Sampling Algorithms for ℓ_2 Regression and Applications

Petros Drineas * Michael W. Mahoney † S. Muthukrishnan ‡

Abstract

We present and analyze a sampling algorithm for the basic linear-algebraic problem of ℓ_2 regression. The ℓ_2 regression (or least-squares fit) problem takes as input a matrix $A \in \mathbb{R}^{n\times d}$ (where we assume $n \gg d$) and a target vector $b \in \mathbb{R}^n$, and it returns as output $\mathcal{Z} = \min_{x\in\mathbb{R}^d} |b - Ax|_2$. Also of interest is $x_{opt} = A^+ b$, where A^+ is the Moore-Penrose generalized inverse, which is the minimum-length vector achieving the minimum. Our algorithm randomly samples r rows from the matrix A and vector b to construct an induced ℓ_2 regression problem with many fewer rows, but with the same number of columns. A crucial feature of the algorithm is the nonuniform sampling probabilities. These probabilities depend in a sophisticated manner on the lengths, i.e., the Euclidean norms, of the rows of the left singular vectors of A and the manner in which b lies in the complement of the column space of A. Under appropriate assumptions, we show relative error approximations for both $\mathcal{Z}$ and x_{opt}. Applications of this sampling methodology are briefly discussed.

1 Introduction

One of the common paradigms in computing with large data sets is the use of "sampling." In this approach, one uses only a small portion of the data, and one performs computations of interest for the full dataset by using that small portion as a surrogate. For many problems it is provably impossible to compute accurately the answer without touching each of the input elements at least once [13]. Thus, sampling methods typically produce an approximation to the quantity of interest. A question arises: for what problems do sampling methods provide accurate estimates?

In this paper, we study sampling algorithms for the basic linear-algebraic problem of ℓ_2 regression. This is one of the most fundamental regression problems, and it has found many applications in mathematics and statistical data analysis. Recall the ℓ_2 regression (or least-squares fit) problem: given as input a matrix $A \in \mathbb{R}^{n\times d}$ and a target vector $b \in \mathbb{R}^n$, compute

$$\mathcal{Z} = \min_{x\in\mathbb{R}^d} |b - Ax|_2 . \tag{1.1}$$

Also of interest is the computation of vectors that achieve the minimum $\mathcal{Z}$. If $n > d$ there are more constraints than variables and the problem is an *overconstrained* least-squares fit problem; in this case, there does not in general exist a vector x such that $Ax = b$. It is well-known that the minimum-length vector among those minimizing $|b - Ax|_2$ is

$$x_{opt} = A^+ b, \tag{1.2}$$

where A^+ denotes the Moore-Penrose generalized inverse of the matrix A. The classical algorithm for solving the ℓ_2 regression problem takes time $O(nd^2)$, assuming, as we will, that $n \gg d$ [10].

It is well-known that certain linear-algebraic problems depend quite sensitively on perturbations of individual matrix entries, while other problems are much more robust to perturbations [17]. Thus, a fundamental algorithmic question is: can the input to a linear-algebraic problem like the ℓ_2 regression problem even be represented by a smaller-sized input for the purpose of accurately approximating the ℓ_2 regression problem? A related and no less important algorithmic question is: can the smaller-sized sample be found efficiently?

The main result of this paper is an elaborate sampling algorithm that represents a matrix by a small (nearly constant) number of rows so that the ℓ_2 regression problem can be solved to accuracy $1 \pm \epsilon$ for any $\epsilon > 0$. To our knowledge, this sampling method yields the first known sublinear representation (or sample) for the accurate approximation of the ℓ_2 regression problem. More precisely, we present and analyze an algorithm that constructs and solves an induced subproblem of the ℓ_2 regression problem of Equations (1.1) and (1.2). Let $DS^T A$ be the $r \times d$ matrix consisting of the sampled and appropriately rescaled rows of the original matrix A, and let $DS^T b$ be the r-vector consisting of

*Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York 12180, drinep@cs.rpi.edu.

†Yahoo Research Labs, Sunnyvale, California 94089, mahoney@yahoo-inc.com. Part of this work was performed while at the Department of Mathematics, Yale University, New Haven, Connecticut, USA 06520.

‡Department of Computer Science, Rutgers University, New Brunswick NJ 08854, e-mail: muthu@cs.rutgers.edu. This work was supported by NSF DMS-0354690.

SODA '06, January 22-26, Miami, FL
 ISBN 0-89871-605-5/06/01

the sampled and appropriately rescaled rows (i.e., elements) of b. (This notation is defined more precisely in Section 2.) Then consider the problem

$$\tilde{\mathcal{Z}} = \min_{x \in \mathbb{R}^d} \left| DS^T b - DS^T A x \right|_2 . \tag{1.3}$$

The minimum-length vector $\tilde{x}_{opt} \in \mathbb{R}^d$ among those that achieve the minimum value $\tilde{\mathcal{Z}}$ in the *sampled* ℓ_2 regression problem of Equation (1.3) is

$$\tilde{x}_{opt} = \left(DS^T A\right)^+ DS^T b. \tag{1.4}$$

Since we will sample a number of rows $r \ll n$ of the original problem, we will compute (1.4), and thus (1.3), exactly. Our main theorem, Theorem 3.1, states that under appropriate assumptions on the original problem and on the sampling probabilities, the computed quantities $\tilde{\mathcal{Z}}$ and $\tilde{x}_{opt}$ will provide very accurate *relative error* approximations to the exact solution $\mathcal{Z}$ and the optimal vector x_{opt}. Since $r(= O(d^2))$ will always be chosen to be at least d, solving the sampled ℓ_2 regression problem will take $O(rd^2)$ time, i.e., time polynomial in just d and independent of n.

The main technical contribution of this paper has to do with the nonuniform probabilities that we use for sampling, which depend in a sensitive manner on the Singular Value Decomposition of A, and not on A itself. Existing methods employ sampling probabilities that depend on the Euclidean norms of rows and/or columns of the matrix [8, 9, 4, 5, 6, 15, 7, 3] or on the magnitudes of the individual elements [1]. Although these methods are appropriate for capturing coarse structure such as approximating matrix multiplication or computing low-rank matrix approximations, they seem inadequate for solving problems such as ℓ_2 regression. Intuitively, this is since approximating ℓ_2 regression depends in a more sensitive manner on the way in which A disperses information to its column space and the manner in which the target vector b interacts with that column space and its complement. Our sampling probabilities will depend on the lengths, i.e., the Euclidean norms, of the *rows of the left singular vectors* of A as well as the manner in which b lies in the complement of the column space of A. More precisely, they will be required to satisfy the conditions of the form: (3.8), (3.9), and (3.10). Note that, as described in Section 3, $O(nd^2)$ time suffices to compute nontrivial probabilities satisfying these conditions, and it is an open problem whether such probabilities can be computed more rapidly. Thus, our result may be viewed as showing the existence of a small sample for the accurate approximation of the ℓ_2 regression problem, although producing the sample is currently no faster than solving the original ℓ_2 regression problem.

2 Review of Linear Algebra and Approximating Matrix Multiplication

Let $[n]$ denote the set $\{1, 2, \ldots, n\}$. For any matrix $A \in \mathbb{R}^{m \times n}$, let $A_{(i)}, i \in [m]$ denote the i-th row of A as a row vector, and let $A^{(j)}, j \in [n]$ denote the j-th column of A as a column vector. Let the rank of A be $\rho \leq \min\{m, n\}$. The Singular Value Decomposition (SVD) of A is denoted by $A = U_A \Sigma_A V_A^T$, where $U_A \in \mathbb{R}^{m \times \rho}$, $\Sigma_A \in \mathbb{R}^{\rho \times \rho}$, and $V_A \in \mathbb{R}^{n \times \rho}$. Let $\sigma_i(A), i \in [\rho]$ denote the i-th singular value of A, and $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ denote the maximum and minimum singular value of A. The condition number of A is $\kappa(A) = \sigma_{\max}(A)/\sigma_{\min}(A)$. The Moore-Penrose generalized inverse, or pseudoinverse, of A may be expressed in terms of the SVD as $A^+ = V_A \Sigma_A^{-1} U_A^T$ [14]. Finally, for any orthogonal matrix $U \in \mathbb{R}^{m \times \ell}$, let $U^\perp \in \mathbb{R}^{m \times (m-\ell)}$ denote an orthogonal matrix whose columns are an orthonormal basis spanning the subspace of $\mathbb{R}^m$ that is orthogonal to the column space of U. In terms of $U_A^\perp$, the solution of the ℓ_2 regression problem (1.1) is

$$\mathcal{Z} = \min_{x \in \mathbb{R}^d} |b - Ax|_2 = \left| U_A^\perp {U_A^\perp}^T b \right|_2 . \tag{2.5}$$

For more details on linear algebra, see [11, 10, 2].

The following result on approximating the product of two matrices by random sampling will be used in an essential manner in our main result; it is described in more detail in [4]. Algorithm 1 takes as input two matrices A and B, a number $c \leq n$, and a probability distribution over $[n]$. It returns as output two matrices C and R, where the columns of C are a small number of sampled and rescaled columns of A and where the rows of R are a small number of sampled and rescaled rows of B. To state Algorithm 1, we have used the following sampling matrix formalism which was introduced in [4]. Assume that the i_t-th column of A (and thus also the i_t-th row of B) is chosen in the t-th (for $t = 1, \ldots, c$) independent random trial. Then, define the sampling matrix $S \in \mathbb{R}^{n \times c}$ to be the zero-one matrix where $S_{i_t t} = 1$ and $S_{ij} = 0$ otherwise, and define the diagonal rescaling matrix $D \in \mathbb{R}^{c \times c}$ to be the diagonal matrix with $D_{tt} = 1/\sqrt{cp_{i_t}}$, where p_{i_t} is the probability of choosing the i_t-th column-row pair. Then, clearly, $C = ASD$ is an $m \times c$ matrix consisting of sampled and rescaled copies of the columns of A, and $R = (SD)^T B = DS^T B$ is a $c \times n$ matrix consisting of sampled and rescaled copies of the rows of B.

Theorem 2.1 is a basic quality-of-approximation result for Algorithm 1. Its proof may be found in [4], and it states that, under appropriate assumptions,

$$CR = ASDDS^T B \approx AB.$$

The most interesting of these assumptions is that the

Data : $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$, $\{p_i\}_{i=1}^n$ such that $\sum_{i=1}^n p_i = 1$, $c \le n$.
Result : $C \in \mathbb{R}^{m\times c}$, $R \in \mathbb{R}^{c\times p}$.
$(n \times c)$ matrix $S = \mathbf{0}_{n\times c}$;
$(c \times c)$ matrix $D = \mathbf{0}_{c\times c}$;
for $t = 1, \ldots, c$ **do**
  Pick $i_t \in [n]$, where $\mathbf{Pr}(i_t = i) = p_i$;
  $D_{tt} = 1/\sqrt{cp_{i_t}}$;
  $S_{i_t t} = 1$;
end
$C = ASD$;
$R = DS^T B$;

Algorithm 1: A fast Monte-Carlo algorithm for approximate matrix multiplication described in [4].

sampling probabilities used to randomly sample the columns of A and the corresponding rows of B are nonuniform and depend on the product of the Euclidean norm of the columns of A and the corresponding rows of B. Note that sampling probabilities of the form (2.6), with $\beta = 1$, are optimal for approximating AB by CR in a sense made precise in [4]. Allowing more general probabilities of the form (2.6) leads to only a small β-dependent loss in accuracy, but it provides important flexibility that will be used in an essential manner in our main result.

THEOREM 2.1. *Suppose $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$, and the sampling probabilities $\{p_i\}_{i=1}^n$ are such that*

$$p_i \ge \beta \frac{\left|A^{(i)}\right|_2 \left|B_{(i)}\right|_2}{\sum_{j=1}^n \left|A^{(j)}\right|_2 \left|B_{(j)}\right|_2} \tag{2.6}$$

for some $\beta \in (0,1]$. Construct C and R with Algorithm 1, and assume that $\delta \in (0,1/3)$. Then, with probability at least $1-\delta$:

$$\|AB - CR\|_F \le \frac{4\sqrt{\ln(1/\delta)}}{\beta\sqrt{c}} \|A\|_F \|B\|_F . \tag{2.7}$$

Note that for simplicity of presentation we have slightly modified the statement of the theorem for the purposes of this paper. (Theorem 2.1 follows immediately from the corresponding theorem in [4] since

$$\left(1 + \sqrt{(8/\beta)\log(1/\delta)}\right)/(\sqrt{\beta c}) \le 4\sqrt{\ln(1/\delta)}/(\beta\sqrt{c})$$

if $\beta \in (0,1]$ and $\delta \in (0,1/3)$.) Note also that if the sampling probabilities of the form (2.6) are used, then Algorithm 1 may be implemented in two passes over the data matrices from external storage and $O(c(m+n+p))$ additional storage space and computation time; see [4] for more details.

3 Our Main Result for ℓ_2 Regression

In this section, we first present Algorithm 2, our main random sampling algorithm for approximating the solution to the ℓ_2 regression problem, as defined in Equations (1.1) and (1.2). Then, we discuss the sufficient conditions on the nonuniform sampling probabilities used by the algorithm that will guarantee that the algorithm returns a good approximation to the original problem. Then, we state Theorem 3.1, which provides our main quality-of-approximation result for Algorithm 2. Finally, we provide a discussion of several observations related to the conditions we impose on the sampling probabilities. The proof of Theorem 3.1 is deferred to Section 4.

3.1 Description of our main algorithm Algorithm 2 takes as input an $n \times d$ (where $d \ll n$) matrix A, an n-vector b, a set of sampling probabilities $\{p_i\}_{i=1}^n$, and a positive integer $r \le n$. It returns as output a number $\tilde{\mathcal{Z}}$ and a d-vector $\tilde{x}_{opt}$. Using the sampling matrix formalism described in Section 2, the algorithm (implicitly) forms a sampling matrix S, the transpose of which samples with replacement a few rows of A and also the corresponding elements of b, and a rescaling matrix D, which is a diagonal matrix scaling the sampled rows of A and the elements of b. Since r rows of A and the corresponding r elements of b are sampled, the algorithm randomly samples with replacement r of the n constraints in the original overconstrained ℓ_2 regression problem. Thus, intuitively, the algorithm approximates the solution of $Ax \approx b$ with the exact solution of the downsampled problem $DS^TAx \approx DS^Tb$. Note that it is the high-dimensional space of constraints that is sampled and that the dimension of the unknown vector x is the same in both problems. Note also that, as we will see below, $r(= O(d^2))$ will always be chosen to be at least d, and thus (assuming that we have access to the sampling probabilities – see below) solving the sampled ℓ_2 regression problem takes $O(rd^2)$ time, i.e., time polynomial just in d and independent of n, rather than $O(nd^2)$ time.

3.2 Conditions on the sampling probabilities An important aspect of the algorithm will be the nonuniform sampling probabilities. The nonuniform probabilities that we will consider and that will be sufficient for the bounds we obtain will be any probabilities

Data : $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, $\{p_i\}_{i=1}^n$ such that $\sum_{i=1}^n p_i = 1$, $r \leq n$.
Result : $\tilde{x}_{opt} \in \mathbb{R}^d$, $\tilde{\mathcal{Z}} \in \mathbb{R}$.
$(n \times r)$ matrix $S = \mathbf{0}_{n\times r}$;
$(r \times r)$ matrix $D = \mathbf{0}_{r\times r}$;
for $t = 1, \ldots, r$ **do**
  Pick $i_t \in [n]$, where $\mathbf{Pr}(i_t = i) = p_i$;
  $D_{tt} = 1/\sqrt{rp_{i_t}}$;
  $S_{i_t t} = 1$;
end
$\tilde{\mathcal{Z}} = \min_{x\in\mathbb{R}^d} \left|DS^Tb - DS^TAx\right|_2$;
$\tilde{x}_{opt} = \left(DS^TA\right)^+ DS^Tb$;

Algorithm 2: A Monte-Carlo algorithm for approximating ℓ_2 regression.

that satisfy the following three conditions:

$$(3.8)\quad p_i \geq \beta_1 \frac{\left|(U_A)_{(i)}\right|_2^2}{\sum_{j=1}^n \left|(U_A)_{(j)}\right|_2^2},$$

$$(3.9)\quad p_i \geq \beta_2 \frac{\left|(U_A)_{(i)}\right|_2 \left(U_A^\perp U_A^{\perp\,T} b\right)_i}{\sum_{j=1}^n \left|(U_A)_{(j)}\right|_2 \left(U_A^\perp U_A^{\perp\,T} b\right)_j},$$

$$(3.10)\quad p_i \geq \beta_3 \frac{\left(U_A^\perp U_A^{\perp\,T} b\right)_i^2}{\sum_{j=1}^n \left(U_A^\perp U_A^{\perp\,T} b\right)_j^2}.$$

Several things should be noted about these conditions on the sampling probabilities. First, probabilities satisfying these three conditions clearly exist. For example, for all $i \in [n]$, let

$$(3.11)\quad p_i = \frac{(1/3)\left|(U_A)_{(i)}\right|_2^2}{\sum_{j=1}^n \left|(U_A)_{(j)}\right|_2^2} + \frac{(1/3)\left|(U_A)_{(i)}\right|_2 \left(U_A^\perp U_A^{\perp\,T} b\right)_i}{\sum_{j=1}^n \left|(U_A)_{(j)}\right|_2 \left(U_A^\perp U_A^{\perp\,T} b\right)_j} + \frac{(1/3)\left(U_A^\perp U_A^{\perp\,T} b\right)_i^2}{\sum_{j=1}^n \left(U_A^\perp U_A^{\perp\,T} b\right)_j^2}.$$

Then (3.8), (3.9), and (3.10) are satisfied with $\beta_1 = \beta_2 = \beta_3 = 1/3$. Second, note that almost all probability distributions satisfy these conditions (in the sense that all but a measure-zero set do satisfy them). In particular, by choosing sufficiently small values for β_1, β_2, and β_3, almost all sets of probabilities satisfy (3.8), (3.9), and (3.10). Of course, as was seen in Theorem 2.1, relaxing the β_i's has a direct and adverse effect on the sampling complexity. Third, probabilities satisfying these three conditions, e.g., the probabilities of (3.11), can be computed in $O(d^2n)$ time, which is also sufficient for exactly solving the original unsampled ℓ_2 regression problem. Fourth, it is an open question whether we can compute sampling probabilities that satisfy the three constraints with *constant* values for β_1, β_2, and β_3 in $o(nd^2)$ time.

3.3 Statement of our main theorem Theorem 3.1 is our main quality-of-approximation result for Algorithm 2. Its proof may be found in Section 4.

THEOREM 3.1. *Suppose $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, and that the sampling probabilities $\{p_i\}_{i=1}^n$ are given. Let*

$$\mathcal{Z} = \min_{x\in\mathbb{R}^d} |b - Ax|_2 = |b - Ax_{opt}|_2,$$

where $x_{opt} = A^+b$, let Algorithm 2 return as output a number $\tilde{\mathcal{Z}}$ and a d-vector $\tilde{x}_{opt}$, and let $\epsilon \in (0,1]$. Then,

- *If the sampling probabilities satisfy (3.8) and (3.10), and if $r \geq \frac{64d^2\ln(3/\delta)}{\epsilon^4 \min\{\beta_1^2,\beta_3^2\}}$ then with probability at least $1-\delta$:*

  $$(3.12)\qquad \tilde{\mathcal{Z}} \leq (1+\epsilon)\,\mathcal{Z}.$$

- *If the sampling probabilities satisfy (3.8), (3.9), and (3.10), and if $r \geq \frac{388d^2\ln(3/\delta)}{\epsilon^2 \min\{\beta_1^2,\beta_2^2,\beta_3^2\}}$ then with probability at least $1-\delta$:*

  $$(3.13)\qquad |b - A\tilde{x}_{opt}|_2 \leq (1+\epsilon)\,\mathcal{Z},$$

  $$(3.14)\qquad |x_{opt} - \tilde{x}_{opt}|_2 \leq \frac{\epsilon}{\sigma_{\min}(A)}\mathcal{Z}.$$

- *If, in addition, we assume that*

  $$(3.15)\qquad \left|U_AU_A^Tb\right|_2 \geq \gamma\,|b|_2$$

  for some fixed $\gamma \in (0,1]$, then with probability at least $1-\delta$:

  $$(3.16)\qquad |x_{opt} - \tilde{x}_{opt}|_2 \leq \epsilon\left(\kappa(A)\sqrt{\gamma^{-2}-1}\right)|x_{opt}|_2.$$

Equation (3.12) states that solving the sampled ℓ_2 regression problem provides a minimum value that is an accurate approximation to the minimum value of the original ℓ_2 regression problem, and Equation (3.13)

states that if the minimum-length vector achieving the minimum in the sampled problem is substituted back into the original problem then a good approximation to the original ℓ_2 regression problem is obtained. Both provide a relative error approximation to $\mathcal{Z}$.

Equation (3.14) provides a bound for $|x_{opt} - \tilde{x}_{opt}|_2$ in terms of $\sigma_{\min}(A)$ and $\mathcal{Z}$. If most of the "weight" of b lies in the complement of the column space of A then this will provide a very poor approximation in terms of $|x_{opt}|_2$. However, if we also assume (3.15), i.e., if a constant fraction of the "weight" of b lies in the subspace spanned by the columns of A, then we obtain the relative error approximation of Equation (3.16). Thus, Theorem 3.1 returns a good bound for $|x_{opt} - \tilde{x}_{opt}|_2$ if A is well-conditioned and if b lies "reasonably well" in the column space of A.

Note that if the target vector b lies completely within the column space of A, then $\mathcal{Z} = 0$ and $\gamma = 1$. In this case, Theorem 3.1 shows that Algorithm 2 returns $\tilde{\mathcal{Z}}$ and $\tilde{x}_{opt}$ that are exact solutions of the original ℓ_2 regression problem, independent of $\kappa(A)$.

3.4 Discussion Before concluding this section, we briefly discuss at an informal level several observations regarding sampling probabilities that satisfy the each of the three conditions (3.8), (3.9), and (3.10).

3.4.1 Discussion of the three conditions Condition (3.8) is the most interesting condition on the sampling probabilities, as will be seen in Section 4. This condition states that the sampling probabilities should be close to, or rather not much less than, the lengths, i.e., the Euclidean norms, of the rows of the left singular vectors of the matrix A. (Recall that A is an $n \times d$ matrix, and thus U_A is an $n \times \rho$ matrix, where $\rho = \text{rank}(A) \leq d \ll n$. Thus, the Euclidean norm of every *column* of U_A equals 1, but the Euclidean norm of every *row* of U_A is in general not equal and is only bounded above by 1.) These lengths may be interpreted as capturing a notion of information dispersal by the matrix A since they indicate to which part of the n-dimensional vector space the singular value information of A is being dispersed. In this case, condition (3.8) ensures that the sampling probabilities provide a bias toward the part of the high-dimensional constraint space to which A disperses its singular value information.

Condition (3.10) has information about where the target vector b is positioned relative to the matrix A of constraint vectors. Sampling probabilities satisfying condition (3.10) provide a bias toward the part of the complement of the column space of A where b has significant weight. Although it may seem counterintuitive that the bias is not toward, e.g., the part of the column space of A where b has significant weight, note that information about the column space of A has already been taken into account by condition (3.8). Relatedly, since (3.12), (3.13), and (3.14) all provide bounds in terms of

$$\mathcal{Z} = \min_{x \in \mathbb{R}^d} |b - Ax|_2 = \left| U_A^{\perp} {U_A^{\perp}}^T b \right|_2,$$

a condition like (3.10) seems necessary.

Condition (3.9) captures a combination of the two previous effects. Note that this condition is not needed to prove (3.12). Note also that this condition is not needed to prove that each of the statements of Theorem 3.1 holds with probability at least $1-\delta$, if we are willing to sample a number r of rows that is proportional to $1/\delta$, rather than proportional to $\sqrt{\ln(1/\delta)}$, as in Theorem 3.1. This follows by using Markov's inequality and Lemma 8 of [4].

3.4.2 Intuition behind the sampling probabilities Sampling probabilities satisfying the three conditions (3.8), (3.9), and (3.10) should be contrasted with sampling probabilities that depend on the Euclidean norms of the rows or columns of $A = U_A \Sigma_A V_A^T$ and that have received much attention recently; see, e.g., [8, 9] and more recently [4, 5, 6, 7]. Sampling probabilities with this latter form depend in a complicated manner on a mixture of subspace information (as found in U_A and V_A) and "size-of-A" information (as found in Σ_A). This convolution of information may account for their ability to capture coarse statistics such as approximating matrix multiplication or computing low-rank matrix approximations, but it also accounts for their difficulty in dealing with problems such as ℓ_2 regression.

Since the solution of the ℓ_2 regression problem involves the computation of a pseudoinverse, the problem is not well-conditioned with respect to a perturbation (such as that introduced by sampling) that entails a change in dimensionality, even if (actually, especially if) that change in dimensionality corresponds to a small singular value. Since sampling probabilities satisfying (3.8) allow us to disentangle subspace information and "size-of-A" information, we will see that they will allow us to capture (with high probability) the *entire* subspace of interest by sampling. More precisely, as we will see in Lemma 4.1, by using sampling probabilities that satisfy condition (3.8) and by choosing r appropriately, then with high probability it will follow that

$$\text{rank}(DS^T U_A) = \text{rank}(U_A) = \text{rank}(A).$$

Then, we will go to the low-dimensional, i.e., the r-dimensional rather than then n-dimensional, space and approximate the ℓ_2 regression problem by doing computations that involve "size-of-A" information on the random sample.

4 Proof of Our Main Theorem

In this section we provide a proof of Theorem 3.1, which is our main quality-of-approximation result for Algorithm 2. For simplicity of notation in this section, we will let $\mathcal{S} = DS^T$ denote the $r \times n$ rescaled row-sampling matrix.

Let the rank of the $n \times d$ matrix A be $\rho \leq d$, and let its SVD be

$$A = U_A \Sigma_A V_A^T,$$

where $U_A \in \mathbb{R}^{n \times \rho}$, $\Sigma_A \in \mathbb{R}^{\rho \times \rho}$, and $V_A \in \mathbb{R}^{d \times \rho}$. In addition, let the rank of the $r \times \rho$ matrix $\mathcal{S}U_A = DS^T U_A$ be $\tilde{\rho}$, and let its SVD be

$$\mathcal{S}U_A = U_{\mathcal{S}U_A} \Sigma_{\mathcal{S}U_A} V_{\mathcal{S}U_A}^T,$$

where $U_{\mathcal{S}U_A} \in \mathbb{R}^{r \times \tilde{\rho}}$, $\Sigma_{\mathcal{S}U_A} \in \mathbb{R}^{\tilde{\rho} \times \tilde{\rho}}$, and $V_{\mathcal{S}U_A} \in \mathbb{R}^{\rho \times \tilde{\rho}}$. Recall that $\mathcal{S} = DS^T$ denotes the $r \times n$ row-sampling-and-rescaling matrix, and that $\tilde{\rho} \leq \rho \leq d \leq r$.

In order to illustrate the essential difficulty in constructing a sampling algorithm to approximate the solution of the ℓ_2 regression problem, consider inserting $\tilde{x}_{opt} = (\mathcal{S}A)^+ \mathcal{S}b$ into $b - Ax$:

$$\begin{aligned} b - A\tilde{x}_{opt} &= b - A\left(\mathcal{S}A\right)^+ \mathcal{S}b \\ &= b - U_A \Sigma_A V_A^T \left(\mathcal{S}U_A \Sigma_A V_A^T\right)^+ \mathcal{S}b \\ &= b - U_A \Sigma_A \left(\mathcal{S}U_A \Sigma_A\right)^+ \mathcal{S}b \\ &= b - U_A \Sigma_A \left(U_{\mathcal{S}U_A} \Sigma_{\mathcal{S}U_A} V_{\mathcal{S}U_A}^T \Sigma_A\right)^+ \mathcal{S}b \\ &= b - U_A \Sigma_A \left(\Sigma_{\mathcal{S}U_A} V_{\mathcal{S}U_A}^T \Sigma_A\right)^+ U_{\mathcal{S}U_A}^T \mathcal{S}b. \end{aligned}$$

To proceed further, we must deal with the pseudoinverse, which is not well-behaved with respect to perturbations that involve a change in dimensionality. To deal with this, we will focus on probabilities that depend on the subspace that we are downsampling, i.e., that depend on U_A, in order to guarantee that we capture the full subspace of interest.

4.1 Several lemmas of general interest In this subsection, we will present several lemmas, each of which will be used in the proof of (most or) all of the claims of Theorem 3.1. Then, in the next four subsections, we will use these lemmas to provide a proof of each of the four claims of Theorem 3.1.

For the first lemma of this subsection, r depends quadratically on d, and the only assumption on the sampling probabilities is that they satisfy condition (3.8).

LEMMA 4.1. *Let $\epsilon \in (0, 1]$. If the sampling probabilities satisfy equation (3.8) and if $r \geq 64d^2 \ln(3/\delta)/\left(\beta_1^2 \epsilon^2\right)$, then with probability at least $1 - \delta/3$:*

$$\tilde{\rho} = \rho \text{ i.e., } \mathrm{rank}(\mathcal{S}U_A) = \mathrm{rank}(U_A) = \mathrm{rank}(A) \tag{4.17}$$

$$\left\|\Sigma_{\mathcal{S}U_A} - \Sigma_{\mathcal{S}U_A}^{-1}\right\|_2 \leq \epsilon/\sqrt{2} \tag{4.18}$$

$$(\mathcal{S}A)^+ = V_A \Sigma_A^{-1} \left(\mathcal{S}U_A\right)^+. \tag{4.19}$$

Proof. To prove the first claim, note that for all $i \in [\rho]$

$$\begin{aligned} \left|1 - \sigma_i^2\left(\mathcal{S}U_A\right)\right| &= \left|\sigma_i\left(U_A^T U_A\right) - \sigma_i\left(U_A^T \mathcal{S}^T \mathcal{S} U_A\right)\right| \\ &\leq \left\|U_A^T U_A - U_A^T \mathcal{S}^T \mathcal{S} U_A\right\|_2 \\ &\leq \left\|U_A^T U_A - U_A^T \mathcal{S}^T \mathcal{S} U_A\right\|_F. \end{aligned} \tag{4.20}$$

To bound the error of approximating $U_A^T U_A$ by $U_A^T \mathcal{S}^T \mathcal{S} U_A$ we apply our main theorem for approximating the product of two matrices. Since the sampling probabilities p_i satisfy equation (3.8), it follows from Theorem 2.1 that with probability at least $1 - \delta/3$:

$$\begin{aligned} \left\|U_A^T U_A - U_A^T \mathcal{S}^T \mathcal{S} U_A\right\|_F &\leq \frac{4\sqrt{\ln(3/\delta)}}{\beta_1 \sqrt{r}} \left\|U_A\right\|_F^2 \\ &\leq \frac{4d\sqrt{\ln(3/\delta)}}{\beta_1 \sqrt{r}}, \end{aligned} \tag{4.21}$$

where (4.21) follows since $\|U_A\|_F^2 = \rho \leq d$. By combining (4.20) and (4.21), and using the assumed choice of r it follows that

$$\left|1 - \sigma_i^2\left(\mathcal{S}U_A\right)\right| \leq \epsilon/2 \leq 1/2$$

since $\epsilon \leq 1$. This implies that all singular values of $\mathcal{S}U_A$ are strictly positive, and thus that $\mathrm{rank}(\mathcal{S}U_A) = \mathrm{rank}(U_A) = \mathrm{rank}(A)$, which establishes the first claim.

To prove the second claim, recall that under the assumptions of the lemma $\rho = \tilde{\rho}$ with high probability, and thus $\sigma_i\left(\mathcal{S}U_A\right) > 0$ for all $i \in [\rho]$. Thus,

$$\begin{aligned} \left\|\Sigma_{\mathcal{S}U_A}^{-1} - \Sigma_{\mathcal{S}U_A}\right\|_2 &= \max_{i,j \in [\rho]} \left|\sigma_i\left(\mathcal{S}U_A\right) - \frac{1}{\sigma_j\left(\mathcal{S}U_A\right)}\right| \\ &= \max_{i,j \in [\rho]} \frac{\left|\sigma_i\left(\mathcal{S}U_A\right)\sigma_j\left(\mathcal{S}U_A\right) - 1\right|}{\left|\sigma_j\left(\mathcal{S}U_A\right)\right|} \\ &\leq \max_{j \in [\rho]} \frac{\left|\sigma_j^2\left(\mathcal{S}U_A\right) - 1\right|}{\left|\sigma_j\left(\mathcal{S}U_A\right)\right|}. \end{aligned} \tag{4.22}$$

Using that fact that for all $i \in [\rho]$,

$$\left|1 - \sigma_i^2\left(\mathcal{S}U_A\right)\right| \leq \left\|U_A^T U_A - U_A^T \mathcal{S}^T \mathcal{S} U_A\right\|_2,$$

it follows that for all $i \in [\rho]$

$$\frac{1}{\sigma_i\left(\mathcal{S}U_A\right)} \leq \frac{1}{\sqrt{1 - \left\|U_A^T U_A - U_A^T \mathcal{S}^T \mathcal{S} U_A\right\|_2}}.$$

When these are combined with (4.22) it follows that

$$\left\|\Sigma_{\mathcal{S}U_A} - \Sigma_{\mathcal{S}U_A}^{-1}\right\|_2 \leq \frac{\left\|U_A^T U_A - U_A^T \mathcal{S}^T \mathcal{S} U_A\right\|_2}{\sqrt{1 - \left\|U_A^T U_A - U_A^T \mathcal{S}^T \mathcal{S} U_A\right\|_2}}.$$

Combining this with the Frobenius norm bound of (4.21), and noticing that our choice for r guarantees that $1 - \left\|U_A^T U_A - U_A^T \mathcal{S}^T \mathcal{S} U_A\right\|_2 \geq 1/2$, concludes the proof of the second claim.

Finally, to prove the third claim, note that

$$\begin{aligned}
(\mathcal{S}A)^+ &= \left(\mathcal{S}U_A \Sigma_A V_A^T\right)^+ \\
&= \left(U_{\mathcal{S}U_A} \Sigma_{\mathcal{S}U_A} V_{\mathcal{S}U_A}^T \Sigma_A V_A^T\right)^+ \\
&= V_A \left(\Sigma_{\mathcal{S}U_A} V_{\mathcal{S}U_A}^T \Sigma_A\right)^+ U_{\mathcal{S}U_A}^T. \qquad (4.23)
\end{aligned}$$

To remove the pseudoinverse in the above derivations, notice that since $\rho = \tilde{\rho}$ with high probability, all three matrices $\Sigma_{\mathcal{S}U_A}$, $V_{\mathcal{S}U_A}$, and Σ_A are full rank square $\rho \times \rho$ matrices, and thus are invertible. In this case,

$$\begin{aligned}
\left(\Sigma_{\mathcal{S}U_A} V_{\mathcal{S}U_A}^T \Sigma_A\right)^+ &= \left(\Sigma_{\mathcal{S}U_A} V_{\mathcal{S}U_A}^T \Sigma_A\right)^{-1} \\
&= \Sigma_A^{-1} V_{\mathcal{S}U_A} \Sigma_{\mathcal{S}U_A}^{-1}. \qquad (4.24)
\end{aligned}$$

By combining (4.23) and (4.24) we have that

$$\begin{aligned}
(\mathcal{S}A)^+ &= V_A \Sigma_A^{-1} V_{\mathcal{S}U_A} \Sigma_{\mathcal{S}U_A}^{-1} U_{\mathcal{S}U_A}^T \\
&= V_A \Sigma_A^{-1} \left(\mathcal{S}U_A\right)^+,
\end{aligned}$$

which establishes the third claim. This concludes the proof of the lemma.

The previous lemma showed that, in terms of its singular values, the matrix $\mathcal{S}U_A$, i.e., the row sampled and rescaled version of U_A, is almost an orthogonal matrix. A useful property of a matrix U with orthonormal columns is that $U^+ = U^T$. The next lemma (especially in combination with the previous lemma) shows that, although this property does not hold for $\mathcal{S}U_A$, the difference between $(\mathcal{S}U_A)^+$ and $(\mathcal{S}U_A)^T$ can be bounded.

LEMMA 4.2. *Define* $\Omega = (\mathcal{S}U_A)^+ - (\mathcal{S}U_A)^T$. *Then,*

$$\|\Omega\|_2 = \left\|\Sigma_{\mathcal{S}U_A}^{-1} - \Sigma_{\mathcal{S}U_A}\right\|_2.$$

Proof. Using the SVD of $\mathcal{S}U_A$, we have that

$$\begin{aligned}
\|\Omega\|_2 &= \left\|\left(\mathcal{S}U_A\right)^+ - \left(\mathcal{S}U_A\right)^T\right\|_2 \\
&= \left\|\left(U_{\mathcal{S}U_A} \Sigma_{\mathcal{S}U_A} V_{\mathcal{S}U_A}^T\right)^+ - \left(U_{\mathcal{S}U_A} \Sigma_{\mathcal{S}U_A} V_{\mathcal{S}U_A}^T\right)^T\right\|_2 \\
&= \left\|V_{\mathcal{S}U_A} \left(\Sigma_{\mathcal{S}U_A}^{-1} - \Sigma_{\mathcal{S}U_A}\right) U_{\mathcal{S}U_A}^T\right\|_2.
\end{aligned}$$

The lemma follows since $V_{\mathcal{S}U_A}$ and $U_{\mathcal{S}U_A}$ are matrices with orthonormal columns.

The next two lemmas provide two different approximate matrix multiplication bounds that are also useful in the proof of the claims of Theorem 3.1. For the next lemma, r depends linearly on d, and the only assumption on the sampling probabilities is that they satisfy condition (3.9).

LEMMA 4.3. *Let* $\epsilon \in (0,1]$. *If the sampling probabilities satisfy equation (3.9) and if* $r \geq 16 d \ln(3/\delta)/\left(\beta_2^2 \epsilon^2\right)$, *then with probability at least* $1 - \delta/3$:

$$\left|U_A^T \mathcal{S}^T \mathcal{S} U_A^\perp {U_A^\perp}^T b\right|_2 \leq \epsilon \left|U_A^\perp {U_A^\perp}^T b\right|_2.$$

Proof. First, note that since U_A is an orthogonal matrix, we have that

$$\left|U_A^T \mathcal{S}^T \mathcal{S} U_A^\perp {U_A^\perp}^T b\right|_2 = \left|U_A U_A^T \mathcal{S}^T \mathcal{S} U_A^\perp {U_A^\perp}^T b\right|_2.$$

Thus, since $U_A U_A^T U_A^\perp {U_A^\perp}^T b = 0$ we may write

$$\begin{aligned}
&\left|U_A^T \mathcal{S}^T \mathcal{S} U_A^\perp {U_A^\perp}^T b\right|_2 \qquad (4.25) \\
&\quad = \left|U_A U_A^T U_A^\perp {U_A^\perp}^T b - U_A U_A^T \mathcal{S}^T \mathcal{S} U_A^\perp {U_A^\perp}^T b\right|_2.
\end{aligned}$$

Thus, we can estimate the Euclidean norm of the vector $U_A^T \mathcal{S}^T \mathcal{S} U_A^\perp {U_A^\perp}^T b$ by bounding the error of approximating the product $U_A U_A^T U_A^\perp {U_A^\perp}^T b$ by $U_A U_A^T \mathcal{S}^T \mathcal{S} U_A^\perp {U_A^\perp}^T b$. To do so, note that since $\left|\left(U_A U_A^T\right)_{(i)}\right|_2 = \left|\left(U_A^T\right)_{(i)}\right|_2$ sampling probabilities appropriate for bounding the right hand side of (4.25) are also those satisfying satisfying (3.9). Thus, since the sampling probabilities p_i satisfy equation (3.9), it follows from Theorem 2.1 that with probability at least $1 - \delta/3$

$$\left|U_A^T \mathcal{S}^T \mathcal{S} U_A^\perp {U_A^\perp}^T b\right|_2 \leq \frac{4\sqrt{\ln(3/\delta)}}{\beta_2 \sqrt{r}} \left\|U_A U_A^T\right\|_F \left|U_A^\perp {U_A^\perp}^T b\right|_2.$$

The lemma follows by the choice of r and since $\left\|U_A U_A^T\right\|_F = \sqrt{\rho} \leq \sqrt{d}$.

The final lemma of this subsection relates the norm of the n-vector $U_A^\perp {U_A^\perp}^T b$ to the norm of the r-vector $\mathcal{S} U_A^\perp {U_A^\perp}^T b$, i.e., the sampled and rescaled version of the original n-vector. For this lemma, r is independent of d, and the only assumption on the sampling probabilities is that they satisfy condition (3.10).

LEMMA 4.4. *Let* $\epsilon \in (0,1]$. *If the sampling probabilities satisfy equation (3.10) and if* $r \geq 16 \ln(3/\delta)/\left(\beta_3^2 \epsilon^2\right)$, *then with probability at least* $1 - \delta/3$:

$$\left|\left|U_A^\perp {U_A^\perp}^T b\right|^2 - \left|\mathcal{S} U_A^\perp {U_A^\perp}^T b\right|^2\right| \leq \epsilon \left|U_A^\perp {U_A^\perp}^T b\right|_2^2.$$

Proof. First recall that

$$\begin{aligned}\left|U_A^\perp {U_A^\perp}^T b\right|^2 &= b^T U_A^\perp {U_A^\perp}^T U_A^\perp {U_A^\perp}^T b \\ \left|\mathcal{S}U_A^\perp {U_A^\perp}^T b\right|^2 &= b^T U_A^\perp {U_A^\perp}^T \mathcal{S}^T \mathcal{S} U_A^\perp {U_A^\perp}^T b.\end{aligned}$$

Thus, to prove this lemma it suffices to bound the error of approximating the product $b^T U_A^\perp {U_A^\perp}^T U_A^\perp {U_A^\perp}^T b$ by $b^T U_A^\perp {U_A^\perp}^T \mathcal{S}^T \mathcal{S} U_A^\perp {U_A^\perp}^T b$. Since the sampling probabilities p_i satisfy equation (3.10) and by the choice of r, it follows from Theorem 2.1 that with probability at least $1-\delta/3$:

$$\begin{aligned}&\left|b^T U_A^\perp {U_A^\perp}^T U_A^\perp {U_A^\perp}^T b - b^T U_A^\perp {U_A^\perp}^T \mathcal{S}^T \mathcal{S} U_A^\perp {U_A^\perp}^T b\right| \\ &\quad \leq \epsilon \left|U_A^\perp {U_A^\perp}^T b\right|_2^2,\end{aligned}$$

from which the lemma follows.

4.2 Proof of Equation (3.12) In this subsection, we will bound $\mathcal{S}b - \mathcal{S}A\tilde{x}_{opt}$, thus proving (3.12).

For the moment, let us assume that $r = 64d^2 \ln(3/\delta)/\left(\epsilon^2 \min\{\beta_1^2, \beta_3^2\}\right)$. Note that since $d \geq 1$ and since we have taken $\min\{\beta_1^2, \beta_3^2\}$ in the denominator, the assumption on r is satisfied for both Lemma 4.1 and Lemma 4.4. Thus, the claims of both lemmas hold simultaneously with probability at least $1 - 2(\delta/3) \geq 1-\delta$, and so let us condition on this event. Note also that Lemma 4.3 is not necessary to establish (3.12), and thus we do not assume that the probabilities satisfy condition (3.9).

First, we have that

$$\begin{aligned}\mathcal{S}b - \mathcal{S}A\tilde{x}_{opt} &= \mathcal{S}b - \mathcal{S}A\left(\mathcal{S}A\right)^+ \mathcal{S}b \\ (4.26) \qquad &= \mathcal{S}b - \mathcal{S}U_A \Sigma_A V_A^T V_A \Sigma_A^{-1} \left(\mathcal{S}U_A\right)^+ \mathcal{S}b \\ &= \mathcal{S}b - \mathcal{S}U_A \left(\mathcal{S}U_A\right)^+ \mathcal{S}b,\end{aligned}$$

where (4.26) follows from Lemma 4.1. Then, since $U_A U_A^T + U_A^\perp {U_A^\perp}^T = I_n$, we have that

$$\begin{aligned}\mathcal{S}b - \mathcal{S}A\tilde{x}_{opt} &= \mathcal{S}b - \mathcal{S}U_A \left(\mathcal{S}U_A\right)^+ \mathcal{S}U_A U_A^T b \\ &\quad - \mathcal{S}U_A \left(\mathcal{S}U_A\right)^+ \mathcal{S}U_A^\perp {U_A^\perp}^T b \\ &= \mathcal{S}b - \mathcal{S}U_A U_A^T b \\ &\quad - \mathcal{S}U_A \left(\mathcal{S}U_A\right)^+ \mathcal{S}U_A^\perp {U_A^\perp}^T b \\ &= \mathcal{S}U_A^\perp {U_A^\perp}^T b \\ &\quad - \mathcal{S}U_A \left(\mathcal{S}U_A\right)^+ \mathcal{S}U_A^\perp {U_A^\perp}^T b.\end{aligned}$$

Thus, it follows that

$$\begin{aligned}\tilde{\mathcal{Z}} &= \left|\mathcal{S}b - \mathcal{S}A\tilde{x}_{opt}\right|_2 \\ &= \left|\mathcal{S}U_A^\perp {U_A^\perp}^T b - \mathcal{S}U_A \left(\mathcal{S}U_A\right)^+ \mathcal{S}U_A^\perp {U_A^\perp}^T b\right|_2 \\ &= \left|\left(I - \mathcal{S}U_A \left(\mathcal{S}U_A\right)^+\right) \mathcal{S}U_A^\perp {U_A^\perp}^T b\right|_2 \\ (4.27) \qquad &\leq \left|\mathcal{S}U_A^\perp {U_A^\perp}^T b\right|_2,\end{aligned}$$

where (4.27) follows since $\mathcal{S}U_A \left(\mathcal{S}U_A\right)^+ = U_{\mathcal{S}U_A} U_{\mathcal{S}U_A}^T$ is a projection. By combining (4.27) with the bound on $\left|\mathcal{S}U_A^\perp {U_A^\perp}^T b\right|_2$ provided by Lemma 4.4, and recalling (2.5), it follows that

$$\tilde{\mathcal{Z}} \leq \left(1 + \sqrt{\epsilon}\right) \mathcal{Z}.$$

Equation (3.12) follows by setting $\epsilon' = \sqrt{\epsilon}$ and using the value of r assumed by the theorem.

4.3 Proof of Equation (3.13) In this subsection, we will bound $b - A\tilde{x}_{opt}$, thus proving (3.13).

For the moment, let us assume that $r = 64d^2 \ln(3/\delta)/\left(\epsilon^2 \min\{\beta_1^2, \beta_2^2, \beta_3^2\}\right)$. Note that since we have taken $\min\{\beta_1^2, \beta_2^2, \beta_3^2\}$ in the denominator, the assumption on r is satisfied for each of Lemma 4.1, Lemma 4.3, and Lemma 4.4. Thus, the claims of all three lemmas hold simultaneously with probability at least $1 - 3(\delta/3) \geq 1-\delta$, and so let us condition on this event.

First, we have that

$$\begin{aligned}b - A\tilde{x}_{opt} &= b - A\left(\mathcal{S}A\right)^+ \mathcal{S}b \\ (4.28) \qquad &= b - U_A \left(\mathcal{S}U_A\right)^+ \mathcal{S}b \\ (4.29) \qquad &= b - U_A \left(\mathcal{S}U_A\right)^+ \mathcal{S}U_A U_A^T b \\ &\quad - U_A \left(\mathcal{S}U_A\right)^+ \mathcal{S}U_A^\perp {U_A^\perp}^T b \\ (4.30) \qquad &= U_A^\perp {U_A^\perp}^T b - U_A \left(\mathcal{S}U_A\right)^+ \mathcal{S}U_A^\perp {U_A^\perp}^T b.\end{aligned}$$

(4.28) follows from Lemma 4.1, (4.29) follows by inserting $U_A U_A^T + U_A^\perp {U_A^\perp}^T = I_n$, and (4.30) follows since $\left(\mathcal{S}U_A\right)^+ \mathcal{S}U_A = I_\rho$ by Lemma 4.1. We emphasize that $\left(\mathcal{S}U_A\right)^+ \mathcal{S}U_A = V_{\mathcal{S}U_A} V_{\mathcal{S}U_A}^T = I_\rho$ does not hold for general sampling methods, but it does hold in this case since $\tilde{\rho} = \rho$, which follows from Lemma 4.1.

By taking the Euclidean norm of both sides of (4.30), by using the triangle inequality, and recalling that $\Omega = \left(\mathcal{S}U_A\right)^+ - \left(\mathcal{S}U_A\right)^T$, we have that

$$\begin{aligned}\left|b - A\tilde{x}_{opt}\right|_2 &\leq \left|U_A^\perp {U_A^\perp}^T b\right|_2 + \left|U_A \left(\mathcal{S}U_A\right)^T \mathcal{S}U_A^\perp {U_A^\perp}^T b\right|_2 \\ &\quad + \left|U_A \Omega \mathcal{S}U_A^\perp {U_A^\perp}^T b\right|_2 \\ (4.31) \qquad &\leq \left|U_A^\perp {U_A^\perp}^T b\right|_2 + \left|U_A^T \mathcal{S}^T \mathcal{S}U_A^\perp {U_A^\perp}^T b\right|_2 \\ &\quad + \|\Omega\|_2 \left|\mathcal{S}U_A^\perp {U_A^\perp}^T b\right|_2,\end{aligned}$$

where (4.31) follows by submultiplicitivity and since U_A is an orthogonal matrix. By combining (4.31) with the bounds provided by Lemma 4.1 through Lemma 4.4, and recalling (2.5), it follows that

$$\begin{aligned} |b - A\tilde{x}_{opt}|_2 &\leq (1 + \epsilon + \epsilon/\sqrt{2} + \epsilon^{3/2}/\sqrt{2})\mathcal{Z} \\ &\leq (1 + 2.5\epsilon)\mathcal{Z}, \end{aligned}$$

where the second inequality follows since $\epsilon \leq 1$. Equation (3.13) follows by setting $\epsilon' = \epsilon/2.5$ and using the value of r assumed by the theorem.

4.4 Proof of Equation (3.14) In this subsection, we will provide a bound for $|\tilde{x}_{opt} - x_{opt}|_2$ in terms of $\mathcal{Z}$, thus proving (3.14).

For the moment, let us assume that $r = 64d^2 \ln(3/\delta)/\left(\epsilon^2 \min\{\beta_1^2, \beta_2^2, \beta_3^2\}\right)$. Note that since we have taken $\min\{\beta_1^2, \beta_2^2, \beta_3^2\}$ in the denominator, the assumption on r is satisfied for each of Lemma 4.1, Lemma 4.3, and Lemma 4.4. Thus, the claims of all three lemmas hold simultaneously with probability at least $1 - 3(\delta/3) \geq 1 - \delta$, and so let us condition on this event.

Since $U_A U_A^T + U_A^\perp U_A^{\perp T} = I_n$ and $(\mathcal{S}U_A)^+ \mathcal{S}U_A = I_\rho$, we have that

$$\begin{aligned} x_{opt} - \tilde{x}_{opt} &= A^+ b - (\mathcal{S}A)^+ \mathcal{S}b \\ &= V_A \Sigma_A^{-1} U_A^T b - V_A \Sigma_A^{-1} (\mathcal{S}U_A)^+ \mathcal{S}b \\ &= V_A \Sigma_A^{-1} U_A^T b - V_A \Sigma_A^{-1} (\mathcal{S}U_A)^+ \mathcal{S}U_A U_A^T b \\ &\qquad - V_A \Sigma_A^{-1} (\mathcal{S}U_A)^+ \mathcal{S}U_A^\perp U_A^{\perp T} b \\ &= -V_A \Sigma_A^{-1} (\mathcal{S}U_A)^+ \mathcal{S}U_A^\perp U_A^{\perp T} b. \end{aligned}$$

Thus, it follows that

$$\begin{aligned} |x_{opt} - \tilde{x}_{opt}|_2 &= \left| V_A \Sigma_A^{-1} (\mathcal{S}U_A)^+ \mathcal{S}U_A^\perp U_A^{\perp T} b \right|_2 \\ &= \left| \Sigma_A^{-1} \left((\mathcal{S}U_A)^T + \Omega \right) \mathcal{S}U_A^\perp U_A^{\perp T} b \right|_2 \\ &\leq \frac{1}{\sigma_{min}(A)} \left| (\mathcal{S}U_A)^T \mathcal{S}U_A^\perp U_A^{\perp T} b \right|_2 \\ &\quad + \frac{1}{\sigma_{min}(A)} \left| \Omega \mathcal{S}U_A^\perp U_A^{\perp T} b \right|_2 \\ &\leq \frac{1}{\sigma_{min}(A)} \left| U_A^T \mathcal{S}^T \mathcal{S}U_A^\perp U_A^{\perp T} b \right|_2 \\ &\quad + \frac{1}{\sigma_{min}(A)} \|\Omega\|_2 \left| \mathcal{S}U_A^\perp U_A^{\perp T} b \right|_2 . \end{aligned} \tag{4.32}$$

By combining (4.32) with Lemma 4.1 through Lemma 4.4, and recalling (2.5), it follows that

$$\begin{aligned} |\tilde{x}_{opt} - x_{opt}|_2 &\leq \sigma_{\min}^{-1}(A) \left(\epsilon + \epsilon/\sqrt{2} + \epsilon^{3/2}/\sqrt{2} \right) \mathcal{Z} \\ &\leq \frac{2.5\epsilon}{\sigma_{\min}(A)} \mathcal{Z}, \end{aligned}$$

where the second inequality follows since $\epsilon \leq 1$. Equation (3.14) follows by setting $\epsilon' = \epsilon/2.5$ and using the value of r assumed by the theorem.

4.5 Proof of Equation (3.16) The error bound provided by (3.14) could be quite weak, since $\min_{x \in \mathbb{R}^d} |b - Ax|_2$ could be quite close or even equal to $|b|_2$, if b has most or all of its "weight" outside of the column space of A. Under a slightly stronger assumption, we will provide a bound $|\tilde{x}_{opt} - x_{opt}|_2$ in terms of $|x_{opt}|_2$, thus proving (3.16).

If we make the additional assumption (3.15), which is satisfied if a *constant fraction* of the "weight" of b lies in the subspace spanned by the columns of A, then it follows that

$$\begin{aligned} \mathcal{Z}^2 &= \left(\min_{x \in \mathbb{R}^d} |b - Ax|_2 \right)^2 \\ &= \left| U_A^\perp U_A^{\perp T} b \right|_2^2 \\ &= |b|_2^2 - \left| U_A U_A^T b \right|_2^2 \\ &\leq (\gamma^{-2} - 1) \left| U_A U_A^T b \right|_2^2 . \end{aligned} \tag{4.33}$$

In order to relate $\left|U_A U_A^T b\right|_2$ and thus $\mathcal{Z}$ to $|x_{opt}|_2$ note that

$$\begin{aligned} |x_{opt}|_2 &= \left| V_A \Sigma_A^{-1} U_A^T b \right|_2 \\ &= \left| \Sigma_A^{-1} U_A^T b \right|_2 \\ &\geq \sigma_{\min}(\Sigma_A^{-1}) \left| U_A^T b \right|_2 \\ &= \frac{\left| U_A U_A^T b \right|_2}{\sigma_{\max}(A)}. \end{aligned} \tag{4.34}$$

By combining (3.14) with (4.33) and (4.34), we get

$$\begin{aligned} |\tilde{x}_{opt} - x_{opt}|_2 &\leq \frac{\epsilon}{\sigma_{\min}(A)} \mathcal{Z} \\ &\leq \frac{\epsilon}{\sigma_{\min}(A)} \sqrt{\gamma^{-2} - 1} \left| U_A U_A^T b \right|_2 \\ &\leq \epsilon \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} \sqrt{\gamma^{-2} - 1} \, |x_{opt}|_2 , \end{aligned}$$

which establishes (3.16).

5 Conclusion

Recently, a row norm-based sampling algorithm has been applied to the ℓ_1 regression problem by Clarkson [3]. Such a sampling method does not appear to work for the ℓ_2 regression problem. Nevertheless, the similarities and differences between the two methods are worth exploring in greater detail.

Two open questions immediately suggest themselves. One is whether we can compute sampling probabilities that satisfy the three constraints (3.8), (3.9),

and (3.10) with constant values for β_1, β_2, and β_3 in $o(nd^2)$ time. A second open question is whether we can improve the sampling complexity, either with respect to ϵ or, more importantly, with respect to d. With regard to the latter, notice that the only place where we need a quadratic dependency on d is in bounding $\left\|U_A^T U_A - U_A^T SDDS^T U_A\right\|_F$ in Lemma 4.1. It seems likely that one could obtain an improved bound that would only require a linear dependency on d by bounding the spectral norm of the error matrix directly by employing techniques related to those of Rudelson and Vershynin [16, 18].

The overconstrained ℓ_2 regression problem is such a fundamental problem in applied mathematics and statistical data analysis that numerous applications of this work suggest themselves. We briefly discuss just two. One application is to statistical learning theory. A difficulty in applying traditional random sampling techniques to statistical learning problems is that the discriminative variables for a particular learning problem are not in general in the dominant part of the singular value spectrum. The sampling methodology described in this paper will have application to learning problems such as regression, classification, and clustering since, e.g., linear Support Vector Machines with ℓ_2-loss functions (such as those used in text classification) may be formulated as a regularized least squares problem [12]. A second application is to computing low-rank matrix approximations that are expressed in terms of the columns and/or rows of the input matrix (see, e.g., [5, 6] and references therein). In these algorithms, a small number of columns and/or rows are randomly sampled and used as a basis with which to express the remaining columns/rows in an approximate least squares sense. These algorithms typically sample columns and/or rows with probabilities depending on the norms of the columns and/or rows of the input matrix and return additive error guarantees. It seems likely that using the sampling methodology described in this paper one will be able to obtain relative error guarantees efficiently.

References

[1] D. Achlioptas and F. McSherry. Fast computation of low rank matrix approximations. In *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing*, pages 611–618, 2001.

[2] R. Bhatia. *Matrix Analysis.* Springer-Verlag, New York, 1997.

[3] K. Clarkson. Subgradient and sampling algorithms for L1 regression. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2005.

[4] P. Drineas, R. Kannan, and M.W. Mahoney. Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. Technical Report YALEU/DCS/TR-1269, Yale University Department of Computer Science, New Haven, CT, February 2004. Accepted for publication in the SIAM Journal on Computing.

[5] P. Drineas, R. Kannan, and M.W. Mahoney. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. Technical Report YALEU/DCS/TR-1270, Yale University Department of Computer Science, New Haven, CT, February 2004. Accepted for publication in the SIAM Journal on Computing.

[6] P. Drineas, R. Kannan, and M.W. Mahoney. Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. Technical Report YALEU/DCS/TR-1271, Yale University Department of Computer Science, New Haven, CT, February 2004. Accepted for publication in the SIAM Journal on Computing.

[7] P. Drineas and M.W. Mahoney. A randomized algorithm for a tensor-based generalization of the Singular Value Decomposition. Technical Report YALEU/DCS/TR-1327, Yale University Department of Computer Science, New Haven, CT, June 2005.

[8] A. Frieze, R. Kannan, and S. Vempala. Fast Monte-Carlo algorithms for finding low-rank approximations. In *Proceedings of the 39th Annual IEEE Symposium on Foundations of Computer Science*, pages 370–378, 1998.

[9] A. Frieze, R. Kannan, and S. Vempala. Fast Monte-Carlo algorithms for finding low-rank approximations. *Journal of the ACM*, 51(6):1025–1041, 2004.

[10] G.H. Golub and C.F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, 1989.

[11] R.A. Horn and C.R. Johnson. *Matrix Analysis.* Cambridge University Press, New York, 1985.

[12] S. S. Keerthi and D. DeCoste. A modified finite Newton method for fast solution of large scale linear SVMs. *Journal of Machine Learning Research*, 6:341–361, 2005.

[13] E. Kushilevitz and N. Nisan. *Communication Complexity.* Cambridge University Press, New York, 1996.

[14] M.Z. Nashed, editor. *Generalized Inverses and Applications.* Academic Press, New York, 1976.

[15] L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via iterative sampling. Technical Report MIT-LCS-TR-983, Massachusetts Institute of Technology, Cambridge, MA, March 2005.

[16] M. Rudelson and R. Vershynin. Approximation of matrices. *manuscript.*

[17] G.W. Stewart and J.G. Sun. *Matrix Perturbation Theory.* Academic Press, New York, 1990.

[18] R. Vershynin. Coordinate restrictions of linear operators in l_2^n. *manuscript.*