Appendix to
An Iterative, Sketching-based Framework for Ridge Regression

A. Preliminary Results

We start by reviewing a result regarding the convergence of a matrix *von Neumann* series for $(\mathbf{I}-\mathbf{P})^{-1}$. This will be an important tool in our analysis.

Proposition 8. *Let* $\mathbf{P}$ *be any square matrix with* $\|\mathbf{P}\|_2 < 1$. *Then* $(\mathbf{I}-\mathbf{P})^{-1}$ *exists and*

$$(\mathbf{I}-\mathbf{P})^{-1} = \mathbf{I} + \sum_{\ell=1}^{\infty} \mathbf{P}^{\ell}.$$

Next, we state and prove another fundamental result. This provides an alternative formulation of the ridge regression solution vector, which will be one of our primary building blocks. The result has previously appeared in Saunders et al. (1998), but we provide a proof here for completeness.

Lemma 9. (Saunders et al., 1998) *Let* $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{b} \in \mathbb{R}^n$, *and* $\lambda > 0$ *be the inputs of the ridge regression problem. The solution to eqn.* (1) *can also be expressed as*

$$\mathbf{x}^* = \mathbf{A}^{\mathsf{T}} \left(\mathbf{A}\mathbf{A}^{\mathsf{T}} + \lambda \mathbf{I}_n\right)^{-1} \mathbf{b}.$$

Proof. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}_f\mathbf{V}_f^{\mathsf{T}}$ be the full SVD representation of $\mathbf{A}$ with $\mathbf{U}\mathbf{U}^{\mathsf{T}} = \mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}_n$ and $\mathbf{V}_f\mathbf{V}_f^{\mathsf{T}} = \mathbf{V}_f^{\mathsf{T}}\mathbf{V}_f = \mathbf{I}_d$. Further, $\mathbf{\Sigma}_f = \begin{pmatrix}\mathbf{\Sigma} & \mathbf{0}\end{pmatrix} \in \mathbb{R}^{n\times d}$ and $\mathbf{V}_f = \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix}$, where $\mathbf{\Sigma}$ and $\mathbf{V}$ are as described in Section 1.3. Additionally, $\mathbf{V}_\perp$ consists the bottom $d-n$ columns of $\mathbf{V}_f$. Note that $\mathbf{U}$ remains the same in both the thin as well as full SVD representations, since we assume the design matrix $\mathbf{A}$ to have full row-rank.

Under this setup, we have

$$\mathbf{A}^{\mathsf{T}}\mathbf{A} + \lambda\mathbf{I}_d = \mathbf{V}_f\mathbf{\Sigma}_f^{\mathsf{T}}\mathbf{U}^{\mathsf{T}}\mathbf{U}\mathbf{\Sigma}_f\mathbf{V}_f^{\mathsf{T}} + \lambda\mathbf{V}_f\mathbf{V}_f^{\mathsf{T}} = \mathbf{V}_f\left(\mathbf{\Sigma}_f^{\mathsf{T}}\mathbf{\Sigma}_f + \lambda\mathbf{I}_d\right)\mathbf{V}_f^{\mathsf{T}},$$

where we used the fact that $\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}_n$. Now, we can rewrite eqn. (2) as

$$\begin{aligned}\mathbf{x}^* &= \left(\mathbf{A}^{\mathsf{T}}\mathbf{A} + \lambda\mathbf{I}_d\right)^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{b} = \left[\mathbf{V}_f\left(\mathbf{\Sigma}_f^{\mathsf{T}}\mathbf{\Sigma}_f + \lambda\mathbf{I}_d\right)\mathbf{V}_f^{\mathsf{T}}\right]^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{b} \\ &= \mathbf{V}_f\left(\mathbf{\Sigma}_f^{\mathsf{T}}\mathbf{\Sigma}_f + \lambda\mathbf{I}_d\right)^{-1}\mathbf{V}_f^{\mathsf{T}}\mathbf{V}_f\mathbf{\Sigma}_f^{\mathsf{T}}\mathbf{U}^{\mathsf{T}}\mathbf{b} = \mathbf{V}_f\left(\mathbf{\Sigma}_f^{\mathsf{T}}\mathbf{\Sigma}_f + \lambda\mathbf{I}_d\right)^{-1}\mathbf{\Sigma}_f^{\mathsf{T}}\mathbf{U}^{\mathsf{T}}\mathbf{b},\end{aligned} \tag{29}$$

where we noticed that $(\mathbf{\Sigma}_f^{\mathsf{T}}\mathbf{\Sigma}_f + \lambda\mathbf{I}_d)^{-1}$ exists since $\mathbf{\Sigma}_f^{\mathsf{T}}\mathbf{\Sigma}_f + \lambda\mathbf{I}_d$ is a diagonal matrix with non-zero entries.

From eqn. (29), we further have

$$\left(\mathbf{\Sigma}_f^{\mathsf{T}}\mathbf{\Sigma}_f + \lambda\mathbf{I}_d\right)^{-1}\mathbf{\Sigma}_f^{\mathsf{T}} = \begin{pmatrix}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_n)^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\lambda}\mathbf{I}_{d-n}\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma} \\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_n)^{-1}\mathbf{\Sigma} \\ \mathbf{0}\end{pmatrix}, \tag{30}$$

where $\mathbf{0}$'s denote null matrices with compatible dimensions.

Combining eqn. (29) and eqn. (30), we obtain

$$\mathbf{x}^* = \mathbf{V}_f\begin{pmatrix}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_n)^{-1}\mathbf{\Sigma} \\ \mathbf{0}\end{pmatrix}\mathbf{U}^{\mathsf{T}}\mathbf{b} = \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix}\begin{pmatrix}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_n)^{-1}\mathbf{\Sigma} \\ \mathbf{0}\end{pmatrix}\mathbf{U}^{\mathsf{T}}\mathbf{b}$$

$$= \mathbf{V}(\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n)^{-1}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\mathbf{b} = \ (\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T})\mathbf{U}\mathbf{\Sigma}^{-1}(\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n)^{-1}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\mathbf{b}$$
$$= \mathbf{A}^\mathsf{T}\mathbf{U}(\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n)^{-1}\mathbf{U}^\mathsf{T}\mathbf{b} = \ \mathbf{A}^\mathsf{T}\left[\mathbf{U}(\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n)\mathbf{U}^\mathsf{T}\right]^{-1}\mathbf{b}$$
$$= \mathbf{A}^\mathsf{T}\left[\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\mathsf{T} + \lambda \mathbf{U}\mathbf{U}^\mathsf{T}\right]^{-1}\mathbf{b} = \ \mathbf{A}^\mathsf{T}\left(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n\right)^{-1}\mathbf{b},$$

where we used the facts that $\mathbf{\Sigma}^{-1}(\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n)^{-1}\mathbf{\Sigma} = (\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n)^{-1}$ and that $\mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\mathsf{T}$ by the thin SVD of $\mathbf{A}$. This completes the proof. □

B. Proof of Theorem 1

The overall proof strategy is similar to that of Theorem 2 (see Section 3). In terms of algebraic manipulation, this proof is simpler as the final bound does not involve any additive term. We begin by providing an alternative expression of $\mathbf{x}^{*(j)}$ that is easier to work with.

Lemma 10. *For $j = 1, 2, \ldots, t$, let $\mathbf{b}^{(j)}$ be the intermediate response vectors in Algorithm 1 and $\mathbf{x}^{*(j)}$ be the vector defined in eqn. (15). Then for any $j = 1, 2, \ldots, t$, $\mathbf{x}^{*(j)}$ can also be expressed as*

$$\mathbf{x}^{*(j)} = \mathbf{V}\mathbf{G}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)},$$

where $\mathbf{G} = \mathbf{I}_n + \lambda\mathbf{\Sigma}^{-2}$.

Proof. Setting $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$ in eqn. (3), we have

$$\begin{aligned}\mathbf{x}^{*(j)} &= \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\left(\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\mathsf{T} + \lambda\mathbf{U}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{b}^{(j)} = \mathbf{V}\mathbf{\Sigma}\left(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_n\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \\ &= \mathbf{V}\mathbf{\Sigma}\left(\mathbf{\Sigma}\left(\mathbf{I}_n + \lambda\mathbf{\Sigma}^{-2}\right)\mathbf{\Sigma}\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} = \mathbf{V}\mathbf{\Sigma}\left(\mathbf{\Sigma}\mathbf{G}\mathbf{\Sigma}\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \\ &= \mathbf{V}\mathbf{G}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)}, \end{aligned} \tag{31}$$

where we note that $\mathbf{G}^{-1}$ exists. This completes the proof. □

Our next result expresses the intermediate vectors $\widetilde{\mathbf{x}}^{(j)}$ of Algorithm 1 in terms of the vectors $\mathbf{x}^{*(j)}$.

Lemma 11. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{b} \in \mathbb{R}^n$, and $\lambda > 0$ be the inputs of the ridge regression problem and $\mathbf{G}$ is as defined in Lemma 10. Further, let $\mathbf{S} \in \mathbb{R}^{d\times s}$ be the sketching matrix and define,*

$$\widehat{\mathbf{E}} = \mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V} - \mathbf{I}_n.$$

If the constraint of eqn. (6) is satisfied i.e. $\|\widehat{\mathbf{E}}\|_2 < 1$, *then for all $j = 1, \ldots, t$,*

$$\widetilde{\mathbf{x}}^{(j)} = \ \mathbf{x}^{*(j)} + \mathbf{V}\widehat{\mathbf{R}}\mathbf{G}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)},$$

where $\widehat{\mathbf{R}} = \sum_{\ell=1}^{\infty}(-1)^\ell(\mathbf{G}^{-1}\widehat{\mathbf{E}})^\ell$.

Proof. Denote $\mathbf{W} = \mathbf{S}\mathbf{S}^\mathsf{T}$. Using the thin SVD of $\mathbf{A}$, we can rewrite $\widetilde{\mathbf{x}}^{(j)}$ as follows:

$$\begin{aligned}\widetilde{\mathbf{x}}^{(j)} &= \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T} + \lambda\mathbf{U}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{b}^{(j)} \\ &= \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\left(\mathbf{U}\mathbf{\Sigma}\left(\mathbf{I}_n + \widehat{\mathbf{E}}\right)\mathbf{\Sigma}\mathbf{U}^\mathsf{T} + \lambda\mathbf{U}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{b}^{(j)} \\ &= \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\left(\mathbf{U}\mathbf{\Sigma}\left(\mathbf{I}_n + \widehat{\mathbf{E}} + \lambda\mathbf{\Sigma}^{-2}\right)\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{b}^{(j)} \\ &= \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{\Sigma}^{-1}\left(\mathbf{I}_n + \widehat{\mathbf{E}} + \lambda\mathbf{\Sigma}^{-2}\right)^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \qquad (32) \\ &= \mathbf{V}\left(\mathbf{I}_n + \widehat{\mathbf{E}} + \lambda\mathbf{\Sigma}^{-2}\right)^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)}, \qquad (33)\end{aligned}$$

where in the second equality we used the fact that $\widehat{\mathbf{E}} = \mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V} - \mathbf{I}_n$. Furthermore, we note that $(\mathbf{I}_n + \widehat{\mathbf{E}} + \lambda\mathbf{\Sigma}^{-2})^{-1}$ exists since $\mathbf{I}_n + \widehat{\mathbf{E}} = \mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}$ is positive semidefinite and $\lambda\mathbf{\Sigma}^{-2}$ is positive definite ($\lambda > 0$).

Proceeding further with eqn. (33), we have

$$\begin{aligned}\widetilde{\mathbf{x}}^{(j)} &= \mathbf{V}\left(\mathbf{I}_n + \widehat{\mathbf{E}} + \lambda\boldsymbol{\Sigma}^{-2}\right)^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} = \mathbf{V}\left(\mathbf{G} + \widehat{\mathbf{E}}\right)^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \\ &= \mathbf{V}\left(\mathbf{G}\left(\mathbf{I}_n + \mathbf{G}^{-1}\widehat{\mathbf{E}}\right)\right)^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)}. \end{aligned} \tag{34}$$

Notice that, since $\|\widehat{\mathbf{E}}\|_2 < 1$, we have

$$\left\|\mathbf{G}^{-1}\widehat{\mathbf{E}}\right\|_2 \leq \left\|\mathbf{G}^{-1}\right\|_2\left\|\widehat{\mathbf{E}}\right\|_2 < \left\|\mathbf{G}^{-1}\right\|_2 = \frac{\sigma_1^2}{\sigma_1^2+\lambda} \leq 1\,. \tag{35}$$

Thus, taking $\mathbf{P} = -\mathbf{G}^{-1}\widehat{\mathbf{E}}$ in Proposition 8 implies that $(\mathbf{I}_n + \mathbf{G}^{-1}\widehat{\mathbf{E}})^{-1}$ exists and

$$\left(\mathbf{I}_n + \mathbf{G}^{-1}\widehat{\mathbf{E}}\right)^{-1} = \mathbf{I}_n + \sum_{\ell=1}^{\infty}(-1)^\ell\left(\mathbf{G}^{-1}\widehat{\mathbf{E}}\right)^\ell = \mathbf{I}_n + \widehat{\mathbf{R}}\,. \tag{36}$$

Finally, combining eqns. (34) and (36), we have

$$\begin{aligned}\widetilde{\mathbf{x}}^{(j)} &= \mathbf{V}\left(\mathbf{I}_n + \mathbf{G}^{-1}\widehat{\mathbf{E}}\right)^{-1}\mathbf{G}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} = \mathbf{V}\left(\mathbf{I}_n + \widehat{\mathbf{R}}\right)\mathbf{G}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \\ &= \mathbf{V}\mathbf{G}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} + \mathbf{V}\widehat{\mathbf{R}}\mathbf{G}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} = \mathbf{x}^{*(j)} + \mathbf{V}\widehat{\mathbf{R}}\mathbf{G}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)}, \end{aligned} \tag{37}$$

where the last equality follows from Lemma 10. This concludes the proof. □

Corollary 12. *Assuming the structural condition of eqn. (6), we further have, for all $j = 1, 2, \ldots t$,*

$$\|\widetilde{\mathbf{x}}^{(j)} - \mathbf{x}^{*(j)}\|_2 \leq \varepsilon\|\mathbf{x}^{*(j)}\|_2\,.$$

In addition, applying Lemma 6 yields

$$\left\|\widetilde{\mathbf{x}}^{(t)} - \mathbf{x}^{*(t)}\right\|_2 \leq \varepsilon\|\mathbf{x}^{*(t)}\|_2\,.$$

Proof. From the structural condition of eqn. (6), we have

$$\left\|\mathbf{G}^{-1}\widehat{\mathbf{E}}\right\|_2 \leq \left\|\mathbf{G}^{-1}\right\|_2\left\|\widehat{\mathbf{E}}\right\|_2 \leq \left\|\mathbf{G}^{-1}\right\|_2\frac{\varepsilon}{2} = \left(\frac{\sigma_1^2}{\sigma_1^2+\lambda}\right)\frac{\varepsilon}{2} \leq \frac{\varepsilon}{2}. \tag{38}$$

Moreover, eqn. (37) gives

$$\begin{aligned}\|\widetilde{\mathbf{x}}^{(j)} - \mathbf{x}^{*(j)}\|_2 &= \|\mathbf{V}\widehat{\mathbf{R}}\mathbf{G}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}\|_2 = \|\widehat{\mathbf{R}}\mathbf{G}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}\|_2 \\ &\leq \|\widehat{\mathbf{R}}\|_2\|\mathbf{G}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}\|_2 = \|\widehat{\mathbf{R}}\|_2\|\mathbf{V}\mathbf{G}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}\|_2 = \|\widehat{\mathbf{R}}\|_2\|\mathbf{x}^{*(j)}\|_2\,, \end{aligned} \tag{39}$$

where we used the unitary invariance and sub-multiplicativity of the spectral norm, as well as eqn. (31).

Next, from eqn. (36) and (38) and we have

$$\begin{aligned}\|\widehat{\mathbf{R}}\|_2 &= \left\|\sum_{\ell=1}^{\infty}(-1)^\ell\left(\mathbf{G}^{-1}\widehat{\mathbf{E}}\right)^\ell\right\|_2 \leq \sum_{\ell=1}^{\infty}\left\|\left(\mathbf{G}^{-1}\widehat{\mathbf{E}}\right)^\ell\right\|_2 \\ &\leq \sum_{\ell=1}^{\infty}\left(\left\|\mathbf{G}^{-1}\widehat{\mathbf{E}}\right\|_2\right)^\ell \leq \sum_{\ell=1}^{\infty}\left(\frac{\varepsilon}{2}\right)^\ell = \frac{\varepsilon/2}{1-\varepsilon/2} \leq \varepsilon\,. \end{aligned} \tag{40}$$

Here, eqn. (40) follows from the triangle inequality, sub-multiplicativity of the 2-norm, and the fact that $\varepsilon/2 < 1/2$. Finally, combining eqns. (39) and (40) we have

$$\|\widetilde{\mathbf{x}}^{(j)} - \mathbf{x}^{*(j)}\|_2 \leq \varepsilon\|\mathbf{x}^{*(j)}\|_2. \tag{41}$$

Note that, as Lemma 6 does not assume any specific structural condition, it holds in this case as well. Thus, repeated application of eqns. (24) and (41) results in the bound

$$\|\widehat{\mathbf{x}}^* - \mathbf{x}^*\|_2 = \Big\| \sum_{j=1}^{t} \widetilde{\mathbf{x}}^{(j)} - \mathbf{x}^* \Big\|_2 = \Big\| \widetilde{\mathbf{x}}^{(t)} - \Big(\mathbf{x}^* - \sum_{j=1}^{t-1} \widetilde{\mathbf{x}}^{(j)}\Big) \Big\|_2 = \Big\| \widetilde{\mathbf{x}}^{(t)} - \mathbf{x}^{*(t)} \Big\|_2 \leq \varepsilon \|\mathbf{x}^{*(t)}\|_2 \,.$$

This concludes the proof. □

The next result provides a critical inequality that can be used recursively in order to establish Theorem 1.

Lemma 13. *Let $\mathbf{x}^{*(j)}$, $j = 1, \ldots, t$, be the vectors of eqn. (15). For any $j = 1, \ldots, t-1$, if the structural condition of eqn. (6) is satisfied, then*

$$\|\mathbf{x}^{*(j+1)}\|_2 \leq \varepsilon \|\mathbf{x}^{*(j)}\|_2 \,. \tag{42}$$

Proof. For any $j = 1, 2, \ldots t$, we have

$$\begin{aligned}
\Big\|\mathbf{x}^{*(j+1)}\Big\|_2 &= \Big\|\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}^{(j+1)}\Big\|_2 \\
&= \Big\|\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\left(\mathbf{b}^{(j)} - \lambda \mathbf{y}^{(j)} - \mathbf{A}\widetilde{\mathbf{x}}^{(j)}\right)\Big\|_2 \\
&= \Big\|\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\left(\mathbf{b}^{(j)} - (\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}^{(j)}\right)\Big\|_2 \\
&= \Big\|\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}^{(j)} - \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}^{(j)}\Big\|_2 \\
&= \Big\|\mathbf{x}^{*(j)} - \widetilde{\mathbf{x}}^{(j)}\Big\|_2 \leq \varepsilon \Big\|\mathbf{x}^{*(j)}\Big\|_2 \,,
\end{aligned} \tag{43}$$

where the last inequality follows from eqn. (41). This completes the proof. □

Proof of Theorem 1. From Corollary 12, we have

$$\|\widehat{\mathbf{x}}^* - \mathbf{x}^*\|_2 \leq \varepsilon \|\mathbf{x}^{*(t)}\|_2 \tag{44}$$

and applying Lemma 13 iteratively yields

$$\Big\|\mathbf{x}^{*(t)}\Big\|_2 \leq \varepsilon \Big\|\mathbf{x}^{*(t-1)}\Big\|_2 \leq \varepsilon^2 \Big\|\mathbf{x}^{*(t-2)}\Big\|_2 \leq \cdots \leq \varepsilon^{t-1} \|\mathbf{x}^*\|_2 \,. \tag{45}$$

Finally, combining eqns. (44) and (45), we conclude

$$\|\widehat{\mathbf{x}}^* - \mathbf{x}^*\|_2 \leq \varepsilon^t \|\mathbf{x}^*\|_2 \,.$$

This completes the proof of Theorem 1. □

C. Proof of Theorem 2

In this section, we will only highlight (and prove) those results which has been either mentioned or stated without proof in Section 3, in order to give reader a complete picture.

Lemma 14. *For $j = 1, 2, \ldots, t$, let $\mathbf{b}^{(j)}$ be the intermediate response vectors in Algorithm 1 and $\mathbf{x}^{*(j)}$ be the vector defined in eqn. (15). then for any $j = 1, 2, \ldots, t$, $\mathbf{x}^{*(j)}$ can also be expressed as*

$$\mathbf{x}^{*(j)} = \mathbf{V}\boldsymbol{\Sigma}_\lambda^2 \boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \,. \tag{46}$$

Proof. From eqn. (15) and the thin SVD representation of $\mathbf{A}$, we have

$$\begin{aligned}
\mathbf{x}^{*(j)} &= \mathbf{A}^\mathsf{T}\left(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n\right)^{-1}\mathbf{b}^{(j)} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\mathsf{T}\left(\mathbf{U}\boldsymbol{\Sigma}^2\mathbf{U}^\mathsf{T} + \lambda \mathbf{U}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{b}^{(j)} \\
&= \mathbf{V}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}^2 + \lambda \mathbf{I}_n\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} = \mathbf{V}\left[\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}^2 + \lambda \mathbf{I}_n\right)^{-1}\boldsymbol{\Sigma}\right]\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} = \mathbf{V}\boldsymbol{\Sigma}_\lambda^2\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)},
\end{aligned} \tag{47}$$

where we used the fact that $\boldsymbol{\Sigma}_\lambda^2 = \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}^2 + \lambda \mathbf{I}_n\right)^{-1}\boldsymbol{\Sigma}$. This concludes the proof. □

Proof of Lemma 5. Denote $\mathbf{W} = \mathbf{S}\mathbf{S}^\mathsf{T}$. Using the thin SVD representation of $\mathbf{A}$, we have

$$\begin{aligned}
\widetilde{\mathbf{x}}^{(j)} &= \mathbf{A}^\mathsf{T}\left(\mathbf{A}\mathbf{W}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n\right)^{-1}\mathbf{b}^{(j)} \\
&= \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T} + \lambda\mathbf{U}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{b}^{(j)} \\
&= \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\left(\mathbf{U}\left(\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma} + \lambda\mathbf{I}_n\right)\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{b}^{(j)} \\
&= \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\mathbf{U}\left(\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma} + \lambda\mathbf{I}_n\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)}\,. \qquad (48)
\end{aligned}$$

Clearly, the matrix $\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma}$ is (symmetric) positive semidefinite and $\lambda\mathbf{I}_n$ is a positive definite matrix (as $\lambda > 0$). Thus, $\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma} + \lambda\mathbf{I}_n$ is positive definite, and the underlying inverse exists.

Now, proceeding with eqn. (48) and noting that $\mathbf{U}\mathbf{U}^\mathsf{T} = \mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}_n$, we have

$$\begin{aligned}
\widetilde{\mathbf{x}}^{(j)} &= \mathbf{V}\mathbf{\Sigma}\left(\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma} + \lambda\mathbf{I}_n\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \\
&= \mathbf{V}\mathbf{\Sigma}\left(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}\left(\mathbf{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma}_\lambda\right)\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma} + \lambda\mathbf{I}_n\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \\
&= \mathbf{V}\mathbf{\Sigma}\left(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}\left(\mathbf{\Sigma}_\lambda^2 + \mathbf{E}\right)\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma} + \lambda\mathbf{I}_n\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \qquad (49) \\
&= \mathbf{V}\mathbf{\Sigma}\left(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}\left(\mathbf{\Sigma}_\lambda^2 + \mathbf{E}\right)\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma} + \lambda\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-2}\mathbf{\Sigma}_\lambda\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma}\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \\
&= \mathbf{V}\mathbf{\Sigma}\left(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}\left(\mathbf{\Sigma}_\lambda^2 + \mathbf{E} + \lambda\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-2}\mathbf{\Sigma}_\lambda\right)\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma}\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \\
&= \mathbf{V}\mathbf{\Sigma}\left(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}\left(\mathbf{I}_n + \mathbf{E}\right)\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma}\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)}, \qquad (50)
\end{aligned}$$

where eqn. (49) used the fact that $\mathbf{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma}_\lambda = \mathbf{\Sigma}_\lambda^2 + \mathbf{E}$ and eqn. (50) follows from the fact that $\mathbf{\Sigma}_\lambda^2 + \lambda\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-2}\mathbf{\Sigma}_\lambda \in \mathbb{R}^{n\times n}$ is a diagonal matrix with i^{th} diagonal element

$$\left(\mathbf{\Sigma}_\lambda^2 + \lambda\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-2}\mathbf{\Sigma}_\lambda\right)_{ii} = \frac{\sigma_i^2}{\sigma_i^2+\lambda} + \frac{\lambda}{\sigma_i^2+\lambda} = 1\,,$$

for any $i = 1, 2, \ldots n$. Thus, we have $\left(\mathbf{\Sigma}_\lambda^2 + \lambda\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-2}\mathbf{\Sigma}_\lambda\right) = \mathbf{I}_n$.

Since $\|\mathbf{E}\|_2 < 1$, taking $\mathbf{P} = -\mathbf{E}$ in Proposition 8 implies that $(\mathbf{I}_n + \mathbf{E})^{-1}$ exists and $(\mathbf{I}_n + \mathbf{E})^{-1} = \mathbf{I}_n + \sum_{\ell=1}^{\infty}(-1)^\ell\mathbf{E}^\ell$. Thus, eqn. (50) can further be expressed as

$$\begin{aligned}
\widetilde{\mathbf{x}}^{(j)} &= \mathbf{V}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda\left(\mathbf{I}_n + \mathbf{E}\right)^{-1}\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \\
&= \mathbf{V}\mathbf{\Sigma}_\lambda\left(\mathbf{I}_n + \sum_{\ell=1}^{\infty}(-1)^\ell\mathbf{E}^\ell\right)\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \\
&= \mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} + \mathbf{V}\mathbf{\Sigma}_\lambda\mathbf{R}\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \\
&= \mathbf{x}^{*(j)} + \mathbf{V}\mathbf{\Sigma}_\lambda\mathbf{R}\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)}, \qquad (51)
\end{aligned}$$

where we applied Lemma 14 in the last line. This concludes the proof. □

Proof of Lemma 6. We prove by induction on t.

For $t = 1$, eqn. (15) boils down to

$$\mathbf{x}^{*(1)} = \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{b}^{(1)} = \mathbf{x}^*.$$

For $t = 2$, we have

$$\begin{aligned}
\mathbf{x}^{*(2)} &= \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{b}^{(2)} \\
&= \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\left(\mathbf{b}^{(1)} - \lambda\mathbf{y}^{(1)} - \mathbf{A}\widetilde{\mathbf{x}}^{(1)}\right) \\
&= \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\left(\mathbf{b}^{(1)} - (\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{b}^{(1)}\right) \\
&= \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{b}^{(1)} - \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{b}^{(1)}
\end{aligned}$$

$$= \mathbf{x}^* - \widetilde{\mathbf{x}}^{(1)}.$$

Now, suppose eqn. (24) is also true for $t = p$, *i.e.*,

$$\mathbf{x}^{*(p)} = \mathbf{x}^* - \sum_{j=1}^{p-1} \widetilde{\mathbf{x}}^{(j)}. \tag{52}$$

Then, for $t = p+1$, we can express $\mathbf{x}^{*(t)}$ as

$$\begin{aligned}
\mathbf{x}^{*(p+1)} &= \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}^{(p+1)} \\
&= \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\left(\mathbf{b}^{(p)} - \lambda \mathbf{y}^{(p)} - \mathbf{A}\widetilde{\mathbf{x}}^{(p)}\right) \\
&= \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\left(\mathbf{b}^{(p)} - (\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}^{(p)}\right) \\
&= \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}^{(p)} - \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}^{(p)} \\
&= \mathbf{x}^{*(p)} - \widetilde{\mathbf{x}}^{(p)} = \left(\mathbf{x}^* - \sum_{j=1}^{p-1} \widetilde{\mathbf{x}}^{(j)}\right) - \widetilde{\mathbf{x}}^{(p)} = \mathbf{x}^* - \sum_{j=1}^{p} \widetilde{\mathbf{x}}^{(j)},
\end{aligned}$$

where the second last equality in the last line follows from eqn. (52).

By the induction principle, we have proven eqn. (24). □

Proof of Lemma 7. From eqn. (15), we have for any $j = 1, 2, \ldots t$,

$$\begin{aligned}
\left\|\mathbf{x}^{*(j+1)}\right\|_2 &= \left\|\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}^{(j+1)}\right\|_2 \\
&= \left\|\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\left(\mathbf{b}^{(j)} - \lambda \mathbf{y}^{(j)} - \mathbf{A}\widetilde{\mathbf{x}}^{(j)}\right)\right\|_2 \\
&= \left\|\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\left(\mathbf{b}^{(j)} - (\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}^{(j)}\right)\right\|_2 \\
&= \left\|\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}^{(j)} - \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}^{(j)}\right\|_2 \\
&= \left\|\mathbf{x}^{*(j)} - \widetilde{\mathbf{x}}^{(j)}\right\|_2 \leq \frac{\varepsilon}{2}\left(\left\|\mathbf{x}^{*(j)}\right\|_2 + \frac{1}{\sqrt{2\lambda}}\left\|\mathbf{U}_{k,\perp}^\mathsf{T}\mathbf{b}^{(j)}\right\|_2\right),
\end{aligned} \tag{53}$$

where the last inequality follows from eqn. (23).

Next, for any $j = 1, 2, \ldots, t-1$, using the thin SVD representation of $\mathbf{A}$, we can rewrite $\mathbf{b}^{(j+1)}$ as

$$\begin{aligned}
\mathbf{b}^{(j+1)} &= \mathbf{b}^{(j)} - \lambda \mathbf{y}^{(j)} - \mathbf{A}\widetilde{\mathbf{x}}^{(j)} \\
&= \mathbf{b}^{(j)} - (\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}^{(j)} \\
&= \mathbf{b}^{(j)} - \mathbf{U}\left(\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n\right)\mathbf{U}^\mathsf{T}\mathbf{U}\left(\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\mathbf{\Sigma} + \lambda \mathbf{I}_n\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \\
&= \mathbf{b}^{(j)} - \mathbf{U}\left(\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n\right)\left(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}\underbrace{\mathbf{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\mathbf{\Sigma}_\lambda}_{\mathbf{E}+\mathbf{\Sigma}_\lambda^2}\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma} + \lambda \mathbf{I}_n\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} \\
&= \mathbf{b}^{(j)} - \mathbf{U}\left(\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n\right)\left(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}(\mathbf{I}_n + \mathbf{E})\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma}\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} && (54) \\
&= \mathbf{b}^{(j)} - \mathbf{U}\left(\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n\right)\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda(\mathbf{I}_n + \mathbf{E})^{-1}\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} && (55) \\
&= \mathbf{b}^{(j)} - \mathbf{U}\left(\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n\right)\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda(\mathbf{I}_n + \mathbf{R})\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)}, && (56)
\end{aligned}$$

where eqn. (54) follows from the same steps performed from eqn. (49) to eqn. (50). Also, eqn. (55) and eqn. (56) follow from Proposition 8 as $\|\mathbf{E}\|_2 \leq \frac{\varepsilon}{4\sqrt{2}} < 1$.

Moreover, note that $\left(\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n\right)\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda^2\mathbf{\Sigma}^{-1} = \mathbf{I}_n$ and using the fact that $\mathbf{U}\mathbf{U}^\mathsf{T} = \mathbf{I}_n$, we can rewrite eqn. (56) as

$$\mathbf{b}^{(j+1)} = \mathbf{b}^{(j)} - \mathbf{U}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)} - \mathbf{U}\left(\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n\right)\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda\mathbf{R}\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)}$$

$$= - \mathbf{U} \left(\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n\right) \mathbf{\Sigma}^{-1} \mathbf{\Sigma}_\lambda \mathbf{R} \mathbf{\Sigma}_\lambda \mathbf{\Sigma}^{-1} \mathbf{U}^\mathsf{T} \mathbf{b}^{(j)} \,. \tag{57}$$

Next, combining eqns. (18) and (57), we have

$$\begin{aligned}
\left\| \mathbf{U}_{k,\perp}^\mathsf{T} \mathbf{b}^{(j+1)} \right\|_2 &= \left\| -\mathbf{U}_{k,\perp}^\mathsf{T} \mathbf{U} (\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n) \mathbf{\Sigma}^{-1} \mathbf{\Sigma}_\lambda \mathbf{R} \mathbf{\Sigma}_\lambda \mathbf{\Sigma}^{-1} \mathbf{U}^\mathsf{T} \mathbf{b}^{(j)} \right\|_2 \\
&\leq \left\| \mathbf{U}_{k,\perp}^\mathsf{T} \mathbf{U} (\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n) \mathbf{\Sigma}^{-1} \mathbf{\Sigma}_\lambda \right\|_2 \|\mathbf{R}\|_2 \left\| \mathbf{\Sigma}_\lambda \mathbf{\Sigma}^{-1} \mathbf{U}^\mathsf{T} \mathbf{b}^{(j)} \right\|_2 \\
&\leq \frac{\varepsilon}{2\sqrt{2}} \left\| \mathbf{U}_{k,\perp}^\mathsf{T} \mathbf{U} (\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n) \mathbf{\Sigma}^{-1} \mathbf{\Sigma}_\lambda \right\|_2 \left\| \mathbf{\Sigma}_\lambda \mathbf{\Sigma}^{-1} \mathbf{U}^\mathsf{T} \mathbf{b}^{(j)} \right\|_2 \\
&= \frac{\varepsilon}{2\sqrt{2}} \left\| \mathbf{U}_{k,\perp}^\mathsf{T} \begin{pmatrix} \mathbf{U}_k & \mathbf{U}_{k,\perp} \end{pmatrix} (\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n) \mathbf{\Sigma}^{-1} \mathbf{\Sigma}_\lambda \right\|_2 \left\| \mathbf{\Sigma}_\lambda \mathbf{\Sigma}^{-1} \mathbf{U}^\mathsf{T} \mathbf{b}^{(j)} \right\|_2 \\
&= \frac{\varepsilon}{2\sqrt{2}} \left\| \begin{pmatrix} \mathbf{0}_{(n-k)\times k} & \mathbf{I}_{n-k} \end{pmatrix} (\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n) \mathbf{\Sigma}^{-1} \mathbf{\Sigma}_\lambda \right\|_2 \left\| \mathbf{\Sigma}_\lambda \mathbf{\Sigma}^{-1} \mathbf{U}^\mathsf{T} \mathbf{b}^{(j)} \right\|_2 .
\end{aligned} \tag{58}$$

Now, similar to equation eqn. (20), we apply triangle inequality and the fact that $\mathbf{\Sigma}_\lambda^{-1} = (\mathbf{\Sigma}_\lambda^{-1})_k + (\mathbf{\Sigma}_\lambda^{-1})_{k,\perp}$ to get the following inequality

$$\|\mathbf{\Sigma}_\lambda^{-1} \mathbf{\Sigma}_\lambda^2 \mathbf{\Sigma}^{-1} \mathbf{U}^\mathsf{T} \mathbf{b}^{(j)}\|_2 \leq \underbrace{\|(\mathbf{\Sigma}_\lambda^{-1})_k \, \mathbf{\Sigma}_\lambda^2 \mathbf{\Sigma}^{-1} \mathbf{U}^\mathsf{T} \mathbf{b}^{(j)}\|_2}_{\Delta_1} + \underbrace{\|(\mathbf{\Sigma}_\lambda^{-1})_{k,\perp} \, \mathbf{\Sigma}_\lambda^2 \mathbf{\Sigma}^{-1} \mathbf{U}^\mathsf{T} \mathbf{b}^{(j)}\|_2}_{\Delta_2} . \tag{59}$$

We now proceed to bound Δ_1 and Δ_2 separately.

Bounding Δ_1. Using $\mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}_n$, we have

$$\begin{aligned}
\Delta_1 = \left\| (\mathbf{\Sigma}_\lambda^{-1})_k \, \mathbf{V}^\mathsf{T} (\mathbf{V} \mathbf{\Sigma}_\lambda^2 \mathbf{\Sigma}^{-1} \mathbf{U}^\mathsf{T} \mathbf{b}^{(j)}) \right\|_2 &= \left\| (\mathbf{\Sigma}_\lambda^{-1})_k \, \mathbf{V}^\mathsf{T} \mathbf{x}^{*(j)} \right\|_2 \leq \left\| (\mathbf{\Sigma}_\lambda^{-1})_k \right\|_2 \left\| \mathbf{V}^\mathsf{T} \right\|_2 \left\| \mathbf{x}^{*(j)} \right\|_2 \\
&= \sqrt{(1 + \lambda/\sigma_k^2)} \left\| \mathbf{x}^{*(j)} \right\|_2 \leq \sqrt{2} \left\| \mathbf{x}^{*(j)} \right\|_2 ,
\end{aligned} \tag{60}$$

where we used the facts that $\mathbf{x}^{*(j)} = \mathbf{V} \mathbf{\Sigma}_\lambda^2 \mathbf{\Sigma}^{-1} \mathbf{U}^\mathsf{T} \mathbf{b}^{(j)}$ (see Lemma 14 in the Appendix), The last inequality follows from our assumption that $\sigma_k^2 \geq \lambda$.

Bounding Δ_2. Rewriting $\mathbf{U} = \begin{pmatrix} \mathbf{U}_k & \mathbf{U}_{k,\perp} \end{pmatrix}$, we have

$$\begin{aligned}
\Delta_2 = \left\| (\mathbf{\Sigma}_\lambda^{-1})_{k,\perp} \, \mathbf{\Sigma}_\lambda^2 \mathbf{\Sigma}^{-1} \mathbf{U}^\mathsf{T} \mathbf{b}^{(j)} \right\|_2 &= \left\| (\mathbf{\Sigma}_\lambda^{-1})_{k,\perp} \, \mathbf{\Sigma}_\lambda^2 \mathbf{\Sigma}^{-1} \begin{pmatrix} \mathbf{U}_k^\mathsf{T} \\ \mathbf{U}_{k,\perp}^\mathsf{T} \end{pmatrix} \mathbf{b}^{(j)} \right\|_2 \\
&= \left\| (\mathbf{\Sigma}_\lambda^{-1})_{k,\perp} \, \mathbf{\Sigma}_\lambda^2 \mathbf{\Sigma}^{-1} \begin{pmatrix} \mathbf{0}^\mathsf{T} \\ \mathbf{U}_{k,\perp}^\mathsf{T} \end{pmatrix} \mathbf{b}^{(j)} \right\|_2 \leq \left\| (\mathbf{\Sigma}_\lambda^{-1})_{k,\perp} \, \mathbf{\Sigma}_\lambda^2 \mathbf{\Sigma}^{-1} \right\|_2 \left\| \begin{pmatrix} \mathbf{0}^\mathsf{T} \\ \mathbf{U}_{k,\perp}^\mathsf{T} \end{pmatrix} \mathbf{b}^{(j)} \right\|_2 \qquad (61) \\
&\leq \left\| (\mathbf{\Sigma}_\lambda^{-1})_{k,\perp} \, \mathbf{\Sigma}_\lambda^2 \mathbf{\Sigma}^{-1} \right\|_2 \left\| \mathbf{U}_{k,\perp}^\mathsf{T} \mathbf{b}^{(j)} \right\|_2 = \frac{1}{\sqrt{\sigma_n^2 + \lambda}} \left\| \mathbf{U}_{k,\perp}^\mathsf{T} \mathbf{b}^{(j)} \right\|_2 \\
&\leq \frac{1}{\sqrt{\lambda}} \left\| \mathbf{U}_{k,\perp}^\mathsf{T} \mathbf{b}^{(j)} \right\|_2 . \qquad (62)
\end{aligned}$$

Equality in eqn. (61) holds because note that $\left((\mathbf{\Sigma}_\lambda^{-1})_{k,\perp} \, \mathbf{\Sigma}_\lambda^2 \mathbf{\Sigma}^{-1}\right) \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose (i,i)th diagonal entry is equal to $\frac{1}{\sqrt{\sigma_i^2 + \lambda}}$ if $i > k$ and zero otherwise.

In order to upper bound $\|\mathbf{\Sigma}_\lambda \mathbf{\Sigma}^{-1} \mathbf{U}^\mathsf{T} \mathbf{b}^{(j)}\|_2$, we combine eqns. (59), (60) and (62) to obtain

$$\left\| \mathbf{\Sigma}_\lambda \mathbf{\Sigma}^{-1} \mathbf{U}^\mathsf{T} \mathbf{b}^{(j)} \right\|_2 \leq \sqrt{2} \left\| \mathbf{x}^{*(j)} \right\|_2 + \frac{1}{\sqrt{\lambda}} \left\| \mathbf{U}_{k,\perp}^\mathsf{T} \mathbf{b}^{(j)} \right\|_2 . \tag{63}$$

Next, it can easily be verified that

$$\left\| \begin{pmatrix} \mathbf{0}_{(n-k)\times k} & \mathbf{I}_{n-k} \end{pmatrix} (\mathbf{\Sigma}^2 + \lambda \mathbf{I}_n) \mathbf{\Sigma}^{-1} \mathbf{\Sigma}_\lambda \right\|_2 = \sqrt{\sigma_{k+1}^2 + \lambda} \leq \sqrt{2\lambda} \,, \tag{64}$$

where the last inequality in eqn. (64) directly follows from the definition of k *i.e.* $\sigma_{k+1}^2 \leq \lambda$.

Further, combining eqns. (58), (63) and eqn. (64), we have

$$\left\|\mathbf{U}_{k,\perp}^{\mathsf{T}}\mathbf{b}^{(j+1)}\right\|_2 \leq \frac{\varepsilon}{2\sqrt{2}}\sqrt{2\lambda}\left(\sqrt{2}\left\|\mathbf{x}^{*(j)}\right\|_2 + \frac{1}{\sqrt{\lambda}}\left\|\mathbf{U}_{k,\perp}^{\mathsf{T}}\mathbf{b}^{(j)}\right\|_2\right). \tag{65}$$

Finally, putting together eqns. (53) and (65), we conclude

$$\left\|\mathbf{x}^{*(j+1)}\right\|_2 + \frac{1}{\sqrt{2\lambda}}\left\|\mathbf{U}_{k,\perp}^{\mathsf{T}}\mathbf{b}^{(j+1)}\right\|_2 \leq \varepsilon\left(\left\|\mathbf{x}^{*(j)}\right\|_2 + \frac{1}{\sqrt{2\lambda}}\left\|\mathbf{U}_{k,\perp}^{\mathsf{T}}\mathbf{b}^{(j)}\right\|_2\right) \tag{66}$$

for any $j = 1, 2, \ldots, t-1$. □

D. Connection to Preconditioned Richardson Iteration

In Algorithm 1, let $\bar{\mathbf{y}}^{(j)} = \sum_{k=1}^{j}\mathbf{y}^{(k)}$. Therefore, after t iterations the final output is given by $\widehat{\mathbf{x}}^* = \mathbf{A}^{\mathsf{T}}\bar{\mathbf{y}}^{(t)}$. Furthermore, from our construction,

$$\begin{aligned}
\mathbf{b}^{(j)} &= \mathbf{b}^{(j-1)} - (\mathbf{A}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n)(\mathbf{A}\mathbf{S}\mathbf{S}^{\mathsf{T}}\mathbf{A} + \lambda\mathbf{I}_n)^{-1}\mathbf{b}^{(j-1)} \\
&= \mathbf{b}^{(j-1)} - (\mathbf{A}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n)\mathbf{y}^{(j-1)} \\
&= \mathbf{b}^{(j-2)} - (\mathbf{A}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n)\mathbf{y}^{(j-2)} - (\mathbf{A}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n)\mathbf{y}^{(j-1)} \\
&= \mathbf{b}^{(j-2)} - (\mathbf{A}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n)\left(\mathbf{y}^{(j-1)} + \mathbf{y}^{(j-2)}\right) \\
&\;\;\vdots \\
&= \mathbf{b}^{(1)} - (\mathbf{A}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n)\left(\mathbf{y}^{(j-1)} + \mathbf{y}^{(j-2)} + \cdots + \mathbf{y}^{(1)}\right) \\
&= \mathbf{b} - (\mathbf{A}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n)\bar{\mathbf{y}}^{(j-1)}.
\end{aligned} \tag{67}$$

Again, repeatedly using the definition of $\bar{\mathbf{y}}^{(j)}$ and eqn. (67), we obtain

$$\begin{aligned}
\bar{\mathbf{y}}^{(j)} = \bar{\mathbf{y}}^{(j-1)} + \mathbf{y}^{(j)} &= \bar{\mathbf{y}}^{(j-1)} + (\mathbf{A}\mathbf{S}\mathbf{S}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n)^{-1}\mathbf{b}^{(j)} \\
&= \bar{\mathbf{y}}^{(j-1)} + (\mathbf{A}\mathbf{S}\mathbf{S}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n)^{-1}\left(\mathbf{b} - (\mathbf{A}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n)\bar{\mathbf{y}}^{(j-1)}\right).
\end{aligned} \tag{68}$$

Thus, our Algorithm 1 can be formulated as a preconditioned Richardson iteration to solve the linear system

$$(\mathbf{A}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n)\mathbf{y} = \mathbf{b} \tag{69}$$

with preconditioner $\mathbf{P}^{-1} = (\mathbf{A}\mathbf{S}\mathbf{S}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n)^{-1}$ and step-size one.

Next, we state an important result on the convergence of preconditioned Richardson iteration and use it to show that subject to our structural conditions in eqns. (6) and (8), $\bar{\mathbf{y}}^{(t)}$ converges to the true solution $\mathbf{y}^* = (\mathbf{A}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n)^{-1}\mathbf{b}$ as t increases.

Lemma 15. (Corollary 2.4.1 of (Quarteroni & Valli, 1994)) *The preconditioned Richardson method of eqn.* (68) *converges if and only if the maximum eigenvalue (spectral radius) of* $\mathbf{P}^{-1}(\mathbf{A}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n)$ *satisfies:*

$$\lambda_{\max}\big(\mathbf{P}^{-1}(\mathbf{A}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n)\big) < 2\,,$$

where $\mathbf{P} = \mathbf{A}\mathbf{S}\mathbf{S}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}} + \lambda\mathbf{I}_n$.

Proof of convergence under the structural condition of eqn. (6). Consider the condition of eqn. (6):

$$\begin{aligned}
&\|\mathbf{V}^{\mathsf{T}}\mathbf{S}\mathbf{S}^{\mathsf{T}}\mathbf{V} - \mathbf{I}_n\|_2 \leq \frac{\varepsilon}{2} \;\Leftrightarrow\; -\frac{\varepsilon}{2}\mathbf{I}_n \preccurlyeq \mathbf{V}^{\mathsf{T}}\mathbf{S}\mathbf{S}^{\mathsf{T}}\mathbf{V} - \mathbf{I}_n \preccurlyeq \frac{\varepsilon}{2}\mathbf{I}_n \\
\Rightarrow\; &-\frac{\varepsilon}{2}\mathbf{A}\mathbf{A}^{\mathsf{T}} \preccurlyeq \mathbf{A}\mathbf{S}\mathbf{S}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}} - \mathbf{A}\mathbf{A}^{\mathsf{T}} \preccurlyeq \frac{\varepsilon}{2}\mathbf{A}\mathbf{A}^{\mathsf{T}}
\end{aligned} \tag{70}$$

$$
\begin{aligned}
&\Rightarrow \left(1-\frac{\varepsilon}{2}\right)\mathbf{A}\mathbf{A}^\mathsf{T} \preccurlyeq \mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} \preccurlyeq \left(1+\frac{\varepsilon}{2}\right)\mathbf{A}\mathbf{A}^\mathsf{T} \\
&\Rightarrow \left(1-\frac{\varepsilon}{2}\right)\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n \preccurlyeq \mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n \preccurlyeq \left(1+\frac{\varepsilon}{2}\right)\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n \\
&\Rightarrow \left(1-\frac{\varepsilon}{2}\right)\left(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n\right) \preccurlyeq \underbrace{\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n}_{\mathbf{P}} \preccurlyeq \left(1+\frac{\varepsilon}{2}\right)\left(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n\right), \qquad (71)
\end{aligned}
$$

where we obtain eqn. (70) by pre- and post-multiplying the previous inequality by $\mathbf{U}\mathbf{\Sigma}$ and $\mathbf{\Sigma}\mathbf{U}^\mathsf{T}$ respectively and using the facts that $\mathbf{A}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$ and $\mathbf{A}\mathbf{A}^\mathsf{T}=\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\mathsf{T}$. Furthermore, eqn. (71) holds as $(1-\varepsilon/2)\leq 1$ and $(1+\varepsilon/2)\geq 1$. Next, pre- and post- multiplying eqn. (71) by $\mathbf{P}^{-1/2}$, we obtain

$$
\left(1+\frac{\varepsilon}{2}\right)^{-1}\mathbf{I}_n \preccurlyeq \mathbf{P}^{-1/2}\left(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n\right)\mathbf{P}^{-1/2} \preccurlyeq \left(1-\frac{\varepsilon}{2}\right)^{-1}\mathbf{I}_n,
$$

which implies that the eigenvalues of $\mathbf{P}^{-1/2}\left(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n\right)\mathbf{P}^{-1/2}$ are bounded between $\left(1+\frac{\varepsilon}{2}\right)^{-1}$ and $\left(1-\frac{\varepsilon}{2}\right)^{-1}$. Moreover, notice that $\mathbf{P}^{-1/2}\left(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n\right)\mathbf{P}^{-1/2}$ is similar to $\mathbf{P}^{-1}\left(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n\right)$ which implies that both matrices have same set of eigenvalues and therefore the eigenvalues of $\mathbf{P}^{-1}\left(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n\right)$ are also bounded between $\left(1+\frac{\varepsilon}{2}\right)^{-1}$ and $\left(1-\frac{\varepsilon}{2}\right)^{-1}$. Finally, using $\varepsilon<1$, we obtain

$$
\lambda_{\max}\big(\mathbf{P}^{-1}(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)\big) \leq \left(1-\frac{\varepsilon}{2}\right)^{-1} < 2.
$$

This concludes the proof. □

Proof of convergence under the structural condition of eqn. (8). Using the SVD of $\mathbf{A}$, it is easy to verify that

$$
\begin{aligned}
&\|\mathbf{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\mathbf{\Sigma}_\lambda-\mathbf{\Sigma}_\lambda^2\|_2 \\
=&\|(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-\frac{1}{2}}\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-\frac{1}{2}}-(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-\frac{1}{2}}\mathbf{A}\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-\frac{1}{2}}\|_2. \qquad (72)
\end{aligned}
$$

Using eqn. (72), we rewrite the structural condition of eqn. (8) as follows:

$$
-\frac{\varepsilon}{4\sqrt{2}}\mathbf{I}_n \preccurlyeq (\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-\frac{1}{2}}\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-\frac{1}{2}}-(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-\frac{1}{2}}\mathbf{A}\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-\frac{1}{2}} \preccurlyeq \frac{\varepsilon}{4\sqrt{2}}\mathbf{I}_n.
$$

Now, pre- and post-multiplying the above inequality by $(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{\frac{1}{2}}$, we obtain

$$
\begin{aligned}
&-\frac{\varepsilon}{4\sqrt{2}}\left(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n\right) \preccurlyeq \mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T}-\mathbf{A}\mathbf{A}^\mathsf{T} \preccurlyeq \frac{\varepsilon}{4\sqrt{2}}\left(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n\right) \\
\Rightarrow\ &-\frac{\varepsilon}{4\sqrt{2}}\left(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n\right) \preccurlyeq \mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n-\left(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n\right) \preccurlyeq \frac{\varepsilon}{4\sqrt{2}}\left(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n\right) \\
\Rightarrow\ &\left(1-\frac{\varepsilon}{4\sqrt{2}}\right)\left(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n\right) \preccurlyeq \underbrace{\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n}_{\mathbf{P}} \preccurlyeq \left(1+\frac{\varepsilon}{4\sqrt{2}}\right)\left(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n\right). \qquad (73)
\end{aligned}
$$

As before, pre- and post-multiplying eqn. (73) by $\mathbf{P}^{-1/2}$, we obtain

$$
\left(1+\frac{\varepsilon}{4\sqrt{2}}\right)^{-1}\mathbf{I}_n \preccurlyeq \mathbf{P}^{-1/2}\left(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n\right)\mathbf{P}^{-1/2} \preccurlyeq \left(1-\frac{\varepsilon}{4\sqrt{2}}\right)^{-1}\mathbf{I}_n.
$$

Now, using a similar argument as in the previous case, we obtain

$$
\lambda_{\max}\big(\mathbf{P}^{-1}(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)\big) \leq \left(1-\frac{\varepsilon}{4\sqrt{2}}\right)^{-1} < 2\ ,\ \text{ as } \varepsilon<1\,.
$$

This concludes the proof. □

Number of Iterations. The above derivations imply that the eigenvalues of $\mathbf{P}^{-1}(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)$ are bounded between $(1+\mathcal{O}(\varepsilon))^{-1}$ and $(1-\mathcal{O}(\varepsilon))^{-1}$ and thus the condition number of $\mathbf{P}^{-1}(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)$ is constant whenever ε is constant. Now, using Theorem 2.3.1 of (Kyng, 2017), we can argue that for any error parameter $\varepsilon'=\mathcal{O}(\varepsilon)$, the preconditioned Richardson iteration needs $\mathcal{O}\left(\ln(1/\varepsilon')\right)$ steps to converge.

E. Bias-Variance Trade-off

Our next result quantifies the bias-variance trade-off for under-constrained ridge regression.

Lemma 16. *Let the data-generation model be given by eqn. (12). Then, the mean squared error (MSE) of $\mathbf{x}^*$ can be expressed as follows*

$$\text{MSE}(\mathbf{x}^*) = \sigma^2 \left\|(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{A}\right\|_F^2 + \left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2^2 . \tag{74}$$

Proof. The covariance matrix of $\mathbf{b}$ is given by $\mathbb{E}\left[(\mathbf{b} - \mathbb{E}(\mathbf{b}))(\mathbf{b} - \mathbb{E}(\mathbf{b}))^\mathsf{T}\right]$ and is denoted $\text{Var}(\mathbf{b})$. Since the ridge regression estimator $\mathbf{x}^*$ of the parameter vector $\mathbf{x}_0$ is given by eqn. (3), we have

$$\begin{aligned}\mathbb{E}(\mathbf{x}^*) &= \mathbb{E}\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}\right) = \mathbf{A}^\mathsf{T}\left(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n\right)^{-1}\mathbb{E}\left(\mathbf{b}\right) \\ &= \mathbf{A}^\mathsf{T}\left(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n\right)^{-1}\mathbf{A}\mathbf{x}_0 = \mathbf{x}_0 + \left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0 = \mathbf{x}_0 + b(\mathbf{x}^*) ,\end{aligned} \tag{75}$$

where

$$b(\mathbf{x}^*) = \left(\mathbf{A}^\mathsf{T}\left(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n\right)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0$$

is the underlying *bias* in estimating $\mathbf{x}_0$ through $\mathbf{x}^*$.

Furthermore, combining second equality in eqn. (75) with eqn. (3), we obtain

$$\mathbf{x}^* - \mathbb{E}(\mathbf{x}^*) = \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\left(\mathbf{b} - \mathbb{E}(\mathbf{b})\right)$$

and thus

$$\left(\mathbf{x}^* - \mathbb{E}(\mathbf{x}^*)\right)\left(\mathbf{x}^* - \mathbb{E}(\mathbf{x}^*)\right)^\mathsf{T} = \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\left(\mathbf{b} - \mathbb{E}(\mathbf{b})\right)\left(\mathbf{b} - \mathbb{E}(\mathbf{b})\right)^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{A} . \tag{76}$$

Taking expectation on both sides of eqn. (76) and using the linearity of expectation, we have

$$\text{Var}(\mathbf{x}^*) = \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\text{Var}(\mathbf{b})\left(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n\right)^{-1}\mathbf{A} = \sigma^2\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-2}\mathbf{A} , \tag{77}$$

where we used the fact that $\text{Var}(\mathbf{b}) = \sigma^2\mathbf{I}_n$.

In order to decompose $\text{MSE}(\mathbf{x}^*)$ into the variance and bias components, we add and subtract $\mathbb{E}(\mathbf{x}^*)$ and proceed as follows:

$$\begin{aligned}\text{MSE}(\mathbf{x}^*) &= \mathbb{E}\left[\|\mathbf{x}^* - \mathbf{x}_0\|_2^2\right] = \mathbb{E}\left[\|\mathbf{x}^* - \mathbb{E}(\mathbf{x}^*) + \mathbb{E}(\mathbf{x}^*) - \mathbf{x}_0\|_2^2\right] \\ &= \mathbb{E}\left[\|\mathbf{x}^* - \mathbb{E}(\mathbf{x}^*) + b(\mathbf{x}^*)\|_2^2\right] = \mathbb{E}\left[\|\mathbf{x}^* - \mathbb{E}(\mathbf{x}^*)\|_2^2\right] + \|b(\mathbf{x}^*)\|_2^2 \qquad (78) \\ &= \sum_{i=1}^{d}\mathbb{E}\left[\left(\mathbf{x}_i^* - \mathbb{E}(\mathbf{x}_i^*)\right)^2\right] + \|b(\mathbf{x}^*)\|_2^2 = \sum_{i=1}^{d}\left[\text{Var}(\mathbf{x}^*)\right]_{ii} + \|b(\mathbf{x}^*)\|_2^2 \\ &= \text{tr}\left(\text{Var}(\mathbf{x}^*)\right) + \|b(\mathbf{x}^*)\|_2^2 . \qquad (79)\end{aligned}$$

Here, $\mathbf{x}_i^*$ is the i^{th} element of $\mathbf{x}^*$. To achieve the second equality in eqn. (78), we used the fact that $\mathbb{E}\left(\mathbf{x}^* - \mathbb{E}(\mathbf{x}^*)\right) = \mathbf{0}$. Further, combining eqn. (75), eqn. (77) and eqn. (79), we have

$$\begin{aligned}\text{MSE}(\mathbf{x}^*) =& \sigma^2\text{tr}\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-2}\mathbf{A}\right) + \left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2^2 \\ =& \underbrace{\sigma^2\left\|(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{A}\right\|_F^2}_{\text{Variance}} + \underbrace{\left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2^2}_{\text{Bias}^2} .\end{aligned}$$

This concludes the proof. □

E.1. Proof of Theorem 4 under eqn. (6)

First, we present the following result showing an alternative formulation of $\left\|(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F$.

Lemma 17. *Let $\mathbf{A}\in\mathbb{R}^{n\times d}$ be the design matrix and $\lambda(>0)$ be the ridge parameter of the ridge regression problem. Then, we have*

$$\text{(a)}\quad \left\|(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F = \left\|\mathbf{\Sigma}^{-1}\mathbf{G}^{-1}\right\|_F, \text{ and} \tag{80}$$

$$\text{(b)}\quad \left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{A}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2 = \left\|\left(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2, \tag{81}$$

where $\mathbf{G}=\mathbf{I}_n+\lambda\mathbf{\Sigma}^{-2}$.

Proof. Part (a): Using the thin SVD representation of $\mathbf{A}$ and putting $\mathbf{I}_n=\mathbf{U}\mathbf{U}^\mathsf{T}$, we have

$$\begin{aligned}\left\|(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F &= \left\|\left(\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\mathsf{T}+\lambda\mathbf{U}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\right\|_F\\ &= \left\|\left(\mathbf{U}\mathbf{\Sigma}(\mathbf{I}_n+\lambda\mathbf{\Sigma}^{-2})\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\right\|_F\\ &= \left\|\left(\mathbf{U}\mathbf{\Sigma}\mathbf{G}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\right\|_F.\end{aligned} \tag{82}$$

Clearly, $\mathbf{G}^{-1}$ exists. Further, using the fact that $\mathbf{U}^\mathsf{T}\mathbf{U}=\mathbf{I}_n$ and exploiting unitary invariance of Frobenius norm, we can rewrite eqn. (82) as

$$\left\|(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F = \left\|\mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{G}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\right\|_F = \left\|\mathbf{\Sigma}^{-1}\mathbf{G}^{-1}\right\|_F, \tag{83}$$

which concludes the proof of part (a).

Part (b): It suffices to show that $\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{A}=\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T}$. From the thin SVD representation of $\mathbf{A}$, we have

$$\begin{aligned}\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{A} &= \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\left(\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\mathsf{T}+\lambda\mathbf{U}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\\ &= \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\left(\mathbf{U}\mathbf{\Sigma}(\mathbf{I}_n+\lambda\mathbf{\Sigma}^{-2})\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\\ &= \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\left(\mathbf{U}\mathbf{\Sigma}\mathbf{G}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\\ &= \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{G}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\\ &= \mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T},\end{aligned}$$

where we used the facts that $\mathbf{G}^{-1}$ exists and that $\mathbf{U}^\mathsf{T}\mathbf{U}=\mathbf{I}_n$. This completes the proof. □

Our next result bounds each term in eqn. (14) separately subject to the structural condition of eqn. (6).

Lemma 18. *Let $\mathbf{A}\in\mathbb{R}^{n\times d}$, $\mathbf{b}\in\mathbb{R}^n$, and $\lambda>0$ be the inputs of the ridge regression problem. Let $\mathbf{S}\in\mathbb{R}^{d\times s}$ be the sketching matrix in Algorithm 1 and define*

$$\widehat{\mathbf{E}}=\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}-\mathbf{I}_n.$$

Further, assume for some constant $0<\varepsilon<1$, if the condition of eqn. (6) is satisfied i.e. $\|\widehat{\mathbf{E}}\|_2\le\varepsilon/2$, then

$$\text{(a)}\quad \sigma^2\left\|(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F^2 \le (1+\varepsilon)^2\,\sigma^2\left\|(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F^2, \text{ and} \tag{84}$$

$$\text{(b)}\quad \left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{A}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2^2 \le (1+\varepsilon\gamma_1)^2\left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{A}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2^2, \tag{85}$$

where $\gamma_1=\left(1+\sigma_1^2/\lambda\right)$.

Proof. Let $\mathbf{W}=\mathbf{S}\mathbf{S}^\mathsf{T}$. As before, we start with the thin SVD representation of $\mathbf{A}$.

Part (a): We have

$$\left\|(\mathbf{A}\mathbf{W}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F^2 = \left\|\left(\mathbf{U}\mathbf{\Sigma}(\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V})\mathbf{\Sigma}\mathbf{U}^\mathsf{T}+\lambda\mathbf{U}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\right\|_F^2$$

$$
\begin{aligned}
&= \left\| \left(\mathbf{U}\mathbf{\Sigma}(\mathbf{I}_n + \widehat{\mathbf{E}})\mathbf{\Sigma}\mathbf{U}^\mathsf{T} + \lambda \mathbf{U}\mathbf{U}^\mathsf{T} \right)^{-1} \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} \right\|_F^2 \\
&= \left\| \left(\mathbf{U}\mathbf{\Sigma}(\mathbf{I}_n + \widehat{\mathbf{E}} + \lambda\mathbf{\Sigma}^{-2})\mathbf{\Sigma}\mathbf{U}^\mathsf{T} \right)^{-1} \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} \right\|_F^2 \\
&= \left\| \mathbf{U}\mathbf{\Sigma}^{-1}(\mathbf{I}_n + \widehat{\mathbf{E}} + \lambda\mathbf{\Sigma}^{-2})^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} \right\|_F^2 . \qquad (86)
\end{aligned}
$$

Now, using the facts that $\mathbf{U}\mathbf{U}^\mathsf{T} = \mathbf{I}_n$ and the unitary invariance of the Frobenius norm, we can rewrite eqn. (86) as

$$
\begin{aligned}
\left\| (\mathbf{A}\mathbf{W}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} \right\|_F^2 &= \left\| \mathbf{\Sigma}^{-1}(\mathbf{I}_n + \widehat{\mathbf{E}} + \lambda\mathbf{\Sigma}^{-2})^{-1} \right\|_F^2 = \left\| \mathbf{\Sigma}^{-1}(\mathbf{G} + \widehat{\mathbf{E}})^{-1} \right\|_F^2 \\
&= \left\| \mathbf{\Sigma}^{-1} \left((\mathbf{I}_n + \widehat{\mathbf{E}}\mathbf{G}^{-1})\mathbf{G} \right)^{-1} \right\|_F^2 = \left\| \mathbf{\Sigma}^{-1}\mathbf{G}^{-1} \left(\mathbf{I}_n + \widehat{\mathbf{E}}\mathbf{G}^{-1} \right)^{-1} \right\|_F^2 , \qquad (87)
\end{aligned}
$$

where $\mathbf{G} = \mathbf{I}_n + \lambda\mathbf{\Sigma}^{-2}$ and is invertible. Further, $(\mathbf{I}_n + \widehat{\mathbf{E}}\mathbf{G}^{-1})^{-1}$ exists because of Proposition 8 and the fact that $\|\widehat{\mathbf{E}}\mathbf{G}^{-1}\|_2 \leq \varepsilon/2$ (the proof is the same as eqn. (38)). Thus, eqn. (87) holds. Moreover, taking $\mathbf{P} = -\widehat{\mathbf{E}}\mathbf{G}^{-1}$ in Proposition 8 yields

$$
\left(\mathbf{I}_n + \widehat{\mathbf{E}}\mathbf{G}^{-1} \right)^{-1} = \sum_{\ell=0}^{\infty} (-1)^\ell \left(\widehat{\mathbf{E}}\mathbf{G}^{-1} \right)^\ell \triangleq \mathbf{T} . \qquad (88)
$$

Next, combining eqns. (87) and (88) and applying strong sub-multiplicativity, we obtain

$$
\left\| (\mathbf{A}\mathbf{W}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} \right\|_F^2 = \left\| \mathbf{\Sigma}^{-1}\mathbf{G}^{-1}\mathbf{T} \right\|_F^2 \leq \|\mathbf{T}\|_2^2 \left\| \mathbf{\Sigma}^{-1}\mathbf{G}^{-1} \right\|_F^2 . \qquad (89)
$$

Next, using eqn. (88) and the fact $\|\widehat{\mathbf{E}}\mathbf{G}^{-1}\|_2 \leq \varepsilon/2$ yields

$$
\begin{aligned}
\|\mathbf{T}\|_2 &= \left\| \sum_{\ell=0}^{\infty} (-1)^\ell \left(\widehat{\mathbf{E}}\mathbf{G}^{-1} \right)^\ell \right\|_2 \leq \sum_{\ell=0}^{\infty} \left\| \left(\widehat{\mathbf{E}}\mathbf{G}^{-1} \right)^\ell \right\|_2 \\
&\leq \sum_{\ell=0}^{\infty} \left(\left\| \widehat{\mathbf{E}}\mathbf{G}^{-1} \right\|_2 \right)^\ell \leq \sum_{\ell=0}^{\infty} \left(\frac{\varepsilon}{2} \right)^\ell = \frac{1}{1 - \varepsilon/2} \leq 1 + \varepsilon , \qquad (90)
\end{aligned}
$$

where the first inequality is due to the triangle inequality, the second one follows from sub-multiplicativity and the last inequality holds as $0 < \varepsilon < 1$.

Finally, combining eqn. (80), eqn. (89), eqn. (90) and multiplying both sides by σ^2, we have

$$
\sigma^2 \left\| (\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} \right\|_F^2 \leq (1 + \varepsilon)^2 \, \sigma^2 \left\| (\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} \right\|_F^2 .
$$

Part (b): We have

$$
\begin{aligned}
&\left\| \left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{W}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d \right) \mathbf{x}_0 \right\|_2 \\
&= \left\| \left(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T} + \lambda\mathbf{U}\mathbf{U}^\mathsf{T})^{-1}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} - \mathbf{I}_d \right) \mathbf{x}_0 \right\|_2 \\
&= \left\| \left(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T} \left(\mathbf{U}\mathbf{\Sigma}(\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V} + \lambda\mathbf{\Sigma}^{-2})\mathbf{\Sigma}\mathbf{U}^\mathsf{T} \right)^{-1} \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} - \mathbf{I}_d \right) \mathbf{x}_0 \right\|_2 \\
&= \left\| \left(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T} \left(\mathbf{U}\mathbf{\Sigma}(\mathbf{I}_n + \widehat{\mathbf{E}} + \lambda\mathbf{\Sigma}^{-2})\mathbf{\Sigma}\mathbf{U}^\mathsf{T} \right)^{-1} \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} - \mathbf{I}_d \right) \mathbf{x}_0 \right\|_2 \\
&= \left\| \left(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T} \left(\mathbf{U}\mathbf{\Sigma}(\mathbf{G} + \widehat{\mathbf{E}})\mathbf{\Sigma}\mathbf{U}^\mathsf{T} \right)^{-1} \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} - \mathbf{I}_d \right) \mathbf{x}_0 \right\|_2 \\
&= \left\| \left(\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{\Sigma}^{-1}(\mathbf{G} + \widehat{\mathbf{E}})^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} - \mathbf{I}_d \right) \mathbf{x}_0 \right\|_2 , \qquad (91)
\end{aligned}
$$

where $\mathbf{G} = \mathbf{I}_n + \lambda\boldsymbol{\Sigma}^{-2}$ and is invertible. Further, using the similar argument as in Lemma 11, $(\mathbf{G} + \widehat{\mathbf{E}})^{-1}$ exists and eqn. (91) holds. Thus, we have

$$\begin{aligned}
&\left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{W}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 \\
&= \left\|\left(\mathbf{V}(\mathbf{G} + \widehat{\mathbf{E}})^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 = \left\|\left(\mathbf{V}\left(\mathbf{G}(\mathbf{I}_n + \mathbf{G}^{-1}\widehat{\mathbf{E}})\right)^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 \\
&= \left\|\left(\mathbf{V}\left(\mathbf{I}_n + \mathbf{G}^{-1}\widehat{\mathbf{E}}\right)^{-1}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 = \left\|\left(\mathbf{V}\left(\mathbf{I}_n + \widehat{\mathbf{R}}\right)\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 \\
&= \left\|\left(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0 + \mathbf{V}\widehat{\mathbf{R}}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T}\mathbf{x}_0\right\|_2 ,
\end{aligned} \tag{92}$$

where

$$\widehat{\mathbf{R}} = \sum_{\ell=1}^{\infty}(-1)^\ell\left(\mathbf{G}^{-1}\widehat{\mathbf{E}}\right)^\ell .$$

Using the same argument as in eqn.(38), we have $\|\mathbf{G}^{-1}\widehat{\mathbf{E}}\|_2 \le \varepsilon/2$, and by Proposition 8, $\mathbf{I}_n + \mathbf{G}^{-1}\widehat{\mathbf{E}}$ is invertible and $(\mathbf{I}_n + \mathbf{G}^{-1}\widehat{\mathbf{E}})^{-1} = \mathbf{I}_n + \widehat{\mathbf{R}}$. Thus eqn. (92) holds. Moreover, from eqn. (40), we have $\|\widehat{\mathbf{R}}\|_2 \le \varepsilon$.

Proceeding further, we have

$$\begin{aligned}
&\left\|\left(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0 + \mathbf{V}\widehat{\mathbf{R}}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T}\mathbf{x}_0\right\|_2 \\
&\quad\le \left\|\left(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 + \left\|\mathbf{V}\widehat{\mathbf{R}}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T}\mathbf{x}_0\right\|_2 \\
&\quad\le \left\|\left(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 + \|\widehat{\mathbf{R}}\|_2\left\|\mathbf{G}^{-1}\right\|_2\|\mathbf{x}_0\|_2 \\
&\quad\le \left\|\left(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 + \varepsilon\|\mathbf{x}_0\|_2 ,
\end{aligned} \tag{93}$$

where the first step is due to the triangle inequality, the second inequality follows from sub-multiplicativity and the last step holds as $\|\widehat{\mathbf{R}}\|_2 \le \varepsilon$ and $\|\mathbf{G}^{-1}\|_2 \le 1$.

Next, we seek to upper-bound $\|\mathbf{x}_0\|_2$ in terms of $\left\|\left(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2$. We begin by noticing that

$$\left\|\left(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 \ge \sigma_{\min}(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d)\|\mathbf{x}_0\|_2 . \tag{94}$$

Now, we need to bound the smallest singular value of $\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d$. We write

$$\begin{aligned}
\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d &= \mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \left(\mathbf{V}\mathbf{V}^\mathsf{T} + \mathbf{V}_\perp\mathbf{V}_\perp^\mathsf{T}\right) \\
&= \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix}\begin{pmatrix}\mathbf{G}^{-1} - \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_{d-n}\end{pmatrix}\begin{pmatrix}\mathbf{V}^\mathsf{T} \\ \mathbf{V}_\perp^\mathsf{T}\end{pmatrix} = \mathbf{V}_f\begin{pmatrix}\mathbf{G}^{-1} - \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_{d-n}\end{pmatrix}\mathbf{V}_f^\mathsf{T} ,
\end{aligned}$$

where $\mathbf{V}_f = \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix} \in \mathbb{R}^{d\times d}$ consisting of the right singular vectors in the full SVD representation of $\mathbf{A}$ with $\mathbf{V}_f\mathbf{V}_f^\mathsf{T} = \mathbf{V}_f^\mathsf{T}\mathbf{V}_f = \mathbf{I}_d$ and thus,

$$(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d)^2 = \mathbf{V}_f\underbrace{\begin{pmatrix}(\mathbf{G}^{-1} - \mathbf{I}_n)^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d-n}\end{pmatrix}}_{\mathbf{H}}\mathbf{V}_f^\mathsf{T} . \tag{95}$$

We observe that eqn. (95) is the SVD representation of $(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d)^2$. Since $\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d$ is symmetric, we have

$$\begin{aligned}
\sigma_{\min}^2(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d) &= \sigma_{\min}\left[(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d)^2\right] = \min_{1\le i\le d}\mathbf{H}_{ii} \\
&= \min_{1\le i\le n}\left\{\left(\frac{\sigma_i^2}{\sigma_i^2 + \lambda} - 1\right)^2, 1\right\} = \min_{1\le i\le n}\left(\frac{\sigma_i^2}{\sigma_i^2 + \lambda} - 1\right)^2 \\
&= \min_{1\le i\le n}\frac{\lambda^2}{(\lambda + \sigma_i^2)^2} = \frac{\lambda^2}{(\lambda + \sigma_1^2)^2} ,
\end{aligned}$$

and hence,

$$\sigma_{\min}(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d) = \frac{\lambda}{\lambda + \sigma_1^2}\,. \tag{96}$$

Therefore, combining eqns. (94) and (96), we have

$$\|\mathbf{x}_0\|_2 \leq \left(1 + \frac{\sigma_1^2}{\lambda}\right) \left\|\left(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2\,. \tag{97}$$

Again, combining eqns. (92), (93) and (97) yields

$$\begin{aligned}\left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{W}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 &\leq \left\|\left(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 + \varepsilon\left(1 + \frac{\sigma_1^2}{\lambda}\right)\left\|\left(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 \\ &= (1+\varepsilon\gamma_1)\left\|\left(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 \\ &= (1+\varepsilon\gamma_1)\left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2\,,\end{aligned} \tag{98}$$

where the last equality follows directly from Lemma 17.

Finally, squaring both sides of eqn. (98) concludes the proof. □

Final bound on the MSE. For $t = 1$, the MSE of the output of Algorithm 1 is given by

$$\begin{aligned}\mathrm{MSE}(\widehat{\mathbf{x}}^*) &= \sigma^2\left\|(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F^2 + \left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2^2 \\ &\leq \sigma^2(1+\varepsilon)^2\left\|(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F^2 + (1+\varepsilon\gamma_1)^2\left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2^2 \\ &\leq (1+\varepsilon\gamma_1)^2\left(\sigma^2\left\|(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F^2 + \left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2^2\right) \\ &= (1+\varepsilon\gamma_1)^2\ \mathrm{MSE}(\mathbf{x}^*) = \left(1 + 2\varepsilon\gamma_1 + \varepsilon^2\gamma_1^2\right)\mathrm{MSE}(\mathbf{x}^*) \\ &\leq \left(1 + 2\varepsilon\gamma_1^2 + \varepsilon\gamma_1^2\right)\mathrm{MSE}(\mathbf{x}^*) = \left(1 + 3\varepsilon\gamma_1^2\right)\mathrm{MSE}(\mathbf{x}^*)\,,\end{aligned}$$

where the first inequality directly follows from Lemma 18 and the second inequality is due to the fact that $\gamma_1 \geq 1$ as well as Lemma 16. The last inequality is again due to the facts that $\gamma_1 \geq 1$ and $\varepsilon < 1$.

E.2. Proof of Theorem 4 under eqn. (8)

First, we provide an alternative formulation of $\|(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\|_F$ and $\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\|_2$ using the thin SVD of $\mathbf{A}$.

Lemma 19. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be the design matrix and $\lambda(>0)$ be the ridge parameter of the ridge regression problem. Then, we have*

$$\textit{(a)}\quad \left\|(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F = \left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda^2\right\|_F \tag{99}$$

$$\textit{(b)}\quad \left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 = \left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2\,. \tag{100}$$

Proof. First, recall the matrix $\mathbf{\Sigma}_\lambda$ defined in eqn. (7). The proof directly follows from Lemma 17. Note that $\mathbf{\Sigma}_\lambda^2 = (\mathbf{I}_n + \lambda\mathbf{\Sigma}^{-2})^{-1}$ is the same as $\mathbf{G}^{-1}$ in Lemma 17. Thus, we have

$$\left\|(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F = \left\|\mathbf{\Sigma}^{-1}\mathbf{G}^{-1}\right\|_F = \left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda^2\right\|_F\,,$$

and

$$\left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 = \left\|\left(\mathbf{V}\mathbf{G}^{-1}\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2\,. = \left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2\,.$$

This concludes the proof. □

Our next result bounds both each term in eqn. (14) separately subject to the structural condition of eqn. (8).

Lemma 20. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{b} \in \mathbb{R}^n$, and $\lambda > 0$ be the inputs of the ridge regression problem. Let $\mathbf{S} \in \mathbb{R}^{d\times s}$ be the sketching matrix in Algorithm 1 and define,*

$$\mathbf{E} = \mathbf{\Sigma}_\lambda \mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\mathbf{\Sigma}_\lambda - \mathbf{\Sigma}_\lambda^2 .$$

Further, assume for some constant $0 < \varepsilon < 1$, if the condition of eqn. (8) is satisfied i.e. $\|\mathbf{E}\|_2 \leq \frac{\varepsilon}{4\sqrt{2}}$, then

$$\textit{(a)}\ \ \sigma^2 \left\|(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F^2 \leq\ (1+\varepsilon\gamma_2)^2\ \sigma^2 \left\|(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F^2 \tag{101}$$

$$\textit{(b)}\ \ \left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2^2 \leq (1+\varepsilon\gamma_2)^2 \left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2^2 , \tag{102}$$

where $\gamma_2 = \max\{\sqrt{1+\lambda/\sigma_n^2},\ 1+\sigma_1^2/\lambda\}$.

Proof. Let $\mathbf{W} = \mathbf{S}\mathbf{S}^\mathsf{T}$. As before, we start with the thin SVD representation of $\mathbf{A}$.

Part (a):

$$\begin{aligned}
\left\|(\mathbf{A}\mathbf{W}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F &= \left\|\left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma}^\mathsf{T}\mathbf{U}^\mathsf{T} + \lambda\mathbf{U}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\right\|_F \\
&= \left\|\mathbf{U}\left(\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma}^\mathsf{T} + \lambda\mathbf{I}_n\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\right\|_F \\
&= \left\|(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}(\mathbf{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma}_\lambda)\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma} + \lambda\mathbf{I}_n)^{-1}\mathbf{\Sigma}\right\|_F \\
&= \left\|(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}(\mathbf{\Sigma}_\lambda^2 + \mathbf{E})\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma} + \lambda\mathbf{I}_n)^{-1}\mathbf{\Sigma}\right\|_F \qquad (103) \\
&= \left\|(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}(\mathbf{\Sigma}_\lambda^2 + \mathbf{E})\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma} + \lambda\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}(\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-2}\mathbf{\Sigma}_\lambda)\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma})^{-1}\mathbf{\Sigma}\right\|_F \\
&= \left\|(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}(\mathbf{\Sigma}_\lambda^2 + \mathbf{E} + \lambda\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-2}\mathbf{\Sigma}_\lambda)\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma})^{-1}\mathbf{\Sigma}\right\|_F \\
&= \left\|(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}(\mathbf{I}_n + \mathbf{E})\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma})^{-1}\mathbf{\Sigma}\right\|_F . \qquad (104)
\end{aligned}$$

In eqn. (103), we used the fact that $\mathbf{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma}_\lambda = \mathbf{\Sigma}_\lambda^2 + \mathbf{E}$. Further, eqn. (104) holds as $(\mathbf{\Sigma}_\lambda^2 + \lambda\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-2}\mathbf{\Sigma}_\lambda) \in \mathbb{R}^{n\times n}$ is a diagonal matrix with i-th diagonal entry equal to

$$\left(\mathbf{\Sigma}_\lambda^2 + \lambda\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-2}\mathbf{\Sigma}_\lambda\right)_{ii} = \frac{\sigma_i^2}{\sigma_i^2+\lambda} + \frac{\lambda}{\sigma_i^2+\lambda} = 1$$

for any $i = 1, 2, \ldots n$. Thus, we have $\left(\mathbf{\Sigma}_\lambda^2 + \lambda\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-2}\mathbf{\Sigma}_\lambda\right) = \mathbf{I}_n$.

Since $\|\mathbf{E}\|_2 < 1$, taking $\mathbf{P} = -\mathbf{E}$ in Proposition 8 implies that $(\mathbf{I}_n + \mathbf{E})^{-1}$ exists and $(\mathbf{I}_n + \mathbf{E})^{-1} = \mathbf{I}_n + \sum_{\ell=1}^\infty (-1)^\ell \mathbf{E}^\ell$. Let $\mathbf{R} = \sum_{\ell=1}^\infty (-1)^\ell \mathbf{E}^\ell$. Then, eqn. (104) can further be simplified as

$$\begin{aligned}
\left\|(\mathbf{A}\mathbf{W}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F &= \left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda(\mathbf{I}_n + \mathbf{E})^{-1}\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\right\|_F = \left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda(\mathbf{I}_n + \mathbf{E})^{-1}\mathbf{\Sigma}_\lambda\right\|_F \\
&= \left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda(\mathbf{I}_n + \mathbf{R})\mathbf{\Sigma}_\lambda\right\|_F = \left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda^2 + \mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda\mathbf{R}\mathbf{\Sigma}_\lambda\right\|_F \\
&\leq \left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda^2\right\|_F + \left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda\mathbf{R}\mathbf{\Sigma}_\lambda\right\|_F = \left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda^2\right\|_F + \left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda^2\mathbf{\Sigma}_\lambda^{-1}\mathbf{R}\mathbf{\Sigma}_\lambda\right\|_F \\
&\leq \left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda^2\right\|_F + \|\mathbf{R}\|_2 \left\|\mathbf{\Sigma}_\lambda^{-1}\right\|_2 \|\mathbf{\Sigma}_\lambda\|_2 \left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda^2\right\|_F , \qquad (105)
\end{aligned}$$

where the first inequality follows from the triangle inequality and the second inequality is due to strong-sub-multiplicativity. For the second term on the right hand side of eqn. (105), we have $\|\mathbf{R}\|_2 \leq \frac{\varepsilon}{2\sqrt{2}}$ (by eqn. (18)), $\|\mathbf{\Sigma}_\lambda^{-1}\|_2 = \sqrt{1+\lambda/\sigma_n^2}$ and $\|\mathbf{\Sigma}_\lambda\|_2 \leq 1$. Using these facts, eqn. (105) boils down to

$$\begin{aligned}
\left\|(\mathbf{A}\mathbf{W}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F &\leq \left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda^2\right\|_F + \frac{\varepsilon}{2\sqrt{2}}\sqrt{1+\frac{\lambda}{\sigma_n^2}}\left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda^2\right\|_F \\
&\leq (1+\varepsilon\gamma_2)\left\|\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda^2\right\|_F = (1+\varepsilon\gamma_2)\left\|(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F , \qquad (106)
\end{aligned}$$

where the second inequality follows from the facts: $\frac{1}{2\sqrt{2}} < 1$ and $\sqrt{1+\frac{\lambda}{\sigma_n^2}} \leq \gamma_2$.The last step is due to Lemma 19. Finally, squaring both sides of eqn.(106) and then pre-multiplying by σ^2 concludes the proof.

Part (b): We have

$$\begin{aligned}
\left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{W}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{A}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2 &= \left\|\left(\mathbf{V}\mathbf{\Sigma}^\mathsf{T}\mathbf{U}^\mathsf{T}\left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma}^\mathsf{T}\mathbf{U}^\mathsf{T}+\lambda\mathbf{U}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2 \\
&= \left\|\left(\mathbf{V}\mathbf{\Sigma}^\mathsf{T}\left(\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma}^\mathsf{T}+\lambda\mathbf{I}_n\right)^{-1}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2 \\
&= \left\|\left(\mathbf{V}\mathbf{\Sigma}^\mathsf{T}\left(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}(\mathbf{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma}_\lambda)\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma}^\mathsf{T}+\lambda\mathbf{I}_n\right)^{-1}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2 \\
&= \left\|\left(\mathbf{V}\mathbf{\Sigma}^\mathsf{T}\left(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}(\mathbf{\Sigma}_\lambda^2+\mathbf{E})\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma}^\mathsf{T}+\lambda\mathbf{I}_n\right)^{-1}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2 , \qquad (107)
\end{aligned}$$

where we used the fact that $\mathbf{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\mathbf{W}\mathbf{V}\mathbf{\Sigma}_\lambda = \mathbf{\Sigma}_\lambda^2+\mathbf{E}$. Proceeding in the same way as in the proof of part (a), we have

$$\begin{aligned}
\left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{W}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{A}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2 &= \left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda(\mathbf{I}_d+\mathbf{R})\mathbf{\Sigma}_\lambda\mathbf{V}^\mathsf{T}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2 \\
&= \left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\mathbf{I}_d+\mathbf{V}\mathbf{\Sigma}_\lambda\mathbf{R}\mathbf{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\right)\mathbf{x}_0\right\|_2 \\
&\leq \left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2 + \left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda\mathbf{R}\mathbf{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\right)\mathbf{x}_0\right\|_2 \\
&\leq \left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2 + \left\|\mathbf{V}\mathbf{\Sigma}_\lambda\mathbf{R}\mathbf{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\right\|_2\|\mathbf{x}_0\|_2 \\
&= \left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2 + \|\mathbf{\Sigma}_\lambda\mathbf{R}\mathbf{\Sigma}_\lambda\|_2\|\mathbf{x}_0\|_2 \\
&\leq \left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2 + \|\mathbf{R}\|_2\|\mathbf{x}_0\|_2 \\
&\leq \left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2 + \frac{\varepsilon}{2\sqrt{2}}\|\mathbf{x}_0\|_2 , \qquad (108)
\end{aligned}$$

where $\mathbf{R} = \sum_{\ell=1}^{\infty}(-1)^\ell\mathbf{E}^\ell$. In the above expression, the first inequality follows from the triangle inequality, the second and third inequalities are due to sub-multiplicativity and the fact that $\|\mathbf{\Sigma}_\lambda\|_2 \leq 1$. The final inequality holds as $\|\mathbf{R}\|_2 \leq \frac{\varepsilon}{2\sqrt{2}}$ by eqn. (18).

Note that

$$\left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\mathbf{I}_d\right)\mathbf{x}_0\right\|_2 \geq \sigma_{\min}(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\mathbf{I}_d)\|\mathbf{x}_0\|_2 . \qquad (109)$$

We seek to bound the smallest singular value of $\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\mathbf{I}_d$ which can be expressed as

$$\begin{aligned}
\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\mathbf{I}_d =& \mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\left(\mathbf{V}\mathbf{V}^\mathsf{T}+\mathbf{V}_\perp\mathbf{V}_\perp^\mathsf{T}\right) \\
=& \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma}_\lambda^2-\mathbf{I}_n & \mathbf{0}\\ \mathbf{0} & -\mathbf{I}_{d-n}\end{pmatrix}\begin{pmatrix}\mathbf{V}^\mathsf{T}\\ \mathbf{V}_\perp^\mathsf{T}\end{pmatrix} = \mathbf{V}_f\begin{pmatrix}\mathbf{\Sigma}_\lambda^2-\mathbf{I}_n & \mathbf{0}\\ \mathbf{0} & -\mathbf{I}_{d-n}\end{pmatrix}\mathbf{V}_f^\mathsf{T} ,
\end{aligned}$$

where $\mathbf{V}_f = \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix} \in \mathbb{R}^{d\times d}$ consisting of the right singular vectors in the full SVD representation of $\mathbf{A}$ with $\mathbf{V}_f\mathbf{V}_f^\mathsf{T} = \mathbf{V}_f^\mathsf{T}\mathbf{V}_f = \mathbf{I}_d$ and thus,

$$(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\mathbf{I}_d)^2 = \mathbf{V}_f\underbrace{\begin{pmatrix}(\mathbf{\Sigma}_\lambda^2-\mathbf{I}_n)^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{d-n}\end{pmatrix}}_{\mathbf{H}}\mathbf{V}_f^\mathsf{T}. \qquad (110)$$

Observe that eqn. (110) is the SVD representation of $(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\mathbf{I}_d)^2$. Since $\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\mathbf{I}_d$ is symmetric, we have

$$\begin{aligned}
\sigma_{\min}^2(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\mathbf{I}_d) &= \sigma_{\min}\left[(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}-\mathbf{I}_d)^2\right] = \min_{1\leq i\leq d}\mathbf{H}_{ii} \\
&= \min_{1\leq i\leq n}\left\{\left(\frac{\sigma_i^2}{\sigma_i^2+\lambda}-1\right)^2, 1\right\} = \min_{1\leq i\leq n}\left(\frac{\sigma_i^2}{\sigma_i^2+\lambda}-1\right)^2 \\
&= \min_{1\leq i\leq n}\frac{\lambda^2}{(\lambda+\sigma_i^2)^2} = \frac{\lambda^2}{(\lambda+\sigma_1^2)^2}
\end{aligned}$$

and hence

$$\sigma_{\min}(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T} - \mathbf{I}_d) = \frac{\lambda}{\lambda + \sigma_1^2}. \tag{111}$$

Therefore, combining eqns. (109) and (111), we have

$$\|\mathbf{x}_0\|_2 \leq \left(1 + \frac{\sigma_1^2}{\lambda}\right) \left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2. \tag{112}$$

Finally, combining eqns. (108) and (112), we obtain

$$\begin{aligned}
\left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{W}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 &\leq \left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 + \frac{\varepsilon}{2\sqrt{2}}\left(1 + \frac{\sigma_1^2}{\lambda}\right)\left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 \\
&\leq \left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 + \varepsilon\gamma_2\left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 \\
&= (1 + \varepsilon\gamma_2)\left\|\left(\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2 \\
&= (1 + \varepsilon\gamma_2)\left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2,
\end{aligned} \tag{113}$$

where the second inequality follows from the facts: $\frac{1}{2\sqrt{2}} < 1$ and $\left(1 + \frac{\sigma_1^2}{\lambda}\right) \leq \gamma_2$. The last step is due to Lemma 19. Finally, squaring both sides of eqn. (113) concludes the proof. □

Final bund on the MSE. For $t = 1$, MSE of the output of Algorithm 1 is given by

$$\begin{aligned}
\mathrm{MSE}(\widehat{\mathbf{x}}^*) &= \sigma^2\left\|(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F^2 + \left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2^2 \\
&\leq \sigma^2(1 + \varepsilon\gamma_2)^2\left\|(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F^2 + (1 + \varepsilon\gamma_2)^2\left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2^2 \\
&= (1 + \varepsilon\gamma_2)^2\left(\sigma^2\left\|(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\right\|_F^2 + \left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2^2\right) \\
&= (1 + \varepsilon\gamma_2)^2\ \mathrm{MSE}(\mathbf{x}^*) = \left(1 + 2\varepsilon\gamma_2 + \varepsilon^2\gamma_2^2\right)\ \mathrm{MSE}(\mathbf{x}^*) \\
&\leq \left(1 + 2\varepsilon\gamma_2^2 + \varepsilon\gamma_2^2\right)\ \mathrm{MSE}(\mathbf{x}^*) = \left(1 + 3\varepsilon\gamma_2^2\right)\ \mathrm{MSE}(\mathbf{x}^*),
\end{aligned}$$

where the first inequality directly follows from Lemma 20, the third equality follows from Lemma 16, and the last inequality is due to the facts that $\gamma_2 \geq 1$ and $\varepsilon < 1$.

F. Ridge Leverge Scores

In this section, we begin by revisiting the definition of ridge leverage scores (Cohen et al., 2017) and then provide an alternative expression that is easier to work with.

Definition 1. *The i-th column ridge leverage score of the matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ with respect to the ridge parameter $\lambda > 0$ is defined as*

$$\tau_i^\lambda \triangleq \left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A}\right)_{ii}, \tag{114}$$

for $i = 1, 2, \ldots, d$.

In the next result, we present a more compact version of eqn. (114) using the thin SVD representation of $\mathbf{A}$.

Lemma 21. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be the design matrix and $\lambda > 0$ be the ridge parameter. Eqn. (114) can also be expressed as*

$$\tau_i^\lambda = \|(\mathbf{V}\mathbf{\Sigma}_\lambda)_{i*}\|_2^2, \tag{115}$$

for $i = 1, 2, \ldots, d$.

Proof. First, using the fact $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$, we have

$$\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{A} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T}\left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\mathsf{T} + \lambda\mathbf{U}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$$

$$
\begin{aligned}
&= \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\mathsf{T}\left(\mathbf{U}\boldsymbol{\Sigma}^2\mathbf{U}^\mathsf{T} + \lambda\mathbf{U}\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T} \\
&= \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\mathsf{T}\left(\mathbf{U}(\boldsymbol{\Sigma}^2 + \lambda\mathbf{I}_n)\mathbf{U}^\mathsf{T}\right)^{-1}\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T} \\
&= \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\mathsf{T}\mathbf{U}(\boldsymbol{\Sigma}^2 + \lambda\mathbf{I}_n)^{-1}\mathbf{U}^\mathsf{T}\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T} \\
&= \mathbf{V}\underbrace{\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^2 + \lambda\mathbf{I}_n)^{-1}\boldsymbol{\Sigma}}_{\boldsymbol{\Sigma}_\lambda^2}\mathbf{V}^\mathsf{T}, \qquad (116)
\end{aligned}
$$

where we used the facts that $\mathbf{U}\mathbf{U}^\mathsf{T} = \mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}_n$, $\mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}_n$, and $(\boldsymbol{\Sigma}^2 + \lambda\mathbf{I}_n)$ is invertible. Now, combining eqn. (114) and eqn. (116), we have

$$
\tau_i^\lambda = \left(\mathbf{V}\boldsymbol{\Sigma}_\lambda^2\mathbf{V}^\mathsf{T}\right)_{ii} = (\mathbf{V})_{i*}\,\boldsymbol{\Sigma}_\lambda^2\left(\mathbf{V}^\mathsf{T}\right)_{*i} = \|\mathbf{V}_{i*}\boldsymbol{\Sigma}_\lambda\|_2^2 = \|(\mathbf{V}\boldsymbol{\Sigma}_\lambda)_{i*}\|_2^2 .
$$

This concludes the proof. □

G. Proof of Theorem 3

This result is similar in spirit to Theorem 4.2 of Holodnak & Ipsen (2015), but our objective and (therefore) the analysis are slightly different in two ways. First, Holodnak & Ipsen (2015) presented a probabilistic bound for the 2-norm of the relative error whereas our bound holds for the 2-norm of the absolute error. Second, we have an additional condition $\|\mathbf{X}\|_2 \leq 1$ which enables us to come up with a minimum value for s that depends only on $\|\mathbf{X}\|_F^2$ and not on the stable rank of $\mathbf{X}$.

We first state two auxiliary results: a stable rank (intrinsic dimension) matrix Bernstein concentration inequality (Theorem 22) and a bound for the singular values of a difference of positive semi-definite matrices (Theorem 23). We then utilize these two results to obtain a proof of Theorem 3.

Theorem 22. (Theorem 7.3.1 of Tropp (2015)) *Let $\mathbf{Y}_j$ be s independent real symmetric random matrices, with $\mathbb{E}(\mathbf{Y}_j) = \mathbf{0}$, $j = 1, 2, \ldots, s$. Let $\max_{1\leq j\leq s}\|\mathbf{Y}_j\|_2 \leq \rho_1$ and $\mathbf{P}$ be a symmetric positive semi-definite matrix such that $\sum_{j=1}^s \mathbb{E}(\mathbf{Y}_j^2) \preccurlyeq \mathbf{P}$. Then, for any $\varepsilon \geq \|\mathbf{P}\|_2^{1/2} + \rho_1/3$, we have*

$$
\mathbb{P}\left(\left\|\sum_{j=1}^s \mathbf{Y}_j\right\|_2 \geq \varepsilon\right) \leq 4\,\mathrm{intdim}(\mathbf{P})\exp\left(-\frac{\varepsilon^2/2}{\|\mathbf{P}\|_2 + \rho_1\varepsilon/3}\right),
$$

where $\mathrm{intdim}(\mathbf{P}) \triangleq \mathrm{tr}(\mathbf{P})/\|\mathbf{P}\|_2$.

Theorem 23. (Theorem 2.1 of Zhan (2001)) *If $\mathbf{M}$ and $\mathbf{N}$ are real symmetric positive semi-definite matrices $\in \mathbb{R}^{m\times m}$, with singular values $\sigma_1(\mathbf{M}) \geq \sigma_2(\mathbf{M}) \geq \cdots \geq \sigma_m(\mathbf{M})$ and $\sigma_1(\mathbf{N}) \geq \sigma_2(\mathbf{N}) \geq \cdots \geq \sigma_m(\mathbf{N})$, then the singular values of the difference $\mathbf{M} - \mathbf{N}$ is bounded by*

$$
\sigma_j(\mathbf{M} - \mathbf{N}) \leq \sigma_j\begin{pmatrix}\mathbf{M} & \mathbf{0}\\ \mathbf{0} & \mathbf{N}\end{pmatrix}, \qquad 1 \leq j \leq m.
$$

In particular, we have $\|\mathbf{M} - \mathbf{N}\|_2 \leq \max\{\|\mathbf{M}\|_2, \|\mathbf{N}\|_2\}$.

Proof of Theorem 3. Let $\mathrm{rank}(\mathbf{X}) = \rho$ and $\mathbf{X} = \mathbf{U}_\mathbf{X}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{V}_\mathbf{X}^\mathsf{T}$ be the thin SVD representation of $\mathbf{X}$ with $\mathbf{U}_\mathbf{X} \in \mathbb{R}^{d\times\rho}$, $\mathbf{V}_\mathbf{X} \in \mathbb{R}^{n\times\rho}$ such that $\mathbf{U}_\mathbf{X}^\mathsf{T}\mathbf{U}_\mathbf{X} = \mathbf{V}_\mathbf{X}^\mathsf{T}\mathbf{V}_\mathbf{X} = \mathbf{I}_\rho$. Also, $\boldsymbol{\Sigma}_\mathbf{X} \in \mathbb{R}^{\rho\times\rho}$ is the diagonal matrix consisting of the non-zero singular values of $\mathbf{X}$ arranged in a non-increasing order *i.e.* $\sigma_1(\mathbf{X}) \geq \sigma_2(\mathbf{X}) \geq \cdots \geq \sigma_\rho(\mathbf{X}) > 0$. Further, according to the statement of the theorem, we have, $\|\mathbf{X}\|_2 = \sigma_1(\mathbf{X}) \leq 1$.

Setting $\mathbf{C} = \mathbf{X}^\mathsf{T}\mathbf{S}$, we have

$$
\begin{aligned}
\mathbf{X}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{X} - \mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{C}\mathbf{C}^\mathsf{T} - \mathbf{X}^\mathsf{T}\mathbf{X} &= \left(\sum_{j=1}^s \mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*}\right) - \mathbf{X}^\mathsf{T}\mathbf{X} \\
&= \sum_{j=1}^s\left(\mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*} - \frac{1}{s}\mathbf{X}^\mathsf{T}\mathbf{X}\right) = \sum_{j=1}^s \mathbf{Y}_j\,, \qquad (117)
\end{aligned}
$$

where $\mathbf{Y}_j = \mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*} - \frac{1}{s}\mathbf{X}^\mathsf{T}\mathbf{X}$.

Clearly,

$$\begin{aligned}\mathbb{E}(\mathbf{Y}_j) &= \mathbb{E}\left(\mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*} - \frac{1}{s}\mathbf{X}^\mathsf{T}\mathbf{X}\right) = \mathbb{E}\left(\mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*}\right) - \frac{1}{s}\mathbf{X}^\mathsf{T}\mathbf{X} \\ &= \sum_{i=1}^{d}\left(\frac{(\mathbf{X}^\mathsf{T})_{*i}}{\sqrt{sp_i}}\frac{\mathbf{X}_{i*}}{\sqrt{sp_i}}\right)p_i - \frac{1}{s}\mathbf{X}^\mathsf{T}\mathbf{X} = \frac{1}{s}\sum_{i=1}^{d}(\mathbf{X}^\mathsf{T})_{*i}\mathbf{X}_{i*} - \frac{1}{s}\mathbf{X}^\mathsf{T}\mathbf{X} \\ &= \frac{1}{s}\mathbf{X}^\mathsf{T}\mathbf{X} - \frac{1}{s}\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{0}, \end{aligned} \tag{118}$$

where the third equality follows from Algorithm 2 and the definition of expectation. Thus, we have shown that $\mathbf{Y}_j$'s have zero mean. Next, we check that the assumptions of Theorem 22 are satisfied.

Bound for $\max_{1\le j\le s}\|\mathbf{Y}_j\|_2$. As per eqn. (117), $\mathbf{Y}_j = \mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*} - \frac{1}{s}\mathbf{X}^\mathsf{T}\mathbf{X}$ is a difference of two positive semi-definite matrices. We apply Theorem 23 to obtain

$$\begin{aligned}\|\mathbf{Y}_j\|_2 &= \left\|\mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*} - \frac{1}{s}\mathbf{X}^\mathsf{T}\mathbf{X}\right\|_2 \le \max\left\{\left\|\mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*}\right\|_2, \left\|\frac{1}{s}\mathbf{X}^\mathsf{T}\mathbf{X}\right\|_2\right\} \\ &\le \max_{1\le i\le d}\left\{\left\|\frac{(\mathbf{X}^\mathsf{T})_{*i}}{\sqrt{sp_i}}\frac{\mathbf{X}_{i*}}{\sqrt{sp_i}}\right\|_2, \left\|\frac{1}{s}\mathbf{X}^\mathsf{T}\mathbf{X}\right\|_2\right\} = \frac{1}{s}\max_{1\le i\le d}\left\{\frac{\|\mathbf{X}_{i*}\|_2^2}{p_i}, \|\mathbf{X}\|_2^2\right\} \\ &= \frac{1}{s}\max_{1\le i\le d}\left\{\frac{\|\mathbf{X}_{i*}\|_2^2}{(\|\mathbf{X}_{i*}\|_2^2/\|\mathbf{X}\|_F^2)}, \|\mathbf{X}\|_2^2\right\} = \frac{\|\mathbf{X}\|_F^2}{s}, \end{aligned} \tag{119}$$

which holds for all $j = 1, 2, \ldots s$.

Thus, we have shown that $\max_{1\le j\le s}\|\mathbf{Y}_j\|_2 \le \frac{\|\mathbf{X}\|_F^2}{s} \triangleq \rho_1$.

The matrix $\mathbf{P}$. From the definition of $\mathbf{Y}_j$ in eqn. (117), we have

$$\begin{aligned}\mathbf{Y}_j = \mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*} - \frac{1}{s}\mathbf{X}^\mathsf{T}\mathbf{X} &\Rightarrow \mathbf{Y}_j + \frac{1}{s}\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*} \\ &\Rightarrow \left(\mathbf{Y}_j + \frac{1}{s}\mathbf{X}^\mathsf{T}\mathbf{X}\right)^2 = \left(\mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*}\right)^2 = \mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*}\mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*} \\ &\Rightarrow \mathbf{Y}_j^2 - \mathbf{Y}_j\mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{Y}_j + \frac{1}{s^2}(\mathbf{X}^\mathsf{T}\mathbf{X})^2 = \mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*}\mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*}. \end{aligned} \tag{120}$$

Taking expectations on both sides of eqn. (120) and noting that $\mathbb{E}(\mathbf{Y}_j) = \mathbf{0}$ gives

$$\begin{aligned}\mathbb{E}(\mathbf{Y}_j^2) + \frac{1}{s^2}(\mathbf{X}^\mathsf{T}\mathbf{X})^2 &= \mathbb{E}(\mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*}\mathbf{C}_{*j}(\mathbf{C}^\mathsf{T})_{j*}) \\ &= \sum_{i=1}^{d}\left(\frac{(\mathbf{X}^\mathsf{T})_{*i}}{\sqrt{sp_i}}\frac{\mathbf{X}_{i*}}{\sqrt{sp_i}}\frac{(\mathbf{X}^\mathsf{T})_{*i}}{\sqrt{sp_i}}\frac{\mathbf{X}_{i*}}{\sqrt{sp_i}}\right)p_i \\ &= \frac{1}{s^2}\sum_{i=1}^{d}(\mathbf{X}^\mathsf{T})_{*i}\left(\frac{\|\mathbf{X}_{i*}\|_2^2}{p_i}\right)\mathbf{X}_{i*} = \frac{1}{s^2}\sum_{i=1}^{d}\left(\frac{\|\mathbf{X}_{i*}\|_2^2}{\|\mathbf{X}_{i*}\|_2^2/\|\mathbf{X}\|_F^2}\right)(\mathbf{X}^\mathsf{T})_{*i}\mathbf{X}_{i*} \\ &= \frac{\|\mathbf{X}\|_F^2}{s^2}\sum_{i=1}^{d}(\mathbf{X}^\mathsf{T})_{*i}\mathbf{X}_{i*} = \frac{\|\mathbf{X}\|_F^2}{s^2}\mathbf{X}^\mathsf{T}\mathbf{X}. \end{aligned} \tag{121}$$

Summing both sides of eqn. (121) over j gives

$$\sum_{j=1}^{s}\mathbb{E}(\mathbf{Y}_j^2) = \frac{\|\mathbf{X}\|_F^2}{s}\mathbf{X}^\mathsf{T}\mathbf{X} - \frac{1}{s}(\mathbf{X}^\mathsf{T}\mathbf{X})^2$$

$$\begin{aligned} &\preccurlyeq \frac{\|\mathbf{X}\|_F^2}{s}\mathbf{X}^\mathsf{T}\mathbf{X} = \frac{\|\mathbf{X}\|_F^2}{s}\mathbf{V}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}^2\mathbf{V}_\mathbf{X}^\mathsf{T} \\ &\preccurlyeq \frac{\|\mathbf{X}\|_F^2}{s}\mathbf{V}_\mathbf{X}\mathbf{D}\mathbf{V}_\mathbf{X}^\mathsf{T} \triangleq \mathbf{P}, \end{aligned} \tag{122}$$

where $\mathbf{D} \in \mathbb{R}^{\rho\times\rho}$ is diagonal matrix whose i-th diagonal entry is equal to

$$\mathbf{D}_{ii} = \begin{cases} 1 & \text{if} \quad i = 1 \\ \sigma_i^2(\mathbf{X}) & \text{otherwise.} \end{cases}$$

The second-to-last inequality in eqn. (122) holds because $\sum_{j=1}^s \mathbb{E}(\mathbf{Y}_j^2)$, $\frac{\|\mathbf{X}\|_F^2}{s}\mathbf{X}^\mathsf{T}\mathbf{X}$ and $\frac{1}{s}(\mathbf{X}^\mathsf{T}\mathbf{X})^2$ are all positive semi-definite matrices. Further, the last inequality follows from the fact that $\mathbf{\Sigma}_\mathbf{X}^2 \preccurlyeq \mathbf{D}$ as $\sigma_1(\mathbf{X}) = \|\mathbf{X}\|_2 \leq 1$.

Note that $\|\mathbf{D}\|_2 = 1$ and

$$\begin{aligned} \operatorname{tr}(\mathbf{D}) &= 1 + \sum_{i=2}^{\rho}\sigma_i^2(\mathbf{X}) = 1 - \sigma_1^2(\mathbf{X}) + \sum_{i=1}^{\rho}\sigma_i^2(\mathbf{X}) \\ &= 1 - \sigma_1^2(\mathbf{X}) + \|\mathbf{X}\|_F^2 \leq 1 + \|\mathbf{X}\|_F^2. \end{aligned} \tag{123}$$

Again,

$$\|\mathbf{P}\|_2 = \frac{\|\mathbf{X}\|_F^2}{s}\left\|\mathbf{V}_\mathbf{X}\mathbf{D}\mathbf{V}_\mathbf{X}^\mathsf{T}\right\|_2 = \frac{\|\mathbf{X}\|_F^2}{s}\|\mathbf{D}\|_2 = \frac{\|\mathbf{X}\|_F^2}{s}, \tag{124}$$

where the second equality follows from the unitary invariance of 2-norm.

Similarly, from eqn. (123)

$$\operatorname{tr}(\mathbf{P}) = \frac{\|\mathbf{X}\|_F^2}{s}\operatorname{tr}\left(\mathbf{V}_\mathbf{X}\mathbf{D}\mathbf{V}_\mathbf{X}^\mathsf{T}\right) = \frac{\|\mathbf{X}\|_F^2}{s}\operatorname{tr}(\mathbf{D}) \leq \frac{\|\mathbf{X}\|_F^2}{s}(1 + \|\mathbf{X}\|_F^2). \tag{125}$$

Combining eqns. (124) and (125) yields

$$\operatorname{intdim}(\mathbf{P}) = \frac{\operatorname{tr}(\mathbf{P})}{\|\mathbf{P}\|_2} \leq \frac{\frac{\|\mathbf{X}\|_F^2}{s}(1 + \|\mathbf{X}\|_F^2)}{\frac{\|\mathbf{X}\|_F^2}{s}} = 1 + \|\mathbf{X}\|_F^2. \tag{126}$$

Application of Theorem 22. From eqn. (117), we have

$$\mathbb{P}\left(\left\|\mathbf{X}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{X} - \mathbf{X}^\mathsf{T}\mathbf{X}\right\|_2 > \varepsilon\right) = \mathbb{P}\left(\left\|\sum_{j=1}^s \mathbf{Y}_j\right\|_2 > \varepsilon\right). \tag{127}$$

Applying Theorem 22 to the right hand side of eqn. (127) yields:

$$\begin{aligned} \mathbb{P}\left(\left\|\sum_{j=1}^s \mathbf{Y}_j\right\|_2 > \varepsilon\right) &\leq 4\operatorname{intdim}(\mathbf{P})\exp\left(-\frac{\varepsilon^2/2}{\|\mathbf{P}\|_2 + \rho_1\varepsilon/3}\right) \\ &\leq 4(1 + \|\mathbf{X}\|_F^2)\exp\left(-\frac{\varepsilon^2/2}{\frac{\|\mathbf{X}\|_F^2}{s} + \frac{\varepsilon\|\mathbf{X}\|_F^2}{3s}}\right) \\ &= 4(1 + \|\mathbf{X}\|_F^2)\exp\left(-\frac{s\varepsilon^2}{\|\mathbf{X}\|_F^2(2 + 2\varepsilon/3)}\right). \end{aligned} \tag{128}$$

Clearly, $\mathbb{P}\left(\left\|\mathbf{X}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{X} - \mathbf{X}^\mathsf{T}\mathbf{X}\right\|_2 > \varepsilon\right) \leq \delta$ holds if the right hand side of eqn. (128) is at most δ, *i.e.*,

$$4(1 + \|\mathbf{X}\|_F^2)\exp\left(-\frac{s\varepsilon^2}{\|\mathbf{X}\|_F^2(2 + 2\varepsilon/3)}\right) \leq \delta \iff \frac{s\varepsilon^2}{\|\mathbf{X}\|_F^2(2 + 2\varepsilon/3)} \geq \ln\left(\frac{4(1 + \|\mathbf{X}\|_F^2)}{\delta}\right)$$

$$\iff s \geq \left(2 + \frac{2\varepsilon}{3}\right) \frac{\|\mathbf{X}\|_F^2}{\varepsilon^2} \ln\left(\frac{4(1+\|\mathbf{X}\|_F^2)}{\delta}\right). \tag{129}$$

As $\varepsilon \leq 1$, eqn. (129) holds if

$$s \geq \frac{8\|\mathbf{X}\|_F^2}{3\,\varepsilon^2} \ln\left(\frac{4(1+\|\mathbf{X}\|_F^2)}{\delta}\right).$$

Finally, it still remains to be shown that the last condition of Theorem 22 is satisfied, *i.e.*, $\varepsilon \geq \|\mathbf{P}\|_2^{1/2} + \rho_1/3$. We solve the following equation for ε:

$$\begin{aligned}
&4\big(1+\|\mathbf{X}\|_F^2\big)\exp\left(-\frac{s\varepsilon^2}{\|\mathbf{X}\|_F^2\,(2+2\varepsilon/3)}\right) = \delta \\
\implies &3s\varepsilon^2 - 2\|\mathbf{X}\|_F^2 \ln\left(\frac{4(1+\|\mathbf{X}\|_F^2)}{\delta}\right)\varepsilon - 6\|\mathbf{X}\|_F^2 \ln\left(\frac{4(1+\|\mathbf{X}\|_F^2)}{\delta}\right) = 0 \\
\implies &\varepsilon = \beta + \sqrt{\beta^2 + 6\beta},
\end{aligned}$$

where

$$\beta = \frac{\|\mathbf{X}\|_F^2 \ln\left(\frac{4(1+\|\mathbf{X}\|_F^2)}{\delta}\right)}{3s}.$$

Observe that $\varepsilon \geq \frac{\rho_1}{3} + \|\mathbf{P}\|_2^{1/2}$ if $\beta \geq \frac{\rho_1}{3}$ and $6\beta \geq \|\mathbf{P}\|_2$. Both conditions will be satisfied if

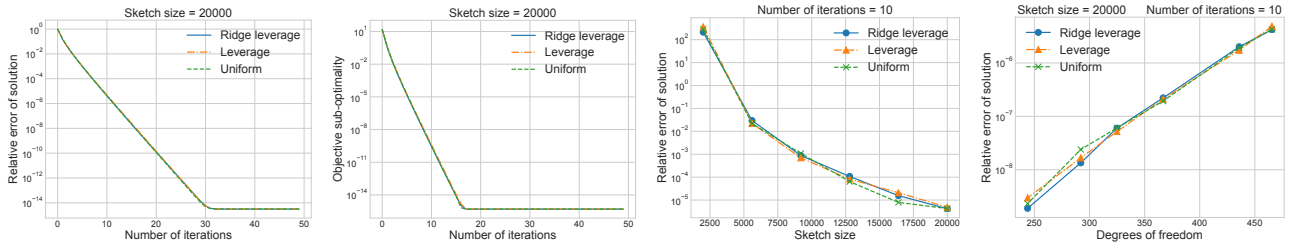
$$\ln\left(\frac{4(1+\|\mathbf{X}\|_F^2)}{\delta}\right) \geq 1 \iff \delta \leq \frac{4(1+\|\mathbf{X}\|_F^2)}{e},$$

which is always true since $\delta < 1$. This concludes the proof.

H. Additional Experiment Results

H.1. Synthetic Data Experiments

We generate synthetic data using the same mechanism as Chen et al. (2015). Specifically, we construct the $n \times d$ design matrix via $\mathbf{A} = \mathbf{M}\mathbf{D}\mathbf{V}^\top + \alpha\mathbf{E}$, where $\mathbf{M}$ is an $n \times s$ matrix with *i.i.d.* standard Gaussian entries; $\mathbf{D}$ is an $s \times s$ diagonal matrix with diagonal entries $D_{ii} = 1 - {}^{(i-1)}/_{d}$ for each $i = 1, \ldots, s$; and $\mathbf{V}$ is a $d \times n$ column-orthonormal matrix containing a random s-dimensional subspace of $\mathbb{R}^d$. Note that $\mathbf{M}\mathbf{D}\mathbf{V}^\top$ is a rank s matrix with linearly decreasing singular values. Further, $\mathbf{E}$ is an $n \times d$ noise matrix with *i.i.d.* standard Gaussian entries; and $\alpha > 0$ balances the strength of the signals $\mathbf{M}\mathbf{D}\mathbf{V}^\top$ with the noises $\mathbf{E}$. Finally, the response vector $\mathbf{b} \in \mathbb{R}^n$ is given by $\mathbf{b} = \mathbf{A}\mathbf{x} + \gamma\mathbf{e}$, where $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{e} \in \mathbb{R}^n$ are *i.i.d.* standard Gaussian vectors. Following Chen et al. (2015), we set $n = 500$, $d = 50,000$, $s = 50$, $\alpha = 0.05$, and $\gamma = 5$.



(a) Solution error vs. iterations (b) Objective error vs. iterations (c) Solution error vs. sketch size (d) Solution error vs. d_λ

Figure 2. Experiment results on synthetic data (errors are on log-scale).

The experiment results on synthetic data are shown in Figure 2, and are consistent with our observations regarding Figure 1. Figures 2a and 2b plot the relative error of the solution vector and the objective sub-optimality (for a fixed sketch size) as the iterative algorithm progresses. Figure 2c plots the relative error of the solution with respect to varying sketch sizes (the plots

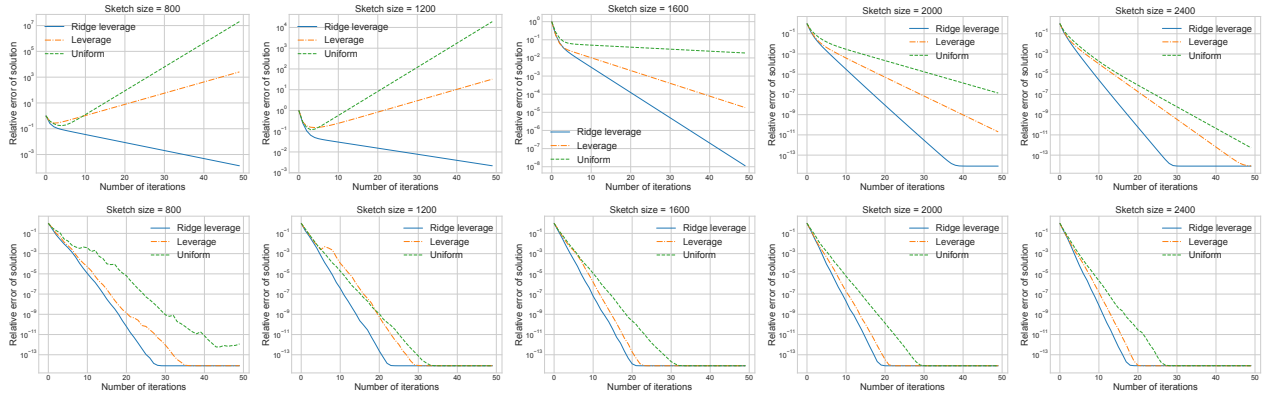
for objective sub-optimality are analogous and are thus omitted). We observe that both the solution error and the objective sub-optimality decay *exponentially* as our iterative algorithm progresses.[5]

In Figure 2d, we keep the design matrix unchanged (n remains fixed), while varying the regularization parameter $\lambda \in \{10, 20, 50, 75, 100, 150\}$, and plot the relative error of the solution against the degrees of freedom d_λ (for a fixed sketch size and number of iterations). We observe that the relative error decreases exponentially as d_λ decreases (as λ increases). Thus, the sketch size or number of iterations necessary to achieve a certain precision in the solution also decreases with d_λ, even though n remains fixed.

H.2. Additional Results on Real Data

As noted in Section 5, we conjecture that using different sampling matrices in each iteration of Algorithm 1 (*i.e.*, introducing new "randomness" in each iteration) could lead to improved bounds for our main theorems. We evaluate this conjecture empirically by comparing the performance of Algorithm 1 using either a single sampling-and-rescaling matrix $\mathbf{S}$ (the setup in the main paper) or drawing (independently) a new sampling-and-rescaling matrix at every iteration j.

Figure 3 shows the relative approximation error vs. number of iterations on the ARCENE dataset for increasing sketch sizes. We observe that using a newly sampled sketching matrix at every iteration enables faster convergence as the iterations progress, and also reduces the minimum sketch size s necessary for Algorithm 1 to converge. Also note that the minimum sketch size requirement is smaller when ridge leverage scores are used to construct $\mathbf{S}$ as compared to leverage score sampling probabilities; this confirms our discussion in Section 2.1: for ridge leverage score sampling, setting $s = \mathcal{O}(\varepsilon^{-2} d_\lambda \ln d_\lambda)$ suffices to satisfy the structural condition of eqn. (8), while for leverage scores, setting $s = \mathcal{O}(\varepsilon^{-2} n \ln n)$ suffices to satisfy the structural condition of eqn. (6) (recall that n can be substantially larger than the effective degrees of freedom d_λ).



(a) $s = 800$ (b) $s = 1200$ (c) $s = 1600$ (d) $s = 2000$ (e) $s = 2400$

Figure 3. Relative approximation error vs. number of iterations on ARCENE dataset for increasing sketch size s (errors are on log-scale). *Top row*: using a single sampling-and-rescaling matrix $\mathbf{S}$ throughout the iterations. *Bottom row*: sampling a new $\mathbf{S}_j$ at every iteration j.

[5] For these experiments, we have set the regularization parameter $\lambda = 10$ in the ridge regression objective as well as when computing the ridge leverage score sampling probabilities.