# Sharper Bounds for Regularized Data Fitting[*]

**Haim Avron[1], Kenneth L. Clarkson[2], and David P. Woodruff[3]**

1 **Tel Aviv University, Tel Aviv, Israel**
`haimav@post.tau.ac.il`

2 **IBM Research – Almaden, San Jose, CA, USA**
`klclarks@us.ibm.com`

3 **IBM Research – Almaden, San Jose, CA, USA**
`dpwoodru.ibm.com`

## Abstract

We study matrix sketching methods for regularized variants of linear regression, low rank approximation, and canonical correlation analysis. Our main focus is on sketching techniques which preserve the objective function value for regularized problems, which is an area that has remained largely unexplored. We study regularization both in a fairly broad setting, and in the specific context of the popular and widely used technique of ridge regularization; for the latter, as applied to each of these problems, we show algorithmic resource bounds in which the *statistical dimension* appears in places where in previous bounds the rank would appear. The statistical dimension is always smaller than the rank, and decreases as the amount of regularization increases. In particular, for the ridge low-rank approximation problem $\min_{Y,X} \|YX - A\|_F^2 + \lambda\|Y\|_F^2 + \lambda\|X\|_F^2$, where $Y \in \mathbb{R}^{n\times k}$ and $X \in \mathbb{R}^{k\times d}$, we give an approximation algorithm needing $O(\mathtt{nnz}(A)) + \tilde{O}((n+d)\varepsilon^{-1}k\min\{k, \varepsilon^{-1}\,\mathtt{sd}_\lambda(Y^*)\}) + \mathrm{poly}(\mathtt{sd}_\lambda(Y^*)\epsilon^{-1})$ time, where $s_\lambda(Y^*) \le k$ is the statistical dimension of $Y^*$, $Y^*$ is an optimal $Y$, $\varepsilon$ is an error parameter, and $\mathtt{nnz}(A)$ is the number of nonzero entries of $A$. This is faster than prior work, even when $\lambda = 0$. We also study regularization in a much more general setting. For example, we obtain sketching-based algorithms for the low-rank approximation problem $\min_{X,Y} \|YX - A\|_F^2 + f(Y, X)$ where $f(\cdot,\cdot)$ is a regularizing function satisfying some very general conditions (chiefly, invariance under orthogonal transformations).

**1998 ACM Subject Classification** G.1.3 Numerical Linear Algebra

**Keywords and phrases** Matrices, Regression, Low-rank approximation, Regularization, Canonical Correlation Analysis

**Digital Object Identifier** 10.4230/LIPIcs.APPROX/RANDOM.2017.27

## 1 Introduction

The technique of matrix sketching, such as the use of random projections, has been shown in recent years to be a powerful tool for accelerating many important statistical learning techniques. Indeed, recent work has proposed highly efficient algorithms for, among other problems, linear regression, low-rank approximation [22, 30] and canonical correlation analysis [3]. In addition to being a powerful theoretical tool, sketching is also an applied one; see [31] for a discussion of state-of-the-art performance for important techniques in statistical learning.

Many statistical learning techniques can benefit substantially, in their quality of results, by using some form of regularization. Regularization can also help by reducing the computing

[*] Condensed version; most proofs omitted. Full version: `https://arxiv.org/abs/1611.03225`.

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017).
Editors: Klaus Jansen, José D. P. Rolim, David Williamson, and Santosh S. Vempala; Article No. 27; pp. 27:1–27:22
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

resources needed for these techniques. While there has been some prior exploration in this area, as discussed in §1.1, commonly it has featured sampling-based techniques, often focused on regression, and often with analyses using distributional assumptions about the input (though such assumptions are not always necessary). Our study considers fast (linear-time) sketching methods, a breadth of problems, and makes no distributional assumptions. Also, where most prior work studied the distance of an approximate solution to the optimum, our guarantees are concerning approximation with respect to a relevant loss function - see below for more discussion.

It is a long-standing theme in the study of randomized algorithms that structures that aid statistical inference can also aid algorithm design, so that for example, VC dimension and sample compression have been applied in both areas, and more recently, in cluster analysis the algorithmic advantages of natural statistical assumptions have been explored. This work is another contribution to this theme. Our high-level goal in this work is to study generic conditions on sketching matrices that can be applied to a wide array of regularized problems in linear algebra, preserving their objective function values, and exploiting the power of regularization.

## 1.1 Results

We study regularization both in a fairly broad setting, and in the specific context of the popular and widely used technique of ridge regularization. We discuss the latter in sections 2, 3 and B; our main results for ridge regularization, Theorem 15, on linear regression, Theorem 26, on low-rank approximation, and Theorem 33, on canonical correlation analysis, show that for ridge regularization, the sketch size need only be a function of the *statistical dimension* of the input matrix, as opposed to its rank, as is common in the analysis of sketching-based methods. Thus, ridge regularization improves the performance of sketching-based methods.

Next, we consider regularizers under rather general assumptions involving invariance under left and/or right multiplication by orthogonal matrices, and show that sketching-based methods can be applied, to regularized multiple-response regression in §C and to regularized low-rank approximation, in §D. Here we obtain running times in terms of the statistical dimension. Along the way, in §D.1, we give a "base case" algorithm for reducing low-rank approximation, via singular value decomposition, to the special case of diagonal matrices.

Throughout we rely on sketching matrix constructions involving *sparse embeddings* [10, 24, 23, 6, 12], and on *Sampled Randomized Hadamard Transforms* (SRHT) [1, 26, 14, 15, 28, 7, 16, 33]. Here for matrix $A$, its sketch is $SA$, where $S$ is a sketching matrix. The sketching constructions mentioned can be combined to yield a sketching matrix $S$ such that the sketch of matrix $A$, which is simply $SA$, can be computed in time $O(\mathtt{nnz}(A))$, which is proportional to the number of nonzero entries of $A$. Moreover, the number of rows of $S$ is small. Corollary 14 summarizes our use of these constructions as applied to ridge regression.

A key property of a sketching matrix $S$ is that it be a *subspace embedding*, so that $\|SAx\|_2 \approx \|Ax\|_2$ for all $x$. Definition 20 gives the technical definition, and Definition 22 gives the definition of the related property of an *affine embedding* that we also use. Lemma 23 summarizes the use of sparse embeddings and SRHT for subspace and affine embeddings.

In the following we give our main results in more detail. However, before doing so, we need the formal definition of the statistical dimension.

▶ **Definition 1** (Statistical Dimension). For real value $\lambda \geq 0$ and rank-$k$ matrix $A$ with singular values $\sigma_i, i \in [k]$, the quantity $\mathtt{sd}_\lambda(A) \equiv \sum_{i\in[k]} 1/(1+\lambda/\sigma_i^2)$ is the *statistical dimension* (or *effective dimension*, or "hat matrix trace") of the ridge regression problem with regularizing weight $\lambda$.

Note that $\mathtt{sd}_\lambda(A)$ is decreasing in $\lambda$, with maximum $\mathtt{sd}_0(A)$ equal to the rank of $A$. Thus a dependence of resources on $\mathtt{sd}_\lambda(A)$ instead of the rank is never worse, and will be much better for large $\lambda$.

In §A, we give an algorithm for estimating $\mathtt{sd}_\lambda(A)$ to within a constant factor, in $O(\mathtt{nnz}(A))$ time, for $\mathtt{sd}_\lambda(A) \leq (n+d)^{1/3}$. Knowing $\mathtt{sd}_\lambda(A)$ to within a constant factor allows us to set various parameters of our algorithms.

### 1.1.1 Ridge Regression

In §2 we apply sketching to reduce from one ridge regression problem to another one with fewer rows.

▶ **Theorem 2** (Less detailed version of Thm. 15)**.** *Given $\varepsilon \in (0,1]$ and $A \in \mathbb{R}^{n\times d}$, there is a sketching distribution over $S \in \mathbb{R}^{m\times n}$, where $m = \tilde{O}(\varepsilon^{-1}\,\mathtt{sd}_\lambda(A))$, such that $SA$ can be computed in $O(\mathtt{nnz}(A)) + d \cdot \operatorname{poly}(\mathtt{sd}_\lambda(A)/\varepsilon)$ time, and with constant probability $\tilde{x} \equiv \operatorname{argmin}_{x\in\mathbb{R}^d} \|S(Ax-b)\|^2 + \lambda\|x\|^2$ satisfies*

$$\|A\tilde{x}-b\|^2 + \lambda\|\tilde{x}\|^2 \leq (1+\varepsilon)\min_{x\in\mathbb{R}^d}\|Ax-b\|^2 + \lambda\|x\|^2.$$

*Here $\operatorname{poly}(\kappa)$ denotes some polynomial function of the value $\kappa$.*

In our analysis (Lemma 10), we map ridge regression to ordinary least squares (by using a matrix with $\sqrt{\lambda}I$ adjoined), and then apply prior analysis of sketching algorithms, but with the novel use of a sketching matrix that is "partly exact"; this latter step is important to obtain our overall bounds. We also show that sketching matrices can be usefully composed in our regularized setting; this is straightforward in the non-regularized case, but requires some work here.

As noted, the statistical dimension of a data matrix in the context of ridge regression is also referred to as the *effective degrees of freedom* of the regression problem in the statistics literature, and the statistical dimension features, as the name suggests, in the statistical analysis of the method. Our results show that the statistical dimension affects not only the statistical capacity of ridge regression, but also its computational complexity.

The reduction of the above theorem is mainly of interest when $n \gg \mathtt{sd}_\lambda(A)$, which holds in particular when $n \gg d$, since $d \geq \mathtt{rank}(A) \geq \mathtt{sd}_\lambda(A)$. We also give a reduction using sketching when $d$ is large, discussed in §2.2. Here algorithmic resources depend on a power of $\sigma_1^2/\lambda$, where $\sigma_1$ is the leading singular value of $A$. This result falls within our theme of improved efficiency as $\lambda$ increases, but in contrast to our other results, performance does not degrade gracefully as $\lambda \to 0$. The difficulty is that we use the product of sketches $AS^\top SA^\top$ to estimate the product $AA^\top$ in the expression $\|AA^\top y - b\|$. Since that expression can be zero, and since we seek a strong notion of relative error, the error of our overall estimate is harder to control, and impossible when $\lambda = 0$.

As for related work on ridge regression, Lu *et al.* [21] apply the SRHT to ridge regression, analyzing the statistical risk under the distributional assumption on the input data that $b$ is a random variable, and not giving bounds in terms of $\mathtt{sd}_\lambda$. El Alaoui *et al.* [17] apply sampling techniques based on the *leverage scores* of a matrix derived from the input, with a different error measure than ours, namely, the statistical risk; here for their error analysis they consider the case when the noise in their ridge regression problem is i.i.d. Gaussian. They give results in terms of $\mathtt{sd}_\lambda(A)$, which arises naturally for them as the sum of the leverage scores. Here we show that this quantity arises also in the context of oblivious subspace embeddings, and with the goal being to obtain a worst-case relative-error guarantee in objective function value rather than for minimizing statistical risk. Chen *et al.* [9] apply sparse embeddings to ridge regression, obtaining solutions $\tilde{x}$ with $\|\tilde{x} - x^*\|_2$ small, where $x^*$ is optimal, and do this in $O(\mathtt{nnz}(A) + d^3/\varepsilon^2)$ time. They also analyze the statistical risk of their output. Yang *et al.* [32] consider slower sketching methods than those here, and analyze their error under distributional assumptions using an incomparable notion of statistical dimension. Frostig *et al.* [18] make distributional assumptions, in particular a kurtosis property. Frostig *et al.* [19] give bounds in terms of a convex condition number that can be much larger than $\mathtt{sd}_\lambda(A)$. Another related work is that of Pilanci *et al.* [25] which we dicuss below.

### 1.1.2 Ridge Low-rank Approximation

In §3 we consider the following problem: for given $A \in \mathbb{R}^{n\times d}$, integer $k$, and weight $\lambda \geq 0$, find:

$$\min_{\substack{Y\in\mathbb{R}^{n\times k}\\ X\in\mathbb{R}^{k\times d}}} \|YX - A\|_F^2 + \lambda\|Y\|_F^2 + \lambda\|X\|_F^2, \tag{1}$$

where, as is well known (and discussed in detail later), this regularization term is equivalent to $2\lambda\|YX\|_*$, where $\|\cdot\|_*$ is the trace (nuclear) norm, the Schatten 1-norm. We show the following.

▶ **Theorem 3** (Less detailed Thm. 26). *Given input $A \in \mathbb{R}^{n\times d}$, there is a sketching-based algorithm returning $\tilde{Y} \in \mathbb{R}^{n\times k}, \tilde{X} \in \mathbb{R}^{k\times d}$ such that with constant probability, $\tilde{Y}$ and $\tilde{X}$ form a $(1+\varepsilon)$-approximate minimizer to (1), that is,*

$$\|\tilde{Y}\tilde{X} - A\|_F^2 + \lambda\|\tilde{Y}\|_F^2 + \lambda\|\tilde{X}\|_F^2 \tag{2}$$

$$\leq (1+\varepsilon) \min_{\substack{Y\in\mathbb{R}^{n\times k}\\ X\in\mathbb{R}^{k\times d}}} \|YX - A\|_F^2 + \lambda\|Y\|_F^2 + \lambda\|X\|_F^2. \tag{3}$$

*The matrices $\tilde{Y}$ and $\tilde{X}$ can be found in $O(\mathtt{nnz}(A)) + \tilde{O}((n+d)\varepsilon^{-1}k\min\{k, \varepsilon^{-1}\,\mathtt{sd}_\lambda(Y^*)\}) + \mathrm{poly}(\varepsilon^{-1}\,\mathtt{sd}_\lambda(Y^*))$ time, where $Y^*$ is an optimum $Y$ in (1) such that $\mathtt{sd}_\lambda(X^*) = \mathtt{sd}_\lambda(Y^*) \leq \mathtt{rank}(Y^*) \leq k$.*

This algorithm follows other algorithms for $\lambda = 0$ with running times of the form $O(\mathtt{nnz}(A)) + (n+d)\mathrm{poly}(k/\varepsilon)$ (e.g. [10]), and has the best known dependence on $k$ and $\varepsilon$ for algorithms of this type, even when $\lambda = 0$.

Our approach is to first extend our ridge regression results to the multiple-response case $\min_Z \|AZ - B\|_F^2 + \lambda\|Z\|_F^2$, and then reduce the multiple-response problem to a smaller one by showing that up to a cost in solution quality, we can assume that each row of $Z$ lies in the rowspace of $SA$, for $S$ a suitable sketching matrix. We apply this observation twice to the low-rank approximation problem, so that $Y$ can be assumed to be of the form $AR\tilde{Y}$, and $X$ of the form $\tilde{X}SA$, for sketching matrix $S$ and (right) sketching matrix $R$. Another round of sketching then reduces to a low-rank approximation problem of size independent of $n$ and $d$, and finally an SVD-based method is applied to that small problem.

Regarding related work: the regularization “encourages” the rank of $YX$ to be small, even when there is no rank constraint ($k$ is large), and this unconstrained problem has been extensively studied; even so, the rank constraint can reduce the computational cost and improve the output quality, as discussed by [8], who also give further background, and who give experimental results on an iterative algorithm. Pilanci *et al.* [25] consider only algorithms where the sketching time is at least $\Omega(nd)$, which can be much slower than our $\mathtt{nnz}(A)$ for sparse matrices, and it is not clear if their techniques can be extended. In the case of low-rank approximation with a nuclear norm constraint (the closest to our work), as the authors note, their paper gives no improvement in running time. While their framework might imply analyses for ridge regression, they did not consider it specifically, and such an analysis may not follow directly.

### 1.1.3 Regularized Canonical Correlation Analysis

Canonical correlation analysis (CCA) is an important statistical technique whose input is a pair of matrices, and whose solution depends on the Gram matrices $A^\top A$ and $B^\top B$. If these Gram matrices are ill-conditioned it is useful to regularize them by instead using $A^\top A + \lambda_1 I_d$ and $B^\top B + \lambda_2 I_{d'}$, for weights $\lambda_1, \lambda_2 \geq 0$. Thus, in this paper we consider a regularized version of CCA, defined as follows (our definition is in the same spirit as the one used by [3]).

▶ **Definition 4.** Let $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{n\times d'}$, and let

$$q = \min(\texttt{rank}(A^\top A + \lambda_1 I_d), \texttt{rank}(B^\top B + \lambda_2 I_{d'})).$$

Let $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$. The $(\lambda_1, \lambda_2)$ *canonical correlations* $\sigma_1^{(\lambda_1,\lambda_2)} \geq \cdots \geq \sigma_q^{(\lambda_1,\lambda_2)}$ and $(\lambda_1, \lambda_2)$ *canonical weights* $u_1, \ldots, u_q \in \mathbb{R}^d$ and $v_1, \ldots, v_q \in \mathbb{R}^{d'}$ are ones that maximize

$$\texttt{tr}(U^\top A^\top B V)$$

subject to

$$\begin{aligned} U^\top(A^\top A + \lambda_1 I_d)U &= I_q \\ V^\top(B^\top B + \lambda_2 I_{d'})V &= I_q \\ U^\top A^\top B V &= \texttt{diag}(\sigma_1^{(\lambda_1,\lambda_2)}, \ldots, \sigma_q^{(\lambda_1,\lambda_2)}) \end{aligned}$$

where $U = [u_1, \ldots, u_q] \in \mathbb{R}^{n\times q}$ and $V = [v_1, \ldots, v_q] \in \mathbb{R}^{d'\times q}$.

One classical way to solve non-regularized CCA ($\lambda_1 = \lambda_2 = 0$) is the Björck-Golub algorithm [5]. In §B we show that regularized CCA can be solved using a variant of the Björck-Golub algorithm.

Avron et al. [3] showed how to use sketching to compute an approximate CCA. In §B we show how to use sketching to compute an approximate regularized CCA.

▶ **Theorem 5** (Loose version of Thm. 33). *There is a distribution over matrices $S \in \mathbb{R}^{m\times n}$ with $m = O(\max(\texttt{sd}_{\lambda_1}(A), \texttt{sd}_{\lambda_2}(B))^2/\epsilon^2)$ such that with constant probability, the regularized CCA of $(SA, SB)$ is an $\epsilon$-approximate CCA of $(A, B)$. The matrices $SA$ and $SB$ can be computed in $O(\texttt{nnz}(A) + \texttt{nnz}(B))$ time.*

Our generalization of the classical Björck-Golub algorithm shows that regularized canonical correlation analysis can be computed via the product of two matrices whose columns are non-orthogonal regularized bases of $A$ and $B$. We then show that these two matrices are easier to sketch than the orthogonal bases that arise in non-regularized CCA. This in turn can be tied to approximation bounds of sketched regularized CCA versus exact CCA.

### 1.1.4 General Regularization

A key property of the Frobenius norm $\|\cdot\|_F$ is that it is invariant under rotations; for example, it satisfies the *right orthogonal invariance* condition $\|AQ\|_F = \|A\|_F$, for any orthogonal matrix $Q$ (assuming, of course, that $A$ and $Q$ having dimensions so that $AQ$ is defined). In

§C and §D, we study conditions under which such an invariance property, and little else, is enough to allow fast sketching-based approximation algorithms.

For regularized multiple-response regression, we have the following.

▶ **Theorem 6** (Implied by Thm. 39). *Let $f(\cdot)$ be a real-valued function on matrices that is right orthogonally invariant, subadditive, and invariant under padding the input matrix by rows or columns of zeros. Let $A \in \mathbb{R}^{n\times d}, B \in \mathbb{R}^{n\times d'}$. Suppose that for $r \equiv \mathtt{rank}\, A$, there is an algorithm that for general $n, d, d', r$ and $\varepsilon > 0$, in time $\tau(d, n, d', r, \varepsilon)$ finds $\tilde{X}$ with*

$$\|A\tilde{X} - B\|_F^2 + f(\tilde{X}) \leq (1+\varepsilon) \min_{X\in\mathbb{R}^{d\times d'}} \|AX - B\|_F^2 + f(X).$$

*Then there is another algorithm that with constant probability finds such an $\tilde{X}$, taking time*

$$O(\mathtt{nnz}(A) + \mathtt{nnz}(B) + (n + d + d')\mathrm{poly}(r/\varepsilon)) + \tau(d, \mathrm{poly}(r/\varepsilon), \mathrm{poly}(r/\varepsilon), r, \varepsilon).$$

That is, sketching can be used to reduce to a problem in which the only remaining large matrix dimension is $d$, the number of columns of $A$.

This reduction is a building block for our results for regularized low-rank approximation. Here the regularizer is a real-valued function $f(Y, X)$ on matrices $Y \in \mathbb{R}^{n\times k}, X \in \mathbb{R}^{k\times d}$. We show that under broad conditions on $f(\cdot, \cdot)$, sketching can be applied to

$$\min_{\substack{Y\in\mathbb{R}^{n\times k}\\ X\in\mathbb{R}^{k\times d}}} \|YX - A\|_F^2 + f(Y, X). \tag{4}$$

Our conditions imply fast algorithms when, for example, $f(Y, X) = \|YX\|_{(p)}$, where $\|\cdot\|_{(p)}$ is a Schatten $p$-norm, or when $f(Y, X) = \min\{\lambda_1\|YX\|_{(1)}, \lambda_2\|YX\|_{(2)}\}$, for weights $\lambda_1, \lambda_2$, and more. Of course, there are norms, such as the entriwise $\ell_1$ norm, that do not satisfy these orthogonal invariance conditions.

▶ **Theorem 7** (Implied by Thm. 44). *Let $f(Y, X)$ be a real-valued function on matrices that in each argument is subadditive and invariant under padding by rows or columns of zeros, and also right orthogonally invariant in its right argument and left orthogonally invariant in its left argument.*

*Suppose there is a procedure that solves (4) when $A$, $Y$, and $X$ are $k \times k$ matrices, and $A$ is diagonal, and $YX$ is constrained to be diagonal, taking time $\tau(k)$ for a function $\tau(\cdot)$.*

*Then for general $A$, there is an algorithm that finds a $(1+\varepsilon)$-approximate solution $(\tilde{Y}, \tilde{X})$ in time $O(\mathtt{nnz}(A)) + \tilde{O}(n + d)\mathrm{poly}(k/\varepsilon) + \tau(k)$.*

The proof involves a reduction to small matrices, followed by a reduction, discussed in §D.1, that uses the SVD to reduce to the diagonal case. This result, Corollary 43, generalizes results of [29], who gave such a reduction for $f(Y, X) = \|X\|_F^2 + \|Y\|_F^2$; also, we give a very different proof.

As for related work, [29] survey and extend work in this setting, and propose iterative algorithms for this problem. The regularizers $f(Y, X)$ they consider, and evaluate experimentally, are more general than we can analyze.

The conditions on $f(Y, X)$ are quite general; it may be that for some instances, the resulting problem is NP-hard. Here our reduction would be especially interesting, because the size of the reduced NP-hard problem depends only on $k$.

## 1.2 Basic Definitions and Notation

We denote scalars using Greek letters. Vectors are denoted by $x, y, \ldots$ and matrices by $A, B, \ldots$. We use the convention that vectors are column-vectors. We use $\texttt{nnz}(\cdot)$ to denote the number of nonzeros in a vector or matrix. We denote by $[n]$ the set $\{1, \ldots, n\}$. The notation $\alpha = (1 \pm \gamma)\beta$ means that $(1-\gamma)\beta \leq \alpha \leq (1+\gamma)\beta$. Throughout the paper, $A$ denotes an $n \times d$ matrix, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(n,d)}$ its singular values.

▶ **Definition 8** (Schatten $p$-norm)**.** The *Schatten $p$-norm* of $A$ is $\|A\|_{(p)} \equiv [\sum_i \sigma_i^p]^{1/p}$. Note that the trace (nuclear) norm $\|A\|_* = \|A\|_{(1)}$, the Frobenius norm $\|A\|_F = \|A\|_{(2)}$, and the spectral norm $\|A\|_2 = \|A\|_{(\infty)}$.

The notation $\|\cdot\|$ without a subscript denotes the $\ell_2$ norm for vectors, and the spectral norm for matrices. We use a subscript for other norms. We use $\texttt{range}(A)$ to denote the subspace spanned by the columns of $A$, i.e. $\texttt{range}(A) \equiv \{Ax \mid x \in \mathbb{R}^d\}$. $I_d$ denotes the $d \times d$ identity matrix, $0_d$ denotes the column vector comprising $d$ entries of zero, and $0_{a\times b} \in \mathbb{R}^{a\times b}$ denotes a zero matrix.

The rank $\texttt{rank}(A)$ of a matrix $A$ is the dimension of the subspace $\texttt{range}(A)$ spanned by its columns (equivalently, the number of its non-zero singular values). Bounds on sketch sizes are often written in terms of the rank of the matrices involved.

▶ **Definition 9** (Stable Rank)**.** The *stable rank* $\texttt{sr}(A) \equiv \|A\|_F^2/\|A\|_2^2$. The stable rank satisfies $\texttt{sr}(A) \leq \texttt{rank}(A)$.

**Paper Outline:** Due to space constraints, most proofs are omitted, and all results except our results for ridge regression and ridge low-rank approximation are deferred to the appendix. The missing proofs and results can also be found in the full version of our paper on arXiv under the same title: `https://arxiv.org/abs/1611.03225`.

## 2 Ridge Regression

Let $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, and $\lambda > 0$. In this section we consider the *ridge regression* problem:

$$\min_{x\in\mathbb{R}^d} \|Ax - b\|^2 + \lambda\|x\|^2, \tag{5}$$

Let $x^* \equiv \text{argmin}_{x\in\mathbb{R}^d}\|Ax - b\|^2 + \lambda\|x\|^2$ and $\Delta_* \equiv \|Ax^* - b\|^2 + \lambda\|x^*\|^2$. In general $x^* = (A^\top A + \lambda I_d)^{-1}A^\top b = A^\top(AA^\top + \lambda I_n)^{-1}b$, so $x^\star$ can be found in $O(\texttt{nnz}(A)\min(n,d))$ time using an iterative method (e.g., LSQR). Our goal in this section is to design faster algorithms that find an approximate $\tilde{x}$ in the following sense:

$$\|A\tilde{x} - b\|^2 + \lambda\|\tilde{x}\|^2 \leq (1+\varepsilon)\Delta_* \,. \tag{6}$$

In our analysis, we distinguish between two cases: $n \gg d$ and $d \gg n$.

▶ Remark. In this paper we consider only approximations of the form (6). Although we do not explore it in this paper, our techniques can also be used to derive preconditioned methods. Analysis of preconditioned kernel ridge regression, which is related to the $d \gg n$ case, is explored in [4].

## 2.1 Large $n$

In this subsection we design an algorithm that is aimed at the case when $n \gg d$. However, the results themselves are correct even when $n < d$. The general strategy is to design a distribution on matrices of size $m$-by-$n$ ($m$ is a parameter), sample an $S$ from that distribution, and solve $\tilde{x} \equiv \operatorname{argmin}_{x\in\mathbb{R}^d} \|S(Ax-b)\|^2 + \lambda\|x\|^2$ .

The following lemma defines conditions on the distribution that guarantee that (6) holds with constant probability (which can be boosted to high probability by repetition and taking the solution with minimum objective value).

▶ **Lemma 10.** *Let $x^* \in \mathbb{R}^d$, $A$ and $b$ as above. Let $U_1 \in \mathbb{R}^{n\times d}$ comprise the first $n$ rows of an orthogonal basis for $\left[\begin{smallmatrix} A \\ \sqrt{\lambda} I_d \end{smallmatrix}\right]$. Let sketching matrix $S \in \mathbb{R}^{m\times n}$ have a distribution such that with constant probability*

$$\|U_1^\top S^\top S U_1 - U_1^\top U_1\|_2 \le 1/4, \tag{7}$$

*and*

$$\|U_1^\top S^\top S(b - Ax^*) - U_1^\top(b - Ax^*)\| \le \sqrt{\varepsilon\Delta_*/2}. \tag{8}$$

*Then with constant probability, $\tilde{x} \equiv \operatorname{argmin}_{x\in\mathbb{R}^d}\|S(Ax-b)\|^2+\lambda\|x\|^2$ has $\|A\tilde{x}-b\|^2+\lambda\|\tilde{x}\|^2 \le (1+\varepsilon)\Delta_*$.*

**Proof.** Omitted in this version. ◀

▶ **Lemma 11.** *For $U_1$ as in Lemma 10, $\|U_1\|_F^2 = \mathtt{sd}_\lambda(A) = \sum_i 1/(1+\lambda/\sigma_i^2)$, where $A$ has singular values $\sigma_i$. Also $\|U_1\|_2 = 1/\sqrt{1+\lambda/\sigma_1^2}$.*

This follows from (3.47) of [20]; for completeness, a proof is given here.

**Proof.** Suppose $A = U\Sigma V^\top$, the full SVD, so that $U \in \mathbb{R}^{n\times n}$, $\Sigma \in \mathbb{R}^{n\times d}$, and $V \in \mathbb{R}^{d\times d}$. Let $D \equiv (\Sigma^\top\Sigma + \lambda I_d)^{-1/2}$. Then $\hat{A} = \left[\begin{smallmatrix} U\Sigma D \\ V\sqrt{\lambda}D \end{smallmatrix}\right]$ has $\hat{A}^\top\hat{A} = I_d$, and for given $x$, there is $y = D^{-1}V^\top x$ with $\hat{A}y = \left[\begin{smallmatrix} A \\ \sqrt{\lambda} I_d \end{smallmatrix}\right]x$. We have $\|U_1\|_F^2 = \|U\Sigma D\|_F^2 = \|\Sigma D\|_F^2 = \sum_i 1/(1+\lambda/\sigma_i^2)$ as claimed. Also $\|U_1\|_2 = \|U\Sigma D\|_2 = \|\Sigma D\|_2 = 1/\sqrt{1+\lambda/\sigma_1^2}$, and the lemma follows. ◀

▶ **Definition 12** (large $\lambda$)**.** Say that $\lambda$ is *large* for $A$ with largest singular value $\sigma_1$, and error parameter $\varepsilon$, if $\lambda/\sigma_1^2 \ge 1/\varepsilon$.

The following lemma implies that if $\lambda$ is large, then $x = 0$ is a good approximate solution, and so long as we include a check that a proposed solution is no worse than $x = 0$, we can assume that $\lambda$ is not large.

▶ **Lemma 13.** *For $\varepsilon \in (0,1]$, large $\lambda$, and all $x$, $\|Ax-b\|^2 + \lambda\|x\|^2 \ge \|b\|^2/(1+\varepsilon)$. If $\lambda$ is not large then $\|U_1\|_2^2 \ge \varepsilon/2$.*

**Proof.** If $\sigma_1\|x\| \ge \|b\|$, then $\lambda\|x\|^2 \ge \sigma_1^2\|x\|^2 \ge \|b\|^2$. Suppose $\sigma_1\|x\| \le \|b\|$. Then:

$$\begin{aligned}
\|Ax-b\|^2 + \lambda\|x\|^2 &= \|Ax\|^2 + \|b\|^2 - 2b^\top Ax + \lambda\|x\|^2 \\
&\ge (\|b\| - \|Ax\|)^2 + \lambda\|x\|^2 && \text{Cauchy-Schwartz} \\
&\ge (\|b\| - \sigma_1\|x\|)^2 + \lambda\|x\|^2 && \text{assumption} \\
&\ge \|b\|^2/(1+\sigma_1^2/\lambda) && \text{calculus} \\
&\ge \|b\|^2/(1+\varepsilon), && \text{large } \lambda
\end{aligned}$$

as claimed. The last statement follows from Lemma 11. ◀

Below we discuss possibilities for choosing the sketching matrix $S$. We want to emphasize that the first condition in Lemma 10 is *not* a subspace embedding guarantee, despite having superficial similarity. Indeed, notice that the columns of $U_1$ are not orthonormal, since we only take the first $n$ rows of an orthogonal basis of $\left[\begin{smallmatrix} A \\ \sqrt{\lambda} I_d \end{smallmatrix}\right]$. Rather, the first condition is an instance of approximate matrix product with a spectral norm guarantee with constant error, for which optimal bounds in terms of the stable rank $\mathtt{sr}(U_1)$ were recently obtained [13]. As we discuss in the proof of part (i) of Corollary 14 below, $\mathtt{sr}(U_1)$ is upper bounded by $\mathtt{sd}_\lambda(A)/\epsilon$.

We only mention a few possibilities of sketching matrix $S$ below, though others are possible with different tradeoffs and compositions.

▶ **Corollary 14.** *Suppose $\lambda$ is not large (Def. 12). There is a constant $K > 0$ such that for*

**(i)** *$m \geq K(\varepsilon^{-1}\, \mathtt{sd}_\lambda(A) + \mathtt{sd}_\lambda(A)^2)$ and $S \in \mathbb{R}^{m\times n}$ a sparse embedding matrix (see [10, 23, 24]) with $SA$ computable in $O(\mathtt{nnz}(A))$ time, or one can choose $m \geq K(\varepsilon^{-1}\, \mathtt{sd}_\lambda(A) + \min((\mathtt{sd}_\lambda(A)/\epsilon)^{1+\gamma}, \mathtt{sd}_\lambda(A)^2))$ an OSNAP (see [24, 6, 12]) with $SA$ computable in $O(\mathtt{nnz}(A))$ time, where $\gamma > 0$ is an arbitrarily small constant, or*

**(ii)** *$m \geq K\varepsilon^{-1}(\mathtt{sd}_\lambda(A) + \log(1/\varepsilon))\log(\mathtt{sd}_\lambda(A)/\varepsilon)$ and $S \in \mathbb{R}^{m\times n}$ a Subsampled Randomized Hadamard Transform (SRHT) embedding matrix (see, e.g., [7]), with $SA$ computable in $O(nd\log n)$ time, or*

**(iii)** *$m \geq K\varepsilon^{-1}\, \mathtt{sd}_\lambda(A)$ and $S \in \mathbb{R}^{m\times n}$ a matrix of i.i.d. subgaussian values with $SA$ computable in $O(ndm)$ time,*

*the conditions (7) and (8) of Lemma 10 apply, and with constant probability the corresponding $\tilde{x} = \operatorname{argmin}_{x\in\mathbb{R}^d}\|S(Ax-b)\| + \lambda\|x\|^2$ is an $\varepsilon$-approximate solution to $\min_{x\in\mathbb{R}^d}\|b-Ax\|^2 + \lambda\|x\|^2$.*

**Proof.** Recall that $\mathtt{sd}_\lambda(A) = \|U_1\|_F^2$. For (i): sparse embedding distributions satisfy the bound for matrix multiplication

$$\|W^\top S^\top SH - W^\top H\|_F \leq C\|W\|_F\|H\|_F/\sqrt{m},$$

for a constant $C$ [10, 23, 24]; this is also true of OSNAP matrices. We set $W = H = U_1$ and use $\|X\|_2 \leq \|X\|_F$ for all $X$ and $m \geq K\|U_1\|_F^4$ to obtain (7), and set $W = U_1$, $H = b - Ax^*$ and use $m \geq K\|U_1\|_F^2/\varepsilon$ to obtain (8). (Here the bound is slightly stronger than (8), holding for $\lambda = 0$.) With (7) and (8), the claim for $\tilde{x}$ from a sparse embedding follows using Lemma 10.

For OSNAP, Theorem 1 in [13] together with [24] imply that for $m = O(\mathtt{sr}(U_1)^{1+\gamma})$, condition (7) holds. Here $\mathtt{sr}(U_1) = \frac{\|U_1\|_F^2}{\|U_1\|_2^2}$, and by Lemma 11 and Lemma 13, $\mathtt{sr}(U_1) \leq \mathtt{sd}_\lambda(A)/\epsilon$. We note that (8) continues to hold as in the previous paragraph. Thus, $m$ is at most the min of $O((\mathtt{sd}_\lambda(A)/\epsilon)^{1+\gamma})$ and $O(\mathtt{sd}_\lambda(A)/\epsilon + \mathtt{sd}_\lambda(A)^2)$.

For (ii): Theorems 1 and 9 of [13] imply that for $\gamma \leq 1$, with constant probability

$$\|W^\top S^\top SH - W^\top H\|_2 \leq \gamma\|W\|_2\|H\|_2 \tag{9}$$

for SRHT $S$, when

$$m \geq C(\mathtt{sr}(W) + \mathtt{sr}(H) + \log(1/\gamma))\log(\mathtt{sr}(W) + \mathtt{sr}(H))/\gamma^2$$

for a constant $C$. We let $W = H = U_1$ and $\gamma = \min\{1, 1/4\|U_1\|^2\}$. We have

$$\|U_1^\top S^\top SU_1 - U_1^\top U_1\|_2 \leq \min\{1, 1/4\|U_1\|^2\}\|U_1\|_2^2 = \min\{\|U_1\|_2^2, 1/4\} \leq 1/4,$$

and

$$\mathtt{sr}(U_1)/\gamma^2 = \frac{\|U_1\|_F^2}{\|U_1\|_2^2}\max\{1, 4\|U_1\|_2^2\} = \|U_1\|_F^2 \max\{1/\|U_1\|_2^2, 4\} \le 2\|U_1\|_F^2/\varepsilon$$

using Lemma 13 and the assumption that $\lambda$ is large. (And assuming $\varepsilon \le 1/2$.) Noting that $\log(1/\gamma) = O(\log(1/\varepsilon))$ and $\log(\mathtt{sr}(U_1)) = O(\log\|U_1\|_F/\varepsilon)$ using Lemma 13, we have that $m$ as claimed suffices for (7).

For (8), we use (9) with $W = U_1$, $H = Ax^* - b$, and $\gamma = \sqrt{\varepsilon/2}/\|U_1\|_2$; note that using Lemma 13 and by the assumption that $\lambda$ is large, $\gamma \le 1$ and so (9) can be applied. We have

$$\|U_1^\top S^\top S(Ax^* - b)\| \le (\sqrt{\varepsilon/2}/\|U_1\|_2)\|U_1\|_2\|Ax^* - b\| \le \sqrt{\varepsilon\Delta_*/2},$$

and

$$\mathtt{sr}(U_1)\log(\mathtt{sr}(U_1))/\gamma^2 \le \frac{\|U_1\|_F^2}{\|U_1\|_2^2}[2\log(\|U_1\|_F/\varepsilon)][2\|U_1\|_2^2/\varepsilon] = 4\|U_1\|_F^2\log(\|U_1\|_F/\varepsilon)/\varepsilon.$$

Noting that since $Ax^* - b$ is a vector, its stable rank is one, we have that $m$ as claimed suffices for (8). With (7) and (8), the claim for $\tilde{x}$ from an SRHT follows using Lemma 10.

The claim for (iii) follows as (ii), with a slightly simpler expression for $m$. ◀

Here we mention the specific case of composing a sparse embedding matrix with an SRHT.

▶ **Theorem 15.** *Given $A \in \mathbb{R}^{n\times d}$, there are dimensions within constant factors of those given in Cor. 14 such that for $S_1$ a sparse embedding and $S_2$ an SRHT with those dimensions,*

$$\tilde{x} \equiv \operatorname*{argmin}_{x\in\mathbb{R}^d}\|S_2S_1(Ax - b)\|^2 + \lambda\|x\|^2,$$

*satisfies $\|A\tilde{x} - b\|^2 + \lambda\|\tilde{x}\|^2 \le (1+\varepsilon)\min_{x\in\mathbb{R}^d}\|Ax - b\|^2 + \lambda\|x\|^2$ with constant probability.*

*Therefore in $O(\mathtt{nnz}(A)) + \tilde{O}(d\,\mathtt{sd}_\lambda(A)/\varepsilon + \mathtt{sd}_\lambda(A)^2)$ time, a ridge regression problem with $n$ rows can be reduced to one with $O(\varepsilon^{-1}(\mathtt{sd}_\lambda(A) + \log(1/\varepsilon))\log(\mathtt{sd}_\lambda(A)/\varepsilon))$ rows, whose solution is a $(1+\varepsilon)$-approximate solution.*

**Proof.** This follows from Corollary 14 and the general comments of Appendix A.3 of [13]; the results there imply that $\|S_iU_1\|_F = \Theta(\|U_1\|_F)$ and $\|S_iU_1\|_2 = \Theta(\|U_1\|_2)$ for $i \in [3]$ with constant probability, which implies that $\mathtt{sr}(S_1U_1)$ and $\mathtt{sr}(S_2S_1U_1)$ are $O(\mathtt{sr}(U_1))$. Moreover, the approximate multiplication bounds of (7) and (8) have versions when using $S_2S_1U_1$ and $S_2S_1(Ax^* - b)$ to estimate products involving $S_1U_1$ and $S_1(Ax^* - b)$, so that for example, using the triangle inequality,

$$\begin{aligned}\|U_1^\top S_1^\top S_2^\top S_2S_1U_1 - U_1^\top U_1\|_2 &\le \|U_1^\top S_1^\top S_2^\top S_2S_1U_1 - U_1^\top S_1^\top S_1U_1\|_2 \\ &\quad + \|U_1^\top S_1^\top S_1U_1 - U_1^\top U_1\|_2 \\ &\le 1/8 + 1/8 = 1/4.\end{aligned}$$

We have that $S = S_2S_1$ satisfies (7) and (8), as desired. ◀

Similar arguments imply that a reduction also using a sketching matrix $S_3$ with subgaussian entries could be used, to reduce to a ridge regression problem with $O(\varepsilon^{-1}\,\mathtt{sd}_\lambda(A))$ rows.

## 2.2 Large d

If the number of columns is larger than the number of rows, it is more attractive to sketch the rows, i.e., to use $AS^\top$. In general, we can express (5) as $\min_{x\in\mathbb{R}^d}\|Ax\|^2-2b^\top Ax+\|b\|^2+\lambda\|x\|^2$. We can assume $x$ has the form $x = A^\top y$, yielding the equivalent problem

$$\min_{y\in\mathbb{R}^n}\|AA^\top y\|^2 - 2b^\top AA^\top y + \|b\|^2 + \lambda\|A^\top y\|^2. \tag{10}$$

Sketching $A^\top$ with $S$ in the first two terms yields

$$\tilde{y} \equiv \operatorname*{argmin}_{y\in\mathbb{R}^n} \lambda\|SA^\top y\|^2 + \|AS^\top SA^\top y\|^2 - 2b^\top AA^\top y + \|b\|^2 \tag{11}$$

Now let $c^\top \equiv b^\top AA^\top$. Note that we can compute $c$ in $O(\mathtt{nnz}(A))$ time. The solution to (11) is, for $B \equiv SA^\top$ with $B^\top B$ invertible, $\tilde{y} = (\lambda B^\top B + B^\top BB^\top B)^+ c/2$.

In the main result of this subsection, we show that provided $\lambda > 0$ then a sufficiently tight subspace embedding to $\mathtt{range}(A^\top)$ suffices.

▶ **Theorem 16.** *Suppose $A$ has rank $k$, and its SVD is $A = U\Sigma V^\top$, with $U \in \mathbb{R}^{n\times k}$, $\Sigma \in \mathbb{R}^{k\times k}$ and $V \in \mathbb{R}^{d\times k}$. If $S \in \mathbb{R}^{m\times d}$ has*
1. *(Subspace Embedding) $E \equiv V^\top S^\top SV - I_k$ with $\|E\|_2 \le \varepsilon/2$*
2. *(Spectral Norm Approximate Matrix Product) for any fixed matrices $C, D$, each with $d$ rows,*

$$\|C^T S^T SD - C^T D\|_2 \le \varepsilon'\|C\|_2\|D\|_2,$$

*where $\varepsilon' \equiv (\varepsilon/2)/(1+3\sigma_1^2/\lambda)$.*

*Then* (11) *has $\tilde{x} \equiv A^\top\tilde{y}$ approximately solving* (5)*, that is, $\|A\tilde{x}-b\|^2 + \lambda\|\tilde{x}\|^2 \le (1+\varepsilon)\Delta_*$.*

**Proof.** To compare the sketched with the unsketched formulations, let $A$ have full SVD $A = U\Sigma V^\top$, and let $w = \Sigma U^\top y$. Using $\|Uz\| = \|z\|$ and $\|Vw\| = \|w\|$ yields the unsketched problem

$$\min_{w\in\mathbb{R}^k}\|\Sigma w\|^2 - 2b^\top AVw + \|b\|^2 + \lambda\|w\|^2, \tag{12}$$

equivalent to (10). The corresponding sketched version is

$$\min_{w\in\mathbb{R}^k}\|\Sigma V^\top S^\top SVw\|^2 - 2b^\top AVw + \|b\|^2 + \lambda\|SVw\|^2.$$

Now suppose $S$ has $E$ satisfying the first property in the theorem statement. This implies $S$ is an $\varepsilon/2$-embedding for $V$:

$$|\|SVw\|^2 - \|w\|^2| = |w^\top(V^\top S^\top SV - I_k)w| \le (\varepsilon/2)\|w\|^2,$$

and, using the second property in the theorem statement with $C^T = \Sigma V^T$ and $D = V$ (which do not depend on $w$),

$$\|\Sigma V^\top S^\top SV - \Sigma\|_2 = f,$$

where $f$ satisfies $|f| \le \varepsilon'\sigma_1$. It follows by the triangle inequality for any $w$ that

$$\|\Sigma V^\top S^\top SVw\| \in [\|\Sigma w\| - f\|w\|, \|\Sigma w\| + f\|w\|].$$

Hence,

$$\begin{aligned}|\|\Sigma V^\top S^\top S V w\|^2 - \|\Sigma w\|^2| &\in |(\|\Sigma w\| \pm f\|w\|)^2 - \|\Sigma w\|^2| \\ &\leq 2f\|\Sigma w\|\|w\| + f^2\|w\|^2 \\ \\ &\leq 3\varepsilon' \sigma_1^2 \|w\|^2\end{aligned}$$

The value of (12) is at least $\lambda\|w\|^2$, so the relative error of the sketch is at most

$$\frac{\lambda(\varepsilon/2)\|w\|^2 + 3\varepsilon'\sigma_1^2\|w\|^2}{\lambda\|w\|^2} \leq \varepsilon.$$

The statement of the theorem follows. ◀

We now discuss which matrices $S$ can be used in Theorem 16. Note that the first property is just the oblivious subspace embedding property, and we can use CountSketch, Subsampled Randomized Hadamard Transform, or Gaussian matrices to achieve this. One can also use OSNAP matrices [24]; note that here, unlike for Corollary 14, the running time will be $O(\mathtt{nnz}(A)/\epsilon)$ (see, e.g., [30] for a survey). For the second property, we use the recent work of [13], where tight bounds for a number of oblivious subspace embeddings $S$ were shown.

In particular, applying the result in Appendix A.3 of [13], it is shown that the *composition* of matrices each satisfying the second property, results in a matrix also satisfying the second property. It follows that we can let $S$ be of the form $\Pi \cdot \Pi'$, where $\Pi'$ is an $r \times d$ CountSketch matrix, where $r = O(n^2/(\epsilon')^2)$, and $\Pi$ is an $\tilde{O}(n/(\epsilon')^2) \times r$ Subsampled Randomized Hadamard Transform. By standard results on oblivious subspace embeddings, the first property of Theorem 16 holds provided $r = \Theta(n^2/\epsilon^2)$ and $\Pi$ has $\tilde{O}(n/\epsilon^2)$ rows. Note that $\epsilon' \leq \epsilon$, so in total we have $O(n/(\epsilon')^2)$ rows.

Thus, we can compute $B = \Pi \cdot \Pi' A^T$ in $O(\mathtt{nnz}(A)) + \tilde{O}(n^3/(\epsilon')^2)$ time, and $B$ has $\tilde{O}(n/(\epsilon')^2)$ rows and $n$ columns. We can thus compute $\tilde{y}$ as above in $\tilde{O}(n^3/(\epsilon')^2)$ additional time. Therefore in $O(\mathtt{nnz}(A)) + \tilde{O}(n^3/(\epsilon')^2)$ time, we can solve the problem of (5).

We note that, using our results in Section 2.1, in particular Theorem 15, we can first replace $n$ in the above time complexities with a function of $\mathtt{sd}_\lambda(A)$ and $\varepsilon$, which can further reduce the overall time complexity.

## 2.3 Multiple-response Ridge Regression

In multiple-response ridge regression one is interested in finding $X^* \equiv \operatorname{argmin}_{X \in \mathbb{R}^{d\times d'}} \|AX - B\|_F^2 + \lambda\|X\|_F^2$, where $B \in \mathbb{R}^{n\times d'}$. It is straightforward to extend the results and algorithms for large $n$ to multiple regression. Since we use these results when we consider regularized low-rank approximation, we state them next. The proofs are omitted as they are entirely analogous to the proofs in subsection 2.1.

▶ **Lemma 17.** *Let $A$, $U_1$, $U_2$ as in Lemma 10, $B \in \mathbb{R}^{n\times d'}$,*

$$X^* \equiv \operatorname*{argmin}_{X \in \mathbb{R}^{d\times d'}} \|AX - B\|_F^2 + \lambda\|X\|_F^2,$$

*and $\Delta_* \equiv \|AX^* - B\|_F^2 + \lambda\|X^*\|_F^2$. Let sketching matrix $S \in \mathbb{R}^{m\times n}$ have a distribution such that with constant probability,*

$$\|U_1^\top S^\top S U_1 - U_1^\top U_1\|_2 \leq 1/4, \tag{13}$$

*and*

$$\|U_1^\top S^\top S(B - AX^*) - U_1^\top (B - AX^*)\|_F \le \sqrt{\varepsilon \Delta_*}. \tag{14}$$

*Then with constant probability,*

$$\tilde{X} \equiv \operatorname*{argmin}_{X \in \mathbb{R}^{d \times d'}} \|S(AX - B)\|_F^2 + \lambda \|X\|_F^2 \tag{15}$$

*has* $\|A\tilde{X} - B\|^2 + \lambda\|\tilde{X}\|_F^2 \le (1+\varepsilon)\Delta_*$.

▶ **Theorem 18.** *There are dimensions within a constant factor of those given in Thm. 15, such that for* $S_1$ *a sparse embedding and* $S_2$ *SRHT with those dimensions,* $S = S_2 S_1$ *satisfies the conditions of Lemma 17, therefore the corresponding* $\tilde{X}$ *does as well. That is, in time*

$$O(\mathtt{nnz}(A) + \mathtt{nnz}(B)) + \tilde{O}((d + d')(\mathtt{sd}_\lambda(A)/\varepsilon + \mathtt{sd}_\lambda(A)^2)$$

*time, a multiple-response ridge regression problem with* $n$ *rows can be reduced to one with* $\tilde{O}(\varepsilon^{-1}\,\mathtt{sd}_\lambda(A))$ *rows, whose solution is a* $(1+\varepsilon)$*-approximate solution.*

▶ Remark. Note that the solution to (15), that is, the solution to $\min_X \|\hat{S}(\hat{A}X - \hat{B})\|_F^2$, where $\hat{S}$ and $\hat{A}$ are as defined in the proof of Lemma 10, and $\hat{B} \equiv \begin{bmatrix} B \\ 0_{d \times d'} \end{bmatrix}$, is $\tilde{X} = (\hat{S}\hat{A})^+ \hat{S}\hat{B}$; that is, the matrix $\hat{A}\tilde{X} = \hat{A}(\hat{S}\hat{A})^+\hat{S}\hat{B}$ whose distance to $\hat{B}$ is within $1+\varepsilon$ of optimal has rows in the rowspace of $\hat{B}$, which is the rowspace of $B$. This property will be helpful building low-rank approximations.

## 3 Ridge Low-Rank Approximation

For an integer $k$ we consider the problem

$$\min_{\substack{Y \in \mathbb{R}^{n \times k} \\ X \in \mathbb{R}^{k \times d}}} \|YX - A\|_F^2 + \lambda\|Y\|_F^2 + \lambda\|X\|_F^2. \tag{16}$$

From [29] (see also Corollary 43 below), this has the solution

$$\begin{aligned} Y^* &= U_k(\Sigma_k - \lambda I_k)_+^{1/2} \\ X^* &= (\Sigma_k - \lambda I_k)_+^{1/2} V_k^\top \\ \implies \mathtt{sd}_\lambda(Y^*) &= \mathtt{sd}_\lambda(X^*) = \sum_{\substack{i \in [k] \\ \sigma_i > \lambda}} (1 - \lambda/\sigma_i) \end{aligned} \tag{17}$$

where $U_k \Sigma_k V_k^\top$ is the best rank-$k$ approximation to $A$, and for a matrix $W$, $W_+$ has entries that are equal to the corresponding entries of $W$ that are nonnegative, and zero otherwise.

While [29] gives a general argument, it was also known (see for example [27]) that when the rank $k$ is large enough not to be an active constraint (say, $k = \mathtt{rank}(A)$), then $Y^*X^*$ for $Y^*, X^*$ from (17) solves

$$\min_{Z \in \mathbb{R}^{n \times d}} \|Z - A\|_F^2 + 2\lambda\|Z\|_*,$$

where $\|Z\|_*$ is the nuclear norm of $X$ (also called the trace norm).

It is also well-known that

$$\|Z\|_* = \frac{1}{2}(\min_{YX=Z} \|Y\|_F^2 + \|X\|_F^2),$$

so that the optimality of (17) follows for large $k$.

▶ **Lemma 19.** *Given integer $k \geq 1$ and $\varepsilon > 0$, $Y^*$ and $X^*$ as in* (17)*, there are*

$$m = \tilde{O}(\varepsilon^{-1}\,\mathtt{sd}_\lambda(Y^*)) = \tilde{O}(\varepsilon^{-1}k) \text{ and } m' = \tilde{O}(\varepsilon^{-1}\min\{k, \varepsilon^{-1}\,\mathtt{sd}_\lambda(Y^*)\}),$$

*such that there is a distribution on $S \in \mathbb{R}^{m\times n}$ and $R \in \mathbb{R}^{d\times m'}$ so that for*

$$Z_S^*, Z_R^* \equiv \operatorname*{argmin}_{\substack{Z_S\in\mathbb{R}^{k\times m}\\ Z_R\in\mathbb{R}^{m'\times k}}} \|ARZ_RZ_SSA - A\|_F^2 + \lambda\|ARZ_R\|_F^2 + \lambda\|Z_SSA\|_F^2,$$

*with constant probability $\tilde{Y} \equiv ARZ_R^*$ and $\tilde{X} \equiv Z_S^*SA$ satisfy*

$$\|\tilde{Y}\tilde{X} - A\|_F^2 + \lambda\|\tilde{Y}\|_F^2 + \lambda\|\tilde{X}\|_F^2 \leq (1+\varepsilon)(\|Y^*X^* - A\|_F^2 + \lambda\|Y^*\|_F^2 + \lambda\|X^*\|_F^2).$$

*The products $SA$ and $AR$ take altogether $O(\mathtt{nnz}(A)) + \tilde{O}((n+d)(\varepsilon^{-2}\,\mathtt{sd}_\lambda(Y^*) + \varepsilon^{-1}\,\mathtt{sd}_\lambda(Y^*)^2)$ to compute.*

**Proof.** Omitted in this version. ◀

We can reduce to an even yet smaller problem, using affine embeddings, which are built using subspace embeddings. These are defined next.

▶ **Definition 20** (subspace embedding)**.** Matrix $S \in \mathbb{R}^{m_S\times n}$ is a *subspace $\varepsilon$-embedding* for $A$ with respect to the Euclidean norm if $\|SAx\|_2 = (1\pm\varepsilon)\|Ax\|_2$ for all $x$.

▶ **Lemma 21.** *There are* sparse embedding *distributions on matrices $S \in \mathbb{R}^{m\times n}$ with $m = O(\varepsilon^{-2}\,\mathtt{rank}(A)^2)$ so that $SA$ can be computed in $\mathtt{nnz}(A)$ time, and with constant probability $S$ is a subspace $\varepsilon$-embedding. The SRHT (of Corollary 14) is a distribution on $S \in \mathbb{R}^{m\times n}$ with $m = \tilde{O}(\varepsilon^{-2}\,\mathtt{rank}(A))$ such that $S$ is a subspace embedding with constant probability.*

**Proof.** The sparse embedding claim is from [10], sharpened by [24, 23]; the SRHT claim is from for example [7]. ◀

▶ **Definition 22** (Affine Embedding)**.** For $A$ as usual and $B \in \mathbb{R}^{n\times d'}$, matrix $S$ is an *affine $\varepsilon$-embedding* for $A, B$ if $\|S(AX-B)\|_F^2 = (1\pm\varepsilon)\|AX-B\|_F^2$ for all $X \in \mathbb{R}^{d\times d'}$. A distribution over $\mathbb{R}^{m_S\times n}$ is a *poly-sized affine embedding* distribution if there is $m_S = \text{poly}(d/\varepsilon)$ such that constant probability, $S$ from the distribution is an affine $\varepsilon$-embedding.

▶ **Lemma 23.** *For $A$ as usual, $B \in \mathbb{R}^{n\times d'}$, suppose there is a distribution over $S \in \mathbb{R}^{m\times n}$ so that with constant probability, $S$ is a subspace embedding for $A$ with parameter $\varepsilon$, and for $X^* \equiv \operatorname{argmin}_{X\in\mathbb{R}^{d\times d'}}\|AX-B\|_F^2$ and $B^* \equiv AX^* - B$, $\|SB*\|_F^2 = (1\pm\varepsilon)\|B^*\|_F^2$ and $\|U^\top S^\top SB^* - U^\top B^*\| \leq \varepsilon\|B^*\|_F^2$. Then $S$ is an affine embedding for $A, B$. A sparse embedding with $m = O(\mathtt{rank}(A)^2/\varepsilon^2)$ has the needed properties. By first applying a sparse embedding $\Pi$, and then a Subsampled Randomized Hadamard Transform (SHRT) $T$, there is an affine $\varepsilon$-embedding $S = T\Pi$ with $m = \tilde{O}(\mathtt{rank}(A)/\varepsilon^2)$ taking time $O(\mathtt{nnz}(A) + \mathtt{nnz}(B)) + \tilde{O}((d+d')\,\mathtt{rank}(A)^{1+\kappa}/\varepsilon^2)$ time to apply to $A$ and $B$, that is, to compute $SA = T\Pi A$ and $SB$. Here $\kappa > 0$ is any fixed value.*

**Proof.** Shown in [10], sharpened with [24, 23]. ◀

▶ **Theorem 24.** *With notation as in Lemma 19, there are*

$$p' = \tilde{O}(\varepsilon^{-2}m) = \tilde{O}(\varepsilon^{-3}\,\mathtt{sd}_\lambda(Y^*)) = \tilde{O}(\varepsilon^{-3}k) \text{ and}$$
$$p = \tilde{O}(\varepsilon^{-2}m') = \tilde{O}(\varepsilon^{-3}\min\{k, \varepsilon^{-1}\,\mathtt{sd}_\lambda(Y^*)\}),$$

*such that there is a distribution on $S_2 \in \mathbb{R}^{p\times n}$, $R_2 \in \mathbb{R}^{d\times p'}$ so that for*

$$\tilde{Z}_S, \tilde{Z}_R \equiv \operatorname*{argmin}_{\substack{Z_S\in\mathbb{R}^{k\times m}\\ Z_R\in\mathbb{R}^{m'\times k}}} \|S_2ARZ_RZ_SSAR_2 - S_2AR_2\|_F^2 + \lambda\|S_2ARZ_R\|_F^2 + \lambda\|Z_SSAR_2\|_F^2,$$

*with constant probability $\tilde{Y} \equiv AR\tilde{Z}_R$ and $\tilde{X} \equiv \tilde{Z}_SSA$ satisfy*

$$\|\tilde{Y}\tilde{X} - A\|_F^2 + \lambda\|\tilde{Y}\|_F^2 + \lambda\|\tilde{X}\|_F^2 \leq (1+\varepsilon)(\|Y^*X^* - A\|_F^2 + \lambda\|Y^*\|_F^2 + \lambda\|X^*\|_F^2).$$

*The matrices $S_2AR$, $SAR$, and $SAR_2$ can be computed in $O(\mathtt{nnz}(A)) + \mathrm{poly}(\mathtt{sd}_\lambda(Y^*)/\varepsilon)$ time.*

**Proof.** Omitted in this version. ◀

▶ **Lemma 25.** *For $C \in \mathbb{R}^{p\times m'}$, $D \in \mathbb{R}^{m\times p'}$, $G \in \mathbb{R}^{p\times p'}$, the problem of finding*

$$\min_{\substack{Z_S\in\mathbb{R}^{k\times m}\\ Z_R\in\mathbb{R}^{m'\times k}}} \|CZ_RZ_SD - G\|_F^2 + \lambda\|CZ_R\|_F^2 + \lambda\|Z_SD\|_F^2, \tag{18}$$

*and the minimizing $CZ_R$ and $Z_SD$, can be solved in*

$$O(pm'r_C + p'mr_D + r_Dp(p' + r_C))$$

*time, where $r_C \equiv \mathtt{rank}(C) \leq \min\{m', p\}$, and $r_D \equiv \mathtt{rank}(D) \leq \min\{m, p'\}$.*

**Proof.** Please see §E. ◀

▶ **Theorem 26.** *The matrices $\tilde{Z}_S, \tilde{Z}_R$ of Theorem 24 can be found in*

$$O(\mathtt{nnz}(A)) + \mathrm{poly}(\mathtt{sd}_\lambda(Y^*)/\varepsilon)$$

*time, in particular $O(\mathtt{nnz}(A)) + \tilde{O}(\varepsilon^{-7}\,\mathtt{sd}_\lambda(Y^*)^2\ \min\{k, \varepsilon^{-1}\,\mathtt{sd}_\lambda(Y^*)\})$ time, such that with constant probability, $AR\tilde{Z}_R, \tilde{Z}_SSA$ is an $\varepsilon$-approximate minimizer to (16), that is,*

$$\|(AR\tilde{Z}_R)(\tilde{Z}_SSA) - A\|_F^2 + \lambda\|AR\tilde{Z}_R\|_F^2 + \lambda\|\tilde{Z}_SSA\|_F^2 \tag{19}$$

$$\leq (1+\varepsilon) \min_{\substack{Y\in\mathbb{R}^{n\times k}\\ X\in\mathbb{R}^{k\times d}}} \|YX - A\|_F^2 + \lambda\|Y\|_F^2 + \lambda\|X\|_F^2. \tag{20}$$

*With an additional $O(n+d)\mathrm{poly}(\mathtt{sd}_\lambda(Y^*)/\varepsilon)$ time, and in particular*

$$\tilde{O}(\varepsilon^{-1}k\,\mathtt{sd}_\lambda(Y^*)(n + d + \min\{n, d\}\min\{k/\,\mathtt{sd}_\lambda(Y^*), \varepsilon^{-1}\}))$$

*time, the solution matrices $\tilde{Y} \equiv AR\tilde{Z}_R, \tilde{X} \equiv \tilde{Z}_SSA$ can be computed and output.*

An expression for $\mathtt{sd}_\lambda(Y^*)$ is given at (17).

**Proof.** Follows from Theorem 24 and Lemma 25, noting that for efficiency's sake we can use the transpose of $A$ instead of $A$. ◀

**Acknowledgement.** We acknowledge the support from XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323.

## References

1 Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *ACM Symposium on Theory of Computing (STOC)*, 2006.

2 Alexandr Andoni and Huy L. Nguyen. Eigenvalues of a matrix in the streaming model. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1729–1737. Society for Industrial and Applied Mathematics, 2013.

3 Haim Avron, Christos Boutsidis, Sivan Toledo, and Anastasios Zouzias. Efficient dimensionality reduction for canonical correlation analysis. *SIAM Journal on Scientific Computing*, 36(5):S111–S131, 2014. doi:10.1137/130919222.

4 Haim Avron, Kenneth L. Clarkson, and David P. Woodruff. Faster kernel ridge regression using sketching and preconditioning. *CoRR*, abs/1611.03220, 2016. URL: http://arxiv.org/abs/1611.03220.

5 A. Björck and G. H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

6 Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in euclidean space. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 499–508, 2015.

7 C. Boutsidis and A. Gittens. Improved matrix algorithms via the Subsampled Randomized Hadamard Transform. *ArXiv e-prints*, March 2012. arXiv:1204.0062.

8 Ricardo Cabral, Fernando De la Torre, João P Costeira, and Alexandre Bernardino. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2488–2495. IEEE, 2013.

9 Shouyuan Chen, Yang Liu, Michael Lyu, Irwin King, and Shengyu Zhang. Fast relative-error approximation algorithm for ridge regression. In *31st Conference on Uncertainty in Artificial Intelligence*, 2015.

10 Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *STOC*, 2013. Full version at http://arxiv.org/abs/1207.6365.

11 M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu. Dimensionality Reduction for k-Means Clustering and Low Rank Approximation. *ArXiv e-prints*, October 2014. arXiv:1410.6801.

12 Michael B. Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 278–287, 2016.

13 Michael B. Cohen, Jelani Nelson, and David P. Woodruff. Optimal approximate matrix product in terms of stable rank. *CoRR*, abs/1507.02268, 2015.

14 P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Sampling algorithms for $\ell_2$ regression and applications. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1127–1136, 2006.

15 P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlos. Faster least squares approximation, Technical Report, arXiv:0710.1435, 2007. URL: http://www.citebase.org/abstract?id=oai:arXiv.org:0710.1435.

16 Petros Drineas, Michael W. Mahoney, Malik Magdon-Ismail, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 – July 1, 2012*, 2012.

17 Ahmed El Alaoui and Michael W. Mahoney. Fast randomized kernel methods with statistical guarantees. *stat*, 1050:2, 2014.

18 R. Frostig, R. Ge, S. M. Kakade, and A. Sidford. Competing with the Empirical Risk Minimizer in a Single Pass. *ArXiv e-prints*, December 2014. Appeared in COLT 2015. `arXiv:1412.6606`.

19 R. Frostig, R. Ge, S. M. Kakade, and A. Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *International Conference on Machine Learning (ICML)*, 2015.

20 Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2013.

21 Yichao Lu, Paramveer Dhillon, Dean P Foster, and Lyle Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *Advances in Neural Information Processing Systems*, pages 369–377, 2013.

22 Michael W. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, 3(2):123–224, February 2011. `doi:10.1561/2200000035`.

23 Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *STOC*, pages 91–100, 2013.

24 Jelani Nelson and Huy L. Nguyen. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *FOCS*, pages 117–126, 2013.

25 Mert Pilanci and Martin J. Wainwright. Randomized sketches of convex programs with sharp guarantees. *CoRR*, abs/1404.7203, 2014. URL: `http://arxiv.org/abs/1404.7203`.

26 T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2006.

27 Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *Learning Theory*, pages 545–560. Springer, 2005.

28 Joel Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Adv. Adapt. Data Anal., Special Issue, "Sparse Representation of Data and Images"*, 2011.

29 M. Udell, C. Horn, R. Zadeh, and S. Boyd. Generalized Low Rank Models. *ArXiv e-prints*, October 2014. `arXiv:1410.0342`.

30 David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014. `doi:10.1561/0400000060`.

31 Jiyan Yang, Xiangrui Meng, and M. W. Mahoney. Implementing randomized matrix algorithms in parallel and distributed environments. *Proceedings of the IEEE*, 104(1):58–92, Jan 2016. `doi:10.1109/JPROC.2015.2494219`.

32 Y. Yang, M. Pilanci, and M. J. Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *ArXiv e-prints*, January 2015. `arXiv:1501.06195`.

33 Dean Foster Yichao Lu, Paramveer Dhillon and Lyle Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *Proceedings of the Neural Information Processing Systems (NIPS) Conference*, 2013.

## A Estimation of statistical dimension

▶ **Theorem 27.** *If the statistical dimension* $\mathtt{sd}_\lambda(A)$ *is at most*

$$M \equiv \min\{n, d, \lfloor (n+d)^{1/3}/\mathrm{poly}(\log(n+d))\rfloor\},$$

*it can be estimated to within a constant factor in* $O(\mathtt{nnz}(A))$ *time, with constant probability.*

**Proof.** From Lemma 18 of [11], generalizing the machinery of [2], the first $z$ squared singular values of $A$ can be estimated up to additive $\frac{\varepsilon}{z}\|A_{-z}\|_F^2$ in time $O(\mathtt{nnz}(A)) + \tilde{O}(z^3/\mathrm{poly}(\varepsilon))$, where $A_{-z} \equiv A - A_z$ denotes the residual error of the best rank-$z$ approximation $A_z$ to $A$. Therefore $\|A_z\|_F^2$ can be estimated up to additive $\varepsilon\|A_{-z}\|_F^2$, and the same for $\|A_{-z}\|_F^2$. This implies that for small enough constant $\varepsilon$, $\|A_{-z}\|_F^2$ can be estimated up to constant relative error, using the same procedure.

Thus in $O(\mathtt{nnz}(A))$ time, the first $6M$ singular values of $A$ can be estimated up to additive $\frac{1}{6M}\|A_{-6M}\|_F^2$ error, and there is an estimator $\hat{\gamma}_z$ of $\|A_{-z}\|_F^2$ up to relative error $1/3$, for $z \in [6M]$.

Since $1/(1+\lambda/\sigma_i^2) \leq \min\{1, \sigma_i^2/\lambda\}$, for any $z$ the summands of $\mathtt{sd}_\lambda(A)$ for $i \leq z$ are at most 1, while those for $i > z$ are at most $\sigma_i^2/\lambda$, and so $\mathtt{sd}_\lambda(A) \leq z + \|A_{-z}\|_F^2/\lambda$.

When $\sigma_z^2 \leq \lambda$, the summands of $\mathtt{sd}_\lambda(A)$ for $i \geq z$ are at least $\frac{1}{2}\frac{\sigma_i^2}{\lambda}$, and so $\mathtt{sd}_\lambda(A) \geq \frac{1}{2}\|A_{-z}\|_F^2/\lambda$. When $\sigma_z^2 \geq \lambda$, the summands of $\mathtt{sd}_\lambda(A)$ for $i \leq z$ are at least $1/2$. Therefore $\mathtt{sd}_\lambda(A) \geq \frac{1}{2}\min\{z, \|A_{-z}\|_F^2/\lambda\}$.

Under the constant-probability assumption that $\hat{\gamma}_z = (1 \pm 1/3)\|A_{-z}\|_F^2$, we have

$$\frac{3}{8}\min\{z, \hat{\gamma}_z/\lambda\} \leq \mathtt{sd}_\lambda(A) \leq \frac{3}{2}(z + \hat{\gamma}_z/\lambda). \tag{21}$$

Let $z'$ be the smallest $z$ of the form $2^j$ for $j = 0, 1, 2, \ldots$, with $z' \leq 6M$, such that $z' \geq \hat{\gamma}_{z'}/\lambda$. Since $M \geq \mathtt{sd}_\lambda(A) \geq \frac{3}{8}z$ for $z \leq \hat{\gamma}_z/\lambda$, there must be such a $z'$. Then by considering the lower bound of (21) for $z'$ and for $z'/2$, we have $\mathtt{sd}_\lambda(A) \geq \frac{3}{8}\max\{z'/2, \hat{\gamma}_{z'}/\lambda\} \geq \frac{1}{16}(z' + \hat{\gamma}_{z'}/\lambda)$, which combined with the upper bound of (21) implies that $z' + \hat{\gamma}_{z'}/\lambda$ is an estimator of $\mathtt{sd}_\lambda(A)$ up to a constant factor. ◀

## B Regularized Canonical Correlation Analysis

First, we show how to compute regularized CCA using a modified Björck-Golub algorithm.

▶ **Definition 28.** Let $A \in \mathbb{R}^{n\times d}$ with $n \geq d$ and let $\lambda \geq 0$. $A = QR$ is a *$\lambda$-QR factorization* if $Q$ is full rank, $R$ is upper triangular and $R^\top R = A^\top A + \lambda I_d$.

▶ Remark. A $\lambda$-QR factorization always exists, and $R$ will be invertible for $\lambda > 0$. $Q$ has orthonormal columns for $\lambda = 0$.

▶ **Fact 29.** *For a $\lambda$-QR factorization $A = QR$ we have $Q^\top Q + \lambda R^{-\top}R^{-1} = I_d$.*

**Proof.** A direct consequence of $R^\top R = A^\top A + \lambda I_d$ (multiply from the right by $R^{-1}$ and the left by $R^{-\top}$). ◀

▶ **Fact 30.** *For a $\lambda$-QR factorization $A = QR$ we have $\mathtt{sd}_\lambda(A) = \|Q\|_F^2$.*

**Proof.** Omitted in this version. ◀

▶ **Theorem 31** (Regularized Björck-Golub). *Let $A = Q_A R_A$ be a $\lambda_1$-QR factorization of $A$, and $B = Q_B R_B$ be a $\lambda_2$-QR factorization of $B$. Assume that $\lambda_1 > 0$ and $\lambda_2 > 0$. The $(\lambda_1, \lambda_2)$ canonical correlations are exactly the singular values of $Q_A^\top Q_B$. Furthermore, if $Q_A^\top Q_B = M\Sigma N^T$ is a thin SVD of $Q_A^\top Q_B$, then the columns of $R_A^{-1}M$ and $R_B^{-1}N$ are canonical weights.*

**Proof.** Omitted in this version. ◀

We now consider how to approximate the computation using sketching. The basic idea is similar to the one used in [3] to accelerate the computation of non-regularized CCA: compute the regularized canonical correlations and canonical weights of the pair $(SA, SB)$ for a sufficiently large subspace embedding matrix $S$. Similarly to [3], we define the notion of approximate regularized CCA, and show that for large enough $S$ we find an approximate CCA with high probability.

▶ **Definition 32** (Approximate $(\lambda_1, \lambda_2)$ regularized CCA)**.** For $0 \leq \eta \leq 1$, an $\eta$-approximate $(\lambda_1, \lambda_2)$ regularized CCA of $(A, B)$ is a set of positive numbers $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_q$, and vectors $\hat{u}_1, \ldots, \hat{u}_q \in \mathbb{R}^d$ and $\hat{v}_1, \ldots, \hat{v}_q \in \mathbb{R}^{d'}$ such that

**(a)** For every $i$,

$$\left|\hat{\sigma}_i - \sigma_i^{(\lambda_1,\lambda_2)}\right| \leq \eta\,.$$

**(b)** Let $\hat{U} = [\hat{u}_1, \ldots, \hat{u}_q] \in \mathbb{R}^{n \times q}$ and $\hat{V} = [\hat{v}_1, \ldots, \hat{v}_q] \in \mathbb{R}^{d' \times q}$. We have,

$$\left|\hat{U}^\top (A^\top A + \lambda_1 I_d)\hat{U} - I_q\right| \leq \eta$$

and

$$\left|\hat{V}^\top (B^\top B + \lambda_2 I_{d'})\hat{V}s - I_q\right| \leq \eta\,.$$

In the above, the notation $|X| \leq \alpha$ should be understood as entry-wise inequality.

**(c)** For every $i$,

$$\left|\hat{u}_i^\top A^\top B \hat{v}_i - \sigma_i^{(\lambda_1,\lambda_2)}\right| \leq \eta\,.$$

▶ **Theorem 33.** *If $S$ is a sparse embedding matrix with $m = \Omega(\max(\mathtt{sd}_{\lambda_1}(A), \mathtt{sd}_{\lambda_2}(B))^2/\epsilon^2)$ rows, then with high probability the $(\lambda_1, \lambda_2)$ canonical correlations and canonical weights of $(SA, SB)$ form an $\epsilon$-approximate $(\lambda_1, \lambda_2)$ regularized CCA for $(A, B)$.*

**Proof.** Omitted in this version. ◀

Taking an optimization point of view, the following Corollary shows that the suboptimality in the objective is not too big (the fact that the constraints are approximately held is established in the previous theorem).

▶ **Corollary 34.** *Let $U_L$ and $V_L$ (respectively, $\hat{U}_L$ and $\hat{V}_L$) denote the first $L$ columns of $U$ and $V$ (respectively, $\hat{U}$ and $\hat{V}$. Then,*

$$\mathtt{tr}(\hat{U}_L^\top A^\top B \hat{V}_L) \leq \mathtt{tr}(U_L^\top A^\top B V_L) + \epsilon L\,.$$

## C General Regularization: Multiple-response Regression

In this section we consider the problem

$$X^* \equiv \operatorname*{argmin}_{X \in \mathbb{R}^{d \times d'}} \|AX - B\|_F^2 + f(X)$$

for a real-valued function $f$ on matrices. We show that under certain assumptions on $f$ (generalizing from $f(X) = \|X\|_h$ for some orthogonally invariant norm $\|\cdot\|_h$), if we have an approximation algorithm for the problem, then via sketching the running time dependence of the algorithm on $n$ can be improved.

▶ **Definition 35** ((left/right) orthogonal invariance($\mathtt{loi}$/$\mathtt{roi}$))**.** A matrix measure $f()$ is *left orthogonally invariant* (or `loi` for short) if $f(UA) = f(A)$ for all $A$ and orthogonal $U$. Similarly define *right orthogonal invariance* (`roi`). Note that $f()$ is orthogonally invariant if it is both left and right orthogonally invariant.

When norm $\|\cdot\|_g$ is orthogonally invariant, it can be expressed as $\|A\|_g = g(\sigma_1, \sigma_2, \ldots, \sigma_r)$, where the $\sigma_i$ are the singular values of $A$, and $g()$ is a *symmetric gauge function*: a function that is even in each argument, and symmetric, meaning that its value depends only on the set of input values and not their order.

▶ **Definition 36** (padding invariance)**.** Say that a matrix measure $f()$ is *padding invariant* if it is preserved by padding $A$ with rows or columns of zeroes: $f(\left[\begin{smallmatrix} A \\ 0_{z\times d}\end{smallmatrix}\right]) = f\left(\begin{smallmatrix} A & 0_{n\times z'}\end{smallmatrix}\right) = f(A)$.

▶ **Lemma 37.** *Unitarily invariant norms and $v$-norms are padding invariant.*

**Proof.** Omitted in this version. ◀

▶ **Definition 38** (`piloi`, `piroi`)**.** Say that a matrix measure is `piloi` if it is padding invariant and left orthogonally invariant, and `piroi` if it is padding invariant and right orthogonally invariant.

The following is the main theorem of this section.

▶ **Theorem 39.** *Let $f()$ be a real-valued function on matrices that is `piroi` and subadditive. Let $B \in \mathbb{R}^{n\times d'}$. Let*

$$X^* \equiv \operatorname*{argmin}_{X\in\mathbb{R}^{d\times d'}} \|AX - B\|_F^2 + f(X), \tag{22}$$

*and $\Delta_* \equiv \|AX^* - B\|_F^2 + f(X^*)$. Suppose that for $r \equiv \texttt{rank}\, A$, there is an algorithm that for general $n, d, d', r$ and $\varepsilon > 0$, finds $\tilde{X}$ with $\|A\tilde{X} - B\|_F^2 + f(\tilde{X}) \leq (1+\varepsilon)\Delta_*$ in time $\tau(d, n, d', r, \varepsilon)$. Then there is an algorithm that with constant probability finds such a $\tilde{X}$, taking time*

$$O(\texttt{nnz}(A) + \texttt{nnz}(B) + (n + d + d')\text{poly}(r/\varepsilon)) + \tau(d, \text{poly}(r/\varepsilon), \text{poly}(r/\varepsilon), r, \varepsilon).$$

Although earlier results for constrained least squares (e.g. [10]) can be applied to obtain approximation algorithms for regularized multiple-response least squares, via the solution of $\min_{X\in\mathbb{R}^{d\times d'}} \|AX - B\|_F^2$, subject to $f(X) \leq C$ for a chosen constant $C$, such a reduction yields a slower algorithm if properties of $f(X)$ are not exploited, as here.

**Proof.** Omitted in this version. ◀

## D General Regularization: Low-rank Approximation

For an integer $k$ we consider the problem

$$\min_{\substack{Y\in\mathbb{R}^{n\times k}\\ X\in\mathbb{R}^{k\times d}}} \|YX - A\|_F^2 + f(Y, X), \tag{23}$$

where $f(\cdot,\cdot)$ is a real-valued function that is `piloi` in the left argument, `piroi` in the right argument, and left and right reduced by contraction in its left and right arguments, respectively.

For example $\hat{f}(\|Y\|_\ell, \|X\|_r)$ for `piloi` $\|\cdot\|_\ell$ and `piroi` $\|\cdot\|_r$ would satisfy these conditions, as would $\|YX\|_g$ for orthogonally invariant norm $\|\cdot\|_g$. The function $\hat{f}$ could be zero for arguments whose maximum is less than some $\mu$, and infinity otherwise.

## D.1 Via the SVD

First, a solution method relying on the singular value decomposition for a slightly more general problem than (23).

▶ **Theorem 40.** *Let $k$ be a positive integer, $f_1 : \mathbb{R} \mapsto \mathbb{R}$ increasing, and $f : \mathbb{R}^{n \times k} \times \mathbb{R}^{k \times d} \mapsto \mathbb{R}$, where $f$ is* `piloi` *and subadditive in its left argument, and* `piroi` *and subadditive in in its right argument.*

*Let $A$ have full SVD $A = U\Sigma V^\top$, $\Sigma_k \in \mathbb{R}^{k\times k}$ the diagonal matrix of top $k$ singular values of $A$. Let matrices $W^*, Z^* \in \mathbb{R}^{k\times k}$ solve*

$$\min_{\substack{W\in\mathbb{R}^{k\times k}\\ Z\in\mathbb{R}^{k\times k}\\ WZ \text{ diagonal}}} f_1(\|WZ-\Sigma_k\|_{(p)}) + f(W,Z), \tag{24}$$

*and suppose there is a procedure taking $\tau(k)$ time to find $W^*$ and $Z^*$. Then the solution to*

$$\min_{\substack{Y\in\mathbb{R}^{n\times k}\\ X\in\mathbb{R}^{k\times d}}} f_1(\|YX-A\|_{(p)}) + f(Y,X) \tag{25}$$

*is $Y^* = U\begin{bmatrix} W^* \\ 0_{(n-k)\times k}\end{bmatrix}$ and $X^* = \begin{bmatrix} Z^* & 0_{k\times(d-k)}\end{bmatrix} V^\top$. Thus for general $A$, (25) can be solved in time $O(nd\min\{n,d\}) + \tau(k)$.*

**Proof.** Omitted in this version. ◀

We sharpen this result for the case that the regularization term comes from orthogonally invariant norms.

▶ **Theorem 41.** *Consider (25) when $f(\cdot,\cdot)$ has the form $\hat{f}(\|Y\|_\ell, \|X\|_r)$, where $\|\cdot\|_\ell$ and $\|\cdot\|_r$ are orthogonally invariant, and $\hat{f} : \mathbb{R}\times\mathbb{R} \mapsto \mathbb{R}$ increasing in each argument. Suppose in that setting there is a procedure that solves (25) when $A$, $Y$, and $X$ are diagonal matrices, taking time $\tau(r)$ for a function $\tau(\cdot)$, with $r \equiv \texttt{rank}(A)$. Then for general $A$, (25) can be solved by finding the SVD of $A$, and applying the given procedure to $k\times k$ diagonal matrices, taking altogether time $O(nd\min\{n,d\}) + \tau(k)$.*

**Proof.** Omitted in this version. ◀

▶ **Definition 42** (clipping to nonnegative $(\cdot)_+$)**.** For real number $a$, let $(a)_+$ denote $a$, if $a \geq 0$, and zero otherwise. For matrix $A$, let $(A)_+$ denote coordinatewise application.

▶ **Corollary 43.** *If the objective function in (25) is $\|YX-A\|_F^2 + 2\lambda\|YX\|_{(1)}$ or $\|YX-A\|_F^2 + \lambda(\|Y\|_F^2 + \|X\|_F^2)$, then the diagonal matrices $W^*$ and $Z^*$ from Theorem 41 yielding the solution are $W^* = Z^* = \sqrt{(\Sigma_k - \lambda I_k)_+}$, where $\Sigma_k$ is the $k\times k$ diagonal matrix of top $k$ singular values of $A$ [29].*

*If the objective function is $\|YX-A\|_{(p)} + \lambda\|YX\|_{(1)}$ for $p\in[1,\infty]$, then $W^* = Z^* = \sqrt{(\Sigma_k - \alpha I_k)_+}$, for an appropriate value $\alpha$.*

*If the objective function is $\|YX-A\|_F^2 + \lambda\|YX\|_F^2$, then $W^* = Z^* = \sqrt{\Sigma_k/(1+\lambda)}$.*

**Proof.** Omitted in this version. ◀

## D.2 Reduction to a small problem via sketching

▶ **Theorem 44.** *Suppose there is a procedure that solves* (23) *when $A$, $Y$, and $X$ are $k \times k$ matrices, and $A$ is diagonal, and $YX$ is constrained to be diagonal, taking time $\tau(k)$ for a function $\tau(\cdot)$. Let $f$ also inherit a sketching distribution on the left in its left argument, and on the right in its right argument. Then for general $A$, there is an algorithm that finds $\varepsilon$-approximate solution $(\tilde{Y}, \tilde{X})$ in time*

$$O(\mathtt{nnz}(A)) + \tilde{O}(n+d)\mathrm{poly}(k/\varepsilon) + \tau(k).$$

**Proof.** Omitted in this version. ◀

# E Proof of Lemma 25

**Proof.** Let $U_C$ be an orthogonal basis for $\mathtt{colspace}(C)$, so that every matrix of the form $CZ_R$ is equal to $U_C Z'_R$ for some $Z'_R$. Similarly let $U_D^\top$ be an orthogonal basis for $\mathtt{rowspan}(D)$, so that every matrix of the form $Z_S D$ is equal to one of the form $Z'_S U_D$. Let $P_C \equiv U_C U_C^\top$ and $P_D \equiv U_D U_D^\top$. Then using $P_C(I - P_C) = 0$, $P_D(I - P_D) = 0$, and matrix Pythagoras,

$$\begin{aligned}
&\|CZ_R Z_S D - G\|_F^2 + \lambda\|CZ_R\|_F^2 + \lambda\|Z_S D\|_F^2 \\
&\qquad = \|P_C U_C Z'_R Z'_S U_D^\top P_D - G\|_F^2 + \lambda\|U_C Z'_R\|_F^2 + \lambda\|Z'_S U_D^\top\|_F^2 \\
&\qquad = \|P_C U_C Z'_R Z'_S U_D^\top P_D - P_C G P_D\|_F^2 + \|(I - P_C)G\|_F^2 \\
&\qquad\quad + \|P_C G(I - P_D)\|_F^2 + \lambda\|Z'_R\|_F^2 + \lambda\|Z'_S\|_F^2.
\end{aligned}$$

So minimizing (18) is equivalent to minimizing

$$\begin{aligned}
&\|P_C U_C Z'_R Z'_S U_D^\top P_D - P_C G P_D\|_F^2 + \lambda\|Z'_R\|_F^2 + \lambda\|Z'_S\|_F^2 \\
&\qquad = \|U_C Z'_R Z'_S U_D^\top - U_C U_C^\top G U_D U_D^\top\|_F^2 + \lambda\|Z'_R\|_F^2 + \lambda\|Z'_S\|_F^2 \\
&\qquad = \|Z'_R Z'_S - U_C^\top G U_D\|_F^2 + \lambda\|Z'_R\|_F^2 + \lambda\|Z'_S\|_F^2.
\end{aligned}$$

This has the form of (16), mapping $Y$ of (16) to $Z'_R$, $X$ to $Z'_S$, and $A$ to $U_C^\top G U_D$, from which a solution of the form (17) can be obtained.

To recover $Z_R$ from $Z'_R$: we have $C = U_C \left[\, T_C \;\; T'_C \,\right]$, for matrices $T_C$ and $T'_C$, where upper triangular $T_C \in \mathbb{R}^{r_C \times r_C}$. We recover $Z_R$ as $\left[\begin{smallmatrix} T_C^{-1}\hat{Z}'_R \\ 0_{m-r_C \times k} \end{smallmatrix}\right]$, since then $U_C Z'_R = CZ_R$. A similar back-substitution allows recovery of $Z_S$ from $Z'_S$.

Running times: to compute $U_C$ and $U_D$, $O(pm'r_C + mp'r_D)$; to compute $U_C^\top G U_D$, $O(r_D p(p' + r_C))$; to compute and use the SVD of $U_C^\top G U_D$ to to solve (16) via (17), $O(r_C r_D \min\{r_C, r_D\})$; to recover $Z_R$ and $Z_S$, $O(k(r_C^2 + r_D^2))$. Thus, assuming $k \le \min\{p, p'\}$ and using $r_C \le \min\{p, m'\}$ and $r_D \le \min\{m, p'\}$, the total running time is $O(pm'r_C + p'mr_D + pp'(r_C + r_D))$, as claimed. ◀