

A Note on Sparse Least-squares Regression

Christos Boutsidis

Mathematical Sciences Department

IBM T.J. Watson Research Center

cboutsi@us.ibm.com

Malik Magdon-Ismail

Computer Science Department

Rensselaer Polytechnic Institute

magdon@cs.rpi.edu

December 31, 2013

Abstract

We compute a *sparse* solution to the classical least-squares problem $\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2$, where \mathbf{A} is an arbitrary matrix. We describe a novel algorithm for this sparse least-squares problem. The algorithm operates as follows: first, it selects columns from \mathbf{A} , and then solves a least-squares problem only with the selected columns. The column selection algorithm that we use is known to perform well for the well studied column subset selection problem. The contribution of this article is to show that it gives favorable results for sparse least-squares as well. Specifically, we prove that the solution vector obtained by our algorithm is close to the solution vector obtained via what is known as the “SVD-truncated regularization approach”.

1 Introduction

Fix inputs $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$. We study least-squares regression: $\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{Ax} - \mathbf{b}\|_2$. It is well known that the minimum norm solution vector can be found using the pseudo-inverse of \mathbf{A} : $\mathbf{x}^* = \mathbf{A}^\dagger \mathbf{b} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}$. When \mathbf{A} is ill-conditioned, \mathbf{A}^\dagger becomes unstable to perturbations and overfitting can become a serious problem. For example, when the smallest non-zero singular value of \mathbf{A} is close to zero, the largest singular value of \mathbf{A}^\dagger can be extremely large and the solution vector $\mathbf{x}^* = \mathbf{A}^\dagger \mathbf{b}$ obtained via a numerical algorithm is not the optimal, due to numerical instability issues. Practitioners deal with such situations using *regularization*.

Popular regularization techniques are the Lasso [8], the Tikhonov regularization [4], and the truncated SVD [6]. The lasso minimizes $\|\mathbf{Ax} - \mathbf{b}\|_2 + \lambda \|\mathbf{x}\|_1$, and Tikhonov regularization minimizes $\|\mathbf{Ax} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_2^2$ (in both cases $\lambda > 0$ is the regularization parameter). The truncated SVD minimizes $\|\mathbf{A}_k \mathbf{x} - \mathbf{b}\|_2$, where $k < \text{rank}(\mathbf{A})$ is a rank parameter and $\mathbf{A}_k \in \mathbb{R}^{m \times n}$ is the best rank- k

approximation to \mathbf{A} obtained via the SVD. So, the truncated SVD solution is $\mathbf{x}_k^* = \mathbf{A}_k^\dagger \mathbf{b}$. Notice that these regularization methods impose parsimony on \mathbf{x} in different ways. A combinatorial approach to regularization is to explicitly impose the sparsity constraint on \mathbf{x} , requiring it to have few non-zero elements. We give a new deterministic algorithm which, for $r = O(k)$, computes an $\hat{\mathbf{x}}_r \in \mathbb{R}^n$ with at most r non-zero entries such that $\|\mathbf{A}\hat{\mathbf{x}}_r - \mathbf{b}\|_2 \approx \|\mathbf{A}\mathbf{x}_k^* - \mathbf{b}\|_2$.

1.1 Preliminaries

The compact (or thin) Singular Value Decomposition (SVD) of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ of rank ρ is

$$\mathbf{A} = \underbrace{\begin{pmatrix} \mathbf{U}_k & \mathbf{U}_{\rho-k} \end{pmatrix}}_{\mathbf{U}_\mathbf{A} \in \mathbb{R}^{m \times \rho}} \underbrace{\begin{pmatrix} \mathbf{\Sigma}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{\rho-k} \end{pmatrix}}_{\mathbf{\Sigma}_\mathbf{A} \in \mathbb{R}^{\rho \times \rho}} \underbrace{\begin{pmatrix} \mathbf{V}_k^\top \\ \mathbf{V}_{\rho-k}^\top \end{pmatrix}}_{\mathbf{V}_\mathbf{A}^\top \in \mathbb{R}^{\rho \times n}},$$

Here, $\mathbf{U}_k \in \mathbb{R}^{m \times k}$ and $\mathbf{U}_{\rho-k} \in \mathbb{R}^{m \times (\rho-k)}$ contain the left singular vectors of \mathbf{A} . Similarly, $\mathbf{V}_k \in \mathbb{R}^{n \times k}$ and $\mathbf{V}_{\rho-k} \in \mathbb{R}^{n \times (\rho-k)}$ contain the right singular vectors. The singular values of \mathbf{A} , which we denote as $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \dots \geq \sigma_\rho(\mathbf{A}) > 0$ are contained in $\mathbf{\Sigma}_k \in \mathbb{R}^{k \times k}$ and $\mathbf{\Sigma}_{\rho-k} \in \mathbb{R}^{(\rho-k) \times (\rho-k)}$. We use $\mathbf{A}^\dagger = \mathbf{V}_\mathbf{A} \mathbf{\Sigma}_\mathbf{A}^{-1} \mathbf{U}_\mathbf{A}^\top \in \mathbb{R}^{n \times m}$ to denote the Moore-Penrose pseudo-inverse of \mathbf{A} with $\mathbf{\Sigma}_\mathbf{A}^{-1}$ denoting the inverse of $\mathbf{\Sigma}_\mathbf{A}$. Let $\mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^\top \in \mathbb{R}^{m \times n}$ and $\mathbf{A}_{\rho-k} = \mathbf{A} - \mathbf{A}_k = \mathbf{U}_{\rho-k} \mathbf{\Sigma}_{\rho-k} \mathbf{V}_{\rho-k}^\top \in \mathbb{R}^{m \times n}$. For $k < \text{rank}(\mathbf{A})$, the SVD gives the best rank k approximation to \mathbf{A} in both the spectral and the

- 1: **Input:** $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, target rank $k < \text{rank}(\mathbf{A})$, and parameter $0 < \varepsilon < 1/2$.
- 2: Obtain $\mathbf{V}_k \in \mathbb{R}^{n \times k}$ from the SVD of \mathbf{A} and compute $\mathbf{E} = \mathbf{A} - \mathbf{A}\mathbf{V}_k\mathbf{V}_k^\top \in \mathbb{R}^{m \times n}$.
- 3: Set $\mathbf{C} = \mathbf{A}\mathbf{\Omega}\mathbf{S} \in \mathbb{R}^{m \times r}$, with $r = \lceil \frac{9k}{\varepsilon^2} \rceil$ and
 $[\mathbf{\Omega}, \mathbf{S}] = \text{DeterministicSampling}(\mathbf{V}_k^\top, \mathbf{E}, r)$,
- 4: Set $\mathbf{x}_r = \mathbf{C}^\dagger \mathbf{b} \in \mathbb{R}^r$, and $\hat{\mathbf{x}}_r = \mathbf{\Omega}\mathbf{S}\mathbf{x}_r \in \mathbb{R}^n$ ($\hat{\mathbf{x}}_r$ has at most r non-zeros at the indices of the selected columns in \mathbf{C}).
- 5: **Return** $\hat{\mathbf{x}}_r \in \mathbb{R}^n$.

Algorithm 1: Deterministic Sparse Regression

Frobenius norm: for $\tilde{\mathbf{A}} \in \mathbb{R}^{m \times n}$, let $\text{rank}(\tilde{\mathbf{A}}) \leq k$; then, for $\xi = 2, \text{F}$, $\|\mathbf{A} - \mathbf{A}_k\|_\xi \leq \|\mathbf{A} - \tilde{\mathbf{A}}\|_\xi$. Also, $\|\mathbf{A} - \mathbf{A}_k\|_2 = \|\mathbf{\Sigma}_{\rho-k}\|_2 = \sigma_{k+1}(\mathbf{A})$, and $\|\mathbf{A} - \mathbf{A}_k\|_\text{F}^2 = \|\mathbf{\Sigma}_{\rho-k}\|_\text{F}^2 = \sum_{i=k+1}^\rho \sigma_i^2(\mathbf{A})$. The Frobenius and the spectral norm of \mathbf{A} are defined as: $\|\mathbf{A}\|_\text{F}^2 = \sum_{i,j} \mathbf{A}_{ij}^2 = \sum_{i=1}^\rho \sigma_i^2(\mathbf{A})$; and $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$. Let \mathbf{X} and \mathbf{Y} be matrices of appropriate dimensions; then, $\|\mathbf{X}\mathbf{Y}\|_\text{F} \leq \min\{\|\mathbf{X}\|_\text{F}\|\mathbf{Y}\|_2, \|\mathbf{X}\|_2\|\mathbf{Y}\|_\text{F}\}$. This is a stronger version of the standard submultiplicativity property $\|\mathbf{X}\mathbf{Y}\|_\text{F} \leq \|\mathbf{X}\|_\text{F}\|\mathbf{Y}\|_\text{F}$, which we will refer to as “spectral submultiplicativity”.

- 1: **Input:** $\mathbf{V}^\top = [\mathbf{v}_1, \dots, \mathbf{v}_n] \in \mathbb{R}^{k \times n}$; $\mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_n] \in \mathbb{R}^{m \times n}$; and $r > k$.
- 2: **Output:** Sampling and rescaling matrices $\mathbf{\Omega} \in \mathbb{R}^{n \times r}$, $\mathbf{S} \in \mathbb{R}^{r \times r}$.
- 3: Initialize $\mathbf{B}_0 = \mathbf{0}_{k \times k}$, $\mathbf{\Omega} = \mathbf{0}_{n \times r}$, $\mathbf{S} = \mathbf{0}_{r \times r}$.
- 4: **for** $\tau = 0$ **to** $r - 1$ **do**
- 5: Set $L_\tau = \tau - \sqrt{rk}$.
- 6: Pick index $i \in \{1, 2, \dots, n\}$ and t such that $U(\mathbf{e}_i) \leq \frac{1}{t} \leq L(\mathbf{v}_i, \mathbf{B}_\tau, L_\tau)$.
- 7: Update $\mathbf{B}_{\tau+1} = \mathbf{B}_\tau + t\mathbf{v}_i\mathbf{v}_i^\top$. Set $\mathbf{\Omega}_{i,\tau+1} = 1$ and $\mathbf{S}_{\tau+1,\tau+1} = 1/\sqrt{t}$.
- 8: **end for**
- 9: **Return:** $\mathbf{\Omega} \in \mathbb{R}^{n \times r}$, $\mathbf{S} \in \mathbb{R}^{r \times r}$.

Algorithm 2: DeterministicSampling (from [1])

Given $k < \rho = \text{rank}(\mathbf{A})$, the truncated rank- k SVD regularized weights are

$$\mathbf{x}_k^* = \mathbf{A}_k^\dagger \mathbf{b} = \mathbf{V}_k \mathbf{\Sigma}_k^{-1} \mathbf{U}_k^\top \mathbf{b} \in \mathbb{R}^n,$$

and note that $\|\mathbf{b} - \mathbf{A}_k \mathbf{A}_k^\dagger \mathbf{b}\|_2 = \|\mathbf{b} - \mathbf{U}_k \mathbf{U}_k^\top \mathbf{b}\|_2$

Finally, for $r < n$, let $\mathbf{\Omega} = [\mathbf{z}_{i_1}, \dots, \mathbf{z}_{i_r}] \in \mathbb{R}^{n \times r}$ where $\mathbf{z}_i \in \mathbb{R}^n$ are standard basis vectors; $\mathbf{\Omega}$ is a *sampling matrix* because $\mathbf{A}\mathbf{\Omega} \in \mathbb{R}^{m \times r}$ is a matrix whose columns are sampled (with possible repetition) from the columns of \mathbf{A} . Let $\mathbf{S} \in \mathbb{R}^{r \times r}$ be a diagonal *rescaling matrix* with positive entries; then, we define the sampled and rescaled columns from \mathbf{A} by $\mathbf{C} = \mathbf{A}\mathbf{\Omega}\mathbf{S}$: $\mathbf{\Omega}$ samples some columns from \mathbf{A} and then \mathbf{S} rescales them.

2 Results

Our sparse solver to minimize $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ takes as input the sparsity parameter r (i.e., the solution vector \mathbf{x} is allowed at most r non-zero entries), and selects r rescaled columns from \mathbf{A} (denoted by \mathbf{C}). We then solve the least-squares problem to minimize $\|\mathbf{C}\mathbf{x} - \mathbf{b}\|_2$. The result is a dense vector $\mathbf{C}^\dagger \mathbf{b}$ with r dimensions. The sparse solution $\hat{\mathbf{x}}_r$ will be zero at indices corresponding to columns not selected in \mathbf{C} , and we use $\mathbf{C}^\dagger \mathbf{b}$ to compute the other entries of $\hat{\mathbf{x}}_r$.

Theorem 1. Let $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, rank $k < \text{rank}(\mathbf{A})$, and $0 < \varepsilon < 1/2$. Algorithm 1 runs in time $O(mn \min\{m, n\} + nk^3/\varepsilon^2)$ and returns $\hat{\mathbf{x}}_r \in \mathbb{R}^n$ with at most $r = \lceil 9k/\varepsilon^2 \rceil$ non-zero entries such that:

$$\|\mathbf{A}\hat{\mathbf{x}}_r - \mathbf{b}\|_2 \leq \|\mathbf{A}\mathbf{x}_k^* - \mathbf{b}\|_2 + (1 + \varepsilon) \cdot \|\mathbf{b}\|_2 \cdot \frac{\|\mathbf{A} - \mathbf{A}_k\|_F}{\sigma_k(\mathbf{A})}.$$

This upper bound is “small” when \mathbf{A} is “effectively” low-rank, i.e., $\|\mathbf{A} - \mathbf{A}_k\|_F/\sigma_k(\mathbf{A}) \ll 1$. Also,

a trivial bound is $\|\mathbf{A}\hat{\mathbf{x}}_r - \mathbf{b}\|_2 \leq \|\mathbf{b}\|_2$ (error when $\hat{\mathbf{x}}_r$ is the all-zeros vector), because $\|\mathbf{C}\mathbf{C}^\dagger\mathbf{b} - \mathbf{b}\|_2 \leq \|\mathbf{C}\mathbf{0}_{r \times 1} - \mathbf{b}\|_2 = \|\mathbf{b}\|_2$.

In the heart of Algorithm 1 lies a method for selecting columns from \mathbf{A} (Algorithm 2), which was originally developed in [1] for column subset selection, where one selects columns \mathbf{C} from \mathbf{A} to minimize $\|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger\mathbf{A}\|_F$. Here, we adopt the same algorithm for least-squares.

The main tool used to prove Theorem 1 is a new “structural” result that may be of independent interest.

Lemma 2. *Fix $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^n$, rank $k < \text{rank}(\mathbf{A})$, and sparsity $r > k$. Let $\mathbf{x}_k^* = \mathbf{A}_k^\dagger \mathbf{b} \in \mathbb{R}^n$, where $\mathbf{A}_k \in \mathbb{R}^{m \times n}$ is the rank- k SVD approximation to \mathbf{A} . Let $\mathbf{\Omega} \in \mathbb{R}^{n \times r}$ and $\mathbf{S} \in \mathbb{R}^{r \times r}$ be any sampling and rescaling matrices with $\text{rank}(\mathbf{V}_k^T \mathbf{\Omega} \mathbf{S}) = k$. Let $\mathbf{C} = \mathbf{A} \mathbf{\Omega} \mathbf{S} \in \mathbb{R}^{m \times r}$ be a matrix of sampled rescaled columns of \mathbf{A} and let $\hat{\mathbf{x}}_r = \mathbf{\Omega} \mathbf{S} \mathbf{C}^\dagger \mathbf{b} \in \mathbb{R}^n$ (having at most r non-zeros). Then,*

$$\|\mathbf{A}\hat{\mathbf{x}}_r - \mathbf{b}\|_2 \leq \|\mathbf{A}\mathbf{x}_k^* - \mathbf{b}\|_2 + \|(\mathbf{A} - \mathbf{A}_k) \mathbf{\Omega} \mathbf{S} (\mathbf{V}_k^T \mathbf{\Omega} \mathbf{S})^\dagger \mathbf{\Sigma}_k \mathbf{U}_k^T \mathbf{b}\|_2.$$

The lemma says that if the sampling matrix satisfies a simple rank condition, then solving the regression on the sampled columns gives a sparse solution to the original problem with a performance guarantee.

2.1 Algorithm Description

Algorithm 1 selects r columns from \mathbf{A} to form \mathbf{C} and the corresponding sparse vector $\hat{\mathbf{x}}_r$. The core of Algorithm 1 is the subroutine `DeterministicSampling`, which is a method to simultaneously sample the columns of two matrices, while controlling their spectral and Frobenius norms. `DeterministicSampling` takes inputs $\mathbf{V}^T \in \mathbb{R}^{k \times n}$ and $\mathbf{E} \in \mathbb{R}^{m \times n}$; the matrix \mathbf{V} is orthonormal, $\mathbf{V}^T \mathbf{V} = \mathbf{I}_k$. (In our application, $\mathbf{V}^T = \mathbf{V}_k^T$ and $\mathbf{E} = \mathbf{A} - \mathbf{A}_k$.) We view \mathbf{V}^T and \mathbf{E} as two sets of n column vectors, $\mathbf{V}^T = [\mathbf{v}_1, \dots, \mathbf{v}_n]$, and $\mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_n]$.

Given k and r and the iterator $\tau = 0, 1, 2, \dots, r-1$, define $L_\tau = \tau - \sqrt{rk}$. For a symmetric matrix $\mathbf{B} \in \mathbb{R}^{k \times k}$ with eigenvalues $\lambda_1, \dots, \lambda_k$ and $L \in \mathbb{R}$, define functions $\phi(L, \mathbf{B}) = \sum_{i=1}^k \frac{1}{\lambda_i - L}$, and $L(\mathbf{v}, \mathbf{B}, L) = \frac{\mathbf{v}^T (\mathbf{B} - L' \mathbf{I}_k)^{-2} \mathbf{v}}{\phi(L', \mathbf{B}) - \phi(L, \mathbf{B})} - \mathbf{v}^T (\mathbf{B} - L' \mathbf{I}_k)^{-1} \mathbf{v}$, where $L' = L + 1$. Also, for a column \mathbf{e} , define $U(\mathbf{e}) = \frac{\mathbf{e}^T \mathbf{e}}{\|\mathbf{A}\|_F^2} \left(1 - \sqrt{k/r}\right)$. At step τ , the algorithm selects any column i for which $U(\mathbf{e}_i) \leq L(\mathbf{v}_i, \mathbf{B}, L_\tau)$ and computes a weight t such that $U(\mathbf{e}_i) \leq t^{-1} \leq L(\mathbf{v}_i, \mathbf{B}, L_\tau)$; Any t^{-1} in the interval is acceptable. (There is always at least one such index i (see Lemma 8.1 in [1]).)

The running time is dominated by the search for a column which satisfies $U \leq L$. To compute L , one needs $\phi(L, \mathbf{B})$, and hence the eigenvalues of \mathbf{B} , and $(\mathbf{B} - L' \mathbf{I}_k)^{-1}$. This takes $O(k^3)$ time once

per iteration, for a total of $O(rk^3)$. Then, for $i = 1, \dots, n$, we need to compute L for every \mathbf{v}_i . This takes $O(nk^2)$ per iteration, for a total of $O(nrk^2)$. To compute U , we need $\mathbf{e}_i^T \mathbf{e}_i$ for $i = 1, \dots, n$ which takes $O(mn)$. So, in total, `DeterministicSampling` takes $O(nrk^2 + mn)$ time, hence Algorithm 1 needs $O(mn \min\{m, n\} + nk^3/\varepsilon^2)$ time.

`DeterministicSampling` uses a greedy procedure to sample columns of \mathbf{V}_k^T that satisfy the next Lemma.

Lemma 3 ([1]). *On $\mathbf{V}^T \in \mathbb{R}^{k \times n}$, $\mathbf{E} \in \mathbb{R}^{m \times n}$, and $r > k$ `DeterministicSampling` returns $\mathbf{\Omega}, \mathbf{S}$ satisfying*

$$\sigma_k(\mathbf{V}^T \mathbf{\Omega} \mathbf{S}) \geq 1 - \sqrt{k/r}, \quad \|\mathbf{E} \mathbf{\Omega} \mathbf{S}\|_F \leq \|\mathbf{E}\|_F.$$

By Lemma 3, Algorithm 1 returns $\mathbf{\Omega}, \mathbf{S}$ that satisfy the rank condition in Lemma 2, so the structural bound applies. Lemma 3 also bounds two key terms in the bound which ultimately allow us to prove Theorem 1.

2.2 Proofs

Proof of Theorem 1 By Lemma 3, $\text{rank}(\mathbf{V}_k^T \mathbf{\Omega} \mathbf{S}) = k$ so the bound in Lemma 2 holds. Recall $\mathbf{E} = \mathbf{A} - \mathbf{A}_k$. By submultiplicativity, $\|\mathbf{E} \mathbf{\Omega} \mathbf{S} (\mathbf{V}_k^T \mathbf{\Omega} \mathbf{S})^\dagger \mathbf{\Sigma}_k^{-1} \mathbf{U}_k^T \mathbf{b}\|_2$ is at most

$$\|\mathbf{E} \mathbf{\Omega} \mathbf{S}\|_2 \|(\mathbf{V}_k^T \mathbf{\Omega} \mathbf{S})^\dagger\|_2 \|\mathbf{\Sigma}_k^{-1} \mathbf{U}_k^T \mathbf{b}\|_2.$$

We now bound each term to obtain Theorem 1:

$$\|\mathbf{E} \mathbf{\Omega} \mathbf{S}\|_2 \leq \|\mathbf{E} \mathbf{\Omega} \mathbf{S}\|_F \leq \|\mathbf{E}\|_F = \|\mathbf{A} - \mathbf{A}_k\|_F \tag{a}$$

$$\|(\mathbf{V}_k^T \mathbf{\Omega} \mathbf{S})^\dagger\|_2 = \frac{1}{\sigma_k(\mathbf{V}_k^T \mathbf{\Omega} \mathbf{S})} \leq \frac{1}{1 - \sqrt{k/r}} \leq 1 + \varepsilon \tag{b}$$

$$\|\mathbf{\Sigma}_k^{-1} \mathbf{U}_k^T \mathbf{b}\|_2 \leq \|\mathbf{\Sigma}_k^{-1}\|_2 \|\mathbf{U}_k^T\|_2 \|\mathbf{b}\|_2 = \|\mathbf{b}\|_2 / \sigma_k(\mathbf{A}) \tag{c}$$

(a) follows from Lemma 3; (b) also follows from Lemma 3 using $r = \lceil 9k/\varepsilon^2 \rceil$ and $\varepsilon < 1/2$; (c) follows from submultiplicativity. ■

Proof of Lemma 2 We will prove a more general result, and Lemma 2 will be a simple corollary. We first introduce a general matrix approximation problem and present an algorithm for this problem (Lemma 4). Lemma 2 is a corollary of Lemma 4.

Let $\mathbf{B} \in \mathbb{R}^{m \times \omega}$ be a matrix which we would like to approximate; let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be the matrix

which we will use to approximate \mathbf{B} . Specifically, we want a *sparse* approximation of \mathbf{B} from \mathbf{A} , which means that we would like to choose $\mathbf{C} \in \mathbb{R}^{m \times r}$ consisting of $r < n$ columns from \mathbf{A} such that $\|\mathbf{B} - \mathbf{C}\mathbf{C}^\dagger\mathbf{B}\|_F$ is small. If $\mathbf{A} = \mathbf{B}$, then, this is the column based matrix approximation problem, which has received much interest recently [2, 1]. The more general problem which we study here, with $\mathbf{A} \neq \mathbf{B}$, takes on a surprisingly more difficult flavor. Our motivation is regression, but the problem could be of more general interest. We will approach the problem through the use of matrix factorizations. For $\mathbf{Z} \in \mathbb{R}^{n \times k}$, with $\mathbf{Z}^\top \mathbf{Z} = \mathbf{I}_k$, let $\mathbf{A} = \mathbf{H}\mathbf{Z}^\top + \mathbf{E}$, where $\mathbf{H} \in \mathbb{R}^{m \times k}$; and, $\mathbf{E} \in \mathbb{R}^{m \times n}$ is the residual error. For fixed \mathbf{A} and \mathbf{Z} , $\|\mathbf{E}\|_\xi$ ($\xi = 2, F$) is minimized when $\mathbf{H} = \mathbf{A}\mathbf{Z}$. Let $\mathbf{\Omega} \in \mathbb{R}^{n \times r}$, $\mathbf{S} \in \mathbb{R}^{r \times r}$, and $\mathbf{C} = \mathbf{A}\mathbf{\Omega}\mathbf{S} \in \mathbb{R}^{m \times r}$.

Lemma 4. *If $\text{rank}(\mathbf{Z}^\top \mathbf{\Omega}\mathbf{S}) = k$, then,*

$$\|\mathbf{B} - \mathbf{C}\mathbf{C}^\dagger\mathbf{B}\|_\xi \leq \|\mathbf{B} - \mathbf{H}\mathbf{H}^\dagger\mathbf{B}\|_\xi + \|\mathbf{E}\mathbf{\Omega}(\mathbf{Z}^\top \mathbf{\Omega})^\dagger \mathbf{H}^\dagger \mathbf{B}\|_\xi.$$

Proof. $\|\mathbf{B} - \mathbf{C}\mathbf{C}^\dagger\mathbf{B}\|_\xi$

$$\leq \|\mathbf{B} - \mathbf{C}(\mathbf{Z}^\top \mathbf{\Omega}\mathbf{S})^\dagger \mathbf{H}^\dagger \mathbf{B}\|_\xi \tag{a}$$

$$= \|\mathbf{B} - \mathbf{A}\mathbf{\Omega}\mathbf{S}(\mathbf{Z}^\top \mathbf{\Omega}\mathbf{S})^\dagger \mathbf{H}^\dagger \mathbf{B}\|_\xi$$

$$= \|\mathbf{B} - (\mathbf{H}\mathbf{Z}^\top + \mathbf{E})\mathbf{\Omega}\mathbf{S}(\mathbf{Z}^\top \mathbf{\Omega}\mathbf{S})^\dagger \mathbf{H}^\dagger \mathbf{B}\|_\xi$$

$$= \|\mathbf{B} - \mathbf{H}(\mathbf{Z}^\top \mathbf{\Omega}\mathbf{S})(\mathbf{Z}^\top \mathbf{\Omega}\mathbf{S})^\dagger \mathbf{H}^\dagger \mathbf{B} + \mathbf{E}\mathbf{\Omega}(\mathbf{Z}^\top \mathbf{\Omega}\mathbf{S})^\dagger \mathbf{H}^\dagger \mathbf{B}\|_\xi$$

$$= \|\mathbf{B} - \mathbf{H}\mathbf{H}^\dagger \mathbf{B} + \mathbf{E}\mathbf{\Omega}\mathbf{S}(\mathbf{Z}^\top \mathbf{\Omega}\mathbf{S})^\dagger \mathbf{H}^\dagger \mathbf{B}\|_\xi \tag{b}$$

$$\leq \|\mathbf{B} - \mathbf{H}\mathbf{H}^\dagger \mathbf{B}\|_\xi + \|\mathbf{E}\mathbf{\Omega}\mathbf{S}(\mathbf{Z}^\top \mathbf{\Omega}\mathbf{S})^\dagger \mathbf{H}^\dagger \mathbf{B}\|_\xi. \tag{c}$$

(a) follows by the optimality of $\mathbf{C}^\dagger \mathbf{B}$; (b) follows because $\text{rank}(\mathbf{Z}^\top \mathbf{\Omega}\mathbf{S}) = k$ and so $\mathbf{Z}^\top \mathbf{\Omega}\mathbf{S}(\mathbf{Z}^\top \mathbf{\Omega}\mathbf{S})^\dagger = \mathbf{I}_k$; (c) follows by the triangle inequality of matrix norms. \blacksquare

Lemma 4 is a general tool for the general matrix approximation problem. The bound has two terms which highlight some trade offs: the first term is the approximation of \mathbf{B} using \mathbf{H} (\mathbf{H} is used in the factorization to approximate \mathbf{A}); the second term is related to \mathbf{E} , the residual error in approximating \mathbf{A} . Ideally, one should choose \mathbf{H} and \mathbf{Z} to simultaneously approximate \mathbf{B} with \mathbf{H} and have small residual error \mathbf{E} . In general, these are two competing goals, and a balance should be struck. Here, we focus on the Frobenius norm, and will consider only one extreme of this trade off, namely choosing the factorization to minimize $\|\mathbf{E}\|_F$. Specifically, since \mathbf{Z} has rank k , the best choice for $\mathbf{H}\mathbf{Z}^\top$ which minimizes $\|\mathbf{E}\|_F$ is \mathbf{A}_k . In this case, $\mathbf{E} = \mathbf{A} - \mathbf{A}_k$. Via the SVD, $\mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^\top$,

and so $\mathbf{A} = (\mathbf{U}_k \boldsymbol{\Sigma}_k) \mathbf{V}_k^T + \mathbf{A} - \mathbf{A}_k$. We apply Lemma 4, with $\mathbf{B} = \mathbf{b}$, $\mathbf{H} = \mathbf{U}_k \boldsymbol{\Sigma}_k$, $\mathbf{Z} = \mathbf{V}_k$ and $\mathbf{E} = \mathbf{A} - \mathbf{A}_k$, obtaining the next corollary.

Corollary 5. *If $\text{rank}(\mathbf{V}_k^T \boldsymbol{\Omega} \mathbf{S}) = k$, then,*

$$\|\mathbf{b} - \mathbf{C}\mathbf{C}^\dagger \mathbf{b}\|_2 \leq \|\mathbf{b} - \mathbf{U}_k \mathbf{U}_k^T \mathbf{b}\|_2 + \|\mathbf{E} \boldsymbol{\Omega} \mathbf{S} (\mathbf{V}_k^T \boldsymbol{\Omega} \mathbf{S})^\dagger \boldsymbol{\Sigma}_k^{-1} \mathbf{U}_k^T \mathbf{b}\|_2.$$

Setting $\|\mathbf{b} - \mathbf{C}\mathbf{C}^\dagger \mathbf{b}\|_2 = \|\mathbf{b} - \mathbf{A} \hat{\mathbf{x}}_r\|_2$, with $\hat{\mathbf{x}}_r = \boldsymbol{\Omega} \mathbf{S} \mathbf{C}^\dagger \mathbf{b}$ and $\|\mathbf{b} - \mathbf{U}_k \mathbf{U}_k^T \mathbf{b}\|_2 = \|\mathbf{b} - \mathbf{A} \mathbf{x}_k^*\|_2$, we get Lemma 2.

3 Related work

A bound can be obtained using the Rank-Revealing QR (RRQR) factorization [3] which only applies to $r = k$: a QR-like decomposition is used to select exactly k columns of \mathbf{A} to obtain a sparse solution $\hat{\mathbf{x}}_k$. Combining Eqn. (12) of [3] with Strong RRQR [5] one gets a bound $\|\mathbf{x}_k^* - \hat{\mathbf{x}}_k\|_2 \leq \sqrt{4k(n-k)+1}/\sigma_k(\mathbf{A}) \cdot (2\|\mathbf{b}\|_2 + \|\mathbf{b} - \mathbf{A} \mathbf{x}_k^*\|_2)$. We compare $\|\mathbf{A} \hat{\mathbf{x}}_r - \mathbf{b}\|_2$ and $\|\mathbf{A} \mathbf{x}_k^* - \mathbf{b}\|_2$ and our bound is generally stronger and applies to any user specified $r > k$.

Sparse Approximation Literature The problem studied in this paper is NP-hard [7]. Sparse approximation has important applications and many approximation algorithms have been proposed. The proposed algorithms are typically either greedy or are based on convex optimization relaxations of the objective. We refer the reader to [9, 10, 11] and references therein for more details. In general, these results try to reconstruct \mathbf{b} to within an error using the sparsest possible solution \mathbf{x} . In our setting, we fix the sparsity r as a constraint and compare our solution $\hat{\mathbf{x}}_r$ with the benchmark \mathbf{x}_k^* .

4 Numerical illustration

We implemented our algorithm in Matlab and tested it on a sparse approximation problem $\|\mathbf{A} \mathbf{x} - \mathbf{b}\|_2$, where \mathbf{A} and \mathbf{b} are $m \times n$ and $m \times 1$, respectively, with $m = 2000$ and $n = 1000$. Each element of \mathbf{A} and \mathbf{b} are i.i.d. Gaussian random variables with zero mean and unit variance. We chose $k = 20$ and experimented with different values of $r = 20, 30, 40, \dots, 200$. Figure 1 shows the additive error $\|\mathbf{A} \hat{\mathbf{x}}_r - \mathbf{b}\|_2 - \|\mathbf{A} \mathbf{x}_k^* - \mathbf{b}\|_2$. This experiment illustrates that the proposed algorithm computes a sparse solution vector with small approximation error. In this case, $\|\mathbf{b}\|_2 \approx 25$ and $\|\mathbf{A} - \mathbf{A}_k\|_F / \sigma_k(\mathbf{A}) \approx 18$, so the algorithm performs empirically better than what the worst-case bound of our main theorem predicts.

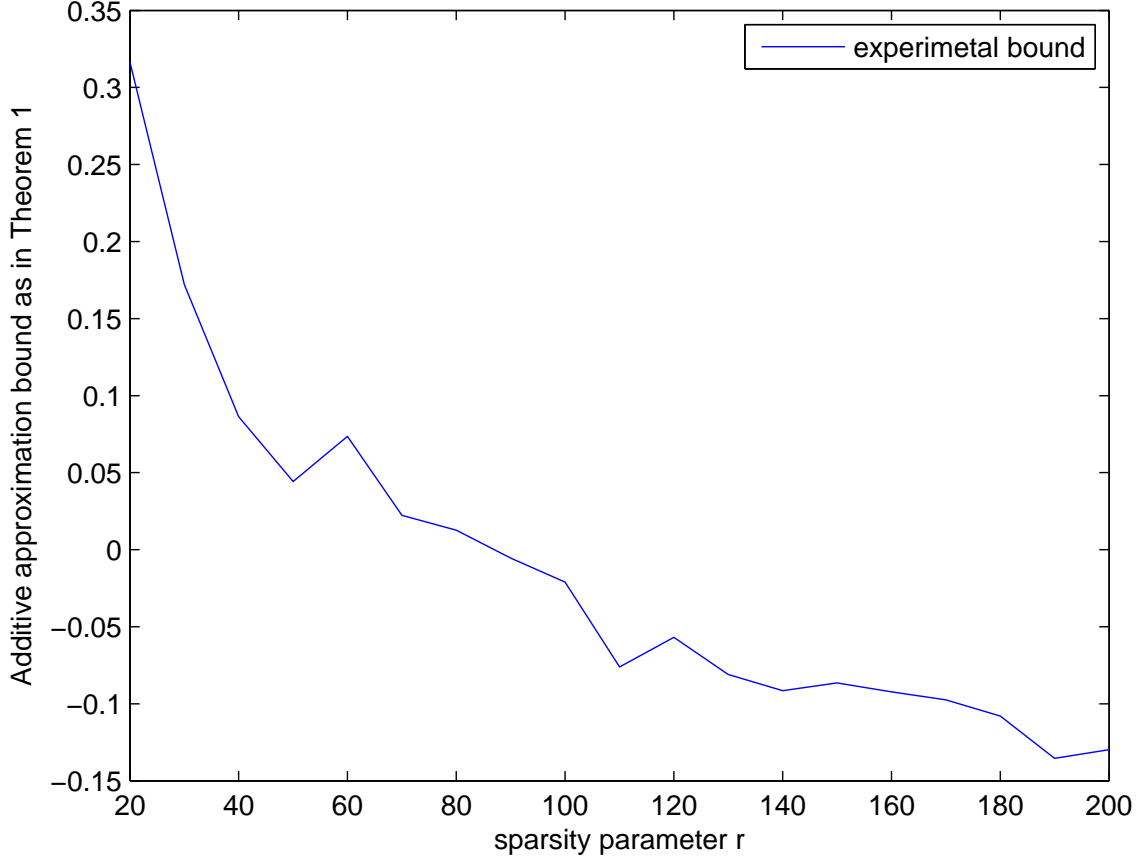


Figure 1: Residual error on a problem with a 2000×1000 matrix \mathbf{A} . The non-monotonic decrease arises because the algorithm chooses columns given r , which means that the columns chosen for a small r are not necessarily a subset of the columns chosen for a larger r .

5 Concluding Remarks

We observe that our bound involves $\|\mathbf{A} - \mathbf{A}_k\|_F$. This can be converted to a bound in terms of $\|\mathbf{A} - \mathbf{A}_k\|_2$ using $\|\mathbf{A} - \mathbf{A}_k\|_F \leq \sqrt{n-k} \cdot \|\mathbf{A} - \mathbf{A}_k\|_2$. The better bound $\|\mathbf{A} - \mathbf{A}_k\|_F \leq O(1 + \varepsilon\sqrt{n/k}) \cdot \|\mathbf{A} - \mathbf{A}_k\|_2$ can be obtained by using a more expensive variant of Deterministic sampling in [1] that bounds the spectral norm of the sampled \mathbf{E} : $\|\mathbf{E}\Omega\mathbf{S}\|_2 \leq (1 + \sqrt{n/r})\|\mathbf{E}\|_2$.

Sparsity in our algorithm is enforced in an unsupervised way: the columns \mathbf{C} are selected obliviously to \mathbf{b} . An interesting open question is whether the use of different factorizations in Lemma 5, together with choosing the columns \mathbf{C} in a \mathbf{b} -dependent way can give an error bound in terms of the optimal error $\|\mathbf{b} - \mathbf{A}\mathbf{A}^\dagger\mathbf{b}\|_2$?

Acknowledgements

Christos Boutsidis acknowledges the support from XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323. Malik Magdon-Ismael was partially supported by the Army Research Laboratory's NS-CTA program under Cooperative Agreement Number W911NF-09-2-0053 and an NSF CDI grant NSF-IIS 1124827.

References

- [1] C. Boutsidis, P. Drineas, and M. Magdon-Ismael. Near-optimal column based matrix reconstruction. In *Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2011.
- [2] C. Boutsidis, M. W. Mahoney, and P. Drineas. An improved approximation algorithm for the column subset selection problem. In *SODA*, 2009.
- [3] T. F. Chan and P. C. Hansen. Some applications of the rank revealing QR factorization. *SIAM Journal on Scientific and Statistical Computing*, 13:727–741, 1992.
- [4] G.H. Golub, P.C. Hansen, and D. O’Leary. Tikhonov regularization and total least squares. *SIAM Journal on Matrix Analysis and Applications*, 21(1):185–194, 2000.
- [5] M. Gu and S. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, 17:848–869, 1996.
- [6] P.C. Hansen. The truncated svd as a method for regularization. *BIT Numerical Mathematics*, 27(4):534–553, 1987.
- [7] B. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227234, 1995.
- [8] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, pages 267–288, 1996.
- [9] J. Tropp. Greed is good: Algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on* 50:10 (2004): 2231-2242.
- [10] J. Tropp, A. Gilbert, and M. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing* 86.3 (2006): 572-588.

- [11] J. Tropp. Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. *Signal Processing* 86.3 (2006): 589-602.