Proceedings of Machine Learning Research vol 98:1–39, 2019 30th International Conference on Algorithmic Learning Theory

Old Techniques in Differentially Private Linear Regression

Or Sheffet OSHEFFET@UALBERTA.CA
Dept. of Computing Science, University of Alberta, Edmonton, AB Canada [*]

Editors: Aurélien Garivier and Satyen Kale

Abstract

We introduce three novel differentially private algorithms that approximate the 2^{nd}-moment matrix of the data. These algorithms, which in contrast to existing algorithms always output positive-definite matrices, correspond to existing techniques in linear regression literature. Thus these techniques have an immediate interpretation and all results known about these techniques are straight-forwardly applicable to the outputs of these algorithms. More 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 batch of $d + O(\epsilon^{-2})$ random samples to our data preserves differential privacy. (iii) We show that sampling the 2^{nd}-moment matrix from a Bayesian posterior inverse-Wishart distribution is differentially private. We also give utility bounds for our algorithms and compare them with the existing "Analyze Gauss" algorithm of Dwork et al (2014).

Keywords: Differential Privacy; Linear Regression; Second-Moment Matrix; Wishart Distribution

1. Introduction

Differentially private algorithms (Dwork et al., 2006b,a) are data analysis algorithms that give a strong guarantee of privacy, roughly stated as: by altering 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 and the uses of such approximations in linear regression. Recall, given a dataset $A \in \mathbb{R}^{n\times d}$, its 2*nd-moment matrix* is the matrix $A^\mathsf{T} A$ (also referred to as the *Gram* matrix of data or the *scatter matrix* if the mean of A is $\mathbf{0}$). Not surprisingly, since the 2nd-moment matrix of the data plays a major role in many data-analysis techniques, numerous differentially private algorithms that involve an approximation of the 2nd-moment matrix have already been designed. These algorithms include privately approximating the 2^{nd}-moment matrix for approximating the PCA of the data (Dwork et al., 2014), techniques for approximating the rank-k PCA of the data directly (Chaudhuri et al., 2012; Hardt and Roth, 2012, 2013; Hardt, 2013; Kapralov and Talwar, 2013), or differentially private algorithms for linear regressions (Chaudhuri et al., 2011; Kifer et al., 2012; Thakurta and Smith, 2013; Bassily et al., 2014), the latter were also studied empirically (Zhang et al., 2012; Chen et al., 2016).

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

[*] We gratefully acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) for supporting O.S. with grant #2017–06701; O.S. is also an unpaid collaborator on NSF grant #1565387.

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 A with d attributes we are free to pick any single attribute as a label, and any subset of the remaining attributes as features. Therefore, a database with d attributes yields $\exp(d)$ potential linear regression problems; and running these algorithms for each linear regression problem separately simply introduces far too much random noise.[1] We stress that running multiple regressions on the data is a highly prevalent technique in data analysis.[2] In contrast, the differentially private techniques that approximate the 2^{nd}-moment matrix of the data, such as the Analyze Gauss paper of Dwork et al (2014), allow us to run as many regressions on the data as we want. Yet, to the best of our knowledge, the utility of this approach for the purpose of linear regression has never been analyzed. Furthermore, the Analyze Gauss algorithm suffers from the drawback that it does not necessarily output a positive-definite matrix. This, as discussed in Xi et al. (2011) 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.[3] Finally, the notion of adding Gaussian noise to the 2nd-moment matrix is foreign to existing literature on linear regression, and statisticians and data-analysts have no "natural" interpretation to such Gaussian noise.

Our Contribution and Organization. In this work, we introduce three differentially private algorithms that approximate the 2^{nd}-moment matrix, which are all based on the Wishart distribution and thus all three *always* output a positive-definite matrix. All of which are also based on a novel approach to prove differential privacy that utilizes χ^2-distribution concentration bounds (see Proof Technique below).[4] Most importantly, these three algorithms are in direct correspondence with *existing techniques in linear regression* — techniques which are (not surprisingly) based on regularization and have been extensively studied and successfully applied since the 1970s. Thus our work contributes to an increasing line of works (Blocki et al., 2012; Dimitrakakis et al., 2014; Vadhan and Zheng, 2015; Wang et al., 2015; Zhang et al., 2016; Geumlek et al., 2017) that shows how differential privacy can rise from existing techniques, devised far before privacy in data analysis became a concern. (Hopefully this explains the "old techniques" title of this paper — emphasizing the notion of preserving privacy using techniques that predate the invention of differential privacy.) In addition to proving these techniques are differentially private, we also analyze their utility, both theoretically and empirically. Despite the provided utility analysis, we emphasize that our motivating question isn't surpassing the baseline of Analyze Gauss, which is known to match the lower bounds of private PCA (Dwork et al., 2014). Rather, the main focus of this work is to understand the privacy preserving properties of existing algorithms.

(The overview of our techniques requires some notation first. 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

[1] Indeed, Ullman's iterative mechanism (Ullman, 2015) allows us to answer $\exp(d)$ queries, but in the more-cumbersome online model which may require exponential runtime.

[2] Any paper whose main result is "there exists a positive correlation between feature x and label y, even when we control for variables a, b and c" is based on running multiple regressions.

[3] 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.

[4] A recent work (Jiang et al., 2016) also uses the Wishart distribution. Unfortunately, their main claim, stating that additive Wishart noise yields ϵ-differential privacy, has been proven wrong.

features are the first $p \stackrel{\text{def}}{=} d-1$ columns — and denote its regressor as $\boldsymbol{\beta}$. We use $\sigma_{\min}(A)$ to denote the least singular value of A.)

1. The Johnson-Lindenstrauss Transform and Ridge Regression. Blocki et al (2012) have shown that projecting the data using a Gaussian Johnson-Lindenstrauss transform preserves privacy if $\sigma_{\min}(A)$ is sufficiently large. 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 (Sarlós, 2006) 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 (Tikhonov, 1963; Hoerl and Kennard, 1970), 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 (Hastie et al., 2009) to determining the penalty coefficient w^2, approaches that are based on the data itself and on minimizing risk. Here we propose a fundamentally different approach — set w as to preserve (ϵ,δ)-differential privacy. Admittedly, the need for regularization for privacy was presented in prior works (Chaudhuri et al., 2011; Kifer et al., 2012) and subsequent works[5] (Minami et al., 2016; Wang, 2018) yet never from the purposes of approximating the Gram-matrix. Further details 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.[6] Note, as opposed to Analyze Gauss (Dwork et al., 2014), 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 and theorems regarding the additive Wishart noise mechanism, including dealing with the special case of a all-1 column (intercept), appear in Section 4.

3. Sampling from an inverse-Wishart distribution. The Bayesian approach for estimating the 2^{nd}-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. We begin with some prior belief on V, each datapoint causes us to update our belief on V until finally we infer some posterior distribution for V. Though often one just outputs the MAP of the posterior belief (in this case, 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 in practice, as the inverse-Wishart distribution is conjugate prior), then sampling

[5] After posting this work on arXiv.

[6] 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.

from the posterior is (ϵ, δ)-differentially private, provided the prior is sufficiently well-spread. 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 was already introduced in beautiful works such as Dimitrakakis et al. (2014); Vadhan and Zheng (2015); Zhang et al. (2016). But whereas their work focuses on estimating the mean of the sample, we focus on estimating the variance / 2^{nd}-moment. Details and theorems regarding sampling from the inverse-Wishart distribution appear in Section 5.

Section 6 completes the picture by stating the utility guarantees of all 4 algorithms (the above-mentioned three plus the existing "Analyze Gauss" mechanism). We comment that similar results were obtained in a subsequent work (Wang, 2018). These bounds however are somewhat "all over the place" as each depends on slightly different assumptions. This should not come as a surprise — our algorithms are in direct correspondence with existing techniques in linear regression, and the literature on linear regression / 2^{nd}-moment matrix estimation is replete with numerous variations, each working under slightly different premise or setting. Therefore, in addition to providing theoretical bounds, we also compare empirically all 4 algorithms on both synthetic and real data.

Our proof technique. To prove that each algorithm preserves (ϵ, δ)-differential privacy we state and prove three separate theorems, but their proofs all follow a similar high-level approach. This approach is best explained in comparison to the work of Blocki et al (2012), 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-row matrix of Gaussians, we effectively repeat the same one-dimensional projection r independent times. They proved that each row in this projection yields a privacy-loss of at most ϵ, and using the off-the-shelf composition theorem (2010), they got an overall privacy-loss of roughly $O(\epsilon\sqrt{r}\log(r))$. Moreover, the bound of ϵ privacy-loss per row was itself derived from two terms, each of ratio at most $e^{\epsilon/2}$. Blocki et al studied the ratio of the PDFs of two multivariate Gaussians, 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). Through careful analysis, Blocki et al bounded the ratio of each of the terms (w.h.p) by $e^{\epsilon/2}$.

Here, we shave-off a $\log(r)$ factor and derive a bound of $O(\epsilon\sqrt{r})$ by studying the specific r-fold composition of the projection rather than by appealing to existing composition theorems. The ratio of the r-fold composition is still composed of two terms as before (the determinant term and the exponential term). Yet each of the two terms is bounded by roughly $e^{r\epsilon}$ so we cannot mimic the approach of Blocki et al. Instead, we observe that the two terms are of *opposite signs*. So we use the Matrix Determinant Lemma and the Sherman-Morrison Lemma (see Theorem 10) to combine both terms into a single exponent term, and bound its size using tight concentration bounds on the χ^2-distribution; and so we have a privacy loss of $\epsilon(r + O(\sqrt{r})) - \epsilon r = O(\epsilon\sqrt{r})$. The main lemma we use in our analysis is Lemma 8 . This lemma, in addition to giving tight bounds for the Gaussian JL-transform (mimicking the approach of Dasgupta and Gupta (2003)), also gives a result that may be of independent interest. 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 8 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 $r = d + O(\eta^{-2})$. [7] This result is the reason why the number of added samples in the additive Wishart noise case is $k = d + O(\epsilon^{-2})$. We comment that our analysis proves (ϵ, δ)-differential privacy, but in the case of additive Wishart noise the technique does *not* abide the stronger notions of concentrated differential privacy (or zCDP) (Bun and Steinke, 2016; Dwork and Rothblum, 2016).

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 all zero vector is denoted $\mathbf{0}_l$, and the $(l \times m)$-matrix of all zeros is denoted $0_{l\times m}$. The l-dimensional identity matrix is denoted $I_{l\times l}$. (Subscripts are omitted when the dimension is clear.) For two matrices M, N with same number of rows we use $[M; N]$ to denote the concatenation of M and N. For a given matrix, $\|M\|$ denotes the spectral norm $(= \sigma_{\max}(M))$ and $\|M\|_F$ denotes the Frobenious norm $(\sum_{j,k} M_{j,k}^2)^{1/2}$; and use $\sigma_{\max}(M)$ and $\sigma_{\min}(M)$ to denote its largest and smallest singular value 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 $\mathcal{N}(\mu, \sigma^2)$ 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 $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ for some positive semi-definite 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 $\Sigma_{j,k}$. The PDF of such Gaussian is defined only on the subspace $colspan(\Sigma)$, where for every $x \in colspan(\Sigma)$ we have $\mathsf{PDF}(\boldsymbol{x}) = \left((2\pi)^{rank(\Sigma)} \cdot \tilde{\det}(\Sigma)\right)^{-1/2} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\mathsf{T}\Sigma^\dagger(\boldsymbol{x} - \boldsymbol{\mu})\right)$, where $\tilde{\det}(\Sigma)$ is the multiplication of all non-zero singular values of Σ and $\Sigma^\dagger$ denotes the Moore-Penrose Inverse of Σ. We repeatedly use the rules regarding linear operations on Gaussians. That is, for any scalar c, it holds that $c\mathcal{N}(\mu, \sigma^2) = \mathcal{N}(c \cdot \mu, c^2\sigma^2)$, and for any matrix C it holds that $C \cdot \mathcal{N}(\boldsymbol{\mu}, \Sigma) = \mathcal{N}\left(C\boldsymbol{\mu}, C\Sigma C^\mathsf{T}\right)$.

The χ^2_k-distribution, where k denotes the degrees of freedom of the distribution, is the distribution over the norm of the sum of k i.i.d normal Gaussians. That is, given $X_1, \ldots, X_k \sim \mathcal{N}(0, 1)$ it holds that $\|(X_1, X_2, \ldots, X_k)\|^2 \overset{\text{iid}}{\sim} \chi^2_k$. Standard tail bounds on the χ^2-distribution give that for any $\nu \in (0, \frac{1}{2})$ we have $\mathbf{Pr}_{X\sim\chi^2_k}[X \in \left(\sqrt{k} \pm \sqrt{2\ln(\frac{2}{\nu})}\right)^2] \geq 1 - \nu$. (We present them in Section B for completeness.) 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. For $m > d - 1$ we have that $\mathsf{PDF}_{\mathcal{W}_d(V,m)}(X) \propto \det(V)^{-\frac{m}{2}}\det(X)^{\frac{m-d-1}{2}}\exp(-\frac{1}{2}\mathrm{tr}(V^{-1}X))$. 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 $X^{-1} \sim \mathcal{W}_d(V^{-1}, m)$. For $m > d - 1$ it holds that $\mathsf{PDF}_{\mathcal{W}_d^{-1}(V,m)}(X) \propto \det(V)^{\frac{m}{2}}\det(X)^{-\frac{m+d+1}{2}} \cdot \exp(-\frac{1}{2}\mathrm{tr}(VX^{-1}))$.

[7] 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 unit-length vectors in the $\mathbb{R}^d$ space, see (Sarlós, 2006) Corollary 11.

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 1 (Dwork et al., 2006b,a) *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* $\mathbf{Pr}[\mathsf{ALG}(A) \in \mathcal{S}] \leq e^{\epsilon}\mathbf{Pr}\left[\mathsf{ALG}(A') \in \mathcal{S}\right] + \delta$. *When* $\delta = 0$ *we say the algorithm is* ϵ*-differentially private.*

It was shown in Dwork et al. (2006b) that for any f where $\|f(A) - f(A')\|_1 \leq \Delta_1$ then adding iid Laplace noise $Lap(\frac{\Delta_1}{\epsilon})$ to each coordinate of $f(A)$ is ϵ-differentially private. It was shown in Dwork et al. (2006a) that for any f where $\|f(A) - f(A')\|_2 \leq \Delta_2$ then adding iid Gaussian noise $\mathcal{N}\left(0, {}^{2\Delta_2^2 \ln(2/\delta)}/_{\epsilon}\right)$ to each coordinate of $f(A)$ is (ϵ, δ)-differentially private. This is precisely the Analyze Gauss algorithm of Dwork et al (2014) analyzed in Section 6. Dwork et al observed that in our setting we have that $\|A^{\mathsf{T}}A - A'^{\mathsf{T}}A'\|_F = B^2$, 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). Also, composing two (ϵ, δ)-differentially private algorithms yields an algorithm which is $(2\epsilon, 2\delta)$-differentially private.

3. Ridge Regression: Choose Regularization to Preserve Privacy

In standard linear regression, the uniqueness of the regressor $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|X\boldsymbol{\beta} - \boldsymbol{y}\|^2$ 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, introduced by Tikhonov (1963); Hoerl and Kennard (1970) in the 60s and 70s. Ridge regression always has a closed form solution: $\boldsymbol{\beta}^R = (X^{\mathsf{T}}X + w^2 I_{p\times p})^{-1}X^{\mathsf{T}}\boldsymbol{y}$. The problem of setting w has been well-studied (Hastie et al., 2009) where existing techniques are data-driven, often proposing to set w as to empirically minimize the risk of $\boldsymbol{\beta}^R$. Conversely, we propose a fundamentally different approach: set w based on the privacy-loss you wish to incur (after using Gaussian JL projection).

Observe, the Ridge regression problem can be written as: $\min \|X\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \|wI_{p\times p}\boldsymbol{\beta} - \mathbf{0}_p\|^2$. So, denote X' as 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 $\boldsymbol{\beta}^R = \arg\min \|X'\boldsymbol{\beta} - \boldsymbol{y}'\|^2$. 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 $f(\boldsymbol{\beta}) \stackrel{\text{def}}{=} \left\| A' \begin{pmatrix} \boldsymbol{\beta} \\ -1 \end{pmatrix} \right\|^2 = \|X'\boldsymbol{\beta} - \boldsymbol{y}'\|^2 + w^2$. Hence $\boldsymbol{\beta}^R = \arg\min f(\boldsymbol{\beta})$. Hence, a differentially private approximation of $A'^{\mathsf{T}}A'$ results in the ability to approximate any of the (exponentially many) Ridge regression problems based on the data. Here we propose approximating $A'^{\mathsf{T}}A'$ via the Johnson-Lindenstrauss transform, which is known to be differentially private if all the singular values of the given input are sufficiently large (Blocki et al., 2012). And that is precisely why we appended A with wI to create A' — as it must be the case that *all* singular values of $A'^{\mathsf{T}}A'$ are greater then w^2. Therefore, applying the JLT to A' gives a differentially private approximation of $A'^{\mathsf{T}}A'$ (see Theorem 2 below). As discussed in the Introduction, the following theorem also improves on the original bounds of Blocki et al. (2012).

Theorem 2 *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.*

```
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 $\mathcal{N}(0,1)$.    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

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 2 and above-mentioned discussion, it is clear that Algorithm 1 is (ϵ,δ)-differentially private. In Section A (deferred for brevity), we also present a variation on Algorithm 1, where we first use some of the privacy budget to estimate $\sigma_{\min}$ of the data, and adjusts w accordingly.

4. Additive Wishart Noise: Regression with Random Regularization

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 technique follows this approach — we introduce abount $d + O(\epsilon^{-2})$ datapoints that are *random* and independent of the data.[8] Formally, we give the details in Algorithm 2 and immediately following it we present the theorem proving it is (ϵ,δ)-differentially private.

```
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 $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$ i.i.d examples from $\mathcal{N}(\boldsymbol{0}_d, B^2 I_{d\times d})$.
 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

Theorem 3 *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 W 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., $W \sim \mathcal{W}_d(B^2 \cdot I_{d\times d}, k)$) for $k \geq \lfloor d + \frac{14}{\epsilon^2} \cdot 2\ln(4/\delta) \rfloor$. Then outputting $X = A^{\mathsf{T}}A + W$ is (ϵ,δ)-differentially private.*

[8] Independent of the data itself, but dependent of its properties. Our noise *does* depend on the l_2-bound B.

The One-Dimensional Version: The full proof of Theorem 3, though technical and hairy, is a multi-dimensional analog of the following unidimensional back-of-the-envelop calculation. Suppose A is just a $\{0,1\}$-vector (a 1-dimensional input), so $A^{\mathsf{T}}A = \#1\text{s in } A$. Fix a neighboring A' so that $A'^{\mathsf{T}}A' = A^{\mathsf{T}}A + 1$. Thus Algorithm 2 outputs a scalar sampled by $A^{\mathsf{T}}A + X$ where $X \sim \chi^2_k$ for $k = O(\ln(\frac{1}{\delta})/\epsilon^2)$. Concentration bounds give that $X \in \left(\sqrt{k} \pm \sqrt{2\ln(2/\delta)}\right)^2$ w.p $\geq 1-\delta$. As the PDF of the χ^2-distribution is $\propto x^{\frac{k-2}{2}} e^{-\frac{x}{2}}$ we have for any a in the appropriate interval

$$\frac{\mathsf{PDF}[A^{\mathsf{T}}A + X = a]}{\mathsf{PDF}[A'^{\mathsf{T}}A' + X = a]} = \frac{\mathsf{PDF}_{\chi^2_k}[a - A^{\mathsf{T}}A]}{\mathsf{PDF}_{\chi^2_k}[a - A^{\mathsf{T}}A - 1]} = \left(1 + \frac{1}{a - A^{\mathsf{T}}A - 1}\right)^{\frac{k-2}{2}} \cdot e^{-\frac{1}{2}} \leq \exp\left(\frac{k-2}{2\cdot(a - A^{\mathsf{T}}A - 1)} - \frac{1}{2}\right)$$

Based on the concentration of the χ^2_k-distribution we have that on the likely values of a

$$\frac{1}{2} \cdot \frac{k-2}{a - A^{\mathsf{T}}A - 1} - \frac{1}{2} \approx \frac{1}{2} \cdot \frac{k-2-\left(\sqrt{k}\pm\sqrt{2\ln(2/\delta)}\right)^2}{\left(\sqrt{k}\pm\sqrt{2\ln(2/\delta)}\right)^2} \approx \frac{\sqrt{2k\ln(2/\delta)}}{k} = \sqrt{\frac{2\ln(2/\delta)}{k}}$$

so by plugging in the value of k, we see that the log of the PDF-ratio is at most ϵ.

This one dimensional version contains the flavor of most of the proof, and hopefully illustrates the discussion at the Introduction ("Our Proof Technique"): the log of the ratio is composed of two terms – both are roughly $\frac{1}{2}$ yet are of opposite sign – and we show that only w.p. $< \delta$ it holds that the magnitude of the difference between them is $> \epsilon$. In the full proof we replace scalars with determinants and matrix product, and so the simple arithmetic manipulations in this one-dimensional example are replaced with the Sherman-Morison Lemma and the matrix inverse lemma. More importantly, the χ^2-concentration bounds are replaced with our lemma regarding the JLT. This 1-dimensional example also illustrates that our technique is (ϵ,δ)-differentially private yet isn't zCDP (Dwork and Rothblum, 2016; Bun and Steinke, 2016) as the Reyni divergence between the two PDFs is unbounded: if A is the all-0 vector then it is possible (though highly unlikely) to have $A^{\mathsf{T}}A + X < 1$, an impossible outcome under the neighboring A' that holds a single 1 entry where $A'^{\mathsf{T}}A' = 1$.

Remarks. (1) 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 3 implies that solving the regression problem with a random prior preserves privacy. (I.e., output the MAP of $\boldsymbol{\beta}$ after setting its prior to a random sample from the Wishart distribution.) However, notice that our algorithm also adds random noise to $X^{\mathsf{T}}\boldsymbol{y}$. Indeed, just using random prior doesn't guarantee privacy: with a random prior, if $\boldsymbol{y} = \mathbf{0}_n$ then the MAP has to be $\mathbf{0}_p$, thus we can differentiate between the input in which $\boldsymbol{y} = \mathbf{0}_n$ to a neighboring input where $\boldsymbol{y} \neq \mathbf{0}_n$. We leave the (very interesting) question of whether Wishart additive random noise can be interpreted as a Bayesian prior for future work.

(2) Observe that Wishart noise has non-zero mean, but rather $\mathbf{E}[W] = kB^2 \cdot I_{d\times d}$. It thus stands to reason that we post-process the output by zeroing the mean and releasing $A^{\mathsf{T}}A + W - kB^2 \cdot I_{d\times d}$. Note that when $\sigma_{\min}(A^{\mathsf{T}}A)$ is small, it might be the case that some of the eigenvalues of $A^{\mathsf{T}}A + W$ are smaller then kB^2, in which case Lemma 9 assures us that w.h.p we *can* release $A^{\mathsf{T}}A + W - B^2\left(\sqrt{k} - (\sqrt{d} + \sqrt{2\ln(4/\delta)})\right)^2 I$ and the output remains a positive definite matrix. This is the algorithm we set to evaluate empirically in the experiments detailed in Sections 6 and E.

(3) In Section A we present a variation of Algorithm 2 where all examples, including the added ones, have a constant all-1 column, namely the intercept. The intercept plays a key

role in linear regression, especially in the ability to shift features, and we would like to maintain this feature as 1 on the new examples as well. Our analysis shows this is doable, but the number of added samples increases (by a constant factor) and we must make the data-size n publicly known.

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 where: $\mathbf{Pr}[V \mid A] \propto \mathbf{Pr}[A \mid V] \cdot \mathbf{Pr}_{\mathcal{D}}[V]$. 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 section 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.[9] 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 beautiful work of Vadhan and Zheng (2015), who showed that many Bayesian techniques for estimating the means are differentially private, provided the prior is set correctly.[10] 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 4 *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 $w^2 = 2B^2\left(2\sqrt{2\nu\ln(4/\delta)} + 2\ln(4/\delta)\right)/\epsilon$. If we have that $\sigma_{\min}(A) \geq w$ then sampling a matrix from $\mathcal{W}_d^{-1}(A^{\mathsf{T}}A, \nu)$ is (ϵ,δ)-differentially private.*

We comment on the similarities between Theorem 4 and Theorem 2. Indeed, the Algorithm 1 essentially samples a matrix from $\mathcal{W}(A^{\mathsf{T}}A + w^2 I, k)$ for some choice of w

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

[10] The result however is in sharp contrast to the exponential mechanism for PCA sampling in (Chaudhuri et al., 2012; Kapralov and Talwar, 2013). For example, for input of all zeros, the exponential mechanism samples uniformly among unitary matrices; whereas the inverse-Wishart distribution would be centered on the all-0 matrix. This is why we require the input to be well-spread and can guarantee only (ϵ, δ)-differential privacy rather than pure differential privacy.

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 appear in Section A.

6. Utility Analysis and Comparison to the Analyze Gauss Baseline

Theoretical Guarantees. In this paper we discuss multiple ways for outputting a differentially private approximation of $A^\mathsf{T}A$, in addition to the additive Gaussian noise technique presented in Dwork et al. (2014) ("Analyze Gauss"). In this section we provide theoretical utility guarantees for our algorithms. We begin with the "Analyze Gauss" baseline. (We are the first to analyze the performance of "Analyze Gauss" for the purpose of linear regression.) Not surprisingly, as we repeatedly stated throughout the paper that the output of Analyze Gauss isn't necessarily a PSD matrix, we are able to give reasonable utility bounds only in the case that all singular values of the input are sufficiently large, guaranteeing the output is a PSD (as the eigenvalues of the add noise matrix aren't large enough to flip the sign of any of $A^\mathsf{T}A$'s eigenvalues). Formally, we argue the following.

Theorem 5 *Fix $X \in \mathbb{R}^{n\times p}$ and $\boldsymbol{y} \in \mathbb{R}^n$ s.t. $X^\mathsf{T}X$ is invertible. Fix $\eta \in (0,1)$ and $\nu \in (0, 1/e)$. Denote $\widetilde{X^\mathsf{T}X} = X^\mathsf{T}X + N$ and $\widetilde{X^\mathsf{T}\boldsymbol{y}} = X^\mathsf{T}\boldsymbol{y} + \boldsymbol{n}$ where each entry of N and $\boldsymbol{n}$ is sampled i.i.d from $\mathcal{N}(0, \sigma^2)$. Denote also $\widehat{\boldsymbol{\beta}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{y}$ and $\widetilde{\boldsymbol{\beta}} = (\widetilde{X^\mathsf{T}X})^{-1}\widetilde{X^\mathsf{T}\boldsymbol{y}}$. Then, if there exists some constant $C \geq 1$ s.t. we have that $\sigma_{\min}(X^\mathsf{T}X) \geq \frac{2C}{\eta}\cdot\sigma\sqrt{p}\log(1/\nu)$, then w.p. $\geq 1-\nu$ we have $\left\|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right\| \leq 2\eta\|\widehat{\boldsymbol{\beta}}\| + \frac{\eta}{C}$.*

In the Analyze Gauss algorithm, the variance σ^2 of each entry is $B^4\ln(\frac{1}{\delta})/\epsilon^2$. In this case, concentration bounds from Tao (2012) give that w.p. $\geq 1-\nu$ all eigenvalues of the noise matrix used in the Analyze Gauss algorithm are $\leq B^2\sqrt{p\ln(1/\delta)}\cdot\ln(1/\nu)/\epsilon$. Hence Theorem 5 effectively requires the eigenvalues of $X^\mathsf{T}X$ be $O(\frac{1}{\eta})$ times greater than then eigenvalues of the noise matrix.

We now turn to the utility of the Additive Wishart noise algorithm (Algorithm 2). Known concentration bounds on the eigenvalues of the Wishart noise W (Tao, 2012) (see also Lemma 9) give that $\|W\| \leq B^2(\sqrt{k}+\sqrt{p}+\sqrt{2\ln(4/\nu)})^2$ w.p. $\geq 1-\nu$. Therefore, it is evident that the difference between any eigenvalue of $A^\mathsf{T}A$ and $A^\mathsf{T}A+W$ is $B^2(\sqrt{k}+\sqrt{p}+\sqrt{2\ln(4/\nu)})^2$. Next, we also give a bound on the difference in the linear regression estimators.

Theorem 6 *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 denote $\widehat{\boldsymbol{\beta}} = (X^\mathsf{T}X)^{-1}X\boldsymbol{y}$.*

Denote $\widetilde{X^\mathsf{T}X} = X^\mathsf{T}X \ + \ N$, $\widetilde{X^\mathsf{T}\boldsymbol{y}} = X^\mathsf{T}\boldsymbol{y} + \boldsymbol{n}$ and $\widetilde{\boldsymbol{\beta}} = \widetilde{X^\mathsf{T}X}^{-1}\widetilde{X^\mathsf{T}\boldsymbol{y}}$; and also denote $C \stackrel{\text{def}}{=} \frac{\sigma_{\min}(X^\mathsf{T}X)}{\sigma^2\left(\sqrt{k}+\sqrt{p}+\sqrt{2\ln(4/\nu)}\right)^2}$. Then $\left\|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right\| \leq \frac{1}{C+1}\|\widehat{\boldsymbol{\beta}}\| + \left(1 - \frac{1}{C+1}\right)\cdot\frac{2\sigma^2}{\sigma_{\min}(X^\mathsf{T}X)}\sqrt{2kp\cdot\ln(4p/\nu)}$.

The similarities between Algorithm 1 and 3 have been discussed already. Since both algorithms are JL-based, and thus we can rely on the JL lemma for their utility analysis. Based on the work of Sarlos, we can bound the difference between the non-private Ridge-regressor $\boldsymbol{\beta}^R$ and the private approximation of it $\widetilde{\boldsymbol{\beta}}^R$.

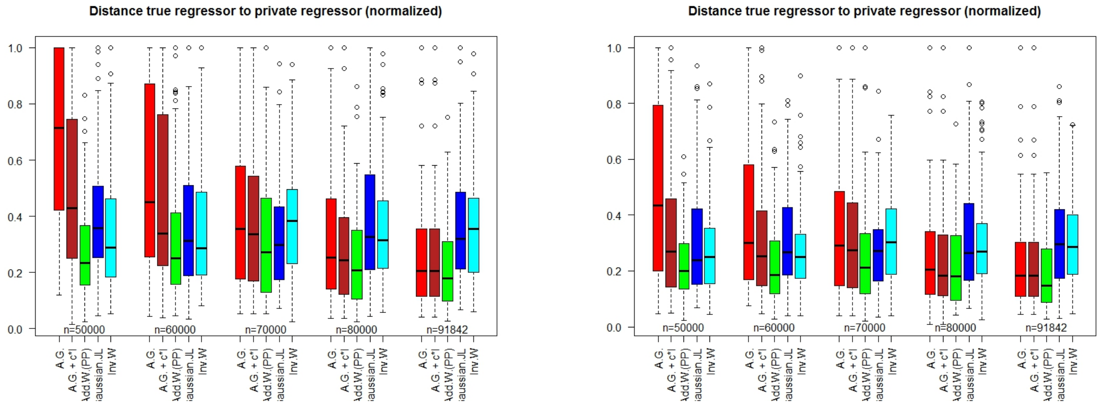
Theorem 7 *[(Sarlós, 2006), Theorems 11& 12] Fix any $\eta, \nu \in (0, \frac{1}{2})$. Apply Algorithm 1 with $r = O(d \log(d) \ln(1/\nu)/\eta^2)$. Then, w.p $\geq 1 - \nu$ it holds that for every i, the i-th singular value of the output M satisfies that $\sigma_i(M) \in (1 \pm \eta)(\sigma_i(A^\mathsf{T} A) + w^2)$, and it also 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 Dhillon et al. (2013)) can be used together with Theorem 7 to give a bound on $\|\widetilde{\boldsymbol{\beta}}^R - \widehat{\boldsymbol{\beta}}\|^2$. These results use the Mahalanobis distance of the input, which we can approximate using the projected output of either algorithm, but due to the hairiness of such results we omit their final form. We also comment that all utility guarantees result in various dependencies on $\sigma_{\min}(X^\mathsf{T} X)$ (or $\sigma_{\min}(A^\mathsf{T} A)$). If we assume that the data is drawn i.i.d from, say, a multivariate Gaussian, such bounds translate to sample complexity bounds, where w.h.p n draws from a multivariate Gaussian of variance Σ have least eigenvalue of $n\sigma_{\min}(\Sigma)$.

Empirical Comparison. The lower bound in Dwork et al. (2014) proves that 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 PSD (and it bears repeating that additive Gaussian noise isn't "interpretable" by classic techniques in regression). Moreover, our utility guarantees depend on a variety of factors, such as the data's least-singular value (or rather the sub-matrix of the data we care about), so comparing all 4 algorithms and determining which one is "best" is a laborious task at best. In fact, this shouldn't come as a surprise considering the vast literature on all the different techniques for regression, varying in their performance based on different assumptions regarding the data and even the analyst's belief (frequentist vs. bayesian). And so, we compare the performance of the 4 algorithms empirically on both synthetic and real data. Due to space constraints, the bulk of the discussion of the experiments is deferred to Section E, so here we provide here just a summary of the results. First, over synthetic data, we show that when all features are uncorrelated indeed the magnitude of the noise is the most influential factor determining the distance between the non-private and private regressors (hence Analyze Gauss has the smallest error); yet, when features are correlated it is far more difficult to discern which algorithm is better.

In addition, we also assess the performance of the four algorithms on real-life data.
The Data: We ran our algorithms over diabetes data collected over ten years (1999-2008) taken from the UCI repository (Strack et al., 2014). We truncated the data to 9 attributes: sex (binary), age (in buckets of 10 years), time in hospital (numeric, in days), number lab procedures (numeric, 0-100), number procedures (numeric, 0-20), number medications (numeric, 0-100), and 3 different diagnoses (numeric, 0-1000), and a 10th column of all-1 (intercept). Omitting any entry with missing or non-numeric values on these nine attributes we were left with $N = 91842$ entries.
The experiments: We shuffled the entries randomly and used different size prefixes of the random dataset. We set $\epsilon = 0.1$ and $\delta = e^{-10}$. We also linearly converted each attribute independently to reside in the range $[-1, 1]$ to set our row-wise bound as $\sqrt{10}$, before running our algorithms (and rescaled each attribute to its original range after the execution of each algorithm). We tried to predict the 3rd diagnosis as a linear function of the other attributes, in three different settings: (i) using all 9 attributes; (ii) omitting the first two diagnoses from the input and using only non-diagnoses attributes (after all, it is reasonable to conjecture one

would want to estimate the value of the diagnosis based on other attributes); (iii) running the algorithm on the entire data, but omitting the first two diagnoses from *the output* so that the regressor must assign zero-value to the two other diagnoses. We believe setting (iii) captures the benefit of outputting the 2^{nd}-moment matrix rather than a private linear-regression algorithm: we get to to choose the features for the problem by ourselves and rather than be constraint to the curator's choice of features. Denoting $\boldsymbol{\beta}$ as the predictor with all 91842 entries and $\widetilde{\boldsymbol{\beta}}$ as the predictor returned by a differentially private algorithm, we measured the performance of the algorithm by $\max\{\frac{\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|}{\|\boldsymbol{\beta}\|}, 1\}$. We ran each algorithm 100 times.



(a) Experiment on Real Data, Setting (i) (3 diagnoses, regression with all attributes)

(b) Experiment on Real Data, Setting (iii) (3 diagnoses, regression doesn't use diagnoses as features)

Figure 1: Experiment on Real Data

Results: The experiment's results in settings (i) and (iii) appear in Figure 1 (full results appear in Section E, Figure 4). In both of these settings one can observe that the Additive Wishart Algorithm outperforms the Analyze Gauss baseline (bright red), *even after* post-processing the output of Analyze Gauss so it is a PSD (dark red). (We comment that we experimented extensively with various post-processing techniques for Analyze Gauss and the one report is the best one of all.) We note that while the improvement isn't necessarily substantial, it is consistent throughout all experiments. We observe in this experiment the same phenomena as in the synthetic data: if the data's feature are not correlated, Analyze Gauss produces the best results; whereas if there are correlations in the data, it under performs in comparison to the Additive Wishart noise algorithm.

More strikingly is the comparison between settings (ii) and (iii) — in both setting we study the exact same regression problem, only in setting (iii) the algorithms also output correlations with two additional unused features. In our full set of results one can see that in setting (ii) Analyze Gauss outperforms Additive Wishart, and (as Figure 1b shows) in setting (iii) it is the other way around. This is because in setting (iii) the least singular value of the input matrix is noticeably smaller than the least singular value of the submatrix considered in setting (ii). Therefore the additive Gaussian noise tends to output a non-PSD matrix even for fairly large values of n. (E.g., even for $n = 80000$ we have that Analyze Gauss has non-negligible probability to output a non-PSD.)

Acknowledgments

The bulk of this work was done when the author was a member of the Harvard's "Privacy Tools for Sharing Research Data" project. The author wholeheartedly thanks Prof. Salil Vadhan for his tremendous help in shaping this paper. The author also thanks numerous other members of this project — Prof. Kobbi Nissim, James Honaker, Vito D'Orazio, Vishesh Karwa and Prof. Gary King, for many helpful discussions and suggestions; and also Prof. Jelani Nelson for a few clarifications regarding certain properties of the JLT. In addition the author thanks the anonymous referees for various helpful suggestions and references.

References

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

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

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *TCC*, pages 635–658, 2016.

K. Chaudhuri, A. Sarwate, and K. Sinha. Near-optimal differentially private principal components. In *NIPS*, 2012.

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

Y. Chen, A. Machanavajjhala, J. P. Reiter, and A. F. Barrientos. Differentially private regression diagnostics. In *ICDM*, pages 81–90, 2016.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms*, 22(1), January 2003.

Kenneth R. Davidson and Stanislaw J. Szarek. Local operator theory, random matrices and banach spaces. In *Handbook of the geometry of Banach spaces*, volume 1. 2001.

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.

Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin I. P. Rubinstein. Robust and private bayesian inference. In *Algorithmic Learning Theory*, pages 291–305, 2014.

C. Dwork, G. Rothblum, and S. Vadhan. Boosting and differential privacy. In *FOCS*, 2010.

Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. *CoRR*, 2016.

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

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

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

Joseph Geumlek, Shuang Song, and Kamalika Chaudhuri. Renyi differential privacy mechanisms for posterior sampling. In *NIPS*, pages 5295–5304, 2017.

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

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

Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *STOC*, 2013.

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.

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

Wuxuan Jiang, Cong Xie, and Zhihua Zhang. Wishart mechanism for differentially private principal components analysis. In *AAAI*, 2016.

M. Kapralov and K. Talwar. On differentially private low rank approximation. In *SODA*, 2013.

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

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate analysis.* Probability and mathematical statistics. Academic Press, 1979.

Kentaro Minami, Hiromi Arai, Issei Sato, and Hiroshi Nakagawa. Differential privacy without sensitivity. In *NIPS*, pages 956–964, 2016.

T. Sarlós. Improved approx. algs for large matrices via random projections. In *FOCS*, 2006.

B. Strack, J. DeShazo, C. Gennings, J. Olmo, S. Ventura, K. Cios, and J. Clore. Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014:11 pages, 2014.

T. Tao. *Topics in Random Matrix Theory.* American Mathematical Soc., 2012.

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

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

Jonathan Ullman. Private multiplicative weights beyond linear queries. In *PODS*, 2015.

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.

Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In *UAI*, pages 93–103, 2018.

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

Oliver Williams and Frank McSherry. Probabilistic inference and differential privacy. In *NIPS*, 2010.

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

Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett. Functional mechanism: Regression analysis under differential privacy. *VLDB*, 5(11):1364–1375, 2012.

Zuhe Zhang, Benjamin I. P. Rubinstein, and Christos Dimitrakakis. On the differential privacy of bayesian inference. In *AAAI*, 2016.

Appendix A. Algorithms deferred from the main text

For completeness, we specify here the formal description of algorithms deferred from the main text, as well as the figures corresponding to our various experiments.

Variation of the Ridge Regression Algorithm. 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 Theorem 2. Details appear in Algorithm 4.

We comment that even though both Algorithms 1 and 4 are written as though they are solving one specific regression (for the ease of exposition). However, their output is clearly a private approximation of $A'^\mathsf{T} A'$ (or $A^\mathsf{T} A$) and regression is merely a post-processing of the output.

```
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 $r_0$ indicating the minimal number of rows in the resulting matrix.
Set $w = \left(\frac{8B^2}{\epsilon}\left(\sqrt{2r_0\ln(\frac{8}{\delta})} + \ln(\frac{8}{\delta})\right)\right)^{1/2}$.
Set $s \leftarrow \max\left\{0, \sigma_{\min}(A^\mathsf{T} A) - \frac{2B^2\ln(2/\delta)}{\epsilon} + Z\right\}$ where $Z \sim Lap(\frac{2B^2}{\epsilon})$.
Adjust $w \leftarrow \sqrt{\max\{0, w^2 - s\}}$.
if $w > 0$ then
    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 $M = \frac{1}{r_0}(RA')^\mathsf{T}(RA')$, $w$ and the approximation $\widetilde{\boldsymbol{\beta}}^R = \arg\min_{\beta_d=-1} \boldsymbol{\beta}^\mathsf{T} M\boldsymbol{\beta}$.
else
    Set $r^*$ as the largest integer $r$ satisfying $\frac{8B^2}{\epsilon}\left(\sqrt{2r\ln(\frac{8}{\delta})} + \ln(\frac{8}{\delta})\right) \le s$
    Sample a $(r^* \times n)$-matrix $R$ whose entries are i.i.d samples from a normal Gaussian.
    return $M = \frac{1}{r^*}(RA)^\mathsf{T}(RA)$, $r^*$ and the approximation $\widetilde{\boldsymbol{\beta}} = \arg\min_{\beta_d=-1} \boldsymbol{\beta}^\mathsf{T} M\boldsymbol{\beta}$.
end
```

Algorithm 4: Approximating Regression (Ridge or standard) while Preserving Privacy.

Variation of the Additive Wishart Noise Algorithm for the Intercept. In linear regression, it is common to have a column of all ones in A (or append such a column to A). Wlog, this all-1 column is the first column of A, and so the first coordinate of $\boldsymbol{\beta}$ is the *intercept* of the regression. Thus the first coordinate in each datapoint is 1 and this fact data-independent. So it stands to reason that the k random points that we add to the data should also have 1 in their first coordinate. Indeed, in Theorem 14 we prove that it is possible to pad the data with k' random examples who also have 1 in the intercept column, provided that n, the number of entries, is known a-priori.[11] We comment that k', the number of examples added with first coordinate set to 1, is indeed greater than our previous parameter k (see definition in Algorithm 2 and Theorem 3) yet k' remains on the order of $d + O(\log(1/\delta)/\epsilon^2)$ (we made no effort to optimize constants). However, since our analysis is based (in part) on the standard additive Gaussian noise mechanism, we

[11] To see why n needs to be public, observe that the padded matrix A' has $n + k'$ examples whose first coordinate is 1, and so the coordinate $(A'^\mathsf{T} A')_{1,1}$ is deterministically set to $n + k'$.

are forced to increase the scale matrix of the examples, from $B^2 \cdot I_{d\times d}$ in Algorithm 2, to $(B^2 \frac{\ln(1/\delta)}{\epsilon^2}) \cdot I_{d\times d}$. We conjecture that a tighter analysis could allow for using the same scale matrix without increasing the number of examples significantly, but we leave it as an open problem. Details appear in Algorithm 5, and its privacy proof appear in Section C.3.

```
Input: A matrix $A \in \mathbb{R}^{n\times p}$ whose first column is all-1 and a bound $B > 0$ on the $l_2$-norm of any row in $A$.
    Privacy parameters: $\epsilon, \delta > 0$.
Set $k' \leftarrow \lfloor p + 1 + \frac{250}{\epsilon^2} \cdot \ln(20/\delta) \rfloor$. Sample $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{k'}$ i.i.d, each coordinate picked i.i.d
  from $\mathcal{N}\left(0, B^2 \ln(1/\delta)/\epsilon^2\right)$. Set the first coordinate of each $\boldsymbol{v}_i$ to be 1. Let $\bar{A}$ denote
  the $((n + k') \times p)$-matrix one gets by concatenating $A$ with the new $k$ samples. return
  $M = \bar{A}^\mathsf{T}\bar{A}$ and the approximation $\widetilde{\boldsymbol{\beta}} = \arg\min_{\beta_d=-1} \boldsymbol{\beta}^\mathsf{T} M \boldsymbol{\beta}$.
```

Algorithm 5: Additive Wishart Noise Algorithm where new examples also have 1 on their intercept column

Variation of the Inverse-Wishart Sampling Algorithm. Similar to the version of the Ridge-Regressions that uses some of the privacy budget to estimate the least singular value of the data and sets w accordingly, we present Algorithm 6 below, which does essentially the same thing for sampling from the inverse-Wishart distribution.

```
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: $\epsilon, \delta > 0$.
    A parameter $k_0$ indicating the minimal degrees of freedom.
Set $\psi \leftarrow \frac{4B^2}{\epsilon}\left(2\sqrt{2k_0 \ln(8/\delta)} + 2\ln(8/\delta)\right)$. Set $s \leftarrow \max\left\{0, \sigma_{\min}(A^\mathsf{T}A) - \frac{2B^2\ln(2/\delta)}{\epsilon} + Z\right\}$
  where $Z \sim Lap(\frac{2B^2}{\epsilon})$. Adjust $\psi \leftarrow \max\{0, \psi - s\}$. if $w > 0$ then
    Sample $M \sim \mathcal{W}_d^{-1}((A^\mathsf{T}A + \psi \cdot I_{d\times d}), k_0)$.
else
    Set $k^*$ as the largest integer $k$ satisfying $\frac{4B^2}{\epsilon}\left(2\sqrt{2k\ln(8/\delta)} + 2\ln(8/\delta)\right) \leq s$
     Sample $M \sim \mathcal{W}_d^{-1}(A^\mathsf{T}A, k^*)$.
end
return $M$ and the approximation $\widetilde{\boldsymbol{\beta}} = \arg\min_{\beta_d=-1} \boldsymbol{\beta}^\mathsf{T} M \boldsymbol{\beta}$.
```

Algorithm 6: Sampling from an Inverse-Wishart Distribution whose degrees of freedom are determined by the input.

Appendix B. Useful Lemmas

In this section we detail the main lemmas that we use in our privacy proofs in the following section. The lemmas and theorems presented here, for the most part, were known prior to our work. We chose to include so that the uninformed reader can have their full proof, but, with the exception of Equation (1), we do not claim any originality to the proofs of the lemmas. The proofs of Lemma 8 and Claim 1 are based in part on the result Dasgupta and Gupta (2003) and in part about results regarding the Wishart distribution given in (Mardia et al., 1979) (Theorem 3.4.7). We encourage the reader who is familiar with lemmas and claims in this section to skip their proofs and turn to Section C where we prove our privacy theorems.

Lemma 8 *Let X be a $(r \times d)$-matrix of i.i.d normal Gaussians (i.e., $x_{i,j} \sim \mathcal{N}(0,1)$). Fix $\beta \in (0, \frac{1}{e})$. Then, for any vector $\boldsymbol{v}$ it holds that*

$$\mathbf{Pr}\left[\left|\boldsymbol{v}^\mathsf{T}(\tfrac{1}{r}X^\mathsf{T}X - I)\boldsymbol{v}\right| \ \le \ \left(\sqrt{\tfrac{8\ln(2/\beta)}{r}} + \tfrac{2\ln(2/\beta)}{r}\right)\|\boldsymbol{v}\|^2\right] \ge 1-\beta$$

Furthermore, if $r \ge d$ then denote $t = \sqrt{\frac{2\ln(2/\beta)}{r-d+1}}$ and assume $t < 1$. Then

$$\mathbf{Pr}\left[\left|\boldsymbol{v}^\mathsf{T}(I - (\tfrac{1}{r-d+1}X^\mathsf{T}X)^{-1})\boldsymbol{v}\right| \ \le \ \frac{2t-t^2}{(1-t)^2}\|\boldsymbol{v}\|^2\right] \ge 1-\beta \tag{1}$$

Proof Fix $\boldsymbol{v}$. Each entry of $X\boldsymbol{v}$ is distributed like $\mathcal{N}(0, \|v\|^2)$ and so $\boldsymbol{v}^\mathsf{T}X^\mathsf{T}X\boldsymbol{v}$ is just the sum of r i.i.d Gaussians with variance $\|v\|^2$. In other words, $\frac{1}{\|\boldsymbol{v}\|^2}\boldsymbol{v}^\mathsf{T}X^\mathsf{T}X\boldsymbol{v} \sim \chi^2_r$. Concentration bounds (see Claim 1) give therefore that w.p. $\ge 1-\beta$ we have

$$(\sqrt{r} - \sqrt{2\ln(2/\beta)})^2 \le \tfrac{1}{\|\boldsymbol{v}\|^2}\boldsymbol{v}^\mathsf{T}X^\mathsf{T}X\boldsymbol{v} \le (\sqrt{r} + \sqrt{2\ln(2/\beta)})^2$$

which implies

$$\begin{aligned}\boldsymbol{v}^\mathsf{T}(\tfrac{1}{r}X^\mathsf{T}X - I)\boldsymbol{v} \ &\ge \ \left(-2\sqrt{\tfrac{2\ln(2/\beta)}{r}} + \tfrac{2\ln(2/\beta)}{r}\right)\|\boldsymbol{v}\|^2\\ \boldsymbol{v}^\mathsf{T}(\tfrac{1}{r}X^\mathsf{T}X - I)\boldsymbol{v} \ &\le \ \left(2\sqrt{\tfrac{2\ln(2/\beta)}{r}} + \tfrac{2\ln(2/\beta)}{r}\right)\|\boldsymbol{v}\|^2\end{aligned}$$

and so we get the bound on $\boldsymbol{v}^\mathsf{T}(\frac{1}{r}X^\mathsf{T}X - I)\boldsymbol{v}$.

We now argue that $\frac{\boldsymbol{v}^\mathsf{T}\boldsymbol{v}}{\boldsymbol{v}(X^\mathsf{T}X)^{-1}\boldsymbol{v}} \sim \chi^2_{r-d+1}$. To see this, we argue that specifically for the vector $\boldsymbol{e}_d$ (the indicator of the d-th coordinate) we have $\frac{1}{\boldsymbol{e}_d(X^\mathsf{T}X)^{-1}\boldsymbol{e}_d} \sim \chi^2_{r-d+1}$, and the results for any $\boldsymbol{v}$ follows from taking any unitary function s.t. $U^\mathsf{T}\boldsymbol{v} = \|\boldsymbol{v}\|\boldsymbol{e}_d$, and the observation that the distributions of X and XU^T are identical.

Now, clearly $\boldsymbol{e}_d(X^\mathsf{T}X)^{-1}\boldsymbol{e}_d = (X^\mathsf{T}X)^{-1}_{d,d}$. Now, if we denote the last column of X as $\boldsymbol{x}_d$ and the first $d-1$ columns of X as X_{-d} then $X^\mathsf{T}X = \left[\begin{array}{c|c} {X_{-d}}^\mathsf{T}X_{-d} & {X_{-d}}^\mathsf{T}\boldsymbol{x}_d \\ \hline {\boldsymbol{x}_d}^\mathsf{T}X_{-d} & \|\boldsymbol{x}_d\|^2 \end{array}\right]$. Thus, the formula for the entries of the inverse give

$$\begin{aligned}\frac{1}{(X^\mathsf{T}X)^{-1}_{d,d}} &= \|\boldsymbol{x}_d\|^2 - {\boldsymbol{x}_d}^\mathsf{T}X_{-d}({X_{-d}}^\mathsf{T}X_{-d})^{-1}{X_{-d}}^\mathsf{T}\boldsymbol{x}_d\\ &= \boldsymbol{x}_d\left(I - X_{-d}({X_{-d}}^\mathsf{T}X_{-d})^{-1}{X_{-d}}^\mathsf{T}\right)\boldsymbol{x}_d \overset{\text{def}}{=} {\boldsymbol{x}_d}^\mathsf{T}P\,\boldsymbol{x}_d\end{aligned}$$

Now, w.p. 1 we have that X_{-d} has full rank $(d-1)$. For any choice of X_{-d} with full rank we get a matrix P which has rank $r-(d-1)$ and its eigenvalues are either 1 or 0. Hence, for any X_{-d} we get $\frac{1}{(X^\mathsf{T}X)^{-1}_{d,d}} \sim \chi^2_{r-d+1}$. Since this distribution is independent of X_{-d} we therefore have that this result holds w.p. 1. I.e.:

$$\begin{aligned}
&\mathsf{PDF}\left(\tfrac{1}{(X^\mathsf{T}X)^{-1}_{d,d}} = z\right)\\
&= \int\limits_P \mathsf{PDF}\left(\tfrac{1}{(X^\mathsf{T}X)^{-1}_{d,d}} = z \,\middle|\, I - X_{-d}({X_{-d}}^\mathsf{T}X_{-d})^{-1}{X_{-d}}^\mathsf{T} = P\right) \mathsf{PDF}\left(I - X_{-d}({X_{-d}}^\mathsf{T}X_{-d})^{-1}{X_{-d}}^\mathsf{T} = P\right) dP\\
&= \int\limits_P \mathsf{PDF}\left(I - X_{-d}({X_{-d}}^\mathsf{T}X_{-d})^{-1}{X_{-d}}^\mathsf{T} = P\right) \cdot \mathsf{PDF}_{\chi^2_{r-d+1}}(z) dP\\
&= 1 \cdot \mathsf{PDF}_{\chi^2_{r-d+1}}(z) = \mathsf{PDF}_{\chi^2_{r-d+1}}(z)
\end{aligned}$$

Therefore, same argument from Claim 1 gives that with probability $\geq 1-\beta$ we have

$$\frac{\boldsymbol{v}^\mathsf{T}\boldsymbol{v}}{\boldsymbol{v}^\mathsf{T}(X^\mathsf{T}X)^{-1}\boldsymbol{v}} \in \left(\sqrt{r-d+1} \pm \sqrt{2\ln(2/\beta)}\right)^2$$

so

$$\begin{aligned}
\boldsymbol{v}^\mathsf{T}\left(\tfrac{1}{r-d+1}X^\mathsf{T}X\right)^{-1}\boldsymbol{v} &\geq \left(\frac{\sqrt{r-d+1}}{\sqrt{r-d+1}+\sqrt{2\ln(2/\beta)}}\right)^2 \|\boldsymbol{v}\|^2\\
\boldsymbol{v}^\mathsf{T}\left(\tfrac{1}{r-d+1}X^\mathsf{T}X\right)^{-1}\boldsymbol{v} &\leq \left(\frac{\sqrt{r-d+1}}{\sqrt{r-d+1}-\sqrt{2\ln(2/\beta)}}\right)^2 \|\boldsymbol{v}\|^2
\end{aligned}$$

which implies

$$\boldsymbol{v}^\mathsf{T}\left(I - (\tfrac{1}{r-d+1}X^\mathsf{T}X)^{-1}\right)\boldsymbol{v} \leq \frac{2\sqrt{\frac{2\ln(2/\beta)}{r-d-1}} + \frac{2\ln(2/\beta)}{r-d-1}}{(1+\sqrt{\frac{2\ln(2/\beta)}{r-d-1}})^2}$$

$$\boldsymbol{v}^\mathsf{T}\left(I - (\tfrac{1}{r-d+1}X^\mathsf{T}X)^{-1}\right)\boldsymbol{v} \geq -\frac{2\sqrt{\frac{2\ln(2/\beta)}{r-d-1}} - \frac{2\ln(2/\beta)}{r-d-1}}{(1-\sqrt{\frac{2\ln(2/\beta)}{r-d-1}})^2}$$

Some arithmetic manipulations show that when $\frac{2\ln(2/\beta)}{r-d-1} < 1$ we have that

$$\left|\boldsymbol{v}^\mathsf{T}\left(I - (\tfrac{1}{r-d+1}X^\mathsf{T}X)^{-1}\right)\boldsymbol{v}\right| \leq \frac{2\sqrt{\frac{2\ln(2/\beta)}{r-d-1}} - \frac{2\ln(2/\beta)}{r-d-1}}{(1-\sqrt{\frac{2\ln(2/\beta)}{r-d-1}})^2}$$

as this is the larger term of the two. ■

Claim 1 *Fix k and let $X_1, \ldots, X_k$ be iid samples from $\mathcal{N}(0,1)$. Then, for any $0 < \Delta < k$ we have that $\mathbf{Pr}[\sum_i X_i^2 > (\sqrt{k}+\sqrt{\Delta})^2] < e^{-\Delta/2}$ and $\mathbf{Pr}[\sum_i X_i^2 < (\sqrt{k}-\sqrt{\Delta})^2] < e^{-\Delta/2}$.*

Proof We start with the following calculation. For any $X \sim \mathcal{N}(0,1)$ and any $s < 1/2$ it holds that

$$\mathbf{E}[e^{sX^2}] = \int\limits_{-\infty}^{\infty} \tfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} e^{sx^2} dx = \int\limits_{-\infty}^{\infty} \tfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2(1-2s)}{2}} dx$$

$$\overset{\substack{y=x\sqrt{1-2s}\\ \text{so } dy=dx\sqrt{1-2s}}}{=} \int\limits_{-\infty}^{\infty} \tfrac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \tfrac{dy}{\sqrt{1-2s}} = \tfrac{1}{\sqrt{1-2s}}$$

We now use Markov's inequality, to deduce that for any $\lambda \in (0, 1/2)$

$$\begin{aligned}
\mathbf{Pr}[\sum_i X_i^2 > (\sqrt{k}+\sqrt{\Delta})^2] &= \mathbf{Pr}[e^{\lambda \sum_i X_i^2} > e^{\lambda(\sqrt{k}+\sqrt{\Delta})^2}] \leq \frac{\mathbf{E}[e^{\lambda \sum_i X_i^2}]}{e^{\lambda(\sqrt{k}+\sqrt{\Delta})^2}} \\
&= \prod_i \mathbf{E}[e^{\lambda X_i^2}] e^{-\lambda(\sqrt{k}+\sqrt{\Delta})^2} = \left(\frac{1}{1-2\lambda}\right)^{\frac{k}{2}} e^{-\lambda(\sqrt{k}+\sqrt{\Delta})^2} \\
&= \left(1 + \frac{2\lambda}{1-2\lambda}\right)^{\frac{k}{2}} e^{-\lambda(\sqrt{k}+\sqrt{\Delta})^2} \leq \exp\left(\frac{\lambda k}{1-2\lambda} - \lambda(\sqrt{k}+\sqrt{\Delta})^2\right)
\end{aligned}$$

Setting $\lambda = \frac{\sqrt{\Delta}}{2(\sqrt{k}+\sqrt{\Delta})}$ so that $1-2\lambda = \frac{\sqrt{k}}{\sqrt{k}+\sqrt{\Delta}}$ we have

$$\mathbf{Pr}[\sum_i X_i^2 > (\sqrt{k}+\sqrt{\Delta})^2] \leq \exp\left(\tfrac{1}{2}\sqrt{k\Delta} - \tfrac{1}{2}\sqrt{\Delta}(\sqrt{k}+\sqrt{\Delta})\right) = \exp(-\tfrac{\Delta}{2})$$

A similar calculation gives the lower bound. ■

Lemma 9 *Fix $\delta \in (0, e^{-1})$. Let X be a matrix sampled from a Wishart distribution $\mathcal{W}_d(V, m)$ where $\sqrt{m} > \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right)$. Then, w.p. $\geq 1-\delta$ we have that for every $j = 1, 2, \ldots, d$ it holds that*

$$\sigma_j(X) \in (\sqrt{m} \pm \left(\sqrt{d} + \sqrt{2\ln(\tfrac{2}{\delta})}\right))^2 \sigma_j(V)$$

Furthermore, we also have that for any $0 < \alpha \leq m$ it holds $\|\alpha V - X\| \leq \|V\| \cdot |\alpha - (\sqrt{m} - \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right))^2|$ and $\|(\alpha V)^{-1} - X^{-1}\| \leq$

$$\sigma_{\min}^{-1}(V) \cdot |\alpha^{-1} - (\sqrt{m} + \left(\sqrt{d} + \sqrt{2\ln(\tfrac{2}{\delta})}\right))^{-2}|.$$

Proof In order to sample $X \sim \mathcal{W}_d(V, m)$ we first sample a matrix $Y \in \mathbb{R}^{m \times d}$ in which every entry is i.i.d normal Gaussian. We then multiply Y by $V^{1/2}$, s.t. every row in $YV^{1/2}$ is sampled i.i.d from $\mathcal{N}(\mathbf{0}_d, V)$. We then set $X = V^{1/2} Y^\mathsf{T} Y V^{1/2}$.

Now, we invoke a theorem of Davidson and Szarek (2001) (Theorem II.13) that states that for any $t > 1$ we have $\mathbf{Pr}[\sigma_{\max}(Y) > \sqrt{m} + \sqrt{d} + t] < e^{-t^2/2}$ and $\mathbf{Pr}[\sigma_{\min}(Y) < \sqrt{m} - \sqrt{d} - t] < e^{-t^2/2}$. So we deduce that w.p. $\geq 1-\delta$ it holds that all of the singular values of Y lie on the interval $\left(\sqrt{m} - \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right), \sqrt{m} + \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right)\right)$. Next, we let $\boldsymbol{u}_j$ denote the j-th eigenvector of V, corresponding to the j-th eigenvalue $\sigma_j(V)$. Therefore, for any j we have

$$\begin{aligned}
\boldsymbol{u}_j{}^\mathsf{T} X \boldsymbol{u}_j &= (V^{1/2}\boldsymbol{u}_j)^\mathsf{T} Y^\mathsf{T} Y (V^{1/2}\boldsymbol{u}_j) \leq (\sqrt{m} + \left(\sqrt{d} + \sqrt{2\ln(\tfrac{2}{\delta})}\right))^2 \|V^{1/2}\boldsymbol{u}_j\|^2 \\
&= \sigma_j(V)(\sqrt{m} + \left(\sqrt{d} + \sqrt{2\ln(\tfrac{2}{\delta})}\right))^2 \\
\boldsymbol{u}_j{}^\mathsf{T} X \boldsymbol{u}_j &\geq (\sqrt{m} - \left(\sqrt{d} + \sqrt{2\ln(\tfrac{2}{\delta})}\right))^2 \|V^{1/2}\boldsymbol{u}_j\|^2 = \sigma_j(V)(\sqrt{m} - \left(\sqrt{d} + \sqrt{2\ln(\tfrac{2}{\delta})}\right))^2
\end{aligned}$$

and furthermore, for any subspace S we have that

$$\max_{\boldsymbol{u}\in S\colon \|u\|=1} \boldsymbol{u}^{\mathsf{T}} X \boldsymbol{u} \le (\sqrt{m} + \left(\sqrt{d} + \sqrt{2\ln(\tfrac{2}{\delta})}\right))^2 \cdot \left(\max_{\boldsymbol{u}\in S\colon \|u\|=1} \|V^{1/2}\boldsymbol{u}_j\|^2\right)$$

$$\min_{\boldsymbol{u}\in S\colon \|u\|=1} \boldsymbol{u}^{\mathsf{T}} X \boldsymbol{u} \ge (\sqrt{m} - \left(\sqrt{d} + \sqrt{2\ln(\tfrac{2}{\delta})}\right))^2 \cdot \left(\min_{\boldsymbol{u}\in S\colon \|u\|=1} \|V^{1/2}\boldsymbol{u}_j\|^2\right)$$

Thus, to complete the first part of the proof, we invoke the Courant-Fischer Min-Max Theorem that state that

$$\begin{aligned}\sigma_j(X) &= \max_{\{S\subset\mathbb{R}^d\colon\ \dim(S)=j\}} \min_{\{\boldsymbol{u}\in S\colon\ \|\boldsymbol{u}\|=1\}} \boldsymbol{u}^{\mathsf{T}} X \boldsymbol{u} \\ &= \min_{\{S\subset\mathbb{R}^d\colon\ \dim(S)=d-j+1\}} \max_{\{\boldsymbol{u}\in S\colon\ \|\boldsymbol{u}\|=1\}} \boldsymbol{u}^{\mathsf{T}} X \boldsymbol{u}\end{aligned}$$

Therefore, we can pick $S' = \mathit{span}\{\boldsymbol{u}_1, \dots \boldsymbol{u}_j\}$ and $S'' = \mathit{span}\{\boldsymbol{u}_j, \dots, \boldsymbol{u}_d\}$ to deduce

$$\sigma_j(X) \ge \min_{\boldsymbol{u}\in S'\colon\|\boldsymbol{u}\|=1} \boldsymbol{u}^{\mathsf{T}} X \boldsymbol{u} \ge (\sqrt{m} - \left(\sqrt{d} + \sqrt{2\ln(\tfrac{2}{\delta})}\right))^2 \sigma_j(V)$$

$$\sigma_j(X) \le \max_{\boldsymbol{u}\in S''\colon\|\boldsymbol{u}\|=1} \boldsymbol{u}^{\mathsf{T}} X \boldsymbol{u} \le (\sqrt{m} + \left(\sqrt{d} + \sqrt{2\ln(\tfrac{2}{\delta})}\right))^2 \sigma_j(V)$$

As for the second part of the claim, it follows from the fact that $\alpha V - X = V^{1/2}\left(\alpha I - Y^{\mathsf{T}} Y\right) V^{1/2}$. Now, if we denote $Y = U\Sigma U^{\mathsf{T}}$ as the SVD decomposition of Y, then we have $\alpha I - Y^{\mathsf{T}} Y = U\left(\alpha I - \Sigma\right) U^{\mathsf{T}}$ and all the entries on the diagonal of $(\alpha I - \Sigma)$ lie in the range $|\alpha - (\sqrt{m} \pm \left(\sqrt{d} + \sqrt{2\ln(\tfrac{2}{\delta})}\right))^2|$. As $\alpha \le m$ we have that all eigenvalues are upper bounded by $(m-\alpha) + 2\sqrt{m}\left(\sqrt{d} + \sqrt{2\ln(\tfrac{2}{\delta})}\right)$ and the claim follows. Similarly, for $(\alpha V)^{-1} - X^{-1} = V^{-1/2}\left(\alpha I - Y^{\mathsf{T}} Y\right) V^{-1/2}$ all eigenvalues lie in the range $|\alpha^{-1} - (\sqrt{m} \pm \left(\sqrt{d} + \sqrt{2\ln(\tfrac{2}{\delta})}\right))^{-2}|$, which in this case is upper bounded by $|\alpha^{-1} - (\sqrt{m} + \left(\sqrt{d} + \sqrt{2\ln(\tfrac{2}{\delta})}\right))^{-2}|$. (We comment that the bounds on $\|\alpha V - X\|$ and on $\|(\alpha V)^{-1} - X^{-1}\|$ require we use both the upper- and lower-bounds on the eigenvalues of Y.) ■

The other two useful tools we use are the formula for rank-1 updates of the determinant and the inverse (the Sherman-Morrison lemma).

Theorem 10 *Let A be a $(d \times d)$-invertible matrix and fix any two d-dimensional vectors $\boldsymbol{u}, \boldsymbol{v}$ s.t. $\boldsymbol{v}^{\mathsf{T}} A^{-1} \boldsymbol{u} \neq -1$. Then:*

$$\det(A + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}) = \det(A)(1 + \boldsymbol{v}^{\mathsf{T}} A^{-1} \boldsymbol{u})$$

$$(A + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}})^{-1} = A^{-1} - \frac{A^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}} A^{-1}}{1 + \boldsymbol{v}^{\mathsf{T}} A^{-1}\boldsymbol{u}}$$

Proof *Since we have $A + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}} = A(I + A^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}})$, we analyze the spectrum of the matrix $I + A^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}$. Clearly, for any $\boldsymbol{x} \perp \boldsymbol{v}$ we have $(I + A^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}})\boldsymbol{x} = \boldsymbol{x} + 0 \cdot A^{-1}\boldsymbol{u} = \boldsymbol{x}$, so $d-1$ of the eigenvalues of $I + A^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}$ are exactly 1. As for the last one, take a unit length vector $\boldsymbol{z} = \frac{1}{\|\boldsymbol{v}\|}\boldsymbol{v}$, and we have $\boldsymbol{z}^{\mathsf{T}}(I + A^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}})\boldsymbol{z} = 1 + \|\boldsymbol{v}\| \cdot \boldsymbol{z}^{\mathsf{T}} A^{-1} u = 1 + \boldsymbol{v}^{\mathsf{T}} A^{-1}\boldsymbol{u}$. Therefore, $\det(A + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}) = \det(A)\det(I + A^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}) = \det(A)(1 + \boldsymbol{v}^{\mathsf{T}} A^{-1} u)$.*

As for the Sherman-Morrison formula, we can simply check and see that indeed:

$$(A+\boldsymbol{u}\boldsymbol{v}^\mathsf{T})(A^{-1}-\frac{A^{-1}\boldsymbol{u}\boldsymbol{v}^\mathsf{T}A^{-1}}{1+\boldsymbol{v}^\mathsf{T}A^{-1}\boldsymbol{u}}) = I+\boldsymbol{u}\boldsymbol{v}^\mathsf{T}A^{-1} - \frac{\boldsymbol{u}\boldsymbol{v}^\mathsf{T}A^{-1}}{1+\boldsymbol{v}^\mathsf{T}A^{-1}\boldsymbol{u}} - \frac{\boldsymbol{u}\boldsymbol{v}^\mathsf{T}A^{-1}\boldsymbol{u}\boldsymbol{v}^\mathsf{T}A^{-1}}{1+\boldsymbol{v}^\mathsf{T}A^{-1}\boldsymbol{u}}$$

$$= I+\boldsymbol{u}\cdot\left(1-\frac{1}{1+\boldsymbol{v}^\mathsf{T}A^{-1}\boldsymbol{u}}-\frac{\boldsymbol{v}^\mathsf{T}A^{-1}\boldsymbol{u}}{1+\boldsymbol{v}^\mathsf{T}A^{-1}\boldsymbol{u}}\right)\cdot\boldsymbol{v}^\mathsf{T}A^{-1} = I$$

■

Appendix C. Privacy Theorems

In this section, we provide the formal proofs the our algorithms are differential privacy. We comment that, because we hope these algorithms will be implemented, we took the time to analyze the exact constants in our proofs rather than settling for $O(\cdot)$-notation. In addition to the three algorithms we provide, we give another theorem about the privacy of an algorithm that adds Gaussian noise to the inverse of the data, which may be of independent interest.

C.1. Privacy Proof for Algorithm 1

Theorem 11 *Fix $\epsilon > 0$ and $\delta \in (0, \frac{1}{e})$. Fix $B > 0$. Fix a positive integer r and let w be such that*

$$w^2 = B^2\left(1+\frac{1+\frac{\epsilon}{\ln(4/\delta)}}{\epsilon}\left(2\sqrt{2r\ln(\tfrac{4}{\delta})}+2\ln(\tfrac{4}{\delta})\right)\right)$$

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 $\sigma_{\min}(A)\ge w$, the algorithm that picks a $(r\times n)$-matrix R whose entries are iid samples from a normal distribution $\mathcal{N}(0,1)$ and publishes $R\cdot A$ is (ϵ,δ)-differentially private.

Corollary 12 *assuming $\epsilon<1$ and $\delta<e^{-1}$, if it holds that $r\ge 2\ln(\frac{4}{\delta})$ then it suffices to have $w^2\ge 8B^2\frac{\sqrt{r\ln(4/\delta)}}{\epsilon}$ for the results of Theorem 11 to hold. Alternatively, given input where its least singular value is publicly known to w, we can set $r=\left\lceil\left(\frac{\epsilon w^2}{8B^2\ln(\frac{4}{\delta})}\right)^2\right\rceil$, if indeed $r>2\ln(\frac{4}{\delta})$ and satisfy (ϵ,δ)-differential privacy. Therefore, if the rows of A are i.i.d draws from a $\mathbf{0}$-mean multivariate Gaussian with variance Σ, then we may set r as $\left\lceil\left(n\frac{\epsilon\sigma_{\min}(\Sigma)}{8B^2\ln(\frac{4}{\delta})}\right)^2\right\rceil = \Omega(n^2)$.*

Proof Fix A and A' be two neighboring $(n\times d)$ matrices, s.t. $A-A'$ is a rank-1 matrix of the form $E\overset{\text{def}}{=}A-A'=e_i(\boldsymbol{v}-\boldsymbol{v}')^\mathsf{T}$. We thus denote M as the matrix with the i-th row zeroed out, and have $M^\mathsf{T}M=A^\mathsf{T}A-\boldsymbol{v}\boldsymbol{v}^\mathsf{T}=A'^\mathsf{T}A'-\boldsymbol{v}'\boldsymbol{v}'^\mathsf{T}$. Recall that we assume that $\sigma_{\min}(A),\sigma_{\min}(A')\ge w$ and $\|E\|=\|\boldsymbol{v}-\boldsymbol{v}'\|\le 2B$. We transpose A and R and denote $X=A^\mathsf{T}R^\mathsf{T}$ and $X'=(A')^\mathsf{T}R^\mathsf{T}$. For each column $\boldsymbol{y}_j$ of R^T it holds that $\boldsymbol{y}_j{}^\mathsf{T}\sim\mathcal{N}(\mathbf{0}_n,I_{n\times n})$, and therefore the j-th column of X is distributed like a random variable from $\mathcal{N}\left(\mathbf{0}_r, A^\mathsf{T}A\right)$. Furthermore, as the columns of R are independently chosen, so are the columns of X are independent of one another. Therefore, for any r vectors $\boldsymbol{x}_1, ..., \boldsymbol{x}_r \in \mathbb{R}^d$ it holds that

$$\begin{aligned}
\mathsf{PDF}_X(\boldsymbol{x}_1, ..., \boldsymbol{x}_r) &= \prod_{j=1}^{r} \left(\sqrt{(2\pi)^d \det(A^\mathsf{T}A)}\right)^{-1} \exp\left(-\tfrac{1}{2}{\boldsymbol{x}_j}^\mathsf{T}(A^\mathsf{T}A)^{-1}\boldsymbol{x}_j\right) \\
\mathsf{PDF}_{X'}(\boldsymbol{x}_1, ..., \boldsymbol{x}_r) &= \prod_{j=1}^{r} \left(\sqrt{(2\pi)^d \det(A'^\mathsf{T}A')}\right)^{-1} \exp\left(-\tfrac{1}{2}{\boldsymbol{x}_j}^\mathsf{T}(A'^\mathsf{T}A')^{-1}\boldsymbol{x}_j\right)
\end{aligned}$$

We apply the Matrix Determinant Lemma, and the Sherman-Morrison Lemma, and deduce:

$$\begin{aligned}
\det(A^\mathsf{T}A) &= \det(M^\mathsf{T}M)\left(1 + \boldsymbol{v}^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}\right) \\
\det(A'^\mathsf{T}A') &= \det(M^\mathsf{T}M)\left(1 + \boldsymbol{v}'^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}'\right) \\
(A^\mathsf{T}A)^{-1} &= (M^\mathsf{T}M)^{-1} - \frac{(M^\mathsf{T}M)^{-1}\boldsymbol{v}\boldsymbol{v}^\mathsf{T}(M^\mathsf{T}M)^{-1}}{1 + \boldsymbol{v}^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}} \\
(A'^\mathsf{T}A')^{-1} &= (M^\mathsf{T}M)^{-1} - \frac{(M^\mathsf{T}M)^{-1}\boldsymbol{v}'\boldsymbol{v}'^\mathsf{T}(M^\mathsf{T}M)^{-1}}{1 + \boldsymbol{v}'^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}'}
\end{aligned}$$

Together with the inequality $\frac{1+x}{1+y} = (1+x)(1-\frac{y}{1+y}) \leq \exp(x - \frac{y}{1-y})$ for any $x, y \neq 1$ we have

$$\begin{aligned}
&\frac{\mathsf{PDF}_X(\boldsymbol{x}_1, ..., \boldsymbol{x}_r)}{\mathsf{PDF}_{X'}(\boldsymbol{x}_1, ..., \boldsymbol{x}_r)} = \prod_{j=1}^{r} \sqrt{\frac{\det(A'^\mathsf{T}A')}{\det(A^\mathsf{T}A)}} \exp\left(-\tfrac{1}{2}{\boldsymbol{x}_j}^\mathsf{T}((A^\mathsf{T}A)^{-1} - (A'^\mathsf{T}A')^{-1})\boldsymbol{x}_j\right) \\
&= \prod_{j=1}^{r} \left(\frac{1 + \boldsymbol{v}'^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}'}{1 + \boldsymbol{v}^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}}\right)^{\frac{1}{2}} \exp\left(-\tfrac{1}{2}{\boldsymbol{x}_j}^\mathsf{T}((A^\mathsf{T}A)^{-1} - (A'^\mathsf{T}A')^{-1})\boldsymbol{x}_j\right) \\
&\leq \prod_{j=1}^{r} \exp\Big(\frac{1}{2}\Big(\boldsymbol{v}'^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}' - \frac{{\boldsymbol{x}_j}^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}'\boldsymbol{v}'^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{x}_j}{1 + \boldsymbol{v}'^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}'}\Big) \\
&\qquad\qquad + \frac{1}{2}\Big(-\frac{\boldsymbol{v}^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}}{1 + \boldsymbol{v}^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}} + \frac{{\boldsymbol{x}_j}^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}\boldsymbol{v}^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{x}_j}{1 + \boldsymbol{v}^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}}\Big)\Big) \\
&= \exp\Big(\frac{1}{2}\Big(r \cdot \boldsymbol{v}'^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}' - \frac{\boldsymbol{v}'^\mathsf{T}(M^\mathsf{T}M)^{-1}\left(\sum_{j=1}^{r}\boldsymbol{x}_j{\boldsymbol{x}_j}^\mathsf{T}\right)(M^\mathsf{T}M)^{-1}\boldsymbol{v}'}{1 + \boldsymbol{v}'^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}'}\Big)\Big) \\
&\qquad \cdot \exp\Big(\frac{1}{2}\Big(-\frac{r \cdot \boldsymbol{v}^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}}{1 + \boldsymbol{v}^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}} + \frac{\boldsymbol{v}^\mathsf{T}(M^\mathsf{T}M)^{-1}\Big(\sum_{j=1}^{r}\boldsymbol{x}_j{\boldsymbol{x}_j}^\mathsf{T}\Big)(M^\mathsf{T}M)^{-1}\boldsymbol{v}}{1 + \boldsymbol{v}^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}}\Big)\Big)
\end{aligned} \tag{2}$$

Denote

$$\begin{aligned}
z_1 &\overset{\text{def}}{=} \boldsymbol{v}'^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v}' - \boldsymbol{v}'^\mathsf{T}(M^\mathsf{T}M)^{-1}\left(\tfrac{1}{r}\sum_{j=1}^{r}\boldsymbol{x}_j{\boldsymbol{x}_j}^\mathsf{T}\right)(M^\mathsf{T}M)^{-1}\boldsymbol{v}' \\
z_2 &\overset{\text{def}}{=} \boldsymbol{v}^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v} - \boldsymbol{v}^\mathsf{T}(M^\mathsf{T}M)^{-1}\left(\tfrac{1}{r}\sum_{j=1}^{r}\boldsymbol{x}_j{\boldsymbol{x}_j}^\mathsf{T}\right)(M^\mathsf{T}M)^{-1}\boldsymbol{v}
\end{aligned}$$

we have that

$$\begin{aligned}\ln\left(\frac{\mathsf{PDF}_X(\boldsymbol{x}_1,...,\boldsymbol{x}_r)}{\mathsf{PDF}_{X'}(\boldsymbol{x}_1,...,\boldsymbol{x}_r)}\right) &\le \frac{r}{2}\Big(\frac{z_1}{1+\boldsymbol{v}'(M^\top M)^{-1}\boldsymbol{v}'} + \frac{-z_2}{1+\boldsymbol{v}(M^\top M)^{-1}\boldsymbol{v}} + \frac{(\boldsymbol{v}'(M^\top M)^{-1}\boldsymbol{v}')^2}{1+\boldsymbol{v}'(M^\top M)^{-1}\boldsymbol{v}'}\Big)\\ &\le \frac{r}{2}\left(|z_1|+|z_2|+\boldsymbol{v}'(M^\top M)^{-1}\boldsymbol{v}'\right)\end{aligned}$$

We now turn to analyze each of the above three terms separately. The easiest to bound are the terms $\boldsymbol{v}(M^\top M)^{-1}\boldsymbol{v}$ and $\boldsymbol{v}'(M^\top M)^{-1}\boldsymbol{v}'$. Weyl's inequality yields that $\sigma_{\min}(M^\top M) \ge \sigma_{\min}(A^\top A) - B^2$, and so we give both terms that bound $\frac{B^2}{w^2-B^2} = \left(\frac{w^2}{B^2}-1\right)^{-1}$. We turn to bounding $|z_1|$, $|z_2|$.

We continue assuming that $\boldsymbol{x}_1,\dots,\boldsymbol{x}_r$ were sampled from $A^\top A$. If they were sampled from $A'^\top A'$ then the proof is analogous. Denote X as the matrix whose columns are $\boldsymbol{x}_1,\dots,\boldsymbol{x}_r$. We have

$$\begin{aligned}z_2 &= ((M^\top M)^{-1}\boldsymbol{v})^\top\left(M^\top M - \left(\tfrac{1}{r}X^\top X\right)\right)(M^\top M)^{-1}\boldsymbol{v}\\ &= ((M^\top M)^{-1}\boldsymbol{v})^\top\left(A^\top A - \boldsymbol{v}\boldsymbol{v}^\top - \left(\tfrac{1}{r}X^\top X\right)\right)(M^\top M)^{-1}\boldsymbol{v}\\ &= ((M^\top M)^{-1}\boldsymbol{v})^\top(A^\top A)^{1/2}\left(I - (A^\top A)^{-1/2}\left(\tfrac{1}{r}X^\top X\right)(A^\top A)^{-1/2}\right)(A^\top A)^{1/2}(M^\top M)^{-1}\boldsymbol{v}\\ &\qquad - (\boldsymbol{v}(M^\top M)^{-1}\boldsymbol{v})^2\end{aligned}$$

Recall that X is a matrix whose rows are i.i.d samples from the multivariate Gaussian $\mathcal{N}\left(\mathbf{0}, A^\top A\right)$. Therefore, the rows of the matrix $X(A^\top A)^{-1/2}$ are i.i.d samples from $\mathcal{N}\left(\mathbf{0}, I_{d\times d}\right)$. In other words, the distribution of $X(A^\top A)^{-1/2}$ is the same as a matrix whose entries are i.i.d samples from $\mathcal{N}\left(0,1\right)$. We can therefore invoke Lemma 8 and have that w.p. $\ge 1-\delta/2$.

$$\begin{aligned}|z_2| &\le (2\sqrt{\tfrac{2\ln(4/\delta)}{r}} + \tfrac{2\ln(4/\delta)}{r})\left\|(A^\top A)^{1/2}(M^\top M)^{-1}\boldsymbol{v}\right\|^2 + (\boldsymbol{v}(M^\top M)^{-1}\boldsymbol{v})^2\\ &\le (2\sqrt{\tfrac{2\ln(4/\delta)}{r}} + \tfrac{2\ln(4/\delta)}{r})\cdot\left(\boldsymbol{v}^\top(M^\top M)^{-1}(M^\top M + \boldsymbol{v}\boldsymbol{v}^\top)(M^\top M)^{-1}\boldsymbol{v}\right) + (\boldsymbol{v}(M^\top M)^{-1}\boldsymbol{v})^2\\ &= (\boldsymbol{v}(M^\top M)^{-1}\boldsymbol{v})\left(2\sqrt{\tfrac{2\ln(4/\delta)}{r}} + \tfrac{2\ln(4/\delta)}{r}\right) + (\boldsymbol{v}(M^\top M)^{-1}\boldsymbol{v})^2\left(2\sqrt{\tfrac{2\ln(4/\delta)}{r}} + \tfrac{2\ln(4/\delta)}{r} + 1\right)\\ &\le \left(\tfrac{w^2}{B^2}-1\right)^{-1}\left(2\sqrt{\tfrac{2\ln(4/\delta)}{r}} + \tfrac{2\ln(4/\delta)}{r}\right) + \left(\tfrac{w^2}{B^2}-1\right)^{-2}\left(2\sqrt{\tfrac{2\ln(4/\delta)}{r}} + \tfrac{2\ln(4/\delta)}{r} + 1\right)\end{aligned}$$

As the bound on $|z_1|$ is the same as the bound on $|z_2|$ we conclude that

$$\begin{aligned}\ln\left(\frac{\mathsf{PDF}_X(\boldsymbol{x}_1,...,\boldsymbol{x}_r)}{\mathsf{PDF}_{X'}(\boldsymbol{x}_1,...,\boldsymbol{x}_r)}\right) &\le \frac{r}{2}\left(|z_1|+|z_2|+\boldsymbol{v}'(M^\top M)^{-1}\boldsymbol{v}'\right)\\ &\le \left(\tfrac{w^2}{B^2}-1\right)^{-1}\left(2\sqrt{2r\ln(4/\delta)} + 2\ln(4/\delta)\right) + \left(\tfrac{w^2}{B^2}-1\right)^{-2}\left(2\sqrt{2r\ln(4/\delta)} + 2\ln(4/\delta) + \frac{3r}{2}\right)\\ &\le \frac{\epsilon}{1+\frac{\epsilon}{\ln(4/\delta)}} + \epsilon^2\left(\frac{2\sqrt{2r\ln(4/\delta)} + 2\ln(4/\delta)}{(2\sqrt{2r\ln(4/\delta)} + 2\ln(4/\delta))^2} + \frac{3r}{16r\ln(4/\delta)}\right)\\ &\le \frac{\epsilon}{1+\frac{\epsilon}{\ln(4/\delta)}}\left(1+\frac{\epsilon}{\ln(4/\delta)}\left(\frac{1}{2}+\frac{3}{16}\right)\right) < \epsilon\end{aligned}$$

by plugging in the value of w^2. ■

C.2. Privacy Proof for Algorithm 2

Theorem 13 *Fix $\epsilon \in (0,1)$ and $\delta \in (0, \frac{1}{e})$. Fix $B > 0$. Let C_1 and C_2 be such that they satisfy*

$$\frac{2\sqrt{C_2}}{C_1(\sqrt{C_2}-1)^2} \leq \frac{\epsilon}{B^2}$$

(E.g., $C_1 = B^2$ and $C_2 = \frac{14}{\epsilon^2}$.) 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 ν-degrees of freedom using the scale matrix V (i.e., $N \sim \mathcal{W}_d(V, \nu)$) for any matrix V with least singular value $\sigma_{\min}(V) \geq C_1$ (e.g. $V = C_1 I_{d\times d}$) and $\nu \geq \lfloor d + 2C_2 \ln(4/\delta) \rfloor$. Then outputting $X = A^\mathsf{T}A + N$ is (ϵ, δ)-differentially private.

We comment that in order to sample such an N, one can sample a matrix $N' \in \mathbb{R}^{\nu \times d}$ of i.i.d normal Gaussians, multiple all entries by $B/\sqrt{\epsilon}$ and set $N' = N^\mathsf{T}N$.

Proof Fix A and A' that are two neighboring datasets that differ on the i-th row, denoted as $\boldsymbol{v}^\mathsf{T}$ in A and $\boldsymbol{v}'^\mathsf{T}$ in A'. Let M denote A or A' without the i-th row, i.e. $M^\mathsf{T}M = A^\mathsf{T}A - \boldsymbol{v}\boldsymbol{v}^\mathsf{T} = A'^\mathsf{T}A' - \boldsymbol{v}'\boldsymbol{v}'^\mathsf{T}$. Therefore, denoting $\sigma_{\min}(M)$ and $\sigma_{\min}(A)$ as the least singular value of M and A resp., we have that $\sigma^2_{\min}(M) \leq \sigma^2_{\min}(A) \leq \sigma^2_{\min}(M) + B^2$. Same holds for the least singular value of M and A'.

Recall that

$$\mathsf{PDF}_{\mathcal{W}_d(V,\nu)}(N) \propto \det(N)^{\frac{\nu-d-1}{2}} \exp\left(-\tfrac{1}{2}\mathrm{tr}(V^{-1}N)\right)$$

We argue that Wishart-matrix additive noise is (ϵ, δ)-different-ially private, using the explicit formulation of the PDF. For the time being, we ignore the issue of outputting a matrix X s.t. either $X - A^\mathsf{T}A$, $X - A'^\mathsf{T}A'$ or $X - M^\mathsf{T}M$ are non-invertible. (Note, if our input matrix is A, then $\mathbf{Pr}[X - A^\mathsf{T}A \text{ non invertible}] = \mathbf{Pr}_{N\sim\mathcal{W}_d(V,\nu)}[N \text{ non invertible}] = 0$. However, it is not a-priori clear why we should also have that $\mathbf{Pr}[X - A'^\mathsf{T}A' \text{ non invertible}] = 0$ and also have that $\mathbf{Pr}[X - M^\mathsf{T}M \text{ non invertible}] = 0$.) Later, we justify why such events can be ignored. We now bound the appropriate ratios. If we denote the output of the mechanism as a matrix X, then we compare

$$\begin{aligned}
\frac{\mathsf{PDF}_{\mathcal{W}_d(V,\nu)}(X - A^\mathsf{T} A)}{\mathsf{PDF}_{W_d(V,\nu)}(X - A'^\mathsf{T} A')} &= \left(\frac{\det(X - A^\mathsf{T} A)}{\det(X - A'^\mathsf{T} A')}\right)^{\frac{\nu-d-1}{2}} \\
&\qquad \cdot \exp\left(-\tfrac{1}{2}\left(\mathrm{tr}(V^{-1}(X - A^\mathsf{T} A)) - \mathrm{tr}(V^{-1}(X - A'^\mathsf{T} A'))\right)\right) \\
&= \left(\frac{\det(X - M^\mathsf{T} M - \boldsymbol{v}\boldsymbol{v}^\mathsf{T})}{\det(X - M^\mathsf{T} M - \boldsymbol{v}'\boldsymbol{v}'^\mathsf{T})}\right)^{\frac{\nu-d-1}{2}} \\
&\qquad \cdot \exp\left(-\tfrac{1}{2}\left(\mathrm{tr}(V^{-1}(X - A^\mathsf{T} A - X + A'^\mathsf{T} A'^\mathsf{T}))\right)\right) \\
&= \left(\frac{1 - \boldsymbol{v}^\mathsf{T}(X - M^\mathsf{T} M)^{-1}\boldsymbol{v}}{1 - \boldsymbol{v}'^\mathsf{T}(X - M^\mathsf{T} M)^{-1}\boldsymbol{v}'}\right)^{\frac{\nu-d-1}{2}} \\
&\qquad \cdot \exp\left(-\tfrac{1}{2}\mathrm{tr}(V^{-1}\boldsymbol{v}'\boldsymbol{v}'^\mathsf{T}) + \tfrac{1}{2}\mathrm{tr}(V^{-1}\boldsymbol{v}\boldsymbol{v}^\mathsf{T})\right) \\
&\overset{\mathrm{tr}(AB)=\mathrm{tr}(BA)}{=} \left(\frac{1 - \boldsymbol{v}^\mathsf{T}(X - M^\mathsf{T} M)^{-1}\boldsymbol{v}}{1 - \boldsymbol{v}'^\mathsf{T}(X - M^\mathsf{T} M)^{-1}\boldsymbol{v}'}\right)^{\frac{\nu-d-1}{2}} \\
&\qquad \cdot \exp\left(-\tfrac{1}{2}\boldsymbol{v}'^\mathsf{T} V^{-1}\boldsymbol{v}' + \tfrac{1}{2}\boldsymbol{v}^\mathsf{T} V^{-1}\boldsymbol{v}\right)
\end{aligned}$$

We can now use the inequality $\frac{1-x}{1-y} = (1-x)(1+\frac{y}{1-y}) \leq \exp(-x + \frac{y}{1-y})$ for any x and any $y \neq 1$ to deduce

$$\begin{aligned}
\ln\left(\frac{\mathsf{PDF}_{A^\mathsf{T} A+N}(X)}{\mathsf{PDF}_{A'^\mathsf{T} A'+N}(X)}\right) &\leq \frac{1}{2}\cdot \boldsymbol{v}^\mathsf{T}\left(V^{-1} - (\nu-d-1)(X - M^\mathsf{T} M)^{-1}\right)\boldsymbol{v} \\
&\quad + \frac{1}{2}\cdot \boldsymbol{v}'^\mathsf{T}\left(\frac{\nu-d-1}{1 - \boldsymbol{v}'^\mathsf{T}(X - M^\mathsf{T} M)^{-1}\boldsymbol{v}'}(X - M^\mathsf{T} M)^{-1} - V^{-1}\right)\boldsymbol{v}'
\end{aligned}$$

Note that we either have $X - M^\mathsf{T} M = X - A^\mathsf{T} A + \boldsymbol{v}\boldsymbol{v}^\mathsf{T} = N + \boldsymbol{v}\boldsymbol{v}^\mathsf{T}$ or $X - M^\mathsf{T} M = N + \boldsymbol{v}'\boldsymbol{v}'^\mathsf{T}$. And so, we continue assuming X was sampled using $A^\mathsf{T} A$, but the case X was sampled from $A'^\mathsf{T} A'$ is symmetric. Further, we only show a bound for the first term of the two above, as the other term will have the same upper bound.

Note that $(X - M^\mathsf{T} M)^{-1} = (X - A^\mathsf{T} A + \boldsymbol{v}\boldsymbol{v}^\mathsf{T})^{-1} = (X - A^\mathsf{T} A)^{-1} - \frac{(X-A^\mathsf{T} A)^{-1}\boldsymbol{v}\boldsymbol{v}^\mathsf{T}(X-A^\mathsf{T} A)^{-1}}{1+\boldsymbol{v}(X-A^\mathsf{T} A)^{-1}\boldsymbol{v}}$, hence

$$\begin{aligned}
\boldsymbol{v}^\mathsf{T}(X - M^\mathsf{T} M)^{-1}\boldsymbol{v} &= \boldsymbol{v}^\mathsf{T}(X - A^\mathsf{T} A)^{-1}\boldsymbol{v} - \frac{(\boldsymbol{v}^\mathsf{T}(X - A^\mathsf{T} A)^{-1}\boldsymbol{v})^2}{1 + \boldsymbol{v}^\mathsf{T}(X - A^\mathsf{T} A)^{-1}\boldsymbol{v}} \\
&= \frac{\boldsymbol{v}^\mathsf{T}(X - A^\mathsf{T} A)^{-1}\boldsymbol{v}}{1 + \boldsymbol{v}^\mathsf{T}(X - A^\mathsf{T} A)^{-1}\boldsymbol{v}} \\
\boldsymbol{v}'^\mathsf{T}(X - M^\mathsf{T} M)^{-1}\boldsymbol{v}' &= \boldsymbol{v}'^\mathsf{T}(X - A^\mathsf{T} A)^{-1}\boldsymbol{v}' - \frac{(\boldsymbol{v}'^\mathsf{T}(X - A^\mathsf{T} A)^{-1}\boldsymbol{v})^2}{1 + \boldsymbol{v}^\mathsf{T}(X - A^\mathsf{T} A)^{-1}\boldsymbol{v}}
\end{aligned}$$

And so we have:

$$\begin{aligned}
&\boldsymbol{v}^\intercal\left(V^{-1} - (\nu-d-1)(X - M^\intercal M)^{-1}\right)\boldsymbol{v}\\
&\quad = \boldsymbol{v}^\intercal\left(V^{-1} - (\nu-d+1)(X - M^\intercal M)^{-1}\right)\boldsymbol{v} + 2\boldsymbol{v}^\intercal(X - M^\intercal M)^{-1}\boldsymbol{v}\\
&\quad \le \boldsymbol{v}^\intercal\left(V^{-1} - (\nu-d+1)(X - A^\intercal A)^{-1}\right)\boldsymbol{v} + 2\boldsymbol{v}^\intercal(X - A^\intercal A)^{-1}\boldsymbol{v}\\
&\qquad\qquad + (\nu-d+1)(\boldsymbol{v}^\intercal(X - A^\intercal A)^{-1}\boldsymbol{v})^2
\end{aligned}$$

Now, note that $(X - A^\intercal A) \sim \mathcal{W}_d(V,\nu)$, and so $V^{-1/2}(X - A^\intercal A)V^{-1/2} \sim \mathcal{W}_d(I_{d\times d},\nu)$. So we invoke Lemma 8 on the term $\boldsymbol{v}^\intercal\left(V^{-1} - (\frac{1}{\nu-d+1}(X - A^\intercal A))^{-1}\right)\boldsymbol{v}$ $= (V^{-1/2}\boldsymbol{v})^\intercal\left(I - \left(\frac{V^{-1/2}(X-A^\intercal A)V^{-1/2}}{\nu-d+1}\right)^{-1}\right)(V^{-1/2}\boldsymbol{v})$ and infer that w.p. $\ge 1-\delta/2$ we have the following bound

$$\begin{aligned}
&\boldsymbol{v}^\intercal\left(V^{-1} - (\nu-d-1)(X - M^\intercal M)^{-1}\right)\boldsymbol{v}\\
&\quad \le \Big(\frac{2\sqrt{2(\nu-d+1)\ln(4/\delta)} - 2\ln(4/\delta)}{(\sqrt{\nu-d+1} - \sqrt{2\ln(4/\delta)})^2} + \frac{2}{(\sqrt{\nu-d+1} - \sqrt{2\ln(4/\delta)})^2} + \frac{2(\nu-d+1)}{(\sqrt{\nu-d+1} - \sqrt{2\ln(4/\delta)})^4}\Big)\cdot\|V^{-1/2}\boldsymbol{v}\|^2\\
&\quad = \frac{\|V^{-1/2}\boldsymbol{v}\|^2}{(\sqrt{\nu-d+1} - \sqrt{2\ln(4/\delta)})^2}\cdot\Big(2\sqrt{2(\nu-d+1)\ln(4/\delta)} - 2\ln(4/\delta) + 2 + \frac{2}{(1-\sqrt{\frac{2\ln(4/\delta)}{\nu-d-1}})^2}\Big)\\
&\quad = \frac{2\sqrt{2(\nu-d+1)\ln(4/\delta)} - 2\ln(4/\delta) + 6}{(\sqrt{\nu-d+1} - \sqrt{2\ln(4/\delta)})^2}\|V^{-1/2}\boldsymbol{v}\|^2
\end{aligned}$$

Analogously, w.p. $\ge 1-\delta/2$ the following bound holds as well:

$$\begin{aligned}
&\boldsymbol{v}'^\intercal\left(\frac{\nu-d-1}{1-\boldsymbol{v}'^\intercal(X - M^\intercal M)^{-1}\boldsymbol{v}'}(X - M^\intercal M)^{-1} - V^{-1}\right)\boldsymbol{v}'\\
&\quad = \boldsymbol{v}'^\intercal\left((\nu-d-1)(X - M^\intercal M)^{-1} - V^{-1}\right)\boldsymbol{v}' + \frac{(\nu-d-1)(\boldsymbol{v}'^\intercal(X - M^\intercal M)^{-1}\boldsymbol{v}')^2}{1-\boldsymbol{v}'^\intercal(X - M^\intercal M)^{-1}\boldsymbol{v}'}\\
&\quad \le \boldsymbol{v}'^\intercal\left((\nu-d+1)(X - M^\intercal M)^{-1} - V^{-1}\right)\boldsymbol{v}' + \frac{(\nu-d-1)(\boldsymbol{v}'^\intercal(X - M^\intercal M)^{-1}\boldsymbol{v}')^2}{1-\boldsymbol{v}'^\intercal(X - M^\intercal M)^{-1}\boldsymbol{v}'}\\
&\quad \le \boldsymbol{v}'^\intercal\left((\nu-d+1)(X - A^\intercal A)^{-1} - V^{-1}\right)\boldsymbol{v}' + \frac{(\nu-d-1)(\boldsymbol{v}'^\intercal(X - A^\intercal A)^{-1}\boldsymbol{v}')^2}{1-\boldsymbol{v}'^\intercal(X - M^\intercal M)^{-1}\boldsymbol{v}'}\\
&\quad \le \Big(\frac{2\sqrt{2(\nu-d+1)\ln(4/\delta)} - 2\ln(4/\delta)}{(\sqrt{\nu-d+1} - \sqrt{2\ln(4/\delta)})^2} + \frac{2(\nu-d+1)}{(\sqrt{\nu-d+1} - \sqrt{2\ln(4/\delta)})^4}\Big)\cdot\|V^{-1/2}\boldsymbol{v}'\|^2
\end{aligned}$$

Combining the two upper bounds we get

$$\begin{aligned}
&\ln\left(\frac{\mathsf{PDF}_{A^{\mathsf{T}}A+N}(X)}{\mathsf{PDF}_{A'^{\mathsf{T}}A'+N}(X)}\right)\\
&\leq \frac{1}{2}\boldsymbol{v}^{\mathsf{T}}\Big(V^{-1} - \frac{\nu-d-1}{1-\boldsymbol{v}'^{\mathsf{T}}(X-M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'}(X-M^{\mathsf{T}}M)^{-1}\Big)\boldsymbol{v}\\
&\qquad + \frac{1}{2}\boldsymbol{v}'^{\mathsf{T}}\left(\frac{\nu-d-1}{1-\boldsymbol{v}'^{\mathsf{T}}(X-M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'}(X-M^{\mathsf{T}}M)^{-1} - V^{-1}\right)\boldsymbol{v}'\\
&\leq \frac{2\sqrt{2(\nu-d+1)\ln(4/\delta)} - 2\ln(4/\delta)+6}{(\sqrt{\nu-d+1}-\sqrt{2\ln(4/\delta)})^2}\cdot\frac{\|V^{-1/2}\boldsymbol{v}\|^2+\|V^{-1/2}\boldsymbol{v}'\|^2}{2}\\
&\overset{\delta<\frac{1}{6}}{\leq} \frac{B^2}{\sigma_{\min}(V)}\cdot\frac{2\sqrt{2(\nu-d+1)\ln(4/\delta)}}{(\sqrt{\nu-d+1}-\sqrt{2\ln(4/\delta)})^2}
\end{aligned}$$

All we now need to do is to plug in the fact that $\nu = \lfloor d + C_2\cdot 2\ln(4/\delta)\rfloor \geq d-1+C_2\cdot 2\ln(4/\delta)$, and that $\sigma_{\min}(V)\geq C_1$ to deduce

$$\ln\left(\frac{\mathsf{PDF}_{A^{\mathsf{T}}A+N}(X)}{\mathsf{PDF}_{A'^{\mathsf{T}}A'+N}(X)}\right) \leq \frac{B^2}{C_1}\cdot\frac{2\cdot 2\ln(4/\delta)\cdot\sqrt{C_2}}{(\sqrt{C_2\cdot 2\ln(4/\delta)}-\sqrt{2\ln(4/\delta)})^2} \leq \frac{2B^2\sqrt{C_2}}{C_1(\sqrt{C_2}-1)^2} \leq \epsilon$$

■

C.3. Privacy Proof for Algorithm 5

While the proof of the Additive Wishart Algorithm is proven in the main body, we now prove that the version of the Additive Wishart Algorithm that maintains the intercept is also privacy-preserving.

Variation: Intercept. As we mentioned earlier, it will be useful for us to deal with the case of the input has a all-1 column (representing the intercept in linear regression). In this case, the random example we pad the input with will also have 1 in that column. And so, the matrix which is originally $[A;\mathbf{1}]$ with A being our input $(n\times p)$-matrix [12] is padded with random examples turns into $\begin{bmatrix} A & \mathbf{1}\\ R & \mathbf{1}\end{bmatrix}$, where R is a random $(\nu\times p)$-matrix. We then output the 2nd moment matrix of the padded matrix, i.e. the matrix: $\begin{bmatrix} A^{\mathsf{T}}A + R^{\mathsf{T}}R & A^{\mathsf{T}}\mathbf{1}+R^{\mathsf{T}}\mathbf{1}\\ (A^{\mathsf{T}}\mathbf{1}+R^{\mathsf{T}}\mathbf{1})^{\mathsf{T}} & n+\nu\end{bmatrix}$. It is evident that the last coordinate in the output matrix is deterministically always $n+\nu$, and therefore this algorithm reveals the number of entries in the data. However, we assume n is known in advance and do not consider this a privacy violation. For simplicity, we analyze solely the case where every entry of R is chosen i.i.d from $\mathcal{N}(0, C_1)$. Let $\mathcal{P}_A$ be the distribution over outputs we get given that the input is A. (Since the output is a symmetric matrix, we denote it as a distribution over pairs $(X,\boldsymbol{u})$ with X being a $(p\times p)$-matrix and $\boldsymbol{u}$ being a p-dimensional vector.) We argue the following.

Claim 2 *In our setting, we have*

$$\mathsf{PDF}_{\mathcal{P}_A}(A^{\mathsf{T}}A + X, A^{\mathsf{T}}\mathbf{1}+\boldsymbol{u}) \propto \exp\left(-\tfrac{\|\boldsymbol{u}\|^2}{2C_1\cdot p}\right)\cdot\det(X-\tfrac{1}{\nu}\boldsymbol{u}\boldsymbol{u}^{\mathsf{T}})^{\frac{\nu-p-2}{2}}\exp(-\tfrac{\mathrm{tr}(X-\frac{1}{\nu}\boldsymbol{u}\boldsymbol{u}^{\mathsf{T}})}{2C_1})$$

[12] We keep to the convention of defining p as $d = p + 1$.

That is, the density function of $\mathcal{P}_A$ at $(A^\mathsf{T}A + X, A^\mathsf{T}\mathbf{1} + u)$ is the product of sampling $\boldsymbol{u}$ from a Guassian $\mathcal{N}(\mathbf{0}, C_1\nu \cdot I_{p\times p})$ times sampling the matrix $(X - \frac{1}{\nu}\boldsymbol{u}\boldsymbol{u}^T)$ from a Wishart distribution $\mathcal{W}_p(C_1 I_{p\times p}, \nu - 1)$ with $(\nu-1)$-degrees of freedom. I.e.

$$\mathsf{PDF}_{\mathcal{P}_A}(A^\mathsf{T}A + X, A^\mathsf{T}\mathbf{1} + \boldsymbol{u}) = \mathsf{PDF}_{\mathcal{N}(\mathbf{0}, C_1\nu\cdot I_{p\times p})}(\boldsymbol{u}) \cdot \mathsf{PDF}_{\mathcal{W}_p(C_1 I_{p\times p}, \nu-1)}(X - \tfrac{1}{\nu}\boldsymbol{u}\boldsymbol{u}^\mathsf{T})$$

Proof By definition, $\mathsf{PDF}_{\mathcal{P}_A}(A^\mathsf{T}A + X, A^\mathsf{T}\mathbf{1} + \boldsymbol{u}) = \mathsf{PDF}[R^\mathsf{T}\mathbf{1} = \boldsymbol{u}] \cdot \mathsf{PDF}(R^\mathsf{T}R = X \mid R^\mathsf{T}\mathbf{1} = \boldsymbol{u})$. Each column of R is sampled independently from $\mathcal{N}(\mathbf{0}, C_1 \cdot I_{\nu\times\nu})$, thus each of the p coordinates of $R^\mathsf{T}\mathbf{1}$ is sampled i.i.d from $N(0, C_1\cdot\nu)$, hence $\mathsf{PDF}[R^\mathsf{T}\mathbf{1} = \boldsymbol{u}] = (2\pi C_1\cdot\nu)^{-p/2}\cdot \exp\left(-\frac{\|\boldsymbol{u}\|^2}{2C_1\cdot\nu}\right)$.

Let $\boldsymbol{v}_1 = \frac{1}{\sqrt{\nu}}\mathbf{1}$ be a unit length vector, and denote $\boldsymbol{v}_2, \ldots, \boldsymbol{v}_\nu$ a completion of $\boldsymbol{v}_1$ into an orthonormal basis for $\mathbb{R}^\nu$. Denote the $(\nu\times\nu)$-matrix they form as V s.t. $VV^\mathsf{T} = I_{\nu\times\nu}$. Hence $R^\mathsf{T}R = R^\mathsf{T}VV^\mathsf{T}R = \left[\begin{array}{c} R^\mathsf{T}\boldsymbol{v}_1 \ldots R^\mathsf{T}\boldsymbol{v}_\nu \end{array}\right]\left[\begin{array}{c} {\boldsymbol{v}_1}^\mathsf{T}R \\ \vdots \\ {\boldsymbol{v}_\nu}^\mathsf{T}R \end{array}\right] = \sum_i (R^\mathsf{T}\boldsymbol{v}_i)(R^\mathsf{T}\boldsymbol{v}_i)^\mathsf{T} = \frac{1}{\nu}\boldsymbol{u}\boldsymbol{u}^\mathsf{T} + \sum_{i>1}(R^\mathsf{T}\boldsymbol{v}_i)(R^\mathsf{T}\boldsymbol{v}_i)^\mathsf{T}$. We thus have that given $R^\mathsf{T}\mathbf{1} = \boldsymbol{u}$ then $R^\mathsf{T}R = X$ if and only if $\sum_{i>1}(R^\mathsf{T}\boldsymbol{v}_i)(R^\mathsf{T}\boldsymbol{v}_i)^\mathsf{T} = X - \frac{1}{\nu}\boldsymbol{u}\boldsymbol{u}^\mathsf{T}$. The key point is that each row of R is chosen from a spherical Gaussian, and so its projection onto $\boldsymbol{v}_i$ is independent of its projection onto $\boldsymbol{v}_j$ for any $i \neq j$. Moreover, since each row of R is sampled i.i.d from $\mathcal{N}(\mathbf{0}, C_1 I_{\nu\times nu})$ we have that each coordinate of $R^\mathsf{T}\boldsymbol{v}_i$ is sampled i.i.d from $\mathcal{N}(0, C_1)$. Hence, $\sum_{i>1}(R^\mathsf{T}\boldsymbol{v}_i)(R^\mathsf{T}\boldsymbol{v}_i)^\mathsf{T}$ is the sum of $(\nu-1)$ independent p-dimensional outer products of vectors sampled from $\mathcal{N}(\mathbf{0}, C_1 I_{p\times p})$. By definition this is the Wishart distribution with $(\nu-1)$ degrees of freedom $\mathcal{W}_p(C_1 I_{p\times p}, \nu-1)$. ■

We can now prove that padding A with $\nu = p + O(\ln(1/\delta)/\epsilon^2)$ random examples from $\mathcal{N}(\mathbf{0}, (B^2\ln(1/\delta)/\epsilon^2)I_{p\times p})$ *while keeping the intercept column filled with 1s* is still differentially private.

Theorem 14 *Fix $\epsilon \in (0,1)$ and $\delta \in (0, \frac{1}{e})$. Fix integers n and p and a scalar $B > 0$. Let A be a $(n\times p)$-matrix where each row of A has bounded L_2-norm of B. Set an integer $\nu \geq p + 1 + 250\ln(20/\delta)/\epsilon^2$ and let R denote a Gaussian matrix whose ν rows are sampled from $\mathcal{N}(\mathbf{0}, (B^2\ln(1/\delta)/\epsilon^2)I_{p\times p})$. Then outputting the 2^{nd} moment matrix $\left[\begin{array}{cc} A^\mathsf{T}A + R^\mathsf{T}R & A^\mathsf{T}\mathbf{1} + R^\mathsf{T}\mathbf{1} \\ (A^\mathsf{T}\mathbf{1} + R^\mathsf{T}\mathbf{1})^\mathsf{T} & n+\nu \end{array}\right]$ is (ϵ,δ)-differentially private.*

Proof Fix A and A' that are two neighboring datasets that differ on the i-th row, denoted as $\boldsymbol{v}^\mathsf{T}$ in A and $\boldsymbol{v}'^\mathsf{T}$ in A'. Let M denote A or A' without the i-th row, i.e. $M^\mathsf{T}M = A^\mathsf{T}A - \boldsymbol{v}\boldsymbol{v}^\mathsf{T} = A'^\mathsf{T}A' - \boldsymbol{v}'\boldsymbol{v}'^\mathsf{T}$. Therefore, denoting $\sigma_{\min}(M)$ and $\sigma_{\min}(A)$ as the least singular value of M and A resp., we have that $\sigma^2_{\min}(M) \leq \sigma^2_{\min}(A) \leq \sigma^2_{\min}(M) + B^2$. Same holds for the least singular value of M and A'. Given that the output of that algorithm is $(M^\mathsf{T}M + X; M^\mathsf{T}\mathbf{1} + \boldsymbol{y})$ we are upper bounding the ratio

$$\begin{aligned}
&\frac{\mathsf{PDF}_{P_A}(M^\mathsf{T}M + X; M^\mathsf{T}\mathbf{1} + \boldsymbol{y})}{\mathsf{PDF}_{P_{A'}}(M^\mathsf{T}M + X; M^\mathsf{T}\mathbf{1} + \boldsymbol{y})} \\
&= \frac{\mathsf{PDF}_{\mathcal{N}(\mathbf{0}, B^2\nu I_{p\times p})}(\boldsymbol{y} - \boldsymbol{v})}{\mathsf{PDF}_{\mathcal{N}(\mathbf{0}, B^2\nu I_{p\times p})}(\boldsymbol{y} - \boldsymbol{v}')} \cdot \frac{\mathsf{PDF}_{\mathcal{W}_p(\sigma^2 I_{p\times p}, \nu-1)}(X - \boldsymbol{v}\boldsymbol{v}^\mathsf{T} - \frac{1}{\nu}(\boldsymbol{y} - \boldsymbol{v})(\boldsymbol{y} - \boldsymbol{v})^\mathsf{T})}{\mathsf{PDF}_{\mathcal{W}_p(\sigma^2 I_{p\times p}, \nu-1)}(X - \boldsymbol{v}'\boldsymbol{v}'^\mathsf{T} - \frac{1}{\nu}(\boldsymbol{y} - \boldsymbol{v}')(\boldsymbol{y} - \boldsymbol{v}')^\mathsf{T})}
\end{aligned}$$

Standard results from differential privacy through additive Gaussian noise assure that w.p. $\geq 1-\delta/5$ we have that $\|\boldsymbol{y}\| \leq B\sqrt{\nu}(\sqrt{p}+\sqrt{2\ln(20/\delta)})$ and as a result $\frac{\mathsf{PDF}_{\mathcal{N}(\mathbf{0},B^2\nu I_{p\times p})}(\boldsymbol{y}-\boldsymbol{v})}{\mathsf{PDF}_{\mathcal{N}(\mathbf{0},B^2\nu I_{p\times p})}(\boldsymbol{y}-\boldsymbol{v}')} \leq \exp(\epsilon/5)$, provided $B^2\nu \geq 50B^2\ln(6/\delta)/\epsilon^2$.

We now note that

$$\begin{aligned} & X - \boldsymbol{v}\boldsymbol{v}^\mathsf{T} - \tfrac{1}{\nu}(\boldsymbol{y}-\boldsymbol{v})(\boldsymbol{y}-\boldsymbol{v})^\mathsf{T} \\ &\quad = X - \boldsymbol{v}'\boldsymbol{v}'^\mathsf{T} + (\boldsymbol{v}'\boldsymbol{v}'^\mathsf{T} - \boldsymbol{v}\boldsymbol{v}^\mathsf{T}) \\ &\qquad - \tfrac{1}{\nu}\Big[(\boldsymbol{y}-\boldsymbol{v}')(\boldsymbol{y}-\boldsymbol{v}')^\mathsf{T} + (\boldsymbol{v}'-\boldsymbol{v})(\boldsymbol{y}-\boldsymbol{v}')^\mathsf{T} + (\boldsymbol{y}-\boldsymbol{v}')(\boldsymbol{v}'-\boldsymbol{v})^\mathsf{T} + (\boldsymbol{v}-\boldsymbol{v}')(\boldsymbol{v}-\boldsymbol{v}')^\mathsf{T}\Big] \\ &= X - \boldsymbol{v}'\boldsymbol{v}'^\mathsf{T} + (\boldsymbol{v}'\boldsymbol{v}'^\mathsf{T} - \boldsymbol{v}\boldsymbol{v}^\mathsf{T}) \\ &\qquad - \tfrac{1}{\nu}(\boldsymbol{y}-\boldsymbol{v}')(\boldsymbol{y}-\boldsymbol{v}')^\mathsf{T} - \tfrac{1}{\nu}(\boldsymbol{v}'-\boldsymbol{v})(\boldsymbol{y}-\boldsymbol{v}')^\mathsf{T} - \tfrac{1}{\nu}(\boldsymbol{y}-\boldsymbol{v}')(\boldsymbol{v}'-\boldsymbol{v})^\mathsf{T} - \tfrac{1}{\nu}(\boldsymbol{v}-\boldsymbol{v}')(\boldsymbol{v}-\boldsymbol{v}')^\mathsf{T} \end{aligned}$$

Denoting $\Delta_1 = (\boldsymbol{v}'\boldsymbol{v}'^\mathsf{T} - \boldsymbol{v}\boldsymbol{v}^\mathsf{T})$ and $\Delta_2 = \frac{1}{\nu}(\boldsymbol{v}'-\boldsymbol{v})(\boldsymbol{y}-\boldsymbol{v}')^\mathsf{T} + \frac{1}{\nu}(\boldsymbol{y}-\boldsymbol{v}')(\boldsymbol{v}'-\boldsymbol{v})^\mathsf{T} + \frac{1}{\nu}(\boldsymbol{v}-\boldsymbol{v}')(\boldsymbol{v}-\boldsymbol{v}')^\mathsf{T}$. From definition, we have that Δ_2 is a symmetric rank-3 matrix. Moreover, as $\|\boldsymbol{v}\|$ and $\|\boldsymbol{v}'\|$ are upper bounded by B and $\|\boldsymbol{y}-\boldsymbol{v}\|$ and $\|\boldsymbol{y}-\boldsymbol{v}'\|$ are upper bounded by $B\sqrt{\nu}(\sqrt{p}+\sqrt{2\ln(12/\delta)}+1)$, it holds that the Frobenius norm of Δ_2 is upper bounded by $B^2(2\sqrt{2}\frac{\sqrt{p}+\sqrt{2\ln(12/\delta)}+1}{\sqrt{\nu}} + \frac{4}{\nu})$. From the definition of ν we thus have that $\|\Delta_2\|_F \leq 3B^2$. Hence we can write Δ_2 as the sum of 3 symmetric rank-1 matrices, $\boldsymbol{u}_i\boldsymbol{u}_i^\mathsf{T}$ with $\|\boldsymbol{u}_i\| \leq \sqrt{3}B$.

We now apply Theorem 13, twice: once for a single change in the form of Δ_1 (replacing $\boldsymbol{v}$ with $\boldsymbol{v}'$) and once for the 3-changes the form Δ_2. (Namely, we use the property of group privacy.) This is why we choose ν s.t. $\nu - 1 \geq p + 250\ln(20/\delta)/\epsilon^2$ — so that we can upper bound the ratio $\frac{\mathsf{PDF}_{\mathcal{W}_p(\sigma^2 I_{p\times p},\nu-1)}(X-\boldsymbol{v}\boldsymbol{v}^\mathsf{T}-\frac{1}{\nu}(\boldsymbol{y}-\boldsymbol{v})(\boldsymbol{y}-\boldsymbol{v})^\mathsf{T})}{\mathsf{PDF}_{\mathcal{W}_p(\sigma^2 I_{p\times p},\nu-1)}(X-\boldsymbol{v}'\boldsymbol{v}'^\mathsf{T}-\frac{1}{\nu}(\boldsymbol{y}-\boldsymbol{v}')(\boldsymbol{y}-\boldsymbol{v}')^\mathsf{T})}$ by $\exp(4\epsilon/5)$ w.p. $\geq 1-4\delta/5$. We thus get an upper bound of e^ϵ overall, w.p. $\geq 1-\delta$. ■

C.4. Privacy Proof for Algorithm 3

Theorem 15 *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*

$$w^2 = \frac{B^2}{\epsilon(1-\frac{\epsilon}{2\ln(4/\delta)})}\left(2\sqrt{2\nu\ln(4/\delta)} + 2\ln(4/\delta)\right)$$

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 similarity between the bounds of Theorem 11 and Theorem 15. This is after all quite natural, since the JL-theorem is a way to sample from a Wishart distribution $\mathcal{W}_d(A^\mathsf{T}A, r)$ (since every row in the matrix RA is an i.i.d sample from $\mathcal{N}\Big(\mathbf{0}, A^\mathsf{T}A\Big)$). Clearly, one can sample a matrix from $\mathcal{W}_d(A^\mathsf{T}A, r)$ and invert it, to get a sample from $\mathcal{W}_d^{-1}((A^\mathsf{T}A)^{-1}, r)$ and vice-versa. Therefore, we get similar bounds. The only slight difference lies in the fact that we require in Theorem 15 that $\nu \geq d$, s.t. the matrix we sample is indeed invertible, whereas we do not require any such lower bound for sampling from $\mathcal{W}_d(A^\mathsf{T}A, r)$.

Proof As always, we denote A' as a neighbor of A that differs just on a single row, which we denote $\boldsymbol{v}$ for A and $\boldsymbol{v}'$ for A', and as before, the matrix M is the matrix A with the i-th row all zeroed out. Therefore, $A^\mathsf{T}A - \boldsymbol{v}\boldsymbol{v}^\mathsf{T} = A'^\mathsf{T}A' - \boldsymbol{v}'\boldsymbol{v}'^\mathsf{T} = M^\mathsf{T}M$. So, denoting $\sigma_{\min}(M)$ and $\sigma_{\min}(A)$ as the least singular value of M and A resp., we have that $\sigma^2_{\min}(M) \le \sigma^2_{\min}(A) \le \sigma^2_{\min}(M) + B^2$. Same holds for the least singular value of M and A'.

Recall that $\mathsf{PDF}_{\mathcal{W}_d^{-1}(A^\top A,\nu)}(X) \propto \det(A^\top A)^{\frac{\nu}{2}} \det(X)^{\frac{\nu+p+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}((A^\top A)X^{-1})\right)$. So like before, we invoke the determinant update lemma, the Sherman Morisson lemma and the inequality $\frac{1+x}{1+y} \le \exp(x - \frac{y}{1+y})$ and deduce:

$$\begin{aligned}
\frac{\mathsf{PDF}_{\mathcal{W}_d^{-1}(A^\top A,\nu)}(X)}{\mathsf{PDF}_{\mathcal{W}_d^{-1}(A'^\top A',\nu)}(X)} &= \frac{\det(A^\top A)^{\nu/2}\exp\left(-\frac{1}{2}\mathrm{tr}((A^\top A)X^{-1})\right)}{\det(A'^\top A')^{\nu/2}\exp\left(-\frac{1}{2}\mathrm{tr}((A'^\top A')X^{-1})\right)} \\
&= \left(\frac{1+\boldsymbol{v}^\top(M^\top M)^{-1}\boldsymbol{v}}{1+\boldsymbol{v}'^\top(M^\top M)^{-1}\boldsymbol{v}'}\right)^{\nu/2}\exp\left(-\frac{1}{2}\mathrm{tr}((A^\top A - A'^\top A')X^{-1}\right) \\
&\le \exp\left(\frac{\nu}{2}\left(\boldsymbol{v}^\top(M^\top M)^{-1}\boldsymbol{v} - \frac{\boldsymbol{v}'^\top(M^\top M)^{-1}\boldsymbol{v}'}{1+\boldsymbol{v}'^\top(M^\top M)^{-1}\boldsymbol{v}'}\right)\right)\cdot\exp\left(-\frac{1}{2}\left(\mathrm{tr}((\boldsymbol{v}\boldsymbol{v}^\top - \boldsymbol{v}'\boldsymbol{v}'^\top)X^{-1})\right)\right) \\
&= \exp\left(\frac{1}{2}\left(\nu\cdot\boldsymbol{v}^\top(M^\top M)^{-1}\boldsymbol{v} - \boldsymbol{v}^\top X^{-1}\boldsymbol{v}\right)\right)\cdot\exp\left(-\frac{1}{2}\left(\frac{\nu\cdot\boldsymbol{v}'^\top(M^\top M)^{-1}\boldsymbol{v}'}{1+\boldsymbol{v}'^\top(M^\top M)^{-1}\boldsymbol{v}'} - \boldsymbol{v}'^\top X^{-1}\boldsymbol{v}'\right)\right) \\
&\le \exp\left(\frac{1}{2}\boldsymbol{v}^\top\left(\nu(M^\top M)^{-1} - X^{-1}\right)\boldsymbol{v}\right)\cdot\exp\left(-\frac{1}{2}\boldsymbol{v}'^\top\left(\frac{\nu}{1+\boldsymbol{v}'^\top(M^\top M)^{-1}\boldsymbol{v}'}(M^\top M)^{-1} - X^{-1}\right)\boldsymbol{v}'\right)
\end{aligned}$$

We continue assuming $X \sim \mathcal{W}_d^{-1}(A^\top A,\nu)$ (the case $X \sim \mathcal{W}_d^{-1}(A'^\top A',\nu)$ is symmetric). By definition, we have that $X^{-1} \sim \mathcal{W}_d((A^\top A)^{-1},\nu)$. Hence $(A^\top A)^{1/2}X^{-1}(A^\top A)^{-1/2} \sim \mathcal{W}_d(I_{d\times d},\nu)$, which implies that the distribution of $(A^\top A)^{1/2}X^{-1}(A^\top A)^{-1/2}$ is the same as sampling a $(\nu\times d)$-matrix of i.i.d $\mathcal{N}(0,1)$ samples and take its cross-product with itself.

We continue using the Sherman-Morrison formula, and derive the bound

$$\begin{aligned}
\boldsymbol{v}^\top\left(\nu(M^\top M)^{-1} - X^{-1}\right)\boldsymbol{v} &= \boldsymbol{v}^\top\left(\nu(A^\top A)^{-1} - X^{-1}\right)\boldsymbol{v} - \frac{\nu\cdot(\boldsymbol{v}^\top(A^\top A)^{-1}\boldsymbol{v})^2}{1-\boldsymbol{v}^\top(A^\top A)^{-1}\boldsymbol{v}} \\
&\le ((A^\top A)^{-1/2}\boldsymbol{v})^\top\cdot\left(\nu I_{d\times d} - (A^\top A)^{1/2}X^{-1}(A^\top A)^{1/2}\right)\cdot((A^\top A)^{-1/2}\boldsymbol{v}) \\
&\le \|(A^\top A)^{-1/2}\boldsymbol{v}\|^2\left(2\sqrt{2\nu\ln(4/\delta)} + 2\ln(4/\delta)\right)
\end{aligned}$$

which holds w.p. $\ge 1-\delta/2$ due to Lemma 8. Similarly, we have

$$\begin{aligned}
&-\boldsymbol{v}'^\top\left(\frac{\nu}{1+\boldsymbol{v}'^\top(M^\top M)^{-1}\boldsymbol{v}'}(M^\top M)^{-1} - X^{-1}\right)\boldsymbol{v}' \\
&\quad= -\boldsymbol{v}'^\top\left(\nu(M^\top M)^{-1} - X^{-1}\right)\boldsymbol{v}' + \frac{\nu\cdot(\boldsymbol{v}'^\top(M^\top M)^{-1}\boldsymbol{v}')^2}{1-\boldsymbol{v}'^\top(M^\top M)^{-1}\boldsymbol{v}'} \\
&\quad= -\boldsymbol{v}'^\top\left(\nu(A^\top A)^{-1} - X^{-1}\right)\boldsymbol{v}' + \frac{\nu\cdot(\boldsymbol{v}'^\top(M^\top M)^{-1}\boldsymbol{v}')^2}{1-\boldsymbol{v}'^\top(M^\top M)^{-1}\boldsymbol{v}'} + \frac{\nu\cdot(\boldsymbol{v}'^\top(A^\top A)^{-1}\boldsymbol{v})^2}{1-\boldsymbol{v}'^\top(A^\top A)^{-1}\boldsymbol{v}'} \\
&\quad\le -\boldsymbol{v}'^\top\left(\nu(A^\top A)^{-1} - X^{-1}\right)\boldsymbol{v}' + \frac{\nu\cdot(\boldsymbol{v}'^\top(A^\top A)^{-1}\boldsymbol{v})^2}{1-\boldsymbol{v}'^\top(M^\top M)^{-1}\boldsymbol{v}'} + \frac{\nu\cdot(\boldsymbol{v}'^\top(A^\top A)^{-1}\boldsymbol{v})^2}{1-\boldsymbol{v}'^\top(A^\top A)^{-1}\boldsymbol{v}'} \\
&\quad\le \|(A^\top A)^{-1/2}\boldsymbol{v}'\|^2\left(2\sqrt{2\nu\ln(4/\delta)} + 2\ln(4/\delta)\right) \\
&\qquad + \nu\cdot\|(A^\top A)^{-1/2}\boldsymbol{v}'\|^2\|(A^\top A)^{-1/2}\boldsymbol{v}\|^2\cdot\left(\frac{1}{1-\boldsymbol{v}'^\top(M^\top M)^{-1}\boldsymbol{v}'} + \frac{1}{1-\boldsymbol{v}'^\top(A^\top A)^{-1}\boldsymbol{v}'}\right)
\end{aligned}$$

Denoting the least singular value of $(A^\mathsf{T}A)$ as w^2, and using the fact that $\|\boldsymbol{v}\|, \|\boldsymbol{v'}\| \leq B$ and crudely upper bounding $\boldsymbol{v'}^\mathsf{T}(M^\mathsf{T}M)^{-1}\boldsymbol{v'}$ and $\boldsymbol{v'}^\mathsf{T}(A^\mathsf{T}A)^{-1}\boldsymbol{v'}$ by $\frac{1}{2}$ we get

$$\ln\left(\frac{\mathsf{PDF}_{\mathcal{W}_d^{-1}(A^\mathsf{T}A,\nu)}(X)}{\mathsf{PDF}_{\mathcal{W}_d^{-1}(A'^\mathsf{T}A',\nu)}(X)}\right) \leq \frac{1}{2}\cdot 2\cdot\frac{B^2}{w^2}\left(2\sqrt{2\nu\ln(4/\delta)}+2\ln(4/\delta)\right)+\frac{1}{2}\cdot\frac{B^4}{w^4}(4\nu+4\nu)$$

As we have $w^2 = \frac{B^2}{\epsilon(1-\frac{\epsilon}{2\ln(4/\delta)})}\left(2\sqrt{2\nu\ln(4/\delta)}+2\ln(4/\delta)\right)$ we get that

$$\ln\left(\frac{\mathsf{PDF}_{\mathcal{W}_d^{-1}(A^\mathsf{T}A,\nu)}(X)}{\mathsf{PDF}_{\mathcal{W}_d^{-1}(A'^\mathsf{T}A',\nu)}(X)}\right) \leq \epsilon - \frac{\epsilon^2}{2\ln(4/\delta)} + \frac{\epsilon^2\cdot 4\nu}{8\nu\ln(4/\delta)} \leq \epsilon$$

■

Appendix D. Utility Theorems

In this section we provide the utility guarantees of the additive Wishart noise algorithm and the "Analyze Gauss" algorithm. Throughout this section we assume our database $D \in \mathbb{R}^{n\times d}$ is in fact composed of $D = [X; \boldsymbol{y}]$ where $X \in \mathbb{R}^{n\times p}$ and $\boldsymbol{y} \in \mathbb{R}^n$ (so we denote $p = d-1$). Clearly, to assume $\boldsymbol{y}$ is the last column of D simplifies the notation, but $\boldsymbol{y}$ can be any single column of D and X can be any subset of the other columns of D.

In this section we will repeatedly use the Woodbury formula, which states that for any invertible $A \in \mathbb{R}^{p\times p}$ and $U \in \mathbb{R}^{p\times k}$ and $V \in \mathbb{R}^{k\times p}$ of corresponding dimension we have

$$(A+UV)^{-1} = A^{-1} - A^{-1}U\left(I_{k\times k}+VA^{-1}U\right)^{-1}VA^{-1}$$

which implies that for any $B \in \mathbb{R}^{p\times p}$ we have the binomial inverse formula:

$$(A+B)^{-1} = A^{-1} - A^{-1}(I_{p\times p} - BA^{-1})^{-1}BA^{-1} \tag{3}$$

Our goal is to compare the distance between our predictor to the predictor one gets without noise, i.e. to $\widehat{\boldsymbol{\beta}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{y}$. Since we release a matrix $\widetilde{D^\mathsf{T}D}$ that approximates $D^\mathsf{T}D$, we can decompose it into the $p\times p$ matrix $\widetilde{X^\mathsf{T}X}$ and the p-dimensional vector $\widetilde{X^\mathsf{T}\boldsymbol{y}}$ and compute $\widetilde{\boldsymbol{\beta}} = (\widetilde{X^\mathsf{T}X})^{-1}\widetilde{X^\mathsf{T}\boldsymbol{y}}$. We thus give bounds on

$$\left\|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right\| = \left\|(\widetilde{X^\mathsf{T}X})^{-1}\widetilde{X^\mathsf{T}\boldsymbol{y}} - (X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{y}\right\|$$

Our analysis presents utility analysis that depends on the input parameters. This is in contrast to previous works on DP ERM that give a uniform bound and obtain it via *regularization* of the problem. (This is natural, as for $X = 0_{n\times p}$ clearly $\widehat{\boldsymbol{\beta}}$ is ill-defined unless we regularize the problem.) We begin with the utility of the additive Wishart noise mechanism.

Theorem 16 *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 denote $\widehat{\boldsymbol{\beta}} = (X^\mathsf{T}X)^{-1}X\boldsymbol{y}$.*
Denote $\widetilde{X^\mathsf{T}X} = X^\mathsf{T}X\ +\ N$, $\widetilde{X^\mathsf{T}\boldsymbol{y}} = X^\mathsf{T}\boldsymbol{y} + \boldsymbol{n}$ and $\widetilde{\boldsymbol{\beta}} = \widetilde{X^\mathsf{T}X}^{-1}\widetilde{X^\mathsf{T}\boldsymbol{y}}$; and also denote $C \overset{\text{def}}{=} \frac{\sigma_{\min}(X^\mathsf{T}X)}{\sigma^2\left(\sqrt{k}+\sqrt{p}+\sqrt{2\ln(4/\nu)}\right)^2}$. Then $\left\|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right\| \leq \frac{1}{C+1}\|\widehat{\boldsymbol{\beta}}\| + \left(1-\frac{1}{C+1}\right)\cdot\frac{2\sigma^2}{\sigma_{\min}(X^\mathsf{T}X)}\sqrt{2kp\cdot\ln(4p/\nu)}$.

Proof Because $\sigma^2 I$ is a diagonal matrix, standard results on the Wishart distribution give that $N \sim \mathcal{W}_p(\sigma^2 I_{p\times p}, k)$. We therefore denote R as a $(k \times p)$-matrix of i.i.d samples from a normal Gaussian $\mathcal{N}(0,1)$, and have $N = \sigma^2 R^\top R$. The Woodbury formula gives that

$$(X^\top X + N)^{-1} = (X^\top X)^{-1} - \sigma^2 (X^\top X)^{-1} R^\top (I + \sigma^2 R (X^\top X)^{-1} R^\top)^{-1} R (X^\top X)^{-1}$$

Denoting $Q = \sigma R (X^\top X)^{-1/2}$ we get

$$= (X^\top X)^{-1} - (X^\top X)^{-1/2} \left[Q^\top (I + QQ^\top)^{-1} Q \right] (X^\top X)^{-1/2}$$

Now, if we denote $Q = U\Lambda V^\top$ where Q's singular values are $\lambda_1, \ldots, \lambda_d$, we get $Q^\top(I + QQ^\top)^{-1}Q = V \cdot \mathrm{diag}(\left(\frac{\lambda_i^2}{1+\lambda_i^2}\right)_i) \cdot V^\top = V \cdot \mathrm{diag}(\left(1 - \frac{1}{1+\lambda_i^2}\right)_i) \cdot V^\top$. Note that

$$Q^\top Q = \sigma^2 (X^\top X)^{-1/2} R^\top R (X^\top X)^{-1/2}$$

and so, due to Lemma 9 we have $\lambda_1^2 = \sigma_{\max}(Q^\top Q) \leq \frac{\sigma^2(\sqrt{k}+\sqrt{p}+\sqrt{2\ln(4/\nu)})^2}{\sigma_{\min}(X^\top X)} \leq C^{-1}$ w.p. $\geq 1 - \nu/2$. Therefore, w.p. $\geq 1-\nu/2$ we have $\sigma_{\max}(Q^\top(I+QQ^\top)^{-1}Q) \leq \frac{1}{C+1}$. And so we have that both (i) $(X^\top X)^{-1} - (X^\top X + N)^{-1} \preceq \frac{1}{C+1}(X^\top X)^{-1}$ and (ii) $(X^\top X + N)^{-1} \preceq \frac{C}{C+1}(X^\top X)^{-1}$.

Next we turn to bound $\|\boldsymbol{n}\|$. One easy bound, given Lemma 9, is to show that w.p. $\geq 1 - \nu/2$ it holds that

$$\|\boldsymbol{n}\| \leq \|W \boldsymbol{e}_d\| \leq \|W\| \cdot 1 \leq \sigma^2(\sqrt{k} + \sqrt{p} + \sqrt{2\ln(4/\nu)})^2$$

Alternatively we can derive the following bound. Each coordinate in $\boldsymbol{n}$ is the result of the dot-product between the j-th column of R, denoted $\boldsymbol{r}_j$ with the d-th column of R, denoted $\boldsymbol{r}_d$. Each coordinate in R is sampled i.i.d from $\mathcal{N}(0, \sigma^2)$. Next, we use the fact that for two independent Gaussians *with the same variance* $X, Y \sim \mathcal{N}(0, \sigma^2)$ it holds that $XY = \frac{(X+Y)^2}{2} - \frac{(X-Y)^2}{2}$ with $\frac{1}{2}(X+Y)$ and $\frac{1}{2}(X-Y)$ are two independent[13] Gaussians $\mathcal{N}\left(0, \frac{\sigma^2}{2}\right)$. And so $\boldsymbol{r}_j \cdot \boldsymbol{r}_d = Z_{j_1} - Z_{j_2}$ where $Z_{j_1}, Z_{j_2} \sim \frac{\sigma}{\sqrt{2}} \cdot \chi_k^2$. Tail bounds for the χ^2-distribution (see Claim 1) give that w.p. $\geq 1 - \nu/2$ it holds that each coordinate of $\boldsymbol{n}$ is bounded in absolute value by $\frac{\sigma^2}{2}(\sqrt{k} + \sqrt{2\ln(4p/\nu)})^2 - \frac{\sigma^2}{2}(\sqrt{k} - \sqrt{2\ln(4p/\nu)})^2 = 4\sqrt{2k\ln(4p/\nu)}$, which means $\|\boldsymbol{n}\| \leq 2\sigma^2\sqrt{2pk \cdot \ln(4p/\nu)}$.[14]

Combining both bounds, we have that w.p. $\geq 1 - \nu$ it holds that

$$\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}} = \left((X^\top X)^{-1} - (X^\top X + N)^{-1}\right) X^\top \boldsymbol{y} - (X^\top X + N)^{-1}\boldsymbol{n}$$

therefore

$$\begin{aligned} \|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\| &\leq \frac{1}{C+1}\|(X^\top X)^{-1}X^\top y\| + \frac{2C\sigma^2}{C+1}\|(X^\top X)^{-1}\|\sqrt{2kp \cdot \ln(4p/\nu)} \\ &= \frac{1}{C+1}\|\widehat{\boldsymbol{\beta}}\| + \frac{C}{C+1} \cdot \frac{2\sigma^2}{\sigma_{\min}(X^\top X)}\sqrt{2kp \cdot \ln(4p/\nu)} \end{aligned}$$

■

[13] This is where we need to use the fact that X and Y have the same variance. We have $\begin{pmatrix} X+Y \\ X-Y \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix}$ and so the variance of $\begin{pmatrix} X+Y \\ X-Y \end{pmatrix}$ is diagonal iff X and Y have the same variance.

[14] We conjecture that the true bound in $\log(p)$-factor smaller, i.e. $O(\sigma^2\sqrt{2kp \cdot \ln(4/\nu)})$.

Next, we discuss the utility of the "Analyze Gauss" mechanism of Dwork et al. (2014).

Theorem 17 *Fix $X \in \mathbb{R}^{n\times p}$ and $\boldsymbol{y} \in \mathbb{R}^n$ s.t. $X^\top X$ is invertible. Fix $\eta \in (0,1)$ and $\nu \in (0, 1/e)$. Denote $\widetilde{X^\top X} = X^\top X + N$ and $\widetilde{X^\top \boldsymbol{y}} = X^\top \boldsymbol{y} + \boldsymbol{n}$ where each entry of N and $\boldsymbol{n}$ is sampled i.i.d from $\mathcal{N}\left(0, \sigma^2\right)$. Denote also $\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}$ and $\widetilde{\boldsymbol{\beta}} = (\widetilde{X^\top X})^{-1}\widetilde{X^\top \boldsymbol{y}}$. Then, if there exists some constant $C \geq 1$ s.t. we have that $\sigma_{\min}(X^\top X) \geq \frac{2C}{\eta} \cdot \sigma\sqrt{p}\log(1/\nu)$, then w.p. $\geq 1 - \nu$ we have $\left\|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right\| \leq 2\eta\|\widehat{\boldsymbol{\beta}}\| + \frac{\eta}{C}$.*

We comment that this is not precisely the same as the behavior of the "Analyze Gauss" algorithm. The difference lies in the fact that Analyze Gauss outputs $X^\top X + M$ where M is a symmetric matrix whose entries along and above the main diagonal are sampled i.i.d from a suitable $\mathcal{N}\left(0, \sigma^2\right)$. However, one can denote $M = \frac{1}{\sqrt{2}}(N + N^\top)$ for a matrix N whose entries are i.i.d samples from $\mathcal{N}\left(0, \sigma^2\right)$, and so the same result, up to a factor of $\sqrt{2}$, holds for Analyze Gauss.

Proof Plugging in (3) we get

$$
\begin{aligned}
&(\widetilde{X^\top X})^{-1}\widetilde{X^\top \boldsymbol{y}} = \\
&\left(I_{p\times p} - (X^\top X)^{-1}(I_{p\times p} - N(X^\top X)^{-1})^{-1}N\right)(X^\top X)^{-1}X^\top \boldsymbol{y} \\
&+ \left(I_{p\times p} - (X^\top X)^{-1}(I_{p\times p} - N(X^\top X)^{-1})^{-1}N\right)(X^\top X)^{-1}\boldsymbol{n}
\end{aligned}
$$

Denoting $Z = (X^\top X)^{-1}(I_{p\times p} - N(X^\top X)^{-1})^{-1}N$, we bound on $\|Z\|$, $\|I - Z\|$ and $\|\boldsymbol{n}\|$ so that we can derive a bound on $\left\|(\widetilde{X^\top X})^{-1}\widetilde{X^\top \boldsymbol{y}} - (X^\top X)^{-1}X^\top \boldsymbol{y}\right\|$.

Standard bounds on a symmetric ensemble of Gaussians (used also in (Dwork et al., 2014)) give that $\|N\| \leq C \cdot \sigma\sqrt{p}\log(1/\nu)$ w.p. $\geq 1 - \frac{\nu}{2}$ for some suitable constant $C > 0$. Hence we have that $\|N\| \cdot \|(X^\top X)^{-1}\| \leq \eta$. Hence, all singular values of $N(X^\top X)^{-1}$ are upper bounded in absolute value by η, and so all singular values of $I - N(X^\top X)^{-1}$ lie in the range $[1-\eta, 1+\eta]$. This implies that $\|Z\| \leq \frac{\eta}{1-\eta}$ and $\|I - Z\| \leq 1 + \frac{\eta}{1-\eta} = \frac{1}{1-\eta}$. Next we note that $\|\boldsymbol{n}\|^2 \sim \sigma^2 \cdot \chi^2_p$, and so, w.p. $\geq 1 - \frac{\nu}{2}$ it holds that $\|\boldsymbol{n}\| \leq \sigma(\sqrt{p} + \sqrt{2\ln(2/\nu)})$.

Thus, we get

$$
\left\|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right\| \leq \frac{\eta}{1-\eta}\|\widehat{\boldsymbol{\beta}}\| + \frac{1}{1-\eta} \cdot \frac{\sqrt{\sigma^2 p} + \sqrt{2\sigma^2\ln(2/\nu)}}{\sigma_{\min}(X^\top X)} \leq \frac{\eta}{1-\eta}\|\widehat{\boldsymbol{\beta}}\| + \frac{\eta}{C}
$$

■

Corollary 18 *Denote $\rho = \frac{\sigma_{\min}(\widetilde{X^\top X})}{2\sigma\sqrt{p}\log(1/\nu)}$. Then, for the same constant C in Theorem 17, if $\rho \geq 2C$ we have*

$$
\left\|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right\| \leq \frac{2C}{\rho}\|\widetilde{\boldsymbol{\beta}}\| + \frac{1}{\rho}
$$

Proof The proof follows from Theorem 17, and the observation that we can flip the role of $X^\top X$ and $\widetilde{X^\top X}$ because the Gaussian distribution is symmetric. And so, we just use the notation $\rho = \frac{C}{\eta}$. ■

Appendix E. Experiments: Comparison with the "Analyze Gauss" Baseline

Experimental Comparison between Algorithms. In this section we compare between the following 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 (bright red).
2. Post processing of Analyze Gauss: if the output of Analyze Gauss is not positive definite, add $cI_{d\times d}$ to it with $c = \mathbf{E}[\|N\|]$ (dark red).[15]
3. The additive Wishart noise algorithm given by Algorithm 2 - with post processing. Namely, 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 (green).
4. The JL-based algorithm, Algorithm 4 (blue).
5. Algorithm 6, which, as we commented in the experiments of Section 5, is analogous to Algorithm 4 and seems to consistently do better than Algorithm 4 (light blue).

Experiments over synthetic data. First, we compare the algorithms a simple setting using only a *a single regression.* We pick $p = 20$ i.i.d. independent features sampled from a normal Gaussian, a pick some $\boldsymbol{\beta} \in_R [-1,1]^{p+1}$ (the last coordinate denotes the regression's intercept), and set $\boldsymbol{y}$ as the linear combination of the features and the intercept (the all-1 column) plus random noise sampled from $\mathcal{N}(0, 0.5)$. Hence our data had dimension $d = p + 2 = 22$. We vary n to take any of the values in $\{2^{14} = 4,096, 2^{15}, 2^{16}, \ldots, 2^{25} = 33,554,432\}$. We also vary ϵ to take any of the values $\{0.05, 0.1, 0.15, 0.2, 0.25, 0.5\}$, and fixed[16] $\delta = e^{-10}$, and use the l_2-bound of $B = \sqrt{2.5d}$. (As preprocessing, each datapoint whose length is $> B$ is shrunk to have length B.) For each estimator we experimented with, we run it $t = 15$ times, and report the mean and standard variation of the 15 experiments. In all experiments we measure the l_2-distance between the outputted estimator of each algorithm to the true $\boldsymbol{\beta}$ we used to generate the data. After all, the algorithms we give are aimed at learning the $\boldsymbol{\beta}$ that generated the given samples, and so they should return an estimator close to the true $\boldsymbol{\beta}$. We coded all experiments R and ran the experiments on standard laptop.

In this case the bottom line is fairly clear: once n is sufficiently large, the Analyze Gauss baseline outperforms all other algorithms. We plotted the relative distance of the private regressor to the non-private regressor and the results for the case of $\epsilon = 0.5$ are presented in Figure 2. (The results for different values of ϵ are similar.)

We also compare the performances of all algorithms where there are multiple regressions of interest (and so the data has multiple small singular values). Here is it far more complicated to declare "a clear winner." In this experimental setting the data A is composed of 30 features: the first $p = 20$ columns are independent of one another (sampled i.i.d from a normal Gaussian); the latter 10 columns are the result of some linear combination of the

[15] 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 (which is the equivalent in this case to the general post-processing technique of Williams and McSherry (2010)), or setting different values for c. This other techniques did not seem to do better than technique *2* above. In fact, their utility was just as bad as the standard Analyze Gauss algorithm (with no post-processing).

[16] We are aware that it is a good standard practice to set $\delta < \frac{1}{n}$ since otherwise, sampling from the data is (ϵ, δ)-differentially private. However, as we vary n drastically, we aim to keep all other parameters equal.

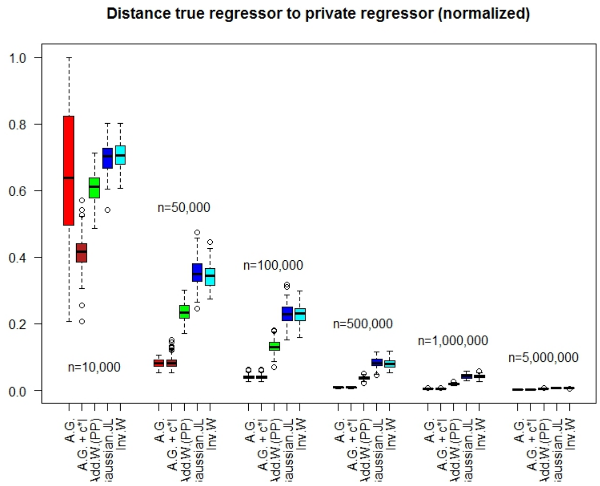


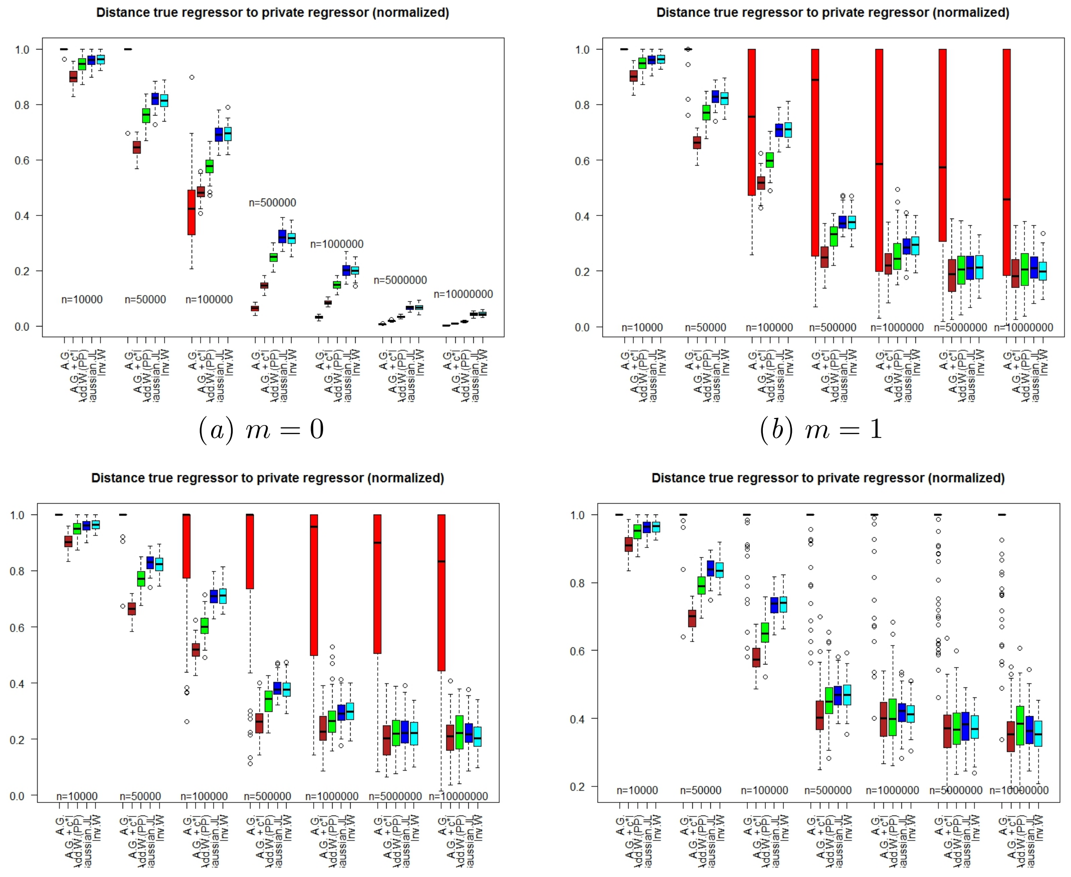
Figure 2: Single-Regression experiment on synthetic data ($\epsilon = 0.5$)

first p ones. And so $A = [X; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{10}]$ 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 $\mathcal{N}(0, 0.25)$. We fix the privacy-loss at $\epsilon = 0.5$, but we vary the number of $\boldsymbol{y}$-features used in our regression. Specifically, 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 p columns (plus an addition intercept column) plus some m additional $\boldsymbol{y}$-columns. (I.e.: $\{\boldsymbol{x}_1, \ldots \boldsymbol{x}_p\} \cup \{\boldsymbol{1}\} \cup \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$ where the latter m columns are disjoint of $\boldsymbol{y}_{i_0}$.) 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 (or irrelevant) for regression.

The results in this case are far less conclusive and are given in Figure 3. When $m = 0$, we are back to the case of a single regression (with no redundant features) and Analyze Gauss out-performs all other algorithms once n is large enough (same results as in Figure2). Yet, it is enough to set $m = 1$ to get very different results, where Analyze Gauss does fairly poorly. Still, post-processing Analyze Gauss still does fairly as well as the other three techniques, given in this paper.

Experiments over Real Data. *The Data:* We ran the 5 algorithms over diabetes data collected over ten years (1999-2008) taken from the UCI repository (Strack et al., 2014). We truncated the data to 9 attributes: sex (binary), age (in buckets of 10 years), time in hospital (numeric, in days), number lab procedures (numeric, 0-100), number procedures (numeric, 0-20), number medications (numeric, 0-100), and 3 different diagnoses (numeric, 0-1000), and a 10th column of all-1 (intercept). Omitting any entry with missing or non-numeric values on these nine attributes we were left with $N = 91842$ entries.

The experiments: We shuffled the entries randomly and used different size prefixes of the random dataset. We set $\epsilon = 0.1$ and $\delta = e^{-10}$. We also linearly converted each attribute independently to reside in the range $[-1, 1]$ to set our row-wise bound as $\sqrt{10}$, before running our algorithms (and re-converted each attribute to its original range after the execution of each algorithm). We tried to predict the 3rd diagnosis as a linear function of the other attributes, in three different settings: (i) using all 9 attributes; (ii) omitting the first two



$(a)\ m = 0$

$(b)\ m = 1$

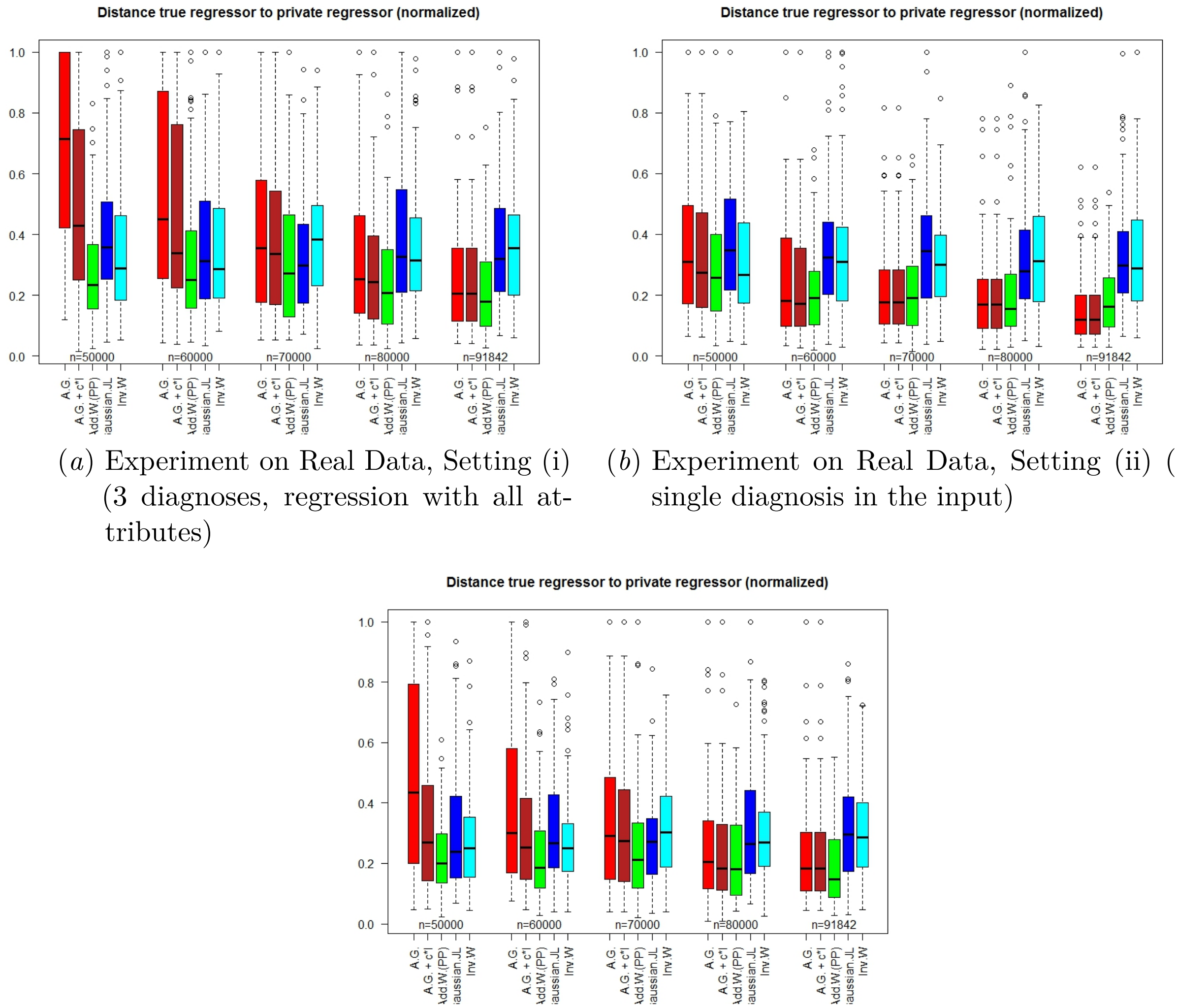
$(c)\ m = 2$

$(d)\ m = 5$

Figure 3: (best seen in color) The results of running our algorithm on synthetic data composed of multiple linear regressions. We use m to denote the number of additional columns that are linearly dependent of the first $p = 20$ columns.

diagnoses from the input and using only non-diagnoses attributes (after all, it is reasonable to conjecture one would want to estimate the value of the diagnosis based on other attributes); (iii) running the algorithm on the entire data, but omitting the first two diagnoses from *the output* so that the regressor must assign zero-value to the two other diagnoses. We believe setting (iii) captures the benefit of outputting the 2$^{\text{nd}}$-moment matrix rather than a private linear-regression algorithm: we can choose the features for the problem by ourselves and not be constraint by a curator's choice of features. Denoting $\boldsymbol{\beta}$ as the predictor with all 91842 entries and $\widetilde{\boldsymbol{\beta}}$ as the predictor returned by a differentially private algorithm, we measured the performance of the algorithm by $\max\{\frac{\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|}{\|\boldsymbol{\beta}\|}, 1\}$. We ran each algorithm 100 times.

Results: Results appear in Figure 4, where we contrast our experiments in settings (i), (ii) and (iii). Like before, comparing settings (i) and (ii) (Figures 4a and 4b resp.) we observe the same phenomena as in the synthetic data: if the data's feature are not correlated, Analyze Gauss produces the best results; whereas if there are correlations in the data, it under performs in comparison to the Additive Wishart noise algorithm. More strikingly is



(*a*) Experiment on Real Data, Setting (i) (3 diagnoses, regression with all attributes)

(*b*) Experiment on Real Data, Setting (ii) (a single diagnosis in the input)

(*c*) Experiment on Real Data, Setting (iii) (3 diagnoses, regression doesn't use diagnoses as features)

Figure 4: (best seen in color) Results for Our Experiment on Real Data

the comparison between settings (ii) and (iii) — in both setting we study the exact same regression problem, only in setting (iii) the algorithms also output correlations with two additional unused features. Indeed, in setting (ii) (Figure 4b) there's little difference between Analyze Gauss and the Additive Wishart algorithm.

In contrast, in setting (iii) (Figure 4c) the input matrix least singular value is smaller, and the additive Gaussian noise tends to output a non-PSD even for fairly large values of n. E.g., even for $n = 80000$ we have that Analyze Gauss has non-negligible probability to output a non-PSD, which means we have to post-process the output — hence the difference between standard Analyze Gauss (bright red) and post-processing Analyze Gauss (dark red). It is thus no surprise that in setting (iii) (and in setting (ii) as well), Additive Wishart outperforms the Analyze Gauss algorithm. Note that for various values of n both JL

Algorithm (blue) and the Inverse-Wishart Algorithm (light-blue) exhibit roughly the same performance. The reason is that their estimation the least singular value of the data remains fairly small for all different values of n. As a result, they run using roughly the same l_2-penalty term regardless of the data size. Also, these algorithms tend to do slightly worse than the additive noise algorithm because they are forced to spend some of the privacy budget on privately estimating the least singular value, thus only use a budget of 0.75ϵ for outputting the approximated 2^{nd}-moment matrix itself.