An Iterative, Sketching-based Framework for Ridge Regression

Agniva Chowdhury [1] **Jiasen Yang** [1] **Petros Drineas** [2]

Abstract

Ridge regression is a variant of regularized least squares regression that is particularly suitable in settings where the number of predictor variables greatly exceeds the number of observations. We present a simple, iterative, sketching-based algorithm for ridge regression that guarantees high-quality approximations to the optimal solution vector. Our analysis builds upon two simple structural results that boil down to randomized matrix multiplication, a fundamental and well-understood primitive of randomized linear algebra. An important contribution of our work is the analysis of the behavior of sub-sampled ridge regression problems when the ridge leverage scores are used: we prove that accurate approximations can be achieved by a sample whose size depends on the degrees of freedom of the ridge-regression problem rather than the dimensions of the design matrix. Our empirical evaluations verify our theoretical results on both real and synthetic data.

1. Introduction

In statistics and machine learning, ridge regression (Gunst & Mason, 1977; Hoerl & Kennard, 1970) (also known as Tikhonov regularization or weight decay) is a variant of regularized least squares problems where the choice of the penalty function is the squared ℓ_2-norm. Formally, let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be the design matrix and let $\mathbf{b} \in \mathbb{R}^n$ be the response vector. Then, the linear algebraic formulation of the ridge regression problem is as follows:

$$\mathcal{Z}^* = \min_{\mathbf{x}\in\mathbb{R}^d} \left\{ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_2^2 \right\}, \tag{1}$$

where $\lambda > 0$ is the regularization parameter. There are two fundamental motivations underlying the use of ridge regression. First, when $d \gg n$, *i.e.*, the number of predictor variables d greatly exceeds the number of observations n, fitting the full model without regularization (*i.e.*, setting λ to zero) will result in large prediction intervals and a non-unique regression estimator. Second, if the design matrix $\mathbf{A}$ is ill-conditioned, solving the standard least-squares problem without regularization would depend on $(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}$. This inversion would be problematic if $\mathbf{A}^\mathsf{T}\mathbf{A}$ were singular or nearly singular and thus adding even a little noise to the elements of $\mathbf{A}$ could result in large changes in $(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}$. Due to these two considerations, solving standard least-squares problems without regularization may provide a good fit to the training data but may not generalize well to test data.

Ridge regression abandons the requirement of an unbiased estimator in order to address the aforementioned problems. At the cost of introducing bias, ridge regression reduces the variance and thus might reduce the overall mean squared error (MSE). The minimizer of eqn. (1) is

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

or, equivalently (see Saunders et al. (1998) and Lemma 9 in Appendix A),

$$\mathbf{x}^* = \mathbf{A}^\mathsf{T}\left(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n\right)^{-1} \mathbf{b}. \tag{3}$$

Both formulations work for any $\lambda > 0$ for either under-constrained or over-constrained ridge regression problems, regardless of the rank of the design matrix $\mathbf{A}$. It is easy to see that $\mathbf{x}^*$ can be computed in time

$$\mathcal{O}(nd\min\{n,d\} + \min\{n^3, d^3\}) = \mathcal{O}(nd\min\{n,d\}).$$

In our work, we will focus on design matrices $\mathbf{A} \in \mathbb{R}^{n\times d}$ with $d \gg n$, which is the most common setting for ridge regression. For simplicity of exposition, we will assume that the rank of $\mathbf{A}$ is equal to n.[1] In the context of ridge regression, a much more important quantity than the rank of the design matrix is the effective *degrees of freedom*:

$$d_\lambda = \sum_{i=1}^{n} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \leq n, \tag{4}$$

where σ_i are the singular values of $\mathbf{A}$.

[1]Department of Statistics, Purdue University, West Lafayette, IN [2]Department of Computer Science, Purdue University, West Lafayette, IN. Correspondence to: Agniva Chowdhury <chowdhu5@purdue.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[1]Our results can be slightly improved to depend on the rank ρ of the matrix $\mathbf{A}$ instead of n.

The recent flurry of activity on Randomized Linear Algebra (RLA) (Drineas & Mahoney, 2016) and the widespread use of *sketching* as a tool for matrix computations (Woodruff, 2014), resulted in many novel results for ridge regression. In Section 1.2 we discuss relevant prior work.

1.1. Our Contributions

We present a novel iterative algorithm (Algorithm 1) for *sketched* ridge regression and two simple sketching-based structural conditions under which Algorithm 1 guarantees highly accurate approximations to the optimal solution $\mathbf{x}^*$. More precisely, Algorithm 1 guarantees that, as long as a simple structural constraint is satisfied, the resulting approximate solution vector $\widehat{\mathbf{x}}^*$ satisfies (after t iterations)

$$\|\mathbf{x}^* - \widehat{\mathbf{x}}^*\|_2 \leq \varepsilon^t \|\mathbf{x}^*\|_2. \tag{5}$$

Prior to discussing the aforementioned constraint, we note that error guarantees of the above form are highly desirable. Indeed, beyond being a relative error guarantee, the dependency on ε drops exponentially fast as the number of iterations increases. It is easy to see that by setting $\varepsilon^t = \varepsilon'$, $\mathcal{O}(\ln(1/\varepsilon'))$ iterations would suffice to provide a relative error guarantee with accuracy parameter ε'. This means that converging to, say, ten decimal digits of accuracy would necessitate only a constant number of iterations. See Section 1.2 for a comparison of this bound with prior work.

Let $\mathbf{V} \in \mathbb{R}^{d\times n}$ be the matrix of right singular vectors of $\mathbf{A}$; recall that $\mathbf{A}$ has rank n. For eqn. (5) to hold, a sketching matrix $\mathbf{S} \in \mathbb{R}^{d\times s}$ is to be constructed such that (for an appropriate choice of the sketching dimension $s \ll d$)

$$\|\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V} - \mathbf{I}_n\|_2 \leq \frac{\varepsilon}{2}. \tag{6}$$

We note that the constraint of eqn. (6) has been the topic of intense research in the RLA literature; this is precisely the reason why we use eqn. (6) as the building block in our analysis. Indeed, assuming that $n \ll d$, one can use the (exact or approximate) column leverage scores (Mahoney & Drineas, 2009; Mahoney, 2011) of $\mathbf{A}$ to satisfy the aforementioned constraint, in which case $\mathbf{S}$ is a sampling-and-rescaling matrix. Perhaps more interestingly, a variety of oblivious sketching matrix constructions for $\mathbf{S}$ can be used to satisfy eqn. (6). We discuss various constructions for $\mathbf{S}$ in Section 2.1.

One deficiency of the structural constraint of eqn. (6) is that all known constructions for $\mathbf{S}$ that satisfy the constraint need a number of columns s that is proportional to n. As a result, the running time of any algorithm that computes the sketch $\mathbf{AS}$ is also proportional to n. It would be much better to design algorithms whose running time depends on the *degrees of freedom* d_λ, which is upper bounded by n, but could be significantly smaller depending on the distribution of the singular values and the choice of λ.

Towards that end, we analyze Algorithm 1 under a second structural constraint. We define a *diagonal* matrix $\mathbf{\Sigma}_\lambda \in \mathbb{R}^{n\times n}$ whose i-th diagonal entry is given by

$$(\mathbf{\Sigma}_\lambda)_{ii} = \sqrt{\frac{\sigma_i^2}{\sigma_i^2 + \lambda}}, \quad i = 1, \ldots, n. \tag{7}$$

Notice that $\|\mathbf{\Sigma}_\lambda\|_F^2 = d_\lambda$. Our second structural condition is given by

$$\|\mathbf{\Sigma}_\lambda \mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\mathbf{\Sigma}_\lambda - \mathbf{\Sigma}_\lambda^2\|_2 \leq \frac{\varepsilon}{4\sqrt{2}}. \tag{8}$$

Similarly to the constraint of eqn. (6), the constraint of eqn. (8) can also be satisfied by, for example, sampling with respect to the *ridge leverage scores* of Alaoui & Mahoney (2015); Cohen et al. (2017) or by oblivious sketching matrix constructions for $\mathbf{S}$. The difference is that, instead of having the column size s of the matrix $\mathbf{S}$ depend on n, it now depends on d_λ, which could be considerably smaller. Indeed, it follows that by sampling-and-rescaling $\mathcal{O}(d_\lambda \ln d_\lambda)$ predictor variables from the design matrix $\mathbf{A}$ (using either exact or approximate *ridge* leverage scores (Alaoui & Mahoney, 2015; Cohen et al., 2017) we can satisfy the constraint of eqn. (8). Similarly, oblivious sketching matrix constructions for $\mathbf{S}$ can be used to satisfy eqn. (8). We discuss constructions for $\mathbf{S}$ in Section 2.1.

However, this improved dependency on d_λ instead of n comes with a mild loss in accuracy. For simplicity, we only state a result when λ satisfies $\sigma_{k+1}^2 \leq \lambda \leq \sigma_k^2$ for some integer k, $1 \leq k \leq n$.[2] In words, λ can be thought of as "regularizing" the bottom $n-k$ singular values of the design matrix $\mathbf{A}$, since it dominates them. In this case, we prove that the approximation $\widehat{\mathbf{x}}^*$ returned by Algorithm 1 satisfies

$$\|\mathbf{x}^* - \widehat{\mathbf{x}}^*\|_2 \leq \frac{\varepsilon^t}{2}\left(\|\mathbf{x}^*\|_2 + \frac{1}{\sqrt{2\lambda}}\left\|\mathbf{U}_{k,\perp}^\mathsf{T}\mathbf{b}\right\|_2\right). \tag{9}$$

Here $\mathbf{U}_{k,\perp} \in \mathbb{R}^{n\times(n-k)}$ denotes the matrix of the bottom $n-k$ left singular vectors of the design matrix $\mathbf{A}$. In words, we achieve an additive-relative error approximation, where the additive error part depends on the norm of the "piece" of the response vector $\mathbf{b}$ that lies on the regularized component of the design matrix $\mathbf{A}$. As this piece grows, the quality of the approximation worsens. The error decreases exponentially fast with the number of iterations.

Another contribution of our work is Theorem 4, which proves that the mean-square-error (MSE) of the approximate solution $\widehat{\mathbf{x}}^*$ is a relative error approximation to the MSE of $\mathbf{x}^*$, under the structural assumptions of eqns. (6) or (8), even after a single iteration.

[2] The bound of eqn. (9) can be easily generalized to hold when $c_1\sigma_{k+1}^2 \leq \lambda \leq c_2\sigma_k^2$ for some constants $c_1, c_2 > 0$. For simplicity of exposition, we assume that both c_1 and c_2 equal one.

To the best of our knowledge, our bounds are a first attempt to provide general structural results that guarantee high-quality approximations to the optimal solution vector of ridge regression. Our first structural result can be satisfied by sampling with respect to the leverage scores or by the use of oblivious sketching matrices whose size depends on the rank of the design matrix and guarantees relative error approximations. Our second structural result presents the first accuracy analysis for ridge regression when the *ridge leverage scores are used to sample predictor variables*. Interestingly, the ridge leverage scores have been used in a number of applications that have to do with matrix approximation, cost-preserving projections, clustering, etc. (Cohen et al., 2017), but their performance in the context of ridge regression has not been analyzed in prior work. Our work here argues that the second structural condition can be satisfied by sampling with respect to the ridge leverage scores. The number of predictor variables to be sampled depends on the degrees of freedom of the ridge-regression problem rather than the dimensions of the design matrix, and results in a relative-additive error guarantee.

1.2. Prior Work

In this section, we discuss our contributions in the context of the large and ever-growing body of prior work on sketching-based algorithms for regression and ridge regression. The work more closely related to ours is Chen et al. (2015), which (in our notation) returns an approximation $\widehat{\mathbf{x}}^*$ to $\mathbf{x}^*$ that satisfies (with high probability) a relative error guarantee of the form

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

The running time of the proposed approach is $\mathcal{O}(\mathrm{nnz}(\mathbf{A}) + \varepsilon^{-2} n^3 \ln(n/\varepsilon))$. The proposed approach is also based on sketching $\mathbf{A}$ using RLA tools such as the count-min sketch of Clarkson & Woodruff (2013) and the sub-sampled Randomized Hadamard Transform of Ailon & Chazelle (2009); Sarlós (2006); Drineas et al. (2011). Compared to our work, notice that their dependency on ε is exponentially higher: our approach has a running time that grows with $\ln(1/\varepsilon)$ whereas the above bound grows proportionally to $1/\varepsilon^2$. Additionally, our analysis can be made to depend on the degrees of freedom of the ridge-regression problem (see Theorem 2 and Section 2.1). Finally, we complement the bounds on the MSE for the response vector presented in Theorem 6 of Chen et al. (2015) with a relative-error guarantee on the MSE of the solution vector (see Theorem 4). We should also mention that prior to Chen et al. (2015); Lu et al. (2013) proposed a fast approximation algorithm for the computation of the kernel matrix using the sub-sampled randomized Hadamard transformation (SRHT).

Recently, Wang et al. (2017) presented many results on ridge-regression problems assuming $n \gg d$. In this setting, the main motivation for ridge regression is to deal with the potential ill-conditioning of the design matrix $\mathbf{A}$. Wang et al. (2017) presented sketching-based approaches that guarantee relative error approximations to the value of the objective $\mathcal{Z}^*$, as opposed to the actual solution vector. Our approach and analysis is quite different and is applicable where $d \gg n$; the results of Wang et al. (2017) do not generalize to this setting. However, recent work by Avron et al. (2017a;b) also focused on $d \gg n$: for example, Theorem 17 of Avron et al. (2017b) presents structural conditions under which the value of the objective $\mathcal{Z}^*$ can be estimated up to relative error accuracy, but no bounds are presented for the approximate solution vector. This last result seems to necessitate two structural conditions: the first one is identical to the condition of eqn. (6), but the second one is on the spectral norm of an approximate matrix product that is not needed in our analysis.

Our work was partially motivated by Pilanci & Wainwright (2016), where an iterative algorithm (the so-called Iterative Hessian Sketch) was presented for standard (*i.e.*, $\lambda = 0$), over-constrained ($n \gg d$) regression problems. Indeed, the authors provide strong motivation that clarifies the need for algorithms for regression problems whose running times depends on $\ln(1/\varepsilon)$ in order to achieve ε-relative-error approximations. We emphasize that the transition from standard to regularized regression problems as well as from the over- to the under-constrained case is far from trivial. Indeed, algorithms and structural results for over-constrained regression problems date back to 2006 (Drineas et al., 2006b), whereas the analogous results for ridge-regression problems appeared after 2015. Similarly, the only result that we know for under-constrained regression problems ($\lambda = 0$, $n \ll d$) appeared in Section 6.2 of Drineas et al. (2012).

Another line of research that motivated our approach was the recent introduction of ridge leverage scores (Alaoui & Mahoney, 2015; Cohen et al., 2017). Indeed, our Theorem 2 presents a structural result that can be satisfied (with high probability) by sampling columns of $\mathbf{A}$ with probabilities proportional to (exact or approximate) ridge leverage scores (see Section 2.1). The number of sampled predictor variables (columns of $\mathbf{A}$) is proportional to $\mathcal{O}(d_\lambda \ln d_\lambda)$. To the best of our knowledge, this is the first result showing a strong accuracy guarantee for ridge regression problems when the ridge leverage scores are used to sample predictor variables, in one or more iterations. We also note a recent application of ridge leverage scores (Calandriello et al., 2017a;b) where the authors presented a row sampling algorithm in order to construct a kernel sketch which is eventually used in a second-order gradient-based method for online kernel convex optimization.

In yet another relevant line of work, much research recently focused on the computation and inversion of the kernel ma-

trix $\mathbf{A}\mathbf{A}^\mathsf{T}$ (or $\mathbf{A}^\mathsf{T}\mathbf{A}$). A number of recent papers have considered the problem of fast kernel approximation for large datasets (Zhang et al., 2015; Avron et al., 2017b; Musco & Musco, 2017; Calandriello et al., 2017c; Wang et al., 2017). However, direct comparison of the bounds presented in the aforementioned papers and our work is not straightforward, since our objective (accuracy of the approximate solution vector) is different than the objective of the above papers. In this context, there are also several recent works (Cutajar et al., 2016; Rudi et al., 2017; Ma & Belkin, 2017) that considered preconditioned gradient-based methods to develop fast and scalable approaches for approximating kernels.

Finally, Gonen et al. (2016) presented a sketching-based preconditioned SVRG approach for ridge regression problems that converges to the optimal solution in a number of iterations that depends on $\ln(1/\varepsilon)$, returning an ε-relative-error approximation to the objective value $\mathcal{Z}^*$. However, no such bounds were presented for the actual solution vector.

1.3. Notation

We use $\mathbf{a}, \mathbf{b}, \ldots$ to denote vectors and $\mathbf{A}, \mathbf{B}, \ldots$ to denote matrices. For a matrix $\mathbf{A}$, $\mathbf{A}_{*i}$ ($\mathbf{A}_{i*}$) denotes the i-th column (row) of $\mathbf{A}$ as a column (row) vector. For vector $\mathbf{a}$, $\|\mathbf{a}\|_2$ denotes its Euclidean norm; for a matrix $\mathbf{A}$, $\|\mathbf{A}\|_2$ denotes its spectral norm and $\|\mathbf{A}\|_F$ denotes its Frobenius norm. We refer the reader to Golub & Van Loan (1996) for properties of norms that will be quite useful in our work. For a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ with $d > n$ of rank n, its (thin) Singular Value Decomposition (SVD) is equal to the product $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$, with $\mathbf{U} \in \mathbb{R}^{n\times n}$ (the matrix of the left singular vectors), $\mathbf{V} \in \mathbb{R}^{d\times n}$ (the matrix of the right singular vectors), and $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$ a diagonal matrix whose diagonal entries are the singular values of $\mathbf{A}$. Computation of the SVD takes, in this setting, $\mathcal{O}(n^2 d)$ time. We will use the notation $\mathbf{U}_k \in \mathbb{R}^{n\times k}$ to denote the matrix of the top k left singular vectors and $\mathbf{U}_{k,\perp} \in \mathbb{R}^{n\times(n-k)}$ to denote the matrix of the bottom $n-k$ left singular vectors. We will often use σ_i to denote the singular values of a matrix implied by context. Additional notation will be introduced as needed.

2. Iterative, Sketching-based Ridge Regression

Algorithm 1 iteratively computes a sequence of vectors $\widetilde{\mathbf{x}}^{(j)} \in \mathbb{R}^d$ for $j = 1, \ldots, t$ and returns the estimator $\widehat{\mathbf{x}}^* = \sum_{j=1}^t \widetilde{\mathbf{x}}^{(j)}$ to the true solution vector $\mathbf{x}^*$ of eqn. (3).

In words, Algorithm 1 is quite simple: roughly, it solves ridge regression problems with the residual vector $\mathbf{b}^{(j)}$ (*i.e.*, the part of the vector $\mathbf{b}^{(j-1)}$ that was *not* captured in the previous iteration) as the new response vector for $i = 1, \ldots, t$. Our main quality-of-approximation results (Theorems 1 and 2) argue that returning the *sum* of those intermediate solutions results in a highly accurate approximation

Algorithm 1 Iterative, sketching-based ridge regression

Input: $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{b} \in \mathbb{R}^n$, $\lambda > 0$; number of iterations $t > 0$; sketching matrix $\mathbf{S} \in \mathbb{R}^{d\times s}$;
Initialize: $\mathbf{b}^{(0)} \leftarrow \mathbf{b}$, $\widetilde{\mathbf{x}}^{(0)} \leftarrow \mathbf{0}_d$, $\mathbf{y}^{(0)} \leftarrow \mathbf{0}_n$;
for $j = 1$ **to** t **do**
  $\mathbf{b}^{(j)} \leftarrow \mathbf{b}^{(j-1)} - \lambda\mathbf{y}^{(j-1)} - \mathbf{A}\widetilde{\mathbf{x}}^{(j-1)}$;
  $\mathbf{y}^{(j)} \leftarrow (\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)^{-1}\mathbf{b}^{(j)}$;
  $\widetilde{\mathbf{x}}^{(j)} \leftarrow \mathbf{A}^\mathsf{T}\mathbf{y}^{(j)}$;
end for
Output: Approximate solution vector $\widehat{\mathbf{x}}^* = \sum_{j=1}^t \widetilde{\mathbf{x}}^{(j)}$;

to the optimal solution vector $\mathbf{x}^*$. Theorem 1 presents a quality-of-approximation result under the assumption that the sketching matrix $\mathbf{S}$ satisfies the constraint of eqn. (6).

Theorem 1. *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. Assume that for some constant $0 < \varepsilon < 1$, the sketching matrix $\mathbf{S} \in \mathbb{R}^{d\times s}$ satisfies the constraint of eqn. (6). Then, the estimator $\widehat{\mathbf{x}}^*$ returned by Algorithm 1 satisfies*

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

Here $\mathbf{x}^*$ is the true solution of the ridge regression problem.

Similarly, Theorem 2 presents a quality-of-approximation result under the assumption that the sketching matrix $\mathbf{S}$ satisfies the constraint of eqn. (8).

Theorem 2. *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. Assume that for some constant $0 < \varepsilon < 1$, the sketching matrix $\mathbf{S} \in \mathbb{R}^{d\times s}$ satisfies the constraint of eqn. (8). Then, the estimator $\widehat{\mathbf{x}}^*$ returned by Algorithm 1 satisfies*

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

Here $k \in \{1, \ldots, n\}$ is an integer such that $\sigma_{k+1} \leq \lambda \leq \sigma_k$ and $\mathbf{x}^*$ is the true solution of the ridge regression problem.

As we have already discussed, the bound of Theorem 2 is weaker. However, the structural condition of eqn. (8) on which the above theorem depends, can be satisfied with a sketching matrix $\mathbf{S}$ whose dimensionality depends only on the degrees of freedom d_λ of the underlying ridge regression problem, as opposed to the dimensions of the design matrix. This could result in significant savings (see Section 2.1).

Our algorithm can also be viewed as a *preconditioned Richardson iteration* (see *e.g.*, Chapter 2 of Quarteroni & Valli (1994)) for solving the linear system $(\mathbf{A}\mathbf{A}^\mathsf{T}+\lambda\mathbf{I}_n)\mathbf{y} = \mathbf{b}$ with pre-conditioner $\mathbf{P}^{-1} = (\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I}_n)^{-1}$ and step-size equal to one. More precisely, Algorithm 1 can be formulated as

$$\bar{\mathbf{y}}^{(j)} = \bar{\mathbf{y}}^{(j-1)} + \mathbf{P}^{-1}\left(\mathbf{b} - (\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda\mathbf{I}_n)\bar{\mathbf{y}}^{(j-1)}\right),$$

where $\bar{\mathbf{y}}^{(j)} = \sum_{k=1}^{j} \mathbf{y}^{(k)}$ (see Appendix D for the derivation). Further, subject to the structural conditions of eqns. (6) and (8), it can be shown that $\bar{\mathbf{y}}^{(t)}$ converges to the true solution $\mathbf{y}^* = (\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}$ in $\mathcal{O}(\ln(1/\varepsilon))$ steps (see Appendix D) and, consequently, the output of Algorithm 1 (which can be expressed as $\widehat{\mathbf{x}}^* = \mathbf{A}^\mathsf{T}\bar{\mathbf{y}}^{(t)}$) also converges to $\mathbf{x}^* = \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}$, the true solution of the ridge regression problem. Our analysis offers several advantages over preconditioned Richardson iteration. In our case, $\mathbf{P}^{-1}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)$ is not symmetric positive definite which, according to existing literature, implies that the convergence of Richardson's method is monotone in terms of the energy-norm induced by $\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n$, but not the Euclidean norm (see eqn. (2.4.17) in Quarteroni & Valli (1994)). Additionally, standard convergence analysis of the Richardson iteration is with respect to $\bar{\mathbf{y}}^{(t)}$, whereas our vector of interest is $\widehat{\mathbf{x}}^*$ (which is $\bar{\mathbf{y}}^{(t)}$ premultiplied by $\mathbf{A}^\mathsf{T}$). The equality $\|\bar{\mathbf{y}}^{(t)} - \mathbf{y}^*\|_2 = \|\widehat{\mathbf{x}}^* - \mathbf{x}^*\|_2$ holds if $\mathbf{A}$ has orthonormal rows, which is not true in general.

We now discuss the running time of Algorithm 1. First, we need to compute $\mathbf{A}\widetilde{\mathbf{x}}^{(j-1)}$ which takes time $\mathcal{O}(\mathrm{nnz}(\mathbf{A}))$. Next, computing the sketch $\mathbf{AS} \in \mathbb{R}^{n\times s}$ takes $T(\mathbf{A},\mathbf{S})$ time and depends on the particular construction of $\mathbf{S}$ (see Section 2.1). Then, in order to invert the matrix $\mathbf{\Theta} = \mathbf{AS}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n$ it suffices to compute the SVD of the matrix $\mathbf{AS}$. Notice that given the singular values of $\mathbf{AS}$ we can compute the singular values of $\mathbf{\Theta}$; also note that the left and right singular vectors of $\mathbf{\Theta}$ are the same as the left singular vectors of $\mathbf{AS}$. Interestingly, we do not need to compute $\mathbf{\Theta}^{-1}$: we can store it implicitly by storing its left (and right) singular vectors $\mathbf{U}_{\mathbf{\Theta}}$ and its singular values $\mathbf{\Sigma}_{\mathbf{\Theta}}$. Then, we can compute all necessary matrix-vector products using this implicit representation of $\mathbf{\Theta}^{-1}$. Thus, inverting $\mathbf{\Theta}$ takes $\mathcal{O}(sn^2)$ time. Updating the vectors $\mathbf{b}^{(j)}$, $\mathbf{y}^{(j)}$, and $\widetilde{\mathbf{x}}^{(j)}$ is dominated by the aforementioned running times, as all updates amount to just matrix-vector products. Thus, summing over all t iterations, the running time of Algorithm 1 is given by

$$\mathcal{O}(t \cdot \mathrm{nnz}(\mathbf{A})) + \mathcal{O}(sn^2) + T(\mathbf{A},\mathbf{S}). \tag{10}$$

We conclude this section by noting that our results remain valid when *different* sampling matrices $\mathbf{S}_j$ are used in each iteration $j = 1, \ldots, t$, as long as they satisfy the constraints of eqns. (6) or (8). As a matter of fact, the sketching matrices $\mathbf{S}_j$ do not even need to have the same number of columns. See Section 5 for an interesting open problem in this setting.

2.1. Satisfying the Conditions of Eqns. (6) or (8)

The conditions of eqns. (6) and (8) essentially boil down to randomized, approximate matrix multiplication (Drineas & Kannan, 2001; Drineas et al., 2006a), a task that has received much attention in the RLA community. We start by discussing *sketching-based* approaches: a particularly useful result for our purposes appeared in Cohen et al. (2016). Using our notation, Cohen et al. (2016) proved that for $\mathbf{X} \in \mathbb{R}^{d\times n}$ and for a (suitably constructed) sketching matrix $\mathbf{S} \in \mathbb{R}^{d\times s}$, with probability at least $1-\delta$,

$$\left\|\mathbf{X}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{X} - \mathbf{X}^\mathsf{T}\mathbf{X}\right\|_2 \leq \varepsilon \left(\|\mathbf{X}\|_2^2 + \frac{\|\mathbf{X}\|_F^2}{r}\right), \tag{11}$$

for any arbitrary $r \geq 1$. The above bound holds for a very broad family of constructions for the sketching matrix $\mathbf{S}$ (see Cohen et al. (2016) for details). In particular, Cohen et al. (2016) demonstrated a construction for $\mathbf{S}$ with $s = \mathcal{O}(r/\varepsilon^2)$ columns such that, for any $n \times d$ matrix $\mathbf{A}$, the product $\mathbf{AS}$ can be computed in time $\mathcal{O}(\mathrm{nnz}(\mathbf{A})) + \widetilde{\mathcal{O}}((r^3 + r^2 n)/\varepsilon^\gamma)$ for some constant γ. Thus, starting with eqn. (6) and using this particular construction for $\mathbf{S}$, let $\mathbf{X} = \mathbf{V}$ and note that $\|\mathbf{V}\|_F^2 = n$ and $\|\mathbf{V}\|_2 = 1$. Setting $r = n$, eqn. (11) implies that

$$\left\|\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V} - \mathbf{I}_n\right\|_2 \leq 2\,\varepsilon.$$

In this case, the running time of the sketch computation is equal to $T(\mathbf{A},\mathbf{S}) = \mathcal{O}(\mathrm{nnz}(\mathbf{A})) + \widetilde{\mathcal{O}}(n^3/\varepsilon^\gamma)$. The running time of the overall algorithm follows from eqn. (10) and our choices for s and r:

$$\mathcal{O}(t \cdot \mathrm{nnz}(\mathbf{A})) + \widetilde{\mathcal{O}}(n^3/\varepsilon^{\max\{2,\gamma\}}).$$

The failure probability (hidden in the polylogarithmic terms) can be easily controlled using a union bound. Finally, a simple change of variables (using $\varepsilon/4$ instead of ε) suffices to satisfy the structural condition of eqn. (6) without changing the above running time.

Similarly, starting with eqn. (8), let $\mathbf{X} = \mathbf{V}\mathbf{\Sigma}_\lambda$ and note that $\|\mathbf{V}\mathbf{\Sigma}_\lambda\|_F^2 = d_\lambda$ and $\|\mathbf{V}\mathbf{\Sigma}_\lambda\|_2 \leq 1$. Setting $r = d_\lambda$, eqn. (11) implies that $\left\|\mathbf{\Sigma}_\lambda\mathbf{V}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{V}\mathbf{\Sigma}_\lambda - \mathbf{\Sigma}_\lambda^2\right\|_2 \leq 2\varepsilon$. In this case, the running time of the sketch computation is equal to $T(\mathbf{A},\mathbf{S}) = \mathcal{O}(\mathrm{nnz}(\mathbf{A})) + \tilde{O}(d_\lambda^2 n/\varepsilon^\gamma)$. The running time of the overall algorithm follows from eqn. (10) and our choices for s and r:

$$\mathcal{O}(t \cdot \mathrm{nnz}(\mathbf{A})) + \widetilde{\mathcal{O}}(d_\lambda n^2/\varepsilon^{\max\{2,\gamma\}}).$$

Again, a change of variables suffices to satisfy the structural condition of eqn. (8) without changing the running time.

We now discuss how to satisfy the conditions of eqns. (6) or (8) by *sampling*, *i.e.*, by selecting a small number of predictor variables. Towards that end, consider Algorithm 2 for the construction of the sampling-and-rescaling matrix $\mathbf{S}$.

The following theorem (see Appendix G for its proof) is of independent interest and is a strengthening of Theorem 4.2 of Holodnak & Ipsen (2015), since the sampling complexity s is improved to depend only on $\|\mathbf{X}\|_F^2$ instead of the stable rank of $\mathbf{X}$ for the special case where $\|\mathbf{X}\|_2 \leq 1$.[3]

[3] We do note that Theorem 3 is implicit in Cohen et al. (2017).

Algorithm 2 Construct sampling-and-rescaling matrix

Input: Probabilities p_i, $i = 1, \ldots, d$; integer $s \ll d$;
$\mathbf{S} \leftarrow \mathbf{0}_{d \times s}$;
for $j = 1$ **to** s **do**
  Pick $i_j \in \{1, \ldots, d\}$ with $\mathbb{P}(i_j = i) = p_i$;
  $\mathbf{S}_{i_j j} \leftarrow (s\, p_{i_j})^{-\frac{1}{2}}$;
end for
Output: Sampling-and-rescaling matrix $\mathbf{S}$;

Theorem 3. *Let $\mathbf{X} \in \mathbb{R}^{d \times n}$ with $\|\mathbf{X}\|_2 \leq 1$ and let $\mathbf{S}$ be constructed by Algorithm 2 with $p_i = \|\mathbf{X}_{i*}\|_2^2 / \|\mathbf{X}\|_F^2$ for $i = 1, \ldots, d$. Let δ be a failure probability and let $\varepsilon \in (0, 1]$ be an accuracy parameter. If the number of sampled columns s satisfies*

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

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

$$\left\|\mathbf{X}^\mathsf{T}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{X} - \mathbf{X}^\mathsf{T}\mathbf{X}\right\|_2 \leq \varepsilon.$$

Using Theorem 3 with $\mathbf{X} = \mathbf{V}$ we can satisfy the condition of eqn. (6) by simply using the sampling probabilities $p_i = \|\mathbf{V}_{i*}\|_2^2 / n$ (recall that $\|\mathbf{V}\|_F^2 = n$ and $\|\mathbf{V}\|_2 = 1$), which are the column *leverage scores* of the design matrix $\mathbf{A}$. Setting $s = \mathcal{O}(\varepsilon^{-2} n \ln n)$ suffices to satisfy the condition of eqn. (6). We note that approximate leverage scores also suffice and that their computation can be done efficiently without computing $\mathbf{V}$ (Drineas et al., 2012).

Finally, using Theorem 3 with $\mathbf{X} = \mathbf{V}\mathbf{\Sigma}_\lambda$ we can satisfy the condition of eqn. (8) 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 proportional to the column *ridge leverage scores* of the design matrix $\mathbf{A}$ (see Lemma 21 in Appendix F). Setting $s = \mathcal{O}(\varepsilon^{-2} d_\lambda \ln d_\lambda)$ suffices to satisfy the condition of eqn. (8). We note that approximate ridge leverage scores also suffice and that their computation can be done efficiently without computing $\mathbf{V}$ (Cohen et al., 2017).

2.2. Bounding the MSE of $\widehat{\mathbf{x}}^*$

Consider the data-generation model

$$\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \boldsymbol{\varepsilon}, \tag{12}$$

where $\mathbf{b} \in \mathbb{R}^n$ is the response vector, $\mathbf{A} \in \mathbb{R}^{n \times d}$ is the design matrix, $\mathbf{x}_0 \in \mathbb{R}^n$ is the "true" parameter vector, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the noise satisfying $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\mathsf{T}) = \sigma^2 \mathbf{I}_n$, $\sigma > 0$. Then, the ridge regression estimator $\mathbf{x}^*$ of the parameter vector $\mathbf{x}_0$ can be expressed as in eqn. (3), with mean squared error (MSE) given by (see Lemma 16 in Appendix E for the derivation)

$$\mathrm{MSE}(\mathbf{x}^*) = \sigma^2 \left\|(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{A}\right\|_F^2 + \left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2^2. \tag{13}$$

Similarly, we can prove that the MSE of $\widehat{\mathbf{x}}^*$ for the special case where $t = 1$ in Algorithm 1 is equal to

$$\mathrm{MSE}(\widehat{\mathbf{x}}^*) = \sigma^2 \left\|(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{A}\right\|_F^2 + \left\|\left(\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{A} - \mathbf{I}_d\right)\mathbf{x}_0\right\|_2^2. \tag{14}$$

We present bounds on the MSE of $\widehat{\mathbf{x}}^*$ for the special case where Algorithm 1 is run for a single iteration ($t = 1$) under the assumptions of eqns. (6) or (8). Bounds for $t > 1$ (more than one iteration) are delegated to future work.

Theorem 4. *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be the design matrix and let $\widehat{\mathbf{x}}^*$ be the output of Algorithm 1 for $t = 1$. If the condition of eqn. (6) is satisfied for some constant $0 < \varepsilon < 1$, then,*

$$\mathrm{MSE}(\widehat{\mathbf{x}}^*) \leq (1 + 3\varepsilon\gamma_1^2)\,\mathrm{MSE}(\mathbf{x}^*),$$

where $\gamma_1 = 1 + \frac{\sigma_1^2}{\lambda}$. If the condition of eqn. (8) is satisfied for some constant $0 < \varepsilon < 1$, then,

$$\mathrm{MSE}(\widehat{\mathbf{x}}^*) \leq (1 + 3\varepsilon\gamma_2^2)\,\mathrm{MSE}(\mathbf{x}^*),$$

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

3. Sketching the Proof of Theorem 2

Due to space considerations, essentially all our proofs have been deferred to the Appendix. However, to give a flavor of the mathematical derivations underlying our contributions, we present an outline of the proof of Theorem 2, starting with the special case where Algorithm 1 is run for a single iteration ($t = 1$).

Using the quantities defined in Algorithm 1, let

$$\mathbf{x}^{*(j)} = \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \lambda \mathbf{I}_n)^{-1}\mathbf{b}^{(j)} \tag{15}$$

for $j = 1, \ldots, t$. Notice that $\mathbf{x}^* = \mathbf{x}^{*(1)}$. Our next result expresses the intermediate vectors $\widetilde{\mathbf{x}}^{(j)}$ of Algorithm 1 in terms of the vectors $\mathbf{x}^{*(j)}$. We remind the reader that $\mathbf{U} \in \mathbb{R}^{n \times n}$ and $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$ are, respectively, the matrices of the left singular vectors and singular values of $\mathbf{A}$. We will make extensive use of the matrix $\mathbf{\Sigma}_\lambda$ defined in eqn. (7).

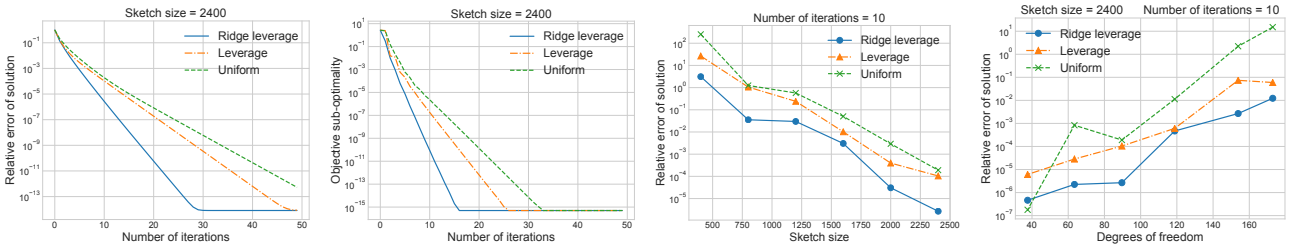
Lemma 5. *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{b} \in \mathbb{R}^n$, and $\lambda > 0$ be the inputs of the ridge regression problem. Let $\mathbf{S} \in \mathbb{R}^{d \times s}$ be the sketching matrix and define*

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

If $\|\mathbf{E}\|_2 < 1$, then for all $j = 1, \ldots, t$,

$$\widetilde{\mathbf{x}}^{(j)} = \mathbf{x}^{*(j)} + \mathbf{V}\mathbf{\Sigma}_\lambda \mathbf{R}\mathbf{\Sigma}_\lambda \mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}^{(j)}, \tag{16}$$

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



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

Figure 1. Experiment results on real data (errors are on log-scale).

Now, consider the case when $t = 1$. Algorithm 1 returns $\widehat{\mathbf{x}}^* = \widetilde{\mathbf{x}}^{(1)}$; also recall that $\mathbf{x}^* = \mathbf{x}^{*(1)}$ and $\mathbf{b} = \mathbf{b}^{(1)}$. Therefore, applying Lemma 5 yields

$$\widehat{\mathbf{x}}^* = \mathbf{x}^* + \mathbf{V}\mathbf{\Sigma}_\lambda \mathbf{R} \mathbf{\Sigma}_\lambda \mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}. \tag{17}$$

Further, for any $j = 1, \ldots, t$,

$$\begin{aligned}\|\mathbf{R}\|_2 = \Big\| \sum_{\ell=1}^{\infty} (-1)^\ell \mathbf{E}^\ell \Big\|_2 \le \sum_{\ell=1}^{\infty} \|\mathbf{E}^\ell\|_2 \le \sum_{\ell=1}^{\infty} \|\mathbf{E}\|_2^\ell \\ \le \sum_{\ell=1}^{\infty} \left(\frac{\varepsilon}{4\sqrt{2}}\right)^\ell = \frac{\frac{\varepsilon}{4\sqrt{2}}}{1 - \frac{\varepsilon}{4\sqrt{2}}} \le \frac{\varepsilon}{2\sqrt{2}}.\end{aligned} \tag{18}$$

where we used the triangle inequality, sub-multiplicativity of the spectral norm, and the fact that $\frac{\varepsilon}{4\sqrt{2}} \le \frac{1}{2}$. Now, using eqn. (17), we have

$$\begin{aligned}\|\widehat{\mathbf{x}}^* - \mathbf{x}^*\|_2 &= \|\mathbf{V}\mathbf{\Sigma}_\lambda \mathbf{R} \mathbf{\Sigma}_\lambda \mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}\|_2 \\ &\le \|\mathbf{\Sigma}_\lambda\|_2 \|\mathbf{R}\|_2 \|\mathbf{\Sigma}_\lambda \mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}\|_2 \\ &\le \frac{\varepsilon}{2\sqrt{2}} \|\mathbf{\Sigma}_\lambda \mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}\|_2 \\ &= \frac{\varepsilon}{2\sqrt{2}} \|\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma}_\lambda^2 \mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}\|_2.\end{aligned} \tag{19}$$

where the first inequality follows from the unitary invariance and sub-multiplicativity of the spectral norm, and the second inequality is due to eqn. (18) and the fact that $\|\mathbf{\Sigma}_\lambda\|_2 \le 1$.

Now, let $(\mathbf{\Sigma}_\lambda^{-1})_k$ denote the diagonal matrix whose first k diagonal entries are equal to the first k diagonal entries of $\mathbf{\Sigma}_\lambda^{-1}$ and the bottom $n-k$ diagonal entries are set to zero. Let $(\mathbf{\Sigma}_\lambda^{-1})_{k,\perp} = \mathbf{\Sigma}_\lambda^{-1} - (\mathbf{\Sigma}_\lambda^{-1})_k$. Then, we have

$$\begin{aligned}\|\mathbf{\Sigma}_\lambda^{-1}\mathbf{\Sigma}_\lambda^2 \mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}\|_2 \le \underbrace{\|(\mathbf{\Sigma}_\lambda^{-1})_k \, \mathbf{\Sigma}_\lambda^2 \mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}\|_2}_{\Delta_1} \\ + \underbrace{\|(\mathbf{\Sigma}_\lambda^{-1})_{k,\perp} \, \mathbf{\Sigma}_\lambda^2 \mathbf{\Sigma}^{-1}\mathbf{U}^\mathsf{T}\mathbf{b}\|_2}_{\Delta_2}.\end{aligned} \tag{20}$$

where eqn. (20) follows from the triangle inequality and the fact that $\mathbf{\Sigma}_\lambda^{-1} = (\mathbf{\Sigma}_\lambda^{-1})_k + (\mathbf{\Sigma}_\lambda^{-1})_{k,\perp}$.

Next, we bound Δ_1 and Δ_2 separately using eqns. (60) and (62) in Appendix C:

$$\Delta_1 \le \sqrt{2}\,\|\mathbf{x}^*\|_2, \quad \Delta_2 \le \frac{1}{\sqrt{\lambda}} \left\|\mathbf{U}_{k,\perp}^\mathsf{T}\mathbf{b}\right\|_2. \tag{21}$$

Finally, combining eqns. (19), (20) and (21), we obtain

$$\begin{aligned}\|\widehat{\mathbf{x}}^* - \mathbf{x}^*\|_2 &\le \frac{\varepsilon}{2\sqrt{2}} \left(\sqrt{2}\,\|\mathbf{x}^*\|_2 + \frac{1}{\sqrt{\lambda}} \|\mathbf{U}_{k,\perp}^\mathsf{T}\mathbf{b}\|_2\right) \\ &= \frac{\varepsilon}{2} \left(\|\mathbf{x}^*\|_2 + \frac{1}{\sqrt{2\lambda}} \|\mathbf{U}_{k,\perp}^\mathsf{T}\mathbf{b}\|_2\right),\end{aligned} \tag{22}$$

which concludes the proof for the $t = 1$ case.

Interestingly, the eqn. (22) holds more generally and can be used to bound the distance between the intermediate approximate solution vectors $\widetilde{\mathbf{x}}^{(j)}$ and the intermediate true solution vectors $\mathbf{x}^{*(j)}$ of eqn. (15). Indeed, for $j = 1, \ldots, t$, we have

$$\|\widetilde{\mathbf{x}}^{(j)} - \mathbf{x}^{*(j)}\|_2 \le \frac{\varepsilon}{2} \left(\|\mathbf{x}^{*(j)}\|_2 + \frac{1}{\sqrt{2\lambda}} \|\mathbf{U}_{k,\perp}^\mathsf{T}\mathbf{b}^{(j)}\|_2\right). \tag{23}$$

The next lemma (see Appendix C for its proof) presents a structural result for the optimal solution $\mathbf{x}^*$.

Lemma 6. *Let $\widetilde{\mathbf{x}}^{(j)}$, $j = 1, \ldots, t$ be the sequence of vectors introduced in Algorithm 1 and let $\mathbf{x}^{*(t)} \in \mathbb{R}^d$ be defined as in eqn. (15). Then,*

$$\mathbf{x}^* = \mathbf{x}^{*(t)} + \sum_{j=1}^{t-1} \widetilde{\mathbf{x}}^{(j)}, \tag{24}$$

where $\mathbf{x}^*$ is the true solution of the ridge regression problem.

Repeated application of eqns. (23) and (24) yields

$$\begin{aligned}\|\widehat{\mathbf{x}}^* - \mathbf{x}^*\|_2 &= \Big\| \sum_{j=1}^{t} \widetilde{\mathbf{x}}^{(j)} - \mathbf{x}^* \Big\|_2 \\ &= \Big\|\widetilde{\mathbf{x}}^{(t)} - \Big(\mathbf{x}^* - \sum_{j=1}^{t-1} \widetilde{\mathbf{x}}^{(j)}\Big)\Big\|_2 = \left\|\widetilde{\mathbf{x}}^{(t)} - \mathbf{x}^{*(t)}\right\|_2 \\ &\le \frac{\varepsilon}{2} \left(\|\mathbf{x}^{*(t)}\|_2 + \frac{1}{\sqrt{2\lambda}} \|\mathbf{U}_{k,\perp}^\mathsf{T}\mathbf{b}^{(t)}\|_2\right).\end{aligned} \tag{25}$$

The next bound (see Appendix C for its proof) provides a critical inequality that can be used recursively in order to establish Theorem 2.

Lemma 7. *Let $\mathbf{b}^{(j)}$, $j = 1, \ldots, t$, be the intermediate response vectors of Algorithm 1 and let $\mathbf{x}^{*(j)}$ be the vector defined in eqn. (15) for $j = 1, \ldots, t-1$. If the structural condition of eqn. (8) is satisfied, then*

$$\|\mathbf{x}^{*(j+1)}\|_2 + \frac{1}{\sqrt{2\lambda}}\|\mathbf{U}_{k,\perp}^{\mathsf{T}}\mathbf{b}^{(j+1)}\|_2 \leq \varepsilon\left(\|\mathbf{x}^{*(j)}\|_2 + \frac{1}{\sqrt{2\lambda}}\|\mathbf{U}_{k,\perp}^{\mathsf{T}}\mathbf{b}^{(j)}\|_2\right). \quad (26)$$

Applying eqn. (26) iteratively, we obtain

$$\begin{aligned}\|\mathbf{x}^{*(t)}\|_2 + \frac{1}{\sqrt{2\lambda}}\|\mathbf{U}_{k,\perp}^{\mathsf{T}}\mathbf{b}^{(t)}\|_2 &\leq \varepsilon\left(\|\mathbf{x}^{*(t-1)}\|_2 + \frac{1}{\sqrt{2\lambda}}\|\mathbf{U}_{k,\perp}^{\mathsf{T}}\mathbf{b}^{(t-1)}\|_2\right) \\ &\leq \cdots \leq \varepsilon^{t-1}\left(\|\mathbf{x}^{*}\|_2 + \frac{1}{\sqrt{2\lambda}}\|\mathbf{U}_{k,\perp}^{\mathsf{T}}\mathbf{b}\|_2\right). \quad (27)\end{aligned}$$

Finally, combining eqns. (25) and (27), we conclude that

$$\|\widehat{\mathbf{x}}^{*} - \mathbf{x}^{*}\|_2 \leq \frac{\varepsilon^t}{2}\left(\|\mathbf{x}^{*}\|_2 + \frac{1}{\sqrt{2\lambda}}\|\mathbf{U}_{k,\perp}^{\mathsf{T}}\mathbf{b}\|_2\right). \quad (28)$$

4. Empirical Evaluation

We perform experiments on the ARCENE dataset (Guyon et al., 2005) from the UCI repository (Lichman, 2013). The design matrix contains 200 samples with $10,000$ real-valued features; we normalize the entries to be within the interval $[0, 1]$. The response vector consists of ± 1 labels. We also perform experiments on synthetic data generated as in Chen et al. (2015); see Appendix H for details.

In our experiments, we compare three different choices of sampling probabilities: selecting columns (*i*) uniformly at random, (*ii*) proportional to their leverage scores, or (*iii*) proportional to their ridge leverage scores. For each sampling method, we run Algorithm 1 for 50 iterations with a variety of sketch sizes, and measure (*i*) the relative error of the solution vector $\frac{\|\widehat{\mathbf{x}}^{*}-\mathbf{x}^{*}\|_2}{\|\mathbf{x}^{*}\|_2}$, where $\mathbf{x}^{*}$ is the true optimal solution and (*ii*) the objective sub-optimality $\frac{f(\widehat{\mathbf{x}}^{*})}{f(\mathbf{x}^{*})} - 1$, where $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_2^2$ is the objective function for the ridge-regression problem.

The results are shown in Figure 1. Figures 1a and 1b plot the relative error of the solution vector and the objective sub-optimality (for a fixed sketch size) as the iterative algorithm progresses. Figure 1c plots the relative error of the solution with respect to varying sketch sizes (the plots for objective sub-optimality are analogous and thus omitted). We observe that both the solution error and the objective sub-optimality decay *exponentially* as our iterative algorithm progresses.[4]

[4]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.

Next, we show that the approximation quality depends directly on the *degrees of freedom* d_λ of the ridge-regression problem (eqn. (4)), rather than the dimensions of the design matrix. To this end, we keep the design matrix unchanged (n remains fixed), and vary the regularization parameter $\lambda \in \{1, 2, 5, 10, 20, 50\}$. Figure 1d plots the relative solution error against the degrees of freedom d_λ (for a fixed sketch size and number of iterations); we observe that the relative error decreases roughly 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.

5. Conclusion and Open Problems

We have presented simple structural results that guarantee high-quality approximations to the optimal solution vector of ridge regression. In particular, our second structural result presents the first accuracy analysis for ridge regression when the ridge leverage scores are used to sample predictor variables. The sample size depends on the degrees of freedom of the ridge regression problem and not the dimensions of the design matrix. An obvious open problem is to either improve the sample size or present lower bounds showing that our bounds are tight. Additionally, the results of Theorem 4 should be generalized to cover the $t > 1$ case.

Finally, an interesting open problem would be to investigate whether the use of different sampling matrices in each iteration of Algorithm 1 (*i.e.*, introducing new "randomness" in each iteration) could lead to *provably* improved bounds for our main theorems. We conjecture that this is indeed the case, and we present further experiment results in Appendix H which support our conjecture. In particular, the results show that using a newly sampled sketching matrix at every iteration enables faster convergence as the iterations progress, and also reduces the minimum sketch size necessary for Algorithm 1 to converge.

Acknowledgements. We thank an anonymous reviewer for pointing out the connection between our method and the preconditioned Richardson iteration. AC and PD were partially supported by NSF IIS-1661760 and IIS-1661756. JY was supported by NSF IIS-1149789 and IIS-1618690.

References

Ailon, N. and Chazelle, B. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

Alaoui, A. E. and Mahoney, M. W. Fast randomized kernel ridge regression with statistical guarantees. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pp. 775–783, 2015.

Avron, H., Clarkson, K. L., and Woodruff, D. P. Sharper bounds for regularized data fitting. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pp. 27:1–27:22, 2017a.

Avron, H., Clarkson, K. L., and Woodruff, D. P. Faster kernel ridge regression using sketching and preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 38 (4):1116–1138, 2017b.

Calandriello, D., Lazaric, A., and Valko, M. Second-order kernel online convex optimization with adaptive sketching. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 645–653, 2017a.

Calandriello, D., Lazaric, A., and Valko, M. Efficient second-order online kernel learning with adaptive embedding. In *Advances in Neural Information Processing Systems 30*, pp. 6142–6151. 2017b.

Calandriello, D., Lazaric, A., and Valko, M. Distributed adaptive sampling for kernel matrix approximation. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1421–1429, 2017c.

Chen, S., Liu, Y., Lyu, M. R., King, I., and Zhang, S. Fast relative-error approximation algorithm for ridge regression. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pp. 201–210, 2015.

Clarkson, K. L. and Woodruff, D. P. Low rank approximation and regression in input sparsity time. In *Proceedings of the 45th annual ACM symposium on Symposium on Theory of Computing*, pp. 81, 2013.

Cohen, M. B., Nelson, J., and Woodruff, D. P. Optimal approximate matrix product in terms of stable rank. In *43rd International Colloquium on Automata, Languages, and Programming*, pp. 11:1–11:14, 2016.

Cohen, M. B., Musco, C., and Musco, C. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1758–1777, 2017.

Cutajar, K., Osborne, M. A., Cunningham, J. P., and Filippone, M. Preconditioning kernel matrices. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, pp. 2529–2538, 2016.

Drineas, P. and Kannan, R. Fast monte-carlo algorithms for approximate matrix multiplication. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science*, pp. 452–459, 2001.

Drineas, P. and Mahoney, M. W. RandNLA: Randomized Numerical Linear Algebra. *Communications of the ACM*, 59(6):80–90, 2016.

Drineas, P., Kannan, R., and Mahoney, M. W. Fast monte carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1):132–157, 2006a.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1127–1136, 2006b.

Drineas, P., Mahoney, M. W., Muthukrishnan, S., and Sarlós, T. Faster least squares approximation. *Numerische Mathematik*, 117:219–249, 2011.

Drineas, P., Magdon-Ismail, M., Mahoney, M. W., and Woodruff, D. P. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(1):3475–3506, 2012.

Golub, G. H. and Van Loan, C. F. *Matrix Computations*. Johns Hopkins University Press, 1996.

Gonen, A., Orabona, F., and Shalev-Shwartz, S. Solving ridge regression using sketched preconditioned SVRG. In *Proceedings of the 33nd International Conference on Machine Learning*, pp. 1397–1405, 2016.

Gunst, R. F. and Mason, R. L. Biased estimation in regression: An evaluation using mean squared error. *Journal of the American Statistical Association*, 72(359):616–628, 1977.

Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G. Result Analysis of the NIPS 2003 Feature Selection Challenge. In *Advances in Neural Information Processing Systems 17*, pp. 545–552. 2005.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

Holodnak, J. T. and Ipsen, I. C. F. Randomized approximation of the gram matrix: Exact computation and probabilistic bounds. *SIAM Journal on Matrix Analysis and Applications*, 36(1):110–137, 2015.

Kyng, R. *Approximate Gaussian Elimination*. Ph.D Thesis, Yale University, 2017.

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Lu, Y., Dhillon, P., Foster, D. P., and Ungar, L. Faster ridge regression via the subsampled randomized hadamard transform. In *Advances in Neural Information Processing Systems 26*, pp. 369–377. 2013.

Ma, S. and Belkin, M. Diving into the shallows: a computational perspective on large-scale shallow learning. In *Advances in Neural Information Processing Systems 30*, pp. 3778–3787. 2017.

Mahoney, M. W. *Randomized algorithms for matrices and data*. Foundations and Trends in Machine Learning. 2011.

Mahoney, M. W. and Drineas, P. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 106(3), 2009.

Musco, C. and Musco, C. Recursive sampling for the nystrom method. In *Advances in Neural Information Processing Systems 30*, pp. 3836–3848. 2017.

Pilanci, M. and Wainwright, M. J. Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, 17 (53):1–38, 2016.

Quarteroni, A. M. and Valli, A. *Numerical Approximation of Partial Differential Equations*. Springer, 1994.

Rudi, A., Carratino, L., and Rosasco, L. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems 30*, pp. 3888–3898. 2017.

Sarlós, T. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pp. 143–152, 2006.

Saunders, C., Gammerman, A., and Vovk, V. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 515–521, 1998.

Tropp, J. A. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

Wang, S., Gittens, A., and Mahoney, M. W. Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3608–3616, 2017.

Woodruff, D. P. Sketching as a Tool for Numerical Linear Algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

Zhan, X. Singular values of differences of positive semidefinite matrices. *SIAM Journal on Matrix Analysis and Applications*, 22(3):819–823, 2001.

Zhang, Y., Duchi, J., and Wainwright, M. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

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

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

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 $\mathbf{P}^{-1} = (\mathbf{A}\mathbf{S}\mathbf{S}^{\mathsf{T}}\mathbf{A} + \lambda\mathbf{I}_n)^{-1}$ 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:*

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

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

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*

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

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 $\mathrm{Var}(\mathbf{b})$. 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

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

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

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

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

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

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

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

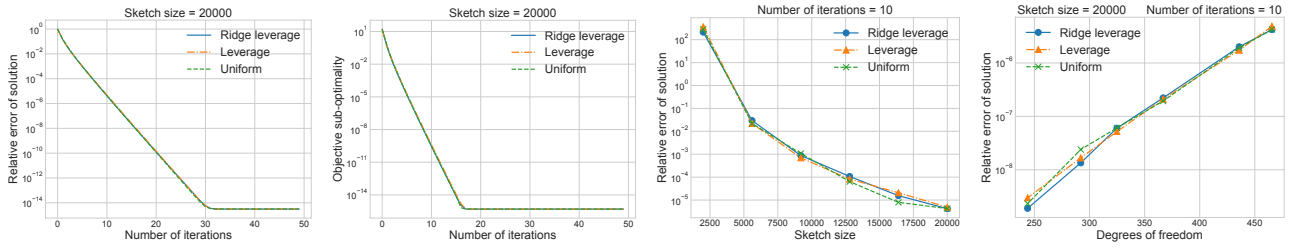
$$\ln\left(\frac{4(1+\|\mathbf{X}\|_F^2)}{\delta}\right) \geq 1 \iff \delta \leq \frac{4(1+\|\mathbf{X}\|_F^2)}{e},$$

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

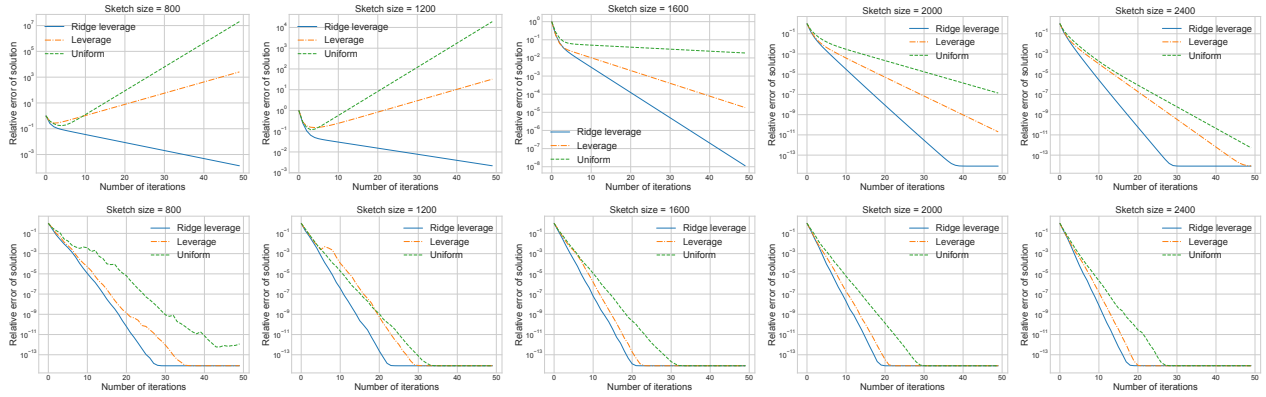
for objective sub-optimality are analogous and are thus omitted). We observe that both the solution error and the objective sub-optimality decay *exponentially* as our iterative algorithm progresses.[5]

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.