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

Or Sheffet

Center for Research on Computation and Society Harvard University Cambridge, MA osheffet@seas.harvard.edu

August 18, 2018

#### Abstract

We introduce three differentially-private algorithms that approximate the 2nd-moment matrix of the data. These algorithm, which in contrast to existing algorithms output positive-definite matrices, correspond to existing techniques in linear regression literature. Specifically, we discuss the following three techniques. (i) For Ridge Regression, we propose setting the regularization coefficient so that by approximating the solution using Johnson-Lindenstrauss transform we preserve privacy. (ii) We show that adding a small batch of random samples to our data preserves differential privacy. (iii) We show that sampling the 2nd-moment matrix from a Bayesian posterior inverse-Wishart distribution is differentially private provided the prior is set correctly. We also evaluate our techniques experimentally and compare them to the existing "Analyze Gauss" algorithm of Dwork et al [DTTZ14].

# 1 Introduction

Differentially private algorithms [DMNS06, DKM<sup>+</sup>06] are data analysis algorithms that give a strong guarantee of privacy, roughly stated as: by adding to or removing from the data a single datapoint we do not significantly change the probability of any outcome of the algorithm. The focus of this paper is on differentially private approximations of the 2nd-moment matrix of the data — given a dataset  $D \in \mathbb{R}^{n \times d}$ , its 2nd-moment matrix (also referred to as the Gram matrix of data or the scatter matrix if the mean of D is **0**) is the matrix  $D^{\mathsf{T}}D$  — and the uses of such approximations in linear regression. Indeed, since the 2nd-moment matrix of the data plays a major role in many data-analysis techniques, we already have differentially private algorithms that approximate the 2nd-moment matrix [DTTZ14] for the purpose of approximating the PCA, techniques for approximating the rank-k PCA of the data directly [HR12, Har13, KT13], or differentially private algorithms for linear regressions [CMS11, KST12, TS13, BST14].

However, existing techniques for differentially private linear regression suffer from the drawback that they approximate a single regression. That is, they assume that each datapoint is composed of a vector of features x and a label y and find the best linear combination of the features that predicts y. Yet, given a dataset D with d attributes we are free to pick any single attribute as a label, and any subset of the remaining attributes as features. 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.<sup>1</sup>

In contrast, the differentially private techniques that approximate the 2nd-moment matrix of the data, such as the Analyze Gauss paper of Dwork et al [DTTZ14], allow us to run as many regressions on the data

<sup>&</sup>lt;sup>1</sup>Indeed, Ullman [Ull15] have devised a solution to this problem, but this solution works in the more-cumbersome online model and requires exponential running-time for the curator; whereas our techniques follow the more efficient offline approach.

as we want. Yet, to the best of our knowledge, they have never been analyzed for the purpose of linear regression. Furthermore, the Analyze Gauss algorithm suffers from the drawback that it does not necessarily output a positive-definite matrix. This, as discussed in [XKI11] and as we show in our experiments, can be very detrimental — even if we do project the output back onto the set of positive definite matrices. And though the focus of this work is on linear regression, one can postulate additional reasons why releasing a positive definite matrix is of importance, such as using the output as a kernel matrix or doing statistical inference on top of the linear regression.

**Our Contribution.** In this work, we give three differentially private techniques for approximating the 2nd-moment matrix of the data that output a positive-definite matrix. We analyze their utility, both theoretically and empirically, and more importantly — show how they correspond to *existing techniques in linear regression*. And so we contribute to an increasing line of works [BBDS12, VZ15, WFS15] that shows that differential privacy may rise from existing techniques, provided parameters are set properly. We also compare our algorithms to the existing Analyze Gauss technique.

(Some notation before we introduce our techniques. We assume the data is a matrix  $A \in \mathbb{R}^{n \times d}$  with n sample points in d dimensions. For the ease of exposition, we focus on a single regression problem, given by A = [X; y] — i.e., the label is the d-th column and the features are the remaining p = d - 1 columns. We use  $\sigma_{\min}(A)$  to denote the least singular value of A.)

1. The Johnson-Lindenstrauss Transform and Ridge Regression. Blocki et al [BBDS12] have shown that projecting the data using a Gaussian Johnson-Lindenstrauss transform preserves privacy if  $\sigma_{\min}(A)$ is sufficiently large and it has been applied for linear regression [Upa14]. 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 [Sar06] 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 [Tik63, HK70], 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 [HTF09] to determining the penalty coefficient  $w^2$ , approaches that are based on the data itself and on minimizing risk. Here we give a fundamentally different approach — set w as to preserve  $(\epsilon, \delta)$ -differential privacy. Details, utility analysis and experiments regarding this approach appear in Section 3.

2. Additive Wishart noise. Whereas the Analyze Gauss algorithm adds Gaussian noise to  $A^{\mathsf{T}}A$ , here we show that we can sample a positive definite matrix W from a suitably chosen Wishart distribution  $\mathcal{W}_d(V, k)$ , and output  $A^{\mathsf{T}}A + W$ . This in turn corresponds to appending A with k i.i.d samples from a multivariate Gaussian  $\mathcal{N}(\mathbf{0}_d, V)$ . One is able to view this too as an extension of Ridge regression, where instead of appending A with d fixed examples, we append A with  $k \approx d + O(1/\epsilon^2)$  random examples.<sup>2</sup> Note, as opposed to Analyze Gauss [DTTZ14], where the noise has 0-mean, here the expected value of the noise is kV. This yields a useful way of post-processing the output:  $A^{\mathsf{T}}A + W - kV$ . Details, theorems and experiments with additive Wishart noise appear in Section 4.

3. Sampling from an inverse-Wishart distribution. The Bayesian approach for estimating the 2ndmoment matrix of the data assumes that the *n* sample points are sampled i.i.d from some  $\mathcal{N}(\mathbf{0}_d, V)$  for some unknown *V*, where we have a prior distribution on *V*. Each sample point causes us to update our belief on *V* which results in a posterior distribution on *V*. Though often one just outputs the MAP of the posterior belief (the mean of the posterior distribution), it is also common to output a sample drawn randomly from the posterior distribution. We show that if one uses the inverse-Wishart distribution as a prior (which is common, as the inverse-Wishart distribution is a conjugate prior), then sampling from

 $<sup>^{2}</sup>$ 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.

the posterior is  $(\epsilon, \delta)$ -diffrentially private, provided the prior is spread enough. This gives rise to our third approach of approximating  $A^{\mathsf{T}}A$  — sampling from a suitable inverse Wishart distribution. We comment that the idea that existing techniques in Bayesian analysis, and specifically sampling from the posterior distribution, are differentially-private on their own was originally introduced in the beautiful and elegant work of Vadhan and Zheng [VZ15]. But whereas their work focuses on estimating the mean of the sample, we focus on estimating the variances/2nd-moment. Details, theorems and experiments on sampling from the inverse-Wishart distribution appear in Section 5.

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

**Our proof technique.** Before continuing to preliminaries and the formal details of our algorithms, we give an overview of the proof technique. (All of the proofs are deferred to Appendix B.) To prove that each algorithm preserves  $(\epsilon, \delta)$ -differential privacy we state and prove 3 corresponding theorems, whose proofs follow the same high-level approach. As mentioned above, one theorem improves on a theorem of Blocki et al [BBDS12], who were the first to show that the JL-transform is differentially private. Blocki et al observed that by projecting the data using a  $(r \times n)$ -matrix of i.i.d normal Gaussians, we effectively repeat the same one-dimensional projection r independent times. So they proved that each one-dimensional projection is  $(\epsilon, \delta)$ -differentially private, and to show the entire projection preserves privacy they used the off-the-shelf composition of Dwork et al [DRV10], getting a bound that depends on  $O(\sqrt{r} \log(r))$ . In order to derive a bound depending only on  $O(\sqrt{r})$ , we do not use the composition theorem of [DRV10] but rather study the specific r-fold composition of the projection. As a result, we cannot follow the approach of Blocki et al.

To show that a one-dimensional projection is  $(\epsilon, \delta)$ -differentially private, Blocki et al compared the PDFs of two multivariate Gaussians. The PDF of a multivariate Gaussian is given by the multiplication of two terms: the first depends on the determinant of the variance, and the second depends on some exponent (see exact definition in Section 2). Blocki et al compared the ratio of each of the terms and showed that w.h.p each term's ratio is bounded by  $e^{\epsilon/2}$ . Unfortunately, following the same approach of Blocki et al yields a bound of  $e^{r\epsilon/2}$  for each of the terms and an overall bound that depends on O(r). Instead, we observe that the contributions of the determinant term and the exponent term to the ratio of the PDFs are of opposite signs. So we use the Matrix Determinant Lemma and the Sherman-Morrison Lemma (see Theorem A.4) to combine both terms into a single exponent term, and bound its size using the Johnson-Lindenstrauss transform (or rather, tight bounds on the  $\chi^2$ -distribution). The main lemma we use in our analysis is detailed in Lemma A.1. This lemma, in addition to giving tight bounds for the Gaussian JL-transform (mimicking the approach of Dasgupta and Gupta [DG03]), also gives a result that might be of independent interest. The standard JL lemma shows that for a  $(r \times d)$ -matrix R of i.i.d normal Gaussians and any fixed vector  $\mathbf{v}$  it holds w.h.p that  $\mathbf{v}^{\mathsf{T}}\mathbf{v} \in (1 \pm \eta)\mathbf{v}^{\mathsf{T}}(\frac{1}{r}R^{\mathsf{T}}R)\mathbf{v}$  provided  $r = O(\eta^{-2})$ . In Lemma A.1 we also show that for any fixed  $\mathbf{v}$  we have w.h.p. that  $\mathbf{v}^{\mathsf{T}}\mathbf{v} \in (1 \pm \eta)\mathbf{v}^{\mathsf{T}}(\frac{1}{r}R^{\mathsf{T}}R)^{-1}\mathbf{v}$  provided  $r = d + O(\eta^{-2})$ .

# 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  $\mathbf{0}_{l \times m}$ . The *l*-dimensional identity matrix is denoted  $I_{l \times l}$ . For two

<sup>&</sup>lt;sup>3</sup>To the best of our knowledge, for a general JLT, this is known to hold only when  $r = O(d \cdot \eta^{-2})$  and the transform preserves the lengths of all vectors in the  $\mathbb{R}^d$  space, see [Sar06] Corollary 11.

matrices M, N with the same number of row we use [M; N] to denote the concatenation of M and N. We use  $\epsilon, \delta$  to denote the privacy parameters. 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  $\mu$  and variance  $\sigma^2$ . Standard concentration bounds on Gaussians give that  $\Pr[x > \mu + \sigma \sqrt{\ln(1/\nu)}] < \nu$ . A multivariate Gaussian  $\mathcal{N}(\mu, \Sigma)$  for some positive semi-definite  $\Sigma$  denotes the multivariate Gaussian distribution where the mean of the *j*-th coordinate is the  $\mu_j$  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  $\mathsf{PDF}(x) = \left((2\pi)^{rank(\Sigma)} \cdot \tilde{\det}(\Sigma)\right)^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^{\mathsf{T}} \Sigma^{\dagger}(x-\mu)\right)$  and  $\tilde{\det}(\Sigma)$  is the multiplication of all non-zero singular values of  $\Sigma$ . We will repeatedly use the rules regarding linear operations on Gaussians. That is, for any scalar *c*, it holds that  $c\mathcal{N}(\mu, \sigma^2) = \mathcal{N}(c \cdot \mu, c^2\sigma^2)$ . For any matrix *C* it holds that  $C \cdot \mathcal{N}(\mu, \Sigma) = \mathcal{N}(C\mu, C\Sigma C^{\mathsf{T}})$ .

The  $\chi_k^2$ -distribution, where k is referred to as the degrees of freedom of the distribution, is the distribution over the  $l_2$ -norm of the sum of k independent normal Gaussians. That is, given  $X_1, \ldots, X_k \sim \mathcal{N}(0, 1)$ it holds that  $\zeta \stackrel{\text{def}}{=} (X_1, X_2, \ldots, X_k) \sim \mathcal{N}(\mathbf{0}_k, I_{k \times k})$ , and  $\|\zeta\|^2 \sim \chi_k^2$ . Standard tail bounds on the  $\chi^2$ distribution give that for any  $\nu \in (0, \frac{1}{2})$  we have  $\Pr_{\mathbf{x} \sim \chi_k^2} [x \in (\sqrt{k} \pm \sqrt{2\ln(2/\nu)})^2] \ge 1 - \nu$ . (We present them in Section A for completeness.) The Wishart-distribution  $\mathcal{W}_d(V, m)$  is the multivariate extension of the  $\chi^2$ -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  $\mathsf{PDF}_{\mathcal{W}_d(V,m)}(X) \propto \det(V)^{-\frac{m}{2}} \det(X)^{\frac{m-d-1}{2}} \exp(-\frac{1}{2}\operatorname{tr}(V^{-1}X))$ . 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  $\mathsf{PDF}_{\mathcal{W}_d^{-1}(V,m)}(X) \propto \det(V)^{\frac{m}{2}} \det(X)^{-\frac{m+d+1}{2}} \exp(-\frac{1}{2}\operatorname{tr}(VX^{-1}))$ .

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

**Definition 2.1** ([DMNS06, DKM<sup>+</sup>06]). An algorithm ALG which maps  $(n \times d)$ -matrices into some range  $\mathcal{R}$  is  $(\epsilon, \delta)$ -differential privacy if for all pairs of neighboring inputs A and A' and all subsets  $\mathcal{S} \subset \mathcal{R}$  it holds that  $\mathbf{Pr}[\mathsf{ALG}(A) \in \mathcal{S}] \leq e^{\epsilon} \mathbf{Pr}[\mathsf{ALG}(A') \in \mathcal{S}] + \delta$ . When  $\delta = 0$  we say the algorithm is  $\epsilon$ -differentially private.

It was shown in [DMNS06] that for any f where  $||f(A) - f(A')||_1 \leq \Delta$  then the algorithm that adds Laplace noise  $Lap(\frac{\Delta}{\epsilon})$  to f(A) is  $\epsilon$ -differential privacy. It was shown in [DKM<sup>+</sup>06] that for any f where  $||f(A) - f(A')||_2 \leq \Delta$  then adding Laplace noise  $\mathcal{N}\left(0, \frac{2\Delta^2 \ln(2/\delta)}{\epsilon}\right)$  to f(A) is  $(\epsilon, \delta)$ -differential privacy. This is precisely the algorithm of Dwork et al in their "Analyze Gauss" paper [DTTZ14]. They observed that in our setting, for the function  $f(A) = A^{\mathsf{T}}A$  we have that  $||f(A) - f(A')||_F^2 = B^4$ . And so they add i.i.d Gaussian noise to each coordinate of  $A^{\mathsf{T}}A$  (forcing the noise to be symmetric, as  $A^{\mathsf{T}}A$  is symmetric). We therefore refer to this benchmark as the Analyze Gauss algorithm. In addition, it is known that the composition of two algorithms, each of which is  $(\epsilon, \delta)$ -differentially private, yields an algorithm which is  $(2\epsilon, 2\delta)$ -differentially private.

# 3 Ridge Regression — Set the Regularization Coefficient to Preserve Privacy

The standard problem of linear regression, finding  $\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} ||X\boldsymbol{\beta} - \boldsymbol{y}||^2$ , relies on the fact that X is of full-rank. This clearly isn't always the case, and  $X^T X$  may be singular or close to singular. To that end, as well as for the purpose of preventing over-fitting, regularization is introduced. One way to regularize the linear regression problem is to introduce a  $l_2$ -penalty term: finding  $\boldsymbol{\beta}^R = \arg \min_{\boldsymbol{\beta}} ||X\boldsymbol{\beta} - \boldsymbol{y}||^2 + w^2 ||\boldsymbol{\beta}||^2$ . This is known as the *Ridge regression* problem, introduce by [Tik63, HK70] in the 60s and 70s. Ridge regression has a closed form solution:  $\boldsymbol{\beta}^R = (X^T X + w^2 I_{p \times p}) X^T y$ . The problem of setting w has been well-studied [HTF09] where existing techniques are data-driven, often proposing to set w as to minimize the risk of  $\boldsymbol{\beta}^R$ . Here, we propose a fundamentally different approach to the problem of setting w: set it so that we can satisfy  $(\epsilon, \delta)$ -differential privacy (via the Johnson-Lindenstrauss transform).

Observe, the Ridge regression problem can be written as: minimize  $||X\beta - y||^2 + ||wI_{p\times p}\beta - \mathbf{0}_p||^2$ . So, denote X' are the  $((n + p) \times p)$ -matrix which we get by concatenating X and  $wI_{p\times p}$ , and denote y' as the concatenation of y with p zeros. Then  $\beta^R = \arg \min ||X'\beta - y'||^2$ . Since p = d - 1 and we denote A = [X; 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  $f(\beta) \stackrel{\text{def}}{=} \left\|A'\begin{pmatrix}\beta\\-1\end{pmatrix}\right\|^2 = ||X'\beta - y'||^2 + w^2$ . Hence  $\beta^R = \arg \min f(\beta)$ . Hence, an approximation of  $A'^T A'$  yields an approximation of the Ridge regression problem. One way to approximate  $A'^T A'$  is via the Johnson-Lindenstrauss transform, which is known to be differentially private if all the singular values of the given input are sufficiently large [BBDS12]. And that is precisely why we use A' — all the singular values of  $A'^T A'$  are greater by  $w^2$  than the singular values of  $A^T A$ , and in particular are always  $\geq w^2$ . Therefore, applying the JLT to A' gives an approximation of  $A'^T A'$ , and furthermore, due to the work of Sarlos [Sar06] the JLT also approximates the linear regression. The following theorem improves on the original theorem of Blocki et al [BBDS12].

**Theorem 3.1.** Fix  $\epsilon > 0$  and  $\delta \in (0, \frac{1}{e})$ . Fix B > 0. Fix a positive integer r and let w be such that  $w^2 = 4B^2\left(\sqrt{2r\ln(\frac{4}{\delta})} + \ln(\frac{4}{\delta})\right)/\epsilon$ . Let A be a  $(n \times d)$ -matrix with d < r and where each row of A has bounded  $L_2$ -norm of B. Given that  $\sigma_{\min}(A) \ge w$ , 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  $(\epsilon, \delta)$ -differentially private.

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

**Theorem 3.2.** [[Sar06], Theorem 12] Fix any  $\eta > 0$  and  $\nu \in (0, \frac{1}{2})$ . Apply Algorithm 1 with  $r = O(d\log(d)\ln(1/\nu)/\eta^2)$ . Then,  $w.p \ge 1 - \nu$  it holds that  $\|\boldsymbol{\beta}^R - \tilde{\boldsymbol{\beta}}^R\| \le \frac{\eta}{\sqrt{w^2 + \sigma_{\min}(A^{\mathsf{T}}A)}} f(\boldsymbol{\beta}^R)$ .

Existing results about the expected distance  $\mathbf{E}[\|\boldsymbol{\beta}^R - \hat{\boldsymbol{\beta}}\|^2]$  (see [DFKU13]) can be used together with Theorem 3.2 to give a bound on  $\|\boldsymbol{\tilde{\beta}}^R - \boldsymbol{\hat{\beta}}\|^2$ .

In addition to Algorithm 1, we can use part of the privacy budget to look at the least singular-value of  $A^{\mathsf{T}}A$ . If it happens to be the case that  $\sigma_{\min}(A^{\mathsf{T}}A)$  is large, then we can adjust w by decreasing it by the appropriate factor. In fact, one can completely invert the algorithm and, in case  $\sigma_{\min}(A^{\mathsf{T}}A)$  is really large, not only set the regularization coefficient to be any arbitrary non-negative number, but also determine r based on Thm 3.1. Details appear in Algorithm 2.

To measure the effect of regularization we ran the following experiment. (Since the same experimental setting is used in the following sections we describe it here lengthly, and refer to it in later sections.)

**Input**: A matrix  $A \in \mathbb{R}^{n \times d}$  and a bound B > 0 on the  $l_2$ -norm of any row in A. Privacy parameters:  $\epsilon, \delta > 0$ .

Parameter r indicating the number of rows in the resulting matrix.

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

Set A' as the concatenation of A with  $wI_{d\times d}$ . Sample a  $r \times (n+d)$ -matrix R whose entries are i.i.d samples from a normal Gaussian. **return**  $M = \frac{1}{r} (RA')^{\mathsf{T}} (RA')$  and the approximation  $\widetilde{\boldsymbol{\beta}}^R = \arg \min_{\beta_d = -1} \boldsymbol{\beta}^{\mathsf{T}} M \boldsymbol{\beta}$ .

Algorithm 1: Approximating Ridge Regression while Preserving Privacy

Algorithm 2: Approximating Regression (Ridge or standard) while Preserving Privacy.

### 3.1 The Basic Single-Regression Experiment — Setting

To compare between the various algorithms we introduce and to analyze their utility we ran experiments testing their performance over data generated from a multivariate Gaussian. The experiments all share the same common setting, but each experiment studied a different set of estimators. In this section we detail the common setting, and in the next one we details the specific estimators and results of each experiment separately.

We pick p = 20 i.i.d. features sampled from a normal Gaussian, and pick some  $\beta \in_R [-1, 1]^{p+1}$  (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 and the 21-dimensional vector  $\boldsymbol{\beta}$  has  $l_2$  of about 3. 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 vary  $\epsilon$  to take any of the values  $\{0.05, 0.1, 0.15, 0.2, 0.25, 0.5\}$ , and fix<sup>4</sup>  $\delta = e^{-9}$ , and use the  $l_2$ -bound of  $B = \sqrt{2.5d}$ . (As preprocessing, each datapoint whose length is > Bis 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

<sup>&</sup>lt;sup>4</sup>We are aware that it is a good standard practice to set  $\delta < \frac{1}{n}$  since otherwise, sampling from the data is  $(\epsilon, \delta)$ -differentially private. However, as we vary *n* drastically, we aim to keep all other parameters equal.

return an estimator close to the true  $\beta$ . We coded all experiments in R and ran the experiments on standard laptop.

### 3.2 Experiment on Ridge Regression — Measuring the Effect of Regularization

To measure the effect of regularization we ran the following experiment in the setting detailed in Section 3.1. For each choice of  $\epsilon$  and n we ran three predictors. The first one is based on Algorithm 2 with  $r_0 = 2d$ . The second one is Algorithm 1 where we fed it the parameter r that the first one used, just so all predictors will be comparable. The last one is the *non-private* version that projected the data itself, without appending it with the  $w \cdot I_{d \times d}$  matrix (again, using the same parameter r as the other two predictors). The results are given in Figure 1.



Figure 1: (best seen in color) A comparison of the average  $l_2$ -error of the JL-based estimators. Algorithm 2 in blue, Algorithm 1 in red, and the non-private version in black. The *x*-axis is the size of the data in log-scale.

The results are strikingly similar across all values of  $\epsilon$ . Initially the error of the predictors is very high (for the value  $\beta$  we used to generate the data,  $\|\beta\| \approx 2.786$ , so such levels on noise mean in fact zero utility). Furthermore, it takes a while until Algorithm 2 (in blue) outperforms the more naïve Algorithm 1 (in red). (In most experiments, it happens only once  $n \geq 2^{19}$  or  $n \geq 2^{20}$ .) This implies that the privacy-budget "wasted" on the private estimation of the least singular value of the data actually ends up reducing our utility but not by a large factor. Towards the largest value of n, Algorithm 2 actually does noticeably better than Algorithm 1 by a multiplicative factor of  $\approx 3.5$  to  $\approx 11$  (for  $n = 2^{25}$  when  $\epsilon = 0.1$  we have mean accuracy of 0.0192 vs. 0.0671; when  $\epsilon = 0.5$  we have mean accuracy of 0.0058 vs. 0.0639). In all experiments, the non-private estimator (in black) was clearly the best for all values of n.

# 4 Additive Wishart Noise — Regression with Additional Random Examples

As discussed in the previous section, Ridge regression can be viewed as regression where in addition to the sample points given by [X; y] we see d additional datapoints given by  $wI_{d\times d}$ . Our second techniques follows this approach, only, instead of introducing these d fixed datapoints, we introduce a few more than d datapoints which are *random* and independent of the data.<sup>5</sup> Formally, we give the details in Algorithm 3 and immediately following — the theorem proving it is  $(\epsilon, \delta)$ -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  $v_1, v_2, \ldots, v_k$  i.i.d examples from  $\mathcal{N}(\mathbf{0}_d, B^2 I_{d \times d})$ . return  $M = A^{\mathsf{T}}A + \sum_{i=1}^{k} v_i v_i^{\mathsf{T}}$  and the approximation  $\widetilde{\boldsymbol{\beta}} = \arg \min_{\beta_d = -1} \boldsymbol{\beta}^{\mathsf{T}} M \boldsymbol{\beta}$ . Algorithm 3: Additive Wishart Noise Algorithm

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

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

We give a bound on the utility of the estimator we get with this technique in Theorem C.3. However, we are more interested in the utility of this approach after we *remove* some of the noise we add in this technique. Note,  $\mathbf{E}[N] = kB^2 \cdot I_{d \times d}$ , and so it stands to reason that we output  $A^TA + N - kB^2 \cdot I_{d \times d}$ . Now, when  $\sigma_{\min}(A^TA)$  is small, we run the risk that some of the eigenvalues of  $A^TA + N$  are smaller then  $kB^2$ , causing some of the eigenvalue of  $A^TA + N - kB^2 \cdot I_{d \times d}$  to be negative (which means we no longer output a PSD). In such a case, Lemma A.3 assures us that w.h.p we *can* decrease  $A^TA + N$  by  $B^2\left(\sqrt{k} - (\sqrt{d} + \sqrt{2\ln(4/\delta)})\right)^2 \cdot I_{d \times d}$  and maintain the property that the output is positive definite matrix. This is the algorithm we set to evaluate empirically.

#### 4.1 Experiment on Additive Wishart Noise

To evaluate the utility of the additive random Wishart noise algorithm we implemented and ran the algorithm in the same setting as detailed in Section 3.1. For each choice of  $\epsilon$  and n we ran three predictors. The first one is the naïve and non-private linear regression, that uses the data with no additive noise (i.e.,  $\hat{\beta}$ ). The second one is given by Algorithm 3. The last one is the estimator we get using the output of Algorithm 3 minus either  $kB^2 \cdot I_{d \times d}$  or  $B^2 \left(\sqrt{k} - (\sqrt{d} + \sqrt{2\ln(4/\delta)})\right)^2 \cdot I_{d \times d}$  (whichever of the two we can use and maintain positive definiteness). We repeat each experiment t = 15 times, measuring the  $l_2$ -distance between the outputted estimator of each algorithm to the true  $\beta$  we used to generate the data. (This yields randomness

 $<sup>^{5}</sup>$ Independent of the data itself, but dependent of its properties. Our noise *does* depend on the  $l_{2}$ -bound *B*.

in  $\|\widehat{\beta} - \beta\|$ , since every time we re-sample the data.) We report the mean and standard variation of the 15 experiments. The results are given in Figure 2. The results are again consistent across the board — reducing the noise also reduces the error, and indeed the second estimator is consistently doing better than the naïve estimator.



Figure 2: (best seen in color) A comparison of the average  $l_2$ -error of the Wishart additive noise estimators. Algorithm 3 in blue, deducting the expected shift from the output of Algorithm 3 and then running the regression is in red, and the non-private version in black. The *x*-axis is the size of the data in log-scale.

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

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

One of the most common priors used for positive definite matrices is the inverse-Wishart distribution. This is mainly due to the fact that the inverse-Wishart distribution is conjugate prior.<sup>6</sup> Specifically, if

 $<sup>^{6}</sup>$ A family of distributions is called conjugate prior if the prior distribution and the posterior distribution both belong to this family.

our prior belief is that  $V \sim W_d^{-1}(\Psi, k)$ , then after viewing *n* examples in the dataset *A* our posterior is  $V \sim W_d^{-1}((A^{\mathsf{T}}A + \Psi), n + k)$ . Here we show that sampling such a positive definite matrix *V* from our posterior inverse-Wishart distribution is  $(\epsilon, \delta)$ -differentially private, provided the prior distribution's scale matrix,  $\Psi$ , has a sufficiently large  $\sigma_{\min}(\Psi)$ . This result is in line with the recent beautiful work of Vadhan and Zheng [VZ15], who showed that many Bayesian techniques for estimating the means are differentially private, provided the prior is set correctly. The formal description of our algorithm and its privacy statement are given below.

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

Algorithm 4: Sampling from an Inverse-Wishart Distribution

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

We comment on the similarities between Theorem 5.1 and Theorem 3.1. Indeed, the Algorithm 1 essentially samples a matrix from  $\mathcal{W}(A^{\mathsf{T}}A + w^2I, k)$  for some choice of w and k (and then normalizes the sample by  $\frac{1}{k}$ ); and Algorithm 4 samples a matrix from  $\mathcal{W}^{-1}(A^{\mathsf{T}}A + w^2I, k)$  for a very similar choice of w. In fact, in Algorithm 1, instead of sampling R and then multiplying it with A, we can sample the same R and multiply it with  $(A^{\mathsf{T}}A)^{1/2}$ ; or even sample a  $(r \times d)$ -matrix  $\tilde{R}$  where each of its rows is sampled i.i.d from  $\mathcal{N}(\mathbf{0}_d, A^{\mathsf{T}}A)$ . (All of those have the same distribution over the output.) 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  $\psi$  accordingly. Details appear in Algorithm 5.

**Algorithm 5:** Sampling from an Inverse-Wishart Distribution whose degrees of freedom are determined by the input.

### 5.1 Experiments on Sampling from the Inverse Wishart Distribution

To estimate the utility of Algorithms 4 and 5, we conduct similar experiments to before, in the same setting detailed in Section 3.1. For each choice of  $\epsilon$  and n we ran 5 predictors. (i) The first one (in black) is the naïve and non-private Bayesian posterior sampling from the inverse Wishart distrbituion. (ii) The second one is given by Algorithm 4 (in blue). (iii) The third one is given by Algorithm 5 (in red) where the min-degreesof-freedom parameter is set to n + d (so that we have a direct way to compare between Algorithm 4 and 5). (iv) The fourth is given by Algorithm 2 where the min-number-of-rows parameter is set 2d (in green), and (v) the fifth one is Algorithm 5 when the min-degrees-of-freedom parameter is set 2d. This gives us a direct comparison between Algorithms 2 and 5. We repeat each experiment t = 15 times, measuring the  $l_2$ -distance between the outputted estimator of each algorithm to the true  $\beta$  we used to generate the data. We report the mean and standard variation of the 15 experiments in Figure 3. The results are again consistent among the various choices of  $\epsilon$ . Both Algorithm 4 and 5 (techniques (ii) and (iii)) exhibit fairly large errors throughout, mainly due to the fact that the parameter  $\psi$  used in each algorithm depends in n, as opposed to any other algorithm we present. We were surprised to see how little variance there exists in the results (the variance is too small to be visible in the figure). We did find it surprising that for the most part, the fact we split the privacy budget in Algorithm 5 turns out to be consistently costlier than Algorithm 4, even for very large values of n. Another result that we found interesting is that technique (v) outperforms the JL-technique (iv) (and it is holds for all values of n). Initially we conjectured that the gap can be explained by Lemma A.1, where the bound for the inverse-JL has a slightly better second order term than the bound for the standard JL. However, for some values of n the gap is fairly noticeable, and we leave it as an open problem to see if this holds for any projection matrix (and not just JL).

# 6 Comparison to the Analyze Gauss Baseline

In this paper we discuss multiple ways for outputting a differentially private approximation of  $A^{\mathsf{T}}A$ . One such way was already given by Dwork et al in their "Analyze Gauss" paper [DTTZ14]. As mentioned already, Dwork et al simply add to  $A^{\mathsf{T}}A$  a symmetric matrix N whose entries are sampled i.i.d from a suitable Gaussian. Furthermore, the magnitude of the noise introduced by the Analyze Gauss algorithm is the smallest out of all algorithms. Yet, as we stressed before, the output of Analyze Gauss isn't necessarily a positive definite matrix. In this work we investigate the effect of these fact on the problem of linear regression. We study the utility of the Analyze Gauss algorithm for the linear regression problem both theoretically (the theorem regarding the utility of Analyze Gauss is deferred to Appendix C) and experimentally, in comparison to the other algorithms we introduce in this work. The high-level message from the experiments we show here as follows. In the simple case, Analyze Gauss is the best algorithm to use,,<sup>7</sup> and when it returns "unreasonable" answers — so do all other algorithms we use (details to follow). However, there do exist cases where it under performs in comparison to the additive Wishart noise algorithm (Algorithm 3) and the Wishart (Algorithm 2) or inverse-Wishart (Algorithm 5) sampling algorithms.

In this section we compare between the following 6 techniques.

1. Analyze Gauss algorithm: output  $A^{\mathsf{T}}A + N$  with N a symmetric matrix whose entries are i.i.d samples from a Gaussian (black line, squares.)

2. The JL-based algorithm, Algorithm 2 (blue line, squares.)

3. The additive Wishart noise algorithm given by Algorithm 3 (magenta line, squares.)

4. A scaling version of Analyze Gauss: if the output of Analyze Gauss is not positive definite, add  $cI_{d\times d}$  to it with  $c = \mathbf{E}[||N||]$  (black line, circles.)

5. Algorithm 5, which, as we commented in the experiments of Section 5, is analogous to Algorithm 2 and seems to consistently do better than Algorithm 2. Both Algorithm 2 and 5 were given the same min-degrees-of-freedom parameter: 2d (blue line, circles.)

6. The scaling version of the additive Wishart random noise, as detailed in the experiment of Section 4.

 $<sup>^7\</sup>mathrm{In}$  our opinion, this result is of interest by itself.



Figure 3: (best seen in color) A comparison of the average  $l_2$ -error for the estimators based on inverse-Wishart distribution sampling. The non-private sampler is in black, Algorithm 4 is in blue and Algorithm 5 in red. The JL-based algorithm (Algorithm 2) that effectively samples from the Wishart distribution is in green; and its analogous algorithm that samples from the inverse-Wishart distribution (Algorithm 5) is in magenta. The *x*-axis is the size of the data in log-scale.

I.e., outputting  $A^{\mathsf{T}}A + W - k \cdot V$  (if this leaves the output positive definite) or  $A^{\mathsf{T}}A + W - (\sqrt{k} - (\sqrt{d} + \sqrt{2\ln(4/\delta)}))^2 \cdot V$  otherwise (magenta line, circles.)

**Post-processing the Analyze Gauss output.** We have experimented extensively with multiple ways to project the output of the Analyze Gauss algorithm onto the manifold of PSD matrices. Indeed, the most naïve approach is to find a PSD matrix M as to minimize  $||M - (A^TA + N)||_F^2$ . Such M effectively turns to be the result of zeroing out all negative eigenvalue of  $(A^TA + N)$ . The utility of this approach turns out to be just as bad as the standard Analyze Gauss algorithm (with no post-processing), returning estimations of size 12 or 9 when the true  $\beta$  has  $||\beta|| \approx 3$ . Other approached we have experimented with were to try other values of c for a post-processing of the form  $A^TA + N + cI_{d\times d}$ . (Such as setting c to be the upper- and lower-bound on the singular values of N w.p.  $\geq 1 - \delta$ .) The performance of such approaches was, overall, comparable to the chosen technique (setting  $c = \mathbf{E}[||N||]$ ) but with worse performance then our choice of c. Therefore, in our experiment, we used the *best* of all techniques *we were able to come up* with to post-process Analyze Gauss. This, however, does not mean that there isn't another post-process technique for Analyze Gauss that we didn't think of which out-performs our own approach.

#### 6.1 The Basic Single-Regression Experiment

In the same experiment setting from before (see Section 3.1) we compare our 6 estimators based on the  $l_2$ distance to the true  $\beta$  that generates our observations. The results, given in Figure 4 are pretty conclusive: Analyze Gauss is the better of all algorithms. Indeed, for smaller values of n its output is completely out of scale (while  $\|\beta\| \approx 3$ , the average error of Analyze Gauss is about 9, 12, and sometimes 30). In fact, the error of Analyze Gauss for small values of n is so large that we don't even present it in our graphs (and the standard deviation is so large, that the error bar of Analyze Gauss results in a big spike for such values of n). However, it is important to notice that for such values of n all other techniques also have a fairly large error (recall,  $\|\beta\|$  is roughly 3, so errors > 2.5 essentially give no information about  $\beta$ ). Once n reaches a certain size, then there is a sharp shift transition, and Analyze Gauss becomes the algorithm with the smallest error for all greater values of n. Eventually, the errors of all algorithms becomes smaller than the error between  $\hat{\beta}$ and  $\beta$  ( $\hat{\beta}$  is the non-private estimator of  $\beta$ ). We also comment that, like before, technique 5 (Algorithm 5) in consistently better than technique #2 (Algorithm 2), but also note that both technique have the largest variances in comparison to all other techniques.



Figure 4: (best seen in color) A comparison of the average  $l_2$ -error for our 6 estimators. Analyze Gauss (squares) and its scaled version (circles) are in black; JL algorithm (squares) and the JL variant that samples from the inverse Wishart distribution (circles) are in blue; and the additive Wishart noise (squares) and its scaled version (circles) are in magenta. The x-axis is the size of the data in log-scale.

### 6.2 The Multiple-Regressions Experiment

In this paper we argue that it is important to use algorithms that inherently output a positive definite matrix. To that end, we now investigate a more complex case, where the data is close to being singular, such that additive Gaussian noise is likely to introduce much error. The example we focus on is when the data A is composed of 2p features: the first p = 20 columns are independent of one another (sampled i.i.d from

a normal Gaussian); the latter p = 20 columns are the result of some linear combination of the first p ones. And so  $A = [X; \mathbf{y}_1, \ldots, \mathbf{y}_p]$  where for every i we have  $\mathbf{y}_i = X\beta_i + \mathbf{e}_i$  where each coordinate of  $\mathbf{e}_i$  is sampled i.i.d from  $\mathcal{N}(0, \sigma^2)$  for  $\sigma = 0.5$  (fixed for all i). In our experiments, we vary n (from  $2^{12}$  to  $2^{27}$  in powers of 2), but fix  $\epsilon = 0.1$ . What we also vary is the number of  $\mathbf{y}$ -features we use in our regression.

Recall, our algorithms approximate the Gram matrix of the data. Once such an approximation is published, it is possible to run as many linear regressions on it as we want — fixing any one column of the data as a label and any *subset* of the remaining columns as the features of the problem. This is precisely what we analyze here. We look at the linear regression problem where the label is some  $y_{i_0}$ , and the features of the problem are the first d columns plus some m additional y-columns.<sup>8</sup> (I.e.:  $\{x_1, \ldots, x_p\} \cup \{y_1, \ldots, y_m\}$  where the latter are disjoint to  $y_{i_0}$ .) A good approximation of  $\beta$  should therefore return some  $\tilde{\beta}$  which is 0 (or roughly 0) on the latter m coordinates. This corresponds to what we believe to be a high-level task a data-analyst might want to perform: finding out which features are relevant and which are irrelevant for regression.

The results in this case are far less conclusive and are given in Figure 5. When m = 0, we are back to the case of a single regression (with no redundant features), and here Analyze Gauss (black, squares) outperforms all other algorithms once n is large enough (in our case,  $n \ge 2^{16}$ ). Yet, it is enough to set m = 1to get very different results. When m > 0 it is evident that Analyze Gauss really performs badly — in fact, in most cases its values were far beyond the range of a reasonable approximation for  $\beta$  (taking values like 26 and 45 where  $\|\beta\| \approx 3.2$ ). The scaled version of Analyze Gauss (black, circles) does perform significantly better, yet — it is not the best out of all algorithms. In fact, it is consistently worse than the JL-based algorithms (blue, circles and squares) and from the scaled version of the additive Wishart noise (magenta, circles) for  $n < 2^{22} = 4,194,304$ . Note that as m increases, all algorithms' errors become fairly large. In addition, Figure 6 shows the variance of our estimators. It is clear that the scaled version of Analyze Gauss has the smallest variance.<sup>9</sup> However, the scaled additive Wishart noise algorithm (magenta, circle) seems to have a good variance as well, and, as discussed, does out-perform the scaled Analyze Gauss algorithm for a wide range of values of n.

**Discussion.** It is possible to interpret the results of this experiment, especially for the larger values of m, as a detriment for *all* the algorithms that approximate the Gram matrix of the data. Indeed, we pose the question of running regression over data where there does exist a large correlation between multiple columns as an open question. One approach could be to find a differentially private analogues to the techniques of [Mah11] for choosing a subset of the coordinates that approximate the k-PCA. An alternatively approach is to analyze the Lasso regression over the output of the algorithms that approximate the 2nd-moment matrix. In fact, we did experiment (though not extensively) with the Lasso regression. Using off-the-shelf Lasso regression packages (R package named glmnet), it seems that *all* algorithms give estimators that are indeed sparse, but *not* specifically over the latter m coordinates. Rather, the estimator is sparse both on the the first p coordinates and on the latter m coordinates. In contract, running the Lasso regression on the data without additional randomness (non-privately) gives sparsity over the latter m coordinates. We leave both problems for future work.

# References

[BBDS12] J. Blocki, A. Blum, A. Datta, and O. Sheffet. The Johnson-Lindenstrauss transform itself preserves differential privacy. In *FOCS*, 2012.

 $<sup>^8\</sup>mathrm{We}$  actually used one more column, of all 1s, representing the intercept.

<sup>&</sup>lt;sup>9</sup>There is a spike up for the largest value of  $n = 2^{27}$ , which is recurring throughout our experiments — most estimators have reasonable performance, but some have a really large error. One possible explanation could be related to the fact that we shrink sample points to have  $l_2$ -norm of at most B.



Figure 5: (best seen in color) A comparison of the average  $l_2$ -error for 6 estimators based. Analyze Gauss (squares) and its scaled version (circles) are in black; JL algorithm (squares) and the JL variant that samples from the inverse Wishart distribution (circles) are in blue; and the additive Wishart noise (squares) and its scaled version (circles) are in magenta. The x-axis is the size of the data in log-scale.

- [BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In 55th IEEE Annual Symposium on Foundations of Computer Science, FOCS, 2014.
- [CMS11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12, 2011.
- [DFKU13] 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.
- [DG03] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Struct. Algorithms*, 22(1), January 2003.
- [DKM<sup>+</sup>06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, 2006.
- [DMNS06] Cynthia Dwork, Frank Mcsherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.
- [DRV10] Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *FOCS*, 2010.
- [DS01] 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.
- [DTTZ14] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss optimal bounds for privacy preserving principal component analysis. In *STOC*, 2014.
- [Har13] Moritz Hardt. Robust subspace iteration and privacy-preserving spectral analysis. In 51st Annual Allerton Conference on Communication, Control, and Computing, 2013.



Figure 6: (best seen in color) A comparison of the standard error of the  $l_2$ -error for 6 estimators based, which are shown in Figure 5. Analyze Gauss (squares) and its scaled version (circles) are in black; JL algorithm (squares) and the JL variant that samples from the inverse Wishart distribution (circles) are in blue; and the additive Wishart noise (squares) and its scaled version (circles) are in magenta. The x-axis is the size of the data in log-scale.

- [HK70] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.
- [HR12] Moritz Hardt and Aaron Roth. Beating randomized response on incoherent matrices. In *STOC*, 2012.
- [HTF09] 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.
- [KST12] Daniel Kifer, Adam D. Smith, and Abhradeep Thakurta. Private convex optimization for empirical risk minimization with applications to high-dimensional regression. In *COLT*, 2012.
- [KT13] Michael Kapralov and Kunal Talwar. On differentially private low rank approximation. In SODA, 2013.
- [Mah11] Michael W. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, 3(2), February 2011.
- [MKB79] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate analysis*. Probability and mathematical statistics. Academic Press, 1979.
- [Sar06] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In FOCS, 2006.
- [Tao12] T. Tao. Topics in Random Matrix Theory. American Mathematical Soc., 2012.

- [Tik63] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. Soviet Math. Dokl., 4, 1963.
- [TS13] Abhradeep Thakurta and Adam Smith. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *COLT*, 2013.
- [Ull15] Jonathan Ullman. Private multiplicative weights beyond linear queries. In Proceedings of the 34th ACM Symposium on Principles of Database Systems, PODS, 2015.
- [Upa14] Jalaj Upadhyay. Differentially private linear algebra in the streaming model. *CoRR*, abs/1409.5414, 2014.
- [VZ15] 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.
- [WFS15] Yu-Xiang Wang, Stephen E. Fienberg, and Alexander J. Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *ICML*, 2015.
- [XKI11] Bowei Xi, Murat Kantarcioglu, and Ali Inan. Mixture of gaussian models and bayes error under differential privacy. In *CODASPY*. ACM, 2011.

# A 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 we make no claim as to the originality of the proofs of the lemmas. The proofs of Lemma A.1 and Claim A.2 are based in part on the result Dasgupta and Gupta [DG03] and in part about results regarding the Wishart distribution given in [MKB79] (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 B where we prove our privacy theorems.

**Lemma A.1.** 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

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

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

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

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

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

which implies

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

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  $\frac{\boldsymbol{v}^{\mathsf{T}}\boldsymbol{v}}{\boldsymbol{v}(X^{\mathsf{T}}X)^{-1}\boldsymbol{v}} \sim \chi_{r-d+1}^{2}$ . To see this, we argue that specifically for the vector  $\boldsymbol{e}_{d}$  (the indicator of the *d*-th coordinate) we have  $\frac{1}{\boldsymbol{e}_{d}(X^{\mathsf{T}}X)^{-1}\boldsymbol{e}_{d}} \sim \chi_{r-d+1}^{2}$ , 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  $XU^{\mathsf{T}}$ are identical.

Now, clearly  $e_d(X^{\mathsf{T}}X)^{-1}e_d = (X^{\mathsf{T}}X)^{-1}_{d,d}$ . Now, if we denote the last column of X as  $x_d$  and the first

 $d-1 \text{ columns of } X \text{ as } X_{-d} \text{ then } X^{\mathsf{T}}X = \begin{bmatrix} X_{-d}^{\mathsf{T}}X_{-d} & X_{-d}^{\mathsf{T}}x_{d} \\ \hline x_{d}^{\mathsf{T}}X_{-d} & \|x_{d}\|^{2} \end{bmatrix}.$  Thus, the formula for the entries

of the inverse give

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

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  $\frac{1}{(X^{\intercal}X)_{d,d}^{-1}} \sim \chi_{r-d+1}^2$ . Since this distribution is independent of  $X_{-d}$  we therefore have that this result holds w.p. 1. I.e.:

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

Therefore, with probability  $\geq 1 - \beta$  we have

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

 $\mathbf{SO}$ 

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

which implies

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

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

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

as this is the larger term of the two.

 $\begin{array}{l} \textbf{Claim A.2. 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 $\mathbf{Pr}[\sum_i X_i^2 < (\sqrt{k} - \sqrt{\Delta})^2] < e^{-\Delta/2}$.} \end{array}$ 

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

$$\mathbf{E}[e^{sX^2}] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} e^{sx^2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2(1-2s)}{2}} dx \qquad so \quad \frac{y=x\sqrt{1-2s}}{so \quad dy=dx\sqrt{1-2s}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \frac{dy}{\sqrt{1-2s}} = \frac{1}{\sqrt{1-2s}}$$

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

$$\begin{split} \mathbf{Pr}[\sum_{i} X_{i}^{2} > (\sqrt{k} + \sqrt{\Delta})^{2}] &\leq \exp\left(\lambda\sqrt{k}(\sqrt{k} + \sqrt{\Delta}) - \lambda(\sqrt{k} + \sqrt{\Delta})^{2}\right) \\ &= \exp\left(\frac{1}{2}\sqrt{k\Delta} - \frac{1}{2}\sqrt{\Delta}(\sqrt{k} + \sqrt{\Delta})\right) = \exp(-\frac{\Delta}{2}) \end{split}$$

A similar calculation shows the lower bound.

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

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

$$\begin{aligned} \mathbf{Pr}[\sum_{i} X_{i}^{2} > (\sqrt{k} + \sqrt{\Delta})^{2}] &\leq \exp\left(-\lambda\sqrt{k}(\sqrt{k} - \sqrt{\Delta}) + \lambda(\sqrt{k} - \sqrt{\Delta})^{2}\right) \\ &= \exp\left(-\frac{1}{2}\sqrt{k\Delta} + \frac{1}{2}\sqrt{\Delta}(\sqrt{k} - \sqrt{\Delta})\right) = \exp(-\frac{\Delta}{2}) \end{aligned}$$

**Lemma A.3.** 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

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

Furthermore, we also have that for any  $0 < \alpha \leq m$  it holds

$$\|\alpha V - X\| \le \|V\| \cdot |\alpha - (\sqrt{m} - \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right))^2| \quad and$$
$$\|(\alpha V)^{-1} - X^{-1}\| \le \sigma_{\min}^{-1}(V) \cdot |\alpha^{-1} - (\sqrt{m} + \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right))^{-2}|$$

*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  $X = V^{1/2}Y^{\mathsf{T}}YV^{1/2}$ .

Now, we invoke a theorem of Davidson and Szarek [DS01] (Theorem II.13) that states that for any t > 1 we have

$$\mathbf{Pr}[\sigma_{\max}(Y) > \sqrt{m} + \sqrt{d} + t] < e^{-t^2/2} \text{ and } \mathbf{Pr}[\sigma_{\min}(Y) < \sqrt{m} - \sqrt{d} - t] < e^{-t^2/2}$$

to 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

$$\boldsymbol{u}_{j}^{\mathsf{T}} X \boldsymbol{u}_{j} = (V^{1/2} \boldsymbol{u}_{j})^{\mathsf{T}} Y^{\mathsf{T}} Y (V^{1/2} \boldsymbol{u}_{j})$$

$$\leq (\sqrt{m} + \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right))^{2} \|V^{1/2} \boldsymbol{u}_{j}\|^{2} = \sigma_{j}(V) (\sqrt{m} + \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right))^{2}$$

$$\boldsymbol{u}_{j}^{\mathsf{T}} X \boldsymbol{u}_{j} \geq (\sqrt{m} - \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right))^{2} \|V^{1/2} \boldsymbol{u}_{j}\|^{2} = \sigma_{j}(V) (\sqrt{m} - \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right))^{2}$$

and furthermore, for any subspace S we have that

$$\max_{\boldsymbol{u}\in S: \, \|\boldsymbol{u}\|=1} \boldsymbol{u}^{\mathsf{T}} \boldsymbol{X} \boldsymbol{u} \leq (\sqrt{m} + \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right))^2 \left(\max_{\boldsymbol{u}\in S: \, \|\boldsymbol{u}\|=1} \|V^{1/2}\boldsymbol{u}_j\|^2\right)$$
$$\min_{\boldsymbol{u}\in S: \, \|\boldsymbol{u}\|=1} \boldsymbol{u}^{\mathsf{T}} \boldsymbol{X} \boldsymbol{u} \geq (\sqrt{m} - \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right))^2 \left(\min_{\boldsymbol{u}\in S: \, \|\boldsymbol{u}\|=1} \|V^{1/2}\boldsymbol{u}_j\|^2\right)$$

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

$$\sigma_j(X) = \max_{\{S \subset \mathbb{R}^d: \dim(S) = j\}} \min_{\{u \in S: \|u\|=1\}} u^\mathsf{T} X u = \min_{\{S \subset \mathbb{R}^d: \dim(S) = d-j+1\}} \max_{\{u \in S: \|u\|=1\}} u^\mathsf{T} X u$$

Therefore, we can pick  $S' = span\{u_1, \dots, u_j\}$  and  $S'' = span\{u_j, \dots, u_d\}$  to deduce

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

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, we have  $\alpha I - Y^{\mathsf{T}}Y = U \left(\alpha I - \Sigma\right) U^{\mathsf{T}}$ . Since all the entries on the diagonal lie in the range  $|\alpha - (\sqrt{m} \pm \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right))^2|$ . As  $\alpha \leq m$  we have that all eigenvalues are upper bounded by  $(m - \alpha) + 2\sqrt{m} \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right)$  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  $|\alpha^{-1} - (\sqrt{m} \pm \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right))^{-2}|$ , which in this case is upper bounded by  $|\alpha^{-1} - (\sqrt{m} + \left(\sqrt{d} + \sqrt{2\ln(\frac{2}{\delta})}\right))^{-2}|$ . 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 A.4.** Let A be a  $(d \times d)$ -invertible matrix and fix any two d-dimensional vectors u, v s.t.  $v^{\mathsf{T}} A^{-1} u \neq d$ -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  $A + uv^{\mathsf{T}} = A(I + A^{-1}uv^{\mathsf{T}})$ , we analyze the spectrum of the matrix  $I + A^{-1}uv^{\mathsf{T}}$ . 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  $\boldsymbol{z}^{\mathsf{T}}(I + A^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}})\boldsymbol{z} =$  $1 + \|\boldsymbol{v}\| \cdot \boldsymbol{z}^{\mathsf{T}} A^{-1} \boldsymbol{u} = 1 + \boldsymbol{v}^{\mathsf{T}} A^{-1} \boldsymbol{u}.$  Therefore,  $\det(A + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}) = \det(A) \det(I + A^{-1} \boldsymbol{u} \boldsymbol{v}^{\mathsf{T}}) = \det(A)(1 + \boldsymbol{v}^{\mathsf{T}} A^{-1} \boldsymbol{u}).$ As for the Sherman-Morrison formula, we can simply check and see that indeed:

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

#### В **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.

#### B.1Privacy Proof for Algorithm 1

**Theorem B.1.** Fix  $\epsilon > 0$  and  $\delta \in (0, \frac{1}{2})$ . Fix B > 0. Fix a positive integer r and let w be such that

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

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  $\sigma_{\min}(A) \geq w$ , 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  $(\epsilon, \delta)$ -differentially private.

**Corollary B.2.** assuming  $\epsilon < 1$  and  $\delta < e^{-1}$ , if it holds that  $r \geq 2\ln(\frac{4}{\delta})$  then it suffices to have  $w^2 \geq 2\ln(\frac{4}{\delta})$  $8B^2 \sqrt{r \ln(4/\delta)}$  for the results of Theorem B.1 to hold. Alternatively, given input where its least singular value is publicly known to w, we can set

$$r = \left\lceil \left( \frac{\epsilon w^2}{8B^2 \ln(\frac{4}{\delta})} \right)^2 \right\rceil, \quad if \ indeed \ r > 2 \ln(\frac{4}{\delta})$$

and satisfy  $(\epsilon, \delta)$ -differential privacy. Therefore, if the rows of A are i.i.d draws from a **0**-mean multivariate Gaussian with variance  $\Sigma$ , then we may set r as  $\left[\left(n\frac{\epsilon\sigma_{\min}(\Sigma)}{8B^2\ln(\frac{4}{\delta})}\right)^2\right] = \Omega(n^2).$ 

Proof. Fix A and A' be two neighboring  $(n \times d)$  matrices, s.t. A - A' is a rank-1 matrix of the form  $E \stackrel{\text{def}}{=} A - A' = e_i(\boldsymbol{v} - \boldsymbol{v}')^{\mathsf{T}}$ . 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  $R^{\mathsf{T}}$  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  $\mathcal{N}(\mathbf{0}_r, A^{\mathsf{T}}A)$ . 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

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

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  $\frac{1+x}{1+y} = (1+x)(1-\frac{y}{1+y}) \le \exp(x-\frac{y}{1-y})$  for any  $x, y \ne 1$  we have

$$\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(-\frac{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(-\frac{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}^{d} \exp\left(\frac{1}{2}\left(\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}'}\right) \\
+ \frac{1}{2}\left(-\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}^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{x}_{j}}{1 + \boldsymbol{v}'^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}}\right)\right) \\
= \exp\left(\frac{1}{2}\left(r \cdot \boldsymbol{v}'^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}' - \frac{\boldsymbol{v}'^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}}{1 + \boldsymbol{v}'^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'}\right)\right)\right) \\
\cdot \exp\left(\frac{1}{2}\left(-\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}\boldsymbol{v}'}{1 + \boldsymbol{v}^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}}\right)\right)\right)(2)$$

Denote

$$z_{1} \stackrel{\text{def}}{=} \boldsymbol{v}^{\prime \mathsf{T}} (M^{\mathsf{T}} M)^{-1} \boldsymbol{v}^{\prime} - \boldsymbol{v}^{\prime \mathsf{T}} (M^{\mathsf{T}} M)^{-1} \left( \frac{1}{r} \sum_{j=1}^{r} \boldsymbol{x}_{j} \boldsymbol{x}_{j}^{\mathsf{T}} \right) (M^{\mathsf{T}} M)^{-1} \boldsymbol{v}^{\prime}$$
$$z_{2} \stackrel{\text{def}}{=} \boldsymbol{v}^{\mathsf{T}} (M^{\mathsf{T}} M)^{-1} \boldsymbol{v} - \boldsymbol{v}^{\mathsf{T}} (M^{\mathsf{T}} M)^{-1} \left( \frac{1}{r} \sum_{j=1}^{r} \boldsymbol{x}_{j} \boldsymbol{x}_{j}^{\mathsf{T}} \right) (M^{\mathsf{T}} M)^{-1} \boldsymbol{v}$$

we have that

$$\ln\left(\frac{\mathsf{PDF}_{X}(\boldsymbol{x}_{1},...,\boldsymbol{x}_{r})}{\mathsf{PDF}_{X'}(\boldsymbol{x}_{1},...,\boldsymbol{x}_{r})}\right) \leq \frac{r}{2} \left(\frac{z_{1}}{1+\boldsymbol{v}'(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'} + \frac{-z_{2}}{1+\boldsymbol{v}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}} + \frac{(\boldsymbol{v}'(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}')^{2}}{1+\boldsymbol{v}'(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'}\right) \\ \leq \frac{r}{2} \left(|z_{1}| + |z_{2}| + (\boldsymbol{v}'(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}')^{2}\right)$$

We now turn to analyze each of the above three terms separately. The easiest to bound are the terms  $\boldsymbol{v}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}$  and  $\boldsymbol{v}'(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'$ . Weyl's inequality yields that  $\sigma_{\min}(M^{\mathsf{T}}M) \geq \sigma_{\min}(A^{\mathsf{T}}A) - B^2$ , 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  $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r$  were sampled from  $A^{\mathsf{T}}A$ . If they were sampled from  $A'^{\mathsf{T}}A'$  then the proof is analogous. Denote X as the matrix whose columns are  $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r$ . We have

$$z_{2} = ((M^{\mathsf{T}}M)^{-1}\boldsymbol{v})^{\mathsf{T}} (M^{\mathsf{T}}M - (\frac{1}{r}X^{\mathsf{T}}X)) (M^{\mathsf{T}}M)^{-1}\boldsymbol{v} = ((M^{\mathsf{T}}M)^{-1}\boldsymbol{v})^{\mathsf{T}} (A^{\mathsf{T}}A - \boldsymbol{v}\boldsymbol{v}^{\mathsf{T}} - (\frac{1}{r}X^{\mathsf{T}}X)) (M^{\mathsf{T}}M)^{-1}\boldsymbol{v} = ((M^{\mathsf{T}}M)^{-1}\boldsymbol{v})^{\mathsf{T}} (A^{\mathsf{T}}A)^{1/2} (I - (A^{\mathsf{T}}A)^{-1/2} (\frac{1}{r}X^{\mathsf{T}}X) (A^{\mathsf{T}}A)^{-1/2}) (A^{\mathsf{T}}A)^{1/2} (M^{\mathsf{T}}M)^{-1}\boldsymbol{v} - (\boldsymbol{v}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v})^{2}$$

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

$$\begin{aligned} |z_{2}| &\leq \left(2\sqrt{\frac{2\ln(4/\delta)}{r}} + \frac{2\ln(4/\delta)}{r}\right) \left\| (A^{\mathsf{T}}A)^{1/2} (M^{\mathsf{T}}M)^{-1} \boldsymbol{v} \right\|^{2} + \left(\boldsymbol{v}(M^{\mathsf{T}}M)^{-1} \boldsymbol{v}\right)^{2} \\ &\leq \left(2\sqrt{\frac{2\ln(4/\delta)}{r}} + \frac{2\ln(4/\delta)}{r}\right) \left(\boldsymbol{v}^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1} (M^{\mathsf{T}}M + \boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}) (M^{\mathsf{T}}M)^{-1} \boldsymbol{v}\right) + \left(\boldsymbol{v}(M^{\mathsf{T}}M)^{-1} \boldsymbol{v}\right)^{2} \\ &= \left(\boldsymbol{v}(M^{\mathsf{T}}M)^{-1} \boldsymbol{v}\right) \left(2\sqrt{\frac{2\ln(4/\delta)}{r}} + \frac{2\ln(4/\delta)}{r}\right) + \left(\boldsymbol{v}(M^{\mathsf{T}}M)^{-1} \boldsymbol{v}\right)^{2} \left(2\sqrt{\frac{2\ln(4/\delta)}{r}} + \frac{2\ln(4/\delta)}{r} + 1\right) \\ &\leq \left(\frac{w^{2}}{B^{2}} - 1\right)^{-1} \left(2\sqrt{\frac{2\ln(4/\delta)}{r}} + \frac{2\ln(4/\delta)}{r}\right) + \left(\frac{w^{2}}{B^{2}} - 1\right)^{-2} \left(2\sqrt{\frac{2\ln(4/\delta)}{r}} + \frac{2\ln(4/\delta)}{r} + 1\right) \end{aligned}$$

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

$$\ln \left( \frac{\mathsf{PDF}_{X}(\boldsymbol{x}_{1},...,\boldsymbol{x}_{r})}{\mathsf{PDF}_{X'}(\boldsymbol{x}_{1},...,\boldsymbol{x}_{r})} \right) \leq \frac{r}{2} \left( |z_{1}| + |z_{2}| + (\boldsymbol{v}'(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}')^{2} \right)$$

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

$$+ \left( \frac{w^{2}}{B^{2}} - 1 \right)^{-2} \left( 2\sqrt{2r\ln(4/\delta)} + 2\ln(4/\delta) + \frac{3r}{2} \right)$$

$$\leq \frac{\epsilon}{1 + \frac{\epsilon}{\ln(4/\delta)}} + \epsilon^{2} \left( \frac{2\sqrt{2r\ln(4/\delta)} + 2\ln(4/\delta)}{(2\sqrt{2r\ln(4/\delta)} + 2\ln(4/\delta))^{2}} + \frac{3r}{16r\ln(4/\delta)} \right)$$

$$\leq \frac{\epsilon}{1 + \frac{\epsilon}{\ln(4/\delta)}} \left( 1 + \frac{\epsilon}{\ln(4/\delta)} \left( \frac{1}{2} + \frac{3}{16} \right) \right) < \epsilon$$

by plugging in the value of  $w^2$ .

### **B.2** Privacy Proof for Algorithm 3

**Theorem B.3.** 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

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

(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  $\nu$ -degrees of freedom using the scale matrix V (i.e.,  $N \sim W_d(V, \nu)$ ) for any matrix V with least singular value  $\sigma_{\min}(V) \ge C_1$  (e.g.  $V = C_1 I_{d \times d}$ ) and  $\nu \ge \lfloor d + 2C_2 \ln(4/\delta) \rfloor$ . Then outputting  $X = A^T A + N$  is  $(\epsilon, \delta)$ -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  $\sigma_{\min}^2(M) \leq \sigma_{\min}^2(A) \leq \sigma_{\min}^2(M) + B^2$ . Same holds for the least singular value of M and A'. Recall that

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

We argue that Wishart-matrix additive noise is  $(\epsilon, \delta)$ -differentially 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  $\Pr[X - A^{\mathsf{T}}A$  non invertible] =  $\Pr_{N \sim W_d(V,\nu)}[N \text{ non invertible}] = 0$ . However, it is not a-priori clear why we should also have  $\Pr[X - A'^{\mathsf{T}}A']$  $A'^{\mathsf{T}}A'$  non invertible] = 0 or  $\Pr[X - M^{\mathsf{T}}M \text{ non invertible}] = 0$ .) 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

$$\begin{aligned} \frac{\mathsf{PDF}_{\mathcal{W}_{d}(V,\nu)}(X-A^{\mathsf{T}}A)}{\mathsf{PDF}_{W_{d}(V,\nu)}(X-A'^{\mathsf{T}}A')} &= \left(\frac{\det(X-A^{\mathsf{T}}A)}{\det(X-A'^{\mathsf{T}}A')}\right)^{\frac{\nu-d-1}{2}} e^{-\frac{1}{2}\left(\operatorname{tr}(V^{-1}(X-A^{\mathsf{T}}A)) - \operatorname{tr}(V^{-1}(X-A'^{\mathsf{T}}A'))\right)} \\ &= \left(\frac{\det(X-M^{\mathsf{T}}M-\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}})}{\det(X-M^{\mathsf{T}}M-\boldsymbol{v}'\boldsymbol{v}'^{\mathsf{T}})}\right)^{\frac{\nu-d-1}{2}} e^{-\frac{1}{2}\left(\operatorname{tr}(V^{-1}(X-A^{\mathsf{T}}A-X+A'^{\mathsf{T}}A')\right)} \\ &= \left(\frac{1-\boldsymbol{v}^{\mathsf{T}}(X-M^{\mathsf{T}}M)^{-1}\boldsymbol{v}}{1-\boldsymbol{v}'^{\mathsf{T}}(X-M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'}\right)^{\frac{\nu-d-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}(V^{-1}\boldsymbol{v}'\boldsymbol{v}'^{\mathsf{T}}) + \frac{1}{2}\operatorname{tr}(V^{-1}\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}})\right) \\ & \operatorname{tr}(AB) \stackrel{=}{=} \operatorname{tr}(BA)} \left(\frac{1-\boldsymbol{v}^{\mathsf{T}}(X-M^{\mathsf{T}}M)^{-1}\boldsymbol{v}}{1-\boldsymbol{v}'^{\mathsf{T}}(X-M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'}\right)^{\frac{\nu-d-1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{v}'^{\mathsf{T}}V^{-1}\boldsymbol{v}' + \frac{1}{2}\boldsymbol{v}^{\mathsf{T}}V^{-1}\boldsymbol{v}\right) \end{aligned}$$

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

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

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

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

And so we have:

$$\boldsymbol{v}^{\mathsf{T}} \left( V^{-1} - (\nu - d - 1)(X - M^{\mathsf{T}}M)^{-1} \right) \boldsymbol{v}$$
  
=  $\boldsymbol{v}^{\mathsf{T}} \left( V^{-1} - (\nu - d + 1)(X - M^{\mathsf{T}}M)^{-1} \right) \boldsymbol{v} + 2\boldsymbol{v}^{\mathsf{T}}(X - M^{\mathsf{T}}M)^{-1} \boldsymbol{v}$   
 $\leq \boldsymbol{v}^{\mathsf{T}} \left( V^{-1} - (\nu - d + 1)(X - A^{\mathsf{T}}A)^{-1} \right) \boldsymbol{v} + 2\boldsymbol{v}^{\mathsf{T}}(X - A^{\mathsf{T}}A)^{-1} \boldsymbol{v} + (\nu - d + 1)(\boldsymbol{v}^{\mathsf{T}}(X - A^{\mathsf{T}}A)^{-1} \boldsymbol{v})^{2}$ 

Now, note that  $(X - A^{\mathsf{T}}A) \sim \mathcal{W}_d(V, \nu)$ , and so  $V^{-1/2}(X - A^{\mathsf{T}}A)V^{-1/2} \sim \mathcal{W}_d(I_{d \times d}, \nu)$ . This allows us to invoke Lemma A.1 to

$$\boldsymbol{v}^{\mathsf{T}}\left(V^{-1} - \left(\frac{1}{\nu - d + 1}(X - A^{\mathsf{T}}A)\right)^{-1}\right)\boldsymbol{v} = (V^{-1/2}\boldsymbol{v})^{\mathsf{T}}\left(I - \left(\frac{V^{-1/2}(X - A^{\mathsf{T}}A)V^{-1/2}}{\nu - d + 1}\right)^{-1}\right)(V^{-1/2}\boldsymbol{v})$$

and infer that w.p.  $\geq 1-\delta/2$  we have the following bound

$$\begin{split} \mathbf{v}^{\mathsf{T}} \left( V^{-1} - (\nu - d - 1)(X - M^{\mathsf{T}}M)^{-1} \right) \mathbf{v} \\ &\leq \left( \frac{2\sqrt{2(\nu - d + 1)\ln(4/\delta)} - 2\ln(4/\delta)}{(\sqrt{\nu - d + 1} - \sqrt{2\ln(4/\delta)})^2} + \frac{2}{(\sqrt{\nu - d + 1} - \sqrt{2\ln(4/\delta)})^2} + \frac{2(\nu - d + 1)}{(\sqrt{\nu - d + 1} - \sqrt{2\ln(4/\delta)})^4} \right) \|V^{-1/2}\mathbf{v}\|^2 \\ &= \frac{\|V^{-1/2}\mathbf{v}\|^2}{(\sqrt{\nu - d + 1} - \sqrt{2\ln(4/\delta)})^2} \left( 2\sqrt{2(\nu - d + 1)\ln(4/\delta)} - 2\ln(4/\delta) + 2 + \frac{2}{(1 - \sqrt{\frac{2\ln(4/\delta)}{\nu - d - 1}})^2} \right) \\ &= \frac{2\sqrt{2(\nu - d + 1)\ln(4/\delta)} - 2\ln(4/\delta) + 6}{(\sqrt{\nu - d + 1} - \sqrt{2\ln(4/\delta)})^2} \|V^{-1/2}\mathbf{v}\|^2 \end{split}$$

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

$$\begin{split} \mathbf{v}'^{\mathsf{T}} & \left( \frac{\nu - d - 1}{1 - \mathbf{v}'^{\mathsf{T}} (X - M^{\mathsf{T}} M)^{-1} \mathbf{v}'} (X - M^{\mathsf{T}} M)^{-1} - V^{-1} \right) \mathbf{v}' \\ &= \mathbf{v}'^{\mathsf{T}} \left( (\nu - d - 1) (X - M^{\mathsf{T}} M)^{-1} - V^{-1} \right) \mathbf{v}' + \frac{(\nu - d - 1) (\mathbf{v}'^{\mathsf{T}} (X - M^{\mathsf{T}} M)^{-1} \mathbf{v}')^2}{1 - \mathbf{v}'^{\mathsf{T}} (X - M^{\mathsf{T}} M)^{-1} \mathbf{v}'} \\ &\leq \mathbf{v}'^{\mathsf{T}} \left( (\nu - d + 1) (X - M^{\mathsf{T}} M)^{-1} - V^{-1} \right) \mathbf{v}' + \frac{(\nu - d - 1) (\mathbf{v}'^{\mathsf{T}} (X - M^{\mathsf{T}} M)^{-1} \mathbf{v}')^2}{1 - \mathbf{v}'^{\mathsf{T}} (X - M^{\mathsf{T}} M)^{-1} \mathbf{v}'} \\ &\leq \mathbf{v}'^{\mathsf{T}} \left( (\nu - d + 1) (X - A^{\mathsf{T}} A)^{-1} - V^{-1} \right) \mathbf{v}' + \frac{(\nu - d - 1) (\mathbf{v}'^{\mathsf{T}} (X - M^{\mathsf{T}} A)^{-1} \mathbf{v}')^2}{1 - \mathbf{v}'^{\mathsf{T}} (X - M^{\mathsf{T}} M)^{-1} \mathbf{v}'} \\ &\leq \left( \frac{2\sqrt{2(\nu - d + 1) \ln(4/\delta)} - 2\ln(4/\delta)}{(\sqrt{\nu - d + 1} - \sqrt{2\ln(4/\delta)})^2} + \frac{2(\nu - d + 1)}{(\sqrt{\nu - d + 1} - \sqrt{2\ln(4/\delta)})^4} \right) \| V^{-1/2} \mathbf{v}' \|^2 \end{split}$$

Combining the two upper bounds we get

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

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

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

## **B.3** Privacy Proof for Algorithm 4

**Theorem B.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  $\nu \ge d$ . Let w be such that

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

Then, given that  $\sigma_{\min}(A) \ge w$ , the algorithm that samples a matrix from  $\mathcal{W}_d^{-1}(A^{\mathsf{T}}A,\nu)$  is  $(\epsilon, \delta)$ -differentially private.

We comment on the similarity between the bounds of Theorem B.1 and Theorem B.4. 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  $\mathcal{N}(\mathbf{0}, A^{\mathsf{T}}A)$ ). 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 B.4 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 v for A and v' for A', and as before, the matrix M is the matrix A with the *i*-th row all zeroed out. Therefore,  $A^{\mathsf{T}}A - vv^{\mathsf{T}} = A'^{\mathsf{T}}A' - v'v'^{\mathsf{T}} = M^{\mathsf{T}}M$ . So, denoting  $\sigma_{\min}(M)$  and  $\sigma_{\min}(A)$  as the least singular value of M and A resp., we have that  $\sigma_{\min}^2(M) \leq \sigma_{\min}^2(M) \leq \sigma_{\min}^2(M) + B^2$ . Same holds for the least singular value of M and A'.

Recall that

$$\mathsf{PDF}_{\mathcal{W}_{d}^{-1}(A^{\mathsf{T}}A,\nu)}(X) \propto \det(A^{\mathsf{T}}A)^{\frac{\nu}{2}} \det(X)^{\frac{\nu+p+1}{2}} \exp\left(-\frac{1}{2} \mathrm{tr}((A^{\mathsf{T}}A)X^{-1})\right)$$

We invoke the determinant update lemma, the Sherman Morisson lemma and the inequality  $\frac{1+x}{1+y} \leq \exp(x - \frac{y}{1+y})$  yet again to deduce:

$$\begin{split} \frac{\mathsf{PDF}_{\mathcal{W}_{d}^{-1}(A^{\mathsf{T}}A,\nu)}(X)}{\mathsf{PDF}_{\mathcal{W}_{d}^{-1}(A'^{\mathsf{T}}A',\nu)}(X)} &= \frac{\det(A^{\mathsf{T}}A)^{\nu/2}\exp\left(-\frac{1}{2}\mathrm{tr}((A^{\mathsf{T}}A)X^{-1})\right)}{\det(A'^{\mathsf{T}}A')^{\nu/2}\exp\left(-\frac{1}{2}\mathrm{tr}((A'^{\mathsf{T}}A')X^{-1})\right)} \\ &= \left(\frac{1+\boldsymbol{v}^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}}{1+\boldsymbol{v}'^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'}\right)^{\nu/2}\exp\left(-\frac{1}{2}\mathrm{tr}((A^{\mathsf{T}}A-A'^{\mathsf{T}}A')X^{-1}\right)\right) \\ &\leq \exp\left(\frac{\nu}{2}\left(\boldsymbol{v}^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v} - \frac{\boldsymbol{v}'^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'}{1+\boldsymbol{v}'^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'}\right) - \frac{1}{2}\left(\mathrm{tr}((\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}-\boldsymbol{v}'\boldsymbol{v}'^{\mathsf{T}})X^{-1})\right)\right) \\ &= \exp\left(\frac{1}{2}\left(\nu\cdot\boldsymbol{v}^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v} - \boldsymbol{v}^{\mathsf{T}}X^{-1}\boldsymbol{v}\right) - \frac{1}{2}\left(\frac{\nu\cdot\boldsymbol{v}'^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'}{1+\boldsymbol{v}'^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'} - \boldsymbol{v}'^{\mathsf{T}}X^{-1}\boldsymbol{v}'\right)\right) \\ &\leq \exp\left(\frac{1}{2}\boldsymbol{v}^{\mathsf{T}}\left(\nu(M^{\mathsf{T}}M)^{-1} - X^{-1}\right)\boldsymbol{v} - \frac{1}{2}\boldsymbol{v}'^{\mathsf{T}}\left(\frac{\nu}{1+\boldsymbol{v}'^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}}\boldsymbol{v}'(M^{\mathsf{T}}M)^{-1} - X^{-1}\right)\boldsymbol{v}'\right) \end{split}$$

We continue assuming  $X \sim \mathcal{W}_d^{-1}(A^{\mathsf{T}}A,\nu)$  (the case  $X \sim \mathcal{W}_d^{-1}(A'^{\mathsf{T}}A',\nu)$  is symmetric). By definition, we have that  $X^{-1} \sim \mathcal{W}_d((A^{\mathsf{T}}A)^{-1},\nu)$ . Hence  $(A^{\mathsf{T}}A)^{1/2}X^{-1}(A^{\mathsf{T}}A)^{-1/2} \sim \mathcal{W}_d(I_{d\times d},\nu)$ , which implies that the distribution of  $(A^{\mathsf{T}}A)^{1/2}X^{-1}(A^{\mathsf{T}}A)^{-1/2}$  is the same as generating 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

$$\boldsymbol{v}^{\mathsf{T}} \left( \nu (M^{\mathsf{T}} M)^{-1} - X^{-1} \right) \boldsymbol{v} = \boldsymbol{v}^{\mathsf{T}} \left( \nu (A^{\mathsf{T}} A)^{-1} - X^{-1} \right) \boldsymbol{v} - \frac{\nu \cdot (\boldsymbol{v}^{\mathsf{T}} (A^{\mathsf{T}} A)^{-1} \boldsymbol{v})^{2}}{1 - \boldsymbol{v}^{\mathsf{T}} (A^{\mathsf{T}} A)^{-1} \boldsymbol{v}} \\ \leq \left( (A^{\mathsf{T}} A)^{-1/2} \boldsymbol{v} \right)^{\mathsf{T}} \left( \nu I_{d \times d} - (A^{\mathsf{T}} A)^{1/2} X^{-1} (A^{\mathsf{T}} A)^{1/2} \right) \left( (A^{\mathsf{T}} A)^{-1/2} \boldsymbol{v} \right) \\ \leq \| (A^{\mathsf{T}} A)^{-1/2} \boldsymbol{v} \|^{2} \left( 2 \sqrt{2\nu \ln(4/\delta)} + 2 \ln(4/\delta) \right)$$

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

$$\begin{aligned} &- \boldsymbol{v}'^{\mathsf{T}} \left( \frac{\nu}{1 + \boldsymbol{v}'^{\mathsf{T}} (M^{\mathsf{T}} M)^{-1} \boldsymbol{v}'} (M^{\mathsf{T}} M)^{-1} - X^{-1} \right) \boldsymbol{v}' \\ &= - \boldsymbol{v}'^{\mathsf{T}} \left( \nu (M^{\mathsf{T}} M)^{-1} - X^{-1} \right) \boldsymbol{v}' + \frac{\nu \cdot (\boldsymbol{v}'^{\mathsf{T}} (M^{\mathsf{T}} M)^{-1} \boldsymbol{v}')^{2}}{1 - \boldsymbol{v}'^{\mathsf{T}} (M^{\mathsf{T}} M)^{-1} \boldsymbol{v}'} \\ &= - \boldsymbol{v}'^{\mathsf{T}} \left( \nu (A^{\mathsf{T}} A)^{-1} - X^{-1} \right) \boldsymbol{v}' + \frac{\nu \cdot (\boldsymbol{v}'^{\mathsf{T}} (M^{\mathsf{T}} M)^{-1} \boldsymbol{v}')^{2}}{1 - \boldsymbol{v}'^{\mathsf{T}} (M^{\mathsf{T}} M)^{-1} \boldsymbol{v}'} + \frac{\nu \cdot (\boldsymbol{v}'^{\mathsf{T}} (A^{\mathsf{T}} A)^{-1} \boldsymbol{v})^{2}}{1 - \boldsymbol{v}'^{\mathsf{T}} (M^{\mathsf{T}} M)^{-1} \boldsymbol{v}'} \\ &\leq - \boldsymbol{v}'^{\mathsf{T}} \left( \nu (A^{\mathsf{T}} A)^{-1} - X^{-1} \right) \boldsymbol{v}' + \frac{\nu \cdot (\boldsymbol{v}'^{\mathsf{T}} (A^{\mathsf{T}} A)^{-1} \boldsymbol{v})^{2}}{1 - \boldsymbol{v}'^{\mathsf{T}} (M^{\mathsf{T}} M)^{-1} \boldsymbol{v}'} + \frac{\nu \cdot (\boldsymbol{v}'^{\mathsf{T}} (A^{\mathsf{T}} A)^{-1} \boldsymbol{v})^{2}}{1 - \boldsymbol{v}'^{\mathsf{T}} (A^{\mathsf{T}} A)^{-1} \boldsymbol{v}'} \\ &\leq \| (A^{\mathsf{T}} A)^{-1} 2 \boldsymbol{v}' \|^{2} \left( 2 \sqrt{2\nu \ln(4/\delta)} + 2 \ln(4/\delta) \right) \\ &+ \nu \cdot \| (A^{\mathsf{T}} A)^{-1/2} \boldsymbol{v}' \|^{2} \| (A^{\mathsf{T}} A)^{-1/2} \boldsymbol{v} \|^{2} \left( \frac{1}{1 - \boldsymbol{v}'^{\mathsf{T}} (M^{\mathsf{T}} M)^{-1} \boldsymbol{v}'} + \frac{1}{1 - \boldsymbol{v}'^{\mathsf{T}} (A^{\mathsf{T}} A)^{-1} \boldsymbol{v}'} \right) \end{aligned}$$

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

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

### B.4 An Additional Privacy Theorem — Gaussian Noise for the Inverse

**Theorem B.5.** Fix  $\epsilon \in (0,1)$  and  $\delta \in (0, e^{-1})$ . Let A be a  $(n \times d)$ -matrix where the  $l_2$ -norm of each row is bounded by B, where it is publicly known that  $\frac{\sigma_{\min}(A^{\mathsf{T}}A)}{B^2} \geq 1 + \rho$  with  $\rho > 0$ . Then the algorithm that outputs  $(A^{\mathsf{T}}A)^{-1} + N$  where N is a symmetric matrix with each entry on or above the main diagonal of N is sampled i.i.d from  $\mathcal{N}\left(0, \frac{8\log(2/\delta)}{\rho^2\epsilon^2}\right)$  is  $(\epsilon, \delta)$ -differentially private.

*Proof.* The proof of the theorem just bounds the  $l_2$ -global sensitivity of the inverse, using the Sherman Morrison formula. We then use the fact that by independently adding noise to each entry in  $(A^{\mathsf{T}}A)_{j,k}^{-1}$  for  $j \leq k$  where the noise is sampled i.i.d from  $\mathcal{N}\left(0, GS_2^2 \cdot \frac{2\log(2/\delta)}{\epsilon^2}\right)$  is  $(\epsilon, \delta)$ -differentially private.

Denote A and A', two matrices that differ on a single row, which is denoted  $\boldsymbol{v}$  in A and  $\boldsymbol{v}'$  in A. Therefore,  $A'^{\mathsf{T}}A' = A^{\mathsf{T}}A + \boldsymbol{v}'\boldsymbol{v}'^{\mathsf{T}} - \boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}$ , so Weyl's inequality gives that  $|\sigma_{\min}(A^{\mathsf{T}}A) - \sigma_{\min}(A'^{\mathsf{T}}A')| \leq B^2$ . Denoting M as the matrix we get by zeroing out the *i*-th row on A or A', we have

$$(A^{\mathsf{T}}A)^{-1} = (M^{\mathsf{T}}M)^{-1} - \frac{(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}}{1 + \boldsymbol{v}^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}}$$
$$(A'^{\mathsf{T}}A')^{-1} = (M^{\mathsf{T}}M)^{-1} - \frac{(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'\boldsymbol{v}'^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}}{1 + \boldsymbol{v}'^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'}$$

Hence,

As

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

Let  $x_1, x_2, \ldots, x_d$  be the eigenvectors of M, corresponding to the eigenvalues  $\mu_1, \ldots, \mu_d$ . Then, for any j, k we have

$$\boldsymbol{x}_{j}^{\mathsf{T}}\left((A'^{\mathsf{T}}A')^{-1} - (A^{\mathsf{T}}A)^{-1}\right)\boldsymbol{x}_{k} = \mu_{j}^{-1}\mu_{k}^{-1}\left(\frac{(\boldsymbol{v}\cdot\boldsymbol{x}_{j})(\boldsymbol{v}\cdot\boldsymbol{x}_{k})}{1 + \boldsymbol{v}^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}} - \frac{(\boldsymbol{v}'\cdot\boldsymbol{x}_{j})(\boldsymbol{v}'\cdot\boldsymbol{x}_{k})}{1 + \boldsymbol{v}'^{\mathsf{T}}(M^{\mathsf{T}}M)^{-1}\boldsymbol{v}'}\right)$$

Due to Weyl's inequality,  $\sigma_{\min}(M^{\mathsf{T}}M) \ge \sigma_{\min}(A^{\mathsf{T}}A) - \|\boldsymbol{v}\|^2 \ge \rho \cdot B^2$ . And so, together with the inequality  $(a-b)^2 \le 2a^2 + 2b^2$  we get

$$\begin{split} \left\| (A'^{\mathsf{T}}A')^{-1} - (A^{\mathsf{T}}A)^{-1} \right\|_{F}^{2} &= \sum_{j,k=1}^{d} \left( \mathbf{x}_{j}^{\mathsf{T}} \left( (A'^{\mathsf{T}}A')^{-1} - (A^{\mathsf{T}}A)^{-1} \right) \mathbf{x}_{k} \right)^{2} \\ &= \frac{2}{1 + \mathbf{v}^{\mathsf{T}} (M^{\mathsf{T}}M)^{-1} \mathbf{v}} \sum_{j,k} \frac{(\mathbf{v} \cdot \mathbf{x}_{j})^{2} (\mathbf{v} \cdot \mathbf{x}_{k})^{2}}{\mu_{j}^{2} \mu_{k}^{2}} \\ &+ \frac{2}{1 + \mathbf{v}'^{\mathsf{T}} (M^{\mathsf{T}}M)^{-1} \mathbf{v}'} \sum_{j,k} \frac{(\mathbf{v}' \cdot \mathbf{x}_{j})^{2} (\mathbf{v}' \cdot \mathbf{x}_{k})^{2}}{\mu_{j}^{2} \mu_{k}^{2}} \\ &\leq \frac{2}{(\rho B^{2})^{2}} \sum_{j,k} \left( \mathbf{v} \cdot \mathbf{x}_{j} \right)^{2} (\mathbf{v} \cdot \mathbf{x}_{k})^{2} + (\mathbf{v}' \cdot \mathbf{x}_{j})^{2} (\mathbf{v}' \cdot \mathbf{x}_{k})^{2} \\ &= \frac{2}{(\rho B^{2})^{2}} \sum_{j,k} \left( \left( \sum_{j} (\mathbf{v} \cdot \mathbf{x}_{j})^{2} \right) \left( \sum_{k} (\mathbf{v} \cdot \mathbf{x}_{k})^{2} \right) + \left( \sum_{j} (\mathbf{v}' \cdot \mathbf{x}_{j})^{2} \right) \left( \sum_{k} (\mathbf{v}' \cdot \mathbf{x}_{k})^{2} \right) \right) \\ &= \frac{2 ||\mathbf{v}||^{4} + 2 ||\mathbf{v}'||^{4}}{(\rho B^{2})^{2}} \leq \frac{4B^{4}}{(\rho B^{2})^{2}} = \frac{4}{\rho^{2}} \end{split}$$

# C Utility Theorems

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

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

Our goal is to compare the distance between our predictor to the predictor one gets without noise, i.e. to  $\hat{\boldsymbol{\beta}} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\boldsymbol{y}$ . 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  $\hat{\beta}$  is ill-defined unless we regularize the problem.)

**Theorem C.1.** Fix  $X \in \mathbb{R}^{n \times p}$  and  $\mathbf{y} \in \mathbb{R}^n$  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  $\widetilde{X^{\mathsf{T}}y} = X^{\mathsf{T}}y + n$  where each entry of N and n is sampled i.i.d from  $\mathcal{N}(0,\sigma^2)$ . Then, there exists some constant  $C \geq 1$  s.t. if 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

$$\left\|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right\| = \left\| (\widetilde{X^{\mathsf{T}}X})^{-1} \widetilde{X^{\mathsf{T}}\boldsymbol{y}} - (X^{\mathsf{T}}X)^{-1} X^{\mathsf{T}}\boldsymbol{y} \right\| \le 2\eta \widehat{\boldsymbol{\beta}} + \frac{\eta}{C}$$

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  $X^{\mathsf{T}}X + M$  where M is a symmetric matrix whose entries along and above the main diagonal are sampled i.i.d from a suitable  $\mathcal{N}(0, \sigma^2)$ . However, one can denote  $M = \frac{1}{\sqrt{2}}(N + N^{\mathsf{T}})$  for a matrix N whose entries are i.i.d samples from  $\mathcal{N}(0, \sigma^2)$ , and so the same result, up to a factor of  $\sqrt{2}$ , holds for Analyze Gauss.

*Proof.* Plugging in (3) we get

$$(\widetilde{X^{\mathsf{T}}X})^{-1}\widetilde{X^{\mathsf{T}}y} = (I_{p\times p} - (X^{\mathsf{T}}X)^{-1}(I_{p\times p} - N(X^{\mathsf{T}}X)^{-1})^{-1}N) (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}y + (I_{p\times p} - (X^{\mathsf{T}}X)^{-1}(I_{p\times p} - N(X^{\mathsf{T}}X)^{-1})^{-1}N) (X^{\mathsf{T}}X)^{-1}n$$

Denoting  $Z = (X^{\mathsf{T}}X)^{-1}(I_{p\times p} - N(X^{\mathsf{T}}X)^{-1})^{-1}N$ , we derive a bound on  $\left\| (\widetilde{X^{\mathsf{T}}X})^{-1}\widetilde{X^{\mathsf{T}}y} - (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}y \right\|$  using bounds on  $\|Z\|$ ,  $\|I - Z\|$  and  $\|n\|$ .

Standard bounds on a symmetric ensemble of Gaussians [Tao12] 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  $||N|| \cdot ||(X^{\mathsf{T}}X)^{-1}|| \leq \eta$ . Hence, all singular values of  $N(X^{\mathsf{T}}X)^{-1}$  are upper bounded in absolute value by  $\eta$ , and so all singular values of  $I - N(X^{\mathsf{T}}X)^{-1}$  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  $||\mathbf{n}||^2 \sim \sigma^2 \cdot \chi_p^2$ , and so, w.p.  $\geq 1 - \frac{\nu}{2}$  it holds that  $||\mathbf{n}|| \leq \sigma(\sqrt{p} + \sqrt{2\ln(2/\nu)})$ .

Thus, we get

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

**Corollary C.2.** Denote  $\rho = \frac{\sigma_{\min}(\widetilde{X^{\intercal}X)}}{2\sigma\sqrt{p}\log(1/\nu)}$ . Then, for the same constant *C* in Theorem C.1, if  $\rho \ge 2C$  then 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 C.1, and the observation that we can flip the role of  $X^{\mathsf{T}}X$  and  $X^{\mathsf{T}}X$  because the Gaussian distribution is symmetric. And so, we just use the notation  $\rho = \frac{C}{n}$ .

**Theorem C.3.** Let  $W \sim \mathcal{W}_{p+1}(\sigma^2 I, k)$ , and denote  $N \in \mathbb{R}^{p \times p}$  and  $\mathbf{n} \in \mathbb{R}^p$  s.t.  $W = \begin{pmatrix} N & \mathbf{n} \\ \mathbf{n}^{\mathsf{T}} & * \end{pmatrix}$ . Let  $X \in \mathbb{R}^{n \times p}$  be a matrix s.t.  $X^{\mathsf{T}}X$  is invertible and let  $\mathbf{y} \in \mathbb{R}^n$ , and such that there exists a  $C \geq 2$  s.t.  $\sigma_{\min}(X^{\mathsf{T}}X) = C \cdot \sigma^2(\sqrt{k} + \sqrt{p} + \sqrt{2\ln(4/\nu)})^2$ . Denote  $X^{\mathsf{T}}X = X^{\mathsf{T}}X + N$  and  $X^{\mathsf{T}}\mathbf{y} = X^{\mathsf{T}}\mathbf{y} + \mathbf{n}$ . Then

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

Proof. Because  $\sigma^2 I$  is a diagonal matrix, standard results on the Wishart distribution give that  $N \sim W_p(\sigma^2 I_{p \times p}, k)$ . 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  $N = \sigma^2 R^{\mathsf{T}} R$ . The Woodbury formula gives that

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

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

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

Now, if we denote  $Q = U\Lambda V^{\mathsf{T}}$  where Q's singular values are  $\lambda_1, \ldots, \lambda_d$ , we get  $Q^{\mathsf{T}}(I - QQ^{\mathsf{T}})^{-1}Q = V$ .  $\operatorname{diag}\left(\frac{\lambda_i^2}{1-\lambda_i^2}\right) \cdot V^{\mathsf{T}} = V \cdot \operatorname{diag}\left(\frac{1}{1-\lambda_i^2}-1\right) \cdot V^{\mathsf{T}}.$  Note that  $Q^{\mathsf{T}}Q = \sigma^2 (X^{\mathsf{T}}X)^{-1/2} R^{\mathsf{T}} R (X^{\mathsf{T}}X)^{-1/2}$  and so, due to Lemma A.3 we have  $\lambda_1^2 = \sigma_{\max}(Q^{\mathsf{T}}Q) \leq \frac{\sigma^2(\sqrt{k}+\sqrt{p}+\sqrt{2\ln(4/\nu)})^2}{\sigma_{\min}(X^{\mathsf{T}}X)} \leq C^{-1}$  w.p.  $\geq 1-\nu/2$ . Which means that w.p.  $\geq 1-\nu/2$  we have  $\sigma_{\max}(Q^{\mathsf{T}}(I-QQ^{\mathsf{T}})^{-1}Q) \leq \frac{1}{C-1}$ . And so we have that both (i)  $(X^{\mathsf{T}}X)^{-1} - (X^{\mathsf{T}}X+N)^{-1} \leq \frac{1}{C-1}(X^{\mathsf{T}}X)^{-1}$  and (ii)  $(X^{\mathsf{T}}X+N)^{-1} \leq \frac{C-2}{C-1}(X^{\mathsf{T}}X)^{-1}$ . Next we turn to bound  $\|\boldsymbol{n}\|$ . One easy bound, given Lemma A.3, is to show that w.p.  $\geq 1-\nu/2$  it holds

that

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

Alternatively we can derive the following bound. Each coordinate in n is the result of the dot-product between the j-th column of R, denoted  $r_j$  with the d-th column of R, denoted  $r_d$ . 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<sup>10</sup> Gaussians  $\mathcal{N}\left(0, \frac{\sigma^2}{2}\right)$ . And so  $r_j \cdot r_d = Z_{j_1} - Z_{j_2}$  where  $Z_{j_1}, Z_{j_2} \sim \frac{\sigma}{\sqrt{2}} \cdot \chi_k^2$ . Tail bounds for the  $\chi^2$ -distribution (see Claim A.2) 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  $||n|| \leq 2\sigma^2 \sqrt{2k \cdot \ln(4p/\nu)}$ .<sup>11</sup>

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

$$\begin{split} \widehat{\boldsymbol{\beta}} &- \widetilde{\boldsymbol{\beta}} = \left( (X^{\mathsf{T}}X)^{-1} - (X^{\mathsf{T}}X + N)^{-1} \right) X^{\mathsf{T}} \boldsymbol{y} - (X^{\mathsf{T}}X + N)^{-1} \boldsymbol{n} \\ \Rightarrow \|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\| &\leq \frac{1}{C-1} \| (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}y\| + \frac{2\sigma^2(C-2)}{C-1} \| (X^{\mathsf{T}}X)^{-1} \| \sqrt{2kp \cdot \ln(4p/\nu)} \\ &= \frac{1}{C-1} \|\widehat{\boldsymbol{\beta}}\| + \frac{2\sigma^2(C-2)}{(C-1)\sigma_{\min}(X^{\mathsf{T}}X)} \sqrt{2kp \cdot \ln(4p/\nu)} \\ \|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\| &\leq \frac{1}{C-1} \|\widehat{\boldsymbol{\beta}}\| + \frac{\sigma^2(C-2)}{(C-1)\sigma_{\min}(X^{\mathsf{T}}X)} (\sqrt{k} + \sqrt{p} + \sqrt{2\ln(4/\nu)})^2 \end{split}$$

or:

<sup>&</sup>lt;sup>10</sup>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.

<sup>&</sup>lt;sup>11</sup>We conjecture that the true bound in log(p)-factor smaller, i.e.  $O(\sigma^2 \sqrt{2kp \cdot \ln(4/\nu)})$ .