Optimal Sketching Bounds for Sparse Linear Regression

Tung Mai[*]   Alexander Munteanu[†]   Cameron Musco[‡]   Anup B. Rao[§]

Chris Schwiegelshohn[¶]   David P. Woodruff[‖]

April 6, 2023

Abstract

We study oblivious sketching for k-sparse linear regression under various loss functions. In particular, we are interested in a distribution over sketching matrices $S \in \mathbb{R}^{m \times n}$ that does not depend on the inputs $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$, such that, given access to SA and Sb, we can recover a k-sparse $\tilde{x} \in \mathbb{R}^d$ with $\|A\tilde{x} - b\|_f \leq (1 + \varepsilon) \min_{k\text{-sparse } x \in \mathbb{R}^d} \|Ax - b\|_f$. Here $\|\cdot\|_f : \mathbb{R}^n \to \mathbb{R}$ is some loss function – such as an ℓ_p norm, or from a broad class of hinge-like loss functions, which includes the logistic and ReLU losses.

We show that for sparse ℓ_2 norm regression, there is a distribution over oblivious sketches with $m = \Theta(k \log(d/k)/\varepsilon^2)$ rows, which is tight up to a constant factor. This extends to ℓ_p loss with an additional additive $O(k \log(k/\varepsilon)/\varepsilon^2)$ term in the upper bound. This establishes a surprising separation from the related sparse recovery problem, which is an important special case of sparse regression, where A is the identity matrix. For this problem, under the ℓ_2 norm, we observe an upper bound of $m = O(k \log(d)/\varepsilon + k \log(k/\varepsilon)/\varepsilon^2)$, showing that sparse recovery is strictly easier to sketch than sparse regression.

For sparse regression under hinge-like loss functions including sparse logistic and sparse ReLU regression, we give the first known sketching bounds that achieve $m = o(d)$ showing that $m = O(\mu^2 k \log(\mu n d/\varepsilon)/\varepsilon^2)$ rows suffice, where μ is a natural complexity parameter needed to obtain relative error bounds for these loss functions. We again show that this dimension is tight, up to lower order terms and the dependence on μ.

Finally, we show that similar sketching bounds can be achieved for LASSO regression, a popular convex relaxation of sparse regression, where one aims to minimize $\|Ax - b\|_2^2 + \lambda \|x\|_1$ over $x \in \mathbb{R}^d$. We show that sketching dimension $m = O(\log(d)/(\lambda \varepsilon)^2)$ suffices and that the dependence on d and λ is tight.

---

[*]Adobe Research, tumai@adobe.com

[†]TU Dortmund University, alexander.munteanu@tu-dortmund.de

[‡]University of Massachusetts Amherst, cmusco@cs.umass.edu

[§]Adobe Research, anuprao@adobe.com

[¶]Aarhus University, cschwiegelshohn@gmail.com

[‖]Carnegie Mellon University, dwoodruf@cs.cmu.edu

Contents

1 Introduction

We study oblivious sketching for k-sparse regression. Given a data matrix $A \in \mathbb{R}^{n \times d}$, and a target vector $b \in \mathbb{R}^n$, linear regression problems aim at finding a vector $x \in \mathbb{R}^d$ such that $Ax \approx b$. The deviation from Ax to b is quantified via a loss function $f(Ax, b)$, where popular examples include the loss in terms of an ℓ_p norm, logistic loss, and ReLU. Sketching techniques for these problems have been widely and successfully applied. Here, one samples a sketching matrix $S \in \mathbb{R}^{m \times n}$ from some distribution and attempts to solve the problem on SA and Sb. If S can be sampled from a distribution that does not depend on either A or b, we call the sketch *oblivious*. Our aim is to minimize the target dimension m, called the *sketching complexity* or *sketching dimension*, while retaining the ability to extract a $(1 + \varepsilon)$-approximate solution $\tilde{x}$ by only using the smaller sketch rather than the original data. A particularly desirable property for the model parameter x is *sparsity*, i.e., where x is restricted to have at most k non-zero elements. Sparse linear regression is an important technique for handling very high dimensional data sets, such as those arising in biostatistics [Mba+21]. It produces a linear model that depends on just a small number of parameters, and thus is more interpretable and can be learned accurately from a relatively small data sample. From a computational point of view, imposing sparsity constraints makes the problem significantly harder. Most unconstrained regression problems are convex and thus we can draw on a wide array of gradient-based methods, but sparse linear regression is NP-hard [Nat95]. Many heuristics and relaxations have been developed to solve the problem in practice, [see Tib96; Mil02; DK08; DK18]. Their running time typically depends at least polynomially on the number n of observations. It is thus natural to study methods for reducing the number of rows in the data so that computations become more efficient and an approximately optimal solution is retained.

While computational savings are immediate from our results for most of the mentioned heuristics and approximation algorithms covered in the related work (see Section 1.1), potential speed-ups need to be verified on a case-wise basis and contrasted with the cost of applying the sketch to the data. Additionally, sketching has other motivations such as saving memory, processing (turnstile) data streams [CW09], and aggregating distributed data [KVW14]. We refer to [Mun23] for a brief introduction. We stress that our bounds refer only to the reduced sketching dimension m, not to other complexity measures.

While the performance of sketching techniques is well understood for *unconstrained* regression problems, we know little about the complexity for sparse regression problems. It is clear that a guarantee for unconstrained regression also applies to sparse regression, but it is not at all clear that these bounds are tight. In particular, the special case of sparse recovery has many celebrated results exploiting sparsity to reconstruct a target signal from a few measurements. Thus we ask the following question:

> *Can oblivious sketching techniques benefit from*
> *model-sparsity for various regression problems?*

To our knowledge, the above question has not been investigated previously. In this paper we answer this question in the affirmative for a large class of loss functions, including the ℓ_p loss $\|Ax - b\|_p = \sqrt[p]{\sum_{i=1}^n |A_i x - b_i|^p}$ for $p \in [1, 2]$, logistic loss $\sum_{i=1}^n \ln(1 + \exp(-b_i \cdot A_i x))$, ReLU $\sum_{i=1}^n \max(0, -b_i \cdot A_i x)$ and further hinge-like loss functions f, where we define $\|Ax - b\|_f = \sum_{i=1}^n f(A_i x - b_i)$. We also investigate the sketching complexity for popular heuristics used to impose sparsity, such as LASSO regression [Tib96], where instead of strictly forcing x to be k-sparse, we use the ℓ_1 penalized objective function $\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2 + \lambda \|x\|_1$ to find a sparse solution. For all of these problems, we obtain optimal or nearly optimal sketching bounds. Crucially, the dependence on d is at least linear for unconstrained regression, but it appears only logarithmically for

sparse regression. Thus our paper makes significant steps towards exploiting the power of sparsity for sketching regression problems.

Let $\Psi_k = \{x \in \mathbb{R}^d \mid \|x\|_0 \leq k\}$ denote the set of k-sparse vectors in $\mathbb{R}^d$. We also study the related problem of k-sparse affine embeddings, where given the sketch and an estimator $E(SA, Sb, x)$ we require that

$$\forall x \in \Psi_k : (1 - \varepsilon)\|Ax - b\|_f \leq E(SA, Sb, x) \leq (1 + \varepsilon)\|Ax - b\|_f.$$

A k-sparse affine embedding is stronger than a sketch for sparse linear regression since it implies the latter. It is a generalization of the restricted isometry property (RIP) studied in the context of compressed sensing. The RIP is a special case where $A = I$ is the identity matrix, $b = 0$, and we seek to preserve the norm of any k-sparse $x \in \Psi_k$ up to $1 \pm \varepsilon$ distortion.

The problem of ℓ_2 sparse recovery can also be seen as a special case of sparse ℓ_2 regression, where $A = I$, and we are given access only to a sketch Sb of the vector b. The goal is to recover a k-sparse vector $\tilde{x}$ that is within $(1 + \varepsilon)$ error to the minimizing x, i.e., it satisfies $\|\tilde{x} - b\|_2 \leq (1 + \varepsilon) \min_{x \in \Psi_k} \|x - b\|_2$. Sketching sparse regression seems very similar to sparse recovery, since similar methods are available that yield similar upper bounds for both problems. However, our studies imply a surprising separation result. Namely, the sparse regression problem is *strictly harder* to sketch than sparse recovery.

One might wonder why we do not consider data dependent sampling algorithms in addition to oblivious linear sketches, since sampling techniques are important tools for approximating regression problems in the non-sparse setting. This is because it can be observed (see Theorems 5 and 6) that sampling does not help in the case of sparse regression; sampling roughly the entire input is necessary to achieve any non-trivial bound.

The crucial advantage of sketching over sampling seems to be the property of *obliviousness* to the subspaces that need to be embedded, which allows us to take a union bound over all k-subsets of coordinates. Sampling algorithms, however, would need a different measure for each possibility. While data dependent importance sampling techniques are widely successful for the unconstrained non-sparse regression problems, they do not give any non-trivial bounds in the sparse setting. This underlines the importance of oblivious sketching techniques in the sparse context.

1.1 Related work

An important special case of the k-sparse regression problem is compressed sensing [CRT06; Don06; Bar+08] where the matrix A is the $n \times n$ identity matrix and one seeks to find a sparse vector x that represents the non-sparse signal b well using a linear sketching (or sensing) matrix. It is well known that a matrix for that problem requires $m = \Omega(k \log(d/k))$ rows [Ba+10], which was improved to $m = \Omega(k \log(d/k)/\varepsilon + k/\varepsilon^2)$ by Price and Woodruff [PW11]. As for upper bounds, it is known that a (Gaussian) RIP matrix can be constructed with $m = O(k \log(d/k))$ rows, which suffices to solve the problem for constant ε. Recently, new and tighter proofs for the Gaussian construction appeared [LXG20] which implicitly yield $m = \Theta(k \log(d/k)/\varepsilon^2)$. However, to solve sparse recovery, neither the Gaussian matrix nor the RIP are necessary. For instance, when sparsity $2k$ is allowed, the problem can be solved, with constant probability for any particular input, using CountSketch [CCF04] with $m = O(k \log(d/k)/\varepsilon)$ rows [PW11].

To our knowledge the generalization of oblivious linear sketching for sparse linear regression has not been investigated before. However, there is a body of work on the column selection problem for sparse linear ℓ_2 regression. The sparse regression problem, i.e., minimizing the regression cost over all k-sparse vectors, is NP-hard [Nat95], and under reasonable complexity-theoretic assumptions it

is even hard to approximate within a significantly stronger bound than the trivial $\|b\|_2$ in quasipolynomial time [FKT15]. In light of this, some authors have identified and characterized instances for which widely applied heuristics have performance guarantees [DK08; DK18; GMS03; Tro06a; Tro06b; TGS06]. Another direction that is closer to the used heuristics, is the online version of the column selection problem. It was shown by [FKT15] that an online algorithm whose iterations run in polynomial time would imply $NP \subseteq BPP$, even if it is allowed to increase the number of columns by an $O(\log d)$ factor. In light of these impossibility results, research has focused on *inefficient* algorithms. The online algorithm of [FKT15] runs in roughly $O(k\binom{d}{k})$ time per iteration. This was improved by Har-Peled, Indyk, and Mahabadi [HIM18] to roughly $\tilde{O}(d^{k-1})$ by giving a data structure that approximates, to within a $(1 + \varepsilon)$ factor, the geometric distance query to the closest $(k-1)$-dimensional flat spanned by the input points, leveraging the geometric interpretation of sparse linear regression. The same reference gives an impossibility result of roughly $\tilde{\Omega}(d^{k/2}/e^k)$ for any multiplicative error approximation by reducing from the k-SUM problem. Assuming the RIP property, [Kal+17] give an efficient online algorithm with guarantees.

The heuristic that is arguably most used in practice for solving sparse regression is the least absolute shrinkage and selection operator (LASSO) by Tibshirani [Tib96]. It introduces a convex relaxation of the ℓ_0 constraint, replacing it by an ℓ_1 constraint. It was shown for well-behaved matrices A, that the LASSO algorithm recovers a (nearly optimal) sparse solution [CRT06; Don06]. The resulting Lagrangian form is $\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2 + \lambda\|x\|_1$. Since the $\lambda\|x\|_1$ regularization term is non-negative, it is immediate that an ℓ_2-subspace embedding is sufficient for preserving the cost of any $x \in \mathbb{R}^d$ within $(1 \pm \varepsilon)$ multiplicative error with $m = \Theta(d/\varepsilon^2)$ rows [Sar06; NN14]. To our knowledge there are no results on sketching this objective with fewer rows by exploiting the sparsity induced by the regularizer.

For dense linear ℓ_p regression there are numerous sketching results. Starting with ℓ_2, [Sar06] showed that $m = O(d/\varepsilon)$ rows suffice to preserve the minimizer up to $1 + \varepsilon$ error. This was complemented by a matching lower bound by [CW09]. Extensions to ℓ_p regression via oblivious ℓ_p subspace embeddings and sampling were given in [Cla05; Das+09; SW11; WZ13; CW15; WW19; LWY21]. Recent works [MOW21; MOW23] gave the first oblivious linear sketches for logistic regression and (implicitly) for the ReLU function. Importance sampling algorithms for those generalized linear regression problems were developed by [Mun+18] and further improved and generalized [MMR21; MOP22; WY23].

2 Our techniques and results

Our results are summarized in Table 1 and cover bounds on the reduced sketching dimension m, not on other complexity measures such as computational cost. Our upper bounds are for affine embeddings, so an algorithm using our sketch enjoys approximation bounds over the entire search space. This straightforwardly implies matching minimization upper bounds. Our lower bounds are for the more challenging minimization variants, except for the hinge-like losses. Our bounds are tight for sparse ℓ_2 regression. The generalization to ℓ_p is tight up to an *additive* $O(k \log(k/\varepsilon)/\varepsilon^2)$ term; specifically, this means our result is tight for reasonably small $k = O(\sqrt{\varepsilon d})$. The ReLU upper bound has another $\tilde{O}(\mu^2)$ factor, in addition to the gap reported for ℓ_1. Here μ is a natural parameter that is needed to parameterize the complexity of compressing data below $\Omega(n)$ for those losses [Mun+18]. This also holds for hinge-like loss functions (including logistic loss) which adds another additive $\log n$, and the complementing lower bound is slightly weaker in the sense that it holds for k-sparse affine embeddings instead of the minimization problem.

Any subsampling approach has a matching $\Theta(n)$ bound for any loss function. The lower bound is

Table 1: Summary of results. Here μ is a data dependent complexity parameter [Mun+18], and λ is the regularization parameter for LASSO, see the main text for details.

Loss function / type	Bound	Reference
ℓ_2	$\Theta(k\log(d/k)/\varepsilon^2)$	Theorems 1 and 7
$\ell_p, p \in [1,2)$		
Lower bound	$\Omega(k\log(d/k)/\varepsilon^2)$	Theorem 2
Upper bound	$O(k\log(d/k)/\varepsilon^2 + k\log(k/\varepsilon)/\varepsilon^2)$	Theorem 8
ReLU & hinge-like		
Lower bound	$\Omega(k\log(d/k)/\varepsilon^2)$	Corollaries 3 and 4
Upper bound for ReLU	$O(\mu^2 k\log(\mu d/\varepsilon)/\varepsilon^2)$	Theorem 9
Upper bound for hinge-like	$O(\mu^2 k\log(\mu nd/\varepsilon)/\varepsilon^2)$	Theorem 10
Any loss via sampling	$\Theta(n)$	Theorems 5 and 6
LASSO with $\lambda\|x\|_1$		
Lower bound	$\Omega(\log(\lambda d)/\lambda^2)$	Theorem 13
Upper bound	$O(\log(d)/(\lambda\varepsilon)^2)$	Theorem 12
ℓ_2 sparse recovery		
Lower bound	$\Omega(k\log(d/k)/\varepsilon + k/\varepsilon^2)$	[PW11]
Upper bound	$O(k\log(d)/\varepsilon + k\log(k/\varepsilon)/\varepsilon^2)$	Theorem 14

given in the minimization setting and for subspace embeddings with different levels of obliviousness to the data. For LASSO with regularization parameter λ, our upper bound is $m = O(\log(d)/(\lambda\varepsilon)^2)$, which we complement by a lower bound of $m = \Omega(\log(\lambda d)/\lambda^2)$. Finally, our upper bound for ℓ_2 sparse recovery leaves only a small additive gap of $O(k\log(k/\varepsilon))$ to the best known lower bound. More interestingly, the bound is sufficient to yield a separation from the *strictly harder* sparse regression problem. In summary, our bounds are tight up to lower order (additive) terms with general parameterizations, and they are tight for reasonably small values of k/ε.

In the remainder we present our main results and the main ideas and technical challenges behind their proofs. The formal proofs and details are moved to their respective appendices for a concise presentation.

2.1 Lower bounds

We obtain our lower bounds by giving a sequence of reductions. Our main lower bound for sparse ℓ_2 regression (from which the further bounds will be derived) is obtained by a reduction from approximate (constant fraction) support recovery for sparse PCA (sPCA).

We note that a related reduction was given in [BPP18], whereas the hardness of the approximate sPCA was covered in [CMW13]. However, the combination of these prior works does not give the desired hardness result for our problem for the following reasons. Their reduction requires d exact sparse regression solves, where *each* column is regressed on all remaining columns. The decision if the column in iteration i is included in the sparse support is done by comparing the projected norm onto the optimal sparse subspace to a certain fixed threshold. The main issue with this is that they get only a k^2 dependence in the reduction from sPCA to sparse regression rather than k, which is necessary in their analysis to separate between columns being in the support or not. It is unclear how to replace those steps with randomized decisions without inflating the dependence on k or other parameters even more. Further, introducing randomization in each iteration would yield

only a lower bound against $1/d$ error probability as opposed to our lower bound against constant probability.

To prove our optimal lower bound against constant error, we give a reduction by solving only one single regression problem. The arguments and support set constructions from prior work [AW09; CMW13] are not directly applicable for the following reasons: we plant the information on the unknown support onto an additional column which we regress onto the standard columns. This additional information is weighted sufficiently high such that it allows us to recover (a constant fraction of) the support.

Crucially, the weight is also sufficiently low, such that the support recovery problem remains hard to solve. But this needs to be reproven using our techniques, which then implies the hardness of the sparse regression problem. We note that full recovery has an $\Omega(k \log(d - k))$ lower bound [AW09], which is larger than our upper bound on the sparse regression problem. We thus rely on a relaxation to approximate constant fraction support recovery using error correcting codes, which requires us to prove a novel lower bound for the *simplified* problem, where the additional planted information is given to the algorithm.

We build on [AW09] as our starting point. The authors studied sPCA for a *spiked covariance model*, where we take measurements from a Gaussian with a covariance matrix $(I_d + vv^T)$ and v is a k-sparse vector. Here, to find the largest eigenvector means that based on vectors drawn from the Gaussian distribution, we need to find the vector v. Since all non-zero entries have the same value, this reduces to finding the k-sparse support of v. The authors show that based on a small number of measurements, this problem is impossible to solve with good probability by information-theoretic arguments. Here we adapt their high level intuition and prove novel hardness bounds for our adapted variant of sPCA.

Sparse ℓ_2 regression We sketch the proof of our main theorem. Several technical details are omitted for brevity of presentation. The detailed technical derivations are in the appendix.

Theorem 1. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$. Suppose $S \in \mathbb{R}^{m \times n}$ is an oblivious linear sketch for k-sparse ℓ_2 regression with an arbitrary estimator $E(SA, Sb, x)$, such that $\tilde{x} \in \operatorname{argmin}_{x \in \Psi_k} E(SA, Sb, x)$ satisfies $\|A\tilde{x} - b\|_2 \leq (1 + \varepsilon) \min_{x \in \Psi_k} \|Ax - b\|_2$ with constant probability. Then $m = \Omega(k \log(d/k)/\varepsilon^2)$.*

Our proof is structured as follows: First, we construct a suitable (hard) distribution over k-sparse supports, which is used to define our input distribution. Second, we prove the impossibility of recovering a constant fraction of the support with a small number of measurements (rows) from the input distribution below an information-theoretic lower bound. Third, we construct an ℓ_2-regression instance for which any $1 + \Theta(\varepsilon)$ approximation derived from an oblivious sketch, paired with an arbitrary estimator, reveals a constant fraction of the support. The hardness result thus turns over to the regression problem.

For constructing the hard distribution (first proof step), we construct an error correcting code $\mathcal{C}$ of roughly size $|\mathcal{C}| = (\frac{d}{k})^k$ consisting of k-sparse binary vectors that overlap in at most ck indexes. The code is also exactly balanced in the following sense. Every single index i appears in exactly the same number of codewords as any other index i', i.e., roughly $(\frac{d}{k})^{k-1} = \frac{k}{d}|\mathcal{C}|$ times, and each pair (i, j), for $i \neq j$, appears exactly the same number of times as any other pair (i', j'), for $i' \neq j'$, i.e., roughly $(\frac{d}{k})^{k-2} = \frac{k^2}{d^2}|\mathcal{C}|$ times. We augment each codeword by another (w.l.o.g. first) coordinate, which is fixed to 1.

We pick a codeword $c \in \mathcal{C}$ uniformly at random and we let our distribution over $n \times d$ inputs be $Z = [b, A] = G(I_d + vv^T)^{1/2}$, where G is a random Gaussian matrix, i.e., each $G_{ij} \overset{i.i.d.}{\sim} N(0, 1)$, and

for all $j \in [d] \setminus \{1\}$ we set $v_j = \frac{\varepsilon}{\sqrt{k}} c_j$ whereas for the first, augmented coordinate, we set $v_1 = c_1 = 1$. This concludes the description of the hard input distribution and the situation before sketching.

Now, since we prove a lower bound against arbitrary estimators, which could for example change the basis and rescale in an arbitrary but appropriate way, we can assume w.l.o.g. that the sketching matrix $S \in \mathbb{R}^{m \times n}$ has orthonormal rows. The sketch thus takes the form

$$SZ = SG(I_d + vv^T)^{1/2} = H(I_d + vv^T)^{1/2},$$

where H is again a Gaussian matrix. SZ thus has the same Gaussian distribution as the input matrix Z but the number of rows is reduced from n to $m < n$.

We move to the second proof step, i.e., the impossibility of recovering a constant fraction of the support with a small number of measurements (rows) from the above input distribution. To this end, we let X be a random variable that has the distribution of one row of our sketch. We use Fano's inequality to bound the failure probability in terms of the size of the code $\mathcal{C}$ and the mutual information that quantifies how much information the rows of the sketch, denoted by X^m, reveal about the unknown support U of $c \in \mathcal{C}$:

$$\mathbb{P}[\texttt{error}] \geq 1 - \frac{I(U; X^m) + \log 2}{\log(|\mathcal{C}| - 1)}.$$

In order to bound the mutual information, we observe by the chain rule for entropy and the maximum entropy property of the Gaussian distribution [CT06] that the mutual information can be bounded in terms of log determinants:

$$I(U, X^m) = H(X^m) - H(X^m | U)$$
$$\leq \frac{m}{2} \log \det \mathbb{E}[xx^T] - \frac{m}{2} \log \det \mathbb{E}[xx^T | U]$$

Leveraging the balanced structure of our code construction, the matrices involved have a nice block structure which we exploit to bound the mutual information by $O(\varepsilon^2 m)$. Further note that the logarithmic code size satisfies $\log(|\mathcal{C}|) = O(k \log(d/k))$. Plugging this back into Fano's inequality we obtain a lower bound $m = \Omega(k \log(d/k)/\varepsilon^2)$ against any constant error probability support recovery algorithm, which concludes the second step of our proof. We refer to the appendix for formal details.

Finally, we describe the reduction to the sparse ℓ_2 regression problem, i.e., the third and last step of our proof. Given $Z = [b, A]$ as described above, we consider the following instance:

$$\min_{x \in \Psi_k} \left\| \begin{bmatrix} M\, M \, \ldots \, M \\ A \end{bmatrix} x - \begin{bmatrix} \sqrt{k}M \\ b \end{bmatrix} \right\|_2,$$

where M is sufficiently large such as to enforce $\sum_i x_i$ to be close to $\sqrt{k}$. In particular this is needed to prevent the trivial solution $x = 0$, and more precisely to impose $\|x\|_2 \geq \frac{1}{\sqrt{k}} \|x\|_1 \geq \frac{1}{\sqrt{k}} \sum_i x_i \approx 1$. Since we prove a lower bound against an arbitrary estimator, we can assume that it is given the first column of any sketching matrix $S' = [s_1, S]$ and the structure of the additional row $r_1 = [M, M, \ldots, \sqrt{k}M]$ including the value of M. This enables the estimator to remove the influence of the tensor product $s_1 \cdot r_1$, such that it can proceed with the estimation on $SZ = S[b, A]$ only. Finally, we show that if we solve the above problem (on the sketch) up to a factor of $1 + \Theta(\varepsilon)$, then the resulting solution $\tilde{x}$ shares a constant fraction of its support with the actual support U of the random codeword $c \in \mathcal{C}$. Due to the error correcting code construction, this uniquely identifies the *full* unknown support U, which concludes the third step of our proof. We hereby obtain our $m = \Omega(k \log(d/k)/\varepsilon^2)$ lower bound against constant error probability oblivious sketching for sparse ℓ_2 regression.

Sparse ℓ_p regression   Our next aim is to extend the $m = \Omega(k \log(d/k)/\varepsilon^2)$ result to the minimization version of ℓ_p norm regression for arbitrary $p \geq 1$.

Theorem 2. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$. Suppose $S \in \mathbb{R}^{m \times n}$ is an oblivious linear sketch for k-sparse ℓ_p regression for any $p \geq 1$ with an estimator $E_p(SA, Sb, x)$, such that the minimizer $\tilde{x} \in \mathrm{argmin}_{x \in \Psi_k} E_p(SA, Sb, x)$ satisfies $\|A\tilde{x} - b\|_p \leq (1 + \varepsilon) \min_{x \in \Psi_k} \|Ax - b\|_p$. Then $m = \Omega(k \log(d/k)/\varepsilon^2)$.*

We start with ℓ_1 and discuss the general case $p \in [1, \infty)$ below. By Dvoretzky's theorem, we can embed ℓ_2 into ℓ_1 with distortion $1 \pm \varepsilon$ using a random Gaussian mapping G. More precisely, $\|Ax - b\|_2 = (1 \pm \varepsilon)\|GAx - Gb\|_1$ for all k-sparse x, where G has $O(n \log(1/\varepsilon)/\varepsilon^2)$ rows and consists of i.i.d. Gaussians. In particular it is an oblivious embedding. Now suppose we had a sketch S for the k-sparse ℓ_1 regression problem. Then we can show that the minimizer $\tilde{x}$ for $\|SGAx - SGb\|_1$ is also a $1 + O(\varepsilon)$ approximation for $\|Ax - b\|_2$. This implies that SG is an oblivious sketch for the ℓ_2-norm problem (with an ℓ_1-norm estimator), and thus S requires $m = \Omega(k \log(d/k)/\varepsilon^2)$ rows.

Similarly, ℓ_2 embeds obliviously and up to $1 \pm \varepsilon$ into ℓ_p for all $p \geq 1$ using a random Gaussian matrix G with a number of rows depending on $O(n)$ for $1 \leq p \leq 2$ and on $n^{O(p)}$ for $p > 2$, [see Mat13, p. 30], but the number of rows of G does not matter in our context since it is reduced by the sketch S. It follows that the $m = \Omega(k \log(d/k)/\varepsilon^2)$ lower bound holds for ℓ_p, for every $p \geq 1$.

Sparse ReLU and hinge-like loss regression   We further extend the lower bound to the ReLU loss function. This time we reduce from the sparse ℓ_1 regression problem by designing an exact embedding of ℓ_1 into $\| \cdot \|_{\mathrm{ReLU}}$.

Corollary 3. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$. Suppose $S \in \mathbb{R}^{m \times n}$ is an oblivious linear sketch for k-sparse ReLU regression with an estimator $E_{\mathrm{ReLU}}(SA, Sb, x)$, such that $\tilde{x} \in \mathrm{argmin}_{x \in \Psi_k} E_{\mathrm{ReLU}}(SA, Sb, x)$ satisfies $\|A\tilde{x} - b\|_{\mathrm{ReLU}} \leq (1 + \varepsilon) \min_{x \in \Psi_k} \|Ax - b\|_{\mathrm{ReLU}}$. Then $m = \Omega(k \log(d/k)/\varepsilon^2)$.*

More precisely, it holds for all $x \in \mathbb{R}^d$ that $\|x\|_1 = \|x\|_{\mathrm{ReLU}} + \| - x\|_{\mathrm{ReLU}}$. A similar argument as in the case of the Gaussian $\ell_2 \to \ell_p$ embedding yields the lower bound for sketching $\| \cdot \|_{\mathrm{ReLU}}$, when we replace G by the embedding matrix $P = [I, -I]^T$, which duplicates and negates the input vector. It follows that SP is an oblivious sketch for ℓ_1 and thus has the same lower bound of $m = \Omega(k \log(d/k)/\varepsilon^2)$.

Finally, in this line of reductions, we deduce a lower bound for a k-sparse affine embedding for the class of hinge-like loss functions [MMR21], which are close to the ReLU function up to an additive deviation a (e.g., logistic loss with $a \leq \ln(2)$), see Definition 1.

Definition 1. *We say $f(\cdot)$ is an (L, a_1, a_2) hinge-like loss function if f is L Lipschitz, $\forall x \geq 0 : f(x) \geq a_2 > 0$, and $\forall x : |f(x) - \mathrm{ReLU}(x)| \leq a_1$.*

Corollary 4. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$. Suppose $S \in \mathbb{R}^{m \times n}$ is an oblivious subspace embedding for some hinge-like loss function f with an estimator $E_f(SA, Sb, x)$, such that we have $\forall x \in \Psi_k : (1 - \varepsilon)\|Ax - b\|_f \leq E_f(SA, Sb, x) \leq (1 + \varepsilon)\|Ax - b\|_f$. Then $m = \Omega(k \log(d/k)/\varepsilon^2)$.*

Since our minimization lower bound for ReLU implies an affine embedding lower bound as a direct consequence, it is sufficient to show that any k-sparse affine embedding for f yields a k-sparse affine embedding for ReLU. The difficulty is that we have an approximate multiplicative embedding of ReLU into f only for the positive part. Hence, the remaining part where ReLU evaluates to zero needs to be taken care of separately, and the minimization lower bound does not follow directly as in our previous arguments.

7

Sampling fails for sparse regression Sampling based algorithms are important tools in sketching for non-sparse regression [DMM06; Das+09; Mun+18; MMR21]. However, they do not give any non-trivial results in the sparse setting.

To corroborate why our upper bounds all build on oblivious linear sketches, rather than sampling, we prove lower bounds for matrices S that subsample (and reweight) rows of the input. Here we show that if A is the identity and b is a random standard basis vector, then any algorithm that has access only to SA and Sb, must fail with probability $> 1/2$ already in the case $k = 1$.

Theorem 5. *Consider any bounded approximation factor $\alpha \geq 1$ and any $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}^{\geq 0}$ which evaluates to 0 on the all zeros vector and to some positive number on any other vector. For any $n > 9$, there is some input matrix $A \in \mathbb{R}^{n \times n}$ and distribution over vectors $b \in \mathbb{R}^n$ such that for any sampling matrix $S \in \mathbb{R}^{m \times n}$ with $m < n/3$, no algorithm that accesses just SA and Sb can output a k-sparse $\tilde{x} \in \Psi_k$ with $\|A\tilde{x} - b\| \leq \alpha \cdot \min_{x \in \Psi_k} \|Ax - b\|$ with probability at least $1/2$ (over the choice of b and any possible randomness in the algorithm).*

The reason is that by the random choice of b, if $m < n/3$ then the sketch contains only rows where $b = 0$, but to obtain a bounded approximation error, it is crucial to retain the non-zero row. By construction, A cannot help to find this coordinate and thus all possibilities that are not in the sample have equal probability of roughly $1/n$ to succeed. A bound of $m = \Omega(n)$ thus follows.

It is crucial for this bound that the algorithm has no access to b when S is being constructed, supporting the fact that it is the property of *obliviousness* that separates sketching from sampling for sparse regression. However, if our aim is to obtain a k-sparse affine embedding via sampling then an $m = \Omega(n)$ bound follows even if the algorithm has full access to the data and even if $k = 1, A = I$ and $b = 0$, which also means that the RIP property cannot be obtained via sampling.

Theorem 6. *Consider any bounded approximation factor $\alpha \geq 1$ and any $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}^{\geq 0}$ which evaluates to 0 on the all zeros vector and to some positive number on any other vector. For any $n > 9$, there is some input matrix $A \in \mathbb{R}^{n \times n}$ such that there is no sampling matrix $S \in \mathbb{R}^{m \times n}$ with $m < n$, which satisfies for all k-sparse $x \in \Psi_k$, $\alpha^{-1}\|Ax\| \leq \|SAx\| \leq \alpha\|Ax\|$.*

We note that in contrast to the impossibility results on sampling, sketching succeeds by relatively simple union bound arguments, e.g., [Bar+08] for the ℓ_2 case, as we will see in the next section.

2.2 Upper Bounds

Sparse ℓ_2 regression Again we start with ℓ_2. The upper bound is similar to the known constructions [CRT06; Don06; Bar+08] of RIP matrices via Johnson-Lindenstrauss embeddings [JL84], i.e., Gaussian matrices. The main difference to these works is that the subspaces formed by any fixed k-sparse solution space need not be orthogonal or aligned with the standard basis vectors.

Theorem 7. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$. There exists a distribution over random matrices $S \in \mathbb{R}^{m \times n}$ with $m = O(k \log(d/k)/\varepsilon^2)$ such that it holds with constant probability that $\forall x \in \Psi_k : (1 - \varepsilon)\|Ax - b\|_2 \leq \|S(Ax - b)\|_2 \leq (1 + \varepsilon)\|Ax - b\|_2$.*

The idea is that there are at most $\binom{d}{k} \leq (ed/k)^k$ different k-sparse supports and each of them corresponds to one choice of k columns of A. Every such choice spans a k-dimensional linear subspace of $\mathbb{R}^n$. By the subspace embedding construction of [Sar06], each subspace formed by one choice of k columns can be handled by embedding the points in a net of size $(3/\varepsilon)^k$ covering the unit ball within the subspace. The remaining vectors can be related to the net points by the triangle inequality and the embedding extends to vectors of arbitrary norm outside the unit sphere by

linearity. Indeed, by a more sophisticated argument (see, e.g., [Woo14]) the net can be constructed with an absolute constant $\varepsilon_0 := 1/2$ instead of ε. So the total number of points to embed up to $(1 \pm \varepsilon)$ distortion is bounded by $|\mathcal{N}| \leq (d/k)^k \cdot c^k$ for an absolute constant c. The embedding can be accomplished via the Johnson-Lindenstrauss lemma followed by a union bound, by using a Gaussian matrix with $m = O(\log(|\mathcal{N}|)/\varepsilon^2) = O(k \log(d/k)/\varepsilon^2)$ rows, matching the lower bound.

Sparse ℓ_p regression   Towards an extension to ℓ_p we focus on $p = 1$ first. Unlike the ℓ_2 case, it is known that a direct embedding of ℓ_1^n into ℓ_1^m is, even for a fixed constant ε, either exponential, i.e., $m \in \Omega(2^{\sqrt{d}})$ or must incur a distortion of $\Omega(d/\mathrm{polylog}(d))$ [WW19; LWY21], and so $(1 \pm \varepsilon)$-approximation seems out of reach. Another alternative is the non-linear median estimator of [Ind06] which is usually avoided since it leads to a non-convex and usually hard optimization problem in the sketch space. However, we note that k-sparse regression is already non-convex and NP-hard. So this is a suitable choice in our setting. The sketching matrix of [Ind06] is a linear sketch $C \in \mathbb{R}^{m \times n}$ whose entries are i.i.d. Cauchy random variables. Those are known to be 1-stable, meaning that their dot product with a vector x is again a Cauchy random variable with scale $\|x\|_1$ so that each row yields a reasonable estimator. However, to achieve concentration, the overall estimator is the median of all row estimators instead of their ℓ_1-norm. This sketch has been combined with a net argument for ℓ_1 in [Bac+16] to obtain a sketching dimension of $O(k \log(k/\varepsilon\delta)/\varepsilon^2)$ for a k-dimensional subspace. However, simply taking a union bound over the $\binom{d}{k}$ choices of k columns would result in $m = O((k^2 \log(d/k) + k \log(k/\varepsilon))/\varepsilon^2)$, which is far from the lower bound. We open up the proof to improve this to $m = O((k \log(d/k) + k \log(k/\varepsilon))/\varepsilon^2)$, which matches our lower bound unless k is relatively large (in the order of $k = \omega(\sqrt{\varepsilon d})$).

Theorem 8. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n, p \in [1, 2)$. There exists a distribution over random matrices $S \in \mathbb{R}^{m \times n}$ with $m = O(k(\log(d/k) + \log(k/(\varepsilon\delta)))/\varepsilon^2)$ such that it holds with probability at least $1 - \delta$ that $\forall x \in \Psi_k : (1 - \varepsilon)\|Ax - b\|_p \leq \|S(Ax - b)\|_{\mathrm{med}} \leq (1 + \varepsilon)\|Ax - b\|_p$, where for arbitrary $y \in \mathbb{R}^d, \|y\|_{\mathrm{med}} := \mathrm{median}\{|y_i| \mid i \in [d]\}$.*

We obtain similar upper bounds by generalizing this result to $\ell_p, p \in [1, 2)$. The sketching matrix is again a linear sketch $C \in \mathbb{R}^{m \times n}$ whose entries are i.i.d. random variables drawn from a p-stable distribution, generalizing the 1-stable Cauchy distribution. Such an extension has been proposed by [Ind06] for sketching single vectors but to our knowledge has never been worked out due to the lack of closed form expressions for the cumulative density function (cdf) and probability density function (pdf), except for $p \in \{1, 2\}$. Here we show how to obtain directly a subspace embedding for all k-sparse vectors. To this end we leverage bounds on the tails of p-stable distributions [BŁM18]. We note that p-stable distributions are leptokurtic. More specifically, they are heavy-tailed with decay $\Pr[|X| > \tau] \leq 1/\tau^p$ except for $p = 2$. Therefore we need to rely on a non-linear quantile estimator to achieve concentration for any $p \in [1, 2)$ to construct and apply the net argument, as in the case $p = 1$ described above. A more intriguing question is how to analyze the cdf of p-stables without closed form expressions. Our solution to this problem comes from the fact that the characteristic function of any p-stable distribution is known in closed form and equals the pdf of a p-generalized normal distribution up to a normalizing constant [Dyt+18]. I.e., it is given by $\phi(t) = \exp(-|\gamma_p t|^p)$, where γ_p is a constant scale parameter that depends on p. Using an inversion theorem of Lévy, we can analyze the cdf and its derivative via an integral involving the characteristic function. As a side result we affirm a conjecture by [Ind06] that a $(1 \pm \varepsilon)$ approximation can be obtained via the median estimator for all $p \in [1, 2]$[1].

---

[1]The reference [Ind06] gives a non-constructive proof showing that there exists some (unknown) quantile, possibly depending on p and ε, that yields a good estimator.

ReLU and hinge-like loss functions   It is well-known [Mun+18] that without any assumptions, these types of functions do not admit relative error sketches with $o(n)$ rows. To address this issue, [Mun+18] introduce a natural notion for the complexity of sketching the matrix A, which we also use to parameterize our results. Intuitively, the parameter $\mu := \sup_{x \in \mathbb{R}^d \setminus \{0\}} \|(Ax)^+\|_1 / \|(Ax)^-\|_1$ is large when there is some x that produces a significant imbalance between the ℓ_1-norm of all positive and the ℓ_1-norm of all negative entries. This can occur, e.g., when the data admits perfect linear separability. However, as [Mun+18] argued, we typically expect μ to be small. These assumptions were recently leveraged to develop the first oblivious linear sketches for logistic regression [MOW21; MOW23], which led to an efficient algorithm for the minimization problem in the non-sparse regime. Another recent contribution of [MMR21] led to small dependencies on d, μ and ε in the regime of sampling and coreset algorithms. However, we already argued that sampling would not work in our setting of sparse regression.

We therefore combine and extend those results to *sketching* a wider class of loss functions and with better dependencies on the approximation parameters, based on the median sketch for ℓ_1. We note that this sacrifices the efficiency of optimization in the sketch space, but as we have argued before, in the context of sparse regression, finding the right support is already a hard problem, which motivates us to focus on the best possible parameterization.

Theorem 9. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$. There exists an oblivious sketch S with $O(\frac{\mu^2 k}{\varepsilon^2} \log(\frac{\mu d}{\varepsilon \delta}))$ rows and an estimator $g_{\mathrm{ReLU}}(SA, Sb, x)$, such that with probability at least $1 - \delta$, we have $\forall x \in \Psi_k$: $(1 - \varepsilon)\|Ax - b\|_{\mathrm{ReLU}} \leq g_{\mathrm{ReLU}}(SA, Sb, x) \leq (1 + \varepsilon)\|Ax - b\|_{\mathrm{ReLU}}$.*

Our k-sparse affine embedding sketch leverages the fact that $\mathrm{ReLU}(x) = (\sum_i x_i + \|x\|_1)/2$, since the negative entries are contained negatively in the sum and again positively in the norm, so they cancel. The positive values are positive in both parts and thus counted twice, so dividing by 2 yields the exact value of $\mathrm{ReLU}(x)$. The sum of entries can be sketched exactly using only one row vector and the ℓ_1 norm is sketched via the (1-stable) Cauchy sketch with median estimator, as detailed above in the previous paragraph. Now, the error of this estimate is $\varepsilon\|x\|_1$ but by the μ-complexity assumption, the ℓ_1 norm is within roughly a μ-factor of the positive entries, so folding μ into ε yields an error of $\varepsilon \mathrm{ReLU}(x)$. We extend this result to an even richer class of hinge-like loss functions, including logistic regression. Those functions are additively close to the ReLU function, see Definition 1. The logistic regression loss $\zeta(x) = \ln(1 + \exp(x))$, for instance, has asymptotes equal to ReLU in the limit of $\pm\infty$. However, close to zero, the two functions differ more significantly, attaining a bounded maximum deviation of $\zeta(0) - \mathrm{ReLU}(0) = \ln(2)$.

Theorem 10. *Let $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$, and let f be an (L, a_1, a_2) hinge-like loss function. There exists an oblivious sketch S with $m = O(\frac{c^{10} \mu^2 k}{\varepsilon^2} \log(\frac{cn\mu d}{\varepsilon \delta}))$ rows, where $c = \max(1, L, a_1, 1/a_2)$, and an estimator $g_f(SA, Sb, x)$, such that, with probability at least $1 - \delta$, we have $\forall x \in \Psi_k$: $(1 - \varepsilon)\|Ax - b\|_f \leq g_f(SA, Sb, x) \leq (1 + \varepsilon)\|Ax - b\|_f$.*

The idea is now to split the loss function into two components $\sum_i(\zeta(x_i) - \mathrm{ReLU}(x_i)) + \sum_i \mathrm{ReLU}(x_i)$. The ReLU function can be dealt with as described in the paragraph above and the remainder is a sum over bounded terms. This again enables us to achieve concentration and union bound over the net of k-sparse vectors up to an additive error of roughly $\varepsilon \frac{n}{\mu}$. This error can be charged by a complementing lower bound which follows from leveraging the μ-complexity assumption, akin to [MMR21; MOW21], and finally yields a relative error guarantee.

For completeness, we have the following simple result that yields a connection between minimizing in the sketch space of a k-sparse affine embedding, as in all upper bounds above, and the original minimization problem.

Corollary 11. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$. Let $\| \cdot \| : \mathbb{R}^n \to \mathbb{R}^{\geq 0}$ be any loss function. Let S be an oblivious linear sketch of $[A, b]$, and let $E(SA, Sb, x)$ be an estimator that satisfies $\forall x \in \Psi_k$: $(1 - \varepsilon)\|Ax - b\| \leq E(SA, Sb, x) \leq (1 + \varepsilon)\|Ax - b\|$. Then $\tilde{x} \in \operatorname{argmin}_{x \in \Psi_k} E(SA, Sb, x)$ satisfies $\|A\tilde{x} - b\| \leq (1 + O(\varepsilon)) \min_{x \in \Psi_k} \|Ax - b\|$.*

2.3 LASSO Regression

LASSO regression [Tib96] is a convex relaxation of k-sparse ℓ_2 regression and enjoys large popularity as a heuristic for inducing sparsity and feature selection. LASSO regression is a special subject of our investigation, since here we do not assume that any solution is k-sparse. In this case, a subspace embedding for the dense ℓ_2 problem is possible for sketching down to $m = \Theta(d/\varepsilon^2)$ dimensions [Sar06; NN14]. Also in cases where the regularization parameter is very small and thus the norm of the solution is actually unconstrained, the problem becomes equivalent to least squares regression, in which case $\Theta(d/\varepsilon)$ is necessary and sufficient [Sar06; CW09]. Usually, however, the ℓ_1 regularization is imposed to yield a sparse minimizer, for which we can again hope to be able to take advantage of the induced sparsity, parameterized by the value of the regularization parameter λ such as to reduce to $\operatorname{poly}(1/\lambda, \log d)$ rows. Here we give an upper bound for sketching that depends on an ℓ_1 regularization parameter λ.

Theorem 12. *Consider $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$, and $\lambda \in (0, 1)$. Assume that $\|A\|_2 \leq 1$ and $\|b\|_2 \leq 1$, If $S \in \mathbb{R}^{m \times n}$ is a random Gaussian matrix (i.e., each entry is sampled i.i.d. from $N(0, 1/m)$) then for any $\varepsilon, \delta \in (0, 1)$ and $m = O(\frac{\log d/\delta}{\lambda^2 \cdot \varepsilon^2})$, with probability at least $1 - \delta$, if $\tilde{x} = \operatorname{argmin}_{x \in \mathbb{R}^d} \|SAx - Sb\|_2^2 + \lambda\|x\|_1$ then $\|A\tilde{x} - b\|_2^2 + \lambda\|\tilde{x}\|_1 \leq (1 + \varepsilon) \cdot \min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2 + \lambda\|x\|_1$.*

Observe that our constraints on $\|A\|_2, \|b\|_2$ are necessary, since LASSO regression is *not* scale invariant. The result follows by first showing that the optimizer must have a bounded norm in terms of the optimal objective value $\|x\|_1 \leq \frac{2 \cdot OPT}{\lambda}$. This allows us to focus on the set $\mathcal{T} = \{y = Ax - b : \|x\|_1 \leq \frac{2 \cdot OPT}{\lambda}\}$, which by the bounded norm, can be expressed as the convex hull of $2d + 1$ points in the unit ball. For this set, we can use an embedding result of [NN19] to obtain an additive error of $\varepsilon\lambda$ for all vectors in the set $\mathcal{T}$, which allows us to relate the sketching error to $O(\varepsilon \cdot OPT)$, and which finally yields our $1 + \varepsilon$ relative error approximation result.

We complement the upper bound by the following lower bound that matches the dependence on d and λ. The proof builds on our new techniques developed for sparse ℓ_2 regression. The condition of bounded norm inputs A, b, however, does not allow us to plant the additional row gadget. Thus, we need to choose a smaller λ by a factor of ε, which unfortunately cancels the ε dependence in the previous ℓ_2 lower bound. Still, our result shows that $\log(\lambda d)/\lambda^2$ rows are necessary for any sketch with an estimator that allows to solve LASSO to within a $1 + \varepsilon$ approximation.

Theorem 13. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$ with bounded $\|A\|_2 \leq 1$ and $\|b\|_2 \leq 1$, and let $\lambda \in (0, 1)$. Suppose $S \in \mathbb{R}^{m \times n}$ is an oblivious linear sketch for LASSO regression with an estimator $E_\lambda(SA, Sb, x)$, such that with constant probability $\tilde{x} \in \operatorname{argmin}_{x \in \Psi_k} E_\lambda(SA, Sb, x)$ satisfies $\|A\tilde{x} - b\|_2^2 + \lambda\|\tilde{x}\|_1 \leq (1 + \varepsilon) \cdot \min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2 + \lambda\|x\|_1$. Then $m = \Omega(\frac{\log(\lambda d)}{\lambda^2})$.*

2.4 Separation of Sparse Recovery from Sparse Regression

Here we give an upper bound of $m = O(k \log(d)/\varepsilon + k \log(k/\varepsilon))/\varepsilon^2)$ that gets very close to the lower bound $m = \Omega(k \log(d/k)/\varepsilon + k/\varepsilon^2)$ of [PW11]. Surprisingly, this provides a separation between the k-sparse recovery problem and the k-sparse regression problem. Combined with our main $\Omega(k \log(d/k)/\varepsilon^2)$ lower bound, it shows that k-sparse regression is *strictly harder* to sketch than

sparse recovery. To obtain the new upper bound, the issue is that we need to figure out which subset to use, but cannot afford to estimate every subset's cost, even though for each subset we can get a good estimate. To cope with this problem we run in parallel a two-stage estimation procedure: one CountSketch that gives a rough estimate of the entries, which yields a superset I of the k coordinates of interest, i.e., $I \subseteq [d]$ of size $|I| = O(k/\varepsilon)$; and another CountSketch that has a higher precision, but is required only to recover estimates for the relatively small number of k-subsets of the set I. Now, if we output the vector that is supported on the top k entries in I together with their estimates of the entries obtained from the high precision sketch, this gives the desired $1 + \varepsilon$ approximation for k-sparse recovery with fewer rows than necessary to solve the k-sparse regression problem.

Theorem 14. *On input $x \in \mathbb{R}^d$, the above sparse recovery scheme uses $O(k \log(k/\varepsilon)/\varepsilon^2 + k \log(d)/\varepsilon)$ measurements and, with probability at least $1 - 1/\operatorname{poly}(d) - 1/\operatorname{poly}(k/\varepsilon)$, returns a k-sparse vector $\hat{x} \in \Psi_k$ satisfying $\|x - \hat{x}\|_2^2 \le (1 + \varepsilon) \min_{x_k \in \Psi_k} \|x - x_k\|_2^2$.*

3 Conclusion

In this paper we study the complexity of oblivious linear sketching for sparse regression problems under various regression loss functions such as ℓ_p regression, logistic regression, ReLU loss, and hinge-like loss functions. Our results are *essentially*[2] tight bounds of $\Theta(k \log(d/k)/\varepsilon^2)$ for all those problems. We further study the sketching complexity of LASSO, a popular convex relaxation often used as a heuristic for solving sparse linear regression. We give the first bound of $O(\log(d)/(\lambda\varepsilon)^2)$ going below the linear dependence on d, where λ is the regularization parameter. Furthermore we provide a separation result from the sparse recovery problem studied in compressed sensing. Surprisingly, we find that the sparse regression problem requires $m = \Omega(k \log(d/k)/\varepsilon^2)$ and is thus strictly harder to sketch than sparse recovery, for which we show a new $m = O(k \log(d)/\varepsilon + k \log(k/\varepsilon)/\varepsilon^2)$ upper bound. We also show that while data dependent importance sampling techniques are widely successful for the unconstrained non-sparse regression problems, they do not give any non-trivial bounds in the sparse setting. This underlines the importance of oblivious sketching techniques in the sparse context. For future directions we aim at closing remaining gaps, especially for hinge-like loss functions. It will also be an interesting avenue to develop more scalable and faster heuristics by incorporating our sketching techniques and evaluate their performance in practice. Finally, since our sketches are optimized for a smallest possible target dimension, it will be interesting to study the trade-off between the speed of applying them to data and an increase in their target dimension.

Acknowledgements

---

[2]up to minor polylogarithmic terms and problem specific parameters such as μ; see Table 1 for the exact terms.

References

[Aue30]   Herman Auerbach. "On the area of convex curves with conjugate diameters". PhD thesis. University of Lwów, 1930.

[AW09]   Arash A. Amini and Martin J. Wainwright. "High-dimensional analysis of semidefinite relaxations for sparse principal components". In: *The Annals of Statistics* 37.5B (2009), pp. 2877–2921. DOI: 10.1214/08-AOS664. URL: https://doi.org/10.1214/08-AOS664.

[Ba+10]   Khanh Do Ba et al. "Lower Bounds for Sparse Recovery". In: *Proc. of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2010, pp. 1190–1197.

[Bac+16]   Arturs Backurs et al. "Nearly-optimal bounds for sparse recovery in generic norms, with applications to k-median sketching". In: *Proc. of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2016, pp. 318–337.

[Bar+08]   Richard Baraniuk et al. "A simple proof of the restricted isometry property for random matrices". In: *Constructive Approximation* 28.3 (2008), pp. 253–263.

[BHW05]   Szymon Borak, Wolfgang Härdle, and Rafał Weron. "Stable distributions". In: *Statistical tools for finance and insurance*. Springer, 2005, pp. 21–44.

[BŁM18]   W. Bednorz, R. Łochowski, and R. Martynek. "On tails of symmetric and totally asymmetric α-stable distributions". In: *arXiv:Probability* (2018). https://arxiv.org/abs/1802.00612.

[BPP18]   Guy Bresler, Sung Min Park, and Madalina Persu. "Sparse PCA from Sparse Linear Regression". In: *Advances in Neural Information Processing Systems 31, (NeurIPS)*. 2018, pp. 10965–10975.

[CCF04]   Moses Charikar, Kevin C. Chen, and Martin Farach-Colton. "Finding frequent items in data streams". In: *Theor. Comput. Sci.* 312.1 (2004), pp. 3–15.

[Cla05]   Kenneth L Clarkson. "Subgradient and sampling algorithms for ℓ_1 regression". In: *Proc. of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, (SODA)*. 2005, pp. 257–266.

[CMW13]   T. Tony Cai, Zongming Ma, and Yihong Wu. "Sparse PCA: Optimal rates and adaptive estimation". In: *The Annals of Statistics* 41.6 (2013), pp. 3074–3110. URL: https://doi.org/10.1214/13-AOS1178.

[CP19]   Xue Chen and Eric Price. "Active regression via linear-sample sparsification". In: *Proc. of the 32nd Annual Conference on Computational Learning Theory (COLT)*. 2019.

[CRT06]   Emmanuel J Candes, Justin K Romberg, and Terence Tao. "Stable signal recovery from incomplete and inaccurate measurements". In: *Communications on Pure and Applied Mathematics* 59.8 (2006), pp. 1207–1223.

[CT06]   Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006. ISBN: 0471241954.

[CW09]   Kenneth L. Clarkson and David P. Woodruff. "Numerical linear algebra in the streaming model". In: *Proc. of the 41st Annual ACM Symposium on Theory of Computing (STOC)*. 2009, pp. 205–214.

[CW15] Kenneth L. Clarkson and David P. Woodruff. "Sketching for M-Estimators: A Unified Approach to Robust Regression". In: *Proc. of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2015, pp. 921–939.

[CW79] J Lawrence Carter and Mark N Wegman. "Universal classes of hash functions". In: *Journal of computer and system sciences* 18.2 (1979), pp. 143–154.

[Das+09] Anirban Dasgupta et al. "Sampling Algorithms and Coresets for ℓ_p Regression". In: *SIAM J. Comput.* 38.5 (2009), pp. 2060–2078. DOI: 10.1137/070696507.

[DK08] Abhimanyu Das and David Kempe. "Algorithms for subset selection in linear regression". In: *Proc.of the 40th Annual ACM Symposium on Theory of Computing, (STOC)*. 2008, pp. 45–54. DOI: 10.1145/1374376.1374384. URL: https://doi.org/10.1145/1374376.1374384.

[DK18] Abhimanyu Das and David Kempe. "Approximate Submodularity and its Applications: Subset Selection, Sparse Approximation and Dictionary Selection". In: *J. Mach. Learn. Res.* 19 (2018), 3:1–3:34. URL: http://jmlr.org/papers/v19/16-534.html.

[DMM06] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. "Sampling algorithms for ℓ_2 regression and applications". In: *Proc. of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2006, pp. 1127–1136.

[Don06] David L. Donoho. "Compressed sensing". In: *IEEE Transactions on information theory* 52.4 (2006), pp. 1289–1306.

[Dyt+18] Alex Dytso et al. "Analytical properties of generalized Gaussian distributions". In: *Journal of Statistical Distributions and Applications* 5 (Dec. 2018), pp. 1–40.

[FKT15] Dean P. Foster, Howard J. Karloff, and Justin Thaler. "Variable Selection is Hard". In: *Proc. of the 28th Conference on Learning Theory (COLT)*. 2015, pp. 696–709.

[GMS03] Anna C. Gilbert, S. Muthukrishnan, and Martin Strauss. "Approximation of functions over redundant dictionaries using coherence". In: *Proc. of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, (SODA)*. 2003, pp. 243–252. URL: http://dl.acm.org/citation.cfm?id=644108.644149.

[HIM18] Sariel Har-Peled, Piotr Indyk, and Sepideh Mahabadi. "Approximate Sparse Linear Regression". In: *45th International Colloquium on Automata, Languages, and Programming (ICALP)*. Vol. 107. 2018, 77:1–77:14.

[Ind06] Piotr Indyk. "Stable distributions, pseudorandom generators, embeddings, and data stream computation". In: *Journal of the ACM* 53.3 (2006), pp. 307–323.

[JL84] William B. Johnson and Joram Lindenstrauss. "Extensions of Lipschitz mappings into a Hilbert space". In: *Contemporary Mathematics* 26.189-206 (1984), pp. 1–1.

[Kal+17] Satyen Kale et al. "Adaptive Feature Selection: Computationally Efficient Online Sparse Linear Regression under RIP". In: *Proc. of the 34th International Conference on Machine Learning, (ICML)*. 2017, pp. 1780–1788. URL: http://proceedings.mlr.press/v70/kale17a.html.

[KVW14] Ravi Kannan, Santosh Vempala, and David P. Woodruff. "Principal Component Analysis and Higher Correlations for Distributed Data". In: *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*. 2014, pp. 1040–1057.

[Lév25] Paul Lévy. *Calcul des probabilités*. Gauthier-Villars, 1925.

[LWY21]   Yi Li, David P. Woodruff, and Taisuke Yasuda. "Exponentially Improved Dimensionality Reduction for ℓ_1: Subspace Embeddings and Independence Testing". In: *Conference on Learning Theory (COLT)*. 2021, pp. 3111–3195.

[LXG20]   Gen Li, Xingyu Xu, and Yuantao Gu. "Lower Bound for RIP Constants and Concentration of Sum of Top Order Statistics". In: *IEEE Trans. Signal Process.* 68 (2020), pp. 3169–3178.

[Mas77]   Pesi Rustom Masani. "Generalizations of P. Lévy's inversion theorem". In: *Journal of Multivariate Analysis* 7 (1977), pp. 292–335.

[Mat13]   Jiří Matoušek. *Lecture notes on metric embeddings.* https://kam.mff.cuni.cz/~matousek/ba-a4.pdf. Accessed: 2021-10-20. 2013.

[Mba+21] Joelle Mbatchou et al. "Computationally efficient whole-genome regression for quantitative and binary traits". In: *Nature Genetics* 53 (2021), pp. 1097–1103. DOI: 10.1038/s41588-021-00870-7. URL: https://doi.org/10.1038/s41588-021-00870-7.

[Mil02]   A. Miller. *Subset Selection in Regression.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2002. ISBN: 9781420035933. URL: https://books.google.de/books?id=7p59iir822sC.

[MMR21]   Tung Mai, Cameron Musco, and Anup Rao. "Coresets for Classification - Simplified and Strengthened". In: *Advances in Neural Information Processing Systems 34 (NeurIPS)*. 2021, pp. 11643–11654.

[MOP22]   Alexander Munteanu, Simon Omlor, and Christian Peters. "p-Generalized Probit Regression and Scalable Maximum Likelihood Estimation via Sketching and Coresets". In: *International Conference on Artificial Intelligence and Statistics, (AISTATS)*. 2022, pp. 2073–2100. URL: https://proceedings.mlr.press/v151/munteanu22a.html.

[MOW21]   Alexander Munteanu, Simon Omlor, and David P. Woodruff. "Oblivious Sketching for Logistic Regression". In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 2021, pp. 7861–7871.

[MOW23]   Alexander Munteanu, Simon Omlor, and David P. Woodruff. "Almost Linear Constant-Factor Sketching for ℓ_1 and Logistic Regression". In: *Proc. of the 11th International Conference on Learning Representations (ICLR)*. 2023.

[Mun+18] Alexander Munteanu et al. "On Coresets for Logistic Regression". In: *Advances in Neural Information Processing Systems 31 (NeurIPS)*. 2018, pp. 6562–6571. URL: https://proceedings.neurips.cc/paper/2018/hash/63bfd6e8f26d1d3537f4c5038264ef36-Abstract.html.

[Mun23]   Alexander Munteanu. "Coresets and Sketches for Regression Problems on Data Streams and Distributed Data". In: *Machine Learning under Resource Constraints, Volume 1 - Fundamentals*. De Gruyter, 2023, pp. 85–98. ISBN: 9783110785944. DOI: doi:10.1515/9783110785944-003. URL: https://doi.org/10.1515/9783110785944-003.

[Nat95]   B. K. Natarajan. "Sparse Approximate Solutions to Linear Systems". In: *SIAM J. Comput.* 24.2 (1995), pp. 227–234.

[NN14]   Jelani Nelson and Huy L. Nguyên. "Lower Bounds for Oblivious Subspace Embeddings". In: *Automata, Languages, and Programming - 41st International Colloquium (ICALP)*. 2014, pp. 883–894.

[NN19]   Shyam Narayanan and Jelani Nelson. "Optimal terminal dimensionality reduction in Euclidean space". In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing (STOC)*. 2019, pp. 1064–1069. DOI: 10.1145/3313276.3316307. URL: https://doi.org/10.1145/3313276.3316307.

[PW11]   Eric Price and David P Woodruff. "$(1 + \varepsilon)$-approximate sparse recovery". In: *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science (FOCS)*. 2011, pp. 295–304.

[Sar06]   Tamás Sarlós. "Improved Approximation Algorithms for Large Matrices via Random Projections". In: *Proc. of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. 2006, pp. 143–152.

[SW11]   Christian Sohler and David P. Woodruff. "Subspace embeddings for the ℓ_1-norm with applications". In: *Proc. of the 43rd ACM Symposium on Theory of Computing, (STOC)*. 2011, pp. 755–764.

[TGS06]   Joel A. Tropp, Anna C. Gilbert, and Martin J. Strauss. "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit". In: *Signal Process.* 86.3 (2006), pp. 572–588. DOI: 10.1016/j.sigpro.2005.05.030. URL: https://doi.org/10.1016/j.sigpro.2005.05.030.

[Tib96]   Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 00359246. URL: http://www.jstor.org/stable/2346178.

[Tro06a]   Joel A. Tropp. "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation". In: *Signal Process.* 86.3 (2006), pp. 589–602. DOI: 10.1016/j.sigpro.2005.05.031. URL: https://doi.org/10.1016/j.sigpro.2005.05.031.

[Tro06b]   Joel A. Tropp. "Just relax: convex programming methods for identifying sparse signals in noise". In: *IEEE Trans. Inf. Theory* 52.3 (2006), pp. 1030–1051. DOI: 10.1109/TIT.2005.864420. URL: https://doi.org/10.1109/TIT.2005.864420.

[Ver18]   Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

[Woo14]   David P. Woodruff. "Sketching as a Tool for Numerical Linear Algebra". In: *Foundations and Trends in Theoretical Computer Science* 10.1-2 (2014), pp. 1–157. DOI: 10.1561/0400000060. URL: https://doi.org/10.1561/0400000060.

[WW19]   Ruosong Wang and David P. Woodruff. "Tight Bounds for ℓ_p Oblivious Subspace Embeddings". In: *Proc. of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2019, pp. 1825–1843.

[WW22]   Ruosong Wang and David P. Woodruff. "Tight Bounds for ℓ_1 Oblivious Subspace Embeddings". In: *ACM Trans. Algorithms* 18.1 (2022), 8:1–8:32. DOI: 10.1145/3477537. URL: https://doi.org/10.1145/3477537.

[WY23]   David P. Woodruff and Taisuke Yasuda. "Online Lewis Weight Sampling". In: *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, (SODA)*. 2023, pp. 4622–4666.

[WZ13]   David P. Woodruff and Qin Zhang. "Subspace Embeddings and ℓ_p-Regression Using Exponential Random Variables". In: *Proc. of the 26th Annual Conference on Learning Theory (COLT)*. 2013, pp. 546–567.

A Lower bounds for k-sparse regression

A.1 Lower Bounds for the ℓ_p-norm Loss Function for $p \geq 1$

In this section we prove our main Theorem 1 on ℓ_2 followed by our extension to $\ell_p, p \geq 1$, see Theorem 2 below.

Theorem 1. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$. Suppose $S \in \mathbb{R}^{m \times n}$ is an oblivious linear sketch for k-sparse ℓ_2 regression with an arbitrary estimator $E(SA, Sb, x)$, such that $\tilde{x} \in \operatorname{argmin}_{x \in \Psi_k} E(SA, Sb, x)$ satisfies $\|A\tilde{x} - b\|_2 \leq (1 + \varepsilon) \min_{x \in \Psi_k} \|Ax - b\|_2$ with constant probability. Then $m = \Omega(k \log(d/k)/\varepsilon^2)$.*

The outline is as follows:

1. We construct a suitable (hard) distribution over k-sparse supports, which is used to define our input distribution.

2. We prove the impossibility of recovering a constant fraction of the support with a small number of measurements (rows) from the input distribution below an information-theoretic lower bound.

3. We construct an ℓ_2-regression instance for which any $1 + \Theta(\varepsilon)$ approximation derived from an oblivious sketch, paired with an arbitrary estimator, reveals a constant fraction of the support. The hardness result thus turns over to the regression problem.

For the first proof step, we begin with the construction of an error correcting code, which will be used in the main argument to construct a hard input distribution.

Definition 2 (Balanced and Correctable Support Set). *Consider a set $\mathcal{U} \subset [d]^{k+1}$ of sets of $k + 1$ indices, such that for all $U \in \mathcal{U}$, $1 \in U$. $\mathcal{U}$ is said to be balanced if, letting $c_i = |\{U \in \mathcal{U} : i \in U\}|$, we have $c_i = c_j$ for all $i, j \in [d] \setminus \{1\}$, and further, letting $c_{ij} = |\{U \in \mathcal{U} : i, j \in U\}|$, $c_{ij} = c_{kl}$ for all $i \neq j$ and $k \neq l$ with $i, j, k, l \in [d] \setminus \{1\}$. The set is said to be correctable if for all $U_1, U_2 \in \mathcal{U}$, $|U_1 \cap U_2| \leq 9/10 \cdot k$.*

We prove the existence of a suitably large balanced and correctable support set as follows. Suppose we choose t sets $S_1, \ldots, S_t$, each of size k, uniformly at random, and each from $[n] = \{1, 2, \ldots, n\}$. Let $\mathcal{H}$ be a pairwise independent family of $n \cdot (n - 1)$ hash functions $h : [n] \to [n]$; it is well-known that such a family exists when n is prime [CW79], which we can assume without loss of generality. For each S_i and $h \in \mathcal{H}$, let $h(S_i)$ denote the image of S_i under h.

Lemma 1. *For any constant $0 < c < 1$, there exists a constant $C > 0$ and $t = \exp(Ck \log(n/k))$ subsets $S_1, \ldots, S_t$, each of size k, such that $|h(S_i) \cap h'(S_j)| < ck$ for all $1 \leq i < j \leq t$ and all $h \neq h' \in \mathcal{H}$.*

Proof. Note that for $i < j$, $h(S_i)$ and $h'(S_j)$ are each random and independent subsets of size k. To calculate their intersection size, we can fix $h(S_i)$. Then the probability that $|h(S_i) \cap h'(S_j)|$ is at least ck is at most the probability that some subset of $h(S_i)$ of size ck is also a subset of $h'(S_j)$. This probability is in turn bounded by

$$\binom{k}{ck} \cdot \frac{\binom{n-ck}{k-ck}}{\binom{n}{k}} \quad \leq \quad \binom{k}{ck} \cdot \frac{(n-ck)! \cdot k!}{(k-ck)! \cdot n!}$$

$$\leq \quad 2^k \cdot \left(\frac{k}{n}\right)^{ck} \leq \exp(-Ck \log(n/k)),$$

where $C > 0$ is a suitable constant. Consequently, by a union bound over all pairs $1 \leq i < j \leq t$ and all choices of $h \neq h' \in \mathcal{H}$, we conclude there exists a choice of $t = \exp(Ck \log(n/k))$ such sets, for a different choice of constant $C > 0$. $\qquad \square$

The lemma above implies that if we have a set T of size k which intersects some $h(S_i)$ in at least ck positions, then S_i is uniquely determined.

Moreover, by pairwise independence of $\mathcal{H}$, for any $a \neq b \in [n]$, the number of sets in the union $\cup_{h \in \mathcal{H}, 1 \leq i \leq t} h(S_i)$ for which a and b occur together is the same. Also, the number of sets in $\cup_{h \in \mathcal{H}, 1 \leq i \leq t} h(S_i)$ containing any particular value $a \in [n]$ is the same as for any other particular value $b \in [n]$. Using this argument we construct the desired set by simply applying Lemma 1 with k and $n = d$ for indexing a family of sets over $\{2, \ldots, d+1\}$ and append the element 1 to each of the sets. The hard instance will later be constructed from a uniform element of this set. This concludes the first part of our proof.

As a second step, we prove the impossibility of recovering a constant fraction of the support with a small number of measurements (rows) from the input distribution below an information-theoretic lower bound. To this end, we need to bound the mutual information first and then plug it into Fano's inequality.

Lemma 2 (Mutual Information Bound). *Let U be selected uniformly at random from a balanced and correctable support set $\mathcal{U}$ (Def. 2) and for some $\varepsilon > 0$, let $z \in \mathbb{R}^d$ have $z(1) = 1$, $z(i) = \varepsilon/\sqrt{k}$ for all $i \in U \setminus \{1\}$ and $z(i) = 0$ for all $i \notin U$. Let $X \in \mathbb{R}^{n \times d}$ have rows drawn independently from a mean zero multivariate Gaussian distribution with covariance $I + zz^T$. Then:*

$$I(U; X) \leq 7n \cdot \varepsilon^2. \tag{1}$$

Proof. Starting from the high level outline of [AW09], since the rows of X are independent, we can write:

$$
\begin{aligned}
I(U; X) &= H(X) - H(X|U) \\
&\leq n \cdot [H(x) - H(x|U)] = \frac{n}{2} \cdot \left[\log \det(\mathbb{E}[xx^T]) - \log \det(\mathbb{E}[xx^T|U]) \right]. \tag{2}
\end{aligned}
$$

We now compute the needed log determinants. First observe that $\mathbb{E}[xx^T|U] = \mathbb{E}[I + zz^T|U] = I + \mathbb{E}[zz^T|U]$. We can observe that $\mathbb{E}[zz^T|U] = DMD$ where $M_{ij} = \varepsilon^2/k$ for $i, j \in U$ and $M_{ij} = 0$ otherwise, and where D is diagonal all with $D_{11} = \sqrt{k}/\varepsilon$ and $D_{ii} = 1$ for all $i \neq 1$. Observe that M is rank-1 and positive semidefinite. Thus, so is DMD. Thus DMD has one non-zero eigenvalue, equal to its trace, which is $1 + \varepsilon^2/k \cdot k = 1 + \varepsilon^2$. Thus, $\mathbb{E}[xx^T|U] = I + DMD$ has one eigenvalue equal to $2 + \varepsilon^2$ and $n - 1$ eigenvalues equal to 1, so

$$\log \det(\mathbb{E}[xx^T|U]) = \log(2 + \varepsilon^2) \geq \log(2). \tag{3}$$

Next consider $\mathbb{E}[xx^T]$. Again we have $\mathbb{E}[xx^T] = \mathbb{E}[I + zz^T] = I + \mathbb{E}[zz^T]$. Since $\mathcal{U}$ is balanced, for all $i \in [d] \setminus \{1\}$, $z(i) = \varepsilon/\sqrt{k}$ with probability $k/(d-1)$ and for $i, j \in [d] \setminus \{1\}$ with $i \neq j$, $z(i) = z(j) = \varepsilon/\sqrt{k}$ with probability $\frac{k(k-1)}{(d-1)(d-2)}$. We can write $\mathbb{E}[zz^T] = D + E$. Here, $E_{11} = 0$, $E_{ij} = \frac{\varepsilon^2}{k} \cdot \frac{k(k-1)}{(d-1)(d-2)} = \frac{\varepsilon^2(k-1)}{(d-1)(d-2)}$ for $i, j \neq 1$, and $E_{i1} = E_{1i} = \frac{\varepsilon}{\sqrt{k}} \cdot \frac{k}{d-1} = \frac{\varepsilon\sqrt{k}}{d-1}$ for $i \neq 1$. $D_{11} = 1$, $D_{ii} = \frac{\varepsilon^2}{d-1} - \frac{\varepsilon^2}{k} \cdot \frac{k(k-1)}{(d-1)(d-2)} \leq \frac{\varepsilon^2}{d-1}$.

Observe that $\|E\|_F \leq \sqrt{2(d-1) \cdot \frac{\varepsilon^2 k}{(d-1)^2} + (d-1)^2 \cdot \frac{\varepsilon^4(k-1)^2}{(d-1)^2(d-2)^2}} \leq \sqrt{3} \cdot \varepsilon^2$. Further, E is rank-3 and thus has just 3 non-zero eigenvalues. By Weyl's inequality,

$$\lambda_1(\mathbb{E}[zz^T]) = \lambda_1(D + E) \leq \lambda_1(D) + \lambda_1(E) \leq 1 + \sqrt{3}\varepsilon^2.$$

For $i = 2, 3, 4, 5$,

$$\lambda_i(\mathbb{E}[zz^T]) = \lambda_i(D + E) \leq \lambda_2(D) + \lambda_1(E) \leq \frac{\varepsilon^2}{d-1} + \sqrt{3}\varepsilon^2 \leq (\sqrt{3} + 1)\varepsilon^2.$$

Finally, for $i \geq 5$,

$$\lambda_i(\mathbb{E}[zz^T]) = \lambda_i(D + E) \leq \lambda_2(D) + \lambda_4(E) = \frac{\varepsilon^2}{d-1} + 0 = \frac{\varepsilon^2}{d-1}.$$

Thus,

$$
\begin{aligned}
\log \det(\mathbb{E}[xx^T]) &= \log \det(I + \mathbb{E}[zz^T]) \\
&\leq \log(2 + \sqrt{3}\varepsilon^2) + 4\log(1 + (\sqrt{3} + 1)\varepsilon^2) + (d - 5)\log\left(1 + \frac{\varepsilon^2}{d-1}\right) \\
&\leq \log(2) + (5\sqrt{3} + 4 + 1)\varepsilon^2 \\
&\leq \log(2) + 14\varepsilon^2. \tag{4}
\end{aligned}
$$

Combined with (3) we have $\log \det(\mathbb{E}[xx^T]) - \log \det(\mathbb{E}[xx^T|U]) \leq 14\varepsilon^2$, and plugging back into (2), we have $I(U; X) \leq 7n \cdot \varepsilon^2$ as desired. □

Our mutual information bound can be plugged into Fano's inequality to obtain a lower bound on the sample complexity needed for an approximate, i.e., partial, recovery of the support set. This will later translate into the number of rows of our sketch.

Corollary 15 (Sample Complexity Lower Bound). *Let U, X be distributed as in Lemma 2 with $n \leq \frac{ck \log(d/k)}{\varepsilon^2}$ for sufficiently small constant c. Then no algorithm that takes just X as input can output a set $\tilde{U}$ with $|\tilde{U}| = k$ and $|\tilde{U} \cap U| > 19k/20$ with probability $\geq 2/3$.*

Proof. Suppose such an algorithm existed. Since all $U, U' \in \mathcal{U}$ have $|U \cap U'| \leq k \cdot 9/10$, if $|\tilde{U} \cap U| > 19k/20$, then $\tilde{U}$ must contain $> k/20$ elements not in U' for any $U' \in \mathcal{U}$ with $U' \neq U$. Thus, we must have $|\tilde{U} \cap U'| < 19k/20$. So $\tilde{U}$ can be used to uniquely identify U. That is, the algorithm identifies U with probability $\geq 2/3$. However, by Fano's inequality (Lemma 6), the algorithm fails with probability at least

$$1 - \frac{I(U; X) + \log 2}{\log |\mathcal{U}|} \geq 1 - \frac{7\varepsilon^2 n + \log 2}{\log |\mathcal{U}|},$$

where the bound on $I(U; X)$ follows from Lemma 2. Since $\log |\mathcal{U}| = \Theta(k \log(d/k))$ this failure probability is $> 1/3$ if $n = \frac{ck \log(d/k)}{\varepsilon^2}$ for small enough c and d, k are bigger than large enough constants. This gives a contradiction to the assumption that the algorithm succeeds with probability $\geq 2/3$, and hence the corollary. □

This concludes the second part of our proof regarding the hardness of support recovery. For the third part, i.e., the reduction of this hard problem to sparse linear regression, we first need a few technical lemmas, before we can finally prove Theorem 1.

The first technical result establishes a connection between approximating a loss function L to within $(1 + O(\varepsilon))$ error and revealing a constant fraction of the support. We note that L will represent the regression cost in our subsequent reduction.

Lemma 3. *Let $v \in \mathbb{R}^d$ be a k-sparse vector with k non-zero entries equal to $1/\sqrt{k}$. Let x be another k-sparse vector. Let $M > \sqrt{n/(\varepsilon k)}$, $\alpha = |\operatorname{supp}(v) \cap \operatorname{supp}(x)|$, and*

$$L = 1 + \|x\|_2^2 + (1 - \varepsilon x^T v)^2 + \frac{M^2}{n} \left(\sum_i x(i) - \sqrt{k} \right)^2.$$

There exists a constant c such that any x with $\alpha < 19k/20$ is not a $1 + c\varepsilon$ approximation solution of L.

Proof. Let $S = \operatorname{supp}(x) \cap \operatorname{supp}(v)$ and $M' = M/\sqrt{n}$. Let $\beta = \sum_i x(i)$ and γ be such that $\sum_{i \in S} x(i) = \gamma \beta$. We will optimize L over all possible values of β and γ in $\mathbb{R}$. Note that any x minimizing L must have the form

$$x(i) = \begin{cases} \frac{\gamma \beta}{\alpha} & \text{for } i \in S, \\ \frac{(1-\gamma)\beta}{k-\alpha} & \text{for } i \in \operatorname{supp}(x) \setminus S. \end{cases}$$

This is because for fixed S, β, γ, making x have the above form minimizes $\|x\|$ without affecting $x^T v$. Therefore,

$$L = 1 + \frac{\gamma^2 \beta^2}{\alpha} + \frac{(1-\gamma)^2 \beta^2}{k-\alpha} + \left(1 - \varepsilon \alpha \frac{1}{\sqrt{k}} \frac{\gamma \beta}{\alpha} \right)^2 + M'^2 \left(\beta - \sqrt{k} \right)^2$$

$$= \gamma^2 \left(\frac{\beta^2}{\alpha} + \frac{\beta^2}{k-\alpha} + \varepsilon^2 \frac{\beta^2}{k} \right) - 2\gamma \left(\frac{\beta^2}{k-\alpha} + \varepsilon \frac{\beta}{\sqrt{k}} \right) + 2 + \frac{\beta^2}{k-\alpha} + M'^2(\beta - \sqrt{k})^2.$$

Minimizing over all $\gamma \in \mathbb{R}$ gives

$$\min_\gamma L = 2 + \frac{\beta^2}{k-\alpha} + M'^2(\beta - \sqrt{k})^2 - \frac{\left(\frac{\beta^2}{k-\alpha} + \varepsilon \frac{\beta}{\sqrt{k}} \right)^2}{\frac{\beta^2}{\alpha} + \frac{\beta^2}{k-\alpha} + \varepsilon^2 \frac{\beta^2}{k}}$$

$$= 2 + \frac{\beta^2}{k-\alpha} + M'^2(\beta - \sqrt{k})^2 - \frac{\frac{\beta^2}{(k-\alpha)^2} + \frac{2\beta\varepsilon}{(k-\alpha)\sqrt{k}} + \frac{\varepsilon^2}{k}}{\frac{k}{\alpha(k-\alpha)} + \frac{\varepsilon^2}{k}}$$

$$= 2 + M'^2(\beta - \sqrt{k})^2 - \frac{\frac{\beta^2}{(k-\alpha)^2} \left(1 - \frac{k}{\alpha} \right) + \frac{\beta^2 \varepsilon^2}{k(k-\alpha)} + \frac{2\beta\varepsilon}{(k-\alpha)\sqrt{k}} + \frac{\varepsilon^2}{k}}{\frac{k}{\alpha(k-\alpha)} + \frac{\varepsilon^2}{k}}$$

$$= 2 + M'^2(\beta - \sqrt{k})^2 + \frac{\beta^2 k - \beta^2 \varepsilon^2 \alpha - 2\beta\sqrt{k}\alpha\varepsilon - \varepsilon^2(k-\alpha)\alpha}{k^2 + \varepsilon^2(k-\alpha)\alpha}$$

$$\approx_\varepsilon 2 + M'^2(\beta - \sqrt{k})^2 + \frac{\beta^2 k - 2\beta\sqrt{k}\alpha\varepsilon}{k^2}$$

$$= \beta^2 \left(M'^2 + \frac{1}{k} \right) - \beta \left(2M'^2 \sqrt{k} + \frac{2\alpha\varepsilon}{k\sqrt{k}} \right) + 2 + M'^2 k.$$

Minimizing over β gives

$$\min_\beta \min_\gamma L \approx_\varepsilon 2 + M'^2 k - \frac{\left(M'^2 \sqrt{k} + \frac{\alpha\varepsilon}{k\sqrt{k}} \right)^2}{M'^2 + \frac{1}{k}}$$

$$= 2 + \frac{M'^2 - \frac{2M'^2 \alpha \varepsilon}{k} - \frac{\alpha^2 \varepsilon^2}{k^3}}{M'^2 + \frac{1}{k}}$$

$$\approx_\varepsilon 2 + \frac{M'^2 - \frac{2M'^2 \alpha \varepsilon}{k}}{M'^2 + \frac{1}{k}}.$$

Since $M' = M/\sqrt{n} > 1/\sqrt{\varepsilon k}$, we have

$$\min_\beta \min_\gamma L \approx_\varepsilon 3 - 2\frac{\alpha}{k}\varepsilon.$$

We can observe that the RHS is minimized when $\alpha = k$ at $3 - 2\varepsilon$. Moreover, if $\alpha < c_1 k$ for $c_1 < 1$, it is at least

$$3 - 2c_1\varepsilon \geq (3 - 2\varepsilon)\left(1 + \frac{2(1 - c_1)}{3}\varepsilon\right).$$

Therefore, there exists a small enough constant c such that if $\alpha < 19k/20$, it is not possible to minimize L within $1 + c\varepsilon$ factor. $\qquad\square$

The next ingredient will help us analyze the regression cost up to $1 \pm \varepsilon$ error deterministically by removing the influence of a random Gaussian matrix in our input distribution. By a standard tail bound for Gaussian matrices (for example, Exercise 4.7.3 in [Ver18]) we have the following lemma.

Lemma 4. *Suppose X is a $n \times d$ Gaussian matrix with covariance Σ. Then there exists a constant C such that we have with probability $\geq 1 - \delta$*

$$(1 - \varepsilon)\Sigma \preceq \frac{1}{n}X^T X \preceq (1 + \varepsilon)\Sigma,$$

when $n \geq C\frac{d + \log(1/\delta)}{\varepsilon^2}$.

By invoking the above lemma to a fixed k-dimensional subspace and applying a union bound over $\binom{d}{k}$ k-dimensional subspaces, we have with probability at least $1 - \delta$

$$\forall k\text{-sparse vectors } v \in \Psi_k: (1 - \varepsilon)v^T\Sigma v \preceq \frac{1}{n}v^T X^T X v \preceq (1 + \varepsilon)v^T\Sigma v,$$

when $n \geq C'\frac{k \log(d/(k\delta))}{\varepsilon^2}$ for some $C' > C$.

Using the previous results, we are now we are ready to give the main proof of Theorem 1, and hereby conclude the third and final part of our proof outline.

We construct an ℓ_2-regression instance for which any $1 + \Theta(\varepsilon)$ approximation derived from an oblivious sketch, paired with an arbitrary estimator, reveals a constant fraction of the support. The hardness result thus turns over to the regression problem.

Proof. (of Theorem 1) Let $X = [b\ A]$ be distributed as $G(I + zz^T)^{1/2}$, where z is distributed as described in Lemma 2. We will construct a k-sparse ℓ_2 regression problem based on X such that an $1 + \varepsilon$ approximation of the constructed problem allows us to recover a large enough fraction, i.e., greater than $19/20$, of the support of z. We can assume without loss of generality that the sketching matrix has orthonormal rows, and since X is a Gaussian matrix, the sketch has rows sampled from the same distribution as the rows of X. This is so because we are proving lower bounds against any estimator on the sketch.

By Corollary 15, if the number of samples is smaller than $ck \log(k/d)/\varepsilon^2$, no algorithm can recover more than a 19/20 fraction of the support of z with probability larger than 2/3. Therefore, we have a lower bound of $\Omega(k \log(d/k)/\varepsilon^2)$ against a $1 + \varepsilon$ approximation of k-sparse ℓ_2 regression.

Consider the following ℓ_2 sparse regression problem

$$\min_{x \in \Psi_k} \left\| \begin{bmatrix} M\,M\,\dots\,M \\ A \end{bmatrix} x - \begin{bmatrix} \sqrt{k}M \\ b \end{bmatrix} \right\|_2.$$

We will let M be a very large number, which enforces $\sum_i x(i)$ to be close to $\sqrt{k}$. The squared loss of the above regression problem is

$$L = \|Ax - b\|_2^2 + M^2 \left(\sum_i x(i) - \sqrt{k} \right)^2 = \|X\tilde{x}\|_2^2 + M^2 \left(\sum_i x(i) - \sqrt{k} \right)^2,$$

where $\tilde{x} = (-1, x)$. By matrix concentration in Lemma 4, for $n = \Omega(k \log(d/(k\delta))/\varepsilon^2)$, with probability at least $1 - \delta$,

$$X^T X \approx_\varepsilon n(I + zz^T).$$

Recall that z has $z(1) = 1$, $z(i) = \varepsilon/\sqrt{k}$ for all $i \in U \setminus \{1\}$ and $z(i) = 0$ for all $i \notin U$. We have

$$\frac{1}{n}\|X\tilde{x}\|_2^2 \approx_\varepsilon \|\tilde{x}\|_2^2 + (\tilde{x}^T z)^2 = (1 + \|x\|_2^2) + (1 - \varepsilon x^T v)^2,$$

where v is such that $z = (1, v)$.

Therefore,

$$\frac{L}{n} \approx_\varepsilon 1 + \|x\|_2^2 + (1 - \varepsilon x^T v)^2 + \frac{M^2}{n} \left(\sum_i x(i) - \sqrt{k} \right)^2.$$

Note that x is a $1 + \varepsilon$ approximation of the ℓ_2 loss iff x is an $1 + \Theta(\varepsilon)$ approximation of the ℓ_2^2 loss.

Let $\alpha = |\operatorname{supp}(x) \cap \operatorname{supp}(v)|$. By Lemma 3, there exists a constant c such that if we can approximate L within a factor of $1 + c\varepsilon$, we must have $\alpha > 19k/20$. Rescaling ε and combining with Corollary 15 and a union bound (with δ set to a small constant) proves the theorem. $\qquad \square$

Next, we extend our ℓ_2 lower bound to ℓ_p for all $p \geq 1$.

Theorem 2. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$. Suppose $S \in \mathbb{R}^{m \times n}$ is an oblivious linear sketch for k-sparse ℓ_p regression for any $p \geq 1$ with an estimator $E_p(SA, Sb, x)$, such that the minimizer $\tilde{x} \in \operatorname{argmin}_{x \in \Psi_k} E_p(SA, Sb, x)$ satisfies $\|A\tilde{x} - b\|_p \leq (1 + \varepsilon) \min_{x \in \Psi_k} \|Ax - b\|_p$. Then $m = \Omega(k \log(d/k)/\varepsilon^2)$.*

Proof. (of Theorem 2) We reduce from the ℓ_2 case by leveraging the fact that ℓ_2 embeds obliviously up to $(1 \pm \varepsilon)$ distortion into ℓ_p for all $p \geq 1$ by Dvoretzky's theorem. Indeed, such an embedding can be constructed using a random mapping $G \in \mathbb{R}^{r \times n}$ whose entries are appropriately rescaled i.i.d. Gaussians. In particular G is an oblivious linear map. The number of rows is $r = O(n \log(1/\varepsilon)/\varepsilon^2)$ for $1 \leq p \leq 2$ and $r = n^{O(p)}$ for $p > 2$; [see Mat13, p. 30]. We note, however, that the number of rows of G does not matter in our context since it is reduced by an application of S in what follows.

More precisely, we have for all $x \in \mathbb{R}^d$ that $(1 - \varepsilon)\|Ax - b\|_2 \leq \|GAx - Gb\|_p \leq (1 + \varepsilon)\|Ax - b\|_2$. Now suppose that S is an oblivious sketching matrix for the k-sparse regression problem in ℓ_p. Then we can find $\tilde{x} \in \operatorname{argmin}_{x \in \Psi_k} E_p(SGA, SGb, x)$. By definition it holds that $\|G(A\tilde{x} - b)\|_p \leq$

22

$(1+\varepsilon)\|G(Ax_G^* - b)\|_p$ where $x_G^* \in \mathrm{argmin}_{x\in\Psi_k}\|G(Ax-b)\|_p$. Also let $x^* \in \mathrm{argmin}_{x\in\Psi_k}\|Ax-b\|_2$ be the minimizer for the ℓ_2 problem. Now it follows that

$$
\begin{aligned}
\|A\tilde{x} - b\|_2 &\le \|G(A\tilde{x}-b)\|_p/(1-\varepsilon) \le \|G(Ax_G^*-b)\|_p(1+\varepsilon)/(1-\varepsilon)\\
&\le \|G(Ax^*-b)\|_p(1+\varepsilon)/(1-\varepsilon) \le \|Ax^*-b\|_2(1+\varepsilon)^2/(1-\varepsilon)\\
&\le (1+7\varepsilon)\|Ax^*-b\|_2,
\end{aligned}
$$

which by rescaling ε means that SG is an oblivious linear sketch for the ℓ_2-norm problem with an ℓ_p-norm minimization estimator and at most $1+\varepsilon$ error. Using the ℓ_2 lower bound given in Theorem 1, it follows that SG and thus also S has $m = \Omega(k\log(d/k)/\varepsilon^2)$ rows. $\qquad\square$

A.2 Lower bounds for ReLU and hinge-like loss functions

Here we give further reductions similar to Theorem 2 in order to extend our main result to ReLU and hinge-like loss functions.

Corollary 3. *Let $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$. Suppose $S \in \mathbb{R}^{m\times n}$ is an oblivious linear sketch for k-sparse ReLU regression with an estimator $E_{\mathrm{ReLU}}(SA, Sb, x)$, such that $\tilde{x} \in \mathrm{argmin}_{x\in\Psi_k} E_{\mathrm{ReLU}}(SA, Sb, x)$ satisfies $\|A\tilde{x}-b\|_{\mathrm{ReLU}} \le (1+\varepsilon)\min_{x\in\Psi_k}\|Ax-b\|_{\mathrm{ReLU}}$. Then $m = \Omega(k\log(d/k)/\varepsilon^2)$.*

Proof. Note that $\|Ax-b\|_1 = \|Ax-b\|_{\mathrm{ReLU}} + \|-(Ax-b)\|_{\mathrm{ReLU}}$, $\forall x$. Therefore, given A, b we have $\|Ax-b\|_1 = \|PAx - Pb\|_{\mathrm{ReLU}}$, where

$$
P = \begin{bmatrix} I_n \\ -I_n \end{bmatrix}
$$

is a $2n \times n$ matrix. Suppose $S \in \mathbb{R}^{m\times 2n}$ is an oblivious linear sketch for k-sparse ReLU regression with an estimator $E_{\mathrm{ReLU}}(SA, Sb, x)$. This implies that $\tilde{x} \in \mathrm{argmin}_{x\in\Psi_k} E_{\mathrm{ReLU}}(SPA, SPb, x)$ satisfies

$$
\|A\tilde{x}-b\|_1 = \|P(A\tilde{x}-b)\|_{\mathrm{ReLU}} \le (1+\varepsilon)\min_{x\in\Psi_k}\|P(Ax-b)\|_{\mathrm{ReLU}} = (1+\varepsilon)\min_{x\in\Psi_k}\|Ax-b\|_1.
$$

Therefore, $S \cdot P$ is an oblivious linear sketch for ℓ_1 regression. By Theorem 2, we have that $m = \Omega(k\log(d/k)/\varepsilon^2)$. $\qquad\square$

Definition 1. *We say $f(\cdot)$ is an (L, a_1, a_2) hinge-like loss function if f is L Lipschitz, $\forall x \ge 0\colon f(x) \ge a_2 > 0$, and $\forall x\colon |f(x) - \mathrm{ReLU}(x)| \le a_1$.*

The logistic loss function $\log(1+e^{-x})$ and the hinge loss function $\max(0, 1-x)$ are $(1, \ln(2), \ln(2))$ and $(1, 1, 1)$ hinge-like loss functions, respectively.

We will use the notation $a = (1\pm\varepsilon)b$ to denote $(1-\varepsilon)b \le a \le (1+\varepsilon)b$. We note that for $\varepsilon > 0$ if $\|x\|_{\mathrm{ReLU}} > 0$, then there is a constant $c > 0$ such that

$$
c\|x\|_{\mathrm{ReLU}} = \|cx\|_{\mathrm{ReLU}} = (1\pm\varepsilon)\|cx\|_f.
$$

Corollary 4. *Let $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n$. Suppose $S \in \mathbb{R}^{m\times n}$ is an oblivious subspace embedding for some hinge-like loss function f with an estimator $E_f(SA, Sb, x)$, such that we have $\forall x \in \Psi_k\colon (1-\varepsilon)\|Ax-b\|_f \le E_f(SA, Sb, x) \le (1+\varepsilon)\|Ax-b\|_f$. Then $m = \Omega(k\log(d/k)/\varepsilon^2)$.*

Proof. Suppose $S \in \mathbb{R}^{m\times n}$ is an oblivious subspace embedding for f with an estimator $E_f(SA, Sb, x)$.

Let $x \in \Psi_k$ be some k-sparse vector. We will show that $S \in \mathbb{R}^{m \times n}$ is an oblivious subspace embedding for ReLU with the estimator

$$E_{\text{ReLU}}(SA, Sb, x) = \lim_{c \to \infty} \frac{E_f(ScA, Scb, x)}{c}.$$

First, consider the case where $\|Ax - b\|_{\text{ReLU}} > 0$. For c large enough, we have

$$\begin{aligned} \|Ax - b\|_{\text{ReLU}} &= \frac{\|c(Ax - b)\|_{\text{ReLU}}}{c} \\ &= (1 \pm \varepsilon) \frac{\|c(Ax - b)\|_f}{c} \\ &= (1 \pm 3\varepsilon) \frac{E_f(ScA, Scb, x)}{c}. \end{aligned}$$

Here, we assumed that ε is small enough so that $(1 + \varepsilon)^2 \leq 1 + 3\varepsilon$.

Next we consider the case where $\|Ax - b\|_{\text{ReLU}} = 0$.

We have

$$\begin{aligned} E_f(ScA, Scb, x) &\leq (1 + \varepsilon)\|c(Ax - b)\|_f \\ &\leq (1 + \varepsilon)na_1. \end{aligned}$$

Therefore,

$$\lim_{c \to \infty} \frac{E_f(ScA, Scb, x)}{c} \leq \lim_{c \to \infty} \frac{2na_1}{c} = 0$$

as desired. $\qquad\square$

A.3 Sampling fails for sparse regression

In this section we argue that sampling based algorithms, which are important tools in sketching for non-sparse regression [DMM06; Das+09; Mun+18; MMR21], do not give any non-trivial results in the sparse setting. Specifically, these algorithms cannot compress beyond what is possible in the non-sparse case – roughly beyond the rank of the input matrix. This is the reason that our upper bounds all build on general linear sketches, rather than sampling.

Theorem 5. *Consider any bounded approximation factor $\alpha \geq 1$ and any $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}^{\geq 0}$ which evaluates to 0 on the all zeros vector and to some positive number on any other vector. For any $n > 9$, there is some input matrix $A \in \mathbb{R}^{n \times n}$ and distribution over vectors $b \in \mathbb{R}^n$ such that for any sampling matrix $S \in \mathbb{R}^{m \times n}$ with $m < n/3$, no algorithm that accesses just SA and Sb can output a k-sparse $\tilde{x} \in \Psi_k$ with $\|A\tilde{x} - b\| \leq \alpha \cdot \min_{x \in \Psi_k} \|Ax - b\|$ with probability at least $1/2$ (over the choice of b and any possible randomness in the algorithm).*

Proof. Let $k = 1$ and A be the $n \times n$ identity matrix. Let b be set to the ith standard basis vector with probability $1/n$. Note that $\min_{\text{1-sparse } x \in \mathbb{R}^n} \|Ax - b\| = 0$, so to achieve any bounded approximation factor, the algorithm must output x with 0 cost – i.e., $x = b$. Any algorithm that accesses any $m < n/3$ rows of A and b will see only zero entries in Sb with probability at least $2/3$. Let $\mathcal{X}_0$ be the distribution over outputs of the algorithm given that $Sb = 0$. The algorithm achieves a bounded approximation factor only if $x \sim \mathcal{X}_0$ satisfies $x = b$. This occurs with probability at most $\frac{3}{2n}$ since after seeing $Sb = 0$, any of the remaining $2/3 \cdot n$ possibilities for b are equally likely. Thus, the algorithm succeeds with probability at most $1/3 + 3/(2n) < 1/2$ for $n > 9$. $\qquad\square$

Note that the sampling matrix S in Theorem 5 may depend on A but not on b. This is necessary. If the sampling matrix can depend on b, then, without bounded computation, a sampling algorithm can in theory compress the problem to $O(k/\varepsilon)$ rows. In particular, it can simply solve for the optimal $x^* = \operatorname{argmin}_{k\text{-sparse } x \in \Psi_k} \|Ax - b\|$ and only consider the k columns of A within the support of x^*. In e.g., the ℓ_2 case, by simply applying variants of standard leverage score sampling, [Woo14; CP19] to these columns, it can output a sampling matrix $S \in \mathbb{R}^{m \times n}$ for $m = O(k/\varepsilon)$ with $\tilde{x} = \operatorname{argmin}_{k\text{-sparse } x \in \Psi_k} \|SAx - Sb\|$ satisfying $\|A\tilde{x} - b\| \leq (1 + \varepsilon)\min_{k\text{-sparse } x \in \Psi_k} \|Ax - b\|$.

If the sampling matrix is required to preserve $\|Ax - b\|$ for every k-sparse $x \in \Psi_k$, then even if a sampling algorithm can read b, it is easy to see that it must sample at least n rows. This is true even in the case that $b = 0$.

Theorem 6. *Consider any bounded approximation factor $\alpha \geq 1$ and any $\| \cdot \| : \mathbb{R}^n \to \mathbb{R}^{\geq 0}$ which evaluates to 0 on the all zeros vector and to some positive number on any other vector. For any $n > 9$, there is some input matrix $A \in \mathbb{R}^{n \times n}$ such that there is no sampling matrix $S \in \mathbb{R}^{m \times n}$ with $m < n$, which satisfies for all k-sparse $x \in \Psi_k$, $\alpha^{-1}\|Ax\| \leq \|SAx\| \leq \alpha\|Ax\|$.*

Proof. Let A be the $n \times n$ identity matrix. Then if we sample $m < n$ rows, we will have $\|SAx\| = 0$ when x is at least one of the standard basis vectors. This violates the approximation bound. $\square$

B   Upper bounds for k-sparse regression

B.1   Upper bounds for the ℓ_p-norm loss function for $p \in [1, 2]$

We prove k-sparse affine embedding upper bounds and note that as a corollary we obtain the same bounds for minimization. We begin with ℓ_2.

Theorem 7. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$. There exists a distribution over random matrices $S \in \mathbb{R}^{m \times n}$ with $m = O(k\log(d/k)/\varepsilon^2)$ such that it holds with constant probability that $\forall x \in \Psi_k : (1 - \varepsilon)\|Ax - b\|_2 \leq \|S(Ax - b)\|_2 \leq (1 + \varepsilon)\|Ax - b\|_2$.*

Proof. (of Theorem 7) The upper bound is similar to the known constructions [Bar+08] of RIP matrices via Johnson-Lindenstrauss embeddings [JL84], i.e., appropriately rescaled Gaussian matrices. The main difference is that the subspaces formed by any fixed k-sparse support of x need not be orthogonal or aligned with the standard basis vectors. The first idea is that there are at most $\binom{d}{k} \leq (ed/k)^k$ different k-sparse supports and each of them corresponds to one choice of k columns of A. Every such choice spans a k-dimensional linear subspace of dimension $\leq k$. By the subspace embedding construction in [Sar06], every subspace formed by one choice of k columns can be handled by embedding the points in a net of size $(3/\varepsilon)^k$ covering the unit ball in the subspace. The remaining vectors can be related to the net points by triangle inequality and the embedding extends to vectors of arbitrary norm outside the unit sphere by linearity. A slightly more sophisticated argument in [Woo14, pp. 13] states that the net can be constructed with ε replaced by an absolute constant $\varepsilon_0 := 1/2$. So the total number of points to embed up to $(1 \pm \varepsilon)$ distortion is bounded by $|\mathcal{N}| \leq (d/k)^k \cdot c^k$ for an absolute constant $c = 3e/\varepsilon_0$. The embedding can be accomplished via the Johnson-Lindenstrauss lemma followed by a union bound over $\mathcal{N}$, which yields a matrix $S = \frac{1}{\sqrt{m}}G$ whose entries are scaled i.i.d. standard Gaussians $G_{ij} \sim N(0, 1)$ with $m = O(\log(|\mathcal{N}|)/\varepsilon^2) = O(k\log(d/k)/\varepsilon^2)$ rows.[3] $\square$

---

[3] We note that the same result can be achieved by random sign (Rademacher) matrices [CW09] which is more convenient in streaming and other space constrained settings.

We continue with $\ell_p, p \in [1,2)$. We note that the outline is similar to [Bac+16, appendix F.1] but our result is non-trivially adapted to the sparse setting and generalized to ℓ_p.

Theorem 8. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n, p \in [1,2)$. There exists a distribution over random matrices $S \in \mathbb{R}^{m \times n}$ with $m = O(k(\log(d/k) + \log(k/(\varepsilon\delta)))/\varepsilon^2)$ such that it holds with probability at least $1 - \delta$ that $\forall x \in \Psi_k : (1 - \varepsilon)\|Ax - b\|_p \leq \|S(Ax - b)\|_{\mathrm{med}} \leq (1 + \varepsilon)\|Ax - b\|_p$, where for arbitrary $y \in \mathbb{R}^d, \|y\|_{\mathrm{med}} := \mathrm{median}\{|y_i| \mid i \in [d]\}$.*

Proof. (of Theorem 8) We choose $S \in \mathbb{R}^{m \times n}$ to be a matrix whose entries are i.i.d. p-stable random variables with scale parameter $\gamma_p = c^{1-1/p}$, where $c \approx 1.099055$; see below. We show that this matrix has the desired property. First note that $Ax - b = [A, b][x^T, -1]^T$, so we can simply assume that the input consist only of $A' = [A, b]$ and it suffices to show that $\|SA'x\|_{\mathrm{med}} = (1 \pm \varepsilon)\|A'x\|_p$ for all $x \in \Psi_{k'} \subseteq \mathbb{R}^{d+1}$ for $k' = k + 1$. In what follows we re-substitute A for A' and k for k' for the sake of presentation.

Fix any k-sparse support indexed by $I \subseteq [d]$ with $|I| = k$. Let A_I be the matrix whose k columns are the columns A_{*i} of A such that $i \in I$. From a classic result of Auerbach [cf. Aue30; Das+09] it follows that there exists a basis $L \in \mathbb{R}^{n \times k}$ for the p-normed subspace spanned by those columns [WW22, Lemma 2.22] that satisfies the following properties: the ℓ_p norm of each column $i \in [k]$ is exactly $\|L_{*i}\|_p = k^{1/q}$, and for all $x \in \mathbb{R}^k$ it holds that $\|Lx\|_p \geq \|x\|_q$, where $q = \infty$ for $p = 1$ and $q = \frac{p}{p-1}$ for $p \in (1,2)$ denotes the *dual* norm of p.

Further, by a property of the p-stable random variables we have that any entry

$$(SL)_{ij} = \sum_{h=1}^{n} S_{ih} L_{hj} \sim C \cdot \|L_{*j}\|_p = C \cdot k^{1/q}$$

is again a scaled p-stable random variable; [cf. Ind06; Dyt+18]. It follows that for any threshold τ the probability that any entry of SL has absolute value larger than τ is bounded by $O((k^{1/q}/\tau)^p) = O(k/\tau^p)$, [cf. BLM18].

Setting $\tau = O((mk^2/\delta)^{1/p}) = \tilde{O}((k^3 \log(d)/\delta)^{1/p})$, we have that all entries of SL are simultaneously bounded by τ with probability $1 - \delta/2$. Suppose this event is true. Then for all $x \in \Psi_k$ with the fixed support indexed by I, we have by Hoelder's inequality (applied only to the k-sparse support, since other terms are zero) and by the properties discussed above, that

$$\|SLx\|_\infty = \max_{i \in [m]} |S_{i*}Lx| \leq \max_{i \in [m]} \|(SL)_{i*}\|_\infty \|x\|_1 \leq \tau k^{1 - \frac{1}{q}} \|x\|_q \leq \tau k^{\frac{1}{p}} \|Lx\|_p$$
$$\leq \tilde{O}((k^4 \log(d)/\delta)^{1/p}) \|Lx\|_p.$$

Let $\tau' = \tilde{O}((k^4 \log(d)/\delta)^{1/p}) = \tilde{O}(k^4 \log(d)/\delta)$. We construct an $\frac{\varepsilon}{\tau'}$-net $\mathcal{N}_I^k$ in the ℓ_p norm for the unit ℓ_p ball intersect the subspace spanned by L. By linearity, the restriction to the unit ball is w.l.o.g. We repeat this construction for each k-sparse support and define our net to be the union over all supports, i.e., $\mathcal{N} = \bigcup_{I \subseteq [d], |I|=k} \mathcal{N}_I^k$. There are at most $\binom{d}{k} \leq (ed/k)^k$ different subspaces, each of which is covered by a net of size at most $|\mathcal{N}_I^k| \leq (3\tau'/\varepsilon)^k$ by the standard volume argument. Consequently for an absolute constant c_1 we have $|\mathcal{N}| \leq (ed/k)^k \cdot (c_1 k^4 \log(d)/(\varepsilon\delta))^k = \exp(O(k \log(d/k) + k \log(k/(\varepsilon\delta))))$.

Next, we investigate the cdf $F_p(x)$ of the random variable $|X|$, where X follows a p-stable distribution. Except for the cases $p \in \{1, 2\}$, which correspond to the Cauchy and Normal distribution, no analytic/closed form expression is known for the cdfs and pdfs. We thus take a detour and leverage the inversion theorem of Lévy [Lév25; Mas77] based on the characteristic function, for

26

which a closed form expression is known [BHW05; Dyt+18]: $\phi_p(t) = \exp -|\gamma_p t|^p$, where γ_p is the constant scale parameter defined above. More precisely, it holds that

$$F_p(x) = P(|X| \le x) = P(X \le x) - P(X \le -x)$$
$$= \frac{1}{2\pi} \lim_{T \to \infty} \int_{-T}^{T} \frac{e^{itx} - e^{-itx}}{it} \phi_p(t) \, dt$$
$$= \frac{1}{2\pi} \lim_{T \to \infty} \int_{-T}^{T} \frac{e^{itx} - e^{-itx}}{it} e^{-|\gamma_p t|^p} \, dt$$
$$= \frac{1}{2\pi} \lim_{T \to \infty} \int_{-T}^{T} \frac{2i \sin(tx)}{it} e^{-|\gamma_p t|^p} \, dt$$
$$= \frac{1}{\pi} \lim_{T \to \infty} \int_{-T}^{T} \frac{\sin(tx)}{t} e^{-|\gamma_p t|^p} \, dt.$$

It follows that

$$F_p(1) = \frac{1}{\pi} \lim_{T \to \infty} \int_{-T}^{T} \frac{\sin t}{t} e^{-|\gamma_p t|^p} \, dt = \frac{1}{2} \tag{5}$$

It remains to show that the derivative of F is bounded at $F(1) = \frac{1}{2}$. To this end we observe that

$$F_p'(x) = \frac{1}{\pi} \lim_{T \to \infty} \int_{-T}^{T} \frac{\cos(tx) \cdot t}{t} e^{-|\gamma_p t|^p} \, dt$$
$$= \frac{1}{\pi} \lim_{T \to \infty} \int_{-T}^{T} \cos(tx) \, e^{-|\gamma_p t|^p} \, dt.$$

Consequently,

$$F_p'(1) = \frac{1}{\pi} \lim_{T \to \infty} \int_{-T}^{T} \cos(t) \, e^{-|\gamma_p t|^p} \, dt.$$

Now by the symmetry of the integrand and monotonicity of the characteristic function w.r.t. the exponent p we have

$$\frac{1}{\pi} = \frac{1}{\pi} \lim_{T \to \infty} \int_{-T}^{T} \cos(t) \, e^{-|\gamma_1 t|} \, dt \le F_p'(1) \le \frac{1}{\pi} \lim_{T \to \infty} \int_{-T}^{T} \cos(t) \, e^{-|\gamma_2 t|^2} \, dt$$
$$\le \frac{1}{\pi} \lim_{T \to \infty} \int_{-T}^{T} \cos(t) \, e^{-|t|^2} \, dt = \frac{1}{\pi^{1/2} e^{1/4}} \tag{6}$$

where in particular we note that $\gamma_1 = 1$, which is used in the lower bound, and $\gamma_2 \ge 1$ is used for the upper bound.

Generalizing [Ind06, Lemma 2], it follows from Equations (5) and (6) that if $F_p(z) \in [1/2 - c\varepsilon, 1/2 + c\varepsilon]$ for some absolute constant c then $z \in [1 - \varepsilon, 1 + \varepsilon]$, which we will use in what follows. For any $x \in \Psi_k$, we say SAx is *good* if only a $\frac{1}{2} - c_2\varepsilon$ fraction of coordinates in the sketch space are too large or too small, i.e.

$$|\{i : |(SAx)_i| < (1 - \varepsilon)\|Ax\|_p\}| \le \left(\frac{1}{2} - c_2\varepsilon\right) m$$
$$|\{i : |(SAx)_i| > (1 + \varepsilon)\|Ax\|_p\}| \le \left(\frac{1}{2} - c_2\varepsilon\right) m$$

27

for some small constant c_2. If SAx is *good*, then for any y with at most $c_2\varepsilon m$ coordinates larger than $\varepsilon\|Ax\|_p$, we have

$$(1 - 2\varepsilon)\|Ax\|_p \leq \|SAx + y\|_{\mathrm{med}} \leq (1 + 2\varepsilon)\|Ax\|_p. \tag{7}$$

By the p-stability property, $(SAx)_i$ is a p-stable random variable with scale $\|Ax\|_p$, we have that

$$\Pr[|(SAx)_i| < (1 - \varepsilon)\|Ax\|_p] < 1/2 - \Omega(\varepsilon)$$
$$\Pr[|(SAx)_i| > (1 + \varepsilon)\|Ax\|_p] < 1/2 - \Omega(\varepsilon).$$

By a Chernoff bound, for sufficiently small c_2 we have that SAx is *good* with probability at least $1 - \exp(-\Omega(\varepsilon^2 m))$. For our choice of m, we can union bound to get that SAx is *good* simultaneously for all $x \in \mathcal{N}$ with probability at least $1 - \exp(-\Omega(\varepsilon^2 m)) \cdot |\mathcal{N}| \geq 1 - \delta^{\Omega(k)}$. Suppose this event is true.

Then every $y = Ax$ for $x \in \Psi_k$ with $\|y\|_p = 1$ can be expressed as $y = z + \eta$ where $z \in \mathcal{N}$ and $\|\eta\|_p \leq \varepsilon/\tau'$. We have that Sz is *good* and that $\|S\eta\|_\infty \leq \tau'\|\eta\|_p \leq \varepsilon$. Hence by (7),

$$(1 - 2\varepsilon)\|z\|_p \leq \|S(z + \eta)\|_{\mathrm{med}} \leq (1 + 2\varepsilon)\|z\|_p.$$

which implies

$$(1 - 3\varepsilon)\|y\|_p \leq \|Sy\|_{\mathrm{med}} \leq (1 + 3\varepsilon)\|y\|_p.$$

Since S is linear, the restriction to $\|y\|_p = 1$ is not necessary. Rescaling ε concludes the proof. □

B.2 Upper bounds for the ReLU loss function

Notation For a function $f : \mathbb{R} \to \mathbb{R}$ and a vector $y \in \mathbb{R}^n$, we let $f(y) \in \mathbb{R}^n$ denote the entry-wise application of f to y. Let y_i denote the ith entry of y. So $f(y)_i = f(y_i)$. Moreover, let $\|y\|_f = \sum_{i=1}^n f(y_i)$. Let y^+ and y^- denote y restricted to the set of positive and negative entries respectively. Finally, for $A \in \mathbb{R}^{n \times d}$ and $x \in \mathbb{R}^d$, let $\mu(A) = \sup_{x \neq 0} \frac{(Ax)^+}{(Ax)^-}$. When A is clear from the context, we drop A from the notation of μ.

Theorem 9. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$. There exists an oblivious sketch S with $O(\frac{\mu^2 k}{\varepsilon^2} \log(\frac{\mu d}{\varepsilon\delta}))$ rows and an estimator $g_{\mathrm{ReLU}}(SA, Sb, x)$, such that with probability at least $1 - \delta$, we have $\forall x \in \Psi_k :$ $(1 - \varepsilon)\|Ax - b\|_{\mathrm{ReLU}} \leq g_{\mathrm{ReLU}}(SA, Sb, x) \leq (1 + \varepsilon)\|Ax - b\|_{\mathrm{ReLU}}.$*

Proof. By appending b to A and increasing k by 1, if suffices to prove the statement of the theorem for the case $b = 0$. Note that

$$\|Ax\|_{\mathrm{ReLU}} = \|(Ax)^+\|_1 = \frac{\|Ax\|_1 + \mathbf{1}^T Ax}{2}$$

since $\|Ax\|_1 = \|(Ax)^+\|_1 + \|(Ax)^-\|_1$ and $\mathbf{1}^T Ax = \|(Ax)^+\|_1 - \|(Ax)^-\|_1$. From Theorem 8, there exists a sketch S_{ℓ_1} with $O\left((k/\varepsilon^2)\log(d/\varepsilon\delta)\right)$ rows and an estimator $g_{\ell_1}(S_{\ell_1}A, x) = \|S_{\ell_1}Ax\|_{\mathrm{med}}$ such that, with probability at least $1 - \delta$,

$$\forall x \in \Psi_k : (1 - \varepsilon)\|Ax\|_1 \leq g_{\ell_1}(S_{\ell_1}A, x) \leq (1 + \varepsilon)\|Ax\|_1. \tag{8}$$

Moreover, $\mathbf{1}^T A$ can be computed exactly using a single row. Let S be the sketch obtained by combining S_{ℓ_1} and the single row $\mathbf{1}^T$, and let

$$g_{\mathrm{ReLU}}(SA, x) \overset{\mathrm{def}}{=} \frac{g_{\ell_1}(S_{\ell_1}A, x) + \mathbf{1}^T Ax}{2} = \frac{\|S_{\ell_1}Ax\|_{\mathrm{med}} + \mathbf{1}^T Ax}{2}.$$

Now, (8) implies

$$\left| g_{\text{ReLU}}(SA, x) - \|Ax\|_{\text{ReLU}} \right| = \left| \frac{g_{\ell_1}(S_{\ell_1}A, x) + \mathbf{1}^T Ax}{2} - \frac{\|Ax\|_1 + \mathbf{1}^T Ax}{2} \right|$$

$$= \left| \frac{g_{\ell_1}(S_{\ell_1}A, x) - \|Ax\|_1}{2} \right|$$

$$\leq \frac{\varepsilon}{2} \|Ax\|_1. \tag{9}$$

By the definition of μ,

$$\mu + 1 \geq \frac{\|(Ax)^-\|_1}{\|(Ax)^+\|_1} + 1 = \frac{\|Ax\|_1}{\|Ax\|_{\text{ReLU}}}. \tag{10}$$

From (9) and (10), we have

$$\left| g_{\text{ReLU}}(SA, x) - \|Ax\|_{\text{ReLU}} \right| \leq \frac{\varepsilon(\mu + 1)}{2} \|Ax\|_{\text{ReLU}} \leq \varepsilon\mu\|Ax\|_{\text{ReLU}},$$

where the last inequality holds because $\mu \geq 1$. The theorem follows by scaling ε by a factor of $1/\mu$. $\qquad\square$

B.3 Upper bounds for hinge-like loss functions

First we give a lemma on hinge-like functions, which has a similar role to (10) for the ReLU function, and was proven in Corollary 9 of [MMR21].

Lemma 5. *Let f be an (L, a_1, a_2) hinge-like loss function, and let $C = 16\max(1, L, a_1, 1/a_2)^4$. Let $A \in \mathbb{R}^{n \times d}$. Then for any $x \in \mathbb{R}^d$,*

$$\|Ax\|_f \geq \frac{n + \|Ax\|_1}{C\mu}.$$

Proof. We have

$$\|Ax\|_f \geq \sum_{i:[Ax]_i \in [0, 2a_1]} f(Ax)_i + \sum_{i:[Ax]_i \geq 2a_1} f(Ax)_i$$

$$\geq \sum_{i:[Ax]_i \in [0, 2a_1]} a_2 + \sum_{i:[Ax]_i \geq 2a_1} \text{ReLU}(Ax)_i - a_1$$

$$\geq \min\left(\frac{a_2}{2a_1}, \frac{1}{2}\right) \cdot \|(Ax)^+\|_1$$

$$\geq \min\left(\frac{a_2}{2a_1}, \frac{1}{2}\right) \cdot \frac{\|Ax\|_1}{\mu + 1}, \tag{11}$$

where the second inequality holds because f is (L, a_1, a_2) hinge-like and last inequality follows from (10) in Theorem 9.

Let $\gamma \overset{\text{def}}{=} \min\left(\frac{a_2}{2a_1}, \frac{1}{2}\right)$. Now we claim that

$$\|Ax\|_f = \sum_{i=1}^n f(Ax)_i \geq \frac{na_2\gamma}{4\mu \cdot \max(1, L)}.$$

29

If $\sum_{i=1}^{n} f(Ax)_i \geq \frac{na_2}{4}$ then this holds immediately since $\mu(X) \geq 1$, $\max(1, L) \geq 1$ and $\gamma \leq 1$. Otherwise, assume that $\sum_{i=1}^{n} f(Ax)_i \leq \frac{na_2}{4}$. Since $f(z) \geq a_2$ for all $z \geq 0$ and since f is L-Lipschitz, $f(z) \geq \frac{a_2}{2}$ for all $z \geq -\frac{a_2}{2L}$. This implies that Ax has at most $\frac{na_2/4}{a_2/2} = \frac{n}{2}$ entries $\geq -\frac{a_2}{2L}$. Thus, Ax has at least $\frac{n}{2}$ entries $\leq -\frac{a_2}{2L}$ and so $\|(Ax)^-\|_1 \geq \frac{na_2}{4L}$. Thus, by the definition of μ along with (11),

$$\|Ax\|_f \geq \gamma \cdot \|(Ax)^+\|_1 \geq \frac{na_2\gamma}{4\mu L} \geq \frac{na_2\gamma}{4\mu \cdot \max(1, L)}. \tag{12}$$

Combining (11) with (12) gives that

$$\begin{aligned}
\|Ax\|_f &\geq \frac{\gamma \cdot \|Ax\|_1}{2\mu + 2} + \frac{na_2\gamma}{8\mu \cdot \max(1, L)} \\
&\geq (\|Ax\|_1 + n) \cdot \frac{\gamma \cdot \min(1, a_2)}{8\mu \cdot \max(1, L)} \\
&\geq (\|Ax\|_1 + n) \frac{1}{8\mu \cdot \max(1, L) \cdot \max(1, 1/a_2) \cdot \max(2, 2a_1/a_2)} \\
&\geq (\|Ax\|_1 + n) \frac{1}{16\mu \cdot \max(1, L, a_1, 1/a_2)^4}.
\end{aligned}$$

Substituting $C = 16\max(1, L, a_1, 1/a_2)^4$ completes the proof. $\qquad\square$

From Lemma 5, it suffices to approximate $\|Ax\|_f$ within $O((\varepsilon/\mu)(n+\|Ax\|_1))$ to obtain a relative error guarantee. Theorem 9 provides a method to approximate $\|Ax\|_{\mathrm{ReLU}}$ within $O((\varepsilon/\mu)\|Ax\|_1)$. In Theorem 10, we will show that uniform sampling can approximate the difference between $\|Ax\|_f$ and $\|Ax\|_{\mathrm{ReLU}}$ within $O((\varepsilon/\mu)n)$.

Theorem 10. *Let $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$, and let f be an (L, a_1, a_2) hinge-like loss function. There exists an oblivious sketch S with $m = O(\frac{c^{10}\mu^2 k}{\varepsilon^2} \log(\frac{cn\mu d}{\varepsilon\delta}))$ rows, where $c = \max(1, L, a_1, 1/a_2)$, and an estimator $g_f(SA, Sb, x)$, such that, with probability at least $1 - \delta$, we have $\forall x \in \Psi_k$: $(1 - \varepsilon)\|Ax - b\|_f \leq g_f(SA, Sb, x) \leq (1 + \varepsilon)\|Ax - b\|_f$.*

Proof. Again, we may assume that $b = 0$. We have

$$\|Ax\|_f = \sum_{i=1}^{n} f(Ax)_i = \sum_{i=1}^{n} (f(Ax)_i - \mathrm{ReLU}(Ax)_i) + \|Ax\|_{\mathrm{ReLU}}.$$

By (9) in the proof of Theorem 9, there exists a sketch S_{ReLU} with $m_1 = O\left(k\log(d/\varepsilon\delta)/\varepsilon^2\right)$ rows and a function g_{ReLU} such that with probability at least $1 - \delta$,

$$\left|g_{\mathrm{ReLU}}(S_{\mathrm{ReLU}}A, x) - \|Ax\|_{\mathrm{ReLU}}\right| \leq \varepsilon\|Ax\|_1. \tag{13}$$

We give a sketch with m_2 rows to approximate $R(Ax) = \sum_{i=1}^{n} (f(Ax)_i - \mathrm{ReLU}(Ax)_i)$. Consider uniformly sampling the rows of A with replacement. Let $S_u \in \mathbb{R}^{m_2 \times n}$ be the sketching matrix corresponding to uniformly sampling m_2 rows. We will show that

$$R(S_u Ax) = \frac{n}{m_2} \sum_{i=1}^{m_2} (f(S_u Ax)_i - \mathrm{ReLU}(S_u Ax)_i)$$

can approximate $R(Ax)$ within error $O(\varepsilon(n + \|Ax\|_1))$ for a suitable value of m_2, i.e.,

$$|R(S_u Ax) - R(Ax)| \leq O(\varepsilon(n + \|Ax\|_1)). \tag{14}$$

Fix any k-sparse support indexed by $I \subseteq [d]$ with $|I| = k$. We will show (14) for all x in this fixed support and then union bound over all $\binom{d}{k}$ supports. Let A_I be the matrix whose k columns are the columns A_{*i} of A such that $i \in I$. We may assume w.l.o.g. that the columns of A_I are orthonormal, since the set of vectors $\{A_I x \mid x \text{ has support } I\}$ remains unchanged by making them orthonormal. With the assumption, $\|Ax\|_2 = \|A_I x\|_2 = \|x\|_2$ for all x having support I. We consider two cases.

Large Norm In this case, we consider x such that $\|Ax\|_1 \geq na_1/\varepsilon$. Since f is an (L, a_1, a_2) hinge-like loss function, $|f(Ax)_i - \mathrm{ReLU}(Ax)_i| \leq a_1$ for all $1 \leq i \leq n$. Therefore, it holds that $|f(S_u Ax)_i - \mathrm{ReLU}(S_u Ax)_i| \leq a_1$ for all $1 \leq i \leq m_2$ as well. We have

$$\left| R(S_u Ax) - R(Ax) \right| = \left| \frac{n}{m_2} \sum_{i=1}^{m_2} (f(S_u Ax)_i - \mathrm{ReLU}(S_u Ax)_i) - \sum_{i=1}^{n} (f(Ax)_i - \mathrm{ReLU}(Ax)_i) \right|$$

$$\leq \frac{n}{m_2} m_2 a_1 + na_1 = 2na_1 \leq 2\varepsilon \|Ax\|_1. \tag{15}$$

Small Norm Now we consider x such that $\|Ax\|_1 < na_1/\varepsilon$. This implies $\|x\|_2 = \|Ax\|_2 \leq \|Ax\|_1 < na_1/\varepsilon$. We construct an $\frac{\varepsilon}{L+1}$-net $\mathcal{N}_I^k$ in the ℓ_2 norm for all x in the ℓ_2 ball $\mathcal{B}_I = \{x \text{ has support } I \mid \|x\|_2 \leq na_1/\varepsilon\}$. By a standard volume argument, $|\mathcal{N}_I^k| < (3(L+1)na_1/\varepsilon^2)^k$. We next consider a fixed vector x with support I and then union bound over all vectors in the net $\mathcal{N}_I^k$.

Let $X_1, X_2, \ldots, X_{m_2}$ be random variables such that $X_i = f(S_u Ax)_i - \mathrm{ReLU}(S_u Ax)_i$. Since f is an (L, a_1, a_2) hinge-like loss function, $|X_i| \leq a_1$ for all i. By Hoeffding's inequality,

$$\mathbb{P}\left(\left| \frac{n \sum_{i=1}^{m_2} X_i}{m_2} - R(Ax) \right| \geq \varepsilon n \right) = \mathbb{P}\left(\left| \frac{\sum_{i=1}^{m_2} X_i}{m_2} - \frac{R(Ax)}{n} \right| \geq \varepsilon \right) \leq 2\exp\left(-\frac{m_2 \varepsilon^2}{4a_1^2} \right). \tag{16}$$

Note that (16) holds for a fixed vector x. Letting

$$m_2 = O\left(\frac{a_1^2}{\varepsilon^2} \left(k \log \frac{d(L+1)na_1}{\varepsilon k} + \log \frac{1}{\delta} \right) \right)$$

and union bounding over at most $(3(L+1)na_1/\varepsilon^2)^k$ points in $\mathcal{N}_I^k$, we have the probability that (16) holds for all points in $\mathcal{N}_I^k$ is $1 - \delta'$, where

$$\delta' = \left(\frac{3(L+1)na_1}{\varepsilon^2} \right)^k \exp\left(-\frac{m_2 \varepsilon^2}{4a_1^2} \right) \leq \delta \left(\frac{k}{ed} \right)^k.$$

Next we will show that, $\left| R(S_u Ax') - R(Ax') \right| \leq \varepsilon n$ for all x' in $\mathcal{N}_I^k$ implies

$$\left| R(S_u Ax) - R(Ax) \right| \leq 3\varepsilon n \tag{17}$$

for all $x \in B_I$. Let $x' \in \mathcal{N}_I^k$ such that $\|x - x'\|_2 \leq \frac{\varepsilon}{L+1}$. By the triangle inequality,

$$\left| R(S_u Ax) - R(Ax) \right| \leq \left| R(Ax) - R(Ax') \right| + \left| R(S_u Ax) - R(S_u Ax') \right| + \left| R(S_u Ax') - R(Ax') \right|.$$

Since x' is in $\mathcal{N}$, the last term is at most εn. We will bound the other two terms using the fact that f is L-Lipschitz and $\|x - x'\|_2 \leq \frac{\varepsilon}{L+1}$. Applying the triangle inequality again, we have

$$\left| R(Ax) - R(Ax') \right| \leq \left| \sum_{i=1}^{n} \left(\mathrm{ReLU}(Ax)_i - \mathrm{ReLU}(Ax')_i \right) \right| + \left| \sum_{i=1}^{n} \left(f(Ax)_i - f(Ax')_i \right) \right|$$

31

$$\leq \|A(x - x')\|_1 + L\|A(x - x')\|_1$$
$$\leq (1 + L)\sqrt{n}\|x - x'\|_2$$
$$\leq \varepsilon\sqrt{n}.$$

Similarly,

$$\left| R(S_u A x) - R(S_u A x') \right|$$
$$\leq \left| \frac{n}{m_2} \sum_{i=1}^{m_2} \left(\text{ReLU}(S_u A x)_i - \text{ReLU}(S_u A x')_i \right) \right| + \left| \frac{n}{m_2} \sum_{i=1}^{m_2} \left(f(S_u A x)_i - f(S_u A x')_i \right) \right|$$
$$\leq \frac{n}{m_2}\|S_u A(x - x')\|_1 + \frac{n}{m_2} L\|S_u A(x - x')\|_1$$
$$\leq (L+1)\frac{n}{\sqrt{m_2}}\|S_u A(x - x')\|_2$$
$$\leq (L+1)\frac{n}{\sqrt{m_2}}\|A(x - x')\|_2$$
$$\leq (L+1)\frac{n}{\sqrt{m_2}}\|x - x'\|_2$$
$$\leq \varepsilon\frac{n}{\sqrt{m_2}}.$$

This completes the proof of (17) and our argument for the small norm case. Combining (15) and (17), we have that for a fixed support I, with probability at least $1 - \delta \left(\frac{k}{ed}\right)^k$,

$$\left| R(S_u A x) - R(A x) \right| \leq 3\varepsilon n + 2\varepsilon\|A x\|_1.$$

Union bounding over $\binom{d}{k} < (ed/k)^k$ k-sparse supports gives a success probability of at least $1 - \delta$. Let $S = \begin{bmatrix} S_{\text{ReLU}} \\ S_u \end{bmatrix}$. Define

$$g_f(SA, x) \stackrel{\text{def}}{=} R(S_u A x) + g_{\text{ReLU}}(S_{\text{ReLU}} A, x).$$

From (13), with probability at least $1 - 2\delta$,

$$\left| g_f(SA, x) - \|A x\|_f \right| \leq \left| R(S_u A x) - R(A x) \right| + \left| g_{\text{ReLU}}(S_{\text{ReLU}} A, x) - \|A x\|_{\text{ReLU}} \right|$$
$$\leq 3\varepsilon \left(n + \|A x\|_1 \right) \tag{18}$$

By Lemma 5, $\|A x\|_f \geq (n + \|A x\|_1)/(C\mu)$, where $C = 16 \max(1, L, a_1, 1/a_2)^4$. Combining this with (18) gives

$$\left| g_f(SA, x) - \|A x\|_f \right| \leq 3C\mu\varepsilon\|A x\|_f.$$

Finally, scaling ε by $O(1/(C\mu))$, the number of rows in S is

$$m_1 + m_2 = O\left(\frac{C^2\mu^2 k(1 + a_1^2)}{\varepsilon^2} \cdot \log\left(\frac{Cna_1(L+1)\mu d}{\varepsilon\delta} \right) \right)$$
$$= O\left(\frac{c^{10}\mu^2 k}{\varepsilon^2} \cdot \log\left(\frac{cn\mu d}{\varepsilon\delta} \right) \right)$$

with $c = \max(1, L, a_1, 1/a_2)$. $\qquad\square$

For completeness we have the following simple result that yields a connection between minimizing in the sketch space using a k-sparse affine embedding – as in all upper bounds above – and the original problem.

Corollary 11. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$. Let $\| \cdot \| : \mathbb{R}^n \to \mathbb{R}^{\geq 0}$ be any loss function. Let S be an oblivious linear sketch of $[A, b]$, and let $E(SA, Sb, x)$ be an estimator that satisfies $\forall x \in \Psi_k :$ $(1 - \varepsilon)\|Ax - b\| \leq E(SA, Sb, x) \leq (1 + \varepsilon)\|Ax - b\|$. Then $\tilde{x} \in \arg\min_{x \in \Psi_k} E(SA, Sb, x)$ satisfies $\|A\tilde{x} - b\| \leq (1 + O(\varepsilon)) \min_{x \in \Psi_k} \|Ax - b\|$.*

Proof. (of Corollary 11) Let $x^* \in \arg\min_{x \in \Psi_k} \|Ax - b\|$. Then

$$\|A\tilde{x} - b\| \leq E(SA, Sb, \tilde{x})/(1 - \varepsilon) \leq E(SA, Sb, x^*)/(1 - \varepsilon)$$
$$\leq \|Ax^* - b\|(1 - \varepsilon)/(1 - \varepsilon) \leq (1 + 4\varepsilon)\|Ax^* - b\|.$$

$\square$

C   Oblivious sketching for LASSO regression

LASSO regression is a convex relaxation of k-sparse ℓ_2 regression and enjoys large popularity as a heuristic for inducing sparsity and feature selection [Tib96]. Here we give an upper bound for sketching that depends on an ℓ_1 regularization parameter λ, and $\log(d)$.

C.1   Upper bound

Theorem 12. *Consider $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$, and $\lambda \in (0, 1)$. Assume that $\|A\|_2 \leq 1$ and $\|b\|_2 \leq 1$, If $S \in \mathbb{R}^{m \times n}$ is a random Gaussian matrix (i.e., each entry is sampled i.i.d. from $N(0, 1/m)$) then for any $\varepsilon, \delta \in (0, 1)$ and $m = O(\frac{\log d/\delta}{\lambda^2 \cdot \varepsilon^2})$, with probability at least $1 - \delta$, if $\tilde{x} = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2^2 + \lambda\|x\|_1$ then $\|A\tilde{x} - b\|_2^2 + \lambda\|\tilde{x}\|_1 \leq (1 + \varepsilon) \cdot \min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2 + \lambda\|x\|_1$.*

Observe that our constraints on $\|A\|_2, \|b\|_2$ are necessary, since classic LASSO regression is *not scale invariant*. If, for an arbitrarily large factor α, we scale A up by a factor of α then we can keep the same error $\|Ax - b\|_2^2$ while scaling x down by a factor of $1/\alpha$. Thus, the $\lambda \cdot \|x\|_1$ term becomes negligible, and the problem reduces to ordinary least squares regression. Similarly, if we scale b up by a factor α, then the minimal achievable $\|Ax - b\|_2^2$ scales by a factor α^2, while the x achieving this minimum has $\lambda\|x\|_1$ scaled by just a factor α. So again, as α grows arbitrarily large, the problem becomes ordinary least squares regression. It is known that sketching dimension $\Theta(d/\varepsilon)$ is necessary and sufficient for ordinary least squares regression [Sar06; CW09]. Thus, going beyond this requires bounding $\|A\|_2, \|b\|_2$ to ensure that the regularization $\lambda\|x\|_1$ has a non-negligible effect.

Also note that we cannot hope to achieve an $o(d/\varepsilon^2)$ bound for preserving the LASSO cost for all $x \in \mathbb{R}^d$, since for x with large enough $\|x\|_2^2$, and for $b = 0$, the problem becomes equivalent to preserving $\|Ax\|_2^2$, which requires $\Theta(d/\varepsilon^2)$ sketch size [NN14].

Proof of Theorem 12. Let $OPT = \min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2 + \lambda\|x\|_1$. Observe that since $\|b\|_2 \leq 1$, $OPT \leq \|b\|_2^2 + \lambda\|0\|_1 \leq 1$. Via the standard Johnson-Lindenstrauss lemma, we have that with high probability, for $x^* = \arg\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2 + \lambda\|x\|_1$, $\|SAx^* - Sb\|_2^2 + \lambda\|x^*\|_1 \leq (1 + \varepsilon) \cdot OPT$.

Thus, we must have $\|\tilde{x}\|_1 \leq \frac{(1+\varepsilon) \cdot OPT}{\lambda} \leq \frac{2 \cdot OPT}{\lambda}$, as otherwise we would have $\|SA\tilde{x} - Sb\|_2^2 + \lambda\|\tilde{x}\|_1 \geq \|SAx^* - Sb\|_2^2 + \lambda\|x^*\|_1$, contradicting the fact that $\tilde{x}$ is a minimizer for the sketched problem. For the same reason, we must have $\|SA\tilde{x} - Sb\|_2 \leq \|Sb\|_2 \leq 2$.

Let $\mathcal{T} = \{y = Ax - b : \|x\|_1 \leq \frac{2 \cdot OPT}{\lambda}\}$. Let $\mathcal{T}' = \{y = Ax - b : \|x\|_1 = 1\}$. Observe that by our assumption that $\|A\|_2 \leq 1$, each column a_i of A has $\|a_i\|_2 \leq 1$. Thus, $\mathcal{T}'$ is the convex hull

33

of $2d + 1$ points in the unit ball: $a_1, -a_1, a_2, -a_2, \ldots a_d, -a_d, b$. By Corollary 3.2 of [NN19], for $m = O\left(\frac{\log(d/\delta)}{\lambda^2 \cdot \varepsilon^2}\right)$, with probability at least $1 - \delta$, for all $y' \in \mathcal{T}'$,

$$|\|Sy\|_2 - \|y\|_2| \leq \varepsilon\lambda.$$

Note that any $y \in \mathcal{T}$ can be written as $\alpha \cdot y'$ for $y' \in \mathcal{T}'$, for some $\alpha \leq \frac{2 \cdot OPT}{\lambda}$. Thus, we have that with probability at least $1 - \delta$, for all $y \in \mathcal{T}$,

$$|\|Sy\|_2 - \|y\|_2| \leq 2\varepsilon \cdot OPT.$$

In particular, for $\tilde{x} = \operatorname{argmin}_{x \in \mathbb{R}^d} \|SAx - Sb\|_2^2 + \lambda\|x\|_1$, we have $A\tilde{x} - b \in \mathcal{T}$ so this gives

$$\left|\|SAx - Sb\|_2^2 - \|Ax - b\|_2^2\right| \leq |\|SAx - Sb\|_2 - \|Ax - b\|_2| \cdot |\|SAx - Sb\|_2 + \|Ax - b\|_2|$$
$$\leq 12\varepsilon \cdot OPT, \tag{19}$$

where the second inequality uses that $\|SA\tilde{x} - Sb\|_2 \leq 2 \cdot OPT \leq 2$ and thus $\|A\tilde{x} - b\|_2 \leq (2 + 2\varepsilon) \cdot OPT \leq 4$. Finally, using (19) we have:

$$\|A\tilde{x} - b\|_2^2 + \lambda\|\tilde{x}\|_1 \leq \|SA\tilde{x} - Sb\|_2^2 + \lambda\|\tilde{x}\|_1 + 5\varepsilon \cdot OPT$$
$$\leq \|SA\tilde{x}^* - Sb\|_2^2 + \lambda\|x^*\|_1 + 12\varepsilon \cdot OPT$$
$$\leq (1 + 13\varepsilon) \cdot [\|SA\tilde{x}^* - Sb\|_2^2 + \lambda\|x^*\|_1].$$

This completes the theorem after adjusting ε by a constant factor. $\qquad\square$

C.2 Lower bound

Theorem 13. *Let $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$ with bounded $\|A\|_2 \leq 1$ and $\|b\|_2 \leq 1$, and let $\lambda \in (0, 1)$. Suppose $S \in \mathbb{R}^{m \times n}$ is an oblivious linear sketch for LASSO regression with an estimator $E_\lambda(SA, Sb, x)$, such that with constant probability $\tilde{x} \in \operatorname{argmin}_{x \in \Psi_k} E_\lambda(SA, Sb, x)$ satisfies $\|A\tilde{x} - b\|_2^2 + \lambda\|\tilde{x}\|_1 \leq (1 + \varepsilon) \cdot \min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2 + \lambda\|x\|_1$. Then $m = \Omega(\frac{\log(\lambda d)}{\lambda^2})$.*

Proof of Theorem 13. We will prove this similar to the lower bound in Theorem 1. We will take $[b\ A] = X \sim \frac{1}{\sqrt{n}}G(I + zz^T)^{1/2}$, where z is as in Corollary 15 with $\varepsilon = 1/2$. Without loss of generality, we can assume that S has orthonormal rows and so $SG(I + zz^T)^{1/2}$ has the same distribution as $G(I + zz^T)^{1/2}$ with fewer rows. We also let $\lambda = 1/(2\sqrt{k})$. The normalization factor $\frac{1}{\sqrt{n}}$ ensures that norm condition $\|A\|_2, \|b\|_2 \leq 1$ holds with high probability. We then have with high probability

$$\|Ax - b\|_2^2 + \lambda\|x\|_1 \approx 1 + \|x\|_2^2 + (1 - \varepsilon v^T x)^2 + \lambda\|x\|_1 =: L(x).$$

The approximation error (and the probability) above can be made arbitrarily small by taking n to be sufficiently large.

We note that $L(x)$ is a 1-strongly convex function, and so

$$L(\hat{x}) \geq L(x^*) + \|\hat{x} - x^*\|_2^2.$$

for any $\hat{x}$, where x^* is the minimizer of $L(x)$. By a straightforward computation, we also get that $x^* = v/5$ and $L(x^*) = 1.95$. Suppose that $L(\hat{x}) \leq (1 + c_1)L(x^*)$ for a sufficiently small c_1. Then we have

$$(1 + c_1)L(x^*) \geq L(x^*) + \|\hat{x} - x^*\|_2^2.$$

This implies that $\|\hat{x} - x^*\|_2^2 \leq c_1 L(x^*) \leq 1.95 \cdot c_1$. Therefore, by choosing c_1 to be a small enough constant, we can recover a $19/20$ fraction of $\operatorname{supp}(x^*) = \operatorname{supp}(v)$. Corollary 15 then gives the required lower bound of $\Omega(k \log(d/k))$ on the size of the sketch, where k is set to $1/(4\lambda^2)$ corresponding to our choice of λ. $\qquad\square$

D   Separating sparse recovery and sparse regression

We now give a separation between the k-sparse recovery problem and k-sparse regression problems. Combined with our lower bounds for k-sparse regression, this shows that the k-sparse recovery problem is a strictly easier problem. Our upper bound below matches the lower bound for k-sparse outputs of [PW11] up to a $\log(k/\varepsilon)$ factor, improving the naïve bound of $O(k(\log d)/\varepsilon^2)$.

In the k-sparse recovery problem, one seeks to sketch a vector $x \in \mathbb{R}^d$ so as to output a k-sparse $\hat{x} \in \mathbb{R}^d$ so that

$$\|x - \hat{x}\|_2^2 \leq (1 + \varepsilon)\|x - x_k\|_2^2, \tag{20}$$

where x_k consists of the top k entries in magnitude of x, breaking ties arbitrarily. We need the following theorem about CountSketch.

Theorem 16. *([CCF04]) There is a distribution on sketching matrices $S \in \mathbb{R}^{bt \times d}$, called Count-Sketch matrices, which is parameterized by the number b of buckets and the number t of tables. For a vector x, there is a procedure which, given $S \cdot x$ and a coordinate $i \in [d]$, returns an estimate $\hat{x}_i$ for which*

$$|\hat{x}_i - x_i| \leq C \cdot \|x - x_k\|_2/\sqrt{b},$$

with failure probability at most $2^{-C't}$, where $C, C' > 0$ are absolute constants.

Consider the following procedure:

1. Run CountSketch with $b = O(k/\varepsilon)$ buckets and $t = O(\log d)$ tables, and let A be the set of indices $i \in \{1, 2, \ldots, d\}$ for which the corresponding estimates $\hat{x}_i'$ are among the largest $O(k/\varepsilon)$ in magnitude. Here we use $\hat{x}_i'$ to denote the estimates returned by Theorem 16 to distinguish them from the estimates returned in the next step.

2. In parallel, run CountSketch with $b = O(k/\varepsilon^2)$ buckets and $t = O(\log(k/\varepsilon))$ tables, and compute estimates $\hat{x}_i$ for each $i \in A$. Let B be the set of the top k magnitude estimates $\hat{x}_i$, restricted to $i \in A$.

3. Return the vector $\hat{x}$ supported on B with corresponding estimates $\hat{x}_i$ for each $i \in B$.

Theorem 14. *On input $x \in \mathbb{R}^d$, the above sparse recovery scheme uses $O(k\log(k/\varepsilon)/\varepsilon^2 + k\log(d)/\varepsilon)$ measurements and, with probability at least $1 - 1/\operatorname{poly}(d) - 1/\operatorname{poly}(k/\varepsilon)$, returns a k-sparse vector $\hat{x} \in \Psi_k$ satisfying $\|x - \hat{x}\|_2^2 \leq (1 + \varepsilon)\min_{x_k \in \Psi_k} \|x - x_k\|_2^2$.*

Proof. Since the sketch S is linear, we can assume, w.l.o.g., that $\|x - x_k\|_2^2 = k$. Then by Theorem 16, with probability $1 - 1/\operatorname{poly}(d)$, the CountSketch in Step 1 returns estimates $\hat{x}_i'$ satisfying

$$|\hat{x}_i' - x_i| \leq \sqrt{\varepsilon}$$

simultaneously for all $i = 1, 2, \ldots, d$. It follows that for every i for which $|x_i| \geq 3\sqrt{\varepsilon}$, we have that $i \in A$. To see this, note that if $|x_i| \geq 3\sqrt{\varepsilon}$, then $|\hat{x}_i'| \geq 2\sqrt{\varepsilon}$. On the other hand, if $|\hat{x}_j'| \geq 2\sqrt{\varepsilon}$ for some $j \in [d]$, then $|x_j| \geq \sqrt{\varepsilon}$. The number of j for which $|x_j| \geq \sqrt{\varepsilon}$ is at most $k + k/\varepsilon$, given that $\|x - x_k\|_2^2 = k$. Also, the number of i for which $|\hat{x}_i| \geq 2\sqrt{\varepsilon}$ is at most $k + k/(4\varepsilon)$. So if $|A| = O(k/\varepsilon)$ for a sufficiently large constant in the big-Oh, we have that if $|x_i| \geq 3\sqrt{\varepsilon}$, then $i \in A$.

In Step 2, and by a union bound over all $i \in A$, we have that with probability $1 - 1/\operatorname{poly}(k/\varepsilon)$, simultaneously for every $i \in A$, $|\hat{x}_i - x_i| \leq \varepsilon$. Letting H be the set of the top k magnitude

35

coordinates of x, we have $\|x_{H\setminus A} - \hat{x}_{H\setminus A}\|_2^2 \le k \cdot 9 \cdot \varepsilon$, since any $i \in H \setminus A$ necessarily satisfies $|x_i| < 3\sqrt{\varepsilon}$ and $\hat{x}_i = 0$.

We also have $\|x_{H\cap B} - \hat{x}_{H\cap B}\|_2^2 \le k \cdot \varepsilon^2$, since for each $i \in H \cap B$, we have $|\hat{x}_i - x_i| \le \varepsilon$.

Finally, consider those $i \in H \cap (A \setminus B)$. For each such i, there necessarily exists a $j = j(i) \in A \setminus H$ for which $|\hat{x}_{j(i)}| \ge |\hat{x}_i|$ and since $|\hat{x}_{j(i)} - x_{j(i)}| \le \varepsilon$ and $|\hat{x}_i - x_i| \le \varepsilon$, this implies $|x_{j(i)}| \ge |x_i| - 2\varepsilon$. Note also by definition of H that $|x_{j(i)}| \le |x_i|$. Consequently, the sketch solution $\hat{x}$ pays at most

$$(x_{j(i)} - \hat{x}_{j(i)})^2 + x_i^2 \le \varepsilon^2 + (x_{j(i)} + 2\varepsilon)^2 = O(\varepsilon^2 + x_{j(i)}^2 + |x_{j(i)}|\varepsilon),$$

on this coordinate, whereas the optimal non-sketched solution x_k pays $x_{j(i)}^2$. As $|A \setminus B| \le k$, the total additional cost the sketched solution pays over the optimal solution is at most

$$O\left(k\varepsilon^2 + \varepsilon \sum_{i \in H \cap (A\setminus B)} |x_{j(i)}|\right). \tag{21}$$

Finally, note that $\sum_{i \in H \cap (A\setminus B)} |x_{j(i)}|$ is at most the ℓ_1-norm of the largest k coordinates in magnitude not in H. Since the ℓ_2-norm of such coordinates is at most $\sqrt{k}$, the ℓ_1-norm of such coordinates is at most k. Combining with (21), the total additional error the sketched solution pays is $O(\varepsilon k)$.

It follows that

$$\|x - \hat{x}\|_2^2 \le \|x_{H\setminus A} - \hat{x}_{H\setminus A}\|_2^2 + \|x_{H\cap B} - \hat{x}_{H\cap B}\|_2^2 + \|x - x_k\|_2^2 + O(\varepsilon k) \le (1 + O(\varepsilon))\|x - x_k\|_2^2,$$

and (20) follows by rescaling ε by a constant factor. The total number of measurements and overall failure probability follow by Theorem 16. $\qquad\square$

E   Information theoretic basics

We require the following notions from information theory.

Definition 3 (Entropy and Mutual Information). *The entropy of a random variable X over some support S is*

$$H(X) = \sum_{i \in S} p_i \log_2 \frac{1}{p_i}.$$

Given two random variables X and Y, the conditional entropy is

$$H(X|Y) = \sum_y H(X|Y = y) \cdot \mathbb{P}[Y = y]$$

and their joint entropy is

$$H(X, Y) = \sum_{x,y} \mathbb{P}\left[X = x \bigwedge Y = y\right] \log_2 \frac{1}{\mathbb{P}[X = x \wedge Y = y]}.$$

The mutual information of two random variables is

$$I(X;Y) = H(X) - H(X|Y)$$

Finally, we require Fano's inequality.

Lemma 6 (Fano's Inequality). *Let X be a random variable chosen from domain $\mathcal{X}$ according to distribution μ_X and let Y be a random variable chosen from domain $\mathcal{Y}$ according to distribution μ_Y. Then for any reconstruction function $g : Y \to X$ with error ε_g, it holds that*

$$H(X|Y) \le H(\varepsilon_g) + \varepsilon_g \log(|\mathcal{X}| - 1).$$