Abstract

Ridge leverage scores provide a balance between low-rank approximation and regularization, and are ubiquitous in randomized linear algebra and machine learning. Deterministic algorithms are also of interest in the moderately big data regime, because deterministic algorithms provide interpretability to the practitioner by having no failure probability and always returning the same results. We provide provable guarantees for deterministic column sampling using ridge leverage scores. The matrix sketch returned by our algorithm is a column subset of the original matrix, yielding additional interpretability. Like the randomized counterparts, the deterministic algorithm provides \((1 + \epsilon)\) error column subset selection, \((1 + \epsilon)\) error projection-cost preservation, and an additive-multiplicative spectral bound. We also show that under the assumption of power-law decay of ridge leverage scores, this deterministic algorithm is provably as accurate as randomized algorithms. Lastly, ridge regression is frequently used to regularize ill-posed linear least-squares problems. While ridge regression provides shrinkage for the regression coefficients, many of the coefficients remain small but non-zero. Performing ridge regression with the matrix sketch returned by our algorithm and a particular regularization parameter forces coefficients to zero and has a provable \((1 + \epsilon)\) bound on the statistical risk. As such, it is an interesting alternative to elastic net regularization.

1 Introduction

Classical leverage scores quantify the importance of each column \(i\) for the range space of the sample-by-feature data matrix \(A \in \mathbb{R}^{n \times d}\). Classical leverage scores have been used in regression diagnostics, outlier detection, and randomized matrix algorithms (Velleman and Welsch, 1981; Chatterjee and Hadi, 1986; Drineas et al., 2008). Historically, leverage scores were used to select informative samples (rows, in our matrix orientation). More recently, as datasets with \(d > n\) have become more common, leverage scores have been used to select informative features (columns, in our matrix orientation). There are many different flavors of leverage scores, and we will focus on a variation called ridge leverage scores. However, to appreciate the advantages of ridge leverage scores, we also briefly review classical and rank-\(k\) subspace leverage scores.

Ridge leverage scores were introduced by Alaoui and Mahoney (2015) to give statistical bounds for the Nyström approximation for kernel ridge regression. Alaoui and Mahoney (2015) argue that ridge leverage scores provide the relevant notion of leverage in the context of kernel ridge regression. Ridge leverage scores have been successfully used in kernel ridge regression to approximate the symmetric kernel matrix \((\in \mathbb{R}^{n \times n})\) by selecting informative samples (Alaoui and Mahoney, 2015; Rudi et al., 2015). Cohen et al. (2017) provide a definition for ridge leverage scores for selecting informative features from the non-symmetric sample-by-feature data matrix \(A \in \mathbb{R}^{n \times d}\). The ridge
leverage score \( \tau_i(A) \) for the \( i^{th} \) column of \( A \) is,

\[
\tau_i(A) = a_i^T (AA^T + \lambda_2 I)^+ a_i,
\]

where the \( i^{th} \) column of \( A \) is an \((n \times 1)\)-vector denoted by \( a_i \), \( M^+ \) denotes the Moore-Penrose pseudoinverse of \( M \), and \( \lambda_2 \) is the regularization parameter. We will always choose \( \lambda_2 = \frac{1}{k} \|A - A_k\|_F^2 \), where \( A_k \) is the rank-\( k \) SVD approximation to \( A \), defined in Sec.1.2, because this choice of regularization parameter gives the stated guarantees. In contrast to ridge leverage scores, the rank-\( k \) subspace leverage score \( \tau(A_k) \) is,

\[
\tau(A_k) = a_i^T (A_k A_k^T + \lambda_2 I)^+ a_i,
\]

The classical leverage score is the ridge leverage score (and also the rank-1 subspace leverage score) evaluated at \( k = \text{rank}(A) \leq n \).

Ridge leverage scores and rank-\( k \) subspace leverage scores take two different approaches to mitigating the small singular values components of \( AA^T \) in classical leverage scores. Ridge leverage scores diminish the importance of small principle components through regularization, as opposed to rank-\( k \) subspace leverage scores, which omit the small principle components entirely. Papailiopoulos et al. (2014) argue that regularization is a more natural and stable alternative to omission. For randomized algorithms with ridge leverage score sampling, Cohen et al. (2017) prove bounds for the spectrum, column subset selection, and projection-cost preservation (counterparts to our Theorems 1, 2, and 3 for deterministic ridge leverage scores, respectively). The first and the last bounds hold for a weighted column subset of the full data matrix. These bounds require \( O(k \log(k/\delta)/\epsilon^2) \) columns, where \( \delta \) is the failure probability and \( \epsilon \) is the error.

In the "big data" era, much attention has been paid to randomized algorithms due to improved algorithm performance and ease of generalization to the streaming setting. However, for moderately big data (i.e. the feature set is too large for inspection by humans, but the algorithm performance is not a limitation), deterministic algorithms provide more interpretability to the practitioner than randomized algorithms, since they always provide the same results and have no failure probability.

The usefulness of deterministic algorithms has already been recognized. Papailiopoulos et al. (2014) introduce a deterministic algorithm for sampling columns from rank-\( k \) subspace leverage scores and provide a columns subset selection bound (the counterpart to our Theorem 2 for deterministic ridge leverage scores). McCurdy et al. (2017) prove a \((1+\epsilon)\) spectral bound for Papailiopoulos et al. (2014)’s deterministic algorithm and for random sampling with rank-\( k \) subspace leverage scores (the counterpart to our Theorem 1 for deterministic ridge leverage scores). One major drawback of using the rank-\( k \) subspace leverage scores is that their relative spectral bound is limited to the rank-\( k \) subspace projection of the column subset matrix \( C \) and the full data matrix \( A \), so to get a relative spectral bound on the complete subspace requires \( k = n \). A consequence of this is that projection-cost preservation also requires \( k = n \) (the counterpart to our Theorem 3). One advantage of using deterministic rather than randomized rank-\( k \) subspace leverage score algorithms is that under the condition of power-law decay in the sorted rank-\( k \) subspace leverage scores, the deterministic algorithm chooses fewer columns than random sampling with the same error for the column subset selection bound when \( \max((2k/\epsilon)^{1/2} - 1, (2k/(a-1)\epsilon))^{1/2} - 1, k < Ck \log(k/\delta)/\epsilon^2 \), where \( a \) is the decay power and \( C \) is an absolute constant Papailiopoulos et al. (2014) (this is the counterpart to our Theorem 5). In addition, Papailiopoulos et al. (2014) show that many real datasets display power-law decay in the sorted rank-\( k \) subspace leverage scores, illustrating the deterministic algorithm’s real-world utility.

Ridge regression (Hoerl and Kennard 1970) is a commonly used method to regularize ill-posed linear least-squares problems. The ridge regression minimization problem is, for outcome \( y \in \mathbb{R}^n \), features \( A \in \mathbb{R}^{n \times d} \), and coefficients \( x \in \mathbb{R}^d \),

\[
\mathcal{S} = \arg\min_x (||y - Ax||_2^2 + \lambda_2 ||x||_2^2)
\]

\[
= (A^T A + \lambda_2 I)^{-1} A^T y.
\]

(3)

where the regularization parameter \( \lambda_2 \) penalizes the size of the coefficients in the minimization problem. We will always choose \( \lambda_2 = \frac{1}{k} ||A - A_k||_F^2 \) for ridge regression with matrix \( A \).

In ridge regression, the underlying statistical model for data generation is,

\[
y = y^* + \sigma^2 \xi,
\]

(4)
where \( y^* = Ax^* \) is a deterministic linear function of the fixed design features \( A \) and \( \xi \sim N(0, I) \) is the random error. The mean squared error is a measure of statistical risk \( R(\hat{y}) \) for the squared error loss function and estimator \( \hat{y} \) and is,

\[
R(\hat{y}) = \frac{1}{n}E\xi ||\hat{y} - y^*||_2^2 .
\]

Ridge regression is often chosen over regression subset selection procedures for regularization because, as a continuous shrinkage method, it exhibits lower variability ([Breiman, 1996]). However many ridge regression coefficients can be small but non-zero, leading to a lack of interpretability for moderately big data (\( d > n \)). The lasso method ([Tibshirani, 1994]) provides continuous shrinkage and automatic feature selection using an \( L_1 \) penalty function instead of the \( L_2 \) penalty function in ridge regression, but for \( d > n \) case, lasso saturates at \( n \) features. The elastic net algorithm combines lasso (\( L_1 \) penalty function) and ridge regression (\( L_2 \) penalty function) for continuous shrinkage and automatic feature selection ([Zou and Hastie, 2005]).

1.1 Contributions

We explore deterministic ridge leverage score (DRLS) sampling for matrix approximation and for feature selection in concert with ridge regression. This work has two main motivations: (1) the advantages of ridge leverage scores over rank-\( k \) subspace leverage scores, and (2) the advantages of deterministic algorithms in some practical settings. This work complements [Papailiopoulos et al., 2014], who considered deterministic rank-\( k \) subspace leverage sampling and experiments on real data, but did not consider DRLS sampling or uses beyond matrix approximation. This work also complements [Cohen et al., 2017], who considered randomized RLS sampling but did not consider DRLS sampling, the uses of RLS sampling beyond matrix approximation (e.g. ridge regression), or experiments on real data.

We introduce a deterministic algorithm (Algorithm 1) for ridge leverage score sampling inspired by the deterministic algorithm for rank-\( k \) subspace leverage score sampling ([Papailiopoulos et al., 2014]). By using ridge leverage scores instead of rank-\( k \) subspace scores in the deterministic algorithm, we prove significantly better bounds for the column subset matrix \( C \) (see Table 1 for a comparison). We prove that the same additive-multiplicative spectral bounds (Theorem 1), \( (1 + \epsilon) \) columns subset selection (Theorem 2), and \( (1 + \epsilon) \) projection-cost preservation (Theorem 3) hold for DRLS column sampling as for random sampling as in [Cohen et al., 2017]. We show that under the condition of power-law decay in the ridge leverage scores, the deterministic algorithm chooses fewer columns than random sampling with the same error when \( \max((4k/\epsilon)^2 - 1, (4k/((\alpha - 1)\epsilon))/\epsilon^2 - 1, k) < Ck \log(k/\delta)/\epsilon^2 \), where \( \alpha \) is the decay power and \( C \) is an absolute constant (Theorem 5).

We combine deterministic ridge leverage score column subset selection with ridge regression for a particular value of the regularization parameter, providing automatic feature selection and continuous shrinkage. This procedure has a provable \((1 + \epsilon)\) bound on the statistical risk (Theorem 4). The proof techniques are such that a \((1 + \epsilon)\) bound on the statistical risk also holds for randomized ridge leverage score sampling. Our ridge regression theorem is novel to both deterministic and randomized sampling with ridge leverage scores (as far as we know, this has never been considered for any leverage score), another demonstrable advance of the state of the art, and one of our main results.

We also provide a proof-of-concept illustration on real biological data, with figures included in the Supplementary Materials. Our real-data illustration makes a strong case for the empirical usefulness of the DRLS algorithm and bounds. The real data exhibits striking power law decay of the ridge leverage scores (Figure 7), justifying the assumptions underlying the use of DRLS sampling (Theorem 3).

Our work is triply beneficial from the interpretability standpoint; it is deterministic, it chooses a subset of representative columns, and it comes with four desirable error guarantees for all rank-\( k \), three of which stem from naturalness of the low-rank ridge regularization.

1.2 Notation

The singular value decomposition (SVD) of any complex matrix \( A \) is \( A = USV^† \), where \( U \) and \( V \) are square unitary matrices \((U^†U = UU^† = I, V^†V = VV^† = I)\), \( \Sigma \) is a rectangular diagonal matrix with real non-negative non-increasingly ordered entries. \( U^† \) is the complex conjugate and
A

Algorithm 1. The DRLS algorithm selects for the submatrix \( C \) all columns \( i \) with ridge leverage score \( \bar{\tau}_i(A) \) above a threshold \( \bar{\tau} \), determined by the error tolerance \( \epsilon \). This algorithm is deeply indebted to the deterministic algorithm of Papailiopoulos et al. (2014). It substitutes ridge leverage scores for rank-\( k \) subspace scores, and has a different stopping parameter. The algorithm is as follows.

1. Choose the error tolerance, \( \epsilon \).
2. For every column \( i \), calculate the ridge leverage scores \( \bar{\tau}_i(A) \) (Eqn. 7).
3. Sort the columns by \( \bar{\tau}_i(A) \), from largest to smallest. The sorted column indices are \( \pi_i \).
4. Define an empty set \( \Theta = \{ \} \). Starting with the largest sorted column index \( \pi_0 \), add the corresponding column index \( i \) to the set \( \Theta \), in decreasing order, until

\[
\sum_{i \in \Theta} \bar{\tau}_i(A) > \bar{\tau} - \epsilon,
\]

and then stop. Note that \( \bar{\tau} = \sum_{i=1}^{d} \bar{\tau}_i(A) \leq 2k \) (see Sec. 5.2 for proof). It will be useful to define \( \bar{\tau} = \sum_{i \in \Theta} \bar{\tau}_i(A) \). Eqn. 6 can equivalently be written as \( \epsilon > \bar{\tau} \).
5. If the set size \( |\Theta| < k \), continue adding columns in decreasing order until \( |\Theta| = k \).
6. The leverage score \( \bar{\tau}_i(A) \) of the last column \( i \) included in \( \Theta \) defines the leverage score threshold \( \tau \).
7. Introduce a rectangular selection matrix \( S \) of size \( d \times |\Theta| \). If the column indexed by \( (i, \pi_i) \) is in \( \Theta \), then \( S_{i,\pi_i} = 1 \). \( S_{i,\pi_i} = 0 \) otherwise. The DRLS submatrix is \( C = AS \).
Note that when the ridge leverage scores on either side of the threshold are not equal, the algorithm returns a unique solution. Otherwise, there are as many solutions as there are columns with equal ridge leverage scores at the threshold.

Algorithm requires $O(\min(d, n)nd)$ arithmetic operations.

3 Approximation Guarantees

3.1 Bounds for DRLS

We derive a new additive-multiplicative spectral approximation bound (Eqn. 7) for the square of the submatrix $C$ selected with DRLS.

**Theorem 1.** Additive-Multiplicative Spectral Bound: Let $A \in \mathbb{R}^{n \times d}$ be a matrix of at least rank $k$ and $\bar{\tau}_i(A)$ be defined as in Eqn. 7. Construct $C$ following the DRLS algorithm described in Sec. 2. Then $C$ satisfies,

$$(1 - \epsilon) A A^T - \epsilon_k^2 ||A_{\backslash k}||_F^2 I \preceq C C^T \preceq A A^T. \tag{7}$$

The symbol $\preceq$ denotes the Loewner partial ordering which is reviewed in Sec 5.1 (see Horn and Johnson 2013) for a thorough discussion).

Conceptually, the Loewner ordering in Eqn. 7 is the generalization of the ordering of real numbers (e.g. $1 < 1.5$) to Hermitian matrices. Statements of Loewner ordering are quite powerful; important consequences include inequalities for the eigenvalues. We will use Eqn. 7 to prove Theorems 2, 3 and 4. Note that our additive-multiplicative bound holds for an un-weighted column subset of $A$.

**Theorem 2.** Column Subset Selection: Let $A \in \mathbb{R}^{n \times d}$ be a matrix of at least rank $k$ and $\bar{\tau}_i(A)$ be defined as in Eqn. 7. Construct $C$ following the DRLS algorithm described in Sec. 2. Then $C$ satisfies,

$$||A - C C^+ A||_F^2 \leq ||A - \Pi_{C, k}^C(A)||_F^2 \leq (1 + 4\epsilon) ||A_{\backslash k}||_F^2, \tag{8}$$

with $0 < \epsilon < \frac{1}{2}$ and where $\Pi_{C, k}^C(A) = (C^+ C)_k$ is the best rank-$k$ approximation to $A$ in the column space of $C$ with the respect to the Frobenius norm.

Column subset selection algorithms are widely used for feature selection for high-dimensional data, since the aim of the column subset selection problem is to find a small number of columns of $A$ that approximate the column space nearly as well as the top $k$ singular vectors.

**Theorem 3.** Rank-$k$ Projection-Cost Preservation: Let $A \in \mathbb{R}^{n \times d}$ be a matrix of at least rank $k$ and $\bar{\tau}_i(A)$ be defined as in Eqn. 7. Construct $C$ following the DRLS algorithm described in Sec. 2. Then $C$ satisfies, for any rank-$k$ orthogonal projection $X \in \mathbb{R}^{n \times n},$

$$(1 - \epsilon)||A - X A||_F^2 \leq ||C - XC||_F^2 \leq ||A - X A||_F^2. \tag{9}$$

To simplify the bookkeeping, we prove the lower bound of Theorem 3 with $(1 - \alpha \epsilon)$ error ($\alpha = 2(2 + \sqrt{2})i$) and assume $0 < \epsilon < \frac{1}{2}$.

Projection-cost preservation bounds were formalized recently in [Feldman et al., 2013] and [Cohen et al., 2015]. Bounds of this type are important because it means that low-rank projection problems can be solved with $C$ instead of $A$ while maintaining the projection cost. Furthermore, the projection-cost preservation bound has implications for $k$-means clustering, because the $k$-means objective function can be written in terms of the orthogonal rank-$k$ cluster indicator matrix [Boutsidis et al., 2009]. Note that our rank-$k$ projection-cost preservation bound holds for an un-weighted column subset of $A$.

A useful lemma on an approximate ridge leverage score kernel comes from combining Theorem 1 and 3.

**Lemma 1.** Approximate ridge leverage score kernel: Let $A \in \mathbb{R}^{n \times d}$ be a matrix of at least rank $k$ and $\bar{\tau}_i(A)$ be defined as in Eqn. 7. Construct $C$ following the DRLS algorithm described in

\[\text{Thanks to Michael Mahoney for this point.}\]
Let $\alpha$ be the coefficient in the lower bound of Theorem 3 and assume $0 < \epsilon < \frac{1}{2}$. Let $K(M) = (MM^T + \frac{1}{\lambda}||M||^2 I)^+$ for matrix $M \in \mathbb{R}^{n \times d}$. Then $K(C)$ and $K(A)$ satisfy,

$$K(A) \preceq K(C) \preceq \frac{1}{1 - (\alpha + 1)\epsilon} K(A).$$ (10)

**Theorem 4.** Approximate Ridge Regression with DRLS: Let $A \in \mathbb{R}^{n \times d}$ be a matrix of at least rank $k$ and $\tilde{\tau}_i(A)$ be defined as in Eqn. [4]. Construct $C$ following the DRLS algorithm described in Sec. 2.1 let $\alpha$ be the coefficient in the lower bound of Theorem 3 and assume $0 < \epsilon < \frac{1}{2\alpha} < \frac{1}{2}$. Choose the regularization parameter $\lambda_2 = \frac{||M||_F^2}{k}$ for ridge regression with a matrix $M$ (Eqn. 3). Under these conditions, the statistical risk $R(y_C)$ of the ridge regression estimator $\hat{y}_C$ is bounded by the statistical risk $R(y_A)$ of the ridge regression estimator $\hat{y}_A$:

$$R(y_C) \leq (1 + \beta \epsilon)R(y_A),$$ (11)

where $\beta = \frac{2\alpha(-1 + 2\alpha + 3\alpha^2)}{(1 - \alpha)^2}$.

Theorem 4 means that there are bounds on the statistical risk for substituting the DRLS selected column subset matrix for the complete matrix when performing ridge regression with the appropriate regularization parameter. Performing ridge regression with the column subset effectively forces coefficients to be zero and adds the benefits of automatic feature selection to the $L_2$ regularization problem. We also note that the proof of Theorem 4 relies only on Theorem 1 and Theorem 3 and facts from linear algebra, so a randomized selection of weighted column subsets that obey similar conditions, the statistical risk $R(y_C)$ of the ridge regression estimator $\hat{y}_C$ has the following bound in terms of the risk of ridge regression $\hat{y}_A$:

$$\hat{y}^E = \arg\min_x \left( ||y - A\hat{x}||^2 + \frac{1}{k}||A_{\setminus \hat{C}}||_F^2 ||\hat{x}||_2^2 + \lambda_1 \sum_{j=1}^d |\hat{x}_j| \right).$$ (12)

The risk of elastic net $R(\hat{y}^E)$ has the following bound in terms of the risk of ridge regression $R(\hat{y}_A)$:

$$R(\hat{y}^E) = A\hat{x}^E = R(\hat{y}_A) + \lambda_1 \frac{4d||A||_F^2}{1 - 2\alpha ||A_{\setminus \hat{C}}||_F^2}.$$ (13)

This comes from a slight re-working of Theorem 3.1 of [Zou and Zhang, 2009]. The bounds for the elastic net risk and $R(\hat{y}_C)$ are comparable when $\lambda_1 \approx \frac{2\alpha}{k} ||A_{\setminus \hat{C}}||_F^2 \frac{R(\hat{y}_A)}{4d||A||_F^2}$.

Ridge regression is a special case of kernel ridge regression with a linear kernel. While previous work in kernel ridge regression has considered the use of ridge leverage scores to approximate the symmetric kernel matrix by selecting a subset of $n$ informative samples [Alaoui and Mahoney, 2015; Rudi et al., 2015], to our knowledge, no previous work has used ridge leverage scores to approximate the symmetric kernel matrix using ridge leverage scores to select a subset of the $f$ informative features (after the feature mapping of the $d$-dimensional data points). The latter case would be the natural generalization of Theorem 4 to non-linear kernels, and remains an interesting open question. Lastly, we note that placing statistical assumptions on $A$ in the spirit of [Rudi et al., 2015] may lead to an improved bound for random designs for $A$.

**Theorem 5.** Ridge Leverage Power-law Decay: Let $A \in \mathbb{R}^{n \times d}$ be a matrix of at least rank $k$ and $\tilde{\tau}_i(A)$ be defined as in Eqn. 4. Furthermore, let the ridge leverage scores exhibit power-law decay in the sorted column index $\pi_i$,

$$\tilde{\tau}_{\pi_i}(A) = \pi_i^{-\alpha} \tau_{\pi_i}(A) \quad \alpha > 1.$$ (14)

Construct $C$ following the DRLS algorithm described in Sec. 2.1. The number of sample columns selected by DRLS is,

$$|\Theta| \leq \max \left( \left( \frac{4k}{\epsilon} \right)^{\frac{1}{\alpha}} - 1, \left( \frac{4k}{(\alpha - 1)\epsilon} \right)^{\frac{1}{\alpha - 1}} - 1, k \right).$$ (15)
Theorem 3 of Papailiopoulos et al. (2014) introduces the concept of power-law decay behavior for leverage scores for rank-$k$ subspace leverage scores. Our Theorem 5 is an adaptation of Papailiopoulos et al. (2014)’s Theorem 3 for ridge leverage scores. An obvious extension of Eqn. 7 is the following bound, 

$$
(1 - \epsilon)\mathbf{A}\mathbf{A}^T - \frac{\epsilon}{k}||\mathbf{A}_{(k)}||^2_F \mathbf{I} \preceq \mathbf{C}\mathbf{C}^T \preceq (1 + \epsilon)\mathbf{A}\mathbf{A}^T + \frac{\epsilon}{k}||\mathbf{A}_{(k)}||^2_F \mathbf{I},
$$

(16)

which also holds for $\mathbf{C}$ selected by ridge leverage random sampling methods with $O\left(\frac{k}{\epsilon^2} \ln \left(\frac{k}{\delta}\right)\right)$ weighted columns and failure probability $\delta$ (Cohen et al., 2017). Thus, DRLS selects fewer columns with the same accuracy $\epsilon$ in Eqn. 16 for power-law decay in the ridge leverage scores when,

$$
\max\left(\left(\frac{4k}{\epsilon^2}\right)^{\frac{1}{a}} - 1, \left(\frac{4k}{\epsilon^2}(a-1)\right)^{\frac{1}{a}} - 1, k\right) < C \frac{k}{\epsilon^2} \ln \left(\frac{k}{\delta}\right),
$$

(17)

where $C$ is an absolute constant. In particular, when $a \geq 2$, the number of columns deterministically sampled is $O(k)$.

4 Biological Data Illustration

We provide a biological data illustration of ridge leverage scores and ridge regression with multi-omic data from lower-grade glioma (LGG) tumor samples collected by the TCGA Research Network (http://cancergenome.nih.gov/). Diffuse lower-grade gliomas are infiltrative brain tumors that occur most frequently in the cerebral hemisphere of adults.

The data is publicly available and hosted by the Broad Institute’s GDAC Firehose (Broad Institute of MIT and Harvard, 2016). We download the data using the R tool TCGA2STAT (Wan et al., 2016). TCGA2STAT imports the latest available version-stamped standardized Level 3 dataset on Firehose. The data collection and data platforms are discussed in detail in the original paper (The Cancer Genome Atlas Research Network, 2015).

We use the following multi-omic data types: mutations ($d = 4845$), DNA copy number (alteration ($d = 22618$) and variation ($d = 22618$)), messenger RNA (mRNA) expression ($d = 20501$), and microRNA expression ($d = 1046$). Methylation data is also available, but we omit it due to memory constraints. The mRNA and microRNA data is normalized. DNA copy number (variation and alteration) has an additional pre-processing step: the segmentation data reported by TCGA is turned into copy number using the R tool CNtools (Zhang, 2015) that is imbedded in TCGA2STAT. The mutation data is filtered based on status and variant classification and then aggregated at the gene level (Wan et al., 2016).

There are 280 tumor samples and $d = 71628$ multi-omic features in the downloaded dataset. We are interested in performing ridge regression with the biologically meaningful outcome variables relating to mutations of the "IDH1" and "IDH2" gene and deletions of the "1p/19q" chromosome arms ("codel"). These variables were shown to be predictive of favorable clinical outcomes and can be found in the supplemental tables (The Cancer Genome Atlas Research Network, 2015). We restrict to samples with these outcome variables (275 tumor samples), and we drop an additional sample ("TCGA-CS-4944") because it is an outlier with respect to the $k = 3$ SVD projection of the samples. This leaves a total of 274 tumor samples with outcome variables "IDH" (a mutation in either "IDH1" or "IDH2") and "codel" for the analysis.

Lastly, we drop all multi-omic features that have zero columns and greater than 10% missing data on the 274 tumor samples. We replace missing values with the mean of the column. This leaves a final multi-omic feature set of $d = 68522$ for the 274 tumor samples. Our final matrix $\mathbf{A} \in \mathbb{R}^{274 \times 68522}$ is column mean-centered. Figure 1 shows a pie chart of the breakdown of the final matrix $\mathbf{A}$’s multi-omic feature types.

4.1 Ridge leverage score sampling

Figure 2 shows the spectrum of eigenvalues of $\mathbf{A}\mathbf{A}^T$ for LGG. The eigenvalues range of multiple orders of magnitude. We choose $k = 3$ for the DRLS algorithm because these components are...
meaningful for the "IDH" and "codel" outcome variables (see Figures 3, 4, and 5). The top three components capture 79% of the Frobenius norm $|A_k|^F$. Applying the DRLS algorithm with $k = 3, \epsilon = 0.1$ leads to $|\Theta| = 1512$, selecting approximately 0.02% of the total multi-omic features for the column subset matrix $C$. The majority of the features selected are mRNA (1473 features), and the remainder are microRNA (39 features). Figure 6 shows the relationship between the number of columns kept, $|\Theta|$, and $\bar{\epsilon} = \sum_{i \in \Theta} \bar{t}_i(A)$ for the $k = 3$ ridge leverage scores. Only a small error penalty is incurred by a dramatic reduction in the number of columns kept according to Algorithm 1.

Figure 7 shows the power-law decay of the LGG $k = 3$ ridge leverage scores with sorted column index. This LGG multi-omic data example shows that ridge leverage score power-law decay occurs in the wild. Figure 8 shows a histogram of the ratio of $|C - XC|^F/||A - XA||^F_2$ for 1000 random rank-$k = 3$ orthogonal projections $X$. The projections are chosen as the first 3 directions from an orthogonal basis randomly selected with respect to the Haar measure for the orthogonal group $O(n)$ (Mezzadri, 2006). This confirms that the projection cost empirically has very small error. Lastly, Figure 2 illustrates the $k = 3$ ridge leverage score regularization of the classical leverage score for the LGG multi-omic features. As expected, many of the columns’ ridge leverage scores exhibit shrinkage when compared to the classical leverage scores. Table 2 includes ratios derived from the full data matrix $A$ and the column subset matrix $C$ selected by the DRLS algorithm with $k = 3, \epsilon = 0.1$.

### 4.2 Ridge regression with ridge leverage score sampling

We perform ridge regression with the appropriate regularization parameter for two biologically meaningful outcome variables; the first is whether either the "IDH1" or the "IDH2" gene is mutated and the second whether the "1p/19q" chromosome arms have deletions ("codel"). We encode the status of each event as ±1. Figures 3, 4, and 5 show the top three SVD projections for the tumor samples, colored by the combined status for "IDH" and "codel". No tumor samples have the "1p/19q" codeletion and no "IDH" mutation. Visual inspection of the SVD plot confirms that this is a reasonable regression problem for "IDH" and a difficult regression problem for "codel"; also, logistic regression for these two outcome variables.

Lastly, we simulate 274 samples $y$ according to the linear model (Eqn. 4), where $y^* = Ax^*$, the coefficients $x^* \sim \mathcal{N}(0, I)$, and $A$ is the LGG multi-omic feature matrix. We choose $\sigma^2 = \{10^{-3}, 1.10^3\}$. We perform ridge regression with $A$ and then again with $C$ in accordance with Theorem 4. We calculate the risks $\mathcal{R}(\hat{y}_A)$ and $\mathcal{R}(\hat{y}_C)$ and find that Theorem 4 is not violated. Table 2 shows the risk ratios $\mathcal{R}(\hat{y}_C)/\mathcal{R}(\hat{y}_A)$ along with other relevant ratios for the ridge leverage scores.

| Algorithm 1 | $\text{ave}(\Sigma_C^2/\Sigma^2)$ | $||C_{\backslash k}||^2_F/||A_{\backslash k}||^2_F$ | $\text{ave}(\Sigma_C^2/\Sigma_C^2)$ | $\sigma^2$ | $\mathcal{R}(\hat{y}_C)/\mathcal{R}(\hat{y}_A)$ |
|-------------|------------------|------------------|------------------|--------|------------------|
| $k = 3, \epsilon = 0.1$ | 0.85 | 0.97 | 1.03 | $10^{-3}$ | 0.99 |
| | | | | $10^0$ | 0.99 |
| | | | | $10^3$ | 0.99 |

### Acknowledgements

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number F32HG008713. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. SRM thanks Michael Mahoney, Ahmed El Alaoui, Elaine Angelino, and Kai Rothauge for thoughtful comments and the Barcellos and Pachter Labs.
Supporting Information

Software in the form of python and R code is available at https://github.com/srmcc/deterministic-ridge-leverage-sampling Code for downloading the data and reproducing all of the figures is included. Proofs and figures are included in the Supplementary Material.

References

Ahmed El Alaoui and Michael W. Mahoney. 2015. Fast Randomized Kernel Ridge Regression with Statistical Guarantees. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS’15). MIT Press, Cambridge, MA, USA, 775–783. http://dl.acm.org/citation.cfm?id=2969239.2969326 http://arxiv.org/abs/1411.0306.

Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. 2011. Near-Optimal Column-Based Matrix Reconstruction. arXiv:1103.0995 [cs] (March 2011). http://arxiv.org/abs/1103.0995.

Christos Boutsidis, Petros Drineas, and Michael W Mahoney. 2009. Unsupervised Feature Selection for the k-means Clustering Problem. In Advances in Neural Information Processing Systems 22, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Eds.). Curran Associates, Inc., 153–161. http://papers.nips.cc/paper/3724-unsupervised-feature-selection-for-the-k-means-clustering-problem.pdf.

Leo Breiman. 1996. Heuristics of instability and stabilization in model selection. The Annals of Statistics 24, 6 (Dec. 1996), 2350–2383. https://doi.org/10.1214/aos/1032181158

Broad Institute of MIT and Harvard. 2016. Broad Institute TCGA Genome Data Analysis Center (2016): Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run. (Jan. 2016). https://doi.org/10.7908/C11G0KM9 Dataset.

Samprit Chatterjee and Ali S. Hadi. 1986. Influential Observations, High Leverage Points, and Outliers in Linear Regression. Statist. Sci. 1, 3 (Aug. 1986), 379–393. https://doi.org/10.1214/ss/1177013622

Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. 2015. Dimensionality Reduction for k-Means Clustering and Low Rank Approximation. In Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing (STOC ’15). ACM, New York, NY, USA, 163–172. https://doi.org/10.1145/2746539.2746569

Michael B. Cohen, Cameron Musco, and Christopher Musco. 2017. Input Sparsity Time Low-rank Approximation via Ridge Leverage Score Sampling. In Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA ’17). Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1758–1777. http://dl.acm.org/citation.cfm?id=3039686.3039801

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. 2008. Relative-Error $CUR$ Matrix Decompositions. SIAM J. Matrix Anal. Appl. 30, 2 (Sept. 2008), 844–881. https://doi.org/10.1137/07070471X

C. Eckart and G. Young. 1936. The approximation of one matrix by another of lower rank. Psychometrika 1 (1936), 211–218. https://doi.org/10.1007/BF02283367

D. Feldman, M. Schmidt, and C. Sohler. 2013. Turning Big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering. In Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, 1434–1453. https://doi.org/10.1137/1.9781611973105.103

Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12, 1 (Feb. 1970), 55–67. https://doi.org/10.1080/00401706.1970.10488834
5 Proofs

5.1 Brief Review

See Horn and Johnson (2013) for a linear algebra review. A square complex matrix $F$ is Hermitian if $F = F^\dagger$. Symmetric positive semi-definite (SPSD) matrices are Hermitian matrices. The set of $n \times n$ Hermitian matrices is a real linear space. As such, it is possible to define a partial ordering (also called a Loewner partial ordering, denoted by $\preceq$) on the real linear space. One matrix is "greater" than another if their difference lies in the closed convex cone of SPSD matrices. Let $F$ and $G$ be Hermitian and the same size, and $x$ a complex vector of compatible dimension. Then,

$$F \preceq G \iff x^\dagger F x \leq x^\dagger G x \quad \forall x \neq 0.$$  \hspace{1cm} (18)

If $F$ is Hermitian with smallest and largest eigenvalues $\lambda_{\min}(F), \lambda_{\max}(F)$, respectively, then,

$$\lambda_{\min}(F)I \preceq F \preceq \lambda_{\max}(F)I.$$  \hspace{1cm} (19)

Let $F$, $G$ be Hermitian and the same size, and let $H$ be any complex rectangular matrix of compatible dimension. The conjugation rule is,

If $F \preceq G$, then $HFH^\dagger \preceq HGH^\dagger$.  \hspace{1cm} (20)
In addition, let \( \lambda_i(F) \) and \( \lambda_i(G) \) be the non-decreasingly ordered eigenvalues of \( F, G \). Then,
\[
\text{If } F \preceq G, \text{ then } \forall i, \lambda_i(F) \leq \lambda_i(G).
\] (21)
Since the trace of a matrix \( F \) is the sum of its eigenvalues, \( \text{tr } F = \sum_i \lambda_i(F) \), and the Loewner ordering implies the ordering of eigenvalues (Eqn. 21), the Loewner ordering also implies the ordering of their sum,
\[
\text{If } F \preceq G, \text{ then } \text{tr } F \leq \text{tr } G.
\] (22)

5.2 Proof of ridge leverage score sum

The sum of ridge leverage scores is,
\[
\sum_{i=1}^{d} \tau_i(A) = \text{tr} A^T \left( A A^T + \frac{1}{k} \|A - A_k\|_F^2 I \right) ^+ A
\]
where \( \Sigma \) is diagonal and \( \Sigma^2_{i,i} = \Sigma_{i,i}^2 + \frac{1}{k} \|A_{i,k}\|_F^2 \). We split the sum into two parts,
\[
\sum_{i=1}^{k} \Sigma^2_{i,i} + \sum_{i=k+1}^{n} \Sigma^2_{i,i} \leq k + k.
\] (24)
This proof is due to Cohen et al. (2017).

5.3 Proof of Theorem 1

The upper bound in Eqn. 7 in Theorem 1 follows from the fact that \( 0 \preceq I - S S^T \) and the conjugation rule (Eqn. 20),
\[
0 \preceq A(I - S S^T)A^T = A A^T - C C^T.
\] (25)
This upper bound is true for any column selection of \( A \).

For the lower bound in Eqn. 7 consider the quantity,
\[
Y = \Sigma^{-1} U^T A(I - S S^T) A^T U \Sigma^{-1} = \Sigma^{-1} U V^T (I - S S^T) V \Sigma \Sigma^{-1}.
\]
By the conjugation rule (Eqn. 20) on Eqn. 25, \( 0 \preceq Y \), so \( Y \) is S.P.S.D. By the construction of DRLS (Eqn. 6) \( \text{tr } Y = \sum_{i \in \Theta} \sum_{i=1}^{n} \sum_{i \in \Theta} \sum_{i \in \Theta} \Sigma_{i,j}^2 V_{i,j}^2 = \epsilon < \epsilon \), and because \( Y \) is S.P.S.D., \( \lambda_{\max}(Y) \leq \text{tr } Y \). By Eqn. 19 and the previous facts, \( Y \preceq \lambda_{\max}(Y) I \preceq \epsilon I \). As a result of the conjugation rule applied to this upper bound,
\[
U \Sigma Y \Sigma U^T = A A^T - C C^T \preceq \epsilon U \Sigma^2 U^T = \epsilon \left( \Sigma^2 + \frac{1}{k} \|A_{i,k}\|_F^2 I \right) ,
\]
and rearrangement leads to the lower bound of Eqn. 7.

5.4 Proof of Theorem 2

To prove Eqn. 8 we will use the following lemma.

Lemma 2. [Boutsidis et al. (2017)] Lemma 3.1, Eqn. 3.2, specialized to the Frobenius norm: Consider \( A = AZZ^T + E \in \mathbb{R}^{n \times d} \), where \( Z \in \mathbb{R}^{n \times k} \) and \( Z^T Z = I_k \). Let \( S \in \mathbb{R}^{n \times |\Theta|} (k \leq |\Theta|) \) be any matrix such that rank \( Z^T S = k \), and let \( C = AS \). Then,
\[
\|A - C C^T A\|_F^2 \leq \|A - \Pi_{C,k}^F (A)\|_F^2 \leq \|E\|_F^2 \| S(Z^T S)^+ \|_2,
\] (26)
where \( \Pi_{C,k}^F (A) = (C C^T A)_{k} \) is the best rank-\( k \) approximation to \( A \) in the column space of \( C \) with the respect to the Frobenius norm.
We’ll use a slightly relaxed version of Lemma 2,
\[ \| A - CC^+ A \|_F^2 \leq \| A - \Pi_{G,k}^F(A) \|_F^2 \leq \| E \|_F^2 \| S \|_2^2 \| (Z^T S)^+ \|_2^2. \]  
Choosing \( Z = V_k \), and noting that \( \| S \|_2^2 = 1 \), we have
\[ \| A - CC^+ A \|_F^2 \leq \| A - \Pi_{G,k}^F(A) \|_F^2 \leq \| V_k \|_F^2 \| (V_k^T S)^+ \|_2^2. \]
It remains to calculate \( \| (V_k^T S)^+ \|_2^2 = \sigma_k^{-2}(V_k SS^T V_k^T) \), where \( k \leq \text{rank}(A) = r \). As a consequence of the conjugation rule (Eqn. 20) applied to Eqn. 7 with pre- (post-)multiplication by \( \Sigma_k^{-1} U_k^T \) (\( U_k \Sigma_k^{-1} \)), we have,
\[ (1 - \epsilon) I_k - \frac{\epsilon}{k} \| A \|_k^2 \Sigma_k^{-2} \leq V_k SS^T V_k^T \]  
From Eqn. 29 the \( k^{th} \) eigenvalue \( \sigma_k^2(V_r SS^T V_r^T) \) obeys,
\[ (1 - 2\epsilon) \leq \sigma_k^2(V_r SS^T V_r^T), \]  
after using the fact that \( \frac{1}{k} \| A \|_k^2 \Sigma_k^{-2} \leq 1 \) (by definition). Eqn. 30 shows that, for \( 0 < \epsilon < \frac{1}{2} \), \( \text{rank}(V_k^T S) = k \). Combining Eqn. 28 and Eqn. 30 gives Eqn. 8. This proof illustrates the power of the spectral bound (Eqn. 7), since the column subset selection bound (Eqn. 8) is a direct consequence of Eqn. 7.

5.5 Proof of Theorem 3

It will be convenient to define the projection matrix \( Y = I - X \). Then the upper bound in Eqn. 9 follows directly from upper spectral bound (Eqn. 25), the conjugation rule (Eqn. 20), and the trace rule (Eqn. 22).

The lower projection-cost preservation bound is considerably more involved to prove, and also relies primarily on the spectral bound (Eqn. 7) and facts from linear algebra. Our proof is nearly identical to the proof of Cohen et al. (2017)'s Theorem 4, with only one large deviation and several small differences in constants. We include the full proof for completeness, and point out the major difference.

We split \( AA^T \) and \( CC^T \) into their projections on the top \( m \) "head" singular vectors and the remaining "tail" singular vectors. Choose \( m \) such that \( \Sigma_{m,m}^2 \) is the smallest singular value that obeys \( \Sigma_{m,m}^2 \geq \frac{1}{k} \| A \|_k^2 \). Let \( P_m = U_m U_m^T \) and \( P_{\perp m} = U_{\perp m} U_{\perp m}^T \) be projection matrices. Note that,
\[ \text{tr} \left( YAA^T Y \right) = \text{tr} \left( YA_m A_m^T Y \right) + \text{tr} \left( YA_{\perp m} A_{\perp m}^T Y \right) \]
\[ \text{tr} \left( YCC^T Y \right) = \text{tr} \left( YP_m CC^T P_m^T Y \right) + \text{tr} \left( YP_{\perp m} CC^T P_{\perp m}^T Y \right) + 2 \text{tr} \left( YP_m CC^T P_{\perp m}^T Y \right). \]  

5.5.1 Head terms

First we bound the terms involving only \( P_m \). Consider Eqn. 7 and the vector \( y = P_m x \) for any vector \( x \). Eqn. 7 gives,
\[ (1 - \epsilon) x^T P_m AA^T P_m x - \frac{\epsilon}{k} \| A \|_k^2 y^T y \leq y^T CC^T y \]
\[ (1 - 2\epsilon) x^T P_m AA^T P_m x \leq x^T P_m CC^T P_m x, \]
because by the artful definition of \( m, x^T P_m AA^T P_m x \geq \frac{1}{k} \| A \|_k^2 y^T y \). Therefore we have,
\[ (1 - 2\epsilon) P_m AA^T P_m \leq P_m CC^T P_m, \]
and finally,
\[ (1 - 2\epsilon) \text{tr} \left( YA_m A_m^T Y \right) \leq \text{tr} \left( YP_m CC^T P_m y \right). \]
5.5.2 Tail terms

To bound the lower singular directions of $A$, we decompose $\text{tr} \left( YA_{\setminus m}A_{\setminus m}^T Y \right)$ further as,

$$\text{tr} \left( YA_{\setminus m}A_{\setminus m}^T Y \right) = \text{tr} \left( A_{\setminus m}A_{\setminus m}^T \right) - \text{tr} \left( XA_{\setminus m}A_{\setminus m}^T X \right),$$

and analogously for $C$.

The upper spectral bound (Eqn. 25), the conjugation rule (Eqn. 20) gives,

$$P_{\setminus m}CC^TP_{\setminus m} \preceq P_{\setminus m}AA^TP_{\setminus m},$$

The conjugation rule (Eqn. 20), and the trace rule (Eqn. 22) give,

$$\text{tr} \left( XP_{\setminus m}CC^TP_{\setminus m}X \right) \leq \text{tr} \left( XP_{\setminus m}AA^TP_{\setminus m}X \right).$$

Next we consider $||P_{\setminus m}A||_F^2 - ||P_{\setminus m}C||_F^2$. In Cohen et al. (2017)’s proof, a scalar Chernoff bound is used for $||P_{\setminus m}A||_F^2 - ||P_{\setminus m}C||_F^2$ (Cohen et al. (2017) Section 4.3, Eqn. 17). Since our matrix $C$ is constructed deterministically, we will prove and substitute the following bound (Eqn. 37),

$$0 \leq ||P_{\setminus m}A||_F^2 - ||P_{\setminus m}C||_F^2 \leq 2\epsilon ||A_{\setminus k}||_F^2.$$  

for the scalar Chernoff bound.

To prove Eqn. 37 we first note the lower bound follows directly from upper spectral bound (Eqn. 25), the conjugation rule (Eqn. 20), and the trace rule (Eqn. 22). To prove the upper bound, we rewrite the difference in Frobenius norms as a difference in sums over column norms:

$$||P_{\setminus m}A||_F^2 - ||P_{\setminus m}C||_F^2 = \sum_{j \notin \Theta} ||P_{\setminus m}a_j||_F^2 = \sum_{j \notin \Theta} \text{tr} a_j^T P_{\setminus m}P_{\setminus m}a_j$$

$$\leq \frac{2}{k} ||A_{\setminus k}||_F^2 \sum_{j \notin \Theta} \text{tr} a_j^T P_{\setminus m}U_{\setminus m} \Sigma^{-2} U_{\setminus m}^T P_{\setminus m}a_j$$

$$\leq \frac{2}{k} ||A_{\setminus k}||_F^2 \sum_{j \notin \Theta} \text{tr} a_j^T U \Sigma^{-2} Ua_j$$

$$= \frac{2\epsilon}{k} ||A_{\setminus k}||_F^2 \leq 2\epsilon ||A_{\setminus k}||_F^2.$$  

(38)

The first inequality follows from the fact that $\Sigma_{i,i}^2 = \Sigma_{i,i}^2 + \frac{1}{2} ||A_{\setminus k}||_F^2 \leq \frac{2}{k} ||A_{\setminus k}||_F^2$ for $i \geq m$. As a result, $P_{\setminus m} \preceq \frac{2}{k} ||A_{\setminus k}||_F^2 U_{\setminus m} \Sigma^{-2} U_{\setminus m}^T$.

Combining the upper bound of Eqn. 37 with Eqn. 38 gives,

$$\text{tr} \left( YA_{\setminus m}A_{\setminus m}^T Y \right) - 2\epsilon ||A_{\setminus k}||_F^2 \leq \text{tr} \left( YP_{\setminus m}CC^TP_{\setminus m}Y \right).$$

(39)

5.5.3 Cross terms

Finally, we show $\text{tr} \left( YP_{\setminus m}CC^TP_{\setminus m} \right)$ is small. We rewrite it as,

$$\text{tr} \left( YP_{\setminus m}CC^TP_{\setminus m} \right) = \text{tr} \left( YAA^T(AA^T)^+P_{\setminus m}CC^TP_{\setminus m} \right),$$

(40)

using the cyclic property of the trace and the fact that $P_{\setminus m}CC^TP_{\setminus m}$ is in $A$’s column span. Because $(AA^T)^+$ is semi-positive definite and defines a semi-inner product, we can use the Cauchy-Schwarz inequality,

$$|\text{tr} \left( YAA^T(AA^T)^+P_{\setminus m}CC^TP_{\setminus m} \right)| \leq \left( \text{tr} \left( YAA^T(AA^T)^+AA^TY \right) \right)^{\frac{1}{2}}$$

$$\times \left( \text{tr} \left( P_{\setminus m}CC^TP_{\setminus m}(AA^T)^+P_{\setminus m}CC^TP_{\setminus m} \right) \right)^{\frac{1}{2}}$$

$$= \left( \text{tr} \left( YAA^TY \right) \right)^{\frac{1}{2}} ||P_{\setminus m}CC^TU_{\setminus m} \Sigma^{-1}||_F.$$
The square of the second term decomposes as,

\[ ||P_m CC^T U_m \Sigma_m^{-1}||_F^2 = \sum_{i=1}^{m} ||P_m CC^T u_i||_F^2 \Sigma_i^{-2}, \]  

which is small for every \( i \). To show this, we define two convenient vectors. The first is the unit vector \( p_i = \frac{P_m CC^T u_i}{||P_m CC^T u_i||_F} \). Note that \( p_i^T u_i = 0 \). This is convenient because \( (p_i^T CC^T u_i)^2 = ||P_m CC^T u_i||_F^2 \). The second vector is \( m = \Sigma_i^{-1} u_i + \frac{\sqrt{\varepsilon}}{||A_{\ell k}||_F} p_i \). From Eqn. 25

\[
m^T CC^T m \leq m^T AA^T m
\]

\[
\Sigma_i^{-2} u_i^T CC^T u_i + \frac{k}{||A_{\ell k}||_F^2} p_i^T CC^T p_i + 2\frac{\sqrt{k}}{||A_{\ell k}||_F} p_i^T CC^T u_i \leq \Sigma_i^{-2} u_i^T AA^T u_i + \frac{k}{||A_{\ell k}||_F^2} p_i^T AA^T p_i = 1 + \frac{k}{||A_{\ell k}||_F^2} p_i^T AA^T p_i.
\]

From Eqn. 32 we have,

\[
(1 - 2\varepsilon) \Sigma_i, i = (1 - 2\varepsilon) u_i AA^T u_i \leq u_i CC^T u_i.
\]

From Eqn. 35 we have,

\[
p_i CC^T p_i \leq p_i AA^T p_i.
\]

Using these facts, Eqn. 43 becomes,

\[
(1 - 2\varepsilon) + \frac{k}{||A_{\ell k}||_F^2} p_i^T CC^T p_i + 2\frac{\sqrt{k}}{||A_{\ell k}||_F} p_i^T CC^T u_i \leq 1 + \frac{2\Sigma_i^{-2} u_i^T AA^T u_i}{k}
\]

\[
(p_i^T CC^T u_i)^2 \leq \frac{k}{\varepsilon^2||A_{\ell k}||_F^2}.
\]

Returning to Eqn. 42 with this, we have

\[
||P_m CC^T U_m \Sigma_m^{-1}||_F^2 \leq \sum_{i=1}^{m} \varepsilon^2 \frac{||A_{\ell k}||_F^2}{k} \leq 2\varepsilon^2||A_{\ell k}||_F^2.
\]

The factor of 2 comes from the fact that \( m \leq 2k \). After recalling that the Eckart-Young-Mirsky theorem [Eckart and Young 1936] gives \( ||A_{\ell k}||_F^2 \leq \text{tr} (Y AA^T Y) \), we have for Eqn. 41.

\[
|\text{tr} \left( YAA^T( AA^T) + P_m CC^T P_m^T \right) | \leq \sqrt{2\varepsilon} \text{tr} (YAA^T Y).
\]

Combining Eqn. 31, Eqn. 33, Eqn. 39, Eqn. 48 leads to,

\[
(1 - 2\varepsilon)\text{tr} (YA m A_m^T Y) + \text{tr} (YA m A_m^T Y) - 2\varepsilon||A_{\ell k}||_F^2
\]

\[
-2\sqrt{2\varepsilon} \text{tr} (YAA^T Y) \leq \text{tr} (YCC^T Y).
\]

Again applying \( ||A_{\ell k}||_F^2 \leq \text{tr} (YAA^T Y) \) and subtracting an extra \( 0 \leq 2\varepsilon \text{tr} (YA m A_m^T Y) \) from the lefthand side gives,

\[
(1 - 2(2 + \sqrt{2})\varepsilon) \text{tr} (YAA^T Y) \leq \text{tr} (YCC^T Y),
\]

proving Theorem 5.
5.6 Proof of Lemma

Let the SVD of $C$ be $C = W \Sigma Z^T$. Set $X = W_k W_k^T$. From the left-hand side of Eqn. 9 and the Eckart-Young-Mirsky theorem (Eckart and Young [1936]),

$$(1 - \alpha) \|A_{\chi k}\|_F^2 \leq (1 - \alpha) \|A - XA\|_F^2 \leq \|C_{\chi k}\|_F^2. \quad (51)$$

Similarly, set $X = U_k U_k^T$. From the right-hand side of Eqn. 9

$$\|C_{\chi k}\|_F^2 \leq \|C - XC\|_F^2 \leq \|A_{\chi k}\|_F^2. \quad (52)$$

This means that,

$$(1 - \alpha) \|A_{\chi k}\|_F^2 \leq \|C_{\chi k}\|_F^2 \leq \|A_{\chi k}\|_F^2. \quad (53)$$

Adding $1/k$ times Eqn. 53 to Eqn. 7 gives,

$$(1 - \alpha) \|A_{\chi k}\|_F^2 \leq \|C_{\chi k}\|_F^2 \leq \|A_{\chi k}\|_F^2 - \frac{1}{k} \alpha \|A\|_F^2 \|A\|_F^2 \leq K(\alpha)^{-1} - \frac{1}{k} \alpha \|A\|_F^2 \leq \|A\|_F^2 \|A\|_F^2. \quad (54)$$

Noting that we can subtract $\alpha \|A\|_F^2$ from the left-most side of Eqn. 54 and that all of the matrices are invertible gives the lemma.

5.7 Proof of Theorem

For ridge regression with matrix $A$ and estimator $\hat{y} = A \hat{x}$, the statistical risk of the estimator $R(\hat{y}_A)$ is,

$$R(\hat{y}_A) = \frac{1}{n} \| 2 \sigma^2 \| A \left( A^T A + \frac{\|A_{\chi k}\|_F^2}{k} I \right)^{-1} A^T (y^* + \sigma^2 \xi) - y^* \|_2^2. \quad (55)$$

Decomposing into bias and variance terms, taking the expectation, and using the Woodbury matrix inversion formula gives,

$$R(\hat{y}_A) = \frac{1}{n} \| \left( A \left( A^T A + \frac{\|A_{\chi k}\|_F^2}{k} I \right)^{-1} A^T - I \right) y^* \|_2^2 + \text{bias}(A)^2 + \text{variance}(A). \quad (56)$$

We begin with the variance term. Using the SVD of $A$ on the variance term gives,

$$\text{variance}(A) = \frac{\sigma^2}{n} \text{tr} \left( \Sigma^4 \Sigma^{-4} \right). \quad (57)$$

As a consequence of Eqn. 21 and Eqn. 7

$$\Sigma^2_C \leq \Sigma^2. \quad (58)$$

As a consequence of Eqn. 21 and Eqn. 10

$$\Sigma^{-2} \leq \Sigma^{-2}_C \leq \frac{1}{1 - (\alpha + 1)\epsilon} \Sigma^{-2}, \quad (59)$$

where $\Sigma^2_C = \Sigma^2_C + \frac{1}{k} \| C_{\chi k} \|_F^2 I.$ Because the matrices in Eqn. 58 and Eqn. 59 are diagonal and (semi-)positive definite, the relationships can be squared. Finally, because the product of $\frac{1}{1 - (\alpha + 1)\epsilon} \Sigma^4 \Sigma^{-4}$ is bigger or equal to the product of $\Sigma^4_C \Sigma^{-4}_C$ for every element along the diagonal,

$$\text{variance}(C) \leq \frac{1}{(1 - (\alpha + 1)\epsilon)^2} \text{variance}(A). \quad (60)$$
Next we analyze the bias. This part of the proof follows the structure of the proof of Theorem 1 in [Alaoui and Mahoney (2015)]. It will be useful to analyze the quantity,

\[
\| (1 - \epsilon) \mathbf{K}(C) \mathbf{y}^* \|_2 = \| (1 - \epsilon) \mathbf{K}(C) - \mathbf{K}(A) \|_2 \mathbf{y}^* + \| \mathbf{K}(A) \mathbf{y}^* \|_2,
\]

where the last line is due to the triangle inequality. Furthermore,

\[
(1 - \epsilon) \mathbf{K}(C) - \mathbf{K}(A) = \mathbf{K}(C) (1 - \epsilon) \mathbf{K}(A)^{-1} - \mathbf{K}(C)^{-1} \mathbf{K}(A),
\]

so

\[
\| (1 - \epsilon) \mathbf{K}(C) - \mathbf{K}(A) \|_2^2 = \| \mathbf{K}(C) \|_2^2 \| (1 - \epsilon) \mathbf{K}(A)^{-1} - \mathbf{K}(C)^{-1} \mathbf{K}(A) \|_2^2 \leq \| \mathbf{K}(C) \|_2^2 \| (1 - \epsilon) \mathbf{K}(A)^{-1} - \mathbf{K}(C)^{-1} \mathbf{K}(A) \|_2^2,
\]

where \( \| \cdot \|_2 \) refers to the operator norm \( \| A \|_2 = \inf \{ c \geq 0 : \| Av \| \leq c \| v \| \forall v \in V \} \), where \( V, W \) are normed vector spaces and \( A \) is a linear map from \( A : V \rightarrow W \). We also have,

\[
\| (1 - \epsilon) \mathbf{K}(A)^{-1} - \mathbf{K}(C)^{-1} \mathbf{K}(C) \|_2^2 = \| \mathbf{K}(C) \|_2^2 \| (1 - \epsilon) \mathbf{K}(A)^{-1} - \mathbf{K}(C)^{-1} \mathbf{K}(C) \|_2^2.
\]

From Eqn. 54 we have

\[
(1 - \epsilon) \mathbf{K}(A)^{-1} - \mathbf{K}(C)^{-1} \leq \frac{\alpha \epsilon}{k} \frac{\| A \|_{op}}{\| A \|_{op}}^2 \mathbf{I},
\]

which is squareable because \( (1 - \epsilon) \mathbf{K}(A)^{-1} - \mathbf{K}(C)^{-1} \) commutes with the identity. So,

\[
\| \mathbf{K}(C) \|_2^2 \| (1 - \epsilon) \mathbf{K}(A)^{-1} - \mathbf{K}(C)^{-1} \mathbf{K}(A) \|_2^2 \leq \frac{\alpha^2 \epsilon^2}{k^2} \| A \|_{op}^2 \| \mathbf{K}(C) \|_2^2 \leq \frac{\alpha^2 \epsilon^2}{k^2} \| A \|_{op}^2 \| \mathbf{K}(C) \|_2^2 \leq \frac{\alpha^2 \epsilon^2}{k^2} \| \mathbf{K}(C) \|_2^2 \leq \frac{\alpha^2 \epsilon^2}{k^2} \frac{\| \mathbf{K}(C) \|_2^2}{\| \mathbf{K}(C) \|_2^2} \leq \frac{\alpha^2 \epsilon^2}{k^2} \frac{\| \mathbf{K}(C) \|_2^2}{\| \mathbf{K}(C) \|_2^2}.
\]

The second to last line follows from the definition of the operator norm and \( \mathbf{K}(C) \). The last line follows from Eqn. 53.

Returning to Eqn. 61 gives,

\[
\| \mathbf{K}(C) \mathbf{y}^* \|_2 \leq \frac{1}{(1 - \epsilon)} \left( \frac{\alpha \epsilon}{1 - \alpha \epsilon} + 1 \right) \| \mathbf{K}(A) \mathbf{y}^* \|_2
\]

\[
\| \mathbf{K}(C) \mathbf{y}^* \|_2 \leq \frac{1}{(1 - \epsilon)(1 - \alpha \epsilon)} \| \mathbf{K}(A) \mathbf{y}^* \|_2.
\]

Therefore, using Eqn. 53 again,

\[
\text{bias}(C) \leq \frac{1}{(1 - \epsilon)^2(1 - \alpha \epsilon)^2} \text{bias}(A).
\]

Combining Eqn. 60 and Eqn. 68 gives,

\[
\mathcal{R}(\hat{y}_C) \leq \max \left( \left( \frac{1}{1 - (\alpha + 1)\epsilon} \right)^2, \left( \frac{1}{1 - \alpha \epsilon} \right)^2 \left( \frac{1}{1 - \epsilon} \right)^2 \right) \mathcal{R}(\hat{y}_A).
\]

On the interval \( 0 < \epsilon < \frac{1}{2\alpha} \), and for \( \alpha > 1 \),

\[
\max \left( \left( \frac{1}{1 - (\alpha + 1)\epsilon} \right)^2, \left( \frac{1}{1 - \alpha \epsilon} \right)^2 \left( \frac{1}{1 - \epsilon} \right)^2 \right) < 1 + \frac{2\alpha(-1 + 2\alpha + 3\alpha^2)}{(1 - \alpha)^2} \epsilon.
\]

Theorem 4 follows immediately.
5.8 Proof of Theorem 5

The proof is nearly identical to the proof of Theorem 3 of Papailiopoulos et al. (2014), so we do not repeat all of the algebra. It will be convenient to change notation. Without loss of generality, we can sort our column indices $i$ in order of decreasing leverage score. We will also start this index at 1 instead of 0. With this change of notation, the power-law decay formula (Eqn. 15) becomes,

$$\bar{\tau}_i(A) = (i)^{-a} \bar{\tau}_1(A) \quad a > 1,$$

(71)

By the definition of the sum of ridge leverage scores and power-law decay (Eqn. 71),

$$\bar{\tau} = \bar{\tau}_1(A) \sum_{i=1}^{d} i^{-a} \quad \bar{\tau}_1 = \frac{\bar{\tau}}{\sum_{i=1}^{d} i^{-a}}.$$

(72)

By the definition of the DRLS algorithm, we collect $|\Theta|$ leverage scores such that $\bar{\tau} - \epsilon < \sum_{i \in \Theta} \bar{\tau}_i(A)$. This gives,

$$\bar{\tau} - \epsilon < \bar{\tau}_1 \sum_{i=1}^{d} i^{-a} \quad \Rightarrow \quad \bar{\tau}_1 = \frac{\bar{\tau}}{\sum_{i=1}^{d} i^{-a}}.$$

(73)

Papailiopoulos et al. (2014) show that,

$$\sum_{i=1}^{d} i^{-a} \leq \max \left( \frac{2}{1(\Theta+1)^{\alpha}}, \frac{2}{(a-1)(\Theta+1)^{\alpha-1}} \right) \quad a > 1.$$

(74)

Noting that $\bar{\tau} \leq 2k$, substituting Eqn. 74 into Eqn. 73, solving for $|\Theta|$, and noting that the algorithm collects a minimum of $k$ columns results in the expression for the number of columns collected when ridge leverage scores exhibit a power-law decay (Eqn. 15).

Figures

Figure 1: Pie chart showing multi-omic feature types of matrix $A$ for LGG tumor data.
Figure 2: Eigenvalues of matrix $AA^T$ for LGG tumor multi-omic data. The eigenvalues range over multiple orders of magnitude.

Figure 3: SVD projection of LGG tumor multi-omic data, colored by the combined status for "IDH" and "codel" outcome variables.

Figure 4: SVD projection of LGG tumor multi-omic data, colored by the combined status for "IDH" and "codel" outcome variables.
Figure 5: SVD projection of LGG tumor multi-omic data, colored by the combined status for "IDH" and "codel" outcome variables.

Figure 6: The error $\tilde{\epsilon}$ vs. the number of columns kept for LGG tumor multi-omic data $k = 3$ ridge leverage scores. A dramatic reduction in the number of columns kept incurs only a small error penalty.
Figure 7: Power-law decay of LGG tumor multi-omic data $k = 3$ ridge leverage scores with sorted column index. The fit is to $\text{Score} = b \times (\text{Index})^a$ on the first $10^3$ ridge leverage scores.

Figure 8: Histogram of $\frac{||C - XC||_F^2}{||A - XA||_F^2}$ for 1000 random rank-$k = 3$ orthogonal projections $X$ and $C$ selected by the $k = 3, \epsilon = 0.1$ DRLS algorithm for LGG tumor multi-omic data.
Figure 9: Classical leverage scores vs. \( k = 3 \) ridge leverage scores for LGG tumor multi-omic data. Most columns with large classical leverage scores have small ridge leverage scores; there is significant shrinkage.

Figure 10: Histogram of \( \hat{y}_A - \hat{y}_C \) for the outcome "codel."
Figure 11: Histogram of $\hat{y}_A - \hat{y}_C$ for the outcome "IDH."

Figure 12: Histogram of $\hat{x}_A - \hat{x}_C$ for the outcome "codel."
Figure 13: Histogram of $\hat{x}_A - \hat{x}_C$ for the outcome "IDH."

Figure 14: Histograms of the predictions $\hat{y}_C$ conditioned the outcome $y$ for "codel."
Figure 15: Histograms of the predictions $\hat{y}_C$ conditioned on the outcome $y$ for "IDH."