A. Extended Introductory Discussion

Due to space constraint, a few details from the introductory parts (Sections 1,2) were omitted. We bring them in this appendix. We especially recommend the uninformed reader to go over the extended OLS background we provide in Appendix A.3.

A.1. Proof Of Privacy of Algorithm 1

Theorem A.1. *Algorithm 1 is (ϵ, δ)-differentially private.*

Proof. The proof of the theorem is based on the fact the Algorithm 1 is the result of composing the differentially private Propose-Test-Release algorithm of (Dwork & Lei, 2009) with the differentially private analysis of the Johnson-Lindenstrauss transform of (Sheffet, 2015).

More specifically, we use Theorem B.1 from (Sheffet, 2015) that states that given a matrix A whose all of its singular values at greater than $T(\epsilon,\delta)$ where $T(\epsilon,\delta)^2 = \frac{2B^2}{\epsilon}\left(\sqrt{2r\ln(4/\delta)} + 2\ln(4/\delta)\right)$, publishing RA is (ϵ,δ)-differentially private for a r-row matrix R whose entries sampled are i.i.d normal Gaussians. Since we have that all of the singular values of A' are greater than w (as specified in Algorithm 1), outputting RA' is $(\epsilon/2, \delta/2)$-differentially private. The rest of the proof boils down to showing that (i) the if-else-condition is $(\epsilon/2, 0)$-differentially private and that (ii) w.p. $\leq \delta/2$ any matrix A whose smallest singular value is smaller than w passes the if-condition (step 3). If both these facts hold, then knowing whether we pass the if-condition or not is $(\epsilon/2)$-differentially private and the output of the algorithm is $(\epsilon/2, \delta)$-differentially private, hence basic composition gives the overall bound of (ϵ,δ)-differential privacy.

To prove (i) we have that for any pair of neighboring matrices A and B that differ only on the i-th row, denoted $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ resp., we have $B^\mathsf{T}B - \boldsymbol{b}_i\boldsymbol{b}_i^\mathsf{T} = A^\mathsf{T}A - \boldsymbol{a}_i\boldsymbol{a}_i^\mathsf{T}$. Applying Weyl's inequality we have

$$\begin{aligned}\sigma_{\min}(B^\mathsf{T}B) &\leq \sigma_{\min}(B^\mathsf{T}B - \boldsymbol{b}_i\boldsymbol{b}_i^\mathsf{T}) + \sigma_{\max}(\boldsymbol{b}_i\boldsymbol{b}_i^\mathsf{T}) \\ &\leq \sigma_{\min}(A^\mathsf{T}A) + \sigma_{\max}(\boldsymbol{a}_i\boldsymbol{a}_i^\mathsf{T}) + \sigma_{\max}(\boldsymbol{b}_i\boldsymbol{b}_i^\mathsf{T}) \\ &\leq \sigma_{\min}(A^\mathsf{T}A) + 2B^2\end{aligned}$$

hence $|\sigma_{\min}(A)^2 - \sigma_{\min}(B)^2| \leq 2B^2$, so adding $Lap(\frac{4B^2}{\epsilon})$ is $(\epsilon/2)$-differentially private.

To prove (ii), note that by standard tail-bounds on the Laplace distribution we have that $\mathbf{Pr}[Z < -\frac{4B^2\ln(1/\delta)}{\epsilon}] \leq \frac{\delta}{2}$. Therefore, w.p. $1-\delta/2$ it holds that any matrix A that passes the if-test of the algorithm must have $\sigma_{\min}(A)^2 > w^2$. Also note that a similar argument shows that for any $0 < \beta < 1$, any matrix A s.t. $\sigma_{\min}(A)^2 > w^2 + \frac{4B^2\ln(1/\beta)}{\epsilon}$ passes the if-condition of the algorithm w.p. $1-\beta$. □

A.2. Omitted Preliminary Details

Linear Algebra and Pseudo-Inverses. Given a matrix M we denote its SVD as $M = USV^\mathsf{T}$ with U and V being orthonormal matrices and S being a non-negative diagonal matrix whose entries are the singular values of M. We use $\sigma_{\max}(M)$ and $\sigma_{\min}(M)$ to denote the largest and smallest singular value resp. Despite the risk of confusion, we stick to the standard notation of using σ^2 to denote the variance of a Gaussian, and use $\sigma_j(M)$ to denote the j-th singular value of M. We use M^+ to denote the Moore-Penrose inverse of M, defined as $M^+ = VS^{-1}U^T$ where S^{-1} is a matrix with $S^{-1}_{j,j} = 1/S_{j,j}$ for any j s.t. $S_{j,j} > 0$.

The Gaussian Distribution. A univariate Gaussian $\mathcal{N}(\mu,\sigma^2)$ denotes the Gaussian distribution whose mean is μ and variance σ^2, with $\mathsf{PDF}(x) = (\sqrt{2\pi\sigma^2})^{-1}\exp(-\frac{x-\mu}{2\sigma^2})$. Standard concentration bounds on Gaussians give that $\mathbf{Pr}[x > \mu + 2\sigma\sqrt{\ln(1/\nu)}] < \nu$ for any $\nu \in (0, \frac{1}{e})$. A multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ for some positive semi-definite Σ denotes the multivariate Gaussian distribution where the mean of the j-th coordinate is the μ_j and the co-variance between coordinates j and k is $\Sigma_{j,k}$. The PDF of such Gaussian is defined only on the subspace $colspan(\Sigma)$, where for every $x \in colspan(\Sigma)$ we have $\mathsf{PDF}(\boldsymbol{x}) = \left((2\pi)^{rank(\Sigma)}\cdot\tilde{\det}(\Sigma)\right)^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\Sigma^+(\boldsymbol{x}-\boldsymbol{\mu})\right)$ and $\tilde{\det}(\Sigma)$ is the multiplication of all non-zero singular values of Σ. A matrix Gaussian distribution denoted $\mathcal{N}(M_{a\times b}, U, V)$ has mean M, variance U on its rows and variance V on its columns. For full rank U and V it holds that $\mathsf{PDF}_{\mathcal{N}(M,U,V)}(X) = (2\pi)^{-ab/2}(\det(U))^{-b/2}(\det(V))^{-a/2}\cdot\exp(-\frac{1}{2}\mathrm{trace}\left(V^{-1}(X-M)^\mathsf{T}U^{-1}(X-M)\right))$. In our case, we will only use matrix Gaussian distributions with $\mathcal{N}(M_{a\times b}, I_{a\times a}, V)$ and so each row in this matrix is an i.i.d sample from a b-dimensional multivariate Gaussian $\mathcal{N}((M)_{j\to}, V)$.

We will repeatedly use the rules regarding linear operations on Gaussians. That in, for any c, it holds that $c\mathcal{N}(\mu,\sigma^2) = \mathcal{N}(c\cdot\mu, c^2\sigma^2)$. For any C it holds that $C\cdot\mathcal{N}(\boldsymbol{\mu},\Sigma) = \mathcal{N}(C\boldsymbol{\mu}, C\Sigma C^\mathsf{T})$. And for any C is holds that $\mathcal{N}(M,U,V)\cdot C = \mathcal{N}(MC, U, C^\mathsf{T}VC)$. In particular, for any $\boldsymbol{c}$ (which can be viewed as a $b\times 1$-matrix) it holds that $\mathcal{N}(M,U,V)\cdot\boldsymbol{c} = \mathcal{N}(M\boldsymbol{c}, U, \boldsymbol{c}^TV\boldsymbol{c}) = \mathcal{N}(M\boldsymbol{c}, \boldsymbol{c}^TV\boldsymbol{c}\cdot U)$.

We will also require the following proposition.

Proposition A.2. *Given σ^2, λ^2 s.t. $1 \leq \frac{\sigma^2}{\lambda^2} \leq c^2$ for some constant c, let X and Y be two random Gaussians s.t. $X \sim \mathcal{N}(0,\sigma^2)$ and $Y \sim \mathcal{N}(0,\lambda^2)$. It follows that $\frac{1}{c}\mathsf{PDF}_Y(x) \leq \mathsf{PDF}_X(x) \leq c\mathsf{PDF}_{cY}(x)$ for any x.*

Corollary A.3. *Under the same notation as in Proposition A.2, for any set $S \subset \mathbb{R}$ it holds that $\frac{1}{c}\mathbf{Pr}_{x\leftarrow Y}[x\in S] \leq \mathbf{Pr}_{x\leftarrow X}[x\in S] \leq c\mathbf{Pr}_{x\leftarrow cY}[x\in S] =$*

$c\mathbf{Pr}_{x\leftarrow Y}[x \in S/c]$

Proof. The proof is mere calculation.

$$\frac{\mathsf{PDF}_X(x)}{\mathsf{PDF}_{cY}(x)} = \sqrt{\frac{c^2\lambda^2}{\sigma^2}} \cdot \frac{\exp(-\frac{x^2}{2\sigma^2})}{\exp(-\frac{x^2}{2c^2\lambda^2})}$$
$$\leq c \cdot \exp(\frac{x^2}{2}(\frac{1}{c^2\lambda^2} - \frac{1}{\sigma^2})) \leq c \cdot \exp(0) = c$$
$$\frac{\mathsf{PDF}_X(x)}{\mathsf{PDF}_Y(x)} = \sqrt{\frac{\lambda^2}{\sigma^2}} \cdot \frac{\exp(-\frac{x^2}{2\sigma^2})}{\exp(-\frac{x^2}{2\lambda^2})}$$
$$\geq c^{-1}\exp(\tfrac{x^2}{2}(\tfrac{1}{\lambda^2} - \tfrac{1}{\sigma^2})) \geq \tfrac{\exp(0)}{c} = c^{-1}$$

□

The T_k-Distribution. The T_k-*distribution*, where k is referred to as the degrees of freedom of the distribution, denotes the distribution over the reals created by *independently* sampling $Z \sim \mathcal{N}(0,1)$ and $\|\zeta\|^2 \sim \chi^2_k$, and taking the quantity $\frac{Z}{\sqrt{\|\zeta\|^2/k}}$. Its PDF is given by $\mathsf{PDF}_{T_k}(x) \propto \left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}$. It is a known fact that as k increases, T_k becomes closer and closer to a normal Gaussian. The T-distribution is often used to determine suitable bounds on the rate of converges, as we illustrate in Section A.3. As the T-distribution is heavy-tailed, existing tail bounds on the T-distribution (which are of the form: if $\tau_\nu = C\sqrt{k((1/\nu)^{2/k}-1)}$ for some constant C then $\int_{\tau_\nu}^{\infty} \mathsf{PDF}_{T_k}(x)dx < \nu$) are often cumbersome to work with. Therefore, in many cases in practice, it common to assume $\nu = \Theta(1)$ (most commonly, $\nu = 0.05$) and use existing tail-bounds on normal Gaussians.

Differential Privacy facts. It is known (Dwork et al., 2006b) that if ALG outputs a vector in $\mathbb{R}^d$ such that for any A and A' it holds that $\|\mathsf{ALG}(A) - \mathsf{ALG}(A')\|_1 \leq B$, then adding Laplace noise $Lap(1/\epsilon)$ to each coordinate of the output of $\mathsf{ALG}(A)$ satisfies ϵ-differential privacy. Similarly, (2006b) showed that if for any neighboring A and A' it holds that $\|\mathsf{ALG}(A) - \mathsf{ALG}(A')\|_2^2 \leq \Delta^2$ then adding Gaussian noise $\mathcal{N}(0, \Delta^2 \cdot \frac{2\ln(2/\delta)}{\epsilon^2})$ to each coordinate of the output of $\mathsf{ALG}(A)$ satisfies (ϵ,δ)-differential privacy.

Another standard result (Dwork et al., 2006a) gives that the composition of the output of a (ϵ_1,δ_1)-differentially private algorithm with the output of a (ϵ_2,δ_2)-differentially private algorithm results in a $(\epsilon_1+\epsilon_2, \delta_1+\delta_2)$-differentially private algorithm.

A.3. Detailed Background on Ordinary Least Squares

For the unfamiliar reader, we give a short description of the model under which OLS operates as well as the confidence bounds one derives using OLS. This is by no means an exhaustive account of OLS and we refer the interested reader to (Rao, 1973; Muller & Stewart, 2006).

Given n observations $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ where for all i we have $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, we assume the existence of a p-dimensional vector $\boldsymbol{\beta} \in \mathbb{R}^p$ s.t. the label y_i was derived by $y_i = \boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_i + e_i$ where $e_i \sim \mathcal{N}(0,\sigma^2)$ independently (also known as the homoscedastic Gaussian model). We use the matrix notation where X denotes the $(n \times p)$-matrix whose rows are $\boldsymbol{x}_i$, and use $\boldsymbol{y}, \boldsymbol{e} \in \mathbb{R}^n$ to denote the vectors whose i-th entry is y_i and e_i resp. To simplify the discussion, we assume X has full rank.

The parameters of the model are therefore $\boldsymbol{\beta}$ and σ^2, which we set to discover. To that end, we minimize $\min_{\boldsymbol{z}} \|\boldsymbol{y} - X\boldsymbol{z}\|^2$ and solve

$$\hat{\boldsymbol{\beta}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{y} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}(X\boldsymbol{\beta}+\boldsymbol{e}) = \boldsymbol{\beta}+X^+\boldsymbol{e}$$

As $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 I_{n\times n})$, it holds that $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(X^\mathsf{T}X)^{-1})$, or alternatively, that for every coordinate j it holds that $\hat{\beta}_j = \boldsymbol{e}_j^\mathsf{T}\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\beta_j, \sigma^2(X^\mathsf{T}X)^{-1}_{j,j})$. Hence we get $\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{(X^\mathsf{T}X)^{-1}_{j,j}}} \sim \mathcal{N}(0,1)$. In addition, we denote the vector

$$\boldsymbol{\zeta} = y - X\hat{\boldsymbol{\beta}} = (X\boldsymbol{\beta}+\boldsymbol{e}) - X(\boldsymbol{\beta}+X^+\boldsymbol{e}) = (I - XX^+)\boldsymbol{e}$$

and since XX^+ is a rank-p (symmetric) projection matrix, we have $\boldsymbol{\zeta} \sim \mathcal{N}(0, \sigma^2(I - XX^+))$. Therefore, $\|\boldsymbol{\zeta}\|^2$ is equivalent to summing the squares of $(n-p)$ i.i.d samples from $\mathcal{N}(0,\sigma^2)$. In other words, the quantity $\|\boldsymbol{\zeta}\|^2/\sigma^2$ is sampled from a χ^2-distribution with $(n-p)$ degrees of freedom.

We sidetrack from the OLS discussion to give the following bounds on the l_2-distance between $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$, as the next claim shows.

Claim A.4. *For any $0 < \nu < 1/2$, the following holds w.p. $\geq 1-\nu$ over the* randomness of the model *(the randomness over $\boldsymbol{e}$)*

$$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|^2 = \|X^+\boldsymbol{e}\|^2$$
$$= O\left(\sigma^2 \log(p/\nu) \cdot \|X^+\|_F^2\right) \quad (6)$$
$$\|\hat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{\beta} + X^+\boldsymbol{e}\|^2$$
$$= O(\left(\|\boldsymbol{\beta}\| + \sigma \cdot \|X^+\|_F \cdot \sqrt{\log(p/\nu)}\right)^2)$$
$$\left|\tfrac{1}{n-p}\|\boldsymbol{\zeta}\|^2 - \sigma^2\right| = O(\sqrt{\tfrac{\ln(1/\nu)}{n-p}})$$

Proof. Since $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 I_{n\times n})$ then $X^+\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2(X^\mathsf{T}X)^{-1})$. Denoting the SVD decomposition $(X^\mathsf{T}X)^{-1} = VSV^\mathsf{T}$ with S denoting the diagonal matrix whose entries are $\sigma^{-2}_{\max}(X), \ldots, \sigma^{-2}_{\min}(X)$, we have that $V^\mathsf{T}X^+\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 S)$. And so, each coordinate of $V^\mathsf{T}X^+\boldsymbol{e}$ is distributed like an i.i.d Gaussian. So

w.p. $\geq 1-\nu/2$ non of these Gaussians is a factor of $O(\sigma\sqrt{\ln(p/\nu)})$ greater than its standard deviation. And so w.p. $\geq 1-\nu/2$ it holds that $\|X^+\boldsymbol{e}\|^2 = \|V^\mathsf{T}X^+\boldsymbol{e}\|^2 \leq O(\sigma^2\log(p/\nu)\left(\sum_i \sigma_i^{-2}(X)\right))$. Since $\sum_i \sigma_i^{-2}(X) = \mathrm{trace}((X^\mathsf{T}X)^{-1}) = \mathrm{trace}(X^+(X^+)^\mathsf{T}) = \|X^+\|_F^2$, the bound of (6) is proven.

The bound on $\|\hat{\boldsymbol{\beta}}\|^2$ is an immediate corollary of (6) using the triangle inequality.[8] The bound on $\|\boldsymbol{\zeta}\|^2$ follows from tail bounds on the χ^2_{n-p} distribution, as detailed in Section 2. □

Returning to OLS, it is important to note that $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\zeta}$ are independent of one another. (Note, $\hat{\boldsymbol{\beta}}$ depends solely on $X^+\boldsymbol{e} = (X^+X)X^+\boldsymbol{e} = X^+P_U\boldsymbol{e}$, whereas $\boldsymbol{\zeta}$ depends on $(I-XX^+)\boldsymbol{e} = P_{U^\perp}\boldsymbol{e}$. As $\boldsymbol{e}$ is spherically symmetric, the two projections are independent of one another and so $\hat{\boldsymbol{\beta}}$ is independent of $\boldsymbol{\zeta}$.) As a result of the above two calculations, we have that the quantity

$$t_{\hat{\beta}_j}(\beta_j) \stackrel{\mathrm{def}}{=} \frac{\hat{\beta}_j-\beta_j}{\sqrt{(X^\mathsf{T}X)^{-1}_{j,j}}\cdot\frac{\|\boldsymbol{\zeta}\|}{\sqrt{n-p}}} = \frac{\hat{\beta}_j-\beta_j}{\sigma\sqrt{(X^\mathsf{T}X)^{-1}_{j,j}}}\Big/\frac{\|\boldsymbol{\zeta}\|}{\sigma\sqrt{n-p}}$$

is distributed like a T-distribution with $(n-p)$ degrees of freedom. Therefore, we can compute an exact probability estimation for this quantity. That is, for any measurable $S\subset\mathbb{R}$ we have

$$\mathbf{Pr}\left[\hat{\boldsymbol{\beta}} \text{ and } \boldsymbol{\zeta} \text{ satisfying } t_{\hat{\beta}_j}(\beta_j)\in S\right] = \int_S \mathsf{PDF}_{T_{n-p}}(x)dx$$

The importance of the *t-value* $t(\beta_j)$ lies in the fact that it can be fully estimated from the observed data X and y (for any value of β_j), which makes it a *pivotal quantity*. Therefore, given X and $\boldsymbol{y}$, we can use $t(\beta_j)$ to describe the likelihood of any β_j — for any $z\in\mathbb{R}$ we can now give an estimation of how likely it is to have $\beta_j = z$ (which is $\mathsf{PDF}_{T_{n-p}}(t(z))$). The t-values enable us to perform multitude of statistical inferences. For example, we can say which of two hypotheses is more likely and by how much (e.g., we are 5-times more likely that the hypothesis $\beta_j=3$ is true than the hypothesis $\beta_j=14$ is true); we can compare between two coordinates j and j' and report we are more confident that $\beta_j>0$ than $\beta_{j'}>0$; or even compare among the t-values we get across multiple datasets (such as the datasets we get from subsampling rows from a single dataset).

In particular, we can use $t(\beta_j)$ to α-reject unlikely values of β_j. Given $0<\alpha<1$, we denote c_α as the number for which the interval $(-c_\alpha, c_\alpha)$ contains a probability mass of $1-\alpha$ from the T_{n-p}-distribution. And so we derive a corresponding *confidence interval* I_α centered at $\hat{\beta}_j$ where $\beta_j\in I_\alpha$ with confidence of level of $1-\alpha$.

We comment as to the actual meaning of this confidence interval. Our analysis thus far applied w.h.p to a vector $\boldsymbol{y}$ derived according to this model. Such X and $\boldsymbol{y}$ will result in the quantity $t_{\hat{\beta}_j}(\beta_j)$ being distributed like a T_{n-p}-distribution — where β_j is given as the model parameters and $\hat{\beta}_j$ is the random variable. We therefore have that guarantee that for X and $\boldsymbol{y}$ derived according to this model, the event $E_\alpha \stackrel{\mathrm{def}}{=} \hat{\beta}_j \in \left(\beta_j \pm c_\alpha\cdot\sqrt{(X^\mathsf{T}X)^{-1}_{j,j}\cdot\frac{\|\boldsymbol{\zeta}\|^2}{n-p}}\right)$ happens w.p. $1-\alpha$. However, the analysis done over a *given* dataset X and $\boldsymbol{y}$ (once $\boldsymbol{y}$ has been drawn) views the quantity $t_{\hat{\beta}_j}(\beta_j)$ with $\hat{\beta}_j$ given and β_j unknown. Therefore the event E_α either holds or does not hold. That is why the alternative terms of *likelihood* or *confidence* are used, instead of probability. We have a confidence level of $1-\alpha$ that indeed $\beta_j\in\hat{\beta}_j\pm c_\alpha\cdot\sqrt{(X^\mathsf{T}X)^{-1}_{j,j}\cdot\frac{\|\boldsymbol{\zeta}\|^2}{n-p}}$, because this event does happen in $1-\alpha$ fraction of all datasets generated according to our model.

Rejecting the Null Hypothesis. One important implication of the quantity $t(\beta_j)$ is that we can refer specifically to the hypothesis that $\beta_j=0$, called the *null hypothesis*. This quantity, $t_0 \stackrel{\mathrm{def}}{=} t_{\hat{\beta}_j}(0) = \frac{\hat{\beta}_j\sqrt{n-p}}{\|\boldsymbol{\zeta}\|\sqrt{(X^\mathsf{T}X)^{-1}_{j,j}}}$, represents how large is $\hat{\beta}_j$ relatively to the empirical estimation of standard deviation σ. Since it is known that as the number of degrees of freedom of a T-distribution tends to infinity then the T-distribution becomes a normal Gaussian, it is common to think of t_0 as a sample from a normal Gaussian $\mathcal{N}(0,1)$. This allows us to associate t_0 with a p-value, estimating the event "β_j and $\hat{\beta}_j$ have different signs." Formally, we define $p_0 = \int_{|t_0|}^\infty \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx$. It is common to reject the null hypothesis when p_0 is sufficiently small (typically, below 0.05).[9]

Specifically, given $\alpha\in(0,1/2)$, we say we *α-reject the null hypothesis* if $p_0<\alpha$. Let τ_α be the number s.t. $\Phi(\tau_\alpha) = \int_{\tau_\alpha}^\infty \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = \alpha$. (Standard bounds give that $\tau_\alpha < 2\sqrt{\ln(1/\alpha)}$.) This means we α-reject the null hypothesis if $t_0>\tau_\alpha$ or $t_0<-\tau_\alpha$, meaning if $|\hat{\beta}_j| > \tau_\alpha\sqrt{(X^\mathsf{T}X)^{-1}_{j,j}}\frac{\|\boldsymbol{\zeta}\|}{\sqrt{n-p}}$.

We can now lower bound the number of i.i.d sample points needed in order to α-reject the null hypothesis. This bound will be our basis for comparison — between standard OLS and the differentially private version.[10]

[8] Observe, though $\boldsymbol{e}$ is spherically symmetric, and is likely to be approximately-orthogonal to $\boldsymbol{\beta}$, this does not necessarily hold for $X^+\boldsymbol{e}$ which isn't spherically symmetric. Therefore, we result to bounding the l_2-norm of $\hat{\boldsymbol{\beta}}$ using the triangle bound.

[9] Indeed, it is more accurate to associate with t_0 the value $\int_{|t_0|}^\infty \mathsf{PDF}_{T_{n-p}}(x)dx$ and check that this value is $<\alpha$. However, as most uses take α to be a constant (often $\alpha=0.05$), asymptotically the threshold we get for rejecting the null hypothesis are the same.

[10] This theorem is far from being new (except for maybe fo-

Theorem A.5 (Theorem 2.2 restated.)**.** *Fix any positive definite matrix $\Sigma \in \mathbb{R}^{p\times p}$ and any $\nu \in (0, \frac{1}{2})$. Fix parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ and σ^2 and a coordinate j s.t. $\beta_j \neq 0$. Let X be a matrix whose n rows are i.i.d samples from $\mathcal{N}(\mathbf{0}, \Sigma)$, and $\boldsymbol{y}$ be a vector where $y_i - (X\boldsymbol{\beta})_i$ is sampled i.i.d from $\mathcal{N}(0, \sigma^2)$. Fix $\alpha \in (0, 1)$. Then w.p. $\geq 1 - \nu$ we have that the $(1 - \alpha)$-confidence interval is of length $O(c_\alpha\sqrt{\sigma^2/(n\sigma_{\min}(\Sigma))})$ provided $n \geq C_1(p + \ln(1/\nu))$ for some sufficiently large constant C_1. Furthermore, there exists a constant C_2 such that w.p. $\geq 1 - \alpha - \nu$ we (correctly) reject the null hypothesis provided*

$$n \geq \max\left\{C_1(p + \ln(1/\nu)),\ C_2\frac{\sigma^2}{\beta_j^2}\cdot\frac{c_\alpha^2 + \tau_\alpha^2}{\sigma_{\min}(\Sigma)}\right\}$$

Here c_α denotes the number for which $\int_{-c_\alpha}^{c_\alpha} \mathsf{PDF}_{T_{n-p}}(x)dx = 1 - \alpha$. (If we are content with approximating T_{n-p} with a normal Gaussian than one can set $c_\alpha \approx \tau_\alpha < 2\sqrt{\ln(1/\alpha)}$.)

Proof. The discussion above shows that w.p. $\geq 1 - \alpha$ we have $|\beta_j - \hat{\beta}_j| \leq c_\alpha\sqrt{(X^\mathsf{T}X)^{-1}_{j,j}\frac{\|\boldsymbol{\zeta}\|^2}{n-p}}$; and in order to α-reject the null hypothesis we must have $|\hat{\beta}_j| > \tau_\alpha\sqrt{(X^\mathsf{T}X)^{-1}_{j,j}\frac{\|\boldsymbol{\zeta}\|^2}{n-p}}$. Therefore, a sufficient condition for OLS to α-reject the null-hypothesis is to have n large enough s.t. $|\beta_j| > (c_\alpha + \tau_\alpha)\sqrt{(X^\mathsf{T}X)^{-1}_{j,j}\frac{\|\boldsymbol{\zeta}\|^2}{n-p}}$. We therefore argue that w.p.$\geq 1 - \nu$ this inequality indeed holds.

We assume each row of X i.i.d vector $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}_p, \Sigma)$, and recall that according to the model $\|\boldsymbol{\zeta}\|^2 \sim \sigma^2\chi^2(n - p)$. Straightforward concentration bounds on Gaussians and on the χ^2-distribution give:
(i) W.p. $\leq \alpha$ it holds that $\|\boldsymbol{\zeta}\| > \sigma\left(\sqrt{n-p} + 2\ln(2/\alpha)\right))$. (This is part of the standard OLS analysis.)
(ii) W.p. $\leq \nu$ it holds that $\sigma_{\min}(X^\mathsf{T}X) \leq \sigma_{\min}(\Sigma)(\sqrt{n} - (\sqrt{p} + \sqrt{2\ln(2/\nu)}))^2$. (Rudelson & Vershynin, 2009)
Therefore, due to the lower bound $n = \Omega(p + \ln(1/\nu))$, w.p.$\geq 1 - \nu - \alpha$ we have that none of these events hold. In such a case we have $\sqrt{(X^\mathsf{T}X)^{-1}_{j,j}} \leq \sqrt{\sigma_{\max}((X^\mathsf{T}X)^{-1})} = O(\frac{1}{\sqrt{n\sigma_{\min}(\Sigma)}})$ and $\|\boldsymbol{\zeta}\| = O(\sigma\sqrt{n-p})$. This implies that the confidence interval of level $1 - \alpha$ has length of $c_\alpha\sqrt{(X^\mathsf{T}X)^{-1}_{j,j}\cdot\frac{\|\boldsymbol{\zeta}\|^2}{n-p}} = O\left(c_\alpha\sqrt{\frac{\sigma^2}{n\sigma_{\min}(\Sigma)}}\right)$; and that in order to α-reject that null-hypothesis it suffices to have $|\beta_j| = \Omega\left((c_\alpha + \tau_\alpha)\sqrt{\frac{\sigma^2}{n\sigma_{\min}(\Sigma)}}\right)$. Plugging in the lower bound on n, we see that this inequality holds.

We comment that for sufficiently large constants C_1, C_2, it holds that all the constants hidden in the O- and Ω-notations of the proof are close to 1. I.e., they are all within the interval $(1 \pm \eta)$ for some small $\eta > 0$ given $C_1, C_2 \in \Omega(\eta^{-2})$. □

cusing on the setting where every row in X is sampled from an i.i.d multivariate Gaussians), it is just stated in a non-standard way, discussing solely the power of the t-test in OLS. For further discussions on sample size calculations see (Muller & Stewart, 2006).

B. Projecting the Data using Gaussian Johnson-Lindenstrauss Transform

B.1. Main Theorem Restated and Further Discussion

Theorem B.1 (Theorem 3.1 restated.)**.** *Let X be a $n \times p$ matrix, and parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ and σ^2 are such that we generate the vector $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$ with each coordinate of $\boldsymbol{e}$ sampled independently from $\mathcal{N}(0, \sigma^2)$. Assume $\sigma_{\min}(X) \geq C \cdot w$ and that n is sufficiently large s.t. all of the singular values of the matrix $[X; \boldsymbol{y}]$ are greater than $C \cdot w$ for some large constant C, and so Algorithm 1 projects the matrix $A = [X; \boldsymbol{y}]$ without altering it, and publishes $[RX; R\boldsymbol{y}]$.*
Fix $\nu \in (0, 1/2)$ and $r = p + \Omega(\ln(1/\nu))$. Fix coordinate j. Then w.p. $\geq 1 - \nu$ we have that deriving $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\zeta}}$ and $\tilde{\sigma}^2$ as follows

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= (X^\mathsf{T}R^\mathsf{T}RX)^{-1}(RX)^\mathsf{T}(R\boldsymbol{y}) = \boldsymbol{\beta} + (RX)^+R\boldsymbol{e}\\ \tilde{\boldsymbol{\zeta}} &= \tfrac{1}{\sqrt{r}}R\boldsymbol{y} - \tfrac{1}{\sqrt{r}}(RX)\tilde{\boldsymbol{\beta}}\\ &= \tfrac{1}{\sqrt{r}}\left(I - (RX)(X^\mathsf{T}R^\mathsf{T}RX)^{-1}(RX)^\mathsf{T})\right)R\boldsymbol{e}\\ \tilde{\sigma}^2 &= \frac{r}{r-p}\|\tilde{\boldsymbol{\zeta}}\|^2\end{aligned}$$

then the pivot quantity

$$\tilde{t}(\beta_j) = \frac{\tilde{\beta}_j - \beta_j}{\tilde{\sigma}\sqrt{(X^\mathsf{T}R^\mathsf{T}RX)^{-1}_{j,j}}}$$

has a distribution $\mathcal{D}$ satisfying $e^{-a}\mathsf{PDF}_{T_{r-p}}(x) \leq \mathsf{PDF}_{\mathcal{D}}(x) \leq e^a\mathsf{PDF}_{T_{r-p}}(e^{-a}x)$ for any $x \in \mathbb{R}$, where we denote $a = \frac{r-p}{n-p}$.

Comparison with Existing Bounds. Sarlos' work (2006) utilizes the fact that when r, the numbers of rows in R, is large enough, then $\frac{1}{\sqrt{r}}R$ is a Johnson-Lindenstrauss matrix. Specifically, given r and $\nu \in (0, 1)$ we denote $\eta = \Omega(\sqrt{\frac{p\ln(p)\ln(1/\nu)}{r}})$, and so $r = O(\frac{p\ln(p)\ln(1/\nu)}{\eta^2})$. Let us denote $\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{z}}\frac{1}{r}\|RX\boldsymbol{z} - R\boldsymbol{y}\|^2$. In this setting, Sarlos' work (Sarlós, 2006) (Theorem 12(3)) guarantees that w.p. $\geq 1 - \nu$ we have $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_2 \leq \eta\|\boldsymbol{\zeta}\|/\sigma_{\min}(X) = O\left(\sqrt{\frac{p\log(p)\log(1/\nu)}{r\sigma_{\min}(X^\mathsf{T}X)}}\|\boldsymbol{\zeta}\|\right)$. Naïvely bounding $|\hat{\beta}_j - \tilde{\beta}_j| \leq \|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|$ and using the confidence interval for $\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j$ from Section A.3[11]

[11] Where we approximate c_α, the tail bound of the T_{n-p}-distribution with the tail bound on a Gaussian, i.e., use the approximation $c_\alpha \approx O(\sqrt{\ln(1/\alpha)})$.

gives a confidence interval of level $1-(\alpha+\nu)$ centered at $\tilde{\beta}_j$ with length of $O\left(\sqrt{\frac{p\ln(p)\log(1/\nu)}{r\sigma_{\min}(X^\mathsf{T}X)}}\|\boldsymbol{\zeta}\|\right)+O\left(\sqrt{(X^\mathsf{T}X)^{-1}_{j,j}\frac{\log(1/\alpha)}{n-p}}\|\boldsymbol{\zeta}\|\right)=O\left(\sqrt{\frac{p\ln(p)\log(1/\nu)+\log(1/\alpha)}{r\sigma_{\min}(X^\mathsf{T}X)}}\|\boldsymbol{\zeta}\|\right)$. This implies that our confidence interval has decreased its degrees of freedom from $n-p$ to roughly $r/p\ln(p)$, and furthermore, that it no longer depends on $(X^TX)^{-1}_{j,j}$ but rather on $1/\sigma_{\min}(X^\mathsf{T}X)$. It is only due to the fact that we rely on Gaussians and by mimicking carefully the original proof that we can deduce that the $\tilde{t}$-value has (roughly) $r-p$ degrees of freedom and depends solely on $(X^TX)^{-1}_{j,j}$.

(In the worst case, we have that $(X^\mathsf{T}X)^{-1}_{j,j}$ is proportional to $\sigma_{\min}(X^\mathsf{T}X)^{-1}$, but it is not uncommon to have matrices where the former is much larger than the latter.) As mentioned in the introduction, alternative techniques ((Chaudhuri et al., 2011; Bassily et al., 2014; Ullman, 2015)) for finding a DP estimator $\boldsymbol{\beta}^{dp}$ of the linear regression give a data-independent[12] bound of $\|\boldsymbol{\beta}^{dp}-\hat{\boldsymbol{\beta}}\|=\tilde{O}(p/\epsilon)$. Such bounds are harder to compare with the interval length given by Corollary 3.2. Indeed, as we discuss in Section 3 under "Rejecting the null-hypothesis," enough samples from a multivariate Gaussian whose covariance-matrix is well conditioned give a bound which is well below the worst-upper bound of $O(p/\epsilon)$. (Yet, it is possible that these techniques also do much better on such "well-behaved" data.) What the works of Sarlos and alternative works regrading differentially private linear regression do not take into account are questions such as generating a likelihood for β_j nor do they discuss rejecting the null hypothesis.

B.2. Proof of Theorem 3.1

We now turn to our analysis of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\zeta}}$, where our goal is to show that the distribution of the $\tilde{t}$-values as specified in Theorem 3.1 is well-approximated by the T_{r-p}-distribution. For now, we assume the existence of fixed vectors $\boldsymbol{\beta}\in\mathbb{R}^p$ and $\boldsymbol{e}\in\mathbb{R}^n$ s.t. $\boldsymbol{y}=X\boldsymbol{\beta}+\boldsymbol{e}$. (Later, we will return to the homoscedastic model where each coordinate of $\boldsymbol{e}$ is sampled i.i.d from $\mathcal{N}(0,\sigma^2)$ for some σ^2.) In other words, we first examine the case where R is the sole source of randomness in our estimation. Based on the assumption that $\boldsymbol{e}$ is fixed, we argue the following.

Claim B.2. *In our model, given X and the output $M=RX$, we have that $\tilde{\boldsymbol{\beta}}\sim\mathcal{N}\left(\boldsymbol{\beta}+X^+\boldsymbol{e},\|P_{U^\perp}\boldsymbol{e}\|^2(M^\mathsf{T}M)^{-1}\right)$ and $\tilde{\boldsymbol{\zeta}}\sim\mathcal{N}\left(\mathbf{0}_n,\frac{\|P_{U^\perp}\boldsymbol{e}\|^2}{r}(I_{r\times r}-M(M^\mathsf{T}M)^{-1}M^\mathsf{T})\right)$. Where $P_{U^\perp}$ denotes the projection operator onto the subspace orthogonal to colspan(X); i.e., $P_U=XX^+$ and $P_{U^\perp}=(I_{r\times r}-XX^+)$.*

[12] In other words, independent of $X,\boldsymbol{\zeta}$.

Proof. The matrix R is sampled from $\mathcal{N}(0_{r\times p},I_{r\times r},I_{p\times p})$. Given X and $RX=M$, we learn the projection of each row in R onto the subspace spanned by the columns of X. That is, denoting $\boldsymbol{u}^\mathsf{T}$ as the i-th row of R and $\boldsymbol{v}^\mathsf{T}$ as the i-th row of M, we have that $X^\mathsf{T}\boldsymbol{u}=\boldsymbol{v}$. Recall, initially $\boldsymbol{u}\sim\mathcal{N}(\mathbf{0}_n,I_{n\times n})$ – a spherically symmetric Gaussian. As a result, we can denote $\boldsymbol{u}=P_U\boldsymbol{u}\times P_{U^\perp}\boldsymbol{u}$ where the two projections are independent samples from $\mathcal{N}(\mathbf{0}_n,P_U)$ and $\mathcal{N}(\mathbf{0}_n,P_{U^\perp})$ resp. However, once we know that $\boldsymbol{v}=X^\mathsf{T}\boldsymbol{u}$ we have that $P_U\boldsymbol{u}=X(X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{u}=X(X^\mathsf{T}X)^{-1}\boldsymbol{v}$ so we learn $P_U\boldsymbol{u}$ exactly, whereas we get no information about $P_{U^\perp}$ so $P_{U^\perp}\boldsymbol{u}$ is still sampled from a Gaussian $\mathcal{N}(\mathbf{0}_n,P_{U^\perp})$. As we know for each row of R that $\boldsymbol{u}^\mathsf{T}P_U=\boldsymbol{v}^\mathsf{T}X^+$, we therefore have that

$$R=RP_U+RP_{U^\perp}=MX^++RP_{U^\perp}$$

where $RP_{U^\perp}\sim\mathcal{N}(0_{r\times n},I_{r\times r},P_{U^\perp})$. From here on, we just rely on the existing results about the linearity of Gaussians.

$$\begin{aligned}&R\sim\mathcal{N}(MX^+,I_{r\times r},P_{U^\perp})\\&\Rightarrow R\boldsymbol{e}\sim\mathcal{N}(MX^+\boldsymbol{e},\|P_{U^\perp}\boldsymbol{e}\|^2I_{r\times r})\\&\Rightarrow M^+R\boldsymbol{e}\sim\mathcal{N}(X^+\boldsymbol{e},\|P_{U^\perp}\boldsymbol{e}\|^2(M^\mathsf{T}M)^{-1})\end{aligned}$$

so $\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}+M^+R\boldsymbol{e}$ implies $\tilde{\boldsymbol{\beta}}\sim\mathcal{N}(\boldsymbol{\beta}+X^+\boldsymbol{e},\|P_{U^\perp}\boldsymbol{e}\|^2(M^\mathsf{T}M)^{-1})$. And as $\tilde{\boldsymbol{\zeta}}=\frac{1}{\sqrt{r}}(I_{r\times r}-M(M^\mathsf{T}M)^{-1}M^\mathsf{T})R\boldsymbol{e}$ then we have $\tilde{\boldsymbol{\zeta}}\sim\mathcal{N}(\mathbf{0}_r,\frac{\|P_{U^\perp}\boldsymbol{e}\|^2}{r}(I_{r\times r}-MM^+))$ as $(I_{r\times r}-MM^+)M=0_{r\times p}$. □

Claim B.2 was based on the assumption that $\boldsymbol{e}$ is fixed. However, given X and $\boldsymbol{y}$ there are many different ways to assign vectors $\boldsymbol{\beta}$ and $\boldsymbol{e}$ s.t. $\boldsymbol{y}=X\boldsymbol{\beta}+\boldsymbol{e}$. However, the distributions we get in Claim B.2 are *unique*. To see that, recall Equations (1) and (2): $\boldsymbol{\beta}+X^+\boldsymbol{e}=X^+\boldsymbol{y}=\hat{\boldsymbol{\beta}}$ and $P_{U^\perp}\boldsymbol{e}=P_{U^\perp}\boldsymbol{y}=(I-XX^+)\boldsymbol{y}=\boldsymbol{\zeta}$. We therefore have $\tilde{\boldsymbol{\beta}}\sim\mathcal{N}(\hat{\boldsymbol{\beta}},\|\boldsymbol{\zeta}\|^2(M^\mathsf{T}M)^{-1})$ and $\tilde{\boldsymbol{\zeta}}\sim\mathcal{N}(\mathbf{0}_n,\frac{\|\boldsymbol{\zeta}\|^2}{r}(I-MM^+))$. We will discuss this further, in Section 4, where we will not be able to better analyze the explicit distributions of our estimators. But in this section, we are able to argue more about the distributions of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\zeta}}$.

So far we have considered the case that $\boldsymbol{e}$ is fixed, whereas our goal is to argue about the case where each coordinate of $\boldsymbol{e}$ is sampled i.i.d from $\mathcal{N}(0,\sigma^2)$. To that end, we now switch to an intermediate model, in which $P_U\boldsymbol{e}$ is sampled from a multivariate Gaussian while $P_{U^\perp}\boldsymbol{e}$ is fixed as some arbitrary vector of length l. Formally, let $\mathcal{D}_l$ denote the distribution where $P_U\boldsymbol{e}\sim\mathcal{N}(0,\sigma^2P_U)$ and $P_{U^\perp}\boldsymbol{e}$ is fixed as some specific vector whose length is denoted by $\|P_{U^\perp}\boldsymbol{e}\|=l$.

Claim B.3. *Under the same assumptions as in Claim B.2, given that $\boldsymbol{e}\sim\mathcal{D}_l$, we have that*

$\tilde{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1} + l^2(M^\mathsf{T}M)^{-1}\right)$ *and* $\tilde{\boldsymbol{\zeta}} \sim \mathcal{N}\left(\mathbf{0}_n, \frac{l^2}{r}(I - MM^+)\right)$.

Proof. Recall, $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + M^+R\boldsymbol{e} = \boldsymbol{\beta} + M^+(MX^+ + RP_{U^\perp})\boldsymbol{e} = \boldsymbol{\beta} + X^+\boldsymbol{e} + M^+R(P_{U^\perp}\boldsymbol{e})$. Now, under the assumption $\boldsymbol{e} \sim \mathcal{D}_l$ we have that β is the sum of two *independent* Gaussians:

$$\begin{aligned}
\boldsymbol{\beta} + X^+\boldsymbol{e} &\sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2\left(X^+ \cdot P_U \cdot (X^+)^\mathsf{T}\right)) \\
&= \mathcal{N}(\boldsymbol{\beta}, \sigma^2(X^\mathsf{T}X)^{-1}) \\
RP_{U^\perp}\boldsymbol{e} &\sim \mathcal{N}(\mathbf{0}_r, \|P_{U^\perp}\boldsymbol{e}\|^2 I_{r\times r}) \\
\Rightarrow M^+R\boldsymbol{e} &\sim \mathcal{N}(\mathbf{0}_p, \|P_{U^\perp}\boldsymbol{e}\|^2 (M^\mathsf{T}M)^{-1})
\end{aligned}$$

Summing the two independent Gaussians' means and variances gives the distribution of $\tilde{\boldsymbol{\beta}}$. Furthermore, in Claim B.2 we have already established that for any fixed $\boldsymbol{e}$ we have $\tilde{\boldsymbol{\zeta}} \sim \mathcal{N}\left(\mathbf{0}_n, \frac{\|P_{U^\perp}\boldsymbol{e}\|^2}{r}(I - MM^+)\right)$. Hence, for $\boldsymbol{e} \sim \mathcal{D}_l$ we still have $\tilde{\boldsymbol{\zeta}} \sim \mathcal{N}\left(\mathbf{0}_n, \frac{l^2}{r}(I - MM^+)\right)$. (It is easy to verify that the same chain of derivations is applicable when $\boldsymbol{e} \sim \mathcal{D}_l$.) □

Corollary B.4. *Given that $\boldsymbol{e} \sim \mathcal{D}_l$ we have that $\tilde{\beta}_j \sim \mathcal{N}(\beta_j,\ \sigma^2(X^\mathsf{T}X)^{-1}_{j,j} + l^2(M^\mathsf{T}M)^{-1}_{j,j})$ for any coordinate j, and that $\|\tilde{\boldsymbol{\zeta}}\|^2 \sim \frac{l^2}{r} \cdot \chi^2_{r-p}$.*

Proof. The corollary follows immediately from the fact that $\beta_j = \boldsymbol{e}_j^\mathsf{T}\tilde{\boldsymbol{\beta}}$, and from the definition of the χ^2-distribution, as $\tilde{\boldsymbol{\zeta}}$ is a spherically symmetric Gaussian defined on the subspace $colspan(M)^\perp$ of dimension $r - p$. □

To continue, we need the following claim.

Claim B.5. *Given X and $M = RX$, and given that $\boldsymbol{e} \sim \mathcal{D}_l$ we have that $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\zeta}}$ are independent.*

Proof. Recall, $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + X^+\boldsymbol{e} + M^+R(P_{U^\perp}\boldsymbol{e})$. And so, given X, M and a specific vector $P_{U^\perp}\boldsymbol{e}$ we have that the distribution of $\tilde{\boldsymbol{\beta}}$ depends on (i) the projection of $\boldsymbol{e}$ on $U = colspan(X)$ and on (ii) the projection of each row in R onto $\tilde{U} = colspan(M)$. The distribution of $\tilde{\boldsymbol{\zeta}} = \frac{1}{\sqrt{r}}P_{\tilde{U}^\perp}R\boldsymbol{e} = \frac{1}{\sqrt{r}}P_{\tilde{U}^\perp}(MX^+ + RP_{U^\perp})\boldsymbol{e} = \frac{1}{\sqrt{r}}P_{\tilde{U}^\perp}RP_{U^\perp}\boldsymbol{e}$ depends on (i) the projection of $\boldsymbol{e}$ onto $U^\perp$ (which for the time being is fix to some specific vector of length l) and on (ii) the projection of each row in R onto $\tilde{U}^\perp$. Since $P_U\boldsymbol{e}$ is independent from $P_{U^\perp}\boldsymbol{e}$, and since for any row $\boldsymbol{u}^\mathsf{T}$ of R we have that $P_{\tilde{U}}\boldsymbol{u}$ is independent of $P_{\tilde{U}^\perp}\boldsymbol{u}$, and since $\boldsymbol{e}$ and R are chosen independently, we have that $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\zeta}}$ are independent.

Formally, consider any pair of coordinates $\tilde{\beta}_j$ and $\tilde{\zeta}_k$, and we have

$$\begin{aligned}
\tilde{\beta}_j - \beta_j &= \boldsymbol{e}_j^\mathsf{T}X^+\boldsymbol{e} + \boldsymbol{e}_j^\mathsf{T}M^+(RP_{U^\perp}\boldsymbol{e}) \\
\tilde{\zeta}_k &= \boldsymbol{e}_k^\mathsf{T}P_{\tilde{U}^\perp}(RP_{U^\perp}\boldsymbol{e})
\end{aligned}$$

Recall, we are given X and $M = RX$. Therefore, we know P_U and $P_{\tilde{U}}$. And so

$$\begin{aligned}
&\mathrm{Cov}[\tilde{\beta}_j, \tilde{\zeta}_k] \\
&= \mathbf{E}[(\tilde{\beta}_j - \beta_j)(\tilde{\zeta}_k - 0)] \\
&= \mathbf{E}[\boldsymbol{e}_j^\mathsf{T}X^+\boldsymbol{e}(RP_{U^\perp}\boldsymbol{e})^\mathsf{T}P_{\tilde{U}^\perp}\boldsymbol{e}_k] \\
&\quad + \mathbf{E}[\boldsymbol{e}_j^\mathsf{T}M^+(RP_{U^\perp}\boldsymbol{e})(RP_{U^\perp}\boldsymbol{e})^\mathsf{T}P_{\tilde{U}^\perp}\boldsymbol{e}_k] \\
&= \boldsymbol{e}_j^\mathsf{T}X^+\mathbf{E}[\boldsymbol{e}\boldsymbol{e}^\mathsf{T}P_{U^\perp}]\mathbf{E}[R^\mathsf{T}]P_{\tilde{U}^\perp}\boldsymbol{e}_k \\
&\quad + \boldsymbol{e}_j^\mathsf{T}M^+\mathbf{E}[(RP_{U^\perp}\boldsymbol{e})(RP_{U^\perp}\boldsymbol{e})^\mathsf{T}]P_{\tilde{U}^\perp}\boldsymbol{e}_k \\
&= \boldsymbol{e}_j^\mathsf{T}X^+\mathbf{E}[\boldsymbol{e}\boldsymbol{e}^\mathsf{T}P_{U^\perp}]\left((MX^+)^\mathsf{T} + \mathbf{E}[(RP_{U^\perp})^\mathsf{T}]\right)P_{\tilde{U}^\perp}\boldsymbol{e}_k \\
&\quad + \boldsymbol{e}_j^\mathsf{T}M^+\left(\|P_{U^\perp}\boldsymbol{e}\|^2 I_{r\times r}\right)P_{\tilde{U}^\perp}\boldsymbol{e}_k \\
&= \boldsymbol{e}_j^\mathsf{T}X^+\mathbf{E}[\boldsymbol{e}\boldsymbol{e}^\mathsf{T}P_{U^\perp}](X^+)^\mathsf{T}\left(M^\mathsf{T}P_{\tilde{U}^\perp}\right)\boldsymbol{e}_k + 0 \\
&\quad + l^2 \cdot \boldsymbol{e}_j^\mathsf{T}\left(M^+P_{\tilde{U}^\perp}\right)\boldsymbol{e}_k \\
&= 0 + 0 + 0 = 0
\end{aligned}$$

And as $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\zeta}}$ are Gaussians, having their covariance $= 0$ implies independence. □

Having established that $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\zeta}}$ are independent Gaussians and specified their distributions, we continue with the proof of Theorem 3.1. We assume for now that there exists some small $a > 0$ s.t.

$$\begin{aligned}
l^2(M^\mathsf{T}M)^{-1}_{j,j} &\le \sigma^2(X^TX)^{-1}_{j,j} + l^2(M^\mathsf{T}M)^{-1}_{j,j} \\
&\le e^{2a} \cdot l^2(M^\mathsf{T}M)^{-1}_{j,j} \qquad (7)
\end{aligned}$$

Then, due to Corollary A.3, denoting the distributions $\mathcal{N}_1 = \mathcal{N}(0,\ l^2(M^\mathsf{T}M)^{-1}_{j,j})$ and $\mathcal{N}_2 = \mathcal{N}(0,\ \sigma^2(X^\mathsf{T}X)^{-1}_{j,j} + l^2(M^\mathsf{T}M)^{-1}_{j,j})$, we have that for any $S \subset \mathbb{R}$ it holds that[13]

$$e^{-a}\mathbf{Pr}_{\tilde{\beta}_j\sim\mathcal{N}_1}[S] \le \mathbf{Pr}_{\tilde{\beta}_j\sim\mathcal{N}_2}[S] \le e^a\mathbf{Pr}_{\tilde{\beta}_j\sim\mathcal{N}_1}[S/e^a] \qquad (8)$$

More specifically, denote the function

$$\begin{aligned}
\tilde{t}(\psi, \|\boldsymbol{\xi}\|, \beta_j) &= \frac{\psi - \beta_j}{\|\boldsymbol{\xi}\|\sqrt{\frac{r}{r-p}(M^\mathsf{T}M)^{-1}_{j,j}}} \\
&= \frac{\psi - \beta_j}{l\sqrt{(M^\mathsf{T}M)^{-1}_{j,j}}} \Big/ \frac{\|\boldsymbol{\xi}\|\sqrt{\frac{r}{r-p}}}{l}
\end{aligned}$$

and observe that when we sample $\psi, \boldsymbol{\xi}$ independently s.t. $\psi \sim \mathcal{N}(\beta_j,\ l^2(M^\mathsf{T}M)^{-1}_{j,j})$ and $\|\boldsymbol{\xi}\|^2 \sim \frac{l^2}{r}\chi^2_{r-p}$ then $\tilde{t}(\psi, \|\boldsymbol{\xi}\|, \beta_j)$ is distributed like a T-distribution with $r - p$

[13] In fact, it is possible to use standard techniques from differential privacy, and argue a similar result — that the probabilities of any event that depends on some function $f(\beta_j)$ under $\beta_j \sim \mathcal{N}_1$ and under $\beta_j \sim \mathcal{N}_2$ are close in the differential privacy sense.

degrees of freedom. And so, for any $\tau > 0$ we have that under such way to sample $\psi, \boldsymbol{\xi}$ we have $\mathbf{Pr}[\tilde{t}(\psi, \|\boldsymbol{\xi}\|, \beta_j) > \tau] = 1 - \mathsf{CDF}_{T_{r-p}}(\tau)$.

For any $\tau \geq 0$ and for any non-negative real value z let S_z^τ denote the suitable set of values s.t.

$$\mathbf{Pr}_{\left\{\begin{smallmatrix}\psi \sim \mathcal{N}(\beta_j,\, l^2(M^\mathsf{T}M)^{-1}_{j,j}) \\ \|\boldsymbol{\xi}\|^2 \sim \frac{l^2}{r}\chi^2_{r-p}\end{smallmatrix}\right\}}[\tilde{t}(\psi, \|\boldsymbol{\xi}\|, \beta_j) > \tau] = \int_0^\infty \mathsf{PDF}_{\frac{l^2}{r}\chi^2_{r-p}}(z) \cdot \mathbf{Pr}_{\{\psi - \beta_j \sim \mathcal{N}(0,\, l^2(M^\mathsf{T}M)^{-1}_{j,j})\}}[S_z^\tau]\, dz$$

That is, $S_z^\tau = \left(\tau \cdot z\sqrt{\frac{r}{r-p}(M^\mathsf{T}M)^{-1}_{j,j}},\ \infty\right)$.

We now use Equation (8) (Since $\mathcal{N}(0,\ l^2(M^\mathsf{T}M)^{-1}_{j,j})$ is precisely $\mathcal{N}_1$) to deduce that

$$\begin{aligned}
&\mathbf{Pr}_{\left\{\begin{smallmatrix}\psi \sim \mathcal{N}(\beta_j,\, l^2(M^\mathsf{T}M)^{-1}_{j,j} + \sigma^2(X^\mathsf{T}X)^{-1}_{j,j}) \\ \|\boldsymbol{\xi}\|^2 \sim \frac{l^2}{r}\chi^2_{r-p}\end{smallmatrix}\right\}}[\tilde{t}(\psi, \|\boldsymbol{\xi}\|, \beta_j) > \tau] \\
&= \int_0^\infty \mathsf{PDF}_{\frac{l^2}{r}\chi^2_{r-p}}(z) \mathop{\mathbf{Pr}}_{\psi - \beta_j \sim \mathcal{N}(0,\, l^2(M^\mathsf{T}M)^{-1}_{j,j} + \sigma^2(X^\mathsf{T}X)^{-1}_{j,j})}[S_z^\tau] dz \\
&\leq e^a \int_0^\infty \mathsf{PDF}_{\frac{l^2}{r}\chi^2_{r-p}}(z) \mathop{\mathbf{Pr}}_{\psi - \beta_j \sim \mathcal{N}(0,\, l^2(M^\mathsf{T}M)^{-1}_{j,j})}[S_z^\tau / e_a] dz \\
&\overset{(*)}{=} e^a \int_0^\infty \mathsf{PDF}_{\frac{l^2}{r}\chi^2_{r-p}}(z) \mathop{\mathbf{Pr}}_{\psi - \beta_j \sim \mathcal{N}(0,\, l^2(M^\mathsf{T}M)^{-1}_{j,j})}[S_z^{\tau/e^a}] dz \\
&= e^a \mathbf{Pr}_{\left\{\begin{smallmatrix}\psi \sim \mathcal{N}(\beta_j,\, l^2(M^\mathsf{T}M)^{-1}_{j,j}) \\ \|\boldsymbol{\xi}\|^2 \sim \frac{l^2}{r}\chi^2_{r-p}\end{smallmatrix}\right\}}[\tilde{t}(\psi, \|\boldsymbol{\xi}\|, \beta_j) > \tau/e^a] \\
&= e^a \left(1 - \mathsf{CDF}_{T_{r-p}}(\tau/e^a)\right)
\end{aligned}$$

where the equality $(*)$ follows from the fact that $S_z^\tau / c = S_z^{\tau/c}$ for any $c > 0$, since it is a non-negative interval. Analogously, we can also show that

$$\begin{aligned}
&\mathbf{Pr}_{\left\{\begin{smallmatrix}\psi \sim \mathcal{N}(\beta_j,\, l^2(M^\mathsf{T}M)^{-1}_{j,j} + \sigma^2(X^\mathsf{T}X)^{-1}_{j,j}) \\ \|\boldsymbol{\xi}\|^2 \sim \frac{l^2}{r}\chi^2_{r-p}\end{smallmatrix}\right\}}[\tilde{t}(\psi, \|\boldsymbol{\xi}\|, \beta_j) > \tau] \\
&\quad \geq e^{-a} \mathbf{Pr}_{\left\{\begin{smallmatrix}\psi \sim \mathcal{N}(\beta_j,\, l^2(M^\mathsf{T}M)^{-1}_{j,j}) \\ \|\boldsymbol{\xi}\|^2 \sim \frac{l^2}{r}\chi^2_{r-p}\end{smallmatrix}\right\}}[\tilde{t}(\psi, \|\boldsymbol{\xi}\|, \beta_j) > \tau] \\
&= e^{-a} \left(1 - \mathsf{CDF}_{T_{r-p}}(\tau)\right)
\end{aligned}$$

In other words, we have just shown that for any interval $I = (\tau, \infty)$ with $\tau \geq 0$ we have that $\mathbf{Pr}_{\left\{\begin{smallmatrix}\psi \sim \mathcal{N}(\beta_j,\, l^2(M^\mathsf{T}M)^{-1}_{j,j} + \sigma^2(X^\mathsf{T}X)^{-1}_{j,j}) \\ \|\boldsymbol{\xi}\|^2 \sim \frac{l^2}{r}\chi^2_{r-p}\end{smallmatrix}\right\}}[\tilde{t}(\psi, \|\boldsymbol{\xi}\|, \beta_j) \in I]$ is lower bounded by $e^a \int_I \mathsf{PDF}_{T_{r-p}}(z)dz$ and upper bounded by $e^a \int_{I/e^a} \mathsf{PDF}_{T_{r-p}}(z)dz$. We can now repeat the same argument for $I = (\tau_1, \tau_2)$ with $0 \leq \tau_1 < \tau_2$ (using an analogous definition of $S_z^{\tau_1,\tau_2}$), and again for any $I = (\tau_1, \tau_2)$ with $\tau_1 < \tau_2 \leq 0$, and deduce that the PDF of the function $\tilde{t}(\psi, \|\boldsymbol{\xi}\|, \beta_j)$ at x — where we sample $\psi \sim \mathcal{N}(\beta_j,\ l^2(M^\mathsf{T}M)^{-1}_{j,j} + \sigma^2(X^\mathsf{T}X)^{-1}_{j,j})$ and $\|\boldsymbol{\xi}\|^2 \sim \frac{l^2}{r}\chi^2_{r-p}$ independently — lies in the range $\left(e^{-a}\mathsf{PDF}_{T_{r-p}}(x), e^a\mathsf{PDF}_{T_{r-p}}(x/e^a)\right)$. And so, using Corollary B.4 and Claim B.5, we have that when $\boldsymbol{e} \sim \mathcal{D}_l$, the distributions of $\tilde{\beta}_j$ and $\|\tilde{\boldsymbol{\zeta}}\|^2$ are precisely as stated above, and so we have that the distribution of $\tilde{t}(\beta_j) \overset{\text{def}}{=} \tilde{t}(\tilde{\beta}_j, \|\tilde{\boldsymbol{\zeta}}\|, \beta_j)$ has a PDF that at the point x is "sandwiched" between $e^{-a}\mathsf{PDF}_{T_{r-p}}(x)$ and $e^a\mathsf{PDF}_{T_{r-p}}(x/e^a)$.

Next, we aim to argue that this characterization of the PDF of $\tilde{t}(\beta_j)$ still holds when $e \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 I_{n\times n})$. It would be convenient to think of $\boldsymbol{e}$ as a sample in $\mathcal{N}(\mathbf{0}_n, \sigma^2 P_U) \times \mathcal{N}(\mathbf{0}_n, \sigma^2 P_{U^\perp})$. (So while in $\mathcal{D}_l$ we have $P_U\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 P_U)$ but $P_{U^\perp}\boldsymbol{e}$ is fixed, now both $P_U\boldsymbol{e}$ and $P_{U^\perp}\boldsymbol{e}$ are sampled from spherical Gaussians.) The reason why the above still holds lies in the fact that $\tilde{t}(\beta_j)$ does not depend on l. In more details:

$$\begin{aligned}
&\mathbf{Pr}_{\boldsymbol{e}\sim\mathcal{N}(\mathbf{0}_n,\sigma^2 I_{n\times n})}\left[\tilde{t}(\beta_j) \in I\right] \\
&= \int_{\boldsymbol{v}} \mathbf{Pr}_{\boldsymbol{e}\sim\mathcal{N}(\mathbf{0}_n,\sigma^2 I_{n\times n})}\left[\tilde{t}(\beta_j) \in I \mid P_{U^\perp}\boldsymbol{e} = \boldsymbol{v}\right] \mathsf{PDF}_{P_{U^\perp}\boldsymbol{e}}(\boldsymbol{v}) d\boldsymbol{v} \\
&= \int_{\boldsymbol{v}} \mathop{\mathbf{Pr}}_{\boldsymbol{e}\sim\mathcal{D}_l}\left[\tilde{t}(\beta_j) \in I \mid l = \|\boldsymbol{v}\|\right] \mathsf{PDF}_{P_{U^\perp}\boldsymbol{e}}(\boldsymbol{v}) d\boldsymbol{v} \\
&\leq \int_{\boldsymbol{v}} \left(e^a \int_{I/e^a} \mathsf{PDF}_{T_{r-p}}(z)dz\right) \mathsf{PDF}_{P_{U^\perp}\boldsymbol{e}}(\boldsymbol{v}) d\boldsymbol{v} \\
&= \left(e^a \int_{I/e^a} \mathsf{PDF}_{T_{r-p}}(z)dz\right) \int_{\boldsymbol{v}} \mathsf{PDF}_{P_{U^\perp}\boldsymbol{e}}(\boldsymbol{v}) d\boldsymbol{v} \\
&= e^a \int_{I/e^a} \mathsf{PDF}_{T_{r-p}}(z)dz
\end{aligned}$$

where the last transition is possible precisely because $\tilde{t}$ is independent of l (or $\|\boldsymbol{v}\|$) — which is precisely what makes this t-value a pivot quantity. The proof of the lower bound is symmetric.

To conclude, we have shown that if Equation (7) holds, then for every interval $I \subset \mathbb{R}$ we have that $\mathbf{Pr}_{\boldsymbol{e}\sim\mathcal{N}(\mathbf{0}_n,\sigma^2 I_{n\times n})}\left[\tilde{t}(\beta_j) \in I\right]$ is lower bounded by $e^{-a}\mathbf{Pr}_{z\sim T_{r-p}}[z \in I]$ and upper bounded by $e^a\mathbf{Pr}_{z\sim T_{r-p}}[z \in (I/e^a)]$. So to conclude the proof of Theorem 3.1, we need to show that w.h.p such a as in Equation (7) exists.

Claim B.6. *In the homoscedastic model with Gaussian noise, if both n and r satisfy $n, r \geq p + \Omega(\log(1/\nu))$, then we have that $\sigma^2(X^TX)^{-1}_{j,j} + l^2(M^\mathsf{T}M)^{-1}_{j,j} \geq l^2(M^\mathsf{T}M)^{-1}_{j,j}$ and*

$$\sigma^2(X^TX)^{-1}_{j,j} + l^2(M^\mathsf{T}M)^{-1}_{j,j} \leq (1 + \tfrac{2(r-p)}{n-p}) \cdot l^2(M^\mathsf{T}M)^{-1}_{j,j}$$

Using $(1 + \frac{2(r-p)}{n-p}) \leq e^{\frac{2(r-p)}{n-p}}$, Theorem 3.1 now follows from plugging $a = \frac{r-p}{n-p}$ to our above discussion.

Proof. The lower bound is immediate from non-negativity of σ^2 and of $(X^\mathsf{T}X)^{-1}_{j,j} = \|(X^\mathsf{T}X)^{-1/2}\boldsymbol{e}_j\|^2$. We therefore prove the upper bound.

First, observe that $l^2 = \|P_{U^\perp}\boldsymbol{e}\|^2$ is sampled from $\sigma^2 \cdot \chi^2_{n-p}$ as $U^\perp$ is of dimension $n-p$. Therefore, it holds that w.p. $\geq 1-\nu/2$ that

$$\sigma^2\left(\sqrt{n-p} - \sqrt{2\ln(2/\nu)}\right)^2 \leq l^2$$

and assuming $n > p + 100\ln(2/\nu)$ we therefore have $\sigma^2 \leq \frac{4}{3(n-p)}l^2$.

Secondly, we argue that when $r > p + 300\ln(4/\nu)$ we have that w.p. $\geq 1-\nu/2$ it holds that $\frac{3}{4}(X^\mathsf{T}X)^{-1}_{j,j} \leq (r-p)(X^\mathsf{T}R^\mathsf{T}RX)^{-1}_{j,j}$. To see this, first observe that by picking $R \sim \mathcal{N}(0_{r\times n}, I_{r\times r}, I_{n\times n})$ the distribution of the product $RX \sim \mathcal{N}(0_{r\times d}, I_{r\times r}, X^\mathsf{T}X)$ is identical to picking $Q \sim \mathcal{N}(0_{r\times d}, I_{r\times r}, I_{d\times d})$ and taking the product $Q(X^\mathsf{T}X)^{1/2}$. Therefore, the distribution of $(X^\mathsf{T}R^TRX)^{-1}$ is identical to $\left((X^\mathsf{T}X)^{1/2}Q^\mathsf{T}Q(X^\mathsf{T}X)^{1/2}\right)^{-1} = (X^\mathsf{T}X)^{-1/2}(Q^\mathsf{T}Q)^{-1}(X^\mathsf{T}X)^{-1/2}$. Denoting $\boldsymbol{v} = (X^\mathsf{T}X)^{-1/2}\boldsymbol{e}_j$ we have $\|\boldsymbol{v}\|^2 = (X^\mathsf{T}X)^{-1}_{j,j}$. Claim A.1 from (Sheffet, 2015) gives that w.p. $\geq 1-\nu/2$ we have

$$\begin{aligned}(r-p)\cdot \boldsymbol{e}_j^\mathsf{T}&\left((X^\mathsf{T}X)^{1/2}Q^\mathsf{T}Q(X^\mathsf{T}X)^{1/2}\right)^{-1}\boldsymbol{e}_j \\ &= \boldsymbol{v}^\mathsf{T}(\tfrac{1}{r-p}Q^\mathsf{T}Q)^{-1}\boldsymbol{v} \geq \tfrac{3}{4}\boldsymbol{v}^\mathsf{T}\boldsymbol{v} = \tfrac{3}{4}(X^\mathsf{T}X)^{-1}_{j,j}\end{aligned}$$

which implies the required.

Combining the two inequalities we get:

$$\begin{aligned}\sigma^2(X^\mathsf{T}X)^{-1}_{j,j} &\leq \tfrac{16l^2(r-p)}{n-p}(X^\mathsf{T}R^\mathsf{T}RX)^{-1}_{j,j} \\ &\leq \tfrac{2(r-p)}{n-p}l^2(X^\mathsf{T}R^\mathsf{T}RX)^{-1}_{j,j}\end{aligned}$$

and as we denote $M = RX$ we are done. □

We comment that our analysis in the proof of Claim B.6 implicitly assumes $r \ll n$ (as we do think of the projection R as dimensionality reduction), and so the ratio $\frac{r-p}{n-p}$ is small. However, a similar analysis holds for r which is comparable to n — in which we would argue that $\frac{\sigma^2(X^TX)^{-1}_{j,j}+l^2(M^\mathsf{T}M)^{-1}_{j,j}}{\sigma^2(X^TX)^{-1}} \in [1, 1+\eta]$ for some small η.

B.3. Proof of Theorem 3.3

Theorem B.7 (Theorem 3.3 restated.)**.** *Fix a positive definite matrix $\Sigma \in \mathbb{R}^{p\times p}$. Fix parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\sigma^2 > 0$ and a coordinate j s.t. $\beta_j \neq 0$. Let X be a matrix whose n rows are sampled i.i.d from $\mathcal{N}(\mathbf{0}_p, \Sigma)$. Let $\boldsymbol{y}$ be a vector s.t. $y_i - (X\boldsymbol{\beta})_i$ is sampled i.i.d from $\mathcal{N}(0,\sigma^2)$. Fix $\nu \in (0,1/2)$ and $\alpha \in (0,1/2)$. Then there exist constants C_1, C_2, C_3 and C_4 such that when we run Algorithm 1 over $[X;\boldsymbol{y}]$ with parameter r w.p. $\geq 1-\nu$ we correctly α-reject the null hypothesis using $\tilde{p}_0$ (i.e., w.p. $\geq 1-\nu$ Algorithm 1 returns* `matrix unaltered` *and we can estimate $\tilde{t}_0$ and verify that indeed $\tilde{p}_0 < \alpha \cdot e^{-\frac{r-p}{n-p}}$) provided*

$$r \geq p + \max\left\{C_1\frac{\sigma^2(\tilde{c}_\alpha^2 + \tilde{\tau}_\alpha^2)}{\beta_j^2\sigma_{\min}(\Sigma)},\ C_2\ln(1/\nu)\right\}$$

and

$$n \geq \max\left\{r,\ C_3\frac{w^2}{\min\{\sigma_{\min}(\Sigma),\sigma^2\}},\ C_4(p+\ln(1/\nu))\right\}$$

where $\tilde{c}_\alpha$, $\tilde{\tau}_\alpha$ denote the numbers s.t. $\int_{\tilde{c}_\alpha/e^{\frac{r-p}{n-p}}}^{\infty} \mathsf{PDF}_{T_{r-p}}(x)dx = \frac{\alpha}{2}e^{-\frac{r-p}{n-p}}$ and $\int_{\tilde{\tau}_\alpha/e^{\frac{r-p}{n-p}}}^{\infty} \mathsf{PDF}_{\mathcal{N}(0,1)}(x)dx = \frac{\alpha}{2}e^{-\frac{r-p}{n-p}}$ resp.

Proof. First we need to use the lower bound on n to show that indeed Algorithm 1 does not alter A, and that various quantities are not far from their expected values. Formally, we claim the following.

Proposition B.8. *Under the same lower bounds on n and r as in Theorem 3.3, w.p. $1-\alpha-\nu$ we have that Theorem 3.1 holds and also that*

$$\tilde{\boldsymbol{\zeta}}\|^2 = \Theta(\tfrac{r-p}{r}\|P_{U^\perp}\boldsymbol{e}\|^2) = \Theta(\tfrac{r-p}{r}(n-p)\sigma^2)$$

and

$$(X^\mathsf{T}R^\mathsf{T}RX)^{-1}_{j,j} = \Theta(\tfrac{1}{r-p}(X^\mathsf{T}X)^{-1}_{j,j})$$

Proof of Proposition B.8. First, we need to argue that we have enough samples as to have the gap $\sigma^2_{\min}([X;y]) - w^2$ sufficiently large.

Since $\boldsymbol{x}_i \sim \mathcal{N}(0,\Sigma)$, and $y_i = \boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_i + e_i$ with $e_i \sim \mathcal{N}(0,\sigma^2)$, we have that the concatenation $(\boldsymbol{x}_i \circ y_i)$ is also sampled from a Gaussian. Clearly, $\mathbf{E}[y_i] = \boldsymbol{\beta}^\mathsf{T}\mathbf{E}[\boldsymbol{x}_i] + \mathbf{E}[e_i] = 0$. Similarly, $\mathbf{E}[x_{i,j}y_i] = \mathbf{E}[x_{i,j}\cdot(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_i + e_i)] = (\Sigma\boldsymbol{\beta})_j$ and $\mathbf{E}[y_i^2] = \mathbf{E}[e_i^2] + \mathbf{E}[\|X\boldsymbol{\beta}\|^2] = \sigma^2 + \mathbf{E}[\boldsymbol{\beta}^\mathsf{T}X^\mathsf{T}X\boldsymbol{\beta}] = \sigma^2 + \boldsymbol{\beta}^\mathsf{T}\Sigma\boldsymbol{\beta}$. Therefore, each row of A is an i.i.d sample of $\mathcal{N}(\mathbf{0}_{p+1}, \Sigma_A)$, with

$$\Sigma_A = \left(\begin{array}{c|c} \Sigma & \Sigma\boldsymbol{\beta} \\ \hline \boldsymbol{\beta}^\mathsf{T}\Sigma & \sigma^2+\boldsymbol{\beta}^\mathsf{T}\Sigma\boldsymbol{\beta}\end{array}\right)$$

Denote $\lambda^2 = \sigma_{\min}(\Sigma)$. Then, to argue that $\sigma_{\min}(\Sigma_A)$ is large we use the lower bound from (Ma & Zarowski, 1995) (Theorem 3.1) combining with some simple arithmetic manipulations to deduce that $\sigma_{\min}(\Sigma_A) \geq \min\{\sigma_{\min}(\Sigma),\sigma^2\}$.

Having established a lower bound on $\sigma_{\min}(\Sigma_A)$, it follows that with $n = \Omega(p\ln(1/\nu))$ i.i.d draws from $\mathcal{N}(\mathbf{0}_{p+1}, \Sigma_A)$ we have w.p. $\leq \nu/4$ that $\sigma_{\min}(A^\mathsf{T}A) = o(n)\cdot\min\{\sigma_{\min}(\Sigma), \sigma^2\}$. Conditioned on $\sigma_{\min}(A^\mathsf{T}A) = \Omega(n\sigma_{\min}(\Sigma_A)) = \Omega(w^2)$ being large enough, we have that w.p. $\leq \nu/4$ over the randomness of Algorithm 1 the matrix A does not pass the if-condition and the output of the algorithm is not RA. Conditioned on Algorithm 1 outputting RA, and due to the lower bound $r = p + \Omega(\ln(1/\nu))$, we have that the result of Theorem 3.1 does not hold w.p. $\leq \alpha + \nu/4$. All in all we deduce that w.p. $\geq 1 - \alpha - 3\nu/4$ the result of Theorem 3.1 holds. And since we argue Theorem 3.1 holds, then the following two bounds that are used in the proof[14] also hold:

$$(X^\mathsf{T}R^\mathsf{T}RX)^{-1}_{j,j} = \Theta(\tfrac{1}{r-p}(X^\mathsf{T}X)^{-1}_{j,j})$$
$$\|P_{U^\perp}\boldsymbol{e}\|^2 = \Theta((n-p)\sigma^2)$$

Lastly, in the proof of Theorem 3.1 we argue that for a given $P_{U^\perp}\boldsymbol{e}$ the length $\|\tilde{\boldsymbol{\zeta}}\|^2$ is distributed like $\frac{\|P_{U^\perp}\boldsymbol{e}\|^2}{r}\chi^2_{r-p}$. Appealing again to the fact that $r = p + \Omega(\ln(1/\nu)$ we have that w.p. $\geq \nu/4$ it holds that $\|\tilde{\boldsymbol{\zeta}}\|^2 > 2(r-p)\frac{\|P_{U^\perp}\boldsymbol{e}\|^2}{r}$. Plugging in the value of $\|P_{U^\perp}\boldsymbol{e}\|^2$ concludes the proof of the proposition. □

Based on Proposition B.8, we now show that we indeed reject the null-hypothesis (as we should). When Theorem 3.1 holds, reject the null-hypothesis iff $\tilde{p}_0 < \alpha \cdot e^{-\frac{r-p}{n-p}}$ which holds iff $|\tilde{t}_0| > e^{\frac{r-p}{n-p}}\tilde{\tau}_\alpha$. This implies we reject that null-hypothesis when $|\tilde{\beta}_j| > e^{\frac{r-p}{n-p}}\tilde{\tau}_\alpha \cdot \tilde{\sigma}\sqrt{(X^\mathsf{T}R^\mathsf{T}RX)^{-1}_{j,j}})$. Note that this bound is based on Corollary 3.2 that determines that $|\tilde{\beta}_j - \beta_j| = O\left(e^{\frac{r-p}{n-p}}\tilde{c}_\alpha \cdot \tilde{\sigma}\sqrt{(X^\mathsf{T}R^\mathsf{T}RX)^{-1}_{j,j}}\right)$. And so we have that w.p. $\geq 1-\nu$ we α-reject the null hypothesis when it holds that $|\beta_j| > 3(\tilde{c}_\alpha + \tilde{\tau}_\alpha)\cdot\tilde{\sigma}\sqrt{(X^\mathsf{T}R^\mathsf{T}RX)^{-1}_{j,j}}) \geq e^{\frac{r-p}{n-p}}(\tilde{c}_\alpha + \tilde{\tau}_\alpha)\tilde{\sigma}\sqrt{(X^\mathsf{T}R^\mathsf{T}RX)^{-1}_{j,j}}$ (due to the lower bound $n \geq r$).

Based on the bounds stated above we have that

$$\tilde{\sigma} = \|\tilde{\boldsymbol{\zeta}}\|\sqrt{\tfrac{r}{r-p}} = \Theta(\sigma\sqrt{n-p}\sqrt{\tfrac{r-p}{r}}\sqrt{\tfrac{r}{r-p}}) = \Theta(\sigma\sqrt{n-p})$$

and that

$$(X^\mathsf{T}R^\mathsf{T}RX)^{-1}_{j,j} = \Theta(\tfrac{1}{r-p}(X^\mathsf{T}X)^{-1}_{j,j}) = O\left(\tfrac{1}{r-p}\cdot\tfrac{1}{n\sigma_{\min}(\Sigma)}\right)$$

And so, a sufficient condition for rejecting the null-hypothesis is to have

$$\begin{aligned}|\beta_j| &= \Omega\left((\tilde{c}_\alpha + \tilde{\tau}_\alpha)\sigma\sqrt{\frac{n-p}{r-p}}\cdot\sqrt{\tfrac{1}{n\sigma_{\min}(\Sigma)}}\right)\\ &= \Omega(e^{\frac{r-p}{n-p}}(\tilde{c}_\alpha + \tilde{\tau}_\alpha)\tilde{\sigma}\sqrt{(X^\mathsf{T}R^\mathsf{T}RX)^{-1}_{j,j}}))\end{aligned}$$

which, given the lower bound $r = p + \Omega\left(\frac{(\tilde{c}_\alpha+\tilde{\tau}_\alpha)^2\sigma^2}{\beta_j^2\sigma_{\min}(\Sigma)}\right)$ indeed holds. □

[14] More accurately, both are bounds shown in Claim B.6.

C. Projected Ridge Regression

In this section we deal with the case that our matrix does not pass the if-condition of Algorithm 1. In this case, the matrix is appended with a $d \times d$-matrix which is $wI_{d\times d}$. Denoting $A' = \begin{bmatrix} A \\ w\cdot I_{d\times d}\end{bmatrix}$ we have that the algorithm's output is RA'.

Similarly to before, we are going to denote $d = p+1$ and decompose $A = [X; \boldsymbol{y}]$ with $X \in \mathbb{R}^{n\times p}$ and $\boldsymbol{y} \in \mathbb{R}^n$, with the standard assumption of $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$ and e_i sampled i.i.d from $\mathcal{N}(0,\sigma^2)$.[15] We now need to introduce some additional notation. We denote the appended matrix and vectors X' and $\boldsymbol{y}'$ s.t. $A' = [X'; \boldsymbol{y}']$. Meaning:

$$X' = \begin{bmatrix} X \\ wI_{p\times p} \\ \mathbf{0}_p^\mathsf{T}\end{bmatrix}$$

and

$$\boldsymbol{y}' = \begin{bmatrix}\boldsymbol{y}\\ \mathbf{0}_p \\ w\end{bmatrix} = X'\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{e}\\ -w\boldsymbol{\beta}\\ w\end{bmatrix} \stackrel{\text{def}}{=} X'\boldsymbol{\beta} + \boldsymbol{e}'$$

And so we respectively denote $R = [R_1; R_2; R_3]$ with $R_1 \in \mathbb{R}^{r\times n}$, $R_2 \in \mathbb{R}^{r\times p}$ and $R_3 \in \mathbb{R}^{r\times 1}$ (so R_3 is a vector denoted as a matrix). Hence:

$$M' = RX' = R_1X + wR_2$$

and

$$R\boldsymbol{y}' = RX'\boldsymbol{\beta} + R\boldsymbol{e}' = R_1\boldsymbol{y} + wR_3 = R_1X\boldsymbol{\beta} + R_1\boldsymbol{e} + wR_3$$

And so, using the output RA' of Algorithm 1, we solve the linear regression problem derived from $\frac{1}{\sqrt{r}}RX'$ and $\frac{1}{\sqrt{r}}R\boldsymbol{y}'$. I.e., we set

$$\begin{aligned}\boldsymbol{\beta}' &= \arg\min_{\boldsymbol{z}} \tfrac{1}{r}\|R\boldsymbol{y}' - RX'\boldsymbol{z}\|^2\\ &= (X'^\mathsf{T}R^\mathsf{T}RX')^{-1}(RX')^\mathsf{T}(R\boldsymbol{y}')\end{aligned}$$

Sarlos' results (2006) regarding the Johnson Lindenstrauss transform give that, when R has sufficiently many rows, solving the latter optimization problem gives a good approximation for the solution of the optimization problem

$$\boldsymbol{\beta}^R = \arg\min_{\boldsymbol{z}}\|\boldsymbol{y}' - X'\boldsymbol{z}\|^2 = \arg\min_{\boldsymbol{z}}\left(\|\boldsymbol{y} - X\boldsymbol{z}\|^2 + w^2\|\boldsymbol{z}\|^2\right)$$

[15] Just as before, it is possible to denote any single column as $\boldsymbol{y}$ and any subset of the remaining columns as X.

The latter problem is known as the Ridge Regression problem. Invented in the 60s (Tikhonov, 1963; Hoerl & Kennard, 1970), Ridge Regression is often motivated from the perspective of penalizing linear vectors whose coefficients are too large. It is also often applied in the case where X doesn't have full rank or is close to not having full-rank. That is because the Ridge Regression problem is always solvable. One can show that the minimizer $\boldsymbol{\beta}^R = (X^\mathsf{T}X + w^2 I_{p\times p})^{-1}X^\mathsf{T}\boldsymbol{y}$ is the unique solution of the Ridge Regression problem and that the RHS is always defined (even when X is singular).

The original focus of Ridge Regression is on penalizing $\boldsymbol{\beta}^R$ for having large coefficients. Therefore, Ridge Regression actually poses a family of linear regression problems: $\min_{\boldsymbol{z}} \|y - X\boldsymbol{z}\| + \lambda\|\boldsymbol{z}\|^2$, where one may set λ to be any non-negative scalar. And so, much of the literature on Ridge Regression is devoted to the art of fine-tuning this penalty term — either empirically or based on the λ that yields the best risk: $\|\mathbf{E}[\boldsymbol{\beta}^R] - \boldsymbol{\beta}\|^2 + \text{Var}(\boldsymbol{\beta}^R)$.[16] Here we propose a fundamentally different approach for the choice of the normalization factor — we set it so that solution of the regression problem would satisfy (ϵ, δ)-differential privacy (by projecting the problem onto a lower dimension).

While the solution of the Ridge Regression problem might have smaller risk than the OLS solution, it is not known how to derive t-values and/or reject the null hypothesis under Ridge Regression (except for using X to manipulate $\boldsymbol{\beta}^R$ back into $\hat{\boldsymbol{\beta}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{y}$ and relying on OLS). In fact, prior to our work there was no need for such analysis! For confidence intervals one could just use the standard OLS, because access to X and $\boldsymbol{y}$ was given.

Therefore, much for the same reason, we are unable to derive t-values under projected Ridge Regression.[17] Clearly, there are situations where such confidence bounds simply cannot be derived. (Consider for example the case where $X = 0_{n\times p}$ and $\boldsymbol{y}$ is just i.i.d draws from $\mathcal{N}(0, \sigma^2)$, so obviously $[X; y]$ gives no information about $\boldsymbol{\beta}$.) Nonetheless, under additional assumptions about the data, our work can give confidence intervals for β_j, and in the case where the interval doesn't intersect the origin — assure us that $sign(\beta'_j) = sign(\beta_j)$ w.h.p.

Clearly, Sarlos' work (2006) gives an upper bound on the distance $\|\boldsymbol{\beta}' - \boldsymbol{\beta}^R\|$. However, such distance bound doesn't come with the coordinate by coordinate confidence guarantee we would like to have. In fact, it is not even clear from Sarlos' work that $\mathbf{E}[\boldsymbol{\beta}'] = \boldsymbol{\beta}^R$ (though it is obvious to see that $\mathbf{E}[(X'^\mathsf{T}R^\mathsf{T}RX')]\boldsymbol{\beta}^R = \mathbf{E}[(RX')^\mathsf{T}R\boldsymbol{y}']$). Here, we show that $\mathbf{E}[\boldsymbol{\beta}'] = \hat{\boldsymbol{\beta}}$ which, more often than not, does not equal $\boldsymbol{\beta}^R$.

[16] Ridge Regression, as opposed to OLS, does not yield an unbiased estimator. I.e., $\mathbf{E}[\boldsymbol{\beta}^R] \neq \boldsymbol{\beta}$.

[17] Note: The naïve approach of using RX' and $R\boldsymbol{y}'$ to interpolate RX and $R\boldsymbol{y}$ and then apply Theorem 3.1 using these estimations of RX and $R\boldsymbol{y}$ ignores the noise added from appending the matrix A into A', and it is therefore bound to produce inaccurate estimations of the t-values.

Comment about notation. Throughout this section we assume X is of full rank and so $(X^\mathsf{T}X)^{-1}$ is well-defined. If X isn't full-rank, then one can simply replace any occurrence of $(X^\mathsf{T}X)^{-1}$ with $X^+(X^+)^\mathsf{T}$. This makes all our formulas well-defined in the general case.

C.1. Running OLS on the Projected Data

In this section, we analyze the projected Ridge Regression, under the assumption (for now) that $\boldsymbol{e}$ is fixed. That is, for now we assume that the only source of randomness comes from picking the matrix $R = [R_1; R_2; R_3]$. As before, we analyze the distribution over $\boldsymbol{\beta}'$ (see Equation (9)), and the value of the function we optimize at $\boldsymbol{\beta}'$. Denoting $M' = RX'$, we can formally express the estimators:

$$\boldsymbol{\beta}' = (M'^\mathsf{T}M')^{-1}M'^\mathsf{T}R\boldsymbol{y}' \tag{9}$$

$$\boldsymbol{\zeta}' = \tfrac{1}{\sqrt{r}}(R\boldsymbol{y}' - RX'\boldsymbol{\beta}') \tag{10}$$

Claim C.1. *Given that $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$ for a fixed $\boldsymbol{e}$, and given X and $M' = RX' = R_1X + wR_2$ we have that*

$$\boldsymbol{\beta}' \sim \mathcal{N}\Big(\boldsymbol{\beta} + X^+\boldsymbol{e}, \\ (w^2(\|\boldsymbol{\beta} + X^+\boldsymbol{e}\|^2 + 1) + \|P_{U^\perp}\boldsymbol{e}\|^2)(M'^\mathsf{T}M')^{-1}\Big)$$

$$\boldsymbol{\zeta}' \sim \mathcal{N}\Big(\mathbf{0}_r, \\ \frac{w^2(\|\boldsymbol{\beta}+X^+\boldsymbol{e}\|^2+1)+\|P_{U^\perp}\boldsymbol{e}\|^2}{r}(I_{r\times r} - M'M'^+)\Big)$$

and furthermore, $\boldsymbol{\beta}'$ and $\boldsymbol{\zeta}'$ are independent of one another.

Proof. First, we write $\boldsymbol{\beta}'$ and $\boldsymbol{\zeta}'$ explicitly, based on $\boldsymbol{e}$ and projection matrices:

$$\begin{aligned}\boldsymbol{\beta}' &= (M'^\mathsf{T}M')^{-1}M'^\mathsf{T}R\boldsymbol{y}' \\ &= M'^+(R_1X)\boldsymbol{\beta} + M'^+(R_1\boldsymbol{e} + wR_3) \\ \boldsymbol{\zeta}' &= \tfrac{1}{\sqrt{r}}(R\boldsymbol{y}' - RX'\boldsymbol{\beta}') \\ &= \tfrac{1}{\sqrt{r}}(I_{r\times r} - M'M'^+)R\boldsymbol{e}' \\ &= \tfrac{1}{\sqrt{r}}P_{U'^\perp}(R_1\boldsymbol{e} - wR_2\boldsymbol{\beta} + wR_3)\end{aligned}$$

with U' denoting $colspan(M')$ and $P_{U'^\perp}$ denoting the projection onto the subspace $U'^\perp$.

Again, we break $\boldsymbol{e}$ into an orthogonal composition: $\boldsymbol{e} = P_U\boldsymbol{e} + P_{U^\perp}\boldsymbol{e}$ with $U = colspan(X)$ (hence $P_U = XX^+$) and $U^\perp = colspan(X)^\perp$. Therefore,

$$\begin{aligned}\boldsymbol{\beta}' &= M'^+(R_1X)\boldsymbol{\beta} + M'^+(R_1XX^+\boldsymbol{e} + R_1P_{U^\perp}\boldsymbol{e} + wR_3) \\ &= M'^+(R_1X)(\boldsymbol{\beta} + X^+\boldsymbol{e}) + M'^+(R_1P_{U^\perp}\boldsymbol{e} + wR_3)\end{aligned}$$

whereas $\boldsymbol{\zeta}'$ is essentially

$$\begin{aligned}
&\tfrac{1}{\sqrt{r}}(I_{r\times r} - M'M'^{+})(R_1XX^{+}\boldsymbol{e} + R_1P_{U^{\perp}}\boldsymbol{e} - wR_2\boldsymbol{\beta} + wR_3)\\
&\overset{(*)}{=} \tfrac{1}{\sqrt{r}}(I_{r\times r} - M'M'^{+})\cdot\\
&\qquad (R_1XX^{+}\boldsymbol{e} + R_1P_{U^{\perp}}\boldsymbol{e} + (M' - wR_2)\boldsymbol{\beta} + wR_3)\\
&= \tfrac{1}{\sqrt{r}}(I_{r\times r} - M'M'^{+})\cdot\\
&\qquad (R_1X(\boldsymbol{\beta} + X^{+}\boldsymbol{e}) + R_1P_{U^{\perp}}\boldsymbol{e} + wR_3)
\end{aligned}$$

where equality $(*)$ holds because $(I - M'M'^{+})M'\boldsymbol{v} = \mathbf{0}$ for any $\boldsymbol{v}$.

We now aim to describe the distribution of R given that we know X' and $M' = RX'$. Since

$$\begin{aligned}
M' &= R_1X + wR_2 + 0\cdot R_3 = R_1X(X^{+}X) + wR_2\\
&= (R_1P_U)X + wR_2
\end{aligned}$$

then M' is independent of R_3 and independent of $R_1P_{U^{\perp}}$. Therefore, given X and M' the induced distribution over R_3 remains $R_3 \sim \mathcal{N}(\mathbf{0}_r, I_{r\times r})$, and similarly, given X and M' we have $R_1P_{U^{\perp}} \sim \mathcal{N}(0_{r\times n}, I_{r\times r}, P_{U^{\perp}})$ (rows remain independent from one another, and each row is distributed like a spherical Gaussian in $colspan(X)^{\perp}$). And so, we have that $R_1X = R_1P_UX = M' - wR_2$, which in turn implies:

$$R_1X \sim \mathcal{N}\left(M',\ \ I_{r\times r},\ \ w^2\cdot I_{p\times p}\right)$$

multiplying this random matrix with a vector, we get

$$R_1X(\boldsymbol{\beta}+X^{+}\boldsymbol{e}) \sim \mathcal{N}\left(M'\boldsymbol{\beta} + M'X^{+}\boldsymbol{e},\ \ w^2\|\boldsymbol{\beta} + X^{+}\boldsymbol{e}\|^2 I_{r\times r}\right)$$

and multiplying this random vector with a matrix we get

$$M'^{+}R_1X(\boldsymbol{\beta}+X^{+}\boldsymbol{e}) \sim \mathcal{N}\left(\boldsymbol{\beta} + X^{+}\boldsymbol{e},\ \ w^2\|\boldsymbol{\beta} + X^{+}\boldsymbol{e}\|^2(M'^{\mathsf{T}}M)^{-1}\right)$$

I.e.,

$$M'^{+}R_1X(\boldsymbol{\beta}+X^{+}\boldsymbol{e}) \sim \|\boldsymbol{\beta}+X^{+}\boldsymbol{e}\|\cdot\mathcal{N}(\boldsymbol{u}, w^2(M'^{\mathsf{T}}M)^{-1})$$

where $\boldsymbol{u}$ denotes a unit-length vector in the direction of $\boldsymbol{\beta} + X^{+}\boldsymbol{e}$.

Similar to before we have

$$\begin{aligned}
&RP_{U^{\perp}} \sim \mathcal{N}(0_{r\times n}, I_{r\times r}, P_{U^{\perp}})\\
&\quad \Rightarrow M'^{+}(RP_{U^{\perp}}\boldsymbol{e}) \sim \mathcal{N}(\mathbf{0}_d,\ \|P_{U^{\perp}}e\|^2(M'^{\mathsf{T}}M')^{-1})\\
&wR_3 \sim \mathcal{N}(\mathbf{0}_r, w^2I_{r\times r})\\
&\quad \Rightarrow M'^{+}(wR_3) \sim \mathcal{N}(\mathbf{0}_d, w^2(M'^{+}M')^{-1})
\end{aligned}$$

Therefore, the distribution of $\boldsymbol{\beta}'$, which is the sum of the 3 independent Gaussians, is as required.

Also, $\boldsymbol{\zeta}' = \frac{1}{\sqrt{r}}P_{U'^{\perp}}\left(R_1X(\boldsymbol{\beta} + X^{+}\boldsymbol{e}) + R_1P_{U^{\perp}}\boldsymbol{e} + wR_3\right)$ is the sum of 3 independent Gaussians, which implies its distribution is

$$\mathcal{N}\Big(\tfrac{1}{\sqrt{r}}P_{U'^{\perp}}M'(\boldsymbol{\beta} + X^{+}\boldsymbol{e}),\ \ \tfrac{1}{r}(w^2(\|\boldsymbol{\beta} + X^{+}\boldsymbol{e}\|^2 + 1) + \|P_{U^{\perp}}\boldsymbol{e}\|^2)P_{U'^{\perp}}\Big)$$

I.e., $\mathcal{N}\left(\mathbf{0}_r,\ \ \frac{1}{r}(w^2(\|\boldsymbol{\beta} + X^{+}\boldsymbol{e}\|^2 + 1) + \|P_{U^{\perp}}\boldsymbol{e}\|^2)P_{U'^{\perp}}\right)$ as $P_{U'^{\perp}}M' = 0_{r\times r}$.

Finally, observe that $\boldsymbol{\beta}'$ and $\boldsymbol{\zeta}'$ are independent as the former depends on the projection of the spherical Gaussian $R_1X(\beta + X^{+}\boldsymbol{e}) + R_1P_{U^{\perp}}\boldsymbol{e} + wR_3$ on U', and the latter depends on the projection of the same multivariate Gaussian on $U'^{\perp}$. □

Observe that Claim C.1 assumes $\boldsymbol{e}$ is given. This may seem somewhat strange, since without assuming anything about $\boldsymbol{e}$ there can be many combinations of $\boldsymbol{\beta}$ and $\boldsymbol{e}$ for which $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$. However, we always have that $\boldsymbol{\beta} + X^{+}\boldsymbol{e} = X^{+}\boldsymbol{y} = \hat{\boldsymbol{\beta}}$. Similarly, it is always the case the $P_{U^{\perp}}\boldsymbol{e} = (I - XX^{+})\boldsymbol{y} = \boldsymbol{\zeta}$. (Recall OLS definitions of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\zeta}$ in Equation (1) and (2).) Therefore, the distribution of $\boldsymbol{\beta}'$ and $\boldsymbol{\zeta}'$ is unique (once $\boldsymbol{y}$ is set):

$$\begin{aligned}
\boldsymbol{\beta}' &\sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}, (w^2(\|\hat{\boldsymbol{\beta}}\|^2 + 1) + \|\boldsymbol{\zeta}\|^2)(M'^{\mathsf{T}}M')^{-1}\right)\\
\boldsymbol{\zeta}' &\sim \mathcal{N}\left(\mathbf{0}_r, \frac{w^2(\|\hat{\boldsymbol{\beta}}\|^2 + 1) + \|\boldsymbol{\zeta}\|^2}{r}(I_{r\times r} - M'M'^{+})\right)
\end{aligned}$$

And so for a given dataset $[X;\boldsymbol{y}]$ we have that $\boldsymbol{\beta}'$ serves as an approximation for $\hat{\boldsymbol{\beta}}$.

An immediate corollary of Claim C.1 is that for any fixed $\boldsymbol{e}$ it holds that the quantity $t'(\beta_j) = \frac{\beta'_j - (\beta_j + (X^{+}\boldsymbol{e})_j)}{\|\boldsymbol{\zeta}'\|\sqrt{\frac{r}{r-p}\cdot(M'^{\mathsf{T}}M')^{-1}_{j,j}}} = \frac{\beta'_j - \hat{\beta}_j}{\|\boldsymbol{\zeta}'\|\sqrt{\frac{r}{r-p}\cdot(M'^{\mathsf{T}}M')^{-1}_{j,j}}}$ is distributed like a T_{r-p}-distribution. Therefore, the following theorem follows immediately.

Theorem C.2. *Fix $X \in \mathbb{R}^{n\times p}$ and $\boldsymbol{y} \in \mathbb{R}$. Define $\hat{\boldsymbol{\beta}} = X^{+}\boldsymbol{y}$ and $\zeta = (I - XX^{+})\boldsymbol{y}$. Let RX' and $R\boldsymbol{y}'$ denote the result of applying Algorithm 1 to the matrix $A = [X;\boldsymbol{y}]$ when the algorithm appends the data with a $w\cdot I$ matrix. Fix a coordinate j and any $\alpha \in (0, 1/2)$. When computing $\boldsymbol{\beta}'$ and $\boldsymbol{\zeta}'$ as in Equations (9) it and (10), we have that w.p. $\geq 1 - \alpha$ it holds that*

$$\hat{\beta}_j \in \left(\beta'_j \pm c'_{\alpha}\|\boldsymbol{\zeta}'\|\sqrt{\tfrac{r}{r-p}\cdot(M'^{\mathsf{T}}M')^{-1}_{j,j}}\right)$$

where c'_{α} denotes the number such that $(-c'_{\alpha}, c'_{\alpha})$ contains $1 - \alpha$ mass of the T_{r-p}-distribution.

Note that Theorem C.2, much like the rest of the discussion in this Section, builds on $\boldsymbol{y}$ being fixed, which means β'_j serves as an approximation for $\hat{\beta}_j$. Yet our goal is to argue about similarity (or proximity) between β'_j and β_j. To that end, we combine the standard OLS confidence interval — which says that w.p. $\geq 1 - \alpha$ over the randomness of picking $\boldsymbol{e}$ in the homoscedastic model we have $|\beta_j - \hat{\beta}_j| \leq c_{\alpha}\|\boldsymbol{\zeta}\|\sqrt{\frac{(X^{\mathsf{T}}X)^{-1}_{j,j}}{n-p}}$ — with the confidence interval of Theorem C.2 above, and deduce that w.p. $\geq 1 - \alpha$

we have that $|\beta'_j - \beta_j|$ is at most

$$O\left(c_\alpha \frac{\|\boldsymbol{\zeta}\|\sqrt{(X^\mathsf{T}X)^{-1}_{j,j}}}{\sqrt{n-p}} + c'_\alpha \frac{\|\boldsymbol{\zeta}'\|\sqrt{r(M'^\mathsf{T}M')^{-1}_{j,j}}}{\sqrt{r-p}}\right) \tag{11}$$

[18]And so, in the next section, our goal is to give conditions under which the interval of Equation (11) isn't much larger in comparison to the interval length of $c'_\alpha \frac{\|\boldsymbol{\zeta}'\|}{\sqrt{r-p}}\sqrt{r(M'^\mathsf{T}M')^{-1}_{j,j}}$ we get from Theorem C.2; and more importantly — conditions that make the interval of Theorem C.2 useful and not too large. (Note, in expectation $\frac{\|\boldsymbol{\zeta}'\|}{\sqrt{r-p}}$ is about $\sqrt{(w^2 + w^2\|\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{\zeta}\|^2)/r}$. So, for example, in situations where $\|\hat{\boldsymbol{\beta}}\|$ is very large, this interval isn't likely to inform us as to the sign of β_j.)

Motivating Example. A good motivating example for the discussion in the following section is when $[X; \boldsymbol{y}]$ is a strict submatrix of the dataset A. That is, our data contains many variables for each entry (i.e., the dimensionality d of each entry is large), yet our regression is made only over a modest subset of variables out of the d. In this case, the least singular value of A might be too small, causing the algorithm to alter A; however, $\sigma_{\min}(X^\mathsf{T}X)$ could be sufficiently large so that had we run Algorithm 1 only on $[X; \boldsymbol{y}]$ we would not alter the input. (Indeed, a differentially private way for finding a subset of the variables that induce a submatrix with high $\sigma_{\min}$ is an interesting open question, partially answered — for a single regression — in the work of Thakurta and Smith (Thakurta & Smith, 2013).) Indeed, the conditions we specify in the following section depend on $\sigma_{\min}(\frac{1}{n}X^\mathsf{T}X)$, which, for a zero-mean data, the minimal variance of the data in any direction. For this motivating example, indeed such variance isn't necessarily small.

C.2. Conditions for Deriving a Confidence Interval for Ridge Regression

Looking at the interval specified in Equation (11), we now give an upper bound on the the random quantities in this interval: $\|\boldsymbol{\zeta}\|$, $\|\boldsymbol{\zeta}'\|$, and $(M'^\mathsf{T}M')^{-1}_{j,j}$. First, we give bound that are dependent on the randomness in R (i.e., we continue to view $\boldsymbol{e}$ as fixed).

Proposition C.3. *For any $\nu \in (0, 1/2)$, if we have $r = p + \Omega(\ln(1/\nu))$ then with probability $\geq 1 - \nu$ over the randomness of R we have $(r-p)(M'^\mathsf{T}M)^{-1}_{j,j} = \Theta\left((w^2 I_{p\times p} + X^\mathsf{T}X)^{-1}_{j,j}\right)$ and $\frac{\|\boldsymbol{\zeta}'\|^2}{r-p} = \Theta(\frac{w^2 + w^2\|\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{\zeta}\|^2}{r})$.*

Proof. The former bound follows from known results on the Johnson-Lindenstrauss transform (as were shown in the proof of Claim B.6). The latter bound follows from standard concentration bounds of the χ^2-distribution. □

Plugging in the result of Proposition C.3 to Equation (11) we get that w.p. $\geq 1 - \nu$ the difference $|\beta'_j - \beta_j|$ is at most

$$O\Big(c_\alpha \frac{\|\boldsymbol{\zeta}\|}{\sqrt{n-p}}\sqrt{(X^\mathsf{T}X)^{-1}_{j,j}} + c'_\alpha \sqrt{\frac{w^2 + w^2\|\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{\zeta}\|^2}{r-p}}\sqrt{(w^2 I_{p\times p} + X^\mathsf{T}X)^{-1}_{j,j}}\Big) \tag{12}$$

We will also use the following proposition.

Proposition C.4.

$$(X^\mathsf{T}X)^{-1}_{j,j} \leq \left(1 + \frac{w^2}{\sigma_{\min}(X^\mathsf{T}X)}\right)(w^2 I_{p\times p} + X^\mathsf{T}X)^{-1}_{j,j}$$

Proof. We have that

$$\begin{aligned}
&(X^\mathsf{T}X)^{-1} \\
&= (X^\mathsf{T}X)^{-1}(X^\mathsf{T}X + w^2 I_{p\times p})(X^\mathsf{T}X + w^2 I_{p\times p})^{-1} \\
&= (X^\mathsf{T}X + w^2 I_{p\times p})^{-1} + w^2(X^\mathsf{T}X)^{-1}(X^\mathsf{T}X + w^2 I_{p\times p})^{-1} \\
&= (I_{p\times p} + w^2(X^\mathsf{T}X)^{-1})(X^\mathsf{T}X + w^2 I_{p\times p})^{-1} \\
&= (X^\mathsf{T}X + w^2 I_{p\times p})^{-1/2} \cdot \\
&\qquad (I_{p\times p} + w^2(X^\mathsf{T}X)^{-1}) \cdot \\
&\qquad\qquad (X^\mathsf{T}X + w^2 I_{p\times p})^{-1/2}
\end{aligned}$$

where the latter holds because $(I_{p\times p} + w^2(X^\mathsf{T}X)^{-1})$ and $(X^\mathsf{T}X + w^2 I_{p\times p})^{-1}$ are diagonalizable by the same matrix V (the same matrix for which $(X^\mathsf{T}X) = VS^{-1}V^\mathsf{T}$). Since we have $\|I_{p\times p} + w^2(X^\mathsf{T}X)^{-1}\| = 1 + \frac{w^2}{\sigma^2_{\min}(X)}$, it is clear that $(I_{p\times p} + w^2(X^\mathsf{T}X)^{-1}) \preceq (1 + \frac{w^2}{\sigma^2_{\min}(X)})I_{p\times p}$. We deduce that $(X^\mathsf{T}X)^{-1}_{j,j} = \boldsymbol{e}_j^\mathsf{T}(X^\mathsf{T}X)^{-1}\boldsymbol{e}_j \leq (1 + \frac{w^2}{\sigma^2_{\min}(X)})(X^\mathsf{T}X + w^2 I_{p\times p})^{-1}_{j,j}$. □

Based on Proposition C.4 we get from Equation (12) that

[18]Observe that w.p. $\geq 1 - \alpha$ over the randomness of $\boldsymbol{e}$ we have that $|\beta_j - \hat{\beta}_j| \leq c_\alpha\|\boldsymbol{\zeta}\|\sqrt{\frac{(X^\mathsf{T}X)^{-1}_{j,j}}{n-p}}$, and w.p. $\geq 1 - \alpha$ over the randomness of R we have that $|\beta'_j - \hat{\beta}_j| \leq c'_\alpha\|\boldsymbol{\zeta}'\|\sqrt{\frac{r}{r-p}\cdot(M'^\mathsf{T}M')^{-1}_{j,j}}$. So technically, to give a $(1-\alpha)$-confidence interval around β'_j that contains β_j w.p. $\geq 1 - \alpha$, we need to use $c_{\alpha/2}$ and $c'_{\alpha/2}$ instead of c_α and c'_α resp. To avoid overburdening the reader with what we already see as too many parameters, we switch to asymptotic notation.

$|\beta'_j - \beta_j|$ is at most

$$O(\Big(c_\alpha\sqrt{\frac{\|\boldsymbol{\zeta}\|^2(1+\frac{w^2}{\sigma_{\min}(X^\mathsf{T}X)})}{n-p}}+ c'_\alpha\sqrt{\frac{w^2+w^2\|\hat{\boldsymbol{\beta}}\|^2+\|\boldsymbol{\zeta}\|^2}{r-p}}\Big)\sqrt{(w^2I_{p\times p}+X^\mathsf{T}X)^{-1}_{j,j}}) \tag{13}$$

And so, if it happens to be the case that exists some small $\eta > 0$ for which $\hat{\boldsymbol{\beta}}, \boldsymbol{\zeta}$ and w^2 satisfy

$$\frac{\|\boldsymbol{\zeta}\|^2(1+\frac{w^2}{\sigma_{\min}(X^\mathsf{T}X)})}{n-p} \leq \eta^2\left(\frac{w^2+w^2\|\hat{\boldsymbol{\beta}}\|^2+\|\boldsymbol{\zeta}\|^2}{r-p}\right) \tag{14}$$

then we have that $\mathbf{Pr}[\beta_j \in \left(\beta'_j \pm O((1+\eta)\cdot c'_\alpha\|\boldsymbol{\zeta}'\|\sqrt{\frac{r}{r-p}\cdot(M'^\mathsf{T}M')^{-1}_{j,j}})\right)] \geq 1-\alpha$.[19] Moreover, if in this case $|\beta_j| > c'_\alpha(1+\eta)\sqrt{\frac{w^2+w^2\|\hat{\boldsymbol{\beta}}\|^2+\|\boldsymbol{\zeta}\|^2}{r-p}}\sqrt{(w^2I_{p\times p}+X^\mathsf{T}X)^{-1}_{j,j}}$ then $\mathbf{Pr}[sign(\beta'_j) = sign(\beta_j)] \geq 1-\alpha$. This is precisely what Claims C.5 and C.6 below do.

Claim C.5. *If there exists $\eta > 0$ s.t. $n - p \geq \frac{2}{\eta^2}(r-p)$ and $n^2 = \Omega\left(r^{3/2}\cdot\frac{B^2\ln(1/\delta)}{\epsilon}\cdot\frac{1}{\eta^2\sigma_{\min}(\frac{1}{n}X^\mathsf{T}X)}\right)$, then $\mathbf{Pr}[\beta_j \in \left(\beta'_j \pm O((1+\eta)\cdot c'_\alpha\|\boldsymbol{\zeta}'\|\sqrt{\frac{r}{r-p}\cdot(M'^\mathsf{T}M')^{-1}_{j,j}})\right)] \geq 1-\alpha$.*

Proof. Based on the above discussion, it is enough to argue that under the conditions of the claim, the constraint of Equation (14) holds. Since we require $\frac{\eta^2}{2} \geq \frac{r-p}{n-p}$ then it is evident that $\frac{\|\boldsymbol{\zeta}\|^2}{n-p} \leq \frac{\eta^2\|\boldsymbol{\zeta}\|^2}{2(r-p)}$. So we now show that $\frac{\|\boldsymbol{\zeta}\|^2}{n-p}\cdot\frac{w^2}{\sigma_{\min}(X^\mathsf{T}X)} \leq \frac{\eta^2\|\boldsymbol{\zeta}\|^2}{2(r-p)}$ under the conditions of the claim, and this will show the required. All that is left is some algebraic manipulations. It suffices to have:

$$\frac{\eta^2}{2}\cdot\frac{n-p}{r-p}\sigma_{\min}(X^\mathsf{T}X) \geq \frac{\eta^2}{2}\cdot\frac{n^2}{r}\sigma_{\min}(\frac{1}{n}X^\mathsf{T}X) \geq \frac{32B^2\sqrt{r}\ln(8/\delta)}{\epsilon} \geq w^2$$

which holds for $n^2 \geq r^{3/2}\cdot\frac{64B^2\ln(1/\delta)}{\epsilon\eta^2}\sigma_{\min}(\frac{1}{n}X^\mathsf{T}X)^{-1}$, as we assume to hold. □

Claim C.6. *Fix $\nu \in (0, \frac{1}{2})$. If (i) $n = p + \Omega(\ln(1/\nu))$, (ii) $\|\boldsymbol{\beta}\|^2 = \Omega(\sigma^2\|X^+\|_F^2\ln(\frac{p}{\nu}))$ and (iii) $r - p = \Omega\left(\frac{(c'_\alpha)^2(1+\eta)^2}{\beta_j^2}\left(1+\|\boldsymbol{\beta}\|^2+\frac{\sigma^2}{\sigma_{\min}(\frac{1}{n}X^\mathsf{T}X)}\right)\right)$, then in the homoscedastic model, with probability $\geq 1-\nu-\alpha$ we have that $sign(\beta_j) = sign(\beta'_j)$.*

[19] We assume $n \geq r$ so $c_\alpha < c'_\alpha$ as the T_{n-p}-distribution is closer to a normal Gaussian than the T_{r-p}-distribution.

Proof. Based on the above discussion, we aim to show that in the homoscedastic model (where each coordinate $e_i \sim \mathcal{N}(0,\sigma^2)$ independently) w.p. $\geq 1-\nu$ it holds that the magnitude of β_j is greater than

$$c'_\alpha(1+\eta)\sqrt{\frac{w^2+w^2\|\hat{\boldsymbol{\beta}}\|^2+\|\boldsymbol{\zeta}\|^2}{r-p}}\sqrt{(w^2I_{p\times p}+X^\mathsf{T}X)^{-1}_{j,j}}$$

To show this, we invoke Claim A.4 to argue that w.p. $\geq 1-\nu$ we have (i) $\|\boldsymbol{\zeta}\|^2 \leq 2\sigma^2(n-p)$ (since $n = p + \Omega(\ln(1/\nu))$), and (ii) $\|\hat{\boldsymbol{\beta}}\|^2 \leq 2\|\boldsymbol{\beta}\|^2$ (since $\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|^2 \leq \sigma^2\|X^+\|_F^2\ln(\frac{p}{\nu})$ whereas $\|\boldsymbol{\beta}\|^2 = \Omega(\sigma^2\|X^+\|_F^2\ln(\frac{p}{\nu}))$). We also use the fact that $(w^2I_{p\times p}+X^\mathsf{T}X)^{-1}_{j,j} \leq (w^2+\sigma^{-1}_{\min}(X^\mathsf{T}X))$, and then deduce that

$$\begin{aligned}(1+\eta)c'_\alpha\sqrt{\frac{w^2+w^2\|\hat{\boldsymbol{\beta}}\|^2+\|\boldsymbol{\zeta}\|^2}{r-p}}\sqrt{(w^2I_{p\times p}+X^\mathsf{T}X)^{-1}_{j,j}} &\leq \frac{(1+\eta)c'_\alpha}{\sqrt{r-p}}\sqrt{2\frac{w^2(1+\|\boldsymbol{\beta}\|^2)+\sigma^2(n-p)}{w^2+\sigma_{\min}(X^\mathsf{T}X)}} \\ &\leq \frac{(1+\eta)c'_\alpha}{\sqrt{r-p}}\sqrt{2(1+\|\boldsymbol{\beta}\|^2)+\frac{2\sigma^2(n-p)}{\sigma_{\min}(X^\mathsf{T}X)}} \leq |\beta_j|\end{aligned}$$

due to our requirement on $r-p$. □

Observe, out of the 3 conditions specified in Claim C.6, condition (i) merely guarantees that the sample is large enough to argue that estimations are close to their expect value; and condition (ii) is there merely to guarantee that $\|\hat{\boldsymbol{\beta}}\| \approx \|\boldsymbol{\beta}\|$. It is condition (iii) which is non-trivial to hold, especially together with the conditions of Claim C.5 that pose other constraints in regards to r, n, η and the various other parameters in play. It is interesting to compare the requirements on r to the lower bound we get in Theorem 3.3 — especially the latter bound. The two bounds are strikingly similar, with the exception that here we also require $r-p$ to be greater than $\frac{1+\|\boldsymbol{\beta}\|^2}{\beta_j^2}$. This is part of the unfortunate effect of altering the matrix A: we cannot give confidence bounds only for the coordinates j for which β_j^2 is very small *relative to* $\|\boldsymbol{\beta}\|^2$.

In summary, we require to have $n = p + \Omega(\ln(1/\nu))$ and that X contains enough sample points to have $\|\hat{\boldsymbol{\beta}}\|$ comparable to $\|\boldsymbol{\beta}\|$, and then set r and η such that (it is convenient to think of η as a small constant, say, $\eta = 0.1$)

- $r - p = O(\eta^2(n-p))$ (which implies $r = O(n)$)
- $r = O(\Big(\eta^2\frac{\epsilon n^2}{B^2\ln(1/\delta)}\sigma_{\min}(\frac{1}{n}X^\mathsf{T}X)\Big)^{\frac{2}{3}})$
- $r - p = \Omega(\frac{1+\|\boldsymbol{\beta}\|^2}{\beta_j^2} + \frac{\sigma^2}{\beta_j^2}\cdot\sigma^{-1}_{\min}(\frac{1}{n}X^\mathsf{T}X))$

to have that the $(1-\alpha)$-confidence interval around β'_j does not intersect the origin. Once again, we comment that these conditions are sufficient but not necessary, and furthermore — even with these conditions holding — we do not make any claims of optimality of our confidence bound. That is because from Proposition C.4 onwards our discussion uses upper bounds that do not have corresponding lower bounds, to the best of our knowledge.

D. Confidence Intervals for "Analyze Gauss" Algorithm

To complete the picture, we now analyze the "Analyze Gauss" algorithm of Dwork et al (Dwork et al., 2014). Algorithm 2 works by adding random Gaussian noise to $A^\mathsf{T}A$, where the noise is symmetric with each coordinate above the diagonal sampled i.i.d from $\mathcal{N}(0,\Delta^2)$ with $\Delta^2 = O\left(B^4 \frac{\log(1/\delta)}{\epsilon^2}\right)$.[20] Using the same notation for a sub-matrix of A as $[X;\boldsymbol{y}]$ as before, with $X \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^n$, we denote the output of Algorithm 2 as

$$\left(\begin{array}{c|c} \widetilde{X^\mathsf{T}X} & \widetilde{X^\mathsf{T}\boldsymbol{y}} \\ \hline \widetilde{\boldsymbol{y}^\mathsf{T}X} & \widetilde{\boldsymbol{y}^\mathsf{T}\boldsymbol{y}} \end{array}\right) = \left(\begin{array}{c|c} X^\mathsf{T}X + N & X^\mathsf{T}\boldsymbol{y} + \boldsymbol{n} \\ \hline \boldsymbol{y}^\mathsf{T}X + \boldsymbol{n}^\mathsf{T} & \boldsymbol{y}^\mathsf{T}\boldsymbol{y} + m \end{array}\right) \tag{15}$$

where N is a symmetric $p\times p$-matrix, $\boldsymbol{n}$ is a p-dimensional vector and m is a scalar, whose coordinates are sampled i.i.d from $\mathcal{N}(0,\Delta^2)$.

Using the output of Algorithm 2, it is simple to derive analogues of $\hat{\boldsymbol{\beta}}$ and $\|\boldsymbol{\zeta}\|^2$ (Equations (1) and (2))

$$\widetilde{\boldsymbol{\beta}} = \left(\widetilde{X^\mathsf{T}X}\right)^{-1}\widetilde{X^\mathsf{T}\boldsymbol{y}} = \left(X^\mathsf{T}X + N\right)^{-1}\left(X^T\boldsymbol{y} + \boldsymbol{n}\right) \tag{16}$$

$$\begin{aligned}\widetilde{\|\boldsymbol{\zeta}\|^2} &= \widetilde{\boldsymbol{y}^\mathsf{T}\boldsymbol{y}} - 2\widetilde{\boldsymbol{y}^T X}\,\widetilde{\boldsymbol{\beta}} + \widetilde{\boldsymbol{\beta}}^\mathsf{T}\widetilde{X^\mathsf{T}X}\,\widetilde{\boldsymbol{\beta}} \\ &= \widetilde{\boldsymbol{y}^T\boldsymbol{y}} - \widetilde{\boldsymbol{y}^T X}\,\widetilde{X^\mathsf{T}X}^{-1}\widetilde{X^T\boldsymbol{y}}\end{aligned} \tag{17}$$

We now argue that it is possible to use $\widetilde{\beta_j}$ and $\widetilde{\|\boldsymbol{\zeta}\|^2}$ to get a confidence interval for β_j under certain conditions.

Theorem D.1. *Fix $\alpha,\nu \in (0,\frac{1}{2})$. Assume that there exists $\eta \in (0,\frac{1}{2})$ s.t. $\sigma_{\min}(X^\mathsf{T}X) > \Delta\sqrt{p\ln(1/\nu)}/\eta$. Under the homoscedastic model, given $\boldsymbol{\beta}$ and σ^2, if we assume also that $\|\boldsymbol{\beta}\| \le B$ and $\|\hat{\boldsymbol{\beta}}\| = \|(X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{y}\| \le B$, then w.p. $\ge 1-\alpha-\nu$ it holds that $|\beta_j - \widetilde{\beta_j}|$ it at most*

$$O\left(\rho\cdot\sqrt{\left(\widetilde{X^\mathsf{T}X}^{-1}_{j,j} + \Delta\sqrt{p\ln(1/\nu)}\cdot\widetilde{X^\mathsf{T}X}^{-2}_{j,j}\right)\ln(1/\alpha)}\right.$$
$$\left. + \Delta\sqrt{\widetilde{X^\mathsf{T}X}^{-2}_{j,j}\cdot\ln(1/\nu)}\cdot(B\sqrt{p}+1)\right)$$

where ρ is such that ρ^2 is w.h.p an upper bound on σ^2, defined as

$$\rho^2 \stackrel{\text{def}}{=} \left(\frac{1}{\sqrt{n-p}-2\sqrt{\ln(4/\alpha)}}\right)^2\cdot$$
$$\left(\widetilde{\|\boldsymbol{\zeta}\|^2} - C\cdot\left(\Delta\frac{B^2\sqrt{p}}{1-\eta}\sqrt{\ln(1/\nu)} + \Delta^2\|\widetilde{X^\mathsf{T}X}^{-1}\|_F\cdot\ln(p/\nu)\right)\right)$$

for some large constant C.

We comment that in practice, instead of using ρ, it might be better to use the MLE of σ^2, namely:

$$\widetilde{\sigma^2} \stackrel{\text{def}}{=} \frac{1}{n-p}\left(\widetilde{\|\boldsymbol{\zeta}\|^2} + \Delta^2\|\widetilde{X^\mathsf{T}X}^{-1}\|_F\right)$$

instead of ρ^2, the upper bound we derived for σ^2. (Replacing an unknown variable with its MLE estimator is a common approach in applied statistics.) Note that the assumption that $\|\boldsymbol{\beta}\| \le B$ is fairly benign once we assume each row has bounded l_2-norm. The assumption $\|\hat{\boldsymbol{\beta}}\| \le B$ simply assumes that $\hat{\boldsymbol{\beta}}$ is a reasonable estimation of $\boldsymbol{\beta}$, which is likely to hold if we assume that $X^\mathsf{T}X$ is well-spread. The assumption about the magnitude of the least singular value of $X^\mathsf{T}X$ is therefore the major one. Nonetheless, in the case we considered before where each row in X is sampled i.i.d from $\mathcal{N}(\mathbf{0},\Sigma)$, this assumption merely means that n is large enough s.t. $n = \tilde{\Omega}(\frac{\Delta\sqrt{p\ln(1/\nu)}}{\eta\cdot\sigma_{\min}(\Sigma)})$.

In order to prove Theorem D.1, we require the following proposition.

Proposition D.2. *Fix any $\nu \in (0,\frac{1}{2})$. Fix any matrix $M \in \mathbb{R}^{p\times p}$. Let $\boldsymbol{v} \in \mathbb{R}^p$ be a vector with each coordinate sampled independently from a Gaussian $\mathcal{N}(0,\Delta^2)$. Then we have that $\mathbf{Pr}\left[\|M\boldsymbol{v}\| > \Delta\cdot\|M\|_F\sqrt{2\ln(2p/\nu)}\right] < \nu$.*

Proof. Given M, we have that $M\boldsymbol{v} \sim \mathcal{N}(\mathbf{0},\Delta^2\cdot MM^\mathsf{T})$. Denoting M's singular values as $sv_1,\ldots,sv_p$, we can rotate $M\boldsymbol{v}$ without affecting its l_2-norm and infer that $\|M\boldsymbol{v}|^2$ is distributed like a sum on p independent Gaussians, each sampled from $\mathcal{N}(0,\Delta^2\cdot sv_i^2)$. Standard union bound gives that w.p. $\ge 1-\nu$ non of the p Gaussians exceeds its standard deviation by a factor of $\sqrt{2\ln(2p/\nu)}$. Hence, w.p. $\ge 1-\nu$ it holds that $\|M\boldsymbol{v}\|^2 \le 2\Delta^2\sum_i sv_i^2\ln(2p/\nu) = 2\Delta^2\cdot\mathrm{trace}(MM^\mathsf{T})\cdot\ln(2p/\nu)$. □

Our proof also requires the use of the following equality, that holds for any invertible A and any matrix B s.t. $I + B\cdot A^{-1}$ is invertible:

$$(A+B)^{-1} = A^{-1} - A^{-1}\left(I + BA^{-1}\right)^{-1}BA^{-1}$$

[20] It is easy to see that the l_2-global sensitivity of the mapping $A \mapsto A^\mathsf{T}A$ is $\propto B^4$. Fix any A_1, A_2 that differ on one row which is some vector $\boldsymbol{v}$ with $\|\boldsymbol{v}\| = B$ in A_1 and the all zero vector in A_2. Then $GS_2^2 = \|A_1^\mathsf{T}A_1 - A_2^\mathsf{T}A_2\|_F^2 = \|\boldsymbol{v}\boldsymbol{v}^T\|_F^2 = \mathrm{trace}(\boldsymbol{v}\boldsymbol{v}^\mathsf{T}\cdot\boldsymbol{v}\boldsymbol{v}^\mathsf{T}) = (\boldsymbol{v}^\mathsf{T}\boldsymbol{v})^2 = B^4$.

In our case, we have

$$\begin{aligned}
&\widetilde{X^\mathsf{T}X}^{-1} \\
&= (X^\mathsf{T}X + N)^{-1} \\
&= (X^\mathsf{T}X)^{-1} - (X^\mathsf{T}X)^{-1}\left(I + N(X^\mathsf{T}X)^{-1}\right)^{-1} N(X^\mathsf{T}X)^{-1} \\
&= (X^\mathsf{T}X)^{-1}\left(I - \left(I + N(X^\mathsf{T}X)^{-1}\right)^{-1} N(X^\mathsf{T}X)^{-1}\right) \\
&\stackrel{\text{def}}{=} (X^\mathsf{T}X)^{-1}\left(I - Z\cdot(X^\mathsf{T}X)^{-1}\right) \qquad (18)
\end{aligned}$$

Proof of Theorem D.1. Fix $\nu > 0$. First, we apply to standard results about Gaussian matrices, such as (Tao, 2012) (used also by (Dwork et al., 2014) in their analysis), to see that w.p. $\geq 1 - \nu/6$ we have $\|N\| = O(\Delta\sqrt{p\ln(1/\nu)})$. And so, for the remainder of the proof we fix N subject to having bounded operator norm. Note that by fixing N we fix $\widetilde{X^\mathsf{T}X}$.

Recall that in the homoscedastic model, $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$ with each coordinate of $\boldsymbol{e}$ sampled i.i.d from $\mathcal{N}(0,\sigma^2)$. We therefore have that

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}} &= \widetilde{X^\mathsf{T}X}^{-1}(X^\mathsf{T}\boldsymbol{y} + \boldsymbol{n}) = \widetilde{X^\mathsf{T}X}^{-1}(X^\mathsf{T}X\boldsymbol{\beta} + X^\mathsf{T}\boldsymbol{e} + \boldsymbol{n}) \\
&= \widetilde{X^\mathsf{T}X}^{-1}(\widetilde{X^\mathsf{T}X} - N)\boldsymbol{\beta} + \widetilde{X^\mathsf{T}X}^{-1}X^\mathsf{T}\boldsymbol{e} + \widetilde{X^\mathsf{T}X}^{-1}\boldsymbol{n} \\
&= \boldsymbol{\beta} - \widetilde{X^\mathsf{T}X}^{-1}N\boldsymbol{\beta} + \widetilde{X^\mathsf{T}X}^{-1}X^\mathsf{T}\boldsymbol{e} + \widetilde{X^\mathsf{T}X}^{-1}\boldsymbol{n}
\end{aligned}$$

Denoting the j-th row of $\widetilde{X^\mathsf{T}X}^{-1}$ as $\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}$ we deduce:

$$\widetilde{\beta}_j = \beta_j - \widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}N\boldsymbol{\beta} + \widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}X^\mathsf{T}\boldsymbol{e} + \widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}\boldsymbol{n} \qquad (19)$$

We naïvely bound the size of the term $\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}N\boldsymbol{\beta}$ by $\left\|\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}\right\|\|N\|\|\boldsymbol{\beta}\| = O\left(\left\|\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}\right\|\cdot B\Delta\sqrt{p\ln(1/\nu)}\right)$.

To bound $\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}X^\mathsf{T}\boldsymbol{e}$ note that $\boldsymbol{e}$ is chosen independently of $\widetilde{X^\mathsf{T}X}$ and since $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0},\sigma^2 I)$ we have $\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}X^\mathsf{T}\boldsymbol{e} \sim \mathcal{N}\left(\mathbf{0},\sigma^2\cdot\boldsymbol{e}_j^\mathsf{T}\widetilde{X^\mathsf{T}X}^{-1}\cdot X^\mathsf{T}X\cdot\widetilde{X^\mathsf{T}X}^{-1}\boldsymbol{e}_j\right)$. Since we have

$$\begin{aligned}
&\widetilde{X^\mathsf{T}X}^{-1}\cdot X^\mathsf{T}X\cdot\widetilde{X^\mathsf{T}X}^{-1} \\
&= \widetilde{X^\mathsf{T}X}^{-1}\cdot(\widetilde{X^\mathsf{T}X} - N)\cdot\widetilde{X^\mathsf{T}X}^{-1} \\
&= \widetilde{X^\mathsf{T}X}^{-1} - \widetilde{X^\mathsf{T}X}^{-1}\cdot N\cdot\widetilde{X^\mathsf{T}X}^{-1}
\end{aligned}$$

we can bound the variance of $\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}X^\mathsf{T}\boldsymbol{e}$ by $\sigma^2\left(\widetilde{X^\mathsf{T}X}^{-1}_{j,j} + \|N\|\cdot\left\|\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}\right\|^2\right)$. Appealing to Gaussian concentration bounds, we have that w.p. $\geq 1-\alpha/2$ the absolute value of this Gaussian is at most $O\left(\sqrt{\left(\widetilde{X^\mathsf{T}X}^{-1}_{j,j} + \Delta\sqrt{p\ln(1/\nu)}\cdot\left\|\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}\right\|^2\right)\sigma^2\ln(1/\alpha)}\right)$.

To bound $\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}\boldsymbol{n}$ note that $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0},\Delta^2 I)$ is sampled independently of $\widetilde{X^\mathsf{T}X}$. We therefore have that $\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}\boldsymbol{n} \sim \mathcal{N}(0,\Delta^2\left\|\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}\right\|^2)$. Gaussian concentration bounds give that w.p $\geq 1-\nu/6$ we have $|\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}\boldsymbol{n}| = O\left(\Delta\left\|\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}\right\|\sqrt{\ln(1/\nu)}\right)$.

Plugging this into our above bounds on all terms that appear in Equation (19) we have that w.p. $\geq 1-\nu/2-\alpha/2$ we have that $\left|\widetilde{\beta}_j - \beta_j\right|$ is at most

$$\begin{aligned}
&O\left(\left\|\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}\right\|\cdot B\Delta\sqrt{p\ln(1/\nu)}\right) \\
&+ O\left(\sigma\sqrt{\left(\widetilde{X^\mathsf{T}X}^{-1}_{j,j} + \Delta\sqrt{p\ln(1/\nu)}\cdot\left\|\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}\right\|^2\right)\ln(1/\alpha)}\right) \\
&+ O\left(\Delta\left\|\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}\right\|\sqrt{\ln(1/\nu)}\right)
\end{aligned}$$

Note that due to the symmetry of $\widetilde{X^\mathsf{T}X}$ we have $\left\|\widetilde{X^\mathsf{T}X}^{-1}_{j\rightarrow}\right\|^2 = \widetilde{X^\mathsf{T}X}^{-2}_{j,j}$ (the (j,j)-coordinate of the matrix $\widetilde{X^\mathsf{T}X}^{-2}$), thus $|\widetilde{\beta}_j - \beta_j|$ is at most

$$\begin{aligned}
&O\Big(\sigma\cdot\sqrt{\left(\widetilde{X^\mathsf{T}X}^{-1}_{j,j} + \Delta\sqrt{p\ln(1/\nu)}\cdot\widetilde{X^\mathsf{T}X}^{-2}_{j,j}\right)\ln(1/\alpha)} \\
&\quad + \Delta\sqrt{\widetilde{X^\mathsf{T}X}^{-2}_{j,j}\cdot\ln(1/\nu)}\cdot(B\sqrt{p}+1)\Big) \qquad (20)
\end{aligned}$$

All of the terms appearing in Equation (20) are known given $\widetilde{X^\mathsf{T}X}$, except for σ — which is a parameter of the model. Next, we derive an upper bound on σ which we can then plug into Equation (20) to complete the proof of the theorem and derive a confidence interval for β_j.

Recall Equation (17), according to which we have

$$\begin{aligned}\widetilde{\|\boldsymbol{\zeta}\|^2} &= \widetilde{\boldsymbol{y}^T\boldsymbol{y}} - \widetilde{\boldsymbol{y}^T X}\,\widetilde{X^\mathsf{T}X}^{-1}\,\widetilde{X^T\boldsymbol{y}} \\ &\overset{(18)}{=} \boldsymbol{y}^\mathsf{T}\boldsymbol{y} + m \\ &\quad - (\boldsymbol{y}^\mathsf{T}X + \boldsymbol{n}^\mathsf{T})(X^\mathsf{T}X)^{-1}(I - Z\cdot(X^\mathsf{T}X)^{-1})(X^\mathsf{T}\boldsymbol{y} + \boldsymbol{n}) \\ &= \boldsymbol{y}^\mathsf{T}\boldsymbol{y} + m \\ &\quad - \boldsymbol{y}^\mathsf{T}X(X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{y} \\ &\quad + \boldsymbol{y}^\mathsf{T}X(X^\mathsf{T}X)^{-1}Z(X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{y} \\ &\quad - 2\boldsymbol{y}^\mathsf{T}X(X^\mathsf{T}X)^{-1}\boldsymbol{n} \\ &\quad + 2\boldsymbol{y}^\mathsf{T}X(X^\mathsf{T}X)^{-1}Z(X^\mathsf{T}X)^{-1}\boldsymbol{n} \\ &\quad - \boldsymbol{n}^\mathsf{T}(X^\mathsf{T}X)^{-1}(I - Z\cdot(X^\mathsf{T}X)^{-1})\boldsymbol{n}\end{aligned}$$

Recall that $\hat{\boldsymbol{\beta}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{y}$, and so we have

$$\begin{aligned}&= \boldsymbol{y}^\mathsf{T}\left(I - X(X^\mathsf{T}X)^{-1}X^\mathsf{T}\right)\boldsymbol{y} + m - \hat{\boldsymbol{\beta}}^\mathsf{T}Z\hat{\boldsymbol{\beta}} \\ &\quad - 2\hat{\boldsymbol{\beta}}^\mathsf{T}(I - Z(X^\mathsf{T}X)^{-1})\boldsymbol{n} - \boldsymbol{n}^\mathsf{T}\widetilde{X^\mathsf{T}X}^{-1}\boldsymbol{n} \qquad (21)\end{aligned}$$

and of course, both $\boldsymbol{n}$ and m are chosen independently of $\widetilde{X^\mathsf{T}X}$ and $\boldsymbol{y}$.

Before we bound each term in Equation (21), we first give a bound on $\|Z\|$. Recall, $Z = \left(I + N(X^\mathsf{T}X)^{-1}\right)^{-1}N$. Recall our assumption (given in the statement of Theorem D.1) that $\sigma_{\min}(X^\mathsf{T}X) \geq \frac{\Delta}{\eta}\sqrt{p\ln(1/\nu)}$. This implies that $\|N(X^\mathsf{T}X)^{-1}\| \leq \|N\|\cdot\sigma_{\min}(X^\mathsf{T}X)^{-1} = O(\eta)$. Hence

$$\|Z\| \leq (\|I + N(X^\mathsf{T}X)^{-1}\|)^{-1}\cdot\|N\| = O\left(\frac{\Delta\sqrt{p\ln(1/\nu)}}{1-\eta}\right)$$

Moreover, this implies that $\|Z(X^\mathsf{T}X)^{-1}\| \leq O\left(\frac{\eta}{1-\eta}\right)$ and that $\|I - Z(X^\mathsf{T}X)^{-1}\| \leq O\left(\frac{1}{1-\eta}\right)$.

Armed with these bounds on the operator norms of Z and $(I - Z(X^\mathsf{T}X)^{-1})$ we bound the magnitude of the different terms in Equation (21).

- The term $\boldsymbol{y}^T\left(I - XX^+\right)\boldsymbol{y}$ is the exact term from the standard OLS, and we know it is distributed like $\sigma^2\cdot\chi^2_{n-p}$ distribution. Therefore, it is greater than $\sigma^2(\sqrt{n-p} - 2\sqrt{\ln(4/\alpha)})^2$ w.p. $\geq 1 - \alpha/2$.
- The scalar m sampled from $m \sim \mathcal{N}(0, \Delta^2)$ is bounded by $O(\Delta\sqrt{\ln(1/\nu)})$ w.p. $\geq 1 - \nu/8$.
- Since we assume $\|\hat{\boldsymbol{\beta}}\| \leq B$, the term $\hat{\boldsymbol{\beta}}^\mathsf{T}Z\hat{\boldsymbol{\beta}}$ is upper bounded by $B^2\|Z\| = O\left(\frac{B^2\Delta\sqrt{p\ln(1/\nu)}}{1-\eta}\right)$.
- Denote $\boldsymbol{z}^\mathsf{T}\boldsymbol{n} = 2\hat{\boldsymbol{\beta}}^\mathsf{T}(I - Z(X^\mathsf{T}X)^{-1})\boldsymbol{n}$. We thus have that $\boldsymbol{z}^\mathsf{T}\boldsymbol{n} \sim \mathcal{N}(0, \Delta^2\|\boldsymbol{z}\|^2)$ and that its magnitude is at most $O(\Delta\cdot\|\boldsymbol{z}\|\sqrt{\ln(1/\nu)})$ w.p. $\geq 1 - \nu/8$. We can upper bound $\|\boldsymbol{z}\| \leq 2\|\hat{\boldsymbol{\beta}}\|\,\|I - Z(X^\mathsf{T}X)^{-1}\| = O(\frac{B}{1-\eta})$, and so this term's magnitude is upper bounded by $O\left(\frac{\Delta\cdot B\sqrt{\ln(1/\nu)}}{1-\eta}\right)$.
- Given our assumption about the least singular value of $X^\mathsf{T}X$ and with the bound on $\|N\|$, we have that $\sigma_{\min}(\widetilde{X^\mathsf{T}X}) \geq \sigma_{\min}(X^\mathsf{T}X) - \|N\| > 0$ and so the symmetric matrix $\widetilde{X^\mathsf{T}X}$ is a PSD. Therefore, the term $\boldsymbol{n}^\mathsf{T}\widetilde{X^\mathsf{T}X}^{-1}\boldsymbol{n} = \|\widetilde{X^\mathsf{T}X}^{-1/2}\boldsymbol{n}\|^2$ is strictly positive. Applying Proposition D.2 we have that w.p. $\geq 1 - \nu/8$ it holds that $\boldsymbol{n}^\mathsf{T}\widetilde{X^\mathsf{T}X}^{-1}\boldsymbol{n} \leq O\left(\Delta^2\|\widetilde{X^\mathsf{T}X}^{-1}\|_F\cdot\ln(p/\nu)\right)$.

Plugging all of the above bounds into Equation (21) we get that w.p. $\geq 1 - \nu/2 - \alpha/2$ it holds that

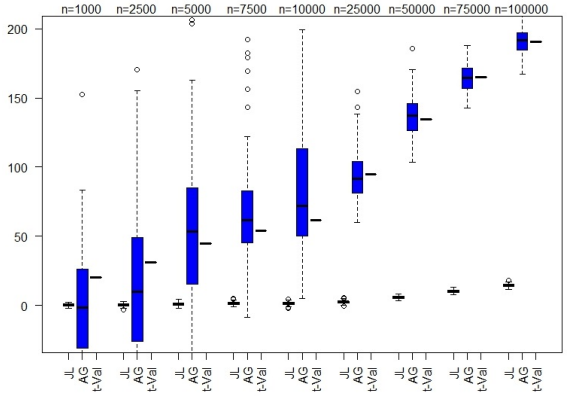
$$\sigma^2 \leq \left(\frac{1}{\sqrt{n-p} - 2\sqrt{\ln(4/\alpha)}}\right)^2 \cdot \left(\widetilde{\|\boldsymbol{\zeta}\|^2} + O\left((1 + \tfrac{B^2\sqrt{p}+B}{1-\eta})\Delta\sqrt{\ln(1/\nu)} + \Delta^2\|\widetilde{X^\mathsf{T}X}^{-1}\|_F\cdot\ln(p/\nu)\right)\right)$$

and indeed, the RHS is the definition of ρ^2 in the statement of Theorem D.1. □

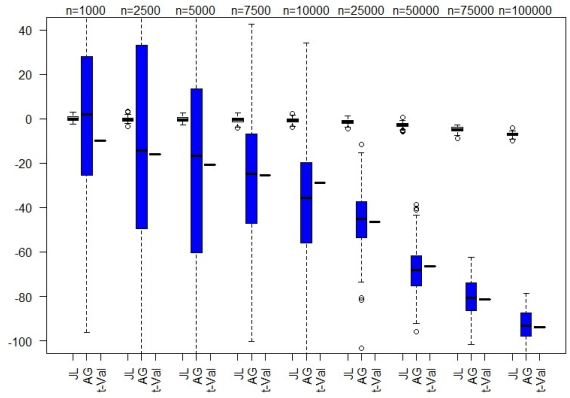
E. Experiment: Additional Figures

To complete our discussion about the experiments we have conducted, we attach here additional figures, plotting both the t-value approximations we get from both algorithms, and the "high-level decision" of whether correctly reject or not-reject the null hypothesis (and with what sign). First, we show the distribution of the t-value approximation for coordinates that should be rejected, in Figure 2, and then the decision of whether to reject or not based on this t-value — and whether it was right, conservative (we didn't reject while we needed to) or wrong (we rejected with the wrong sign, or rejected when we shouldn't have rejected) in Figure 3. As one can see, Algorithm 1 has far lower t-values (as expected) and therefore is much more conservative. In fact, it tends to not-reject coordinate 1 of the real-data even on the largest value of n (Figure 3c).

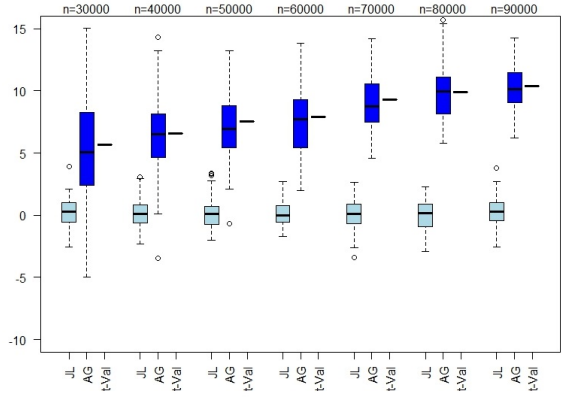
However, because Algorithm 1 also has much smaller variance, it also does not reject when it ought to not-reject, whereas Algorithm 2 erroneiously rejects the null-hypotheses. This can be seen in Figures 4 and 5.



(a) Synthetic data, coordinate $\beta_1 = 0.5$

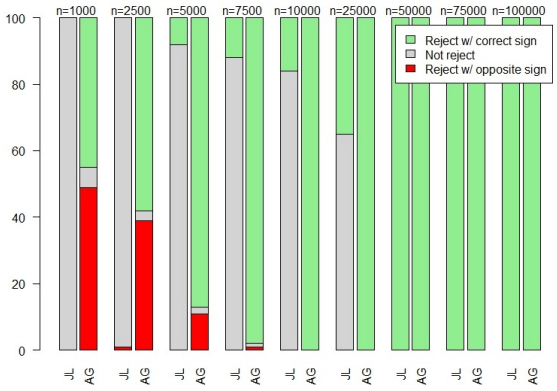


(b) Synthetic data, coordinate $\beta_2 = -0.25$

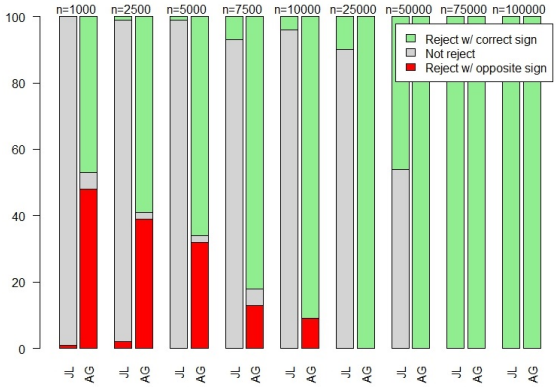


(c) real-life data, coordinate $\beta_1 = 14.07$

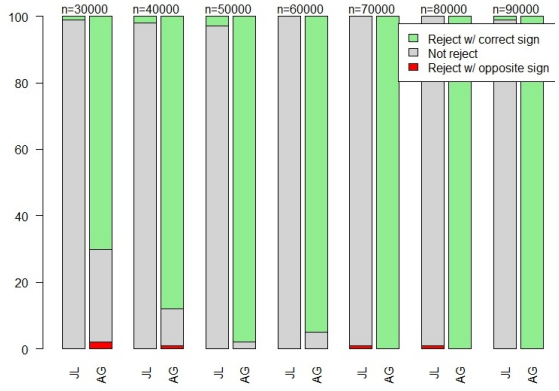
Figure 2. The distribution of the t-value approximations from selected experiments on synthetic and real-life data where the null hypothesis should be rejected



(a) Synthetic data, coordinate $\beta_1 = 0.5$

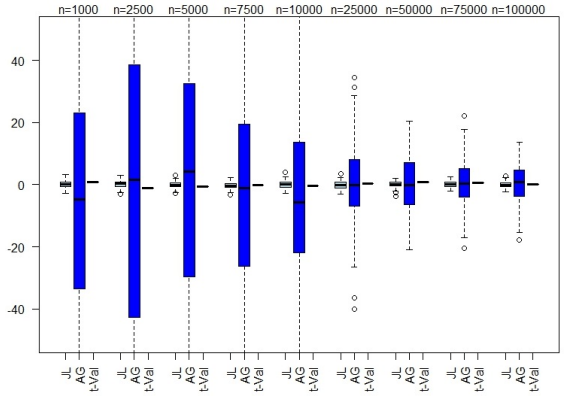


(b) Synthetic data, coordinate $\beta_2 = -0.25$

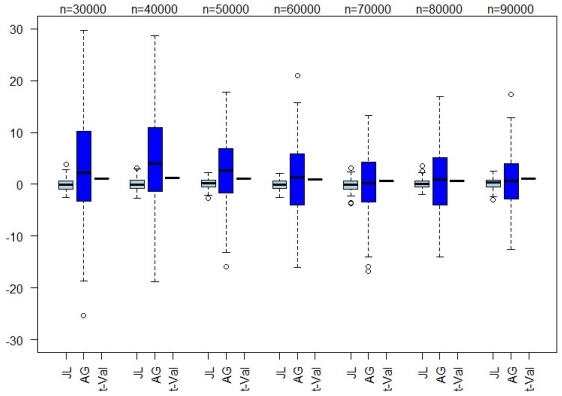


(c) real-life data, coordinate $\beta_1 = 14.07$

Figure 3. The correctness of our decision to reject the null-hypothesis based on the approximated t-value where the null hypothesis should be rejected



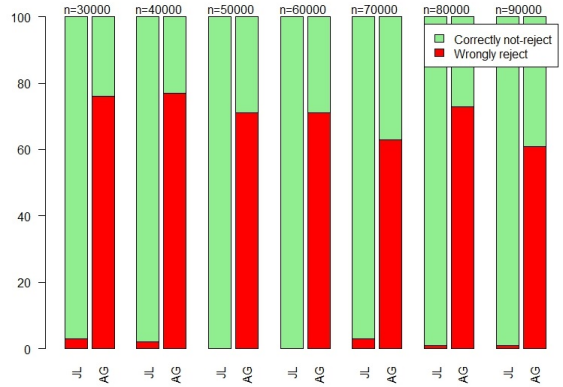
(a) Synthetic data, coordinate $\beta_3 = 0$



(b) real-life data, coordinate $\beta_2 = 0.57$

Figure 4. The distribution of the t-value approximations from selected experiments on synthetic and real-life data when the null hypothesis is (essentially) true

(a) Synthetic data, coordinate $\beta_3 = 0$



(b) Synthetic data, coordinate $\beta_2 = 0.57$

Figure 5. The correctness of our decision to reject the null-hypothesis based on the approximated t-value when the null hypothesis is (essentially) true