Efficient Private Empirical Risk Minimization for High-dimensional Learning

Shiva Prasad Kasiviswanathan KASIVISW@GMAIL.COM
Samsung Research America, Mountain View, CA 94043

Hongxia Jin HONGXIA.JIN@SAMSUNG.COM
Samsung Research America, Mountain View, CA 94043

Abstract

Dimensionality reduction is a popular approach for dealing with high dimensional data that leads to substantial computational savings. Random projections are a simple and effective method for universal dimensionality reduction with rigorous theoretical guarantees. In this paper, we theoretically study the problem of differentially private empirical risk minimization in the projected subspace (compressed domain). We ask: *is it possible to design differentially private algorithms with small excess risk given access to only projected data?* In this paper, we answer this question in affirmative, by showing that for the class of generalized linear functions, given only the projected data and the projection matrix, we can obtain excess risk bounds of $\tilde{O}(w(\mathcal{C})^{2/3}/n^{1/3})$ under ϵ-differential privacy, and $\tilde{O}(\sqrt{w(\mathcal{C})/n})$ under (ϵ, δ)-differential privacy, where n is the sample size and $w(\mathcal{C})$ is the Gaussian width of the parameter space $\mathcal{C}$ that we optimize over. A simple consequence of these results is that, for a large class of ERM problems, in the traditional setting (i.e., with access to the original data), under ϵ-differential privacy, we improve the worst-case risk bounds of (Bassily et al., 2014).

1. Introduction

Curse of dimensionality is a well-established obstacle in machine learning affecting aspects such as the accuracy, computation time, storage space, and communication cost of many common tasks. Therefore, a general strategy for tackling many high-dimensional learning tasks is first transforming from the data domain to some appropriate compressed measurement domain, and then performing the learning task in the measurement domain. This idea has led to the development of the field of *compressed learning* (Eldar & Kutyniok, 2012), where the general goal is to perform a machine learning task in the compressed measurement domain with almost the same performance guarantee as that achievable in the original data domain. Efficiency gains achievable through compressed learning have been well documented for variety of common machine learning tasks (Arriaga & Vempala, 2006; Davenport et al., 2006; Maillard & Munos, 2009; Fard et al., 2012; Kabán, 2014; Calderbank et al., 2009; Wright et al., 2009). In addition to the efficiency gains, these techniques also work successfully in situations where one cannot observe the data domain because collecting data might be difficult or expensive. In this case, by learning in the compressed domain, one avoids the cost of recovering back the data in the high-dimensional domain.

A common approach for dimensionality reduction (compression) is using the technique of *random projections* (Vempala, 2005; Mahoney, 2011), where the original high-dimensional data is projected onto a lower-dimensional subspace using some appropriately chosen random matrix. Random projections, such as the popular Johnson-Lindenstrauss transform and its variants, preserve the structure of the original data space, and hence can be used as a way to reduce the *cost* of the learning process perceptibly, while preserving the performance approximately. For high-dimensional data, a random projection can lead to substantial savings in resources such as storage space, transmission bandwidth, and processing time.

Machine learning algorithms are frequently run on sensitive data, and this has motivated the study of learning algorithms that have good performance guarantees while satisfying a rigorous notion of privacy called *differential privacy* (Dwork et al., 2006b). There currently exist differentially private algorithms for many statistical and machine-learning tasks such as classification, regression, PCA, clustering, density estimation, among others. We refer the reader to recent surveys by Sarwate and Chaudhuri (Sarwate & Chaudhuri, 2013), and by Dwork and Roth (Dwork & Roth, 2013) for an overview of recent developments in machine learning with differential privacy.

In this paper, we theoretically study private compressed learning in the framework of *Empirical Risk Minimization* (ERM). In the traditional (uncompressed) empirical risk minimization framework, we are given n datapoints

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

$\mathbf{z}_1, \ldots, \mathbf{z}_n$ from some domain $\mathcal{Z}$, and a closed, convex set $\mathcal{C} \subseteq \mathbb{R}^d$, and the goal is minimize $\frac{1}{n}\sum_{i=1}^n f(\theta; \mathbf{z}_i)$ over $\theta \in \mathcal{C}$. The loss function $f : \mathcal{C} \times \mathcal{Z} \to \mathbb{R}$ is the loss associated with a single datapoint. We will generally assume that $f(\cdot; \mathbf{z})$ is convex and λ-Lipschitz[1] for all $\mathbf{z} \in \mathcal{Z}$.

In this paper, we focus on a popular class of loss functions, called *generalized linear functions* (Shalev-Shwartz et al., 2009; Jain & Thakurta, 2014; Ullman, 2015), where the loss function $f(\theta; \mathbf{z})$ has a generalized linear form: $f(\theta; \mathbf{z}) = \ell(\langle \mathbf{x}, \theta \rangle; y)$, with $\mathbf{z} = (\mathbf{x}, y)$ and $\mathcal{Z} \subseteq \mathbb{R}^d \times \mathbb{R}$. We will assume that $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is convex with respect to its first argument (over domain $\mathbb{R}$). This formulation, that has also been used in the previous works on differentially private convex ERM (Kifer et al., 2012; Jain & Thakurta, 2014; Bassily et al., 2014; Ullman, 2015), captures the supervised learning of a linear predictor with a convex loss function, where $\ell(\langle \mathbf{x}, \theta \rangle; y)$ is the loss of predicting $\langle \mathbf{x}, \theta \rangle$ when the true target is y.

Given n datapoints (samples), $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ drawn from some universe $\mathbb{R}^d \times \mathbb{R}$, the M-estimator $\hat{\theta}$ associated with a given a function

$$\mathcal{L}(\theta; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) = \frac{1}{n}\sum_{i=1}^n \ell(\langle \mathbf{x}_i, \theta \rangle; y_i)$$

is defined as:

$$\hat{\theta} \in \text{argmin}_{\theta \in \mathcal{C}} \mathcal{L}(\theta; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)). \quad (1)$$

As mentioned above, this type of program captures a variety of important learning problems, e.g., the MLE (Maximum Likelihood Estimators) for linear regression is captured by setting $\ell(\langle \mathbf{x}, \theta \rangle; y) = (y - \langle \mathbf{x}, \theta \rangle)^2$. Similarly, the MLE for logistic regression is captured by setting $\ell(\langle \mathbf{x}, \theta \rangle; y) = \ln(1 + \exp(-y\langle \theta, \mathbf{x} \rangle))$. Another common example is the support vector machine (SVM), where $\ell(\langle \mathbf{x}, \theta \rangle; y) = \text{hinge}(y\langle \theta, \mathbf{x} \rangle)$, where $\text{hinge}(a) = 1 - a$ if $a \leq 1$ and 0 otherwise.

In the empirical risk minimization setting, the success of an algorithm is measured by worst-case (over inputs $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$) excess empirical risk, defined as:

$$\mathcal{L}(\tilde{\theta}; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)), \quad (2)$$

where $\tilde{\theta}$ is the output of the algorithm. We would like to guarantee that with high probability (or expectation over coin tosses of the algorithm) the excess empirical risk is small. For generalized linear functions, a bound on excess empirical risk also translates into a bound on the generalization error, using a generic conversion theorem of (Shalev-Shwartz et al., 2009) (details omitted here).

[1]Throughout this paper, we measure Lipschitzness with respect to L_2-norm in the input space.

In the traditional (uncompressed) setting, when the algorithm gets access to the true data, private convex ERM mechanisms have been investigated extensively in the literature under both ϵ- and (ϵ, δ)-differential privacy notions. Starting with the results of Chaudhuri *et al.* (Chaudhuri & Monteleoni, 2009; Chaudhuri et al., 2011) private convex ERM problems have been studied in various settings including the low-dimensional setting (Rubinstein et al., 2009; Kifer et al., 2012), high-dimensional sparse regression setting (Kifer et al., 2012; Smith & Thakurta, 2013), online learning setting (Jain et al., 2012; Thakurta & Smith, 2013; Jain & Thakurta, 2014; Mishra & Thakurta, 2015), local privacy setting (Duchi et al., 2013), interactive setting (Jain & Thakurta, 2013; Ullman, 2015), and streaming setting (Kasiviswanathan et al., 2016). Bassily *et al.* (Bassily et al., 2014) showed that for a general convex loss function $f(\theta; \mathbf{z})$ that for every $\mathbf{z}$ is 1-Lipschitz the expected excess empirical is at most $\tilde{O}(\sqrt{d}/n)$ under (ϵ, δ)-differential privacy and $O(d/n)$ under ϵ-differential privacy (ignoring the dependence on other parameters for simplicity).[2] They also showed that these bounds cannot be improved in general, even for generalized linear functions. Talwar *et al.* (Talwar et al., 2015a;b) recently showed that for a large class of ERM problems, under (ϵ, δ)-differential privacy, the above worst-case bound can be improved by exploiting properties of the parameter space $\mathcal{C}$. Somewhat surprisingly, even though based on different techniques, like our results, the bounds presented in (Talwar et al., 2015a;b) also depend on the, Gaussian width of $\mathcal{C}$ (a geometric quantity, defined below). Understanding the exact role of Gaussian width in differentially private ERM is an interesting open question. In the non-private world, Gaussian width (sometimes called, Gaussian complexity), plays an important role in statistical learning theory, and has been used as a measure of complexity of a function class in statistical learning theory, see (Mendelson, 2004).

Compressed Learning Problem Formulation. Let $\Phi \in \mathbb{R}^{m \times d}$ be a random projection matrix. In this paper, we use the popular and wide class of subgaussian matrices for random projection. We define the measurement domain $\mathcal{M}$ as: $\mathcal{M} = \{(\Phi\mathbf{x}, y) : (\mathbf{x}, y) \in \mathcal{Z}\}$. In other words, $\mathcal{M}$ is a compressed representation of $\mathcal{Z}$. Our goal is to output an estimator (while satisfying the constraints of differential privacy) that minimizes the empirical risk (2) given access to only to the compressed representation of $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ (i.e., $(\Phi\mathbf{x}_1, y_1), \ldots, (\Phi\mathbf{x}_n, y_n)$) and Φ.[3] This compressed setting is a strict generalization of the traditional setting where the algorithm gets access to $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$. Our results also trivially hold for the traditional setting, as given $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, we can pick Φ and generate the input $(\Phi\mathbf{x}_1, y_1), \ldots, (\Phi\mathbf{x}_n, y_n)$ for our proposed algorithm.

[2]Better bounds can be achieved for strongly convex loss functions (Bassily et al., 2014; Talwar et al., 2015a).

[3]In general, given $\Phi\mathbf{x}$, it is not possible to reconstruct $\mathbf{x}$ without further assumptions on $\mathbf{x}$.

Problem Setting	ϵ-differential privacy	(ϵ, δ)-differential privacy
Compressed Learning (under random projections)	$\tilde{O}\left(\frac{\psi^{4/3}\lambda_\ell w(\mathcal{C})^{2/3}\|\mathcal{C}\|_2}{(\epsilon n)^{1/3}}\right)$	$\tilde{O}\left(\frac{\psi\lambda_\ell\sqrt{w(\mathcal{C})}\|\mathcal{C}\|_2\log^2(1/\delta)}{\sqrt{\epsilon n}}\right)$
Traditional (uncompressed) Learning	$\tilde{O}\left(\frac{\lambda_\ell w(\mathcal{C})^{2/3}\|\mathcal{C}\|_2}{(\epsilon n)^{1/3}}\right)$	$\tilde{O}\left(\frac{\Gamma_\ell^{1/3}(\lambda_\ell w(\mathcal{C}))^{2/3}\log^2(1/\delta)}{(\epsilon n)^{2/3}}\right)$ (Talwar et al., 2015b)

Table 1. Upper bounds on the excess risk under differential privacy for the convex ERM problem: $\min_{\theta\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^n \ell(\langle \mathbf{x}_i, \theta\rangle; y_i)$ for λ_ℓ-Lipschitz convex ℓ and general convex set $\mathcal{C}$. Here $w(\mathcal{C})$ is the Gaussian width of $\mathcal{C}$, $\|\mathcal{C}\|_2$ is the diameter of $\mathcal{C}$, ψ is the subgaussian norm of random vectors used in the projection matrix, and Γ_ℓ is the curvature constant of ℓ. For traditional (uncompressed) learning, under (ϵ, δ)-differential privacy, better bounds can be obtained for some specific $\mathcal{C}$'s (Talwar et al., 2015a;b). All the uncited results appear in this paper.

Our Contributions. We primarily make two contributions in this paper. A summary of our results appear in Table 1. All the presented algorithms run in time polynomial in n and d.

(a) Compressed Learning. A natural first question is whether *non-trivial*[4] private ERM algorithms exist that operate only with compressed data. In this paper, we answer this question in affirmative. Using techniques from convex geometry and high-dimensional estimation, we present a generic mechanism that transforms any differentially private ERM algorithm to provide excess risk bounds from compressed data. The idea is to add noise for privacy in the compressed domain, and then "lift" the result back to the original domain. Our analysis is based on exploiting the geometric structure of $\mathcal{C}$. The geometric parameter, *Gaussian width*, defined as $w(\mathcal{C}) = \mathbb{E}_{\mathbf{g}\in\mathcal{N}(0,1)^d}[\sup_{\theta\in\mathcal{C}}\langle\theta, \mathbf{g}\rangle]$,[5] plays an important role in our analysis, and shows up repeatedly as a geometric measure of the size of the set $\mathcal{C}$. For the important class of generalized linear functions, using the private ERM algorithms from (Bassily et al., 2014), we obtain an excess empirical risk bound of $\approx w(\mathcal{C})^{2/3}/n^{1/3}$ for achieving ϵ-differential privacy, and $\approx \sqrt{w(\mathcal{C})/n}$ for achieving (ϵ, δ)-differential privacy.

(b) Traditional Learning (under ϵ-differential privacy). Our bounds on compressed learning also directly translate to the traditional ERM setting. We obtain the first excess risk bounds, under ϵ-differential privacy, that go beyond the worst-case bounds of $O(d/n)$ from (Bassily et al., 2014). The techniques presented by (Talwar et al., 2015a;b) provide *only* (ϵ, δ)-differential privacy guarantees (i.e., δ has to be greater than 0). There are many simple consequences of our result. For example, we obtain an ϵ-differentially private Lasso algorithm with excess risk $\tilde{O}(1/n^{1/3})$, assuming all the input datapoints have bounded L_2-norm. Previous best bounds here (from (Bassily et al., 2014)) had a polynomial dependence on the dimension.

[4]There is a trivial private ERM algorithm that ignores the input and outputs any $\theta \in \mathcal{C}$. The excess empirical risk of this algorithm is $2\lambda_{\mathcal{L}}\|\mathcal{C}\|_2$, where $\lambda_{\mathcal{L}}$ is the Lipschitz constant of $\mathcal{L}$. All bounds presented in this paper, as also true for all other previous results in the private ERM literature, are only interesting in the regime where they are less than this trivial bound.

[5]For a $\mathcal{C}$ contained in a unit L_2-ball, $w(\mathcal{C})$ is at most $O(\sqrt{d})$, and in general is significantly smaller, see Section 2.

Other Related Work. (Jain & Thakurta, 2014) gave dimension-independent expected excess risk bounds for the case generalized linear functions with a strongly convex regularizer and assuming that $\mathcal{C} = \mathbb{R}^d$ (unconstrained optimization). Note that in this paper, we do not make any strong convexity assumptions and our results hold for the more interesting constrained optimization.

Kifer *et al.* (Kifer et al., 2012), and Smith and Thakurta (Smith & Thakurta, 2013) studied the problem of releasing Lasso estimator privately under certain assumptions about the instance (*restricted strong convexity* and *mutual incoherence*). Under these assumptions, they obtain excess risk of $O(\text{polylog}(d)/n)$. However, these assumptions on the data are rather strong and may not hold in practice (Talwar et al., 2015a;b). Without these data dependent assumptions, (Talwar et al., 2015a;b) present an (ϵ, δ)-differentially private Lasso algorithm with excess risk of $\tilde{O}(\log(d)/n^{1/3})$. However, the question of designing a private Lasso estimator under ϵ-differential privacy whose excess risk grows logarithmically in the dimension size d was still *open*, which we resolve here. We perform a more detailed comparison with the results of Talwar *et al.* in Section 4.

Random projections have been used as a tool to design (differentially) private algorithms in many other problem settings. Blocki *et al.* (Blocki et al., 2012) have shown that if Φ is a Gaussian Johnson-Lindenstrauss matrix of appropriate dimension, then ΦX is differentially private if the least singular value of X is "sufficiently" large. Here $X \in \mathbb{R}^{n\times d}$ is a data matrix of n points in d dimensions. Note that ΦX does not reduce the dimensionality of the data but rather generates a set of fewer datapoints. The bound on the least singular value was recently improved by (Sheffet, 2015). Wang *et al.* (Wang et al., 2015) show that for the problem of subspace clustering achieving, random projections can be helpful for both computational efficiency and privacy protection. Kenthapadi *et al.* (Kenthapadi et al., 2013) use Johnson-Lindenstrauss transform to publish a private sketch that enables estimation of the distance between users. Zhou *et al.* (Zhou et al., 2009b) provide a technique of generating synthetic data using random linear

or affine transformations. Random projections have also been used to achieve various other weaker privacy guarantees (Duncan et al., 1991; Zhou et al., 2009a).

2. Preliminaries

Notation and Data Normalization. We denote $[n] = \{1, \ldots, n\}$. Vectors are in column-wise fashion, denoted by boldface letters. For a vector $\mathbf{v}$, $\|\mathbf{v}\|_p$ for $(1 \leq p \leq \infty)$ denotes its L_p-norm. For $p = 2$, we drop the subscript and use $\|\mathbf{v}\|$ to denote its Euclidean (L_2-) norm. The d-dimensional unit ball in L_p-norm centered at origin is denoted by B_p^d, and the d-dimensional Euclidean unit sphere is denoted by S^{d-1}. For a variable n, we use $\mathrm{poly}(n)$ to denote a polynomial function of n and $\mathrm{polylog}(n)$ to denote $\mathrm{poly}(\log(n))$.

Throughout this paper, we assume that all the inputs are L_2-normalized with $\|\mathbf{x}\| \leq 1$ and $|y| \leq 1$. This is for simplicity, and our results extend to the case without this normalization (with probably an increased Lipschitz constant). Also as is typical in random projection based analyses, we assume that the $\mathbf{x}_i$'s are selected independent of the projection matrix Φ.

We refer the reader to (Boyd & Vandenberghe, 2004) for standard definitions in convex optimization. For a set of vectors, we define its diameter as the maximum attained L_2-norm in the set.

Definition 1. *(Diameter of Set) The diameter of a closed set $\mathcal{C} \subseteq \mathbb{R}^d$, is defined as $\|\mathcal{C}\|_2 = \sup_{\theta \in \mathcal{C}} \|\theta\|$.*

Definition 2. *(Lipschitz Functions over θ) A loss function $f : \mathcal{C} \times \mathcal{Z} \to \mathbb{R}$ is λ_f-Lipschitz with respect to θ over the domain $\mathcal{C}$, if for any $\mathbf{z} \in \mathcal{Z}$, and $\theta_a, \theta_b \in \mathcal{C}$, we have $|f(\theta_a; \mathbf{z}) - f(\theta_b; \mathbf{z})| \leq \lambda_f \|\theta_a - \theta_b\|$.*

For a generalized linear function $\ell(\langle \mathbf{x}, \theta \rangle; y)$, let λ_ℓ denote the Lipschitz constant of ℓ in the first argument. The following claim easily follows from our normalization (proof in Appendix A, Supplementary material).

Claim 2.1. *Let function $\ell(\langle \mathbf{x}, \theta \rangle; y)$ be λ_ℓ-Lipschitz in the first argument over the domain $\mathbb{R}$. Then ℓ is λ_ℓ-Lipschitz with respect to θ.*

For common loss functions, λ_ℓ is small. For squared loss, used in linear regression, the loss function is $2(\|\mathcal{C}\|_2 + 1)$-Lipschitz; for logistic loss, used in logistic regression, the loss function is 1-Lipschitz; for hinge loss, used in SVM, the loss function is 1-Lipschitz.

Subgaussian matrices are a popular choice for random projections (matrix Φ).

Definition 3 (Subgaussian Random Variable and Vector)**.** *We call a random variable $a \in \mathbb{R}$ subgaussian if there exists a constant $C > 0$ if $\Pr[|a| > t] \leq 2\exp(-t^2/C^2)$ for all $t > 0$. We say that a random vector $\mathbf{a} \in \mathbb{R}^d$ is subgaussian if the one-dimensional marginals $\langle \mathbf{a}, \mathbf{b} \rangle$ are subgaussian random variables for all $\mathbf{b} \in \mathbb{R}^d$.*

The class of subgaussian random variables includes many random variables that arise naturally in data analysis, such as Gaussian, Bernoulli, spherical, bounded (where the random variable a satisfies $|a| \leq M$ *almost surely* for some fixed M). The natural generalizations of these random variables to higher dimension are all subgaussian random vectors. For many *isotropic convex sets*[6] $\mathcal{K}$ (such as the hypercube), a random vector $\mathbf{a}$ uniformly distributed in $\mathcal{K}$ is subgaussian.

Definition 4 (Norm of Subgaussian Random Variable and Vector)**.** *The ψ_2-norm of a subgaussian random variable $a \in \mathbb{R}$, denoted by $\|a\|_{\psi_2}$ is:*

$$\|a\|_{\psi_2} = \inf \left\{ t > 0 : \mathbb{E}[\exp(|a|^2/t^2)] \leq 2 \right\}.$$

The ψ_2-norm of a subgaussian random vector $\mathbf{a} \in \mathbb{R}^d$ is:

$$\|\mathbf{a}\|_{\psi_2} = \sup_{\mathbf{b} \in S^{d-1}} \|\langle \mathbf{a}, \mathbf{b} \rangle\|_{\psi_2}.$$

Our analysis is based on exploiting the geometric properties of $\mathcal{C}$. We use the well-studied quantity of *Gaussian width* that captures the L_2-geometric complexity of $\mathcal{C}$.

Definition 5 (Gaussian Width)**.** *Given a closed set $S \subseteq \mathbb{R}^d$, its Gaussian width $w(S)$ is defined as:*

$$w(S) = \mathbb{E}_{\mathbf{g} \in \mathcal{N}(0,1)^d} [\sup_{\mathbf{a} \in S} \langle \mathbf{a}, g \rangle].$$

In particular, $w(S)^2$ can be thought as the "effective dimension" of S. Many popular convex sets have low Gaussian width, e.g., the width of both the unit L_1-ball in $\mathbb{R}^d$ (B_1^d) and the standard d-dimensional probability simplex are both $\Theta(\sqrt{\log d})$, and the width of any ball B_p^d for $1 \leq p \leq \infty$ is $\approx d^{1-1/p}$. For a set $\mathcal{C}$ contained in the B_2^d, $w(\mathcal{C})$ is always $O(\sqrt{d})$.

Differential Privacy Background. Differential privacy is a rigorous notion of privacy that emerged from a long line of work in theoretical computer science (Dwork et al., 2006b). We say two datasets D and D' of size n are neighbors if they differ in one entry.

Definition 6. *(Dwork et al., 2006b;a) A randomized algorithm Alg is (ϵ, δ)-differentially private if for all neighboring datasets D, D' and for all events $\mathcal{R}$ in the output space of Alg, we have $\Pr[Alg(D) \in \mathcal{R}] \leq \exp(\epsilon) \cdot \Pr[Alg(D') \in \mathcal{R}] + \delta$, where the probability is taken over the randomness of the algorithm. When $\delta = 0$, the Algorithm is ϵ-differentially private.*

3. Algorithm for the Compressed Setting

In this section, we present a generic mechanism (Mechanism PROJERM) for private risk minimization on compressed data. Instantiating the generic mechanism with

[6] A convex set $\mathcal{K}$ in $\mathbb{R}^d$ is called isotropic if a random vector chosen uniformly from $\mathcal{K}$ according to the volume is isotropic. A random vector $\mathbf{a} \in \mathbb{R}^d$ is isotropic if for all $\mathbf{b} \in \mathbb{R}^d$, $\mathbb{E}[\langle \mathbf{a}, \mathbf{b} \rangle^2] = \|\mathbf{b}\|^2$.

recent differentially private algorithms of (Bassily et al., 2014) provides our main upper bounds (Theorem 3.11). Remember that in the compressed setting the algorithm gets access (to only) $(\Phi\mathbf{x}_1, y_1), \ldots, (\Phi\mathbf{x}_n, y_n)$ and the matrix Φ, and the goal of the algorithm is to privately output an estimator that minimizes the empirical risk (2). In the next section, we discuss a simple extension of our results to the traditional (uncompressed) setting and its consequences. Missing proofs from this section are presented in Appendix B (Supplementary material).

We focus on a large class of optimization problems expressed as a sum of a generalized linear loss function over individual datapoints.

In the following, let $\hat{\theta}$ be the true minimizer of $\mathcal{L}(\theta; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) = \frac{1}{n}\sum_{i=1}^{n} \ell(\langle \mathbf{x}_i, \theta\rangle; y_i)$, i.e.,

$$\hat{\theta} \in \text{argmin}_{\theta\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} \ell(\langle \mathbf{x}_i, \theta\rangle; y_i). \tag{3}$$

Let ℓ be λ_ℓ-Lipschitz in the first argument. By Claim 2.1, ℓ is λ_ℓ-Lipschitz with respect to θ over the domain $\mathcal{C}$. The Lipschitz constant of $\mathcal{L}$ ($\lambda_{\mathcal{L}}$) satisfies: $\lambda_{\mathcal{L}} \leq \lambda_\ell$.

One of the most celebrated result in geometry, the Johnson-Lindenstrauss (JL) Lemma, states that for any set $S \subseteq \mathbb{R}^d$, given $\gamma > 0$ and $m = \Omega(\log|S|/\gamma^2)$, there exists a map that embeds the set into $\mathbb{R}^m$, distorting all pairwise distances within at most $1 \pm \gamma$ factor. This transform has become a fundamental tool in dimensionality reduction, and there are several constructions of the map based on random projections (Bourgain et al., 2015).

The bound on m in JL lemma is optimal for an arbitrary set S, however if S, has a special structure then the dependence of m can be improved. This was first investigated by Gordon (Gordon, 1988), who showed that one can embed a set of points S in $\mathbb{R}^d$ into a much lower-dimensional space $\mathbb{R}^m$ using a Gaussian matrix while approximately preserving the Euclidean norms of the vectors in S, provided that Gaussian width of S is small. This result has been used in several interesting applications in high-dimensional convex geometry, statistics, and compressed sensing.

We use the following generalization of Gordon's theorem by Dirksen (Dirksen, 2014), which is based on using subgaussian matrices for the projection.

Theorem 3.1 ((Dirksen, 2014)). *Let $\tilde{\Phi}$ be an $m \times d$ random matrix, whose rows $\phi_1^\top, \ldots, \phi_m^\top$ are i.i.d., mean-zero, isotropic, subgaussian random vectors in $\mathbb{R}^d$ with $\psi = \|\phi_i\|_{\psi_2}$. Let $\Phi = \tilde{\Phi}/\sqrt{m}$. Let S be a set of points in $\mathbb{R}^d$. There is a constant $C > 0$ such that for any $0 < \gamma, \beta < 1$,*

$$\Pr\left[\sup_{\mathbf{a}\in S} \left|\|\Phi\mathbf{a}\|^2 - \|\mathbf{a}\|^2\right| \geq \gamma\|\mathbf{a}\|^2\right] \leq \beta,$$

provided that $m \geq \frac{C\psi^4}{\gamma^2}\max\left\{w(S)^2, \log\left(\frac{1}{\beta}\right)\right\}$.

As a simple corollary to the above theorem it also follows that,

Corollary 3.2. *Under the setting of Theorem 3.1, there exists a constant $C' > 0$ such that for any $0 < \gamma, \beta < 1$*

$$\Pr\left[\sup_{\mathbf{a},\mathbf{b}\in S} |\langle \Phi\mathbf{a}, \Phi\mathbf{b}\rangle - \langle \mathbf{a}, \mathbf{b}\rangle| \geq \gamma\|\mathbf{a}\|\|\mathbf{b}\|\right] \leq \beta,$$

provided that $m \geq \frac{C'\psi^4}{\gamma^2}\max\left\{w(S)^2, \log\left(\frac{1}{\beta}\right)\right\}$.

Throughout the rest of this paper, we are going to assume that $\gamma < 1$ (γ in fact will be set much smaller). Consider the following modified version of the optimization objective from (3),

$$\mathcal{L}_{\text{comp}}(\theta; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n); \Phi) = \frac{1}{n}\sum_{i=1}^{n} \ell(\langle \Phi\mathbf{x}_i, \Phi\theta\rangle; y_i). \tag{4}$$

Since in the compressed setting, we have access to $(\Phi\mathbf{x}_i, y_i)$'s, we can solve (4). Our idea will be to privately minimize (4) problem over the domain $\Phi\mathcal{C}$, and use the result to bound the excess risk. Let $\Phi\mathcal{C} = \{\vartheta \in \Phi\theta : \theta \in \mathcal{C}\}$. Note for a convex $\mathcal{C}$, $\Phi\mathcal{C}$ is also convex.

Let $\lambda_{\mathcal{L}_{\text{comp}}}$ denote the Lipschitz constant of the function $\ell(\langle \Phi\mathbf{x}, \vartheta\rangle; y)$ with respect to ϑ over the domain $\Phi\mathcal{C}$:

$$\lambda_{\mathcal{L}_{\text{comp}}} = \sup_{\substack{(\mathbf{x},y)\in(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)\\ \vartheta_a,\vartheta_b\in\Phi\mathcal{C}}} \frac{|\ell(\langle \Phi\mathbf{x}, \vartheta_a\rangle; y) - \ell(\langle \Phi\mathbf{x}, \vartheta_b\rangle; y)|}{\|\vartheta_a - \vartheta_b\|}.$$

Note that in the above definition of $\lambda_{\mathcal{L}_{\text{comp}}}$, $(\mathbf{x}, y)$ is restricted to any *fixed* collection of n elements. The following lemma shows for large enough m, with probability at least $1 - \beta$, that $\lambda_{\mathcal{L}_{\text{comp}}}$ is at most $2\lambda_\ell$. Remember that the Lipschitz constant of $\mathcal{L}$ is at most λ_ℓ.

Lemma 3.3. *Let Φ be a random matrix as defined in Theorem 3.1 with $m = \Theta((\psi^4/\gamma^2)\log(n/\beta))$ for $\beta > 0$. Then with probability, at least $1 - \beta$, the Lipschitz constant of $\mathcal{L}_{\text{comp}}$ ($\lambda_{\mathcal{L}_{\text{comp}}}$) is at most $2\lambda_\ell$ with respect to ϑ over the domain $\Phi\mathcal{C}$.*

The following lemma follows from Lipschitz properties of ℓ and Theorem 3.1.

Lemma 3.4. *Let Φ be a random matrix as defined in Theorem 3.1 with $m = \Theta((\psi^4/\gamma^2)\log(n/\beta))$ for $\beta > 0$. Then with probability at least $1 - \beta$, for every $(\mathbf{x}_i, y_i)$*

$$\ell(\langle \Phi\mathbf{x}_i, \Phi\hat{\theta}\rangle; y_i) \leq \ell(\langle \mathbf{x}_i, \hat{\theta}\rangle; y_i) + \lambda_\ell\gamma\|\mathcal{C}\|_2.$$

From Lemma 3.4, the following lemma is immediate.

Lemma 3.5. *Let Φ be a random matrix as defined in Theorem 3.1 with $m = \Theta((\psi^4/\gamma^2)\log(n/\beta))$ for $\beta > 0$. Then*

Mechanism 1 PROJERM

Input: A random subgaussian matrix $\Phi \in \mathbb{R}^{m\times d}$, and a dataset $D = (\Phi\mathbf{x}_1, y_1), \ldots, (\Phi\mathbf{x}_n, y_n)$ of n datapoints from the domain $\mathcal{M}_\Phi = \{(\Phi\mathbf{x}, y) : \mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\| \leq 1, y \in \mathbb{R}, |y| \leq 1\}$

Output: θ^{priv} a differentially private estimate of $\hat{\theta} \in \text{argmin}_{\theta\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^n \ell(\langle \mathbf{x}_i, \theta\rangle; y_i)$

1. Let $\vartheta^{\text{priv}} \leftarrow$ Output of an (ϵ, δ)-differentially private or an ϵ-differentially private ERM algorithm solving the following problem:
$$\text{argmin}_{\vartheta\in\Phi\mathcal{C}} \frac{1}{n}\sum_{i=1}^n \ell(\langle \Phi\mathbf{x}_i, \vartheta\rangle; y_i)$$
2. $\theta^{\text{priv}} \leftarrow \text{argmin}_{\theta\in\mathbb{R}^d} \|\theta\|_{\mathcal{C}}$ subject to $\Phi\theta = \vartheta^{\text{priv}}$ (can be solved with any convex programming technique)
3. **Return:** θ^{priv}

with probability at least $1-\beta$,

$$\min_{\theta\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^n \ell(\langle \Phi\mathbf{x}_i, \Phi\theta\rangle; y_i) \leq \frac{1}{n}\sum_{i=1}^n \ell(\langle \mathbf{x}_i, \hat{\theta}\rangle; y_i) + \lambda_\ell\gamma\|\mathcal{C}\|_2.$$

We now present a generic mechanism (Mechanism PROJERM) for privately releasing M-estimator based on minimizing (4) under differential privacy. The mechanism is simple: it "solves" the private ERM problem in the projected subspace, and then "lifts" back the result to the original dimension by solving a linear estimation problem. As we will see, working in the compressed domain has the added advantage that the noise added for privacy scales as $\approx \sqrt{m}$.

Mechanism PROJERM is (ϵ, δ)- or ϵ-differentially private based on the algorithm used in Step 1. Post-processing of the output does not affect the differential privacy guarantee. Also note that Mechanism PROJERM is differentially private independent of the choice of Φ. However, the utility guarantee of Mechanism PROJERM (Theorem 3.11) will require Φ to be a subgaussian matrix (as in Theorem 3.1) with a lower bound on m that will depend among various parameters, the geometry of $\mathcal{C}$.

In the Step 1 of Mechanism PROJERM any (ϵ, δ)- or ϵ-differentially private ERM algorithm can be used. In Proposition 3.7, we provide bounds obtained by using the differentially private ERM algorithms of (Bassily et al., 2014). Their (ϵ, δ)-differentially private ERM algorithm is based on a noisy stochastic variant of the classic gradient descent algorithm. While their ϵ-differentially private ERM algorithm is based on a polynomial time implementation of the exponential mechanism of McSherry and Talwar (McSherry & Talwar, 2007).

Theorem 3.6 ((Bassily et al., 2014), Theorem 2.4)**.** *Let $f(\theta; \mathbf{z})$ be a convex loss function that is λ-Lipschitz with respect to θ over the domain $\mathcal{C} \subseteq \mathbb{R}^d$.*

1. There exists an (ϵ, δ)-differentially private algorithm $\mathcal{A}_{\epsilon,\delta}$ that on an input dataset $\mathbf{z}_1, \ldots, \mathbf{z}_n$ outputs $\tilde{\theta} \in \mathcal{C}$ such that for any $\beta > 0$, with probability at least $1-\beta$,

$$\frac{1}{n}\sum_{i=1}^n f(\tilde{\theta}; \mathbf{z}_i) - \min_{\theta\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^n f(\theta; \mathbf{z}_i) = O\left(\frac{\lambda\sqrt{d}\|\mathcal{C}\|_2 \log^{3/2}(n/\delta)\sqrt{\log(1/\delta)}\,\text{polylog}(1/\beta)}{n\epsilon}\right).$$

2. There exists an ϵ-differentially private algorithm $\mathcal{A}_\epsilon$ that on an input dataset $\mathbf{z}_1, \ldots, \mathbf{z}_n$ outputs $\tilde{\theta} \in \mathcal{C}$ such that for any $\beta > 0$, with probability at least $1-\beta$,

$$\frac{1}{n}\sum_{i=1}^n f(\tilde{\theta}; \mathbf{z}_i) - \min_{\theta\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^n f(\theta; \mathbf{z}_i) = O\left(\frac{\lambda d\|\mathcal{C}\|_2\,\text{polylog}(1/\beta)}{n\epsilon}\right).$$

The following proposition follows by combining Lemma 3.5 with Theorem 3.6. The proposition shows that, even for a small projected dimension m, the value of the function $\mathcal{L}_{\text{comp}}(\theta; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n); \Phi)$ at $\theta = \theta^{\text{priv}}$ approximates the minimum empirical risk.

Proposition 3.7. *Let Φ be a random matrix as defined in Theorem 3.1 with $m = \Theta((\psi^4/\gamma^2)\log(n/\beta))$ for $\beta > 0$. Then the output θ^{priv} of Mechanism* PROJERM*,*

1. When invoked (in Step 1) with Algorithm $\mathcal{A}_{\epsilon,\delta}$ from Theorem 3.6, with probability at least $1-\beta$ satisfies:

$$\mathcal{L}_{\text{comp}}(\theta^{\text{priv}}; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n); \Phi) - \mathcal{L}(\hat{\theta}; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) = O\left(\frac{\lambda_{\mathcal{L}_{\text{comp}}}\sqrt{m}\|\Phi\mathcal{C}\|_2 \log^{3/2}(n/\delta)\sqrt{\log(1/\delta)}\,\text{polylog}(1/\beta)}{n\epsilon}\right) + \lambda_\ell\gamma\|\mathcal{C}\|_2.$$

2. When invoked (in Step 1) with Algorithm $\mathcal{A}_\epsilon$ from Theorem 3.6, with probability at least $1-\beta$ satisfies:

$$\mathcal{L}_{\text{comp}}(\theta^{\text{priv}}; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n); \Phi) - \mathcal{L}(\hat{\theta}; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) = O\left(\frac{\lambda_{\mathcal{L}_{\text{comp}}} m\|\Phi\mathcal{C}\|_2\,\text{polylog}(1/\beta)}{n\epsilon}\right) + \lambda_\ell\gamma\|\mathcal{C}\|_2.$$

Existence of θ^{priv}. To lift ϑ^{priv} into the original d-dimensional space, we use recent results about reconstruction from linear observations (Vershynin, 2014). Since $\vartheta^{\text{priv}} \in \Phi\mathcal{C}$, we know that there exists a $\theta^{\text{true}} \in \mathcal{C}$, such that $\Phi\theta^{\text{true}} = \vartheta^{\text{priv}}$. Then the goal is to estimate θ^{true} from $\Phi\theta^{\text{true}}$. Again geometry of $\mathcal{C}$ (Gaussian width) plays an important role, as it controls the diameter of high-dimensional random sections of $\mathcal{C}$ (referred to as $M^\star$ bound (Ledoux & Talagrand, 2013; Vershynin, 2014)). We refer the reader to the excellent tutorial by Vershynin (Vershynin, 2014) for more details.

We define Minkowski functional, as commonly used in geometric functional analysis and convex analysis.

Definition 7 (Minkowski functional)**.** *For any vector $\theta \in \mathbb{R}^d$, the Minkowski functional of $\mathcal{C}$ (a closed set) is the non-negative number $\|\theta\|_{\mathcal{C}}$ defined by the rule: $\|\theta\|_{\mathcal{C}} = \inf\{\tau \in \mathbb{R} : \theta \in \tau\mathcal{C}\}$.*

For the typical situation in ERM problems, where $\mathcal{C}$ is a symmetric convex body, then $\|\cdot\|_{\mathcal{C}}$ defines a norm.

The optimization problem solved in Step 2 of Mechanism ProjERM is convex if $\mathcal{C}$ is convex, and in fact is a linear program is $\mathcal{C}$ is a polytope, and hence can be efficiently solved. The existence of θ^{priv} follows from Theorem 3.8, which in fact can be used to bound the distance between θ^{priv} and θ^{true} (Corollary 3.9).

Theorem 3.8. *(Vershynin, 2014; Mendelson et al., 2007) Let Φ be an $m \times d$ matrix, whose rows $\phi_1^\top, \ldots, \phi_m^\top$ are i.i.d., mean zero, isotropic and subgaussian random vectors in $\mathbb{R}^d$ with $\psi = \|\phi_i\|_{\psi_2}$. Let $\mathcal{C}$ be a convex set. Given $\mathbf{v} = \Phi\mathbf{u}$ and Φ, let $\hat{\mathbf{u}}$ be the solution to the following convex program: $\min_{\mathbf{u}' \in \mathbb{R}^d} \|\mathbf{u}'\|_{\mathcal{C}}$ subject to $\Phi\mathbf{u}' = \mathbf{v}$. Then for any $\beta > 0$, with probability at least $1 - \beta$,*

$$\sup_{\mathbf{u}:\mathbf{v}=\Phi\mathbf{u}} \|\mathbf{u} - \hat{\mathbf{u}}\| = O\left(\frac{\psi^4 w(\mathcal{C})}{\sqrt{m}} + \frac{\psi^4 \|\mathcal{C}\|_2 \sqrt{\log(1/\beta)}}{\sqrt{m}}\right).$$

Corollary 3.9. *If $\Phi \in \mathbb{R}^{m\times d}$ is a subgaussian matrix (as in Theorem 3.1), then for any $\beta > 0$, with probability at least $1 - \beta$, $\|\theta^{\text{true}} - \theta^{\text{priv}}\| = O\left(\frac{\psi^4 w(\mathcal{C})}{\sqrt{m}} + \frac{\psi^4 \|\mathcal{C}\|_2 \sqrt{\log(1/\beta)}}{\sqrt{m}}\right)$.*

One last thing to be verified is that θ^{priv} generated by Mechanism ProjERM is in $\mathcal{C}$. This is simple as by definition of Minkowski functional, as for any closed set $\mathcal{C} = \{\theta \in \mathbb{R}^d : \|\theta\|_{\mathcal{C}} \le 1\}$. Hence, $\|\theta^{\text{true}}\|_{\mathcal{C}} \le 1$. By choice of θ^{priv} in Step 2, ensures that $\|\theta^{\text{priv}}\|_{\mathcal{C}} \le \|\theta^{\text{true}}\|_{\mathcal{C}} \le 1$, which (by definition) shows that $\theta^{\text{priv}} \in \mathcal{C}$.

Now that we have established that $\theta^{\text{priv}} \in \mathcal{C}$ exists, we lower bound $\mathcal{L}_{\text{comp}}(\theta^{\text{priv}}; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n); \Phi)$ in terms of $\mathcal{L}(\theta^{\text{priv}}; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$. Later we will use this lower bound along with Proposition 3.7 to bound the excess empirical risk.

Lemma 3.10. *Let Φ be a random matrix as defined in Theorem 3.1 with $m = \Theta((\psi^4/\gamma^2)(w(\mathcal{C}) + \sqrt{\log n})^2 \log(n/\beta))$ for $\beta > 0$. Then with probability at least $1 - \beta$,*

$$\mathcal{L}(\theta^{\text{priv}}; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) \le \mathcal{L}_{\text{comp}}(\theta^{\text{priv}}; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n); \Phi) + \lambda_\ell \gamma \|\mathcal{C}\|_2.$$

Using Lemma 3.10 and Proposition 3.7 provides the following bounds on excess empirical risk. The parameter γ is chosen to balance various opposing factors. We will assume that $\gamma = o(1)$, which in the following theorem happens when $w(\mathcal{C}) \ll n$ (for achieving (ϵ, δ)-differential privacy) and $w(\mathcal{C}) \ll \sqrt{n}$ (for achieving ϵ-differential privacy). The running time for both these cases is polynomial in n and m.

Theorem 3.11 (Utility of Mechanism ProjERM)**.** *Let Φ be a random matrix as defined in Theorem 3.1.*

1. Then the output θ^{priv} of Mechanism ProjERM *when invoked with Algorithm $\mathcal{A}_{\epsilon,\delta}$ from Theorem 3.6 and $m = \Theta\left(\frac{\psi^2 \epsilon n (w(\mathcal{C}) + \sqrt{\log n})^2 \log(n/\beta)}{w(\mathcal{C})}\right)$ for $\beta > 0$, with probability at least $1 - \beta$ satisfies:*

$$\mathcal{L}(\theta^{\text{priv}}; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) - \mathcal{L}(\hat{\theta}; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) = O\left(\frac{\psi \sqrt{w(\mathcal{C})} \lambda_\ell \|\mathcal{C}\|_2 \log^3 n \log^2(\frac{1}{\delta}) \operatorname{polylog}(\frac{1}{\beta})}{\sqrt{\epsilon n}}\right).$$

2. Then the output θ^{priv} of Mechanism ProjERM *when invoked with Algorithm $\mathcal{A}_{\epsilon}$ from Theorem 3.6 and $m = \Theta\left(\frac{\psi^{4/3} (n\epsilon)^{2/3} (w(\mathcal{C}) + \sqrt{\log n})^2 \log(n/\beta)}{w(\mathcal{C})^{4/3}}\right)$ for $\beta > 0$, with probability at least $1 - \beta$ satisfies:*

$$\mathcal{L}(\theta^{\text{priv}}; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) - \mathcal{L}(\hat{\theta}; (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) = O\left(\frac{\psi^{4/3} w(\mathcal{C})^{2/3} \lambda_\ell \|\mathcal{C}\|_2 \log^2 n \operatorname{polylog}(\frac{1}{\beta})}{(\epsilon n)^{1/3}}\right).$$

4. Algorithm for the Traditional Setting

Mechanism ProjERM can also be directly utilized for solving the traditional private empirical risk minimization problem, where the algorithm gets access to $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$. This approach is illustrated in Mechanism StdERM.

Mechanism StdERM is again (ϵ, δ)- or ϵ-differentially private based on how Mechanism ProjERM is invoked. The excess empirical risk bound of Mechanism StdERM follows from Theorem 3.11.

Corollary 4.1 (of Theorem 3.11)**.** *The output θ^{priv} of Mechanism* StdERM *has the same privacy and excess empirical risk guarantees as that of Mechanism* ProjERM *detailed in Theorem 3.11 with $\psi = O(1)$.*

Under ϵ-differential privacy, these are the first bounds that take the geometry of $\mathcal{C}$ into account to obtain lower excess

Mechanism 2 StdERM

Input: A dataset $D = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ of n datapoints from the domain $\mathcal{Z} = \{(\mathbf{x}, y) : \mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\| \leq 1, y \in \mathbb{R}, |y| \leq 1\}$

Output: θ^{priv} a differentially private estimate of $\hat{\theta} \in \text{argmin}_{\theta \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \ell(\langle \mathbf{x}_i, \theta \rangle; y_i)$

1. Generate $\tilde{\Phi} \in \mathbb{R}^{m \times d} \leftarrow$ matrix whose m rows are i.i.d., mean-zero, isotropic, subgaussian random vectors in $\mathbb{R}^d$ with ψ_2-norm equaling $O(1)$
2. $\Phi \leftarrow \frac{\tilde{\Phi}}{\sqrt{m}}$
3. $\theta^{\text{priv}} \leftarrow$ Output of Mechanism ProjERM invoked on the input $(\Phi \mathbf{x}_1, y_1), \ldots, (\Phi \mathbf{x}_n, y_n)$
4. **Return:** θ^{priv}

risk than the worst case bounds of (Bassily et al., 2014). A simple consequence is improved excess risk bound for releasing the Lasso estimator under ϵ-differential privacy. Given a dataset $D = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ of n datapoints from the domain $\mathcal{Z} = \{(\mathbf{x}, y) : \mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\| \leq 1, y \in \mathbb{R}, |y| \leq 1\}$,[7] the Lasso estimator is defined as $\text{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \mathbf{x}_i, \theta \rangle)^2$ subject to $\theta \in cB_1^d$ (where B_1^d is the unit L_1-ball in $\mathbb{R}^d$, and $c \in \mathbb{R}$ with $c > 0$). Under this setting, Theorem 3.11 (Part 2) implies that there exists an ϵ-differentially private algorithm for releasing the Lasso estimator, that with probability at least $1-\beta$, has excess risk of $O(c^{8/3} \log^{1/3}(d) \log^2(n) \text{polylog}(1/\beta)/(\epsilon n)^{1/3})$, i.e., has only a logarithmic dependence on the dimension d.

We now compare our results with that of (Bassily et al., 2014; Talwar et al., 2015a;b).

Comparison with the Lower Bound on Private ERM Risk from (Bassily et al., 2014). Bassily *et al.* showed that a simple generalized linear problem of the form $\min_{\theta \in S^{d-1}} \frac{1}{n} \sum_{i=1}^{n} \langle \mathbf{x}_i, \theta \rangle$, has an (ϵ, δ)-differentially private excess empirical risk lower bound of $\Omega(\min\{\sqrt{d}/n, 1\})$ (ignoring other parameters). For this generalized linear problem our upper bound of $\approx \sqrt{w(\mathcal{C})/n}$ (ignoring other parameters) from Theorem 3.11 (Part 1) equals $d^{1/4}/\sqrt{n}$ as $w(S^{d-1}) = \Omega(\sqrt{d})$, which is always greater than the lower bound of $\Omega(\min\{\sqrt{d}/n, 1\})$, thereby is consistent with the lower bound. A similar situation also holds for the case of ϵ-differential privacy.

Comparison with the Upper Bounds on Private ERM Risk of (Talwar et al., 2015a;b). Talwar *et al.* presented two (ϵ, δ)-differentially private algorithms, one based on mirror descent (Talwar et al., 2015a) and the other one based on Frank-Wolfe optimization technique. In both these cases, the upper bound on the excess empirical risk depends on the Gaussian width of $\mathcal{C}$. We list the comparison of these results with ours below.

(1) Compressed setting: To our best knowledge, these bounds are the first for private compressed learning, which as discussed in the Introduction (Section 1) is a desired framework for designing efficient algorithms for high-dimensional problems. All previous private ERM algorithms, including that of Talwar *et al.* operate on the original d-dimensional input space.

(2) Traditional setting (under ϵ-differential privacy): The bounds of Talwar *et al.* hold only under the weaker notion of (ϵ, δ)-differential privacy (Thakurta, 2016), whereas in this paper we establish bounds under both ϵ- and (ϵ, δ)-differential privacy.

(3) Traditional setting (under (ϵ, δ)-differential privacy): Firstly note that the bound of Talwar *et al.* holds for all Lipschitz convex functions not just generalized linear functions (the focus of this paper). Now translating the results of Talwar *et al.* to our setting, their mirror descent based algorithm has an expected excess empirical risk bound of $O(\lambda_\ell w(\mathcal{C}) \sqrt{\max_{\theta \in \mathcal{C}} \Psi(\theta) \log(n/\delta)/\epsilon n})$, where $\Psi : \mathcal{C} \rightarrow \mathbb{R}$ needs to be picked to be 1-strongly convex with respect to $\|\cdot\|_{\mathcal{C}}$ norm. Talwar *et al.* also provide few example instantiations of this algorithm by carefully picking Ψ based on $\mathcal{C}$. The Frank-Wolfe based algorithm of Talwar *et al.* has an expected excess empirical risk bound of $O(\Gamma_\ell^{1/3} (\lambda_\ell w(\mathcal{C}))^{2/3} \log^2(n/\delta)/(\epsilon n)^{2/3})$, where Γ_ℓ is the curvature constant of ℓ over the domain $\mathcal{C}$. In general, because of the dependence on slightly different parameters, it is hard to precisely compare the various excess risk bounds. Therefore, as an example, consider the problem of linear regression: $\min_{\theta \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \mathbf{x}_i, \theta \rangle)^2$. In this case, the expected excess risk bound (using the Frank-Wolfe approach) from Talwar *et al.* is $\tilde{O}(\|\mathcal{C}\|_2^{4/3} w(\mathcal{C})^{2/3}/(\epsilon n)^{2/3})$, whereas the bound from Theorem 3.11 (Part 1) is $\tilde{O}(\|\mathcal{C}\|_2^2 \sqrt{w(\mathcal{C})}/\sqrt{\epsilon n})$ (for simplification: setting $\delta \approx 1/\text{poly}(n)$). For bounded $\|\mathcal{C}\|_2, \epsilon$, and in the regime of $w(\mathcal{C}) = o(n)$, when both our result and that of Talwar *et al.* are better than the trivial private risk bound of $O(1)$, the bound of Talwar *et al.* is better than our bound by a factor of $\approx n^{1/6}/w(\mathcal{C})^{1/6}$. So even under (ϵ, δ)-differential privacy, our bounds are close to that achieved by Talwar *et al.*. This along with previous points underscore the importance of our results.

[7] The (ϵ, δ)-differentially private excess risk bounds of (Talwar et al., 2015a;b) for the Lasso estimator hold under a weaker L_∞-normalization of $\mathbf{x}_i$'s.

Acknowledgments

The first author would like to thank his son, Tarun, who thoughtfully prolonged his stay *in utero* until after his father timely submitted this manuscript.

References

Arriaga, Rosa I and Vempala, Santosh. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63(2):161–182, 2006.

Bassily, Raef, Smith, Adam, and Thakurta, Abhradeep. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds. In *FOCS*. IEEE, 2014.

Blocki, Jeremiah, Blum, Avrim, Datta, Anupam, and Sheffet, Or. The johnson-lindenstrauss transform itself preserves differential privacy. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pp. 410–419. IEEE, 2012.

Bourgain, Jean, Sjoerd, Dirksen, and Nelson, Jelani. Toward a unified theory of sparse dimensionality reduction in euclidean space. In *Proceedings of the 47th ACM Symposium on Theory of Computing*. Association for Computing Machinery, 2015.

Boyd, Stephen and Vandenberghe, Lieven. *Convex optimization*. Cambridge university press, 2004.

Calderbank, Robert, Jafarpour, Sina, and Schapire, Robert. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. *preprint*, 2009.

Chaudhuri, Kamalika and Monteleoni, Claire. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems*, pp. 289–296, 2009.

Chaudhuri, Kamalika, Monteleoni, Claire, and Sarwate, Anand D. Differentially private empirical risk minimization. *The Journal of Machine Learning Research*, 12: 1069–1109, 2011.

Davenport, Mark A, Wakin, Michael B, and Baraniuk, Richard G. Detection and estimation with compressive measurements. *Dept. of ECE, Rice University, Tech. Rep*, 2006.

Dirksen, Sjoerd. Dimensionality reduction with subgaussian matrices: a unified theory. *arXiv preprint arXiv:1402.3973*, 2014.

Duchi, John C, Jordan, Michael, Wainwright, Martin J, et al. Local privacy and statistical minimax rates. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pp. 429–438. IEEE, 2013.

Duncan, George T, Pearson, Robert W, et al. Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, 6(3):219–232, 1991.

Dwork, Cynthia and Roth, Aaron. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2013.

Dwork, Cynthia, Kenthapadi, Krishnaram, McSherry, Frank, Mironov, Ilya, and Naor, Moni. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, LNCS, pp. 486–503. Springer, 2006a.

Dwork, Cynthia, McSherry, Frank, Nissim, Kobbi, and Smith, Adam. Calibrating noise to sensitivity in private data analysis. In *TCC*, volume 3876 of *LNCS*, pp. 265–284. Springer, 2006b.

Eldar, Yonina C and Kutyniok, Gitta. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

Fard, Mahdi Milani, Grinberg, Yuri, Pineau, Joelle, and Precup, Doina. Compressed least-squares regression on sparse spaces. In *AAAI*, 2012.

Gordon, Yehoram. *On Milman's inequality and random subspaces which escape through a mesh in* $\mathbb{R}^n$. Springer, 1988.

Jain, Prateek and Thakurta, Abhradeep. Differentially private learning with kernels. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 118–126, 2013.

Jain, Prateek and Thakurta, Abhradeep Guha. (near) dimension independent risk bounds for differentially private learning. In *Proceedings of The 31st International Conference on Machine Learning*, pp. 476–484, 2014.

Jain, Prateek, Kothari, Pravesh, and Thakurta, Abhradeep. Differentially private online learning. In *COLT 2012*, pp. 24.1–24.34, 2012.

Kabán, Ata. New bounds on compressive linear least squares regression. In *The 17-th International Conference on Artificial Intelligence and Statistics (AISTATS 2014)*, volume 33, pp. 448–456, 2014.

Kasiviswanathan, Shiva, Nissim, Kobbi, and Jin, Hongxia. Private incremental regression, 2016.

Kenthapadi, Krishnaram, Korolova, Aleksandra, Mironov, Ilya, and Mishra, Nina. Privacy via the johnson-lindenstrauss transform. *Journal of Privacy and Confidentiality*, 5(1):39–71, 2013.

Kifer, Daniel, Smith, Adam, and Thakurta, Abhradeep. Private convex empirical risk minimization and high-dimensional regression. *Journal of Machine Learning Research*, 1:41, 2012.

Ledoux, Michel and Talagrand, Michel. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 2013.

Mahoney, Michael W. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

Maillard, Odalric and Munos, Rémi. Compressed least-squares regression. In *Advances in Neural Information Processing Systems*, pp. 1213–1221, 2009.

McSherry, Frank and Talwar, Kunal. Mechanism design via differential privacy. In *FOCS*, pp. 94–103. IEEE, 2007.

Mendelson, Shahar. Geometric parameters in learning theory. In *Geometric aspects of functional analysis*, pp. 193–235. Springer, 2004.

Mendelson, Shahar, Pajor, Alain, and Tomczak-Jaegermann, Nicole. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4):1248–1282, 2007.

Mishra, Nikita and Thakurta, Abhradeep. (nearly) optimal differentially private stochastic multi-arm bandits. In *UAI*, pp. 592–601, 2015.

Rubinstein, Benjamin IP, Bartlett, Peter L, Huang, Ling, and Taft, Nina. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *arXiv preprint arXiv:0911.5708*, 2009.

Sarwate, Anand D and Chaudhuri, Kamalika. Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *Signal Processing Magazine, IEEE*, 30(5):86–94, 2013.

Shalev-Shwartz, Shai, Shamir, Ohad, Srebro, Nathan, and Sridharan, Karthik. Stochastic convex optimization. In *COLT*, 2009.

Sheffet, Or. Private approximations of the 2nd-moment matrix using existing techniques in linear regression. *arXiv preprint arXiv:1507.00056*, 2015.

Smith, Adam and Thakurta, Abhradeep Guha. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, pp. 819–850, 2013.

Talwar, Kunal, Thakurta, Abhradeep, and Zhang, Li. Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry. *arXiv preprint arXiv:1411.5417*, 2015a.

Talwar, Kunal, Thakurta, Abhradeep, and Zhang, Li. Nearly optimal private lasso. In *Advances in Neural Information Processing Systems*, pp. 3007–3015, 2015b.

Thakurta, Abhradeep. Personal communication, 2016.

Thakurta, Abhradeep Guha and Smith, Adam. (nearly) optimal algorithms for private online learning in full-information and bandit settings. In *Advances in Neural Information Processing Systems*, pp. 2733–2741, 2013.

Ullman, Jonathan. Private multiplicative weights beyond linear queries. In *Proceedings of the 34th ACM Symposium on Principles of Database Systems*, pp. 303–312. ACM, 2015.

Vempala, Santosh S. *The random projection method*, volume 65. American Mathematical Soc., 2005.

Vershynin, Roman. Estimation in high dimensions: a geometric perspective. *arXiv preprint arXiv:1405.5103*, 2014.

Wang, Yining, Wang, Yu-Xiang, and Singh, Aarti. A deterministic analysis of noisy sparse subspace clustering for dimensionality-reduced data. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 1422–1431, 2015.

Wright, John, Yang, Allen Y, Ganesh, Arvind, Sastry, Shankar S, and Ma, Yi. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.

Zhou, Shuheng, Lafferty, John, and Wasserman, Larry. Compressed and privacy-sensitive sparse regression. *Information Theory, IEEE Transactions on*, 55(2):846–866, 2009a.

Zhou, Shuheng, Ligett, Katrina, and Wasserman, Larry. Differential privacy with compression. In *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, pp. 2718–2722. IEEE, 2009b.