Randomized Iterative Algorithms for Fisher Discriminant Analysis (Appendix)

Appendix A PRELIMINARIES

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

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

Appendix B EVD-BASED ALGORITHMS FOR FDA

For RFDA, we quote an EVD-based algorithm along with an important result from [36] which together are the building blocks of our iterative framework. Let $\mathbf{M} \in \mathbb{R}^{c\times c}$ be the matrix such that $\mathbf{M} = \mathbf{\Omega}^{\mathsf{T}}\mathbf{A}\mathbf{G}$. Clearly, $\mathbf{M}$ is symmetric and positive semi-definite.

Algorithm 2 Algorithm for RFDA problem (3)

Input: $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{\Omega} \in \mathbb{R}^{n\times c}$ and $\lambda > 0$;
$\mathbf{G} \leftarrow (\mathbf{A}^{\mathsf{T}}\mathbf{A} + \lambda \mathbf{I}_d)^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{\Omega}$;
$\mathbf{M} \leftarrow \mathbf{\Omega}^{\mathsf{T}}\mathbf{A}\mathbf{G}$;
Compute thin SVD: $\mathbf{M} = \mathbf{V}_{\mathbf{M}}\mathbf{\Sigma}_{\mathbf{M}}\mathbf{V}_{\mathbf{M}}^{\mathsf{T}}$;
Output: $\mathbf{X} = \mathbf{G}\,\mathbf{V}_{\mathbf{M}}$

Theorem 8. *Using Algorithm 2, let $\mathbf{X}$ be the solution of problem* (3) *, then we have*

$$\mathbf{X}\mathbf{X}^{\mathsf{T}} = \mathbf{G}\,\mathbf{G}^{\mathsf{T}}.$$

For any two data points $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$, Theorem 8 implies

$$\begin{aligned} &(\mathbf{w}_1-\mathbf{w}_2)^{\mathsf{T}}\mathbf{X}\mathbf{X}^{\mathsf{T}}(\mathbf{w}_1-\mathbf{w}_2) = (\mathbf{w}_1-\mathbf{w}_2)^{\mathsf{T}}\mathbf{G}\,\mathbf{G}^{\mathsf{T}}(\mathbf{w}_1-\mathbf{w}_2) \\ \iff &\|(\mathbf{w}_1-\mathbf{w}_2)^{\mathsf{T}}\mathbf{X}\|_2 = \|(\mathbf{w}_1-\mathbf{w}_2)^{\mathsf{T}}\mathbf{G}\|_2. \end{aligned}$$

Theorem 8 indicates that if we use any distance-based classification method such as k-nearest neighbors, both $\mathbf{X}$ and $\mathbf{G}$ shares the same property. Thus, we may shift our interest from $\mathbf{X}$ to $\mathbf{G}$.

Appendix C PROOF OF THEOREM 1

Proof of Lemma 3. Using the full SVD representation of $\mathbf{A}$ we have

$$\begin{aligned} \mathbf{G}^{(j)} &= \mathbf{V}_f\mathbf{\Sigma}_f^{\mathsf{T}}\mathbf{U}_f^{\mathsf{T}}(\mathbf{U}_f\mathbf{\Sigma}_f\mathbf{\Sigma}_f^{\mathsf{T}}\mathbf{U}_f^{\mathsf{T}} + \lambda\mathbf{U}_f\mathbf{U}_f^{\mathsf{T}})^{-1}\mathbf{L}^{(j)} \\ &= \mathbf{V}_f\mathbf{\Sigma}_f^{\mathsf{T}}(\mathbf{\Sigma}_f\mathbf{\Sigma}_f^{\mathsf{T}} + \lambda\mathbf{I}_n)^{-1}\mathbf{U}_f^{\mathsf{T}}\mathbf{L}^{(j)} \\ &= \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\left[\begin{pmatrix}\mathbf{\Sigma}^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} + \lambda\mathbf{I}_n\right]^{-1}\begin{pmatrix}\mathbf{U}^{\mathsf{T}}\\ \mathbf{U}_\perp^{\mathsf{T}}\end{pmatrix}\mathbf{L}^{(j)} \\ &= \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\left[\begin{pmatrix}\mathbf{\Sigma}^2+\lambda\mathbf{I}_\rho & \mathbf{0}\\ \mathbf{0} & \lambda\mathbf{I}_{n-\rho}\end{pmatrix}\right]^{-1}\begin{pmatrix}\mathbf{U}^{\mathsf{T}}\\ \mathbf{U}_\perp^{\mathsf{T}}\end{pmatrix}\mathbf{L}^{(j)} \\ &= \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}(\mathbf{\Sigma}^2+\lambda\mathbf{I}_\rho)^{-1} & \mathbf{0}\\ \mathbf{0} & \frac{1}{\lambda}\mathbf{I}_{n-\rho}\end{pmatrix}\begin{pmatrix}\mathbf{U}^{\mathsf{T}}\\ \mathbf{U}_\perp^{\mathsf{T}}\end{pmatrix}\mathbf{L}^{(j)} \end{aligned}$$

$$
\begin{aligned}
&= \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix} \begin{pmatrix}\mathbf{\Sigma}(\mathbf{\Sigma}^2+\lambda\mathbf{I}_\rho)^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} \begin{pmatrix}\mathbf{U}^\mathsf{T}\\ \mathbf{U}_\perp^\mathsf{T}\end{pmatrix} \mathbf{L}^{(j)}\\
&= \mathbf{V}\mathbf{\Sigma}(\mathbf{\Sigma}^2+\lambda\mathbf{I}_\rho)^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\\
&= \mathbf{V}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}(\mathbf{I}_\rho+\lambda\mathbf{\Sigma}^{-2})^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\\
&= \mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}, \qquad (29)
\end{aligned}
$$

which completes the proof. □

Detailed proof of Lemma 4. First, using SVD of $\mathbf{A}$, we express $\widetilde{\mathbf{G}}^{(j)}$ in terms of $\mathbf{G}^{(j)}$.

$$
\begin{aligned}
\widetilde{\mathbf{G}}^{(j)} &= \mathbf{V}_f\mathbf{\Sigma}_f^\mathsf{T}\mathbf{U}_f^\mathsf{T}(\mathbf{U}_f\mathbf{\Sigma}_f\mathbf{V}_f^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}_f\mathbf{\Sigma}_f^\mathsf{T}\mathbf{U}_f^\mathsf{T}+\lambda\mathbf{U}_f\mathbf{U}_f^\mathsf{T})^{-1}\mathbf{L}^{(j)}\\
&= \mathbf{V}_f\mathbf{\Sigma}_f^\mathsf{T}(\mathbf{\Sigma}_f\mathbf{V}_f^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}_f\mathbf{\Sigma}_f^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{U}_f^\mathsf{T}\mathbf{L}^{(j)}\\
&= \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\left[\begin{pmatrix}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\mathbf{\Sigma} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}+\lambda\mathbf{I}_n\right]^{-1}\begin{pmatrix}\mathbf{U}^\mathsf{T}\\ \mathbf{U}_\perp^\mathsf{T}\end{pmatrix}\mathbf{L}^{(j)}\\
&= \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\left[\begin{pmatrix}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\mathbf{\Sigma}+\lambda\mathbf{I}_\rho & \mathbf{0}\\ \mathbf{0} & \lambda\mathbf{I}_{n-\rho}\end{pmatrix}\right]^{-1}\begin{pmatrix}\mathbf{U}^\mathsf{T}\\ \mathbf{U}_\perp^\mathsf{T}\end{pmatrix}\mathbf{L}^{(j)}\\
&= \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}(\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\mathbf{\Sigma}+\lambda\mathbf{I}_\rho)^{-1} & \mathbf{0}\\ \mathbf{0} & \frac{1}{\lambda}\mathbf{I}_{n-\rho}\end{pmatrix}\begin{pmatrix}\mathbf{U}^\mathsf{T}\\ \mathbf{U}_\perp^\mathsf{T}\end{pmatrix}\mathbf{L}^{(j)}\\
&= \begin{pmatrix}\mathbf{V} & \mathbf{V}_\perp\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma}(\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\mathbf{\Sigma}+\lambda\mathbf{I}_\rho)^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{U}^\mathsf{T}\\ \mathbf{U}_\perp^\mathsf{T}\end{pmatrix}\mathbf{L}^{(j)}\\
&= \mathbf{V}\mathbf{\Sigma}(\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\mathbf{\Sigma}+\lambda\mathbf{I}_\rho)^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \qquad (30)\\
&= \mathbf{V}\mathbf{\Sigma}\left(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}\left(\mathbf{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\mathbf{\Sigma}_\lambda\right)\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma}+\lambda\mathbf{I}_\rho\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(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}_\rho\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \qquad (31)\\
&= \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{L}^{(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{L}^{(j)}\\
&= \mathbf{V}\mathbf{\Sigma}\left(\mathbf{\Sigma}\mathbf{\Sigma}_\lambda^{-1}\left(\mathbf{I}_\rho+\mathbf{E}\right)\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma}\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}. \qquad (32)
\end{aligned}
$$

Eqn. (31) used the fact that $\mathbf{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\mathbf{\Sigma}_\lambda = \mathbf{\Sigma}_\lambda^2+\mathbf{E}$. Eqn. (32) 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 $i = 1\ldots\rho$. Thus, we have $\left(\mathbf{\Sigma}_\lambda^2+\lambda\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-2}\mathbf{\Sigma}_\lambda\right) = \mathbf{I}_\rho$. Since $\|\mathbf{E}\|_2 < 1$, Proposition 7 implies that $(\mathbf{I}_\rho+\mathbf{E})^{-1}$ exists and

$$
(\mathbf{I}_\rho+\mathbf{E})^{-1} = \mathbf{I}_\rho+\sum_{\ell=1}^{\infty}(-1)^\ell\mathbf{E}^\ell = \mathbf{I}_\rho+\mathbf{Q}.
$$

Thus, eqn. (32) can further be expressed as

$$
\begin{aligned}
\widetilde{\mathbf{G}}^{(j)} &= \mathbf{V}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_\lambda\left(\mathbf{I}_\rho+\mathbf{E}\right)^{-1}\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\\
&= \mathbf{V}\mathbf{\Sigma}_\lambda\left(\mathbf{I}_\rho+\mathbf{Q}\right)\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\\
&= \mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}+\mathbf{V}\mathbf{\Sigma}_\lambda\mathbf{Q}\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\\
&= \mathbf{G}^{(j)}+\mathbf{V}\mathbf{\Sigma}_\lambda\mathbf{Q}\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}, \qquad (33)
\end{aligned}
$$

where the last line follows from Lemma 3. Further, we have

$$
\|\mathbf{Q}\|_2 = \|\sum_{\ell=1}^{\infty}(-1)^\ell\mathbf{E}^\ell\|_2 \le \sum_{\ell=1}^{\infty}\|\mathbf{E}^\ell\|_2 \le \sum_{\ell=1}^{\infty}\|\mathbf{E}\|_2^\ell \le \sum_{\ell=1}^{\infty}\left(\frac{\varepsilon}{2}\right)^\ell = \frac{\varepsilon/2}{1-\varepsilon/2} \le \varepsilon\,, \qquad (34)
$$

where we used the triangle inequality, the sub-multiplicativity of the spectral norm, and the fact that $\varepsilon \leq 1$. Next, we combine eqns. (33) and (34) to get

$$\begin{aligned}\|(\mathbf{w}-\mathbf{m})^\mathsf{T}(\widetilde{\mathbf{G}}^{(j)}-\mathbf{G}^{(j)})\|_2 &= \|(\mathbf{w}-\mathbf{m})^\mathsf{T}\mathbf{V}\boldsymbol{\Sigma}_\lambda\mathbf{Q}\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2\\ &\leq \|(\mathbf{w}-\mathbf{m})^\mathsf{T}\mathbf{V}\|_2\|\boldsymbol{\Sigma}_\lambda\|_2\|\mathbf{Q}\|_2\|\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2\\ &\leq \varepsilon\,\|(\mathbf{w}-\mathbf{m})^\mathsf{T}\mathbf{V}\|_2\|\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2\\ &= \varepsilon\,\|\mathbf{V}\mathbf{V}^\mathsf{T}(\mathbf{w}-\mathbf{m})\|_2\|\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2,\end{aligned} \tag{35}$$

which completes the proof. □

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

Detailed proof of Lemma 6. From Algorithm 1, we have for $j = 1 \dots t-1$

$$\begin{aligned}\mathbf{L}^{(j+1)} &= \mathbf{L}^{(j)} - \lambda\mathbf{Y}^{(j)} - \mathbf{A}\widetilde{\mathbf{G}}^{(j)}\\ &= \mathbf{L}^{(j)} - (\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{L}^{(j)}.\end{aligned} \tag{36}$$

Now, starting with the full SVD of $\mathbf{A}$, we get

$$\begin{aligned}&(\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{L}^{(j)}\\
&= \left(\mathbf{U}_f\boldsymbol{\Sigma}_f\boldsymbol{\Sigma}_f^\mathsf{T}\mathbf{U}_f^\mathsf{T}+\lambda\mathbf{U}_f\mathbf{U}_f^\mathsf{T}\right)\left(\mathbf{U}_f\boldsymbol{\Sigma}_f\mathbf{V}_f^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}_f\boldsymbol{\Sigma}_f^\mathsf{T}\mathbf{U}_f^\mathsf{T}+\lambda\mathbf{U}_f\mathbf{U}_f^\mathsf{T}\right)^{-1}\mathbf{L}^{(j)}\\
&= \mathbf{U}_f\left(\boldsymbol{\Sigma}_f\boldsymbol{\Sigma}_f^\mathsf{T}+\lambda\mathbf{I}_n\right)\mathbf{U}_f^\mathsf{T}\mathbf{U}_f\left(\boldsymbol{\Sigma}_f\mathbf{V}_f^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}_f\boldsymbol{\Sigma}_f^\mathsf{T}+\lambda\mathbf{I}_n\right)^{-1}\mathbf{U}_f^\mathsf{T}\mathbf{L}^{(j)}\\
&= \mathbf{U}_f\left(\boldsymbol{\Sigma}_f\boldsymbol{\Sigma}_f^\mathsf{T}+\lambda\mathbf{I}_n\right)\left(\boldsymbol{\Sigma}_f\mathbf{V}_f^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}_f\boldsymbol{\Sigma}_f^\mathsf{T}+\lambda\mathbf{I}_n\right)^{-1}\mathbf{U}_f^\mathsf{T}\mathbf{L}^{(j)}\\
&= \mathbf{U}_f\begin{pmatrix}\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho & \mathbf{0}\\ \mathbf{0} & \lambda\mathbf{I}_{n-\rho}\end{pmatrix}\begin{pmatrix}(\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\boldsymbol{\Sigma}+\lambda\mathbf{I}_\rho)^{-1} & \mathbf{0}\\ \mathbf{0} & \frac{1}{\lambda}\mathbf{I}_{n-\rho}\end{pmatrix}\mathbf{U}_f^\mathsf{T}\mathbf{L}^{(j)}\\
&= \mathbf{U}_f\begin{pmatrix}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)(\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\boldsymbol{\Sigma}+\lambda\mathbf{I}_\rho)^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{n-\rho}\end{pmatrix}\mathbf{U}_f^\mathsf{T}\mathbf{L}^{(j)}\\
&= \begin{pmatrix}\mathbf{U} & \mathbf{U}_\perp\end{pmatrix}\begin{pmatrix}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)(\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\boldsymbol{\Sigma}+\lambda\mathbf{I}_\rho)^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{n-\rho}\end{pmatrix}\begin{pmatrix}\mathbf{U}^\mathsf{T}\\ \mathbf{U}_\perp^\mathsf{T}\end{pmatrix}\mathbf{L}^{(j)}\\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)}+\mathbf{U}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)(\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\boldsymbol{\Sigma}+\lambda\mathbf{I}_\rho)^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \qquad (37)\\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)}+\mathbf{U}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\lambda^{-1}\left(\boldsymbol{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\boldsymbol{\Sigma}_\lambda\right)\boldsymbol{\Sigma}_\lambda^{-1}\boldsymbol{\Sigma}+\lambda\mathbf{I}_\rho\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)}+\mathbf{U}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\lambda^{-1}\left(\boldsymbol{\Sigma}_\lambda^2+\mathbf{E}\right)\boldsymbol{\Sigma}_\lambda^{-1}\boldsymbol{\Sigma}+\lambda\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\lambda^{-1}\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-2}\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}_\lambda^{-1}\boldsymbol{\Sigma}\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)}+\mathbf{U}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\lambda^{-1}\left(\boldsymbol{\Sigma}_\lambda^2+\mathbf{E}+\lambda\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-2}\boldsymbol{\Sigma}_\lambda\right)\boldsymbol{\Sigma}_\lambda^{-1}\boldsymbol{\Sigma}\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)}+\mathbf{U}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\lambda^{-1}\left(\mathbf{I}_\rho+\mathbf{E}\right)\boldsymbol{\Sigma}_\lambda^{-1}\boldsymbol{\Sigma}\right)^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}. \qquad (38)\end{aligned}$$

Here, eqn. (38) holds because $\boldsymbol{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\boldsymbol{\Sigma}_\lambda = \boldsymbol{\Sigma}_\lambda^2+\mathbf{E}$ and the fact that $\boldsymbol{\Sigma}_\lambda^2+\lambda\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-2}\boldsymbol{\Sigma}_\lambda \in \mathbb{R}^{n\times n}$ is a diagonal matrix whose ith diagonal element satisfies

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

for any $i = 1\dots\rho$. Thus, we have $\left(\boldsymbol{\Sigma}_\lambda^2+\lambda\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-2}\boldsymbol{\Sigma}_\lambda\right) = \mathbf{I}_\rho$. Since $\|\mathbf{E}\|_2 < 1$, Proposition 7 implies that $(\mathbf{I}_\rho+\mathbf{E})^{-1}$ exists and

$$(\mathbf{I}_\rho+\mathbf{E})^{-1} = \mathbf{I}_\rho+\sum_{\ell=1}^{\infty}(-1)^\ell\mathbf{E}^\ell = \mathbf{I}_\rho+\mathbf{Q},$$

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

Thus, we rewrite eqn. (38) as

$$(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{L}^{(j)}$$

$$
\begin{aligned}
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_\lambda\left(\mathbf{I}_\rho+\mathbf{E}\right)^{-1}\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_\lambda\left(\mathbf{I}_\rho+\mathbf{Q}\right)\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_\lambda^2\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_\lambda\mathbf{Q}\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_\lambda\mathbf{Q}\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \qquad (39)\\
&= (\mathbf{U}\mathbf{U}^\mathsf{T}+\mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T})\mathbf{L}^{(j)} + \mathbf{U}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_\lambda\mathbf{Q}\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \\
&= \mathbf{U}_f\mathbf{U}_f^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_\lambda\mathbf{Q}\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}. \qquad (40)
\end{aligned}
$$

Eqn. (39) holds as $(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_\lambda^2\boldsymbol{\Sigma}^{-1} = \mathbf{I}_\rho$. Further, using the fact that $\mathbf{U}_f\mathbf{U}_f^\mathsf{T} = \mathbf{I}_n$, we rewrite eqn. (40) as

$$
(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)^{-1}\mathbf{L}^{(j)} = \mathbf{L}^{(j)} + \mathbf{U}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_\lambda\mathbf{Q}\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}. \qquad (41)
$$

Thus, combining eqns. (36) and (41), we have

$$
\mathbf{L}^{(j+1)} = \; -\mathbf{U}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_\lambda\mathbf{Q}\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}. \qquad (42)
$$

Finally, using eqn. (42), we obtain

$$
\begin{aligned}
\|\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j+1)}\|_2 &= \|\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{U}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_\lambda\mathbf{Q}\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2 \\
&= \|\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_\lambda\mathbf{Q}\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2 \\
&= \|\mathbf{Q}\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2 \le \|\mathbf{Q}\|_2\|\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2 \\
&\le \varepsilon\,\|\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2\,,
\end{aligned}
$$

where the third equality holds as $\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^2+\lambda\mathbf{I}_\rho)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_\lambda = \mathbf{I}_\rho$ and the last two steps follow from sub-multiplicativity and eqn. (34) respectively. This concludes the proof. □

Proof of Theorem 1. Applying Lemma 6 iteratively, we get

$$
\|\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(t)}\|_2 \le \varepsilon\,\|\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(t-1)}\|_2 \le \ldots \le \varepsilon^{t-1}\|\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(1)}\|_2. \qquad (43)
$$

Now, from eqn (43), we apply sub-multiplicativity to obtain

$$
\|\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(1)}\|_2 = \|\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\boldsymbol{\Omega}\|_2 \le \|\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\|_2\|\mathbf{U}^\mathsf{T}\|_2\|\boldsymbol{\Omega}\|_2 = \max_{1\le i\le\rho}(\sigma_i^2+\lambda)^{-\frac{1}{2}} \le \lambda^{-\frac{1}{2}}, \qquad (44)
$$

Notice that $\mathbf{L}^{(1)} = \boldsymbol{\Omega}$ by definition. Also, $\boldsymbol{\Omega}^\mathsf{T}\boldsymbol{\Omega} = \mathbf{I}_c$ and thus $\|\boldsymbol{\Omega}\|_2 = 1$. Furthermore, we know that $\|\mathbf{U}^\mathsf{T}\|_2 = 1$ and $\|\boldsymbol{\Sigma}_\lambda\boldsymbol{\Sigma}^{-1}\|_2 = \max_{1\le i\le\rho}(\sigma_i^2+\lambda)^{-\frac{1}{2}}$ and the last inequality holds since $(\sigma_i^2+\lambda)^{-\frac{1}{2}} \le \lambda^{-\frac{1}{2}}$ for all $i = 1\ldots\rho$.

Finally, combining eqns. (22), (43) and (44), we conclude

$$
\|(\mathbf{w}-\mathbf{m})^\mathsf{T}(\widehat{\mathbf{G}}-\mathbf{G})\|_2 \le \frac{\varepsilon^t}{\sqrt{\lambda}}\,\|\mathbf{V}\mathbf{V}^\mathsf{T}(\mathbf{w}-\mathbf{m})\|_2,
$$

which completes the proof. □

Appendix D PROOF OF THEOREM 2

Lemma 9. *For $j = 1\ldots t$, let $\mathbf{L}^{(j)}$ and $\widetilde{\mathbf{G}}^{(j)}$ be the intermediate matrices in Algorithm 1, $\mathbf{G}^{(j)}$ be the matrix defined in eqn.* (12) *and $\mathbf{R}$ be defined as in Lemma 3. Further, let $\mathbf{S} \in \mathbb{R}^{d\times s}$ be the sketching matrix and define $\widehat{\mathbf{E}} = \mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V} - \mathbf{I}_\rho$. If eqn.* (8) *is satisfied, i.e., $\|\widehat{\mathbf{E}}\|_2 \le \frac{\varepsilon}{2}$, then for all $j = 1,\ldots,t$, we have*

$$
\|(\mathbf{w}-\mathbf{m})^\mathsf{T}(\widetilde{\mathbf{G}}^{(j)}-\mathbf{G}^{(j)})\|_2 \le \varepsilon\,\|\mathbf{V}\mathbf{V}^\mathsf{T}(\mathbf{w}-\mathbf{m})\|_2\,\|\mathbf{R}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2\,, \qquad (45)
$$

where $\mathbf{R} = \mathbf{I}_\rho + \lambda\boldsymbol{\Sigma}^{-2}$.

Proof. Note that $\boldsymbol{\Sigma}_\lambda^2 = \mathbf{R}^{-1}$. Applying Lemma 3, we can express $\mathbf{G}^{(j)}$ as

$$\mathbf{G}^{(j)} = \mathbf{V}\mathbf{R}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}. \tag{46}$$

Next, rewriting eqn. (30) gives

$$\begin{aligned}
\widetilde{\mathbf{G}}^{(j)} &= \mathbf{V}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\boldsymbol{\Sigma} + \lambda\mathbf{I}_\rho)^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} && (47)\\
&= \mathbf{V}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}(\mathbf{I}_\rho + \widehat{\mathbf{E}})\boldsymbol{\Sigma} + \lambda\mathbf{I}_\rho)^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} = \mathbf{V}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}(\mathbf{I}_\rho + \widehat{\mathbf{E}} + \lambda\boldsymbol{\Sigma}^{-2})^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\\
&= \mathbf{V}(\mathbf{R} + \widehat{\mathbf{E}})^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} = \mathbf{V}(\mathbf{R}(\mathbf{I}_\rho + \mathbf{R}^{-1}\widehat{\mathbf{E}}))^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\,. && (48)
\end{aligned}$$

Further, notice that

$$\|\mathbf{R}^{-1}\widehat{\mathbf{E}}\|_2 \le \|\mathbf{R}^{-1}\|_2\|\widehat{\mathbf{E}}\|_2 \le \|\mathbf{R}^{-1}\|_2 \cdot \frac{\varepsilon}{2} = \left(\frac{\sigma_1^2}{\sigma_1^2 + \lambda}\right)\frac{\varepsilon}{2} \le \frac{\varepsilon}{2} < 1. \tag{49}$$

Now, Proposition 7 implies that $(\mathbf{I}_\rho + \mathbf{R}^{-1}\widehat{\mathbf{E}})^{-1}$ exists. Let $\widehat{\mathbf{Q}} = \sum_{\ell=1}^{\infty}(-1)^\ell(\mathbf{R}^{-1}\widehat{\mathbf{E}})^\ell$, we have

$$(\mathbf{I}_\rho + \mathbf{R}^{-1}\widehat{\mathbf{E}})^{-1} = \mathbf{I}_\rho + \sum_{\ell=1}^{\infty}(-1)^\ell(\mathbf{R}^{-1}\widehat{\mathbf{E}})^\ell = \mathbf{I}_\rho + \widehat{\mathbf{Q}}.$$

Thus, we can rewrite eqn. (48) as

$$\begin{aligned}
\widetilde{\mathbf{G}}^{(j)} &= \mathbf{V}(\mathbf{I}_\rho + \widehat{\mathbf{Q}})\mathbf{R}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\\
&= \mathbf{V}\mathbf{R}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{V}\widehat{\mathbf{Q}}\mathbf{R}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\\
&= \mathbf{G}^{(j)} + \mathbf{V}\widehat{\mathbf{Q}}\mathbf{R}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}, && (50)
\end{aligned}$$

where eqn. (50) follows eqn. (46). Further, using eqn. (49), we have

$$\|\widehat{\mathbf{Q}}\|_2 = \|\sum_{\ell=1}^{\infty}(-1)^\ell(\mathbf{R}^{-1}\widehat{\mathbf{E}})^\ell\|_2 \le \sum_{\ell=1}^{\infty}\|(\mathbf{R}^{-1}\widehat{\mathbf{E}})^\ell\|_2 \le \sum_{\ell=1}^{\infty}\|\mathbf{R}^{-1}\widehat{\mathbf{E}}\|_2^\ell \le \sum_{\ell=1}^{\infty}\left(\frac{\varepsilon}{2}\right)^\ell = \frac{\varepsilon/2}{1-\varepsilon/2} \le \varepsilon, \tag{51}$$

where we used the triangle inequality, sub-multiplicativity of the spectral norm, and the fact that $\varepsilon \le 1$. Next, we combine eqns. (50) and (51) to get

$$\begin{aligned}
\|(\mathbf{w}-\mathbf{m})^\mathsf{T}(\widetilde{\mathbf{G}}^{(j)} - \mathbf{G}^{(j)})\|_2 &= \|(\mathbf{w}-\mathbf{m})^\mathsf{T}\mathbf{V}\widehat{\mathbf{Q}}\mathbf{R}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2\\
&\le \|(\mathbf{w}-\mathbf{m})^\mathsf{T}\mathbf{V}\|_2\|\widehat{\mathbf{Q}}\|_2\|\mathbf{R}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2\\
&\le \varepsilon\,\|(\mathbf{w}-\mathbf{m})^\mathsf{T}\mathbf{V}\|_2\|\mathbf{R}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2\\
&= \varepsilon\,\|(\mathbf{w}-\mathbf{m})^\mathsf{T}\mathbf{V}\mathbf{V}^\mathsf{T}\|_2\|\mathbf{R}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2\\
&= \varepsilon\,\|\mathbf{V}\mathbf{V}^\mathsf{T}(\mathbf{w}-\mathbf{m})\|_2\|\mathbf{R}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2, && (52)
\end{aligned}$$

where the first inequality follows from sub-multiplicativity and the second last equality holds due to the unitary invariance of the spectral norm. This concludes the proof. □

Remark 10. *Repeated application of Lemmas 5 and 9 yields:*

$$\begin{aligned}
\|(\mathbf{w}-\mathbf{m})^\mathsf{T}(\widehat{\mathbf{G}} - \mathbf{G})\|_2 &= \|(\mathbf{w}-\mathbf{m})^\mathsf{T}\big(\sum_{j=1}^{t}\widetilde{\mathbf{G}}^{(j)} - \mathbf{G}\big)\|_2 = \|(\mathbf{w}-\mathbf{m})^\mathsf{T}(\widetilde{\mathbf{G}}^{(t)} - \big(\mathbf{G} - \sum_{j=1}^{t-1}\widetilde{\mathbf{G}}^{(j)}\big))\|_2\\
&= \|(\mathbf{w}-\mathbf{m})^\mathsf{T}(\widetilde{\mathbf{G}}^{(t)} - \mathbf{G}^{(t)})\|_2 \le \varepsilon\,\|\mathbf{V}\mathbf{V}^\mathsf{T}(\mathbf{w}-\mathbf{m})\|_2\,\|\mathbf{R}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(t)}\|_2. && (53)
\end{aligned}$$

The next bound provides a critical inequality that can be used recursively in order to establish Theorem 2.

Lemma 11. *Let $\mathbf{L}^{(j)}$, $j = 1,\dots,t$, be the matrices of Algorithm 1 and $\mathbf{R}$ is as defined in Lemma 3. For any $j = 1,\dots,t-1$, define $\widehat{\mathbf{E}} = \mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V} - \mathbf{I}_\rho$. If eqn.* (8) *is satisfied* i.e.$\|\widehat{\mathbf{E}}\|_2 \le \frac{\varepsilon}{2}$, *then*

$$\|\mathbf{R}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j+1)}\|_2 \le \varepsilon\,\|\mathbf{R}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2. \tag{54}$$

Proof. From Algorithm 1, we have for $j = 1, \ldots, t-1$,

$$\mathbf{L}^{(j+1)} = \mathbf{L}^{(j)} - \lambda\mathbf{Y}^{(j)} - \mathbf{A}\widetilde{\mathbf{G}}^{(j)} = \mathbf{L}^{(j)} - (\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{L}^{(j)}. \tag{55}$$

Rewriting eqn. (37), we have

$$\begin{aligned}
&(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{L}^{(j)} \\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)(\mathbf{\Sigma}\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\mathbf{\Sigma} + \lambda\mathbf{I}_\rho)^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)(\mathbf{\Sigma}(\mathbf{I}_\rho + \widehat{\mathbf{E}})\mathbf{\Sigma} + \lambda\mathbf{I}_\rho)^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)\mathbf{\Sigma}^{-1}(\mathbf{I}_\rho + \widehat{\mathbf{E}} + \lambda\mathbf{\Sigma}^{-2})^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}.
\end{aligned} \tag{56}$$

Here, eqn. (56) holds because $(\mathbf{I}_\rho + \widehat{\mathbf{E}} + \lambda\mathbf{\Sigma}^{-2})$ is invertible since it is a positive definite matrix. In addition, using the fact that $\mathbf{R} = (\mathbf{I}_\rho + \lambda\mathbf{\Sigma}^{-2})$, we rewrite eqn. (56) as

$$\begin{aligned}
&(\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{L}^{(j)} \\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)\mathbf{\Sigma}^{-1}(\mathbf{R} + \widehat{\mathbf{E}})^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)\mathbf{\Sigma}^{-1}\left(\mathbf{R}(\mathbf{I}_\rho + \mathbf{R}^{-1}\widehat{\mathbf{E}})\right)^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)\mathbf{\Sigma}^{-1}(\mathbf{I}_\rho + \mathbf{R}^{-1}\widehat{\mathbf{E}})^{-1}\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)\mathbf{\Sigma}^{-1}(\mathbf{I}_\rho + \widehat{\mathbf{Q}})\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \\
&= \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)\mathbf{\Sigma}^{-1}\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)\mathbf{\Sigma}^{-1}\widehat{\mathbf{Q}}\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \\
&= (\mathbf{U}\mathbf{U}^\mathsf{T} + \mathbf{U}_\perp\mathbf{U}_\perp^\mathsf{T})\mathbf{L}^{(j)} + \mathbf{U}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)\mathbf{\Sigma}^{-1}\widehat{\mathbf{Q}}\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)} \\
&= \mathbf{U}_f\mathbf{U}_f^\mathsf{T}\mathbf{L}^{(j)} + \mathbf{U}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)\mathbf{\Sigma}^{-1}\widehat{\mathbf{Q}}\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}.
\end{aligned} \tag{57}$$

The second and third equalities follow from Proposition 7 (using eqn. (49)) and the fact that $\mathbf{R}^{-1}$ exists. Further, $\widehat{\mathbf{Q}}$ is as defined as in Lemma 9. Moreover, the second last equality holds as $(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)\mathbf{\Sigma}^{-1}\mathbf{R}^{-1}\mathbf{\Sigma}^{-1} = \mathbf{I}_\rho$. Now, using the fact that $\mathbf{U}_f\mathbf{U}_f^\mathsf{T} = \mathbf{I}_n$, we rewrite eqn. (57) as

$$(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{L}^{(j)} = \mathbf{L}^{(j)} + \mathbf{U}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)\mathbf{\Sigma}^{-1}\widehat{\mathbf{Q}}\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}. \tag{58}$$

Thus, combining, eqns. (55) and (58), we have

$$\mathbf{L}^{(j+1)} = \; -\mathbf{U}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)\mathbf{\Sigma}^{-1}\widehat{\mathbf{Q}}\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}. \tag{59}$$

Finally, from eqn. (59), we obtain

$$\begin{aligned}
\|\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j+1)}\|_2 &= \|\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{U}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)\mathbf{\Sigma}^{-1}\widehat{\mathbf{Q}}\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2 \\
&= \|\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)\mathbf{\Sigma}^{-1}\widehat{\mathbf{Q}}\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2 \\
&= \|\widehat{\mathbf{Q}}\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2 \le \|\widehat{\mathbf{Q}}\|_2\|\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2 \\
&\le \varepsilon\,\|\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(j)}\|_2,
\end{aligned} \tag{60}$$

where the third equality holds as $\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_\rho)\mathbf{\Sigma}^{-1} = \mathbf{I}_\rho$ and the last two steps follow from sub-multiplicativity and eqn. (51) respectively. This concludes the proof. □

Proof of Theorem 2. Applying Lemma 11 iteratively, we have

$$\|\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(t)}\|_2 \le \varepsilon\,\|\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(t-1)}\|_2 \le \ldots \le \varepsilon^{t-1}\,\|\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(1)}\|_2\,. \tag{61}$$

Now, from eqn (61) and noticing that $\mathbf{L}^{(1)} = \mathbf{\Omega}$ by definition, we have

$$\|\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{L}^{(1)}\|_2 \le \|\mathbf{R}^{-1}\mathbf{\Sigma}^{-1}\|_2\|\mathbf{U}^\mathsf{T}\|_2\|\mathbf{\Omega}\|_2 = \max_{1\le i\le\rho}\left\{\frac{\sigma_i}{\sigma_i^2 + \lambda}\right\} \le \frac{1}{2\sqrt{\lambda}}\,, \tag{62}$$

where we used sub-multiplicativity and the facts that $\|\mathbf{U}^\mathsf{T}\|_2 = 1$, $\mathbf{\Omega}^\mathsf{T}\mathbf{\Omega} = \mathbf{I}_c$, and $\|\mathbf{\Omega}\|_2 = 1$. The last step in eqn. (62) holds since for all $i = 1 \ldots \rho$,

$$(\sigma_i - \sqrt{\lambda})^2 \geq 0 \quad \Rightarrow \quad \sigma_i^2 + \lambda \geq 2\sigma_i\sqrt{\lambda} \quad \Rightarrow \quad \frac{\sigma_i}{\sigma_i^2 + \lambda} \leq \frac{1}{2\sqrt{\lambda}}\,. \tag{63}$$

Finally, combining eqns. (53), (61) and (62), we obtain

$$\|(\mathbf{w} - \mathbf{m})^\mathsf{T}(\widehat{\mathbf{G}} - \mathbf{G})\|_2 \leq \frac{\varepsilon^t}{2\sqrt{\lambda}}\,\|\mathbf{V}\mathbf{V}^\mathsf{T}(\mathbf{w} - \mathbf{m})\|_2\,,$$

which concludes the proof. □

Appendix E SAMPLING-BASED CONSTRUCTIONS

We now discuss how to satisfy the conditions of eqns. (5) or (8) by *sampling*, *i.e.*, selecting a small number of features. Towards that end, consider Algorithm 3 for the construction of the sampling-and-rescaling matrix $\mathbf{S}$. Finally, the next result appeared in [6] as Theorem 3 and is a strengthening of Theorem 4.2 of [20], since the sampling complexity s is improved to depend only on $\|\mathbf{Z}\|_F^2$ instead of the stable rank of $\mathbf{Z}$ when $\|\mathbf{Z}\|_2 \leq 1$. We also note that Lemma 12 is implicit in [8] .

Algorithm 3 Sampling-and-rescaling matrix

Input: Sampling probabilities p_i, $i = 1, \ldots, d$;
number of sampled columns $s \ll d$;
$\mathbf{S} \leftarrow \mathbf{0}_{d\times s}$;
for $t = 1$ **to** s **do**
Pick $i_t \in \{1, \ldots, d\}$ with $\mathbb{P}(i_t = i) = p_i$;
$\mathbf{S}_{i_t t} = 1/\sqrt{s\, p_{i_t}}$;
end for
Output: Return $\mathbf{S}$;

Lemma 12. *Let $\mathbf{Z} \in \mathbb{R}^{d\times n}$ with $\|\mathbf{Z}\|_2 \leq 1$ and let $\mathbf{S}$ be constructed by Algorithm 3 with*

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

then, with probability at least $1 - \delta$,

$$\|\mathbf{Z}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{Z} - \mathbf{Z}^\mathsf{T}\mathbf{Z}\|_2 \leq \varepsilon.$$

Applying Lemma 12 with $\mathbf{Z} = \mathbf{V}\mathbf{\Sigma}_\lambda$, we can satisfy the condition of eqn. (5) using the sampling probabilities $p_i = \|(\mathbf{V}\mathbf{\Sigma}_\lambda)_{i*}\|_2^2/d_\lambda$ (recall that $\|\mathbf{V}\mathbf{\Sigma}_\lambda\|_F^2 = d_\lambda$ and $\|\mathbf{V}\mathbf{\Sigma}_\lambda\|_2 \leq 1$). It is easy to see that these probabilities are exactly proportional to the column ridge leverage scores of the design matrix $\mathbf{A}$. Setting $s = \mathcal{O}(\varepsilon^{-2}d_\lambda \ln d_\lambda)$ suffices to satisfy the condition of eqn. (5). We note that approximate ridge leverage scores also suffice and that their computation can be done efficiently without computing $\mathbf{V}$ [8]. Finally, applying Lemma 12 with $\mathbf{Z} = \mathbf{V}$ we can satisfy the condition of eqn. (8) by simply using the sampling probabilities $p_i = \|\mathbf{V}_{i*}\|_2^2/\rho$ (recall that $\|\mathbf{V}\|_F^2 = \rho$ and $\|\mathbf{V}\|_2 = 1$), which correspond to the column leverage scores of the design matrix $\mathbf{A}$. Setting $s = \mathcal{O}(\varepsilon^{-2}\rho \ln \rho)$ suffices to satisfy the condition of eqn. (8). We note that approximate leverage scores also suffice and that their computation can be done efficiently without computing $\mathbf{V}$ [13].

Appendix F SKETCH-SIZE REQUIREMENTS FOR STRUCTURAL CONDITIONS

We provide details on the sketch-size requirements for satisfying the structual conditions of eqns. (5) or (8) when various constructions of the sketching matrix $\mathbf{S}$ are used. It was shown in [9] that eqn. (11) can be achieved using a count-sketch matrix $\mathbf{S}$ with $s = \mathcal{O}(\frac{r}{\delta\varepsilon^2})$ columns or an SRHT matrix $\mathbf{S}$ with $s = \mathcal{O}(\varepsilon^{-2}(r + \log(1/\varepsilon\delta)) \log \frac{r}{\delta})$ columns (here, δ is the failure probability). As discussed in Section 2.2, setting $r = d_\lambda$ or $r = \rho$ in eqn. (11) for eqns. (5) or (8), respectively, we obtain the sketch-size requirements summarized in Table 1.

Appendix G ADDITIONAL EXPERIMENT RESULTS

Table 2 shows the CPU wall-clock times for running RFDA (on a single-core Intel Xeon E5-2660 CPU at 2.6GHz) by either computing $\mathbf{G}$ exactly in eqn. (3) or via our iterative algorithm. For both datasets, we report the per-iteration runtime of our algorithm with various sketching-matrix constructions using a sketch size of $s = 5,000$.

	Count-sketch	SRHT	Sampling (Appendix E)
Eqn. (5)	$s = \mathcal{O}\left(\frac{d_\lambda}{\delta\varepsilon^2}\right)$	$s = \mathcal{O}\left(\frac{d_\lambda + \log(1/\varepsilon\delta)}{\varepsilon^2} \log \frac{d_\lambda}{\delta}\right)$	$s = \mathcal{O}\left(\frac{d_\lambda \log(d_\lambda/\delta)}{\varepsilon^2}\right)$
Eqn. (8)	$s = \mathcal{O}\left(\frac{\rho}{\delta\varepsilon^2}\right)$	$s = \mathcal{O}\left(\frac{\rho + \log(1/\varepsilon\delta)}{\varepsilon^2} \log \frac{\rho}{\delta}\right)$	$s = \mathcal{O}\left(\frac{\rho \log(\rho/\delta)}{\varepsilon^2}\right)$

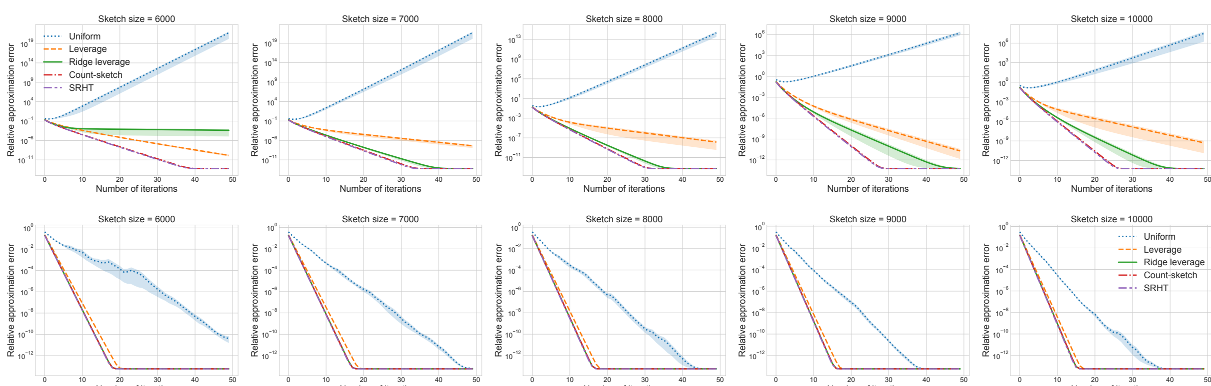
Table 1: Sketch-size requirements for satisfying eqns. (5) or (8) with probability at least $1-\delta$.

Dataset	SVD	Exact	*Uniform*	*Leverage*	*Ridge leverage*	*Count-sketch*
ORL	1.335	0.232	0.101	0.101	0.101	0.103
PEMS	35.781	3.770	0.917	0.892	0.899	0.970

Table 2: CPU wall-clock times (in seconds) for RFDA on ORL and PEMS.

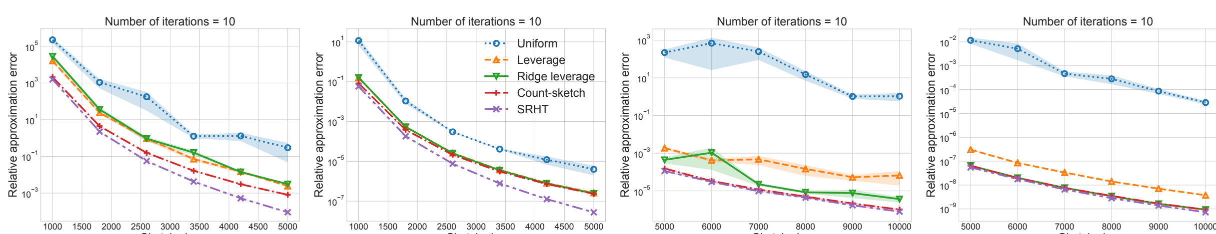
As noted in Section 5, we conjecture that using independent sampling matrices in each iteration of Algorithm 1 (*i.e.*, introducing new "randomness" in each iteration) could lead to improved bounds for our main theorems. We evaluate this conjecture empirically by comparing the performance of Algorithm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

Figure 3 shows the relative approximation error vs. number of iterations on the PEMS dataset for increasing sketch sizes. Figure 4 plots the relative approximation error vs. sketch size after 10 iterations of Algorithm 1 were run. We observe that using a newly sampled sketching matrix at every iteration enables faster convergence as the iterations progress, and also reduces the sketch size s necessary for Algorithm 1 to converge.



(a) $s = 6000$ (b) $s = 7000$ (c) $s = 8000$ (d) $s = 9000$ (e) $s = 10000$

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



(a) ORL; single $\mathbf{S}$ (b) ORL; multiple $\mathbf{S}_j$ (c) PEMS; single $\mathbf{S}$ (d) PEMS; multiple $\mathbf{S}_j$

Figure 4: Relative approximation error vs. sketch size on ORL and PEMS after 10 iterations. *Single* $\mathbf{S}$: using a single sketching matrix $\mathbf{S}$ throughout the iterations. *Multiple* $\mathbf{S}_j$: sample a new $\mathbf{S}_j$ at every iteration j. Errors are on log-scale; note the difference in magnitude of the approximation errors across plots.