}\|)^{2/3}(n_{priv}\epsilon)^{2/3}}{G^{2/3}(\log(1/\delta))^{1/3}} + \frac{\sqrt{H}n_{priv}\epsilon\sqrt{\widehat{L}(\widehat{w}^*;S_{priv})}}{G\sqrt{\log(1/\delta)}}\Big)$, with probability at least $1-\beta$,*

$$L(\widehat{w};\mathcal{D}) - \widehat{L}(\widehat{w}^*;S_{priv}) = \tilde{O}\left(\left(\frac{\sqrt{H}D\|\mathcal{X}\|}{\sqrt{n_{priv}\epsilon}} + \sqrt{\frac{B}{n_{priv}}}\right)\sqrt{\widehat{L}(\widehat{w}^*;S_{priv})} + \frac{H^{1/4}D\|\mathcal{X}\|\sqrt{G}\widehat{L}(\widehat{w}^*;S_{priv})^{1/4}}{\sqrt{n_{priv}\epsilon}}\right)$$
$$+ \tilde{O}\left(\frac{GD\|\mathcal{X}\|}{n_{priv}\epsilon} + \left(\frac{\sqrt{H}D^2\|\mathcal{X}\|^2 G}{n_{priv}\epsilon}\right)^{2/3} + \frac{H\|\mathcal{X}\|^2 D^2}{n_{priv}\epsilon} + \frac{B}{n_{priv}}\right)$$

where $\widehat{w}^*$ is the minimizer of $\widehat{L}$ w.r.t S_{priv} and $\tilde{O}$ hides poly$(\log(1/\delta), \log(1/\beta))$ terms.

A similar result as above can be obtained with $\widehat{L}(\widehat{w}^*;S_{\text{priv}})$ replaced by $L(w^*;\mathcal{D})$ above – see Theorem 14 for the full theorem statement. This rate, in the worst-case, is essentially the same as that of Theorem 3, which is $\tilde{O}\left(\frac{1}{\sqrt{n_{\text{priv}}}} + \frac{1}{\sqrt{n_{\text{priv}}\epsilon}}\right)$. However, optimistically, when $L(w^*;\mathcal{D})$ or $\widehat{L}(\widehat{w}^*;S_{\text{priv}})$ is small, we get a faster rate of $\tilde{O}\big(\frac{1}{n_{\text{priv}}} + \frac{1}{(n_{\text{priv}}\epsilon)^{2/3}}\big)$. We note that this is seemingly weaker than what is known in the unconstrained setting, where [ABG+22] obtained a worst-case rate of $\tilde{O}\big(\frac{1}{\sqrt{n_{\text{priv}}}} + \frac{1}{(n_{\text{priv}}\epsilon)^{2/3}}\big)$. We show that we can recover this faster rate under an extra assumption that the global minimizer of the risk, lies in the constraint set $\mathcal{W}$ – note that this is trivially true in the unconstrained setup; see Theorem 15 for the statement.

We note that projected noisy SGD [BST14] and the regularized exponential mechanism [GLL22], both of which can be implemented efficiently, are possible choices for the private sub-routine $\tilde{\mathcal{A}}$ that realize the above theorem statements.

4.2 PA-DP Supervised learning of Fat-Shattering Classes

In this section, we consider a general supervised learning setting with fat-shattering hypothesis classes and potentially non-convex losses, with unlabeled public data. Our proposed algorithm is similar to that of [ABM19], which uses the pubic unlabeled data to construct a small finite, yet representative, subset of the hypothesis class. Our construction uses a cover of the hypothesis class with respect to the ℓ_2 distance of predictions on the public data points. We then use the exponential mechanism to privately select a hypothesis using the empirical loss on private data as the score function.

We note that we operate under the pure DP setting (as opposed to approximate DP). Our techniques are based on selection which do not exhibit improved guarantees under approximate DP. Further, we note that, without public data, with non-convex losses, there is no separation of optimal rates between pure and approximate DP [GTU23].

Algorithm 2 Supervised private learning with public unlabeled data

Input: Datasets X_{pub} and S_{priv}, privacy parameter $\epsilon > 0$, scale of cover $\alpha > 0$, $\gamma > 0$.

1: Construct $\tilde{\mathcal{H}}$, a minimal α-cover of $\mathcal{H}$, with respect to the following metric

$$\|h_1 - h_2\|_{2,X_{\text{pub}}} = \sqrt{\frac{1}{n_{\text{pub}}}\sum_{x\in X_{\text{pub}}}(h_1(x) - h_2(x))^2}$$

2: Return $\widehat{h}$ sampled with probability $p(h) \propto \exp\left(-\gamma\widehat{L}(h;S_{\text{priv}})\right)$ over $h \in \tilde{\mathcal{H}}$

Our main result for the Lispchitz setting is the following.

Theorem 7. *Algorithm 2 with $\gamma = \frac{2\min(B,GR)}{n_{priv}\epsilon}$ satisfies ϵ-PA-DP. For any $\alpha > 0$ and $n_{pub} = O\left(\max\left(\frac{R^2\log(2/\beta)}{\alpha^2}, \min\left\{m : log^3(m)\mathfrak{R}_m^2(\mathcal{H}) \leq \alpha^2\right\}\right)\right) < \infty$, with probability at least $1-\beta$, we have $L(\widehat{h};\mathcal{D}) - \min_{h\in\mathcal{H}} L(h;\mathcal{D})$ is at most*

$$2G\mathfrak{R}_{n_{priv}}(\mathcal{H}) + O\left(\frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{priv}}}\right) + \tilde{O}\left(\frac{\min(B, GR)(fat_{c\alpha}(\mathcal{H}) + \log(4/\beta))}{n_{priv}\epsilon}\right) + 2G\alpha,$$

where c is an absolute constant.

Our result shows that the model of PA-DP with unlabeled public data allows for obtaining non-trivial rates for supervised learning with any fat-shattering class, as is the case in the non-private setting. Further, in many standard settings, such as that of (Euclidean) GLMs, the Rademacher complexity is $\mathfrak{R}_m(\mathcal{H}) = O\left(\frac{1}{\sqrt{m}}\right)$ which implies that $\text{fat}_\alpha(\mathcal{H}) = O\left(\frac{1}{\alpha^2}\right)$ (see Theorem 9). In those cases, our guarantee simplifies to essentially yield a rate of $O\left(\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H}) + \frac{1}{(n_{\text{priv}}\epsilon)^{1/3}} + \frac{1}{\sqrt{n_{\text{priv}}}}\right)$ – see Corollary 4 for the exact statement for GLMs.

Proof Idea. We briefly discuss some main ideas in the proof. The key is to show that if $\tilde{\mathcal{H}}$ is a cover of $\mathcal{H}$ with respect to the *empirical distance* on public feature vectors, $\|\cdot\|_{2,X_{\text{pub}}}$, then with enough public feature vectors, it is also a cover with respect to the *population distance* $\|\cdot\|_{2,\mathcal{D}_\mathcal{X}}$. This is captured in the following result.

Lemma 1. *Let $\tilde{\mathcal{H}}$ be a τ-cover of $\mathcal{H}$ with respect to $\|\cdot\|_{2,X_{pub}}$. For $n_{pub} = O\big(\max\left(\frac{R^2\log(1/\beta)}{\alpha^2}, \min\left\{m : log^3(m)\mathfrak{R}_m^2(\mathcal{H}) \le \alpha^2\right\}\right)\big) < \infty$, for every $h \in \mathcal{H}$, with probability at least $1-\beta$, there exists $\tilde{h} \in \tilde{\mathcal{H}}$ such that $\|h - \tilde{h}\|_{2,\mathcal{D}_\mathcal{X}} \le \alpha + \tau$.*

This result allows us to appropriately approximate a hypothesis class with enough public unlabeled points. This approximation roughly translates to the same additive error in the final bound while concurrently allowing for the use of the smaller finite hypothesis class $\tilde{\mathcal{H}}$ of size $|\tilde{\mathcal{H}}| = \tilde{O}(\text{fat}_\tau(\mathcal{H}))$.

For linear predictors with convex losses, as in Theorem 3, we show that the $\text{span}(X_{\text{pub}}) \cap \mathcal{W}$ is a valid 0-cover w.r.t. $\|\cdot\|_{2,X_{\text{pub}}}$. However, the cover being continuous and convex allows application of convex optimization techniques (as opposed to selection, as above), thereby obtaining stronger results with efficient procedures. The above procedure yields optimistic rates for non-negative and smooth losses; see Theorem 17 for details.

4.2.1 Application: Neural Networks

In this section, we instantiate our general result to give a guarantee for learning feed-forward neural networks in the PA-DP setting. We use the result of [GRS18] but note that other results which give bounds on the Rademacher complexity of neural networks, such as [BFT17, Sel23] can also be used.

We consider a depth M feed-forward neural network which implements the function $x \mapsto W_M(\sigma(W_{M-1}\ldots\sigma(W_1x))\ldots)$. Here, $W_1, W_2, \ldots, W_M$ are the weight matrices and σ is a (non-linear) activation function. We consider 1-Lipschtiz positive-homogeneous activation such as the ReLU function, $\sigma(z) = \max(0, z)$, applied coordinate-wise. Our main result is the following.

Corollary 2. *Let $(R_j)_{j=1}^M$ be a sequence of scalars and $M \in \mathbb{N}$. In the setting of Theorem 7 with $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \le \|\mathcal{X}\|\}$ and $\mathcal{H}$ being the class of depth M feed-forward neural networks, with 1-Lipschitz positive-homogenous activation, and weight matrices, bounded as $\|W_j\|_F \le R_j$, with $n_{pub} = \tilde{O}\left((\|\mathcal{X}\|(\prod_{j=1}^M R_j))^{2/3}(n_{priv}\epsilon)^{2/3}M^{1/3}\log(2/\beta)\right)$, with probability at least $1-\beta$, $L(\widehat{h}; \mathcal{D}) - \min_{h\in\mathcal{H}} L(h; \mathcal{D})$ is at most*

$$O\left(\frac{G\|\mathcal{X}\|\sqrt{M}\prod_{j=1}^M R_j}{\sqrt{n_{priv}}} + \frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{priv}}} + \frac{B\log(4/\beta)}{n_{priv}\epsilon}\right) + \tilde{O}\left(\left(\frac{BG^2M\|\mathcal{X}\|^2(\prod_{j=1}^M R_j)^2}{n_{priv}\epsilon}\right)^{1/3}\right).$$

We note that the above result has a polynomial dependence on the depth M, which is a consequence of the (non-private) Rademacher complexity of [GRS18]. It is also possible to get fully size-independent bounds by utilizing such existing results, however they require more stringent norm bounds on the weight matrices [Sel23]. Further, a similar result follows for non-negative smooth losses from [SST10], but we omit this extension for brevity.

4.2.2 Application: Non-Euclidean GLMs

In the non-Euclidean GLM setting, we consider $(\mathbb{X}, \|\cdot\|)$ as a d dimensional (where $d \in \mathbb{N} \cup \{\infty\}$) Banach space, and $(\mathbb{W}, \|\cdot\|_*)$ is its dual space. The feature vectors x are bounded as $\mathcal{X} = \{x \in \mathbb{X} : \|x\| \leq \|\mathcal{X}\|\}$ and $\mathcal{W} \subseteq \{w \in \mathbb{W} : \Delta(w) \leq D^r\}$ where Δ is a r-uniformly convex function[6] with respect to $\|\cdot\|_*$. A canonical example is the (ℓ_p, ℓ_q)-setup [KST08, FGV17], wherein the functions $\Delta(w) = \frac{\log(d)}{2}\|w\|^2_{1+(1/\log(d))}$, $\Delta(w) = \frac{1}{2(p-1)}\|w\|_p^2$ and $\Delta(w) = \frac{2^{p-2}}{p}\|w\|_p^p$ are 2, 2 and p-uniformly convex with respect to $\|\cdot\|_p$ for $p = 1$, $1 < p \leq 2$ and $p \geq 2$ respectively.

The GLM loss function $\ell(w; x, y) = \phi_y(\langle w, x\rangle)$ where $\langle\cdot,\cdot\rangle : \mathcal{X} \times \mathcal{W} \to \mathbb{R}$ is a duality pairing. In this case, the Rademacher complexity of linear functions, is bounded as $O\left(\frac{D\|\mathcal{X}\|}{m^{1/r}}\right)$, where s is the conjugate of r i.e. $\frac{1}{r} + \frac{1}{s} = 1$ (see, e.g. [FGV17]). We obtain the following result by instantiating Theorem 7 with the Rademacher complexity and fat-shattering dimension of non-Euclidean GLMs.

Corollary 3. *In the setting of Theorem 7, together with* $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\| \leq \|\mathcal{X}\|\right\}$ *and* $\mathcal{H} = \{x \mapsto \langle w, x\rangle, x \in \mathcal{X}, \Delta(w) \leq D^r\}$. *Given* $n_{pub} = \tilde{O}\left((n_{priv}\epsilon)^{r/(r+1)}\log(2/\beta)\right)$, *with probability at least* $1-\beta$, $L(\widehat{w}; \mathcal{D}) - \min_{w\in\mathcal{W}} L(w; \mathcal{D})$ *is at most*

$$\tilde{O}\left(GD\|\mathcal{X}\|\left(\frac{1}{{n_{priv}}^{1/r}} + \frac{\sqrt{\log(4/\beta)}}{\sqrt{n_{priv}}} + \frac{\log(2/\beta)}{(n_{priv}\epsilon)^{\frac{1}{r+1}}} + \frac{\log(4/\beta)}{n_{priv}\epsilon}\right) + \frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{priv}}}\right).$$

The above yields guarantees for the special case of (ℓ_p, ℓ_q)-setup with $r = \max\{2, p\}$. We remind that in the (constrained) *convex* Euclidean GLM setting, our dimension-independent rate in Theorem 3, with public unlabeled data recover the rates which were known to be achievable in the unconstrained setting. Further the above rate for $p = 1$ case can be used to obtain guarantees for the polyhedral setting with $\|w\|_1 \leq D$ constraint, resulting in a $O\left(\sqrt{\frac{\log(d)}{n_{\text{priv}}}} + \left(\frac{\log(d)}{n_{\text{priv}}\epsilon}\right)^{1/3}\right)$ rate. We note that [BGM21] showed a rate of $\tilde{O}\left(\sqrt{\frac{\log(d)}{n}} + \frac{\sqrt{\log(d)}}{\sqrt{n\epsilon}}\right)$ for this setting, with convex losses without public data. Importantly, for the other cases, i.e. $p > 1, p \neq 2$, there are no such (nearly) dimension-independent analogs of our result without public data, as of yet.

Acknowledgments and Disclosure of Funding

R. Arora's and E. Ullah's research was supported, in part, by NSF BIGDATA award IIS-1838139 and NSF CAREER award IIS-1943251. R. Bassily's and M. Menart's research was supported by NSF CAREER Award 2144532 and, in part, by NSF Award 2112471. C. Guzmán's research was partially supported by INRIA Associate Teams project, ANID FONDECYT 1210362 grant, ANID Anillo ACT210005 grant, and National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

References

[ABDCBH97] Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

[ABG+22] Raman Arora, Raef Bassily, Cristóbal Guzmán, Michael Menart, and Enayat Ullah. Differentially private generalized linear models revisited. *Advances in Neural Information Processing Systems*, 35:22505–22517, 2022.

[ABM19] Noga Alon, Raef Bassily, and Shay Moran. Limits of private learning with access to public data. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[6] A differentiable map Ψ is r-uniformly convex w.r.t. $\|\cdot\|$, if $\Psi(u') \geq \Psi(u) + \langle\nabla\Psi(u), u' - u\rangle + \frac{1}{r}\|u - u'\|^r$.

[AGM+22] Ehsan Amid, Arun Ganesh, Rajiv Mathews, Swaroop Ramaswamy, Shuang Song, Thomas Steinke, Thomas Steinke, Vinith M Suriyakumar, Om Thakkar, and Abhradeep Thakurta. Public data-assisted mirror descent for private model training. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 517–535. PMLR, 17–23 Jul 2022.

[BCM+20] Raef Bassily, Albert Cheu, Shay Moran, Aleksandar Nikolov, Jonathan Ullman, and Steven Wu. Private query release assisted by public data. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 695–703. PMLR, 13–18 Jul 2020.

[BDBC+23] Shai Ben-David, Alex Bie, Clement Louis Canonne, Gautam Kamath, and Vikrant Singhal. Private distribution learning with public data: The view from sample compression. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[BFT17] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

[BFTGT19] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in Neural Information Processing Systems*, 32, 2019.

[BGM21] Raef Bassily, Cristóbal A Guzmán, and Michael Menart. Differentially private stochastic optimization: New results in convex and non-convex settings. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[BKS22] Alex Bie, Gautam Kamath, and Vikrant Singhal. Private estimation with public data. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[BLW94] Peter L Bartlett, Philip M Long, and Robert C Williamson. Fat-shattering and the learnability of real-valued functions. In *Proceedings of the seventh annual conference on Computational learning theory*, pages 299–310, 1994.

[BNS13] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In Prasad Raghavendra, Sofya Raskhodnikova, Klaus Jansen, and José D. P. Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 363–378, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

[BTGT18] Raef Bassily, Om Thakkar, and Abhradeep Guha Thakurta. Model-agnostic private learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[BUV14] Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '14, page 1–10, New York, NY, USA, 2014. Association for Computing Machinery.

[BWZK22] Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Differentially private bias-term only fine-tuning of foundation models. In *NeurIPS 2022 Workshop on Trustworthy and Socially Responsible Machine Learning (TSRML)*, 2022.

[CH11] Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In Sham M. Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 155–186, Budapest, Hungary, 09–11 Jun 2011. PMLR.

[CWZ21] T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825 – 2850, 2021.

[DMNS06] 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, 2006.

[DSS+15] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 650–669, 2015.

[Duc23] John Duchi. Lecture notes on statistics and information theory, May 2023.

[FGV17] Vitaly Feldman, Cristobal Guzman, and Santosh Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1265–1277. SIAM, 2017.

[FS17] Vitaly Feldman and Thomas Steinke. Generalization for adaptively-chosen estimators via stable median. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 728–757. PMLR, 07–10 Jul 2017.

[GHN+23] Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Why is public pretraining necessary for private model training? In *International Conference on Machine Learning*, 2023.

[GLL22] Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. In *Conference on Learning Theory*, pages 1948–1989. PMLR, 2022.

[GRS18] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

[GTU23] Arun Ganesh, Abhradeep Thakurta, and Jalaj Upadhyay. Universality of langevin diffusion for private optimization, with applications to sampling from rashomon sets. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1730–1773. PMLR, 2023.

[KDRT21] Peter Kairouz, Monica Ribero Diaz, Keith Rush, and Abhradeep Thakurta. (nearly) dimension independent private erm with adagrad rates
via publicly estimated subspaces. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2717–2746. PMLR, 15–19 Aug 2021.

[KST08] Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21, 2008.

[KU20] Gautam Kamath and Jonathan Ullman. A primer on private statistics. *arXiv preprint arXiv:2005.00010*, 2020.

[LLHR23] Andrew Lowy, Zeman Li, Tianjian Huang, and Meisam Razaviyayn. Optimal differentially private learning with public data. *arXiv preprint arXiv:2306.15056v1*, 2023.

[LLHR24] Andrew Lowy, Zeman Li, Tianjian Huang, and Meisam Razaviyayn. Optimal differentially private model training with public data. *arXiv preprint arXiv:2306.15056v2*, 2024.

[LTLH22] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022.

[LW19] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation, 2019.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[NMT+23] Milad Nasr, Saeed Mahloujifar, Xinyu Tang, Prateek Mittal, and Amir Houmansadr. Effectively using public data in privacy preserving machine learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25718–25732. PMLR, 23–29 Jul 2023.

[PAE+17] Nicolas Papernot, Martin Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations*, 2017.

[PHYS24] Francesco Pinto, Yaxi Hu, Fanny Yang, and Amartya Sanyal. PILLAR: How to make semi-private learning more effective. In *2nd IEEE Conference on Secure and Trustworthy Machine Learning*, 2024.

[PSM+18] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with PATE. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[PVX+23] Natalia Ponomareva, Sergei Vassilvitskii, Zheng Xu, Brendan McMahan, Alexey Kurakin, and Chiyaun Zhang. How to dp-fy ml: A practical tutorial to machine learning with differential privacy. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5823–5824, New York, NY, USA, 2023. Association for Computing Machinery.

[RS12] Alexander Rakhlin and Karthik Sridharan. Statistical learning theory and sequential prediction. *Lecture Notes in University of Pennsyvania*, 44, 2012.

[SCZ+20] Dianbo Sui, Yubo Chen, Jun Zhao, Yantao Jia, Yuantao Xie, and Weijian Sun. FedED: Federated learning via ensemble distillation for medical relation extraction. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2118–2128, Online, November 2020. Association for Computational Linguistics.

[Sel23] Mark Sellke. On size-independent sample complexity of relu networks. *arXiv preprint arXiv:2306.01992*, 2023.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[SST10] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*, 2010.

[SSTT21] Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private glms. In *International Conference on Artificial Intelligence and Statistics*, pages 2638–2646. PMLR, 2021.

[SU15] Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7, 01 2015.

[VC71] N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264, 1971.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[YNB+22a] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2022.

[YNB+22b] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2022.

[YZCL21] Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. In *International Conference on Learning Representations*, 2021.

[ZWB21] Yingxue Zhou, Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private sgd with gradient subspace identification. In *International Conference on Learning Representations*, 2021.

A Additional Preliminaries

Theorem 8. *[RS12, Theorem 12.7] For $\mathcal{H} \subseteq [-R,R]^{\mathcal{X}}$, $m \in \mathbb{N}$, $p \geq 1$, $0 < \alpha \leq R$, we have that*

$$\mathcal{N}_2(\mathcal{H}, \alpha, m) \leq \left(\frac{2R}{\alpha}\right)^{Cfat_{c\alpha}(\mathcal{H})}.$$

Further, for any $\tau \in (0,1)$,

$$\log\left(\mathcal{N}_\infty(\mathcal{H}, \alpha, m)\right) \leq C' fat(\mathcal{H}, c'\tau\alpha) \log\left(\frac{Rm}{fat(\mathcal{H}, c'\tau\alpha)\alpha}\right) log^\tau\left(\frac{m}{fat(\mathcal{H}, c'\tau\alpha)}\right),$$

where c, c', C and C' are absolute constants.

Theorem 9. *[SST10, Lemma A.3] For any hypothesis class $\mathcal{H}$, any sample size m and any $\alpha > \mathfrak{R}_m(\mathcal{H})$, we have that,*

$$fat_\alpha(\mathcal{H}) \leq \frac{4m\mathfrak{R}_m(\mathcal{H})^2}{\alpha^2}.$$

Theorem 10. *[GRS18, Theorem 1] Let $M \in \mathbb{N}$ and $(R_j)_{j=1}^M$ be a sequence of scalars. The Rademacher complexity of the class of depth M neural networks with 1-Lipschitz, positive-homogeneous activation function, $\mathcal{H}$, with weights $\|W_j\|_F \leq R_j$ is bounded as,*

$$\mathfrak{R}_m(\mathcal{H}) \leq \frac{\|\mathcal{X}\|\left(\sqrt{2\log(2)\,M}+1\right)\prod_{j=1}^M R_j}{\sqrt{m}}.$$

Here $\|\cdot\|_F$ denotes the Frobenius norm.

Lemma 2. *[DSS+15, Implied by Lemmas 5 and 14] Let $f : \{\pm 1\}^n \mapsto \mathbb{R}$ and define $g : [-1,1] \to \mathbb{R}$ as $g(p) = \mathop{\mathbb{E}}_{S\sim\mathcal{D}_p^n}[f(S)]$, where $\mathcal{D}_p$ is as defined in Appendix B. Then for $a, b \in \mathbb{R}$, $b > a$, and $\mu \sim \mathsf{Unif}([a,b])$,*

$$\mathop{\mathbb{E}}_{\mu,S}\left[f(S)\cdot\sum_{x\in S}(x-\mu)\right] = \mathop{\mathbb{E}}_{\mu}\left[g'(\mu)(1-\mu^2)\right]$$

$$= 1 - \mathop{\mathbb{E}}_{\mu}\left[\mu^2\right] + (g(b)-b)(1-b^2)\frac{1}{|b-a|} - (g(a)-a)(1-a^2)\frac{1}{|b-a|} + 2\mathop{\mathbb{E}}_{\mu}\left[(g(\mu)-\mu)\mu\right].$$

Lemma 3. *[FS17, Lemma A.1] Fix $\mu, \epsilon, \delta, \Delta \in \mathbb{R}$. Let X and Y be random variables supported on $[\mu-\Delta, \mu+\Delta]$. Suppose that X and Y are (ϵ,δ)-indistinguishable, that is for any $\mathcal{E} \subseteq \mathbb{R}$, $e^{-\epsilon}\left(\mathbb{P}\left[X \in \mathcal{E}\right] - \delta\right) \leq \mathbb{P}\left[Y \in \mathcal{E}\right] \leq e^{\epsilon}\mathbb{P}\left[X \in \mathcal{E}\right] + \delta$. Then*

$$\left|\mathbb{E}\left[X\right] - \mathbb{E}\left[Y\right]\right| \leq (e^\epsilon - 1)\mathbb{E}\left[|X-\mu|\right] + 2\delta\Delta.$$

B Missing proofs from Section 3

B.1 Lower Bounds for Mean Estimation

As stated previously, our SCO result follows primarily from new lower bound for PA-DP mean estimation. Here, we consider the setting where $\mathcal{D}$ is supported on the ℓ_2 ball of radius $R > 0$. We define the mean of $\mathcal{D}$ as $\mu(D)$. Our lower bound for PA-DP mean estimation follows much the same form as our SCO bound.

Theorem 11. *Let $\delta \leq \frac{1}{16nd}$, $\epsilon \leq 1$. For any (ϵ,δ)-PA-DP algorithm, there exists a distribution $\mathcal{D}$ such that $\mathbb{E}\left[\|\mathcal{A}(S_{pub}, S_{priv}) - \mu(\mathcal{D})\|\right] = \Omega\left(R\cdot\Psi(n_{pub}, n, \epsilon, \delta)\right)$.*

We present the full proof momentarily and provide a more detailed discussion on the challenges of establishing this lower bound in Appendix B.2. We highlight key ideas here. As with many other lower bounds in differential privacy, we leverage a construct known as fingerprinting codes [BUV14, DSS+15]. A key aspect of our analysis is showing that fingerprinting distributions can be

used to recover the optimal *non-private* lower bound for mean estimation. This allows us to create a problem which is "hard" both privately and non-privately. The analysis works by first showing that any sufficiently accurate algorithm must strongly correlate with the sampled datapoints. Next, we show upper bounds on how strongly the output of the algorithm correlates with the sampled dataset. The method for upper bounding this correlation varies depending on whether a given datapoint is considered public or private. Combining these upper and lower bounds on correlation yields the claimed result.

To obtain the $\sqrt{\log(1/\delta)}$ factor term in the lower bound, we use similar ideas to those in [SU15, CWZ21]. However, the introduction of public data leads to complications in prior methods. As such, we show that by analyzing the correlation of the coordinate wise clipping of the algorithms output, we are able to get bounds that appropriately scale with the accuracy.

Proof of Theorem 11 Before proceeding, we introduce the so-called fingerprinting distribution which will be the basis of our hard instance for mean estimation [BUV14, DSS+15]. Towards this end, for any vector $\mu \in [-1,1]^d$ we define $\mathcal{D}_\mu$ as the product distribution where, for any $j \in [d]$, a sample has its j'th coordinate as 1 with probability $(1+\mu_j)/2$ and as -1 with probability $(1-\mu_j)/2$. As shorthand, we denote $\frac{R}{\sqrt{d}}\mathcal{D}_\mu$ as the distribution which samples a vector from $\mathcal{D}_\mu$ and then scales it by $\frac{R}{\sqrt{d}}$. For notational convenience, for a set $\mathcal{E}$, we will also use $\mathsf{Unif}(\mathcal{E})$ to denote the uniform distribution over elements of the set.

The theorem follows from two theorems which have different restrictions on the problem parameters. In addition to the following two theorems, Theorem 11 incorporates the classic $\frac{R}{\sqrt{n}}$ statistical lower bound that holds even non-privately. The first theorem we present holds for a larger range of parameters but does not achieve the dependence on $\log(1/\delta)$.

Theorem 12. *Let $\epsilon > 0$, $\delta \le \frac{1}{16n}$ and $\mathcal{A}$ be an (ϵ,δ)-PA-DP algorithm. For any setting of $\min\left\{\frac{\sqrt{d}}{n_{priv}\epsilon}, \frac{1}{\sqrt{n_{pub}}}\right\} \le M \le 1$, if $\mu \sim \mathsf{Unif}([-M,M]^d)$ and $(S_{pub}, S_{priv}) \sim \frac{R}{\sqrt{d}}\mathcal{D}_\mu^n$ it holds that*

$$\mathop{\mathbb{E}}_{\mathcal{A},S,\mu}\left[\|\mathcal{A}(S_{pub},S_{priv}) - \mu(\mathcal{D})\|\right] = \Omega\left(R\min\left\{\frac{1}{\sqrt{n_{pub}}}, \frac{\sqrt{d}}{n(e^\epsilon-1)}\right\}\right).$$

In application to Theorem 11, we use $(e^\epsilon - 1) \le 2\epsilon$ whenever $\epsilon \le 1$. The second theorem requires $d \ge n\epsilon$ but has the benefit of scaling with $\log(1/\delta)$.

Theorem 13. *Let $\delta \le \frac{1}{3dn}$, $\epsilon \le 1$, $d \ge 120^2 n\epsilon$, and $n_{pub} \le \frac{n\epsilon}{120^2\log\left(1/[\sqrt{nd}\delta]\right)}$, and $\mathcal{A}$ an (ϵ,δ)-PA-DP algorithm. Then there exists $M > 0$ such for $\mu \sim \mathsf{Unif}([-M,M]^d)$ and $(S_{priv}, S_{pub}) \sim \frac{R}{\sqrt{d}}\mathcal{D}_\mu^n$ it holds that*

$$\mathop{\mathbb{E}}_{\mathcal{A},S,\mu}\left[\|\mathcal{A}(S_{pub},S_{priv}) - \mu(\mathcal{D})\|\right] = \Omega\left(R\min\left\{\frac{1}{\sqrt{n_{pub}}}, \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right\}\right).$$

A crucial part of the analysis is leveraging the so called fingerprinting lemma, which roughly states that any accurate algorithm given a dataset sampled from $\mathcal{D}_\mu$ must strongly correlate with vectors in the dataset. Particularly pertinent to our analysis is achieving such a correlation even when the components of the mean μ are much smaller than 1. Towards this end, we leverage the robust distribution framework of [DSS+15] to achieve the following version of the fingerprinting lemma.

Lemma 4 (Fingerprinting Lemma)**.** *Let $M \in [0,1]$ and μ be sampled uniformly from $[-M,M]^d$. Let $\mathcal{A}$ satisfy $\mathop{\mathbb{E}}_{S\sim\mathcal{D}_\mu^n}\left[\|\mathcal{A}(S)-\mu\|\right] \le \alpha$ (for any $\mu \in [-1,1]^d$). Then one has*

$$\mathop{\mathbb{E}}_{\mathcal{A},S,\mu}\left[\sum_{i=1}^n \langle \mathcal{A}(S), x_i - \mu\rangle\right] \ge \frac{2d}{3} - \frac{\alpha\sqrt{d}}{M} - 2M\sqrt{d}\alpha.$$

Proof. In the following we treat $\mathcal{A}$ as a deterministic function and bound $\mathop{\mathbb{E}}_{S,\mu}\left[\sum_{i=1}^n \langle \mathcal{A}(S), x_i - \mu\rangle\right]$. This is sufficient to bound $\mathop{\mathbb{E}}_{\mathcal{A},S,\mu}\left[\sum_{i=1}^n \langle \mathcal{A}(S), x_i - \mu\rangle\right]$ for randomized $\mathcal{A}$, since the analysis holds for

any function (i.e. the distribution does not depend on $\mathcal{A}$). Further, we start with the one dimensional case such that $\mu \in \mathbb{R}$. Define $g(\mu) = \mathop{\mathbb{E}}_{S\sim\mathcal{D}_\mu^n}[\mathcal{A}(S)]$. We start by applying results developed in [DSS+15],

$$\begin{aligned}
\mathop{\mathbb{E}}_{S,\mu}\left[\mathcal{A}(S)\sum_{i=1}^n(x_i-\mu)\right] &\overset{(i)}{=} \mathop{\mathbb{E}}_{\mu}\left[g'(\mu)(1-\mu^2)\right] \\
&\overset{(ii)}{\geq} 1 - \mathbb{E}\left[\mu^2\right] + 2\mathop{\mathbb{E}}_{\mu}\left[(g(\mu)-\mu)\mu\right] - \frac{|g(-M)+M| + |g(M)-M|}{2M} \\
&\geq 2/3 + 2\mathop{\mathbb{E}}_{\mu}\left[(g(\mu)-\mu)\mu\right] - \frac{|g(-M)+M| + |g(M)-M|}{2M}.
\end{aligned}$$

Above, (i) comes from [DSS+15, Lemma 5] and (ii) comes from [DSS+15, Lemma 14], which we have collectively restated in Lemma 2. We now have

$$\begin{aligned}
\mathop{\mathbb{E}}_{S,\mu}\left[\mathcal{A}(S)\sum_{i=1}^n(x_i-\mu)\right] &\geq 2/3 + \frac{|g(-M)+M| + |g(M)-M|}{2M} + 2\mathop{\mathbb{E}}_{\mu}\left[(g(\mu)-\mu)\mu\right] \\
&\geq 2/3 + \frac{|g(-M)-M| + |g(M)-M|}{2M} - 2\mathop{\mathbb{E}}_{\mu}\left[|g(\mu)-\mu|\cdot|\mu|\right] \\
&\geq 2/3 - \frac{|\mathbb{E}_{S\sim\mathcal{D}_{-M}}[\mathcal{A}(S)]+M| + |\mathbb{E}_{S\sim\mathcal{D}_M}[\mathcal{A}(S)]-M|}{2M} \\
&\quad - 2M\mathop{\mathbb{E}}_{\mu}\left[\left|\mathop{\mathbb{E}}_{S\sim\mathcal{D}_\mu}[\mathcal{A}(S)]-\mu\right|\right].
\end{aligned}$$

Above we use the fact that $|\mu| \leq M$ and the definition of g.

We can now extend the above analysis to higher dimensions. For $\mu \in \mathbb{R}^d$, the above holds for each $\mu_j, j \in [d]$. For convenience define $\bar{M} = (M, \ldots, M) \in \mathbb{R}^d$. Summing over d dimensions we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S,\mu}\left[\left\langle\mathcal{A}(S), \sum_{i=1}^n(x_i-\mu)\right\rangle\right] \\
&\geq \frac{2d}{3} - \frac{1}{2M}\left\|\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{-\bar{M}}}[\mathcal{A}(S)]+\bar{M}\right\|_1 - \frac{1}{2M}\left\|\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\bar{M}}}[\mathcal{A}(S)]-\bar{M}\right\|_1 - 2M\mathop{\mathbb{E}}_{\mu}\left[\left\|\mathop{\mathbb{E}}_{S\sim\mathcal{D}_\mu}[\mathcal{A}(S)]-\mu\right\|_1\right] \\
&\geq \frac{2d}{3} - \frac{1}{2M}\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{-\bar{M}}}\left[\|\mathcal{A}(S)+\bar{M}\|_1\right] - \frac{1}{2M}\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\bar{M}}}\left[\|\mathcal{A}(S)-\bar{M}\|_1\right] - 2M\mathop{\mathbb{E}}_{S,\mu}\left[\|\mathcal{A}(S)-\mu\|_1\right] \\
&\geq \frac{2d}{3} - \frac{\sqrt{d}}{2M}\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{-\bar{M}}}\left[\|\mathcal{A}(S)+\bar{M}\|_2\right] - \frac{\sqrt{d}}{2M}\mathop{\mathbb{E}}_{S\sim\mathcal{D}_{\bar{M}}}\left[\|\mathcal{A}(S)-\bar{M}\|_2\right] - 2\sqrt{d}M\mathop{\mathbb{E}}_{S,\mu}\left[\|\mathcal{A}(S)-\mu\|_2\right] \\
&\geq \frac{2d}{3} - \frac{\alpha\sqrt{d}}{M} - 2M\sqrt{d}\alpha.
\end{aligned}$$

This proves the claim. □

We now turn towards proving Theorems 12 and 13. We start with the simpler proof of Theorem 12.

Proof of Theorem 12. For our proof we will use a dataset of vectors in $\{\pm1\}^d$, and as such the ℓ_2 bound on the data is $\sqrt{d}$. The final result will follow from rescaling by $\frac{R}{\sqrt{d}}$.

Let $S = \{x_1, x_2, \ldots x_n\} = (S_{\text{pub}}, S_{\text{priv}}) \sim \mathcal{D}_\mu^n$ be the concatenation of the public and private datasets. We also define $\alpha = \mathbb{E}\left[\|\mathcal{A}(S_{\text{pub}}, S_{\text{priv}}) - \mu\|\right]$ for notational convenience.

Define the following statistics,

$$\begin{aligned}
Z_i &= \langle\mathcal{A}(S_{\text{pub}}, S_{\text{priv}}) - \mu, x_i - \mu\rangle \\
Z_i' &= \langle\mathcal{A}(S_{\text{pub}}, S_{\sim i}) - \mu, x_i - \mu\rangle.
\end{aligned}$$

where $S_{\sim i}$ is the dataset formed by replacing i-th data point of S_{priv} with $x_i' \sim \mathcal{D}_\mu$. We have,

$$\mathop{\mathbb{E}}_{\mathcal{A},S,\mu}\Big[\sum_{i=1}^{n} Z_i\Big] = \mathbb{E}\Big[\sum_{i=1}^{n_{\text{pub}}} Z_i\Big] + \mathbb{E}\Big[\sum_{i=n_{\text{pub}}+1}^{n} Z_i\Big]. \tag{2}$$

The lower bound proceeds by providing upper and lower bounds on the above sum. We first have

$$\begin{aligned}\mathbb{E}\left[\left\langle \mathcal{A}(S_{\text{pub}}, S_{\text{priv}}) - \mu, \sum_{i=1}^{n_{\text{pub}}} (x_i - \mu)\right\rangle\right] &\le \sqrt{\mathbb{E}[\|\mathcal{A}(S_{\text{pub}}, S_{\text{priv}}) - \mu\|^2]\mathbb{E}\left\|\sum_{i=1}^{n_{\text{pub}}}(x_i-\mu)\right\|^2} \\ &\le \alpha\sqrt{dn}.\end{aligned}$$

where the first inequality used Cauchy-Schwartz.

For the second term in Equation (2), we utilize differential privacy. Specifically, [FS17, Lemma A.1], restated in Lemma 3, gives that

$$\begin{aligned}\mathbb{E}\Big[\sum_{i=n_{\text{pub}}+1}^{n} Z_i\Big] &\le \sum_{i=n_{\text{pub}}+1}^{n}\left(\mathbb{E}[Z_i'] + 2(e^\epsilon - 1)\sqrt{\text{Var}(Z_i')} + 8\delta d\right) \\ &\le 4n_{\text{priv}}(e^\epsilon - 1)\alpha + 8n_{\text{priv}}\delta d.\end{aligned}$$

Above we use that $\text{Var}(Z_i') \le 4\alpha^2$ since $\|x_i - \mu\|_\infty \le 4$. Plugging the above two in Equation (2) yields,

$$\mathbb{E}\Big[\sum_{i=1}^{n} Z_i\Big] \le (4n_{\text{priv}}(e^\epsilon - 1)\alpha + 8n_{\text{priv}}\delta d) + \alpha\sqrt{dn_{\text{pub}}}.$$

We now use the fingerprinting lemma, Lemma 4, to lower bound the correlation. In this regard, note $\mathop{\mathbb{E}}_{S,\mu}[\langle \mu, \sum_{i=1}^{n}(x_i - \mu)\rangle] = 0$. Thus

$$\mathbb{E}\Big[\sum_{i=1}^{n} Z_i\Big] = \mathbb{E}\left[\left\langle \mathcal{A}(S_{\text{pub}}, S_{\text{priv}}), \sum_{i=1}^{n} x_i - \mu\right\rangle\right] \ge \frac{2d}{3} - \frac{\alpha\sqrt{d}}{2M} - 2M\sqrt{d}\alpha.$$

Plugging the obtained upper bound on the left hand side gives us,

$$\begin{aligned}&4n_{\text{priv}}(e^\epsilon - 1)\alpha + 8n_{\text{priv}}\delta d + \alpha\sqrt{dn_{\text{pub}}} \ge \frac{2d}{3} - \frac{\alpha\sqrt{d}}{M} - 2M\sqrt{d}\alpha \\ \implies &4n_{\text{priv}}(e^\epsilon - 1)\alpha + \alpha\sqrt{dn_{\text{pub}}} \ge \frac{d}{6} - \frac{\alpha\sqrt{d}}{M} - 2M\sqrt{d}\alpha \\ \implies &\alpha\left(4n_{\text{priv}}(e^\epsilon - 1) + \sqrt{dn_{\text{pub}}} + \frac{\sqrt{d}}{M} + 2M\sqrt{d}\right) \ge \frac{d}{6} \\ \implies &\alpha \ge \frac{1}{24}\min\left\{\frac{d}{n_{\text{priv}}(e^\epsilon - 1)}, \frac{\sqrt{d}}{\sqrt{n_{\text{pub}}}}, M\sqrt{d}, \frac{\sqrt{d}}{M}\right\} \\ \implies &\alpha \ge \frac{1}{24}\min\left\{\frac{d}{n_{\text{priv}}(e^\epsilon - 1)}, \frac{\sqrt{d}}{\sqrt{n_{\text{pub}}}}, M\sqrt{d}\right\}.\end{aligned}$$

Above the first implication uses the assumption that $\delta \le \frac{1}{16n_{\text{priv}}}$. The last implication uses the fact that $M\sqrt{d} \le \frac{\sqrt{d}}{M}$ since $M \le 1$. Rescaling by a $\frac{R}{\sqrt{d}}$ factor yields the bound

$$\mathop{\mathbb{E}}_{\mathcal{A},S,\mu}[\|\mathcal{A}(S_{\text{pub}}, S_{\text{priv}}) - \mu\|] \ge \frac{R}{24}\min\left\{\frac{\sqrt{d}}{n_{\text{priv}}(e^\epsilon - 1)}, \frac{1}{\sqrt{n_{\text{pub}}}}, M\right\}.$$

Observe that any setting of $M \ge \min\left\{\frac{\sqrt{d}}{n_{\text{priv}}\epsilon}, \frac{1}{\sqrt{n_{\text{pub}}}}\right\}$ realizes the bound claimed in the theorem statement. □

We now turn towards achieving a dependence on δ to prove Theorem 13. To do this, we leverage the the idea of filling a dataset with copies of each fingerprinting code seen in previous work [SU15, CWZ21]. However, in our case the introduction of public data makes this argument more delicate and leads to modified techniques for upper bounding the correlation statistics. See our discussion in Section B.2 for more details on why this is necessary.

Proof of Theorem 13. Let $\alpha = \mathbb{E}\left[\|\mathcal{A}(S_{\text{pub}}, S_{\text{priv}}) - \mu\|\right]$, $\alpha^* = \frac{1}{125}\min\left\{\frac{d\sqrt{\log(1/\delta)}}{n\epsilon}, \sqrt{\frac{d}{n_{\text{pub}}}}\right\}$ and assume by way of contradiction that $\alpha < \alpha^*$. Let $k = \frac{1}{3\epsilon}\log\left(1/[\sqrt{dn}\delta]\right)$. Let $m = \frac{n}{k}$. We set $M = \frac{4}{\sqrt{d}}(\alpha^* + \sqrt{\frac{d}{m}})$. Let $\mu \sim \mathsf{Unif}([-M, M]^d)$, $S_z = \{z_1, ..., z_m\} \sim \mathcal{D}_\mu$. Sample $S_{\text{priv}}, S_{\text{pub}} \sim \mathsf{Unif}\left(\{z_1, ..., z_m\}\right)$ and denote the combined dataset as $S = \{x_1, \ldots, x_n\} = (S_{\text{pub}}, S_{\text{priv}})$. Note that as in the proof of Theorem 12, we are starting by showing a lower bound for the case where the data is drawn from $\mathcal{D}_\mu$ instead of $\frac{R}{\sqrt{d}}\mathcal{D}_\mu$, and will rescale at the end of the proof.

To prove our lower bound, we will provide upper and lower bounds on correlation statistics w.r.t. the intermediate dataset S_z. We will also introduce a clipping procedure which helps better control the upper bound on correlation. In this regard, for each $j \in [m]$ define $Z_j = \langle \lfloor \mathcal{A}(S_{\text{pub}}, S_{\text{priv}}) \rfloor_M, z_j - \mu \rangle$, where $\lfloor v \rfloor_M$ denotes the operation of clipping every element of v to $[-M, M]$. In the following, we will provide upper and lower bounds on $\mathbb{E}\left[\sum_{j=1}^m Z_i\right]$ and use this to show that $\alpha \leq \alpha^*$ implies a contradiction.

Lower Bound on Correlation We now want to lower bound $\mathbb{E}\left[\sum_{j=1}^m Z_j\right]$. Towards this end, we can apply fingerprinting lemma, Lemma 4, to the algorithm which outputs the clipping. For $\widehat{\alpha} > 0$, if $\mathop{\mathbb{E}}_{\mathcal{A},S}\left[\|\lfloor \mathcal{A}(S_{\text{pub}}, S_{\text{priv}}) \rfloor_M - \mu\|\right] \leq \widehat{\alpha}$, then this yields,

$$\mathop{\mathbb{E}}_{\mathcal{A},S,\mu}\left[\sum_{j=1}^m Z_i\right] \geq \frac{2d}{3} - \frac{\widehat{\alpha}\sqrt{d}}{M} - 2M\sqrt{d}\widehat{\alpha}.$$

Now observe that

$$\begin{aligned}
\mathbb{E}\left[\|\lfloor \mathcal{A}(S_{\text{pub}}, S_{\text{priv}}) \rfloor_M - \mu\|\right] &\leq \mathbb{E}\left[\|\mathcal{A}(S_{\text{pub}}, S_{\text{priv}}) - \mu\|\right] \\
&\leq \mathbb{E}\left[\left\|\mathcal{A}(S_{\text{pub}}, S_{\text{priv}}) - \frac{1}{m}\sum_{z \in S_z} z\right\|\right] + \mathbb{E}\left[\left\|\frac{1}{m}\sum_{z \in S_z} z - \mu\right\|\right] \\
&\leq \alpha^* + \sqrt{\frac{d}{m}}
\end{aligned}$$

In the last step we use the assumed contradiction that $\alpha \leq \alpha^*$. Thus it suffices to set $\widehat{\alpha} = \alpha^* + \sqrt{\frac{d}{m}}$. Now by the setting $M = \frac{4}{\sqrt{d}}\widehat{\alpha}$ and $\widehat{\alpha} \leq \frac{\sqrt{d}}{12}$, Eqn. (3) implies

$$\mathop{\mathbb{E}}_{\mathcal{A},S,\mu}\left[\sum_{j=1}^m Z_i\right] \geq \frac{2d}{3} - \frac{d}{4} - 8\widehat{\alpha}^2 \geq \frac{d}{3}. \tag{3}$$

Bounding the Number of Copies in the Dataset We now turn towards the more involved process of upper bounding $\mathop{\mathbb{E}}_{S,\mu}\left[\sum_{i=1}^n Z_i\right]$. To do this however, it will be first helpful to show that no datapoint in S_z is copied into S too many times.

The first step is showing that no point is copied too many times into S. For $j \in [m]$, let $\mathcal{Z}_j = \{i \in [n] : x_i = z_j\}$. Observe

$$\begin{aligned}\mathbb{P}\left[\exists j \in [m] : |\mathcal{Z}_j| \geq (\tau+1)k\right] &\leq \sum_{j=1}^{m} \mathbb{P}\left[|\mathcal{Z}_j| \geq (\tau+1)k\right] \\ &\leq m \exp\left(-\frac{3\tau^2 n}{4m(1-1/n)}\right) \\ &\leq m \exp\left(-\frac{3\tau^2 \log(1/\delta)}{8\epsilon}\right).\end{aligned}$$

The second inequality follows from Bernstein's inequality for the sum of n Bernoulli random variables with mean $1/m$ and the fact that $\mathbb{E}\left[|Z_j|\right] = \frac{n}{m} = k$. Set $\tau = \sqrt{\frac{8\epsilon \log(dm)}{3\log(1/\delta)}}$ and note since $\epsilon \leq 1$ and $\log(dm) \leq \log(dn) \leq \log(1/\delta)$ (since $\delta \leq \frac{1}{dn}$), we have that $\tau \leq 2$. Thus, denoting E as the event where no point in S_z is copied into S more than $3k$ times, we establish

$$\mathbb{P}\left[E^c\right] = \mathbb{P}\left[\exists j \in [m] : |\mathcal{Z}_j| \geq 3k\right] \leq \frac{1}{d}. \tag{4}$$

Upper Bound on Correlation Under our model, we assume that $\mathcal{A}$ must treat all data in S_{priv} as private. We will in fact only need to use the privacy property for a subset of samples in S_{priv} to prove the correlation upper bound. Let $\mathcal{I}_{\text{priv}} \subseteq [m]$ denote the set of indices s.t. $j \in \mathcal{I}_{\text{priv}}$ if every copy of z_j sampled into the overall dataset is in the private dataset S_{priv}; that is $\mathcal{I}_{\text{priv}} = \{j : (\forall x \in S_{\text{pub}})\ x \neq z_j\}$. Let $\mathcal{I}_{\text{pub}} = [m] \setminus \mathcal{I}_{\text{priv}}$. Observe that $\mathcal{I}_{\text{priv}}$ may contain indices for points in S_z which are never sampled into S. We will see this does not affect our analysis.

We have

$$\mathbb{E}\left[\sum_{j=1}^{m} Z_i\right] = \mathbb{E}\left[\sum_{j \in \mathcal{I}_{\text{priv}}} \langle \lfloor \mathcal{A}(S_{\text{pub}}, S_{\text{priv}}) \rfloor_M, z_j - \mu \rangle + \sum_{j \in \mathcal{I}_{\text{pub}}} \langle \lfloor \mathcal{A}(S_{\text{pub}}, S_{\text{priv}}) \rfloor_M, z_j - \mu \rangle\right].$$

The first term on the RHS can be bounded using the privacy property of $\mathcal{A}$. For any fixed $j \in \mathcal{I}_{\text{priv}}$, let S'_{priv} denote the dataset which replaces every instance of z_j in S_{priv} with a copied fresh sample from $\mathcal{D}_\mu$. By the above analysis, conditional on the event E, at most $3k$ such points need to be replaced conditional on the event E. Since $\mathcal{A}(S_{\text{pub}}, S'_{\text{priv}})$ is independent of z_j, by the Chernoff Hoeffding bound,

$$\mathbb{P}\left[\left\langle \lfloor \mathcal{A}(S_{\text{pub}}, S'_{\text{priv}}) \rfloor_M, z_j - \mu \right\rangle \geq \tau \mid E\right] \leq \exp\left(-\frac{\tau^2}{8dM^2}\right). \tag{5}$$

Since $\mathcal{A}$ satisfies k-group privacy with parameters $\widehat{\epsilon} \leq 3k\epsilon$ and $\widehat{\delta} = e^{3k\epsilon}\delta$, we have

$$\mathbb{P}\left[\langle \lfloor \mathcal{A}(S_{\text{pub}}, S_{\text{priv}}) \rfloor_M, z_j - \mu \rangle \geq \tau \mid E\right] \leq \exp\left(\widehat{\epsilon} - \frac{\tau^2}{2dM^2}\right) + \widehat{\delta}.$$

Setting $\tau = M\sqrt{d\log(1/\delta)}$, we obtain

$$\begin{aligned}\mathbb{E}\left[Z_j \mid E\right] &\leq \tau + 2d\mathbb{P}\left[Z_j \geq \tau \mid E\right] \\ &\leq M\sqrt{d\log(1/\delta)} + 2de^{\widehat{\epsilon}}\delta + \widehat{\delta} \\ &\leq M\sqrt{d\log(1/\delta)} + 3de^{3k\epsilon}\delta \leq 4M\sqrt{d\log(1/\delta)}.\end{aligned}$$

The last inequality comes from the setting of $k = \frac{1}{3\epsilon}\log\left(\frac{1}{\sqrt{dn}\delta}\right)$ and the fact that $M \geq \frac{1}{\sqrt{n}}$. Repeating this argument for each $j \in \mathcal{I}$ we get

$$\mathbb{E}\Big[\sum_{j \in \mathcal{I}_{\text{priv}}} Z_j\Big] \leq \mathbb{E}\Big[\sum_{j \in \mathcal{I}_{\text{priv}}} Z_j \;\Big|\; E\Big]\mathbb{P}\left[E\right] + mMd\mathbb{P}\left[E^c\right] \leq 5mM\sqrt{d\log(1/\delta)}.$$

The last inequality uses the bound established on each $\mathbb{E}\left[Z_j \mid E\right]$, $j \in \mathcal{I}_{\text{priv}}$, above and the bound on $\mathbb{P}\left[E^c\right]$ from Eqn. (4).

To bound the correlation over the remaining vectors, we have

$$\mathbb{E}\left[\sum_{j\in\mathcal{I}_{\text{pub}}} Z_j\right] \leq \sqrt{\mathbb{E}[\|\lfloor \mathcal{A}(S_{\text{pub}}, S_{\text{priv}})\rfloor_M\|^2]\mathbb{E}\left[\left\|\sum_{j\in\mathcal{I}_{\text{pub}}}(z_j-\mu)\right\|^2\right]} \leq 2Md\sqrt{n_{\text{pub}}}.$$

Above we have used the fact that $|\mathcal{I}_{\text{pub}}| \leq n_{\text{pub}}$ because $i \in \mathcal{I}_{\text{pub}}$ only if at least one copy of z_i is sampled into S_{pub}. Combining the above we have

$$\mathbb{E}\left[\sum_{j=1}^{m} Z_i\right] \leq 5mM\sqrt{d\log\left(1/\delta\right)} + 5Md\sqrt{n_{\text{pub}}}.$$

Combining Bounds: The previously derived lower bound in Eqn. (3) establishes that $\mathbb{E}\left[\sum_{j\in\mathcal{I}_{\text{priv}}} Z_j + \sum_{j\in\mathcal{I}_{\text{pub}}} Z_j\right] \geq \frac{d}{3}$. Using the above derived upper bounds we have the following manipulations,

$$\begin{aligned}
& M\sqrt{d}\left(m\sqrt{\log\left(1/\delta\right)} + \sqrt{dn_{\text{pub}}}\right) \geq \frac{d}{15} \\
\iff \quad & (\alpha^* + \sqrt{d/m})\left(m\sqrt{\log\left(1/\delta\right)} + \sqrt{dn_{\text{pub}}}\right) \geq \frac{d}{60} \\
\iff \quad & m\alpha\sqrt{\log\left(1/\delta\right)} + \alpha^*\sqrt{dn_{\text{pub}}} \geq \frac{d}{60} - \sqrt{d\log\left(1/\delta\right)m} - d\sqrt{\frac{n_{\text{pub}}}{m}}.
\end{aligned}$$

The second line above uses that $M = \frac{4}{\sqrt{d}}(\alpha^* + \sqrt{\frac{d}{m}})$. Under the condition that $n_{\text{pub}} \leq \frac{m}{120^2} \equiv n_{\text{pub}} \leq \frac{3n\epsilon}{120^2\log\left(1/[\sqrt{nd}\delta]\right)}$, which is satisfied under by assumption in the theorem statement, we have

$$m\alpha\sqrt{\log\left(1/\delta\right)} + \alpha^*\sqrt{dn_{\text{pub}}} \geq \frac{d}{120} - \sqrt{d\log\left(1/\delta\right)m}.$$

Now applying the assumption $d \geq 120^2 n\epsilon \implies m \leq \frac{d}{120^2\log(1/\delta)}$ we obtain

$$\begin{aligned}
& m\alpha^*\sqrt{\log\left(1/\delta\right)} + \alpha^*\sqrt{dn_{\text{pub}}} \geq \frac{d}{120} \\
& \alpha^* \geq \frac{1}{120}\min\left\{\frac{d\sqrt{\log\left(1/\delta\right)}}{n\epsilon}, \sqrt{\frac{d}{n_{\text{pub}}}}\right\}.
\end{aligned}$$

This establishes a contradiction, and thus $\alpha \geq \alpha^* = \frac{1}{125}\min\left\{\frac{d\sqrt{\log(1/\delta)}}{n\epsilon}, \sqrt{\frac{d}{n_{\text{pub}}}}\right\}$. Rescaling by $\frac{R}{\sqrt{d}}$ then yields the claimed result. □

B.2 Discussion of Lower Bound Analysis

We here provide more details on why the particular lower bound techniques we present were chosen. Our aim for the following discussion is to elucidate some of the subtleties of leveraging the fingerprinting code framework when public data is present, with the hope that it will aid future work on the characterization of PA-DP problems.

One crucial challenge in developing the mean estimation lower bounds in Appendix B.1 is ensuring that the correlation sum, traditionally defined as $\mathbb{E}\left[\sum_{x\in S}\langle\mathcal{A}(S), x-\mu\rangle\right]$, scales with the accuracy, α, of the algorithm. Previous work, such as [CWZ21], achieves this by setting the underlying distribution, $\mathcal{D}$, to be a mixture distribution which, for some $p = o(1)$, samples a 0 vector with probability $(1-p)$ and samples from the non-trivial distribution, $\mathcal{D}_\mu$, with probability p. However, now the variance satisfies $\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\|x - \mathbb{E}\left[x\right]\|^2\right] \leq 2pR^2$ meaning that when public data is present it

holds that $\mathbb{E}\big[\|\frac{1}{n_{\text{pub}}}\sum_{x\in S_{\text{pub}}} x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[x]\|\big] \le \frac{2pR}{\sqrt{n_{\text{pub}}}} = o\big(\frac{R}{\sqrt{n_{\text{pub}}}}\big)$, and one cannot hope to achieve the desired lower bound. Alternatively, by instead analyzing the sum $\mathbb{E}\left[\sum_{x\in S}\langle\mathcal{A}(S)-\mu, x-\mu\rangle\right]$, as seen for example in [KU20], we are able to avoid sampling from a mixture distribution. Further, by leveraging the flexibility of the strong distribution framework from [DSS+15], we are able to still ensure $\|\mu(\mathcal{D})\| = o(1)$, as needed for the SCO reduction; see Section B.3. These techniques lead to the result in Theorem 12.

Unfortunately, with regards to obtaining the $\sqrt{\log(1/\delta)}$ improvement in Theorem 13, the property $\mathbb{E}[\|\mathcal{A}(S)-\mu\|] \le \alpha$ does little to help establish the needed tail bound; see Eqn. (5). By clipping the components of $\mathcal{A}(S)$ to to the range $[-O(\alpha), O(\alpha)]$, we are able to able to obtain the desired concentration. Unfortunately, this clipping technique in combination with the intermediate distribution, $\mathsf{Unif}(S_z)$, leads to the restrictions that $d \ge n\epsilon$ and $n_{\text{pub}} \le \frac{n}{\log(1/[nd\delta])}$. These restrictions occur because of the need for the "additional error" introduced by the intermediate distribution to be negligible. To see this, observe the intermediate distribution leads to $\mathbb{E}[\|\mathcal{A}(S)-\mu\|] \ge \frac{1}{\sqrt{m}}$ since $\mathcal{A}(S)$ depends on only m vectors from $\mathcal{D}_\mu$, and the analysis in the proof of Theorem 12 (with $n_{\text{priv}} = 0$) shows us that even non-private algorithms cannot do better on this distribution. We remark that [CWZ21] avoids this issue, and hence the restriction on d and n_{pub}, because of the fact that one only actually needs $\|\mathbb{E}[\mathcal{A}(S)]-\mu\| \le \alpha$ for the fingerprinting lemma to hold, and $\|\mathbb{E}[\mathcal{A}(S)]-\mu\| \le \mathbb{E}[\|\mathcal{A}(S)-\mu\|]$. However, after clipping it is possible that $\|\mathbb{E}[\lfloor\mathcal{A}(S)\rfloor_M]-\mu\| \ge \|\mathbb{E}[\mathcal{A}(S)]-\mu\|$.

B.3 Missing proofs from Section 3.1

Proof of Theorem 1. We use the instance in [BST14], $\ell(w;x) = G\langle w, x\rangle$ and $\mathcal{W} = \{x\in\mathbb{R}^d : \|x\| \le 1\}$. By a standard rescaling argument, we only need to consider $G = D = 1$. We will consider the re-scaled data distribution used in Theorem 12, where $\{z_1, ..., z_n\} \overset{i.i.d.}{\sim} \mathcal{D}_\mu$ and the dataset S has $x_j = \frac{1}{\sqrt{d}} z_j$ for $j \in n$. Here $\mu \sim \mathsf{Unif}([-M, M]^d)$ where M will be chosen later.

First note by Lemma 5 we have that $\mathbb{P}[\big|\|\mu\| - \sqrt{\frac{2}{3}}M\big| \ge \frac{M}{256}] \le \frac{1}{512}$ so long as d is larger than some constant. Define this event as E and E' its complement. Thus we have

$$\begin{aligned}\mathbb{E}[L(\mathcal{A}(S);\mathcal{D}) - L(w^*;\mathcal{D})] &= \mathbb{E}[L(\mathcal{A}(S);\mathcal{D}) - L(w^*;\mathcal{D})|E]\,\mathbb{P}[E] \\ &\quad + \mathbb{E}[L(\mathcal{A}(S);\mathcal{D}) - L(w^*;\mathcal{D})|E']\,\mathbb{P}[E'] \\ &\ge \frac{1}{2}\mathbb{E}[L(\mathcal{A}(S);\mathcal{D}) - L(w^*;\mathcal{D})|E]\,.\end{aligned}$$

Thus it suffices to lower bound the conditional excess risk.

The optimal solution under the aforementioned loss is $w^* = -\frac{\mu}{\|\mu\|}$, since the constraint set is a ball of radius 1. We can see that

$$\begin{aligned}L(\mathcal{A}(S);\mathcal{D}) - L(w^*;\mathcal{D}) &= \langle\mathcal{A}(S),\mu\rangle - \left\langle -\frac{\mu}{\|\mu\|}, \mu\right\rangle \\ &= \|\mu\|\,(1 - \langle\mathcal{A}(S), w^*\rangle) \\ &= \|\mu\|\left(1 - \frac{1}{2}\|\mathcal{A}(S)\|^2 - \frac{1}{2}\|w^*\|^2 + \frac{1}{2}\|\mathcal{A}(S) - w^*\|^2\right) \\ &\ge \frac{1}{2}\|\mu\|\,\|\mathcal{A}(S) - w^*\|^2\,. \qquad (6)\end{aligned}$$

We will now lower bound $\|\mathcal{A}(S) - w^*\|$ by using the lower bound for mean estimation developed in Theorem 11. Let the mean estimate candidate is $\bar{\mu}(S) = \bar{\mu} = -\sqrt{\frac{2}{3}}M\mathcal{A}(S)$. Under the event E,

$$\begin{aligned}
\|\bar{\mu} - \mu\|^2 &= \left\|-\sqrt{2/3}M\mathcal{A}(S) - \mu\right\|^2 \\
&= \left\|-\|\mu\|\,\mathcal{A}(S) - \mu + (\sqrt{2/3}M - \|\mu\|)\mathcal{A}(S)\right\|^2 \\
&\leq 2M^2\,\|\mathcal{A}(S) - w^*\|^2 + \frac{M^2}{50} \\
\implies \|\mathcal{A}(S) - w^*\|^2 &\geq \frac{\|\bar{\mu} - \mu\|^2}{2M^2} - \frac{1}{512}. \qquad (7)
\end{aligned}$$

The above follows from the definition of w^* and since the algorithm's output is considered in a ball of radius 1, so $\|\mathcal{A}(S)\| \leq 1$.

Combining the above inequalities (6) and (7) then taking expectation we have,

$$\begin{aligned}
\mathbb{E}\left[L(\mathcal{A}(S);\mathcal{D}) - L(w^*;\mathcal{D})|E\right] &\geq \mathbb{E}\left[\frac{1}{4}\|\mu\|\left(\frac{\|\bar{\mu} - \mu\|^2}{M^2} - \frac{1}{512}\right)\middle| E\right] \\
&\geq \frac{M}{1024}\left(\frac{\mathbb{E}\left[\|\bar{\mu} - \mu\|^2|E\right]}{2M^2} - \frac{1}{512}\right). \qquad (8)
\end{aligned}$$

To bound $\mathbb{E}\left[\|\bar{\mu} - \mu\|^2|E\right]$, observe

$$\begin{aligned}
\mathbb{E}[\|\bar{\mu} - \mu\|^2] &= \mathbb{E}[\|\bar{\mu} - \mu\|^2\,|E]\mathbb{P}[E] + \mathbb{E}_{\mu,S}[\|\bar{\mu} - \mu\|^2\,|E']\mathbb{P}[E'] \\
&\leq \mathbb{E}[\|\bar{\mu} - \mu\|^2\,|E] + 4M^2\mathbb{P}[E'].
\end{aligned}$$

Rearranging we have

$$\mathbb{E}[\|\bar{\mu} - \mu\|^2\,|E] \geq \mathbb{E}[\|\bar{\mu} - \mu\|^2] - \frac{M^2}{128}. \qquad (9)$$

We will finish the bound by applying either Theorem 12 or Theorem 13.

Via Theorem 12: Set $M = \min\left\{\frac{\sqrt{d}}{8n_{\text{priv}}}, \frac{1}{\sqrt{n_{\text{pub}}}}\right\}$. Under this setting of M, Theorem 12 implies that the lower bound on mean estimate distance satisfies $\mathbb{E}\left[\|\bar{\mu} - \mu\|\right] \geq \frac{M}{8}$, and thus $\mathbb{E}[\|\bar{\mu} - \mu\|^2\,|E] \geq \frac{M^2}{128}$ by Eqn. (9) above. Plugging into Eqn. (8) we have

$$\mathbb{E}\left[L(\mathcal{A}(S);\mathcal{D}) - L(w^*;\mathcal{D})\right] = \Omega\left(M\right) = \Omega\left(\min\left\{\frac{\sqrt{d}}{n_{\text{priv}}}, \frac{1}{\sqrt{n_{\text{pub}}}}\right\}\right).$$

Via Theorem 13: In Theorem 13, the setting of M used is

$$\begin{aligned}
M &= 4\left(\frac{1}{125}\min\left\{\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}, \sqrt{\frac{1}{n_{\text{pub}}}}\right\} + \sqrt{\frac{\log\left(1/[\sqrt{dn}\delta]\right)}{n\epsilon}}\right) \\
&\leq \frac{1}{30}\min\left\{\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}, \sqrt{\frac{1}{n_{\text{pub}}}}\right\}.
\end{aligned}$$

The inequality holds under the conditions $d \geq 120^2 n\epsilon$ and $n_{\text{pub}} \leq \frac{n}{120^2\log\left(1/[\sqrt{nd}\delta]\right)}$. Thus we have under this setting of M that $\mathbb{E}\left[\|\bar{\mu} - \mu\|\right] \geq \frac{M}{8}$. Applying Eqns. (9) and (8) as in the previous case we have (providing the above conditions on d and n_{pub} hold)

$$\mathbb{E}\left[L(\mathcal{A}(S);\mathcal{D}) - L(w^*;\mathcal{D})\right] = \Omega\left(M\right) = \Omega\left(\min\left\{\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}, \sqrt{\frac{1}{n_{\text{pub}}}}\right\}\right).$$

□

Lemma 5. *For $z \sim \mathsf{Unif}([-1,1]^d)$, we have that $\|z\| \in \frac{\sqrt{2d}}{\sqrt{3}} \pm \frac{\sqrt{3\ln(1/\gamma)}}{2}$, with probability at least $1-\gamma$.*

Proof. This follows from standard concentration of norm results. We have that,

$$\mathbb{E}\|z\|^2 = d\mathbb{E}z_1^2 = \frac{2d}{3}.$$

As in [Ver18, proof of Theorem 3.1.1], we use the simple fact that $|x-1| > \delta \implies |x^2-1| > \max(\delta, \delta^2)$ for any $x, \delta \geq 0$, to get,

$$\begin{aligned}
&\mathbb{P}\left(\left|\|z\| - \frac{\sqrt{2d}}{\sqrt{3}}\right| > \frac{\sqrt{2d}\delta}{\sqrt{3}}\right) = \mathbb{P}\left(\left|\frac{\sqrt{3}\|z\|}{\sqrt{2d}} - 1\right| > \delta\right) \\
&= \mathbb{P}\left(\left|\frac{3\|z\|^2}{2d} - 1\right| > \max(\delta, \delta^2)\right) \\
&= \mathbb{P}\left(\left|\frac{1}{d}\sum_{i=1}^{d} z_i^2 - \frac{2}{3}\right| > \max\left((2/3)\delta, ((2/3)\delta)^2\right)\right).
\end{aligned}$$

We substitute $\bar{\delta} = \frac{2\delta}{3}$ and apply Bernstein's inequality for i.i.d sub-exponential random variables z_i^2. Since, $z_i \in [-1,1]$, the sub-exponential norm ≤ 1. Applying Corollary 2.8.3 from [Ver18], we get that,

$$\mathbb{P}\left(\left|\frac{1}{d}\sum_{i=1}^{d} z_i^2 - \frac{2}{3}\right| > \max\left(\bar{\delta}, (\bar{\delta})^2\right)\right) \leq \exp\left(-2\bar{\delta}^2 d\right) = \exp\left(-8\delta^2 d/9\right).$$

This gives us that

$$\mathbb{P}\left(\left|\|z\| - \frac{\sqrt{2d}}{\sqrt{3}}\right| > \frac{\sqrt{2d}\delta}{\sqrt{3}}\right) \leq \exp\left(-8\delta^2 d/9\right).$$

Hence, with probability, at least $1-\gamma$, we have that $\|z\| \in \frac{\sqrt{2d}}{\sqrt{3}} \pm \frac{\sqrt{3\ln(1/\gamma)}}{2}$, which completes the proof. □

Proof of Theorem 2. We use the squared loss instance as in [BST14, Section 5.2]; $\ell(w;z) = \frac{\lambda}{2}\|w-z\|^2$, with $\|z\| \leq \frac{G}{2\lambda}$. The loss is G-Lipschitz and λ strongly convex on the domain of unit ball at zero of radius $\frac{G}{2\lambda}$. Given a datasets $S = \{z_1, z_2, \ldots, z_n\}$, the population risk minimizer is simply the population mean $\mu(\mathcal{D})$. Further, it is straightforward to verify that the excess population risk a re-scaling of the mean estimation error

$$\mathbb{E}[L(\mathcal{A}(S)) - \min_w L(w)] = \frac{\lambda}{2}\mathbb{E}\|\mathcal{A}(S) - \mu(\mathcal{D})\|^2.$$

Substituting the mean estimation lower bounds, Theorem 11, completes the proof. □

C Missing Proofs from Section 4

C.1 Proof of Theorem 3

Define the orthogonal projection matrix $P_{X_{\text{pub}}} = UU^\top$. Note that the feature vectors in $\tilde{S}_{\text{priv}} = \left\{(U^\top x_i, y_i)\right\}_{i=1}^{n_{\text{priv}}}$ are bounded. In particular $\left\|U^\top x\right\|^2 = x^\top(UU^\top)x = x^\top P_{X_{\text{pub}}} x \leq \|x\|^2$, since $P_{X_{\text{pub}}}$ is an orthogonal projection onto $\text{span}(\mathcal{W} \cap S_{\text{pub}})$. Further, since $\tilde{w} \in \tilde{\mathcal{W}}$, we have that there exists $\mathring{w} \in \mathcal{W}$ such that $\tilde{w} = U^\top \mathring{w}$. Finally, $\widehat{w} = U\tilde{w} = UU^\top \mathring{w} = P_{X_{\text{pub}}}\mathring{w} \in \mathcal{W}$ since the range of $P_{X_{\text{pub}}} \subseteq \mathcal{W}$.

The privacy guarantee follows from the privacy guarantee of sub-routine $\tilde{\mathcal{A}}$. For utility, we define $w^* \in \arg\min_{w\in\mathcal{W}} L(w;\mathcal{D})$ and $\tilde{w}^* \in \arg\min_{w\in\tilde{\mathcal{W}}} L(w;U^\top\mathcal{D})$, where $U^\top\mathcal{D}$ denotes the distribution which first samples from $\mathcal{D}$ then project using $U^\top$. Let $\mathring{w}^* \in \mathcal{W}$ such that $\mathring{w}^* = U\tilde{w}^*$.

Note that from the GLM structure, $L(\tilde{w}^*;U^\top\mathcal{D}) = L(\mathring{w}^*;\mathcal{D})$. We have,

$$\begin{aligned}
L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D}) &= L(\widehat{w};\mathcal{D}) - \widehat{L}(\widehat{w};S_{\text{priv}}) + L(\mathring{w}^*;\mathcal{D}) - L(w^*;\mathcal{D}) \\
&\quad + \widehat{L}(\tilde{w}^*;S_{\text{priv}}) - L(\mathring{w}^*;\mathcal{D}) + \widehat{L}(\widehat{w};S_{\text{priv}}) - \widehat{L}(\mathring{w}^*;S_{\text{priv}}) \\
&= O\left(G\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H}) + \frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}}\right) \\
&\quad + L(\mathring{w}^*;\mathcal{D}) - L(w^*;\mathcal{D}) + \widehat{L}(\tilde{w};\tilde{S}_{\text{priv}}) - \min_{w\in\tilde{\mathcal{W}}} \widehat{L}(w;\tilde{S}_{\text{priv}}) \qquad (10)
\end{aligned}$$

with probability at least $1-\beta/4$. In the above, we control the generalization gap via uniform convergence and concentration for the fixed $\mathring{w}^*$ with respect to S_{priv}.

The last term $\widehat{L}(\tilde{w};\tilde{S}_{\text{priv}}) - \min_{w\in\tilde{\mathcal{W}}} \widehat{L}(w;\tilde{S}_{\text{priv}})$ is bounded by the guarantee of the private sub-routine with probability at least $1-\beta/4$,

$$\widehat{L}(\tilde{\mathcal{A}}(\tilde{S}_{\text{priv}});\tilde{S}_{\text{priv}}) - \min_{w\in\tilde{\mathcal{W}}} \widehat{L}(\tilde{\mathcal{A}}(\tilde{S}_{\text{priv}});\tilde{S}_{\text{priv}}) = \tilde{O}\left(GD\left\|\mathcal{X}\right\|\left(\frac{\sqrt{n_{\text{pub}}\log(1/\delta)} + \sqrt{\log(4/\beta)}}{n_{\text{priv}}\epsilon}\right)\right).$$

Finally, for any $\bar{w}^*$ such that $\bar{w}^* \in U\tilde{\mathcal{W}}$, with probability at least $1-\beta/2$, from G-Lipschitznes, we have

$$L(\mathring{w}^*;\mathcal{D}) - L(w^*;\mathcal{D}) \le L(\bar{w}^*;\mathcal{D}) - L(w^*;\mathcal{D}) \le G\left\|\bar{w}^* - w^*\right\|_{2,\mathcal{D}_\mathcal{X}} \le G\alpha,$$

where the last inequality follows essentially from Lemma 6 and Lemma 1 for $n_{\text{pub}} = O\left(\max\left(\frac{R^2\log(2/\beta)}{\alpha^2}, \min\left\{m : \log^3(n_{\text{pub}})\mathfrak{R}^2_{n_{\text{pub}}}(\mathcal{H}) \le \alpha^2\right\}\right)\right)$. To elaborate, the first step holds since $\mathring{w}^* = U\tilde{w}^*$ and $\tilde{w}^*$ is the the minimizer of risk over $\tilde{\mathcal{W}}$. Now, Lemma 1 guarantees that for any $w^* \in \mathcal{W}$, there exists a $\bar{w}^*$ in its α-cover with respect to $\left\|\cdot\right\|_{2,X_{\text{pub}}}$, with $\left\|w^* - \bar{w}^*\right\|_{2,\mathcal{D}_\mathcal{X}} \le \alpha$. To argue why span$(X_{\text{pub}})$ is an α-cover, from Lemma 6, we have that from any α-cover $\bar{\mathcal{W}}$, of $\mathcal{W}$ w.r.t. $\left\|\cdot\right\|_{2,X_{\text{pub}}}$, we can remove elements which do not lie in span(S_{pub}) and still have an α-cover. Hence, the superset used in Algorithm 1, which essentially is, $\bar{\mathcal{W}} = P_{X_{\text{pub}}}\mathcal{W}$, is indeed an α-cover.

The n_{pub} we get is,

$$\begin{aligned}
n_{\text{pub}} &= O\left(\max\left(\frac{R^2\log(1/\beta)}{\alpha^2}, \min\left\{m : \log^3(m)\mathfrak{R}^2_m(\mathcal{H}) \le \alpha^2\right\}\right)\right) \\
&= \tilde{O}\left(D^2\left\|\mathcal{X}\right\|^2\max\left(\frac{\log(2/\beta)}{\alpha^2}, \frac{1}{\alpha^2}\right)\right)
\end{aligned}$$

where in the above, we plug in the Rademacher complexity of bounded linear predictor, $\mathfrak{R}_m(\mathcal{H}) = \Theta\left(\frac{D\|\mathcal{X}\|}{m}\right)$. Plugging the above in Equation (10),

$$\begin{aligned}
&L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D}) \\
&= O\left(\frac{GD\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{priv}}}} + \frac{GD\left\|\mathcal{X}\right\|\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}} + \frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}}\right) \\
&+ O\left(GD\left\|\mathcal{X}\right\|\left(\frac{\sqrt{n_{\text{pub}}\log(1/\delta)} + \sqrt{\log(4/\beta)}}{n_{\text{priv}}\epsilon}\right)\right) + G\alpha \\
&= O\left(\frac{GD\left\|\mathcal{X}\right\|\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}} + GD^2\left\|\mathcal{X}\right\|^2\left(\frac{\sqrt{\log(2/\beta) + \log(1/\delta)}}{\alpha n_{\text{priv}}\epsilon}\right) + \frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}}\right) + G\alpha \\
&= O\left(GD\left\|\mathcal{X}\right\|\left(\frac{\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}} + \left(\frac{(\log(2/\beta) + \log(1/\delta))^{1/4}}{\sqrt{n_{\text{priv}}\epsilon}}\right)\right) + \frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}}\right)
\end{aligned}$$

where the above follows by setting $\alpha = \frac{D\|\mathcal{X}\|(\log(1/\delta)+\log(2/\beta))^{1/4}}{\sqrt{n_{\text{priv}}\epsilon}}$. This yields the claimed rate. The resulting public sample complexity is

$$\begin{aligned} n_{\text{pub}} &= \tilde{O}\left(D^2\|\mathcal{X}\|^2 \max\left(\frac{\log(2/\beta)}{\alpha^2}, \frac{1}{\alpha^2}\right)\right) \\ &= \tilde{O}\left(\frac{n_{\text{priv}}\epsilon}{(\log(2/\beta)+\log(1/\delta))^{1/2}}\right). \end{aligned}$$

This completes the proof.

Lemma 6. *Let $\tilde{\mathcal{H}}$ be a α-cover of $\mathcal{H}$ with respect $\|\cdot\|_{2,X_{pub}}$. Then, $\bar{\mathcal{H}} = \tilde{\mathcal{H}} \cap span(X_{pub})$ is also an α-cover.*

Proof. Given two $h_1, h_2 \in \mathcal{H}$, we have,

$$\|h_1 - h_2\|_{2,X_{\text{pub}}} = \sqrt{\frac{1}{n_{\text{pub}}}\sum_{i=1}^{n_{\text{pub}}}(h_1(x_i) - h_2(x_i))^2} = \frac{1}{\sqrt{n_{\text{pub}}}}\sqrt{(w_1 - w_2)^\top X_{\text{pub}}^\top X_{\text{pub}}(w_1 - w_2)}$$

where w_1 and w_2 are the vectors corresponding to linear functions h_1 and h_2 and X_{pub} denote the matrix of public feature vectors.

Given any $h \in \mathcal{H}$, let $\tilde{h}$ denote the element closest to it in the cover $\tilde{\mathcal{H}}$; we have, $\|h - \tilde{h}\|_{2,X_{\text{pub}}} \leq \alpha$. Consider the singular value decomposition, $X_{\text{pub}} = V\Sigma U^\top$, where U and V are orthogonal matrices and Σ is a diagonal matrix. Define $\bar{h} = P_{X_{\text{pub}}}(\tilde{h}) = UU^\top\tilde{h}$. Note that U is an orthogonal projection onto span(X_{pub}) and $P_{X_{\text{pub}}}$ is the corresponding projection matrix. We have,

$$\begin{aligned} \left\|h - \bar{h}\right\|_{2,X_{\text{pub}}}^2 &= \frac{1}{n_{\text{pub}}}(h - \bar{h})^\top X_{\text{pub}}^\top X_{\text{pub}}(h - \bar{h}) \\ &= \frac{1}{n_{\text{pub}}}(h - P_{X_{\text{pub}}}(\tilde{h}))^\top U\Sigma^2U^\top(h - P_{X_{\text{pub}}}(\tilde{h})) \\ &= \frac{1}{n_{\text{pub}}}(h - \tilde{h})^\top U(U^\top U)\Sigma^2(U^\top U)U^\top(h - \tilde{h}) \\ &= \frac{1}{n_{\text{pub}}}(h - \tilde{h})^\top U\Sigma^2U^\top(h - \tilde{h}) \\ &= \frac{1}{n_{\text{pub}}}(h - \tilde{h})^\top X_{\text{pub}}^\top X_{\text{pub}}(h - \tilde{h}) \\ &\leq \alpha^2 \end{aligned}$$

Since by construction $\bar{h}$ also lies in span(X_{pub}), this proves the claim. □

C.2 Proof of Theorem 6

We state the complete version of this theorem and then present its proof.

Theorem 14. *Let $\epsilon > 0, \delta > 0$ and $\epsilon \leq \log(1/\delta)$. For a G-Lipschitz, B-bounded non-negative H-smooth loss function, Algorithm 1 satisfies (ϵ, δ)-DP. If the private subroutine $\tilde{\mathcal{A}}$ guarantees Equation (1) with probability at least $1 - \beta$, then with $n_{pub} = \tilde{O}\left(\frac{(HD\|\mathcal{X}\|)^{2/3}(n_{priv}\epsilon)^{2/3}}{G^{2/3}(\log(1/\delta))^{1/3}} + \frac{\sqrt{H}n_{priv}\epsilon\sqrt{L(w^*;\mathcal{D})}}{G\sqrt{\log(1/\delta)}}\right)$, with probability at least $1-\beta$, $L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D})$ is at most*

$$\tilde{O}\left(\left(\frac{\sqrt{H}D\|\mathcal{X}\|}{\sqrt{n_{priv}\epsilon}} + \sqrt{\frac{B\log(8/\beta)}{n_{priv}}}\right)\sqrt{L(w^*;\mathcal{D})} + \frac{H\|\mathcal{X}\|^2D^2}{n_{priv}\epsilon} + \frac{B\log(8/\beta)}{n_{priv}} + \frac{GD\|\mathcal{X}\|\sqrt{\log(4/\beta)}}{n_{priv}\epsilon}\right)$$

$$+\tilde{O}\left(\left(\frac{\sqrt{H}D^2\|\mathcal{X}\|^2G\sqrt{\log(1/\delta)}}{n_{priv}\epsilon}\right)^{2/3} + \frac{H^{1/4}D\|\mathcal{X}\|\sqrt{G}(\log(1/\delta))^{1/4}L(w^*;\mathcal{D})^{1/4}}{\sqrt{n_{priv}\epsilon}}\right)$$

Further, with $n_{pub} = \tilde{O}\left(\frac{(HD\|\mathcal{X}\|)^{2/3}(n_{priv}\epsilon)^{2/3}}{G^{2/3}(\log(1/\delta))^{1/3}} + \frac{\sqrt{H}n_{priv}\epsilon\sqrt{\widehat{L}(\widehat{w}^*;S_{priv})}}{G\sqrt{\log(1/\delta)}}\right)$*, with probability at least* $1-\beta$*, for any* $\bar{w} \in \mathcal{W}$, $L(\widehat{w};\mathcal{D}) - \widehat{L}(\widehat{w}^*;S_{priv})$ *is at most*

$$\tilde{O}\left(\left(\frac{\sqrt{H}D\|\mathcal{X}\|}{\sqrt{n_{priv}\epsilon}} + \sqrt{\frac{B\log(8/\beta)}{n_{priv}}}\right)\sqrt{\widehat{L}(\widehat{w}^*;S_{priv})} + \frac{H\|\mathcal{X}\|^2 D^2}{n_{priv}\epsilon} + \frac{B\log(8/\beta)}{n_{priv}} + \frac{GD\|\mathcal{X}\|\sqrt{\log(4/\beta)}}{n_{priv}\epsilon}\right)$$
$$+\tilde{O}\left(\left(\frac{\sqrt{H}D^2\|\mathcal{X}\|^2 G\sqrt{\log(1/\delta)}}{n_{priv}\epsilon}\right)^{2/3} + \frac{H^{1/4}D\|\mathcal{X}\|\sqrt{G}(\log(1/\delta))^{1/4}\widehat{L}(\bar{w};S_{priv})^{1/4}}{\sqrt{n_{priv}\epsilon}}\right).$$

where w^* *and* $\widehat{w}^*$ *are population and empirical minimizers with respect to* $\mathcal{D}$ *and* S_{priv} *respectively.*

Proof of Theorem 14. The privacy guarantee follows from the privacy guarantee of sub-routine $\tilde{\mathcal{A}}$. The proof of the utility guarantee proceeds similar to that of Theorem 3. We define $w^* \in \arg\min_{w\in\mathcal{W}} L(w;\mathcal{D})$ and $\tilde{w}^* \in \arg\min_{w\in\tilde{\mathcal{W}}} L(w;U^\top\mathcal{D})$. Let $\mathring{w}^* \in \mathcal{W}$ such that $\mathring{w}^* = U\tilde{w}^*$. From the GLM structure, $L(\tilde{w}^*;U^\top\mathcal{D}) = L(\mathring{w}^*;\mathcal{D})$. We have,

$$\begin{aligned}
L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D}) &= L(\widehat{w};\mathcal{D}) - \widehat{L}(\widehat{w};S_{\text{priv}}) + \widehat{L}(\widehat{w};S_{\text{priv}}) - L(w^*;\mathcal{D}) \\
&\le L(\widehat{w};\mathcal{D}) - \widehat{L}(\widehat{w};S_{\text{priv}}) + \widehat{L}(\mathring{w}^*;S_{\text{priv}}) - L((\mathring{w}^*;\mathcal{D}) \\
&+ L(\mathring{w}^*;\mathcal{D}) - L(w^*;\mathcal{D}) + \widehat{L}(\widehat{w};S_{\text{priv}}) - \widehat{L}(\mathring{w}^*;S_{\text{priv}}) \\
&\le \left|L(\widehat{w};\mathcal{D}) - \widehat{L}(\widehat{w};S_{\text{priv}})\right| + \left|L(\mathring{w}^*;\mathcal{D}) - \widehat{L}(\mathring{w}^*;S_{\text{priv}})\right| \\
&+ L(\mathring{w}^*;\mathcal{D}) - L(w^*;\mathcal{D}) + \widehat{L}(\tilde{w};\tilde{S}_{\text{priv}}) - \min_{w\in\tilde{\mathcal{W}}}\widehat{L}(w;\tilde{S}_{\text{priv}}) \qquad (11)
\end{aligned}$$

The last term $\widehat{L}(\tilde{w};\tilde{S}_{\text{priv}}) - \min_{w\in\tilde{\mathcal{W}}}\widehat{L}(w;\tilde{S}_{\text{priv}})$ is bounded by the guarantees of the private sub-routine with probability at least $1-\beta/4$,

$$\widehat{L}(\widehat{w};\tilde{S}_{\text{priv}}) - \min_{w\in\tilde{\mathcal{W}}}\widehat{L}(w;\tilde{S}_{\text{priv}}) = \tilde{O}\left(GD\|\mathcal{X}\|\left(\frac{\sqrt{n_{\text{pub}}\log(1/\delta)} + \sqrt{\log(4/\beta)}}{n_{\text{priv}}\epsilon}\right)\right). \qquad (12)$$

To bound the term $L(\mathring{w}^*;\mathcal{D}) - L(w^*;\mathcal{D})$ in Equation (11), we apply smoothness to get,

$$\begin{aligned}
&L(\mathring{w}^*;\mathcal{D}) - L(w^*;\mathcal{D}) \\
&\le L(\bar{w}^*;\mathcal{D}) - L(w^*;\mathcal{D}) \\
&\le \mathbb{E}\left[\left\langle\phi_y'(\langle w^*,x\rangle), \langle\bar{w}^*,x\rangle - \langle w^*,x\rangle\right\rangle + \frac{H}{2}\left|\langle\bar{w}^*,x\rangle - \langle w^*,x\rangle\right|^2\right] \\
&\le \mathbb{E}\left[\left|\phi_y'(\langle w^*,x\rangle)\right|\left|\langle\bar{w}^*,x\rangle - \langle w^*,x\rangle\right| + \frac{H}{2}\left|\langle\bar{w}^*,x\rangle - \langle w^*,x\rangle\right|^2\right] \\
&\le \sqrt{\mathbb{E}\left|\phi_y'(\langle w^*,x\rangle)\right|^2}\sqrt{\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left|\langle\bar{w}^*,x\rangle - \langle w^*,x\rangle\right|^2} + \frac{H}{2}\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left|\langle\bar{w}^*,x\rangle - \langle w^*,x\rangle\right|^2 \\
&\le 2\sqrt{H\mathbb{E}_{x\sim\mathcal{D}}\phi_y(\langle w^*,x\rangle)}\sqrt{\mathbb{E}\left|\langle\bar{w}^*,x\rangle - \langle w^*,x\rangle\right|^2} + \frac{H}{2}\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left|\langle\bar{w}^*,x\rangle - \langle w^*,x\rangle\right|^2 \\
&\le 2\sqrt{HL(w^*;\mathcal{D})}\alpha + H\alpha^2 \qquad (13)
\end{aligned}$$

where the above holds for any $\bar{w}^* \in \mathcal{H}$ such that $\bar{w}^* \in U\tilde{\mathcal{W}}$ by optimality of $\tilde{w}^*$ in $\tilde{\mathcal{W}}$. The second inequality holds from H-smoothness, the third and fourth from Cauchy-Schwarz, the fifth from self-bounding property of smooth non-negative losses (Lemma 4.1 in [SST10]). The final step holds with probability $1-\beta/2$ from Lemma 1 with $n_{\text{pub}} = O\left(\max\left(\frac{R^2\log(2/\beta)}{\alpha^2}, \min\left\{m : \log^3(m)\mathfrak{R}_m^2(\mathcal{H}) \le \alpha^2\right\}\right)\right)$ together with that since $\tilde{\mathcal{W}}$ is an α-cover

of $\mathcal{H}$, together with Lemma 6 which shows that $\tilde{\mathcal{W}}$ is a valid α-cover of $\mathcal{W}$. Therefore, there exists $\bar{h}^* \in \tilde{\mathcal{H}}$ with $\left\|\bar{h}^* - h^*\right\|_{2,S_{\text{pub}}} \leq \alpha$.

Further, applying AM-GM inequality, we get

$$L(\mathring{w}^*; \mathcal{D}) \leq 2L(w^*; \mathcal{D}) + 2H\alpha^2 \tag{14}$$

The first two terms in Equation (11) are bound via uniform convergence for smooth non-negative losses, (Theorem 1 in [SST10]) and Bernstein's inequality as follows; with probability at least $1-\beta/4$, we have,

$$\begin{aligned}
&\left|L(\widehat{w}; \mathcal{D}) - \widehat{L}(\widehat{w}; S_{\text{priv}})\right| + \left|L(\mathring{w}^*; \mathcal{D}) - \widehat{L}(\mathring{w}^*; S_{\text{priv}})\right| \\
&= \tilde{O}\left(\sqrt{H}\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H}) + \sqrt{\frac{B\log(8/\beta)}{n_{\text{priv}}}}\right)\left(\sqrt{\widehat{L}(\widehat{w}; S_{\text{priv}})} + \sqrt{L(\mathring{w}^*; \mathcal{D})}\right) \\
&+ \tilde{O}\left(H\mathfrak{R}^2_{n_{\text{priv}}}(\mathcal{H}) + \frac{B\log(8/\beta)}{n_{\text{priv}}}\right) \\
&= \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{priv}}\epsilon}} + \sqrt{\frac{B\log(8/\beta)}{n_{\text{priv}}}}\right)\left(\sqrt{\widehat{L}(\mathring{w}^*; S_{\text{priv}})} + \sqrt{L(\mathring{w}^*; \mathcal{D})}\right) \\
&+ \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{\text{priv}}\epsilon} + \frac{B\log(8/\beta)}{n_{\text{priv}}}\right) + \tilde{O}\left(GD\left\|\mathcal{X}\right\|\left(\frac{\sqrt{n_{\text{pub}}\log(1/\delta)} + \sqrt{\log(4/\beta)}}{n_{\text{priv}}\epsilon}\right)\right) \\
&= \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{priv}}\epsilon}} + \sqrt{\frac{B\log(8/\beta)}{n_{\text{priv}}}}\right)\sqrt{L(\mathring{w}^*; \mathcal{D})} \\
&+ \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{\text{priv}}\epsilon} + \frac{B\log(8/\beta)}{n_{\text{priv}}}\right) + \tilde{O}\left(GD\left\|\mathcal{X}\right\|\left(\frac{\sqrt{n_{\text{pub}}\log(1/\delta)} + \sqrt{\log(4/\beta)}}{n_{\text{priv}}\epsilon}\right)\right) \\
&= \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{priv}}\epsilon}} + \sqrt{H}\alpha + \sqrt{\frac{B\log(8/\beta)}{n_{\text{priv}}}}\right)\sqrt{L(w^*; \mathcal{D})} \\
&+ \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{\text{priv}}\epsilon} + H\alpha^2 + \frac{B\log(8/\beta)}{n_{\text{priv}}}\right) + O\left(GD\left\|\mathcal{X}\right\|\left(\frac{\sqrt{n_{\text{pub}}\log(1/\delta)} + \sqrt{\log(4/\beta)}}{n_{\text{priv}}\epsilon}\right)\right)
\end{aligned} \tag{15}$$

where the second equality follows from Equation (12), instantiating the Rademacher complexity of linear predictors, concavity of $x \mapsto \sqrt{x}$ and AM-GM inequality. The third equality follows concavity of $x \mapsto \sqrt{x}$ and Bernstein's inequality, the fourth follows from Equation (14) and AM-GM inequality.

Plugging the above, Equation (13) and Equation (12) into Equation (11), we get that with $n_{\text{pub}} = O\left(\max\left(\frac{\|\mathcal{X}\|^2 D^2 \log(2/\beta)}{\alpha^2}, \min\left\{m : \log^3(n_{\text{pub}})\mathfrak{R}^2_{n_{\text{pub}}}(\mathcal{H}) \leq \alpha^2\right\}\right)\right)$, the following holds with probability at least $1-\beta$,

$$
\begin{aligned}
&L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D}) \\
&= \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{priv}}\epsilon}} + \sqrt{H}\alpha + \sqrt{\frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}}\right)\sqrt{L(w^*;\mathcal{D})} \\
&+ \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{\text{priv}}\epsilon} + H\alpha^2 + \frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}\right) + O\left(GD\left\|\mathcal{X}\right\|\left(\frac{\sqrt{n_{\text{pub}}\log\left(1/\delta\right)} + \sqrt{\log\left(4/\beta\right)}}{n_{\text{priv}}\epsilon}\right)\right) \\
&= \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{priv}}\epsilon}} + \frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{pub}}}} + \sqrt{\frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}}\right)\sqrt{L(w^*;\mathcal{D})} \\
&+ \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{\text{priv}}\epsilon} + \frac{HD^2\left\|\mathcal{X}\right\|^2}{n_{\text{pub}}} + \frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}\right) + O\left(GD\left\|\mathcal{X}\right\|\left(\frac{\sqrt{n_{\text{pub}}\log\left(1/\delta\right)} + \sqrt{\log\left(4/\beta\right)}}{n_{\text{priv}}\epsilon}\right)\right) \qquad (16) \\
&= \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{priv}}\epsilon}} + \sqrt{\frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}}\right)\sqrt{L(w^*;\mathcal{D})} \\
&+ \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{\text{priv}}\epsilon} + \frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}\right) + O\left(\frac{GD\left\|\mathcal{X}\right\|\sqrt{\log\left(4/\beta\right)}}{n_{\text{priv}}\epsilon}\right) \\
&+ O\left(\left(\frac{\sqrt{H}D^2\left\|\mathcal{X}\right\|^2 G\sqrt{\log\left(1/\delta\right)}}{n_{\text{priv}}\epsilon}\right)^{2/3} + \frac{H^{1/4}D\left\|\mathcal{X}\right\|\sqrt{G}\left(\log\left(1/\delta\right)\right)^{1/4} L(w^*;\mathcal{D})^{1/4}}{\sqrt{n_{\text{priv}}\epsilon}}\right)
\end{aligned}
$$

The public sample complexity is,

$$
n_{\text{pub}} = \tilde{O}\left(\frac{(HD\left\|\mathcal{X}\right\|)^{2/3}(n_{\text{priv}}\epsilon)^{2/3}}{G^{2/3}(\log\left(1/\delta\right))^{1/3}} + \frac{\sqrt{H}n_{\text{priv}}\epsilon\sqrt{L(w^*;\mathcal{D})}}{G\sqrt{\log\left(1/\delta\right))}}\right)
$$

This completes the first part of the theorem. For thee second part, we start from Equation (16),

$$
\begin{aligned}
L(\widehat{w};\mathcal{D}) &\leq L(w^*;\mathcal{D}) + \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{priv}}\epsilon}} + \frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{pub}}}} + \sqrt{\frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}}\right)\sqrt{L(w^*;\mathcal{D})} \\
&+ \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{\text{priv}}\epsilon} + \frac{HD^2\left\|\mathcal{X}\right\|^2}{n_{\text{pub}}} + \frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}\right) \\
&+ O\left(GD\left\|\mathcal{X}\right\|\left(\frac{\sqrt{n_{\text{pub}}\log\left(1/\delta\right)} + \sqrt{\log\left(4/\beta\right)}}{n_{\text{priv}}\epsilon}\right)\right) \\
&\leq L(\widehat{w}^*;\mathcal{D}) + \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{priv}}\epsilon}} + \frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{pub}}}} + \sqrt{\frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}}\right)\sqrt{L(\widehat{w}^*;\mathcal{D})} \\
&+ \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{\text{priv}}\epsilon} + \frac{HD^2\left\|\mathcal{X}\right\|^2}{n_{\text{pub}}} + \frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}\right) \\
&+ O\left(GD\left\|\mathcal{X}\right\|\left(\frac{\sqrt{n_{\text{pub}}\log\left(1/\delta\right)} + \sqrt{\log\left(4/\beta\right)}}{n_{\text{priv}}\epsilon}\right)\right) \\
&\leq \widehat{L}(\widehat{w}^*;S_{\text{priv}}) + \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{priv}}\epsilon}} + \frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{pub}}}} + \sqrt{\frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}}\right)\sqrt{\widehat{L}(\widehat{w}^*;S_{\text{priv}})} \\
&+ \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{\text{priv}}\epsilon} + \frac{HD^2\left\|\mathcal{X}\right\|^2}{n_{\text{pub}}} + \frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}\right) \\
&+ O\left(GD\left\|\mathcal{X}\right\|\left(\frac{\sqrt{n_{\text{pub}}\log\left(1/\delta\right)} + \sqrt{\log\left(4/\beta\right)}}{n_{\text{priv}}\epsilon}\right)\right) \\
&\leq \widehat{L}(\widehat{w}^*;S_{\text{priv}}) + \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{priv}}\epsilon}} + \sqrt{\frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}}\right)\sqrt{\widehat{L}(\widehat{w}^*;\mathcal{D})} \\
&+ \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{\text{priv}}\epsilon} + \frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}\right) + O\left(\frac{GD\left\|\mathcal{X}\right\|\sqrt{\log\left(4/\beta\right)}}{n_{\text{priv}}\epsilon}\right) \\
&+ O\left(\left(\frac{\sqrt{H}D^2\left\|\mathcal{X}\right\|^2 G\sqrt{\log\left(1/\delta\right)}}{n_{\text{priv}}\epsilon}\right)^{2/3} + \frac{H^{1/4}D\left\|\mathcal{X}\right\|\sqrt{G}\left(\log\left(1/\delta\right)\right)^{1/4}L(\widehat{w};\mathcal{D})^{1/4}}{\sqrt{n_{\text{priv}}\epsilon}}\right)
\end{aligned}
$$

where the second inequality holds form optimality of $w*$, the third from uniform convergence, Theorem 1 in [SST10] and AM-GM inequality, and the last by plugging in the following public sample complexity.

$$
n_{\text{pub}} = \tilde{O}\left(\frac{(HD\left\|\mathcal{X}\right\|)^{2/3}(n_{\text{priv}}\epsilon)^{2/3}}{G^{2/3}(\log\left(1/\delta\right))^{1/3}} + \frac{\sqrt{H}n_{\text{priv}}\epsilon\sqrt{\widehat{L}(\widehat{w}^*;S_{\text{priv}})}}{G\sqrt{\log\left(1/\delta\right))}}\right)
$$

This completes the proof.

□

Theorem 15. *In the setting of Theorem 14 with the additional assumption that the global minimizer of risk L, w^* lies in $\mathcal{W}$, we get that with $n_{pub} = \tilde{O}\left(\frac{(HD\|\mathcal{X}\|)^{2/3}(n_{priv}\epsilon)^{2/3}}{G^{2/3}(\log(1/\delta))^{1/3}}\right)$, with probability at least*

$1-\beta$,

$$
\begin{aligned}
&L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D}) \\
&= \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{priv}\epsilon}} + \sqrt{\frac{B\log\left(8/\beta\right)}{n_{priv}}}\right)\sqrt{L(w^*;\mathcal{D})} + \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{priv}\epsilon} + \frac{B\log\left(8/\beta\right)}{n_{priv}}\right) \\
&+ O\left(\frac{GD\left\|\mathcal{X}\right\|\sqrt{\log\left(4/\beta\right)}}{n_{priv}\epsilon}\right) + O\left(\left(\frac{\sqrt{H}D^2\left\|\mathcal{X}\right\|^2 G\sqrt{\log\left(1/\delta\right)}}{n_{priv}\epsilon}\right)^{2/3}\right)
\end{aligned}
$$

Further,

$$
\begin{aligned}
&L(\widehat{w};\mathcal{D}) - \widehat{L}(\widehat{w}^*;S_{priv}) \\
&= \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{priv}\epsilon}} + \sqrt{\frac{B\log\left(8/\beta\right)}{n_{priv}}}\right)\sqrt{\widehat{L}(\widehat{w}^*;S_{priv})} + \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{priv}\epsilon} + \frac{B\log\left(8/\beta\right)}{n_{priv}}\right) \\
&+ O\left(\frac{GD\left\|\mathcal{X}\right\|\sqrt{\log\left(4/\beta\right)}}{n_{priv}\epsilon}\right) + O\left(\left(\frac{\sqrt{H}D^2\left\|\mathcal{X}\right\|^2 G\sqrt{\log\left(1/\delta\right)}}{n_{priv}\epsilon}\right)^{2/3}\right).
\end{aligned}
$$

where w^* and $\widehat{w}^*$ are population and empirical minimizers with respect to $\mathcal{D}$ and S_{priv} respectively.

Proof. The proof is almost identical to that of Theorem 15. We repeat the steps pointing out the differences and how the expressions change. We continue till Equation (12). Next, we apply smoothness which results in the key difference between the analyses,

$$
\begin{aligned}
&L(\mathring{w}^*;\mathcal{D}) - L(w^*;\mathcal{D}) \\
&\leq L(\bar{w}^*;\mathcal{D}) - L(w^*;\mathcal{D}) \\
&\leq \mathbb{E}\left[\left\langle \phi_y'(\langle w^*, x\rangle), \langle \bar{w}^*, x\rangle - \langle w^*, x\rangle\right\rangle + \frac{H}{2}\left|\langle \bar{w}^*, x\rangle - \langle w^*, x\rangle\right|^2\right] \\
&\leq \left\langle \mathbb{E}\left[\phi_y'(\langle w^*, x\rangle)x\right], \bar{w}^* - w^*\right\rangle + \frac{H}{2}\mathbb{E}\left[\left|\langle \bar{w}^*, x\rangle - \langle w^*, x\rangle\right|^2\right] \\
&\leq \langle \nabla L(w^*;\mathcal{D}), \bar{w}^* - w^*\rangle + H\alpha^2 \\
&= H\alpha^2
\end{aligned}
$$

where last equality uses the fact that $\nabla L(w^*;\mathcal{D}) = 0$ since w^* is the unconstrained minimizer. Continuing, we get,

$$
\begin{aligned}
&\left|L(\widehat{w};\mathcal{D}) - \widehat{L}(\widehat{w};S_{\text{priv}})\right| + \left|L(\mathring{w}^*;\mathcal{D}) - \widehat{L}(\mathring{w}^*;S_{\text{priv}})\right| \\
&= \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{priv}}\epsilon}} + \sqrt{\frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}}\right)\sqrt{L(w^*;\mathcal{D})} \\
&+ \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{\text{priv}}\epsilon} + \frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}\right) + O\left(GD\left\|\mathcal{X}\right\|\left(\frac{\sqrt{n_{\text{pub}}\log\left(1/\delta\right)} + H\alpha^2 + \sqrt{\log\left(4/\beta\right)}}{n_{\text{priv}}\epsilon}\right)\right)
\end{aligned}
$$

This yields,

$$
\begin{aligned}
&L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D}) \\
&= \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{priv}}\epsilon}} + \sqrt{\frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}}\right)\sqrt{L(w^*;\mathcal{D})} \\
&+ \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{\text{priv}}\epsilon} + H\alpha^2 + \frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}\right) + O\left(GD\left\|\mathcal{X}\right\|\left(\frac{\sqrt{n_{\text{pub}}\log\left(1/\delta\right)} + \sqrt{\log\left(4/\beta\right)}}{n_{\text{priv}}\epsilon}\right)\right) \\
&= \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{priv}}\epsilon}} + \sqrt{\frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}}\right)\sqrt{L(w^*;\mathcal{D})} \\
&+ \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{\text{priv}}\epsilon} + \frac{HD^2\left\|\mathcal{X}\right\|^2}{n_{\text{pub}}} + \frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}\right) + O\left(GD\left\|\mathcal{X}\right\|\left(\frac{\sqrt{n_{\text{pub}}\log\left(1/\delta\right)} + \sqrt{\log\left(4/\beta\right)}}{n_{\text{priv}}\epsilon}\right)\right) \\
&= \tilde{O}\left(\frac{\sqrt{H}D\left\|\mathcal{X}\right\|}{\sqrt{n_{\text{priv}}\epsilon}} + \sqrt{\frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}}\right)\sqrt{L(w^*;\mathcal{D})} + \tilde{O}\left(\frac{H\left\|\mathcal{X}\right\|^2 D^2}{n_{\text{priv}}\epsilon} + \frac{B\log\left(8/\beta\right)}{n_{\text{priv}}}\right) \\
&+ O\left(\frac{GD\left\|\mathcal{X}\right\|\sqrt{\log\left(4/\beta\right)}}{n_{\text{priv}}\epsilon}\right) + O\left(\left(\frac{\sqrt{H}D^2\left\|\mathcal{X}\right\|^2 G\sqrt{\log\left(1/\delta\right)}}{n_{\text{priv}}\epsilon}\right)^{2/3}\right)
\end{aligned}
$$

The public sample complexity is,

$$
n_{\text{pub}} = \tilde{O}\left(\frac{(HD\left\|\mathcal{X}\right\|)^{2/3}(n_{\text{priv}}\epsilon)^{2/3}}{G^{2/3}(\log\left(1/\delta\right))^{1/3}}\right).
$$

The second part follows similarly.

□

C.3 Lower bounds

C.3.1 Proof of Theorem 4

To establish the $\frac{GD\|\mathcal{X}\|}{\sqrt{n}}$ term in the lower bound, we consider a one-dimenisonal problem where the loss $\phi_y(\widehat{y}) = -Gy\widehat{y}$ and marginal distribution $\mathcal{D}_x$ as the point distribution on $\|\mathcal{X}\|$ such that the overall loss is $\mathop{\mathbb{E}}_{x,y}\left[\ell(w,(x,y))\right] = \mathop{\mathbb{E}}_{y}\left[y\cdot w\|\mathcal{X}\|G\right]$. We further set $\mathcal{W} = [-D, D]$ and consider $\mathcal{D}_y$ to be the distribution which as 1 with probability $\mathbb{P}\left[y=1\right] = (1+\mu)/2$ and $\mathbb{P}\left[y=1\right] = (1-\mu)/2$ for some $\mu \in [-1,1]$. Note the minimizer $w^* = D\frac{\mu}{|\mu|}$ achieves population risk $-\mu GD\|\mathcal{X}\|$. Classic results in information theory establish if μ is sampled uniformly from $\{\pm\frac{1}{\sqrt{6n}}\}$, no algorithm can estimate the sign of μ with probability better than $1/2$ (see [Duc23, Section 8.3]). Thus it must be that for any algorithm $\mathbb{E}_{\mathcal{A},S}\left[L(\mathcal{A}(S);\mathcal{D}) - \min_{w\in\mathbb{R}^d} L(w;\mathcal{D})\right] = \Omega\left(\frac{GD\|\mathcal{X}\|}{\sqrt{n}}\right)$.

The $GD\left\|\mathcal{X}\right\|\min\left\{\frac{1}{\sqrt{n\epsilon}}, \frac{\sqrt{d}}{n\epsilon}\right\}$ term in the lower bound is essentially a corollary of [ABG+22, Theorem 6]. We provide further remarks here. The loss function used is,

$$
\ell(w;(x,y)) = \phi_y(\langle w,x\rangle) = \left|y - \langle w,x\rangle\right|.
$$

Define $d' := \min(d, n\epsilon)$ and $p := \min\left(1, \frac{d'}{n\epsilon}\right)$. The (known) marginal distribution $\mathcal{D}_{\mathcal{X}}$ is described as: with probability $1-p$, $x = \vec{0}$, otherwise, $x \sim \text{Unif}\left(\|\mathcal{X}\|\left\{e_j\right\}_{j=1}^{d'}\right)$ where e_j's are canonical basis vectors. The (unknown) conditional distribution of the response y is as follows. Sample a "fingerprinting code", $z' \in \{0,1\}^{d'}$ with mean $\mu' \in [0,1]^{d'}$ where each co-ordinate $\mu'_j \sim \text{Beta}(0.0625, 0.0625)$ i.i.d. Embed z' in d dimensions as z and let μ be the corresponding mean vector. Finally, define $y = \frac{D\langle z,x\rangle}{\sqrt{d'}}$. The proof in [ABG+22, Theorem 6] then proceeds by lower bounding the loss by bounding the ability of any differential private algorithm to estimate the fingerprinting code z.

Since the rank of $\mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[xx^\top\right] = d'$, the result [ABG+22, Theorem 6] then yields a lower bound on the *unconstrained excess risk*,

$$\mathbb{E}_{\mathcal{A},S}\left[L(\mathcal{A}(S);\mathcal{D}) - \min_{w\in\mathbb{R}^d} L(w;\mathcal{D})\right] = \Omega\left(GD\,\|\mathcal{X}\|\min\left\{\frac{1}{\sqrt{n\epsilon}}, \frac{\sqrt{d}}{n\epsilon}\right\}\right),$$

but also guarantees that the global minimizer has norm at most D. Thus, we achieve the same lower bound for $\mathbb{E}_{\mathcal{A},S}\left[L(\mathcal{A}(S);\mathcal{D}) - \min_{w\in\mathcal{W}} L(w;\mathcal{D})\right]$ by setting $\mathcal{W}$ to be the ball of radius D.

C.3.2 Proof of Theorem 5

The proof uses the lower bound instance in the DP-SCO lower bound with public data, Theorem 1. We consider the case where $\mathcal{D}_y$ is the point distribution on 1. Then for any $y \in \mathcal{Y}$, $\mathcal{Y} = \{1\}$, the loss function is then $\ell(w;(x,y)) = y\langle w, x\rangle = \langle w, x\rangle$, as in Theorem 1. Hence, a labeled and unlabeled sample have the same information. We also set $\mathcal{W}$ to be the ball of radius D.

Assume by contradiction there exists an (ϵ,δ)-PA-DP algorithm, $\mathcal{A}$, which achieves rate $O\left(GD\|\mathcal{X}\|(\frac{1}{\sqrt{n_{\text{priv}}}} + \frac{\sqrt{\log(1/\delta)}}{\sqrt{n_{\text{priv}}\epsilon}})\right)$ with $o(n_{\text{priv}}\epsilon/\log(1/\delta))$ public samples. Since $n_{\text{pub}} = o(n_{\text{priv}}\epsilon/\log(1/\delta))$ and $d = \omega(n\epsilon)$, Theorem 1 gives a lower bound on $\mathbb{E}\left[\mathcal{A}(X_{\text{pub}}, S_{\text{priv}};\mathcal{D}) - \min_{w\in\mathcal{W}}\{L(w;\mathcal{D})\}\right]$ of

$$\Omega\left(GD\|\mathcal{X}\|\min\left\{\frac{1}{\sqrt{n_{\text{pub}}}}, \frac{\sqrt{d\log(1/\delta)}}{n_{\text{priv}}\epsilon}\right\}\right) = \omega\left(GD\|\mathcal{X}\|\frac{\sqrt{\log(1/\delta)}}{\sqrt{n_{\text{priv}}\epsilon}}\right).$$

Since $\epsilon \le 1$, this is a contradiction.

D Missing proofs for Section 4.2

D.1 Proof of Theorem 7

Proof. The privacy proof follows from the guarantee of exponential mechanism [MT07]. In particular, the sensitivity of the score function is at most $\frac{2}{n_{\text{priv}}}\min(B, GR)$ where the first follows from the loss bound of B and the second from the Lipschitzness and bound on predictors. Let $h^* \in \arg\min_{h\in\mathcal{H}} L(h;\mathcal{D})$ and $\tilde{h}^* \in \arg\min_{h\in\tilde{\mathcal{H}}} L(h;\mathcal{D})$. From standard analysis based on uniform convergence, we have

$$\begin{aligned}
L(\widehat{h};\mathcal{D}) - L(h^*;\mathcal{D}) &= L(\widehat{h};\mathcal{D}) - \widehat{L}(\widehat{h};S_{\text{priv}}) + \widehat{L}(\widehat{h};S_{\text{priv}}) - L(h^*;\mathcal{D}) \\
&\le \sup_{h\in\mathcal{H}}\left(L(h;\mathcal{D}) - \widehat{L}(h;S_{\text{priv}})\right) + L(\tilde{h}^*;\mathcal{D}) - L(h^*;\mathcal{D}) \\
&\quad + \widehat{L}(\tilde{h}^*;S_{\text{priv}}) - L(\tilde{h}^*;\mathcal{D}) + \widehat{L}(\widehat{h};S_{\text{priv}}) - \widehat{L}(\tilde{h}^*;S_{\text{priv}}) \qquad (17) \\
&\le 2\sup_{h\in\mathcal{H}}\left|L(h;\mathcal{D}) - \widehat{L}(h;S_{\text{priv}})\right| + L(\tilde{h}^*;\mathcal{D}) - L(h^*;\mathcal{D}) \\
&\quad + \widehat{L}(\widehat{h};S_{\text{priv}}) - \min_{h\in\tilde{\mathcal{H}}}\widehat{L}(h;S_{\text{priv}}) \qquad (18) \\
&\le 2G\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H}) + O\left(\frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}}\right) \\
&\quad + O\left(\frac{\min(B, GR)(\log\left(|\tilde{\mathcal{H}}|\right) + \log(4/\beta))}{n_{\text{priv}}\epsilon}\right) + L(\tilde{h}^*;\mathcal{D}) - L(h^*;\mathcal{D}) \qquad (19)
\end{aligned}$$

where the above holds with probability at least $1 - \beta/2$ and follows from guarantee of exponential mechanism [MT07] and uniform convergence ([SSBD14], see Theorem 16). We further have that $\log\left(|\tilde{\mathcal{H}}|\right) = \tilde{O}(\text{fat}_{c\alpha}(\mathcal{H}))$ from Lemma 7.

For the $L(\tilde{h}^*;\mathcal{D}) - L(h^*;\mathcal{D})$ term, we have

$$\begin{aligned} L(\tilde{h}^*;\mathcal{D}) - L(h^*;\mathcal{D}) &\leq L(\bar{h}^*;\mathcal{D}) - L(h^*;\mathcal{D}) \\ &\leq G\mathbb{E}\left|\bar{h}^*(x) - h^*(x)\right| \\ &\leq G\sqrt{\mathbb{E}\left|\bar{h}^*(x) - h^*(x)\right|^2} \\ &\leq 2G\alpha \end{aligned}$$

where the first step holds for any $\bar{h}^* \in \tilde{\mathcal{H}}$ by optimaility of $\tilde{h}^*$ over $\tilde{\mathcal{H}}$, the second holds from the G-Lipschitzness of the loss function, the third from Jensen's inequality. The final step holds with probability $1 - \beta/2$ from Lemma 1 with $n_{\text{pub}} = O\left(\max\left(\frac{R^2\log(2/\beta)}{\alpha^2}, \min\left\{m : \log^3(m)\mathfrak{R}_m^2(\mathcal{H}) \leq \alpha^2\right\}\right)\right)$ together with the fact that since $\tilde{\mathcal{H}}$ is an α-cover of $\mathcal{H}$, hence there exists $\bar{h}^* \in \tilde{\mathcal{H}}$ with $\left\|\bar{h}^* - h^*\right\|_{2,X_{\text{pub}}} \leq \alpha$.

Plugging the above in Equation (19), we get with probability at least $1 - \beta$,

$$\begin{aligned} L(\widehat{h};\mathcal{D}) - L(h^*;\mathcal{D}) \leq 2G\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H}) &+ O\left(\frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}}\right) \\ &+ O\left(\frac{\min(B, GR)(\text{fat}_{c\alpha}(\mathcal{H}) + \log(4/\beta))}{n_{\text{priv}}\epsilon}\right) + 2G\alpha, \end{aligned}$$

which finishes the proof. □

Theorem 16. *[SSBD14] Let $\mathcal{H} \subseteq [-R, R]^{\mathcal{X}}$. For any G-Lipschitz, B-bounded loss function, any probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, given m i.i.d. samples from S, with probability at least $1 - \beta$, the following holds for all $h \in \mathcal{H}$,*

$$\left|L(h;\mathcal{D}) - \widehat{L}(h;S)\right| \leq 2G\mathfrak{R}_m(\mathcal{H}) + O\left(B\sqrt{\frac{\log(4/\beta)}{m}}\right)$$

Proof. This is a classical result in learning theory which follows directly Theorem 26.5 in [SSBD14] together with the contraction lemma (Lemma 26.9 in [SSBD14]) for Lipschitz losses. □

Lemma 7. *Let $\tilde{\mathcal{H}}$ be an α-cover of $\mathcal{H}$ with respect to $\|\cdot\|_{2,X_{pub}}$. The size of $\tilde{\mathcal{H}}$ is bounded as,*

$$\log\left(\left|\tilde{\mathcal{H}}\right|\right) \leq \mathit{fat}_{c\alpha}(\mathcal{H})\log\left(\frac{2R}{\alpha}\right)$$

where c is an absolute constant.

Proof. This follows directly from Theorem 8,

$$\log\left(\left|\tilde{\mathcal{H}}\right|\right) = \mathcal{N}_2(\mathcal{H}, \alpha, S_{\text{pub}}) \leq \mathcal{N}_2(\mathcal{H}, \alpha, n_{\text{pub}}) \leq \text{fat}_{c\alpha}(\mathcal{H})\log\left(\frac{2R}{\alpha}\right).$$

□

Proof of Lemma 1 For $h \in \mathcal{H}$, let $\tilde{h} \in \arg\min_{\bar{h}\in\tilde{\mathcal{H}}} \|\bar{h} - h\|_{2,X_{\text{pub}}}$. Since $\tilde{\mathcal{H}}$ is an τ-cover, this gives us that $\|h - \tilde{h}\|_{2,X_{\text{pub}}} \leq \tau$. We have,

$$\begin{aligned} \|h - \tilde{h}\|_{2,\mathcal{D}_{\mathcal{X}}}^2 &= \mathbb{E}\left|h(x) - \tilde{h}(x)\right|^2 \\ &= \mathbb{E}\left|h(x) - \tilde{h}(x)\right|^2 - \frac{1}{n_{\text{pub}}}\sum_{x\in S_{\text{pub}}}\left|h(x) - \tilde{h}(x)\right|^2 + \frac{1}{n_{\text{pub}}}\sum_{x\in S_{\text{pub}}}\left|h(x) - \tilde{h}(x)\right|^2 \\ &\leq \sup_{\bar{h}\in\mathcal{H}}\left(\mathbb{E}\left|h(x) - \bar{h}(x)\right|^2 - \frac{1}{n_{\text{pub}}}\sum_{x\in S_{\text{pub}}}\left|h(x) - \bar{h}(x)\right|^2\right) + \tau^2. \end{aligned}$$

The first term above can be seen as uniform deviation between the empirical and population risk, of another prediction problem, with squared loss, in the the realizable setting (with the responses generated by h). The squared loss is $\frac{1}{2}$-smooth and non-negative, so we can apply result of Theorem 1 in [SST10] instantiated in the realizable setting, which gives us that with probability at least $1-\beta$,

$$\|h - \tilde{h}\|^2_{2.\mathcal{D}_\mathcal{X}} = O\left(\log^3(n_{\text{pub}})\mathfrak{R}^2_{n_{\text{pub}}}(\mathcal{H}) + \frac{R^2\log(1/\beta)}{n_{\text{pub}}}\right) + \tau^2.$$

Choosing n_{pub} such that $n_{\text{pub}} = O\left(\max\left(\frac{R^2\log(1/\beta)}{\alpha^2}, \min\left\{m : \log^3(m)\mathfrak{R}^2_m(\mathcal{H}) \le \alpha^2\right\}\right)\right)$, we get the claimed result.

D.2 Optimistic rates with smooth non-negative losses

Algorithm 2 achieves optimistic rates on risk depending on realizability/interpolation conditions, that is, whenever $L(h^*;\mathcal{D})$ or $\widehat{L}(\widehat{h};S_{\text{priv}})$ is small.

Theorem 17. *Algorithm 2 with $\gamma = \frac{2B}{n_{priv}\epsilon}$ satisfies ϵ-PA-DP. For $n_{pub} = O\left(\max\left(\frac{R^2\log(2/\beta)}{\alpha^2}, \min\left\{m : log^3(m)\mathfrak{R}^2_m(\mathcal{H}) \le \alpha^2\right\}\right)\right) < \infty$ and any $\alpha > 0$, with probability at least $1-\beta$, we have*

$$L(\widehat{h};\mathcal{D}) - L(h^*;\mathcal{D}) = \tilde{O}\left(\sqrt{H}\mathfrak{R}_{n_{priv}}(\mathcal{H}) + \sqrt{H}\alpha + \sqrt{\frac{B\log(8/\delta)}{n_{priv}}}\right)\sqrt{L(h^*;\mathcal{D})}$$
$$+ \tilde{O}\left(H\mathfrak{R}^2_{n_{priv}}(\mathcal{H}) + H\alpha^2 + \frac{B\log(8/\beta)}{n_{priv}} + \frac{B(fat_{c\alpha}(\mathcal{H}) + \log(4/\beta))}{n_{priv}\epsilon}\right)$$

$$L(\widehat{h};\mathcal{D}) - \widehat{L}(\widehat{h}^*;S_{priv}) \le \tilde{O}\left(\sqrt{H}\mathfrak{R}_{n_{priv}}(\mathcal{H}) + \sqrt{H}\alpha + \sqrt{\frac{B\log(8/\delta)}{n_{priv}}}\right)\sqrt{\widehat{L}(\widehat{h}^*;S_{priv})}$$
$$+ \tilde{O}\left(H\mathfrak{R}^2_{n_{priv}}(\mathcal{H}) + H\alpha^2 + \frac{B\log(8/\beta)}{n_{priv}} + \frac{B(fat_{c\alpha}(\mathcal{H}) + \log(4/\beta))}{n_{priv}\epsilon}\right)$$

where h^* and $\widehat{h}^*$ are population and empirical minimizers with respect to $\mathcal{D}$ and S_{priv} respectively, and c is an absolute constant.

Proof of Theorem 17. The privacy proof is the same as that of Theorem 7. For utility, let $h^* \in \arg\min_{h\in\mathcal{H}} L(h;\mathcal{D})$ and $\tilde{h}^* \in \arg\min_{h\in\tilde{\mathcal{H}}} L(h;\mathcal{D})$.

We start with the proof of the first part of the theorem. We have,

$$\begin{aligned}
L(\widehat{h};\mathcal{D}) - L(h^*;\mathcal{D}) &= L(\widehat{h};\mathcal{D}) - \widehat{L}(\widehat{h};S_{\text{priv}}) + \widehat{L}(\widehat{h};S_{\text{priv}}) - L(h^*;\mathcal{D}) \\
&\le L(\widehat{h};\mathcal{D}) - \widehat{L}(\widehat{h};S_{\text{priv}}) + \widehat{L}(\tilde{h}^*;S_{\text{priv}}) - L(\tilde{h}^*;\mathcal{D}) \\
&+ L(\tilde{h}^*;\mathcal{D}) - L(h^*;\mathcal{D}) + \widehat{L}(\widehat{h};S_{\text{priv}}) - \widehat{L}(\tilde{h}^*;S_{\text{priv}}) \\
&\le \left|L(\widehat{h};\mathcal{D}) - \widehat{L}(\widehat{h};S_{\text{priv}})\right| + \left|L(\tilde{h}^*;\mathcal{D}) - \widehat{L}(\tilde{h}^*;S_{\text{priv}})\right| \\
&+ L(\tilde{h}^*;\mathcal{D}) - L(h^*;\mathcal{D}) + \widehat{L}(\widehat{h};S_{\text{priv}}) - \widehat{L}(\tilde{h}^*;S_{\text{priv}}) \qquad (20)
\end{aligned}$$

From the guarantee of exponential mechanism together with $\log\left(\left|\tilde{\mathcal{H}}\right|\right) = \tilde{O}(\text{fat}_{c\alpha}(\mathcal{H}))$ from Lemma 7, we have that with probability at least $1-\beta/4$,

$$\widehat{L}(\widehat{h};S_{\text{priv}}) - \widehat{L}(\tilde{h}^*;S_{\text{priv}}) \le \widehat{L}(\widehat{h};S_{\text{priv}}) - \min_{h\in\tilde{\mathcal{H}}}\widehat{L}(h;S_{\text{priv}}) = O\left(\frac{B(\log\left(\left|\tilde{\mathcal{H}}\right|\right) + \log(4/\beta))}{n_{\text{priv}}\epsilon}\right)$$
$$= \tilde{O}\left(\frac{B(\text{fat}_{c\alpha}(\mathcal{H}) + \log(4/\beta))}{n_{\text{priv}}\epsilon}\right) \qquad (21)$$

For the $L(\tilde{h}^*;\mathcal{D}) - L(h^*;\mathcal{D})$ term in Equation (20), we apply smoothness to get,

$$\begin{aligned}
L(\tilde{h}^*;\mathcal{D}) - L(h^*;\mathcal{D}) &\leq L(\bar{h}^*;\mathcal{D}) - L(h^*;\mathcal{D}) \\
&\leq \mathbb{E}\left[\left\langle \phi_y'(h^*(x)), \bar{h}^*(x) - h^*(x)\right\rangle + \frac{H}{2}\left|\bar{h}^*(x) - h^*(x)\right|^2\right] \\
&\leq \mathbb{E}\left[\left|\phi_y'(h^*(x))\right|\left|\bar{h}^*(x) - h^*(x)\right| + \frac{H}{2}\left|\bar{h}^*(x) - h^*(x)\right|^2\right] \\
&\leq \sqrt{\mathbb{E}\left|\phi_y'(h^*(x))\right|^2}\sqrt{\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left|\bar{h}^*(x) - h^*(x)\right|^2} + \frac{H}{2}\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left|\bar{h}^*(x) - h^*(x)\right|^2 \\
&\leq 2\sqrt{H\mathbb{E}_{x\sim\mathcal{D}}\phi_y(h^*(x))}\sqrt{\mathbb{E}\left|\bar{h}^*(x) - h^*(x)\right|^2} + \frac{H}{2}\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left|\bar{h}^*(x) - h^*(x)\right|^2 \\
&\leq 2\sqrt{HL(h^*;\mathcal{D})}\alpha + H\alpha^2 \qquad (22)
\end{aligned}$$

where the above holds for any $\bar{h}^* \in \tilde{\mathcal{H}}$. The second inequality holds from H-smoothness, the third and fourth from Cauchy-Schwarz, the fifth from self-bounding property of smooth non-negative losses [SST10]. The final step holds with probability $1 - \beta/2$ from Lemma 1 with $n_{\text{pub}} = O\left(\max\left(\frac{R^2\log(2/\beta)}{\alpha^2}, \min\left\{m : \log^3(m)\mathfrak{R}_m^2(\mathcal{H}) \leq \alpha^2\right\}\right)\right)$ together with the fact that since $\tilde{\mathcal{H}}$ is an α-cover of $\mathcal{H}$, so there exists $\bar{h}^* \in \tilde{\mathcal{H}}$ with $\left\|\bar{h}^* - h^*\right\|_{2,X_{\text{pub}}} \leq \alpha$.

An application of AM-GM inequality further yields,

$$L(\tilde{h}^*;\mathcal{D}) \leq 2L(h^*;\mathcal{D}) + 2H\alpha^2 \qquad (23)$$

The first two terms in Equation (20) are bound using uniform convergence for smooth non-negative losses, Theorem 1 in [SST10] and Bernstein's inequality as follows; with probability at least $1 - \beta/4$, we have,

$$\begin{aligned}
&\left|L(\widehat{h};\mathcal{D}) - \widehat{L}(\widehat{h};S_{\text{priv}})\right| + \left|L(\tilde{h}^*;\mathcal{D}) - \widehat{L}(\tilde{h}^*;S_{\text{priv}})\right| \\
&= \tilde{O}\left(\sqrt{H}\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H}) + \sqrt{\frac{B\log(8/\beta)}{n_{\text{priv}}}}\right)\left(\sqrt{\widehat{L}(\widehat{h};S_{\text{priv}})} + \sqrt{L(\tilde{h}^*;\mathcal{D})}\right) \\
&+ \tilde{O}\left(\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H})^2 + \frac{B\log(8/\beta)}{n_{\text{priv}}}\right) \\
&= \tilde{O}\left(\sqrt{H}\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H}) + \sqrt{\frac{B\log(8/\beta)}{n_{\text{priv}}}}\right)\left(\sqrt{\widehat{L}(\tilde{h}^*;S_{\text{priv}})} + \sqrt{L(\tilde{h}^*;\mathcal{D})}\right) \\
&+ \tilde{O}\left(H\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H})^2 + \frac{B\log(8/\beta)}{n_{\text{priv}}}\right) + \tilde{O}\left(\frac{B(\text{fat}_{c\alpha}(\mathcal{H}) + \log(4/\beta))}{n_{\text{priv}}\epsilon}\right) \\
&= \tilde{O}\left(\sqrt{H}\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H}) + \sqrt{\frac{B\log(8/\beta)}{n_{\text{priv}}}}\right)\sqrt{L(\tilde{h}^*;\mathcal{D})} \\
&+ \tilde{O}\left(H\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H})^2 + \frac{B\log(8/\beta)}{n_{\text{priv}}}\right) + \tilde{O}\left(\frac{B(\text{fat}_{c\alpha}(\mathcal{H}) + \log(4/\beta))}{n_{\text{priv}}\epsilon}\right) \\
&= \tilde{O}\left(\sqrt{H}\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H}) + \sqrt{\frac{B\log(8/\beta)}{n_{\text{priv}}}}\right)\sqrt{L(h^*;\mathcal{D})} \\
&+ \tilde{O}\left(H\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H})^2 + \frac{B\log(8/\beta)}{n_{\text{priv}}}\right) + \tilde{O}\left(\frac{B(\text{fat}_{c\alpha}(\mathcal{H}) + \log(8/\beta))}{n_{\text{priv}}\epsilon}\right)
\end{aligned}$$

where the second equality follows from Equation (21), concavity of $x \mapsto \sqrt{x}$ and AM-GM inequality. The third equality follows concavity of $x \mapsto \sqrt{x}$ and Bernstein's inequality, the fourth follows from Equation (23) and AM-GM inequality.

Plugging the above, Equation (22) and Equation (21) into Equation (20) yields the following with probability at least $1-\beta$,

$$
\begin{aligned}
&L(\widehat{h};\mathcal{D}) - L(h^*;\mathcal{D}) \\
&= \tilde{O}\left(\sqrt{H}\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H}) + \sqrt{H}\alpha + \sqrt{\frac{B\log(8/\beta)}{n_{\text{priv}}}}\right)\sqrt{L(h^*;\mathcal{D})} \qquad (24)\\
&+ \tilde{O}\left(H\mathfrak{R}^2_{n_{\text{priv}}}(\mathcal{H}) + H\alpha^2 + \frac{B\log(8/\beta)}{n_{\text{priv}}} + \frac{B(\text{fat}_{c\alpha}(\mathcal{H}) + \log(4/\beta))}{n_{\text{priv}}\epsilon}\right) \qquad (25)
\end{aligned}
$$

This completes the first part of the theorem. For the second part, we proceed from Equation (25) onwards

$$
\begin{aligned}
L(\widehat{h};\mathcal{D}) &\leq L(h^*;\mathcal{D}) + \tilde{O}\left(\sqrt{H}\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H}) + \sqrt{H}\alpha + \sqrt{\frac{B\log(8/\beta)}{n_{\text{priv}}}}\right)\sqrt{L(h^*;\mathcal{D})} \\
&+ \tilde{O}\left(H\mathfrak{R}^2_{n_{\text{priv}}}(\mathcal{H}) + H\alpha^2 + \frac{B\log(8/\beta)}{n_{\text{priv}}} + \frac{B(\text{fat}_{c\alpha}(\mathcal{H}) + \log(4/\beta))}{n_{\text{priv}}\epsilon}\right) \\
&\leq L(\widehat{h}^*;\mathcal{D}) + \tilde{O}\left(\sqrt{H}\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H}) + \sqrt{H}\alpha + \sqrt{\frac{B\log(8/\beta)}{n_{\text{priv}}}}\right)\sqrt{L(\widehat{h}^*;\mathcal{D})} \\
&+ \tilde{O}\left(H\mathfrak{R}^2_{n_{\text{priv}}}(\mathcal{H}) + H\alpha^2 + \frac{B\log(8/\beta)}{n_{\text{priv}}} + \frac{B(\text{fat}_{c\alpha}(\mathcal{H}) + \log(4/\beta))}{n_{\text{priv}}\epsilon}\right) \\
&\leq \widehat{L}(\widehat{h}^*;S_{\text{priv}}) + \tilde{O}\left(\sqrt{H}\mathfrak{R}_{n_{\text{priv}}}(\mathcal{H}) + \sqrt{H}\alpha + \sqrt{\frac{B\log(8/\beta)}{n_{\text{priv}}}}\right)\sqrt{\widehat{L}(\widehat{h}^*;S_{\text{priv}})} \\
&+ \tilde{O}\left(H\mathfrak{R}^2_{n_{\text{priv}}}(\mathcal{H}) + H\alpha^2 + \frac{B\log(8/\beta)}{n_{\text{priv}}} + \frac{B(\text{fat}_{c\alpha}(\mathcal{H}) + \log(4/\beta))}{n_{\text{priv}}\epsilon}\right)
\end{aligned}
$$

where the second inequality follows form optimality of h^*, the third from uniform convergence, Theorem 1 in [SST10] and AM-GM inequality. This completes the proof.

□

D.3 Proof of Corollary 2

We use the result from [GRS18], restated as Theorem 10. Further, note that range bound on the hypothesis class is simply $R \leq \|\mathcal{X}\|\prod_{j=1}^{d} R_j$. Instantiating our general result Theorem 7 with the above together with the relation between fat-shattering dimension and Rademacher complexity (Theorem 9) yields the following excess risk bound,

$$
\begin{aligned}
&L(\widehat{h};\mathcal{D}) - \min_{h\in\mathcal{H}} L(h;\mathcal{D}) \\
&= O\left(\frac{G\|\mathcal{X}\|(\sqrt{2\log(2)M}+1)\prod_{j=1}^{M}R_j}{\sqrt{n_{\text{priv}}}} + \frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}} + \frac{B\log(4/\beta)}{n_{\text{priv}}\epsilon}\right) \\
&+ \tilde{O}\left(\left(\frac{BG^2(\sqrt{2\log(2)M}+1)^2\|\mathcal{X}\|^2(\prod_{j=1}^{M}R_j)^2}{n_{\text{priv}}\epsilon}\right)^{1/3}\right).
\end{aligned}
$$

where in the above, we set $\alpha = \left(\frac{B(\sqrt{2\log(2)M}+1)^2\|\mathcal{X}\|^2(\prod_{j=1}^{M}R_j)^2}{Gn_{\text{priv}}\epsilon}\right)^{1/3}$.

The number of public samples then is

$$\begin{aligned} n_{\text{pub}} &= O\left(\max\left(\frac{R^2\log(2/\beta)}{\alpha^2}, \min\left\{m : \log^3(m)\mathfrak{R}_m^2(\mathcal{H}) \le \alpha^2\right\}\right)\right) \\ &= \tilde{O}\left(\|\mathcal{X}\|^2 (\prod_{j=1}^{M} R_j)^2 \max\left(\frac{\log(2/\beta)}{\alpha^2}, \frac{M}{\alpha^2}\right)\right) \\ &= \tilde{O}\left((\|\mathcal{X}\| (\prod_{j=1}^{M} R_j))^{2/3} (n_{\text{priv}}\epsilon)^{2/3} M^{1/3} \log(2/\beta)\right). \end{aligned}$$

D.4 Proof of Corollary 3

Note that $R \le D\|\mathcal{X}\|$. Further, we have [KST08, FGV17],

$$\mathfrak{R}_m(\mathcal{H}) = O\left(\frac{D\|\mathcal{X}\|}{m^{1/r}}\right).$$

Moreover, we have from Theorem 9, for any $\alpha > \mathfrak{R}_m(\mathcal{H})$,

$$\text{fat}_\alpha(\mathcal{H}) \le \frac{4m\mathfrak{R}_m^2(\mathcal{H})}{\alpha^2}.$$

Choose $m = n_{\text{pub}}$ and $\alpha = (\log(n_{\text{pub}}))^{3/2}\mathfrak{R}_{n_{\text{pub}}}(\mathcal{H})$, to get that $\text{fat}_\alpha(\mathcal{H}) = \tilde{O}(n_{\text{pub}})$. Plugging this in Theorem 7, we get,

$$\begin{aligned} L(\widehat{w};\mathcal{D}) - \min_{w\in\mathcal{W}} L(w;\mathcal{D}) &= O\left(\frac{GD\|\mathcal{X}\|}{{n_{\text{priv}}}^{1/r}} + \frac{GD\|\mathcal{X}\|\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}} + \frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}}\right) \\ &\quad + \tilde{O}\left(GD\|\mathcal{X}\|\left(\frac{n_{\text{pub}}}{n_{\text{priv}}\epsilon} + \frac{1}{{n_{\text{pub}}}^{1/r}} + \frac{\log(4/\beta)}{n_{\text{priv}}\epsilon}\right)\right) \\ &= \tilde{O}\left(GD\|\mathcal{X}\|\left(\frac{1}{{n_{\text{priv}}}^{1/r}} + \frac{\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}} + \frac{\log(2/\beta)}{(n_{\text{priv}}\epsilon)^{\frac{1}{r+1}}} + \frac{\log(4/\beta)}{n_{\text{priv}}\epsilon}\right)\right) \\ &\quad + O\left(\frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}}\right), \end{aligned}$$

where in the last step, we plug in $n_{\text{pub}} = O\left((n_{\text{priv}}\epsilon)^{r/(1+r)}\log(2/\beta)\right)$, yielding $\alpha = O\left(\frac{D\|\mathcal{X}\|}{(n_{\text{priv}}\epsilon)^{\frac{1}{r+1}}}\right)$. The number of public samples simplifies as,

$$\begin{aligned} n_{\text{pub}} &= O\left(\max\left(\frac{R^2\log(2/\beta)}{\alpha^2}, \min\left\{m : \log^3(m)\mathfrak{R}_m^2(\mathcal{H}) \le \alpha^2\right\}\right)\right) \\ &= \tilde{O}\left((n_{\text{priv}}\epsilon)^{\frac{2}{r+1}}\log(2/\beta)\right). \end{aligned}$$

which is satisfied from our choice since $r \ge 2$.

D.5 Additional Results

Corollary 4. *In the setting of Theorem 7 together with* $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \le \|\mathcal{X}\|\}$ *and* $\mathcal{H} = \{x \mapsto \langle w, x\rangle, x \in \mathcal{X}, \|w\| \le D\}$. *With* $\alpha = \left(\frac{D^2\|\mathcal{X}\|^2}{n_{priv}\epsilon}\right)^{1/3}$ *and* $n_{pub} = \tilde{O}\left((D\|\mathcal{X}\|)^{2/3}(n_{priv}\epsilon)^{2/3}\log(2/\beta)\right)$, *with probability at least* $1-\beta$,

$$L(\widehat{h};\mathcal{D}) - \min_{h\in\mathcal{H}} L(h;\mathcal{D}) = \tilde{O}\left(\frac{GD\|\mathcal{X}\|}{\sqrt{n_{priv}}} + G\left(\frac{D^2\|\mathcal{X}\|^2}{n_{priv}\epsilon}\right)^{1/3} + \frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{priv}}} + \frac{B\log(4/\beta)}{n_{priv}\epsilon}\right).$$

Proof. In this setting, we have that $R \leq D \|\mathcal{X}\|$. Further, it is known [KST08],

$$\text{fat}_\alpha(\mathcal{H}) = O\left(\frac{D^2 \|\mathcal{X}\|^2}{\alpha^2}\right), \qquad \mathfrak{R}_n(\mathcal{H}) = O\left(\frac{D \|\mathcal{X}\|}{\sqrt{m}}\right).$$

Plugging this in Theorem 7, we get,

$$\begin{aligned}
&L(\widehat{h}; \mathcal{D}) - \min_{h \in \mathcal{H}} L(h; \mathcal{D}) \\
&= O\left(\frac{GD \|\mathcal{X}\|}{\sqrt{n_{\text{priv}}}}\right) + O\left(\frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}}\right) + \tilde{O}\left(\frac{\min(B, GD \|\mathcal{X}\|) D \|\mathcal{X}\|}{\alpha^2 n_{\text{priv}} \epsilon}\right) \\
&+ 2G\alpha + O\left(\frac{B \log(4/\beta)}{n_{\text{priv}} \epsilon}\right) \\
&= \tilde{O}\left(\frac{GD \|\mathcal{X}\|}{\sqrt{n_{\text{priv}}}} + G\left(\frac{D^2 \|\mathcal{X}\|^2}{n_{\text{priv}} \epsilon}\right)^{1/3} + \frac{B\sqrt{\log(4/\beta)}}{\sqrt{n_{\text{priv}}}} + \frac{B \log(4/\beta)}{n_{\text{priv}} \epsilon}\right).
\end{aligned}$$

where in the last step, we plug in $\alpha = \left(\frac{D^2 \|\mathcal{X}\|^2}{n_{\text{priv}} \epsilon}\right)^{1/3}$. The number of public samples simplifies as,

$$\begin{aligned}
n_{\text{pub}} &= O\left(\max\left(\frac{R^2 \log(2/\beta)}{\alpha^2}, \min\left\{m : \log^3(m) \mathfrak{R}_m^2(\mathcal{H}) \leq \alpha^2\right\}\right)\right) \\
&= \tilde{O}\left(D^2 \|\mathcal{X}\|^2 \max\left(\frac{\log(2/\beta)}{\alpha^2}, \frac{1}{\alpha^2}\right)\right) \\
&= \tilde{O}\left((D \|\mathcal{X}\|)^{2/3} (n_{\text{priv}} \epsilon)^{2/3} \log(2/\beta)\right).
\end{aligned}$$

□

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Claims made are given in Sections 3 and 4 proofs are provided in the Appendix

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The specific assumptions made are detailed in the preliminaries and theorem statements.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [Yes]

Justification: Statements are proved or cite a relevant reference. Many of these proofs can be found in the appendix.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

   Answer: [NA]

   Justification: There are no experimental results.

   Guidelines:

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

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

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

Answer: [NA]

Justification: There is no associated code.

Guidelines:

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

6. **Experimental Setting/Details**

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

   Answer: [NA]

   Justification: There are no experiments.

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [NA]

   Justification: There are no experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

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

8. **Experiments Compute Resources**

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

   Answer: [NA]

   Justification: There are no experiments.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: The theoretical nature of the results means there are minimal ethical concerns.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [NA]

    Justification: The theoretical nature of the work means that any societal impact would be very indirect.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: No such assets are used as a part of this research.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [NA]

    Justification: No such assets.

    Guidelines:

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

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

13. **New Assets**

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

    Answer: [NA]

    Justification: No such assets.

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: No human subjects were involved.

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification: No human subjects were involved.

    Guidelines:

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