Differentially Private Learning with Margin Guarantees

Raef Bassily
The Ohio State University
& Google Research NY
bassily.1@osu.edu

Mehryar Mohri
Google Research
& Courant Institute
mohri@google.com

Ananda Theertha Suresh
Google Research, NY
theertha@google.com

Abstract

We present a series of new differentially private (DP) algorithms with dimension-independent margin guarantees. For the family of linear hypotheses, we give a pure DP learning algorithm that benefits from relative deviation margin guarantees, as well as an efficient DP learning algorithm with margin guarantees. We also present a new efficient DP learning algorithm with margin guarantees for kernel-based hypotheses with shift-invariant kernels, such as Gaussian kernels, and point out how our results can be extended to other kernels using oblivious sketching techniques. We further give a pure DP learning algorithm for a family of feed-forward neural networks for which we prove margin guarantees that are independent of the input dimension. Additionally, we describe a general label DP learning algorithm, which benefits from relative deviation margin bounds and is applicable to a broad family of hypothesis sets, including that of neural networks. Finally, we show how our DP learning algorithms can be augmented in a general way to include model selection, to select the best confidence margin parameter.

1 Introduction

Preserving privacy is a crucial objective for machine learning algorithms. A widely adopted criterion in statistical data privacy is the notion of differential privacy (DP) [Dwork et al., 2006, Dwork, 2006, Dwork and Roth, 2014], which ensures that the information gained by an adversary is roughly invariant to the presence or absence of an individual in a dataset. Despite the remarkable theoretical and algorithmic progress in differential privacy over the last decade or more, however, its application to learning still faces several obstacles. A recent series of publications have shown that differentially private PAC learning of infinite hypothesis sets is not possible, even for common hypothesis sets such as that of linear functions. In fact, this is the case for any hypothesis set containing threshold functions [Bun et al., 2015, Alon et al., 2019]. These results imply serious limitations for private agnostic learnability.

Another rich body of literature has studied differentially private empirical risk minimization (DP-ERM) and differentially private stochastic convex optimization (DP-SCO) (e.g., [Chaudhuri et al., 2011, Jain and Thakurta, 2014, Bassily et al., 2014, 2019, Feldman et al., 2020, Song et al., 2021a, Bassily et al., 2021b, Asi et al., 2021, Bassily et al., 2021a]). When the underlying optimization problem is constrained (*constrained setting*), tight upper and lower bounds have been derived for the excess empirical risk of DP-ERM [Bassily et al., 2014] and for the excess population risk for DP-SCO [Bassily et al., 2019, Feldman et al., 2020]. These results show that learning guarantees necessarily admit a dependency on the dimension d of the form $\sqrt{d}/m$, where m is the sample size. This dependency is persistent, even in the special case of *generalized linear losses* (GLLs) [Bassily et al., 2014], which limits the benefit of such guarantees, since learning algorithms typically deal with high-dimensional spaces.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

When the underlying optimization problem is unconstrained (*unconstrained setting*) and the loss is a generalized linear loss, the bounds given by Jain and Thakurta [2014], Song et al. [2021a] and Bassily et al. [2021a] are dimension-independent but they admit a dependency on $\|w^*\|^2$, where w^* is the unconstrained minimizer of the expected loss (population risk), or $\|\widehat{w}\|^2$, where $\widehat{w}$ is the unconstrained minimizer of the empirical loss. Since the problem is unconstrained, the norm of these vectors can be very large, even for classification problems for which the minimizer of the zero-one loss admits a relatively small norm. Thus, in both the constrained and unconstrained settings, the learning guarantees derived from DP-ERM and DP-SCO are weak for hypothesis sets commonly used in machine learning.

The results just mentioned raise some fundamental questions about private learning: is differentially private learning with favorable (dimension-independent) guarantees possible for standard hypothesis sets? Must one resort to distribution-dependent bounds instead? In view of the negative PAC-learning results and other learning bounds mentioned earlier, we will seek instead optimistic margin-based learning bounds.

In the context of classification, learning bounds for linear hypotheses based on the dimension or, more generally, based on the VC-dimension of the hypothesis set are known to be too pessimistic since they deal with the worst case. Instead, margin bounds have been shown to be the most informative and useful guarantees [Koltchinskii and Panchenko, 2002, Schapire et al., 1997]. This motivates our study of differentially private learning algorithms with margin-based guarantees. Note that our *confidence-margin* analysis and guarantees do not require the hard-margin separability assumptions adopted in [Blum et al., 2005, Le Nguyen et al., 2020], which is a strong assumption that typically does not hold in practice. Another existing study that deals with somewhat related questions is that of Chaudhuri et al. [2014]. But, the paper deals with a specific class of maximization problems and adopts a non-standard definition of margin. Another related line of work is that of Rubinstein et al. [2009] and Chaudhuri et al. [2011] on DP Kernel classifiers. These works either provide suboptimal, dimension-dependent learning guarantees or make strong assumptions about the Fourier coefficients of the kernel predictors. We discuss these prior works in more detail in Section 1.1.

Main contributions. We present a series of new differentially private (DP) algorithms for learning linear classifiers, kernel classifiers, and neural-network classifiers with dimension-independent, confidence-margin guarantees. In Section 3, we study the family of linear hypotheses. We first give a pure DP learning algorithm with relative deviation margin guarantees. Next, we present an efficient DP learning algorithm with margin guarantees. Our algorithm is based on a faster construction for the JL-transform and a faster DP-ERM algorithm. While the general structure of our algorithms for linear classifiers is similar to that of Le Nguyen et al. [2020], our results require a new analysis that takes into account the scale-sensitive nature of the margin loss and the ρ-hinge loss. In Section 4, We present a new efficient DP learning algorithm with margin guarantees for kernel-based hypothesis sets, assuming that the positive definite kernel used is shift-invariant, as with the most commonly used Gaussian kernels. Our algorithm combines kernel approximation with the use of the JL-transform. Our result is based on a new style of analysis that uses regularized ERM as a reference. These ideas enable us to attain a bound that nearly matches the non-private margin bound, without resorting to the strong assumptions in prior work. Our confidence-margin bounds for DP learning of linear and kernel classifiers nearly match the standard, non-private confidence-margin bounds. In Section 5, we initiate the study of DP learning of neural networks with margin guarantees. We design a pure DP learning algorithm for a family of feed-forward neural networks for which we prove a confidence-margin bound that is independent of the input dimension and exhibits better dependence on the network parameters than the bounds attained via uniform convergence. Our result entails a new analysis of embedding-based "network compression" technique. Our margin bound for neural networks is the first of its type. The bound is independent of the input dimension and scales only linearly with the number of activation units. In Appendix E, We further present a *label privacy* learning algorithm, which we show benefits from relative deviation margin bounds. The algorithm and its guarantee are applicable to a broad family of hypothesis sets, including that of neural networks. Finally, we show in Appendix F how our DP learning algorithms can be augmented in a general way to include model selection, to select the best confidence margin parameter.

1.1 Related work

Prior work on unconstrained GLLs. Jain and Thakurta [2014] and Song et al. [2021a] showed that it is possible to derive dimension-independent risk bounds for DP-ERM and DP-SCO in the context of linear prediction, when the parameter space is unconstrained and the loss function is convex and

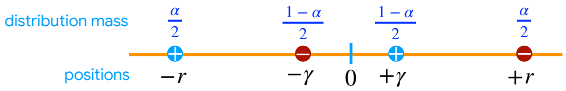


Figure 1: Simple example in dimension one for which the minimizer of the expected hinge loss $\mathbb{E}[\ell^{\text{hinge}}(w)]$ is $w^* = \frac{1}{\gamma}$ and thus $\|w^*\| = \frac{1}{\gamma} \gg 1$ for $\gamma \ll 1$. Here, any other $w > 0$, in particular with a small norm, achieves the same zero-one loss as w^*.

Lipschitz (GLL). However, their bounds scale with $\|w^*\|$, the norm of the optimal unconstrained minimizer of the expected surrogate loss such as the hinge loss. Also, using their techniques for unconstrained DP-ERM for GLLs together with uniform convergence would yield generalization error bounds that scale with the norm of the unconstrained empirical risk minimizer $\widehat{w}$. The first issue with this line of work is that the norms of such unconstrained solutions can be very large, thereby resulting in uninformative bounds. In fact, one can construct simple, low-dimensional examples, where $\|w^*\| = \Omega(m)$ while there is a predictor w with $\|w\| = O(1)$ that attains the same expected zero-one error, see Figure 1 (a detailed analysis of that example is given in Appendix H). More importantly, the paradigm adopted in this line of work is to first devise an algorithm and next derive bounds for its excess risk. In contrast, we start from strong generalization error bounds, which we use to guide the design of our algorithm.

Prior work on DP learning of hard-margin halfspaces. Blum et al. [2005] and Le Nguyen et al. [2020] studied DP learning of linear classifiers in the separable setting, that is with a hard- or *geometric margin*. Blum et al. [2005] gave a construction based on a private version of the Perceptron algorithm, which results in a dimension-dependent bound on the expected error. This result was later improved by Le Nguyen et al. [2020] who gave new constructions with dimension-independent guarantees based their nice idea of using embeddings, namely, the Johnson-Lindenstrauss (JL) transform, to reduce the dimensionality of the problem from d to $1/\gamma$, where γ is the geometric margin. Note that the hard-margin separability is a strong assumption that typically does not hold in practice. Moreover, the constructions proposed in [Le Nguyen et al., 2020] require the knowledge of the margin for their guarantees to be valid. In contrast, our work considers the more general notion of *confidence margin*, which does not require the existence of a geometric margin and applies to realistic scenarios with non-separable data. Moreover, the confidence-margin parameter, ρ, in our algorithms is tunable and can be optimized. Importantly, our algorithms still yield meaningful learning guarantees even if this parameter is not optimized. Our algorithms for linear classifiers also make use of an embedding as a pre-processing step. However despite the similar structure to that of Le Nguyen et al. [2020], our algorithm requires a new analysis and different settings of parameters. This new analysis is necessary to deal with the scale-sensitive nature of our bounds, due to the absence of a hard-margin.

Prior work on DP Kernel classifiers. Rubinstein et al. [2009] were the first to provide differentially private constructions for SVMs in both the finite-dimensional feature space and kernel settings. However, their constructions are suboptimal and the resulting bounds suffer from a polynomial dependence on the dimension of the feature space. In addition, their error bound has a sub-optimal dependence on the sample size m and also has an explicit dependence on the ℓ_1-norm of the dual variables of the SVM. In general, this norm can be as large as $\sqrt{m}$, in which case their error bound becomes vacuous. Chaudhuri et al. [2011] gave a similar construction for shift-invariant kernels. However, their error guarantees are based on the kernel approximation results of Rahimi and Recht [2008] and hence entail a relatively strong condition on the Fourier coefficients of the kernel predictors and the kernel density. We note that the standard assumption of bounded Reproducing Kernel Hilbert Space (RKHS) norm does not imply such a condition. Jain and Thakurta [2013] gave algorithms for DP predictions with kernels, where the goal is to privately generate predictions (labels) on a small test set that admits no privacy constraints. In such scenarios, their algorithms do not output a classifier. Jain and Thakurta [2013] also gave a construction for private learner that outputs a kernel classifier, however, their construction is computationally inefficient and the resulting error guarantees are dimension-dependent.

2 Preliminaries

We consider an input space $\mathcal{X}$, a binary output space $\mathcal{Y} = \{-1, +1\}$ and a hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X}$ to $\mathbb{R}$. We denote by $\mathcal{D}$ a distribution over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and denote by

$R_{\mathcal{D}}(h)$ the generalization error and by $\widehat{R}_S(h)$ the empirical error of a hypothesis $h \in \mathcal{H}$:

$$R_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{z=(x,y)\sim\mathcal{D}}[1_{yh(x)\leq 0}] \qquad \widehat{R}_S(h) = \mathop{\mathbb{E}}_{z=(x,y)\sim S}[1_{yh(x)\leq 0}],$$

where we write $z \sim S$ to indicate that z is randomly drawn from the empirical distribution defined by S. Given $\rho \geq 0$, we similarly define the ρ-margin loss and empirical ρ-margin loss of $h \in \mathcal{H}$:

$$R^{\rho}_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{z=(x,y)\sim\mathcal{D}}[1_{yh(x)\leq \rho}] \qquad \widehat{R}^{\rho}_S(h) = \mathop{\mathbb{E}}_{z=(x,y)\sim S}[1_{yh(x)\leq \rho}].$$

The ρ-margin loss is not convex. Hence, we also consider the ρ-hinge loss to provide computationally-efficient algorithms. For any $\rho > 0$, define the ρ-hinge loss as $\ell^{\rho}(u) \triangleq \max(1 - u/\rho, 0)$, $u \in \mathbb{R}$. Similar to the above definitions, given $\rho > 0$, for a sample S, we define the expected and the empirical ρ-hinge losses as follows:

$$L^{\rho}_{\mathcal{D}}(w) = \mathop{\mathbb{E}}_{z=(x,y)\sim\mathcal{D}}[\ell^{\rho}(y_i \langle w, x_i\rangle)] \qquad \widehat{L}^{\rho}_S(w) = \mathop{\mathbb{E}}_{z=(x,y)\sim S}[\ell^{\rho}(y_i \langle w, x_i\rangle)].$$

In the context of learning, differential privacy is defined as follows.

Definition 2.1 (Differential privacy)**.** *Let $\varepsilon, \delta \geq 0$. Let $\mathcal{A}: (\mathcal{X}\times\mathcal{Y})^m \to \mathcal{H}$ be a randomized algorithm. We say that $\mathcal{A}$ is (ε,δ)-*DP *if for any measurable subset $O \subset \mathcal{H}$ and all $S, S' \in (\mathcal{X}\times\mathcal{Y})^m$ that differ in one sample, the following inequality holds:*

$$\mathbb{P}(\mathcal{A}(S) \in O) \leq e^{\varepsilon}\,\mathbb{P}(\mathcal{A}(S') \in O) + \delta. \qquad (1)$$

If $\delta = 0$, we refer to this guarantee as pure differential privacy.

3 Private Algorithms for Linear Classification with Margin Guarantees

In this section, we present two private learning algorithms for linear classification with margin guarantees: first, a computationally inefficient pure DP algorithm, which we show benefits from relative deviation margin bounds, next, a computationally efficient DP algorithm with a dimension-independent bound expressed in terms of the empirical ρ-hinge loss.

Let $\mathbb{B}^d(r) \triangleq \{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$ denote the Euclidean ball in $\mathbb{R}^d$ of radius r and let $\mathcal{X} \subseteq \mathbb{B}^d(r)$ denote the feature space. We will use the shorthand $\mathbb{B}^d$ for $\mathbb{B}^d(1)$. We consider the class of linear predictors over $\mathcal{X}$ defined by $\mathcal{H}_{\mathsf{Lin}} = \{h_w : x \mapsto \langle w, x\rangle \mid w \in \mathbb{B}^d(\Lambda)\}$. Note that one can represent the general class of affine functions over $\mathbb{R}^d$ as linear functions over $\mathbb{R}^{d+1}$ by simply mapping each $x \in \mathbb{R}^d$ to $\tilde{x} = (x, 1) \in \mathbb{R}^{d+1}$. Thus, without loss of generality, we will consider $\mathcal{H}_{\mathsf{Lin}}$. Here, we view d as possibly much larger than the sample size m. We also note that even though the predictors in the input class $\mathcal{H}_{\mathsf{Lin}}$ admit Λ-bounded norm, we do not constrain the algorithm to output a predictor with bounded norm, which circumvents the necessary dependence on the dimension in constrained DP optimization [Bassily et al., 2014].

3.1 Pure DP Algorithm for Linear Classification

A standard method for designing differentially private algorithms for a continuous hypothesis class is to apply the exponential mechanism [McSherry and Talwar, 2007] to a cover of the hypothesis class. Since $\mathcal{H}_{\mathsf{Lin}}$ is d-dimensional, the size of a useful cover is about $2^{\Omega(d)}$, thus, a direct application of the exponential mechanism yields an $\Omega(d)$-bound; we give a simple example illustrating that in Appendix G. Thus, instead, we seek to reduce the size of the cover without impacting its accuracy, using random projections. This results in a mapping Φ from $\mathbb{R}^d$ to a lower-dimensional space $\mathbb{R}^k$.

For linear classification, we wish to preserve intra-point distances and angles, that is $x \cdot x' \approx \Phi x \cdot \Phi x'$ for points x and x'. It is known that this property can be fulfilled as a corollary of the Johnson-Lindenstrauss lemma [Nelson, 2011, Theorem 109]. For completeness, we provide a full proof in Appendix A. More interestingly, we show that we can reduce the dimension to $\widetilde{O}(\Lambda^2 r^2/\rho^2)$, without the error decreasing significantly. We then run the exponential mechanism in this lower-dimensional space and next compute a classifier $\tilde{w}$ in that space. We finally derive a classifier in the original space by applying the transpose of the original projection matrix $\Phi^T \tilde{w}$. Note that the final output $\Phi^T \tilde{w}$ has expected norm $\widetilde{O}\Big(\frac{\sqrt{d}\rho}{r}\Big)$ and may not lie in $\mathbb{B}^d(\Lambda)$.

Algorithm 1 $\mathcal{A}_{\mathsf{PrivMrg}}$: Private Learner of Linear Classifiers with Margin Guarantees

```
Require: Dataset $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \{\pm 1\})^m$; privacy parameter $\varepsilon > 0$; margin parameter $\rho \in (0, \Lambda r]$; confidence parameter $\beta > 0$.
Let $k = O\left(\frac{\Lambda^2 r^2 \log\left(\frac{m}{\beta}\right)}{\rho^2}\right)$.
Let $\Phi$ be a random $k \times d$ matrix with entries drawn i.i.d. uniformly from $\{\pm \frac{1}{\sqrt{k}}\}$.
Let $S_\Phi = \{(x_\Phi, y) \colon (x, y) \in S\}$, where for any $x \in \mathbb{R}^d$, $x_\Phi \triangleq \Phi x \in \mathbb{R}^k$.
Let $\mathcal{C}$ be a $\frac{\rho}{10\, r}$-cover of $\mathbb{B}^k(2\Lambda)$.
Run the Exponential mechanism over $\mathcal{C}$ with privacy parameter $\varepsilon$, sensitivity $1/m$, and score function $-\widehat{R}^{(k)}_{S_\Phi}(w)$ for $w \in \mathcal{C}$, to select $\widetilde{w} \in \mathcal{C}$.
return $w^{\mathsf{Priv}} = \Phi^\top \widetilde{w}$, where $\Phi^\top$ denotes the transposition of $\Phi$.
```

Algorithm 1 gives the pseudocode of the full algorithm. The algorithm and the analysis in this section include a dimensionality reduction technique for mapping the feature vectors from the input d-dimensional space to a k-dimensional space, where $k = \widetilde{O}(\Lambda^2 r^2/\rho^2)$ for some $\rho \in (0, \Lambda r]$. Hence, we will be dealing with "compressed" parameter vectors in $\mathbb{R}^k$. To distinguish these two spaces, we will denote the empirical error and the empirical ρ-margin error in this k-dimensional space as $\widehat{R}^{(k)}_{S'}(w')$ and $\widehat{R}^{(k),\rho}_{S'}(w')$, respectively, where $w' \in \mathbb{B}^k$ and $S' \in \left(\mathbb{R}^k \times \mathcal{Y}\right)^m$.

Theorem 3.1. *Algorithm 1 is ε-differentially private. For any $\beta \in (0,1)$, with probability at least $1 - \beta$ over the draw of a sample S of size m from $\mathcal{D}$, the solution w^{Priv} it returns satisfies:*

$$R_{\mathcal{D}}(w^{\mathsf{Priv}}) \leq \min_{w \in \mathbb{B}^d(\Lambda)} \left\{ \widehat{R}^\rho_S(w) + \widetilde{O}\left(\sqrt{\widehat{R}^\rho_S(w) \frac{\Lambda^2 r^2}{m\rho^2}} + \frac{\Lambda^2 r^2}{\rho^2 \min(1, \varepsilon) m} \right) \right\}.$$

The proof is given in Appendix B.1. The theorem provides a *margin guarantee* for the private solution. Note that no assumption is made about separability or the existence of a favorable hard-margin. Instead, through the first term, the bound is based on the distribution of the empirical margins $y(w \cdot x)$. The theorem suggests that, when the empirical ρ-margin loss remains small for a relatively large value of the confidence margin parameter ρ, then w^{Priv} benefits from a strong generalization guarantee. These comments hold similarly for other margin bounds presented in this paper.

This result, although given for a computationally inefficient method, is stronger than several previously known ones: First, it is an $(\varepsilon, 0)$-pure differential privacy guarantee; second, it is dimension-independent and furthermore, unlike prior work, the norm of the optimal hypothesis does not appear in the bound. Furthermore, since it is a relative deviation margin bound, it smoothly interpolates between the realizable case of $\widehat{R}^\rho_S(w) = 0$ and the case of $\widehat{R}^\rho_S(w) > 0$. For a sample of size m, the bound is based on an interpolation between a $1/\sqrt{m}$-term that includes the square-root of the empirical margin loss as a factor and another term in $1/m$. In particular, when the empirical margin loss is zero, the bound only admits the $1/m$ fast rate term. As a corollary, note that, up to constants, one can always obtain privacy for $\varepsilon > 1$ essentially for free.

Dependence on Λ/ρ: Our bound depends on the choice of Λ/ρ. We note that this is fundamentally different from the dependence on $\|w^*\|$ in prior works on GLLs [Jain and Thakurta, 2014, Song et al., 2021a] for two reasons: First, unlike $\|w^*\|$, Λ/ρ is a measurable and, more importantly, tunable quantity, which we can select an optimal setting for[1] (as we do in Appendix F). Second, the optimal choice for this parameter can be much smaller than $\|w^*\|$ as demonstrated in Appendix H.

3.2 Efficient Private Algorithm for Linear Classification

The ρ-margin loss is not convex and it is known that its minimization is generally intractable. Instead, we devise a computationally efficient algorithm, whose guarantees are expressed in terms of the empirical ρ-hinge loss. Algorithm 2 shows the pseudocode of our algorithm. We now discuss the key steps of the algorithm.

[1] Note that, without loss of generality, we can set $\Lambda = 1$ and optimize only with respect to ρ.

Algorithm 2 $\mathcal{A}_{\mathsf{EffPrivMrg}}$: Efficient Private Learner of Linear Classifiers with Margin Guarantees

```
Require: Dataset $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in \left(\mathbb{B}^d(R) \times \{\pm 1\}\right)^m$; privacy parameters $\varepsilon, \delta$; norm bound $\Lambda$; margin parameter $\rho \in (0, \Lambda r]$; confidence parameter $\beta > 0$.
1: Let $k = \frac{\varepsilon m \log(m/\beta)}{\log^{\frac{3}{2}}(1/\delta)\log(1/\beta)}$.
2: Let $\Phi$ be a random $k \times d$ matrix from the construction in Lemma A.4.
3: Let $S_\Phi = \left\{\left(\Pi_{\mathbb{B}^k(2r)}(x_\Phi), y\right) : (x, y) \in S\right\}$, where for any $x \in \mathbb{R}^d$, $x_\Phi \triangleq \Phi x \in \mathbb{R}^k$ and $\Pi_{\mathbb{B}^k(2r)}$ is the Euclidean projection on $\mathbb{B}^k(2r)$.
4: Privately solve the convex ERM problem: $\underset{w \in \mathbb{B}^k(2\Lambda)}{\operatorname{argmin}} \widehat{L}^\rho_{S_\Phi}(w)$ via Algorithm 4 (Appendix B.2) and return $\tilde{w} \in \mathbb{B}^k(2\Lambda)$.
5: return $w^{\mathsf{Priv}} = \Phi^\top \tilde{w}$, where $\Phi^\top$ denotes the transposition of $\Phi$.
```

Fast JL-transform. Our algorithm entails a dimensionality reduction step (step 3) as in Algorithm 1. Here, we note that the new dimension k is chosen to be $\widetilde{O}(m\varepsilon)$, which enables us to control the influence of the dimensionality reduction on the empirical hinge loss. We also note that this step is carried out via a fast construction for the JL-transform (Lemma A.4), which takes $O(d\log(d))$ time, assuming $d > \varepsilon m$.

Near linear-time DP convex ERM. After this step, we invoke an efficient algorithm for DP-ERM (step 4 in Algorithm 2) to find an approximate minimizer of the empirical ρ-hinge loss $\widehat{L}^\rho_{S_\Phi}(w)$, rather than using the exponential mechanism to find an approximate minimizer for the empirical zero-one loss $\widehat{R}_{S_\Phi}(w)$. To improve the running time of step 4, we use the construction in [Bassily et al., 2021a, Algorithm 2] to solve DP-ERM in near-linear time and with high-probability guarantee on the excess empirical risk (see Algorithm 4 in Appendix B.2). The algorithm of Bassily et al. [2021a] is devised for DP-SCO with respect to non-smooth generalized linear losses. It is based on a combination of a smoothing technique via proximal steps and the phased SGD algorithm [Feldman et al., 2020, Algorithm 2] for smooth DP-SCO. The algorithm of Bassily et al. [2021a] can be used for DP-ERM if it is fed with a sample from the empirical distribution of the dataset. However, the privacy guarantee requires a careful privacy analysis to deal with the fact that this sample may contain duplicate entries from the original dataset.

Moreover, since the original algorithms in [Feldman et al., 2020, Bassily et al., 2021a] provide only expectation guarantees and we aim at high-probability learning bounds, we use a standard private confidence-boosting technique to provide a high-probability guarantee on the excess risk of our variant. We summarize the guarantees of this variant in the following lemma. The details of the construction and the proof of the lemma below can be found in Appendix B.2.

Lemma 3.1. *Let $m \in \mathbb{N}$, $0 < \delta < \frac{1}{m}$, and $0 < \varepsilon \le \log(1/\delta)$. Algorithm 4 (Appendix B.2) is (ε, δ)-DP. Let $\beta \in (0, 1)$. Let $k \in \mathbb{N}$, and $\tilde{r}, \Lambda > 0$. Let $\widetilde{S} \in \left(\mathbb{B}^k(\tilde{r}) \times \{\pm 1\}\right)^m$ be the input dataset and $\mathbb{B}^k(\Lambda)$ be the parameter space. With probability $1 - \beta$ over the randomness in Algorithm 4, the output $\tilde{w}$ satisfies*

$$\widehat{L}^\rho_{\widetilde{S}}(\tilde{w}) \le \min_{w \in \mathbb{B}^k(\Lambda)} \widehat{L}^\rho_{\widetilde{S}}(w) + \frac{\Lambda \tilde{r}}{\rho} \cdot O\left(\frac{1}{\sqrt{m}} + \frac{\sqrt{k}\log^{\frac{3}{2}}(\frac{1}{\delta})\log(\frac{1}{\beta})}{\varepsilon m}\right).$$

Moreover, Algorithm 2 requires $O(m\log(m)\log(1/\beta))$ gradient computations.

We now state our main result in this section, which we prove in Appendix B.3.

Theorem 3.2. *Let $0 < \delta < \frac{1}{m}$ and $0 < \varepsilon \le \log(1/\delta)$. Algorithm 2 is (ε, δ)-DP. Let $\beta \in (0, 1)$. Let $S \sim \mathcal{D}^m$ for a distribution $\mathcal{D}$ over $\mathbb{B}^d(r) \times \{\pm 1\}$. Algorithm 2 outputs $w^{\mathsf{Priv}} \in \mathbb{R}^d$ such that with probability at least $1 - \beta$, we have*

$$R_{\mathcal{D}}(w^{\mathsf{Priv}}) \le \min_{w \in \mathbb{B}^d(\Lambda)} \widehat{L}^\rho_S(w) + \widetilde{O}\left(\frac{\Lambda r}{\rho\sqrt{\min(1, \varepsilon)\, m}}\right).$$

Moreover, Algorithm 2 runs in time $O\Big(md\log(\max(d, m)) + \varepsilon m^2\log(m)/\log^{\frac{3}{2}}(1/\delta)\Big)$.

4 Private Algorithms for Kernel-Based Classifiers with Margin Guarantees

In this section, we present private algorithms with margin guarantees for kernel-based predictors [Schölkopf and Smola, 2002, Shawe-Taylor and Cristianini, 2004]. We first consider a continuous, positive definite, shift-invariant kernel $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, where $K(x, x) = r^2$ for all $x \in \mathcal{X}$. The associated feature map is defined as $\psi(x) \triangleq K(\cdot, x),\ x \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{B}^d(r)$.

Overview of the technique. Our approach is based on approximating the feature map ψ by a finite-dimensional feature map $\widehat{\psi}: \mathcal{X} \to \mathbb{B}^{2D}(r)$ determined via Random Fourier Features (RFFs). The dimension $2D$ of the approximate feature map is chosen to be sufficiently large to ensure that for all pairs of feature vectors x_i, x_j in a training set $S = ((x_1, y_1), \ldots, (x_m, y_m))$, we have $|\langle \widehat{\psi}(x_i), \widehat{\psi}(x_j) \rangle - K(x_i, x_j)| \lesssim \frac{1}{m}$ with high probability over the randomness of $\widehat{\psi}$ (due to RFFs). This suffices to derive an upper bound (margin bound) on the true error of a finite-dimensional linear predictor trained on the sample made of the labeled points $\big(\widehat{\psi}(x), y\big)$, $(x, y) \in S$, that is essentially the same as the margin bound known for the kernel classifier. Hence, in effect, we reduce the problem to that of learning a linear classifier in a $2D$-dimensional space, which we can solve privately using Algorithm 2. Note that the output predictor in this case is a finite-dimensional linear function rather than a function in the Reproducing Kernel Hilbert Space. A full description of our DP learner of kernel classifiers is given in Algorithm 3 below.

Bochner's Theorem and RFFs. Since the kernel K is shift-invariant, it can be expressed as $K(x, x') = r^2 \bar{K}(x - x'), x, x' \in \mathcal{X}$ for some function $\bar{K}: \mathcal{Z} \to \mathbb{R}$, where $\mathcal{Z} = \{z = x - x': x, x' \in \mathcal{X}\}$. Moreover, since K is positive-definite, $\bar{K}$ is the Fourier transform of a probability distribution $P_{\bar{K}}$:

$$\bar{K}(x) = \int_{\mathcal{X}} e^{i\langle \omega, x \rangle} P_{\bar{K}}(\omega) d\omega.$$

This follows from Bochner's Theorem [Rudin, 2017]. Random Fourier Features (RFFs) provide a simple method introduced in [Rahimi and Recht, 2007] to approximate kernel feature maps in a data-independent fashion. The idea is based on Bochner's theorem. In particular, we first sample $\omega_1, \ldots, \omega_D$ independently from the probability distribution $P_{\bar{K}}$. Then, we define an approximate feature map as follows:

$$\widehat{\psi}(x) \triangleq \frac{r}{\sqrt{D}} \left(\cos\langle \omega_1, x \rangle, \sin\langle \omega_1, x \rangle, \ldots \cos\langle \omega_D, x \rangle, \sin\langle \omega_D, x \rangle \right), \quad \forall x \in \mathcal{X}. \tag{2}$$

For D sufficiently large, it can be shown that $\langle \widehat{\psi}(x), \widehat{\psi}(x') \rangle$ concentrates around $K(x, x')$ for all pairs $x, x' \in \mathcal{X}$ [Mohri et al., 2018, Theorem 6.28]. In our analysis below, we only need that concentration to hold uniformly over pairs x, x' from a fixed training set rather than uniformly over all pairs $x, x' \in \mathcal{X}$. This leads to a simpler approximation guarantee, which we formally state below.

Theorem 4.1. *Let $S_{\mathcal{X}} = (x_1, \ldots, x_m) \in \mathcal{X}^m$. Let K be a shift-invariant, positive definite kernel, where $K(x, x) = r^2,\ \forall x \in \mathcal{X}$. Let $P_{\bar{K}}$ be the probability distribution associated with K. Suppose $\omega_1, \ldots, \omega_D$ are drawn independently from $P_{\bar{K}}$. With probability 1, we have $\|\widehat{\psi}(x)\|_2 = r,\ \forall x \in \mathcal{X}$. For any $\beta \in (0, 1)$, with probability at least $1 - \beta$, for all $i, j \in [m]$ such that $i \neq j$ we have*

$$\left| \langle \widehat{\psi}(x_i), \widehat{\psi}(x_j) \rangle - K(x_i, x_j) \right| \leq 2r^2 \sqrt{\frac{\log\left(\frac{m}{\beta}\right)}{D}}.$$

Proof. The first assertion related to $\|\widehat{\psi}(x)\|_2\ \forall x \in \mathcal{X}$ follows directly from the definition of $\widehat{\psi}(x)$ and a basic trigonometric identity. The proof of the second assertion about the inner products follows from the identity $\mathbb{E}_{\omega_1, \ldots, \omega_D}\left[\langle \widehat{\psi}(x_i), \widehat{\psi}(x_j) \rangle \right] = K(x_i, x_j)$ that holds for all x_i, x_j, the application of Hoeffding's bound combined with the union bound over all $\approx m^2$ pairs $x_i, x_j \in S_{\mathcal{X}}$. The unbiasedness of $\langle \widehat{\psi}(x_i), \widehat{\psi}(x_j) \rangle$ follows from the fact that the expectation is the Fourier transform of $r^2 P_{\bar{K}}(\omega)$, which, by Bochner's Theorem, is $r^2 \bar{K}(x_i - x_j) = K(x_i, x_j)$. □

We now state our main result, which we prove in Appendix C.

Algorithm 3 $\mathcal{A}_{\mathsf{PrivKerMrg}}$: Efficient Private Learner of Kernel Classifiers with Margin Guarantees

```
Require: Dataset $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \{\pm 1\})^m$; shift-invariant, positive definite
    kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with $K(x, x) = r^2$ for all $x \in \mathcal{X}$, privacy parameters $\varepsilon, \delta$; Reproducing kernel Hilbert space (RKHS) norm bound $\Lambda$; margin parameter $\rho \in (0, 2\Lambda r]$; confidence
    parameter $\beta > 0$.
  Let $P_{\bar{K}}$ be the probability distribution associated with $K$.
  Let $D = m^2 \log(2m/\beta)$.
  Draw $\omega_1, \ldots, \omega_D$ independently from $P_{\bar{K}}$.
  Let $S_{\widehat{\psi}} = \big((\widehat{\psi}(x_i), y_i) : i \in [m]\big)$, where $\widehat{\psi}$ is as defined in (2).
  $w^{\mathsf{Priv}} \leftarrow \mathcal{A}_{\mathsf{EffPrivMrg}}\big(S_{\widehat{\psi}},\ \varepsilon, \delta, 2\Lambda,\ \rho, \beta/2\big)$, where $\mathcal{A}_{\mathsf{EffPrivMrg}}$ is the private learner described in
    Algorithm 2. Note that the input dimension to $\mathcal{A}_{\mathsf{EffPrivMrg}}$ is $2D$ rather than $d$, and the norm
    bound parameter is $2\Lambda$.
  return Private predictor $h_{w^{\mathsf{Priv}}} : \mathcal{X} \to \mathbb{R}$ defined as $\forall x \in \mathcal{X},\ h^{\widehat{\psi}}_{w^{\mathsf{Priv}}}(x) \triangleq \big\langle w^{\mathsf{Priv}}, \widehat{\psi}(x) \big\rangle$.
```

Theorem 4.2. *Let $r > 0$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a shift-invariant, positive definite kernel, where $K(x, x) = r^2$ for all $x \in \mathcal{X}$. For any $\varepsilon > 0$ and $\delta \in (0, 1)$, Algorithm 3 is (ε, δ)-differentially private. Define $\mathcal{H}_\Lambda \triangleq \{h \in \mathbb{H} : \|w\|_{\mathbb{H}} \leq \Lambda\}$, where $\|\cdot\|_{\mathbb{H}}$ is the norm corresponding to the reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ associated with the kernel K. Let $\beta \in (0, 1)$. Given an input sample S of m examples drawn i.i.d. from a distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$, Algorithm 3 outputs $h^{\widehat{\psi}}_{w^{\mathsf{Priv}}}$ such that with probability at least $1 - \beta$, we have*

$$R_{\mathcal{D}}(h^{\widehat{\psi}}_{w^{\mathsf{Priv}}}) \leq \min_{h \in \mathcal{H}_\Lambda} \widehat{L}^{\rho}_{S}(h) + \widetilde{O}\left(\frac{\Lambda r}{\rho \sqrt{\min(1, \varepsilon)\, m}}\right),$$

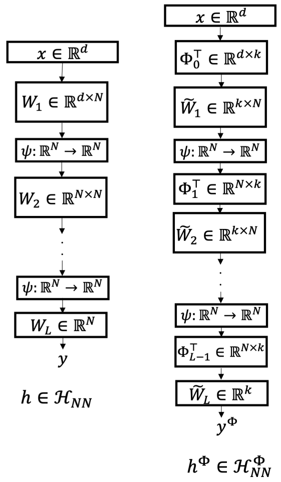
where, for any $h \in \mathcal{H}_\Lambda$, $\widehat{L}^{\rho}_{S}(h) \triangleq \frac{1}{m} \sum_{i=1}^{m} \ell^{\rho}\left(y_i \langle h, \psi(x_i)\rangle_{\mathbb{H}}\right)$, where ψ is the feature map associated with the kernel K and $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ is the inner product associated with the RKHS $\mathbb{H}$.

Polynomial kernels: Our results can be extended to polynomial kernels using a different approach to construct a finite-dimensional approximation of the kernel. A polynomial kernel of degree p, denoted as κ_p, can be expressed as $\kappa_p(x, x') = (\langle x, x' \rangle + c)^p$, $x, x' \in \mathcal{X}$ and $c > 0$ is some constant. Note that a feature map ψ_p associated with such a kernel can be expressed as a vector in $\mathbb{R}^{\bar{d}}$, where $\bar{d} = O(d^p)$. In particular, $\psi_p(x)$ is the vector of all monomials of a p-th degree polynomial. Ignoring computational efficiency considerations (or when p is a small constant), there is a simpler private construction than the one used for shift-invariant kernels. In that case, we can directly use the JL-transform to embed $\{\psi_p(x_1), \ldots, \psi_p(x_m)\}$ into a k-dimensional subspace exactly as in Section 3.2, which would result in a k-dimensional approximation of the kernel (by the properties of the JL-transform). Hence, we can directly use Algorithm 2 on the dataset $S_{\psi_p} \triangleq ((\psi_p(x_1), y_1), \ldots, \psi_p(x_m), y_m))$. We therefore obtain the same bound on the expected error as above except that r would then be r^p. That dependence on r^p is inherent in this case even non-privately since $\kappa_p(x, x)$ can be as large as r^p. However, as discussed below, more efficient solutions can be designed for approximating these and many other kernels.

Further extensions. Our work can directly benefit from the method of Le et al. [2013], which is computationally faster than that of Rahimi and Recht [2007], $O((m + d) \log d)$, instead of $O(md)$. Their technique also extends to any kernel that is a function of an inner product in the input space. We can further use, instead of the JL-transform, the *oblivious sketching* technique of Ahle et al. [2020] from numerical linear algebra, which builds on previous work by Pham and Pagh [2013] and Avron et al. [2014], to design sketches for polynomial kernel with a target dimension that is only polynomially dependent on the degree of the kernel function, as well as a sketch for the Gaussian kernel on bounded datasets that does not suffer from an exponential dependence on the dimensionality of input data points. More recently, Song et al. [2021b] presented new oblivious sketches that further considerably improved upon the running-time complexity of these techniques. Their method also applies to other *slow-growing* kernels such as the neural tangent (NTK) and arc-cosine kernels.

5 Private Algorithms for Learning Neural Networks with Margin Guarantees

In this section, we describe a private learning algorithm that benefits from favorable margin guarantees when run with a family of neural networks with a very large input dimension.



$x \in \mathbb{R}^d$
$W_1 \in \mathbb{R}^{d\times N}$
$\psi: \mathbb{R}^N \to \mathbb{R}^N$
$W_2 \in \mathbb{R}^{N\times N}$
$\psi: \mathbb{R}^N \to \mathbb{R}^N$
$W_L \in \mathbb{R}^N$
y
$h \in \mathcal{H}_{NN}$
$x \in \mathbb{R}^d$
$\Phi_0^\top \in \mathbb{R}^{d\times k}$
$\widetilde{W}_1 \in \mathbb{R}^{k\times N}$
$\psi: \mathbb{R}^N \to \mathbb{R}^N$
$\Phi_1^\top \in \mathbb{R}^{N\times k}$
$\widetilde{W}_2 \in \mathbb{R}^{k\times N}$
$\psi: \mathbb{R}^N \to \mathbb{R}^N$
$\Phi_{L-1}^\top \in \mathbb{R}^{N\times k}$
$\widetilde{W}_L \in \mathbb{R}^k$
y^Φ
$h^\Phi \in \mathcal{H}_{NN}^\Phi$

Figure 2: Illustration of the neural networks before and after JL-transforms.

We consider a family $\mathcal{H}_{\mathsf{NN}}$ of L-layer feed-forward neural networks defined over $\mathbb{B}^d(r)$, with d potentially very large compared to the sample size m. A function h in $\mathcal{H}_{\mathsf{NN}}$ can be viewed as a cascade of linear maps composed with a non-linear activation function (see Figure 2, left column). Here, $W_1, \ldots, W_L$ are the weight matrices defining the network and ψ is a non-linear activation. For simplicity, the width (number of neurons) in each hidden layer, denoted by N, is assumed to be the same for all the layers. Also, we assume that the output of the network is a real scalar and hence we have $W_L \in \mathbb{R}^N$. Furthermore, we assume no activation in the output layer. We also assume the same activation $\psi: \mathbb{R}^N \to \mathbb{R}^N$ for all layers and choose it to be a sigmoid function: for any $u = (u_1, \ldots, u_N) \in \mathbb{R}^N$, $\psi(u) = (\sigma_\eta(u_1), \ldots, \sigma_\eta(u_N))$, for some $\eta > 0$, where $\sigma_\eta(a) = \frac{1-e^{-\frac{\eta a}{2}}}{1+e^{-\frac{\eta a}{2}}}$, $a \in \mathbb{R}$. Note that σ_η is η-Lipschitz and thus ψ is η-Lipschitz with respect to $\|\cdot\|_2$: for any $u, v \in \mathbb{R}^N$, $\|\psi(u) - \psi(v)\|_2 \leq \eta \|u - v\|_2$. A typical choice for η in practice is $\eta = 1$, but we will keep the dependence on η for generality. We define $\mathcal{H}_{\mathsf{NN}^\Lambda}$ as the subset of $\mathcal{H}_{\mathsf{NN}}$ with weight matrices that are Λ-bounded in their Frobenius norm: for all $j \in [L]$, $\|W_j\|_F \leq \Lambda$ for some $\Lambda > 0$.

We design a pure DP algorithm for learning L-layer feed-forward networks in $\mathcal{H}_{\mathsf{NN}^\Lambda}$ that benefits from the following margin-based guarantee.

Theorem 5.1. *Let $\varepsilon > 0, \beta \in (0,1)$, and $\rho > 0$. Then, there is an ε-DP algorithm which returns an L-layer network h^{Priv} with N neurons per layer that with probability at least $1 - \beta$ over the draw of a sample $S \sim \mathcal{D}^m$ and the internal randomness of the algorithm admits the following guarantee:*

$$R_{\mathcal{D}}(h^{\mathsf{Priv}}) \leq \min_{h \in \mathcal{H}_{\mathsf{NN}^\Lambda}} \widehat{R}_S^\rho(h) + O\left(\frac{r(2\eta\Lambda)^L \sqrt{N\theta}}{\rho\sqrt{m}} + \frac{r^2(2\eta\Lambda)^{2L}\, N\theta}{\rho^2 \varepsilon m}\right),$$

where $\theta = \log(Lm/\beta)\log(r(\eta\Lambda)^L/\rho)$.

Note that this guarantee is independent of d and, assuming L is a constant, the bound scales roughly as $\sqrt{\frac{N}{\rho^2 m}} + \frac{N}{\rho^2 \varepsilon m}$, where ρ is the confidence-margin parameter and ε is the privacy parameter. Note that, for $\varepsilon \approx 1$, the bound scales with $\sqrt{\#\text{ neurons}}$, which is more favorable than standard bounds obtained via a uniform convergence argument, which depend on d, as well as the total number of edges $\Omega(N^2)$, in addition to a similar dependence on Λ^L.

Our construction. Our DP learner is based on using L embeddings $\Phi_0 \in \mathbb{R}^{k\times d}, \ldots, \Phi_{L-1} \in \mathbb{R}^{k\times N}$ given by data-independent JL-transform matrices to reduce the dimension of the inputs in each layer, including the input layer, to $k = O\left(\frac{r^2(2\eta\Lambda)^{2L}}{\rho^2}\right)$. We randomly generate a set $\Phi = (\Phi_0, \ldots, \Phi_{L-1})$ of L independent JL matrices whose dimensions are described as above. We let $\mathcal{H}_{\mathsf{NN}}^\Phi$ denote the family of L-layer neural networks, where each network $h^\Phi \in \mathcal{H}_{\mathsf{NN}}^\Phi$ is associated with weight matrices $\Phi_0^\top \widetilde{W}_1, \ldots, \Phi_{L-1}^\top \widetilde{W}_L$ for $\widetilde{W}_j \in \mathbb{R}^{k\times N}, j \in [L-1]$, and $\widetilde{W}_L \in \mathbb{R}^{k\times 1}$ (see Figure 2, right column). We define $\mathcal{H}_{\mathsf{NN}^{2\Lambda}}^\Phi \subset \mathcal{H}_{\mathsf{NN}}^\Phi$ where $\|\widetilde{W}_j\|_F \leq 2\Lambda$ for all $j \in [L]$. We start by creating a γ-cover $\mathcal{C}$ of the product space of the matrices $\widetilde{W}_1, \ldots, \widetilde{W}_L$ associated with $\mathcal{H}_{\mathsf{NN}^{2\Lambda}}^\Phi$, where $\gamma = O\left(\frac{\rho}{r(4\eta\Lambda)^L}\right)$. $\mathcal{C}$ is a γ-cover of $\mathbb{B}^{k\times N}(2\Lambda) \times \ldots \times \mathbb{B}^{k\times N}(2\Lambda) \times \mathbb{B}^k(2\Lambda)$ with respect to $\sqrt{\sum_{j=1}^L \|\cdot\|_F^2}$. We define $\widehat{\mathcal{H}}_{\mathsf{NN}^{2\Lambda}}^\Phi \subset \mathcal{H}_{\mathsf{NN}^{2\Lambda}}^\Phi$ to be the corresponding family of networks whose associated matrices are in $\mathcal{C}$. Given an input dataset $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in \left(\mathbb{B}^d(r) \times \{\pm 1\}\right)^m$, we then run the exponential mechanism over S with privacy parameter ε, the score function being the empirical zero-one loss $\widehat{R}_S(h): h \in \widehat{\mathcal{H}}_{\mathsf{NN}^{2\Lambda}}^\Phi$, and the sensitivity $1/m$, to return a neural network $h^{\mathsf{Priv}} \in \widehat{\mathcal{H}}_{\mathsf{NN}^{2\Lambda}}^\Phi$.

6 Conclusion

We presented a series of new differentially private algorithms with dimension-independent margin guarantees, including algorithms for linear classification, kernel-based classification, or learning with a family of feed-forward neural networks, and label DP learning with general hypothesis sets. Our kernel-based algorithms can be extended to non-linear classification with many other kernels, including a variety of kernels that can be approximated using polynomial kernels, using techniques based on oblivious sketching. Our study of DP algorithms with margin guarantees for a family of neural networks can be viewed as an initiatory step that could serve as the basis for a more extensive analysis of DP algorithms for broader families of neural networks.

Acknowledgments and Disclosure of Funding

This work was done while RB was a visiting scientist at Google, NY. RB's work at OSU on this research is supported by NSF Award AF-1908281 and Google Faculty Research Award.

References

T. D. Ahle, M. Kapralov, J. B. T. Knudsen, R. Pagh, A. Velingker, D. P. Woodruff, and A. Zandieh. Oblivious sketching of high-degree polynomial kernels. In S. Chawla, editor, *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, pages 141–160. SIAM, 2020.

N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563, 2006.

N. Ailon and E. Liberty. Fast dimension reduction using rademacher series on dual BCH codes. *Discrete & Computational Geometry*, 42(4):615–630, 2009.

N. Alon, R. Livni, M. Malliaris, and S. Moran. Private PAC learning implies finite littlestone dimension. In M. Charikar and E. Cohen, editors, *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 852–860. ACM, 2019.

H. Asi, V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: Optimal rates in l1 geometry. *arXiv preprint arXiv:2103.01516*, 2021.

H. Avron, H. L. Nguyen, and D. P. Woodruff. Subspace embeddings for the polynomial kernel. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2258–2266, 2014.

P. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. *Advances in kernel methods: support vector learning*, pages 43–54, 1999.

P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 1998.

R. Bassily, A. D. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pages 464–473. IEEE Computer Society, 2014.

R. Bassily, V. Feldman, K. Talwar, and A. Thakurta. Private stochastic convex optimization with optimal rates. *arXiv preprint arXiv:1908.09970*, 2019.

R. Bassily, C. Guzmán, and M. Menart. Differentially private stochastic optimization: New results in convex and non-convex settings. *arXiv preprint arXiv:2107.05585. Appeared at NeurIPS 2021.*, 2021a.

R. Bassily, C. Guzmán, and A. Nandi. Non-euclidean differentially private stochastic convex optimization. *arXiv preprint arXiv:2103.01278*, 2021b.

A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. In C. Li, editor, *Proceedings of the Twenty-fourth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 13-15, 2005, Baltimore, Maryland, USA*, pages 128–138. ACM, 2005.

M. Bun, K. Nissim, U. Stemmer, and S. P. Vadhan. Differentially private release and learning of threshold functions. In V. Guruswami, editor, *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 634–649. IEEE Computer Society, 2015.

R. I. Busa-Fekete, U. Syed, S. Vassilvitskii, et al. On the pitfalls of label differential privacy. In *NeurIPS 2021 Workshop LatinX in AI*, 2021a.

R. I. Busa-Fekete, U. Syed, S. Vassilvitskii, et al. Population level privacy leakage in binary classification wtih label noise. In *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021b.

K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12:1069–1109, 2011.

K. Chaudhuri, D. J. Hsu, and S. Song. The large margin mechanism for differentially private maximization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1287–1295, 2014.

C. Cortes, S. Greenberg, and M. Mohri. Relative deviation learning bounds and generalization with unbounded loss functions. *Annals of Mathematics and Artificial Intelligence*, 85(1):45–70, 2019.

C. Cortes, M. Mohri, and A. Theertha Suresh. Relative deviation margin bounds. In *International Conference on Machine Learning*, pages 2122–2131. PMLR, 2021.

C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

C. Dwork, F. McSherry, K. Nissim, and A. D. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

H. Esfandiari, V. Mirrokni, U. Syed, and S. Vassilvitskii. Label differential privacy via clustering. *arXiv preprint arXiv:2110.02159*, 2021.

V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: optimal rates in linear time. In K. Makarychev, Y. Makarychev, M. Tulsiani, G. Kamath, and J. Chuzhoy, editors, *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 439–449. ACM, 2020.

B. Ghazi, N. Golowich, R. Kumar, P. Manurangsi, and C. Zhang. On deep learning with label differential privacy. *arXiv preprint arXiv:2102.06062*, 2021.

P. Jain and A. Thakurta. Differentially private learning with kernels. In *International conference on machine learning*, pages 118–126. PMLR, 2013.

P. Jain and A. Thakurta. (near) dimension independent risk bounds for differentially private learning. In *ICML*, 2014.

W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space 26. *Contemporary mathematics*, 26:28, 1984.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.

F. Krahmer and R. Ward. New and improved Johnson–Lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.

K. G. Larsen and J. Nelson. Optimality of the johnson-lindenstrauss lemma. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 633–638. IEEE, 2017.

Q. V. Le, T. Sarlós, and A. J. Smola. Fastfood - computing Hilbert space expansions in loglinear time. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 244–252. JMLR.org, 2013.

H. Le Nguyen, J. Ullman, and L. Zakynthinou. Efficient private algorithms for learning large-margin halfspaces. In *Algorithmic Learning Theory*, pages 704–724. PMLR, 2020.

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

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

J. Nelson. Johnson–Lindenstrauss notes, 2010.

J. Nelson. Dimensionality reduction—notes 2, 2015.

J. J. O. Nelson. *Sketching and streaming high-dimensional vectors*. PhD thesis, Massachusetts Institute of Technology, 2011.

N. Pham and R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–247, 2013.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 1177–1184. Curran Associates, Inc., 2007.

A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21, 2008.

S. Raskhodnikova and A. Smith. Lipschitz extensions for node-private graph statistics and the generalized exponential mechanism. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 495–504. IEEE, 2016.

B. I. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *arXiv preprint arXiv:0911.5708*, 2009.

W. Rudin. *Fourier analysis on groups*. Courier Dover Publications, 2017.

R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *ICML*, pages 322–330, 1997.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press: Cambridge, MA, 2002.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, 2004.

S. Song, T. Steinke, O. Thakkar, and A. Thakurta. Evading the curse of dimensionality in unconstrained private glms. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2638–2646. PMLR, 13–15 Apr 2021a. URL http://proceedings.mlr.press/v130/song21a.html.

Z. Song, D. P. Woodruff, Z. Yu, and L. Zhang. Fast sketching of polynomial kernels of polynomial degree. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9812–9823. PMLR, 2021b.

S. Yuan, M. Shen, I. Mironov, and A. C. Nascimento. Practical, label private deep learning training based on secure multiparty computation and differential privacy. *Cryptology ePrint Archive*, 2021.

Contents of Appendix

A Useful Lemmas

We use empirical Bernstein bounds, properties of exponential mechanism and Johnson-Lindenstrauss lemmas which we state below.

Lemma A.1 (Relative deviation bound)**.** *For any hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X}$ to R, with probability at least $1 - \beta$, the following inequality holds for all $h \in \mathcal{H}$:*

$$R_{\mathcal{D}}(h) \le \widehat{R}_S(h) + 2\sqrt{\widehat{R}_S(h)\frac{V(\mathcal{H})\log(2m) + \log\frac{4}{\beta}}{m}} + 4\frac{V(\mathcal{H})\log(2m) + \log\frac{4}{\beta}}{m},$$

where $V(\mathcal{H})$ is the VC-dimension of class $\mathcal{H}$.

The above lemma is obtained by combining [Cortes et al., 2019, Corollary 7] and VC-dimension bounds.

Lemma A.2 (Relative deviation margin bound [Cortes et al., 2021])**.** *Fix $\rho \ge 0$. Then, for any hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X}$ to $\mathbb{R}$ with $d = \mathrm{fat}_{\frac{\rho}{16}}(\mathcal{H})$, with probability at least $1 - \beta$, the following holds for all $h \in \mathcal{H}$:*

$$R_{\mathcal{D}}(h) \le \widehat{R}^{\rho}_S(h) + 2\sqrt{\widehat{R}^{\rho}_S(h)\frac{M}{m}} + \frac{M}{m},$$

where $M = 1 + d\log_2(2c^2m)\log_2\frac{2cem}{d} + \log\frac{1}{\beta}$ and $c = 17$.

Lemma A.3. *Let $\beta, \gamma \in (0,1)$. Let $T \subset \mathbb{R}^d$ be any set of m vectors. There exists $k = O\left(\frac{\log\left(\frac{m}{\beta}\right)}{\gamma^2}\right)$ such that for any random $k \times d$ matrix Φ with entries drawn i.i.d. uniformly from $\{\pm\frac{1}{\sqrt{k}}\}$, the following inequalities hold simultaneously with probability at least $1 - \beta$ over the choice of Φ:*

- *For any $u \in T$,*

$$\left(1 - \frac{\gamma}{3}\right)\|u\|_2^2 \le \|\Phi u\|_2^2 \le \left(1 + \frac{\gamma}{3}\right)\|u\|_2^2.$$

- *For any $u, v \in T$,*

$$|\langle \Phi u, \Phi v\rangle - \langle u, v\rangle| \le \frac{\gamma}{3}\|u\|_2\,\|v\|_2.$$

Proof. The first property is simply the Johnson-Lindenstrauss (JL) property and follows from the standard JL lemma (see, e.g., [Johnson and Lindenstrauss, 1984, Larsen and Nelson, 2017]). Below we show both first and second property holds simultaneously. Define

$$\widetilde{T} \triangleq \left\{z \in \mathbb{R}^d : \ z = \frac{u}{\|u\|_2} \pm \frac{v}{\|v\|_2}, \ u, v \in T\right\}.$$

Note that the number of non-zero vectors in $\widetilde{T}$ is at most m^2. By the JL lemma [Johnson and Lindenstrauss, 1984, Larsen and Nelson, 2017] over the set $T \cup \widetilde{T}$, there exists $k = O\left(\frac{\log(m^2/\beta)}{\gamma^2}\right) = O\left(\frac{\log\left(\frac{m}{\beta}\right)}{\gamma^2}\right)$ and Φ' such that with probability $\ge 1 - \beta$, for all $u, v \in T$ we have

$$\left(1 - \frac{\gamma}{3}\right)\|\bar{u} + \bar{v}\|_2^2 \le \|\Phi'(\bar{u} + \bar{v})\|_2^2 \le \left(1 + \frac{\gamma}{3}\right)\|\bar{u} + \bar{v}\|_2^2, \tag{3}$$

$$\left(1 - \frac{\gamma}{3}\right)\|\bar{u} - \bar{v}\|_2^2 \le \|\Phi'(\bar{u} - \bar{v})\|_2^2 \le \left(1 + \frac{\gamma}{3}\right)\|\bar{u} - \bar{v}\|_2^2, \tag{4}$$

$$\left(1 - \frac{\gamma}{3}\right)\|u\|_2^2 \le \|\Phi' u\|_2^2 \le \left(1 + \frac{\gamma}{3}\right)\|u\|_2^2. \tag{5}$$

where $\bar{u} = \frac{u}{\|u\|_2}$ and $\bar{v} = \frac{v}{\|v\|_2}$. (5) implies the first result in the lemma. Now, fix any $u, v \in T$. Let $\bar{u} = \frac{u}{\|u\|_2}$ and $\bar{v} = \frac{v}{\|v\|_2}$. Observe that for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, we have $\langle \mathbf{a}, \mathbf{b} \rangle = \frac{1}{4}\left(\|\mathbf{a}+\mathbf{b}\|_2^2 - \|\mathbf{a}-\mathbf{b}\|_2^2\right)$. Hence, we have

$$\begin{aligned} \left|\langle \bar{u}, \bar{v}\rangle - \langle \Phi'\bar{u}, \Phi'\bar{v}\rangle\right| &\le \frac{1}{4}\left|\|\bar{u}+\bar{v}\|_2^2 - \|\Phi'(\bar{u}+\bar{v})\|_2^2\right| + \frac{1}{4}\left|\|\Phi'(\bar{u}-\bar{v})\|_2^2 - \|\bar{u}-\bar{v}\|_2^2\right| \qquad () \\ &\le \frac{\gamma}{12}\left(\|\bar{u}+\bar{v}\|_2^2 + \|\bar{u}-\bar{v}\|_2^2\right) \\ &\le \frac{\gamma}{3}, \end{aligned}$$

where the second inequality follows from (3) and (4) 'and the third inequality follows from the triangle inequality and the fact that $\|\bar{u}\|_2 = \|\bar{v}\|_2 = 1$. Hence, we finally have

$$|\langle u, v\rangle - \langle \Phi' u, \Phi' v\rangle| \le \|u\|_2 \|v\|_2 \left|\langle \bar{u}, \bar{v}\rangle - \langle \Phi'\bar{u}, \Phi'\bar{v}\rangle\right| \le \frac{\gamma}{3}\|u\|_2\|v\|_2.$$

□

The time complexity to apply the random matrix Φ in Lemma A.3 to a vector v is $k \cdot d$, which can be prohibitive in many cases. There are several works which provide Φ that support fast matrix vector products. Ailon and Chazelle [2006] provides a Φ which can be applied in time $c' d \log d + \frac{c' \log(d/\beta)\log^2(1/\beta)}{\gamma^2}$, however the results are stated with constant probability. Nelson [2010] gave a slightly different construction which can be applied in time $c' d \log d + \frac{c' \log(d/\beta)\log^2(1/\beta)}{\gamma^4}$ and the results hold with high probability. Ailon and Liberty [2009] provided a construction which can be applied in time $c' d \log k$ for $k = O(d^{0.499})$. Krahmer and Ward [2011] showed that any RIP matrix can be used for JL-transform and provided JL-transform results for several fast random projections. Since we need high probability bounds without any restrictions, we use the following result, which is computationally efficient, but is suboptimal in the projection dimension up to logarithmic factors.

Lemma A.4. *Let $\beta, \gamma \in (0,1)$ and c and c' be sufficiently large constants. Let $T \subset \mathbb{R}^d$ be any set of m vectors. Let $k = \frac{c \log\left(\frac{m}{\beta}\right)\log\left(\frac{m}{\gamma\beta}\right)}{\gamma^2}$. There exists a matrix Φ which can be applied to any vector v in time $c' d \log d + c' k$, such that the following inequalities hold simultaneously with probability at least $1-\beta$ over the choice of Φ:*

- *For any $u \in T$,*

$$\left(1 - \frac{\gamma}{3}\right)\|u\|_2^2 \le \|\Phi u\|_2^2 \le \left(1 + \frac{\gamma}{3}\right)\|u\|_2^2. \tag{6}$$

- *For any $u, v \in T$,*

$$|\langle \Phi u, \Phi v\rangle - \langle u, v\rangle| \le \frac{\gamma}{3}\|u\|_2\,\|v\|_2. \tag{7}$$

Property (6) in the above result follows directly from Nelson [2015] and the proof for property (7) is similar to that of Lemma A.3 and is omitted.

B DP Algorithms for Linear Classification with Margin Guarantees

B.1 Proof of Theorem 3.1

Theorem 3.1. *Algorithm 1 is ε-differentially private. For any $\beta \in (0,1)$, with probability at least $1-\beta$ over the draw of a sample S of size m from $\mathcal{D}$, the solution w^{Priv} it returns satisfies:*

$$R_{\mathcal{D}}(w^{\mathsf{Priv}}) \le \min_{w \in \mathbb{B}^d(\Lambda)} \left\{ \widehat{R}_S^{\rho}(w) + \widetilde{O}\left(\sqrt{\widehat{R}_S^{\rho}(w)\frac{\Lambda^2 r^2}{m\rho^2}} + \frac{\Lambda^2 r^2}{\rho^2 \min(1,\varepsilon) m} \right) \right\}.$$

A more precise version of the above bound is given as follows:

$$R_{\mathcal{D}}(w^{\mathsf{Priv}}) \le \min_{w \in \mathbb{B}^d(\Lambda)} \left\{ \widehat{R}^{\rho}_{S}(w) + O\left(\sqrt{\widehat{R}^{\rho}_{S}(w) \left(\frac{\Lambda^2 r^2 \log^2\left(\frac{m}{\beta}\right)}{m\rho^2} + \frac{\log\left(\frac{1}{\beta}\right)}{m} \right)} + \Gamma \right) \right\},$$

$$\text{where} \quad \Gamma = \frac{\Lambda^2 r^2 \log\left(\frac{m}{\beta}\right) \log\left(\frac{\Lambda r}{\beta\rho}\right)}{\rho^2 \varepsilon m} + \frac{\Lambda^2 r^2 \log^2\left(\frac{m}{\beta}\right)}{m\rho^2} + \frac{\log\left(\frac{1}{\beta}\right)}{m}.$$

Proof. The proof of privacy follows from combining the following two properties: $\tilde{w}$ is generated via the exponential mechanism, which an ε-differentially private mechanism, and Φ is generated independently of S.

We now prove the accuracy guarantee of Algorithm 1. If $m < \frac{c\Lambda^2 r^2 \log\left(\frac{m}{\beta}\right) \log\left(\frac{20\Lambda r}{\beta\rho}\right)}{\rho^2\varepsilon}$, for some constant c, the bound follow trivially. Hence in the rest of the proof we assume that m is at least $\frac{c\Lambda^2 r^2 \log\left(\frac{m}{\beta}\right) \log\left(\frac{20\Lambda r}{\beta\rho}\right)}{\rho^2\varepsilon}$ for some large constant c. Let

$$\alpha = \frac{c\Lambda^2 r^2 \log\left(\frac{m}{\beta}\right) \log\left(\frac{20\Lambda r}{\beta\rho}\right)}{\rho^2 \varepsilon m}.$$

First, observe that

$$\begin{aligned} \widehat{R}_S(w^{\mathsf{Priv}}) &= \frac{1}{m} \sum_{(x,y)\in S} \mathbf{1}\left(y\langle w^{\mathsf{Priv}}, x\rangle\right) \\ &= \frac{1}{m} \sum_{(x,y)\in S} \mathbf{1}\left(y\langle \Phi^{\mathsf{T}} \tilde{w}, x\rangle\right) \\ &= \frac{1}{m} \sum_{(x,y)\in S} \mathbf{1}\left(y\langle \tilde{w}, \Phi x\rangle\right) \\ &= \frac{1}{m} \sum_{(x_\Phi,y)\in S_\Phi} \mathbf{1}\left(y\langle \tilde{w}, x_\Phi\rangle\right) \\ &= \widehat{R}^{(k)}_{S_\Phi}(\tilde{w}). \end{aligned}$$

Let $\widehat{w} \in \operatorname{argmin}_{w\in\mathcal{C}} \widehat{R}^{(k)}_{S_\Phi}(w)$. Note that $|\mathcal{C}| = \left(\frac{20\,\Lambda\, r}{\rho}\right)^k$, where $k = \frac{c'\,\Lambda^2\, r^2\, \log\left(\frac{m}{\beta}\right)}{\rho^2}$. Hence, by the accuracy properties of the exponential mechanism and the fact that $m \ge \frac{4\log\left(\frac{4|\mathcal{C}|}{\beta}\right)}{\varepsilon\alpha}$, we have that with probability at least $1-\beta/4$,

$$\widehat{R}^{(k)}_{S_\Phi}(\tilde{w}) \le \widehat{R}^{(k)}_{S_\Phi}(\widehat{w}) + \alpha.$$

Combining the above facts, we get that with probability at least $1-\beta/4$,

$$\widehat{R}_S(w^{\mathsf{Priv}}) \le \widehat{R}^{(k)}_{S_\Phi}(\widehat{w}) + \alpha. \tag{8}$$

Let $w^* \in \operatorname{argmin}_{w\in\mathbb{B}^d(\Lambda)} \widehat{R}^{\rho}_S(w)$ and let $w^*_\Phi = \Phi w^*$. Note that

$$\|w^*_\Phi\|_2 \le \sqrt{1 + \frac{\rho}{3\Lambda r}}\|w^*\|_2 \le 2\|w^*\|_2 \le 2\Lambda,$$

and hence $w^*_\Phi \in \mathbb{B}^k(2\Lambda)$. Since $\mathcal{C}$ is $\frac{\rho}{10\,r}$-cover of $\mathbb{B}^k(2\Lambda)$, then there must be $w_c \in \mathcal{C}$ such that $\|w_c - w^*_\Phi\|_2 \le \frac{\rho}{10\,r}$. Hence, observe that for any $(x_\Phi, y) \in S_\Phi$,

$$\begin{aligned} y\langle w_c, x_\Phi\rangle &= y\langle w^*_\Phi, x_\Phi\rangle + y\langle w_c - w^*_\Phi, x_\Phi\rangle \\ &\ge y\langle w^*_\Phi, x_\Phi\rangle - \|w_c - w^*_\Phi\|_2\,\|x_\Phi\|_2 \\ &\ge y\langle w^*_\Phi, x_\Phi\rangle - \frac{\rho}{10\,r}\|x_\Phi\|_2. \end{aligned}$$

Now, by Lemma A.3, with probability at least $1 - \beta/4$, for all x_Φ s.t. $(x_\Phi, y) \in S_\Phi$ we have $\|x_\Phi\|_2 \le \sqrt{1 + \frac{\rho}{3\Lambda r}}\|x\|_2 \le \sqrt{1 + \frac{\rho}{3\Lambda r}}r$. Hence, we get that with probability at least $1 - \beta/4$ for all $(x_\Phi, y) \in S_\Phi$

$$y\langle w_c, x_\Phi\rangle \ge y\langle w^*_\Phi, x_\Phi\rangle - 0.15\rho.$$

The last inequality implies that for any $\rho' > 0$, with probability at least $1 - \beta/4$ (over the choice of Φ), we must have $\widehat{R}^{\rho',(k)}_{S_\Phi}(w_c) \le \widehat{R}^{\rho'+0.15\rho,(k)}_{S_\Phi}(w^*_\Phi)$. In particular, with probability at least $1 - \beta/4$ we have

$$\widehat{R}^{0.5\rho,(k)}_{S_\Phi}(w_c) \le \widehat{R}^{0.65\rho,(k)}_{S_\Phi}(w^*_\Phi). \tag{9}$$

Moreover, by the definition of $\widehat{w}$, we have $\widehat{R}^{(k)}_{S_\Phi}(\widehat{w}) \le \widehat{R}^{(k)}_{S_\Phi}(w_c) \le \widehat{R}^{0.5\rho,(k)}_{S_\Phi}(w_c)$. Combining this fact with (8) and (9), we get that with probability at least $1 - \beta/2$

$$\widehat{R}_S(w^{\mathsf{Priv}}) \le \widehat{R}^{0.65\rho,(k)}_{S_\Phi}(w^*_\Phi) + \alpha. \tag{10}$$

Now, by Lemma A.3 and the fact that $\|w^*\|_2 \le \Lambda$ and $\|x\|_2 \le r$, it follows that with probability at least $1 - \beta/4$ for all $(x_\Phi, y) \in S_\Phi$, we have

$$y\langle w^*, x\rangle \ge \rho \implies y\langle w^*_\Phi, x_\Phi\rangle \ge \rho/3.$$

This directly implies that with probability at least $1 - \beta/4$,

$$\widehat{R}^{0.65\rho,(k)}_{S_\Phi}(w^*_\Phi) \le \widehat{R}^{\rho}_S(w^*).$$

Combining this with (10), we can assert that with probability at least $1 - \frac{3}{4}\beta$, we have

$$\widehat{R}_S(w^{\mathsf{Priv}}) \le \widehat{R}^{\rho}_S(w^*) + \alpha. \tag{11}$$

In the final step of the proof, we rely on a standard uniform convergence argument to bound $R_{\mathcal{D}}(w^{\mathsf{Priv}})$ in terms of $\widehat{R}_S(w^{\mathsf{Priv}})$. Note that the VC-dimension of $\{\operatorname{sgn} \circ h_w : w \in \mathcal{C}\}$ is k. By Lemma A.1, with probability at least $1 - \beta/4$

$$R_{\mathcal{D}}(w^{\mathsf{Priv}}) - \widehat{R}_S(w^{\mathsf{Priv}}) \le 2\sqrt{\widehat{R}_S(w^{\mathsf{Priv}})\frac{k\log(2m) + \log(16/\beta)}{m}} + 4\frac{k\log(2m) + \log(16/\beta)}{m}.$$

Combining the above two equations, we get with probability at least $1 - \beta$,

$$R_{\mathcal{D}}(w^{\mathsf{Priv}}) \le \widehat{R}^{\rho}_S(w^*) + 2\sqrt{\widehat{R}^{\rho}_S(w^*)\frac{k\log(2m) + \log(16/\beta)}{m}} + 2\alpha + 8\frac{k\log(2m) + \log(16/\beta)}{m}.$$

The lemma follows from observing that $w^* \in \underset{w\in\mathbb{B}^d(\Lambda)}{\operatorname{argmin}}\, \widehat{R}^{\rho}_S(w)$ if and only if $w^* \in \underset{w\in\mathbb{B}^d(\Lambda)}{\operatorname{argmin}}\, \widehat{R}^{\rho}_S(w) + 2\sqrt{\widehat{R}^{\rho}_S(w)\frac{k\log(2m)+\log(16/\beta)}{m}}$. □

B.2 Algorithm 4 of Section 3.2 and Proof of Lemma 3.1

Here, we give the details of Algorithm 4 invoked in step 4 of Algorithm 2. We describe here a more general setup where the loss function is any (possibly non-smooth) convex generalized linear loss (GLL). Given a parameter space $\mathcal{W}$, feature space $\mathcal{X}$, and label/target set $\mathcal{Y}$, a GLL is a loss function defined over $\mathcal{W} \times (\mathcal{X} \times \mathcal{Y})$ that can be written as $\ell(\langle w, x\rangle, y)$, $w \in \mathcal{W}, x \in \mathcal{X}, y \in \mathcal{Y}$ for some function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$. Here, we assume that for any $y \in \mathcal{Y}$, $\ell(\cdot, y)$ is convex and $\frac{1}{\rho}$-Lipschitz. We also assume that $\mathcal{W} \subseteq \mathbb{B}^k(\Lambda)$ for some $\Lambda > 0$, $\mathcal{X} \subseteq \mathbb{B}^k(\tilde{r})$ for some $\tilde{r} > 0$, and $\mathcal{Y} \subseteq [-1, 1]$. Given a dataset $\widetilde{S} = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$, we define the empirical risk of $w \in \mathcal{W}$ with respect to $\widetilde{S}$ as $\widehat{L}_{\widetilde{S}}(w) \triangleq \frac{1}{m}\sum_{i=1}^m \ell(\langle w, x_i\rangle, y)$. Note that the setup in Algorithm 2 is a special case of the above.

Given an input dataset $\widetilde{S} \in (\mathcal{X} \times \mathcal{Y})^m$, Algorithm 4 below invokes the "Phased SGD algorithm for GLL" [Bassily et al., 2021a, Algorithm 2] on a set $\widehat{S}$ of m samples drawn uniformly with replacement from $\widetilde{S}$, and hence obtain an output $\tilde{w} \in \mathcal{W}$. In the sequel, we will refer to the algorithm in [Bassily et al., 2021a] as $\mathcal{A}_{\mathsf{GLL}}$. Note that the expected loss with respect to the choice of $\widehat{S} \leftarrow \widetilde{S}$ is the empirical risk with respect to $\widetilde{S}$. Hence, one can derive a bound on the expected excess empirical risk that is roughly the same as the bound in [Bassily et al., 2021a, Theorem 6] on the expected excess population risk. However, we note that ensuring that Algorithm 4 is (ε, δ)-DP does not follow directly from the privacy guarantee of the $\mathcal{A}_{\mathsf{GLL}}$ since the sample $\widehat{S}$ may contain duplicate entries from $\widetilde{S}$. Nonetheless, we show that the privacy guarantee can be attained by appropriately setting the input privacy parameters to $\mathcal{A}_{\mathsf{GLL}}$ together with a careful privacy analysis. To transform the in-expectation bound into a high-probability bound, we perform a standard confidence-boosting procedure [Bassily et al., 2014, Appendix D], where the procedure described above is repeated independently $M = O(\log(1/\beta))$ times to generate $\tilde{w}^1, \ldots, \tilde{w}^M$, and finally, the exponential mechanism (with a score function $-\widehat{L}_{\widetilde{S}}$) is used to privately select a final output $\tilde{w} \in \{\tilde{w}^1, \ldots, \tilde{w}^M\}$.

Algorithm 4 DP-ERM algorithm for GLLs

```
Require: Private dataset $\widetilde{S} = \big((x_1, y_1), \ldots, (x_m, y_m)\big) \in (\mathcal{X} \times \mathcal{Y})^m$, where $\mathcal{X} \subseteq \mathbb{B}^k(\tilde{r})$ and $\mathcal{Y} \subseteq [-1, 1]$; parameter space $\mathcal{W} \subseteq \mathbb{B}^k(\Lambda)$; privacy parameters $(\varepsilon, \delta)$; confidence parameter $\beta \in (0, 1)$; convex, $\frac{1}{\rho}$-Lipschitz loss function $\ell$ for some $\rho > 0$; Oracle access to algorithm $\mathcal{A}_{\mathsf{GLL}}$ [Bassily et al., 2021a, Algorithm 2].
1: Let $M := \log(2/\beta)$.
2: Let $\varepsilon' := \frac{\varepsilon}{4M \log(2M/\delta)}$.
3: Let $\delta' := \frac{\delta^2}{4M \log(2M/\delta)}$.
4: for $t = 1$ to $M$ do
5:     Sample $\widehat{S}^t = \big((\widehat{x}^t_1, \widehat{y}^t_1), \ldots \big(\widehat{x}^t_m, \widehat{y}^t_m\big)\big) \leftarrow \widetilde{S}$ uniformly with replacement.
6:     $\tilde{w}^t = \mathcal{A}_{\mathsf{GLL}}(\widehat{S}^t; \varepsilon', \delta')$, where $\mathcal{A}_{\mathsf{GLL}}$ is [Bassily et al., 2021a, Algorithm 2] (the other obvious inputs to $\mathcal{A}_{\mathsf{GLL}}$ are omitted; the smoothing parameter and the oracle accuracy parameter of $\mathcal{A}_{\mathsf{GLL}}$ are set as in [Bassily et al., 2021a, Theorem 6]).
7: Run the exponential mechanism with privacy parameter $\varepsilon/2$ to select $\tilde{w}$ from the set $\big(\tilde{w}^1, \ldots, \tilde{w}^M\big)$ associated with scores $\big(-\widehat{L}_{\widetilde{S}}(\tilde{w}^t) : t \in [M]\big)$.
8: return $\tilde{w}$
```

Lemma 3.1. *Let $m \in \mathbb{N}$, $0 < \delta < \frac{1}{m}$, and $0 < \varepsilon \leq \log(1/\delta)$. Algorithm 4 (Appendix B.2) is (ε, δ)-DP. Let $\beta \in (0, 1)$. Let $k \in \mathbb{N}$, and $\tilde{r}, \Lambda > 0$. Let $\widetilde{S} \in \big(\mathbb{B}^k(\tilde{r}) \times \{\pm 1\}\big)^m$ be the input dataset and $\mathbb{B}^k(\Lambda)$ be the parameter space. With probability $1 - \beta$ over the randomness in Algorithm 4, the output $\tilde{w}$ satisfies*

$$\widehat{L}^{\rho}_{\widetilde{S}}(\tilde{w}) \leq \min_{w \in \mathbb{B}^k(\Lambda)} \widehat{L}^{\rho}_{\widetilde{S}}(w) + \frac{\Lambda \tilde{r}}{\rho} \cdot O\left(\frac{1}{\sqrt{m}} + \frac{\sqrt{k} \log^{\frac{3}{2}}(\frac{1}{\delta}) \log(\frac{1}{\beta})}{\varepsilon m}\right).$$

Moreover, Algorithm 2 requires $O(m \log(m) \log(1/\beta))$ gradient computations.

Proof. First, we show the privacy guarantee. Fix a round $t \in [M]$ of Algorithm 4. We will show that the t-th round is $(\frac{\varepsilon}{2M}, \frac{\delta}{M})$-DP. Suppose we can do that. Then, by the basic composition property of DP, the entire M rounds of the algorithm is $(\frac{\varepsilon}{2}, \delta)$-DP. Next, we note that step 7 is $(\frac{\varepsilon}{2}, 0)$-DP by the privacy guarantee of the exponential mechanism. Hence, again by basic composition of DP, we conclude that Algorithm 4 is (ε, δ)-DP. Thus, it remains to show that for any fixed $t \in [M]$, the t-th round is $(\hat{\varepsilon}, \hat{\delta})$-DP, where $\hat{\varepsilon} = \frac{\varepsilon}{2M}$ and $\hat{\delta} = \frac{\delta}{M}$. Fix any data point $(x_i, y_i) \in \widetilde{S}$. Let J denote the number of appearances of (x_i, y_i) in $\widehat{S}^t$. Note that $J \sim \mathsf{Bin}(m, 1/m)$. Hence, using the multiplicative Chernoff's bound, $J \leq 2 \log(2/\hat{\delta})$ with probability at least $1 - \hat{\delta}/2$. We will show that conditioned on this event, the t-th round is $(\hat{\varepsilon}, \hat{\delta}/2)$-DP, which suffices to prove our privacy claim for round t. Given that $J \leq 2 \log(2/\hat{\delta})$ and since $\mathcal{A}_{\mathsf{GLL}}$ is $(\frac{\hat{\varepsilon}}{2 \log(2/\hat{\delta})}, \frac{\delta \hat{\delta}}{4 \log(2/\hat{\delta})})$-DP with respect to to its input dataset $\widehat{S}^t$, then by the group-privacy property of DP [Dwork and Roth, 2014], round t is $(\hat{\varepsilon}, \delta'')$-DP with respect to the dataset $\widetilde{S}$, where

$$\begin{aligned}\delta'' &= \frac{\delta\hat{\delta}}{4\log(2/\hat{\delta})} \cdot \sum_{j=0}^{2\log(2/\hat{\delta})} e^{\frac{\hat{\varepsilon}}{2\log(2/\hat{\delta})}j} \\ &= \frac{\delta\hat{\delta}}{4\log(2/\hat{\delta})} \cdot \frac{e^{\hat{\varepsilon}} - 1}{e^{\frac{\hat{\varepsilon}}{2\log(2/\hat{\delta})}} - 1} \\ &\leq \frac{\delta\hat{\delta}\left(e^{\hat{\varepsilon}} - 1\right)}{2\hat{\varepsilon}} \\ &\leq \frac{\hat{\delta}}{2},\end{aligned}$$

where the third inequality follows from the fact that $e^{\frac{\hat{\varepsilon}}{2\log(2/\hat{\delta})}} - 1 \geq \frac{\hat{\varepsilon}}{2\log(2/\hat{\delta})}$, and the last step follows from the fact that $\frac{e^a-1}{a}$ is increasing in $a > 0$ and the assumption that $\varepsilon \leq \log(1/\delta)$ (and hence $\hat{\varepsilon} < \varepsilon \leq \log(1/\delta)$). Hence, we have shown that any given round of the algorithm is $(\varepsilon, \hat{\delta})$-DP. This concludes the proof of the privacy guarantee.

We now prove the bound on the excess empirical risk. Fix any round t. Let $\widehat{\mathcal{D}}_{\widetilde{S}}$ denote the empirical distribution of $\widetilde{S}$. Note that $\widehat{S}^t \sim \widehat{\mathcal{D}}^m_{\widetilde{S}}$, i.e., $\widehat{S}^t$ is comprised of m independent samples from $\widehat{\mathcal{D}}_{\widetilde{S}}$. Hence, $\underset{(x,y)\sim\widehat{\mathcal{D}}_{\widetilde{S}}}{\mathbb{E}}[\ell(\langle w, x\rangle, y)] = \widehat{L}_{\widetilde{S}}(w)$. Thus, by the excess risk guarantee of $\mathcal{A}_{\mathsf{GLL}}$ [Bassily et al., 2021a, Theorem 6], we have

$$\begin{aligned}\mathbb{E}\left[\widehat{L}^{\rho}_{\widetilde{S}}(\tilde{w})\right] - \min_{w\in\mathbb{B}^k(\Lambda)} \widehat{L}^{\rho}_{\widetilde{S}}(w) &= \frac{\Lambda\tilde{r}}{\rho}\cdot O\left(\frac{1}{\sqrt{m}} + \frac{\sqrt{k\log(\frac{1}{\delta'})}}{\varepsilon' m}\right) \\ &= \frac{\Lambda\tilde{r}}{\rho}\cdot O\left(\frac{1}{\sqrt{m}} + \frac{\sqrt{k}\log^{\frac{3}{2}}(\frac{1}{\delta})\log(\frac{1}{\beta})}{\varepsilon m}\right),\end{aligned}$$

where the expectation is with respect to the sampling step (step 5 of Algorithm 4) and the randomness in $\mathcal{A}_{\mathsf{GLL}}$. Note the last step follows from the setting of ε' and δ' in Algorithm 4 and the fact that $\log(\log(1/\beta)/\delta) = O(\log(1/\delta))$, which follows from the assumption $\delta < 1/m$ (in the statement of the lemma) and $\log(1/\beta) < m$ (since the bound would be trivial otherwise). Given this expectation guarantee on the output of each round, the final selection step (step 7) returns a parameter $\tilde{w}$ that satisfies the bound above with probability at least $1 - \beta$. This can be shown by following the same argument in [Bassily et al., 2014, Appendix D] while noting that the sensitivity of the score function $-\widehat{L}_{\widetilde{S}}$ is bounded by $\frac{\Lambda\tilde{r}}{m}$.

Finally, the running time of $\mathcal{A}_{\mathsf{GLL}}$, measured in terms of gradient computations, is $O(m\log(m))$ [Bassily et al., 2021a, Theorem 6]. Hence, the gradient complexity of Algorithm 4 is bounded by $O(m\log(m)\log(1/\beta))$. □

B.3 Proof of Theorem 3.2

Theorem 3.2. *Let $0 < \delta < \frac{1}{m}$ and $0 < \varepsilon \leq \log(1/\delta)$. Algorithm 2 is (ε, δ)-DP. Let $\beta \in (0,1)$. Let $S \sim \mathcal{D}^m$ for a distribution $\mathcal{D}$ over $\mathbb{B}^d(r) \times \{\pm 1\}$. Algorithm 2 outputs $w^{\mathsf{Priv}} \in \mathbb{R}^d$ such that with probability at least $1 - \beta$, we have*

$$R_{\mathcal{D}}(w^{\mathsf{Priv}}) \leq \min_{w\in\mathbb{B}^d(\Lambda)} \widehat{L}^{\rho}_{S}(w) + \widetilde{O}\left(\frac{\Lambda r}{\rho\sqrt{\min(1,\varepsilon)\, m}}\right).$$

Moreover, Algorithm 2 runs in time $O\Big(md\log(\max(d,m)) + \varepsilon m^2\log(m)/\log^{\frac{3}{2}}(1/\delta)\Big)$.

A more precise version of the above bound is given as follows:

$$R_{\mathcal{D}}(w^{\mathsf{Priv}}) \le \min_{w\in\mathbb{B}^d(\Lambda)} \widehat{L}^{\rho}_{S}(w) + O\left(\sqrt{\frac{\log(1/\beta)}{m}} + \frac{\Lambda r}{\rho}\left(\frac{1}{\sqrt{m}} + \frac{\sqrt{\log(\frac{m}{\beta})\log(\frac{1}{\beta})}\log^{\frac{3}{4}}(\frac{1}{\delta})}{\sqrt{\varepsilon m}}\right)\right).$$

Proof. The proof of privacy follows directly from the (ε,δ)-DP guarantee of Algorithm 4 (step 4) and the fact that DP is closed under post-processing.

Next, we will prove the claimed margin bound. For simplicity and without loss of generality, we will set $\Lambda = 1$. For the general setting of Λ, the claimed bound follows by rescaling the parameter vectors in the proof.

Recall that $x_\Phi \triangleq \Phi x$. Let $w^* \in \underset{w\in\mathbb{B}^d}{\operatorname{argmin}}\,\widehat{L}^{\rho}_{S}(w)$. Define $w^*_\Phi \triangleq \Phi w^*$. By Lemma A.4, there is $\gamma = O\left(\sqrt{\frac{\log\left(\frac{m}{\beta}\right)}{k}}\right) = O\left(\sqrt{\frac{\log(1/\beta)}{\varepsilon m}}\log^{\frac{3}{4}}(1/\delta)\right)$ such that with probability at least $1-\beta/3$ over the randomness of Φ, for every feature vector x in the training set S, we have

$$\|x_\Phi\|_2^2 \le \left(1+\frac{\gamma}{3}\right)\|x\|_2^2 \tag{12}$$

$$1-\frac{y\langle w^*_\Phi, x_\Phi\rangle}{\rho} \le 1-\frac{y\langle w^*, x\rangle}{\rho} + \frac{r\gamma}{\rho} \tag{13}$$

We condition on this event for the remainder of the proof.

Note that (12) implies that

$$S_\Phi = \{(x_\Phi, y) : (x,y)\in S\};$$

that is, for all feature vectors x in the dataset S, $x_\Phi \in \mathbb{B}^k(2r)$ (i.e., $\Pi_{\mathbb{B}^k(2r)}(x_\Phi) = x_\Phi$).

Let $\mathcal{D}_\Phi$ denote the distribution of the pair (x_Φ, y), where $(x,y)\sim\mathcal{D}$. Via a standard margin bound [Mohri et al., 2018, Theorems 5.8 & 5.10], with probability at least $1-\beta/3$ over the choice of the training set S, we have

$$\forall w\in\mathbb{B}^k \qquad R_{\mathcal{D}_\Phi}(w) \le \widehat{L}^{\rho}_{S_\Phi}(w) + \frac{4r}{\rho\sqrt{m}} + 2\sqrt{\frac{\log(6/\beta)}{m}}$$

It follows that with probability at least $1-\beta/3$, we have

$$R_{\mathcal{D}_\Phi}(\tilde{w}) \le \widehat{L}^{\rho}_{S_\Phi}(\tilde{w}) + \frac{4r}{\rho\sqrt{m}} + 2\sqrt{\frac{\log(6/\beta)}{m}},$$

where $\tilde{w}$ is the output of step 4 of Algorithm 2. Moreover, note that

$$\begin{aligned} R_{\mathcal{D}}(w^{\mathsf{Priv}}) &= \underset{(x,y)\sim\mathcal{D}}{\mathbb{P}}\left[y\langle w^{\mathsf{Priv}}, x\rangle \le 0\right] \\ &= \underset{(x_\Phi,y)\sim\mathcal{D}_\Phi}{\mathbb{P}}\left[y\langle \tilde{w}, x_\Phi\rangle \le 0\right] \\ &= R_{\mathcal{D}_\Phi}(\tilde{w}) \end{aligned}$$

Thus, we get that with probability at least $1-\beta/3$,

$$R_{\mathcal{D}}(w^{\mathsf{Priv}}) \le \widehat{L}^{\rho}_{S_\Phi}(\tilde{w}) + \frac{4r}{\rho\sqrt{m}} + 2\sqrt{\frac{\log(6/\beta)}{m}}. \tag{14}$$

By Lemma 3.1, with probability at least $1-\beta/3$ over the randomness of Algorithm 4 (step 4 of Algorithm 2), we have that

$$\widehat{L}^{\rho}_{S_\Phi}(\tilde{w}) \le \widehat{L}^{\rho}_{S_\Phi}(\widehat{w}) + \frac{\Lambda r}{\rho}\left(\frac{1}{\sqrt{m}} + \frac{\sqrt{\log(\frac{m}{\beta})\log(\frac{1}{\beta})}\log^{\frac{3}{4}}(\frac{1}{\delta})}{\sqrt{\varepsilon m}}\right), \tag{15}$$

where $\widehat{w} \in \underset{w \in \mathbb{B}^k}{\operatorname{argmin}} \widehat{L}^{\rho}_{S_\Phi}(w)$. Moreover, (13) implies

$$\widehat{L}^{\rho}_{S_\Phi}(w^*_\Phi) \leq \widehat{L}^{\rho}_{S}(w^*) + \frac{r}{\rho} \cdot O\left(\sqrt{\frac{\log(1/\beta)}{\varepsilon m}} \log^{\frac{3}{4}}(1/\delta)\right)$$

Note that by definition of $\widehat{w}$, we have $\widehat{L}^{\rho}(\widehat{w}; S) \leq \widehat{L}^{\rho}_{S_\Phi}(w^*_\Phi)$. Hence, we have

$$\widehat{L}^{\rho}_{S_\Phi}(\widehat{w}) \leq \widehat{L}^{\rho}_{S}(w^*) + \frac{r}{\rho} O\left(\sqrt{\frac{\log(1/\beta)}{\varepsilon m}} \log^{\frac{3}{4}}(1/\delta)\right) \tag{16}$$

Now, by combining (14), (15), and (16), we reach the desired bound.

Finally, concerning the running time, observe that the Fast JL-transform (steps 2 and 3) takes $O(md \log(d) + \varepsilon m^2 \log(m)/\log^{3/2}(1/\delta))$ (follows from Lemma A.4), the DP-ERM algorithm (Algorithm 4) invoked in step 4 has $O(m)$ gradient steps; each of which takes involves $O(k + \log(m)) = O(\varepsilon m \log(m)/\log^{3/2}(1/\delta))$ operations. That is, the total number of operations of this step is $O(\varepsilon m^2 \log(m)/\log^{3/2}(1/\delta))$. Finally, the step 5 requires $O(dk) = O(\varepsilon dm \log(m)/\log^{3/2}(1/\delta))$. Thus, the overall running time is $O\Big(md \log(\max(d, m)) + \varepsilon m^2 \log(m)/\log^{\frac{3}{2}}(1/\delta)\Big)$.

□

C DP Algorithms for Kernel-Based Classification with Margin Guarantees

Theorem 4.2. *Let $r > 0$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a shift-invariant, positive definite kernel, where $K(x, x) = r^2$ for all $x \in \mathcal{X}$. For any $\varepsilon > 0$ and $\delta \in (0, 1)$, Algorithm 3 is (ε, δ)-differentially private. Define $\mathcal{H}_\Lambda \triangleq \{h \in \mathbb{H} : \|w\|_{\mathbb{H}} \leq \Lambda\}$, where $\|\cdot\|_{\mathbb{H}}$ is the norm corresponding to the reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ associated with the kernel K. Let $\beta \in (0, 1)$. Given an input sample S of m examples drawn i.i.d. from a distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$, Algorithm 3 outputs $h^{\widehat{\psi}}_{w^{\mathsf{Priv}}}$ such that with probability at least $1 - \beta$, we have*

$$R_{\mathcal{D}}(h^{\widehat{\psi}}_{w^{\mathsf{Priv}}}) \leq \min_{h \in \mathcal{H}_\Lambda} \widehat{L}^{\rho}_{S}(h) + \widetilde{O}\left(\frac{\Lambda r}{\rho \sqrt{\min(1, \varepsilon)\, m}}\right),$$

where, for any $h \in \mathcal{H}_\Lambda$, $\widehat{L}^{\rho}_{S}(h) \triangleq \frac{1}{m} \sum_{i=1}^{m} \ell^{\rho}\left(y_i \langle h, \psi(x_i)\rangle_{\mathbb{H}}\right)$, where ψ is the feature map associated with the kernel K and $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ is the inner product associated with the RKHS $\mathbb{H}$.

A more precise version of the above bound is given as follows:

$$R_{\mathcal{D}}(h^{\widehat{\psi}}_{w^{\mathsf{Priv}}}) \leq \min_{h \in \mathcal{H}_\Lambda} \widehat{L}^{\rho}_{S}(h) + O\left(\sqrt{\frac{\log(1/\beta)}{m}} + \frac{\Lambda r}{\rho}\left(\frac{1}{\sqrt{m}} + \frac{\sqrt{\log(\frac{m}{\beta}) \log(\frac{1}{\beta}) \log^{\frac{3}{4}}(\frac{1}{\delta})}}{\sqrt{\varepsilon m}}\right)\right),$$

Proof. Let $S_{\widehat{\psi}}$ be as defined in step 4 in Algorithm 3. Note that by Theorem 4.1, we have $\|\widehat{\psi}(x_i)\|_2 = r$ for all $i \in [m]$. Moreover, the output of Algorithm 3 depends only on $S_{\widehat{\psi}}$. Thus, the privacy guarantee follows directly from the privacy guarantee of Algorithm 2 (Theorem 3.2).

Next, we turn to proving the claimed margin bound. First, note that using the margin bound attained by Algorithm 2 (Theorem 3.2), it follows that for any fixed realization of the randomness in $\widehat{\psi}$, with probability $1 - \beta/2$ over the choice of $S \sim \mathcal{D}^m$ and the internal randomness of Algorithm 2, we have

$$R_{\mathcal{D}}(h^{\widehat{\psi}}_{w^{\mathsf{Priv}}}) \leq \min_{w \in \mathbb{B}^{2D}(2\Lambda)} \widehat{L}^{\rho}_{S_{\widehat{\psi}}}(w) + O\left(\sqrt{\frac{\log(1/\beta)}{m}} + \frac{\Lambda r}{\rho}\left(\frac{1}{\sqrt{m}} + \frac{\sqrt{\log(\frac{m}{\beta}) \log(\frac{1}{\beta}) \log^{\frac{3}{4}}(\frac{1}{\delta})}}{\sqrt{\varepsilon m}}\right)\right). \tag{17}$$

The essence of the proof is to show that with probability $\geq 1 - \beta/2$ over the randomness in $\widehat{\psi}$ (i.e., over the choice of $\omega_1, \ldots, \omega_D$), we have

$$\min_{w \in \mathbb{B}^{2D}(2\Lambda)} \widehat{L}^{\rho}_{S_{\widehat{\psi}}}(w) \leq \min_{h \in \mathcal{H}_\Lambda} \widehat{L}^{\rho}_{S}(h) + \frac{2\Lambda r}{\rho\sqrt{m}}. \tag{18}$$

Combining (17) and (18) yields the desired bound.

To prove the bound in (18), we will use the following fact. .

Fact C.1. *Let* $\mu > 0$*. Let* $\psi : \mathcal{X} \to \mathbb{R}$ *denote the feature map associated with the kernel* K*. Let* $h_\mu = \operatorname{argmin}_{h \in \mathcal{H}_\Lambda} \left(\widehat{L}^{\rho}_{S}(h) + \mu \|h\|^2_{\mathbb{H}}\right)$*. Then,* $h_\mu = \sum_{i=1}^{m} \alpha_i \psi(x_i)$ *for some* α_i*,* $i \in [m]$*, that satisfy:* $0 \leq y_i \alpha_i \leq \frac{1}{2\, m\, \mu\, \rho}$*.*

This fact simply follows from the dual formulation of the optimization problem for kernel support vector machines (see, e.g., [Mohri et al., 2018, Section 6.3]) . The fact asserts that the minimizer h_μ of the regularized empirical hinge loss can be expressed as a linear combination of $(\psi(x_i) : i \in [m])$ (such assertion also follows from the representer theorem) where the coefficients of the linear combination (the dual variables) $\alpha = (\alpha_1, \ldots, \alpha_m)$ are bounded; namely, $\|\alpha\|_1 \leq \frac{1}{2\mu\rho}$.

Below, we set $\mu = \frac{r}{\Lambda\rho\sqrt{m}}$. Let $\widehat{w} = \sum_{i=1}^{m} \alpha_i \widehat{\psi}(x_i)$ be a $2D$-dimensional approximation of h_μ. Observe that

$$\begin{aligned}
\widehat{L}^{\rho}_{S_{\widehat{\psi}}}(\widehat{w}) - \widehat{L}^{\rho}_{S}(h_\mu) &= \frac{1}{m}\sum_{i=1}^{m}\left[\ell^{\rho}(y_i\langle \widehat{w}, \widehat{\psi}(x_i)\rangle) - \ell^{\rho}(y_i\langle h_\mu, \psi(x_i)\rangle_{\mathbb{H}}\right] \\
&\leq \frac{1}{\rho m}\sum_{i,j \in [m]} |\alpha_j| |\langle \widehat{\psi}(x_i), \widehat{\psi}(x_j)\rangle - \langle \psi(x_i), \psi(x_j)\rangle_{\mathbb{H}}|
\end{aligned}$$

α_i where the inequality in the second line follows from the fact that ℓ^{ρ} is $\frac{1}{\rho}$-Lipschitz. Hence, by Theorem 4.1, with probability $\geq 1 - \beta/2$ with respect to the randomness in $\widehat{\psi}$, we have

$$\begin{aligned}
\widehat{L}^{\rho}_{S_{\widehat{\psi}}}(\widehat{w}) - \widehat{L}^{\rho}_{S}(h_\mu) &\leq \frac{2r^2}{\rho}\sqrt{\frac{\log(2m/\beta)}{D}}\|\alpha\|_1 \\
&\leq \frac{r^2}{\rho^2\mu}\sqrt{\frac{\log(2m/\beta)}{D}} \\
&\leq \frac{\Lambda r}{\rho\sqrt{m}},
\end{aligned}$$

where the second inequality follows from the fact that $\|\alpha\|_1 \leq \frac{1}{2\mu\rho}$, which follows from Fact C.1, and the third inequality follows from the setting of D in step 2 in Algorithm 3 and the setting of $\mu = \frac{r}{\Lambda\rho\sqrt{m}}$. Moreover, we note that $\widehat{w} \in \mathbb{B}^{2D}(2\Lambda)$. Indeed, conditioned on the same event above (the kernel matrix is well approximated via $\widehat{\psi}$), observe that

$$\begin{aligned}
\|\widehat{w}\|_2^2 &= \sum_{i,j}\alpha_i\alpha_j\langle \widehat{\psi}(x_i), \widehat{\psi}(x_j)\rangle \\
&\leq \sum_{i,j}\alpha_i\alpha_j\langle \psi(x_i), \psi(x_j)\rangle_{\mathbb{H}} + 2r^2\sqrt{\frac{\log(2m/\beta)}{D}}\|\alpha\|_1^2 \\
&\leq \|h\|^2_{\mathbb{H}} + \frac{\Lambda^2}{2} \leq \frac{3}{2}\Lambda^2.
\end{aligned}$$

Thus, we have $\|\widehat{w}\|_2 < 2\Lambda$. Hence, we can assert that with probability $\geq 1 - \beta/2$ over the randomness in $\widehat{\psi}$, we have

$$\min_{w \in \mathbb{B}^{2D}(2\Lambda)} \widehat{L}^{\rho}_{S_{\widehat{\psi}}}(w) \leq \widehat{L}^{\rho}_{S_{\widehat{\psi}}}(\widehat{w}) \leq \widehat{L}(h_\mu; S) + \frac{\Lambda r}{\rho\sqrt{m}}.$$

Finally, note that $\widehat{L}^{\rho}_{S}(h_\mu) \leq \min\limits_{h \in \mathcal{H}_\Lambda} \widehat{L}^{\rho}_{S}(h) + \mu\,\Lambda^2 = \min\limits_{h \in \mathcal{H}_\Lambda} \widehat{L}^{\rho}_{S}(h) + \frac{\Lambda r}{\rho\sqrt{m}}$. Hence, we arrive at the claimed bound (18), and thus, the proof is complete. □

D DP Algorithms for Learning Neural Networks with Margin Guarantees

Theorem 5.1. *Let $\varepsilon > 0, \beta \in (0,1)$, and $\rho > 0$. Then, there is an ε-DP algorithm which returns an L-layer network h^{Priv} with N neurons per layer that with probability at least $1-\beta$ over the draw of a sample $S \sim \mathcal{D}^m$ and the internal randomness of the algorithm admits the following guarantee:*

$$R_{\mathcal{D}}(h^{\mathsf{Priv}}) \leq \min_{h \in \mathcal{H}_{\mathsf{NN}^\Lambda}} \widehat{R}_S^{\rho}(h) + O\left(\frac{r(2\eta\Lambda)^L\sqrt{N\theta}}{\rho\sqrt{m}} + \frac{r^2(2\eta\Lambda)^{2L}N\theta}{\rho^2\varepsilon m}\right),$$

where $\theta = \log(Lm/\beta)\log(r(\eta\Lambda)^L/\rho)$.

Proof. First, note that our construction is indeed ε-DP by the properties of the exponential mechanism. Thus, we now turn to the proof of the margin bound. Our proof relies on the following properties of the JL-transform.

Lemma D.1 (Follows from Theorem 109 in [Nelson, 2010])**.** *Let $p, N, m, k \in \mathbb{N}$. Let $W \in \mathbb{R}^{p\times N}$. Let $z_1, \ldots, z_m \in \mathbb{R}^p$. Let Φ be a random $k \times p$ matrix with entries drawn i.i.d. uniformly from $\{\pm\frac{1}{\sqrt{k}}\}$. Let $\beta \in (0,1)$. There is a constant $c > 0$ such that the following inequalities hold simultaneously with probability at least $1-\beta$:*

$$\|\Phi W\|_F^2 \leq \|W\|_F^2\left(1 + c\sqrt{\frac{\log(m/\beta)}{k}}\right),$$

$$\forall i \in [m]: \ \|W^\top\Phi^\top\Phi z_i - W^\top z_i\|_2 \leq c\|W\|_F\|z_i\|_2\sqrt{\frac{\log(m/\beta)}{k}}$$

Consider the algorithmic construction described earlier. Let $h_* \in \operatorname*{argmin}_{h\in\mathcal{H}_{\mathsf{NN}}} \widehat{R}_S^{\rho}(h)$. Let $W_1^*, \ldots, W_L^*$ denote the weight matrices of h_*. Let $h_*^{\Phi} \in \mathcal{H}_{\mathsf{NN}}^{\Phi}$ be the network specified by the matrices $\widetilde{W}_1 \triangleq \Phi_0 W_1^*, \ldots, \widetilde{W}_L \triangleq \Phi_{L-1}W_L^*$. That is, the weight matrices of h_*^{Φ} are given by $\Phi_0^\top\widetilde{W}_1 = \Phi_0^\top\Phi_0 W_1^*, \ldots, \Phi_{L-1}^\top\widetilde{W}_L = \Phi_{L-1}^\top\Phi_{L-1}W_L^*$.

We make the following four claims. Combining those claims together with the union bound immediately yields the margin bound of the theorem. We first state those claims and then prove them.

Claim D.2. *There is a setting $k = O\Big(\frac{r^2(2\eta\Lambda)^{2L}\log(Lm/\beta)}{\rho^2}\Big)$ such that with probability $1-\beta/4$ over the choice of $\Phi_0, \ldots, \Phi_{L-1}$, we have $h_*^{\Phi} \in \mathcal{H}_{\mathsf{NN}^{2\Lambda}}^{\Phi}$ and for all $i \in [m]$*

$$|h_*(x_i) - h_*^{\Phi}(x_i)| = O\left(r(2\eta\Lambda)^L\sqrt{\frac{\log(Lm/\beta)}{k}}\right).$$

Consequently, with probability $1-\beta/4$,

$$\widehat{R}_S^{0.5\rho}(h_*^{\Phi}) \leq \widehat{R}_S^{\rho}(h_*).$$

Claim D.3. *Let $\widehat{h}^{\Phi} \in \operatorname*{argmin}_{h\in\widehat{\mathcal{H}}_{\mathsf{NN}^{2\Lambda}}^{\Phi}} \widehat{R}_S(h)$. There exists a setting $k = O\Big(\frac{r^2(2\eta\Lambda)^{2L}\log(Lm/\beta)}{\rho^2}\Big)$ for the embedding parameter such that with probability $1-\beta/4$*

$$\widehat{R}_S(\widehat{h}^{\Phi}) \leq \widehat{R}_S^{0.5\rho}(h_*^{\Phi}).$$

Claim D.4. *Let $\widehat{h}^{\Phi} \in \operatorname*{argmin}_{h\in\widehat{\mathcal{H}}_{\mathsf{NN}^{2\Lambda}}^{\Phi}} \widehat{R}_S(h)$. Let $k = O\Big(\frac{r^2(2\eta\Lambda)^{2L}\log(Lm/\beta)}{\rho^2}\Big)$. With probability $1-\beta/4$ over the randomness of the exponential mechanism, we have*

$$\widehat{R}_S(h^{\mathsf{Priv}}) \leq \widehat{R}_S(\widehat{h}^{\Phi}) + O\left(\frac{r^2(2\eta\Lambda)^{2L}N\log(Lm/\beta)\log(r(4\eta\Lambda)^L/\rho)}{\rho^2\varepsilon m}\right).$$

Claim D.5. *Let $k = O\Big(\frac{r^2(2\eta\Lambda)^{2L}\log(Lm/\beta)}{\rho^2}\Big)$. With probability $1 - \beta/4$ over the choice of $S \sim \mathcal{D}^m$, we have*

$$R_{\mathcal{D}}(h^{\mathsf{Priv}}) \le \widehat{R}_S(h^{\mathsf{Priv}}) + O\left(\frac{r(2\eta\Lambda)^L\sqrt{N\log(Lm/\beta)\log(r(\eta\Lambda)^L/\rho)}}{\rho\sqrt{m}}\right).$$

Recall that $\widehat{\mathcal{H}}^{\Phi}_{\mathsf{NN}^{2\Lambda}}$ is a finite approximation of $\mathcal{H}^{\Phi}_{\mathsf{NN}^{2\Lambda}}$ constructed via a γ-cover $\mathcal{C}$ for $\mathbb{B}^{k\times N}(2\Lambda) \times \ldots \times \mathbb{B}^{k}(2\Lambda)$, where we choose $\gamma = \frac{\rho}{10r(4\eta\Lambda)^{L-1}}$. In particular, for any $W = (W_1, \ldots, W_L), W' = (W'_1, \ldots, W'_L) \in \mathcal{C}$, we have $\|W - W'\|_F = \sqrt{\sum_{j=1}^{L}\|W_j - W'_j\|_F^2} \le \gamma$. Given that $\mathcal{C}$ is a γ-cover, we have $|\widehat{\mathcal{H}}^{\Phi}_{\mathsf{NN}^{2\Lambda}}| = |\mathcal{C}| = O\Big(\big(\frac{\sqrt{L}\Lambda}{\gamma}\big)^{k\times N}\Big)$. Namely, $\log(|\widehat{\mathcal{H}}^{\Phi}_{\mathsf{NN}^{2\Lambda}}|) = O\Big(kN\log(\frac{r(\eta\Lambda)^L}{\rho})\Big)$. Given this, together with the setting of k in Claim D.5 and the fact that h^{Priv} is in $\widehat{\mathcal{H}}^{\Phi}_{\mathsf{NN}^{2\Lambda}}$, note that Claim D.5 follows from a straightforward uniform convergence bound for the hypotheses in $\widehat{\mathcal{H}}^{\Phi}_{\mathsf{NN}^{2\Lambda}}$. Note also that the proof of Claim D.4 follows directly from the standard accuracy guarantee of the exponential mechanism when instantiated on $\widehat{\mathcal{H}}^{\Phi}_{\mathsf{NN}^{2\Lambda}}$. In particular, since the score function is $-\widehat{R}_S(\cdot)$, with probability at least $1 - \beta/4$, the excess empirical loss of h^{Priv} is bounded by $O\Big(\frac{\log(|\mathcal{C}|/\beta)}{\varepsilon m}\Big)$, which yields the bound claimed in Claim D.4 given the bound on $|\mathcal{C}|$ above and the setting of k.

We now turn to the proofs of Claims D.2 and D.3. We start with the proof of Claim D.2.

For each $i \in [m]$ and each $j \in [L]$, let $v_{i,j} \in \mathbb{R}^N$ denote the output of the j-th layer of h_* on input x_i prior to activation (i.e., $v_{i,j}$ is the input to the neurons of layer $j+1$ when the input to the network h_* is the i-th feature vector x_i in the dataset S). Analogously, for each $i \in [m]$ and each $j \in [L]$, let $v^{\Phi}_{i,j}$ denote the output of the j-th layer of h^{Φ}_* on input x_i prior to activation. Also, let $u_{i,j} \triangleq \psi(v_{i,j}) - \psi(v^{\Phi}_{i,j}), i \in [m], j \in [L]$.

As a direct corollary of Lemma D.1, by applying the union bound over the choice of $\Phi_0, \ldots, \Phi_{L-1}$, there is a constant $\hat{c} > 0$ such that with probability $1 - \beta/4$ over the choice of $\Phi_0, \ldots, \Phi_{L-1}$, for all $i \in [m], j \in [L]$, we have

$$\|\Phi_{j-1}W^*_j\|_F^2 \le \Lambda^2\left(1 + \hat{c}\sqrt{\frac{\log(Lm/\beta)}{k}}\right), \tag{19}$$

$$\|(W^*_j)^\intercal\Phi^\intercal_{j-1}\Phi_{j-1}\psi(v_{i,j-1}) - (W^*_j)^\intercal\psi(v_{i,j-1})\|_2 \le \hat{c}\Lambda\|\psi(v_{i,j-1})\|_2\sqrt{\frac{\log(Lm/\beta)}{k}}, \quad j \neq 1, \tag{20}$$

$$\|(W^*_j)^\intercal\Phi^\intercal_{j-1}\Phi_{j-1}u_{i,j-1} - (W^*_j)^\intercal u_{i,j-1}\|_2 \le \hat{c}\Lambda\|u_{i,j-1}\|_2\sqrt{\frac{\log(Lm/\beta)}{k}}, \quad j \neq 1, \tag{21}$$

$$\|(W^*_1)^\intercal\Phi^\intercal_0\Phi_0 x_i - (W^*_1)^\intercal x_i\|_2 \le \hat{c}\Lambda r\sqrt{\frac{\log(Lm/\beta)}{k}} \tag{22}$$

We now condition on the event where all the above inequalities are satisfied for the remainder of the proof. Below, we let $\tau = \hat{c}\sqrt{\frac{\log(Lm/\beta)}{k}} < 1$. First, from (19), there is a setting k as indicated in the statement of the claim, where $\|\Phi_{j-1}W^*_j\|_F < 2\Lambda$. Thus, $h^{\Phi}_* \in \mathcal{H}^{\Phi}_{\mathsf{NN}^{2\Lambda}}$.

Now, fix any $i \in [m]$. Define $\Gamma_j \triangleq \|(W_j^*)^\top \psi(v_{i,j-1}) - (W_j^*)^\top \Phi_{j-1}^\top \Phi_{j-1} \psi(v_{i,j-1}^\Phi)\|_2$ for $j \in [L]$. Observe that

$$\begin{aligned}
&|h_*(x_i) - h_*^\Phi(x_i)| = \Gamma_L \\
=&|(W_L^*)^\top \psi(v_{i,L-1}) - (W_L^*)^\top \Phi_{L-1}^\top \Phi_{L-1} \psi(v_{i,L-1}^\Phi)| \\
\le&|(W_L^*)^\top \psi(v_{i,L-1}) - (W_L^*)^\top \Phi_{L-1}^\top \Phi_{L-1} \psi(v_{i,L-1})| \\
&+ |(W_L^*)^\top \Phi_{L-1}^\top \Phi_{L-1} \psi(v_{i,L-1}) - (W_L^*)^\top \Phi_{L-1}^\top \Phi_{L-1} \psi(v_{i,L-1}^\Phi)| \\
\le&\tau\Lambda \|\psi(v_{i,L-1})\|_2 + |(W_L^*)^\top \Phi_{L-1}^\top \Phi_{L-1} \left(\psi(v_{i,L-1}) - \psi(v_{i,L-1}^\Phi)\right)| && \Big(\text{follows from (20) and the fact } W_L^* \in \mathbb{B}^N(\Lambda)\Big) \\
=&\tau\Lambda \|\psi(v_{i,L-1})\|_2 + |(W_L^*)^\top \Phi_{L-1}^\top \Phi_{L-1} u_{i,L-1}| && \Big(\text{by definition of } u_{i,L-1} \text{ given above}\Big) \\
\le&\tau\Lambda \|\psi(v_{i,L-1})\|_2 + (1+\tau)\Lambda \|u_{i,L-1}\|_2 && \Big(\text{follows from (21)}\Big) \\
\le&\tau\Lambda \|\psi(v_{i,L-1})\|_2 + 2\Lambda \|\psi(v_{i,L-1}) - \psi(v_{i,L-1}^\Phi)\|_2 \\
\le&\tau\eta\Lambda \|v_{i,L-1}\|_2 + 2\eta\Lambda \|v_{i,L-1} - v_{i,L-1}^\Phi\|_2 && \Big(\text{since } \psi \text{ is } \eta\text{-Lipschitz and } \psi(0) = 0\Big) \\
=&\tau\eta\Lambda \|v_{i,L-1}\|_2 + 2\eta\Lambda \|(W_{L-1}^*)^\top \psi(v_{i,L-2}) - (W_{L-1}^*)^\top \Phi_{L-2}^\top \Phi_{L-2} \psi(v_{i,L-2}^\Phi)\|_2 \\
=&\tau\eta\Lambda \|v_{i,L-1}\|_2 + 2\eta\Lambda \Gamma_{L-1}
\end{aligned}$$

Hence, we obtain $\Gamma_L \le \tau\eta\Lambda \|v_{i,L-1}\|_2 + 2\eta\Lambda\Gamma_{L-1}$. Before we solve this recurrence, we first unravel the term $\|v_{i,L-1}\|_2$. Note that

$$\begin{aligned}
\|v_{i,L-1}\|_2 &= \|(W_{L-1}^*)^\top \psi(v_{i,L-2})\|_2 \\
&\le \|W_{L-1}^*\|_F \cdot \|\psi(v_{i,L-2})\|_2 \\
&\le \eta\Lambda \|v_{i,L-2}\|_2
\end{aligned}$$

Proceeding recursively, we obtain

$$\|v_{i,L-1}\|_2 \le \eta^{L-2}\Lambda^{L-1} \|x_i\|_2 \le r\eta^{L-2}\Lambda^{L-1}.$$

Plugging this in the recurrence for Γ_L above yields

$$\Gamma_L \le \tau r \eta^{L-1} \Lambda^L + 2\eta\Lambda\Gamma_{L-1}.$$

Unraveling this recursion (and using (22) in the last step of the recursion) yields

$$|h_*(x_i) - h_*^\Phi(x_i)| = \Gamma_L \le r(2\eta\Lambda)^{L-1}\Lambda\,\tau = \hat{c}\sqrt{\frac{\log(Lm/\beta)}{k}}\, r(2\eta\Lambda)^{L-1}\Lambda.$$

Note that choosing $k = \frac{10\hat{c}^2 r^2 (2\eta\Lambda)^{2(L-1)} \Lambda^2 \log(Lm/\beta)}{\rho^2}$ guarantees $|h_*(x_i) - h_*^\Phi(x_i)| < \frac{\rho}{2}$ for all $i \in [m]$. Hence, as in the argument of the proof of Theorem 3.1, this implies that for all $i \in [m]$, $y_i h_*(x_i) > \rho \Rightarrow y_i h_*^\Phi(x_i) > \frac{\rho}{2}$. Thus, $\widehat{R}_S^{0.5\rho}(h_*^\Phi) \le \widehat{R}_S^\rho(h_*)$. This concludes the proof of Claim D.2.

Finally, we prove Claim D.3.

As shown in Claim D.2, we have $h_*^\Phi \in \mathcal{H}_{\mathsf{NN}^{2\Lambda}}^\Phi$. Since $\widehat{\mathcal{H}}_{\mathsf{NN}^{2\Lambda}}^\Phi$ is a γ-cover of $\mathcal{H}_{\mathsf{NN}^{2\Lambda}}^\Phi$, there exists $\tilde{h} \in \widehat{\mathcal{H}}_{\mathsf{NN}^{2\Lambda}}^\Phi$ that "approximates" h_*^Φ. Namely, there is $\tilde{h} \in \widehat{\mathcal{H}}_{\mathsf{NN}^{2\Lambda}}^\Phi$ defined by matrices $(\widetilde{W}_1, \ldots, \widetilde{W}_L) \in \mathcal{C}$ such that

$$\sum_{j=1}^L \|\widetilde{W}_j - \Phi_{j-1} W_j^*\|_F^2 \le \gamma^2,$$

where, as defined before, $(\Phi_0 W_1^*, \ldots, \Phi_{L-1} W_L^*)$ are the matrices defining h_*^Φ. We choose $\gamma = \frac{\rho}{10r(4\eta\Lambda)^{L-1}}$.

To simplify notation, we will denote

$$W_j^{\Phi,*} \triangleq \Phi_{j-1} W_j^*, \ \forall\, j \in [L].$$

As before, for each $i \in [m], j \in [L]$, we let $v_{i,j}^\Phi \in \mathbb{R}^N$ denote the output of the j-th layer of h_*^Φ on input x_i prior to activation, and let $\tilde{v}_{i,j}$ denote the output of the j-th layer of $\tilde{h}$ on input x_i prior to activation.

Again, as a corollary of Lemma D.1 (by applying the union bound over the choice of $\Phi_0, \dots, \Phi_{L-1}$), there is a constant $\hat{c} > 0$ such that with probability $1 - \beta/4$ over the choice of $\Phi_0, \dots, \Phi_{L-1}$, for all $i \in [m], j \in [L]$, we have

$$\|\Phi_{j-1}\psi(\tilde{v}_{i,j-1})\|_2^2 \leq \|\psi(\tilde{v}_{i,j-1})\|_2^2 \left(1 + \hat{c}\sqrt{\frac{\log(Lm/\beta)}{k}}\right), \;\; j \neq 1 \tag{23}$$

$$\|\Phi_{j-1}\left(\psi(\tilde{v}_{i,j-1}) - \psi(v^{\Phi}_{i,j-1})\right)\|_2^2 \leq \|\psi(\tilde{v}_{i,j-1}) - \psi(v^{\Phi}_{i,j-1})\|_2^2 \left(1 + \hat{c}\sqrt{\frac{\log(Lm/\beta)}{k}}\right), \;\; j \neq 1 \tag{24}$$

$$\|\Phi_0 x_i\|_2^2 \leq r^2 \left(1 + \hat{c}\sqrt{\frac{\log(Lm/\beta)}{k}}\right) \tag{25}$$

We will condition on the event above for the remainder of the proof. Note that for the setting of k as in Claim D.2, we have $\left(1 + \hat{c}\sqrt{\frac{\log(Lm/\beta)}{k}}\right) < 2$.

For each $j \in [L]$, define

$$\Delta_j \triangleq \|\widetilde{W}_j^\top \Phi_{j-1}\psi(\tilde{v}_{i,j-1}) - W_j^{\Phi,*}\Phi_{j-1}\psi(v^{\Phi}_{i,j-1})\|_2.$$

Fix any $i \in [m]$. Observe

$$\begin{aligned}
&|\tilde{h}(x_i) - h^{\Phi}_*(x_i)| = \Delta_L \\
\leq&|\widetilde{W}_L^\top \Phi_{L-1}\psi(\tilde{v}_{i,L-1}) - (W_L^{\Phi,*})^\top \Phi_{L-1}\psi(\tilde{v}_{i,L-1})| \\
&+ |(W_L^{\Phi,*})^\top \Phi_{L-1}\psi(\tilde{v}_{i,L-1}) - \widetilde{(}W_L^{\Phi,*})^\top \Phi_{L-1}\psi(v^{\Phi}_{i,L-1})| \\
\leq&\|\widetilde{W}_L - W_L^{\Phi,*}\|_F \|\Phi_{L-1}\psi(\tilde{v}_{i,L-1})\|_2 + \|W_L^{\Phi,*}\|_F \|\Phi_{L-1}\left(\psi(\tilde{v}_{i,L-1}) - \psi(v^{\Phi}_{i,L-1})\right)\|_2 \\
\leq&\gamma\, \|\Phi_{L-1}\psi(\tilde{v}_{i,L-1})\|_2 + 2\Lambda\, \|\Phi_{L-1}\left(\psi(\tilde{v}_{i,L-1}) - \psi(v^{\Phi}_{i,L-1})\right)\|_2 \quad \left(\mathcal{C} \text{ is } \gamma\text{-cover and } W_L^{\Phi,*} \in \mathbb{B}^{k\times N}(2\Lambda)\right) \\
\leq&\sqrt{2}\,\gamma\, \|\psi(\tilde{v}_{i,L-1})\|_2 + 2\sqrt{2}\Lambda\, \|\psi(\tilde{v}_{i,L-1}) - \psi(v^{\Phi}_{i,L-1})\|_2 \qquad \left(\text{follows from (23)-(24)}\right) \\
\leq&\sqrt{2}\,\gamma\eta\, \|\tilde{v}_{i,L-1}\|_2 + 2\sqrt{2}\,\eta\Lambda\, \|\tilde{v}_{i,L-1} - v^{\Phi}_{i,L-1}\|_2 \qquad \left(\psi \text{ is } \eta\text{-Lipschitz and } \psi(0) = 0\right) \\
\leq&\sqrt{2}\,\gamma\eta\, \|\widetilde{W}_{L-1}^\top \Phi_{L-2}\psi(\tilde{v}_{i,L-2})\|_2 + 2\sqrt{2}\,\eta\Lambda\, \|\widetilde{W}_{L-1}^\top \Phi_{L-2}\psi(\tilde{v}_{i,L-2}) - W_{L-1}^{\Phi,*}\Phi_{L-2}\psi(v^{\Phi}_{i,L-2})\|_2 \\
=&\sqrt{2}\,\gamma\eta\, \|\widetilde{W}_{L-1}^\top \Phi_{L-2}\psi(\tilde{v}_{i,L-2})\|_2 + 2\sqrt{2}\,\eta\Lambda\, \Delta_{L-1}
\end{aligned}$$

Hence, we arrive at a recursive bound

$$\Delta_L \leq \sqrt{2}\,\gamma\eta\, \|\widetilde{W}_{L-1}^\top \Phi_{L-2}\psi(\tilde{v}_{i,L-2})\|_2 + 2\sqrt{2}\,\eta\Lambda\, \Delta_{L-1}.$$

Before proceeding, we first unravel the term $\|\widetilde{W}_{L-1}^\top \Phi_{L-2}\psi(\tilde{v}_{i,L-2})\|_2$. Let's denote this term as B_{L-1}. Observe that

$$\begin{aligned}
B_{L-1} =&\|\widetilde{W}_{L-1}^\top \Phi_{L-2}\psi(\tilde{v}_{i,L-2})\|_2 \\
\leq&\|\widetilde{W}_{L-1}\|_F \|\Phi_{L-2}\psi(\tilde{v}_{i,L-2})\|_2 \\
\leq&2\sqrt{2}\Lambda\|\psi(\tilde{v}_{i,L-2})\|_2 \\
\leq&2\sqrt{2}\eta\Lambda\|\tilde{v}_{i,L-2}\|_2 \\
=&2\sqrt{2}\eta\Lambda\|\widetilde{W}_{L-2}^\top \Phi_{L-3}\psi(\tilde{v}_{i,L-3})\|_2 \\
=&2\sqrt{2}\eta\Lambda B_{L-2}
\end{aligned}$$

Thus, continuing recursively, we get $B_{L-1} \leq r\eta^{L-2}(2\sqrt{2}\Lambda)^{L-1}$ (where in the last step of the recursion, we use (25)). Plugging this back in the recursive bound for Δ_L, we get

$$\Delta_L \leq \sqrt{2}\gamma r(2\sqrt{2}\eta\Lambda)^{L-1} + 2\sqrt{2}\,\eta\,\Lambda\,\Delta_{L-1}$$

Unraveling this recurrence yields

$$\begin{aligned}
\Delta_L \leq&\sqrt{2}\gamma r L(2\sqrt{2}\eta\Lambda)^{L-1} \\
\leq&2\sqrt{2}\gamma r(4\eta\Lambda)^{L-1}
\end{aligned}$$

Algorithm 5 $\mathcal{A}_{\mathsf{LabMarg}}$: Private learning algorithm under label-privacy

```
Require: Dataset $S = ((x_1, y_1), \dots, (x_m, y_m)) \in \left(B^d \times \{\pm 1\}\right)^m$; privacy parameter $\varepsilon > 0$; margin parameter $\rho$.
Compute the $\rho/2$ minimal cover $\widehat{\mathcal{H}}_\rho = \mathcal{C}(\mathcal{H}_\rho, \rho/2, x_1^m)$.
Run the Exponential mechanism with privacy parameter $\varepsilon$, sensitivity $1/m$, and score function $-\widehat{R}_S^{\rho/2}(h),\ h \in \widehat{\mathcal{H}}_\rho$ to select $h^{\mathsf{priv}} \in \widehat{\mathcal{H}}_\rho$.
return $h^{\mathsf{priv}}$.
```

Thus, by the choice of γ, we have $|\tilde{h}(x_i) - h_*^\Phi(x_i)| < \frac{\rho}{2}$ for all $i \in [m]$. Hence, as before, we have $\left(y_i h_*^\Phi(x_i) > \rho/2\right) \Rightarrow \left(y_i \tilde{h}(x_i) > 0\right)$ for all $i \in [m]$, which implies that

$$\widehat{R}_S(\tilde{h}) \le \widehat{R}_S^{0.5\rho}(h_*^\Phi).$$

Since $\widehat{h}^\Phi \in \operatorname*{argmin}_{h \in \widehat{\mathcal{H}}^\Phi_{\mathsf{NN}2\Lambda}} \widehat{R}_S(h)$, then we have $\widehat{R}_S(\widehat{h}^\Phi) \le R_S(\tilde{h})$. Therefore, we can write

$$\widehat{R}_S(\widehat{h}^\Phi) \le \widehat{R}_S^{0.5\rho}(h_*^\Phi).$$

This concludes the proof of Claim D.3 and completes the proof of Theorem 5.1. □

E Label-Private Algorithms with Margin Guarantees

In many tasks, the features are public information and only the labels are sensitive and need to be protected. Several recent publications have suggested to train learning models with differential privacy for labels for these tasks, while treating features as public information [Ghazi et al., 2021, Esfandiari et al., 2021]. This motivates the following definition of label differential privacy.

Definition E.1 (Label differential privacy)**.** *Let $\varepsilon, \delta \ge 0$. Let $\mathcal{A}\colon (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$ be a (potentially randomized) mechanism. We say that $\mathcal{A}$ is (ε, δ)-*label-DP *if for any measurable subset $O \subset \mathcal{H}$ and all $S, S' \in (\mathcal{X} \times \mathcal{Y})^m$ that differ in one label of one sample, the following inequality holds:*

$$\mathbb{P}(\mathcal{A}(S) \in O) \le e^\varepsilon\, \mathbb{P}(\mathcal{A}(S') \in O) + \delta. \tag{26}$$

Ghazi et al. [2021] gave an algorithm for deep learning with label differential privacy in the local differential privacy model. Yuan et al. [2021] proposed and evaluated algorithms for label differential privacy in conjunction with secure multiparty computation. Esfandiari et al. [2021] presented a clustering-based algorithm for label differential privacy. There are several other works which show pitfalls on label differential privacy [Busa-Fekete et al., 2021a,b].

Here, we design a simple algorithm for label differential privacy, which we show benefits from margin guarantees for any hypotheses class with finite fat-shattering dimension, including the class of linear classifiers, neural networks, and ensembles [Bartlett and Shawe-Taylor, 1999].

We first introduce some definitions needed to describe our algorithm. Fix $\rho > 0$. Define the ρ-truncation function $\beta_\rho\colon \mathbb{R} \to [-\rho, +\rho]$ by $\beta_\rho(u) = \max\{u, -\rho\} 1_{u \le 0} + \min\{u, +\rho\} 1_{u \ge 0}$, for all $u \in \mathbb{R}$. For any $h \in \mathcal{H}$, we denote by h_ρ the ρ-truncation of h, $h_\rho = \beta_\rho(h)$, and define $\mathcal{H}_\rho = \{h_\rho\colon h \in \mathcal{H}\}$. For any family of functions $\mathcal{F}$, we also denote by $\mathcal{N}_\infty(\mathcal{F}, \varepsilon, x_1^m)$ the empirical covering number of $\mathcal{F}$ over the sample $(x_1, \dots, x_m)$ and by $\mathcal{C}(\mathcal{F}, \varepsilon, x_1^m)$ a minimum empirical cover. With these definitions, the algorithm is given in Algorithm 5. The algorithm uses an exponential mechanism over a cover of truncated hypotheses sets.

Theorem E.1. *Algorithm 5 is ε-label-DP. Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$ and suppose $S \sim \mathcal{D}^m$. Let $c = 17$ and $d = \operatorname{fat}_{\frac{\rho}{32}}(\mathcal{H})$ and $M = 1 + d \log_2(2c^2 m) \log_2 \frac{2cem}{d} + \log \frac{2}{\beta}$. For any $\beta \in (0, 1)$, with probability at least $1 - \beta$, the output w^{Priv} satisfies:*

$$R_{\mathcal{D}}(h^{\mathsf{priv}}) \le \min_{h \in \mathcal{H}} \left(\widehat{R}_S^\rho(h) + 2\sqrt{\min_{h \in \mathcal{H}} \widehat{R}_S^\rho(h)} \sqrt{\frac{M}{m}} \right) + \frac{2M}{m} + \frac{64 M \log\left(\frac{2}{\beta}\right)}{\varepsilon m}.$$

Before we prove the above theorem, we first want to remark that while our algorithm is computationally inefficient, it admits strong theoretical guarantees. First, it is an $(\varepsilon, 0)$ pure label-differential privacy guarantee. Second, it is dimension-independent. Furthermore, our algorithm benefits from a relative deviation margin bound that smoothly interpolates between the realizable case of $R^\rho_S(w) = 0$ and the case of $R^\rho_S(w) > 0$. As a corollary, note that up to constants one can always get privacy for $\varepsilon > 1$ for free. Finally, observe that this bound holds not only for linear classes, but also for any hypothesis set with favorable ρ-fat-shattering dimension. In particular, we can use known upper bounds for the ρ-fat-shattering dimension of feed-forward neural networks [Bartlett and Shawe-Taylor, 1999] to derive label-privacy guarantees for training neural networks.

We now prove Theorem E.1.

Proof. The ε-differential privacy guarantee follows directly from the properties of the exponential mechanism. In particular, given the finite class $\widehat{\mathcal{H}}_\rho$ and the score function $-\widehat{R}^{\rho/2}_S(h)$, $h \in \widehat{\mathcal{H}}_\rho$, the algorithm becomes an instantiation of the exponential mechanism [McSherry and Talwar, 2007].

We focus on proving the utility guarantee in the rest of the proof. If $m < \frac{64M\log(2/\beta)}{\varepsilon}$, then the bound follows trivially. Hence in the rest of the paper, we focus on the regime $m \geq \frac{64M\log(2/\beta)}{\varepsilon}$. By definition of $\widehat{\mathcal{H}}_\rho$, for any $h \in \mathcal{H}$ there exists $g \in \widehat{\mathcal{H}}_\rho$ such that for any $x \in x_1^m$,

$$|g(x) - h(x)| \leq \frac{\rho}{2}.$$

Thus, for any $y \in \{-1, +1\}$ and $x \in x_1^m$, we have $|yg(x) - yh(x)| \leq \rho/2$, which implies:

$$1_{yg(x)\leq\rho/2} \leq 1_{yh(x)\leq\rho}.$$

Let $h^*_S \in \operatorname*{argmin}_{h\in\mathcal{H}} \widehat{R}^\rho_S(h)$. By the construction of $\widehat{\mathcal{H}}_\rho$ and the above argument,

$$\min_{h\in\widehat{\mathcal{H}}_\rho} \widehat{R}^{\rho/2}_S(h) \leq \widehat{R}^\rho_S(h^*_S). \tag{27}$$

We now bound the size $\widehat{\mathcal{H}}_\rho$.

$$|\widehat{\mathcal{H}}_\rho| = \mathcal{N}_\infty(\mathcal{H}_\rho, \rho/2, x_1^m).$$

By [Bartlett, 1998, Proof of theorem 2], we have

$$\log \max_{x_1^m}[\mathcal{N}_\infty(\mathcal{H}_\rho, \tfrac{\rho}{2}, x_1^m)] \leq 1 + d' \log_2(2c^2 m) \log_2 \frac{2cem}{d'},$$

where $d' = \text{fat}_{\frac{\rho}{32}}(\mathcal{H}_\rho) \leq \text{fat}_{\frac{\rho}{32}}(\mathcal{H}) = d$ and $c = 17$. Given the bound on the sample size in the theorem statement and the properties of the exponential mechanism [McSherry and Talwar, 2007], value of m, with probability at least $1 - \beta/2$,

$$\begin{aligned}\widehat{R}^{\rho/2}_S(h^{\mathsf{priv}}) &\leq \min_{h\in\widehat{\mathcal{H}}_\rho} \widehat{R}^\rho_S(h) + \frac{32M\log(2/\beta)}{\varepsilon m} \\ &\leq \widehat{R}^\rho_S(h^*_S) + \frac{32M\log(2/\beta)}{\varepsilon m}.\end{aligned} \tag{28}$$

By Lemma A.2, with probability at least $1 - \beta/2$,

$$R_{\mathcal{D}}(h^{\mathsf{priv}}) \leq \widehat{R}^{\rho/2}_S(h^{\mathsf{priv}}) + 2\sqrt{\widehat{R}^{\rho/2}_S(h)\frac{M}{m}} + \frac{M}{m}, \tag{29}$$

where $M = 1 + d\log_2(2c^2m)\log_2\frac{2cem}{d} + \log\frac{2}{\beta}$, $c = 17$, and $d = \text{fat}_{\frac{\rho}{32}}(\mathcal{H})$. Combining (28) and (29) yields

$$R_{\mathcal{D}}(h^{\mathsf{priv}}) \leq \widehat{R}^\rho_S(h^*_S) + 2\sqrt{\widehat{R}^\rho_S(h^*_S)\frac{M}{m}} + \frac{2M}{m} + \frac{64M\log(2/\beta)}{\varepsilon m}.$$

The lemmas follows by observing that if $h^* \in \operatorname*{argmin}_{h\in\mathcal{H}} \widehat{R}^\rho_S(h)$, if and only if $h^* \in \operatorname*{argmin}_{h\in\mathcal{H}} \widehat{R}^\rho_S(h) + 2\sqrt{\widehat{R}^\rho_S(h)\frac{M}{m}}$. □

Algorithm 6 $\mathcal{A}_{\mathsf{PrivMrg}}$: Algorithm to select confidence margin

```
Require: Dataset $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \{\pm 1\})^m$; algorithm $\mathcal{A}$; bound $F(\rho, g_\rho(S))$;
    $\mathrm{h}_{\max}$ an upper bound on $\rho$; privacy parameters $\varepsilon > 0, \delta \geq 0$; and confidence parameter $\beta > 0$.
1: Let $\mathcal{V} \triangleq \left\{\rho_j \triangleq 2^{-j}\,\mathrm{h}_{\max} : j \in [J]\right\}$, where $J = \frac{1}{2}\log(m)$.
2: Run the generalized exponential mechanism [Raskhodnikova and Smith, 2016, Algorithm 1]
   over $\mathcal{V}$ with privacy parameter $\varepsilon$ and score function $-F(\rho_j; g_{\rho_j}(S))$ for $\rho_j \in \mathcal{V}$, to select $\rho^* \in \mathcal{V}$.
3: Run $\mathcal{A}$ on the dataset $S$ with margin parameter $\rho^*$ and privacy parameters $(\varepsilon, \delta)$, confidence
   parameter $\beta$, and return its output $w^{\mathsf{Priv}}$.
```

F Confidence Margin Parameter Selection

The algorithms of Sections 3, 4, 5 and Appendix E can all be augmented to include the selection of the confidence margin parameter ρ by using an exponential mechanism. All of the proposed algorithms in the previous sections output w^{Priv} such that

$$R_{\mathcal{D}}(w^{\mathsf{Priv}}) \leq F(\rho, g_\rho(S)),$$

where $g_\rho(S)$ is either the minimum ρ-margin loss or the minimum ρ-hinge loss. Furthermore, in all our results, for any fixed t, $F(\rho, t)$ is a non-increasing function of ρ and $g_\rho(S)$ is a non-decreasing function of ρ for any S. Suppose we have an algorithm $\mathcal{A}$ such that the above inequality holds. We can then augment it with an exponential mechanism algorithm to select a near-optimal margin ρ. Let $\mathrm{h}_{\max}$ be an upper bound on $\max_{x \in \mathcal{X}} \max_{h \in \mathcal{H}} |h(x)|$. For example, $\mathrm{h}_{\max} = \Lambda r$ for linear classifiers. If $\mathrm{h}_{\max} > \rho$, then the bound $F(\rho; g_\rho(S))$ becomes trivial (i.e., $\Omega(1)$). Similarly, the bound typically becomes trivial when $\rho \lesssim \frac{\mathrm{h}_{\max}}{\sqrt{m}}$. It is easy to see this property for linear classifiers, for other models such as neural networks with label privacy it can be obtained by bounds on fat-shattering dimension [Bartlett and Shawe-Taylor, 1999]. Hence, without loss of optimality, we will seek an approximation for ρ_{opt} that minimizes $F(\rho; g_\rho(S))$ for $\rho \in \left[\frac{\mathrm{h}_{\max}}{\sqrt{m}}, \mathrm{h}_{\max}\right]$. To do this, we define a finite grid over the above interval: $\mathcal{V} \triangleq \left\{\rho_j \triangleq 2^{-j}\,\mathrm{h}_{\max}\; j \in [J]\right\}$, where $J = \frac{1}{2}\log(m)$. We use an instantiation of the generalized exponential mechanism, with score function $-F(\rho; g_\rho(S))$, $\rho \in \mathcal{V}$ and privacy parameter ε, to select $\rho^* \in \mathcal{V}$ that approximately minimizes $F(\rho; g_\rho(S))$ over $\rho \in \mathcal{V}$. We use the generalized exponential mechanism as the sensitivity of $-F(\rho_j; g_{\rho_j}(S))$ depends on ρ_j. We then run $\mathcal{A}$ with margin parameter $\rho = \rho^*$ to output the final parameter vector w^{Priv}. For clarity, we include a formal description of the full algorithm in Algorithm 6. We now state the guarantee of the augmented algorithms.

Lemma F.1. *Let $\beta \in (0, 1)$. Suppose $S \sim \mathcal{D}^m$ for some distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. Suppose $\mathcal{A}$ is (ε, δ) differentially private and its output satisfies $R_{\mathcal{D}}(w^{\mathsf{Priv}}) \leq F(\rho, g_\rho(S))$ with probability at least $1 - \beta$. Furthermore, for any t, let $F(\rho, t)$ be a non-increasing function of ρ and $g_\rho(S)$ is a non-decreasing function of ρ for any S. Then, Algorithm 6 is $(2\varepsilon, \delta)$-differentially private and with probability at least $1 - 2\beta$, the output w^{Priv} satisfies:*

$$R_{\mathcal{D}}(w^{\mathsf{Priv}}) \leq \min_{\rho \in \left[\frac{h_{\max}}{\sqrt{m}}, h_{\max}\right]} F(\rho/2, g_\rho(S)) + \frac{\Delta_{\rho/2}(F)}{\varepsilon} \cdot \log\left(\frac{\log(m)}{\beta}\right),$$

where $\Delta_\rho(F)$ is a non-decreasing function of ρ and is an upper bound on the sensitivity of F given by $\Delta_\rho(F) = \max_{S, S' : d(S, S') = 1} |F(\rho, g_\rho(S)) - F(\rho, g_\rho(S'))|$ and $d(S, S')$ is the number of samples in which S and S' differ.

Proof. The privacy guarantee follows from the basic composition property of differential privacy together with the fact that the generalized exponential mechanism invoked in step 2 is ε-differentially private and $\mathcal{A}$ is (ε, δ)-differentially private.

We now turn to the proof of the error bound. Note that there exists $\hat{\rho} \in \mathcal{V}$ such that $\hat{\rho} \leq \rho_{\mathsf{opt}} < 2 \cdot \hat{\rho}$. By the properties of the generalized exponential mechanism [Raskhodnikova and Smith, 2016, Theorem

I.4] and the fact that the sensitivity of $F(\rho, g_\rho(S))$ is $\Delta_\rho(F)$, with probability at least $1-\beta$ we have

$$\begin{aligned} F(\rho^*; g_{\rho^*}(S)) &\le \min_{\rho\in\mathcal{V}} F(\rho; g_\rho(S)) + \frac{\Delta_\rho(F)}{\varepsilon}\cdot\log\left(\frac{\log(m)}{\beta}\right) \\ &\le F(\hat{\rho}; g_{\hat{\rho}}(S)) + \frac{\Delta_{\hat{\rho}}(F)}{\varepsilon}\cdot\log\left(\frac{\log(m)}{\beta}\right) \\ &\le F(\hat{\rho}; g_{\hat{\rho}}(S)) + \frac{\Delta_{\rho_{\mathsf{opt}/2}}(F)}{\varepsilon}\cdot\log\left(\frac{\log(m)}{\beta}\right) \\ &\le F(\hat{\rho}; g_{\rho_{\mathsf{opt}}}(S)) + \frac{\Delta_{\rho_{\mathsf{opt}/2}}(F)}{\varepsilon}\cdot\log\left(\frac{\log(m)}{\beta}\right) \\ &\le F(\rho_{\mathsf{opt}}/2; g_{\rho_{\mathsf{opt}}}(S)) + \frac{\Delta_{\rho_{\mathsf{opt}/2}}(F)}{\varepsilon}\cdot\log\left(\frac{\log(m)}{\beta}\right), \end{aligned} \tag{30}$$

where the last two inequalities follow from the fact that for any t, let $F(\rho, t)$ be a non-increasing function of ρ and $g_\rho(S)$ is a non-decreasing function of ρ for any S. By the assumption on $\mathcal{A}$, with probability $1-\beta$,

$$R_{\mathcal{D}}(w^{\mathsf{Priv}}) \le F(\rho^*; g_{\rho^*}(S)). \tag{31}$$

Combining (30) and (31) yields the lemma. The error probability follows by the union bound. □

The above lemma can be combined with any of the algorithms of Section 3, 4 and Appendix E. We instantiate it for $\mathcal{A}_{\mathsf{EffPrivMrg}}$ in the following corollary. Below, we compute sensitivity for the bounds on other algorithms, which can be used to get similar guarantees.

Corollary F.2. *Let $\beta \in (0,1)$ and $m \in \mathbb{N}$. Suppose $S \sim \mathcal{D}^m$ for some distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$. Recall that by Theorem 3.2, the output of Algorithm 2 (denoted by w') with probability at least $1-\beta/2$ satisfies, $R_{\mathcal{D}}(w') \le F(\rho, g_\rho(S))$, where $g_\rho(S) = \min_{w\in\mathbb{B}^d(\Lambda)} \widehat{L}^\rho_S(w)$ and*

$$F(\rho', t) = t + O\left(\sqrt{\frac{\log(1/\beta)}{m}} + \frac{\Lambda r}{\rho'}\left(\frac{1}{\sqrt{m}} + \frac{\sqrt{\log(\frac{m}{\beta})\log(\frac{1}{\beta})\log^{\frac{3}{4}}(\frac{1}{\delta})}}{\sqrt{\varepsilon m}}\right)\right).$$

Let w^{Priv} be the output of Algorithm 6 with inputs S, privacy parameters $\varepsilon/2, \delta$, bound $F(\rho, g_\rho(S))$, algorithm $\mathcal{A}_{\mathsf{EffPrivMrg}}$, confidence parameter $\beta/2$, and $h_{\max} = \Lambda r$. Then w^{Priv} is (ε, δ) differentially private. Furthermore, with probability at least $1-\beta$,

$$R_{\mathcal{D}}(w^{\mathsf{Priv}}) \le \min_{\rho\in\left[\frac{\Lambda r}{\sqrt{m}}, \Lambda r\right]} F(\rho/2, g_\rho(S)) + O\left(\frac{\Lambda r}{m\rho\varepsilon}\cdot\log\left(\frac{\log(m)}{\beta}\right)\right).$$

Lemma F.3. *Fix $\rho > 0$. Let the functions F_1, F_2, F_3, F_4 and F_5 are defined as follows:*

$$F_1(\rho', g_\rho(S)) = \min_{w\in\mathbb{B}^d(\Lambda)} \widehat{R}^\rho_S(w) + O\left(\sqrt{\widehat{R}^\rho_S(w)\left(\frac{\Lambda^2 r^2\log^2\left(\frac{m}{\beta}\right)}{m(\rho')^2} + \frac{\log\left(\frac{1}{\beta}\right)}{m}\right)} + \Gamma\right),$$

$$F_2(\rho', g_\rho(S)) = \min_{w\in\mathbb{B}^d(\Lambda)} \widehat{L}^\rho_S(w) + O\left(\sqrt{\frac{\log(1/\beta)}{m}} + \frac{\Lambda r}{\rho'}\left(\frac{1}{\sqrt{m}} + \frac{\sqrt{\log(\frac{m}{\beta})\log(\frac{1}{\beta})\log^{\frac{3}{4}}(\frac{1}{\delta})}}{\sqrt{\varepsilon m}}\right)\right),$$

$$F_3(\rho', g_\rho(S)) = \min_{h\in\mathcal{H}_\Lambda} \widehat{L}^\rho_S(h) + O\left(\sqrt{\frac{\log(1/\beta)}{m}} + \frac{\Lambda r}{\rho'}\left(\frac{1}{\sqrt{m}} + \frac{\sqrt{\log(\frac{m}{\beta})\log(\frac{1}{\beta})\log^{\frac{3}{4}}(\frac{1}{\delta})}}{\sqrt{\varepsilon m}}\right)\right),$$

$$F_4(\rho', g_\rho(S)) = \min_{h\in\mathcal{H}} \widehat{R}^\rho_S(h) + 2\sqrt{\min_{h\in\mathcal{H}} \widehat{R}^\rho_S(h)}\sqrt{\frac{M}{m}} + \frac{2M}{m} + \frac{64M\log\left(\frac{2}{\beta}\right)}{\varepsilon m},$$

$$F_5(\rho', g_\rho(S)) = \min_{h\in\mathcal{H}_{\mathsf{NN}^\Lambda}} \widehat{R}^\rho_S(h) + O\left(\frac{r(2\eta\Lambda)^L\sqrt{N\theta}}{\rho'\sqrt{m}} + \frac{r^2(2\eta\Lambda)^{2L}N\theta}{(\rho')^2\varepsilon m}\right),$$

where M is defined in Theorem E.1 and Γ is defined in Theorem 3.1. Then

$$\begin{aligned}
\Delta_\rho(F_1) &= O\left(\frac{1}{m} + \frac{1}{m}\sqrt{\frac{\Lambda^2 r^2 \log^2\left(\frac{m}{\beta}\right)}{\rho^2} + \log\left(\frac{1}{\beta}\right)}\right). \\
\Delta_\rho(F_2) &= O\left(\frac{\Lambda r}{m\rho}\right). \\
\Delta_\rho(F_3) &= O\left(\frac{\Lambda r}{m\rho}\right). \\
\Delta_\rho(F_4) &= O\left(\frac{1+\sqrt{M}}{m}\right). \\
\Delta_\rho(F_5) &= O\left(\frac{1}{m}\right).
\end{aligned}$$

Proof. We provide the proof for the bound on $\Delta_\rho(F_2)$. The proof for other quantities is similar and omitted. Let S' and S'' be two samples that differ in at most one sample. Without loss of generality, let $F_1(\rho, g_\rho(S')) \geq F_2(\rho, g_\rho(S''))$. Let

$$w' \in \operatorname*{argmin}_{w \in \mathbb{B}^d(\Lambda)} \widehat{L}^\rho_{S'}(w) + O\left(\sqrt{\frac{\log(1/\beta)}{m}} + \frac{\Lambda r}{\rho}\left(\frac{1}{\sqrt{m}} + \frac{\sqrt{\log(\frac{m}{\beta})\log(\frac{1}{\beta})\log^{\frac{3}{4}}(\frac{1}{\delta})}}{\sqrt{\varepsilon m}}\right)\right)$$

and

$$w'' \in \operatorname*{argmin}_{w \in \mathbb{B}^d(\Lambda)} \widehat{L}^\rho_{S''}(w) + O\left(\sqrt{\frac{\log(1/\beta)}{m}} + \frac{\Lambda r}{\rho}\left(\frac{1}{\sqrt{m}} + \frac{\sqrt{\log(\frac{m}{\beta})\log(\frac{1}{\beta})\log^{\frac{3}{4}}(\frac{1}{\delta})}}{\sqrt{\varepsilon m}}\right)\right).$$

Then

$$\begin{aligned}
F_1(\rho, g_\rho(S')) - F_2(\rho, g_\rho(S'')) &= \widehat{L}^\rho_{S'}(w') - \widehat{L}^\rho_{S''}(w'') \\
&\leq \widehat{L}^\rho_{S'}(w'') - \widehat{L}^\rho_{S''}(w'') \\
&\leq \frac{2}{m\rho} \max_{w \in \mathbb{B}^d(\Lambda), x \in \mathbb{B}^d(r)} |w \cdot x| \\
&\leq \frac{2}{m\rho}\Lambda r.
\end{aligned}$$

□

G Example of high error for exponential mechanism

Lemma G.1. *Let $d \geq c$ for some constant c and $\rho \in [0,1]$. There exists a distribution $\mathcal{D}$ over $\mathbb{B}^d$ and a subset $\mathcal{H} \in \mathcal{H}_{\mathsf{Lin}}$ such that the following hold:*

- ***Realizable setting**: There exists a h^* in $\mathcal{H}$ such that $R_{\mathcal{D}}(h^*) = 0$.*
- ***Only one good hypothesis**: For any h in $\mathcal{H} \setminus \{h^*\}$, $R_{\mathcal{D}}(h) = 1$.*
- ***A good cover**: For any two $h_w, h_{w'} \in \mathcal{H}$, their corresponding weights satisfy $|\langle w, w' \rangle| \geq 1/8$.*
- ***Exponential mechanism incurs high error**: Given $m < c' \cdot d/\varepsilon$ samples $\mathcal{D}$, with probability at least $9/10$, the exponential mechanism on $-\widehat{R}^\rho_S(h)$ will select a h such that $R_{\mathcal{D}}(h) = 1$.*

Proof. Let $\mathcal{D}$ be defined as follows. Let $\mathcal{D}(x)$ be a uniform distribution over $\{-1/\sqrt{d}, 1/\sqrt{d}\}^d$ and $y = 1$ if $x_1 > 0$, 0 otherwise. The optimal hypothesis $h^*(x) = 1_{x_1 > 0}$ and satisfies $R_{\mathcal{D}}(h^*) = 0$. Let $\mathcal{H} = \{h^*\} \cup \{h_w : w \in \mathcal{W}\}$, where $\mathcal{W}$ is the largest set such that for all $w \in \mathcal{W}$, $w_1 = -1/\sqrt{d}$ and for any two $w, w' \in \mathcal{W}$, $|\langle w, w' \rangle| \geq 1/8$. Bythe Gilbert-Varshamov bound, the size of such a set is at least $2^{c \cdot d}$ for some constant c. Note that for any $\mathcal{H} \setminus \{h^*\}$, $R_{\mathfrak{D}}(h) = 1$.

Now suppose we use the exponential mechanism with score $-\widehat{R}_S^{\rho}(h)$. The probability of selecting the correct hypothesis is at most

$$\frac{1}{\sum_{h \in \{h_w : w \in \mathcal{W}\}} \exp(-\widehat{R}_S^{\rho}(h)\varepsilon/2m)} \leq \frac{1}{2^{c \cdot d} e^{-\varepsilon m/2}} = e^{\varepsilon m/2 - c' d}.$$

Hence if $m < c'/d\varepsilon$, then the probability of choosing h^* is at most $e^{-c' d/2} \leq 1/10$ for $d \geq \frac{2}{c'} + 3$.

□

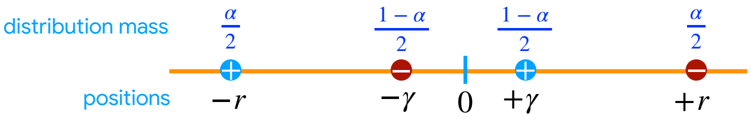


Figure 3: Simple example in dimension one for which the minimizer of the expected hinge loss $\mathbb{E}[\ell^{\text{hinge}}(w)]$ is $w^* = \frac{1}{\gamma}$ and thus $\|w^*\| = \frac{1}{\gamma} \gg 1$ for $\gamma \ll 1$.

H Example of a large norm hinge-loss minimizer

Fix $\alpha \in [0,1]$ and $\gamma \in (0,r)$. Consider the distribution $\mathcal{D}$ on the real line (dimension one) defined as follows: there is a probability mass of $\frac{\alpha}{2}$ at coordinate $(+r,-1)$, a probability mass of $\frac{\alpha}{2}$ at $(-r,+1)$, a probability mass of $\frac{1-\alpha}{2}$ at $(+\gamma,+1)$, and a probability mass of $\frac{1-\alpha}{2}$ at $(-\gamma,-1)$. Figure 3 illustrates this distribution. We first examine the expected hinge loss $\ell^{\text{hinge}}(w)$ of an arbitrary linear classifier $w \in \mathbb{R}$ in dimension one:

$$\begin{aligned}\ell^{\text{hinge}}(w) &= \frac{\alpha}{2}[\max\{0,1+wr\}+\max\{0,1+wr\}] + \frac{1-\alpha}{2}[\max\{0,1-w\gamma\}+\max\{0,1-w\gamma\}]\\ &= \alpha\max\{0,1+wr\} + (1-\alpha)\max\{0,1-w\gamma\}.\end{aligned}$$

Thus, distinguishing cases based on the value of scalar w, we can write:

$$\begin{aligned}\ell^{\text{hinge}}(w) &= \begin{cases}\alpha(1+wr)+(1-\alpha)(1-w\gamma) & \text{if } w \in \left[0,\frac{1}{\gamma}\right]\\ \alpha(1+wr) & \text{if } w \geq \frac{1}{\gamma}\\ \alpha(1+wr)+(1-\alpha)(1-w\gamma) & \text{if } w \in \left[-\frac{1}{r},0\right]\\ (1-\alpha)(1-w\gamma) & \text{if } w \leq -\frac{1}{r}.\end{cases}\\ &= \begin{cases}w(\alpha r-(1-\alpha)\gamma)+1 & \text{if } w \in \left[0,\frac{1}{\gamma}\right]\\ \alpha(1+wr) & \text{if } w \geq \frac{1}{\gamma}\\ w(\alpha r-(1-\alpha)\gamma)+1 & \text{if } w \in \left[-\frac{1}{r},0\right]\\ (1-\alpha)(1-w\gamma) & \text{if } w \leq -\frac{1}{r}.\end{cases}\end{aligned}$$

To simplify the discussion, we will set $r = 1$ and $\alpha = \frac{\gamma}{2}$, with $\gamma \ll 1$. This, implies $(\alpha r - (1-\alpha)\gamma) = \alpha(-1+2\alpha) < 0$. As a result of this negative sign, the best solution for the first two cases above is $w = \frac{1}{\gamma}$, $w = 0$ in the third case, and $w = -\frac{1}{r}$ in the last case. The loss achieved in the two latter cases is 1 and $(1-\alpha)(1+\frac{\gamma}{r})$, both larger than the loss $\alpha\big(1+\frac{r}{\gamma}\big)$ obtained in the first two cases. In view of that, the overall minimizer of $\ell^{\text{hinge}}(w)$ is given by $w^* = \frac{1}{\gamma}$, with $\ell^{\text{hinge}}(w^*) = \frac{1}{2}(1+\gamma)$. Note that the zero-one loss of w^* is $\ell(w^*) = \alpha = \frac{\gamma}{2}$.

Thus, for this example, the norm of the hinge-loss minimizer is arbitrary large: $\|w^*\| = \frac{1}{\gamma} \gg 1$. In particular, for a sample size m, we could choose $\gamma < \frac{1}{m}$, leading to $\|w^*\| > m$. Note that, here, any other positive classifier, $w > 0$, achieves the same zero-one loss as w^*. For example, $w = 1$ achieves the same performance as w^* with a more favorable norm.

Our analysis was presented for the population hinge loss but a similar result holds for the empirical hinge loss.