Supplementary Material for "Fast Relative-Error Approximation Algorithm for Ridge Regression"

A Proof of Theorem 3

For convenience, we restate the theorem in the following.

Theorem 3. *Given $\epsilon \in (0,1)$, $\delta \in (0,1)$ and $r > 0$, select integers $t' \geq 2\delta^{-1}(r^2+r)/(2\epsilon/3-\epsilon^2/9)^2$ and $t \geq 18\epsilon^{-1}[\sqrt{r}+\sqrt{8\log(6p/\delta)}]^2\log(6r/\delta)$. Let $\mathbf{\Phi}_{sparse} \in \mathbb{R}^{t'\times p}$ be a $t' \times p$ sparse embedding matrix and let $\mathbf{\Phi}_{srht} \in \mathbb{R}^{t\times t'}$ be a $t \times t'$ SRHT matrix. Then the product $\mathbf{S} = \mathbf{\Phi}_{srht}\mathbf{\Phi}_{sparse}$ is an (r,δ,ϵ)-OSE.*

Proof. By Theorem 1, $\mathbf{\Phi}_{\text{sparse}} \in \mathbb{R}^{t'\times p}$ is an $(r,\delta/2,\epsilon/3)$-OSE. Similarly, by Theorem 2, $\mathbf{\Phi}_{\text{srht}} \in \mathbb{R}^{t'\times t}$ is also an $(r,\delta/2,\epsilon/3)$-OSE.

Now fix an arbitrary matrix $\mathbf{M} \in \mathbb{R}^{p\times m}$ of rank r. Since $\mathbf{\Phi}_{\text{sparse}}$ is an $(r,\delta/2,\epsilon/3)$-OSE, by Definition 1, we have

$$(1-\epsilon/3)\left\|\mathbf{M}\mathbf{z}\right\|_2 \leq \left\|\mathbf{\Phi}_{\text{sparse}}\mathbf{M}\mathbf{z}\right\|_2 \leq (1+\epsilon/3)\left\|\mathbf{M}\mathbf{z}\right\|_2, \tag{18}$$

holds for all $\mathbf{z} \in \mathbb{R}^m$ simultaneously with probability at least $1-\delta/2$. Eq. (18) also implies that $\operatorname{rank}(\mathbf{\Phi}_{\text{sparse}}\mathbf{M}) = \operatorname{rank}(\mathbf{M}) = r$. Now, conditioning on the event that Eq. (18) holds, and using the fact that $\mathbf{\Phi}_{\text{srht}}$ is an $(r,\delta/2,\epsilon/3)$-OSE, we have

$$(1-\epsilon/3)\left\|\mathbf{\Phi}_{\text{sparse}}\mathbf{M}\mathbf{z}\right\|_2 \leq \left\|\mathbf{\Phi}_{\text{srht}}(\mathbf{\Phi}_{\text{sparse}}\mathbf{M}\mathbf{z})\right\|_2 \leq (1+\epsilon/3)\left\|\mathbf{\Phi}_{\text{sparse}}\mathbf{M}\mathbf{z}\right\|_2, \tag{19}$$

holds for all $\mathbf{z} \in \mathbb{R}^m$ simultaneously with probability at least $1-\delta/2$. When both Eq. (18) and Eq. (19) hold, we have, for any $\mathbf{z} \in \mathbb{R}^m$,

$$\begin{aligned}\left\|\mathbf{\Phi}_{\text{srht}}(\mathbf{\Phi}_{\text{sparse}}\mathbf{M}\mathbf{z})\right\|_2 &\leq (1+\epsilon/3)\left\|\mathbf{\Phi}_{\text{sparse}}\mathbf{M}\mathbf{z}\right\|_2 \\ &\leq (1+\epsilon/3)^2\left\|\mathbf{M}\mathbf{z}\right\|_2 \\ &\leq (1+\epsilon)\left\|\mathbf{M}\mathbf{z}\right\|_2,\end{aligned} \tag{20}$$

where we have used Eq. (18), Eq. (19) and the fact that $\epsilon < 1$. The other direction: $\left\|\mathbf{\Phi}_{\text{srht}}(\mathbf{\Phi}_{\text{sparse}}\mathbf{M}\mathbf{z})\right\|_2 \geq (1-\epsilon)\left\|\mathbf{M}\mathbf{z}\right\|_2$ can be proved using the same method. Now, notice that, by union bound, the probability that both Eq. (18) and Eq. (19) hold simultaneously for any $\mathbf{z}$ is at least $1-\delta$. This concludes our proof. □

B Proof of Theorem 6

Proof. Let $\tilde{\mathbf{b}} = \mathbf{A}\tilde{\mathbf{x}}$ denote the prediction using approximated $\tilde{\mathbf{x}}$ returned by Algorithm 1. Let $\mathbf{b}^* = \mathbf{A}\mathbf{x}^*$ denote the prediction using optimal solution $\mathbf{x}^*$ of Eq. (1). Define $\mathbf{b}_0 = \mathbf{A}\mathbf{x}_0$. Let the thin SVD of $\mathbf{A}$ be $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.

Using the classical bias-variance decomposition [4], for any $\hat{\mathbf{b}} \in \mathbb{R}^n$, we have

$$\operatorname{risk}(\hat{\mathbf{b}}) = \operatorname{bias}(\hat{\mathbf{b}}) + \operatorname{var}(\hat{\mathbf{b}}), \tag{21}$$

where we define

$$\operatorname{bias}(\hat{\mathbf{b}}) \triangleq \frac{1}{n}\left\|\mathbf{E}\left[\hat{\mathbf{b}}\right]-\mathbf{b}_0\right\|_2^2 \quad \text{and} \quad \operatorname{var}(\hat{\mathbf{b}}) \triangleq \frac{1}{n}\mathbf{E}\left[\left\|\hat{\mathbf{b}}-\mathbf{E}\left[\hat{\mathbf{b}}\right]\right\|_2^2\right]$$

as the bias component and the variance component, respectively.

By Lemma 3, we have $\mathbf{x}^* = \mathbf{V}\mathbf{G}^{-1}\mathbf{U}^T\mathbf{b}$, where $\mathbf{G} = \lambda\mathbf{\Sigma}^{-1}+\mathbf{\Sigma}$. And by Lemma 4, we have $\tilde{\mathbf{x}} = \mathbf{V}\tilde{\mathbf{G}}^{-1}\mathbf{U}^T\mathbf{b}$, where $\tilde{\mathbf{G}} = \lambda\mathbf{\Sigma}^{-1}+\mathbf{\Sigma}(\mathbf{S}\mathbf{V})^T(\mathbf{S}\mathbf{V})$. Also notice that $\left\|\mathbf{G}^{-1}\right\|_F^2 = \rho^2$. Using these definition, we first bound the variance component $\operatorname{var}(\tilde{\mathbf{b}})$. We have

$$\begin{aligned}\operatorname{var}(\tilde{\mathbf{b}}) &= \frac{1}{n}\mathbf{E}\left[\left\|\tilde{\mathbf{b}}-\mathbf{E}\left[\hat{\mathbf{b}}\right]\right\|_2^2\right] \\ &= \frac{1}{n}\mathbf{E}\left[\left\|\mathbf{A}\tilde{\mathbf{x}}-\mathbf{E}\left[\mathbf{A}\tilde{\mathbf{x}}\right]\right\|_2^2\right] \\ &= \frac{1}{n}\mathbf{E}\left[\left\|\mathbf{A}\mathbf{V}\tilde{\mathbf{G}}^{-1}\mathbf{U}^T\mathbf{e}\right\|_2^2\right]\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\mathbf{E}\left[\left\|\mathbf{U}\boldsymbol{\Sigma}\tilde{\mathbf{G}}^{-1}\mathbf{U}^T\mathbf{e}\right\|_2^2\right] \\
&= \frac{1}{n}\mathbf{E}\left[\mathrm{tr}\left(\mathbf{e}^T\mathbf{U}(\tilde{\mathbf{G}}^{-1})^T\boldsymbol{\Sigma}^2\tilde{\mathbf{G}}^{-1}\mathbf{U}^T\mathbf{e}\right)\right] \\
&= \frac{1}{n}\mathbf{E}\left[\mathrm{tr}\left(\mathbf{U}(\tilde{\mathbf{G}}^{-1})^T\boldsymbol{\Sigma}^2\tilde{\mathbf{G}}^{-1}\mathbf{U}^T\mathbf{e}\mathbf{e}^T\right)\right] \\
&= \frac{\sigma^2}{n}\,\mathrm{tr}\left(\mathbf{U}(\tilde{\mathbf{G}}^{-1})^T\boldsymbol{\Sigma}^2\tilde{\mathbf{G}}^{-1}\mathbf{U}^T\right) \\
&= \frac{\sigma^2}{n}\,\mathrm{tr}\left((\tilde{\mathbf{G}}^{-1})^T\boldsymbol{\Sigma}^2\tilde{\mathbf{G}}^{-1}\right) \\
&= \frac{\sigma^2}{n}\left\|\boldsymbol{\Sigma}\tilde{\mathbf{G}}^{-1}\right\|_F^2,
\end{aligned}
$$

where we have repeatedly used the cyclic property of trace of matrix product. Similarly, one can show that

$$
\mathrm{var}(\mathbf{b}^*) = \frac{\sigma^2}{n}\left\|\boldsymbol{\Sigma}\mathbf{G}^{-1}\right\|_F^2.
$$

Now we recall the definition $\mathbf{R} = \mathbf{G}^{-1}(\tilde{\mathbf{G}} - \mathbf{G})$ and write $\tilde{\mathbf{G}}^{-1} = (\mathbf{I}+\mathbf{R})^{-1}\mathbf{G}^{-1}$. Then, we have

$$
\begin{aligned}
\left\|\boldsymbol{\Sigma}\tilde{\mathbf{G}}^{-1}\right\|_F &= \left\|\boldsymbol{\Sigma}(\mathbf{I}+\mathbf{R})^{-1}\mathbf{G}^{-1}\right\|_F \\
&= \left\|\boldsymbol{\Sigma}\mathbf{G}^{-1} - \boldsymbol{\Sigma}(\mathbf{I}+\mathbf{R})^{-1}\mathbf{R}\mathbf{G}^{-1}\right\|_F && (22)\\
&\le \left\|\boldsymbol{\Sigma}\mathbf{G}^{-1}\right\|_F + \left\|\boldsymbol{\Sigma}(\mathbf{I}+\mathbf{R})^{-1}\mathbf{R}\mathbf{G}^{-1}\right\|_F && (23)\\
&\le \left\|\boldsymbol{\Sigma}\mathbf{G}^{-1}\right\|_F + \|\boldsymbol{\Sigma}\|_2\left\|(\mathbf{I}+\mathbf{R})^{-1}\mathbf{R}\right\|_2\left\|\mathbf{G}^{-1}\right\|_F \\
&\le \left\|\boldsymbol{\Sigma}\mathbf{G}^{-1}\right\|_F + \epsilon\rho\,\|\mathbf{A}\|_2, && (24)
\end{aligned}
$$

where Eq. (22) follows from Woodbury matrix identity, Eq. (23) follows from triangle inequality and Eq. (24) follows from Lemma 5, the fact that $\mathbf{S}$ is an $(r,\delta,\epsilon/4)$-OSE and the definition that $\rho = \left\|\mathbf{G}^{-1}\right\|_F$.

Now, we can bound $\mathrm{var}(\tilde{\mathbf{b}})$ as follows

$$
\begin{aligned}
\mathrm{var}(\tilde{\mathbf{b}}) &= \frac{\sigma^2}{n}\left\|\boldsymbol{\Sigma}\tilde{\mathbf{G}}^{-1}\right\|_F^2 \\
&\le \frac{\sigma^2}{n}\left(\left\|\boldsymbol{\Sigma}\mathbf{G}^{-1}\right\|_F + \epsilon\rho\,\|\mathbf{A}\|_2\right)^2 \\
&\le \frac{\sigma^2}{n}\left(\left\|\boldsymbol{\Sigma}\mathbf{G}^{-1}\right\|_F^2 + 2\epsilon\rho\,\|\mathbf{A}\|_2\left\|\boldsymbol{\Sigma}\mathbf{G}^{-1}\right\|_F + \epsilon^2\rho^2\,\|\mathbf{A}\|_2^2\right) \\
&\le \frac{\sigma^2}{n}\left(\left\|\boldsymbol{\Sigma}\mathbf{G}^{-1}\right\|_F^2 + 2\epsilon\rho^2\,\|\mathbf{A}\|_2^2 + \epsilon^2\rho^2\,\|\mathbf{A}\|_2^2\right) && (25)\\
&\le \mathrm{var}(\mathbf{b}^*) + (2\epsilon+\epsilon^2)\frac{\sigma^2}{n}\rho^2\,\|\mathbf{A}\|_2^2, && (26)
\end{aligned}
$$

where Eq. (25) follows from the definition $\rho = \left\|\mathbf{G}^{-1}\right\|_F$.

Next, we bound the bias component $\mathrm{bias}(\tilde{\mathbf{b}})$. We can simplify $\mathrm{bias}(\tilde{\mathbf{b}})$ as follows

$$
\begin{aligned}
\mathrm{bias}(\tilde{\mathbf{b}}) &= \frac{1}{n}\left\|\mathbf{E}\left[\tilde{\mathbf{b}}\right] - \mathbf{b}_0\right\|^2 \\
&= \frac{1}{n}\left\|\mathbf{A}\mathbf{V}\tilde{\mathbf{G}}^{-1}\mathbf{U}^T\mathbf{E}\left[\mathbf{b}\right] - \mathbf{b}_0\right\|_2^2 \\
&= \frac{1}{n}\left\|\mathbf{A}\mathbf{V}\tilde{\mathbf{G}}^{-1}\mathbf{U}^T\mathbf{A}\mathbf{x}_0 - \mathbf{A}\mathbf{x}_0\right\|_2^2 \\
&= \frac{1}{n}\left\|\mathbf{U}\boldsymbol{\Sigma}\tilde{\mathbf{G}}^{-1}\mathbf{U}^T\mathbf{A}\mathbf{x}_0 - \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}_0\right\|_2^2 \\
&= \frac{1}{n}\left\|\boldsymbol{\Sigma}\tilde{\mathbf{G}}^{-1}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}_0 - \boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}_0\right\|_2
\end{aligned}
$$

$$= \frac{1}{n} \left\| (\boldsymbol{\Sigma}\tilde{\mathbf{G}}^{-1} - \mathbf{I})\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}_0 \right\|_2^2,$$

where we have dropped $\mathbf{U}$ which does not change l_2 norms. Using the same method, $\mathrm{bias}(\mathbf{b}^*)$ can be simplified as

$$\mathrm{bias}(\mathbf{b}^*) = \frac{1}{n} \left\| (\boldsymbol{\Sigma}\mathbf{G}^{-1} - \mathbf{I})\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}_0 \right\|_2^2.$$

It is easy to see that $\sqrt{n\,\mathrm{bias}(\tilde{\mathbf{b}})} = \left\| (\boldsymbol{\Sigma}\tilde{\mathbf{G}}^{-1} - \mathbf{I})\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}_0 \right\|_2$ and we can bound it as follows

$$\begin{aligned}
\sqrt{n\,\mathrm{bias}(\tilde{\mathbf{b}})} &= \left\| (\boldsymbol{\Sigma}\tilde{\mathbf{G}}^{-1} - \mathbf{I})\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}_0 \right\|_2 \\
&= \left\| (\boldsymbol{\Sigma}(\mathbf{I}+\mathbf{R})^{-1}\mathbf{G}^{-1} - \mathbf{I})\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}_0 \right\|_2 \\
&= \left\| (\boldsymbol{\Sigma}\mathbf{G}^{-1} - \mathbf{I} - \boldsymbol{\Sigma}(\mathbf{I}+\mathbf{R})^{-1}\mathbf{R}\mathbf{G}^{-1})\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}_0 \right\|_2 \qquad (27) \\
&\le \left\| (\boldsymbol{\Sigma}\mathbf{G}^{-1} - \mathbf{I})\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}_0 \right\|_2 + \left\| \boldsymbol{\Sigma}(\mathbf{I}+\mathbf{R})^{-1}\mathbf{R}\mathbf{G}^{-1}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}_0 \right\|_2 \qquad (28) \\
&\le \sqrt{n\,\mathrm{bias}(\mathbf{b}^*)} + \|\boldsymbol{\Sigma}\|_2 \left\| (\mathbf{I}+\mathbf{R})^{-1}\mathbf{R} \right\|_2 \left\| \mathbf{G}^{-1}\boldsymbol{\Sigma} \right\|_2 \left\| \mathbf{V}^T\mathbf{x}_0 \right\|_2 \\
&\le \sqrt{n\,\mathrm{bias}(\mathbf{b}^*)} + \|\mathbf{A}\|_2 \cdot \epsilon \cdot \max_i \frac{\sigma_i^2}{\lambda+\sigma_i^2} \|\mathbf{x}_0\|_2 \qquad (29) \\
&\le \sqrt{n\,\mathrm{bias}(\mathbf{b}^*)} + \epsilon \|\mathbf{A}\|_2 \|\mathbf{x}_0\|_2, \qquad (30)
\end{aligned}$$

where Eq. (27) follows from matrix inversion lemma, Eq. (28) is the triangle inequality and Eq. (29) follows from Lemma 5 and that $\mathbf{S}$ is an $(r, \delta, \epsilon/4)$-OSE. Now, dividing both sides of the above inequality Eq. (30) by $\sqrt{n}$, we have that

$$\sqrt{\mathrm{bias}(\tilde{\mathbf{b}})} \le \sqrt{\mathrm{bias}(\mathbf{b}^*)} + \frac{\epsilon}{\sqrt{n}} \|\mathbf{A}\|_2 \|\mathbf{x}_0\|_2.$$

Finally, we obtain the following bound on $\mathrm{bias}(\tilde{\mathbf{b}})$

$$\begin{aligned}
\mathrm{bias}(\tilde{\mathbf{b}}) &\le \mathrm{bias}(\mathbf{b}^*) + 2\sqrt{\mathrm{bias}(\mathbf{b}^*)} \cdot \frac{\epsilon}{\sqrt{n}} \|\mathbf{A}\|_2 \|\mathbf{x}_0\|_2 + \frac{\epsilon^2}{n} \|\mathbf{A}\|_2^2 \|\mathbf{x}_0\|_2^2 \\
&\le \mathrm{bias}(\mathbf{b}^*) + \frac{2}{\sqrt{n}} \left\| (\boldsymbol{\Sigma}\mathbf{G}^{-1} - \mathbf{I})\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{x}_0 \right\|_2 \cdot \frac{\epsilon}{\sqrt{n}} \|\mathbf{A}\|_2 \|\mathbf{x}_0\|_2 + \frac{\epsilon^2}{n} \|\mathbf{A}\|_2^2 \|\mathbf{x}_0\|_2^2 \\
&\le \mathrm{bias}(\mathbf{b}^*) + 2\left(\frac{1}{\sqrt{n}} \left\| \boldsymbol{\Sigma}\mathbf{G}^{-1} - \mathbf{I} \right\|_2 \|\boldsymbol{\Sigma}\|_2 \left\| \mathbf{V}^T\mathbf{x}_0 \right\|_2 \right) \left(\frac{\epsilon}{\sqrt{n}} \|\mathbf{A}\|_2 \|\mathbf{x}_0\|_2 \right) + \frac{\epsilon^2}{n} \|\mathbf{A}\|_2^2 \|\mathbf{x}_0\|_2^2 \\
&\le \mathrm{bias}(\mathbf{b}^*) + \frac{2\epsilon}{n} \|\mathbf{A}\|_2^2 \|\mathbf{x}_0\|_2^2 + \frac{\epsilon^2}{n} \|\mathbf{A}\|_2^2 \|\mathbf{x}_0\|_2^2 \qquad (31) \\
&= \mathrm{bias}(\mathbf{b}^*) + \frac{(2\epsilon+\epsilon^2)}{n} \|\mathbf{A}\|_2^2 \|\mathbf{x}_0\|_2^2, \qquad (32)
\end{aligned}$$

where Eq. (31) holds since $\|\boldsymbol{\Sigma}\|_2 = \|\mathbf{A}\|_2$, $\left\| \mathbf{V}^T\mathbf{x}_0 \right\|_2 = \|\mathbf{x}_0\|_2$ and $\left\| \boldsymbol{\Sigma}\mathbf{G}^{-1} - \mathbf{I} \right\|_2 = \max_i \frac{\lambda}{\lambda+\sigma_i^2} \le 1$.

The theorem follows immediately from Eq. (21), Eq. (26), Eq. (32) and the fact that $\epsilon < 1$. □

C Proof of Theorem 7

We first prove a generalization of Lemma 1 as follows.

Lemma 7. *Given $\mathbf{A} \in \mathbb{R}^{n\times p}$ of rank r, $\mathbf{B} \in \mathbb{R}^{n\times m}$ and $\lambda > 0$. Suppose that $\mathbf{S} \in \mathbb{R}^{t\times p}$ is an $(r, \delta, \epsilon/4)$-OSE for $\epsilon \in (0,1)$ and $\delta \in (0,1)$. Then, with probability $1-\delta$, we have*

$$\left\| \tilde{\mathbf{X}} - \mathbf{X}^* \right\|_F \le \epsilon \|\mathbf{X}^*\|_F,$$

where $\tilde{\mathbf{X}}$ is given by Eq. (16) *and $\mathbf{X}^*$ is the optimal solution to multiple response ridge regression Eq.* (14)*.*

Proof. Let the thin SVD of $\mathbf{A}$ be $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Fix an arbitrary $i \in [m]$, consider the column vectors $\tilde{\mathbf{X}}^{(i)}$ and $\mathbf{X}^{*(i)}$. By definition, we can see that $\mathbf{X}^{*(i)} = \mathbf{A}^T(\lambda \mathbf{I}_n + \mathbf{A}\mathbf{A}^T)^{-1}\mathbf{B}^{(i)}$ and

$$\tilde{\mathbf{X}}^{(i)} = \mathbf{A}^T(\mathbf{A}\mathbf{S}^T)^{\dagger^T}(\lambda(\mathbf{A}\mathbf{S}^T)^{\dagger^T} + \mathbf{A}\mathbf{S}^T)^{\dagger}\mathbf{B}^{(i)}.$$

Define $\tilde{\mathbf{G}} = \lambda\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}(\mathbf{S}\mathbf{V})^T(\mathbf{S}\mathbf{V})$ and $\mathbf{G} = \lambda\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}$. Now, we can regard $\mathbf{B}^{(i)}$ as the target vector and then apply Lemma 3 on $\mathbf{X}^{*(i)}$ and Lemma 4 on $\tilde{\mathbf{X}}^{(i)}$, respectively. This shows that $\mathbf{X}^{*(i)} = \mathbf{V}\mathbf{G}^{-1}\mathbf{U}^T\mathbf{B}^{(i)}$ and $\tilde{\mathbf{X}}^{(i)} = \mathbf{V}\tilde{\mathbf{G}}^{-1}\mathbf{U}^T\mathbf{B}^{(i)}$. Hence, combining all columns $i \in [m]$, we have that $\mathbf{X}^* = \mathbf{V}\mathbf{G}^{-1}\mathbf{U}^T\mathbf{B}$ and $\tilde{\mathbf{X}} = \mathbf{V}\tilde{\mathbf{G}}^{-1}\mathbf{U}^T\mathbf{B}$.

Similar to the proof of Lemma 1, we recall the definition $\mathbf{R} = \mathbf{G}^{-1}(\tilde{\mathbf{G}} - \mathbf{G})$ and write $\tilde{\mathbf{G}}^{-1} = (\mathbf{I} + \mathbf{R})^{-1}\mathbf{G}^{-1}$. Let $\epsilon' = \epsilon/4$. Then, we have

$$\begin{aligned}
\left\|\tilde{\mathbf{X}} - \mathbf{X}^*\right\|_F &= \left\|\mathbf{V}(\tilde{\mathbf{G}}^{-1} - \mathbf{G}^{-1})\mathbf{U}^T\mathbf{B}\right\|_F \\
&= \left\|(\tilde{\mathbf{G}}^{-1} - \mathbf{G}^{-1})\mathbf{U}^T\mathbf{B}\right\|_F \\
&= \left\|\left((\mathbf{I} + \mathbf{R})^{-1} - \mathbf{I}\right)\mathbf{G}^{-1}\mathbf{U}^T\mathbf{B}\right\|_F \\
&= \left\|(\mathbf{I} + \mathbf{R})^{-1}\mathbf{R}\mathbf{G}^{-1}\mathbf{U}^T\mathbf{B}\right\|_F \qquad (33) \\
&\le \left\|(\mathbf{I} + \mathbf{R})^{-1}\mathbf{R}\right\|_2 \left\|\mathbf{G}^{-1}\mathbf{U}^T\mathbf{B}\right\|_F \\
&= \left\|(\mathbf{I} + \mathbf{R})^{-1}\mathbf{R}\right\|_2 \left\|\mathbf{X}^*\right\|_F \\
&\le \frac{2\epsilon' + \epsilon'^2}{1 - (2\epsilon' + \epsilon'^2)} \left\|\mathbf{X}^*\right\|_F \qquad (34) \\
&\le 4\epsilon' \left\|\mathbf{X}^*\right\|_F \\
&= \epsilon \left\|\mathbf{X}^*\right\|_F . \qquad (35)
\end{aligned}$$

where Eq. (33) follows from matrix inversion lemma and Eq. (34) follows from Lemma 5 and the fact that $\mathbf{S}$ is an $(r, \delta, \epsilon/4)$-OSE. □

Then, Theorem 7 is an immediate consequence of Lemma 7.

Proof of Theorem 7. The bound on $\left\|\tilde{\mathbf{X}} - \mathbf{X}\right\|_F$ follows immediately from Lemma 7 and Theorem 3 which shows that $\mathbf{S} = \mathbf{\Phi}_{\text{srht}}\mathbf{\Phi}_{\text{sparse}}$ is an $(r, \delta, \epsilon/4)$-OSE. And the running time analysis is similar to the one given in Section 3. □

D Structured Ridge Regression: Algorithm Details and Analysis

In this part, we supplement the details of the relative-error approximation algorithm for structured ridge regression.

First, we compute the sketched matrix $\varphi(\mathbf{A})\mathbf{S}^T$, where $\mathbf{S} = \mathbf{\Phi}_{\text{srht}}\mathbf{\Phi}_{\text{sparse}}$. Next, we compute the approximate solution $\tilde{\mathbf{x}}$ as follows

$$\tilde{\mathbf{x}} = \varphi(\mathbf{A})^T(\varphi(\mathbf{A})\mathbf{S}^T)^{\dagger^T}(\lambda(\varphi(\mathbf{A})\mathbf{S}^T)^{\dagger^T} + \varphi(\mathbf{A})\mathbf{S}^T)^{\dagger}\mathbf{b}. \qquad (36)$$

The following theorem shows that the $\tilde{\mathbf{x}}$ obtained by Eq. (36) is a relative-error approximation of $\mathbf{x}^*$.

Theorem 8. *Given $\mathbf{A} \in \mathbb{R}^{n\times p}$, $\mathbf{b} \in \mathbb{R}^n$, $\lambda > 0$, $q \in \mathbb{N}^+$, parameter $\epsilon \in (0,1)$ and $\delta \in (0,1)$. Select integers t', t such that $t' \ge 2\delta^{-1}(n^2+n)/(\epsilon/6 - \epsilon^2/144)^2$ and $t \ge 72\epsilon^{-1}[\sqrt{n} + \sqrt{8\log(6p/\delta)}]^2\log(6n/\delta)$. Let $\mathbf{S} = \mathbf{\Phi}_{srht}\mathbf{\Phi}_{sparse}$, where $\mathbf{\Phi}_{sparse} \in \mathbb{R}^{t'\times pq}$ is a sparse embedding matrix and $\mathbf{\Phi}_{srht} \in \mathbb{R}^{t\times t'}$ is an SRHT matrix. Then, with probability at least $1-\delta$, we have*

$$\left\|\tilde{\mathbf{x}} - \mathbf{x}^*\right\|_2 \le \epsilon \left\|\mathbf{x}^*\right\|_2 ,$$

where $\tilde{\mathbf{x}}$ is given by Eq. (36) *and $\mathbf{x}^*$ is the optimal solution to structured ridge regression Eq.* (17)*. In addition, the total time complexity of computing $\varphi(\mathbf{A})\mathbf{S}^T$ and $\tilde{\mathbf{x}}$ is $O(\text{nnz}(\mathbf{A})\log^2(q) + n^3 q\log^2(q)/\epsilon^2 + n^3\log(\frac{n}{\epsilon})/\epsilon^2)$.*

In order to prove Theorem 8, we only need to calculate the running time of Eq. (36), since the relative error approximation guarantee directly comes from Theorem 4. To analyze the running time, we use the following result of Avron et al. [3].

Lemma 8. [3, Lemma 9, 10, 11] *Let $\mathbf{A} \in \mathbb{R}^{n\times p}$, $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{z} \in \mathbb{R}^{pq}$. Let $\varphi_q : \mathbb{R}^p \to \mathbb{R}^{pq}$ be the kernel expansion function such that $\varphi_q(\mathbf{a}) = \{a_i^{j-1}\}_{(i,j)\in[p]\times[q]}$ for all $\mathbf{a} \in \mathbb{R}^p$. Let $\mathbf{\Phi}_{sparse} \in \mathbb{R}^{t'\times pq}$ be a sparse embedding matrix.*

Then, there exists a fast matrix-vector multiplication algorithm such that $\mathbf{y}^T\varphi_q(\mathbf{A})$ *and* $\varphi_q(\mathbf{A})\mathbf{z}$ *can be computed in* $O(\mathrm{nnz}(\mathbf{A})\log^2(q) + nq\log^2(q))$ *time. And therefore the product* $\varphi_q(\mathbf{A})\mathbf{\Phi}_{sparse}^T$ *can be computed in* $O(\mathrm{nnz}(\mathbf{A})\log^2(q) + nqt'\log^2(q))$ *time.*

Proof of Theorem 8. From Theorem 3 we knows that $\mathbf{S} = \mathbf{\Phi}_{\text{srht}}\mathbf{\Phi}_{\text{sparse}}$ is an $(n, \delta, \epsilon/4)$ OSE. The bound on $\|\tilde{\mathbf{x}} - \mathbf{x}\|_2$ follows immediately from Lemma 1.

From Lemma 8 we know the time cost to compute $\varphi_q(\mathbf{A})\mathbf{\Phi}_{\text{sparse}}^T$ is $O((\mathrm{nnz}(\mathbf{A}) + qn^3/\epsilon^2)\log^2(q))$. Therefore, the total time complexity of computing $\varphi_q(\mathbf{A})\mathbf{S}^T$ and $\tilde{\mathbf{x}}$ is $O(\mathrm{nnz}(\mathbf{A})\log^2(q) + qn^3/\epsilon^2\log^2(q) + n^3/\epsilon^2\log(\frac{n}{\epsilon}))$. □

E Argument and Counterexample for the Proof of [22, Theorem 1]

In this section, we show a counterexample of a claim used in [22, Theorem 1]. For clarity, we will use the notations in [22].

The risk inflation bound of [22] is based on the claim that the following function

$$B(\mathbf{M}) = n\lambda^2\mathbf{z}^T(\mathbf{M} + n\lambda\mathbf{I}_n)^{-2}z$$

is non-increasing in $\mathbf{M}$ for any positive semi-definite matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ and $\mathbf{z} \in \mathbb{R}^n$. It is easy to see that this claim is equivalent to the claim that for any positive definite matrices $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{n\times n}$, if $\mathbf{M}_1 \succeq \mathbf{M}_2 \succ \mathbf{0}_n$, then $\mathbf{M}_1^{-2} \preceq \mathbf{M}_2^{-2}$.

However, this claim is not true and we can construct following counterexample.

Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be the following 2×2 matrices.

$$\mathbf{M}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 100 \end{bmatrix}, \quad \mathbf{M}_2 = \begin{bmatrix} 2/3 & -2 \\ -2 & 12 \end{bmatrix}.$$

It is easy to verify that $\mathbf{M}_1 \succeq \mathbf{M}_2 \succ \mathbf{0}_2$, but $\mathbf{M}_1^{-2} \not\preceq \mathbf{M}_2^{-2}$. This issue also propagates to other parts of their proof. Hence their bound [22, Theorem 1] may be flawed.