$(1+\epsilon)$-approximate Sparse Recovery

Eric Price
MIT CSAIL
ecprice@mit.edu

David P. Woodruff
IBM Almaden
dpwoodru@us.ibm.com

Abstract— The problem central to sparse recovery and compressive sensing is that of *stable sparse recovery*: we want a distribution $\mathcal{A}$ of matrices $A \in \mathbb{R}^{m\times n}$ such that, for any $x \in \mathbb{R}^n$ and with probability $1-\delta > 2/3$ over $A \in \mathcal{A}$, there is an algorithm to recover $\hat{x}$ from Ax with

$$\left\|\hat{x} - x\right\|_p \leq C \min_{k\text{-sparse } x'} \left\|x - x'\right\|_p \tag{1}$$

for some constant $C > 1$ and norm p.

The measurement complexity of this problem is well understood for constant $C > 1$. However, in a variety of applications it is important to obtain $C = 1+\epsilon$ for a small $\epsilon > 0$, and this complexity is not well understood. We resolve the dependence on ϵ in the number of measurements required of a k-sparse recovery algorithm, up to polylogarithmic factors for the central cases of $p = 1$ and $p = 2$. Namely, we give new algorithms and lower bounds that show the number of measurements required is $k/\epsilon^{p/2}\text{polylog}(n)$. For $p = 2$, our bound of $\frac{1}{\epsilon}k\log(n/k)$ is tight up to *constant* factors. We also give matching bounds when the output is required to be k-sparse, in which case we achieve $k/\epsilon^p\text{polylog}(n)$. This shows the distinction between the complexity of sparse and non-sparse outputs is fundamental.

1. Introduction

Over the last several years, substantial interest has been generated in the problem of solving underdetermined linear systems subject to a sparsity constraint. The field, known as *compressed sensing* or *sparse recovery*, has applications to a wide variety of fields that includes data stream algorithms [16], medical or geological imaging [5], [11], and genetics testing [17], [4]. The approach uses the power of a *sparsity* constraint: a vector x' is *k-sparse* if at most k coefficients are non-zero. A standard formulation for the problem is that of *stable sparse recovery*: we want a distribution $\mathcal{A}$ of matrices $A \in \mathbb{R}^{m\times n}$ such that, for any $x \in \mathbb{R}^n$ and with probability $1-\delta > 2/3$ over $A \in \mathcal{A}$, there is an algorithm to recover $\hat{x}$ from Ax with

$$\left\|\hat{x} - x\right\|_p \leq C \min_{k\text{-sparse } x'} \left\|x - x'\right\|_p \tag{2}$$

for some constant $C > 1$ and norm p[1]. We call this a C-*approximate ℓ_p/ℓ_p recovery scheme* with *failure probability* δ. We refer to the elements of Ax as *measurements*.

It is known [5], [13] that such recovery schemes exist for $p \in \{1, 2\}$ with $C = O(1)$ and $m = O(k\log\frac{n}{k})$. Furthermore, it is known [10], [12] that any such recovery scheme requires $\Omega(k\log_{1+C}\frac{n}{k})$ measurements. This means the measurement complexity is well understood for $C = 1+\Omega(1)$, but not for $C = 1+o(1)$.

A number of applications would like to have $C = 1+\epsilon$ for small ϵ. For example, a radio wave signal can be modeled as $x = x^* + w$ where x^* is k-sparse (corresponding to a signal over a narrow band) and the noise w is i.i.d. Gaussian with $\left\|w\right\|_p \approx D\left\|x^*\right\|_p$ [18]. Then sparse recovery with $C = 1+\alpha/D$ allows the recovery of a $(1-\alpha)$ fraction of the true signal x^*. Since x^* is concentrated in a small band while w is located over a large region, it is often the case that $\alpha/D \ll 1$.

The difficulty of $(1+\epsilon)$-approximate recovery has seemed to depend on whether the output x' is required to be k-sparse or can have more than k elements in its support. Having k-sparse output is important for some applications (e.g. the aforementioned radio waves) but not for others (e.g. imaging). Algorithms that output a k-sparse x' have used $\Theta(\frac{1}{\epsilon^p}k\log n)$ measurements [6], [7], [8], [19]. In contrast, [13] uses only $\Theta(\frac{1}{\epsilon}k\log(n/k))$ measurements for $p = 2$ and outputs a non-k-sparse x'.

Our results: We show that the apparent distinction between complexity of sparse and non-sparse outputs is fundamental, for both $p = 1$ and $p = 2$. We show that for sparse output, $\Omega(k/\epsilon^p)$ measurements are necessary, matching the upper bounds up to a $\log n$ factor. For general output and $p = 2$, we show $\Omega(\frac{1}{\epsilon}k\log(n/k))$ measurements are necessary, matching the upper bound up to a constant factor. In the remaining case of general output and $p = 1$, we show $\widetilde{\Omega}(k/\sqrt{\epsilon})$ measurements are necessary. We then give a novel algorithm that uses $O(\frac{\log^3(1/\epsilon)}{\sqrt{\epsilon}}k\log n)$ measurements, beating the $1/\epsilon$ dependence given by all previous algorithms. As a result, all our bounds are tight up to factors logarithmic in n. The full results are shown in Figure 1.

In addition, for $p = 2$ and general output, we show that thresholding the top $2k$ elements of a Count-Sketch [6] estimate gives $(1+\epsilon)$-approximate recovery with $\Theta(\frac{1}{\epsilon}k\log n)$ measurements. This is interesting because it highlights the distinction between sparse output and non-sparse output: [8] showed that thresholding the top k elements of a Count-Sketch estimate requires $m = \Theta(\frac{1}{\epsilon^2}k\log n)$. While [13] achieves $m = \Theta(\frac{1}{\epsilon}k\log(n/k))$ for the same regime, it only

[1]Some formulations allow the two norms to be different, in which case C is not constant. We only consider equal norms in this paper.

		Lower bound	Upper bound
k-sparse output	ℓ_1	$\Omega(\frac{1}{\epsilon}(k \log \frac{1}{\epsilon} + \log \frac{1}{\delta}))$	$O(\frac{1}{\epsilon} k \log n)$[7]
	ℓ_2	$\Omega(\frac{1}{\epsilon^2}(k + \log \frac{1}{\delta}))$	$O(\frac{1}{\epsilon^2} k \log n)$[6], [8], [19]
Non-k-sparse output	ℓ_1	$\Omega(\frac{1}{\sqrt{\epsilon} \log^2(k/\epsilon)} k)$	$O(\frac{\log^3(1/\epsilon)}{\sqrt{\epsilon}} k \log n)$
	ℓ_2	$\Omega(\frac{1}{\epsilon} k \log(n/k))$	$O(\frac{1}{\epsilon} k \log(n/k))$[13]

Figure 1. Our results, along with existing upper bounds. Fairly minor restrictions on the relative magnitude of parameters apply; see the theorem statements for details.

succeeds with constant probability while ours succeeds with probability $1 - n^{-\Omega(1)}$; hence ours is the most efficient known algorithm when $\delta = o(1), \epsilon = o(1)$, and $k < n^{0.9}$.

Related work: Much of the work on sparse recovery has relied on the Restricted Isometry Property [5]. None of this work has been able to get better than 2-approximate recovery, so there are relatively few papers achieving $(1+\epsilon)$-approximate recovery. The existing ones with $O(k \log n)$ measurements are surveyed above (except for [14], which has worse dependence on ϵ than [7] for the same regime).

No general lower bounds were known in this setting but a couple of works have studied the ℓ_∞/ℓ_p problem, where every coordinate must be estimated with small error. This problem is harder than ℓ_p/ℓ_p sparse recovery with sparse output. For $p = 2$, [19] showed that schemes using Gaussian matrices A require $m = \Omega(\frac{1}{\epsilon^2} k \log(n/k))$. For $p = 1$, [9] showed that any sketch requires $\Omega(k/\epsilon)$ bits (rather than measurements).

Our techniques: For the upper bounds for non-sparse output, we observe that the hard case for sparse output is when the noise is fairly concentrated, in which the estimation of the top k elements can have $\sqrt{\epsilon}$ error. Our goal is to recover enough mass from outside the top k elements to cancel this error. The upper bound for $p = 2$ is a fairly straightforward analysis of the top $2k$ elements of a Count-Sketch data structure.

The upper bound for $p = 1$ proceeds by subsampling the vector at rate 2^{-i} and performing a Count-Sketch with size proportional to $\frac{1}{\sqrt{\epsilon}}$, for $i \in \{0, 1, \dots, O(\log(1/\epsilon))\}$. The intuition is that if the noise is well spread over many (more than $k/\epsilon^{3/2}$) coordinates, then the ℓ_2 bound from the first Count-Sketch gives a very good ℓ_1 bound, so the approximation is $(1+\epsilon)$-approximate. However, if the noise is concentrated over a small number k/ϵ^c of coordinates, then the error from the first Count-Sketch is proportional to $1 + \epsilon^{c/2+1/4}$. But in this case, one of the subsamples will only have $O(k/\epsilon^{c/2-1/4}) < k/\sqrt{\epsilon}$ of the coordinates with large noise. We can then recover those coordinates with the Count-Sketch for that subsample. Those coordinates contain an $\epsilon^{c/2+1/4}$ fraction of the total noise, so recovering them decreases the approximation error by exactly the error induced from the first Count-Sketch.

The lower bounds use substantially different techniques for sparse output and for non-sparse output. For sparse output, we use reductions from communication complexity to show a lower bound in terms of bits. Then, as in [10], we embed $\Theta(\log n)$ copies of this communication problem into a single vector. This multiplies the bit complexity by $\log n$; we also show we can round Ax to $\log n$ bits per measurement without affecting recovery, giving a lower bound in terms of measurements.

We illustrate the lower bound on bit complexity for sparse output using $k = 1$. Consider a vector x containing $1/\epsilon^p$ ones and zeros elsewhere, such that $x_{2i} + x_{2i+1} = 1$ for all i. For any i, set $z_{2i} = z_{2i+1} = 1$ and $z_j = 0$ elsewhere. Then successful $(1+\epsilon/3)$-approximate sparse recovery from $A(x+z)$ returns $\hat{z}$ with $\text{supp}(\hat{z}) = \text{supp}(x) \cap \{2i, 2i+1\}$. Hence we can recover each bit of x with probability $1 - \delta$, requiring $\Omega(1/\epsilon^p)$ bits[2]. We can generalize this to k-sparse output for $\Omega(k/\epsilon^p)$ bits, and to δ failure probability with $\Omega(\frac{1}{\epsilon^p} \log \frac{1}{\delta})$. However, the two generalizations do not seem to combine.

For non-sparse output, we split between ℓ_2 and ℓ_1. In ℓ_2, we consider $A(x+w)$ where x is sparse and w has uniform Gaussian noise with $\|w\|_2^2 \approx \|x\|_2^2/\epsilon$. Then each coordinate of $y = A(x+w) = Ax + Aw$ is a Gaussian channel with signal to noise ratio ϵ. This channel has channel capacity ϵ, showing $I(y; x) \le \epsilon m$. Correct sparse recovery must either get most of x or an ϵ fraction of w; the latter requires $m = \Omega(\epsilon n)$ and the former requires $I(y; x) = \Omega(k \log(n/k))$. This gives a tight $\Theta(\frac{1}{\epsilon} k \log(n/k))$ result. Unfortunately, this does not easily extend to ℓ_1, because it relies on the Gaussian distribution being both stable and maximum entropy under ℓ_2; the corresponding distributions in ℓ_1 are not the same.

Therefore for ℓ_1 non-sparse output, we have yet another argument. The hard instances for $k = 1$ must have one large value (or else 0 is a valid output) but small other values (or else the 2-sparse approximation is significantly better than the 1-sparse approximation). Suppose x has one value of size ϵ and d values of size $1/d$ spread through a vector of size d^2. Then a $(1+\epsilon/2)$-approximate recovery scheme must either locate the large element or guess the locations

[2]For $p = 1$, we can actually set $|\text{supp}(z)| = 1/\epsilon$ and search among a set of $1/\epsilon$ candidates. This gives $\Omega(\frac{1}{\epsilon} \log(1/\epsilon))$ bits.

of the d values with $\Omega(\epsilon d)$ more correct than incorrect. The former requires $1/(d\epsilon^2)$ bits by the difficulty of a novel version of the Gap-ℓ_∞ problem. The latter requires ϵd bits because it allows recovering an error correcting code. Setting $d = \epsilon^{-3/2}$ balances the terms at $\epsilon^{-1/2}$ bits. Because some of these reductions are very intricate, this extended abstract does not manage to embed $\log n$ copies of the problem into a single vector. As a result, we lose a $\log n$ factor in a universe of size $n = \text{poly}(k/\epsilon)$ when converting to measurement complexity from bit complexity.

2. Preliminaries

Notation: We use $[n]$ to denote the set $\{1 \dots n\}$. For any set $S \subset [n]$, we use $\overline{S}$ to denote the complement of S, i.e., the set $[n] \setminus S$. For any $x \in \mathbb{R}^n$, x_i denotes the ith coordinate of x, and x_S denotes the vector $x' \in \mathbb{R}^n$ given by $x'_i = x_i$ if $i \in S$, and $x'_i = 0$ otherwise. We use $\text{supp}(x)$ to denote the support of x.

3. Upper bounds

The algorithms in this section are indifferent to permutation of the coordinates. Therefore, for simplicity of notation in the analysis, we assume the coefficients of x are sorted such that $|x_1| \geq |x_2| \geq \ldots \geq |x_n| \geq 0$.

Count-Sketch: Both our upper bounds use the Count-Sketch [6] data structure. The structure consists of $c \log n$ hash tables of size $O(q)$, for $O(cq \log n)$ total space; it can be represented as Ax for a matrix A with $O(cq \log n)$ rows. Given Ax, one can construct x^* with

$$\|x^* - x\|_\infty^2 \leq \frac{1}{q} \left\| x_{\overline{[q]}} \right\|_2^2 \tag{3}$$

with failure probability n^{1-c}.

3.1. Non-sparse ℓ_2

It was shown in [8] that, if x^* is the result of a Count-Sketch with hash table size $O(k/\epsilon^2)$, then outputting the top k elements of x^* gives a $(1+\epsilon)$-approximate ℓ_2/ℓ_2 recovery scheme. Here we show that a seemingly minor change—selecting $2k$ elements rather than k elements—turns this into a $(1+\epsilon^2)$-approximate ℓ_2/ℓ_2 recovery scheme.

Theorem 3.1. *Let $\hat{x}$ be the top $2k$ estimates from a Count-Sketch structure with hash table size $O(k/\epsilon)$. Then with failure probability $n^{-\Omega(1)}$,*

$$\|\hat{x} - x\|_2 \leq (1+\epsilon) \left\| x_{\overline{[k]}} \right\|_2 .$$

Therefore, there is a $1+\epsilon$-approximate ℓ_2/ℓ_2 recovery scheme with $O(\frac{1}{\epsilon} k \log n)$ rows.

Proof: Let the hash table size be $O(ck/\epsilon)$ for constant c, and let x^* be the vector of estimates for each coordinate. Define S to be the indices of the largest $2k$ values in x^*, and $E = \left\| x_{\overline{[k]}} \right\|_2$.

By (3), the standard analysis of Count-Sketch:

$$\|x^* - x\|_\infty^2 \leq \frac{\epsilon}{ck} E^2.$$

so

$$\begin{aligned}
&\|x_S^* - x\|_2^2 - E^2 \\
&= \|x_S^* - x\|_2^2 - \left\| x_{\overline{[k]}} \right\|_2^2 \\
&\leq \|(x^* - x)_S\|_2^2 + \left\| x_{[n]\setminus S} \right\|_2^2 - \left\| x_{\overline{[k]}} \right\|_2^2 \\
&\leq |S| \, \|x^* - x\|_\infty^2 + \left\| x_{[k]\setminus S} \right\|_2^2 - \left\| x_{S\setminus[k]} \right\|_2^2 \\
&\leq \frac{2\epsilon}{c} E^2 + \left\| x_{[k]\setminus S} \right\|_2^2 - \left\| x_{S\setminus[k]} \right\|_2^2
\end{aligned} \tag{4}$$

Let $a = \max_{i \in [k]\setminus S} x_i$ and $b = \min_{i \in S\setminus[k]} x_i$, and let $d = |[k] \setminus S|$. The algorithm passes over an element of value a to choose one of value b, so

$$a \leq b + 2\|x^* - x\|_\infty \leq b + 2\sqrt{\frac{\epsilon}{ck}} E.$$

Then

$$\begin{aligned}
&\left\| x_{[k]\setminus S} \right\|_2^2 - \left\| x_{S\setminus[k]} \right\|_2^2 \\
&\leq da^2 - (k+d)b^2 \\
&\leq d(b + 2\sqrt{\frac{\epsilon}{ck}} E)^2 - (k+d)b^2 \\
&\leq -kb^2 + 4\sqrt{\frac{\epsilon}{ck}} dbE + \frac{4\epsilon}{ck} dE^2 \\
&\leq -k(b - 2\sqrt{\frac{\epsilon}{ck^3}} dE)^2 + \frac{4\epsilon}{ck^2} dE^2 (k-d) \\
&\leq \frac{4d(k-d)\epsilon}{ck^2} E^2 \leq \frac{\epsilon}{c} E^2
\end{aligned}$$

and combining this with (4) gives

$$\|x_S^* - x\|_2^2 - E^2 \leq \frac{3\epsilon}{c} E^2$$

or

$$\|x_S^* - x\|_2 \leq (1 + \frac{3\epsilon}{2c}) E$$

which proves the theorem for $c \geq 3/2$. ■

3.2. Non-sparse ℓ_1

Theorem 3.2. *There exists a $(1+\epsilon)$-approximate ℓ_1/ℓ_1 recovery scheme with $O(\frac{\log^3 1/\epsilon}{\sqrt{\epsilon}} k \log n)$ measurements and failure probability $e^{-\Omega(k/\sqrt{\epsilon})} + n^{-\Omega(1)}$.*

Set $f = \sqrt{\epsilon}$, so our goal is to get $(1+f^2)$-approximate ℓ_1/ℓ_1 recovery with $O(\frac{\log^3 1/f}{f} k \log n)$ measurements.

For intuition, consider 1-sparse recovery of the following vector x: let $c \in [0, 2]$ and set $x_1 = 1/f^9$ and $x_2, \ldots, x_{1+1/f^{1+c}} \in \{\pm 1\}$. Then we have

$$\left\| x_{\overline{[1]}} \right\|_1 = 1/f^{1+c}$$

and by (3), a Count-Sketch with $O(1/f)$-sized hash tables returns x^* with

$$\|x^* - x\|_\infty \leq \sqrt{f}\left\|x_{\overline{[1/f]}}\right\|_2 \approx 1/f^{c/2} = f^{1+c/2}\left\|x_{\overline{[1]}}\right\|_1 .$$

The reconstruction algorithm therefore cannot reliably find any of the x_i for $i > 1$, and its error on x_1 is at least $f^{1+c/2}\left\|x_{\overline{[1]}}\right\|_1$. Hence the algorithm will not do better than a $f^{1+c/2}$-approximation.

However, consider what happens if we subsample an f^c fraction of the vector. The result probably has about $1/f$ non-zero values, so a $O(1/f)$-width Count-Sketch can reconstruct it exactly. Putting this in our output improves the overall ℓ_1 error by about $1/f = f^c\left\|x_{\overline{[1]}}\right\|_1$. Since $c < 2$, this more than cancels the $f^{1+c/2}\left\|x_{\overline{[1]}}\right\|_1$ error the initial Count-Sketch makes on x_1, giving an approximation factor better than 1.

This tells us that subsampling can help. We don't need to subsample at a scale below k/f (where we can reconstruct well already) or above k/f^3 (where the ℓ_2 bound is small enough already), but in the intermediate range we need to subsample. Our algorithm subsamples at all $\log 1/f^2$ rates in between these two endpoints, and combines the heavy hitters from each.

First we analyze how subsampled Count-Sketch works.

Lemma 3.3. *Suppose we subsample with probability* p *and then apply Count-Sketch with* $\Theta(\log n)$ *rows and* $\Theta(q)$*-sized hash tables. Let* y *be the subsample of* x*. Then with failure probability* $e^{-\Omega(q)} + n^{-\Omega(1)}$ *we recover a* y^* *with*

$$\|y^* - y\|_\infty \leq \sqrt{p/q}\left\|x_{\overline{[q/p]}}\right\|_2 .$$

Proof: Recall the following form of the Chernoff bound: if $X_1, \ldots, X_m$ are independent with $0 \leq X_i \leq M$, and $\mu \geq \mathrm{E}[\sum X_i]$, then

$$\Pr[\sum X_i \geq \frac{4}{3}\mu] \leq e^{-\Omega(\mu/M)}.$$

Let T be the set of coordinates in the sample. Then $\mathrm{E}[\left|T \cap [\frac{3q}{2p}]\right|] = 3q/2$, so

$$\Pr\left[\left|T \cap [\frac{3q}{2p}]\right| \geq 2q\right] \leq e^{-\Omega(q)}.$$

Suppose this event does not happen, so $\left|T \cap [\frac{3q}{2p}]\right| < 2q$. We also have

$$\left\|x_{\overline{[q/p]}}\right\|_2 \geq \sqrt{\frac{q}{2p}}\left|x_{\frac{3q}{2p}}\right| .$$

Let $Y_i = 0$ if $i \notin T$ and $Y_i = x_i^2$ if $i \in T$. Then

$$\mathrm{E}[\sum_{i > \frac{3q}{2p}} Y_i] = p\left\|x_{\overline{[\frac{3q}{2p}]}}\right\|_2^2 \leq p\left\|x_{\overline{[q/p]}}\right\|_2^2$$

For $i > \frac{3q}{2p}$ we have

$$Y_i \leq \left|x_{\frac{3q}{2p}}\right|^2 \leq \frac{2p}{q}\left\|x_{\overline{[q/p]}}\right\|_2^2$$

giving by Chernoff that

$$\Pr[\sum Y_i \geq \frac{4}{3}p\left\|x_{\overline{[q/p]}}\right\|_2^2] \leq e^{-\Omega(q/2)}$$

But if this event does not happen, then

$$\left\|y_{\overline{[2q]}}\right\|_2^2 \leq \sum_{i \in T, i > \frac{3q}{2p}} x_i^2 = \sum_{i > \frac{3q}{2p}} Y_i \leq \frac{4}{3}p\left\|x_{\overline{[q/p]}}\right\|_2^2$$

By (3), using $O(2q)$-size hash tables gives a y^* with

$$\|y^* - y\|_\infty \leq \frac{1}{\sqrt{2q}}\left\|y_{\overline{[2q]}}\right\|_2 \leq \sqrt{p/q}\left\|x_{\overline{[q/p]}}\right\|_2$$

with failure probability $n^{-\Omega(1)}$, as desired. ■

Let $r = 2\log 1/f$. Our algorithm is as follows: for $j \in \{0, \ldots, r\}$, we find and estimate the $2^{j/2}k$ largest elements not found in previous j in a subsampled Count-Sketch with probability $p = 2^{-j}$ and hash size $q = ck/f$ for some parameter $c = \Theta(r^2)$. We output $\hat{x}$, the union of all these estimates. Our goal is to show

$$\|\hat{x} - x\|_1 - \left\|x_{\overline{[k]}}\right\|_1 \leq O(f^2)\left\|x_{\overline{[k]}}\right\|_1 .$$

For each level j, let S_j be the $2^{j/2}k$ largest coordinates in our estimate not found in $S_1 \cup \cdots \cup S_{j-1}$. Let $S = \cup S_j$. By Lemma 3.3, for each j we have (with failure probability $e^{-\Omega(k/f)} + n^{-\Omega(1)}$) that

$$\begin{aligned}\left\|(\hat{x} - x)_{S_j}\right\|_1 &\leq |S_j|\sqrt{\frac{2^{-j}f}{ck}}\left\|x_{\overline{[2^j ck/f]}}\right\|_2 \\ &\leq 2^{-j/2}\sqrt{\frac{fk}{c}}\left\|x_{\overline{[2k/f]}}\right\|_2\end{aligned}$$

and so

$$\begin{aligned}\|(\hat{x} - x)_S\|_1 &= \sum_{j=0}^{r}\left\|(\hat{x} - x)_{S_j}\right\|_1 \\ &\leq \frac{1}{(1 - 1/\sqrt{2})\sqrt{c}}\sqrt{fk}\left\|x_{\overline{[2k/f]}}\right\|_2 \end{aligned} \quad (5)$$

By standard arguments, the ℓ_∞ bound for S_0 gives

$$\left\|x_{[k]}\right\|_1 \leq \|x_{S_0}\|_1 + k\|\hat{x}_{S_0} - x_{S_0}\|_\infty \leq \sqrt{fk/c}\left\|x_{\overline{[2k/f]}}\right\|_2 \quad (6)$$

Combining Equations (5) and (6) gives

$$\begin{aligned}
&\|\hat{x}-x\|_1 - \left\|x_{\overline{[k]}}\right\|_1 \qquad (7)\\
&= \|(\hat{x}-x)_S\|_1 + \|x_{\overline{S}}\|_1 - \left\|x_{\overline{[k]}}\right\|_1\\
&= \|(\hat{x}-x)_S\|_1 + \left\|x_{[k]}\right\|_1 - \|x_S\|_1\\
&= \|(\hat{x}-x)_S\|_1 + (\left\|x_{[k]}\right\|_1 - \|x_{S_0}\|_1) - \sum_{j=1}^{r}\|x_{S_j}\|_1\\
&\le \left(\frac{1}{(1-1/\sqrt{2})\sqrt{c}} + \frac{1}{\sqrt{c}}\right)\sqrt{fk}\left\|x_{\overline{[2k/f]}}\right\|_2\\
&\quad - \sum_{j=1}^{r}\|x_{S_j}\|_1\\
&= O(\frac{1}{\sqrt{c}})\sqrt{fk}\left\|x_{\overline{[2k/f]}}\right\|_2 - \sum_{j=1}^{r}\|x_{S_j}\|_1 \qquad (8)
\end{aligned}$$

We would like to convert the first term to depend on the ℓ_1 norm. For any u and s we have, by splitting into chunks of size s, that

$$\begin{aligned}
\left\|u_{\overline{[2s]}}\right\|_2 &\le \sqrt{\frac{1}{s}}\left\|u_{\overline{[s]}}\right\|_1\\
\left\|u_{\overline{[s]}\cap[2s]}\right\|_2 &\le \sqrt{s}\,|u_s|\,.
\end{aligned}$$

Along with the triangle inequality, this gives us that

$$\begin{aligned}
\sqrt{kf}\left\|x_{\overline{[2k/f]}}\right\|_2 &\le \sqrt{kf}\left\|x_{\overline{[2k/f^3]}}\right\|_2\\
&\quad + \sqrt{kf}\sum_{j=1}^{r}\left\|x_{\overline{[2^jk/f]}\cap[2^{j+1}k/f]}\right\|_2\\
&\le f^2\left\|x_{\overline{[k/f^3]}}\right\|_1 + \sum_{j=1}^{r}k2^{j/2}\left|x_{2^jk/f}\right|
\end{aligned}$$

so

$$\begin{aligned}
&\|\hat{x}-x\|_1 - \left\|x_{\overline{[k]}}\right\|_1\\
&\le O(\frac{1}{\sqrt{c}})f^2\left\|x_{\overline{[k/f^3]}}\right\|_1 + \sum_{j=1}^{r}O(\frac{1}{\sqrt{c}})k2^{j/2}\left|x_{2^jk/f}\right|\\
&\quad - \sum_{j=1}^{r}\|x_{S_j}\|_1 \qquad (9)
\end{aligned}$$

Define $a_j = k2^{j/2}\left|x_{2^jk/f}\right|$. The first term grows as f^2 so it is fine, but a_j can grow as $f2^{j/2} > f^2$. We need to show that they are canceled by the corresponding $\left\|x_{S_j}\right\|_1$. In particular, we will show that $\left\|x_{S_j}\right\|_1 \ge \Omega(a_j) - O(2^{-j/2}f^2\left\|x_{\overline{[k/f^3]}}\right\|_1)$ with high probability—at least wherever $a_j \ge \|a\|_1/(2r)$.

Let $U \in [r]$ be the set of j with $a_j \ge \|a\|_1/(2r)$, so that $\|a_U\|_1 \ge \|a\|_1/2$. We have

$$\begin{aligned}
\left\|x_{\overline{[2^jk/f]}}\right\|_2^2 &= \left\|x_{\overline{[2k/f^3]}}\right\|_2^2 + \sum_{i=j}^{r}\left\|x_{\overline{[2^jk/f]}\cap[2^{j+1}k/f]}\right\|_2^2\\
&\le \left\|x_{\overline{[2k/f^3]}}\right\|_2^2 + \frac{1}{kf}\sum_{i=j}^{r}a_j^2 \qquad (10)
\end{aligned}$$

For $j \in U$, we have

$$\sum_{i=j}^{r}a_i^2 \le a_j\|a\|_1 \le 2ra_j^2$$

so, along with $(y^2+z^2)^{1/2} \le y+z$, we turn Equation (10) into

$$\begin{aligned}
\left\|x_{\overline{[2^jk/f]}}\right\|_2 &\le \left\|x_{\overline{[2k/f^3]}}\right\|_2 + \sqrt{\frac{1}{kf}\sum_{i=j}^{r}a_j^2}\\
&\le \sqrt{\frac{f^3}{k}}\left\|x_{\overline{[k/f^3]}}\right\|_1 + \sqrt{\frac{2r}{kf}}a_j
\end{aligned}$$

When choosing S_j, let $T \in [n]$ be the set of indices chosen in the sample. Applying Lemma 3.3 the estimate x^* of x_T has

$$\begin{aligned}
\|x^*-x_T\|_\infty &\le \sqrt{\frac{f}{2^jck}}\left\|x_{\overline{[2^jk/f]}}\right\|_2\\
&\le \sqrt{\frac{1}{2^jc}}\frac{f^2}{k}\left\|x_{\overline{[k/f^3]}}\right\|_1 + \sqrt{\frac{2r}{2^jc}}\frac{a_j}{k}\\
&= \sqrt{\frac{1}{2^jc}}\frac{f^2}{k}\left\|x_{\overline{[k/f^3]}}\right\|_1 + \sqrt{\frac{2r}{c}}\left|x_{2^jk/f}\right|
\end{aligned}$$

for $j \in U$.

Let $Q = [2^jk/f] \setminus (S_0 \cup \cdots \cup S_{j-1})$. We have $|Q| \ge 2^{j-1}k/f$ so $\mathrm{E}[|Q \cap T|] \ge k/2f$ and $|Q \cap T| \ge k/4f$ with failure probability $e^{-\Omega(k/f)}$. Conditioned on $|Q \cap T| \ge k/4f$, since x_T has at least $|Q \cap T| \ge k/(4f) = 2^{r/2}k/4 \ge 2^{j/2}k/4$ possible choices of value at least $\left|x_{2^jk/f}\right|$, x_{S_j} must have at least $k2^{j/2}/4$ elements at least $\left|x_{2^jk/f}\right| - \|x^*-x_T\|_\infty$. Therefore, for $j \in U$,

$$\|x_{S_j}\|_1 \ge -\frac{1}{4\sqrt{c}}f^2\left\|x_{\overline{[k/f^3]}}\right\|_1 + \frac{k2^{j/2}}{4}(1-\sqrt{\frac{2r}{c}})\left|x_{2^jk/f}\right|$$

and therefore

$$\begin{aligned}
&\sum_{j=1}^{r}\|x_{S_j}\|_1 \ge \sum_{j\in U}\|x_{S_j}\|_1\\
&\ge \sum_{j\in U} -\frac{1}{4\sqrt{c}}f^2\left\|x_{\overline{[k/f^3]}}\right\|_1 + \frac{k2^{j/2}}{4}(1-\sqrt{\frac{2r}{c}})\left|x_{2^jk/f}\right|\\
&\ge -\frac{r}{4\sqrt{c}}f^2\left\|x_{\overline{[k/f^3]}}\right\|_1 + \frac{1}{4}(1-\sqrt{\frac{2r}{c}})\|a_U\|_1\\
&\ge -\frac{r}{4\sqrt{c}}f^2\left\|x_{\overline{[k/f^3]}}\right\|_1 + \frac{1}{8}(1-\sqrt{\frac{2r}{c}})\sum_{j=1}^{r}k2^{j/2}\left|x_{2^jk/f}\right| \qquad (11)
\end{aligned}$$

Using (9) and (11) we get

$$\begin{aligned}&\|\hat{x}-x\|_1 - \left\|x_{\overline{[k]}}\right\|_1\\ \le&\left(\frac{r}{4\sqrt{c}}+O(\frac{1}{\sqrt{c}})\right)f^2\left\|x_{\overline{[k/f^3]}}\right\|_1\\ &+\sum_{j=1}^{r}\left(O(\frac{1}{\sqrt{c}})+\frac{1}{8}\sqrt{\frac{2r}{c}}-\frac{1}{8}\right)k2^{j/2}\left|x_{2^jk/f}\right|\\ \le&f^2\left\|x_{\overline{[k/f^3]}}\right\|_1 \le f^2\left\|x_{\overline{[k]}}\right\|_1\end{aligned}$$

for some $c = O(r^2)$. Hence we use a total of $\frac{rc}{f}k\log n = \frac{\log^3 1/f}{f}k\log n$ measurements for $1+f^2$-approximate ℓ_1/ℓ_1 recovery.

For each $j \in \{0,\ldots,r\}$ we had failure probability $e^{-\Omega(k/f)} + n^{-\Omega(1)}$ (from Lemma 3.3 and $|Q \cap T| \ge k/2f$). By the union bound, our overall failure probability is at most

$$(\log\frac{1}{f})(e^{-\Omega(k/f)}+n^{-\Omega(1)}) \le e^{-\Omega(k/f)}+n^{-\Omega(1)},$$

proving Theorem 3.2.

4. Lower bounds for non-sparse output and $p = 2$

In this case, the lower bound follows fairly straightforwardly from the Shannon-Hartley information capacity of a Gaussian channel.

We will set up a communication game. Let $\mathcal{F} \subset \{S \subset [n] \mid |S| = k\}$ be a family of k-sparse supports such that:

- $|S\Delta S'| \ge k$ for $S \ne S' \in \mathcal{F}$,
- $\Pr_{S\in\mathcal{F}}[i \in S] = k/n$ for all $i \in [n]$, and
- $\log|\mathcal{F}| = \Omega(k\log(n/k))$.

This is possible; for example, a Reed-Solomon code on $[n/k]^k$ has these properties.

Let $X = \{x \in \{0,\pm1\}^n \mid \operatorname{supp}(x) \in \mathcal{F}\}$. Let $w \sim N(0,\alpha\frac{k}{n}I_n)$ be i.i.d. normal with variance $\alpha k/n$ in each coordinate. Consider the following process:

Procedure: First, Alice chooses $S \in \mathcal{F}$ uniformly at random, then $x \in X$ uniformly at random subject to $\operatorname{supp}(x) = S$, then $w \sim N(0,\alpha\frac{k}{n}I_n)$. She sets $y = A(x+w)$ and sends y to Bob. Bob performs sparse recovery on y to recover $x' \approx x$, rounds to X by $\hat{x} = \arg\min_{\hat{x}\in X}\|\hat{x}-x'\|_2$, and sets $S' = \operatorname{supp}(\hat{x})$. This gives a Markov chain $S \to x \to y \to x' \to S'$.

If sparse recovery works for any $x+w$ with probability $1-\delta$ as a distribution over A, then there is some specific A and random seed such that sparse recovery works with probability $1-\delta$ over $x+w$; let us choose this A and the random seed, so that Alice and Bob run deterministic algorithms on their inputs.

Lemma 4.1. $I(S;S') = O(m\log(1+\frac{1}{\alpha}))$.

Proof: Let the columns of A^T be $v^1,\ldots,v^m$. We may assume that the v^i are orthonormal, because this can be accomplished via a unitary transformation on Ax. Then we have that $y_i = \langle v^i, x+w\rangle = \langle v^i, x\rangle + w'_i$, where $w'_i \sim N(0,\alpha k\left\|v^i\right\|_2^2/n) = N(0,\alpha k/n)$ and

$$\mathrm{E}_x[\langle v^i,x\rangle^2] = \mathrm{E}_S[\sum_{j\in S}(v^i_j)^2] = \frac{k}{n}$$

Hence $y_i = z_i + w'_i$ is a Gaussian channel with power constraint $\mathrm{E}[z_i^2] \le \frac{k}{n}\left\|v^i\right\|_2^2$ and noise variance $\mathrm{E}[(w'_i)^2] = \alpha\frac{k}{n}\left\|v^i\right\|_2^2$. Hence by the Shannon-Hartley theorem this channel has information capacity

$$\max_{v_i} I(z_i;y_i) = C \le \frac{1}{2}\log(1+\frac{1}{\alpha}).$$

By the data processing inequality for Markov chains and the chain rule for entropy, this means

$$\begin{aligned}I(S;S') \le I(z;y) &= H(y) - H(y\mid z) = H(y) - H(y-z\mid z)\\ &= H(y) - \sum H(w'_i \mid z, w'_1,\ldots,w'_{i-1})\\ &= H(y) - \sum H(w'_i) \le \sum H(y_i) - H(w'_i)\\ &= \sum H(y_i) - H(y_i\mid z_i) = \sum I(y_i;z_i)\\ &\le \frac{m}{2}\log(1+\frac{1}{\alpha}). \qquad (12)\end{aligned}$$

■

We will show that successful recovery either recovers most of x, in which case $I(S;S') = \Omega(k\log(n/k))$, or recovers an ϵ fraction of w. First we show that recovering w requires $m = \Omega(\epsilon n)$.

Lemma 4.2. *Suppose $w \in \mathbb{R}^n$ with $w_i \sim N(0,\sigma^2)$ for all i and $n = \Omega(\frac{1}{\epsilon^2}\log(1/\delta))$, and $A \in \mathbb{R}^{m\times n}$ for $m < \delta\epsilon n$. Then any algorithm that finds w' from Aw must have $\|w'-w\|_2^2 > (1-\epsilon)\|w\|_2^2$ with probability at least $1-O(\delta)$.*

Proof: Note that Aw merely gives the projection of w onto m dimensions, giving no information about the other $n-m$ dimensions. Since w and the ℓ_2 norm are rotation invariant, we may assume WLOG that A gives the projection of w onto the first m dimensions, namely $T = [m]$. By the norm concentration of Gaussians, with probability $1-\delta$ we have $\|w\|_2^2 < (1+\epsilon)n\sigma^2$, and by Markov with probability $1-\delta$ we have $\|w_T\|_2^2 < \epsilon n\sigma^2$.

For any fixed value d, since w is uniform Gaussian and $w'_{\overline{T}}$ is independent of $w_{\overline{T}}$,

$$\begin{aligned}\Pr[\|w'-w\|_2^2 < d] &\le \Pr[\|(w'-w)_{\overline{T}}\|_2^2 < d]\\ &\le \Pr[\|w_{\overline{T}}\|_2^2 < d].\end{aligned}$$

Therefore

$$\begin{aligned}&\Pr[\|w'-w\|_2^2 < (1-3\epsilon)\|w\|_2^2]\\ \le&\Pr[\|w'-w\|_2^2 < (1-2\epsilon)n\sigma^2]\\ \le&\Pr[\|w_{\overline{T}}\|_2^2 < (1-2\epsilon)n\sigma^2]\\ \le&\Pr[\|w_{\overline{T}}\|_2^2 < (1-\epsilon)(n-m)\sigma^2] \le \delta\end{aligned}$$

as desired. Rescaling ϵ gives the result. ∎

Lemma 4.3. *Suppose $n = \Omega(1/\epsilon^2 + (k/\epsilon)\log(k/\epsilon))$ and $m = O(\epsilon n)$. Then $I(S; S') = \Omega(k \log(n/k))$ for some $\alpha = \Omega(1/\epsilon)$.*

Proof: Consider the x' recovered from $A(x + w)$, and let $T = S \cup S'$. Suppose that $\|w\|_\infty^2 \le O(\frac{\alpha k}{n}\log n)$ and $\|w\|_2^2/(\alpha k) \in [1 \pm \epsilon]$, as happens with probability at least (say) $3/4$. Then we claim that if recovery is successful, one of the following must be true:

$$\|x'_T - x\|_2^2 \le 9\epsilon \|w\|_2^2 \tag{13}$$

$$\left\|x'_{\overline{T}} - w\right\|_2^2 \le (1 - 2\epsilon)\|w\|_2^2 \tag{14}$$

To show this, suppose $\|x'_T - x\|_2^2 > 9\epsilon \|w\|_2^2 \ge 9\|w_T\|_2^2$ (the last by $|T| = 2k = O(\epsilon n/\log n)$). Then

$$\begin{aligned}\|(x' - (x + w))_T\|_2^2 &> (\|x' - x\|_2 - \|w_T\|_2)^2 \\ &\ge (2\|x' - x\|_2/3)^2 \ge 4\epsilon \|w\|_2^2 .\end{aligned}$$

Because recovery is successful,

$$\|x' - (x + w)\|_2^2 \le (1 + \epsilon)\|w\|_2^2 .$$

Therefore

$$\begin{aligned}\left\|x'_{\overline{T}} - w_{\overline{T}}\right\|_2^2 + \|x'_T - (x + w)_T\|_2^2 &= \|x' - (x + w)\|_2^2 \\ \left\|x'_{\overline{T}} - w_{\overline{T}}\right\|_2^2 + 4\epsilon\|w\|_2^2 &< (1 + \epsilon)\|w\|_2^2 \\ \left\|x'_{\overline{T}} - w\right\|_2^2 - \|w_T\|_2^2 &< (1 - 3\epsilon)\|w\|_2^2 \\ &\le (1 - 2\epsilon)\|w\|_2^2\end{aligned}$$

as desired. Thus with $3/4$ probability, at least one of (13) and (14) is true.

Suppose Equation (14) holds with at least $1/4$ probability. There must be some x and S such that the same equation holds with $1/4$ probability. For this S, given x' we can find T and thus $x'_{\overline{T}}$. Hence for a uniform Gaussian $w_{\overline{T}}$, given $Aw_{\overline{T}}$ we can compute $A(x + w_{\overline{T}})$ and recover $x'_{\overline{T}}$ with $\left\|x'_{\overline{T}} - w_{\overline{T}}\right\|_2^2 \le (1 - \epsilon)\|w_{\overline{T}}\|_2^2$. By Lemma 4.2 this is impossible, since $n - |T| = \Omega(\frac{1}{\epsilon^2})$ and $m = \Omega(\epsilon n)$ by assumption.

Therefore Equation (13) holds with at least $1/2$ probability, namely $\|x'_T - x\|_2^2 \le 9\epsilon\|w\|_2^2 \le 9\epsilon(1 - \epsilon)\alpha k < k/2$ for appropriate α. But if the nearest $\hat{x} \in X$ to x is not equal to x,

$$\begin{aligned}&\|x' - \hat{x}\|_2^2 \\ &= \left\|x'_{\overline{T}}\right\|_2^2 + \left\|x'_T - \hat{x}\right\|_2^2 \ge \left\|x'_{\overline{T}}\right\|_2^2 + (\|x - \hat{x}\|_2 - \left\|x'_T - x\right\|_2)^2 \\ &> \left\|x'_{\overline{T}}\right\|_2^2 + (k - k/2)^2 > \left\|x'_{\overline{T}}\right\|_2^2 + \left\|x'_T - x\right\|_2^2 = \|x' - x\|_2^2 ,\end{aligned}$$

a contradiction. Hence $S' = S$. But Fano's inequality states $H(S|S') \le 1 + \Pr[S' \ne S]\log|\mathcal{F}|$ and hence

$$I(S; S') = H(S) - H(S|S') \ge -1 + \frac{1}{4}\log|\mathcal{F}| = \Omega(k\log(n/k))$$

as desired. ∎

Theorem 4.4. *Any $(1 + \epsilon)$-approximate ℓ_2/ℓ_2 recovery scheme with $\epsilon > \sqrt{\frac{k\log n}{n}}$ and failure probability $\delta < 1/2$ requires $m = \Omega(\frac{1}{\epsilon}k\log(n/k))$.*

Proof: Combine Lemmas 4.3 and 4.1 with $\alpha = 1/\epsilon$ to get $m = \Omega(\frac{k\log(n/k)}{\log(1+\epsilon)}) = \Omega(\frac{1}{\epsilon}k\log(n/k))$, $m = \Omega(\epsilon n)$, or $n = O(\frac{1}{\epsilon}k\log(k/\epsilon))$. For ϵ as in the theorem statement, the first bound is controlling. ∎

5. Bit complexity to measurement complexity

The remaining lower bounds proceed by reductions from communication complexity. The following lemma (implicit in [10]) shows that lower bounding the number of bits for approximate recovery is sufficient to lower bound the number of measurements. Let $B_p^n(R) \subset \mathbb{R}^n$ denote the ℓ_p ball of radius R.

Definition 5.1. *Let $X \subset \mathbb{R}^n$ be a distribution with $x_i \in \{-n^d, \dots, n^d\}$ for all $i \in [n]$ and $x \in X$. We define a $1 + \epsilon$-approximate ℓ_p/ℓ_p sparse recovery* bit scheme *on X with b bits, precision n^{-c}, and failure probability δ to be a deterministic pair of functions $f\colon X \to \{0,1\}^b$ and $g\colon \{0,1\}^b \to \mathbb{R}^n$ where f is linear so that $f(a + b)$ can be computed from $f(a)$ and $f(b)$. We require that, for $u \in B_p^n(n^{-c})$ uniformly and x drawn from X, $g(f(x))$ is a valid result of $1 + \epsilon$-approximate recovery on $x + u$ with probability $1 - \delta$.*

Lemma 5.2. *A lower bound of $\Omega(b)$ bits for such a sparse recovery bit scheme with $p \le 2$ implies a lower bound of $\Omega(b/((1 + c + d)\log n))$ bits for regular $(1 + \epsilon)$-approximate sparse recovery with failure probability $\delta - 1/n$.*

Proof: Suppose we have a standard $(1+\epsilon)$-approximate sparse recovery algorithm $\mathcal{A}$ with failure probability δ using m measurements Ax. We will use this to construct a (randomized) sparse recovery bit scheme using $O(m(1 + c + d)\log n)$ bits and failure probability $\delta + 1/n$. Then by averaging some deterministic sparse recovery bit scheme performs better than average over the input distribution.

We may assume that $A \in \mathbb{R}^{m\times n}$ has orthonormal rows (otherwise, if $A = U\Sigma V^T$ is its singular value decomposition, $\Sigma^+ U^T A$ has this property and can be inverted before applying the algorithm). When applied to the distribution $X + u$ for u uniform over $B_p^n(n^{-c})$, we may assume that $\mathcal{A}$ and A are deterministic and fail with probability δ over their input.

Let A' be A rounded to $t\log n$ bits per entry for some parameter t. Let x be chosen from X. By Lemma 5.1 of [10], for any x we have $A'x = A(x - s)$ for some s with $\|s\|_1 \le n^2 2^{-t\log n}\|x\|_1$, so $\|s\|_p \le n^{2.5-t}\|x\|_p \le n^{3.5+d-t}$. Let $u \in B_p^n(n^{5.5+d-t})$ uniformly at random. With probability at least $1 - 1/n$, $u \in B_p^n((1 - 1/n^2)n^{5.5+d-t})$ because the balls are similar so the ratio of volumes is $(1 - 1/n^2)^n > 1-$

$1/n$. In this case $u + s \in B_p^n(n^{5.5+d-t})$; hence the random variable u and $u+s$ overlap in at least a $1-1/n$ fraction of their volumes, so $x+s+u$ and $x+u$ have statistical distance at most $1/n$. Therefore $\mathcal{A}(A(x+u)) = \mathcal{A}(A'x + Au)$ with probability at least $1 - 1/n$.

Now, $A'x$ uses only $(t+d+1)\log n$ bits per entry, so we can set $f(x) = A'x$ for $b = m(t+d+1)\log n$. Then we set $g(y) = \mathcal{A}(y+Au)$ for uniformly random $u \in B_p^n(n^{5.5+d-t})$. Setting $t = 5.5+d+c$, this gives a sparse recovery bit scheme using $b = m(6.5 + 2d + c)\log n$. ■

6. Non-sparse output Lower Bound for $p = 1$

First, we show that recovering the locations of an ϵ fraction of d ones in a vector of size $n > d/\epsilon$ requires $\widetilde{\Omega}(\epsilon d)$ bits. Then, we show high bit complexity of a distributional product version of the Gap-ℓ_∞ problem. Finally, we create a distribution for which successful sparse recovery must solve one of the previous problems, giving a lower bound in bit complexity. Lemma 5.2 converts the bit complexity to measurement complexity.

6.1. ℓ_1 Lower bound for recovering noise bits

Definition 6.1. *We say a set $C \subset [q]^d$ is a (d, q, ϵ) code if any two distinct $c, c' \in C$ agree in at most ϵd positions. We say a set $X \subset \{0,1\}^{dq}$ represents C if X is C concatenated with the trivial code $[q] \to \{0,1\}^q$ given by $i \to e_i$.*

Claim 6.2. *For $\epsilon \geq 2/q$, there exist (d, q, ϵ) codes C of size $q^{\Omega(\epsilon d)}$ by the Gilbert-Varshamov bound (details in [10]).*

Lemma 6.3. *Let $X \subset \{0,1\}^{dq}$ represent a (d, q, ϵ) code. Suppose $y \in \mathbb{R}^{dq}$ satisfies $\|y - x\|_1 \leq (1-\epsilon)\|x\|_1$. Then we can recover x uniquely from y.*

Proof: We assume $y_i \in [0, 1]$ for all i; thresholding otherwise decreases $\|y - x\|_1$. We will show that there exists no other $x' \in X$ with $\|y - x\|_1 \leq (1-\epsilon)\|x\|_1$; thus choosing the nearest element of X is a unique decoder. Suppose otherwise, and let $S = \mathrm{supp}(x), T = \mathrm{supp}(x')$. Then

$$\begin{aligned}(1-\epsilon)\|x\|_1 &\geq \|x - y\|_1 \\ &= \|x\|_1 - \|y_S\|_1 + \|y_{\overline{S}}\|_1 \\ \|y_S\|_1 &\geq \|y_{\overline{S}}\|_1 + \epsilon d\end{aligned}$$

Since the same is true relative to x' and T, we have

$$\begin{aligned}\|y_S\|_1 + \|y_T\|_1 &\geq \|y_{\overline{S}}\|_1 + \|y_{\overline{T}}\|_1 + 2\epsilon d \\ 2\|y_{S\cap T}\|_1 &\geq 2\|y_{\overline{S\cup T}}\|_1 + 2\epsilon d \\ \|y_{S\cap T}\|_1 &\geq \epsilon d \\ |S \cap T| &\geq \epsilon d\end{aligned}$$

This violates the distance of the code represented by X. ■

Lemma 6.4. *Let $R = [s, cs]$ for some constant c and parameter s. Let X be a permutation independent distribution over $\{0,1\}^n$ with $\|x\|_1 \in R$ with probability p. If y satisfies $\|x - y\|_1 \leq (1-\epsilon)\|x\|_1$ with probability p' with $p' - (1-p) = \Omega(1)$, then $I(x; y) = \Omega(\epsilon s \log(n/s))$.*

Proof: For each integer $i \in R$, let $X_i \subset \{0,1\}^n$ represent an $(i, n/i, \epsilon)$ code. Let $p_i = \Pr_{x\in X}[\|x\|_1 = i]$. Let S_n be the set of permutations of $[n]$. Then the distribution X' given by (a) choosing $i \in R$ proportional to p_i, (b) choosing $\sigma \in S_n$ uniformly, (c) choosing $x_i \in X_i$ uniformly, and (d) outputting $x' = \sigma(x_i)$ is equal to the distribution $(x \in X \mid \|x\|_1 \in R)$.

Now, because $p' \geq \Pr[\|x\|_1 \notin R] + \Omega(1)$, x' chosen from X' satisfies $\|x' - y\|_1 \leq (1-\epsilon)\|x'\|_1$ with $\delta \geq p' - (1-p)$ probability. Therefore, with at least $\delta/2$ probability, i and σ are such that $\|\sigma(x_i) - y\|_1 \leq (1-\epsilon)\|\sigma(x_i)\|_1$ with $\delta/2$ probability over uniform $x_i \in X_i$. But given y with $\|y - \sigma(x_i)\|_1$ small, we can compute $y' = \sigma^{-1}(y)$ with $\|y' - x_i\|_1$ equally small. Then by Lemma 6.3 we can recover x_i from y with probability $\delta/2$ over $x_i \in X_i$. Thus for this i and σ, $I(x; y \mid i, \sigma) \geq \Omega(\log|X_i|) = \Omega(\delta\epsilon s \log(n/s))$ by Fano's inequality. But then $I(x; y) = \mathrm{E}_{i,\sigma}[I(x; y \mid i, \sigma)] = \Omega(\delta^2 \epsilon s \log(n/s)) = \Omega(\epsilon s \log(n/s))$. ■

6.2. Distributional Indexed Gap ℓ_∞

Consider the following communication game, which we refer to as $\mathsf{Gap}\ell_\infty^B$, studied in [2]. The legal instances are pairs (x, y) of m-dimensional vectors, with $x_i, y_i \in \{0, 1, 2, \ldots, B\}$ for all i such that

- NO instance: for all i, $y_i - x_i \in \{0, 1\}$, or
- YES instance: there is a *unique* i for which $y_i - x_i = B$, and for all $j \neq i$, $y_i - x_i \in \{0, 1\}$.

The *distributional* communication complexity $D_{\sigma,\delta}(f)$ of a function f is the minimum over all deterministic protocols computing f with error probability at most δ, where the probability is over inputs drawn from σ.

Consider the distribution σ which chooses a random $i \in [m]$. Then for each $j \neq i$, it chooses a random $d \in \{0, \ldots, B\}$ and (x_i, y_i) is uniform in $\{(d, d), (d, d+1)\}$. For coordinate i, (x_i, y_i) is uniform in $\{(0, 0), (0, B)\}$. Using similar arguments to those in [2], Jayram [15] showed $D_{\sigma,\delta}(\mathsf{Gap}\ell_\infty^B) = \Omega(m/B^2)$ (this is reference [70] on p.182 of [1]) for δ less than a small constant.

We define the one-way distributional communication complexity $D_{\sigma,\delta}^{1-way}(f)$ of a function f to be the smallest distributional complexity of a protocol for f in which only a single message is sent from Alice to Bob.

Definition 6.5 (Indexed $\mathsf{Ind}\ell_\infty^{r,B}$ Problem)**.** *There are r pairs of inputs $(x^1, y^1), (x^2, y^2), \ldots, (x^r, y^r)$ such that every pair (x^i, y^i) is a legal instance of the $\mathsf{Gap}\ell_\infty^B$ problem. Alice is given $x^1, \ldots, x^r$. Bob is given an index $I \in [r]$ and $y^1, \ldots, y^r$. The goal is to decide whether (x^I, y^I) is a NO or a YES instance of $\mathsf{Gap}\ell_\infty^B$.*

Let η be the distribution $\sigma^r \times U_r$, where U_r is the uniform distribution on $[r]$. We bound $D_{\eta,\delta}^{1-way}(\mathsf{Ind}\ell_\infty)^{r,B}$ as follows.

For a function f, let f^r denote the problem of computing r instances of f. For a distribution ζ on instances of f, let $D^{1-way,*}_{\zeta^r,\delta}(f^r)$ denote the minimum communication cost of a deterministic protocol computing a function f with error probability at most δ *in each* of the r copies of f, where the inputs come from ζ^r.

Theorem 6.6. *(special case of Corollary 2.5 of [3]) Assume $D_{\sigma,\delta}(f)$ is larger than a large enough constant. Then $D^{1-way,*}_{\sigma^r,\delta/2}(f^r) = \Omega(rD_{\sigma,\delta}(f))$.*

Theorem 6.7. *For δ less than a sufficiently small constant, $D^{1-way}_{\eta,\delta}(\mathsf{Ind}\ell_\infty^{r,B}) = \Omega(\delta^2 rm/(B^2 \log r))$.*

Proof: Consider a deterministic 1-way protocol Π for $\mathsf{Ind}\ell_\infty^{r,B}$ with error probability δ on inputs drawn from η. Then for at least $r/2$ values $i \in [r]$, $\Pr[\Pi(x^1,\ldots,x^r,y^1,\ldots,y^r,I) = \mathsf{Gap}\ell_\infty^B(x^I,y^I) \mid I = i] \geq 1-2\delta$. Fix a set $S = \{i_1,\ldots,i_{r/2}\}$ of indices with this property. We build a deterministic 1-way protocol Π' for $f^{r/2}$ with input distribution $\sigma^{r/2}$ and error probability at most 6δ in each of the $r/2$ copies of f.

For each $\ell \in [r] \setminus S$, independently choose $(x^\ell, y^\ell) \sim \sigma$. For each $j \in [r/2]$, let Z_j^1 be the probability that $\Pi(x^1,\ldots,x^r,y^1,\ldots,y^r,I) = \mathsf{Gap}\ell_\infty^B(x^{i_j},y^{i_j})$ given $I = i_j$ and the choice of (x^ℓ,y^ℓ) for all $\ell \in [r] \setminus S$.

If we repeat this experiment independently $s = O(\delta^{-2}\log r)$ times, obtaining independent $Z_j^1,\ldots,Z_j^s$ and let $Z_j = \sum_t Z_j^t$, then $\Pr[Z_j \geq s - s\cdot 3\delta] \geq 1 - \frac{1}{r}$. So there exists a set of $s = O(\delta^{-1}\log r)$ repetitions for which for each $j \in [r/2]$, $Z_j \geq s - s\cdot 3\delta$. We hardwire these into Π' to make the protocol deterministic.

Given inputs $((X^1,\ldots,X^{r/2}),(Y^1,\ldots,Y^{r/2})) \sim \sigma^{r/2}$ to Π', Alice and Bob run s executions of Π, each with $x^{i_j} = X^j$ and $y^{i_j} = Y^j$ for all $j \in [r/2]$, filling in the remaining values using the hardwired inputs. Bob runs the algorithm specified by Π for each $i_j \in S$ and each execution. His output for (X^j,Y^j) is the majority of the outputs of the s executions with index i_j.

Fix an index i_j. Let W be the number of repetitions for which $\mathsf{Gap}\ell_\infty^B(X^j,Y^j)$ does not equal the output of Π on input i_j, for a random $(X^j,Y^j) \sim \sigma$. Then, $\mathbf{E}[W] \leq 3\delta$. By a Markov bound, $\Pr[W \geq s/2] \leq 6\delta$, and so the coordinate is correct with probability at least $1-6\delta$.

The communication of Π' is a factor $s = \Theta(\delta^{-2}\log r)$ more than that of Π. The theorem now follows by Theorem 6.6, using that $D_{\sigma,12\delta}(\mathsf{Gap}\ell_\infty^B) = \Omega(m/B^2)$. ■

6.3. Lower bound for sparse recovery

Fix the parameters $B = \Theta(1/\epsilon^{1/2}), r = k,\ m = 1/\epsilon^{3/2}$, and $n = k/\epsilon^3$. Given an instance $(x^1,y^1),\ldots,(x^r,y^r),I$ of $\mathsf{Ind}\ell_\infty^{r,B}$, we define the input signal z to a sparse recovery problem. We allocate a set S^i of m disjoint coordinates in a universe of size n for each pair (x^i,y^i), and on these coordinates place the vector $y^i - x^i$. The locations are important for arguing the sparse recovery algorithm cannot learn much information about the noise, and will be placed uniformly at random.

Let ρ denote the induced distribution on z. Fix a $(1+\epsilon)$-approximate k-sparse recovery bit scheme Alg that takes b bits as input and succeeds with probability at least $1-\delta/2$ over $z \sim \rho$ for some small constant δ. Let S be the set of top k coordinates in z. Alg has the guarantee that if it succeeds for $z \sim \rho$, then there exists a small u with $\|u\|_1 < n^{-2}$ so that $v = Alg(z)$ satisfies

$$\begin{aligned}\|v-z-u\|_1 &\leq (1+\epsilon)\left\|(z+u)_{[n]\setminus S}\right\|_1 \\ \|v-z\|_1 &\leq (1+\epsilon)\left\|z_{[n]\setminus S}\right\|_1 + (2+\epsilon)/n^2 \\ &\leq (1+2\epsilon)\left\|z_{[n]\setminus S}\right\|_1\end{aligned}$$

and thus

$$\|(v-z)_S\|_1 + \left\|(v-z)_{[n]\setminus S}\right\|_1 \leq (1+2\epsilon)\|z_{[n]\setminus S}\|_1. \quad (15)$$

Lemma 6.8. *For $B = \Theta(1/\epsilon^{1/2})$ sufficiently large, suppose that $\Pr_{z\sim\rho}[\|(v-z)_S\|_1 \leq 10\epsilon \cdot \|z_{[n]\setminus S}\|_1] \geq 1-\delta$. Then Alg requires $b = \Omega(k/(\epsilon^{1/2}\log k))$.*

Proof: We show how to use Alg to solve instances of $\mathsf{Ind}\ell_\infty^{r,B}$ with probability at least $1-C$ for some small C, where the probability is over input instances to $\mathsf{Ind}\ell_\infty^{r,B}$ distributed according to η, inducing the distribution ρ. The lower bound will follow by Theorem 6.7. Since Alg is a deterministic sparse recovery bit scheme, it receives a sketch $f(z)$ of the input signal z and runs an arbitrary recovery algorithm g on $f(z)$ to determine its output $v = Alg(z)$.

Given $x^1,\ldots,x^r$, for each $i = 1,2,\ldots,r$, Alice places $-x^i$ on the appropriate coordinates in the block S^i used in defining z, obtaining a vector z_{Alice}, and transmits $f(z_{Alice})$ to Bob. Bob uses his inputs $y^1,\ldots,y^r$ to place y^i on the appropriate coordinate in S^i. He thus creates a vector z_{Bob} for which $z_{Alice}+z_{Bob} = z$. Given $f(z_{Alice})$, Bob computes $f(z)$ from $f(z_{Alice})$ and $f(z_{Bob})$, then $v = Alg(z)$. We assume all coordinates of v are rounded to the real interval $[0,B]$, as this can only decrease the error.

We say that S^i is *bad* if either

- there is no coordinate j in S^i for which $|v_j| \geq \frac{B}{2}$ yet (x^i,y^i) is a YES instance of $\mathsf{Gap}\ell_\infty^{r,B}$, or
- there is a coordinate j in S^i for which $|v_j| \geq \frac{B}{2}$ yet either (x^i,y^i) is a NO instance of $\mathsf{Gap}\ell_\infty^{r,B}$ or j is not the unique j^* for which $y^i_{j^*} - x^i_{j^*} = B$

The ℓ_1-error incurred by a bad block is at least $B/2-1$. Hence, if there are t bad blocks, the total error is at least $t(B/2-1)$, which must be smaller than $10\epsilon \cdot \|z_{[n]\setminus S}\|_1$ with probability $1-\delta$. Suppose this happens.

We bound t. All coordinates in $z_{[n]\setminus S}$ have value in the set $\{0,1\}$. Hence, $\|z_{[n]\setminus S}\|_1 < rm$. So $t \leq 20\epsilon rm/(B-2)$. For $B \geq 6$, $t \leq 30\epsilon rm/B$. Plugging in r, m and B, $t \leq Ck$, where $C > 0$ is a constant that can be made arbitrarily small by increasing $B = \Theta(1/\epsilon^{1/2})$.

If a block S^i is not bad, then it can be used to solve $\mathsf{Gap}\ell_\infty^{r,B}$ on (x^i, y^i) with probability 1. Bob declares that (x^i, y^i) is a YES instance if and only if there is a coordinate j in S^i for which $|v_j| \geq B/2$.

Since Bob's index I is uniform on the m coordinates in $\mathsf{Ind}\ell_\infty^{r,B}$, with probability at least $1 - C$ the players solve $\mathsf{Ind}\ell_\infty^{r,B}$ given that the ℓ_1 error is small. Therefore they solve $\mathsf{Ind}\ell_\infty^{r,B}$ with probability $1 - \delta - C$ overall. By Theorem 6.7, for C and δ sufficiently small Alg requires $\Omega(mr/(B^2 \log r)) = \Omega(k/(\epsilon^{1/2} \log k))$ bits. ■

Lemma 6.9. *Suppose* $\Pr_{z\sim\rho}[\|(v-z)_{[n]\setminus S}\|_1] \leq (1-8\epsilon) \cdot \|z_{[n]\setminus S}\|_1] \geq \delta/2$. *Then* Alg *requires* $b = \Omega(\frac{1}{\sqrt{\epsilon}} k \log(1/\epsilon))$.

Proof: The distribution ρ consists of $B(mr, 1/2)$ ones placed uniformly throughout the n coordinates, where $B(mr, 1/2)$ denotes the binomial distribution with mr events of $1/2$ probability each. Therefore with probability at least $1 - \delta/4$, the number of ones lies in $[\delta mr/8, (1-\delta/8)mr]$. Thus by Lemma 6.4, $I(v; z) \geq \Omega(\epsilon mr \log(n/(mr)))$. Since the mutual information only passes through a b-bit string, $b = \Omega(\epsilon mr \log(n/(mr)))$ as well. ■

Theorem 6.10. *Any* $(1+\epsilon)$*-approximate* ℓ_1/ℓ_1 *recovery scheme with sufficiently small constant failure probability* δ *must make* $\Omega(\frac{1}{\sqrt{\epsilon}} k/\log^2(k/\epsilon))$ *measurements.*

Proof: We will lower bound any ℓ_1/ℓ_1 sparse recovery bit scheme Alg. If Alg succeeds, then in order to satisfy inequality (15), we must either have $\|(v-z)_S\|_1 \leq 10\epsilon \cdot \|z_{[n]\setminus S}\|_1$ or we must have $\|(v-z)_{[n]\setminus S}\|_1 \leq (1-8\epsilon) \cdot \|z_{[n]\setminus S}\|_1$. Since Alg succeeds with probability at least $1 - \delta$, it must either satisfy the hypothesis of Lemma 6.8 or the hypothesis of Lemma 6.9. But by these two lemmas, it follows that $b = \Omega(\frac{1}{\sqrt{\epsilon}} k/\log k)$. Therefore by Lemma 5.2, any $(1+\epsilon)$-approximate ℓ_1/ℓ_1 sparse recovery algorithm requires $\Omega(\frac{1}{\sqrt{\epsilon}} k/\log^2(k/\epsilon))$ measurements. ■

7. Lower bounds for k-sparse output

Theorem 7.1. *Any* $1+\epsilon$*-approximate* ℓ_1/ℓ_1 *recovery scheme with* k*-sparse output and failure probability* δ *requires* $m = \Omega(\frac{1}{\epsilon}(k \log \frac{1}{\epsilon} + \log \frac{1}{\delta}))$*, for* $32 \leq \frac{1}{\delta} \leq n\epsilon^2/k$.

Theorem 7.2. *Any* $1+\epsilon$*-approximate* ℓ_2/ℓ_2 *recovery scheme with* k*-sparse output and failure probability* δ *requires* $m = \Omega(\frac{1}{\epsilon^2}(k + \log \frac{\epsilon^2}{\delta}))$*, for* $32 \leq \frac{1}{\delta} \leq n\epsilon^2/k$.

These two theorems correspond to four statements: one for large k and one for small δ for both ℓ_1 and ℓ_2.

All are fairly similar to the framework of [10]: they use a sparse recovery algorithm to robustly identify x from Ax for x in some set X. This gives bit complexity $\log|X|$, or measurement complexity $\log|X| / \log n$ by Lemma 5.2. They amplify the bit complexity to $\log|X| \log n$ by showing they can recover x_1 from $A(x_1 + \frac{1}{10}x_2 + \ldots + \frac{1}{n}x_{\Theta(\log n)})$ for $x_1, \ldots, x_{\Theta(\log n)} \in X$ and reducing from augmented indexing. This gives a $\log|X|$ measurement lower bound. Due to space constraints, we defer full proof to the full paper.

Acknowledgment: We thank T.S. Jayram for helpful discussions.

References

[1] Z. Bar-Yossef, "The complexity of massive data set computations," Ph.D. dissertation, UC Berkeley, 2002.

[2] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar, "An information statistics approach to data stream and communication complexity," *J. Comput. Syst. Sci.*, vol. 68, no. 4, pp. 702–732, 2004.

[3] M. Braverman and A. Rao, "Information equals amortized communication," in *STOC*, 2011.

[4] A. Bruex, A. Gilbert, R. Kainkaryam, J. Schiefelbein, and P. Woolf, "Poolmc: Smart pooling of mRNA samples in microarray experiments," *BMC Bioinformatics*, 2010.

[5] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. 59, no. 8, pp. 1208–1223, 2006.

[6] M. Charikar, K. Chen, and M. Farach-Colton, "Finding frequent items in data streams," *ICALP*, 2002.

[7] G. Cormode and S. Muthukrishnan, "Improved data stream summaries: The count-min sketch and its applications," *LATIN*, 2004.

[8] ——, "Combinatorial algorithms for compressed sensing," *Sirocco*, 2006.

[9] ——, "Summarizing and mining skewed data streams," in *SDM*, 2005.

[10] K. Do Ba, P. Indyk, E. Price, and D. Woodruff, "Lower bounds for sparse recovery," *SODA*, 2010.

[11] D. L. Donoho, "Compressed Sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[12] S. Foucart, A. Pajor, H. Rauhut, and T. Ullrich, "The gelfand widths of lp-balls for $0 < p \leq 1$," 2010.

[13] A. C. Gilbert, Y. Li, E. Porat, and M. J. Strauss, "Approximate sparse recovery: optimizing time and measurements," in *STOC*, 2010, pp. 475–484.

[14] P. Indyk and M. Ruzic, "Near-optimal sparse recovery in the l1 norm," in *FOCS*, 2008, pp. 199–207.

[15] T. Jayram, "Unpublished manuscript," 2002.

[16] S. Muthukrishnan, "Data streams: Algorithms and applications)," *FTTCS*, 2005.

[17] N. Shental, A. Amir, and O. Zuk, "Identification of rare alleles and their carriers using compressed se(que)nsing," *Nucleic Acids Research*, vol. 38(19), pp. 1–22, 2010.

[18] J. Treichler, M. Davenport, and R. Baraniuk, "Application of compressive sensing to the design of wideband signal acquisition receivers," *In Proc. U.S./Australia Joint Work. Defense Apps. of Signal Processing (DASP)*, 2009.

[19] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Transactions on Information Theory*, vol. 55, no. 12, pp. 5728–5741, 2009.