Efficient Private Algorithms for Learning Large-Margin Halfspaces

Huy Lê Nguyễn HLNGUYEN@CS.PRINCETON.EDU
Jonathan Ullman JULLMAN@CCS.NEU.EDU
Lydia Zakynthinou ZAKYNTHINOU.L@NORTHEASTERN.EDU
Khoury College of Computer Sciences, Northeastern University
Boston, USA

Editors: Aryeh Kontorovich and Gergely Neu

Abstract

We present new differentially private algorithms for learning a large-margin halfspace. In contrast to previous algorithms, which are based on either differentially private simulations of the statistical query model or on private convex optimization, the sample complexity of our algorithms depends only on the margin of the data, and not on the dimension. We complement our results with a lower bound, showing that the dependence of our upper bounds on the margin is optimal.

Keywords: Differential Privacy, PAC Learning, Large-Margin Halfspaces

1. Introduction

In a classification problem, we are given labeled examples from some unknown distribution, and the goal is to learn a classifier that accurately labels future examples from the same distribution. In many applications, each of these examples represents the highly sensitive *privacy* information of some individual. Although the goal of classification is to learn about the distribution, and not about the examples *per se*, many natural learning algorithms have the unfortunate side effect of revealing all or part of some of the labeled examples. For example, support vector machines represent the learned classifier as a set of support vectors, which are just labeled examples from the input!

The now-standard approach for ensuring privacy in machine learning is *differential privacy (DP)* (Dwork et al., 2006), which, informally, requires that no individual labeled example in the input significantly influences the learned classifier. Starting with some of the earliest work in differential privacy (Blum et al., 2005; Kasiviswanathan et al., 2008), there is a large body of literature showing that nearly every classification problem can be solved with differential privacy, albeit with large overheads in both sample complexity and running time. It is thus central to understand for which problems these overheads can be eliminated, and for which they are inherent.

In this paper we study the classical problem of learning a *large-margin halfspace*. That is, the examples are unit vectors $\mathbf{x} \in \mathbb{R}^d$ labeled with $y \in \{\pm 1\}$, and we assume that $y = \operatorname{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)$ for some unknown unit vector $\mathbf{w} \in \mathbb{R}^d$. Further, no example falls too close to the boundary of the halfspace, meaning that $y \cdot \langle \mathbf{w}, \mathbf{x} \rangle \geq \gamma$, where γ is called the *margin*. Any algorithm (private or non-private) for learning halfspaces over arbitrary distributions requires sample complexity growing polynomially in the dimension d. When d is large, assuming a large margin enables learning the halfspace with sample complexity independent of d.

Many results in differential privacy either explicitly or implicitly give private algorithms for learning a large-margin halfspace (see the related work for a detailed discussion). Blum et al.

Table 1: Sample complexity and running time bounds for our algorithms.

	Sample Complexity	Running Time	Privacy
Theorem 6	$\frac{1}{\alpha\varepsilon\gamma^2}$	$\mathrm{poly}(d, \frac{\log(1/\beta\delta)}{\alpha\varepsilon\gamma})$	(ε, δ)
Theorem 11	$\frac{1}{\alpha\varepsilon\gamma^2}$	$2^{\tilde{O}(1/\gamma^2)} \cdot \mathrm{poly}(d, \frac{\log(1/\beta\delta)}{\alpha\varepsilon})$	$(\varepsilon, 0)$

(2005) gave a differentially private implementation of the classical Perceptron algorithm for learning a large-margin halfspace, however their implementation requires sample complexity $\mathrm{poly}(d)$, which is precisely what the large-margin assumption is meant to avoid. A distinct line of work, beginning with Chaudhuri et al. (2011), studies differentially private algorithms for empirical loss minimization problems. Although learning a large-margin halfspace can be achieved via minimizing the *hinge loss*, generic algorithms for differentially private loss minimization inherently require $\mathrm{poly}(d)$ samples (Bun et al., 2014; Bassily et al., 2014).

1.1. Results

In this work we give two new differentially private algorithms for learning a large-margin halfspace. The key feature of our algorithms is that the sample complexity depends only on the margin, the desired accuracy of the learner, and the desired level of privacy, and not on the dimension. More precisely, our sample complexity is (ignoring constants and logarithmic factors) $1/\alpha\varepsilon\gamma^2$ where α is the desired error and ε is the desired privacy. In contrast, without privacy the sample complexity is roughly $1/\alpha\gamma^2$, so our sample complexity is comparable to that of non-private algorithms except when ε is very small.

Our first algorithm runs in polynomial time in all the parameters and satisfies the standard notion of (ε, δ)-DP. Our second algorithm's running time grows exponentially in the inverse-margin $1/\gamma$, but the algorithm satisfies the very strong special case of $(\varepsilon, 0)$-DP (so-called *pure DP*). Our results are described in more detail in Table 1. For simplicity, each of the bounds in Table 1 suppresses polylogarithmic factors of $\alpha, \beta, \varepsilon, \delta, \gamma$.

The main technique in both of our algorithms is to use random projections to reduce the dimensionality of the space to $\approx 1/\gamma^2$. After projection, we can learn using either a differentially private algorithm for minimizing hinge loss or by using the exponential mechanism over a net of possible halfspaces. We note that using either of these techniques on its own, without the projection, would fail to find an accurate classifier without $\mathrm{poly}(d)$ samples. We also note that one could apply a random projection and then run the algorithm of Blum et al. (2005), which would have sample complexity on the order of $1/\alpha\varepsilon\gamma^3$, which would be suboptimal.

We also prove a lower bound showing that any $(\varepsilon, 0)$-differentially private algorithm for learning a large-margin halfspace (with constant classification error) requires $\Omega(1/\varepsilon\gamma^2)$ samples (unless $d = o(1/\gamma^2)$). This lower bound is presented in Theorem 12.

1.2. Related Work

Blum et al. (2005) gave a differentially private implementation of the classical Perceptron algorithm, based on a general differentially private simulation of algorithms in the statistical queries model (Kearns, 1993). Their algorithm can be improved using more recent statistical queries al-

gorithms by Feldman et al. (2017), but this approach still requires $\mathrm{poly}(d)$ samples. The foundational work of Kasiviswanathan et al. (2008) studied differentially private PAC learning, and gave a generic private PAC learner, but they did not consider margin-based learning guarantees. In a recent work, Beimel et al. (2019) give a private learner for halfspaces over an arbitrary finite domain, significantly improving the dependence on the domain size by reducing the problem to the one of privately locating an approximate center point. They prove that their algorithm requires $\mathrm{poly}(d)$ samples and that this technique can not yield a better bound, but again they do not consider the large-margin assumption.

An alternative approach is to leverage algorithms for differentially private convex optimizing to identify a halfspace minimizing the hinge loss. Differentially private convex optimization is now the subject of a large body of literature that is too large to survey here. Notably, Bassily et al. (2014) gave nearly optimal algorithms for private convex optimization in the relevant setting, and showed that such algorithms necessarily require $\mathrm{poly}(d)$ samples. Jain and Thakurta (2014) gave nearly dimension-free results for minimizing *generalized linear models*, which could be used for learning a large-margin halfspace. However the sample complexity obtained by using these results would be $1/\alpha^2\varepsilon^2\gamma^2$, which is again significantly worse than ours.

Blum et al. (2008) gave an algorithm for the related *query release* problem for large-margin halfspaces—they construct a differentially private algorithm that outputs a data structure such that one can input a halfspace such that if the data has large margin with respect to that halfspace, then the structure outputs an estimate of how many points are labeled positively. One could use such a data structure to learn a large-margin halfspace, however, their algorithm has sample complexity $\mathrm{poly}(d)$, and the resulting learning algorithm would also not be computationally efficient.

Random projections have proven to be a very useful tool in learning theory in applications that assume some kind of separability (Vempala, 2004; Blum, 2006)). Similar to our work, there have also been applications of random projections in differential privacy. One example is the above query release algorithm from Blum et al. (2008), which is conceptually similar to our pure differentially private algorithm. In a very different setting, Blocki et al. (2013) demonstrated that certain random projection matrices *automatically* preserve privacy, however there is no technical relationship between their results and ours. Kenthapadi et al. (2012) also used the Johnson-Lindenstrauss transform to achieve better utility and computational efficiency for privately estimating distances between users, but again there is no technical relationship between their results and ours.

2. Preliminaries

2.1. Learning Halfspaces

We consider a distribution D over $\mathcal{X} \times \{\pm 1\}$, where $\mathcal{X} \subseteq \mathbb{R}^d$. We denote by $\mathcal{B}_2^d(r)$ the ball in $\mathbb{R}^d$ with center $\mathbf{0}$ and radius r with respect to the euclidean norm $\|\cdot\|_2$, and $\mathcal{B}_2^d(1) = \mathcal{B}_2^d$. We assume that all examples are normalized so that $\mathcal{X} = \mathcal{B}_2^d$. This assumption is without loss of generality. Since we can normalize each point in the input dataset, this operation would not affect privacy: any two neighboring datasets, as defined in the next section, would remain neighbors. As for utility, the (normalized) margin, as defined in this section, would also remained unchanged. Furthermore, if we consider γ to be the non-normalized margin and assume $\mathcal{X} \subseteq \mathcal{B}_2^d(R)$ for some known R, we can still run the algorithm with the scaled margin γ/R, increasing the sample complexity to $\tilde{O}(R^2/\alpha\varepsilon\gamma^2)$.

A linear threshold function is defined as $f_{\mathbf{w},\theta}(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + \theta)$, where $\mathbf{x}, \mathbf{w} \in \mathbb{R}^d$ and $\theta \in \mathbb{R}$. We assume without loss of generality that $\theta = 0$ so $f_{\mathbf{w}}(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)$.[1] We call a vector $\mathbf{w}$ a hypothesis. The error of a threshold function defined by hypothesis $\mathbf{w}$ on distribution D is

$$\text{err}_D(f_{\mathbf{w}}) = \Pr_{(\mathbf{x},y)\sim D}[f_{\mathbf{w}}(\mathbf{x}) \neq y] = \Pr_{(\mathbf{x},y)\sim D}[\text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) \neq y] = \Pr_{(\mathbf{x},y)\sim D}[y \cdot \langle \mathbf{w}, \mathbf{x} \rangle < 0].$$

As in the PAC (*probably approximately correct*) learning model, introduced by Valiant (1984), the goal is to find a hypothesis $\mathbf{w}$ such that $\text{err}_D(f_{\mathbf{w}}) \leq \alpha$ with probability $1 - \beta$, for given parameters α and β. We assume that there exists a hypothesis with zero error, that is, there exists a $\mathbf{w}^* \in \mathcal{B}_2^d$ such that $y \cdot \langle \mathbf{w}^*, \mathbf{x} \rangle > 0 \; \forall (\mathbf{x}, y)$. More specifically, we assume that $\mathbf{w}^*$ maximizes the margin

$$\gamma = \min_{\mathbf{x}\in\mathcal{X}} \frac{|\langle \mathbf{w}^*, \mathbf{x} \rangle|}{\|\mathbf{w}^*\|_2 \cdot \|\mathbf{x}\|_2},$$

which is assumed to be known in advance. Equivalently, $\gamma \leq |\cos(\mathbf{w}^*, \mathbf{x})| \; \forall \mathbf{x}$, where the right hand side is the distance of a scaled point $\mathbf{x}$ from the halfspace $\langle \mathbf{w}^*, \mathbf{x} \rangle = 0$.

Our goal is to design algorithms which, given enough data points drawn from a distribution D over a linearly separable set with margin γ, return a hypothesis which has error at most α with respect to the distribution, with probability $1-\beta$. More formally, we aim to design an (α, β, γ)-PAC learner with low sample complexity.

Definition 1 ((α, β, γ)-PAC learner) *Let D be a distribution over $\mathcal{B}_2^d \times \{\pm 1\}$ such that there exists $\mathbf{w}^* \in \mathcal{B}_2^d$ for which $\Pr_{(\mathbf{x},y)\sim D}[y\langle \mathbf{w}^*, \mathbf{x} \rangle \geq \gamma] = 1$. We call such a distribution D a* distribution with margin γ. *An algorithm $\mathcal{A}$ is an (α, β, γ)-*PAC learner for halfspaces in $\mathbb{R}^d$ with margin γ and sample complexity n *if, given a sample set $S \sim D^n$ from any distribution D with margin γ, it outputs a classifier $\mathcal{A}(S) = \hat{\mathbf{w}} \in \mathcal{B}_2^d$ such that, with probability at least $1 - \beta$,*

$$\Pr_{(\mathbf{x},y)\sim D}[y = \text{sign}(\langle \hat{\mathbf{w}}, \mathbf{x} \rangle)] \geq 1 - \alpha.$$

2.2. Differential Privacy

We design algorithms which draw a sample set S and output a hypothesis $\mathbf{w} \in \mathbb{R}^d$. In addition to finding a good hypothesis, our algorithms must satisfy *differential privacy (DP)* guarantees. Differential privacy is a property that a randomized algorithm satisfies if its output distribution does not change significantly under the change of a single data point.

More formally, let $S, S' \in \mathcal{S}^n$ be two data sets of the same size. We say that S, S' are *neighbors*, denoted as $S \sim S'$, if they differ in at most one data point.

Definition 2 (Differential Privacy, Dwork et al. (2006)) *A randomized algorithm $\mathcal{A} : \mathcal{S}^n \to \mathcal{O}$ is* (ε, δ)-differentially private *if for all neighboring data sets S, S' and all measurable $O \subseteq \mathcal{O}$,*

$$\Pr[\mathcal{A}(S) \in O] \leq \exp(\varepsilon) \Pr[\mathcal{A}(S') \in O] + \delta.$$

Algorithm $\mathcal{A}$ is $(\varepsilon, 0)$-differentially private *if it satisfies the definition for $\delta = 0$.*

1. If $m_{\mathcal{X}} = \min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_2$ is known, we can run the algorithm with the modified points $\tilde{\mathbf{x}} = [\mathbf{x}, 1] \in \mathbb{R}^{d+1}$, margin $\tilde{\gamma} = \gamma \cdot \min\{1, m_{\mathcal{X}}\}/(2 + 2|\theta|)$, and hypothesis space $\{[\mathbf{w}, \theta] : \mathbf{w} \in \mathcal{B}_2^d\}$.

3. An Efficient Private Algorithm

Both the algorithm of this and the next section draw a sample set $S \sim D^n$ and perform dimension reduction from a d-dimensional to an m-dimensional space, which allows them to run in the reduced space for the remainder of the execution.

Algorithm 1: $\mathcal{A}_{\alpha,\beta,\varepsilon,\delta,\gamma}(S)$

1: Choose a random matrix $A \in \mathbb{R}^{m\times d}$, where $m = O\left(\frac{\log(1/\beta_{JL})}{\gamma^2}\right)$, $\beta_{JL} = \alpha\beta^2/64n$, and

$$A_{ij} = \begin{cases} +1/\sqrt{m} & \text{w.p. } 1/2 \\ -1/\sqrt{m} & \text{w.p. } 1/2. \end{cases}$$

2: Define $S_A \leftarrow \{(A\mathbf{x}/\|A\mathbf{x}\|_2, y) \mid (\mathbf{x}, y) \in S\}$.
3: Define the hypothesis set $\mathcal{C} \leftarrow \mathcal{B}_2^m$.
4: Define the $\frac{100}{86\gamma}$-Lipschitz loss function $\ell : \mathcal{C} \times (\mathcal{B}_2^m \times \{\pm 1\}) \to \mathbb{R}$ as

$$\ell(\mathbf{w}; (\mathbf{x}, y)) = \mathbb{1}\left\{y \cdot \langle \mathbf{w}, \mathbf{x}\rangle < \frac{96\gamma}{100}\right\} \cdot \left(\frac{96}{86} - \frac{y \cdot \langle \mathbf{w}, \mathbf{x}\rangle}{86\gamma/100}\right).$$

5: Let $\hat{\mathbf{w}} \leftarrow \mathcal{F}(S_A, \ell, (\varepsilon, \delta), \mathcal{C})$ and **return** $\hat{\mathbf{w}}^\top A$.

Algorithm $\mathcal{F}$ is any differentially private empirical risk minimization algorithm. We can instantiate it with the noisy stochastic gradient descent algorithm of Bassily et al. (2014) so that it has the following guarantee. The full algorithm is presented in Appendix A for completeness.

Theorem 3 (Bassily et al. (2014)) *Let sample set D, L-Lipschitz loss function ℓ, differential privacy parameters (ε, δ), and convex hypothesis space $\mathcal{C}$ with diameter $\|\mathcal{C}\|_2$. There exists (ε, δ)-differentially private algorithm $\mathcal{F}$, such that with probability $1 - \beta/4$, its returned hypothesis $\hat{\mathbf{w}}$ satisfies*

$$\mathcal{L}(\hat{\mathbf{w}}; D) - \min_{\mathbf{w}\in\mathcal{C}\subseteq\mathbb{R}^m} \mathcal{L}(\mathbf{w}; D) = \frac{\sqrt{m}L\|\mathcal{C}\|_2}{\varepsilon} \cdot \operatorname{polylog}\left(n, \frac{1}{\beta}, \frac{1}{\delta}\right), \tag{1}$$

where $\mathcal{L}(\mathbf{w}; D) = \sum_{(\mathbf{x},y)\in D} \ell(\mathbf{w}; (\mathbf{x}, y))$ is the total loss of a hypothesis $\mathbf{w}$ on the data set D.

For the following proofs, we denote $\mathbf{x}_A := \frac{A\mathbf{x}}{\|A\mathbf{x}\|_2}$ for any $\mathbf{x} \in \mathcal{B}_2^d$. It holds that $\mathbf{x}_A \in \mathcal{B}_2^m$ and the modified sample set can be also written as $S_A = \{(\mathbf{x}_A, y) \mid (\mathbf{x}, y) \in S\}$. The lemma that follows guarantees that the transformation of a point $\mathbf{x} \mapsto A\mathbf{x}$, with high probability, only changes its euclidean norm by a small multiplicative factor.

Lemma 4 (Distributional Johnson-Lindenstrauss Lemma, Achlioptas (2003)) *Let $A \in \mathbb{R}^{m\times d}$ be a random matrix such that $m = O\left(\frac{\log(1/\beta_{JL})}{\gamma^2}\right)$ and $A_{ij} = \begin{cases} +1/\sqrt{m} & \text{w.p. } 1/2 \\ -1/\sqrt{m} & \text{w.p. } 1/2 \end{cases}$. Then, for every $\mathbf{x} \in \mathbb{R}^d$, it holds that $\Pr_A\left[\left|\|A\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \le \frac{\gamma}{100}\|\mathbf{x}\|_2^2\right] \ge 1 - \beta_{JL}$.*

Since the transform leaves the norm of a point $\mathbf{x}$ almost unchanged, one would expect that the corresponding transformed and normalized hypothesis $\mathbf{w}_A^* := A\mathbf{w}^*/\|A\mathbf{w}^*\|_2$ would still have a

large enough margin with respect to the corresponding point $\mathbf{x}_A$. The following lemma defines the probability that a point $\mathbf{x}$ belongs in the set of points $\mathcal{G}_A$, which are "good" for a fixed matrix A, in the sense that their norm remains almost unchanged and the margin of their corresponding points $\mathbf{x}_A$ from $\mathbf{w}_A^*$ is close to the original.

Lemma 5 *For every given matrix A, let $\mathcal{G}_A \subseteq \mathcal{X} \times \{\pm 1\}$ be the set of data points $(\mathbf{x}, y)$ that satisfy the following two statements:*

(i) $\left| \|A\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right| \leq \frac{\gamma}{100}\|\mathbf{x}\|_2^2$ *and*

(ii) $\mathbf{w}_A^* = \frac{A\mathbf{w}^*}{\|A\mathbf{w}^*\|_2}$ *has margin at least* $96\gamma/100$ *on* $(\mathbf{x}_A, y)$*, i.e.* $y \cdot \langle \mathbf{w}_A^*, \mathbf{x}_A \rangle \geq 96\gamma/100$.

It holds that $\Pr_{(\mathbf{x},y)\sim D}[(\mathbf{x}, y) \in \mathcal{G}_A] \geq 1 - 4\beta_{JL}$.

For the proof of Lemma 5, we express an inner product as $\langle \mathbf{w}^*, \mathbf{x} \rangle = \frac{1}{4}\| \mathbf{x} + \mathbf{w}^*\|_2^2 - \frac{1}{4}\| \mathbf{x} - \mathbf{w}^*\|_2^2$ and use the guarantee of Lemma 4 on vectors $\mathbf{x}, \mathbf{w}^*, \mathbf{x} - \mathbf{w}^*, \mathbf{x} + \mathbf{w}^*$. By union bound, with probability $1 - 4\beta_{JL}$, $\mathbf{x}_A$ has margin $96\gamma/100$ with respect to $\mathbf{w}_A^*$. The proof of Lemma 5 follows by four applications of Lemma 4 and is in Appendix A.

In the following, we provide the privacy and sample complexity guarantees of our algorithm.

Theorem 6 (Sample complexity) *Algorithm $\mathcal{A}_{\alpha,\beta,\varepsilon,\delta,\gamma}$ is an (α, β, γ)-learner with sample complexity*

$$n = \frac{1}{\alpha\varepsilon\gamma^2} \cdot \text{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\delta}, \frac{1}{\varepsilon}, \frac{1}{\gamma}\right).$$

Proof [Proof of Theorem 6] The first step of the algorithm is to sample matrix A uniformly at random from $U = \{\pm 1/\sqrt{m}\}^{m\times d}$. By Lemma 5, $\mathop{\mathbb{E}}_{A}\left[\Pr_{(\mathbf{x},y)\sim D}[(\mathbf{x}, y) \notin \mathcal{G}_A]\right] \leq 4\beta_{JL}$. By Markov's inequality,

$$\Pr_{A}\left[\Pr_{(\mathbf{x},y)\sim D}[(\mathbf{x}, y) \notin \mathcal{G}_A] \geq \beta'\right] \leq \mathop{\mathbb{E}}_{A}\left[\Pr_{(\mathbf{x},y)\sim D}[(\mathbf{x}, y) \notin \mathcal{G}_A]\right]/\beta' \leq 4\beta_{JL}/\beta'.$$

We set $\beta' = \alpha\beta/4n$. Then, substituting $\beta_{JL} = \frac{\alpha\beta^2}{64n}$, we get that with probability at least $1-\beta/4$,

$$\Pr_{(\mathbf{x},y)\sim D}[(\mathbf{x}, y) \in \mathcal{G}_A] \geq 1 - \beta'. \tag{2}$$

Therefore, with probability $1 - \beta/4$, the sampled matrix A satisfies inequality (2): a point $(\mathbf{x}, y) \sim D$ is in $\mathcal{G}_A$ with probability at least $1 - \beta'$. By union bound, $S \subseteq \mathcal{G}_A$, with probability $1 - n\beta' \geq 1 - \beta/4$. For the rest of the proof, we condition on the following, occurring with probability $1 - \beta/2$:

1. $\Pr_{(\mathbf{x},y)\sim D}[(\mathbf{x}, y) \in \mathcal{G}_A] \geq 1 - \beta'$ holds for A and
2. $S \subseteq \mathcal{G}_A$, that is, $\mathbf{w}_A^*$ has margin at least $96\gamma/100$ on S_A.

Claim 6.1 *If $n = \frac{1}{\alpha\varepsilon\gamma^2} \cdot \text{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\delta}, \frac{1}{\varepsilon}, \frac{1}{\gamma}\right)$, then for the hypothesis $\hat{\mathbf{w}}$ returned by $\mathcal{F}$, with probability $1 - \beta/4$, it holds that $\frac{1}{n}\sum_{(\mathbf{x}_A,y)\in S_A} \mathbb{1}\{y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A \rangle < \frac{\gamma}{10}\} \leq \frac{\alpha}{4}$.*

Proof [Proof of Claim 6.1] Since $\mathbf{w}_A^*$ has margin at least $96\gamma/100$ for all points in S_A, it holds that $\min_{\mathbf{w}\in\mathcal{C}} \mathcal{L}(\mathbf{w}; S_A) \leq \mathcal{L}(\mathbf{w}_A^*; S_A) = 0$. Substituting $\|\mathcal{C}\|_2 = 2$, $L = 100/86\gamma$, and $m = O\big(\log(n/\alpha\beta)/\gamma^2\big)$ into (1), dividing by n, and simplifying the expression, we get that with probability at least $1 - \beta/4$,

$$\frac{1}{n}\mathcal{L}(\hat{\mathbf{w}}; S_A) = \frac{1}{n\varepsilon\gamma^2} \cdot \mathrm{polylog}\left(n, \frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\delta}\right). \tag{3}$$

By calculations, it also holds that:

$$\begin{aligned}\frac{1}{n}\mathcal{L}(\hat{\mathbf{w}}; S_A) &= \frac{1}{n}\sum_{(\mathbf{x}_A, y)\in S_A} \mathbb{1}\left\{y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A\rangle < \frac{96\gamma}{100}\right\} \cdot \left(\frac{96}{86} - \frac{y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A\rangle}{86\gamma/100}\right) \\ &\geq \frac{1}{n}\sum_{(\mathbf{x}_A, y)\in S_A} \mathbb{1}\left\{y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A\rangle < \frac{\gamma}{10}\right\}.\end{aligned}$$

By the latter and inequality (3), it follows that with probability at least $1 - \beta/4$,

$$\frac{1}{n}\sum_{(\mathbf{x}_A, y)\in S_A} \mathbb{1}\left\{y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A\rangle < \frac{\gamma}{10}\right\} = \frac{1}{n\varepsilon\gamma^2} \cdot \mathrm{polylog}\left(n, \frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\delta}\right).$$

For the stated n, with probability $1 - \frac{\beta}{4}$, it holds that $\frac{1}{n}\sum_{(\mathbf{x}_A, y)\in S_A} \mathbb{1}\left\{y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A\rangle < \frac{\gamma}{10}\right\} \leq \frac{\alpha}{4}$. This completes the proof of the claim. ■

Claim 6.2 *If $n = \frac{1}{\alpha\varepsilon\gamma^2} \cdot \mathrm{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\delta}, \frac{1}{\varepsilon}, \frac{1}{\gamma}\right)$, then with probability $1 - \beta/2$, the error of the returned classifier $\hat{\mathbf{w}}^\top A$ on distribution D is $\Pr_{(\mathbf{x},y)\sim D}[y \cdot \langle \hat{\mathbf{w}}^\top A, \mathbf{x}\rangle < 0] \leq \alpha$.*

Proof [Proof of Claim 6.2] Let D_A denote the probability distribution with domain $\mathcal{B}_2^m \times \{\pm 1\}$, from which a sample $(\mathbf{x}_A, y) \in S_A$ is drawn. Let us also denote by $D_{|\mathcal{G}_A}$ distribution D restricted on $\mathcal{G}_A$. In our conditioned probability space, $S_A \sim D_A^n$, where the probability density function of D_A would be defined as

$$\Pr_{(\mathbf{x}_A, y)\sim D_A}[\mathbf{x}_A = \mathbf{x}' \wedge y = y'] = \Pr_{(\mathbf{x},y)\sim D_{|\mathcal{G}_A}}\left[\frac{A\mathbf{x}}{\|A\mathbf{x}\|_2} = \mathbf{x}' \wedge y = y'\right].$$

Let $\mathcal{H} = \{h : \{\mathbf{x}_A \mid (\mathbf{x}, y) \in \mathcal{G}_A\} \to \{\pm 1\} \text{ s.t. } h(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x}\rangle) \text{ for some } \mathbf{w} \in \mathcal{B}_2^m\}$ be a concept class of threshold functions in $\mathcal{B}_2^m$. We will use the following generalization bound.

Lemma 7 (Anthony and Bartlett (2009)) *Let $\mathcal{H}$ be a set of $\{\pm 1\}$-valued functions defined on a set X and P is a probability distribution on $Z = X \times \{\pm 1\}$. For $\eta \in (0, 1)$, $\zeta > 0$, and $n \in \mathbb{N}^+$, $\Pr_{z\sim P^n}[\exists h \in \mathcal{H} : \mathrm{err}_P(h) > (1 + \zeta)\hat{\mathrm{err}}_z(h) + \eta] \leq 4\Pi_{\mathcal{H}}(2n)\exp\left(-\frac{\eta\zeta n}{4(\zeta+1)}\right)$, where $\hat{\mathrm{err}}_z(h)$ is the empirical error of h on the sample set z and $\Pi_{\mathcal{H}}(\cdot)$ the growth function of $\mathcal{H}$.*

Setting $\eta = \alpha/4$ and $\zeta = 1$, we get that:

$$\Pr_{S_A \sim D_A^n}\left[\exists h \in \mathcal{H} : \mathrm{err}_{D_A}(h) > 2 \cdot \frac{1}{n} \sum_{(\mathbf{x}_A, y) \in S_A} \mathbb{1}\{h(\mathbf{x}_A) \neq y\} + \frac{\alpha}{4}\right] \leq 4\Pi_{\mathcal{H}}(2n) \exp\left(-\frac{\alpha n}{32}\right).$$

By Theorems 3.4 and 3.7 of Anthony and Bartlett (2009), we get that $\mathrm{VCdim}(\mathcal{H}) = m + 1$ and $\Pi_{\mathcal{H}}(2n) \leq (2n)^{m+1} + 1$. Substituting $m = O\big(\log(1/\alpha\beta)/\gamma^2\big)$ and for $n = \frac{1}{\alpha\gamma^2} \cdot \mathrm{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\gamma}\right)$, $4\Pi_{\mathcal{H}}(2n)\exp(-\alpha n/32) \leq \beta/4$. Therefore, with probability at least $1 - \beta/4$,

$$\mathrm{err}_{D_A}(f_{\hat{\mathbf{w}}}) \leq 2 \cdot \frac{1}{n} \sum_{(\mathbf{x}_A, y) \in S_A} \mathbb{1}\{y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A \rangle < 0\} + \frac{\alpha}{4}. \tag{4}$$

By Claim 6.1, $\frac{1}{n} \sum_{(\mathbf{x}_A, y) \in S_A} \mathbb{1}\{y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A \rangle < 0\} \leq \frac{1}{n} \sum_{(\mathbf{x}_A, y) \in S_A} \mathbb{1}\big\{y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A \rangle < \frac{\gamma}{10}\big\} \leq \frac{\alpha}{4}$ holds with probability $1 - \beta/4$, if $n = \frac{1}{\alpha\varepsilon\gamma^2} \cdot \mathrm{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\delta}, \frac{1}{\varepsilon}, \frac{1}{\gamma}\right)$. Therefore, by inequality (4), if $n = \frac{1}{\alpha\varepsilon\gamma^2} \cdot \mathrm{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\delta}, \frac{1}{\varepsilon}, \frac{1}{\gamma}\right)$, then with probability at least $1 - \beta/2$, $\mathrm{err}_{D_A}(f_{\hat{\mathbf{w}}}) \leq 2 \cdot \frac{\alpha}{4} + \frac{\alpha}{4} = \frac{3\alpha}{4}$.

Equivalently, with probability at least $1 - \beta/2$,

$$\Pr_{(\mathbf{x}, y) \sim D_{|\mathcal{G}_A}}[y \cdot \langle \hat{\mathbf{w}}^\top A, \mathbf{x} \rangle < 0] = \Pr_{(\mathbf{x}_A, y) \sim D_A}[y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A \rangle < 0] = \mathrm{err}_{D_A}(f_{\hat{\mathbf{w}}}) \leq \frac{3\alpha}{4}.$$

Since, by Condition 1., $\Pr_{(\mathbf{x}, y) \sim D}[(\mathbf{x}, y) \notin \mathcal{G}_A] \leq \beta' \leq \frac{\alpha}{4}$, it follows that with probability $1 - \beta/2$,

$$\begin{aligned}\Pr_{(\mathbf{x}, y) \sim D}[y \cdot \langle \hat{\mathbf{w}}^\top A, \mathbf{x} \rangle < 0] &\leq \Pr_{(\mathbf{x}, y) \sim D_{|\mathcal{G}_A}}[y \cdot \langle \hat{\mathbf{w}}^\top A, \mathbf{x} \rangle < 0] \cdot (1 - \beta') + 1 \cdot \beta' \\ &\leq \frac{3\alpha}{4} \cdot (1 - \beta') + \beta' \leq \frac{3\alpha}{4} + \frac{\alpha}{4} \leq \alpha.\end{aligned}$$

This completes the proof of the claim. ■

Accounting for the probability that we are not in the conditioned space, we conclude that if $n = \frac{1}{\alpha\varepsilon\gamma^2} \cdot \mathrm{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\delta}, \frac{1}{\varepsilon}, \frac{1}{\gamma}\right)$, then with probability $1 - \beta/2 - \beta/2 = 1 - \beta$, $\mathrm{err}_D(f_{\hat{\mathbf{w}}^\top A}) \leq \alpha$. This completes the proof of the theorem. ■

Theorem 8 (Privacy guarantee) *Algorithm $\mathcal{A}_{\alpha,\beta,\varepsilon,\delta,\gamma}$ is (ε, δ)-differentially private.*

Proof [Proof of Theorem 8] Differential privacy is closed under post-processing Dwork et al. (2006). So it suffices to show that an algorithm $\mathcal{N}$ that is the same as $\mathcal{A}_{\alpha,\beta,\varepsilon,\delta,\gamma}$ except that it returns $\hat{\mathbf{w}}$ instead of $\hat{\mathbf{w}}^\top A$, is (ε, δ)-DP.

Let data sets $S \sim S'$ such that $S = S' \setminus \{(\mathbf{x}', y')\} \cup \{(\mathbf{x}, y)\}$. Let $U = \{\pm 1/\sqrt{m}\}^{m \times d}$. If we fix a matrix $A \in U$, then S and S' correspond to $\mathcal{F}$'s inputs S_A and $S'_A = S_A \setminus \{(\mathbf{x}'_A, y')\} \cup \{(\mathbf{x}_A, y)\}$,

respectively. Recall from Theorem 3, that $\mathcal{F}$ is (ε, δ)-DP. For any measurable set $\mathcal{R} \subseteq \mathbb{R}^m$,

$$\begin{aligned}
\Pr[\mathcal{N}(S) \in \mathcal{R}] &= \sum_{A \in U} \Pr[A] \cdot \Pr[\mathcal{F}(S_A) \in \mathcal{R} \mid A] \\
&\leq \sum_{A \in U} \Pr[A] \cdot \left(\exp(\varepsilon) \Pr[\mathcal{F}(S'_A) \in \mathcal{R} \mid A] + \delta\right) \qquad \text{(by Theorem 3)} \\
&= \exp(\varepsilon) \sum_{A \in U} \Pr[A] \cdot \Pr[\mathcal{F}(S'_A) \in \mathcal{R} \mid A] + \delta \sum_{A \in U} \Pr[A] \\
&= \exp(\varepsilon) \Pr[\mathcal{N}(S') \in \mathcal{R}] + \delta.
\end{aligned}$$

Therefore, $\mathcal{N}$ is (ε, δ)-DP, and so is $\mathcal{A}_{\alpha,\beta,\varepsilon,\delta,\gamma}$. ■

4. A Pure Differentially Private Algorithm

As previously, algorithm $\mathcal{A}_{\alpha,\beta,\varepsilon,\gamma}$ takes as input a sample set $S \sim D^n$ and performs dimension reduction. In this reduced space, it defines a net of hypotheses and uses the *Exponential Mechanism* (McSherry and Talwar, 2007) to choose a good hypothesis with respect to the sample set. The Exponential Mechanism is a well-known algorithm, which serves as a building block for many differentially private algorithms. It is used in cases where we need to choose the optimal output with respect to some utility function on the data set.

Lemma 9 (The Exponential Mechanism, McSherry and Talwar (2007)) *Let data set $S \in \mathcal{S}^n$, range $\mathcal{O}$, and utility function $u : \mathcal{S}^n \times \mathcal{O} \to \mathbb{R}$. The* Exponential Mechanism $\mathcal{M}_E(S, u, \mathcal{O})$ *selects and outputs an element $o \in \mathcal{O}$ with probability proportional to $\exp\left(\frac{\varepsilon \cdot u(S,o)}{2\Delta u}\right)$, where $\Delta u = \max_{o \in \mathcal{O}} \max_{S \sim S'} |u(S, o) - u(S', o)|$. The Exponential Mechanism is $(\varepsilon, 0)$-differentially private and with probability at least $1 - \delta$,*

$$|\max_{o \in \mathcal{O}} u(S, o) - u(\mathcal{M}_E(S, u, \mathcal{O}))| \leq \frac{2\Delta u}{\varepsilon} \ln\left(\frac{|\mathcal{O}|}{\delta}\right).$$

In the following we provide the sample complexity and privacy guarantees of our algorithm.

Theorem 10 (Privacy guarantee) *Algorithm $\mathcal{A}_{\alpha,\beta,\varepsilon,\gamma}$ is $(\varepsilon, 0)$-differentially private.*

Theorem 11 (Sample Complexity) *Algorithm $\mathcal{A}_{\alpha,\beta,\varepsilon,\gamma}$ is an (α, β, γ)-learner with sample complexity*

$$n = \frac{1}{\alpha\varepsilon\gamma^2} \cdot \text{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\varepsilon}, \frac{1}{\gamma}\right).$$

The proofs of Theorems 10 and 11 are similar to those of the previous section, except for the bound on the empirical loss of the learned classifier. We prove this part here, while full proof is in Appendix B.

Claim 11.1 *If $n = \frac{1}{\alpha\varepsilon\gamma^2} \cdot \text{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\varepsilon}, \frac{1}{\gamma}\right)$, then with probability $1 - \beta/4$, for hypothesis $\hat{\mathbf{w}}$ returned by the Exponential Mechanism it holds that $\frac{1}{n} \sum_{(\mathbf{x}_A, y) \in S_A} \mathbb{1}\{y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A \rangle < \frac{\gamma}{10}\} \leq \frac{\alpha}{4}$.*

Algorithm 2: $\mathcal{A}_{\alpha,\beta,\varepsilon,\gamma}(S)$

1: Choose a random matrix $A \in \mathbb{R}^{m\times d}$, where $m = O\Big(\frac{\log(1/\beta_{JL})}{\gamma^2}\Big)$, $\beta_{JL} = \alpha\beta^2/64n$, and

$$A_{ij} = \begin{cases} +1/\sqrt{m} & \text{w.p. } 1/2 \\ -1/\sqrt{m} & \text{w.p. } 1/2. \end{cases}$$

2: Define $S_A \leftarrow \{(A\mathbf{x}/\|A\mathbf{x}\|_2, y) \mid (\mathbf{x}, y) \in S\}$.
3: Let $\mathcal{W}$ be a $\frac{\gamma}{10}-$Net of $\mathcal{B}_2^m$.
4: Define the utility function $u : (\mathcal{B}_2^m \times \{\pm 1\})^n \times \mathcal{W} \to [-1, 0]$:

$$u(D, \mathbf{w}) = -\frac{1}{n} \cdot \sum_{(\mathbf{x},y)\in D} \mathbb{1}\Big\{y \cdot \langle \mathbf{w}, \mathbf{x}\rangle < \frac{\gamma}{10}\Big\}.$$

5: $\hat{\mathbf{w}} \leftarrow \mathcal{M}_E(S_A, u, \mathcal{W})$ and **return** $\hat{\mathbf{w}}^\top A$.

Proof [Proof of Claim 11.1] Every point in $\mathcal{B}_2^m$ is within $\gamma/10$ from a center of $\mathcal{W}$. Let $\mathbf{w}_c^*$ be the center within $\gamma/10$ from $\mathbf{w}_A^*$, that is, $\|\mathbf{w}_A^* - \mathbf{w}_c^*\|_2 \le \gamma/10$. Recall that in our conditioned probability space, for all $(\mathbf{x}_A, y) \in S_A$, $y \cdot \langle \mathbf{w}_A^*, \mathbf{x}_A\rangle \ge 96\gamma/100$. Therefore, by these inequalities, for all $(\mathbf{x}_A, y) \in S_A$,

$$y \cdot \langle \mathbf{w}_c^*, \mathbf{x}_A\rangle = y \cdot \langle \mathbf{w}_A^*, \mathbf{x}_A\rangle - y \cdot \langle \mathbf{w}_A^* - \mathbf{w}_c^*, \mathbf{x}_A\rangle \ge y \cdot \langle \mathbf{w}_A^*, \mathbf{x}_A\rangle - \|\mathbf{w}_A^* - \mathbf{w}_c^*\|_2 \cdot \|\mathbf{x}_A\|_2 > \frac{\gamma}{10}.$$

It follows that $\max_{\mathbf{w}\in\mathcal{W}} u(S_A, \mathbf{w}) \ge u(S_A, \mathbf{w}_c^*) = -\frac{1}{n}\sum_{(\mathbf{x}_A,y)\in S_A} \mathbb{1}\big\{y \cdot \langle \mathbf{w}, \mathbf{x}_A\rangle < \frac{\gamma}{10}\big\} = 0$.

By the latter and Lemma 9, with probability at least $1 - \beta/4$, it holds that:

$$\frac{1}{n} \sum_{(\mathbf{x}_A,y)\in S_A} \mathbb{1}\Big\{y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A\rangle < \frac{\gamma}{10}\Big\} \le \frac{2}{n\varepsilon}(\ln(|\mathcal{W}|) + \ln(4/\beta)) \tag{5}$$

It is a well-known result that the covering number of an m-dimensional unit ball by balls of radius $\gamma/10$ is at most $O((10/\gamma)^m)$. Therefore, substituting $m = O\big(\log(n/\alpha\beta)/\gamma^2\big)$, it follows that $\ln|\mathcal{W}| = \frac{1}{\gamma^2} \cdot \mathrm{polylog}\Big(n, \frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\gamma}\Big)$.Thus, by inequality (5), if $n = \frac{1}{\alpha\varepsilon\gamma^2} \cdot \mathrm{polylog}\Big(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\gamma}, \frac{1}{\varepsilon}\Big)$ then with probability $1 - \beta/4$, $\frac{1}{n}\sum_{(\mathbf{x}_A,y)\in S_A} \mathbb{1}\big\{y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A\rangle < \frac{\gamma}{10}\big\} \le \frac{\alpha}{4}$. This completes the proof of the claim. ■

We note that the running time of algorithm $\mathcal{A}_{\alpha,\beta,\varepsilon,\gamma}$ is exponential in $1/\gamma$, due to the use of the exponential mechanism to select from a net of size $\exp(1/\gamma)$. Although there are efficient implementations of the exponential mechanism for optimizing over *continuous* domains (Bassily et al., 2014), relaxing the domain to be continuous would lead to higher sample complexity.

5. A Sample Complexity Lower Bound for Pure Differential Privacy

In this section we prove a lower bound on the sample complexity of any $(\varepsilon, 0)$-differentially private algorithm for learning a large-margin halfspace.

Theorem 12 *Any $(\varepsilon, 0)$-differentially private $(\frac{1}{10}, \frac{1}{10}, \gamma)$-learner for halfspaces in $\mathbb{R}^{\Omega(1/\gamma^2)}$ requires $\Omega\big(1/\varepsilon\gamma^2\big)$ samples.*

Proof Our proof uses a standard *packing argument*. We construct distributions $D^{(1)}, \ldots, D^{(K)}$ over $\mathcal{B}_2^d \times \{\pm 1\}$ for $d = 1/1000\gamma^2$ and $K = 2^{d/20}$. We will construct these distributions so that no classifier is simultaneously accurate for two distinct distributions $D^{(i)}$ and $D^{(j)}$. This will imply that $n = \Omega(\log(K)/\varepsilon) = \Omega(1/\varepsilon\gamma^2)$ samples are necessary to achieve $(\varepsilon, 0)$-differential privacy.

Each distribution $D^{(i)}$ is defined with respect to a halfspace $\mathbf{w}^{(i)} \in \{\pm 1/\sqrt{d}\}^d$ such that $\Pr_{(\mathbf{x},y)\sim D^{(i)}}[y \cdot \langle \mathbf{w}^{(i)}, \mathbf{x}\rangle \geq \gamma] = 1$. In addition, $\mathbf{x}$ is distributed uniformly at random on the remaining surface of $\mathcal{B}_2^d$ so that is does not violate the margin and $y = \operatorname{sign}(\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle)$. Formally, if U denotes the uniform distribution on $\mathcal{B}_2^d$ and f_U its pdf, then the pdf of X where $(X, y) \sim D^{(i)}$, is

$$f_X(\mathbf{x}') = \begin{cases} f_U(\mathbf{x}')/\Pr_{\mathbf{x}\sim U}[|\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle| \geq \gamma], & \text{if } |\langle \mathbf{w}^{(i)}, \mathbf{x}'\rangle| \geq \gamma \\ 0, & \text{otherwise.} \end{cases}$$

Using standard constructions of error correcting codes, there exists a set $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(K)}$ such that the Hamming distance of any pair $i \neq j$, is $\operatorname{Ham}(\mathbf{w}^{(i)}, \mathbf{w}^{(j)}) \geq \frac{d}{10}$, which implies that $\langle \mathbf{w}^{(i)}, \mathbf{w}^{(j)}\rangle \leq \frac{4}{5}$.

Let $\{\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(K)}\}$ be such a set and let $D^{(1)}, \ldots, D^{(K)}$ be the resulting distributions. The crux of the proof is in establishing the following claim about this set of distributions. For each distribution $D^{(i)}$, we define the set $\mathcal{G}^{(i)} = \big\{\hat{\mathbf{w}} \in \mathcal{B}_2^d : \operatorname{err}_{D^{(i)}}(\hat{\mathbf{w}}) \leq \frac{1}{10}\big\}$ of all classifiers that have error at most $1/10$ on the distribution $D^{(i)}$.

Claim 12.1 *For every $i \neq j$, $\mathcal{G}^{(i)}$ and $\mathcal{G}^{(j)}$ are disjoint.*

We prove this claim in Appendix C. Using this claim, we can complete the proof as follows. Let $S^{(1)} \sim \big(D^{(i)}\big)^n$ denote a random iid sample of n examples from $D^{(i)}$. Let $\mathcal{A}$ be an $(\varepsilon, 0)$-differentially private $(\frac{1}{10}, \frac{1}{10}, \gamma)$ learner. By privacy and accuracy, we have for every $i \in \{2, 3, \ldots, K\}$,

$$\Pr[\mathcal{A}(S^{(1)}) \in \mathcal{G}^{(i)}] \geq \exp(-n\varepsilon)\Pr[\mathcal{A}(S^{(1)}) \in \mathcal{G}^{(1)}] \geq \frac{9}{10}\exp(-n\varepsilon).$$

Since all $\mathcal{G}^{(i)}$ are disjoint,

$$\Pr[\mathcal{A}(S^{(1)}) \notin \mathcal{G}^{(1)}] \geq \sum_{i=2}^{K} \Pr[\mathcal{A}(S^{(1)}) \in \mathcal{G}^{(i)}] \geq \frac{9(K-1)}{10}\exp(-n\varepsilon).$$

Since, by accuracy, $\Pr[\mathcal{A}(S^{(1)}) \notin \mathcal{G}^{(1)}] \leq \frac{1}{10}$, it follows that

$$\frac{9}{10}(K-1)\exp(-n\varepsilon) \leq \frac{1}{10}.$$

Rearranging, and substituting our choice of K, we conclude $n = \Omega(1/\varepsilon\gamma^2)$. ■

Acknowledgments

JU was supported by NSF grants CCF-1718088, CCF-1750640, and CNS-1816028. HN and LZ were supported by NSF grants CCF-1750716 and CCF-1909314.

References

Dimitris Achlioptas. Database-friendly Random Projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671 – 687, 2003. ISSN 0022-0000. doi: https://doi.org/10.1016/S0022-0000(03)00025-4. URL http://www.sciencedirect.com/science/article/pii/S0022000003000254. Special Issue on PODS 2001.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, New York, NY, USA, 1st edition, 2009. ISBN 052111862X, 9780521118620.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, FOCS '14, pages 464–473, Washington, DC, USA, 2014. IEEE Computer Society. ISBN 978-1-4799-6517-5. doi: 10.1109/FOCS.2014.56. URL http://dx.doi.org/10.1109/FOCS.2014.56.

Amos Beimel, Shay Moran, Kobbi Nissim, and Uri Stemmer. Private center points and learning of halfspaces. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 269–282, Phoenix, USA, 25–28 Jun 2019. PMLR. URL http://proceedings.mlr.press/v99/beimel19a.html.

Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. Differentially private data analysis of social networks via restricted sensitivity. In Robert D. Kleinberg, editor, *Innovations in Theoretical Computer Science, ITCS '13, Berkeley, CA, USA, January 9-12, 2013*, pages 87–96. ACM, 2013. ISBN 978-1-4503-1859-4. doi: 10.1145/2422436.2422449. URL https://doi.org/10.1145/2422436.2422449.

Avrim Blum. Random Projection, Margins, Kernels, and Feature-Selection. In Craig Saunders, Marko Grobelnik, Steve Gunn, and John Shawe-Taylor, editors, *Subspace, Latent Structure and Feature Selection*, pages 52–68, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-34138-3.

Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical Privacy: The SuLQ framework. In *Proceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '05, pages 128–138, New York, NY, USA, 2005. ACM. ISBN 1-59593-062-0. doi: 10.1145/1065167.1065184. URL http://doi.acm.org/10.1145/1065167.1065184.

Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, pages 609–618, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-047-0. doi: 10.1145/1374376.1374464. URL http://doi.acm.org/10.1145/1374376.1374464.

Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting Codes and the Price of Approximate Differential Privacy. In *Proceedings of the Forty-sixth Annual ACM Symposium on Theory of*

Computing, STOC '14, pages 1–10, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2710-7. doi: 10.1145/2591796.2591877. URL http://doi.acm.org/10.1145/2591796.2591877.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially Private Empirical Risk Minimization. *J. Mach. Learn. Res.*, 12:1069–1109, July 2011. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1953048.2021036.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

Vitaly Feldman, Cristóbal Guzmán, and Santosh Vempala. Statistical Query Algorithms for Mean Vector Estimation and Stochastic Convex Optimization. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, pages 1265–1277, Philadelphia, PA, USA, 2017. Society for Industrial and Applied Mathematics. URL http://dl.acm.org/citation.cfm?id=3039686.3039768.

Prateek Jain and Abhradeep Thakurta. (Near) Dimension Independent Risk Bounds for Differentially Private Learning. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages I–476–I–484. JMLR.org, 2014. URL http://dl.acm.org/citation.cfm?id=3044805.3044860.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. What Can We Learn Privately? In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA*, pages 531–540. IEEE Computer Society, 2008. doi: 10.1109/FOCS.2008.27. URL https://doi.org/10.1109/FOCS.2008.27.

Michael Kearns. Efficient Noise-tolerant Learning from Statistical Queries. In *Proceedings of the Twenty-fifth Annual ACM Symposium on Theory of Computing*, STOC '93, pages 392–401, New York, NY, USA, 1993. ACM. ISBN 0-89791-591-7. doi: 10.1145/167088.167200. URL http://doi.acm.org/10.1145/167088.167200.

Krishnaram Kenthapadi, Aleksandra Korolova, Ilya Mironov, and Nina Mishra. Privacy via the Johnson-Lindenstrauss transform. *arXiv preprint arXiv:1204.2606*, 2012.

Frank McSherry and Kunal Talwar. Mechanism Design via Differential Privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 94–103, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-3010-9. doi: 10.1109/FOCS.2007.41. URL http://dx.doi.org/10.1109/FOCS.2007.41.

L. G. Valiant. A Theory of the Learnable. *Commun. ACM*, 27(11):1134–1142, November 1984. ISSN 0001-0782. doi: 10.1145/1968.1972.

Santosh Srinivas Vempala. *The Random Projection Method*, volume 65 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. DIMACS/AMS, 2004. ISBN 0-8218-3793-1. URL http://dimacs.rutgers.edu/Volumes/Vol65.html.

Appendix A. Algorithm and Proofs of Section 3

A.1. Differentially Private Empirical Risk Minimization Algorithm $\mathcal{F}$

A complete algorithm, employing the differentially private stochastic gradient descent algorithm by Bassily et al. (2014), is presented below in Algorithm 3. We denote by $\Pi_{\mathcal{C}}(\cdot)$ the euclidean projection on $\mathcal{C}$ and by $\|\mathcal{C}\|_2$ the diameter of $\mathcal{C}$.

Algorithm 3: $\mathcal{A}_{\alpha,\beta,\varepsilon,\delta,\gamma}(S)$

Choose a random matrix $A \in \mathbb{R}^{m\times d}$, where $m = O\Big(\frac{\log(1/\beta_{JL})}{\gamma^2}\Big)$, $\beta_{JL} = \alpha\beta^2/64n$, and

$$A_{ij} = \begin{cases} +1/\sqrt{m} & \text{w.p. } 1/2 \\ -1/\sqrt{m} & \text{w.p. } 1/2. \end{cases}$$

Define $S_A \leftarrow \{(A\mathbf{x}/\|A\mathbf{x}\|_2, y) \mid (\mathbf{x}, y) \in S\}$.
Define the hypothesis set $\mathcal{C} \leftarrow \mathcal{B}_2^m$.
Define the $\frac{100}{86\gamma}$-Lipschitz loss function $\ell : \mathcal{C} \times (\mathcal{B}_2^m \times \{\pm 1\}) \to \mathbb{R}$ as

$$\ell(\mathbf{w}; (\mathbf{x}, y)) = \mathbb{1}\left\{y \cdot \langle \mathbf{w}, \mathbf{x}\rangle < \frac{96\gamma}{100}\right\} \cdot \left(\frac{96}{86} - \frac{y \cdot \langle \mathbf{w}, \mathbf{x}\rangle}{86\gamma/100}\right).$$

Let $\hat{\mathbf{w}} \leftarrow \mathcal{F}(S_A, \ell, (\varepsilon, \delta), \mathcal{C})$.
return $\hat{\mathbf{w}}^\top A$.

procedure $\mathcal{F}(D, \ell, (\varepsilon, \delta), \mathcal{C})$
  for $i \leftarrow 1$ **to** $\lceil \log(8/\beta)\rceil$ **do**
    $\hat{\mathbf{w}}^{(i)} \leftarrow \mathcal{A}_{\text{Noise-GD}}(D, \ell, (\varepsilon/\lceil \log(8/\beta)\rceil, \delta/\lceil \log(8/\beta)\rceil), \mathcal{C})$.
  end
  $\mathcal{W} \leftarrow \{\hat{\mathbf{w}}^{(1)}, \ldots, \hat{\mathbf{w}}^{(\lceil \log(8/\beta)\rceil)}\}$.
  return $\hat{\mathbf{w}} \leftarrow \mathcal{M}_E(D, -\ell, \mathcal{W})$.

procedure $\mathcal{A}_{\text{Noise-GD}}(D, \ell, (\varepsilon', \delta'), \mathcal{C})$
  Noise variance $\sigma^2 \leftarrow \frac{32L^2n^2\log(n/\delta')\log(1/\delta')}{\varepsilon'^2}$, where L is the Lipschitz constant of ℓ.
  Learning rate function $\eta : [n^2] \to \mathbb{R}$: $\eta(t) = \frac{\|\mathcal{C}\|_2}{\sqrt{t(n^2L^2+m\sigma^2)}}$.
  Choose a point from $\mathcal{C}$, $\mathbf{w}_1$.
  for $t \leftarrow 1$ **to** $n^2 - 1$ **do**
    Pick $(\mathbf{x}, y) \sim_u D$ with replacement.
    $\mathbf{w}_{t+1} \leftarrow \Pi_{\mathcal{C}}(\mathbf{w}_t - \eta(t)[n\nabla\ell(\mathbf{w}_t; (\mathbf{x}, y)) + b_t])$, where $b_t \sim \mathcal{N}(0, \mathbb{I}_m\sigma^2)$.
  end
  return $\mathbf{w}_{n^2}$.

To achieve a high-probability guarantee, algorithm $\mathcal{F}$ runs $\mathcal{A}_{\text{Noise–GD}}$ $\lceil \log(8/\beta)\rceil$ times, with privacy parameters $\varepsilon/\lceil \log(8/\beta)\rceil$ and $\delta/\lceil \log(8/\beta)\rceil$, and uses the Exponential Mechanism to pick the best hypothesis $\hat{\mathbf{w}}$, as described in Appendix D of Bassily et al. (2014).

A.2. Proof of Lemma 5

We state Lemma 5 again for convenience.

Lemma 13 (Lemma 5) *For every given matrix A, let $\mathcal{G}_A \subseteq \mathcal{X} \times \{\pm 1\}$ be the set of data points $(\mathbf{x}, y)$ that satisfy the following two statements:*

(i) $\left|\|A\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \leq \frac{\gamma}{100}\|\mathbf{x}\|_2^2$ *and*

(ii) $\mathbf{w}_A^* = \frac{A\mathbf{w}^*}{\|A\mathbf{w}^*\|_2}$ *has margin at least* $96\gamma/100$ *on* $(\mathbf{x}_A, y)$*, i.e.* $y \cdot \langle \mathbf{w}_A^*, \mathbf{x}_A \rangle \geq 96\gamma/100$.

It holds that

$$\Pr_{(\mathbf{x},y)\sim D}[(\mathbf{x}, y) \in \mathcal{G}_A] \geq 1 - 4\beta_{JL}.$$

Proof [Proof of Lemma 5] By Lemma 4,

$$\left|\|A\mathbf{u}\|_2^2 - \|\mathbf{u}\|_2^2\right| \leq \frac{\gamma}{100}\|\mathbf{u}\|_2^2$$

holds for a point $\mathbf{u} \in \mathbb{R}^d$ with probability at least $1 - \beta_{JL}$. By union bound, it holds simultaneously for all points $\mathbf{x} + \mathbf{w}^*$, $\mathbf{x} - \mathbf{w}^*$, $\mathbf{x}$, and $\mathbf{w}^*$, with probability at least $1 - 4\beta_{JL}$. Under this condition, statement (i) is true and for $y = 1$ we have:

$$\begin{aligned}
\langle \mathbf{w}^*, \mathbf{x} \rangle &= \frac{1}{4}\|\, \mathbf{x} + \mathbf{w}^*\|_2^2 - \frac{1}{4}\|\, \mathbf{x} - \mathbf{w}^*\|_2^2 \\
&\leq \frac{1}{4\left(1 - \frac{\gamma}{100}\right)}\|\, A(\mathbf{x} + \mathbf{w}^*)\|_2^2 - \frac{1}{4\left(1 + \frac{\gamma}{100}\right)}\|\, A(\mathbf{x} - \mathbf{w}^*)\|_2^2 \\
&= \frac{1}{4\left(1 - \frac{\gamma^2}{100^2}\right)}\left[\left(1 + \frac{\gamma}{100}\right)\|A(\mathbf{x} + \mathbf{w}^*)\|_2^2 - \left(1 - \frac{\gamma}{100}\right)\|A(\mathbf{x} - \mathbf{w}^*)\|_2^2\right] \\
&= \frac{1}{1 - \frac{\gamma^2}{100^2}}\left(\frac{1}{4}\|\, A\mathbf{x} + A\mathbf{w}^*\|_2^2 - \frac{1}{4}\|\, A\mathbf{x} - A\mathbf{w}^*\|_2^2\right) \\
&\qquad + \frac{\frac{\gamma}{100}}{4\left(1 - \frac{\gamma^2}{100^2}\right)}\left(\|\, A\mathbf{x} + A\mathbf{w}^*\|_2^2 + \|\, A\mathbf{x} - A\mathbf{w}^*\|_2^2\right) \\
&= \frac{1}{1 - \frac{\gamma^2}{100^2}}\langle A\mathbf{w}^*, A\mathbf{x} \rangle + \frac{\frac{\gamma}{100}}{2\left(1 - \frac{\gamma^2}{100^2}\right)}\left(\|\, A\mathbf{x}\|_2^2 + \|\, A\mathbf{w}^*\|_2^2\right) \\
&\leq \frac{1}{1 - \frac{\gamma^2}{100^2}}\langle A\mathbf{w}^*, A\mathbf{x} \rangle + \frac{\frac{\gamma}{100}\left(1 + \frac{\gamma}{100}\right)}{1 - \frac{\gamma^2}{100^2}}
\end{aligned}$$

Equivalently $\langle A\mathbf{w}^*, A\mathbf{x} \rangle \geq \left(1 - \frac{\gamma^2}{100^2}\right)\langle \mathbf{w}^*, \mathbf{x} \rangle - \frac{\gamma}{100}\left(1 + \frac{\gamma}{100}\right)$. Since $y = 1$ and $\langle \mathbf{w}^*, \mathbf{x} \rangle = y \cdot \langle \mathbf{w}^*, \mathbf{x} \rangle \geq \gamma$, it follows that:

$$y \cdot \langle A\mathbf{w}^*, A\mathbf{x} \rangle \geq \left(1 - \frac{\gamma^2}{100^2}\right)\gamma - \frac{\gamma}{100}\left(1 + \frac{\gamma}{100}\right) \geq \frac{98\gamma}{100}$$

Therefore, for $y = 1$,

$$y \cdot \langle \mathbf{w}_A^*, \mathbf{x}_A \rangle = y \cdot \left\langle \frac{A\mathbf{w}^*}{\|A\mathbf{w}^*\|_2}, \frac{A\mathbf{x}}{\|A\mathbf{x}\|_2} \right\rangle \geq \frac{\frac{98\gamma}{100}}{1 + \frac{\gamma}{100}} \geq \frac{96\gamma}{100}.$$

The proof for $y = -1$ is similar. We conclude that with probability at least $1 - 4\beta_{JL}$, statements (i) and (ii) are true. ■

Appendix B. Proofs of Section 4

B.1. Proof of Sample Complexity: $(\varepsilon, 0)$-DP

Theorem 14 (Sample Complexity, Theorem 11) *Algorithm $\mathcal{A}_{\alpha,\beta,\varepsilon,\gamma}$ is an (α, β, γ)-learner with sample complexity*

$$n = \frac{1}{\alpha\varepsilon\gamma^2} \cdot \text{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\varepsilon}, \frac{1}{\gamma}\right).$$

Proof [Proof of Theorem 11] As in the previous section, the first step of the algorithm is to sample matrix A uniformly at random from $U = \left\{\pm\frac{1}{\sqrt{m}}\right\}^{m\times d}$. From Lemma 5, it follows that:

$$\mathop{\mathbb{E}}_{A}\left[\Pr_{(\mathbf{x},y)\sim D}[(\mathbf{x}, y) \notin \mathcal{G}_A]\right] \leq 4\beta_{JL}$$

And, by Markov's inequality,

$$\Pr_{A}\left[\Pr_{(\mathbf{x},y)\sim D}[(\mathbf{x}, y) \notin \mathcal{G}_A] \geq \beta'\right] \leq \frac{\mathop{\mathbb{E}}_{A}\left[\Pr_{(\mathbf{x},y)\sim D}[(\mathbf{x}, y) \notin \mathcal{G}_A]\right]}{\beta'} \leq \frac{4\beta_{JL}}{\beta'}.$$

We set $\beta' = \alpha\beta/4n$. Then, substituting $\beta_{JL} = \frac{\alpha\beta^2}{64n}$, we get that with probability at least $1-\beta/4$,

$$\Pr_{(\mathbf{x},y)\sim D}[(\mathbf{x}, y) \in \mathcal{G}_A] \geq 1 - \beta'.$$

Therefore, with probability $1 - \beta/4$, the sampled matrix A satisfies the above inequality, that is, a point $(\mathbf{x}, y) \sim D$ is in $\mathcal{G}_A$ with probability at least $1 - \beta'$. Furthermore, by union bound, $\forall(\mathbf{x}, y) \in S$ it holds that $(\mathbf{x}, y) \in \mathcal{G}_A$, with probability at least $1 - n\beta' \geq 1 - \beta/4$.

For the remainder of the proof, we condition on the event that

1. $\Pr_{(\mathbf{x},y)\sim D}[(\mathbf{x}, y) \in \mathcal{G}_A] \geq 1 - \beta'$ holds for A and
2. $S \subseteq \mathcal{G}_A$, that is, $\mathbf{w}^*_A$ has margin at least $96\gamma/100$ on S_A.

This event occurs with probability at least $1 - \beta/4 - \beta/4 = 1 - \beta/2$.

Claim 14.1 *If $n = \frac{1}{\alpha\varepsilon\gamma^2} \cdot \text{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\varepsilon}, \frac{1}{\gamma}\right)$, then with probability $1 - \beta/4$, for hypothesis $\hat{\mathbf{w}}$ returned by the Exponential Mechanism it holds that*

$$\frac{1}{n}\sum_{(\mathbf{x}_A,y)\in S_A} \mathbb{1}\left\{y \cdot \langle\hat{\mathbf{w}}, \mathbf{x}_A\rangle < \frac{\gamma}{10}\right\} \leq \frac{\alpha}{4}. \tag{6}$$

Proof [Proof of Claim 14.1] Every point in $\mathcal{B}_2^m$ is within $\gamma/10$ from a center of $\mathcal{W}$. Let $\mathbf{w}_c^*$ be the center within $\gamma/10$ from $\mathbf{w}_A^*$, that is,

$$\|\mathbf{w}_A^* - \mathbf{w}_c^*\|_2 \leq \gamma/10. \tag{7}$$

Recall that in our conditioned probability space,

$$y \cdot \langle \mathbf{w}_A^*, \mathbf{x}_A \rangle \geq 96\gamma/100 \tag{8}$$

holds for all $(\mathbf{x}_A, y) \in S_A$. Therefore, for all $(\mathbf{x}_A, y) \in S_A$,

$$\begin{aligned} y \cdot \langle \mathbf{w}_c^*, \mathbf{x}_A \rangle &= y \cdot \langle \mathbf{w}_A^*, \mathbf{x}_A \rangle - y \cdot \langle \mathbf{w}_A^* - \mathbf{w}_c^*, \mathbf{x}_A \rangle \\ &\geq y \cdot \langle \mathbf{w}_A^*, \mathbf{x}_A \rangle - \|\mathbf{w}_A^* - \mathbf{w}_c^*\|_2 \cdot \|\mathbf{x}_A\|_2 \\ &\geq 96\gamma/100 - \gamma/10 = 86\gamma/10 > \gamma/10. \qquad \text{(by inequalities (7), (8))} \end{aligned}$$

It follows that

$$\max_{\mathbf{w} \in \mathcal{W}} u(S_A, \mathbf{w}) \geq u(S_A, \mathbf{w}_c^*) = -\frac{1}{n} \sum_{(\mathbf{x}_A, y) \in S_A} \mathbb{1}\left\{ y \cdot \langle \mathbf{w}, \mathbf{x}_A \rangle < \frac{\gamma}{10} \right\} = 0. \tag{9}$$

By Lemma 9 and inequality (9), with probability at least $1 - \beta/4$, it holds that:

$$\frac{1}{n} \sum_{(\mathbf{x}_A, y) \in S_A} \mathbb{1}\left\{ y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A \rangle < \frac{\gamma}{10} \right\} \leq \frac{2}{n\varepsilon}(\ln(|\mathcal{W}|) + \ln(4/\beta)) \tag{10}$$

It is a well-known result that the covering number of an m-dimensional unit ball by balls of radius $\gamma/10$ is at most $O\left(\left(\frac{1}{\gamma/10}\right)^m\right)$. Therefore, substituting $m = O\left(\frac{\log(n/\alpha\beta)}{\gamma^2}\right)$, it follows that

$$\ln |\mathcal{W}| = \frac{1}{\gamma^2} \cdot \operatorname{polylog}\left(n, \frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\gamma}\right).$$

Thus, by inequality (10), if $n = \frac{1}{\alpha\varepsilon\gamma^2} \cdot \operatorname{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\gamma}, \frac{1}{\varepsilon}\right)$ then with probability at least $1 - \beta/4$,

$$\frac{1}{n} \sum_{(\mathbf{x}_A, y) \in S_A} \mathbb{1}\left\{ y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A \rangle < \frac{\gamma}{10} \right\} \leq \frac{\alpha}{4}.$$

This concludes the proof of the claim. ■

Claim 14.2 *If $n = \frac{1}{\alpha\varepsilon\gamma^2} \cdot \operatorname{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\varepsilon}, \frac{1}{\gamma}\right)$, then with probability $1 - \beta/2$, the error of the returned classifier $\hat{\mathbf{w}}^\top A$ on distribution D is*

$$\Pr_{(\mathbf{x}, y) \sim D}[y \cdot \langle \hat{\mathbf{w}}^\top A, \mathbf{x} \rangle < 0] \leq \alpha. \tag{11}$$

Proof [Proof of Claim 14.2] Let D_A denote the probability distribution with domain $\mathcal{B}_2^m \times \{\pm 1\}$, from which a sample $(\mathbf{x}_A, y) \in S_A$ is drawn. Let us also denote by $D_{|\mathcal{G}_A}$ distribution D restricted on $\mathcal{G}_A$. In our conditioned probability space, $S_A \sim D_A^n$, where the probability density function of D_A would be defined as

$$\Pr_{(\mathbf{x}_A,y)\sim D_A}[\mathbf{x}_A = \mathbf{x}' \wedge y = y'] = \Pr_{(\mathbf{x},y)\sim D_{|\mathcal{G}_A}}\left[\frac{A\mathbf{x}}{\|A\mathbf{x}\|_2} = \mathbf{x}' \wedge y = y'\right].$$

Let $\mathcal{H} = \{h : \{\mathbf{x}_A \mid (\mathbf{x}, y) \in \mathcal{G}_A\} \to \{\pm 1\}$ s.t. $h(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x}\rangle)$ for some $\mathbf{w} \in \mathcal{B}_2^m\}$ be a concept class of threshold functions in $\mathcal{B}_2^m$. By Theorem 3.4 of Anthony and Bartlett (2009), $\text{VCdim}(\mathcal{H}) = m + 1$.

By the generalization bound of Theorem 5.7 of Anthony and Bartlett (2009), stated in Lemma 7, it holds that:

$$\Pr_{S_A\sim D_A^n}\left[\exists h \in \mathcal{H} : \text{err}_{D_A}(h) > 2 \cdot \frac{1}{n}\sum_{(\mathbf{x}_A,y)\in S_A} \mathbb{1}\{h(\mathbf{x}_A) \neq y\} + \frac{\alpha}{4}\right] \leq 4\Pi_{\mathcal{H}}(2n)\exp(-\alpha n/32)$$

where the growth function $\Pi_{\mathcal{H}}(2n) \leq (2n)^{m+1}+1$, by Theorem 3.7 of Anthony and Bartlett (2009).

Using $m = O\left(\frac{\log(1/\alpha\beta)}{\gamma^2}\right)$, it holds that if $n = \frac{1}{\alpha\gamma^2}\cdot\text{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\gamma}\right)$ then $4\Pi_{\mathcal{H}}(2n)\exp(-\alpha n/32) \leq \beta/4$. Therefore, with probability at least $1 - \beta/4$,

$$\text{err}_{D_A}(f_{\hat{\mathbf{w}}}) \leq 2 \cdot \frac{1}{n}\sum_{(\mathbf{x}_A,y)\in S_A} \mathbb{1}\{y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A\rangle < 0\} + \frac{\alpha}{4} \tag{12}$$

By Claim 14.1, $\frac{1}{n}\sum_{(\mathbf{x}_A,y)\in S_A} \mathbb{1}\{y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A\rangle < 0\} \leq \frac{1}{n}\sum_{(\mathbf{x}_A,y)\in S_A} \mathbb{1}\{y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A\rangle < \frac{\gamma}{10}\} \leq \frac{\alpha}{4}$ holds with probability $1 - \beta/4$, if $n = \frac{1}{\alpha\varepsilon\gamma^2}\cdot\text{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\delta}, \frac{1}{\varepsilon}, \frac{1}{\gamma}\right)$.

Therefore, by inequality (12), if $n = \frac{1}{\alpha\varepsilon\gamma^2}\cdot\text{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\delta}, \frac{1}{\varepsilon}, \frac{1}{\gamma}\right)$, then with probability at least $1 - \beta/4 - \beta/4 = 1 - \beta/2$,

$$\text{err}_{D_A}(f_{\hat{\mathbf{w}}}) \leq 2 \cdot \frac{\alpha}{4} + \frac{\alpha}{4} = \frac{3\alpha}{4}.$$

Equivalently, with probability at least $1 - \beta/2$,

$$\Pr_{(\mathbf{x},y)\sim D_{|\mathcal{G}_A}}[y \cdot \langle \hat{\mathbf{w}}^\top A, \mathbf{x}\rangle < 0] = \Pr_{(\mathbf{x}_A,y)\sim D_A}[y \cdot \langle \hat{\mathbf{w}}, \mathbf{x}_A\rangle < 0] = \text{err}_{D_A}(f_{\hat{\mathbf{w}}}) \leq \frac{3\alpha}{4}.$$

Since, by Condition 1., $\Pr_{(\mathbf{x},y)\sim D}[(\mathbf{x}, y) \notin \mathcal{G}_A] \leq \beta' \leq \frac{\alpha}{4}$, it follows that with probability at least $1 - \beta/2$,

$$\begin{aligned}\Pr_{(\mathbf{x},y)\sim D}[y \cdot \langle \hat{\mathbf{w}}^\top A, \mathbf{x}\rangle < 0] &\leq \Pr_{(\mathbf{x},y)\sim D_{|\mathcal{G}_A}}[y \cdot \langle \hat{\mathbf{w}}^\top A, \mathbf{x}\rangle < 0] \cdot (1 - \beta') + 1 \cdot \beta' \\ &\leq \frac{3\alpha}{4} \cdot (1 - \beta') + \beta' \\ &\leq \frac{3\alpha}{4} + \frac{\alpha}{4} \leq \alpha.\end{aligned}$$

This completes the proof of the claim. ■

Accounting for the probability that we are not in the conditioned space, we conclude that if $n = \frac{1}{\alpha\varepsilon\gamma^2} \cdot \mathrm{polylog}\left(\frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\delta}, \frac{1}{\varepsilon}, \frac{1}{\gamma}\right)$, then with probability at least $1-\beta/2-\beta/2 = 1-\beta$, $\mathrm{err}_D(f_{\hat{\mathbf{w}}^\top A}) \leq \alpha$. This completes the proof of the theorem. ■

B.2. Proof of Privacy Guarantee: $(\varepsilon, 0)$-DP

Theorem 15 (Privacy guarantee, Theorem 10) *Algorithm $\mathcal{A}_{\alpha,\beta,\varepsilon,\gamma}$ is $(\varepsilon, 0)$-differentially private.*

Proof [Proof of Theorem 10] The sensitivity of the utility function is

$$\Delta u = \max_{\mathbf{w}\in\mathcal{W}} \max_{\substack{Z,Z'\in(\mathcal{X}\times\{\pm1\})^n \\ Z\sim Z'}} |u(Z,\mathbf{w}) - u(Z',\mathbf{w})| \leq \frac{1}{n}.$$

It follows by Lemma 9 that $\mathcal{M}_E$ is $(\varepsilon, 0)$-DP.

Differentialy privacy is closed under post-processing Dwork et al. (2006). Therefore it suffices to show that an algorithm $\mathcal{N}$ that is the same as $\mathcal{A}_{\alpha,\beta,\varepsilon,\gamma}$ except that it returns $\hat{\mathbf{w}}$ instead of $\hat{\mathbf{w}}^\top A$, is $(\varepsilon, 0)$-DP.

Let S and S' be two neighboring sample sets such that $S = S' \setminus \{(\mathbf{x}', y')\} \cup \{(\mathbf{x}, y)\}$. Let $U = \left\{\pm\frac{1}{\sqrt{m}}\right\}^{m\times d}$. If we fix a matrix $A \in U$, the sample sets S and S' would correspond to $\mathcal{M}_E$'s inputs S_A and $S'_A = S_A \setminus \{(\mathbf{x}'_A, y')\} \cup \{(\mathbf{x}_A, y)\}$, respectively. For any measurable set $\mathcal{R} \subseteq \mathbb{R}^m$, it holds that

$$\begin{aligned}
\Pr[\mathcal{N}(S) \in \mathcal{R}] &= \sum_{A\in U} \Pr[A] \cdot \Pr[\mathcal{M}_E(S_A) \in \mathcal{R} \mid A] \\
&\leq \sum_{A\in U} \Pr[A] \cdot \exp(\varepsilon) \Pr[\mathcal{M}_E(S'_A) \in \mathcal{R} \mid A] \qquad \text{(since } \mathcal{M}_E \text{ is } (\varepsilon,0)\text{-DP)} \\
&= \exp(\varepsilon) \sum_{A\in U} \Pr[A] \cdot \Pr[\mathcal{M}_E(S'_A) \in \mathcal{R} \mid A] \\
&= \exp(\varepsilon) \Pr[\mathcal{N}(S') \in \mathcal{R}].
\end{aligned}$$

Therefore, $\mathcal{N}$ is $(\varepsilon, 0)$-DP, and so is $\mathcal{A}_{\alpha,\beta,\varepsilon,\gamma}$. ■

Appendix C. Remaining Proofs of Section 5

Claim 15.1 (Claim 12.1) *For every $i \neq j$, $\mathcal{G}^{(i)}$ and $\mathcal{G}^{(j)}$ are disjoint.*

Proof [Proof of Claim 12.1] We will show that for an arbitrary $\hat{\mathbf{w}} \in \mathcal{B}_2^d$, and every $i \neq j$, if $\mathrm{err}_{D^{(i)}}(\hat{\mathbf{w}}) \leq 1/10$ then $\mathrm{err}_{D^{(j)}}(\hat{\mathbf{w}}) > 1/10$. Let U_i be the distribution over $\mathcal{B}_2^d \times \{\pm1\}$ such that $(\mathbf{x}, y) \sim U_i$ if $\mathbf{x} \sim U$ is uniform over the unit sphere in $\mathbb{R}^d$ and $y = \mathrm{sign}(\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle)$. Define the probability

$$p_\gamma = \Pr_{\mathbf{x}\sim U}[|\langle \mathbf{w}, \mathbf{x}\rangle| < \gamma]$$

of a uniformly distributed point on $\mathcal{B}_2^d$ lying within margin γ of a unit vector $\mathbf{w}$. Probability p_γ remains unchanged if we replace $\mathbf{w}$ with any other unit vector and, more specifically,

$$p_\gamma = \Pr_{(\mathbf{x},y)\sim U_i}[|\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle| < \gamma]$$

holds for all $U_i, i \in [K]$.

The next lemma will allow us to show that U_i is not too far from $D^{(i)}$.

Lemma 16 *If $d = \frac{1}{1000\gamma^2}$ then $p_\gamma \leq 0.2$.*

Proof Consider the following sampling process from the uniform distribution on the sphere: choose each coordinate $x_i \sim \mathcal{N}(0,1)$ and normalize with $\|\mathbf{x}\|_2$. By the symmetric property of multi-dimensional gaussian vectors, we know that the projection of $\mathbf{x}$ on a unit vector $\mathbf{w}$ is distributed as $x_1/\|\mathbf{x}\|_2$, where $x_1 \sim \mathcal{N}(0,1)$ is the first coordinate of $\mathbf{x}$ and $\|\mathbf{x}\|_2^2 \sim \tilde{\chi}^2(d)$ is the square of the normalization factor. The probability of a point having margin more than γ from $\mathbf{w}$ is:

$$\begin{aligned}
1 - p_\gamma &= \Pr_{\mathbf{x}\sim U}\left[\frac{|x_1|}{\|\mathbf{x}\|_2} \geq \gamma\right] \\
&\geq \Pr_{\mathbf{x}\sim U}\left[\left(|x_1| \geq \frac{1}{10}\right) \wedge \left(\|\mathbf{x}\|_2 \leq \frac{1}{10\gamma}\right)\right] \\
&= 1 - \Pr_{\mathbf{x}\sim U}\left[\left(|x_1| < \frac{1}{10}\right) \vee \left(\|\mathbf{x}\|_2 > \sqrt{10d}\right)\right] \qquad (d = \tfrac{1}{1000\gamma^2}) \\
&\geq 1 - \Pr_{x_1\sim\mathcal{N}(0,1)}\left[|x_1| < \frac{1}{10}\right] - \Pr_{\|\mathbf{x}\|_2^2\sim\tilde{\chi}^2(d)}\left[\|\mathbf{x}\|_2^2 > 10d\right]
\end{aligned}$$

By the tables of the standard normal distribution we have that $\Pr_{x_1\sim\mathcal{N}(0,1)}\left[|x_1| < \frac{1}{10}\right] \leq 0.08$. Also, the mean of a $\tilde{\chi}^2(d)$ distributed variable is d. By Markov's inequality, it follows that

$$\Pr_{\|\mathbf{x}\|_2^2\sim\tilde{\chi}^2(d)}\left[\|\mathbf{x}\|_2^2 > 10d\right] \leq d/10d = 1/10.$$

Thus, $p_\gamma \leq 0.18$. ■

We can apply the preceding lemma to relate $\mathrm{err}_{U_i}(\hat{\mathbf{w}})$ to $\mathrm{err}_{D^{(i)}}(\hat{\mathbf{w}})$. Specifically, for any $\hat{\mathbf{w}}$ with $\mathrm{err}_{D^{(i)}}(\hat{\mathbf{w}}) \leq 0.10$, it holds that:

$$\begin{aligned}
\mathrm{err}_{U_i}(\hat{\mathbf{w}}) &= \Pr_{(\mathbf{x},y)\sim U_i}[\mathrm{sign}(\langle\hat{\mathbf{w}},\mathbf{x}\rangle) \neq \mathrm{sign}(\langle\mathbf{w}^{(i)},\mathbf{x}\rangle)] \\
&\leq \Pr_{(\mathbf{x},y)\sim U_i}[\mathrm{sign}(\langle\hat{\mathbf{w}},\mathbf{x}\rangle) \neq \mathrm{sign}(\langle\mathbf{w}^{(i)},\mathbf{x}\rangle) \mid |\langle\mathbf{w}^{(i)},\mathbf{x}\rangle| \geq \gamma] \cdot \Pr_{(\mathbf{x},y)\sim U_i}[|\langle\mathbf{w}^{(i)},\mathbf{x}\rangle| \geq \gamma] \\
&\qquad + \Pr_{(\mathbf{x},y)\sim U_i}[|\langle\mathbf{w}^{(i)},\mathbf{x}\rangle| < \gamma] \\
&\leq 0.1 + 0.2 = 0.3
\end{aligned}$$

Therefore,

$$\mathrm{err}_{U_i}(\hat{\mathbf{w}}) \leq 0.3. \tag{13}$$

Next we will argue that the same vector $\hat{\mathbf{w}}$ cannot have low error with respect to some other distribution U_j. Fix any two vectors $\mathbf{w}^{(i)}, \mathbf{w}^{(j)}$ as in our construction. Consider the plane defined by these vectors and let θ be their angle. It holds that

$$\Pr_{\mathbf{x}\sim U}[\mathrm{sign}(\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle) = \mathrm{sign}(\langle \mathbf{w}^{(j)}, \mathbf{x}\rangle)] = \frac{2\theta}{2\pi} = \frac{\theta}{\pi} = \frac{\cos^{-1}(\langle \mathbf{w}^{(i)}, \mathbf{w}^{(j)}\rangle)}{\pi}.$$

Since $\mathrm{Ham}(\mathbf{w}^{(i)}, \mathbf{w}^{(j)}) \geq d/10$, $\langle \mathbf{w}^{(i)}, \mathbf{w}^{(j)}\rangle \leq \frac{1}{d}(9d/10 - d/10) = \frac{8}{10}$. Thus,

$$\Pr_{\mathbf{x}\sim U}[\mathrm{sign}(\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle) = \mathrm{sign}(\langle \mathbf{w}^{(j)}, \mathbf{x}\rangle)] \leq \frac{\cos^{-1}(8/10)}{\pi} = 0.21$$

For the error of $\hat{\mathbf{w}}$ on distribution U_j it holds that:

$$\begin{aligned}
&\Pr_{(\mathbf{x},y)\sim U_j}[\mathrm{sign}(\langle \hat{\mathbf{w}}, \mathbf{x}\rangle) = \mathrm{sign}(\langle \mathbf{w}^{(j)}, \mathbf{x}\rangle)] \\
&\leq \Pr_{(\mathbf{x},y)\sim U_j}[\mathrm{sign}(\langle \hat{\mathbf{w}}, \mathbf{x}\rangle) \neq \mathrm{sign}(\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle)] + \Pr_{(\mathbf{x},y)\sim U_j}[\mathrm{sign}(\langle \mathbf{w}^{(j)}, \mathbf{x}\rangle) = \mathrm{sign}(\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle)] \\
&\leq 0.3 + 0.21 = 0.51 \qquad \text{(by (13))}
\end{aligned}$$

Therefore,

$$\mathrm{err}_{U_j}(\hat{\mathbf{w}}) \geq 0.49. \tag{14}$$

Once again, we can relate this to the error on the distribution $D^{(j)}$ as follows.

$$\begin{aligned}
&\Pr_{(\mathbf{x},y)\sim D^{(j)}}[\mathrm{sign}(\langle \hat{\mathbf{w}}, \mathbf{x}\rangle) \neq \mathrm{sign}(\langle \mathbf{w}^{(j)}, \mathbf{x}\rangle)] \\
&= \Pr_{(\mathbf{x},y)\sim U_j}[\mathrm{sign}(\langle \hat{\mathbf{w}}, \mathbf{x}\rangle) \neq \mathrm{sign}(\langle \mathbf{w}^{(j)}, \mathbf{x}\rangle) \mid |\langle \mathbf{w}^{(j)}, \mathbf{x}\rangle| \geq \gamma] \cdot \Pr_{(\mathbf{x},y)\sim U_j}[|\langle \mathbf{w}^{(j)}, \mathbf{x}\rangle| \geq \gamma] \\
&\geq \Pr_{(\mathbf{x},y)\sim U_j}[\mathrm{sign}(\langle \hat{\mathbf{w}}, \mathbf{x}\rangle) \neq \mathrm{sign}(\langle \mathbf{w}^{(j)}, \mathbf{x}\rangle)] - \Pr_{(\mathbf{x},y)\sim U_j}[|\langle \mathbf{w}^{(j)}, \mathbf{x}\rangle| < \gamma] \\
&\geq 0.49 - p_\gamma \geq 0.29 \qquad \text{(by (14))}
\end{aligned}$$

Therefore, $\mathrm{err}_{D^{(j)}}(\hat{\mathbf{w}}) \geq 0.29$. Thus, for any $\hat{\mathbf{w}}$, if $\mathrm{err}_{D^{(i)}}(\hat{\mathbf{w}}) \leq 0.1$ then $\mathrm{err}_{D^{(j)}}(\hat{\mathbf{w}}) \geq 0.29$. ■