Private Approximations of the 2nd-Moment Matrix Using Existing Techniques in Linear Regression

Or Sheffet
Harvard University
Cambridge, MA
osheffet@seas.harvard.edu

Abstract

We introduce three differentially-private algorithms that approximate the 2nd-moment matrix of the data. These algorithms, which in contrast to existing algorithms output positive-definite matrices, correspond to existing techniques in linear regression literature. Specifically, we discuss the following three techniques. (i) For Ridge Regression, we propose setting the regularization coefficient so that by approximating the solution using Johnson-Lindenstrauss transform we preserve privacy. (ii) We show that adding a small batch of random samples to our data preserves differential privacy. (iii) We show that sampling the 2nd-moment matrix from a Bayesian posterior inverse-Wishart distribution is differentially private provided the prior is set correctly. We also evaluate our techniques experimentally and compare them to the existing "Analyze Gauss" algorithm of Dwork et al [10].

1 Introduction

Differentially private algorithms [8, 7] are data analysis algorithms that give a strong guarantee of privacy, roughly stated as: by adding to or removing from the data a single datapoint we do not significantly change the probability of any outcome of the algorithm. The focus of this paper is on differentially private approximations of the 2nd-moment matrix of the data — given a dataset $D \in \mathbb{R}^{n\times d}$, its *2nd-moment matrix* (also referred to as the *Gram* matrix of data or the *scatter matrix* if the mean of D is $\mathbf{0}$) is the matrix $D^{\mathsf{T}}D$ — and the uses of such approximations in linear regression. Indeed, since the 2nd-moment matrix of the data plays a major role in many data-analysis techniques, we already have differentially private algorithms that approximate the 2nd-moment matrix [10] for the purpose of approximating the PCA, techniques for approximating the rank-k PCA of the data directly [12, 11, 4, 15], or differentially private algorithms for linear regressions [3, 16, 18, 1].

However, existing techniques for differentially private linear regression suffer from the drawback that they approximate a single regression. That is, they assume that each datapoint is composed of a vector of features $\boldsymbol{x}$ and a label y and find the best linear combination of the features that predicts y. Yet, given a dataset D with d attributes we are free to pick any single attribute as a label, and any subset of the remaining attributes as features; and naïvely applying these algorithms to the $\exp(d)$ different linear regression problems simply introduces far too much random noise.[1] In contrast, the differentially private techniques that approximate the 2nd-moment matrix of the data, such as the Analyze Gauss paper of Dwork et al [10], allow us to run as many regressions on the data as we want. Yet, to the best of our knowledge, they have never been analyzed for the purpose of linear regression. Furthermore, the Analyze Gauss algorithm suffers from the drawback that it

[1] Indeed, Ullman [20] have devised a solution to this problem, but this solution works in the more-cumbersome online model and requires exponential running-time for the curator; whereas our techniques follow the more efficient offline approach.

does not necessarily output a positive-definite matrix. This, as discussed in [24] and as we show in our experiments, can be very detrimental — even if we do project the output back onto the set of positive definite matrices. And though the focus of this work is on linear regression, one can postulate additional reasons why releasing a positive definite matrix is of importance, such as using the output as a kernel matrix or doing statistical inference on top of the linear regression.

Our Contribution. In this work, we give three differentially private techniques for approximating the 2nd-moment matrix of the data that output a positive-definite matrix. We analyze their utility, both theoretically and empirically, and more importantly — show how they correspond to *existing techniques in linear regression*. And so we contribute to an increasing line of works [2, 22, 23] that shows that differential privacy may rise from existing techniques, provided parameters are set properly. We also compare our algorithms to the existing Analyze Gauss technique.
(Some notation before we introduce our techniques. We assume the data is a matrix $A \in \mathbb{R}^{n\times d}$ with n sample points in d dimensions. For the ease of exposition, we focus on a single regression problem, given by $A = [X; \boldsymbol{y}]$ — i.e., the label is the d-th column and the features are the remaining $p = d - 1$ columns. We use $\sigma_{\min}(A)$ to denote the least singular value of A.)

1. The Johnson-Lindenstrauss Transform and Ridge Regression. Blocki et al [2] have shown that projecting the data using a Gaussian Johnson-Lindenstrauss transform preserves privacy if $\sigma_{\min}(A)$ is sufficiently large and it has been applied for linear regression [21]. Our first result improves on the analysis of Blocki et al and uses a smaller bound on $\sigma_{\min}(A)$ (shaving off a factor of $\log(r)$ with r denoting the number of rows in the JL transform). This result implies that when $\sigma_{\min}(A)$ is large we can project the data using the JL-transform and output the 2nd-moment matrix of the projected data and preserve privacy. Furthermore, it is also known [17] that the JL-transform gives a good approximation for linear regression problems. However, this is somewhat contradictory to our intuition: for datasets where $\boldsymbol{y}$ is well approximated by a linear combination of X, the least singular value should be small (as A's stretch along the direction $(\boldsymbol{\beta}, -1)^{\mathsf{T}}$ is small). That is why we artificially increase the singular values of A by appending it with a matrix $w \cdot I_{d\times d}$. It turns out that this corresponds to approximating the solution of the *Ridge regression* problem [19, 14], the linear regression problem with l_2-regularization — the problem of finding $\boldsymbol{\beta}^R = \arg\min_{\boldsymbol{\beta}} \sum_i \|y_i - \boldsymbol{\beta}\cdot \boldsymbol{x}_i\|^2 + w^2\|\boldsymbol{\beta}\|^2$. Literature suggests many approaches [13] to determining the penalty coefficient w^2, approaches that are based on the data itself and on minimizing risk. Here we give a fundamentally different approach — set w as to preserve (ϵ, δ)-differential privacy. Details, utility analysis and experiments regarding this approach appear in Section 3.

2. Additive Wishart noise. Whereas the Analyze Gauss algorithm adds Gaussian noise to $A^{\mathsf{T}}A$, here we show that we can sample a positive definite matrix W from a suitably chosen Wishart distribution $\mathcal{W}_d(V, k)$, and output $A^{\mathsf{T}}A + W$. This in turn corresponds to appending A with k i.i.d samples from a multivariate Gaussian $\mathcal{N}(\mathbf{0}_d, V)$. One is able to view this too as an extension of Ridge regression, where instead of appending A with d fixed examples, we append A with $k \approx d + O(1/\epsilon^2)$ random examples.[2] Note, as opposed to Analyze Gauss [10], where the noise has 0-mean, here the expected value of the noise is kV. This yields a useful way of post-processing the output: $A^{\mathsf{T}}A + W - kV$. Details, theorems and experiments with additive Wishart noise appear in Section 4.

3. Sampling from an inverse-Wishart distribution. The Bayesian approach for estimating the 2nd-moment matrix of the data assumes that the n sample points are sampled i.i.d from some $\mathcal{N}(\mathbf{0}_d, V)$ for some unknown V, where we have a prior distribution on V. Each sample point causes us to update our belief on V which results in a posterior distribution on V. Though often one just outputs the MAP of the posterior belief (the mean of the posterior distribution), it is also common to output a sample drawn randomly from the posterior distribution. We show that if one uses the inverse-Wishart distribution as a prior (which is common, as the inverse-Wishart distribution is a conjugate prior), then sampling from the posterior is (ϵ, δ)-differentially private, provided the prior is spread enough. This gives rise to our third approach of approximating $A^{\mathsf{T}}A$ — sampling from a suitable inverse Wishart distribution. We comment that the idea that existing techniques in Bayesian analysis, and specifically sampling from the posterior distribution, are differentially-private on their own was originally introduced in the beautiful and elegant work of Vadhan and Zheng [22]. But whereas their work focuses on estimating the mean of the sample, we focus on estimating the

[2] Though it is also tempting to think of this technique as running Bayesian regression with random prior, this analogy does not fully carry through as we discuss later.

variances/2nd-moment. Details, theorems and experiments on sampling from the inverse-Wishart distribution appear in Section 5.

Finally, in Section 6 we compare our algorithms to the Analyze Gauss algorithm. We show that in the simple case where the data is devised by p independent features concatenated with a single linear combination of the features, the Analyze Gauss algorithm, which introduces the least noise out of all algorithms, is clearly the best algorithm once n is sufficiently large. However, when the data contains multiple such regressions and therefore has small singular values, the situation is far from being clear cut, and indeed, unless n is extremely large, our algorithms achieve smaller errors than the Analyze Gauss baseline. We comment that our experiments should be viewed solely as a proof-of-concept. They are only preliminary, and much more experimentation is needed to fully evaluate the benefits of the various algorithms.

Our proof technique. Before continuing to preliminaries and the formal details of our algorithms, we give an overview of the proof technique. (The proofs themselves are deferred to the supplementary material, Section **??**.) To prove that each algorithm preserves (ϵ,δ)-differential privacy we state and prove 3 corresponding theorems, whose proofs follow the same high-level approach. As mentioned above, we improve on a theorem of Blocki et al [2], who were the first to show that the JL-transform is differentially private. Blocki et al observed that by projecting the data using a $(r\times n)$-matrix of i.i.d normal Gaussians, we effectively repeat the same one-dimensional projection r independent times. So they proved that each one-dimensional projection is (ϵ,δ)-differentially private, and to show the entire projection preserves privacy they used the off-the-shelf composition of Dwork et al [9], getting a bound that depends on $O(\sqrt{r}\log(r))$. In order to derive a bound depending only on $O(\sqrt{r})$, we do not use the composition theorem of [9] but rather study the specific r-fold composition of the projection. As a result, we cannot follow the approach of Blocki et al.

To show that a one-dimensional projection is (ϵ,δ)-differentially private, Blocki et al compared the PDFs of two multivariate Gaussians. The PDF of a multivariate Gaussian is given by the multiplication of two terms: the first depends on the determinant of the variance, and the second depends on some exponent (see exact definition in Section 2). Blocki et al compared the ratio of each of the terms and showed that w.h.p each term's ratio is bounded by $e^{\epsilon/2}$. Unfortunately, following the same approach of Blocki et al yields a bound of $e^{r\epsilon/2}$ for each of the terms and an overall bound that depends on $O(r)$. Instead, we observe that the contributions of the determinant term and the exponent term to the ratio of the PDFs are of opposite signs. So we use the Matrix Determinant Lemma and the Sherman-Morrison Lemma (see full version) to combine both terms into a single exponent term, and bound its size using the Johnson-Lindenstrauss transform (or rather, tight bounds on the χ^2-distribution). The main lemma we use in our analysis, not only gives tight bounds for the Gaussian JL-transform (mimicking the approach of Dasgupta and Gupta [5]), but also gives a result that might be of independent interest. The standard JL-lemma shows that for a $(r\times d)$-matrix R of i.i.d normal Gaussians and any fixed vector $\boldsymbol{v}$ it holds w.h.p that $\boldsymbol{v}^\mathsf{T}\boldsymbol{v}\in(1\pm\eta)\boldsymbol{v}^\mathsf{T}(\frac{1}{r}R^\mathsf{T}R)\boldsymbol{v}$ provided $r=O(\eta^{-2})$. In Lemma **??** we also show that for any fixed $\boldsymbol{v}$ we have w.h.p. that $\boldsymbol{v}^\mathsf{T}\boldsymbol{v}\in(1\pm\eta)\boldsymbol{v}^\mathsf{T}(\frac{1}{r-d}R^\mathsf{T}R)^{-1}\boldsymbol{v}$ provided $r=d+O(\eta^{-2})$. [3]

2 Preliminaries and Notation

Notation. Throughout this paper, we use $lower$-case letters to denote scalars; ***bold*** characters to denote vectors; and UPPER-case letters to denote matrices. The l-dimensional identity matrix is denoted $I_{l\times l}$. For two matrices M,N with the same number of rows we use $[M;N]$ to denote the concatenation of M and N. For a given matrix, $\sigma_{\max}(M)$ and $\sigma_{\min}(M)$ denote its largest and smallest singular values resp.

The Gaussian Distribution and Related Distributions. We denote by $Lap(\sigma)$ the Laplace distribution whose mean is 0 and variance is $2\sigma^2$. A univariate Gaussian $\mathcal{N}\left(\mu,\sigma^2\right)$ denotes the Gaussian distribution whose mean is μ and variance σ^2. Standard concentration bounds on Gaussians give that $\mathbf{Pr}[x>\mu+\sigma\sqrt{\ln(1/\nu)}]<\nu$. A multivariate Gaussian $\mathcal{N}\left(\boldsymbol{\mu},\Sigma\right)$ for some positive semi-definite

[3] To the best of our knowledge, for a general JLT, this is known to hold only when $r=O(d\cdot\eta^{-2})$ and the transform preserves the lengths of all vectors in the $\mathbb{R}^d$ space, see [17] Corollary 11.

Σ denotes the multivariate Gaussian distribution where the mean of the j-th coordinate is the μ_j and the co-variance between coordinates j and k is $\Sigma_{j,k}$. We repeatedly use the rules regarding linear operations on Gaussians. That is, for any scalar c, it holds that $c\mathcal{N}\left(\mu, \sigma^2\right) = \mathcal{N}\left(c \cdot \mu, c^2\sigma^2\right)$. For any matrix C it holds that $C \cdot \mathcal{N}\left(\boldsymbol{\mu}, \Sigma\right) = \mathcal{N}\left(C\boldsymbol{\mu}, C\Sigma C^{\mathsf{T}}\right)$.

The χ^2_k-distribution, where k is referred to as the degrees of freedom of the distribution, is the distribution over the l_2-norm of the sum of k independent normal Gaussians. That is, given $X_1, \ldots, X_k \sim \mathcal{N}(0,1)$ it holds that $\zeta \overset{\text{def}}{=} (X_1, X_2, \ldots, X_k) \sim \mathcal{N}(\mathbf{0}_k, I_{k\times k})$, and $\|\zeta\|^2 \sim \chi^2_k$. The Wishart-distribution $\mathcal{W}_d(V, m)$ is the multivariate extension of the χ^2-distribution. It describes the scatter matrix of a sample of m i.i.d samples from a multivariate Gaussian $\mathcal{N}(\mathbf{0}_d, V)$ and so the support of the distribution is on positive definite matrices. The inverse-Wishart distribution $\mathcal{W}_d^{-1}(V, m)$ describes the distribution over positive definite matrices whose inverse is sampled from the Wishart distribution using the inverse of V; i.e. $X \sim W_d^{-1}(V, m)$ iff $X^{-1} \sim \mathcal{W}_d(V^{-1}, m)$.

Differential Privacy. In this work, we deal with input of the form of a $(n \times d)$-matrix with each row bounded by a l_2-norm of B. Converting A into a linear regression problem, we denote A as the concatenation of the $(n \times p)$-matrix X with the vector $\boldsymbol{y} \in \mathbb{R}^n$ ($A = [X; \boldsymbol{y}]$) where $p = d - 1$. This implies we are tying to predict $\boldsymbol{y}$ as a linear combination of the columns of X. Two matrices A and A' are called *neighbors* if they differ on a single row.

Definition 2.1 ([8, 7])**.** *An algorithm* ALG *which maps* $(n \times d)$*-matrices into some range* $\mathcal{R}$ *is* (ϵ, δ)-differential privacy *if for all pairs of neighboring inputs* A *and* A' *and all subsets* $\mathcal{S} \subset \mathcal{R}$ *it holds that* $\mathbf{Pr}[\mathsf{ALG}(A) \in \mathcal{S}] \leq e^{\epsilon}\mathbf{Pr}[\mathsf{ALG}(A') \in \mathcal{S}] + \delta$*. When* $\delta = 0$ *we say the algorithm is* ϵ*-differentially private.*

It was shown in [8] that for any f where $\|f(A) - f(A')\|_1 \leq \Delta$ then the algorithm that adds Laplace noise $Lap(\frac{\Delta}{\epsilon})$ to $f(A)$ is ϵ-differential privacy. It was shown in [7] that for any f where $\|f(A) - f(A')\|_2 \leq \Delta$ then adding Laplace noise $\mathcal{N}\left(0, \frac{2\Delta^2 \ln(2/\delta)}{\epsilon}\right)$ to $f(A)$ is (ϵ, δ)-differential privacy. This is precisely the algorithm of Dwork et al in their "Analyze Gauss" paper [10]. They observed that in our setting, for the function $f(A) = A^{\mathsf{T}}A$ we have that $\|f(A) - f(A')\|_F^2 = B^4$. And so they add i.i.d Gaussian noise to each coordinate of $A^{\mathsf{T}}A$ (forcing the noise to be symmetric, as $A^{\mathsf{T}}A$ is symmetric). We therefore refer to this benchmark as the Analyze Gauss algorithm. In addition, it is known that the composition of two algorithms, each of which is (ϵ, δ)-differentially private, yields an algorithm which is $(2\epsilon, 2\delta)$-differentially private.

3 Ridge Regression — Set the Regularization Coefficient to Preserve Privacy

The standard problem of linear regression, finding $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|X\boldsymbol{\beta} - \boldsymbol{y}\|^2$, relies on the fact that X is of full-rank. This clearly isn't always the case, and $X^{\mathsf{T}}X$ may be singular or close to singular. To that end, as well as for the purpose of preventing over-fitting, regularization is introduced. One way to regularize the linear regression problem is to introduce a l_2-penalty term: finding $\boldsymbol{\beta}^R = \arg\min_{\boldsymbol{\beta}} \|X\boldsymbol{\beta} - \boldsymbol{y}\|^2 + w^2\|\boldsymbol{\beta}\|^2$. This is known as the *Ridge regression* problem, introduce by [19, 14] in the 60s and 70s. Ridge regression has a closed form solution: $\boldsymbol{\beta}^R = (X^{\mathsf{T}}X + w^2 I_{p\times p})X^{\mathsf{T}}y$. The problem of setting w has been well-studied [13] where existing techniques are data-driven, often proposing to set w as to minimize the empirical risk of $\boldsymbol{\beta}^R$. Here, we propose a fundamentally different approach to the problem of setting w: set it so that we can satisfy (ϵ, δ)-differential privacy (via the Johnson-Lindenstrauss transform).

Observe, the Ridge regression problem can be written as: minimize$\|X\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \|wI_{p\times p}\boldsymbol{\beta} - \mathbf{0}_p\|^2$. So, denote X' are the $((n+p) \times p)$-matrix which we get by concatenating X and $wI_{p\times p}$, and denote $\boldsymbol{y}'$ as the concatenation of $\boldsymbol{y}$ with p zeros. Then $\beta^R = \arg\min \|X'\boldsymbol{\beta} - \boldsymbol{y}'\|^2$. Since $p = d - 1$ and we denote $A = [X; \boldsymbol{y}]$, we can in fact set A' as the concatenation of A with the d-dimensional matrix $wI_{d\times d}$, and we have that $f(\boldsymbol{\beta}) \overset{\text{def}}{=} \left\| A' \begin{pmatrix} \boldsymbol{\beta} \\ -1 \end{pmatrix} \right\|^2 = \|X'\boldsymbol{\beta} - \boldsymbol{y}'\|^2 + w^2$. Hence $\boldsymbol{\beta}^R = \arg\min f(\boldsymbol{\beta})$. Hence, an approximation of $A'^{\mathsf{T}}A'$ yields an approximation of the Ridge regression problem. One way to approximate $A'^{\mathsf{T}}A'$ is via the Johnson-Lindenstrauss transform, which is known to be differentially private if all the singular values of the given input are sufficiently large [2].

And that is precisely why we use A' — all the singular values of $A'^{\mathsf{T}} A'$ are greater by w^2 than the singular values of$A^{\mathsf{T}} A$, and in particular are always $\geq w^2$. Therefore, applying the JLT to A' gives an approximation of $A'^{\mathsf{T}} A'$, and furthermore, due to the work of Sarlos [17] this JLT also approximates the linear regression problem induced by A' (namely, $\arg\min f(\boldsymbol{\beta})$ as defined above). The following theorem improves on the original theorem of Blocki et al [2].

Theorem 3.1. *Fix $\epsilon > 0$ and $\delta \in (0, \frac{1}{e})$. Fix $B > 0$. Fix a positive integer r and let w be such that $w^2 = 4B^2 \left(\sqrt{2r \ln(\frac{4}{\delta})} + \ln(\frac{4}{\delta})\right) / \epsilon$. Let A be a $(n \times d)$-matrix with $d < r$ and where each row of A has bounded l_2-norm of B. Given that $\sigma_{\min}(A) \geq w$, the algorithm that picks a $(r \times n)$-matrix R whose entries are i.i.d samples from a normal distribution $\mathcal{N}(0, 1)$ and publishes $R \cdot A$ is (ϵ, δ)-differentially private.*

This gives rise to our first algorithm. Algorithm 1 gets as input the parameter r — the number of rows in our JLT, and chooses the appropriate regularization coefficient w. Based on Theorem 3.1 and above-mentioned discussion, it is clear that Algorithm 1 is (ϵ, δ)-differentially private. Furthermore, based on the work of Sarlos, we can also argue the following.

```
Input: A matrix $A \in \mathbb{R}^{n \times d}$ and a bound $B > 0$ on the $l_2$-norm of any row in $A$.
    Privacy parameters: $\epsilon, \delta > 0$.
    Parameter $r$ indicating the number of rows in the resulting matrix.
Set $w = \sqrt{4B^2 \left(\sqrt{2r \ln(\frac{4}{\delta})} + \ln(\frac{4}{\delta})\right) / \epsilon}$.
Set $A'$ as the concatenation of $A$ with $wI_{d \times d}$.
Sample a $r \times (n + d)$-matrix $R$ whose entries are i.i.d samples from a normal Gaussian.
return $M = \frac{1}{r}(RA')^{\mathsf{T}}(RA')$ and the approximation $\widetilde{\boldsymbol{\beta}}^R = \arg\min_{\beta_d = -1} \boldsymbol{\beta}^{\mathsf{T}} M \boldsymbol{\beta}$.
```

Algorithm 1: Approximating Ridge Regression while Preserving Privacy

Theorem 3.2. *[[17], Theorem 12] Fix any $\eta > 0$ and $\nu \in (0, \frac{1}{2})$. Apply Algorithm 1 with $r = O(d \log(d) \ln(1/\nu)/\eta^2)$. Then, w.p $\geq 1 - \nu$ it holds that $\|\boldsymbol{\beta}^R - \widetilde{\boldsymbol{\beta}}^R\| \leq \frac{\eta}{\sqrt{w^2 + \sigma_{\min}(A^{\mathsf{T}} A)}} f(\boldsymbol{\beta}^R)$.*

Existing results about the expected distance $\mathbf{E}[\|\boldsymbol{\beta}^R - \widehat{\boldsymbol{\beta}}\|^2]$ (see [6]) can be used together with Theorem 3.2 to give a bound on $\|\widetilde{\boldsymbol{\beta}}^R - \widehat{\boldsymbol{\beta}}\|^2$.

In addition to Algorithm 1, we can use part of the privacy budget to look at the least singular-value of $A^{\mathsf{T}} A$. If it happens to be the case that $\sigma_{\min}(A^{\mathsf{T}} A)$ is large, then we can adjust w by decreasing it by the appropriate factor. In fact, one can completely invert the algorithm and, in case $\sigma_{\min}(A^{\mathsf{T}} A)$ is really large, not only set the regularization coefficient to be any arbitrary non-negative number, but also determine r based on Thm 3.1. To measure the effect of regularization and to analyze the utility of Algorithms 1 we compared them empirically. Details appear in the full version of this work.

4 Additive Wishart Noise — Regression with Additional Random Examples

As discussed in the previous section, Ridge regression can be viewed as regression where in addition to the sample points given by $[X; \boldsymbol{y}]$ we see d additional datapoints given by $wI_{d \times d}$. Our second techniques follows this approach, only, instead of introducing these d fixed datapoints, we introduce a few more than d datapoints which are *random* and independent of the data.[4] Formally, we give the details in Algorithm 2 and immediately following — the theorem proving it is (ϵ, δ)-differentially private.

Theorem 4.1. *Fix $\epsilon \in (0, 1)$ and $\delta \in (0, \frac{1}{e})$. Fix $B > 0$. Let A be a $(n \times d)$-matrix where each row of A has bounded l_2-norm of B. Let N be a matrix sampled from the d-dimensional Wishart distribution with k-degrees of freedom using the scale matrix $B^2 \cdot I_{d \times d}$ (i.e., $N \sim \mathcal{W}_d(B^2 \cdot I_{d \times d}, k)$) for $k \geq \lfloor d + \frac{14}{\epsilon^2} \cdot 2\ln(4/\delta) \rfloor$. Then outputting $X = A^{\mathsf{T}} A + N$ is (ϵ, δ)-differentially private.*

[4] Independent of the data, but dependent of the problem parameters. Our noise *does* depend on the l_2-bound B.

```
Input: A matrix $A \in \mathbb{R}^{n\times d}$ and a bound $B > 0$ on the $l_2$-norm of any row in $A$.
    Privacy parameters: $\epsilon, \delta > 0$.
Set $k \leftarrow \lfloor d + \frac{14}{\epsilon^2} \cdot 2\ln(4/\delta) \rfloor$.
Sample $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_k$ i.i.d examples from $\mathcal{N}\left(\mathbf{0}_d, B^2 I_{d\times d}\right)$.
return $M = A^\mathsf{T} A + \sum_{i=1}^k \boldsymbol{v}_i \boldsymbol{v}_i^\mathsf{T}$ and the approximation $\widetilde{\boldsymbol{\beta}} = \arg\min_{\beta_d = -1} \boldsymbol{\beta}^\mathsf{T} M \boldsymbol{\beta}$.
```

Algorithm 2: Additive Wishart Noise Algorithm

Note: Ridge Regression also has a Bayesian interpretation, as introducing a prior on $\boldsymbol{\beta}$ in regression problem. It is therefore tempting to argue that Theorem 4.1 implies that solving the regression problem with a random prior preserves privacy. (I.e., output the MAP of β after setting its prior to a random sample from the Wishart distribution.) However, this analogy isn't fully accurate, since our algorithm also adds random noise to $X^\mathsf{T}\boldsymbol{y}$. Indeed, regardless of what prior we use for $\boldsymbol{\beta}$, if $\boldsymbol{y} = \mathbf{0}_n$ then we always output $\mathbf{0}_p$ as the estimator of $\boldsymbol{\beta}$, so one can differentiate between the case that $\boldsymbol{y} = \mathbf{0}_n$ and $\boldsymbol{y} \neq \mathbf{0}_n$. We leave the (very interesting) question of whether Wishart additive random noise can be interpreted as a Bayesian prior for future work.

As for utility analysis, we have the following theorem.

Theorem 4.2. *Let $W \sim \mathcal{W}_{p+1}(\sigma^2 I, k)$, and denote $N \in \mathbb{R}^{p\times p}$ and $\boldsymbol{n} \in \mathbb{R}^p$ s.t. $W = \begin{pmatrix} N & \boldsymbol{n} \\ \boldsymbol{n}^\mathsf{T} & * \end{pmatrix}$. Let $X \in \mathbb{R}^{n\times p}$ be a matrix s.t. $X^\mathsf{T}X$ is invertible and let $\boldsymbol{y} \in \mathbb{R}^n$, and such that there exists a $C \geq 2$ s.t. $\sigma_{\min}(X^\mathsf{T}X) = C \cdot \sigma^2(\sqrt{k} + \sqrt{p} + \sqrt{2\ln(4/\nu)})^2$. Denote $\widehat{\boldsymbol{\beta}} = (X^\mathsf{T}X)^{-1}(X^\mathsf{T}\boldsymbol{y})$ and $\widetilde{\boldsymbol{\beta}} = (X^\mathsf{T}X + N)^{-1}(X^\mathsf{T}\boldsymbol{y} + \boldsymbol{n})$. Then*

$$\left\| \widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}} \right\| \leq \frac{1}{C-1}\|\widehat{\boldsymbol{\beta}}\| + \frac{\sigma^2(C-2)}{(C-1)\sigma_{\min}(X^\mathsf{T}X)} \cdot \min\left\{ 2\sqrt{2kp\cdot\ln(4p/\nu)}, (\sqrt{k} + \sqrt{p} + \sqrt{2\ln(4/\nu)})^2 \right\}$$

We are also interested in the utility of this approach after we *remove* some of the noise we add in this technique. Note, $\mathbf{E}[N] = kB^2 \cdot I_{d\times d}$, and so it stands to reason that we output $A^\mathsf{T}A + N - kB^2 \cdot I_{d\times d}$. Now, when $\sigma_{\min}(A^\mathsf{T}A)$ is small, we run the risk that some of the eigenvalues of $A^\mathsf{T}A + N$ are smaller then kB^2. In such a case, we can still decrease $A^\mathsf{T}A + N$ by $B^2\left(\sqrt{k} - (\sqrt{d} + \sqrt{2\ln(4/\delta)})\right)^2 \cdot I_{d\times d}$ and w.h.p get a positive definite matrix. This is the algorithm we set to evaluate empirically, details appear in the full version.

5 Sampling from an Inverse-Wishart Distribution (Bayesian Posterior)

In Bayesian statistics, one estimates the 2nd-moment matrix in question by starting with a prior and updating it based on the examples in the data. More specifically, our dataset A contains n datapoints which we assumed to be drawn i.i.d from some $\mathcal{N}(\mathbf{0}_d, V)$. We assume V was sampled from some distribution $\mathcal{D}$ over positive definite matrices, which is the prior for V. We then update our belief over V using the Bayesian formula: $\mathbf{Pr}[V \mid A] = \frac{\mathbf{Pr}[A \mid V] \cdot \mathbf{Pr}_{\mathcal{D}}[V]}{\int_W \mathbf{Pr}[A \mid W] \cdot \mathbf{Pr}_{\mathcal{D}}[W] dW}$. Finally, with the posterior belief we give an estimation of V — either by outputting the posterior distribution itself, or by outputting the most-likely V according to the posterior, or by sampling from this posterior distribution (maybe multiple times). In this work we assume that our estimator of V is given by sampling from the posterior distribution.

One of the most common priors used for positive definite matrices is the inverse-Wishart distribution. This is mainly due to the fact that the inverse-Wishart distribution is conjugate prior.[5] Specifically, if our prior belief is that $V \sim \mathcal{W}_d^{-1}(\Psi, k)$, then after viewing n examples our posterior is $V \sim \mathcal{W}_d^{-1}\left((\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} + \Psi), n + k\right)$. Here we show that sampling such a positive definite matrix V from our posterior inverse-Wishart distribution is (ϵ, δ)-differentially private, provided the prior distribution's scale matrix, Ψ, has a sufficiently large $\sigma_{\min}(\Psi)$. This result is in line with the recent

[5] A family of distributions is called conjugate prior if the prior distribution and the posterior distribution both belong to this family.

beautiful work of Vadhan and Zheng [22], who showed that many Bayesian techniques for estimating the means are differentially private, provided the prior is set correctly. The formal description of our algorithm and its privacy statement are given below.

```
**Input**: A matrix $A \in \mathbb{R}^{n\times d}$ and a bound $B > 0$ on the $l_2$-norm of any row in $A$.
    Privacy parameters: $\epsilon, \delta > 0$.
Set $\psi \leftarrow \frac{2B^2}{\epsilon}\left(2\sqrt{2(n+d)\ln(4/\delta)} + 2\ln(4/\delta)\right)$.
Sample $M \sim \mathcal{W}_d^{-1}((A^{\mathsf{T}}A + \psi \cdot I_{d\times d}), n+d)$.
**return** $M$ *and the approximation* $\widetilde{\boldsymbol{\beta}} = \arg\min_{\beta_d=-1} \boldsymbol{\beta}^{\mathsf{T}} M \boldsymbol{\beta}$.
```

Algorithm 3: Sampling from an Inverse-Wishart Distribution

Theorem 5.1. *Fix $\epsilon > 0$ and $\delta \in (0, \frac{1}{e})$. Fix $B > 0$. Let A be a $(n \times d)$-matrix and fix an integer $\nu \geq d$. Let w be such that $w^2 = 2B^2\left(2\sqrt{2\nu\ln(4/\delta)} + 2\ln(4/\delta)\right)/\epsilon$. Then, given that $\sigma_{\min}(A) \geq w$, the algorithm that samples a matrix from $\mathcal{W}_d^{-1}(A^{\mathsf{T}}A, \nu)$ is (ϵ, δ)-differentially private.*

We comment on the similarities between Theorem 5.1 and Theorem 3.1. Indeed, the Algorithm 1 essentially samples a matrix from $\mathcal{W}(A^{\mathsf{T}}A + w^2I, k)$ for some choice of w and k (and then normalizes the sample by $\frac{1}{k}$); and Algorithm 3 samples a matrix from $\mathcal{W}^{-1}(A^{\mathsf{T}}A + w^2I, k)$ for a very similar choice of w. And so, much like we did in the Johnson-Lindenstrauss case, we can also use part of the privacy budget to estimate $\sigma_{\min}(A^{\mathsf{T}}A)$ and then set the parameter ψ accordingly. Details and empirical evaluation of this algorithm appear in the full version.

6 Comparison to the "Analyze Gauss" Baseline

In this work we discuss multiple ways for outputting a differentially private approximation of $A^{\mathsf{T}}A$. One such way was given by Dwork et al in their "Analyze Gauss" paper [10]. As mentioned already, Dwork et al simply add to $A^{\mathsf{T}}A$ a symmetric matrix N whose entries are sampled i.i.d from a suitable Gaussian. Furthermore, the magnitude of the noise introduced by the Analyze Gauss algorithm is the smallest out of all algorithms. Yet, as we stressed before, the output of Analyze Gauss isn't necessarily a positive definite matrix. In this section we investigate the effect of these fact on the problem of linear regression.

We study Analyze Gauss' utility empirically, in comparison to the other algorithms we introduce in this work. We compare between the following 6 techniques.
1. Analyze Gauss algorithm: output $A^{\mathsf{T}}A + N$ with N a symmetric matrix whose entries are i.i.d samples from a Gaussian. (Black line, squares.)
2. The JL-based algorithm. (Blue line, squares.)
3. The additive Wishart noise algorithm given by Algorithm 2. (Magenta line, squares.)
4. A scaling version of Analyze Gauss: if the output of Analyze Gauss is not positive definite, add $cI_{d\times d}$ to it with $c = \mathbf{E}[\|N\|]$.[6] (Black line, circles.)
5. The Inverse-Wishart sampling algorithm, which, as we commented in the experiments of Section 5, is analogous to the JL-based algorithm and seems to consistently do better.[7](Blue line, circles.)
6. The scaling version of the additive Wishart random noise, as detailed in the experiment of Section 4. I.e., outputting $A^{\mathsf{T}}A + W - k \cdot V$ (if this leaves the output positive definite) or $A^{\mathsf{T}}A + W - (\sqrt{k} - (\sqrt{d} + \sqrt{2\ln(4/\delta)}))^2 \cdot V$ otherwise. (Magenta line, circles.)

The high-level message from the experiments we show here as follows. In the simple case, Analyze Gauss is the best algorithm to use (and we believe this result is of interest by itself); when Analyze

[6] We have experimented extensively with multiple ways to project the output of the Analyze Gauss algorithm onto the manifold of PSD matrices; including zeroing or setting to 1 all negative eigenvalues, or setting different values for c. This other techniques did not seem to do better than technique *4* above. In fact, their utility was just as bad as the standard Analyze Gauss algorithm (with no post-processing), returning estimations of size 12 or 9 when the true $\boldsymbol{\beta}$ has $\|\boldsymbol{\beta}\| \approx 3$.

[7] Both Algorithms, in (2) and (5) were given the same min-degrees-of-freedom parameter: $2d$.

Gauss returns “unreasonable” answers — so do all other algorithms we use (details below). However, there do exist cases where the Analyze Gauss algorithm under performs in comparison to the additive Wishart noise algorithm, the JL-based algorithm or the Inverse-Wishart sampling algorithm. Due to lack of space, we only discuss the latter case.

We argue that it is important to use algorithms that inherently output a positive definite matrix. To that end, we now investigate a more complex case, where the data is close to being singular, such that additive Gaussian noise is likely to introduce much error. Moreover, this setting illustrates the difference between PCA and linear regression. In the simpler case, running regression on a k-PCA of the data should give a good approximation of the true $\boldsymbol{\beta}$; whereas in this case, even without privacy, a k-PCA of the data drastically distorts the $\widehat{\boldsymbol{\beta}}$ estimators.

In our experiment, the data A is composed of $2p$ features: the first $p = 20$ columns are independent of one another (sampled i.i.d from a normal Gaussian); the latter $p = 20$ columns are the result of some linear combination of the first p ones. And so $A = [X; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_p]$ where for every i we have $\boldsymbol{y}_i = X\boldsymbol{\beta}_i + \boldsymbol{e}_i$ where each coordinate of $\boldsymbol{e}_i$ is sampled i.i.d from $\mathcal{N}\left(0, \sigma^2\right)$ for $\sigma = 0.5$ (fixed for all i). In our experiments, we vary n (from 2^{12} to 2^{27} in powers of 2), but fix $\epsilon = 0.1$. What we also vary is the number of $\boldsymbol{y}$-features we use in our regression. We look at the linear regression problem where the label is some $\boldsymbol{y}_{i_0}$, and the features of the problem are the first d columns plus some m additional $\boldsymbol{y}$-columns. (I.e.: $\{\boldsymbol{x}_1, \ldots \boldsymbol{x}_p\} \cup \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$ where the latter are disjoint to $\boldsymbol{y}_{i_0}$.) And so in our setting we vary m. A good approximation of $\boldsymbol{\beta}$ should therefore return some $\widetilde{\boldsymbol{\beta}}$ which is 0 (or roughly 0)on the latter m coordinates. This corresponds to what we believe to be a high-level task a data-analyst might want to perform: finding out which features are relevant and which are irrelevant for regression.

The results in this case (Figure 1) are far less conclusive. When $m = 0$, we are back to the case of a single regression (with no redundant features), and here Analyze Gauss (black, squares) outperforms all other algorithms once n is large enough (in our case, $n \geq 2^{16}$). Yet, it is enough to set $m = 1$ to get very different results. When $m > 0$ it is evident that Analyze Gauss really performs badly — in fact, in most cases its values were far beyond the range of a reasonable approximation for $\boldsymbol{\beta}$ (taking values like 26 and 45 where $\|\boldsymbol{\beta}\| \approx 3.2$). The scaled version of Analyze Gauss (black, circles) does perform significantly better, yet — it is not the best out of all algorithms. In fact, it is consistently worse than the JL-based algorithms (blue, circles and squares) and from the scaled version of the additive Wishart noise (magenta, circles) for $n < 2^{22} = 4,194,304$. As expected, as m increases, all algorithms’ errors become fairly large.

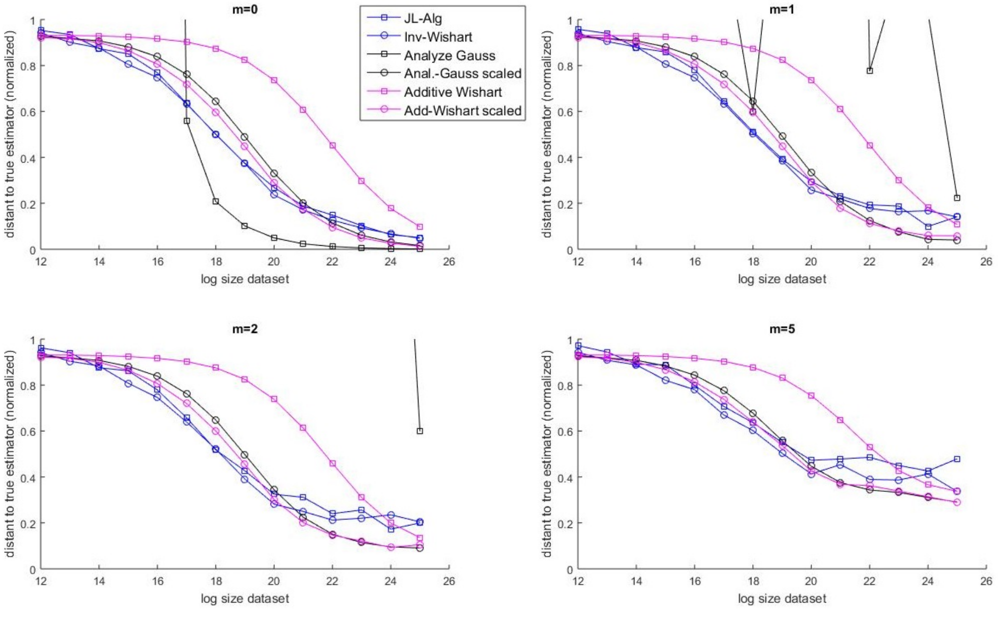

Figure 1: (best seen in color) A comparison of the average l_2-error for 6 estimators. Denoting the true regressor as $\boldsymbol{\beta}$, for each DP-estimator $\widetilde{\boldsymbol{\beta}}$ the y-axis is given by the formula $\frac{\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|}{\|\boldsymbol{\beta}\|}$.

References

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

[2] J. Blocki, A. Blum, A. Datta, and O. Sheffet. The Johnson-Lindenstrauss transform itself preserves differential privacy. In *FOCS*, 2012.

[3] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12, 2011.

[4] Kamalika Chaudhuri, Anand D. Sarwate, and Kaushik Sinha. Near-optimal differentially private principal components. In *NIPS*, 2012.

[5] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Struct. Algorithms*, 22(1), January 2003.

[6] Paramveer S. Dhillon, Dean P. Foster, Sham M. Kakade, and Lyle H. Ungar. A risk comparison of ordinary least squares vs ridge regression. *JMLR*, 14(1), 2013.

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

[8] Cynthia Dwork, Frank Mcsherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

[9] Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *FOCS*, 2010.

[10] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss - optimal bounds for privacy preserving principal component analysis. In *STOC*, 2014.

[11] Moritz Hardt. Robust subspace iteration and privacy-preserving spectral analysis. In *51st Annual Allerton Conference on Communication, Control, and Computing*, 2013.

[12] Moritz Hardt and Aaron Roth. Beating randomized response on incoherent matrices. In *STOC*, 2012.

[13] Trevor J. Hastie, Robert John Tibshirani, and Jerome H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, 2009.

[14] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[15] Michael Kapralov and Kunal Talwar. On differentially private low rank approximation. In *SODA*, 2013.

[16] Daniel Kifer, Adam D. Smith, and Abhradeep Thakurta. Private convex optimization for empirical risk minimization with applications to high-dimensional regression. In *COLT*, 2012.

[17] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *FOCS*, 2006.

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

[19] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 4, 1963.

[20] Jonathan Ullman. Private multiplicative weights beyond linear queries. In *Proceedings of the 34th ACM Symposium on Principles of Database Systems, PODS*, 2015.

[21] Jalaj Upadhyay. Differentially private linear algebra in the streaming model. *CoRR*, abs/1409.5414, 2014.

[22] Salil Vadhan and Joy Zheng. The differential privacy of bayesian inference. Technical report, Faculty of Arts and Sciences, Harvard University, 2015. Available on `http://nrs.harvard.edu/urn-3:HUL.InstRepos:14398533`.

[23] Yu-Xiang Wang, Stephen E. Fienberg, and Alexander J. Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *ICML*, 2015.

[24] Bowei Xi, Murat Kantarcioglu, and Ali Inan. Mixture of gaussian models and bayes error under differential privacy. In *CODASPY*. ACM, 2011.