1 Notation

We will consider the linear regression setting where the number of observations, n, is much larger than the number of variables, p, ie $n \gg p$. Given the $n \times p$ design matrix $\mathbf{X}$, our model for the observations Y is $Y = \mathbf{X}\beta^* + \epsilon$, where β^* is the true parameter that we seek to estimate and ϵ is iid sub-Gaussian noise. Let $Var(\epsilon) = \sigma^2 I$.

Our idea is to replace $\mathbf{X}$ with a sparsified version $\mathring{\mathbf{X}}$. Given a vector of probabilities $\theta \in \mathbb{R}^p$, we set an entry in the jth column to 0 with probability $1 - \theta_j$, and multiply it by $\frac{1}{\theta_j}$ otherwise. In other words

$$\mathring{\mathbf{X}} = \mathbf{X} \circ Z \qquad \text{where } Z_{ij} \overset{iid}{\sim} \frac{1}{\theta_j} Ber(\theta_j) \tag{1}$$

Here we use $\circ$ to denote the Hadamard product. Note that $\mathbb{E}[\mathring{\mathbf{X}}|\mathbf{X}] = \mathbf{X}$ and

$$\frac{1}{n}\mathbb{E}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}} = \frac{1}{n}\mathbb{E}[\mathbf{X}^T\mathbf{X} + diag(\frac{1-\theta_j}{\theta_j}\sum_i \mathbf{X}_{ij}^2)] = \Sigma + diag(\frac{1-\theta_j}{\theta_j}\Sigma_{jj}) \tag{2}$$

where the first equality follows by conditioning on $\mathbf{X}$.

Assumption: Rather than assume that the rows of $\mathbf{X}$ are subgaussian, we will assume that the rows are bounded with $\|\mathbf{X}_i\| \leq m$ for all rows $\mathbf{X}_i$. This trivially implies that $\|\mathring{\mathbf{X}}_i\| \leq \frac{m}{\theta_{\min}}$. Given that $\mathbb{E}[\|\mathring{\mathbf{X}}_i\|^2|\mathbf{X}_i] = \frac{\|\mathbf{X}_i\|^2}{\theta_{\min}}$, we might expect to have $|\mathring{\mathbf{X}}_i\| \leq m/\sqrt{\theta_{\min}}$, but I haven't figured out how to show that yet.

2 Decomposition

From now on, we will assume that $\theta_j = \theta$ for all j. Define the $\widehat{\beta}$ to be the solution to the sparsified system,

$$\widehat{\beta} = (\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\mathring{\mathbf{X}}^T Y \tag{3}$$

Our goal is to bound the expected squared error $\mathbb{E}\|\widehat{\beta} - \beta^*\|^2$, where the expectation is taken over the noise ϵ. We can rewrite this as

$$\mathbb{E}\|\widehat{\beta} - \beta^*\|^2 = \mathbb{E}\|(\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\mathring{\mathbf{X}}^T\mathbf{X}\beta^* - \beta^* + (\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\mathring{\mathbf{X}}^T\epsilon\|^2 \tag{4}$$

$$= \|(\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\mathring{\mathbf{X}}^T\mathbf{X}\beta^* - \beta^*\|^2 + \mathbb{E}\|(\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\mathring{\mathbf{X}}^T\epsilon\|^2 \tag{5}$$

The first term is a bias-squared term and the second is a variance term. We will bound each one and then combine the bounds to bound the expected squared error.

3 Bias

To bound the bias, we will bound the matrix norm $\|(\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\mathring{\mathbf{X}}^T\mathbf{X} - I\|$ as follows:

$$\|(\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\mathring{\mathbf{X}}^T\mathbf{X} - I\| = \|(\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}(\mathring{\mathbf{X}}^T\mathbf{X} - \mathring{\mathbf{X}}^T\mathring{\mathbf{X}})\| \tag{6}$$

$$= \|(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\left((\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma) - (\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}} - \Sigma - \frac{1-\theta}{\theta}diag(\Sigma)) - \frac{1-\theta}{\theta}diag(\Sigma)\right)\| \tag{7}$$

$$\leq \|(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\|\left(\|\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma\| + \|\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}} - \Sigma - \frac{1-\theta}{\theta}diag(\Sigma)\| + \frac{1-\theta}{\theta}\|diag(\Sigma)\|\right) \tag{8}$$

We will bound each of these terms one by one in the following lemmas.

Lemma 1. *If $\theta = O(1)$, then with probability $1 - 2\delta$, we have that*

$$\|\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma\| \leq 2\frac{m}{\theta}\sqrt{\frac{\|\Sigma\|\log(p/\delta)}{n}} \tag{9}$$

N.B. If we have that $\|\mathring{\mathbf{X}}_i\| \leq m/\sqrt{\theta}$, then the θ, in the denominator may be replaced with $\sqrt{\theta}$.

Proof. We can rewrite the matrix difference as $\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma = \sum_{i=1}^n(\frac{1}{n}\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \frac{1}{n}\Sigma)$ and use the Matrix Bernstein inequality (see eg Tropp Cor 6.2.1). Before applying the inequality, we need to bound the norm of the summands and the variance parameter. For the norm, we have

$$\frac{1}{n}\|\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \Sigma\| \leq \frac{1}{n}(\|\mathring{\mathbf{X}}_i\mathbf{X}_i^T\| + \mathbb{E}\|\mathbf{X}_i\mathbf{X}_i^T\|) \leq \frac{1}{n}(m^2/\theta + m^2) = \frac{(1+\theta)m^2}{n\theta} \tag{10}$$

The variance parameter is defined as

$$\max\{\|\mathbb{E}[(\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma)^T(\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma)]\|, \|\mathbb{E}[(\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma)(\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma)^T]\|\} \tag{11}$$

$$= \max\{\|\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}(\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \Sigma)^T(\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \Sigma)\|, \|\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}(\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \Sigma)(\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \Sigma)^T\|\} \tag{12}$$

Beginning with the first term, we have

$$\mathbb{E}[(\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \Sigma)^T(\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \Sigma)] = \mathbb{E}[(\mathbf{X}_i\mathring{\mathbf{X}}_i^T = \Sigma)(\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \Sigma)] \tag{13}$$

$$= \mathbb{E}[\mathbf{X}_i\mathring{\mathbf{X}}_i^T\mathring{\mathbf{X}}_i\mathbf{X}_i] - \Sigma^2 \tag{14}$$

$$\preceq \mathbb{E}[\mathbf{X}_i\mathring{\mathbf{X}}_i^T\mathring{\mathbf{X}}_i\mathbf{X}_i] \preceq \frac{m^2}{\theta^2}\mathbb{E}[\mathbf{X}_i\mathbf{X}_i^T] = \frac{m^2}{\theta^2}\Sigma \tag{15}$$

where $\preceq$ denotes inequality in the sense of the positive semi-definite cone. In other words, $A \preceq B$ if and only if $B - A$ is positive semi-definite. Thus,

$$\|\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}(\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \Sigma)^T(\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \Sigma)\| \leq \frac{m^2}{n\theta^2}\|\Sigma\| \tag{16}$$

The second term is similar to the first:

$$\mathbb{E}[(\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \Sigma)(\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \Sigma)^T] = \mathbb{E}[(\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \Sigma)(\mathbf{X}_i\mathring{\mathbf{X}}_i^T - \Sigma)] \tag{17}$$

$$= \mathbb{E}[\mathring{\mathbf{X}}_i\mathbf{X}_i^T\mathbf{X}_i^T\mathring{\mathbf{X}}_i^T] - \Sigma^2 \tag{18}$$

$$\preceq m^2\mathbb{E}[\mathring{\mathbf{X}}_i\mathring{\mathbf{X}}_i^T] = m^2(\Sigma + \frac{1-\theta}{\theta}diag(\Sigma)) \tag{19}$$

Thus

$$|\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}(\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \Sigma)(\mathring{\mathbf{X}}_i\mathbf{X}_i^T - \Sigma)^T\| \leq \frac{m^2}{n}\|\Sigma + \frac{1-\theta}{\theta}diag(\Sigma)\| \leq \frac{m^2}{n\theta}\|\Sigma\| \tag{20}$$

The last inequality follows from the Schur-Horn Theorem, which says that for a Hermitian matrix, $\max_j \Sigma_{jj} \leq \max_i \lambda_i(\Sigma)$, where λ_i denotes the eigenvalues. Then it follows that $\|\Sigma + \frac{1-\theta}{\theta} diag(\Sigma)\| \leq (1 + \frac{1-\theta}{\theta})\|\Sigma\|$.

Looking at these 2 terms together, we have that the variance parameter is

$$\max\{\|\mathbb{E}[(\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma)^T(\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma)]\|, \|\mathbb{E}[(\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma)(\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma)^T]\|\} \tag{21}$$

$$\leq \max\{\frac{m^2}{n\theta^2}\|\Sigma\|, \frac{m^2}{n\theta}\|\Sigma\|\} = \frac{m^2}{n\theta^2}\|\Sigma\| \tag{22}$$

Plugging this into the Matrix Bernstein inequality gives

$$\mathbb{P}(\|\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma\| > t) \leq 2p \exp\left(-\frac{t^2/2}{\frac{m^2}{n\theta^2}\|\Sigma\| + \frac{(1+\theta)m^2 t}{3n\theta}}\right) \tag{23}$$

We have that $\frac{m^2}{n\theta^2}\|\Sigma\| > \frac{(1+\theta)m^2 t}{3n\theta}$ as long as $t \leq \frac{3\|\Sigma\|}{\theta(1+\theta)}$. Requiring $\theta = O(1)$ is sufficient to satisfy this inequality for the values of t that we are interested in. In this case it follows that,

$$\mathbb{P}(\|\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma\| > t) \leq 2p \exp\left(-\frac{t^2/2}{2\frac{m^2}{n\theta^2}\|\Sigma\|}\right) \tag{24}$$

Setting the right hand side equal to 2δ gives the result. □

If we think of the typical case where $m = O(\sqrt{p})$, this lemma says that

$$\|\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma\| = O_P\left(\frac{1}{\theta}\sqrt{\frac{p \log p}{n}\|\Sigma\|}\right) \tag{25}$$

The extra $\log p$ term can be eliminated if we assume that the rows of $\mathbf{X}$ are subgaussian, rather than simply bounded.

Lemma 2. *Suppose $m = O(\sqrt{p})$ and $\theta = O(1)$. With probability $1 - \delta$*

$$\|\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}} - \Sigma - \frac{1-\theta}{\theta}diag(\Sigma)\| \leq c\frac{m}{\theta^{3/2}}\sqrt{\frac{\|\Sigma\| \log p/\delta}{n}} = O\left(\frac{1}{\theta^{3/2}}\sqrt{\frac{p \log p}{n}\|\Sigma\|}\right) \tag{26}$$

where $c > 0$ is an absolute constant.

N.B. If we have that $\|\mathring{\mathbf{X}}_i\| \leq m/\sqrt{\theta}$, then the $\theta^{3/2}$, in the denominator may be replaced with θ.

Proof. This is simply a bound on the difference between the sample covariance and the true covariance for the case where the data is bounded but not necessarily subgaussian. Then Theorem 5.44 of Vershynin implies that with probability $1 - pe^{-ct^2}$

$$\|\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}} - \Sigma - \frac{1-\theta}{\theta}diag(\Sigma)\| \leq \max\left(t\|\Sigma + \frac{1-\theta}{\theta}diag(\Sigma)\|^{1/2}\frac{m}{\theta\sqrt{n}}, \frac{t^2 m^2}{\theta^2 n}\right) \tag{27}$$

Setting $t = \frac{\sqrt{\log(p/\delta)}}{c}$, we see that the first term is larger as long as

$$\frac{m}{\theta}\sqrt{\frac{\log p}{n}} < \|\Sigma - \frac{1-\theta}{\theta}diag(\Sigma)\|^{1/2} \tag{28}$$

$m = O(\sqrt{p})$ and $\theta = O(1)$ suffices to ensure this. Then the conclusion follows by recalling that $\|\Sigma - \frac{1-\theta}{\theta}diag(\Sigma)\| \leq \frac{1}{\theta}\|\Sigma\|$. □

The final step we need is a lower bound on the smallest eigenvalue of $\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}$. While we can get an additive bound using Weyl's inequality, the following multiplicative bound will be easier to use.

Lemma 3. *Let $\mu_{\min} = \lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}diag(\Sigma))$. Then with probability $1-\delta$, we have*

$$\lambda_{\min}(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}) > \left(1 - c\frac{m}{\theta}\sqrt{\frac{\log(p/\delta)}{n\mu_{\min}}}\right)\mu_{\min} \tag{29}$$

where $c = \sqrt{2}$.

N.B. If we have that $\|\mathring{\mathbf{X}}_i\| \leq m/\sqrt{\theta}$, then the θ, in the denominator may be replaced with $\sqrt{\theta}$.

Proof. We will use the Matrix Chernoff Inequality (see eg Tropp Theorem 5.1.1). We can rewrite $\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^n \mathring{\mathbf{X}}_i\mathring{\mathbf{X}}_i^T$. Each term in the sum is positive semi-definite and is bounded as

$$\|\frac{1}{n}\mathring{\mathbf{X}}_i\mathring{\mathbf{X}}_i^T\| = \frac{1}{n}\|\mathring{\mathbf{X}}_i\|^2 \leq \frac{m^2}{n\theta^2} \tag{30}$$

Using the weaker version of the bound suggested in Tropp, we have that

$$\mathbb{P}(\lambda_{\min}(\frac{1}{n}\mathring{\mathbf{X}}_i\mathring{\mathbf{X}}_i^T) \leq (1-t)\lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}diag(\Sigma))) \leq p\exp\left(\frac{-t^2\mu_{\min}}{2\frac{m^2}{n\theta^2}}\right) \tag{31}$$

Setting the right hand side equal to δ and solving for t gives the result. □

3.1 Putting this together

Now we can combine the results above to bound the bias term. For conciseness we will write the lower bound for $\lambda_{\min}(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})$ as $(1-\gamma)\lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}diag(\Sigma))$. Then combining the lemmas above we have that with probability $1-4\delta$.

$$\|(\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\mathring{\mathbf{X}}^T\mathbf{X} - I\| \tag{32}$$

$$\leq \|(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\|\left(\|\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma\| + \|\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}} - \Sigma - \frac{1-\theta}{\theta}diag(\Sigma)\| + \frac{1-\theta}{\theta}\|diag(\Sigma)\|\right) \tag{33}$$

$$\leq \frac{2\frac{m}{\theta}\sqrt{\frac{\|\Sigma\|\log(p/\delta)}{n}} + c\frac{m}{\theta^{3/2}}\sqrt{\frac{\|\Sigma\|\log(p/\delta)}{n}} + \frac{1-\theta}{\theta}\|diag(\Sigma)\|}{(1-\gamma)\lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}diag(\Sigma))} \tag{34}$$

$$= O\left(\frac{\frac{m}{\theta^{3/2}}\sqrt{\frac{\|\Sigma\|\log(p/\delta)}{n}}}{(1-\gamma)\lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}diag(\Sigma))}\right) + O\left(\frac{\frac{1-\theta}{\theta}\|diag(\Sigma)\|}{(1-\gamma)\lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}diag(\Sigma))}\right) \tag{35}$$

We restate this in the following theorem.

Theorem 4. *Suppose $m = O(\sqrt{p})$ and $\theta = O(1)$. Then the bias is bounded as*

$$\|(\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\mathring{\mathbf{X}}^T\mathbf{X}\beta^* - \beta^*\| \tag{36}$$

$$= O_P\left(\frac{\frac{1}{\theta^{3/2}}\sqrt{\frac{\|\Sigma\|p\log(p/\delta)}{n}}}{(1-\gamma)\lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}diag(\Sigma))}\|\beta^*\|\right) + O_P\left(\frac{\frac{1-\theta}{\theta}\|diag(\Sigma)\|}{(1-\gamma)\lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}diag(\Sigma))}\|\beta^*\|\right) \tag{37}$$

where $\gamma = O_P(\frac{1}{\theta}\sqrt{\frac{p\log(p/\delta)}{n\lambda_{\min}(\Sigma+\frac{1-\theta}{\theta}diag(\Sigma)}})$. This can be further loosened to get

$$\|(\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\mathring{\mathbf{X}}^T\mathbf{X}\beta^*-\beta^*\| = O_P\left(\frac{\kappa(\Sigma)}{(1-\gamma)\theta^{3/2}}\sqrt{\frac{p\log(p/\delta)}{n\|\Sigma\|}}\|\beta^*\|\right)+O_P\left(\frac{\frac{1-\theta}{\theta}\|diag(\Sigma)\|}{(1-\gamma)\lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}diag(\Sigma))}\|\beta^*\|\right) \tag{38}$$

where $\kappa(\cdot)$ denotes the condition number of a matrix.

Proof. The first equation follows from the comments above. The second follows by noticing that

$$\frac{\sqrt{\|\Sigma\|}}{\lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}diag(\Sigma))} \leq \frac{\sqrt{\|\Sigma\|}}{\lambda_{\min}(\Sigma)} = \frac{\kappa(\Sigma)}{\sqrt{\|\Sigma\|}} \tag{39}$$

□

4 Variance

This is similar to the usual variance calculation and relies mostly on Lemma 3 above.

Theorem 5. *Let $\mu_{\min} = \lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}diag(\Sigma))$. With probability $1-\delta$, the variance is bounded by*

$$\mathbb{E}\|(\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\mathring{\mathbf{X}}^T\epsilon\|^2 \leq \frac{\sigma^2 p}{n}\frac{1}{\left(1 - c\frac{m}{\theta}\sqrt{\frac{\log(p/\delta)}{n\mu_{\min}}}\right)\lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}diag(\Sigma))} \tag{40}$$

Proof. Rewriting the norm as a trace we have,

$$\mathbb{E}\|(\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\mathring{\mathbf{X}}^T\epsilon\|^2 = \mathbb{E}[\epsilon^T\mathring{\mathbf{X}}(\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}(\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\mathring{\mathbf{X}}^T\epsilon] \tag{41}$$

$$= \sigma^2 tr((\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}(\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\mathring{\mathbf{X}}^T\mathring{(\mathbf{X})}) \tag{42}$$

$$= \sigma^2 tr((\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}) \tag{43}$$

$$= \frac{\sigma^2}{n}tr((\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}) \tag{44}$$

$$\leq \frac{\sigma^2 p}{n\lambda_{\min}(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})} \tag{45}$$

□

5 Linear Regression Result

We can combine the two theorems to get the following result for the squared error.

Theorem 6. *Suppose* $m = O(\sqrt{p})$ *and* $\theta = O(1)$. *Use the notation* $1-\gamma = 1-c\frac{m}{\theta}\sqrt{\frac{\log(p/\delta)}{n\lambda_{\min}(\Sigma+\frac{1-\theta}{\theta}diag(\Sigma))}}$. *Then with probability* $1-5\delta$ *we have that*

$$\mathbb{E}\|\widehat{\beta}-\beta^*\|^2 \leq O_P\left(\frac{\frac{1}{\theta^3}\frac{\|\Sigma\|p\log(p/\delta)}{n}}{(1-\gamma)^2\lambda_{\min}^2(\Sigma+\frac{1-\theta}{\theta}diag(\Sigma))}\|\beta^*\|^2\right) + O_P\left(\left(\frac{\frac{1-\theta}{\theta}\|diag(\Sigma)\|}{(1-\gamma)\lambda_{\min}(\Sigma+\frac{1-\theta}{\theta}diag(\Sigma))}\right)^2\|\beta^*\|^2\right) \tag{46}$$

$$+\frac{\sigma^2 p}{n}\frac{1}{(1-\gamma)\lambda_{\min}(\Sigma+\frac{1-\theta}{\theta}diag(\Sigma))} \tag{47}$$

6 Ridge Regression

It is straightforward to modify these arguments for the case of ridge regression. Define the sparsified ridge estimator

$$\widehat{\beta}_\lambda = (\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}+\lambda I)^{-1}\frac{1}{n}\mathring{\mathbf{X}}^T Y \tag{48}$$

As before, we can separate the expected squared error into a bias and a variance term:

$$\mathbb{E}\|\widehat{\beta}_\lambda-\beta^*\|^2 = \mathbb{E}\|(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}+\lambda I)^{-1}\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X}\beta^*-\beta^*+(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}+\lambda I)^{-1}\frac{1}{n}\mathring{\mathbf{X}}^T\epsilon\|^2 \tag{49}$$

$$= \|(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}+\lambda I)^{-1}\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X}\beta^*-\beta^*\|^2 + \mathbb{E}\|(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}+\lambda I)^{-1}\frac{1}{n}\mathring{\mathbf{X}}^T\epsilon\|^2 \tag{50}$$

For the bias, we again begin by bounding the matrix norm:

$$\|(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}+\lambda I)^{-1}\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X}-I\| = \|(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}+\lambda I)^{-1}(\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X}-\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}-\lambda I)\| \tag{51}$$

$$\leq \|(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}+\lambda I)^{-1}\|\left(\|\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X}-\Sigma\|+\|\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}-\Sigma-\frac{1-\theta}{\theta}diag(\Sigma)\|+\frac{1-\theta}{\theta}\|diag(\Sigma)\|+\lambda\right) \tag{52}$$

We already have bounds for all but the first term, which we can easily bound by noting that $\lambda_{\min}(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}+\lambda I) = \lambda_{\min}(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})+\lambda$. Thus it follows that

$$\|(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}+\lambda I)^{-1}\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X}\beta^*-\beta^*\| \tag{53}$$

$$= O_P\left(\frac{\frac{1}{\theta^{3/2}}\sqrt{\frac{\|\Sigma\|p\log(p/\delta)}{n}}}{(1-\gamma)\lambda_{\min}(\Sigma+\frac{1-\theta}{\theta}diag(\Sigma))+\lambda}\|\beta^*\|\right) + O_P\left(\frac{\frac{1-\theta}{\theta}\|diag(\Sigma)\|+\lambda}{(1-\gamma)\lambda_{\min}(\Sigma+\frac{1-\theta}{\theta}diag(\Sigma))+\lambda}\|\beta^*\|\right) \tag{54}$$

For the variance we have

$$\mathbb{E}\|(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}+\lambda I)^{-1}\frac{1}{n}\mathring{\mathbf{X}}^T\epsilon\|^2 = \mathbb{E}[\epsilon^T\frac{1}{n}\mathring{\mathbf{X}}(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}+\lambda I)^{-1}(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}+\lambda I)^{-1}\frac{1}{n}\mathring{\mathbf{X}}^T\epsilon] \tag{55}$$

$$= \sigma^2 tr((\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}}+\lambda)^{-2}\frac{1}{n^2}\mathring{\mathbf{X}}^T\mathring{(\mathbf{X})}) \tag{56}$$

$$= \frac{\sigma^2}{n}\sum_{j=1}^{p}\frac{\lambda_j(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})}{(\lambda_j(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})+\lambda)^2} \tag{57}$$

$$\leq \frac{\sigma^2}{n}\sum_{j=1}^{p}\frac{1}{\lambda_j(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})+\lambda} \tag{58}$$

$$\leq \frac{\sigma^2 p}{n(\lambda_{\min}(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})+\lambda)} \tag{59}$$

Putting the bias and variance terms together gives the following result:

Theorem 7. *Suppose* $m = O(\sqrt{p})$ *and* $\theta = O(1)$. *Use the notation* $1-\gamma = 1-c\frac{m}{\theta}\sqrt{\frac{\log(p/\delta)}{n\lambda_{\min}(\Sigma+\frac{1-\theta}{\theta}diag(\Sigma))}}$. *Then with probability* $1-5\delta$ *we have that*

$$\mathbb{E}\|\widehat{\beta}-\beta^*\|^2 \leq O_P\left(\frac{\frac{1}{\theta^3}\frac{\|\Sigma\|p\log(p/\delta)}{n}}{((1-\gamma)\lambda_{\min}(\Sigma+\frac{1-\theta}{\theta}diag(\Sigma))+\lambda)^2}\|\beta^*\|^2\right) \tag{60}$$

$$+ O_P\left(\left(\frac{\frac{1-\theta}{\theta}\|diag(\Sigma)\|+\lambda}{(1-\gamma)\lambda_{\min}(\Sigma+\frac{1-\theta}{\theta}diag(\Sigma))+\lambda}\right)^2\|\beta^*\|^2\right) \tag{61}$$

$$+ \frac{\sigma^2 p}{n}\frac{1}{(1-\gamma)\lambda_{\min}(\Sigma+\frac{1-\theta}{\theta}diag(\Sigma))+\lambda} \tag{62}$$

7 Now with better normalization

As we discussed on Friday, dividing by θ is maybe not actually the correct thing to do. In fact, I did some basic simulations and saw quite a dramatic difference, with not normalizing being much better. So you're right. Now let's try and redo this analysis, but without normalization. Now ideally I would redefine $\mathring{\mathbf{X}}$ to be what you want it to be. But I'm going to leverage some of the stuff I already did, and I don't have the time right now to redo all of that notation. So instead I'll define

$$\widetilde{\mathbf{X}} = \mathbf{X} \circ Z' \qquad where Z'_{ij} \overset{iid}{\sim} Ber(\theta_j) \tag{63}$$

In other words, $\widetilde{\mathbf{X}}_{ij}$ is either $\mathbf{X}_{ij}$ or 0. Assume as before that $\theta_j = \theta$ is constant across variables. Then $\widetilde{\mathbf{X}} = \theta\mathring{\mathbf{X}}$. We also have that $EE[\widetilde{\mathbf{X}}|\mathbf{X}] = \theta\mathbf{X}$ and

$$\frac{1}{n}\mathbb{E}[\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}] = \theta\text{diag}(\Sigma) + \theta^2\text{offidiag}(\Sigma) \tag{64}$$

$$= \theta\Sigma + (\theta^2 - \theta)\text{offdiag}(\Sigma) \tag{65}$$

$$= \theta^2\Sigma + \theta(1-\theta)\text{diag}(\Sigma) \tag{66}$$

We will use both expressions for the sample covariance in deriving the bound. Let

$$\widetilde{\beta} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T Y \tag{67}$$

be the estimator we are interested in. Similarly to before, after taking an expectation over the noise ϵ, we have that the squared error is

$$\mathbb{E}\|\widetilde{\beta} - \beta^*\|^2 = \|(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{X}\beta^* - \beta^*\|^2 + \mathbb{E}\|(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\epsilon\|^2 \tag{68}$$

7.1 Bias

We will proceed similarly to before, by bounding the matrix norm. We do this by adding and subtracting the expected values of the various terms.

$$\|(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{X} - I\| = \|(\frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}(\frac{1}{n}\widetilde{\mathbf{X}}^T\mathbf{X} - \frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})\| \tag{69}$$

$$\leq \|(\frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\|\|(\frac{1}{n}\widetilde{\mathbf{X}}^T\mathbf{X} - \theta\Sigma) + \theta\Sigma - (\frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} - \theta^2\Sigma - \theta(1-\theta)\text{diag}(\Sigma)) - \theta^2\Sigma - \theta(1-\theta)\text{diag}(\Sigma)\| \tag{70}$$

$$\leq \|(\frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\|\left(\|\frac{1}{n}\widetilde{\mathbf{X}}^T\mathbf{X} - \theta\Sigma\| + \|\frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} - \theta^2\Sigma - \theta(1-\theta)\text{diag}(\Sigma)\| + \|\theta(1-\theta)\Sigma - \theta(1-\theta)\text{diag}(\Sigma)\|\right) \tag{71}$$

The first 2 terms are nearly the same, so we'll dispense with them quickly. For the first term,

$$\|(\frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\|\|\frac{1}{n}\widetilde{\mathbf{X}}^T\mathbf{X} - \theta\Sigma\| = \frac{1}{\theta}\|(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\|\|\frac{1}{n}\mathring{\mathbf{X}}^T\mathbf{X} - \Sigma\| \tag{72}$$

$$= O_P\left(\frac{\frac{m}{\theta^2}\sqrt{\frac{\|\Sigma\|\log(p/\delta)}{n}}}{(1-\gamma)\lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}\text{diag}(\Sigma))}\right) \tag{73}$$

Note that the dependence on θ is now like $\frac{1}{\theta^2}$, while previously we had $\frac{1}{\theta}$. Similarly,

$$\|(\frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\|\|\frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} - \theta^2\Sigma - \theta(1-\theta)\text{diag}(\Sigma)\| = \|(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\|\|\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}} - \Sigma - \frac{1-\theta}{\theta}diag(\Sigma)\| \tag{74}$$

$$= O\left(\frac{\frac{m}{\theta^{3/2}}\sqrt{\frac{\|\Sigma\|\log(p/\delta)}{n}}}{(1-\gamma)\lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}diag(\Sigma))}\right) \tag{75}$$

Now we consider the final term. This is the term that has changed and which provides the improvement. Without loss of generality, we will assume that $\text{diag}(\Sigma) = I$.

$$\|(\frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\|\|(\theta)(1-\theta)(\Sigma - \text{diag}(\Sigma))\| = \frac{\theta(1-\theta)}{\theta^2}\|(\frac{1}{n}\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\|\|\Sigma - I\| \tag{76}$$

$$\leq \frac{\theta(1-\theta)\max\{\lambda_{\max}(\Sigma) - 1, 1 - \lambda_{\min}(\Sigma)\}}{\theta^2(1-\gamma)\lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}diag(\Sigma))} \tag{77}$$

$$= \frac{(1-\theta)\max\{\lambda_{\max}(\Sigma) - 1, 1 - \lambda_{\min}(\Sigma)\}}{\theta(1-\gamma)\lambda_{\min}(\Sigma + \frac{1-\theta}{\theta}I)} \tag{78}$$

$$= \frac{(1-\theta)\max\{\lambda_{\max}(\Sigma) - 1, 1 - \lambda_{\min}(\Sigma)\}}{(1-\gamma)(\theta\lambda_{\min}(\Sigma) + 1 - \theta)} \tag{79}$$

In the previous analysis, the numerator was replaced by $(1-\theta)\|diag(\Sigma)\| = 1 - \theta$.

7.2 Final Result

Theorem 8. *Suppose $m = O(\sqrt{p})$ and θ is constant, or at least not dependent on n and p [1]. Assume that $diag(\Sigma) = I$. Then the squared bias is bounded as*

$$\|(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{X}\beta^* - \beta^*\| \tag{80}$$

$$= O_P\left(\frac{\frac{\|\Sigma\|}{\theta^2}\frac{p\log(p/\delta)}{n}}{(1-\gamma)^2(\theta\lambda_{\min}(\Sigma) + 1 - \theta)^2}\|\beta^*\|^2\right) + O_P\left(\frac{(1-\theta)^2\max\{(\lambda_{\max}(\Sigma) - 1)^2, (1 - \lambda_{\min}(\Sigma))^2\}}{(1-\gamma)^2(\theta\lambda_{\min}(\Sigma) + 1 - \theta)^2}\|\beta^*\|\right) \tag{81}$$

The variance is bounded as

$$\mathbb{E}\|(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\epsilon\|^2 \leq \frac{\sigma^2 p}{n}\frac{1}{(1-\gamma)(\theta^2\lambda_{\min}(\Sigma) + \theta(1-\theta))} \tag{82}$$

The dependence on θ in the first term of the bias squared is like $\frac{1}{\theta^4}$, while before it was like $\frac{1}{\theta^3}$. However, the bound on the second term is much better. So we have traded a slightly worse dependence on θ for a better overall bound.

[1] There is some amount of dependence that can be allowed and at some point I should work out what that is

Proof. We have simply written the squared bias as the sum of the squares of the first term and the third term. We do not include the second term because it is dominated by the first. We have also simplified the denominator of the first big-O term by using the fact that under our assumptions

$$(1-\gamma)^2\lambda_{\min}^2(\Sigma+\frac{1-\theta}{\theta}\text{diag}(\Sigma)) = (1-\gamma)^2\frac{1}{\theta^2}(\theta\lambda_{\min}(\Sigma)+1-\theta)^2 \tag{83}$$

and canceling the θ^2.

The claim for the variance follows because

$$\mathbb{E}\|(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\epsilon\|^2 = \frac{1}{\theta^2}\mathbb{E}\|(\mathring{\mathbf{X}}^T\mathring{\mathbf{X}})^{-1}\mathring{\mathbf{X}}^T\epsilon\|^2 \tag{84}$$

$$\leq \frac{\sigma^2 p}{n}\frac{1}{(1-\gamma)\theta^2\lambda_{\min}(\Sigma+\frac{1-\theta}{\theta}diag(\Sigma))} \tag{85}$$

□