Lower Bounds for Sparse Recovery[*]

Khanh Do Ba
MIT CSAIL

Piotr Indyk
MIT CSAIL

Eric Price
MIT CSAIL

David P. Woodruff
IBM Almaden

Abstract

We consider the following k-sparse recovery problem: design an $m \times n$ matrix A, such that for any signal x, given Ax we can efficiently recover $\hat{x}$ satisfying $\|x - \hat{x}\|_1 \le C \min_{k\text{-sparse } x'} \|x - x'\|_1$. It is known that there exist matrices A with this property that have only $O(k \log(n/k))$ rows.

In this paper we show that this bound is tight. Our bound holds even for the more general *randomized* version of the problem, where A is a random variable, and the recovery algorithm is required to work for any fixed x with constant probability (over A).

1 Introduction

In recent years, a new "linear" approach for obtaining a succinct approximate representation of n-dimensional vectors (or signals) has been discovered. For any signal x, the representation is equal to Ax, where A is an $m \times n$ matrix. The vector Ax is often referred to as the *measurement vector* or *sketch* of x. Although m is typically much smaller than n, the sketch Ax contains plenty of useful information about the signal x. A particularly useful and well-studied problem is that of stable sparse recovery: given Ax, recover a k-*sparse* vector $\hat{x}$ (i.e., having at most k non-zero components) such that

$$\|x - \hat{x}\|_p \le C \min_{k\text{-sparse } x'} \|x - x'\|_q \tag{1.1}$$

for some norm parameters p and q and an approximation factor $C = C(k)$. Sparse recovery has applications to numerous areas such as data stream computing [Mut03, Ind07] and compressed sensing [CRT06, Don06, DDT$^+$08].

It is known that there exist matrices A and associated recovery algorithms that produce approximations $\hat{x}$ satisfying Equation (1.1) with $p = q = 1$ (i.e., the "ℓ_1/ℓ_1 guarantee"), constant C and sketch length $m = O(k \log(n/k))$. In particular, a random Gaussian matrix [CRT06][1] or a random sparse binary matrix ([BGI$^+$08], building on [CCFC04, CM05]) has this property with overwhelming probability. In comparison, using a *non-linear* approach, one can obtain a shorter sketch of length $O(k)$: it suffices to store the k coefficients with the largest absolute values, together with their indices.

Surprisingly, it was not known if the $O(k \log(n/k))$ bound for linear sketching could be improved upon[2] , although $O(k)$ sketch length was known to suffice if the signal vectors x are required to be *exactly* k-sparse. This raised hope that the $O(k)$ bound might be achievable even for general vectors x. Such a scheme would have been of major practical interest, since the sketch length determines the compression ratio, and for large n any extra $\log n$ factor worsens that ratio tenfold.

In this paper we show that, unfortunately, such an improvement is not possible. We address two types of recovery schemes:

- A *deterministic* one, which involves a fixed matrix A and a recovery algorithm which work for all signals x. The aforementioned results of [CRT06] and others are examples of such schemes.

- A *randomized* one, where the matrix A is chosen at random from some distribution, and for each signal x the recovery procedure is correct with constant probability. Some of the early schemes proposed in the data stream literature (e.g., [CCFC04, CM05]) belong to this category.

Our main result is that, even in the randomized case, the sketch length m must be at least $\Omega(k \log(n/k))$. By the aforementioned result of [CRT06] this bound is tight.

[*] This research has been supported in part by David and Lucille Packard Fellowship, MADALGO (Center for Massive Data Algorithmics, funded by the Danish National Research Association) and NSF grant CCF-0728645. E. Price has been supported in part by Cisco Fellowship.

[1] In fact, they even achieve a somewhat stronger ℓ_2/ℓ_1 guarantee, see Section 1.2.

[2] The lower bound of $\Omega(k \log(n/k))$ was known to hold for specific recovery algorithms, specific matrix types, or other recovery scenarios. See Section 1.2 for an overview.

Thus, our results show that the linear compression is inherently more costly than the simple non-linear approach.

1.1 Our techniques On a high level, our approach is simple and natural, and utilizes the packing approach: we show that any two "sufficiently" different vectors x and x' are mapped to images Ax and Ax' that are "sufficiently" different themselves, which requires that the image space is "sufficiently" high-dimensional. However, the actual arguments are somewhat subtle.

Consider first the (simpler) deterministic case. We focus on signals $x = y + z$, where y can be thought of as the "head" of the signal and z as the "tail". The "head" vectors y come from a set Y that is a binary error-correcting code, with a minimum distance $\Omega(k)$, where each codeword has weight k. On the other hand, the "tail" vectors z come from an ℓ_1 ball (say B) with a radius that is a small fraction of k. It can be seen that for any two elements $y, y' \in Y$, the balls $y + B$ and $y' + B$, as well as their images, must be disjoint. At the same time, since all vectors x live in a "large" ℓ_1 ball B' of radius $O(k)$, all images Ax must live in a set AB'. The key observation is that the set AB' is a scaled version of $A(y + B)$ and therefore the ratios of their volumes can be bounded by the scaling factor to the power of the dimension m. Since the number of elements of Y is large, this gives a lower bound on m.

Unfortunately, the aforementioned approach does not seem to extend to the randomized case. A natural approach would be to use Yao's principle, and focus on showing a lower bound for a scenario where the matrix A is fixed while the vectors $x = y + z$ are "random". However, this approach fails, in a very strong sense. Specifically, we are able to show that there is a distribution over matrices A with *only* $O(k)$ rows so that for a fixed $y \in Y$ and z chosen uniformly at random from the small ball B, we can recover y from $A(y + z)$ with high probability. In a nutshell, the reason is that a random vector from B has an ℓ_2 norm that is much smaller than the ℓ_2 norm of elements of Y (even though the ℓ_1 norms are comparable). This means that the vector x is "almost" k-sparse (in the ℓ_2 norm), which enables us to achieve the $O(k)$ measurement bound.

Instead, we resort to an altogether different approach, via *communication complexity* [KN97]. We start by considering a "discrete" scenario where both the matrix A and the vectors x have entries restricted to the polynomial range $\{-n^c \dots n^c\}$ for some $c = O(1)$. In other words, we assume that the matrix and vector entries can be represented using $O(\log n)$ bits. In this setting we show the following: there is a method for encoding a sequence of $d = O(k \log(n/k) \log n)$ bits into a vector x, so that any sparse recovery algorithm can recover that sequence given Ax. Since each entry of Ax conveys only $O(\log n)$ bits, it follows that the number m of rows of A must be $\Omega(k \log(n/k))$.

The encoding is performed by taking

$$x = \sum_{j=1}^{\log n} D^j x_j,$$

where $D = O(1)$ and the x_j's are chosen from the error-correcting code Y defined as in the deterministic case. The intuition behind this approach is that a good ℓ_1/ℓ_1 approximation to x reveals most of the bits of $x_{\log n}$. This enables us to identify $x_{\log n}$ exactly using error correction. We could then compute $Ax - Ax_{\log n} = A(\sum_{j=1}^{\log n - 1} D^j x_j)$, and identify $x_{\log n - 1} \dots x_1$ in a recursive manner. The only obstacle to completing this argument is that we would need the recovery algorithm to work for *all* x_i, which would require lower probability of algorithm failure (roughly $1/\log n$). To overcome this problem, we replace the encoding argument by a reduction from a related communication complexity problem called Augmented Indexing. This problem has been used in the data stream literature [CW09, KNW10] to prove lower bounds for linear algebra and norm estimation problems. Since the problem has communication complexity of $\Omega(d)$, the conclusion follows.

We apply the argument to arbitrary matrices A by representing them as a sum $A' + A''$, where A' has $O(\log n)$ bits of precision and A'' has "small" entries. We then show that $A'x = A(x + s)$ for some s with $\|s\|_1 < n^{-\Omega(1)} \|x\|_1$. In the communication game, this means we can transmit $A'x$ and recover $x_{\log n}$ from $A'(\sum_{j=1}^{\log n} D^j x_j) = A(\sum_{j=1}^{\log n} D^j x_j + s)$. This means that the Augmented Indexing reduction applies to arbitrary matrices as well.

1.2 Related Work There have been a number of earlier works that have, directly or indirectly, shown lower bounds for various models of sparse recovery and certain classes of matrices and algorithms. Specifically, one of the most well-known recovery algorithms used in compressed sensing is ℓ_1-minimization, where a signal $x \in \mathbb{R}^n$ measured by matrix A is reconstructed as

$$\hat{x} := \operatorname*{arg\,min}_{x' :\, Ax' = Ax} \|x'\|_1.$$

Kashin and Temlyakov [KT07] gave a characterization of matrices A for which the above recovery algorithm yields the ℓ_2/ℓ_1 guarantee, i.e.,

$$\|x-\hat{x}\|_2 \le Ck^{-1/2} \min_{k\text{-sparse } x'} \|x-x'\|_1$$

for some constant C, from which it can be shown that such an A must have $m = \Omega(k \log(n/k))$ rows.

Note that the ℓ_2/ℓ_1 guarantee is somewhat stronger than the ℓ_1/ℓ_1 guarantee investigated in this paper. Specifically, it is easy to observe that if the approximation $\hat{x}$ itself is required to be $O(k)$-sparse, then the ℓ_2/ℓ_1 guarantee implies the ℓ_1/ℓ_1 guarantee (with a somewhat higher approximation constant). For the sake of simplicity, in this paper we focus mostly on the ℓ_1/ℓ_1 guarantee. However, our lower bounds apply to the ℓ_2/ℓ_1 guarantee as well: see footnote on page 7.

On the other hand, instead of assuming a specific recovery algorithm, Wainwright [Wai07] assumes a specific (randomized) measurement matrix. More specifically, the author assumes a k-sparse binary signal $x \in \{0,\alpha\}^n$, for some $\alpha > 0$, to which is added i.i.d. standard Gaussian noise in each component. The author then shows that with a random Gaussian matrix A, with each entry also drawn i.i.d. from the standard Gaussian, we cannot hope to recover x from Ax with any sub-constant probability of error unless A has $m = \Omega(\frac{1}{\alpha^2} \log \frac{n}{k})$ rows. The author also shows that for $\alpha = \sqrt{1/k}$, this is tight, i.e., that $m = \Theta(k \log(n/k))$ is both necessary and sufficient. Although this is only a lower bound for a specific (random) matrix, it is a fairly powerful one and provides evidence that the often observed upper bound of $O(k \log(n/k))$ is likely tight.

More recently, Dai and Milenkovic [DM08], extending on [EG88] and [FR99], showed an upper bound on superimposed codes that translates to a lower bound on the number of rows in a compressed sensing matrix that deals only with k-sparse signals but can tolerate measurement noise. Specifically, if we assume a k-sparse signal $x \in ([-t,t] \cap \mathbb{Z})^n$, and that arbitrary noise $\mu \in \mathbb{R}^n$ with $\|\mu\|_1 < d$ is added to the measurement vector Ax, then if exact recovery is still possible, A must have had $m \ge Ck \log n / \log k$ rows, for some constant $C = C(t,d)$ and sufficiently large n and k.[3]

2 Preliminaries

In this paper we focus on recovering sparse approximations $\hat{x}$ that satisfy the following C-approximate ℓ_1/ℓ_1 guarantee with sparsity parameter k:

$$\|x-\hat{x}\|_1 \le C \min_{k\text{-sparse } x'} \|x-x'\|_1 . \tag{2.2}$$

We define a C-approximate *deterministic* ℓ_1/ℓ_1 recovery algorithm to be a pair $(A, \mathscr{A})$ where A is an $m \times n$ observation matrix and $\mathscr{A}$ is an algorithm that, for any x, maps Ax (called the *sketch* of x) to some $\hat{x}$ that satisfies Equation (2.2).

We define a C-approximate *randomized* ℓ_1/ℓ_1 recovery algorithm to be a pair $(A, \mathscr{A})$ where A is a *random variable* chosen from some distribution over $m \times n$ measurement matrices, and $\mathscr{A}$ is an algorithm which, for any x, maps a pair (A, Ax) to some $\hat{x}$ that satisfies Equation (2.2) with probability at least 3/4.

We use $B_p^n(r)$ to denote the ℓ_p ball of radius r in $\mathbb{R}^n$; we skip the superscript n if it is clear from the context.

For any vector x, we use $\|x\|_0$ to denote the "ℓ_0 norm of x", i.e., the number of non-zero entries in x.

3 Deterministic Lower Bound

We will prove a lower bound on m for any C-approximate deterministic recovery algorithm. First we use a discrete volume bound (Lemma 3.1) to find a large set Y of points that are at least k apart from each other. Then we use another volume bound (Lemma 3.2) on the images of small ℓ_1 balls around each point in Y. If m is too small, some two images collide. But the recovery algorithm, applied to a point in the collision, must yield an answer close to two points in Y. This is impossible, so m must be large.

Lemma 3.1. *(Gilbert-Varshamov) For any $q, k \in \mathbb{Z}^+, \epsilon \in \mathbb{R}^+$ with $\epsilon < 1 - 1/q$, there exists a set $Y \subset \{0,1\}^{qk}$ of binary vectors with exactly k ones, such that Y has minimum Hamming distance $2\epsilon k$ and*

$$\log |Y| > (1 - H_q(\epsilon)) k \log q$$

where H_q is the q-ary entropy function $H_q(x) = -x \log_q \frac{x}{q-1} - (1-x) \log_q(1-x)$.

See appendix for proof.

Lemma 3.2. *Take an $m \times n$ real matrix A, positive reals ϵ, p, λ, and $Y \subset B_p^n(\lambda)$. If $|Y| > (1+1/\epsilon)^m$, then there exist $z, \overline{z} \in B_p^n(\epsilon\lambda)$ and $y, \overline{y} \in Y$ with $y \ne \overline{y}$ and $A(y+z) = A(\overline{y} + \overline{z})$.*

Proof. If the statement is false, then the images of all $|Y|$ balls $\{y + B_p^n(\epsilon\lambda) \mid y \in Y\}$ are disjoint. However, those balls all lie within $B_p^n((1+\epsilon)\lambda)$, by the bound on the norm of Y. A volume argument gives the result, as follows.

[3] Here A is assumed to have its columns normalized to have ℓ_1-norm 1. This is natural since otherwise we could simply scale A up to make the image points Ax arbitrarily far apart, effectively nullifying the noise.

Let $S = AB_p^n(1)$ be the image of the n-dimensional ball of radius 1 in m-dimensional space. This is a polytope with some volume V. The image of $B_p^n(\epsilon\lambda)$ is a linearly scaled S with volume $(\epsilon\lambda)^m V$, and the volume of the image of $B_p^n((1+\epsilon)\lambda)$ is similar with volume $((1+\epsilon)\lambda)^m V$. If the images of the former are all disjoint and lie inside the latter, we have $|Y|\,(\epsilon\lambda)^m V \leq ((1+\epsilon)\lambda)^m V$, or $|Y| \leq (1+1/\epsilon)^m$. If Y has more elements than this, the images of some two balls $y + B_p^n(\epsilon\lambda)$ and $\overline{y} + B_p^n(\epsilon\lambda)$ must intersect, implying the lemma.

Theorem 3.1. *Any C-approximate deterministic recovery algorithm must have*

$$m \geq \frac{1 - H_{\lfloor n/k \rfloor}(1/2)}{\log(4+2C)} k \log \left\lfloor \frac{n}{k} \right\rfloor .$$

Proof. Let Y be a maximal set of k-sparse n-dimensional binary vectors with minimum Hamming distance k, and let $\gamma = \frac{1}{3+2C}$. By Lemma 3.1 with $q = \lfloor n/k \rfloor$ we have $\log|Y| > (1 - H_{\lfloor n/k \rfloor}(1/2))k \log \lfloor n/k \rfloor$.

Suppose that the theorem is not true; then $m < \log|Y| / \log(4+2C) = \log|Y| / \log(1+1/\gamma)$, or $|Y| > (1+\frac{1}{\gamma})^m$. Hence Lemma 3.2 gives us some $y, \overline{y} \in Y$ and $z, \overline{z} \in B_1(\gamma k)$ with $A(y+z) = A(\overline{y}+\overline{z})$.

Let w be the result of running the recovery algorithm on $A(y+z)$. By the definition of a deterministic recovery algorithm, we have

$$\begin{aligned}
&\|y+z-w\|_1 \leq C \min_{k\text{-sparse } y'} \|y+z-y'\|_1 \\
&\|y-w\|_1 - \|z\|_1 \leq C\,\|z\|_1 \\
&\|y-w\|_1 \leq (1+C)\,\|z\|_1 \leq (1+C)\gamma k = \tfrac{1+C}{3+2C}\,k,
\end{aligned}$$

and similarly $\|\overline{y}-w\|_1 \leq \frac{1+C}{3+2C}\,k$, so

$$\|y-\overline{y}\|_1 \leq \|y-w\|_1 + \|\overline{y}-w\|_1 = \frac{2+2C}{3+2C}k < k.$$

But this contradicts the definition of Y, so m must be large enough for the guarantee to hold.

Corollary 3.1. *If C is a constant bounded away from zero, then $m = \Omega(k\log(n/k))$.*

4 Randomized Upper Bound for Uniform Noise

The standard way to prove a randomized lower bound is to find a distribution of hard inputs, and to show that any deterministic algorithm is likely to fail on that distribution. In our context, we would like to define a "head" random variable y from a distribution Y and a "tail" random variable z from a distribution Z, such that any algorithm given the sketch of $y+z$ must recover an incorrect y with non-negligible probability.

Using our deterministic bound as inspiration, we could take Y to be uniform over a set of k-sparse binary vectors of minimum Hamming distance k and Z to be uniform over the ball $B_1(\gamma k)$ for some constant $\gamma > 0$. Unfortunately, as the following theorem shows, one can actually perform a recovery of such vectors using only $O(k)$ measurements; this is because $\|z\|_2$ is very small (namely, $\tilde{O}(k/\sqrt{n})$) with high probability.

Theorem 4.1. *Let $Y \subset \mathbb{R}^n$ be a set of signals with the property that for every distinct $y_1, y_2 \in Y$, $\|y_1 - y_2\|_2 \geq r$, for some parameter $r > 0$. Consider "noisy signals" $x = y + z$, where $y \in Y$ and z is a "noise vector" chosen uniformly at random from $B_1(s)$, for another parameter $s > 0$. Then using an $m \times n$ Gaussian measurement matrix $A = (1/\sqrt{m})(g_{ij})$, where g_{ij}'s are i.i.d. standard Gaussians, we can recover $y \in Y$ from $A(y+z)$ with probability $1 - 1/n$ (where the probability is over both A and z), as long as*

$$s \leq O\left(\frac{r m^{1/2} n^{1/2-1/m}}{|Y|^{1/m} \log^{3/2} n}\right).$$

To prove the theorem we will need the following two lemmas.

Lemma 4.1. *For any $\delta > 0$, $y_1, y_2 \in Y$, $y_1 \neq y_2$, and $z \in \mathbb{R}^n$, each of the following holds with probability at least $1 - \delta$:*

- $\|A(y_1 - y_2)\|_2 \geq \frac{\delta^{1/m}}{3}\|y_1 - y_2\|_2$, *and*
- $\|Az\|_2 \leq (\sqrt{(8/m)\log(1/\delta)} + 1)\|z\|_2$.

See the appendix for the proof.

Lemma 4.2. *A random vector z chosen uniformly from $B_1(s)$ satisfies*

$$\Pr[\|z\|_2 > \alpha s \log n/\sqrt{n}] < 1/n^{\alpha-1}.$$

See the appendix for the proof.

Proof of theorem. In words, Lemma 4.1 says that A cannot bring faraway signal points too close together, and cannot blow up a small noise vector too much. Now, we already assumed the signals to be far apart, and Lemma 4.2 tells us that the noise is indeed small (in ℓ_2 distance). The result is that in the image space, the noise is not enough to confuse different signals. Quantitatively, applying the second part of Lemma

4.1 with $\delta = 1/n^2$, and Lemma 4.2 with $\alpha = 3$, gives us

(4.3)
$$\|Az\|_2 \leq O\left(\frac{\log^{1/2} n}{m^{1/2}}\right)\|z\|_2 \leq O\left(\frac{s \log^{3/2} n}{(mn)^{1/2}}\right)$$

with probability $\geq 1 - 2/n^2$. On the other hand, given signal $y_1 \in Y$, we know that every other signal $y_2 \in Y$ satisfies $\|y_1 - y_2\|_2 \geq r$, so by the first part of Lemma 4.1 with $\delta = 1/(2n|Y|)$, together with a union bound over every $y_2 \in Y$,

$$\|A(y_1 - y_2)\|_2 \geq \frac{\|y_1 - y_2\|_2}{3(2n|Y|)^{1/m}} \geq \frac{r}{3(2n|Y|)^{1/m}} \tag{4.4}$$

holds for every $y_2 \in Y$, $y_2 \neq y_1$, simultaneously with probability $1 - 1/(2n)$.

Finally, observe that as long as $\|Az\|_2 < \|A(y_1 - y_2)\|_2/2$ for every competing signal $y_2 \in Y$, we are guaranteed that

$$\begin{aligned} \|A(y_1 + z) - Ay_1\|_2 &= \|Az\|_2 \\ &< \|A(y_1 - y_2)\|_2 - \|Az\|_2 \\ &\leq \|A(y_1 + z) - Ay_2\|_2 \end{aligned}$$

for every $y_2 \neq y_1$, so we can recover y_1 by simply returning the signal whose image is closest to our measurement point $A(y_1 + z)$ in ℓ_2 distance. To achieve this, we can chain Equations (4.3) and (4.4) together (with a factor of 2), to see that

$$s \leq O\left(\frac{rm^{1/2}n^{1/2-1/m}}{|Y|^{1/m}\log^{3/2} n}\right)$$

suffices. Our total probability of failure is at most $2/n^2 + 1/(2n) < 1/n$.

The main consequence of this theorem is that for the setup we used in Section 3 to prove a deterministic lower bound of $\Omega(k \log(n/k))$, if we simply draw the noise uniformly randomly from the same ℓ_1 ball (in fact, even one with a much larger radius, namely, polynomial in n), this "hard distribution" can be defeated with just $O(k)$ measurements:

Corollary 4.1. *If Y is a set of binary k-sparse vectors, as in Section 3, and noise z is drawn uniformly at random from $B_1(s)$, then for any constant $\epsilon > 0$, $m = O(k/\epsilon)$ measurements suffice to recover any signal in Y with probability $1 - 1/n$, as long as*

$$s \leq O\left(\frac{k^{3/2+\epsilon}n^{1/2-\epsilon}}{\log^{3/2} n}\right).$$

Proof. The parameters in this case are $r = k$ and $|Y| \leq \binom{n}{k} \leq (ne/k)^k$, so by Theorem 4.1, it suffices to have

$$s \leq O\left(\frac{k^{3/2+k/m}n^{1/2-(k+1)/m}}{\log^{3/2} n}\right).$$

Choosing $m = (k+1)/\epsilon$ yields the corollary.

5 Randomized Lower Bound

Although it is possible to partially circumvent this obstacle by focusing our noise distribution on "high" ℓ_2 norm, sparse vectors, we are able to obtain stronger results via a reduction from a communication game and the corresponding lower bound.

The communication game will show that a message Ax must have a large number of bits. To show that this implies a lower bound on the number of rows of A, we will need A to be discrete. Hence we first show that discretizing A does not change its recovery characteristics by much.

5.1 Discretizing Matrices Before we discretize by rounding, we need to ensure that the matrix is well conditioned. We show that without loss of generality, the rows of A are orthonormal.

We can multiply A on the left by any invertible matrix to get another measurement matrix with the same recovery characteristics. If we consider the singular value decomposition $A = U\Sigma V^*$, where U and V are orthonormal and Σ is 0 off the diagonal, this means that we can eliminate U and make the entries of Σ be either 0 or 1. The result is a matrix consisting of m orthonormal rows. For such matrices, we prove the following:

Lemma 5.1. *Consider any $m \times n$ matrix A with orthonormal rows. Let A' be the result of rounding A to b bits per entry. Then for any $v \in \mathbb{R}^n$ there exists an $s \in \mathbb{R}^n$ with $A'v = A(v - s)$ and $\|s\|_1 < n^2 2^{-b}\|v\|_1$.*

Proof. Let $A'' = A - A'$ be the roundoff error when discretizing A to b bits, so each entry of A'' is less than 2^{-b}. Then for any v and $s = A^T A'' v$, we have $As = A''v$ and

$$\begin{aligned} \|s\|_1 = \|A^T A'' v\|_1 &\leq \sqrt{n}\,\|A''v\|_1 \\ &\leq m\sqrt{n}2^{-b}\|v\|_1 \leq n^2 2^{-b}\|v\|_1. \end{aligned}$$

5.2 Communication Complexity We use a few definitions and results from two-party communication complexity. For further background see the book by Kushilevitz and Nisan [KN97]. Consider the following communication game. There are two parties, Alice and Bob. Alice is given a string $y \in$

$\{0,1\}^d$. Bob is given an index $i \in [d]$, together with $y_{i+1}, y_{i+2}, \ldots, y_d$. The parties also share an arbitrarily long common random string r. Alice sends a single message $M(y, r)$ to Bob, who must output y_i with probability at least 3/4, where the probability is taken over r. We refer to this problem as Augmented Indexing. The communication cost of Augmented Indexing is the minimum, over all correct protocols, of the length of the message $M(y, r)$ on the worst-case choice of r and y.

The next theorem is well-known and follows from Lemma 13 of [MNSW98] (see also Lemma 2 of [BYJKK04]).

THEOREM 5.1. *The communication cost of* Augmented Indexing *is* $\Omega(d)$.

Proof. First, consider the private-coin version of the problem, in which both parties can toss coins, but do not share a random string r (i.e., there is no public coin). Consider any correct protocol for this problem. We can assume the probability of error of the protocol is an arbitrarily small positive constant by increasing the length of Alice's message by a constant factor (e.g., by independent repetition and a majority vote). Applying Lemma 13 of [MNSW98] (with, in their notation, $t = 1$ and $a = c' \cdot d$ for a sufficiently small constant $c' > 0$), the communication cost of such a protocol must be $\Omega(d)$. Indeed, otherwise there would be a protocol in which Bob could output y_i with probability greater than 1/2 without any interaction with Alice, contradicting that $\Pr[y_i = 1/2]$ and that Bob has no information about y_i. Our theorem now follows from Newman's theorem (see, e.g., Theorem 2.4 of [KNR99]), which shows that the communication cost of the best public coin protocol is at least that of the private coin protocol minus $O(\log d)$ (which also holds for one-round protocols).

5.3 Randomized Lower Bound Theorem

THEOREM 5.2. *For any randomized ℓ_1/ℓ_1 recovery algorithm $(A, \mathscr{A})$, with approximation factor $C = O(1)$, A must have $m = \Omega(k \log(n/k))$ rows.*

Proof. We shall assume, without loss of generality, that n and k are powers of 2, that k divides n, and that the rows of A are orthonormal. The proof for the general case follows with minor modifications.

Let $(A, \mathscr{A})$ be such a recovery algorithm. We will show how to solve the Augmented Indexing problem on instances of size $d = \Omega(k \log(n/k) \log n)$ with communication cost $O(m \log n)$. The theorem will then follow by Theorem 5.1.

Let X be the maximal set of k-sparse n-dimensional binary vectors with minimum Hamming distance k. From Lemma 3.1 we have $\log |X| = \Omega(k \log(n/k))$. Let $d = \lfloor \log |X| \rfloor \log n$, and define $D = 2C + 3$.

Alice is given a string $y \in \{0,1\}^d$, and Bob is given $i \in [d]$ together with $y_{i+1}, y_{i+2}, \ldots, y_d$, as in the setup for Augmented Indexing.

Alice splits her string y into $\log n$ contiguous chunks $y^1, y^2, \ldots, y^{\log n}$, each containing $\lfloor \log |X| \rfloor$ bits. She uses y^j as an index into X to choose x_j. Alice defines

$$x = D^1 x_1 + D^2 x_2 + \cdots + D^{\log n} x_{\log n}.$$

Alice and Bob use the common randomness r to agree upon a random matrix A with orthonormal rows. Both Alice and Bob round A to form A' with $b = \lceil 2(1 + \log D) \log n \rceil = O(\log n)$ bits per entry. Alice computes $A'x$ and transmits it to Bob.

From Bob's input i, he can compute the value $j = j(i)$ for which the bit y_i occurs in y^j. Bob's input also contains $y_{i+1}, \ldots, y_n$, from which he can reconstruct $x_{j+1}, \ldots, x_{\log n}$, and in particular can compute

$$z = D^{j+1} x_{j+1} + D^{j+2} x_{j+2} + \cdots + D^{\log n} x_{\log n}.$$

Bob then computes $A'z$, and using $A'x$ and linearity, $A'(x - z)$. Then

$$\|x - z\|_1 \leq \sum_{i=1}^{j} k D^i < k \frac{D^{1+\log n}}{D - 1} < k D^{2 \log n}.$$

So from Lemma 5.1, there exists some s with $A'(x - z) = A(x - z - s)$ and

$$\|s\|_1 < n^2 2^{-2 \log n - 2 \log D \log n} \|x - z\|_1 < k.$$

Set $w = x - z - s$. Bob then runs the estimation algorithm $\mathscr{A}$ on A and Aw, obtaining $\hat{w}$ with the property that with probability at least 3/4,

$$\|w - \hat{w}\|_1 \leq C \min_{k\text{-sparse } w'} \|w - w'\|_1 .$$

Now,

$$\begin{aligned} \min_{k\text{-sparse } w'} \|w - w'\|_1 &\leq \left\| w - D^j x_j \right\|_1 \\ &\leq \|s\|_1 + \sum_{i=1}^{j-1} \left\| D^i x_i \right\|_1 \\ &< k(1 + D + D^2 + \cdots + D^{j-1}) \\ &< k \cdot \frac{D^j}{D - 1}. \end{aligned}$$

Hence

$$\begin{aligned}\left\|D^j x_j - \hat{w}\right\|_1 &\le \left\|D^j x_j - w\right\|_1 + \|w - \hat{w}\|_1 \\ &\le (1+C)\left\|D^j x_j - w\right\|_1 \\ &< \frac{kD^j}{2}.\end{aligned}$$

And since the minimum Hamming distance in X is k, this means $\left\|D^j x_j - \hat{w}\right\|_1 < \left\|D^j x' - \hat{w}\right\|_1$ for all $x' \in X, x' \neq x_j$[4]. So Bob can correctly identify x_j with probability at least 3/4. From x_j he can recover y^j, and hence the bit y_i that occurs in y^j.

Hence, Bob solves Augmented Indexing with probability at least 3/4 given the message $A'x$. The entries in A' and x are polynomially bounded integers (up to scaling of A'), and so each entry of $A'x$ takes $O(\log n)$ bits to describe. Hence, the communication cost of this protocol is $O(m \log n)$. By Theorem 5.1, $m \log n = \Omega(k \log(n/k) \log n)$, or $m = \Omega(k \log(n/k))$.

References

[BGI$^+$08] R. Berinde, A. Gilbert, P. Indyk, H. Karloff, and M. Strauss. Combining geometry and combinatorics: a unified approach to sparse signal recovery. *Allerton*, 2008.

[BYJKK04] Z. Bar-Yossef, T. S. Jayram, R. Krauthgamer, and R. Kumar. The sketching complexity of pattern matching. *RANDOM*, 2004.

[CCFC04] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3–15, 2004.

[CM05] G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *J. Algorithms*, 55(1):58–75, 2005.

[CRT06] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1208–1223, 2006.

[CW09] K. L. Clarkson and D. P. Woodruff. Numerical linear algebra in the streaming model. In *STOC*, pages 205–214, 2009.

[DDT$^+$08] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 2008.

[DM08] W. Dai and O. Milenkovic. Weighted superimposed codes and constrained integer compressed sensing. *Preprint*, 2008.

[Don06] D. L. Donoho. Compressed Sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, Apr. 2006.

[EG88] T. Ericson and L. Györfi. Superimposed codes in $\mathbb{R}^n$. *IEEE Trans. on Information Theory*, 34(4):877–880, 1988.

[FR99] Z. Füredi and M. Ruszinkó. An improved upper bound of the rate of euclidean superimposed codes. *IEEE Trans. on Information Theory*, 45(2):799–802, 1999.

[IN07] P. Indyk and A. Naor. Nearest neighbor preserving embeddings. *ACM Trans. on Algorithms*, 3(3), Aug. 2007.

[Ind07] P. Indyk. Sketching, streaming and sublinear-space algorithms. *Graduate course notes, available at* `http://stellar.mit.edu/S/course/6/fa07/6.895/`, 2007.

[KN97] E. Kushilevitz and N. Nisan. *Communication Complexity.* Cambridge University Press, 1997.

[KNR99] I. Kremer, N. Nisan, and D. Ron. On randomized one-round communication complexity. *Computational Complexity*, 8(1):21–49, 1999.

[KNW10] D. Kane, J. Nelson, and D. Woodruff. On the exact space complexity of sketching and streaming small norms. In *SODA*, 2010.

[KT07] B. S. Kashin and V. N. Temlyakov. A remark on compressed sensing. *Preprint*, 2007.

[MNSW98] P. B. Miltersen, N. Nisan, S. Safra, and A. Wigderson. On data structures and asymmetric communication complexity. *J. Comput. Syst. Sci.*, 57(1):37–49, 1998.

[Mut03] S. Muthukrishnan. Data streams: Algorithms and applications (invited talk at soda'03). *Available at http://athos.rutgers.edu/∼muthu/stream-1-1.ps*, 2003.

[vL98] J.H. van Lint. *Introduction to coding theory.* Springer, 1998.

[Wai07] M. Wainwright. Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting. *IEEE Int'l Symp. on Information Theory*, 2007.

A Proof of Lemma 3.1

Proof. We will construct a codebook T of block length k, alphabet q, and minimum Hamming distance ϵk. Replacing each character i with the q-long standard basis vector e_i will create a binary qk-dimensional codebook S with minimum Hamming distance $2\epsilon k$ of the same size as T, where each element of S has exactly k ones.

The Gilbert-Varshamov bound, based on volumes of Hamming balls, states that a codebook of size L exists for some

$$L \ge \frac{q^k}{\sum_{i=0}^{\epsilon k - 1} \binom{k}{i} (q-1)^i}.$$

Using the claim (analogous to [vL98], p. 21, proven

[4] Note that these bounds would still hold with minor modification if we replaced the ℓ_1/ℓ_1 guarantee with the ℓ_2/ℓ_1 guarantee, so the same result holds in that case.

below) that for $\epsilon < 1 - 1/q$

$$\sum_{i=0}^{\epsilon k} \binom{k}{i} (q-1)^i < q^{H_q(\epsilon)k},$$

we have that $\log L > (1 - H_q(\epsilon))k \log q$, as desired.

Claim A.1. *For $0 < \epsilon < 1 - 1/q$,*

$$\sum_{i=0}^{\epsilon k} \binom{k}{i} (q-1)^i < q^{H_q(\epsilon)k}.$$

Proof. Note that

$$q^{-H_q(\epsilon)} = \left(\frac{\epsilon}{(q-1)(1-\epsilon)} \right)^{\epsilon} (1-\epsilon) < (1-\epsilon).$$

Then

$$\begin{aligned} 1 &= (\epsilon + (1-\epsilon))^k \\ &> \sum_{i=0}^{\epsilon k} \binom{k}{i} \epsilon^i (1-\epsilon)^{k-i} \\ &= \sum_{i=0}^{\epsilon k} \binom{k}{i} (q-1)^i \left(\frac{\epsilon}{(q-1)(1-\epsilon)} \right)^i (1-\epsilon)^k \\ &> \sum_{i=0}^{\epsilon k} \binom{k}{i} (q-1)^i \left(\frac{\epsilon}{(q-1)(1-\epsilon)} \right)^{\epsilon k} (1-\epsilon)^k \\ &= q^{-H_q(\epsilon)k} \sum_{i=0}^{\epsilon k} \binom{k}{i} (q-1)^i \end{aligned}$$

B Proof of Lemma 4.1

Proof. By standard arguments (see, e.g., [IN07]), for any $D > 0$ we have

$$\Pr\left[\|A(y_1 - y_2)\|_2 \le \frac{\|y_1 - y_2\|_2}{D} \right] \le \left(\frac{3}{D} \right)^m$$

and

$$\Pr[\|Az\|_2 \ge D\|z\|_2] \le e^{-m(D-1)^2/8}.$$

Setting both right-hand sides to δ yields the lemma.

C Proof of Lemma 4.2

Proof. Consider the distribution of a single coordinate of z, say, z_1. The probability density of $|z_1|$ taking value $t \in [0, s]$ is proportional to the $(n-1)$-dimensional volume of $B_1^{(n-1)}(s-t)$, which in turn is proportional to $(s-t)^{n-1}$. Normalizing to ensure the probability integrates to 1, we derive this probability as

$$p(|z_1| = t) = \frac{n}{s^n}(s-t)^{n-1}.$$

It follows that, for any $D \in [0, s]$,

$$\Pr[|z_1| > D] = \int_D^s \frac{n}{s^n}(s-t)^{n-1}\, dt = (1 - D/s)^n.$$

In particular, for any $\alpha > 1$,

$$\begin{aligned} \Pr[|z_1| &> \alpha s \log n/n] = (1 - \alpha \log n/n)^n < e^{-\alpha \log n} \\ &= 1/n^{\alpha}. \end{aligned}$$

Now, by symmetry this holds for every other coordinate z_i of z as well, so by the union bound

$$\Pr[\|z\|_\infty > \alpha s \log n/n] < 1/n^{\alpha - 1},$$

and since $\|z\|_2 \le \sqrt{n} \cdot \|z\|_\infty$ for any vector z, the lemma follows.