SIAM J. COMPUT.
Vol. 38, No. 5, pp. 2060–2078

SAMPLING ALGORITHMS AND CORESETS FOR ℓ_p REGRESSION*

ANIRBAN DASGUPTA†, PETROS DRINEAS‡, BOULOS HARB§, RAVI KUMAR†, AND MICHAEL W. MAHONEY¶

Abstract. The ℓ_p *regression problem* takes as input a matrix $A \in \mathbb{R}^{n\times d}$, a vector $b \in \mathbb{R}^n$, and a number $p \in [1,\infty)$, and it returns as output a number $\mathcal{Z}$ and a vector $x_{\text{OPT}} \in \mathbb{R}^d$ such that $\mathcal{Z} = \min_{x\in\mathbb{R}^d} \|Ax-b\|_p = \|Ax_{\text{OPT}}-b\|_p$. In this paper, we construct coresets and obtain an efficient two-stage sampling-based approximation algorithm for the very overconstrained ($n \gg d$) version of this classical problem, for all $p \in [1,\infty)$. The first stage of our algorithm nonuniformly samples $\hat{r}_1 = O(36^p d^{\max\{p/2+1,p\}+1})$ rows of A and the corresponding elements of b, and then it solves the ℓ_p regression problem on the sample; we prove this is an 8-approximation. The second stage of our algorithm uses the output of the first stage to resample $\hat{r}_1/\epsilon^2$ constraints, and then it solves the ℓ_p regression problem on the new sample; we prove this is a $(1+\epsilon)$-approximation. Our algorithm unifies, improves upon, and extends the existing algorithms for special cases of ℓ_p regression, namely, $p = 1, 2$ [K. L. Clarkson, in *Proceedings of the 16th Annual ACM–SIAM Symposium on Discrete Algorithms*, ACM, New York, SIAM, Philadelphia, 2005, pp. 257–266; P. Drineas, M. W. Mahoney, and S. Muthukrishnan, in *Proceedings of the 17th Annual ACM–SIAM Symposium on Discrete Algorithms*, ACM, New York, SIAM, Philadelphia, 2006, pp. 1127–1136]. In the course of proving our result, we develop two concepts—well-conditioned bases and subspace-preserving sampling—that are of independent interest.

Key words. randomized algorithms, sampling algorithms, ℓ_p regression

AMS subject classification. 68W20

DOI. 10.1137/070696507

1. Introduction.

An important question in algorithmic theory is whether there exists a *small* subset of the input such that if computations are performed only on this subset, then the solution to the given problem can be *approximated* well. Such a subset is often known as a *coreset* for the problem. The concept of coresets has been extensively used in solving many problems in optimization and computational geometry; e.g., see the excellent survey by Agarwal, Har-Peled, and Varadarajan [2].

In this paper, we construct coresets and obtain efficient sampling algorithms for the classical ℓ_p regression problem, for all $p \in [1,\infty)$. Recall the ℓ_p *regression problem.*

PROBLEM 1 (ℓ_p regression problem). *Let $\|\cdot\|_p$ denote the p-norm of a vector. Given as input a matrix $A \in \mathbb{R}^{n\times m}$, a target vector $b \in \mathbb{R}^n$, and a real number $p \in [1,\infty)$, find a vector x_{OPT} and a number $\mathcal{Z}$ such that*

$$\mathcal{Z} = \min_{x\in\mathbb{R}^m} \|Ax-b\|_p = \|Ax_{\text{OPT}}-b\|_p . \tag{1}$$

In this paper, we will use the following ℓ_p *regression coreset* concept.

*Received by the editors July 10, 2007; accepted for publication (in revised form) November 5, 2008; published electronically February 6, 2009.
http://www.siam.org/journals/sicomp/38-5/69650.html
†Yahoo! Research, 701 First Ave., Sunnyvale, CA 94089 (anirban@yahoo-inc.com, ravikumar@yahoo-inc.com).
‡Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180 (drinep@cs.rpi.edu).
§Google Inc., New York, NY 10011 (harb@google.com).
¶Department of Mathematics, Stanford University, Stanford, CA 94305 (mmahoney@cs.stanford.edu).

DEFINITION 2 (ℓ_p regression coreset). *Let $0 < \epsilon < 1$. A* coreset *for Problem 1 is a set of indices $\mathcal{I}$ such that the solution $\hat{x}_{\text{OPT}}$ to $\min_{x\in\mathbb{R}^m} ||\hat{A}x - \hat{b}||_p$, where $\hat{A}$ is composed of those rows of A whose indices are in $\mathcal{I}$ and $\hat{b}$ consists of the corresponding elements of b, satisfies $\|A\hat{x}_{\text{OPT}} - b\|_p \leq (1+\epsilon)\min_x \|Ax - b\|_p$.*

If $n \gg m$, i.e., if there are many more constraints than variables, then (1) is an *overconstrained ℓ_p regression problem*. In this case, there does not in general exist a vector x such that $Ax = b$, and thus $\mathcal{Z} > 0$. Overconstrained regression problems are fundamental in statistical data analysis and have numerous applications in applied mathematics, data mining, and machine learning [17, 10]. Even though convex programming methods can be used to solve the overconstrained regression problem in time $O((mn)^c)$ for $c > 1$, this is prohibitive if n is large.[1] This raises the natural question of developing more efficient algorithms that run in time $O(m^c n)$ for $c > 1$, while possibly relaxing the solution to (1). In particular, can we get a κ-approximation to the ℓ_p regression problem, i.e., a vector $\hat{x}$ such that $\|A\hat{x} - b\|_p \leq \kappa\mathcal{Z}$, where $\kappa > 1$? Note that a coreset of small size would strongly satisfy our requirements and result in an efficiently computed solution that is almost as good as the optimal. Thus, the question becomes: Do coresets exist for the ℓ_p regression problem, and if so, can we compute them efficiently?

Our main result is an efficient two-stage sampling-based approximation algorithm that constructs a coreset and thus achieves a $(1+\epsilon)$-approximation for the ℓ_p regression problem. The first stage of the algorithm is sufficient to obtain a (fixed) constant factor approximation. The second stage of the algorithm carefully uses the output of the first stage to construct a coreset and achieve arbitrary constant factor approximation.

1.1. Our contributions. Summary of results. For simplicity of presentation, we summarize the results for the case of $m = d = \text{rank}(A)$. Let $k = \max\{p/2+1,\ p\}$, and let $\phi(r, d)$ be the time required to solve an ℓ_p regression problem with r constraints and d variables. In the first stage of the algorithm, we compute a set of sampling probabilities $p_1, \ldots, p_n$ in time $O(nd^5 \log n)$, sample $\widehat{r_1} = O(36^p d^{k+1})$ rows of A and the corresponding elements of b according to the p_i's, and solve an ℓ_p regression problem on the (much smaller) sample; we prove this is an 8-approximation algorithm with a running time of $O\left(nd^5 \log n + \phi(\widehat{r_1}, d)\right)$. In the second stage of the algorithm, we use the residual from the first stage to compute a new set of sampling probabilities $q_1, \ldots, q_n$, sample an additional $\widehat{r_2} = O(\widehat{r_1}/\epsilon^2)$ rows of A and the corresponding elements of b according to the q_i's, and solve an ℓ_p regression problem on the (much smaller) sample; we prove this is a $(1+\epsilon)$-approximation algorithm with a total running time of $O\left(nd^5 \log n + \phi(\widehat{r_2}, d)\right)$ (section 4). We also show how to extend our basic algorithm to commonly encountered and more general settings of constrained, generalized, and weighted ℓ_p regression problems (section 5).

We note that the l_p regression problem for $p = 1, 2$ has been studied before. For $p = 1$, Clarkson [11] uses a subgradient-based algorithm to preprocess A and b and then samples the rows of the modified problem; these elegant techniques, however, depend crucially on the linear structure of the l_1 regression problem.[2] Furthermore, this algorithm does not yield coresets. For $p = 2$, Drineas, Mahoney, and Muthukrishnan [13] construct coresets by exploiting the singular value decomposition, a property

[1] For the special case of $p = 2$, vector space methods can solve the regression problem in time $O(m^2 n)$, and if $p = 1$ or ∞, linear programming methods can be used.

[2] Two ingredients of [11] use the linear structure: the subgradient-based preprocessing itself and the counting argument for the concentration bound.

peculiar to the l_2 space. Thus, in order to efficiently compute coresets for the ℓ_p regression problem for all $p \in [1, \infty)$, we need tools that capture the geometry of l_p-norms. In this paper we develop the following two tools that may be of independent interest (section 3).

(1) *Well-conditioned bases.* Informally speaking, if the the columns of matrix U form a well-conditioned basis for a d-dimensional subspace of $\mathbb{R}^n$, then for all $z \in \mathbb{R}^d$, $\|z\|_p$ should be close to $\|Uz\|_p$. We will formalize this by requiring[3] that for all $z \in \mathbb{R}^d$, $\|z\|_q$ multiplicatively approximates $\|Uz\|_p$ by a factor that can depend on d but is *independent* of n (where p and q are dual; i.e., $\frac{1}{q}+\frac{1}{p} = 1$). We show that these bases exist and can be constructed in time $O(nd^5 \log n)$. In fact, our notion of a well-conditioned basis can be interpreted as a computational analogue of the Auerbach and Lewis bases studied in functional analysis [28]. They are also related to the barycentric spanners recently introduced by Awerbuch and R. Kleinberg [5] (section 3.1). J. Kleinberg and Sandler [18] defined the notion of an ℓ_1-independent basis, and our well-conditioned basis can be used to obtain an exponentially better "condition number" than their construction. Further, Clarkson [11] defined the notion of an "ℓ_1-conditioned matrix," and he preprocessed the input matrix to an ℓ_1 regression problem so that it satisfies conditions similar to those satisfied by our bases.

(2) *Subspace-preserving sampling.* We show that sampling rows of A according to information in the rows of a well-conditioned basis of A minimizes the sampling variance, and, consequently, the rank of A is not lost by sampling. This is critical for our relative-error approximation guarantees. The notion of subspace-preserving sampling was used in [13] for $p = 2$, but we abstract and generalize this concept for all $p \in [1, \infty)$.

We note that for $p = 2$, our sampling complexity matches that of [13], which is $O(d^2/\epsilon^2)$; and for $p = 1$, it improves that of [11] from $O(d^{3.5}(\log d)/\epsilon^2)$ to $O(d^{2.5}/\epsilon^2)$.

Overview of our methods. Given an input matrix A, we first construct a well-conditioned basis for A and use that to obtain bounds on a slightly nonstandard notion of a p-norm condition number of a matrix. The use of this particular condition number is crucial since the variance in the subspace-preserving sampling can be upper-bounded in terms of it. An ε-net argument then shows that the first stage sampling gives us an 8-approximation. The next twist is to use the output of the first stage as a feedback to fine-tune the sampling probabilities. This is done so that the "positional information" of b with respect to A is also preserved in addition to the subspace. A more careful use of a different ε-net shows that the second stage sampling achieves a $(1 + \epsilon)$-approximation.

1.2. Related work. As mentioned earlier, in the course of providing a sampling-based approximation algorithm for ℓ_1 regression, Clarkson [11] shows that coresets exist and can be computed efficiently for a *controlled* ℓ_1 regression problem. Clarkson first preprocesses the input matrix A to make it well conditioned with respect to the ℓ_1-norm and then applies a subgradient-descent–based approximation algorithm to guarantee that the ℓ_1-norm of the target vector is conveniently bounded. Coresets of size $O(d^{3.5} \log d/\epsilon^2)$ are thereupon exhibited for this modified regression problem. For the ℓ_2 case, Drineas, Mahoney, and Muthukrishnan [13] designed sampling strategies to preserve the subspace information of A and proved the existence of a coreset of

[3]The requirement could equivalently be in terms of $\|z\|_p$, but the above form yields the tightest dependence on d, since we plan to use Hölder's inequality.

rows of size $O(d^2/\epsilon^2)$, for the *original* ℓ_2 regression problem; this leads to a $(1+\epsilon)$-approximation algorithm. Their algorithm used $O(nd^2)$ time to construct the coreset and solve the ℓ_2 regression problem, which is sufficient time to solve the regression problem without resorting to sampling. In a subsequent work, Sarlós [22] improved the running time for the $(1+\epsilon)$-approximation to $\tilde{O}(nd)$ by using random sketches based on the fast Johnson–Lindenstrauss transform of Ailon and Chazelle [3].

More generally, embedding d-dimensional subspaces of L_p into $\ell_p^{f(d)}$ using coordinate restrictions has been extensively studied [21, 23, 8, 25, 26, 24]. Using well-conditioned bases, one can provide a constructive analogue of Schechtman's existential L_1 embedding result [23] (see also [8]) that any d-dimensional subspace of $L_1[0,1]$ can be embedded in ℓ_1^r with distortion $(1+\epsilon)$ with $r = O(d^2/\epsilon^2)$, albeit with an extra factor of $\sqrt{d}$ in the sampling complexity. Coresets have been analyzed by the computational geometry community as a tool for efficiently approximating various extent measures [1, 2]; see also [16, 6, 14] for applications of coresets in combinatorial optimization. An important difference is that most of the coreset constructions are exponential in dimension and thus applicable only to low-dimensional problems, whereas our coresets are polynomial in dimension and thus applicable to high-dimensional problems.

2. Preliminaries. Given a vector $x \in \mathbb{R}^m$, its *p-norm* is $\|x\|_p = \sum_{i=1}^m (|x_i|^p)^{1/p}$, and the *dual norm* of $\|\cdot\|_p$ is denoted $\|\cdot\|_q$, where $1/p + 1/q = 1$. Given a matrix $A \in \mathbb{R}^{n\times m}$, its *generalized p-norm* is $|||A|||_p = (\sum_{i=1}^n \sum_{j=1}^m |A_{ij}|^p)^{1/p}$. This is a submultiplicative matrix norm that generalizes the Frobenius norm from $p = 2$ to all $p \in [1,\infty)$, but it is not a vector-induced matrix norm. The jth column of A is denoted $A_{\star j}$, and the ith row is denoted $A_{i\star}$. In this notation, $|||A|||_p = (\sum_j \|A_{\star j}\|_p^p)^{1/p} = (\sum_i \|A_{i\star}\|_p^p)^{1/p}$. For $x, x', x'' \in \mathbb{R}^m$, it can be shown using Hölder's inequality that $\|x - x'\|_p^p \le 2^{p-1}(\|x - x''\|_p^p + \|x'' - x'\|_p^p)$.

Two crucial ingredients in our proofs are ε-nets and tail inequalities. A subset $\mathcal{N}(D)$ of a set D equipped with a metric $\|\cdot\|$ is called an *ε-net* in D for some $\varepsilon > 0$ if for every $x \in D$ there is a $y \in \mathcal{N}(D)$ with $\|x - y\| \le \varepsilon$. In order to construct an ε-net for D it is enough to choose $\mathcal{N}(D)$ to be the maximal set of points that are pairwise ε apart. It is well known that the unit ball of a d-dimensional space has an ε-net of size at most $(3/\varepsilon)^d$ [8]. We will use the following version of the Bernstein's inequality.

THEOREM 3 (see [20, 7]). *Let $\{X_i\}_{i=1}^n$ be independent random variables with $E[X_i^2] < \infty$ and $X_i \ge 0$. Set $Y = \sum_i X_i$, and let $\gamma > 0$. Then*

$$\Pr[Y \le E[Y] - \gamma] \le \exp\left(\frac{-\gamma^2}{2\sum_i E[X_i^2]}\right). \tag{2}$$

If $X_i - E[X_i] \le \Delta$ for all i, then with $\sigma_i^2 = E[X_i^2] - E[X_i]^2$ we have

$$\Pr[Y \ge E[Y] + \gamma] \le \exp\left(\frac{-\gamma^2}{2\sum_i \sigma_i^2 + 2\gamma\Delta/3}\right). \tag{3}$$

Finally, throughout this paper, we will use the following sampling matrix formalism to represent our sampling operations. Given a set of n probabilities, $p_i \in (0,1]$ for $i = 1,\dots,n$, let S be an $n \times n$ diagonal sampling matrix such that S_{ii} is set to $1/p_i^{1/p}$ with probability p_i and to zero otherwise. Clearly, premultiplying A or b by S determines whether the ith row of A and the corresponding element of b will be included in the sample, and the expected number of rows/elements selected is $r' = \sum_{i=1}^n p_i$.

(In what follows, we will abuse notation slightly by ignoring zeroed-out rows and regarding S as an $r' \times n$ matrix and thus SA as an $r' \times m$ matrix.) Thus, e.g., sampling constraints from (1) and solving the induced subproblem may be represented as solving

$$\hat{\mathcal{Z}} = \min_{\hat{x} \in \mathbb{R}^m} \|SA\hat{x} - Sb\|_p . \tag{4}$$

A vector $\hat{x}$ is said to be a κ-*approximation* to the ℓ_p regression problem of (1) for $\kappa \geq 1$ if $\|A\hat{x} - b\|_p \leq \kappa \mathcal{Z}$.

3. Main technical ingredients.

3.1. Well-conditioned bases. We introduce the following notion of a "well-conditioned" basis.

DEFINITION 4 (well-conditioned basis). *Let A be an $n \times m$ matrix of rank d, let $p \in [1, \infty)$, and let q be its dual norm. Then an $n \times d$ matrix U is an* (α, β, p)–well-conditioned basis *for the column space of A if the columns of U span the column space of A and* (1) $|||U|||_p \leq \alpha$, *and* (2) *for all $z \in \mathbb{R}^d$, $\|z\|_q \leq \beta \|Uz\|_p$. We will say that U is a* p–well-conditioned basis *for the column space of A if α and β are $d^{O(1)}$, independent of m and n.*

Recall that any orthonormal basis U for span(A) satisfies both $|||U|||_2 = \|U\|_F = \sqrt{d}$ and also $\|z\|_2 = \|Uz\|_2$ for all $z \in \mathbb{R}^d$ and thus is a $(\sqrt{d}, 1, 2)$–well-conditioned basis. Thus, Definition 4 generalizes to an arbitrary p-norm for $p \in [1, \infty)$, the notion that an orthogonal matrix is well conditioned with respect to the 2-norm. Observe that the conditions are slightly different from those of the standard definition of a low-distortion embedding for the following reason. If U is a low distortion embedding, that is, if $\|z\|_p / C \leq \|Uz\|_p \leq \|z\|_p$ for some C, then we can easily see that U is a well-conditioned basis according to the above definition with α and β being $Cd^{O(1)}$. The reverse, however, does not hold. The well-conditioned basis definition above is intended to capture the essence of what is required of a basis for our subspace-sampling strategy to hold. Note also that duality is incorporated into Definition 4 since it relates the q-norm of the vector $z \in \mathbb{R}^d$ to the p-norm of the vector $Uz \in \mathbb{R}^n$, where p and q are dual[4] (i.e., $\frac{1}{q} + \frac{1}{p} = 1$).

The existence and efficient construction of these bases are given by the following.

THEOREM 5. *Let A be an $n \times m$ matrix of rank d, let $p \in [1, \infty)$, and let q be its dual norm. Then there exists an* (α, β, p)–well-conditioned basis U *for the column space of A such that if $p < 2$, then $\alpha = d^{\frac{1}{p} + \frac{1}{2}}$ and $\beta = 1$; if $p = 2$, then $\alpha = d^{\frac{1}{2}}$ and $\beta = 1$; and if $p > 2$, then $\alpha = d^{\frac{1}{p} + \frac{1}{2}}$ and $\beta = d^{\frac{1}{q} - \frac{1}{2}}$. Moreover, U can be computed in $O(nmd + nd^5 \log n)$ time (or in just $O(nmd)$ time if $p = 2$).*

Proof. Let $A = QR$, where Q is any $n \times d$ matrix that is an orthonormal basis for span(A) and R is a $d \times m$ matrix. If $p = 2$, then Q is the desired basis U; from the discussion following Definition 4, $\alpha = \sqrt{d}$ and $\beta = 1$, and computing the matrix U requires $O(nmd)$ time [15]. Otherwise, fix Q and p, and define the norm

$$\|z\|_{Q,p} \triangleq \|Qz\|_p .$$

[4] For $p = 2$, Drineas, Mahoney, and Muthukrishnan used this basis, i.e., an orthonormal matrix, to construct probabilities to sample the original matrix. For $p = 1$, Clarkson used a procedure similar to the one we describe in the proof of Theorem 5 to preprocess A such that the 1-norm of z is a $d\sqrt{d}$ factor away from the 1-norm of Az.

A quick check shows that $\|\cdot\|_{Q,p}$ is indeed a norm. ($\|z\|_{Q,p} = 0$ if and only if $z = 0$ since Q has full column rank; $\|\gamma z\|_{Q,p} = \|\gamma Q z\|_p = |\gamma| \|Qz\|_p = |\gamma| \|z\|_{Q,p}$; and $\|z + z'\|_{Q,p} = \|Q(z + z')\|_p \le \|Qz\|_p + \|Qz'\|_p = \|z\|_{Q,p} + \|z'\|_{Q,p}$.)

Consider the set $C = \{z \in \mathbb{R}^d : \|z\|_{Q,p} \le 1\}$, which is the unit ball of the norm $\|\cdot\|_{Q,p}$. In addition, define the $d \times d$ matrix F such that $\mathcal{E}_{\text{LJ}} = \{z \in \mathbb{R}^d : z^T F z \le 1\}$ is the Löwner–John ellipsoid of C. Since C is symmetric about the origin, $(1/\sqrt{d})\mathcal{E}_{\text{LJ}} \subseteq C \subseteq \mathcal{E}_{\text{LJ}}$; thus, for all $z \in \mathbb{R}^d$,

$$\|z\|_{\text{LJ}} \le \|z\|_{Q,p} \le \sqrt{d}\, \|z\|_{\text{LJ}}, \tag{5}$$

where $\|z\|_{\text{LJ}}^2 = z^T F z$ (see, e.g., [9, pp. 413–414]). Since the matrix F is symmetric positive definite, we can express it as $F = G^T G$, where G is full rank and upper triangular. Since Q is an orthogonal basis for span(A) and G is a $d \times d$ matrix of full rank, it follows that $U = QG^{-1}$ is an $n \times d$ matrix that spans the column space of A. Note that

$$A = QR = QG^{-1}GR = U\tau,$$

where $\tau = GR$. We claim that $U = QG^{-1}$ is the desired p–well-conditioned basis.

To establish this claim, let $z' = Gz$. Thus, $\|z\|_{\text{LJ}}^2 = z^T F z = z^T G^T G z = (Gz)^T Gz = z'^T z' = \|z'\|_2^2$. Furthermore, since G is invertible, $z = G^{-1}z'$, and thus $\|z\|_{Q,p} = \|Qz\|_p = \left\|QG^{-1}z'\right\|_p = \|Uz'\|_p$. By combining these expressions with (5), it follows that for all $z' \in \mathbb{R}^d$,

$$\|z'\|_2 \le \|Uz'\|_p \le \sqrt{d}\, \|z'\|_2. \tag{6}$$

Since $\|U\|_p^p = \sum_j \|U_{\star j}\|_p^p = \sum_j \|Ue_j\|_p^p \le \sum_j d^{\frac{p}{2}} \|e_j\|_2^p = d^{\frac{p}{2}+1}$, where the inequality follows from the upper bound in (6), it follows that $\alpha = d^{\frac{1}{p}+\frac{1}{2}}$. If $p < 2$, then $q > 2$ and $\|z\|_q \le \|z\|_2$ for all $z \in \mathbb{R}^d$; by combining this with (6), it follows that $\beta = 1$. On the other hand, if $p > 2$, then $q < 2$ and $\|z\|_q \le d^{\frac{1}{q}-\frac{1}{2}} \|z\|_2$; by combining this with (6), it follows that $\beta = d^{\frac{1}{q}-\frac{1}{2}}$.

In order to construct U, we need to compute Q and G and then invert G. Our matrix A can be decomposed into QR using the compact QR decomposition in $O(nmd)$ time [15]. The matrix F describing the Löwner–John ellipsoid of the unit ball of $\|\cdot\|_{Q,p}$ can be computed in $O(nd^5 \log n)$ time [19]. Finally, computing G from F takes $O(d^3)$ time, and inverting G takes $O(d^3)$ time. □

It is an open question whether the discontinuity at $p = 2$ in Theorem 5 is inherent in the structure of dual norms, or whether it is due to our inability to compute a better set of well-conditioned bases.

Connection to barycentric spanners. A point set $K = \{K_1, \ldots, K_d\} \subseteq D \subseteq \mathbb{R}^d$ is a *barycentric spanner* for the set D if every $z \in D$ may be expressed as a linear combination of elements of K using coefficients in $[-C, C]$ for $C = 1$. When $C > 1$, K is called a C-approximate barycentric spanner. Barycentric spanners were introduced by Awerbuch and R. Kleinberg in [5]. They showed that if a set is compact, then it has a barycentric spanner. Our proof shows that if A is an $n \times d$ matrix, then $B = \tau^{-1}/\sqrt{d} = R^{-1}G^{-1}/\sqrt{d} \in \mathbb{R}^{d \times d}$ is a $\sqrt{d}$-approximate barycentric spanner for $D = \{z \in \mathbb{R}^d : \|Az\|_p \le 1\}$. To see this, first note that each $B_{\star j}$ belongs to D since $\|AB_{\star j}\|_p = \frac{1}{\sqrt{d}} \|Ue_j\|_p \le \|e_j\|_2 = 1$, where the inequality is obtained

from (6). Moreover, since B spans $\mathbb{R}^d$, we can write any $z \in D$ as $z = B\nu$. Thus, $\nu = B^{-1}z = \sqrt{d}\tau z$. Hence,

$$\|\nu\|_\infty \leq \|\nu\|_2 \leq \|U\nu\|_p = \left\|\sqrt{d}U\tau z\right\|_p = \sqrt{d}\,\|Az\|_p \leq \sqrt{d},$$

where the second inequality is also obtained from (6). This shows that our basis has the added property that every element $z \in D$ can be expressed as a linear combination of elements (or columns) of B using coefficients whose ℓ_2-norm is bounded by $\sqrt{d}$.

Connection to Auerbach bases. An *Auerbach basis* $U = \{U_{\star j}\}_{j=1}^d$ for a d-dimensional normed space $\mathcal{A}$ is a basis such that $\|U_{\star j}\|_p = 1$ for all j and such that whenever $y = \sum_j \nu_j U_{\star j}$ is in the unit ball of $\mathcal{A}$, then $|\nu_j| \leq 1$. The existence of such a basis for every finite dimensional normed space was first proved by Auerbach [4] (see also [12, 27]). It can easily be shown that an Auerbach basis is an (α, β, p)–well-conditioned basis, with $\alpha = d$ and $\beta = 1$ for all p. Further, suppose U is an Auerbach basis for span(A), where A is an $n \times d$ matrix of rank d. Writing $A = U\tau$, it follows that τ^{-1} is an *exact* barycentric spanner for $D = \{z \in \mathbb{R}^d : \|Az\|_p \leq 1\}$. Specifically, each $\tau_{\star j}^{-1} \in D$ since $\|A\tau_{\star j}^{-1}\|_p = \|U_{\star j}\|_p = 1$. Now write $z \in D$ as $z = \tau^{-1}\nu$. Since the vector $y = Az = U\nu$ is in the unit ball of span(A), we have $|\nu_j| \leq 1$ for all $1 \leq j \leq d$. Therefore, computing a barycentric spanner for the compact set D—which is the preimage of the unit ball of span(A)—is equivalent (up to polynomial factors) to computing an Auerbach basis for span(A).

3.2. Subspace-preserving sampling. In the previous subsection (and in the notation of the proof of Theorem 5), we saw that given $p \in [1, \infty)$, any $n \times m$ matrix A of rank d can be decomposed as

$$A = QR = QG^{-1}GR = U\tau,$$

where $U = QG^{-1}$ is a p–well-conditioned basis for span(A) and $\tau = GR$. The significance of a p–well-conditioned basis is that we are able to minimize the variance in our sampling process by randomly sampling *rows* of the matrix A and elements of the vector b according to a probability distribution that depends on norms of the *rows* of the matrix U. This will allow us to preserve the subspace structure of span(A) and thus to achieve relative-error approximation guarantees.

More precisely, given $p \in [1, \infty)$ and any $n \times m$ matrix A of rank d decomposed as $A = U\tau$, where U is an (α, β, p)–well-conditioned basis for span(A), consider any set of sampling probabilities p_i for $i = 1, \ldots, n$ that satisfy

$$p_i \geq \min\left\{1, \frac{\|U_{i\star}\|_p^p}{|||U|||_p^p} r\right\}, \tag{7}$$

where $r = r(\alpha, \beta, p, d, \epsilon)$ to be determined below. Let us randomly sample the ith row of A with probability p_i for all $i = 1, \ldots, n$. Recall that we can construct a diagonal sampling matrix S, where each $S_{ii} = 1/p_i^{1/p}$ with probability p_i and 0 otherwise, in which case we can represent the sampling operation as SA.

The following theorem is our main result regarding this subspace-preserving sampling procedure.

THEOREM 6. *Let A be an $n \times m$ matrix of rank d, $\epsilon \leq 1/7$, and let $p \in [1, \infty)$. Let U be an (α, β, p)–well-conditioned basis for* span(A)*, and let us randomly sample rows of A according to the procedure described above using the probability distribution given*

by (7)*, where* $r \geq 16(2^p+2)(\alpha\beta)^p(d\ln(\frac{12}{\epsilon})+\ln(\frac{2}{\delta}))/(p^2\epsilon^2)$*. Then, with probability* $1-\delta$*, the following holds for all* $x \in \mathbb{R}^m$*:*

$$\left| \|SAx\|_p - \|Ax\|_p \right| \leq \epsilon \|Ax\|_p .$$

Proof. For simplicity of presentation, in this proof we will generally drop the subscript from our matrix and vector p-norms; i.e., unsubscripted norms will be p-norms. Note that it suffices to prove that, for all $x \in \mathbb{R}^m$,

$$(1-\epsilon)^p \|Ax\|^p \leq \|SAx\|^p \leq (1+\epsilon)^p \|Ax\|^p \tag{8}$$

with probability $1-\delta$. To this end, fix a vector $x \in \mathbb{R}^m$, define the random variable $X_i = (S_{ii}|A_{i\star}x|)^p$, and recall that $A_{i\star} = U_{i\star}\tau$ since $A = U\tau$. Clearly, $\sum_{i=1}^n X_i = \|SAx\|^p$. In addition, since $E[X_i] = |A_{i\star}x|^p$, it follows that $\sum_{i=1}^n E[X_i] = \|Ax\|^p$. To bound (8), first note that

$$\sum_{i=1}^{n} (X_i - E[X_i]) = \sum_{i:p_i<1} (X_i - E[X_i]) . \tag{9}$$

Equation (9) follows since, according to the definition of p_i in (7), p_i may equal 1 for some rows, and since these rows are always included in the random sample, $X_i = E[X_i]$ for these rows. To bound the right-hand side of (9), note that for all i such that $p_i < 1$,

$$\begin{aligned} |A_{i\star}x|^p / p_i &\leq \|U_{i\star}\|_p^p \|\tau x\|_q^p / p_i && \text{(by Hölder's inequality)} \\ &\leq |||U|||_p^p \|\tau x\|_q^p / r && \text{(by (7))} \\ &\leq (\alpha\beta)^p \|Ax\|^p / r && \text{(by Definition 4 and Theorem 5).} \end{aligned} \tag{10}$$

From (10) it follows that for each i such that $p_i < 1$,

$$X_i - E[X_i] \leq X_i \leq |A_{i\star}x|^p / p_i \leq (\alpha\beta)^p \|Ax\|^p / r.$$

Thus, we may define $\Delta = (\alpha\beta)^p \|Ax\|^p / r$. In addition, it also follows from (10) that

$$\begin{aligned} \sum_{i:p_i<1} E\left[X_i^2\right] &= \sum_{i:p_i<1} |A_{i\star}x|^p \frac{|A_{i\star}x|^p}{p_i} \\ &\leq \frac{(\alpha\beta)^p \|Ax\|^p}{r} \sum_{i:p_i<1} |A_{i\star}x|^p && \text{(by (10))} \\ &\leq (\alpha\beta)^p \|Ax\|^{2p} / r, \end{aligned}$$

from which it follows that $\sum_{i:p_i<1} \sigma_i^2 = \sum_{i:p_i<1} E\left[X_i^2\right] - (E[X_i])^2 \leq \sum_{i:p_i<1} E\left[X_i^2\right] \leq (\alpha\beta)^p \|Ax\|^{2p} / r$.

To apply the upper tail bound in Theorem 3, define $\gamma = ((1+\epsilon/4)^p - 1)\|Ax\|^p$. It follows that $\gamma^2 \geq (p\epsilon/4)^2 \|Ax\|^{2p}$ and also that

$$\begin{aligned} 2\sum_{i:p_i<1} \sigma_i^2 + 2\gamma\Delta/3 &\leq 2(\alpha\beta)^p \|Ax\|^{2p} / r + 2((1+\epsilon/4)^p - 1)(\alpha\beta)^p \|Ax\|^{2p} / 3r \\ &\leq \left(\frac{2}{3}\left(\frac{5}{4}\right)^p + \frac{4}{3}\right)(\alpha\beta)^p \|Ax\|^{2p} / r \\ &\leq (2^p+2)(\alpha\beta)^p \|Ax\|^{2p} / r, \end{aligned}$$

where the second inequality follows by standard manipulations since $\epsilon \leq 1$ and since $p \geq 1$. Thus, by (3) of Theorem 3, it follows that

$$\begin{aligned}\Pr\left[\|SAx\|^p > \|Ax\|^p + \gamma\right] &= \Pr\left[\sum_{i:p_i<1} X_i > E\left[\sum_{i:p_i<1} X_i\right] + \gamma\right] \\ &\leq \exp\left(\frac{-\gamma^2}{2\sum_{i:p_i<1}\sigma_i^2 + 2\gamma\Delta/3}\right) \\ &\leq \exp\left(\frac{-\epsilon^2 p^2 r}{16(2^p+2)(\alpha\beta)^p}\right).\end{aligned}$$

Similarly, to apply the lower tail bound of (2) of Theorem 3, define $\gamma = (1-(1-\epsilon/4)^p)\|Ax\|^p$. Since $\gamma \geq \epsilon\|Ax\|^p/4$, we can follow a similar line of reasoning to show that

$$\begin{aligned}\Pr\left[\|SAx\|^p < \|Ax\|^p - \gamma\right] &\leq \exp\left(\frac{-\gamma^2}{2\sum_{i:p_i<1}\sigma_i^2}\right) \\ &\leq \exp\left(\frac{-\epsilon^2 r}{32(\alpha\beta)^p}\right).\end{aligned}$$

Choosing $r \geq 16(2^p+2)(\alpha\beta)^p(d\ln(\frac{12}{\epsilon}) + \ln(\frac{2}{\delta}))/(p^2\epsilon^2)$, we get that for every fixed x, the following is true with probability at least $1 - \left(\frac{\epsilon}{12}\right)^d \delta$:

$$(1-\epsilon/4)^p\|Ax\|^p \leq \|SAx\|^p \leq (1+\epsilon/4)^p\|Ax\|^p.$$

Now, consider the ball $B = \{y \in \mathbb{R}^n : y = Ax, \|y\| \leq 1\}$, and consider an ε-net for B, with $\varepsilon = \epsilon/4$. The number of points in the ε-net is $\left(\frac{12}{\epsilon}\right)^d$. Thus, by the union bound, with probability $1-\delta$, (8) holds for all points in the ε-net. Now, to show that with the same probability (8) holds for all points $y \in B$, let $y^* \in B$ be such that $|\|Sy\| - \|y\||$ is maximized, and let $\eta = \sup\{|\|Sy\| - \|y\|| : y \in B\}$. Also, let $y_\varepsilon^* \in B$ be the point in the ε-net that is closest to y^*. By the triangle inequality,

$$\begin{aligned}\eta = |\|Sy^*\| - \|y^*\|| &= |\|Sy_\varepsilon^* + S(y^*-y_\varepsilon^*)\| - \|y_\varepsilon^* + (y^*-y_\varepsilon^*)\|| \\ &\leq |\|Sy_\varepsilon^*\| + \|S(y^*-y_\varepsilon^*)\| - \|y_\varepsilon^*\| + 2\|y^*-y_\varepsilon^*\| - \|y^*-y_\varepsilon^*\|| \\ &\leq |\|Sy_\varepsilon^*\| - \|y_\varepsilon^*\|| + |\|S(y^*-y_\varepsilon^*)\| - \|y^*-y_\varepsilon^*\|| + 2\|y^*-y_\varepsilon^*\| \\ &\leq \epsilon/4\|y_\varepsilon^*\| + \epsilon\eta/4 + \epsilon/2,\end{aligned}$$

where the last inequality follows since $\|y^*-y_\varepsilon^*\| \leq \varepsilon$, $(y^*-y_\varepsilon^*)/\varepsilon \in B$, and

$$|\|S(y^*-y_\varepsilon^*)/\varepsilon\| - \|(y^*-y_\varepsilon^*)/\varepsilon\|| \leq \eta.$$

Therefore, $\eta \leq \epsilon$ since $\|y_\varepsilon^*\| \leq 1$ and since we assume $\epsilon \leq 1/7$. Thus, (8) holds for all points $y \in B$, with probability at least $1-\delta$. Similarly, it holds for any $y \in \mathbb{R}^n$ such that $y = Ax$, since $y/\|y\| \in B$ and since $\|S(y/\|y\|) - y/\|y\|\| \leq \epsilon$ implies that $\|Sy - y\| \leq \epsilon\|y\|$, which completes the proof of the theorem. ∎

Several things should be noted about this result. First, it implies that $\text{rank}(SA) = \text{rank}(A)$, since otherwise we could choose a vector $x \in \text{null}(SA)$ and violate the theorem. In this sense, this theorem generalizes the subspace-preservation result of Lemma 4.1 of [13] to all $p \in [1,\infty)$. Second, regarding sampling complexity: if $p < 2$ the sampling complexity is $O(d^{\frac{p}{2}+2})$, if $p = 2$ it is $O(d^2)$, and if $p > 2$ it is $O(d(d^{\frac{1}{p}+\frac{1}{2}}d^{\frac{1}{q}-\frac{1}{2}})^p) = O(d^{p+1})$. Finally, note that this theorem is analogous to the main result of Schechtman [23], which uses the notion of Auerbach bases.

4. The sampling algorithm.

4.1. Statement of our main algorithm and theorem. Our main sampling algorithm for approximating the solution to the ℓ_p regression problem is presented in Figure 1. The algorithm takes as input an $n \times m$ matrix A of rank d, a vector $b \in \mathbb{R}^n$, and a number $p \in [1, \infty)$. It is a two-stage algorithm that returns as output a vector $\hat{x}_{\text{OPT}} \in \mathbb{R}^m$ (or a vector $\hat{x}_c \in \mathbb{R}^m$ if only the first stage is run). In either case, the output is the solution to the induced ℓ_p regression subproblem constructed on the randomly sampled constraints. Note that the set of constraints r_2 extracted by the second stage of the algorithm is a coreset for the ℓ_p regression problem.

Input: An $n \times m$ matrix A of rank d, a vector $b \in \mathbb{R}^n$, and $p \in [1, \infty)$.

Let $0 < \epsilon < 1/7$, and define $k = \max\{p/2 + 1, p\}$.

- Find a p–well-conditioned basis $U \in \mathbb{R}^{n \times d}$ for span(A) (as in the proof of Theorem 5).
- **Stage 1:** Define $p_i = \min\{1, \frac{\|U_{i\star}\|_p^p}{|||U|||_p^p} r_1\}$, where $r_1 = 16(2^p + 2)d^k \left(d \ln(8 \cdot 12) + \ln(200)\right)$.
  - Generate (implicitly) S where $S_{ii} = 1/p_i^{1/p}$ with probability p_i and 0 otherwise.
  - Let $\hat{x}_c$ be the solution to $\min_{x \in \mathbb{R}^m} \|S(Ax - b)\|_p$.
- **Stage 2:** Let $\hat{\rho} = A\hat{x}_c - b$, and unless $\hat{\rho} = 0$, define $q_i = \min\{1, \max\{p_i, \frac{|\hat{\rho}_i|^p}{\|\hat{\rho}\|_p^p} r_2\}\}$ with $r_2 = \frac{150 \cdot 24^p d^k}{\epsilon^2} \left(d \ln(\frac{280}{\epsilon}) + \ln(200)\right)$.
  - Generate (implicitly, a new) T where $T_{ii} = 1/q_i^{1/p}$ with probability q_i and 0 otherwise.
  - Let $\hat{x}_{\text{OPT}}$ be the solution to $\min_{x \in \mathbb{R}^m} \|T(Ax - b)\|_p$.

Output: $\hat{x}_{\text{OPT}}$ (or $\hat{x}_c$ if only the first stage is run).

FIG. 1. *Sampling algorithm for ℓ_p regression.*

The algorithm first computes a p–well-conditioned basis U for span(A), as described in the proof of Theorem 5. Then, in the first stage, the algorithm uses information from the norms of the rows of U to sample constraints from the input ℓ_p regression problem. In particular, roughly $O(d^{p+1})$ rows of A, and the corresponding elements of b, are randomly sampled according to the probability distribution given by

$$p_i = \min\left\{1, \frac{\|U_{i\star}\|_p^p}{|||U|||_p^p} r_1\right\}, \text{ where } r_1 = 16(2^p + 2)d^k \left(d \ln(8 \cdot 12) + \ln(200)\right), \tag{11}$$

implicitly represented by a diagonal sampling matrix S, where each $S_{ii} = 1/p_i^{1/p}$. For the remainder of the paper, we will use S to denote the sampling matrix for the first-stage sampling probabilities. The algorithm then solves, using any ℓ_p solver of one's choice, the smaller subproblem. If the solution to the induced subproblem is denoted $\hat{x}_c$, then, as we will see in Theorem 7, this is an 8-approximation to the original problem.[5]

In the second stage, the algorithm uses information from the residual of the 8-approximation computed in the first stage to refine the sampling probabilities. Define

[5]For $p = 2$, Drineas, Mahoney, and Muthukrishnan show that this first stage actually leads to a $(1+\epsilon)$-approximation. For $p = 1$, Clarkson develops a subgradient-based algorithm and runs it, after preprocessing the input, on all the input constraints to obtain a constant factor approximation in a stage analogous to our first stage. Here, however, we solve an ℓ_p regression problem on a small subset of the constraints to obtain the constant factor approximation. Moreover, our procedure works for all $p \in [1, \infty)$.

the residual $\hat{\rho} = A\hat{x}_c - b$ (and note that $\|\hat{\rho}\|_p \leq 8\,\mathcal{Z}$). Then, roughly $O(d^{p+1}/\epsilon^2)$ rows of A, and the corresponding elements of b, are randomly sampled according to the probability distribution

(12)
$$q_i = \min\left\{1, \max\left\{p_i, \frac{|\hat{\rho}_i|^p}{\|\hat{\rho}\|_p^p} r_2\right\}\right\}, \text{where } r_2 = \frac{150 \cdot 24^p d^k}{\epsilon^2}\left(d\ln\left(\frac{280}{\epsilon}\right) + \ln(200)\right).$$

As before, this can be represented as a diagonal sampling matrix T, where each $T_{ii} = 1/q_i^{1/p}$ with probability q_i and 0 otherwise. For the remainder of the paper, we will use T to denote the sampling matrix for the second-stage sampling probabilities. Again, the algorithm solves, using any ℓ_p solver of one's choice, the smaller subproblem. If the solution to the induced subproblem at the second stage is denoted $\hat{x}_{\text{OPT}}$, then, as we will see in Theorem 7, this is a $(1+\epsilon)$-approximation to the original problem.[6]

The following is our main theorem for the ℓ_p regression algorithm presented in Figure 1 showing that coresets exist for the ℓ_p regression problem and can be efficiently constructed.

THEOREM 7. *Let A be an $n \times m$ matrix of rank d, let $b \in \mathbb{R}^n$, let $p \in [1, \infty)$, and let $k = \max\{p/2+1, p\}$. Recall that $\epsilon \leq 1/7$, $r_1 = 16(2^p+2)d^k\,(d\ln(8\cdot 12) + \ln(200))$, and $r_2 = \frac{150\cdot 24^p d^k}{\epsilon^2}\left(d\ln(\frac{280}{\epsilon}) + \ln(200)\right)$. Then the following hold.*

- Constant factor approximation. *If only the first stage of the algorithm in Figure 1 is run, then with probability at least 0.6 the solution $\hat{x}_c$ to the sampled problem based on the p_i's of (7) is an 8-approximation to the ℓ_p regression problem.*
- Relative-error approximation. *If both stages of the algorithm are run, then with probability at least 0.5 the solution $\hat{x}_{\text{OPT}}$ to the sampled problem based on the q_i's of (12) is a $(1+\epsilon)$-approximation to the ℓ_p regression problem.*
- Running time. *The ith stage of the algorithm runs in time $O(nmd + nd^5 \log n + \phi(20ir_i, m))$, where $\phi(s,t)$ is the time taken to solve the regression problem $\min_{x\in\mathbb{R}^t} \|A'x - b'\|_p$, where $A' \in \mathbb{R}^{s\times t}$ is of rank d and $b' \in \mathbb{R}^s$.*

Note that since the algorithm of Figure 1 constructs the (α, β, p)–well-conditioned basis U using the procedure in the proof of Theorem 5, our sampling complexity depends on α and β. In particular, it will be $O(d(\alpha\beta)^p)$. Thus, if $p < 2$, our sampling complexity is $O(d \cdot d^{\frac{p}{2}+1}) = O(d^{\frac{p}{2}+2})$; if $p > 2$, it is $O(d(d^{\frac{1}{p}+\frac{1}{2}} d^{\frac{1}{q}-\frac{1}{2}})^p) = O(d^{p+1})$; and (although not explicitly stated, our proof will make it clear that) if $p = 2$, it is $O(d^2)$. Note also that we have stated the claims of the theorem as holding with constant probability, but they can be shown to hold with probability at least $1-\delta$ by using standard amplification techniques.

4.2. Proof for first-stage sampling: Constant factor approximation. To prove the claims of Theorem 7 having to do with the output of the algorithm after the first stage of sampling, we begin with two lemmas. First note that, because of our choice of r_1, we can use the subspace-preserving Theorem 6 with only a constant distortion $\epsilon = \frac{1}{8}$ and $\delta = \frac{1}{100}$; i.e., for all x, we have

(13)
$$\frac{7}{8}\|Ax\|_p \leq \|SAx\|_p \leq \frac{9}{8}\|Ax\|_p$$

[6] The subspace-based sampling probabilities (11) are similar to those used by Drineas, Mahoney, and Muthukrishnan [13], while the residual-based sampling probabilities (12) are similar to those used by Clarkson [11].

with probability at least 0.99. The first lemma below now states that the optimal solution to the original problem provides a small (constant factor) residual when evaluated in the sampled problem.

For simplicity of notation, we again drop the p-subscript from the norm notation, except where it might become confusing.

LEMMA 8. *$\|S(Ax_{\text{OPT}} - b)\| \leq 3\mathcal{Z}$, with probability at least $1 - 1/3^p$.*

Proof. Define $X_i = (S_{ii}|A_{i\star}x_{\text{OPT}} - b_i|)^p$. Thus, $\sum_i X_i = \|S(Ax_{\text{OPT}} - b)\|^p$, and the first moment is $E[\sum_i X_i] = \|Ax_{\text{OPT}} - b\|^p = \mathcal{Z}$. The lemma follows since, by Markov's inequality,

$$\Pr\left[\sum_i X_i > 3^p E\left[\sum_i X_i\right]\right] \leq \frac{1}{3^p};$$

i.e., $\|S(Ax_{\text{OPT}} - b)\|^p > 3^p \|Ax_{\text{OPT}} - b\|^p$ with probability no more than $1/3^p$. ∎

The next lemma states that if the solution to the sampled problem provides a constant factor approximation (when evaluated in the sampled problem), then when this solution is evaluated in the original regression problem we get a (slightly weaker) constant factor approximation.

LEMMA 9. *If $\|S(A\hat{x}_c - b)\| \leq 3\,\mathcal{Z}$, then with probability 0.99, $\|A\hat{x}_c - b\| \leq 8\,\mathcal{Z}$.*

Proof. We will prove the contrapositive: If $\|A\hat{x}_c - b\| > 8\,\mathcal{Z}$, then $\|S(A\hat{x}_c - b)\| > 3\,\mathcal{Z}$. To do so, note that, by Theorem 6 and the choice of r_1, we have that, with probability 0.99,

$$\frac{7}{8}\|Ax\|_p \leq \|SAx\|_p \leq \frac{9}{8}\|Ax\|_p.$$

Using this,

$$\begin{aligned}
\|S(A\hat{x}_c - b)\| &\geq \|SA(\hat{x}_c - x_{\text{OPT}})\| - \|S(Ax_{\text{OPT}} - b)\| && \text{(by the triangle inequality)}\\
&\geq \frac{7}{8}\|A\hat{x}_c - Ax_{\text{OPT}}\| - 3\,\mathcal{Z} && \text{(by Theorem 6 and Lemma 8)}\\
&\geq \frac{7}{8}(\|A\hat{x}_c - b\| - \|Ax_{\text{OPT}} - b\|) - 3\,\mathcal{Z} && \text{(by the triangle inequality)}\\
&> \frac{7}{8}(8\,\mathcal{Z} - \mathcal{Z}) - 3\,\mathcal{Z} && \text{(by the premise } \|A\hat{x}_c - b\| > 8\,\mathcal{Z})\\
&> 3\,\mathcal{Z},
\end{aligned}$$

which establishes the lemma. ∎

Clearly, $\|S(A\hat{x}_c - b)\| \leq \|S(Ax_{\text{OPT}} - b)\|$ (since $\hat{x}_c$ is an optimum for the sampled ℓ_p regression problem). Combining this with Lemmas 8 and 9, it follows that the solution $\hat{x}_c$ to the sampled problem based on the p_i's of (7) satisfies $\|A\hat{x}_c - b\| \leq 8\,\mathcal{Z}$; i.e., $\hat{x}_c$ is an 8-approximation to the original $\mathcal{Z}$.

To conclude the proof of the claims for the first stage of sampling, note that by our choice of r_1, Theorem 6 fails to hold for our first-stage sampling with probability no greater than $1/100$. In addition, the inequality in Lemma 8 fails to hold with probability no greater than $1/3^p$, which is no greater than $1/3$ for all $p \in [1, \infty)$. Finally, let $\widehat{r_1}$ be a random variable representing the number of rows chosen by our sampling scheme, and note that $E[\widehat{r_1}] \leq r_1$. By Markov's inequality, it follows that $\widehat{r_1} > 20r_1$ with probability less than $1/20$. Thus, the first stage of our algorithm fails to give an 8-approximation in the specified running time with a probability bounded by $1/3 + 1/20 + 1/100 < 2/5$.

4.3. Proof for second-stage sampling: Relative-error approximation. The proof of the claims of Theorem 7 having to do with the output of the algorithm after the second stage of sampling will parallel that for the first stage, but it will have several technical complexities that arise since the first triangle inequality approximation in the proof of Lemma 9 is too coarse for relative-error approximation. By our construction, since $q_i \geq p_i$, we have a finer result for subspace preservation. Thus, applying Theorem 6 with $\delta = \frac{1}{100}$, and a constant $\epsilon < \frac{1}{8}$, with probability 0.99, the following holds for all x:

$$(1-\epsilon)\left\|Ax\right\|_p \leq \left\|SAx\right\|_p \leq (1+\epsilon)\left\|Ax\right\|_p. \tag{14}$$

As before, we start with a lemma that states that the optimal solution to the original problem provides a small (now a relative-error) residual when evaluated in the sampled problem. This is the analogue of Lemma 8. An important difference is that the second-stage sampling probabilities significantly enhance the probability of success.

LEMMA 10. $\|T(Ax_{\text{OPT}} - b)\| \leq (1+\epsilon)\mathcal{Z}$ *with probability at least* 0.99.

Proof. Define the random variable $X_i = (T_{ii}|A_{i\star}x_{\text{OPT}} - b_i|)^p$, and recall that $A_{i\star} = U_{i\star}\tau$ since $A = U\tau$. Clearly, $\sum_{i=1}^n X_i = \|T(Ax_{\text{OPT}} - b)\|^p$. In addition, since $E[X_i] = |A_{i\star}x_{\text{OPT}} - b_i|^p$, it follows that $\sum_{i=1}^n E[X_i] = \|Ax_{\text{OPT}} - b\|^p$. We will use (3) of Theorem 3 to provide a bound for $\sum_i (X_i - E[X_i]) = \|T(Ax_{\text{OPT}} - b)\|^p - \|Ax_{\text{OPT}} - b\|^p$.

From the definition of q_i in (12), it follows that for some of the rows, q_i may equal 1 (just as in the proof of Theorem 6). Since $X_i = E[X_i]$ for these rows, $\sum_i (X_i - E[X_i]) = \sum_{i:q_i<1} (X_i - E[X_i])$, and thus we will bound this latter quantity with (3). To do so, we must first provide a bound for $X_i - E[X_i] \leq X_i$ and for $\sum_{i:q_i<1} \sigma_i^2 \leq \sum_i E\left[X_i^2\right]$. To that end, note that

$$\begin{aligned}
|A_{i\star}(x_{\text{OPT}} - \hat{x}_c)| &\leq \left\|U_{i\star}\right\|_p \left\|\tau(x_{\text{OPT}} - \hat{x}_c)\right\|_q \qquad \text{(by Hölders inequality)} \\
&\leq \left\|U_{i\star}\right\|_p \beta \left\|U\tau(x_{\text{OPT}} - \hat{x}_c)\right\|_p \qquad \text{(by Definition 4 and Theorem 5)} \\
&\leq \left\|U_{i\star}\right\|_p \beta \quad (\|Ax_{\text{OPT}} - b\| + \|A\hat{x}_c - b\|) \text{ (by the triangle inequality)} \\
&\leq \left\|U_{i\star}\right\|_p \beta 9\mathcal{Z}, \qquad (15)
\end{aligned}$$

where the final inequality follows from the definition of $\mathcal{Z}$ and the results from the first stage of sampling. Next, note that from the conditions on the probabilities q_i in (12), as well as by Definition 4 and the output of the first stage of sampling, it follows that

$$\frac{|\hat{\rho}_i|^p}{q_i} \leq \frac{\|\hat{\rho}\|^p}{r_2} \leq \frac{8^p\mathcal{Z}^p}{r_2} \quad \text{and} \quad \frac{\|U_{i\star}\|^p}{q_i} \leq \frac{\|\|U\|\|^p}{r_2} \leq \frac{\alpha^p}{r_2} \tag{16}$$

for all i such that $q_i < 1$.

Thus, since $X_i - E[X_i] \leq X_i \leq |A_{i\star}x_{\text{OPT}} - b_i|^p/q_i$, it follows that for all i such that $q_i < 1$,

$$\begin{aligned}
X_i - E[X_i] &\leq \frac{2^{p-1}}{q_i}\left(|A_{i\star}(x_{\text{OPT}} - \hat{x}_c)|^p + |\hat{\rho}_i|^p\right) \qquad (\text{since } \hat{\rho} = A\hat{x}_c - b\ ) \qquad (17) \\
&\leq 2^{p-1}\left(\frac{\|U_{i\star}\|_p^p \beta^p 9^p \mathcal{Z}^p}{q_i} + \frac{|\hat{\rho}_i|^p}{q_i}\right) \qquad \text{(by (15))} \\
&\leq 2^{p-1}\left(\alpha^p\beta^p 9^p\mathcal{Z}^p + 8^p\mathcal{Z}^p\right)/r_2 \qquad \text{(by (16))} \\
&\leq c_p(\alpha\beta)^p\mathcal{Z}^p/r_2, \qquad (18)
\end{aligned}$$

where we set $c_p = 2^{p-1}(9^p + 8^p) \le 18^p$. Thus, we may define $\Delta = c_p(\alpha\beta)^p \mathcal{Z}^p / r_2$. In addition, it follows that

$$\begin{aligned}\sum_{i:q_i<1} E\left[X_i^2\right] &= \sum_{i:q_i<1} |A_{i\star}x_{\text{OPT}} - b_i|^p \frac{|A_{i\star}x_{\text{OPT}} - b_i|^p}{q_i} \\ &\le \Delta \sum_i |A_{i\star}x_{\text{OPT}} - b_i|^p \qquad \text{(by (18))} \\ &\le c_p(\alpha\beta)^p \mathcal{Z}^{2p}/r_2. \qquad (19)\end{aligned}$$

To apply the upper tail bound of (3) of Theorem 3, define $\gamma = ((1+\epsilon)^p - 1)\mathcal{Z}^p$. We have $\gamma \ge p\epsilon\, \mathcal{Z}^p$, and since $\epsilon \le 1/7$, we also have $\gamma \le \left(\left(\frac{8}{7}\right)^p - 1\right) \mathcal{Z}^p$. Hence, by (3) of Theorem 3, it follows that

$$\begin{aligned}\ln \Pr\left[\|T(Ax_{\text{OPT}} - b)\|^p > \|Ax_{\text{OPT}} - b\|^p + \gamma\right] &\le \frac{-\gamma^2}{2\sum_{i:q_i<1} \sigma_i^2 + 2\gamma\Delta/3} \\ &\le \frac{-p^2\epsilon^2 r_2}{\left(2c_p + \frac{2c_p}{3}\left(\left(\frac{8}{7}\right)^p - 1\right)\right)(\alpha\beta)^p} \\ &\le \frac{-p^2\epsilon^2 r_2}{3 \cdot 18^p (\alpha\beta)^p}.\end{aligned}$$

Thus, $\Pr\left[\|T(Ax_{\text{OPT}} - b)\| > (1+\epsilon)\mathcal{Z}\right] \le \exp(\frac{-p^2\epsilon^2 r_2}{3\cdot 18^p(\alpha\beta)^p})$, from which the lemma follows by our choice of r_2. ∎

Next we show that if the solution to the sampled problem provides a relative-error approximation (when evaluated in the sampled problem), then when this solution is evaluated in the original regression problem we get a (slightly weaker) relative-error approximation. We first establish two technical lemmas.

The following lemma says that for all optimal solutions $\hat{x}_{\text{OPT}}$ to the second-stage sampled problem, $A\hat{x}_{\text{OPT}}$ is not too far from $A\hat{x}_c$, where $\hat{x}_c$ is the optimal solution from the first stage, in a p-norm sense. Hence, the lemma will allow us to restrict our calculations in Lemmas 12 and 13 to the ball of radius $12\,\mathcal{Z}$ centered at $A\hat{x}_c$.

LEMMA 11. $\|A\hat{x}_{\text{OPT}} - A\hat{x}_c\| \le 12\,\mathcal{Z}$ *with probability* 0.98.

Proof. With probability 0.98, both the inequalities in Lemma 9 and condition (14) hold true. By two applications of the triangle inequality, it follows that

$$\begin{aligned}\|A\hat{x}_{\text{OPT}} - A\hat{x}_c\| &\le \|A\hat{x}_{\text{OPT}} - Ax_{\text{OPT}}\| + \|Ax_{\text{OPT}} - b\| + \|A\hat{x}_c - b\| \\ &\le \|A\hat{x}_{\text{OPT}} - Ax_{\text{OPT}}\| + 9\mathcal{Z},\end{aligned}$$

where the second inequality follows since $\|A\hat{x}_c - b\| \le 8\,\mathcal{Z}$ from the first stage of sampling and since $\mathcal{Z} = \|Ax_{\text{OPT}} - b\|$. In addition, we have that

$$\begin{aligned}\|Ax_{\text{OPT}} - A\hat{x}_{\text{OPT}}\| &\le \frac{1}{(1-\epsilon)} \|T(A\hat{x}_{\text{OPT}} - Ax_{\text{OPT}})\| \qquad \text{(by Theorem 6)} \\ &\le (1+2\epsilon)\left(\|T(A\hat{x}_{\text{OPT}} - b)\| + \|T(Ax_{\text{OPT}} - b)\|\right) \\ &\qquad\qquad \text{(by the triangle inequality)} \\ &\le 2(1+2\epsilon)\|T(Ax_{\text{OPT}} - b)\| \\ &\le 2(1+2\epsilon)(1+\epsilon)\|Ax_{\text{OPT}} - b\| \qquad \text{(by Lemma 10)}\ ,\end{aligned}$$

where the third inequality follows since $\hat{x}_{\text{OPT}}$ is optimal for the sampled problem. The lemma follows since $\epsilon \le 1/8$. ∎

Thus, if we define the affine ball of radius $12\,\mathcal{Z}$ that is centered at $A\hat{x}_c$ and that lies in span(A),

$$B = \{y \in \mathbb{R}^n : y = Ax, x \in \mathbb{R}^m, \|A\hat{x}_c - y\| \leq 12\,\mathcal{Z}\}, \tag{20}$$

then Lemma 11 states that $A\hat{x}_{\text{OPT}} \in B$ for all optimal solutions $\hat{x}_{\text{OPT}}$ to the sampled problem. Let us consider an ε-net, and call it B_ε with $\varepsilon = \epsilon\,\mathcal{Z}$ for this ball B. Using arguments from [8], since B is a ball in a d-dimensional subspace, the size of the ε-net is $\left(\frac{3\cdot 12\,\mathcal{Z}}{\epsilon\,\mathcal{Z}}\right)^d = \left(\frac{36}{\epsilon}\right)^d$. The next lemma states that for all points in the ε-net, if that point provides a relative-error approximation (when evaluated in the sampled problem), then when this point is evaluated in the original regression problem we get a (slightly weaker) relative-error approximation.

LEMMA 12. *For all points Ax_ε in the ε-net, B_ε, if $\|T(Ax_\varepsilon - b)\| \leq (1+3\epsilon)\mathcal{Z}$, then $\|Ax_\varepsilon - b\| \leq (1+6\epsilon)\mathcal{Z}$ with probability* 0.99.

Proof. Fix a given point $y^*_\varepsilon = Ax^*_\varepsilon \in B_\varepsilon$. We will prove the contrapositive for this point; i.e., we will prove that if $\|Ax^*_\varepsilon - b\| > (1+6\epsilon)\mathcal{Z}$, then $\|T(Ax^*_\varepsilon - b)\| > (1+3\epsilon)\mathcal{Z}$ with probability at least $1 - \frac{1}{100}\left(\frac{\epsilon}{36}\right)^d$. The lemma will then follow from the union bound.

To this end, define the random variable $X_i = (T_{ii}|A_{i\star}x^*_\varepsilon - b_i|)^p$, and recall that $A_{i\star} = U_{i\star}\tau$ since $A = U\tau$. Clearly, $\sum_{i=1}^n X_i = \|T(Ax^*_\varepsilon - b)\|^p$. In addition, since $E[X_i] = |A_{i\star}x^*_\varepsilon - b_i|^p$, it follows that $\sum_{i=1}^n E[X_i] = \|Ax^*_\varepsilon - b\|^p$. We will use (2) of Theorem 3 to provide an upper bound for the probability of the event that $\|T(Ax^*_\varepsilon - b)\|^p \leq \|Ax^*_\varepsilon - b\|^p - \gamma$, where $\gamma = \|Ax^*_\varepsilon - b\|^p - (1+3\epsilon)^p\mathcal{Z}^p$, under the assumption that $\|Ax^*_\varepsilon - b\| > (1+6\epsilon)\mathcal{Z}$.

From the definition of q_i in (12), it follows that for some of the rows, q_i may equal 1 (just as in the proof of Theorem 6). Since $X_i = E[X_i]$ for these rows, $\sum_i (X_i - E[X_i]) = \sum_{i:p_i<1}(X_i - E[X_i])$, and thus we will bound this latter quantity with (2). To do so, we must first provide a bound for $\sum_{i:q_i<1} E\left[X_i^2\right]$. To that end, note that

$$\begin{aligned}
|A_{i\star}(x^*_\varepsilon - \hat{x}_c)| &= |U_{i\star}\tau(x^*_\varepsilon - \hat{x}_c)| \\
&\leq \|U_{i\star}\|_p \, \|\tau(x^*_\varepsilon - \hat{x}_c)\|_q && \text{(by Hölders inequality)} \\
&\leq \|U_{i\star}\|_p \, \beta \, \|U\tau(x^*_\varepsilon - \hat{x}_c)\|_p && \text{(by Definition 4 and Theorem 5)} \\
&\leq \|U_{i\star}\| \, \beta 12\mathcal{Z}, && (22)
\end{aligned} \tag{21}$$

where the final inequality follows from the radius of the high-dimensional ball in which the ε-net resides. From this, we can show that

$$\begin{aligned}
\frac{|A_{i\star}x^*_\varepsilon - b_i|}{q_i} &\leq \frac{2^{p-1}}{q_i}\left(|A_{i\star}x^*_\varepsilon - A_{i\star}\hat{x}_c|^p + |\hat{\rho}_i|^p\right) && \text{(since } \hat{\rho} = A\hat{x}_c - b \text{)} \\
&\leq 2^{p-1}\left(\frac{\|U_{i\star}\|^p \, 12^p\beta^p\mathcal{Z}^p}{q_i} + \frac{|\hat{\rho}_i|^p}{q_i}\right) && \text{(by (22))} \\
&\leq 2^{p-1}\left(\alpha^p 12^p\beta^p\mathcal{Z}^p + 8^p\mathcal{Z}^p\right)/r_2 && \text{(by (16))} \\
&\leq 24^p(\alpha\beta)^p\mathcal{Z}^p/r_2. && (23)
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}\sum_{i:q_i<1} E\left[X_i^2\right] &= \sum_{i:q_i<1} |A_{i\star}x_\varepsilon^* - b_i|^p \frac{|A_{i\star}x_\varepsilon^* - b_i|^p}{q_i} \\ &\le \frac{24^p(\alpha\beta)^p \mathcal{Z}^p}{r_2} \sum_i |A_{i\star}x_\varepsilon^* - b_i|^p \qquad \text{(by (23))} \\ &\le 24^p(\alpha\beta)^p \left\|Ax_\varepsilon^* - b\right\|^{2p} / r_2. \qquad (24)\end{aligned}$$

To apply the lower tail bound of (2) of Theorem 3, define $\gamma = \|Ax_\varepsilon^* - b\|^p - (1+3\epsilon)^p\mathcal{Z}^p$. Thus, by (24) and by (2) of Theorem 3 it follows that

$$\begin{aligned}\ln[\|T(Ax_\varepsilon^* - b)\|^p &\le (1+3\epsilon)^p\mathcal{Z}^p] \\ &\le \frac{-r_2(\|Ax_\varepsilon^* - b\|^p - (1+3\epsilon)^p\mathcal{Z}^p)^2}{24^p(\alpha\beta)^p \|Ax_\varepsilon^* - b\|^{2p}} \\ &\le \frac{-r_2}{24^p(\alpha\beta)^p}\left(1 - \frac{(1+3\epsilon)^p\mathcal{Z}^p}{\|Ax_\varepsilon^* - b\|^p}\right)^2 \\ &< \frac{-r_2}{24^p(\alpha\beta)^p}\left(1 - \frac{(1+3\epsilon)^p\mathcal{Z}^p}{(1+6\epsilon)^p\mathcal{Z}^p}\right)^2 \qquad \text{(by the premise)} \\ &\le \frac{-r_2\epsilon^2}{24^p(\alpha\beta)^p}.\end{aligned}$$

The last line can be justified by the fact that $(1+3\epsilon)/(1+6\epsilon) \le 1-\epsilon$ since $\epsilon \le 1/3$, and that $(1-\epsilon)^p$ is maximized at $p=1$. Since $r_2 \ge 24^p(\alpha\beta)^p(d\ln(\frac{36}{\epsilon}) + \ln(200))/\epsilon^2$, it follows that $\|T(Ax_\varepsilon^* - b)\| \le (1+3\epsilon)\mathcal{Z}$ with probability no greater than $\frac{1}{200}\left(\frac{\epsilon}{36}\right)^d$. Since there are no more than $\left(\frac{36}{\epsilon}\right)^d$ such points in the ε-net, the lemma follows by the union bound. □

Finally, the next lemma states that if the solution to the sampled problem (in the second stage of sampling) provides a relative-error approximation (when evaluated in the sampled problem), then when this solution is evaluated in the original regression problem we get a (slightly weaker) relative-error approximation. This is the analogue of Lemma 9, and its proof will use Lemma 12.

LEMMA 13. *If* $\|T(A\hat{x}_{\text{OPT}} - b)\| \le (1+\epsilon)\mathcal{Z}$, *then* $\|A\hat{x}_{\text{OPT}} - b\| \le (1+7\epsilon)\mathcal{Z}$.

Proof. We will prove the contrapositive: If $\|A\hat{x}_{\text{OPT}} - b\| > (1+7\epsilon)\mathcal{Z}$, then $\|T(A\hat{x}_{\text{OPT}} - b)\| > (1+\epsilon)\mathcal{Z}$. Since $A\hat{x}_{\text{OPT}}$ lies in the ball B defined by (20) and since the ε-net is constructed in this ball, there exists a point $y_\varepsilon = Ax_\varepsilon$ (call it Ax_ε^*), such that $\|A\hat{x}_{\text{OPT}} - Ax_\varepsilon^*\| \le \epsilon\,\mathcal{Z}$. Thus,

$$\begin{aligned}\|Ax_\varepsilon^* - b\| &\ge \|A\hat{x}_{\text{OPT}} - b\| - \|Ax_\varepsilon^* - A\hat{x}_{\text{OPT}}\| \qquad \text{(by the triangle inequality)} \\ &\ge (1+7\epsilon)\mathcal{Z} - \epsilon\mathcal{Z} \qquad \text{(by assumption and the definition of } Ax_\varepsilon^*\text{)} \\ &= (1+6\epsilon)\mathcal{Z}.\end{aligned}$$

Next, since Lemma 12 holds for all points Ax_ε in the ε-net, it follows that

$$\|T(Ax_\varepsilon^* - b)\| > (1+3\epsilon)\mathcal{Z}. \qquad (25)$$

Finally, note that

$$\begin{aligned}
\|T(A\hat{x}_{\text{OPT}} - b)\| &\geq \|T(Ax^*_\varepsilon - b)\| - \|TA(x^*_\varepsilon - \hat{x}_{\text{OPT}})\| && \text{(by the triangle inequality)}\\
&> (1+3\epsilon)\mathcal{Z} - (1+\epsilon)\,\|A(x^*_\varepsilon - \hat{x}_{\text{OPT}})\| && \text{(by (25) and Theorem 6)}\\
&> (1+3\epsilon)\mathcal{Z} - (1+\epsilon)\epsilon\,\mathcal{Z} && \text{(by the definition of } A\hat{x}_\varepsilon)\\
&> (1+\epsilon)\mathcal{Z},
\end{aligned}$$

which establishes the lemma. □

Clearly, $\|T(A\hat{x}_{\text{OPT}} - b)\| \leq \|T(Ax_{\text{OPT}} - b)\|$, since $\hat{x}_{\text{OPT}}$ is an optimum for the sampled ℓ_p regression problem. Combining this with Lemmas 10 and 13, it follows that the solution $\hat{x}_{\text{OPT}}$ to the sampled problem based on the q_i's of (12) satisfies $\|A\hat{x}_{\text{OPT}} - b\| \leq (1+7\epsilon)\,\mathcal{Z}$; i.e., $\hat{x}_{\text{OPT}}$ is a $(1+7\epsilon)$-approximation to the original $\mathcal{Z}$.

To conclude the proof of the claims for the second stage of sampling, note that we can actually replace ϵ by $\epsilon/7$, thus getting the $(1+\epsilon)$-approximation with the corresponding bound on r_2 as in Theorem 7. To bound the failure probability, recall that the first stage failed with probability no greater than 2/5. Note also that by our choice of r_2, Theorem 6 fails to hold for our second-stage sampling with probability no greater than 1/100. In addition, Lemma 10 and Lemma 12 each fails to hold with probability no greater than 2/100 and 1/100, respectively. Finally, let $\widehat{r_2}$ be a random variable representing the number of rows actually chosen by our sampling scheme in the second stage, and note that $E[\widehat{r_2}] \leq 2r_2$. By Markov's inequality, it follows that $\widehat{r_2} > 40r_2$ with probability less than 1/20. Thus, the second stage of our algorithm fails with probability less than $1/20 + 1/100 + 2/100 + 1/100 < 1/10$. By combining both stages, our algorithm fails to give a $(1+\epsilon)$-approximation in the specified running time with a probability bounded from above by $2/5 + 1/10 = 1/2$.

Remark. It has been brought to our attention by an anonymous reviewer that one of the main results of this section can be obtained with a simpler analysis. Via an analysis similar to that of section 4.2, one can show that a relative factor (as opposed to a constant factor) approximation can be obtained in one stage by constructing the sampling probabilities using subspace information from both the data matrix A and the target vector b. In particular, we compute the sampling probabilities from a p–well-conditioned basis for the augmented matrix $[A\ \ b]$ as opposed to only from A. Although it simplifies the analysis, this scheme has the disadvantage that a p–well-conditioned basis needs to be constructed for each target vector b. Using our two-stage algorithm, one need only construct one such basis for A which can subsequently be used to compute probabilities for any target vector b (see, e.g., the extension to *generalized ℓ_p regression* in the next section).

5. Extensions. In this section we outline several immediate extensions of our main algorithmic result.

Constrained ℓ_p regression. Our sampling strategies are transparent to constraints placed on x. In particular, suppose we constrain the output of our algorithm to lie within a convex set $\mathcal{C} \subseteq \mathbb{R}^m$. If there is an algorithm to solve the constrained ℓ_p regression problem $\min_{z\in\mathcal{C}} \|A'x - b'\|$, where $A' \in \mathbb{R}^{s\times m}$ is of rank d and $b' \in \mathbb{R}^s$, in time $\phi(s,m)$, then by modifying our main algorithm in a straightforward manner, we can obtain an algorithm that gives a $(1+\epsilon)$-approximation to the constrained ℓ_p regression problem in time $O(nmd + nd^5 \log n + \phi(40r_2, m))$.

Generalized ℓ_p regression. Our sampling strategies extend to the case of generalized ℓ_p regression: given as input a matrix $A \in \mathbb{R}^{n\times m}$ of rank d, a target

matrix $B \in \mathbb{R}^{n\times p}$, and a real number $p \in [1,\infty)$, find a matrix $X \in \mathbb{R}^{m\times p}$ such that $|||AX - B|||_p$ is minimized. To do so, we generalize our sampling strategies in a straightforward manner. The probabilities p_i for the first stage of sampling are the same as before. Then, if $\hat{X}_c$ is the solution to the first-stage sampled problem, we can define the $n \times p$ matrix $\hat{\rho} = A\hat{X}_c - B$ and define the second-stage sampling probabilities to be $q_i = \min\left(1, \max\{p_i, r_2\|\hat{\rho}_{i\star}\|_p^p/|||\hat{\rho}|||_p^p\}\right)$. Then, we can show that the $\hat{X}_{\text{OPT}}$ computed from the second-stage sampled problem satisfies $|||A\hat{X}_{\text{OPT}} - B|||_p \le (1+\epsilon)\min_{X\in\mathbb{R}^{m\times p}} |||AX - B|||_p$ with probability at least 1/2.

Weighted ℓ_p regression. Our sampling strategies also generalize to the case of ℓ_p regression involving weighted p-norms: if $w_1,\dots,w_m$ are a set of nonnegative weights, then the weighted p-norm of a vector $x \in \mathbb{R}^m$ may be defined as $\|x\|_{p,w} = \left(\sum_{i=1}^m w_i|x_i|^p\right)^{1/p}$, and the weighted analogue of the matrix p-norm $|||\cdot|||_p$ may be defined as $|||U|||_{p,w} = \left(\sum_{j=1}^d \|U_{\star j}\|_{p,w}\right)^{1/p}$. Our sampling scheme proceeds as before. First, we compute a well-conditioned basis U for $\text{span}(A)$ with respect to this weighted p-norm. The sampling probabilities p_i for the first stage of the algorithm are then $p_i = \min(1, r_1 w_i \|U_{i\star}\|_p^p / |||U|||_{p,w}^p)$, and the sampling probabilities q_i for the second stage are $q_i = \min\left(1, \max\{p_i, r_2 w_i|\hat{\rho}_i|^p/\|\hat{\rho}\|_{p,w}^p\}\right)$, where $\hat{\rho}$ is the residual from the first stage.

General sampling probabilities. More generally, consider any sampling probabilities of the form $p_i \ge \min\{1, \max\{\frac{\|U_{i\star}\|_p^p}{|||U|||_p^p}, \frac{|(\rho_{\text{OPT}})_i|^p}{\mathcal{Z}^p}\}r\}$, where $\rho_{\text{OPT}} = Ax_{\text{OPT}} - b$ and $r \ge \frac{36^p d^k}{\epsilon^2}\left(d\ln(\frac{36}{\epsilon}) + \ln(200)\right)$ and where we adopt the convention that $\frac{0}{0} = 0$. Then, by an analysis similar to that presented for our two-stage algorithm, we can show that, by picking $O(36^p d^{p+1}/\epsilon^2)$ rows of A and the corresponding elements of b (in a single stage of sampling) according to these probabilities, the solution $\hat{x}_{\text{OPT}}$ to the sampled ℓ_p regression problem is a $(1+\epsilon)$-approximation to the original problem with probability at least 1/2. (Note that these sampling probabilities, if an equality is used in this expression, depend on the entries of the vector $\rho_{\text{OPT}} = Ax_{\text{OPT}} - b$; in particular, they require the solution of the original problem. This is reminiscent of the results of [13]. Our main two-stage algorithm shows that by solving a problem in the first stage based on coarse probabilities, we can refine our probabilities to approximate these probabilities and thus obtain an $(1+\epsilon)$-approximation to the ℓ_p regression problem more efficiently.)

Acknowledgment. We would like to thank Robert Kleinberg for pointing out several useful references.

REFERENCES

[1] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan, *Approximating extent measures of points*, J. ACM, 51 (2004), pp. 606–635.

[2] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan, *Geometric approximation via coresets*, in Combinatorial and Computational Geometry, J. E. Goodman, J. Pach, and E. Welzl, eds., Math. Sci. Res. Inst. Publ. 52, Cambridge University Press, Cambridge, UK, 2005, pp. 1–30.

[3] N. Ailon and B. Chazelle, *Approximate nearest neighbors and the fast Johnson–Lindenstrauss transform*, in Proceedings of the 38th Annual ACM Symposium on Theory of Computing, ACM, New York, 2006, pp. 557–563.

[4] H. Auerbach, *On the Area of Convex Curves with Conjugate Diameters*, Ph.D. thesis, University of Lwów, Lwów, Poland, 1930 (in Polish).

[5] B. Awerbuch and R. D. Kleinberg, *Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches*, in Proceedings of the 36th Annual ACM Symposium on Theory of Computing, ACM, New York, 2004, pp. 45–53.
[6] M. Bădoiu and K. L. Clarkson, *Smaller core-sets for balls*, in Proceedings of the 14th Annual ACM–SIAM Symposium on Discrete Algorithms, ACM, New York, SIAM, Philadelphia, 2003, pp. 801–802.
[7] S. Bernstein, *Theory of Probability*, Moscow, 1927 (in Russian).
[8] J. Bourgain, J. Lindenstrauss, and V. Milman, *Approximation of zonoids by zonotopes*, Acta Math., 162 (1989), pp. 73–141.
[9] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.
[10] S. Chatterjee, A. S. Hadi, and B. Price, *Regression Analysis by Example*, Wiley Series in Probability and Statistics, Wiley, New York, 2000.
[11] K. L. Clarkson, *Subgradient and sampling algorithms for ℓ_1 regression*, in Proceedings of the 16th Annual ACM–SIAM Symposium on Discrete Algorithms, ACM, New York, SIAM, Philadelphia, 2005, pp. 257–266.
[12] M. Day, *Polygons circumscribed about closed convex curves*, Trans. Amer. Math. Soc., 62 (1947), pp. 315–319.
[13] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, *Sampling algorithms for ℓ_2 regression and applications*, in Proceedings of the 17th Annual ACM–SIAM Symposium on Discrete Algorithms, ACM, New York, SIAM, Philadelphia, 2006, pp. 1127–1136.
[14] D. Feldman, A. Fiat, and M. Sharir, *Coresets for weighted facilities and their applications*, in Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, Washington, D.C., 2006, pp. 315–324.
[15] G. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Studies in Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, 1996.
[16] S. Har-Peled and S. Mazumdar, *On coresets for k-means and k-median clustering*, in Proceedings of the 36th Annual ACM Symposium on Theory of Computing, ACM, New York, 2004, pp. 291–300.
[17] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, New York, 2003.
[18] J. Kleinberg and M. Sandler, *Using mixture models for collaborative filtering*, in Proceedings of the 36th Annual ACM Symposium on Theory of Computing, ACM, New York, 2004, pp. 569–578.
[19] L. Lovasz, *An Algorithmic Theory of Numbers, Graphs, and Convexity*, CBMS-NSF Regoinal Conf. Ser. in Appl. Math. 50, SIAM, Philadelphia, 1986.
[20] A. Maurer, *A bound on the deviation probability for sums of non-negative random variables*, JIPAM. J. Inequal. Pure Appl. Math., 4 (2003).
[21] J. Matousek, *Lectures on Discrete Geometry*, Grad. Texts in Math., Springer-Verlag, New York, 2002.
[22] T. Sarlós, *Improved approximation algorithms for large matrices via random projections*, in Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, Washington, D.C., 2006, pp. 143–152.
[23] G. Schechtman, *More on embedding subspaces of L_p in ℓ_r^n*, Compositio Math., 61 (1987), pp. 159–169.
[24] G. Schechtman and A. Zvavitch, *Embedding subspaces of L_p into ℓ_p^N, $0 < p < 1$*, Math. Nachr., 227 (2001), pp. 133–142.
[25] M. Talagrand, *Embedding subspaces of L_1 into ℓ_1^N*, Proc. Amer. Math. Soc., 108 (1990), pp. 363–369.
[26] M. Talagrand, *Embedding subspaces of L_p into ℓ_p^N*, Oper. Theory Adv. Appl., 77 (1995), pp. 311–325.
[27] A. Taylor, *A geometric theorem and its application to biorthogonal systems*, Bull. Amer. Math. Soc., 53 (1947), pp. 614–616.
[28] P. Wojtaszczyk, *Banach Spaces for Analysts*, Cambridge Stud. Adv. Math. 25, Cambridge University Press, Cambridge, UK, 1991.