Optimal Subsampling for High-dimensional Ridge Regression

Hanyu Li*, Chengmei Niu

College of Mathematics and Statistics, Chongqing University, Chongqing 401331, P.R. China

Abstract

We investigate the feature compression of high-dimensional ridge regression using the optimal subsampling technique. Specifically, based on the basic framework of random sampling algorithm on feature for ridge regression and the A-optimal design criterion, we first obtain a set of optimal subsampling probabilities. Considering that the obtained probabilities are uneconomical, we then propose the nearly optimal ones. With these probabilities, a two step iterative algorithm is established which has lower computational cost and higher accuracy. We provide theoretical analysis and numerical experiments to support the proposed methods. Numerical results demonstrate the decent performance of our methods.

Keywords: High-dimensional ridge regression, Optimal subsampling, A-optimal design criterion, Two step iterative algorithm

MSC: 62J07

1. Introduction

For the famous linear model

\[ y = A\beta + \nu, \]

where \( y \in \mathbb{R}^n \) is the response vector, \( A \in \mathbb{R}^{n \times p} \) is the design matrix, \( \beta \in \mathbb{R}^p \) is the parameter vector, and \( \nu \in \mathbb{R}^n \) is the standardized Gaussian noise vector, ridge regression [1], also known as...
the least squares regression with Tikhonov regularisation, has the following form

$$\min_\beta \frac{1}{2} \|y - A\beta\|^2_2 + \frac{\lambda}{2} \|\beta\|^2_2,$$  \hspace{0.5cm} (1.1)

where \(\lambda\) is the regularized parameter, and the corresponding estimator is

$$\hat{\beta}_{rls} = (A^T A + \lambda I)^{-1} A^T y.$$ 

In this paper, we focus only on the case \(p > n\), i.e., the high dimensional ridge regression. For this case, the dominant computational cost of the above estimator is from the matrix inversion which takes \(O(p^3)\) flops. A straightforward way of amelioration is to solve the problem (1.1) in the dual space. Specifically, we first solve the dual problem of (1.1),

$$\min_z \frac{1}{2\lambda} \|A^T z\|^2_2 + \frac{1}{2} \|z\|^2_2 - z^T y,$$  \hspace{0.5cm} (1.2)

and the solution is

$$\hat{z}^* = \lambda(AA^T + \lambda I)^{-1} y.$$  \hspace{0.5cm} (1.3)

Then, setting

$$\hat{\beta}_{rls} = \frac{A^T \hat{z}^*}{\lambda}$$  \hspace{0.5cm} (1.4)

gives the estimator of (1.1) in an alternative form

$$\hat{\beta}_{rls} = A^T (AA^T + \lambda I)^{-1} y.$$  \hspace{0.5cm} (1.5)

More details can be found in [3]. Now, the dominant computational cost is \(O(n^2 p)\) which appears in the computation of \(AA^T\). However, it is still prohibitive when \(p \gg n\).

To reduce the computational cost, some scholars considered the randomized sketching technique [4–9]. The main idea is to compress the design matrix \(A\) to be a small one \(\hat{A}\) by post-multiplying it by a random matrix \(S \in \mathbb{R}^{p \times r}\) with \(r \ll p\), i.e., \(\hat{A} = AS\), and hence the reduced regression can be called the compressed ridge regression. There are two most common ways to generate \(S\): random projection and random sampling. The former can be the (sub)Gaussian matrix [6, 7, 9], the sub-sampled randomized Hadamard transform (SRHT) [4–7], the sub-sampled randomized Fourier transform [7], and the CountSketch (also called the sparse embedding matrix) [6], and the latter can be the uniform sampling and the importance sampling [8].
Specifically, building on (1.3) and (1.4), Lu et al. [4] presented the following estimator

\[ \hat{\beta}_L = \frac{S S^T A^T \tilde{z}_L}{\lambda}, \]

where \( S \) is the SRHT and

\[ \tilde{z}_L = \lambda (A S S^T A^T + \lambda I)^{-1} y \]  

(1.6)

is the solution to the dual problem of the following compressed ridge regression

\[ \min_{\beta_H} \frac{1}{2} \| y - A S \beta_H \|_2^2 + \frac{\lambda}{2} \| \beta_H \|_2^2, \]  

(1.7)

and obtained a risk bound. Soon afterwards, for \( S \) generated by the product of sparse embedding matrix and SRHT, Chen et al. [5] developed an estimator as follows:

\[ \hat{\beta}_C = A^T (A S)^\dagger (\lambda (A S)^\dagger + A S)^\dagger y, \]  

(1.8)

where \( \dagger \) denotes the Moore-Penrose inverse, and provided an estimation error bound and a risk bound. Later, Avron et al. [6] proposed the estimator \( \hat{\beta}_A = A^T \hat{b} \), where

\[ \hat{b} = \arg\min_b \frac{1}{2} \| A S S^T A^T b \|_2^2 - y^T A A^T b + \frac{1}{2} \| y \|_2^2 + \frac{\lambda}{2} \| S^T A^T b \|_2^2 \]

with \( S \) being the CountSketch, SRHT, or Gaussian matrix. The above problem is the sketch of the following regression problem

\[ \min_b \frac{1}{2} \| A A^T b \|_2^2 - y^T A A^T b + \frac{1}{2} \| y \|_2^2 + \frac{\lambda}{2} \| A^T b \|_2^2, \]

which is transformed from (1.1). Additionally, Wang et al. [7] and Lacotte and Pilanci [9] applied the dual random projection proposed in [10, 11] to the high-dimensional ridge regression. By the way, there are some works on compressed least squares regression [12–19], which can be written in the following form

\[ \hat{\alpha}_{ls} = \arg\min_\alpha \frac{1}{2} \| y - A S \alpha \|_2^2, \]  

(1.9)

where \( S \) is typically the (sub)Gaussian matrix.

To the best of our knowledge, there is few work of applying random sampling to high-dimensional ridge regression. We only found a work of [8], which proposed an iterative algorithm by using the random sampling with the column leverage scores or ridge leverage scores as the sampling
probabilities. This algorithm can be viewed as an extension of the method in [6]. However, there are some works on compressed least squares regression via random sampling. As far as we know, Drineas et al. [20] first applied the random sampling with column leverage scores or approximated ones as the sampling probabilities to the least squares regression and established the following estimator

$$\hat{\beta}_D = A^T (AS)^\dagger T (AS)^\dagger y,$$

which can be regarded as a special case of (1.8). Later, Slawski [18] investigated (1.9) using uniform sampling, and discussed the predictive performance.

In this paper, we will consider the application of random sampling on high-dimensional ridge regression further. Inspired by the technique of the optimal subsampling used in e.g., [21–26], we will mainly investigate the optimal subsampling probabilities for compressed ridge regression. The nearly optimal subsampling probabilities and a two step iterative algorithm are also derived.

The remainder of this paper is organized as follows. The basic framework of random sampling algorithm and the optimal subsampling probabilities are presented in Section 2. In Section 3, we propose the nearly optimal subsampling probabilities and a two step iterative algorithm. The detailed theoretical analyses of the proposed methods are also presented in Sections 2 and 3, respectively. In Section 4, we provide some numerical experiments to test our methods. The proofs of all the main theorems are given in the appendix.

Before moving to the next section, we introduce some standard notations used in this paper.

For the matrix $A \in \mathbb{R}^{n \times p}$, $A_i$, $A^j$, $\|A\|_2$ and $\|A\|_F$ denote its $i$-th column, $j$-th row, spectral norm and Frobenius norm, respectively. Also, its thin SVD is given as $A = U \Sigma V^T$, where $U \in \mathbb{R}^{n \times \rho}$, $V \in \mathbb{R}^{p \times \rho}$, and $\Sigma \in \mathbb{R}^{\rho \times \rho}$ with the diagonal elements, i.e., the singular values of $A$, satisfying $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{\rho}(A) > 0$.

For $V$, its row norms $\|V_i\|_2$ with $i = 1, \cdots, p$ are the column leverage scores [8], and for $X = V \Sigma \Lambda$, where $\Sigma \Lambda$ is a diagonal matrix with the diagonal entries being $\sqrt{\frac{\sigma_j(A)^2}{\lambda_+ \sigma_j(A)}}$ ($j = 1, \cdots, \rho$), its row norms $\|X_i\|_2$ are called the ridge leverage scores [8].

In addition, $O_p(1)$ denotes that a sequence of random variables are bounded in probability and $o_p(1)$ represents that the sequence converges to zero in probability. More details can refer to [27, Chap. 2]. In our case, we also use $O_p(\mathcal{F}_n)$ to denote that a sequence of random variables are bounded in conditional probability given the full data matrix $\mathcal{F}_n = (A, y)$. Especially, for any matrix $G$,
\( G = O_p(1) \) \( (G = O_p(1)) \) means that all the elements of \( G \) are bounded in probability \( (\text{given } F_n) \), and \( G = o_p(1) \) symbolizes that its elements are convergence to zero in probability.

2. Optimal Subsampling

In this section, we will present the basic framework of random sampling algorithm, propose the optimal subsampling probabilities, and obtain the corresponding error analysis.

2.1. Algorithm and Optimal Subsampling Probabilities

Given a set of probabilities, i.e., the random sampling matrix \( S \), our approximate estimator
\[
\hat{\beta} = A^T (ASS^TA^T + \lambda I)^{-1} y
\]
(2.1)
of the high-dimensional ridge regression (1.1) is the combination of the solution to the compressed dual problem,
\[
\arg\min_z \frac{1}{2\lambda} \|S^TA^Tz\|^2_2 + \frac{1}{2} \|z\|^2_2 - z^Ty,
\]
(2.2)
i.e.,
\[
\hat{z} = \lambda (ASS^TA^T + \lambda I)^{-1} y,
\]
(2.3)
and (1.4). That is, we first solve the problem (2.2) and then get the approximate estimator through (1.4). The detailed process, i.e., the basic framework of random sampling algorithm, is listed in Algorithm 1.

Remark 2.1. In Algorithm 1, the parameter \( \lambda \) can be determined by \( K \)-fold cross-validation, leave-one-out cross-validation, or generalized cross-validation, see e.g. [28]. Since the main focus of this paper is the performance of subsampling on high-dimensional ridge regression, we omit the investigation of the choice of \( \lambda \).

\(^1\)Note that this approach is different from the one in [4], though the expressions of \( \hat{z} \) in [28] and \( \tilde{z} \) in (1.6) are the same. In fact, the authors in [4] first solve the compressed ridge regression \( \text{(1.7)} \) in the dual space and then find the estimator of the compressed regression via \( \text{(1.4)} \). Finally, the approximate estimator of the original ridge regression is recovered by the random matrix \( S \).
Algorithm 1 Random Sampling Algorithm for High-dimensional Ridge Regression (RS HRR)

Input: \( y \in \mathbb{R}^n, A \in \mathbb{R}^{n \times p} \), the regularized parameter \( \lambda \), the sampling size \( r \) with \( r \ll p \), and the sampling probabilities \( \{ \pi_i \}_{i=1}^p \) with \( \pi_i \geq 0 \) such that \( \sum_{i=1}^p \pi_i = 1 \).

Output: the dual solution \( \hat{z} \) and the primal solution \( \hat{\beta} \).

1. initialize \( S \in \mathbb{R}^{p \times r} \) to an all-zeros matrix.
2. for \( t = 1, \ldots, r \) do
   • pick \( i_t \in [p] \) such that \( \Pr(i_t = i) = \pi_i \).
   • set \( S_{i_t} = \frac{1}{\sqrt{\pi_i}} \).
3. end
4. calculate \( \hat{z} \) as in (2.3).
5. return \( \hat{\beta} = A^T \hat{z} / \lambda \).

Now, we investigate the sampling probabilities \( \{ \pi_i \}_{i=1}^p \) in Algorithm 1, which play a critical role on the performance of the algorithm. Below are some well known probabilities discussed in the literature.

- **Uniform sampling (UNI):** \( \pi_i^{UNI} = \frac{1}{p} \).
- **Column sampling (COL):** \( \pi_i^{COL} = \frac{\|A_i\|_2^2}{\sum_{i=1}^p \|A_i\|_2^2} \).
- **Leverage sampling (LEV):** \( \pi_i^{LEV} = \frac{\|V_i\|_2^2}{\sum_{i=1}^p \|V_i\|_2^2} \).
- **Ridge leverage sampling (RLEV):** \( \pi_i^{RLEV} = \frac{\|X_i\|_2^2}{\sum_{i=1}^p \|X_i\|_2^2} \).

In the following, we discuss a new set of sampling probabilities, i.e., the optimal subsampling probabilities, which can be derived by combining the asymptotic variance of the estimators from Algorithm 1 and the A-optimal design criterion [29]. Considering the property of trace [30, Section 7.7] and the variance \( \text{Var}(\hat{\beta} - \hat{\beta}_{rls}|\mathcal{F}_n) = \frac{1}{\lambda^2}A^T \text{Var}(\hat{z} - \hat{z}^*|\mathcal{F}_n)A \), to let the trace \( \text{tr}(\text{Var}(\hat{\beta} - \hat{\beta}_{rls}|\mathcal{F}_n)) \) attain its minimum, it suffices to make \( \text{tr}(\text{Var}(\hat{z} - \hat{z}^*|\mathcal{F}_n)) \) get its minimum. Thus, we mainly investigate the asymptotic variance of the dual estimator \( \hat{z} \). As done in e.g., [21, 22, 24–26], several conditions are first presented as follows.

**Condition 2.1.** For the design matrix \( A \in \mathbb{R}^{n \times p} \), we assume that

\[
\sum_{i=1}^p \frac{\|A_i\|_2^6}{\pi_i^{3/2} p^3} = O_p(1), \quad (2.4)
\]
\[ \sum_{i=1}^{p} \frac{A_iA_i^T}{p^2 \pi_i} = O_p(1), \quad (2.5) \]
\[ \sum_{i=1}^{p} \frac{\|A_i\|_2^2}{p} = O_p(1), \quad (2.6) \]
\[ \sum_{i=1}^{p} \frac{A_iA_i^T}{p} = O_p(1), \quad (2.7) \]

where \( \pi_i \) with \( i = 1, \cdots, p \) are the given probabilities.

**Remark 2.2.** With respect to uniform sampling, i.e. \( \pi_i = p^{-1} \), the conditions (2.4) and (2.5) are equivalent to

\[ \sum_{i=1}^{p} \frac{\|A_i\|_2^6}{p} = O_p(1), \quad \sum_{i=1}^{p} \frac{A_iA_i^T\|A_i\|_2^2}{p} = O_p(1). \quad (2.8) \]

In this case, to make (2.8) hold, it is sufficient to suppose that \( E(\|A_i\|_2^6) < \infty \). Furthermore, the conditions (2.6) and (2.7) hold if \( E(\|A_i\|_2^2) < \infty \).

**Remark 2.3.** The above moment type conditions are wild. For example, if the entries of \( A \) obey the sub-Gaussian distribution [31], then all the conditions mentioned above are satisfied. The reason is that the sub-Gaussian distribution owns finite moments up to any finite order.

With the above conditions, we can present the following asymptotic distribution theorem.

**Theorem 2.1.** Assume that the conditions (2.4), (2.5), (2.6), and (2.7) are satisfied. Then, as \( p \to \infty, r \to \infty \), conditional on \( F_n \) in probability, the estimator \( \hat{z} \) constructed by Algorithm 1 satisfies

\[ V^{-1/2}(\hat{z} - \hat{z}^*) \xrightarrow{L} N(0, I), \quad (2.9) \]

where the notation \( \xrightarrow{L} \) represents the convergence in distribution, and

\[ V = \left( \frac{M_A}{p} \right)^{-1} V_c \left( \frac{M_A}{p} \right)^{-1} \]

with \( M_A = AA^T + \lambda I \) and \( V_c = \sum_{i=1}^{p} \frac{A_iA_i^T \hat{z}^* \hat{z}^{*T} A_iA_i^T}{p^2 \pi_i} \).

Following the A-optimal design criterion and the asymptotic variance \( V \) in (2.9), we can provide the optimal subsampling probabilities for Algorithm 1 by minimizing the trace \( \text{tr}(V) \). Noting that \( M_A \) does not depend on \( \pi_i \) and is nonnegative definite, we get that \( V_c(\pi_1) \ll V_c(\pi_2) \) is equivalent
Theorem 2.2. For Algorithm 1 when viewed as the L-optimal design criterion \[29\] with L\textsuperscript{V} optimal subsampling probabilities by minimizing \(\text{tr}(\cdot)\) in this case. we can simplify the optimal criterion by avoiding computing \(M\) of \(\pi\). Thus, analogous to Theorem 2.2, we get that when

\[
\hat{\beta}_{rls} = \left( \sum_{i=1}^{p} \hat{\beta}_{rls(i)} A_i \right) \frac{\|A_i\|_2}{\pi_i}, \quad i = 1, \ldots, p, \tag{2.10}
\]

where \(\hat{\beta}_{rls(i)}\) is the \(i\)-th element of the ridge estimator \(\hat{\beta}_{rls}\), \(\text{tr}(\cdot)\) achieves its minimum.

Remark 2.4. When \(\lambda \to 0^+\), \((2.11)\) can be degraded to the optimal subsampling probabilities of the compressed least squares regression.

Remark 2.5. Note that \(V_c = \lambda^2 \sum_{i=1}^{p} \frac{\hat{\beta}_{rls(i)}^2 A_i A_i^T}{\pi_i} \). Thus, by

\[
\text{tr}(V_c) \leq \frac{\lambda^2 \|A\|_2^2}{p^2} \sum_{i=1}^{p} \frac{\|A_i\|_2^2}{\pi_i} \leq \frac{\lambda^2 \|A\|_2^2}{p^2} \sum_{i=1}^{p} \frac{\|A_i\|_2^2}{\pi_i}.
\]

Further, by Cauchy-Schwarz inequality, we obtain

\[
\frac{\lambda^2 \|A\|_2^2}{p^2} \sum_{i=1}^{p} \frac{\|A_i\|_2^2}{\pi_i} \geq \frac{\lambda^2 \|A\|_2^2}{p^2} \sum_{i=1}^{p} \frac{\|A_i\|_2^2}{\pi_i} \left( \sum_{i=1}^{p} \frac{\|A_i\|_2^2}{\pi_i} \right)^2.
\]

Thus, analogous to Theorem 2.2, we get that when

\[
\pi_i = \pi_i^{\text{COL}} = \frac{\|A_i\|_2^2}{\sum_{i=1}^{p} \|A_i\|_2^2}, \tag{2.11}
\]

the upper bound of \(\text{tr}(V_c)\), i.e., \(\frac{\lambda^2 \|A\|_2^2}{p^2} \sum_{i=1}^{p} \frac{\|A_i\|_2^2}{\pi_i}\), reaches the minimum. Obviously, \((2.11)\) is easier to compute compared with \((2.10)\). However, we has to lose some accuracy as expense in this case.

Similarly, based on \(|A|_2^2 \leq \|A\|_2^2\), we have

\[
\text{tr}(V_c) \leq \frac{\lambda^2 \|A\|_2^2}{p^2} \sum_{i=1}^{p} \frac{\hat{\beta}_{rls(i)}^2}{\pi_i} = \frac{\lambda^2 \|A\|_2^2}{p^2} \sum_{i=1}^{p} \frac{\hat{\beta}_{rls(i)}^2}{\pi_i}.
\]
\[ \frac{\chi^2}{p^2} \sum_{i=1}^{p} \pi_i \sum_{i=1}^{p} \frac{\hat{\beta}_{rls(i)}^2}{\pi_i} \geq \frac{\chi^2}{p^2} \left( \sum_{i=1}^{p} | \hat{\beta}_{rls(i)} | \right)^2. \]

Then, we find that when

\[ \pi_i = \pi_i^{RSIS} = \frac{\hat{\beta}_{rls(i)}}{\sum_{i=1}^{p} | \hat{\beta}_{rls(i)} |}, \]

the above upper bound of \( \text{tr}(V_c) \) reaches the minimum. Surprisingly, \( \pi_i^{RSIS} \) corresponds to the screening criteria of iteratively thresholded ridge regression screener given in [32]. This fact implies that the screener with the probabilities in (2.10) may perform better than the one in [32].

### 2.2. Error Analysis for RSHRR

We first give an estimation error bound.

**Theorem 2.3.** Assume that

\[ c_1 \| V_i \|_2 \leq \| A_i \|_2 \leq c_2 \| V_i \|_2 \quad \text{and} \quad s_1 \| V_i \|_2 \| y \|_2 \leq | \hat{\beta}_{rls(i)} | \leq s_2 \| V_i \|_2 \| y \|_2, \quad i = 1, \cdots, p, \tag{2.12} \]

where \( 0 < c_1 \leq c_2 \) and \( 0 < s_1 \leq s_2 \), and let \( r \geq \frac{2s_1 c_2 p}{3s_1 c_1^2} \ln \left( \frac{4p}{\delta} \right) \) with \( \epsilon, \delta \in (0, 1) \). Then, for \( S \) formed by \( \pi_i = \pi_i^{OPL} \) and any \( \epsilon \), with the probability at least \( 1 - \delta \), \( \hat{\beta} \) constructed by Algorithm 1 satisfies

\[ \| \hat{\beta} - \hat{\beta}_{rls} \|_2 \leq \epsilon \| \hat{\beta}_{rls} \|_2, \tag{2.13} \]

where \( \hat{\beta}_{rls} \) is as in (1.5).

**Remark 2.6.** The assumptions in (2.12) are reasonable and reachable due to \( A_i = U \Sigma (V^i)^T \) and

\[ \hat{\beta}_{rls(i)} = A_i^T (A_i A_i^T + \lambda I)^{-1} y = V^i (\Sigma + \lambda \Sigma^{-1})^{-1} U^T y. \]

In fact, for the worst case, \( c_1 = \sigma_1(A), c_2 = \sigma_1(A), \) and \( s_1 \) and \( s_2 \) are controlled by \( \min_{j=1, \cdots, p} \left\{ \frac{\sigma_j(A)}{\sigma_j(A) + \lambda} \right\} \) and \( \max_{j=1, \cdots, p} \left\{ \frac{\sigma_j(A)}{\sigma_j(A) + \lambda} \right\} \), respectively. The aim for introducing the parameters \( c_1, c_2, s_1, \) and \( s_2 \) here is to simplify the expression of \( r \).

In the following, we provide a risk bound, in which the risk function is defined as

\[ \text{risk}(\hat{y}) = \frac{1}{n} E_y(\| \hat{y} - A \beta \|_2^2), \]

where \( E_y \) denotes the expectation on \( y \), and \( \hat{y} \) denotes the prediction of \( A \beta \).
Theorem 2.4. Suppose that the setting is the same as the one in Theorem 2.3, and let
\[ \mu = \sqrt{\sum_{j=1}^{p} \frac{\sigma_j^2(A)}{(\sigma_j^2(A) + \lambda)^2}}. \]
Then, for \( S \) formed by \( \pi_i = \pi_i^{OPL} \) and any \( \epsilon \), with probability at least \( 1 - \delta \),
\[ \text{risk}(\hat{y}) \leq \text{risk}(y^*) + 3\epsilon n \left\| A \right\|_2^2 \left( \mu^2 + \left\| \beta \right\|_2^2 \right), \]
where \( \hat{y} = A\hat{\beta} \) with \( \hat{\beta} \) constructed by Algorithm 1 and \( y^* = A\hat{\beta}_{rls} \).

3. Two Step Iterative Algorithm

Considering that the sampling probabilities (2.10) are uneconomic since they are required to figure out \( \hat{\beta}_{rls} \), we now present the approximate ones. Specifically, we first apply Algorithm 1 with \( \pi_i = \pi_i^{COL} \) and the sampling size being \( r_0 \) to return an approximation \( \tilde{\beta} \) of \( \hat{\beta}_{rls} \). Then, a set of probabilities \( \{\pi_i^{NPL}\}_{i=1}^p \) are obtained by replacing \( \hat{\beta}_{rls}(i) \) in (2.10) with \( \tilde{\beta}(i) \), i.e.,
\[ \pi_i^{NPL} = \frac{|\tilde{\beta}(i)| \left\| A_i \right\|_2}{\sum_{i=1}^{p} |\tilde{\beta}(i)| \left\| A_i \right\|_2}, \quad i = 1, \ldots, p. \tag{3.1} \]
We call them the nearly optimal subsampling probabilities. Moreover, to further reduce the estimation error, we bring in the iterative method. The key motivation is that if \( \|\hat{\beta}_t - \hat{\beta}_{rls}\|_2 \leq \epsilon \|\hat{\beta}_{t-1} - \hat{\beta}_{rls}\|_2 \) holds at the \( t \)-th iteration, then a solution owning the estimation error bound \( \epsilon^m \|\hat{\beta}_{0} - \hat{\beta}_{rls}\|_2 \) will be returned when the approximation process is repeated \( m \) times. Putting the above discussions together, we propose a two step iterative algorithm, i.e., Algorithm 2.

Remark 3.1. The step 2 of Algorithm 2 can be viewed as a variant of iterative Hessian sketch (IHS) [7]. This is because, at the \( t \)-th iteration, applying Algorithm 1 for finding \( \hat{w}_t \) is equivalent to applying Hessian sketch to the residual between \( z \) and \( \hat{z}_{t-1} \). That is, at the \( t \)-th iteration, we need to solve the following problem
\[ \min_{w_t} \frac{1}{2\lambda} \left\| S^T A^T w_t \right\|_2^2 + \frac{1}{2} \left\| w_t \right\|_2^2 - w_t^T b_t, \]
where \( w_t = z - \hat{z}_{t-1} \) and \( S \) is constructed by \( \pi_i^{NPL} \).

In addition, the step 2 of Algorithm 2 is also similar to Algorithm 1 in [8]. However, the key ideas of the two methods are different. The latter can be regarded as the preconditioned Richardson iteration [9, Chap. 2] for solving \( (AA^T + \lambda I)z = \lambda y \) with pre-conditioner \( P^{-1} = (ASS^T A^T + \lambda I)^{-1} \) and the step-size being one. Moreover, its random sampling matrix \( S \) is fixed during the iteration.
Algorithm 2 Two Step Iterative Algorithm for High-dimensional Ridge Regression

Input: $y \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times p}$, the regularized parameter $\lambda$, the iterative number $m$, the sampling size $r$ and $r_0$, where $r_0 \ll r \ll p$.

Output: the dual estimator $\hat{z}_m$ and the recovered solution $\hat{\beta}_m$.

Step1:

1. initialize $S^* \in \mathbb{R}^{p \times r_0}$ to an all-zeros matrix.
2. for $i \in 1, \ldots, p$ do
   
   $\pi_{COL}^i = \frac{\|A_i\|^2}{\sum_{i=1}^p \|A_i\|^2}.$

3. end

4. for $t \in 1, \ldots, r_0$ do
   
   pick $i_t \in [p]$ such that $\Pr(i_t = i) = \pi_i$.

   $S_{i,t}^* = \frac{1}{\sqrt{\pi_i}}.$

5. end

6. compute $A^* = AS^*$.

7. compute $C = (A^*A^{*T} + \lambda I)^{-1}$.

Step2:

1. set $\hat{z}_0 = 0$.

2. for $t \in 1, \ldots, m$ do

   $\hat{\beta}_{t-1} = \frac{1}{\lambda} A^T \hat{z}_{t-1}$.

   $b_t = y - A\hat{\beta}_{t-1} - \hat{z}_{t-1}$.

   $\bar{z} = \lambda C b_t$.

   $\bar{\beta} = \frac{A^T \bar{z}}{\lambda}$.

   compute $\pi_t^{NOLP}$ by (3.1).

   compute $\hat{w}_t$ by applying Algorithm 1 with $y = b_t$ and $\pi = \pi_t^{NOLP}$.

   $\hat{z}_t = \hat{z}_{t-1} + \hat{w}_t$.

3. end

4. return $\hat{z}_m$ and $\hat{\beta}_m = \frac{A^T \hat{z}_m}{\lambda}$.
Next, we show that the difference of $\hat{z}_*$ and $\hat{z}_1$ still obeys asymptotically normal distribution, where $\hat{z}_1$ is returned from Algorithm 2 with $m = 1$.

**Theorem 3.1.** Suppose that the conditions (2.6) and (2.7) hold, and let

$$ N_1\|A_i\|_2\|y\|_2 \leq |\tilde{\beta}_{(i)}| \leq N_2\|A_i\|_2\|y\|_2 \text{ and } N_3\|A_i\|_2\|y\|_2 \leq |\hat{\beta}_{rls(i)}| \leq N_4\|A_i\|_2\|y\|_2, \quad i = 1, \ldots, p, $$

(3.2)

where $\tilde{\beta}_{(i)}$ is as in Algorithm 2. Then, as $p \to \infty$, $r \to \infty$, $r_0 \to \infty$, conditional on $\mathcal{F}_n$ and $\tilde{\beta}$ in probability, the dual estimator $\hat{z}_1$ constructed by Algorithm 2 satisfies

$$ V_{OPL}^{-1/2}(\hat{z}_1 - \hat{z}^*) \overset{L}{\to} N(0, I), $$

(3.3)

where

$$ V_{OPL} = (\frac{M_A}{p})^{-1} V_{OPL} (\frac{M_A}{p})^{-1} $$

with

$$ V_{OPL} = \sum_{i=1}^{p} \frac{A_iA_i^T\hat{z}^*\hat{z}^{*T}A_iA_i^T}{p^2 \pi_i^{OPL}} = \sum_{i=1}^{p} |\hat{\beta}_{rls(i)}| \|A_i\|_2\sum_{i=1}^{p} \frac{A_iA_i^T\hat{z}^*\hat{z}^{*T}A_iA_i^T}{p^2 |\hat{\beta}_{rls(i)}| \|A_i\|_2}. $$

Now, we provide an estimation error bound of our algorithm.

**Theorem 3.2.** To the assumptions of Theorem 2.3, add that

$$ s_3\|V^2\|_2\|y\|_2 \leq |\tilde{\beta}_{(i)}| \leq s_4\|V^2\|_2\|y\|_2, \quad i = 1, \ldots, p, $$

(3.4)

where $\tilde{\beta}_{(i)}$ is as in Algorithm 2 and $0 < s_3 \leq s_4$. Then, for $\tilde{S}$ constructed by $\pi_i^{OPL}$ and any $\epsilon$, with the probability at least $1 - m\delta$, $\hat{\beta}_m$ generated from Algorithm 2 satisfies

$$ \|\hat{\beta}_m - {\hat{\beta}_{rls}}\|_2 \leq \epsilon \|\hat{\beta}_{rls}\|_2. $$

(3.5)

**Remark 3.2.** The bound (3.5) can be used to determine the iteration number. Specifically, it is enough to do $\log\epsilon$ iterations to get $\|\hat{\beta}_m - \hat{\beta}_{rls}\|_2 \leq \epsilon \|\hat{\beta}_{rls}\|_2$.

4. Numerical Experiments

In this section, we provide the numerical results of experiments with simulation data and real data. All experiments are implemented on a laptop running MATLAB software with 16 GB random-access memory (RAM).
4.1. Simulation Data–Example 1

In this example, the simulation data is generated as done in [18]. Specifically, we first produce an $n$-by-$p$ matrix $B$ randomly, whose entries are drawn i.i.d. from the $N(0,1)$ distribution and SVD is denoted as $U_B \Sigma_B V_B^T$ with $U_B \in \mathbb{R}^{n \times n}$, $\Sigma_B \in \mathbb{R}^{n \times n}$ and $V_B \in \mathbb{R}^{p \times n}$. Then, we get $A$ by replacing $\Sigma_B$ with $\Sigma_0$, i.e., $A = U_B \Sigma_0 V_B^T$, where $\Sigma_0$ is a diagonal matrix with polynomial decay diagonal elements $\sigma_j (j = 1, \cdots, n)$, namely, $\sigma_j \propto 9 \times j^{-8}$. Furthermore, we construct the response vector $y$ by $y = A\beta + \varsigma$, where $\beta \in \mathbb{R}^p$ and $\varsigma \in \mathbb{R}^n$ have i.i.d. $N(0,1)$ entries.

In the specific experiments, we set $n = 500$ and $p = 20000$. The description on parameters of the experiments is summarized in Table 1, the explanation on six sampling methods is given in Table 2, and the numerical results on accuracy, i.e., the estimation error $\|\hat{\beta}_{m} - \hat{\beta}_{rls}\|_2 / \|\hat{\beta}_{rls}\|_2$ and the prediction error $\|A\hat{\beta}_{m} - A\hat{\beta}_{rls}\|_2 / \|A\hat{\beta}_{rls}\|_2$, and CPU time are shown in Figures 1-4. Note that all the error results are on log-scale, all the numerical results are based on 50 replications of Algorithm 2, and it suffices to run the step 2 in Algorithm 2 if $\pi_{OPL}^i, \pi_{LEV}^i, \pi_{RLEV}^i, \pi_{UNI}^i$ and $\pi_{COL}^i$ are used to generate $S$. In addition, when $\pi_{OPL}^i$ is employed, $C$ in Algorithm 2 should be $(AA^T + \lambda I)^{-1}$, and when $\pi_{LEV}^i, \pi_{RLEV}^i, \pi_{UNI}^i$ and $\pi_{COL}^i$ are adopted, the lines 5–7 of the step 2 of Algorithm 2 can be omitted.

Table 1: Description of two experiments for example 1.

| Kinds | Comparison | $r$ | $\lambda$ | $m$ | $r_0$ | Results |
|-------|------------|----|-----------|----|------|---------|
| 1     | six methods | 1000 | 1 to 50   | 3  | 100 (NOPL) | Figs. 1-3(b) |
| 1     | six methods | 1000 | 10        | 1 to 15  | 100 (NOPL) | Figs. 1-3(c) |
| 2     | OPL and NOPL | 2000 | 10        | 3  | 100 to 2000 (NOPL) | Fig. 4 |

Table 2: Explanation of sampling methods with different probabilities.

| Method | $\pi_i$ | Expression |
|--------|---------|------------|
| OPL    | $\pi_i^{OPL}$ | $|\tilde{\beta}_{rl(i)}| / |A_i| / \sum_{i=1}^p |\tilde{\beta}_{rl(i)}| / |A_i|_2$ |
| NOPL   | $\pi_i^{NOPL}$ | $|\tilde{\beta}_{rl(i)}| / |A_i| / \sum_{i=1}^p |\tilde{\beta}_{rl(i)}| / |A_i|_2$ |
| LEV    | $\pi_i^{LEV}$ | $|V^*|_2 / \sum_{i=1}^p |V^*|_2$ |
| RLEV   | $\pi_i^{RLEV}$ | $|X^*_i|_2 / \sum_{i=1}^p |X^*_i|_2$ |
| COL    | $\pi_i^{COL}$ | $|A_i|_2 / \sum_{i=1}^p |A_i|_2$ |
| UNI    | $\pi_i^{UNI}$ | $1/p$ |

In the first experiment, we aim to show that the estimators established by OPL and NOPL have better performance. The corresponding numerical results are presented in Figures 1-3.
these figures, it is obvious to find that OPL and NOPL outperform other methods on estimation and prediction accuracy no matter what $r$, $\lambda$ and $m$ are, but they need more computing time than COL and UNI. However, the improvement in accuracy is more than the sacrifice of calculation cost, and fortunately, OPL and NOPL are cheaper than LEV and RLEV. What is more, we can observe that NOPL has extremely similar accuracy to OPL, and the former consumes less running time. In addition, in most cases, the errors of all the methods decrease when $r$, $\lambda$ and $m$ increase.

Figure 1: Comparison of estimation errors using different methods for example 1.
Figure 2: Comparison of prediction errors using different methods for example 1.

Figure 3: Comparison of CPU time using different methods for example 1.

For the second experiment, we compare the methods OPL and NOPL with different $r_0$. According to the numerical results displayed in Figure 4, it is evident to conclude that for different $r_0$, NOPL is able to achieve significantly similar accuracy to OPL but spends less computational cost.
4.2. Simulation Data–Example 2

For this example, we produce the simulation data as done in [8]. Specifically, we construct an $n$-by-$p$ design matrix $A = PDQ^T + \alpha M$, where $P \in \mathbb{R}^{n \times n}$ is a random matrix with i.i.d. $N(0, 1)$ entries, $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix with diagonal entries $D_{ii} = (1 - \frac{i}{p})^5 (i = 1, \ldots, n)$, $Q \in \mathbb{R}^{p \times n}$ is a random column orthonormal matrix, $M \in \mathbb{R}^{n \times p}$ is a noise matrix with i.i.d. $N(0, 1)$ entries, and $\alpha > 0$ is a parameter used to balance $PDQ^T$ and $M$. In addition, the response vector $y \in \mathbb{R}^n$ is generated according to $y = A\beta + \gamma \varsigma$, where $\beta \in \mathbb{R}^p$ and $\varsigma \in \mathbb{R}^p$ are constructed by i.i.d. $N(0, 1)$ entries. In the specific experiments, we set $n = 500$, $p = 20000$, $\alpha = 0.0001$ and $\gamma = 0.5$, and repeat the implementations in Section 4.1 with different $r$, $r_0$, $\lambda$ and $m$ shown in Table 3.

![Figure 4: Comparison of OPL and NOPL with different $r_0$ for example 1.](image)

Table 3: Description of two experiments for example 2.

| Kinds          | Comparison | $r$   | $\lambda$ | $m$   | $r_0$        | Results                  |
|----------------|------------|-------|------------|-------|--------------|--------------------------|
| 1              | six methods| 3000  | 20         | 15    | 2000 (NOPL)  | Figs. 5-7(a)             |
| 1              |            | 5000  | 1 to 200   | 15    | 2000 (NOPL)  | Figs. 5-7(b)             |
| 1              |            | 5000  | 20         | 1 to 30| 2000 (NOPL)  | Figs. 5-7(c)             |
| 2              | OPL and NOPL| 5000 | 20         | 15    | 500 to 20000 (NOPL) | Fig. 8 |

From the numerical results presented in Figures 5-8, we can gain the similar observations to the
ones in Section 4.1. That is, taking different $r$, $\lambda$ and $m$, OPL and NOPL always perform better than other methods on accuracy, however, need more CPU time compared with COL and UNI. And, OPL and NOPL still show better computational efficiency than LEV and RLEV. Besides, when setting a proper $r_0$ or a large $\lambda$, NOPL and OPL have similar accuracy but the former needs less running time. Unfortunately, when $r_0$ is very large, NOPL loses its advantage in CPU time. This is because in this case the computational cost of $\beta$ may not be less than that of $\beta_{rls}$.

![Figure 5: Comparison of estimation errors using different methods for example 2.](image)
Figure 6: Comparison of prediction errors using different methods for example 2.

Figure 7: Comparison of CPU Time using different methods for example 2.
4.3. Real Data–Gene Expression Cancer RNA-Seq Data Set

The data set is from the UCI machine learning repository, which can be found at [http://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq](http://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq). Here, we only take the first 400 samples with 20531 real-valued features, and centralize the design matrix. The response vector consists of 1, 2, 3, 4 and 5 labels, which represent five different types of tumors, i.e., PRAD, LUAD, BRCA, KIRC and COAD. We also centralize it.

We repeat the experiments in Sections 4.1 and 4.2 with different $r, r_0, \lambda$ and $m$. More details are put in Table 4.

| Kinds | Comparison | $r$ | $\lambda$ | $m$ | $r_0$ | Results |
|-------|------------|----|-----------|----|------|---------|
| 1     | six methods | 8000 | 1 to 50 | 16 | 5000 (NOPL) | Figs. 9-11(b) |
| 1     | six methods | 8000 | 10       | 1 to 26 | 5000 (NOPL) | Figs. 9-11(c) |
| 2     | OPL and NOPL | 8000 | 10       | 16 | 1000 to 20531 (NOPL) | Fig. 12 |

The numerical results are displayed in Figures 9-12, and the conclusions summarized from these figures are akin to the ones found in Sections 4.1 and 4.2. Namely, compared with UNI and COL, the accuracy of OPL and NOPL is dramatically improved at the cost of slightly computational efficiency,
and OPL performs better than LEV and RLEV on accuracy and computing time. Although NOPL is only a little better than LEV and RLEV on accuracy, it owns greatly advantage of CPU time. When taking a proper $r_0$, NOPL can be a well approximation of OPL but consumes less computing time. However, when $r_0$ is very large, NOPL will lose its superiority in computational cost. In addition, for this real data, the choice of $\lambda$ has little influence on accuracy.

Figure 9: Comparison of estimation errors for different methods for Gene Expression Cancer RNA-Seq data set.
Figure 10: Comparison of prediction errors for different methods for Gene Expression Cancer RNA-Seq data set.

Figure 11: Comparison of CPU Time for different methods for Gene Expression Cancer RNA-Seq data set.
4.4. Real Data–Gisette Data Set

This data set is also from the UCI machine learning repository, which can be found in [http://archive.ics.uci.edu/ml/datasets/Gisette](http://archive.ics.uci.edu/ml/datasets/Gisette). In our experiments, the first 100 samples of training set with 5000 real-valued features are taken, and the response vector is made up with ±1 labels. Also, we centralize the response vector and design matrix prior to analysis.

As done in Section 4.3, we also repeat the experiments in Sections 4.1 and 4.2 with different $r$, $r_0$, $\lambda$ and $m$. The detailed description can be found in Table 5.

| Kinds          | Comparison | $r$     | $\lambda$ | $m$     | $r_0$       | Results             |
|----------------|------------|---------|-----------|---------|-------------|---------------------|
| 1              | six methods| 1000 to 5000 | 10        | 10      | 900 (NOPL) | Figs. 13-15(a)      |
| 2              | OPL and NOPL| 2000    | 1 to 50  | 10      | 900 (NOPL) | Figs. 13-15(b)      |
|                |            | 2000    | 10       | 1 to 26 | 900 (NOPL) | Figs. 13-15(c)      |
| 2              |            | 2000    | 10       | 10      | 500 to 5000 (NOPL) | Fig. 16 |

The numerical results are shown in Figures 13-16, and are almost identical with the observations in Section 4.3. To be more specific, whatever the values of $r$, $\lambda$ and $m$ are, for accuracy, OPL and NOPL always outperform other methods. Similarly, as for CPU time, OPL and NOPL are
slightly inferior to UNI and COL, but are greatly superior to LEV and RLEV. Only when $r_0$ is not particularly large, NOPL has good performance on both accuracy and computing time, and qualifies as a well alternative to OPL. Besides, the change of $\lambda$ also has little effect on accuracy.

Figure 13: Comparison of estimation errors for different methods for Gisette data set.

Figure 14: Comparison of prediction errors for different methods for Gisette data set.
Figure 15: Comparison of CPU Time for different methods for Gisette data set.

Figure 16: Comparison of OPL and NOPL with different $r_0$ and Gisette data set.

Appendix A Proof of Theorem 2.1

We start by establishing two lemmas.
Lemma A.1. Assuming that the conditions (2.4), (2.5) and (2.6) are satisfied, we have

$$
\sum_{i=1}^{p} \pi_i \left( \frac{e_i}{p} \right)^3 = O_p(1), \tag{A.1}
$$

where $e_i = \left( \frac{A_i A_i^T}{\pi_i} + \lambda I \right) \hat{z}^* - \tilde{y}$ with $\tilde{y} = \lambda y$ and $\hat{z}^*$ being as in (1.3).

Proof. With $e_i = \left( \frac{A_i A_i^T}{\pi_i} + \lambda I \right) \hat{z}^* - \tilde{y}$ and (1.3), it is easy to see that

$$
\sum_{i=1}^{p} \pi_i \left( \frac{e_i}{p} \right)^3 = \frac{1}{p^3} \sum_{i=1}^{p} \pi_i \left( \frac{A_i A_i^T}{\pi_i} + \lambda I \right) (A A^T + \lambda I)^{-1} \tilde{y} \| \tilde{y} \|^2.
$$

Then, considering the basic triangle inequality and the fact that $\sum_{i=1}^{p} \pi_i = 1$, we can have

$$
\sum_{i=1}^{p} \pi_i \left( \frac{e_i}{p} \right)^3 \leq \frac{1}{p^3} \sum_{i=1}^{p} \pi_i \left( \frac{A_i A_i^T}{\pi_i} + \lambda I \right) (A A^T + \lambda I)^{-1} \tilde{y} \| \tilde{y} \|^2 + \frac{\| \tilde{y} \|^3}{p^3}
$$

$$
+ \frac{1}{p^3} \sum_{i=1}^{p} \pi_i \left( \frac{A_i A_i^T}{\pi_i} + \lambda I \right) (A A^T + \lambda I)^{-1} \tilde{y} \| \tilde{y} \|^2 \| \tilde{y} \|_2
$$

$$
+ \frac{1}{p^3} \sum_{i=1}^{p} \pi_i \left( \frac{A_i A_i^T}{\pi_i} + \lambda I \right) (A A^T + \lambda I)^{-1} \tilde{y} \| \tilde{y} \|^2.
$$

$$
\leq \frac{\| \tilde{y} \|^3}{p^3} \pi_n \left( A A^T + \lambda I \right) \left( \sum_{i=1}^{p} \left( \frac{A_i A_i^T}{\pi_i} \right)^2 \pi_i \right) + 3 \lambda \sum_{i=1}^{p} \left( \frac{A_i A_i^T}{\pi_i} \right)^2 \pi_i
$$

$$
+ 3 \lambda^2 \sum_{i=1}^{p} \left( \frac{A_i A_i^T}{\pi_i} \right)^2 \pi_i + \frac{\| \tilde{y} \|^3}{p^3}
$$

$$
+ \frac{3 \| \tilde{y} \|^3}{p^3} \pi_n \left( A A^T + \lambda I \right) \left( \sum_{i=1}^{p} \left( \frac{A_i A_i^T}{\pi_i} \right)^2 \pi_i \right) + 2 \lambda \sum_{i=1}^{p} \left( \frac{A_i A_i^T}{\pi_i} \right)^2 \pi_i + \lambda^2
$$

$$
+ \frac{3 \| \tilde{y} \|^3}{p^3} \pi_n \left( A A^T + \lambda I \right) \left( \sum_{i=1}^{p} \left( \frac{A_i A_i^T}{\pi_i} \right)^2 \pi_i \right) + \lambda.
$$

(A.2)

Following

$$
\frac{\| \tilde{y} \|^3}{p} = o_p(1), \tag{A.3}
$$

which can be derived from $np^{-1} \to 0$, and noting (2.4), (2.5), (2.6) and (A.2), we can get

$$
\sum_{i=1}^{p} \pi_i \left( \frac{e_i}{p} \right)^3 \leq \frac{\| \tilde{y} \|^3}{p^3} \pi_n \left( A A^T + \lambda I \right) \left( \sum_{i=1}^{p} \left( \frac{A_i A_i^T}{\pi_i} \right)^2 \pi_i \right) + o_p(1)
$$

by (2.5), (2.6), (A.2), and (A.3).

$$
= O_p(1). \tag{2.3}
$$

Thus, (A.1) is arrived. \qed

25
Lemma A.2. Suppose that the conditions (2.6) and (2.7) hold. Then, conditional on \( \mathcal{F}_n \) in probability,

\[
\frac{\hat{M}_A - M_A}{p} = O_p(\frac{r^{-1/2}}{p}),
\]

(A.4)

\[
\frac{e^*}{p} = O_p(\frac{r^{-1/2}}{p}),
\]

(A.5)

where \( M_A = AA^T + \lambda I \), \( \hat{M}_A = ASS^T A^T + \lambda I \) with \( S \in \mathbb{R}^{p \times r} \) constructed as in Algorithm 1, and \( e^* = (\hat{M}_A \hat{z}^* - \bar{y}) \) with \( \bar{y} = \lambda y \) and \( \hat{z}^* \) being as in (1.3).

Proof. First, note that

\[
\frac{1}{p^2} \sum_{i=1}^{p} \pi_i \left[ \frac{A_i A_i^T}{\pi_i} + \lambda I \right] - (AA^T + \lambda I) \left[ \frac{A_i A_i^T}{\pi_i} + \lambda I \right] - (AA^T + \lambda I) = O_p(1),
\]

where the last equality is from (2.5) and (2.7). This result implies, for any \( n \)-dimensional vector \( \ell \) with finite elements,

\[
\frac{1}{p^2} \sum_{i=1}^{p} \pi_i \left[ \frac{A_i A_i^T}{\pi_i} + \lambda I \right] - (AA^T + \lambda I) \ell^T \left[ \frac{A_i A_i^T}{\pi_i} + \lambda I \right] - (AA^T + \lambda I) = O_p(1).
\]

(A.6)

Thus, following \( \mathbb{E}(\hat{M}_A | A) = M_A \), it is natural to get

\[
\text{Var}(\frac{\hat{M}_A - M_A}{p} | A) = \mathbb{E}[\frac{\hat{M}_A - M_A}{p} \ell \ell^T (\hat{M}_A - M_A) | A]
\]

\[
= \frac{1}{rp^2} \sum_{i=1}^{p} \pi_i \left[ \frac{A_i A_i^T}{\pi_i} + \lambda I \right] - (AA^T + \lambda I) \ell^T \left[ \frac{A_i A_i^T}{\pi_i} + \lambda I \right] - (AA^T + \lambda I)
\]

\[
= O_p(r^{-1}),
\]

which together with the Markov’s inequality implies (A.3).
Combining (A.6) and (A.3), we can get
\[
\frac{1}{p^2} \sum_{i=1}^{p} \pi_i \hat{z}^T (\frac{A_i A_i^T}{\pi_i} + \lambda I) \ell \ell^T (\frac{A_i A_i^T}{\pi_i} + \lambda I) \hat{z}^* = \frac{1}{p^2} \hat{z}^T (AA^T + \lambda I) \ell \ell^T (AA^T + \lambda I) \hat{z}^* + O_p(1).
\]  
(A.7)

Thus, considering \(e^* = \sum_{t=1}^{r} \frac{1}{r} e_t\) with \(e_t = (\frac{A_i A_i^T}{\pi_i} + \lambda I) \hat{z}^* - \tilde{y}\) and \(E(e_t \mid F_n) = 0\), and (A.7), we can obtain
\[
\text{Var} \left( \frac{\ell^T e^*}{p} \mid F_n \right) = \ell^T E \left( \frac{e^*}{p} \left( \frac{e^*}{p} \right)^T \mid F_n \right) \ell = \frac{1}{r p^2} \ell^T \left( \sum_{i=1}^{p} \pi_i e_t e_t^T \right) \ell
\]
\[
= \frac{1}{r p^2} \ell^T \left[ \sum_{i=1}^{p} \pi_i \left( (\frac{A_i A_i^T}{\pi_i} + \lambda I) \hat{z}^* - \tilde{y} \right) \left( (\frac{A_i A_i^T}{\pi_i} + \lambda I) \hat{z}^* - \tilde{y} \right)^T \right] \ell
\]
\[
= \frac{1}{r p^2} \ell^T \hat{z}^T (AA^T + \lambda I) \ell \ell^T (AA^T + \lambda I) \hat{z}^* + O_p(1) - \frac{\ell^T \tilde{y} \ell}{p^2}
\]
\[
= \frac{1}{r} \frac{\tilde{y}^T \ell^T \ell}{p^2} + O_p(1) - \frac{\ell^T \tilde{y} \ell}{p^2}
\]
\[
= O_p(r^{-1}).
\]

Consequently, by the Markov’s inequality, (A.5) is obtained.

\[\square\]

**Proof of Theorem 2.1** Considering that
\[
\hat{z} = (ASS^T A^T + \lambda I)^{-1} \tilde{y} = \hat{M}_A^{-1} \tilde{y},
\]
\[
\hat{z}^* = (AA^T + \lambda I)^{-1} \tilde{y} = M_A^{-1} \tilde{y},
\]

where \(\tilde{y} = \lambda y\), we can rewrite \(\hat{z} - \hat{z}^*\) as
\[
\hat{z} - \hat{z}^* = (ASS^T A^T + \lambda I)^{-1} (\tilde{y} - (ASS^T A^T + \lambda I) \hat{z}^*)
\]
\[
= \hat{M}_A^{-1} (\tilde{y} - \hat{M}_A \hat{z}^*) = -\hat{M}_A^{-1} e^*
\]
\[
= -(\hat{M}_A^{-1} - M_A^{-1} + M_A^{-1}) e^*
\]
\[
= -M_A^{-1} e^* - (\hat{M}_A^{-1} - M_A^{-1}) e^*
\]
\[
= -M_A^{-1} e^* + \hat{M}_A^{-1} (\hat{M}_A - M_A) M_A^{-1} e^*
\]
\[
= -\left(\frac{MA}{p}\right)^{-1}e^* + \left(\frac{\hat{MA}}{p}\right)^{-1}(\hat{MA} - MA)\left(\frac{MA}{p}\right)^{-1}e^* \\
= -\left(\frac{MA}{p}\right)^{-1}e^* + O_p(f_n(r^{-1})),
\]

where the last equality is derived by (2.7) and Lemma A.2. Thus, to prove (2.9), we first prove
\[
\left(\frac{V_c}{r}\right)^{-1/2}\left(\frac{e^*}{p}\right) \xrightarrow{L} N(0, I).
\]

(A.10)

Recall that
\[
e^* = \sum_{t=1}^T \frac{1}{r} e_{i_t}
\]

with
\[
e_{i_t} = \left(\frac{A_{i_t}A_{i_t}^T}{\pi_{i_t}} + \lambda I\right)\hat{z}^* - \tilde{y}.
\]

Now, we construct the sequence \(\{e_{i_t}\}_{t=1}^T\). These random vectors are independent and identically distributed and it is easy to get that \(E(e_{i_t}|F_n) = 0\). Furthermore, noting that
\[
V_c = \frac{1}{p^2} \sum_{i=1}^p A_iA_i^T \hat{z}^* \hat{z}^*^TA_iA_i^T = O_p(1),
\]

(A.11)

which can be obtained from (2.3), together with (2.7) and (A.3), we have
\[
\text{Var}(\frac{e_{i_t}}{p} | F_n) = E(\frac{e_{i_t}e_{i_t}^T}{p^2} | F_n) = \sum_{i=1}^p \pi_i \frac{e_{i_t}e_{i_t}^T}{p^2}
\]

\[
= \sum_{i=1}^p \pi_i \left[\left(\frac{A_iA_i^T}{\pi_i} + \lambda I\right)\hat{z}^* - \tilde{y}\right]\left[\left(\frac{A_iA_i^T}{\pi_i} + \lambda I\right)\hat{z}^* - \tilde{y}\right]^T
\]

\[
= \sum_{i=1}^p \pi_i p^2 A_iA_i^T \hat{z}^* \hat{z}^*^T \frac{A_iA_i^T}{\pi_i} + \frac{(\lambda \hat{z}^* - \tilde{y})(\lambda \hat{z}^* - \tilde{y})^T}{p^2}
\]

\[
= \sum_{i=1}^p \frac{A_iA_i^T \hat{z}^* \hat{z}^*^T A_iA_i^T}{p^2 \pi_i} + o_p(1) \text{ by (2.7) and (A.3)}
\]

(A.12)

\[
= O_p(1) \text{ by (A.11)}
\]

(A.13)

In addition, for any \(\xi > 0\), we have
\[
\sum_{i=1}^r E\left[\left\| r^{-\frac{1}{2}} p^{-1} e_{i_t}\right\|_2^2 I(\left\| r^{-\frac{1}{2}} p^{-1} e_{i_t}\right\|_2 > \xi) | F_n\right]
\]

28
\[
\begin{align*}
&= \sum_{i=1}^{p} \pi_i \|p^{-1}e_i\|_2^2 I(\|r^{-1/2}p^{-1}e_i\|_2 > \xi) \\
&\leq (r^{1/2}\xi)^{-1} \sum_{i=1}^{p} \pi_i \|p^{-1}e_i\|_2^3 \\
&= o_p(1),
\end{align*}
\]

where the inequality is deduced by the constraint \(I(\|r^{-1/2}p^{-1}e_i\|_2 > \xi)\), and the last equality is from Lemma A.1. Putting the above discussions together, we find that the Lindeberg-Feller conditions are satisfied in probability. Thus, by the Lindeberg-Feller central limit theorem [27, Proposition 2.27], and noting (A.13), we can acquire

\[
\left[\text{Var}\left(\frac{e_i}{p}\big| F_n\right)\right]^{-1/2}(r^{-1/2}p^{-1} \sum_{i=1}^{r} e_{i_{it}}) \xrightarrow{L} N(0, I),
\]

which combined with \(\frac{e_i}{p} = r^{-1}p^{-1} \sum_{t=1}^{r} e_{i_{it}}\) and \(\text{Var}\left(\frac{e_i}{p}\big| F_n\right) = r^{-1}\text{Var}\left(\frac{e_i}{p}\big| F_n\right)\) gives

\[
\left[r^{-1}\text{Var}\left(\frac{e_i}{p}\big| F_n\right)\right]^{-1/2}\left(\frac{e_i}{p}\right) \xrightarrow{L} N(0, I).
\]

Thus, by Lemma A.2, A.12, and the Slutsky’s Theorem [34, Theorem 6], we can get (A.10).

Now, we prove (2.9). Following (2.7) and (A.11), it is easy to get

\[
V = \left(\frac{M_A}{p}\right)^{-1}V_c\left(\frac{M_A}{p}\right)^{-1} = O_p(r^{-1}),
\]

which together with (A.9) yields

\[
V^{-1/2}(\hat{\varepsilon} - \varepsilon^*) = -V^{-1/2}\left(\frac{M_A}{p}\right)^{-1}\frac{e_{i}}{p} + O_p(x_n(r^{-1/2}))
\]

\[
= -V^{-1/2}\left(\frac{M_A}{p}\right)^{-1}\left(\frac{V_c}{r}\right)^{1/2}\left(\frac{V_c}{r}\right)^{-1/2}\frac{e_{i}}{p} + O_p(x_n(r^{-1/2})).
\]

(A.14)

In addition, it is verified that

\[
V^{-1/2}\left(\frac{M_A}{p}\right)^{-1}\left(\frac{V_c}{r}\right)^{1/2}\left[\left(V^{-1/2}\left(\frac{M_A}{p}\right)^{-1}\left(\frac{V_c}{r}\right)^{1/2}\right)^T
\]

\[
= V^{-1/2}\left(\frac{M_A}{p}\right)^{-1}\left(\frac{V_c}{r}\right)^{1/2}\left(\frac{V_c}{r}\right)^{1/2}\left(\frac{M_A}{p}\right)^{-1}V^{-1/2} = I.
\]

(A.15)

Thus, combining (A.10), (A.14), and (A.15), by the Slutsky’s Theorem, we get the desired result (2.9).
Appendix B Proof of Theorem 2.2

According to the Cauchy-Schwarz inequality, we have

$$\text{tr}(V^c) = \sum_{i=1}^{p} \frac{A_i^T \hat{z}_i^* T A_i}{p^2 \pi_i} = \lambda^2 \sum_{i=1}^{p} \frac{\hat{\beta}_{rls(i)}^2 \|A_i\|_2^2}{p^2 \pi_i}$$

where the equality in the last inequality holds if and only if $\pi_i$ is proportional to $|\hat{\beta}_{rls(i)}| \|A_i\|_2$ for some constant $C_0 \geq 0$. Thus, following $\sum_{i=1}^{p} \pi_i = 1$, the desired result (2.10) is obtained.

Appendix C Proof of Theorem 2.3

We first present two auxiliary lemmas.

**Lemma C.1.** [8, Theorem 23] If $J, H \in \mathbb{R}^{m \times m}$ are real symmetric positive semi-definite matrices such that $\sigma_1(J) \geq \sigma_2(J) \geq \cdots \geq \sigma_m(J)$ and $\sigma_1(H) \geq \sigma_2(H) \geq \cdots \geq \sigma_m(H)$, then

$$\sigma_j(J - H) \leq \sigma_j \begin{pmatrix} J & 0 \\ 0 & H \end{pmatrix}, \quad j = 1, \ldots, m.$$  

Especially,

$$\|J - H\|_2 \leq \max\{\|J\|_2, \|H\|_2\}.$$  

**Lemma C.2.** For $S$ established by $\pi_i = \pi_i^{OPL}$, assuming that (2.12) holds and letting $r \geq \frac{R_{s_c} \sqrt{p}}{3\sqrt{\epsilon_1} \epsilon^* \ln(\frac{1}{\delta})}$ with $\epsilon \in (0, \frac{1}{2})$ and $\delta \in (0, 1)$, we have

$$\|V^T S S^T V - I\|_2 \leq \epsilon'$$

with the probability at least $1 - \delta$.

**Proof.** The proof can be accomplished along the line of the proof of [8, Theorem 3]. However, for our case, it is necessary to note that

$$\|F_i\|_2 = \|M_i M_i^T - \frac{V_i^T V_i}{r}\|_2 \leq \max\{\|M_i M_i^T\|_2, \frac{1}{r}\}$$  by Lemma C.1

$$= \frac{1}{r} \max\{\frac{(V_i)^T V_i}{\sqrt{\pi_i^{OPL}}} \|V_i\|_2, 1\} = \frac{1}{r} \max\{\|V_i\|_2^2, 1\}$$

30
To prove (C.1), we define the loss functions

\[ L = \max_{1 \leq i \leq p} \left\{ \frac{\|V_i\|_2^2 \sum_{j=1}^p |\hat{\beta}_{r,i}(i) - \beta_{r,i}(i)| + \|A_i\|_2}{M}, 1 \right\} \quad \text{by (2.11)} \]

\[ \leq \max_{1 \leq i \leq p} \left\{ \frac{\|V_i\|_2^2 \sum_{j=1}^p s_{2c_2} [V_i]^2}{s_{1c_1}[\hat{V}_i]^2}, 1 \right\} \quad \text{by (2.12)} \]

\[ \leq \max_{1 \leq i \leq p} \left\{ \frac{s_{2c_2}}{s_{1c_1}} \sum_{i=1}^p [\hat{V}_i]^2, 1 \right\} \leq \max_{1 \leq i \leq p} \frac{s_{2c_2}}{s_{1c_1}}, 1 \}

\[ \leq \frac{s_{2c_2}}{rs_{1c_1}} \]

and

\[ E(F_t^2) + \frac{(V^TV)^2}{r^2} = E(M_tM_t^T | M_t^2) = \sum_{i=1}^p \pi_i^{OPL} \frac{(V_i)^T V_i^2}{r^2 (s_i^{OPL})^2} \]

\[ = \frac{1}{r^2} \sum_{i=1}^p (V_i)^T V_i^2 \sum_{j=1}^p |\hat{\beta}_{r,i}(i) - \beta_{r,i}(i)| + \|A_i\|_2 \quad \text{by (2.10)} \]

\[ \leq \frac{1}{r^2} \sum_{i=1}^p (V_i)^T V_i^2 \sum_{j=1}^p s_{2c_2} [V_i]^2 \quad \text{by (2.12)} \]

\[ = \frac{s_{2c_2}}{r^2 s_{1c_1}} \sum_{i=1}^p (V_i)^T V_i \sum_{i=1}^p [V_i]^2 \]

\[ = \frac{s_{2c_2} \rho}{r^2 s_{1c_1}} \sum_{i=1}^p (V_i)^T V_i \]

where \( F_t = M_tM_t^T - \frac{V^TV}{r} \) with \( M_t = \frac{(V_i)^T}{\sqrt{s_i^{OPL}}} \) and \( t = 1, \ldots, r \).

**Proof of Theorem 2.3** Noting \( \hat{\beta} = \frac{1}{\lambda} V \Sigma U^T \hat{z} \) and \( \hat{\beta}_{r} = \frac{1}{\lambda} V \Sigma U^T \hat{z}^* \), we can rewrite (2.13) as

\[ \frac{1}{\lambda} \| \Sigma U^T (\hat{z} - \hat{z}^*) \|_2 \leq \frac{\epsilon}{\lambda} \| \Sigma U^T \hat{z}^* \|_2. \quad \text{(C.1)} \]

To prove (C.1), we define the loss functions \( L(z) \) and \( \hat{L}(z) \) as

\[ L(z) = \frac{1}{2\lambda} \| A^T z \|_2^2 + \frac{1}{2} \| z \|_2^2 - z^T y \]

and

\[ \hat{L}(z) = \frac{1}{2\lambda} \| S^T A^T z \|_2^2 + \frac{1}{2} \| z \|_2^2 - z^T y. \]

Thus, by Taylor expansion, we can acquire

\[ \hat{L}(\hat{z}) = \hat{L}(\hat{z}^*) + (\hat{z} - \hat{z}^*)^T \nabla \hat{L}(\hat{z}^*) + (\hat{z} - \hat{z}^*)^T \nabla^2 \hat{L}(z_0)(\hat{z} - \hat{z}^*), \quad \text{(C.2)} \]
where $\hat{z}^*$ and $\hat{z}$ minimize the loss functions $L(z)$ and $\hat{L}(z)$, respectively, and $z_0 \in [\hat{z}, \hat{z}^*]$. Moreover, following $(\nabla^2 \hat{L}(z_0) - \nabla^2 L(z_0))\hat{z}^* = \nabla \hat{L}(\hat{z}^*) - \nabla L(\hat{z}^*)$, which is from

$$
\nabla \hat{L}(\hat{z}^*) = \left(\frac{1}{\lambda} ASST A^T + I\right)\hat{z}^* - y, \quad \nabla L(\hat{z}^*) = \left(\frac{1}{\lambda} AA^T + I\right)\hat{z}^* - y,
$$

and

$$
\nabla^2 \hat{L}(z_0) = \frac{1}{\lambda} ASST A^T + I, \quad \nabla^2 L(z_0) = \frac{1}{\lambda} AA^T + I,
$$

we can obtain

$$
\hat{L}(\hat{z}^*) + (\hat{z} - \hat{z}^*)^T (\nabla^2 \hat{L}(z_0) - \nabla^2 L(z_0))\hat{z}^* = \hat{L}(\hat{z}^*) + (\hat{z} - \hat{z}^*)^T (\nabla \hat{L}(\hat{z}^*) - \nabla L(\hat{z}^*)).
$$

Thus, considering that

$$
\hat{L}(\hat{z}^*) + (\hat{z} - \hat{z}^*)^T (\nabla \hat{L}(\hat{z}^*) - \nabla L(\hat{z}^*)) \leq \hat{L}(\hat{z}) + (\hat{z} - \hat{z}^*)^T \nabla \hat{L}(\hat{z}),
$$

which is derived by the fact $(\hat{z} - \hat{z}^*)^T \nabla L(\hat{z}^*) \geq 0$, and noting (C.2), we can gain

$$
\hat{L}(\hat{z}^*) + (\hat{z} - \hat{z}^*)^T (\nabla^2 \hat{L}(z_0) - \nabla^2 L(z_0))\hat{z}^* \leq \hat{L}(\hat{z}) - (\hat{z} - \hat{z}^*)^T \nabla L(z_0)(\hat{z} - \hat{z}^*).
$$

Further, by $\hat{L}(\hat{z}^*) \geq \hat{L}(\hat{z})$, we have

$$
(\hat{z} - \hat{z}^*)^T (\nabla^2 \hat{L}(z_0) - \nabla^2 L(z_0))\hat{z}^* \geq (\hat{z} - \hat{z}^*)^T \nabla^2 \hat{L}(z_0)(\hat{z} - \hat{z}^*),
$$

which together with

$$
(\hat{z} - \hat{z}^*)^T \nabla^2 \hat{L}(z_0)(\hat{z} - \hat{z}^*) \geq (\hat{z} - \hat{z}^*)^T \frac{1}{\lambda} ASST A^T (\hat{z} - \hat{z}^*)
$$

and (C.3) leads to

$$
(\hat{z} - \hat{z}^*)^T \left(\frac{1}{\lambda} AA^T - \frac{1}{\lambda} ASST A^T\right)\hat{z}^* \geq (\hat{z} - \hat{z}^*)^T \frac{1}{\lambda} ASST A^T (\hat{z} - \hat{z}^*).
$$

Thus, based on $A = USVT$, it is straightforward to get

$$
\frac{1}{\lambda^2} (\hat{z} - \hat{z}^*)^T (US\Sigma^2 U^T - US\Sigma V^T SS^T V\Sigma U^T) \hat{z}^* \geq \frac{1}{\lambda^2} (\hat{z} - \hat{z}^*)^T US\Sigma V^T SS^T V\Sigma U^T (\hat{z} - \hat{z}^*),
$$

which is also allowed to be rewritten as

$$
\frac{1}{\lambda^2} ([\Sigma U^T (\hat{z} - \hat{z}^*)]^T (I - V^T SS^T V) \Sigma U^T \hat{z}^* \geq \frac{1}{\lambda^2} [\Sigma U^T (\hat{z} - \hat{z}^*)]^T V^T SS^T V [\Sigma U^T (\hat{z} - \hat{z}^*)].
$$

(C.4)
Adding $\frac{1}{\lambda^2} [\Sigma U^T (\hat{z} - \hat{z}^*)]^T [\Sigma U^T (\hat{z} - \hat{z}^*)]$ to both sides of \((C.4)\) gives
\[
\frac{1}{\lambda^2} [\Sigma U^T (\hat{z} - \hat{z}^*)]^T (I - V^T S S^T V) \Sigma U^T \hat{z}^* + \frac{1}{\lambda^2} [\Sigma U^T (\hat{z} - \hat{z}^*)]^T (I - V^T S S^T V) [\Sigma U^T (\hat{z} - \hat{z}^*)] \geq \frac{1}{\lambda^2} [\Sigma U^T (\hat{z} - \hat{z}^*)]^T [\Sigma U^T (\hat{z} - \hat{z}^*)].
\]
\((C.5)\)

Taking the Euclidean norm on both sides of \((C.5)\), we obtain
\[
\frac{1}{\lambda^2} \|\Sigma U^T (\hat{z} - \hat{z}^*)\|_2 \|I - V^T S S^T V\|_2 \|\Sigma U^T \hat{z}^*\|_2 + \frac{1}{\lambda^2} \|\Sigma U^T (\hat{z} - \hat{z}^*)\|_2 \|I - V^T S S^T V\|_2 \|\Sigma U^T (\hat{z} - \hat{z}^*)\|_2 \geq \frac{1}{\lambda^2} \|\Sigma U^T (\hat{z} - \hat{z}^*)\|_2^2,
\]
which combined with Lemma \([C.2]\) indicates that
\[
\frac{1}{\lambda^2} \|\Sigma U^T (\hat{z} - \hat{z}^*)\|_2 + \frac{1}{\lambda^2} \|\Sigma U^T (\hat{z} - \hat{z}^*)\|_2 \geq \frac{1}{\lambda^2} \|\Sigma U^T (\hat{z} - \hat{z}^*)\|_2^2.
\]
\((C.6)\)

By rewriting \((C.6)\) as
\[
\frac{1}{\lambda^2} \|\Sigma U^T (\hat{z} - \hat{z}^*)\|_2 \leq \frac{\epsilon'}{1 - \epsilon'} \frac{1}{\lambda^2} \|\Sigma U^T (\hat{z} - \hat{z}^*)\|_2^2
\]
and considering the fact $\epsilon' < \frac{1}{2}$, we have
\[
\frac{1}{\lambda^2} \|\Sigma U^T (\hat{z} - \hat{z}^*)\|_2 \leq \frac{2\epsilon'}{\lambda} \|\Sigma U^T \hat{z}^*\|_2.
\]
Thus, setting $\epsilon = 2\epsilon'$, we get \((C.1)\). That is, \((2.13)\) is arrived.

**Appendix D Proof of Theorem 2.4**

The proof can be completed along the line of the proof of Theorem 6 in \([4]\). However, when we bound $\|R\|_2$ with
\[
R = (\lambda \Sigma^{-1} + \Sigma)^{-1} \Sigma (V^T S S^T V - I),
\]
Lemma \([C.2]\) is adopted but not the oblivious subspace embedding theorem \([5, \text{Theorem 5}]\), namely,
\[
\|R\|_2 \leq \|(\lambda \Sigma^{-1} + \Sigma)^{-1} \Sigma (V^T S S^T V - I)\|_2 \\
\leq \|(\lambda \Sigma^{-1} + \Sigma)^{-1} \Sigma\|_2 \|V^T S S^T V - I\|_2 \\
\leq \epsilon' \|(\lambda \Sigma^{-1} + \Sigma)^{-1} \Sigma\|_2 \quad \text{by Lemma \([C.2]\)} \\
\leq \epsilon',
\]
where $\epsilon'$ satisfies $\epsilon' = \frac{\epsilon}{2}$.
Appendix E  Proof of Theorem 3.1

The proof is similar to the one of Theorem 2.1 (see Appendix A), and we begin by presenting two lemmas.

Lemma E.1. Assume that the condition (2.6) and (3.2) hold. Then, for \( m = 1 \) and \( \pi_1^{\text{NOPL}} \) in (3.1), we have

\[
\sum_{i=1}^{\pi_1^{\text{NOPL}}} \frac{\|e_i\|_2^3}{p} = O_p(1),
\]

where \( e_i = (A_i A_i^T + \lambda I)\hat{z}^* - \bar{y} \) with \( \bar{y} = \lambda y \) and \( \hat{z}^* \) being as in (1.3).

Proof. Similar to the proof of Lemma A.1 based on (2.6), (3.1), (3.2), (A.2), and (A.3), we have

\[
\sum_{i=1}^{\pi_1^{\text{NOPL}}} \frac{\|e_i\|_2^3}{p} \leq \frac{\|\bar{y}\|_2^3 \sigma_i^2 (A_i A_i^T + \lambda I)}{p^3} \left( \sum_{i=1}^{\pi_1^{\text{NOPL}}} \frac{\|A_i A_i^T\|_2^2}{\pi_1^{\text{NOPL}}} \right) + 3\lambda \sum_{i=1}^{\pi_1^{\text{NOPL}}} \|A_i A_i^T\|_2 + \lambda^3 + \frac{\|\bar{y}\|_2^3}{p^3} \left( \sum_{i=1}^{\pi_1^{\text{NOPL}}} \frac{\|A_i A_i^T\|_2^2}{\pi_1^{\text{NOPL}}} \right)
\]

Proof. Similar to the proof of Lemma A.1 based on (2.6), (3.1), (3.2), (A.2), and (A.3), we have
\[ + 3\lambda^2 \sum_{i=1}^{p} \|A_i\|^2 + \lambda] + \frac{\|\tilde{y}\|^2}{p^3} + 3\|\tilde{y}\|^3 \sigma_n^2 (AA^T + \lambda I) (\frac{N_2}{N_1} (\sum_{i=1}^{p} |A_i|^2) + 2\lambda \sum_{i=1}^{p} \|A_i\|^2 + \lambda^2) + 3\|\tilde{y}\|^3 \sigma_n (AA^T + \lambda I) (\sum_{i=1}^{p} |A_i|^2 + \lambda) \text{ by (3.2)} = O_p(1), \text{ by (A.3) and (2.6)}, \]

where the first inequality is gained by replacing \(\pi_i\) in (A.2) with \(\pi_i^{NOLP}\). Then, (E.1) is obtained.

**Lemma E.2.** To the assumption of Lemma E.1, add that the condition (2.7) holds. Then, for \(m = 1\) and \(\pi_i^{NOLP}\) in (E.1), conditional on \(F_n\) and \(\tilde{\beta}\) in probability, we have

\[
\frac{\tilde{M}_A - M_A}{p} = O_p(1), \quad \text{by (A.3) and (2.6)}
\]

\[
\tilde{e}^* = O_p(r^{-1/2}),
\]

where \(M_A = AA^T + \lambda I\), \(\tilde{M}_A = \tilde{A} \tilde{S} \tilde{S}^T A^T + \lambda I\) with \(\tilde{S}\) constructed by \(\pi_i^{NOLP}\), and \(\tilde{e}^* = (\tilde{M}_A \tilde{z}^* - \tilde{y})\) with \(\tilde{y} = \lambda y\) and \(\tilde{z}^*\) being as in (1.3).

**Proof.** The proof can be completed similar to the proof of Lemma A.2. We only need to replace \(\pi_i\) with \(\pi_i^{NOLP}\), and note that

\[
\frac{1}{p^2} \sum_{i=1}^{p} \frac{A_i A_i^T A_i A_i^T}{\pi_i^{NOLP}} = \frac{1}{p^2} \left( \sum_{i=1}^{p} \frac{A_i A_i^T A_i A_i^T}{|\tilde{\beta}(i)| \|A_i\|^2} \right) \left( \sum_{i=1}^{p} |\tilde{\beta}(i)| \|A_i\|^2 \right) \leq \frac{N_2}{N_1 p^2} \left( \sum_{i=1}^{p} \frac{A_i A_i^T A_i A_i^T}{\|A_i\|^2} \right) \left( \sum_{i=1}^{p} \|A_i\|^2 \right) = \frac{N_2}{N_1 p^2} \left( \sum_{i=1}^{p} \|A_i\|^2 \right) = O_p(1), \text{ by (2.0) and (2.7)},
\]

\[
E(\tilde{M}_A | A) = E_{\tilde{\beta}}[E(M_A | A, \tilde{\beta})],
\]

\[
\text{Var}(\frac{(\tilde{M}_A - M_A)^\ell}{p} | A) = E_{\tilde{\beta}}[\text{Var}(\frac{(\tilde{M}_A - M_A)^\ell}{p} | A, \tilde{\beta})],
\]

\[
E(\tilde{e}_i | F_n) = E_{\tilde{\beta}}[E(\tilde{e}_i | F_n, \tilde{\beta})],
\]

\[
\text{Var}(\frac{\ell^T \tilde{e}^*}{p} | F_n) = E_{\tilde{\beta}}[\text{Var}(\frac{\ell^T \tilde{e}^*}{p} | F_n, \tilde{\beta})],
\]

35
where \( E_{\tilde{\beta}} \) denotes the expectation on \( \tilde{\beta} \).

**Remark E.1.** The results (E.2) and (E.3) still hold when \( \tilde{M}_A = AS^*S^TA^T + \lambda I \) with \( S^* \in \mathbb{R}^{p \times r} \) formed by \( \pi_i^{COL} \).

**Corollary E.1.** For \( S^* \in \mathbb{R}^{p \times r} \) formed by \( \pi_i^{COL} \), \( \tilde{z} = (AS^*S^TA^T + \lambda I)^{-1}\tilde{y} \) constructed by Algorithm 2 satisfies

\[
\|\tilde{z} - \hat{z}^*\|_2 = O_p(r_n^{-1/2}). \tag{E.5}
\]

**Proof.** Similar to (A.8), considering (2.7) and Remark E.1, we can get

\[
\tilde{z} - \hat{z}^* = -(MA)^{-1}\tilde{e} + O_p(r_n^{-1/2}),
\]

which suggests that (E.5) holds.

**Proof of Theorem 3.1** Similar to the proof of Theorem 2.1, noting (2.6), (2.7), (E.4), and Lemmas E.1 and E.2, and replacing \( \pi_i \) and \( e_i \) in the proof of Theorem 2.1 with \( \pi_i^{NOPL} \) and \( \tilde{e}_i \), respectively, we first get

\[
\hat{z}_1 - \hat{z}^* = -(MA)^{-1}\tilde{e} + O_p(r_n^{-1}), \tag{E.6}
\]

where

\[
\hat{z}_1 = (A\overline{S}\overline{S}^TA^T + \lambda I)^{-1}\overline{y} = \tilde{M}_A^{-1}\overline{y},
\]

\[
\overline{V}_c = \sum_{i=1}^{p} \frac{A_iA_i^T\hat{z}_i \hat{z}_i^TA_iA_i^T}{p^2\pi_i^{NOPL}} = O_p(1).
\]

To get (3.3), in the following, we need to further prove

\[
V_{OPL}^{-1/2}(\hat{z}_1 - \hat{z}^*) = -V_{OPL}^{-1/2}(MA)^{-1}(\overline{V}_c)^{1/2}(\overline{V}_c)^{-1/2}\tilde{e} + O_p(r_n^{-1/2}), \tag{E.7}
\]

where \( V_{OPL}^{-1/2}(MA)^{-1}(\overline{V}_c)^{1/2} \) satisfies

\[
V_{OPL}^{-1/2}(MA)^{-1}(\overline{V}_c)^{1/2}[V_{OPL}(MA)^{-1}(\overline{V}_c)^{1/2}]^T = I + O_p(r_n^{-1/2}). \tag{E.8}
\]
which indicates

\[
\frac{1}{p^2} \sum_{i=1}^{p} \frac{A_i A_i^T A_i A_i^T}{\pi_i^{\text{OPL}}} = \frac{1}{p^2} (\sum_{i=1}^{p} \frac{A_i A_i^T A_i A_i^T}{\|A_i\|^2}) (\sum_{i=1}^{p} \|\hat{\beta}_{\text{ls}(i)}\| \|A_i\|_2) \quad \text{by (2.10)}
\]

\[
\leq \frac{N_4}{N_3 p^2} (\sum_{i=1}^{p} \frac{A_i A_i^T A_i A_i^T}{\|A_i\|^2}) (\sum_{i=1}^{p} \|A_i\|_2^2) \quad \text{by (3.2)}
\]

\[
= \frac{N_4}{N_3 p^2} (\sum_{i=1}^{p} A_i A_i^T) (\sum_{i=1}^{p} \|A_i\|_2^2)
\]

\[
= O_p(1), \quad \text{by (2.6) and (2.7)}
\]

which indicates

\[
V_{\text{cOPL}} = \sum_{i=1}^{p} \frac{A_i A_i^T \tilde{z}_i \tilde{z}_i^T A_i A_i^T}{p^2 \pi_i^{\text{OPL}}} = O_p(1).
\]

From (2.7) and (3.2), it is evident to get

\[
V_{\text{OPL}} = \left(\frac{M_A}{p}\right)^{-1} V_{\text{cOPL}} \left(\frac{M_A}{p}\right)^{-1} = O_p(r^{-1}),
\]

which combined with (E.6) suggests that (E.7) holds, that is,

\[
V_{\text{OPL}}^{-1/2} (\hat{z}_1 - \tilde{z}) = -V_{\text{OPL}}^{-1/2} \left(\frac{M_A}{p}\right)^{-1} \tilde{z}_0 + O_p(r^{-1/2})
\]

\[
= -V_{\text{OPL}}^{-1/2} \left(\frac{M_A}{p}\right)^{-1} \left(\frac{\tilde{V}_c}{r}\right)^{1/2} (\frac{\tilde{V}_c}{r})^{-1/2} \tilde{z}_0 + O_p(r^{-1/2}).
\]

Now, we need to demonstrate that (E.8) also holds. Evidently, it suffices to show that

\[
V_{\text{OPL}}^{-1/2} \left(\frac{M_A}{p}\right)^{-1} \frac{\tilde{V}_c - V_{\text{cOPL}}}{r} \left(\frac{M_A}{p}\right)^{-1} V_{\text{OPL}}^{-1/2} = O_p(r_0^{-1/2}),
\]

because

\[
= V_{\text{OPL}}^{-1/2} \left(\frac{M_A}{p}\right)^{-1} \left(\frac{\tilde{V}_c}{r}\right)^{1/2} V_{\text{OPL}}^{-1/2} \left(\frac{M_A}{p}\right)^{-1} \left(\frac{\tilde{V}_c}{r}\right)^{1/2}
\]

\[
= V_{\text{OPL}}^{-1/2} \left(\frac{M_A}{p}\right)^{-1} \tilde{V}_c \left(\frac{M_A}{p}\right)^{-1} V_{\text{OPL}}^{-1/2}
\]

\[
= V_{\text{OPL}}^{-1/2} \left(\frac{M_A}{p}\right)^{-1} V_{\text{cOPL}} \left(\frac{M_A}{p}\right)^{-1} V_{\text{OPL}}^{-1/2}
\]

\[
= I + V_{\text{OPL}}^{-1/2} \left(\frac{M_A}{p}\right)^{-1} \tilde{V}_c - V_{\text{cOPL}} \left(\frac{M_A}{p}\right)^{-1} V_{\text{OPL}}^{-1/2}.
\]
Noting
\[ \tilde{V}_c = \frac{1}{p} \left( \sum_{i=1}^{p} A_i A_i^T \hat{\beta}_i \right) \left( \frac{1}{p} \left( \sum_{i=1}^{p} A_i A_i^T \right) \right)^{-1} \],
\[ V_{cOP L} = \frac{1}{p} \left( \sum_{i=1}^{p} A_i A_i^T \hat{\beta}_{rls(i)} \right) \left( \frac{1}{p} \left( \sum_{i=1}^{p} A_i A_i^T \right) \right)^{-1} \]
and the basic triangle inequality, we gain
\[ \| \tilde{V}_c - V_{cOP L} \|_2 = \| \Phi_1 \Phi_2 - \Phi_3 \Phi_4 \|_2 \]
\[ \leq \| \Phi_1 - \Phi_3 \|_2 \| \Phi_2 \|_2 + \| \Phi_2 - \Phi_4 \|_2 \| \Phi_3 \|_2. \]

Following (2.6), (3.2), (A.3), and (E.5), it is evident to gain
\[ \| \Phi_1 - \Phi_3 \|_2 \leq \frac{1}{p} \sum_{i=1}^{p} \frac{\lambda_i^2}{\| A_i \|_2^3} \left( \frac{1}{p} \left( \sum_{i=1}^{p} A_i A_i^T \right) \right)^{-1} \]
\[ \leq \frac{1}{p} \sum_{i=1}^{p} \frac{\lambda_i^2}{\| A_i \|_2^3} \left( \frac{1}{p} \left( \sum_{i=1}^{p} A_i A_i^T \right) \right)^{-1} \]
\[ \leq \frac{\lambda N_1}{p N_1} \sum_{i=1}^{p} \frac{\| A_i \|_2^3}{\| A_i \|_2^3} \left( \frac{1}{p} \left( \sum_{i=1}^{p} A_i A_i^T \right) \right)^{-1} \]
\[ = \| \tilde{z} - \hat{z}^* \|_2 \sum_{i=1}^{p} \frac{\lambda N_1}{p N_1} \frac{\| A_i \|_2^3}{\| A_i \|_2^3} = O_p(F_n(r_0^{-1/2})) \]
\[ \| \Phi_2 \|_2 \leq \frac{N_2 \| y \|_2}{p} \sum_{i=1}^{p} \| A_i \|_2^3 = O_p(1). \]
Similarly, we have \( \| \Phi_2 - \Phi_4 \|_2 = O_p(F_n(r_0^{-1/2})) \) and \( \| \Phi_3 \|_2 = O_p(1) \). Therefore, we get
\[ \| \tilde{V} - V_{cOP L} \|_2 = O_p(F_n(r_0^{-1/2})), \]
which combined with (2.7) and (E.10) yields (E.11). Putting the above discussions and the Slutsky’s Theorem together, the result (3.3) follows.

Appendix F Proof of Theorem 3.2

Before providing the proof of Theorem 3.2, we first present a lemma.
Lemma F.1. To the assumption of Lemma C.2, add that \( r \geq \frac{32 \epsilon s c_2 c_1}{3s c_1 c_3} \ln\left( \frac{4}{\delta} \right) \) with \( \epsilon, \delta \in (0, 1) \). Then, for any \( \epsilon, \hat{w}_t \) obtained from the \( t \)-th iteration of Algorithm 2 satisfies
\[
\left\| A^T \hat{w}_t - \frac{A^T w_t^*}{\lambda} \right\|_2 \leq \epsilon \left\| \frac{A^T w_t^*}{\lambda} \right\|_2,
\]
where \( w_t^* \) is the solution of
\[
\min_{w_t} \frac{1}{2\lambda} \left\| A^T w_t \right\|_2^2 + \frac{1}{2} \left\| w_t \right\|_2^2 - w_t^T b_t.
\]

Proof. The proof can be completed along the line of the proof of Theorem 2.3. Particularly, in this case, Lemma C.2 still holds for \( S = \tilde{S} \), where \( \tilde{S} \) is formed by \( \pi_{i}^{N O P L} \).

Proof of Theorem 3.2. At the \( t \)-th iteration, following the discussion in Remark 3.1 and (F.1), and setting
\[
\Delta_t^* = \frac{A^T w_t^*}{\lambda} = \frac{A^T \hat{z}^*}{\lambda} - \frac{A^T \hat{z}_{t-1}}{\lambda}
\]
and \( \hat{\Delta}_t = \frac{A^T \hat{w}_t}{\lambda} \) as the estimator of \( \Delta_t^* \), we can have
\[
\left\| \hat{\Delta}_t - \Delta_t^* \right\|_2 \leq \epsilon \left\| \Delta_t^* \right\|_2 \quad \text{by (F.1)}
\]
\[
= \epsilon \left\| \frac{A^T \hat{z}^*}{\lambda} - \frac{A^T \hat{z}_{t-1}}{\lambda} \right\|_2
\]
\[
= \epsilon \left\| \frac{A^T (\hat{z}_{t-2} + w_{t-1}')} \lambda - \frac{A^T (\hat{z}_{t-2} + \hat{w}_{t-1})}{\lambda} \right\|_2
\]
\[
\leq \epsilon \left\| \hat{\Delta}_{t-1} - \Delta_{t-1}^* \right\|_2 \leq \epsilon \left\| \Delta_{t-1}^* \right\|_2.
\]

As a result,
\[
\left\| \hat{\Delta}_m - \Delta_m^* \right\|_2 \leq \epsilon \left\| \hat{\Delta}_{m-1} - \Delta_{m-1}^* \right\|_2 \leq \epsilon^m \left\| \Delta_1^* \right\|_2
\]
\[
\leq \epsilon^m \left\| \frac{A^T \hat{z}^*}{\lambda} - \frac{A^T \hat{z}_0}{\lambda} \right\|_2
\]
\[
= \epsilon^m \left\| \frac{A^T \hat{z}^*}{\lambda} \right\|_2 = \epsilon^m \left\| \hat{\beta}_{rls} \right\|_2.
\]

Considering that \( \hat{\beta}_m - \hat{\beta}_{rls} = \hat{\Delta}_m - \Delta_m^* \), the conclusion is arrived.

References

[1] A. E. Hoerl, R. W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970) 55–67.

doi:https://doi.org/10.1080/00401706.1970.10488634
[2] A. N. Tihonov, Solution of incorrectly formulated problems and the regularization method, Soviet Math. Dokl. 5 (1963) 1035–1038.

[3] C. Saunders, A. Gammerman, V. Vovk, Ridge regression learning algorithm in dual variables, in: Proceedings of the 15th International Conference on Machine Learning, 1998, pp. 515–521.

[4] Y. Lu, P. S. Dhillon, D. P. Foster, L. H. Ungar, Faster ridge regression via the subsampled randomized hadamard transform, in: Proceedings of the 26th Annual Conference on Advances in Neural Information Processing Systems, Vol. 26, 2013, pp. 369–377.

[5] S. Chen, Y. Liu, M. R. Lyu, I. King, S. Zhang, Fast relative-error approximation algorithm for ridge regression, in: Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence, 2015, pp. 201–210.

[6] H. Avron, K. L. Clarkson, D. P. Woodruff, Sharper bounds for regularized data fitting, in: Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, Vol. 81, 2017, pp. 27 : 1–27 : 22.

[7] J. Wang, J. D. Lee, M. Mahdavi, M. Kolar, N. Srebro, Sketching meets random projection in the dual: a provable recovery algorithm for big and high-dimensional data, Electron. J. Stat. 11 (2) (2017) 4896–4944. doi:https://doi.org/10.1214/17-EJS1334SI

[8] A. Chowdhury, J. Yang, P. Drineas, An iterative, sketching-based framework for ridge regression, in: Proceedings of the 35th International Conference on Machine Learning, Vol. 80, 2018, pp. 989–998.

[9] J. Lacotte, M. Pilanci, Adaptive and oblivious randomized subspace methods for high-dimensional optimization: sharp analysis and lower bounds, arXiv preprint arXiv:2012.07054 (2020).

[10] L. Zhang, M. Mahdavi, R. Jin, T. Yang, S. Zhu, Recovering the optimal solution by dual random projection, in: Proceedings of the 26th Annual Conference on Learning Theory, Vol. 30, 2013, pp. 135–157.

[11] L. Zhang, M. Mahdavi, R. Jin, T. Yang, S. Zhu, Random projections for classification: a recovery approach, IEEE Trans. Inform. Theory 60 (11) (2014) 7300–7316. doi:https://doi.org/10.1109/TIT.2014.2359204
[12] O.-A. Maillard, R. Munos, Compressed least-squares regression, in: Proceedings of Advances in Neural Information Processing Systems, Vancouver, Canada, 2009, pp. 1213–1221.

[13] M. M. Fard, Y. Grinberg, J. Pineau, D. Precup, Compressed least-squares regression on sparse spaces, in: Proceedings of the 26th AAAI Conference on Artificial Intelligence, Vol. 26, 2012, pp. 1054–1060.

[14] A. Kabán, A new look at compressed ordinary least squares, in: 2013 IEEE 13th International Conference on Data Mining Workshops, 2013, pp. 482–488.

[15] A. Kabán, New bounds on compressive linear least squares regression, in: Proceedings of the 17th International Conference on Artificial Intelligence and Statistics, Vol. 33, 2014, pp. 448–456.

[16] G.-A. Thanei, C. Heinze, N. Meinshausen, Random projections for large-scale regression, in: Big and Complex Data Analysis, 2017, pp. 51–68.

[17] M. Slawski, Compressed least squares regression revisited, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Vol. 54, 2017, pp. 1207–1215.

[18] M. Slawski, On principal components regression, random projections, and column subsampling, Electron. J. Statist. 12 (2) (2018) 3673–3712. doi:https://doi.org/10.1214/18-EJS1486

[19] L. Mor-Yosef, H. Avron, Sketching for principal component regression, SIAM J. Matrix Anal. Appl. 40 (2) (2019) 454–485. doi:https://doi.org/10.1137/18M1188860

[20] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, D. P. Woodruff, Fast approximation of matrix coherence and statistical leverage, J. Mach. Learn. Res. 13 (1) (2012) 3475–3506.

[21] R. Zhu, P. Ma, M. W. Mahoney, B. Yu, Optimal subsampling approaches for large sample linear regression, arXiv preprint arXiv:1509.05111 (2015).

[22] H. Wang, R. Zhu, P. Ma, Optimal subsampling for large sample logistic regression, J. Amer. Statist. Assoc. 113 (522) (2018) 829–844. doi:https://doi.org/10.1080/01621459.2017.1292914
P. Ma, X. Zhang, X. Xing, J. Ma, M. Mahoney, Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms, in: Proceedings of the 23nd International Conference on Artificial Intelligence and Statistics, Vol. 108, 2020, pp. 1026–1035.

Y. Yao, H. Wang, Optimal subsampling for softmax regression, Statist. Papers 60 (2) (2019) 585–599. doi:https://doi.org/10.1007/s00362-018-01068-6

H. Wang, Y. Ma, Optimal subsampling for quantile regression in big data, Biometrika 108 (1) (2021) 99–112. doi:https://doi.org/10.1093/biomet/asaa043

H. Zhang, H. Wang, Distributed subdata selection for big data via sampling-based approach, Comput. Stat. Data Anal. 153 (2021) 107072. doi:https://doi.org/10.1016/j.csda.2020.107072

A. van der Vaart, Asymptotic Statistics, Cambridge University Press, London, 1998.

Y. Chen, N. Zhang, Optimal subsampling for large sample ridge regression, arXiv preprint arXiv:2204.04776 (2022).

F. Pukelsheim, Optimal Design of Experiments, Wiley, New York, 1993.

R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, 2012.

V. V. Buldygin, Y. V. Kozachenko, Sub-Gaussian random variables, Ukrainian Math. J. 32 (6) (1980) 483–489. doi:https://doi.org/10.1007/BF01087176

J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (5) (2008) 849–911. doi:https://doi.org/10.1111/j.1467-9868.2008.00674.x

A. Quarteroni, A. Valli, Numerical Approximation of Partial Differential Equations, Springer-Verlag, Berlin, 2008.

T. S. Ferguson, A Course in Large Sample Theory, Chapman and Hall, London, 1996.