Adaptive Sparse Recovery with Limited Adaptivity

Akshay Kamath
The University of Texas at Austin
kamath@cs.utexas.edu

Eric Price*
The University of Texas at Austin
ecprice@cs.utexas.edu

October 25, 2018

Abstract

The goal of adaptive sparse recovery is to estimate an approximately sparse vector x from a series of linear measurements $A_1x, A_2x, \dots, A_Rx$, where each matrix A_i may depend on the previous observations. With an unlimited number of rounds R, it is known that $O(k \log\log n)$ measurements suffice for $O(1)$-approximate k-sparse recovery in $\mathbb{R}^n$, and that $\Omega(k + \log\log n)$ measurements are necessary. We initiate the study of what happens with a constant number of rounds of adaptivity. Previous techniques could not give nontrivial bounds using less than 5 rounds of adaptivity, and were inefficient for any constant R.

We give nearly matching upper and lower bounds for any constant number of rounds R. Our lower bound shows that $\Omega(k(\log \frac{n}{k})^{1/R})$ measurements are necessary for any $k < 2^{(\log \frac{n}{k})^{1/R}}$; significantly, this is the first lower bound that combines k and n in an adaptive setting.

Our upper bound shows that $O(k(\log \frac{n}{k})^{1/R} \cdot \log^* k)$ measurements suffice. The $O(\log^* k)$ gap between the two bounds comes from a similar gap for nonadaptive sparse recovery in the high-SNR regime, and would be reduced to constant factors with improvements to nonadaptive high-SNR sparse recovery.

1 Introduction

Sparse recovery is the problem of estimating an approximately sparse vector x from a low-dimensional linear sketch Ax. Also known as compressed sensing, sparse recovery is a simple mathematical problem with a diverse collection of applications, including image aquisition [DDT+08], genetic testing [ECG+09], medical imaging [LDSP08], and streaming algorithms [CM06].

We say that an algorithm performs (k, C)-sparse recovery if it recovers a vector x^* such that

$$\|x - x^*\|_2^2 \leq C \min_{k\text{-sparse } x'} \|x - x'\|_2^2 . \tag{1}$$

One could also consider recovery in other norms such as ℓ_1 [CM04, CRT06], but ℓ_2 is the strongest ℓ_p-norm for which efficient sparse recovery is possible [CCF02, BJKS04].

The most common goal in sparse recovery is to achieve (1) for $C = O(1)$ with 90% probability over the choice of matrix $A \in \mathbb{R}^{m \times n}$, with as few "measurements" m as possible. If A is chosen independently of x, it is known that $m = \Theta(k \log n)$ is necessary [DIPW10] and sufficient [CRT06, GLPS10][1]. However, this sample complexity can be improved if A is chosen *adaptively.*

[*] This work was done in part while the author was visiting the Simons Institute for the Theory of Computing.

[1] More precisely, $m = \Theta(k \log \frac{n}{k})$. For simplicity of exposition in the introduction, we assume $k < n^{0.9}$ so these are equivalent.

	Paper	Measurements	Rounds	Comment
Upper	[IPW11]	$\frac{k}{\epsilon} r \log^{1/r} n$	$O(r \log^* k)$	
	[NSWZ18]	$\frac{1}{\epsilon} kr \log^{1/r} \frac{1}{\epsilon} + kr \log^{1/r} n$	$O(r \log^* k)$	
		$\frac{k}{\epsilon} r \log^{1/r} n$	$r + 3$	
	Corollary 3.10	$k \log^{1/r} n \cdot 5^r \log^* k$	r	$\epsilon = O(1)$
Lower	[PW13]	$r \log^{1/r} n$	r	
	[ACD13]	k/ϵ	r	
	Corollary 2.6	$\frac{1}{r} \cdot k \log^{1/r} n$	r	$\log k < \log^{1/r} n$
	Theorem 2.5	$\omega(k)$	r	$k = n^{o(1)}, r = O(1)$

Table 1: Results for adaptive $(k, 1 + \epsilon)$-sparse recovery. The measurements column drops constant factors. The upper bounds above are not explicit in previous papers, which only state the bounds for $r = O(\log \log n)$. However, all previous algorithms reduce to 1-sparse recovery as a black box, and plugging in r-round $O(r \log^{1/r} n)$-sample 1-sparse recovery gives the above.

In adaptive sparse recovery, the algorithm picks $A_1 \in \mathbb{R}^{m_1 \times n}$, observes $A_1 x$, then picks $A_2 \in \mathbb{R}^{m_2 \times n}$ and observes $A_2 x$, and continues until $A_R x$ for some number of rounds R. The goal is still to minimize the total number of measurements $m = \sum_i m_i$. With $O(\log \log n)$ rounds of adaptivity, it is possible to achieve (1) with $m = O(k \log \log n)$ [IPW11, NSWZ18]. On the other hand, we know that $\Omega(k + \log \log n)$ measurements are necessary with unlimited adaptivity [ACD13, PW13].

In this work, we consider sparse recovery with a small constant number of rounds of adaptivity. For example, what is possible with $R = 2$? This is an important question for applications, where adaptivity is typically costly. The number of rounds of adaptivity corresponds to the number of passes of a streaming algorithm, or the number of rounds of mapreduce; thus the overall communication (which is usually the speed bottleneck in such applications) is proportional to R. In other applications such as imaging or genetic testing, parallelism and latency in setting up the measurements can make it difficult to perform many rounds of adaptivity.

For $k = 1$ and $R = O(1)$, it is known that $m = \Theta(\log^{1/R} n)$ is necessary and sufficient [IPW11, PW13]. Thus one expects that the answer for $k \gg 1$ should probably be $k \log^{1/R} n$. However, the best prior algorithm (a variant of [NSWZ18] described in the next section) uses three "extra" rounds, giving only $O(k \log^{1/(R-3)} n)$. This does not benefit from anything less than five rounds of adaptivity. On the lower bound side, existing work shows that $m = \Omega(k + \log^{1/R} n)$ [ACD13, PW13], but cannot connect k and n. For $C = 1 + \epsilon$, one can get an algorithm separating the dependence on n and ϵ [NSWZ18]; perhaps the same could hold for n and k?

We show upper and lower bounds that almost entirely address the problem. First, we show that $\Omega(k \log^{1/R} n)$ samples are necessary, for any k with $k < 2^{\log^{1/R} n}$. This settles the sample complexity for smallish k; for larger k, up to $n^{o(1)}$, we can still show that $\omega(k)$ samples are necessary.

Second, we give an algorithm that uses $O(k \log^{1/R} n \cdot \log^* k)$ samples. The extra $\log^* k$ factor comes from black-box calls to nonadaptive (k, C)-sparse recovery for $C \gg 1$, for which the best current algorithm uses $O(k \log_C \frac{n}{k} \cdot \log^* k)$ samples [PW12]. If that result is improved to match the $\Omega(k \log_C \frac{n}{k})$ lower bound [PW11], the extra $\log^* k$ factor in our algorithm will also be removed.

1.1 Related Work

The adaptive measurement model has been explored in many papers, starting with empirical results [MSW08, JXC08, CHNR08] and theoretical results for $k = 1$ [CHNR08]. Results from the compressed sensing side of the literature have focused on signal approximation accuracy, which corresponds to the behavior for $C = 1+\epsilon$ as $\epsilon \to 0$. With Gaussian noise, nonadaptive algorithms take $m = O(\frac{1}{\epsilon} k \log n)$, while [HCN11, HBCN12] improve this to $O\left(k \log n + \frac{1}{\epsilon} k(\log k + \log\log\log n)\right)$; a corresponding $\Omega(k/\epsilon)$ lower bound appeared in [ACD13]. On the sparse recovery side of the literature, [IPW11] gave a fully adaptive algorithm using $O(\frac{1}{\epsilon} k \log\log n)$ measurements performed in $R = O(\log\log n \log^* k)$ rounds. This was improved by [NSWZ18] in two incomparable ways: either R can be improved to $O(\log\log n)$ or the sample complexity can be improved to $O(\frac{\log\log\frac{1}{\epsilon}}{\epsilon} k + k \log\log n)$, splitting n and ϵ in the sample complexity.

The algorithms in [IPW11] and [NSWZ18] can easily be adapted to use fewer rounds of adaptivity. Each algorithm's round complexity is dominated by black-box applications of the $O(\log\log n)$-round $O(\log\log n)$-sample $O(1)$-approximate 1-sparse recovery algorithm of [IPW11]. By changing this to an r-round $O(\log^{1/r} n)$-sample version, the algorithms can be performed with fewer rounds; see Figure 1. Most relevantly, one of the algorithms in [NSWZ18] would use $O(k \log^{1/r} n)$ samples in $r+3$ rounds. It seems likely that a more careful analysis could reduce this to $r+2$ rounds, but no further: the approach requires an initial round to find the important subproblem instances, and a final round to clean up missing elements.

1.2 Overview of Lower Bound

Prior Work ($k = 1$). We begin by giving an overview of the lower bound for $k = 1$ from [PW13]. The lower bound instance consists of the signal $e_X + w$, where $X \in [n]$ is a uniform random index and $w \sim \mathcal{N}(0, I_n/n)$ is Gaussian. This signal is such that successful 1.1-approximate 1-sparse recovery must return a vector that is close to e_X, and in particular reveals the identity of X. Hence

$$I(X; Y_1, \dots, Y_R) = \Omega(\log n).$$

On the other hand, [PW13] shows that after learning b bits about X, each measurement in the next round reveals only $O(b+1)$ bits. That is, for any set of observations $y_1, \dots, y_{r-1}$ seen so far, if we define

$$b = H(X) - H(X \mid Y_1 = y_1, \dots, Y_{r-1} = y_{r-1}) \tag{2}$$

to be the information revealed so far about X, then it can be shown that the next round has

$$I(X; Y_r \mid Y_1 = y_1, \dots, Y_{r-1} = y_{r-1}) = m_r \cdot O(b+1) \tag{3}$$

where m_r is the number of measurements in round r. It follows that $I(X; Y_1, \dots, Y_R) \leq C^R \prod_{i=1}^{R} m_r$. Then, an application of AM-GM shows $(O(m/R))^R = \Omega(\log n)$, or $m = \Omega(R \log^{1/R} n)$. Thus the key step is to show (3).

The intuition for why (3) should hold is as follows. For any single measurement vector v of unit norm, the corresponding observation is

$$y = \langle v, e_X + w \rangle = v_X + w'$$

where $w' \sim N(0, 1/n)$. This is an additive white gaussian noise channel, so the Shannon-Hartley Theorem bounds the information capacity in terms of the signal-to-noise ratio:

$$I(X; y) \leq \frac{1}{2} \log(1 + n\,\mathbb{E}[v_X^2]).$$

This holds even conditioned on $Y_1 = y_1, \ldots, Y_{r-1} = y_{r-1}$, so we want to bound $\mathbb{E}[v_X^2 \mid Y_1 = y_1, \ldots, Y_{r-1} = y_{r-1}]$. Let $p : [n] \to \mathbb{R}$ denote the probability distribution of $(X \mid Y_1 = y_1, \ldots, Y_{r-1} = y_{r-1})$, so $b = \log n - H(p)$. If p happens to be uniform over its support, then its value is $2^b/n$ at $n/2^b$ locations; then any unit norm v has

$$n \mathop{\mathbb{E}}_{X \sim p}[v_X^2] \leq n \cdot \sum_{i=1}^{n} \frac{2^b}{n} v_i^2 = 2^b$$

or $I(X; y \mid Y_1 = y_1, \ldots, Y_{r-1} = y_{r-1}) \leq \frac{1}{2} \log(1 + 2^b) \lesssim (b + 1)$.

However, p is not necessarily uniform over its support, which necessitates care. For example, consider if $p(1) = 1/\log n$ and p is uniform otherwise. Then $b = O(1)$, yet by setting $v = e_1$ we have

$$n \mathop{\mathbb{E}}_{X \sim p}[v_X^2] = n/\log n$$

so Shannon-Hartley would only show $O(\log n)$ bits per measurement. The problem is that Shannon-Hartley is only a good bound if the signal – in this case v_X – is at a consistent scale. The fix is to partition the indices of X by the scale of $p(X)$; we define $T_j = \{i \mid np(i) \in [2^j, 2^{j+1})\}$ for $j > 0$, and T_0 to have the rest. Let J be the random variable denoting the j such that $X \in T_j$. We can decompose (with implicit conditioning on $y_1, \ldots, y_{r-1}$)

$$I(X; y) \leq I(X; (y, J)) = I(X; y \mid J) + I(X; J). \tag{4}$$

Then $I(X; J) \leq H(J) \lesssim b + 1$ by simple algebra, and since $(X \mid J)$ is roughly uniform over its support the Shannon-Hartley bound can give $I(X; y \mid J) \lesssim b + 1$. This bounds the information content in any single measurement; summing over all m_r measurements in Y_r yields (3).

We now describe how to adapt these techniques to prove a result for $k > 1$.

Problem instance for general k. We use the natural extension of the problem instance, which is to concatenate k copies of the hard instance; that is, for $N = nk$, we draw $X_1, \ldots, X_k \in [n]$, and set the vector to

$$x = \left(\sum_{i=1}^{k} e_{(i-1)n + X_i} \right) + w$$

where $w = N(0, \frac{k}{N} I_N)$. Then successful 1.1-approximate sparse recovery must recover most coordinates X_i, so

$$I(X_1, \ldots, X_k; Y_1, \ldots, Y_R) = \Omega(k \log n).$$

Defining the per-round goal. The first difficulty is how best to define the goal (3). While (3) is true as stated, this is not enough: it would give a lower bound of $(k \log n)^{1/R}$ not $k \log^{1/R} n$. Yet (3) is also tight; given b bits of information about the first coordinate, a single measurement *can* learn $\Omega(b)$ bits about that coordinate.

However, with b bits of information overall, most coordinates will only have $O(b/k)$ of information "about them." Each such coordinate can only be observed with signal-to-noise ratio $2^{O(b/k)}$. Thus we can hope to say that there exists a large set of coordinates, $W \subset [k]$ of size $|W| > 0.99k$, such that

$$I(\{X_i\}_{i \in W}; Y_r \mid Y_1 = y_1, \ldots, Y_{r-1} = y_{r-1}) = m_r \cdot O(\frac{b}{k} + 1).$$

Unfortunately, this is false. Suppose we have learned the parity of $X_i \oplus X_1$ for all i; this is only $b = k - 1$ bits of information. Then the measurement vector v which matches all the parities will have signal-to-noise-ratio k; with a variation on this example[2], the information learned in a single measurement can be $\Omega(\log k)$ bits for every large W even though $b = k$. Thus, the replacement for (3) that we can show is

$$I(\{X_i\}_{i \in W}; Y_r \mid Y_1 = y_1, \ldots, Y_{r-1} = y_{r-1}) = m_r \cdot O(\frac{b}{k} + \log k) + O(b + k). \tag{5}$$

The extra $O(b + k)$ term comes from a term analogous to $I(X; J)$ in (4).

Implications for sample complexity. In the first round we can replace (5) by the straightforward nonadaptive bound

$$I(\{X_i\}_{i \in [k]}; Y_1) \leq O(m_1).$$

Now, for simplicity of exposition suppose each $m_i = m/R = \Theta(m)$. If $m > k \log k$, then after the first round the dominant term in (5) will be $O(b \cdot \frac{m}{k})$. Hence chaining (5) gives a set W_R such that

$$k \log n \lesssim I(\{X_i\}_{i \in W_R}; Y_1, \ldots, Y_R) \leq m \cdot \left(O(\frac{m}{k})\right)^{R-1} = k \left(O(\frac{m}{k})\right)^{R}.$$

Thus $m = \Omega(k \log^{1/R} n)$, as long as this is more than $k \log k$.

Analog of J for general k. The proof of (5) is analogous to that of (3), where we partition the X by "scale", and bound the mutual information conditioned on the scale by Shannon-Hartley. However, the new partition is subtle so we describe it here.

The first coordinate X_1 is partitioned the same way as its marginal would be in the $k = 1$ case: the set T_{j_1} has $\{i \in [n] \mid np(X_1 = i) \in [2^{j_1}, 2^{j_1+1})\}$ for $j_1 > 0$, T_0 has everything else, and J_1 denotes the $j_1 \geq 0$ with $X_1 \in T_{j_1}$. The second coordinate is partitioned as its marginal *conditioned on* J_1. That is, we have sets

$$T_{j_1, j_2} = \{i \in [n] \mid np(X_2 = i \mid X_1 \in T_{j_1}) \in [2^{j_2}, 2^{j_2+1})\}$$

and the random variable J_2 is such that $x_2 \in T_{J_1, J_2}$. This naturally extends to $x_i \in T_{J_1, \ldots, J_i}$.

We show that this partitions $J = (J_1, \ldots, J_k)$ of the coordinates $X_1, \ldots, X_k$ has the following properties. First,$H(J) = O(b)$ so conditioning on J does not reveal too much information. Second, the "signal power" $Z_{i,J}$ that any measurement has about X_i conditioned on J obeys

$$\mathop{\mathbb{E}}_{i \in [k]} \mathop{\mathbb{E}}_{J} \log(1 + Z_{i,J}) \lesssim \frac{b}{k}. \tag{6}$$

Since the Shannon-Hartley theorem implies

$$I(X_1, \ldots, X_k; Y_r \mid J, Y_1 = y_1, \ldots, Y_{r-1} = y_{r-1}) \lesssim m_r \cdot \mathop{\mathbb{E}}_{J} \log(1 + \sum_{i=1}^{k} Z_{i,J})$$

one would get—if (6) held for all i not just on average—that

$$I(X_1, \ldots, X_k; Y_r \mid J, Y_1 = y_1, \ldots, Y_{r-1} = y_{r-1}) \lesssim m_r \cdot \log(1 + k 2^{b/k}) \approx m_r(\frac{b}{k} + \log k)$$

as desired. Using Markov's inequality to choose for each J a large set W of i where (6) is not too far off, we can get (5) and complete the proof.

[2] Partition $[k]$ into $\log k$ pieces, and the prior information reveals the relative parities within each partition.

1.3 Overview of Upper Bound

Prior work for $k = 1$. The high-level intuition for our algorithm is based on the intuition for $k = 1$ from the upper bound in [IPW11] and corresponding lower bound in [PW13]. Suppose the vector x has one large coordinate i^*, of value 1. For $O(1)$-approximate sparse recovery to be nontrivial, the amount of "noise" in other coordinates, i.e. $\left\|x_{[n]\setminus\{i\}}\right\|_2^2$, will be at most a small constant.

At any given round, if we have learned b bits of information in the previous round, we can expect to have located i^* to within a set S of size $n/2^b$. Then our measurement matrix in this round can place zero mass on any coordinate outside S. Effectively, in this round we are trying to find i^* within x_S. This vector still has "signal" 1, but the "noise" $\left\|x_{S\setminus\{i\}}\right\|_2^2$ is likely to be much smaller: if S is random, the noise will be $O(1/2^b)$ on average. With such a high signal-to-noise ratio (SNR), we can hope to learn $\Theta(\log \mathrm{SNR}) = \Theta(b)$ bits per measurement. This will quickly reduce the size of our candidate set S, further enriching the SNR of X_S and increasing the information per measurement.

Given r rounds with t measurements each, we expect to learn t bits in the first round; $\Theta(t^2)$ bits in the second round; $\Theta(t^3)$ bits in the third round; and so on till $\Theta(t^R)$ bits in the Rth round. Setting $t = \log^{1/R} n$, we can learn the desired $\log n$ bits of information in $O(R \log^{1/R} n)$ measurements.

Algorithm for general k. Previous adaptive algorithms with $m = o(k \log n)$ use the $k = 1$ algorithm as a black box [IPW11, NSWZ18]. Unfortunately, such efforts seem to require additional rounds of adaptivity to set up the subproblem instances and/or to clean up coordinates missed in the first pass. Our algorithm avoids this by opening up the $k = 1$ algorithm and extending its techniques to general k.

Our goal is to maintain a candidate set $S \subseteq [n]$ of locations that include the largest k elements of x, known as the "heavy hitters". In each round except for the last, we would like to take a number of measurements that are insufficient to identify the heavy hitters of x_S exactly, but that are sufficient to find a small subset S' of S that contains (almost) all of the heavy hitters. If S' is also fairly random, then $x_{S'}$ will have almost all the signal while only a small fraction of the noise, so it has much higher SNR.

A first attempt for finding such a subset S' could be as follows. Suppose that the SNR is C—that is, the largest k elements of x_S have C times more ℓ_2^2 mass than the other elements. For some parameter $D \gg k$, we construct a vector $y \in \mathbb{R}^D$ by hashing x_S as per Count-Sketch [CCF02]—so each coordinate $i \in S$ is assigned a random coordinate $h(i) \in [D]$ and sign $s_i \in \{\pm 1\}$, and $y_j = \sum_{i:h(i)=j} x_i s_i$. The SNR of y will also be about C, so we can learn a lot about y by performing nonadaptive $C^{0.1}$-approximate sparse recovery of it. This takes $O(k \log_C(D/k) \cdot \log^* k)$ measurements [PW12], so we can set $D = kC^{\log^{1/R} n}$ and fit within our sample complexity budget. The top $O(k)$ elements of y will contain most of the heavy hitters of x, so we can set S' to the preimage of those elements; this has size about $k(|S|/D) = |S|/C^{\log^{1/R} n}$. Hence the C used in the next round will be roughly a $C^{\log^{1/R} n}$ factor larger; after R rounds of this, C will grow from constant to n^{10}, at which point the problem is easy. In fact, the Rth round can estimate x_U directly to avoid needing an extra cleanup round.

This approach *mostly* works, but suffers from one major flaw: in every stage, the set S' can miss a small fraction of the heavy hitters. Even with zero noise, heavy hitters that collide in $[D]$ can cancel out when combining into y, causing them to disappear from S' and from the final reconstruction. Previous algorithms based on the Count-Sketch hashing often run into this issue, and address it by cleaning up the residual afterward [GLPS10, IPW11, LNW18, NSWZ18]. In our context, such a solution would require more rounds of adaptivity.

Triple gaussian hashing. We introduce a new approach to hashing for sparse recovery that lets us avoid any major false negatives, based on replacing the signs s_i with gaussians $g_i \sim N(0,1)$ in the computation of y, so $y_j = \sum_{i:h(i)=j} x_i g_i$. This hash avoids the issue described above with zero noise, since if $x_i \neq 0$ then $y_{h(i)} \neq 0$ with probability 1.

To understand how this hash behaves with noise, consider the following example: $x = v + w$ where $v \in \{0,1\}^n$ is k-sparse and w is gaussian with norm 1. Successful $O(1)$-approximate recovery of x must find all but $O(1)$ elements in $\operatorname{supp}(v)$. In the gaussian hash y of x, the image of w is still very spread out with norm about 1, but the image of v is no longer binary: each entry $\left|y_{h(i)}\right|$ has a $\Theta(\epsilon)$ chance of being less than ϵ. This means about $k^{2/3}$ positions in $h(\operatorname{supp}(v))$ will be smaller than $1/k^{1/3}$. Since these collectively have norm 1, successful $O(1)$-approximate recovery of y could miss all $k^{2/3}$ of these positions, which would be a problem.

We avoid such false negatives by repeating the hash three times, with the same h and different g, and applying sparse recovery separately to the three different y. In the above example, where coordinates are missing from sparse recovery with probability $1/k^{1/3}$, the expected number of coordinates that are missed three times in a row is $O(1)$. In general, the chance q_i that $h(i)$ is recovered by the sparse recovery algorithm may depend on i and x_i in a complicated fashion that we can't control, since the sparse recovery algorithm is a black box. Still, we can show that the (k, C)-approximate recovery guarantee implies that the expected mass lost all three times—$\sum_i q_i^3 x_i^2$—is bounded in terms of the noise level.

Our triple Gaussian hash thus gives a set of locations without any significant false negatives, so we do not need to clean up the missing coordinates. We believe that this technique is likely to have applications in other, nonadaptive, sparse recovery settings.

Decreasing the noise. So far, we have outlined how the algorithm gets a small set S' that does not lose much signal mass. Another key part of the argument is that $x_{S'}$ should have much less noise than x_S. Since S' is much smaller than S, this would be immediate if S' were random. However, since S' is the preimage of the largest coordinates of y, it is biased towards the elements of x containing more noise.

We show that this effect is limited: after dropping $O(k)$ noise coordinates, the rest of the noise shrinks by a factor of $\sqrt{D/k}$. We tolerate the $O(k)$ large noise coordinates by increasing the sparsity k by a constant factor in each round; and the $\sqrt{D/k}$ factor, although not as good as the D/k factor decrease in $|S|$, is still $C^{\Theta(\log^{1/R} k)}$.

By choosing the parameters carefully, we can ensure the total error and total failure probability remain small over all rounds. The sample complexity for constant R is $O(k \log^{1/R}(n/k) \cdot \log^* k)$, which comes from black-box calls to the C-approximate nonadaptive sparse recovery algorithm. If the extra $\log^* k$ factor is removed from that, our sample complexity will become the optimal $O(k \log^{1/R}(n/k))$.

2 Lower Bound

In this section we present a lower bound on the total number of linear measurements for adaptive R-round $(k, O(1))$-sparse recovery.

The instance for which we show a lower bound is as follows: Alice divides the domain $[N]$ into k contiguous "bins" of size n each (indexed by $[k]$) and for every bin i chooses $x_i \in [n]$ uniformly at random. Alice then chooses i.i.d. Gaussian noise $w \in \mathbb{R}^N$ with $\mathbb{E}[\|w\|_2^2] = \sigma^2 = \Theta(k)$, then sets $x = w + \sum_{i=1}^k e_{(i-1)n+x_i}$. Bob performs R adaptive rounds of linear measurements on x, getting $y^r = A^r x = (y_1^r, \dots, y_{m_r}^r)$ in each round r. Let X_i and Y^r denote the random variables from which x_i and y^r are drawn, respectively. In order for sparse recovery to succeed under an appropriate setting of constant for σ^2, at least $k/2$ of the variables $X_1, \dots, X_k$ must be recovered.

For ease of notation, we use j_1^r to denote the tuple $(j_1, \dots, j_r)$. Similarly, j_1^{i-1}, J_i denotes the tuple $(j_1, \dots, j_{i-1}, J_i)$ where the distinction in the context of this proof is that $j_1, \dots, j_{i-1}$ are fixed and J_i is a random variable. We use $(X)_W$ for $W = \{i_1, \dots, i_{|W|}\} \subseteq [k]$ to denote the tuple $(X_{i_1}, \dots, X_{i_W})$.

Definition 2.1. *Given random variables $X_1, \dots, X_k \in [n]$ with joint probability distribution $p(l_1, \dots l_k) = Pr[X_1 = l_1, \dots, X_k = l_k]$, we define the **sequentially conditioned partition** of the domain of X_i as follows*

1. $T_{j_1^i} = \{l \in [n] \mid 2^{j_i} < np_i(l \mid X_1 \in T_{j_1^1}, \dots, X_{i-1} \in T_{j_1^{i-1}}) \le 2^{j_i+1}\}$ *for* $j_i > 0$
2. $T_{j_1^i} = \{l \in [n] \mid np_i(l \mid X_1 \in T_{j_1^1}, \dots, X_{i-1} \in T_{j_1^{i-1}}) \le 2\}$ *for* $j_i = 0$.

where p_i denotes the marginal distribution of X_i. Additionally, we define the probability mass within each partition as $q_{j_1^i} = \sum_{l \in T_{j_1^i}} p_i(l \mid X_1 \in T_{j_1^1}, \dots, X_{i-1} \in T_{j_1^{i-1}})$. So, if we fix $j_1, \dots, j_{i-1}$, we have $\sum_{j_i=0}^{\infty} q_{j_1^i} = 1$.

Denote the event $X_1 \in T_{j_1^1}, \dots, X_i \in T_{j_1^i}$ by $E_{j_1^i}$. These partitions are defined in such a way that $(X_i \mid E_{j_1^i})$ is close to uniform over its support. This allows us to bound the maximum conditional probability within a sequentially conditioned partition of the domain of X_i. So,

$$M_{j_1^i} \overset{\text{def}}{=} n \cdot \max_{l \in T_{j_1^i}} \left(p_i(l \mid E_{j_1^i}) \right) \le \frac{2^{j_i+1}}{q_{j_1^i}} \tag{7}$$

Additionally, for the random variable $(X_i \mid E_{j_1^{i-1}})$ over $[n]$, we define the number of bits that the distribution knows about the location of X_i as:

$$b_i(j_1, \dots, j_{i-1}) = H(\mathcal{U}([n])) - H(X_i \mid E_{j_1^{i-1}}) = \log(n) - H(X_i \mid E_{j_1^{i-1}}).$$

We show for every i and j_1^{i-1} that $M_{j_1^{i-1}, J_i}$ is small on average over J_i:

Lemma 2.2. *Consider random variables $X_1, \dots, X_k \in [n]$ with joint probability distribution $p(l_1, \dots l_k) = \Pr[X_1 = l_1, \dots, X_k = l_k]$ and suppose we know that $X_1 \in T_{j_1}, \dots X_{i-1} \in T_{j_1^{i-1}}$. Suppose that J_i is the discrete random variable that denotes the j_i such that $X_i \in T_{j_1^i}$ conditioned on $X_1 \in T_{j_1^1}, \dots X_{i-1} \in T_{j_1^{i-1}}$. Then,*

$$\mathop{\mathbb{E}}_{J_i}[\log\left(1 + M_{j_1^{i-1}, J_i}\right)] \le O(b_i(j_1, \dots, j_{i-1}) + 1)$$

Proof. Using (7) we get the bound:

$$\begin{aligned}
\mathop{\mathbb{E}}_{J_i}[\log\big(1+M_{j_1^{i-1},J_i}\big)] &\le \mathop{\mathbb{E}}_{J_i}\Big[\log\big(1+\frac{2^{J_i+1}}{q_{j_1^{i-1},J_i}}\big)\Big] \\
&= \sum_{j_i=0}^{\infty} q_{j_1^i}\log\big(1+\frac{2^{j_i+1}}{q_{j_1^i}}\big) \\
&\le \sum_{j_i=0}^{\infty} q_{j_1^i}\log(1+2^{j_i+1}) + \sum_{j=0}^{\infty} q_{j_1^i}\log\big(1+\frac{1}{q_{j_1^i}}\big) \\
&\le \sum_{j_i=0}^{\infty} j_i q_{j_1^i} + \sum_{j_i=0}^{\infty} 2q_{j_1^i} + \sum_{j=0}^{\infty} q_{j_1^i}\log\big(1+\frac{1}{q_{j_1^i}}\big).
\end{aligned}$$

Since $\sum_{j_i=0}^{\infty} q_{j_1^i} = 1$, Lemma A.2 implies:

$$\mathop{\mathbb{E}}_{J_i}[\log\big(1+M_{i,j_1^{i-1},J_i}\big)] \le O(b_i(j_1,\dots,j_{i-1})+1)$$

□

For every i and collection of measurement vectors $v_1,\dots,v_m$, we now show that the amount of "signal energy" for X_i is bounded even conditioned on the partition J_1^k.

Lemma 2.3. *Let $X_1,\dots,X_k$ be random variables over $[n]$ with joint probability distribution $p(l_1,\dots l_k) = Pr[X_1 = l_1,\dots,X_k = l_k]$. For all $i \in [k]$, define $b_i = \log(n) - H(X_i \mid X_1,\dots,X_{i-1})$. Let $v_1,\dots,v_m \in \mathbb{R}^{nk}$ be a fixed set of vectors. Define random variable $Z_{i,j_1^k} \overset{\text{def}}{=} \mathbb{E}_{X_i|E_{j_1^k}}[\sum_{s=1}^m (v_s)^2_{n\cdot(i-1)+X_i}]$ and random variable $M_{i,j_1^k} \overset{\text{def}}{=} n\cdot\max_{l\in T_{j_1^i}}(p_i(l \mid E_{j_1^k}))$. Then, for any $i \in [k]$,*

1. $\log(1+Z_{i,J_1^k}) \le \log\big(1+(\frac{\sum_{s=1}^m \|v_{s|i}\|_2^2}{n})\big) + \log(1+M_{i,J_1^k})$
2. $\mathbb{E}_{J_1,\dots,J_k}\big[\log(1+M_{i,J_1^k})\big] \le O(b_i+1)$

where $v_{s|i}$ denotes the restriction of v_s to the the index set $[n(i-1)+1, ni]$.

Proof. Using the definition of Z_{i,j_1^k} and M_{i,j_1^k}, we can write:

$$Z_{j_1^i} = \sum_{s=1}^m\sum_{t\in[n]}(v_s)^2_{n\cdot(i-1)+t}\cdot\Pr[X_i = t \mid E_{j_1^i}] \le \Big(\frac{\sum_{s=1}^m \|v_{s|i}\|_2^2}{n}\Big)M_{j_1^i}$$

Let J_i be the discrete random variable that denotes the j_i such that $X_i \in T_{j_1^i}$ conditioned on $E_{j_1^{i-1}}$. Then, using Lemma 2.2,

$$\mathop{\mathbb{E}}_{J_i}[\log(1+M_{j_1^{i-1},J_i})] \le O(b_i(j_1,\dots,j_{i-1})+1)$$

We wish to bound $\mathbb{E}_{J_1\dots,J_k}[\log(1+M_{i,J_1^k})]$. Using the concavity of log,

$$\mathop{\mathbb{E}}_{J_1\dots,J_k}[\log(1+Z_{i,J_1^k})] \le \mathop{\mathbb{E}}_{J_1\dots,J_i}[\log\big(1+\mathop{\mathbb{E}}_{J_{i+1},\dots,J_k}[M_{i,J_1^k}]\big)]$$

From the definitions of M_{i,J_1^k} and $M_{J_1^i}$, we know that:

$$\begin{aligned}\mathop{\mathbb{E}}_{J_{i+1},\ldots,J_k}[M_{i,J_1^k}] &= \mathop{\mathbb{E}}_{J_{i+1},\ldots,J_k}\left[\mathop{\mathbb{E}}_{X_i|E_{J_1^k}}\left[n \cdot \max_{l \in T_{j_1^i}}(p_i(l \mid E_{j_1^k})\right]\right] \\ &= \mathop{\mathbb{E}}_{X_i|E_{J_1^i}}\left[n \cdot \max_{l \in T_{J_1^i}}(p_i(l \mid E_{J_1^i})\right] \\ &= M_{J_1^i}\end{aligned}$$

So,

$$\begin{aligned}\mathop{\mathbb{E}}_{J_1\ldots,J_k}[\log(1+M_{i,J_1^k})] &\leq \mathop{\mathbb{E}}_{J_1\ldots,J_i}[\log(1+M_{J_1^i})] \\ &\leq O(\mathop{\mathbb{E}}_{J_1\ldots,J_{i-1}}[b_i(J_1,\ldots,J_{i-1})+1])\end{aligned}$$

Since conditioning decreases entropy, we also know: $\mathbb{E}_{J_1,\ldots,J_{i-1}}[b_i(J_1,\ldots,J_{i-1})] = H(\mathcal{U}([n])) - H(X_i \mid E_{J_1^{i-1}}) \leq H(\mathcal{U}([n])) - H(X_i \mid X_1 \ldots X_{i-1}) = b_i$ and hence,

$$\mathop{\mathbb{E}}_{J_1\ldots,J_k}[\log(1+M_{i,J_1^k})] \leq O(b_i+1)$$

□

We can now show the key lemma, that if b bits of information are known from the previous rounds, the next round will only reveal roughly $m(\frac{b}{k}+\log k)$ more bits of information.

Lemma 2.4. *Let $X_1,\ldots,X_k$ be random variables over $[n]$ and $W = \{l_1,l_2,\ldots,l_{|W|}\} \subseteq [k]$ be a subset such that $|W| = ck$ where $c \leq 1$ is a constant. We define the number of bits of information revealed about the subset W, conditioned on the variables $\{X\}_{[n]\setminus W}$ as*

$$b = |W|\log(n) - H((X)_W \mid (X)_{[n]\setminus W})$$

Define $\tilde{X} = \sum_{i=1}^k e_{(i-1)\cdot n+X_i} + N(0, I_N\sigma^2/N)$ where $\sigma^2 = \Theta(k)$. Consider a fixed set of measurement vectors $v_1,\ldots,v_m \in \mathbb{R}^N$ independent of $X_1,\ldots,X_k$ with $\|v_j\|_2^2 = N$ for all $j \in [m]$, and define $Y_j = \langle v_j, \tilde{X}\rangle$. Then, for all $0 < \alpha < \gamma < 1$, with probability $1-\gamma$, there exists a subset $W' \subseteq W$, $|W'| \geq (1-\frac{\alpha}{\gamma})\,|W|$ such that

$$I((X)_{W'}; Y_1^m \mid (X)_{[n]\setminus W'}, W') \leq c_3\frac{m}{\alpha}\frac{b}{k} + m\log(k) + \frac{c_4 m}{\alpha} + c_2(b+k)$$

for some constants c_2, c_3, c_4.

Proof. Since we wish to condition out the indices not in W, we may perform the analysis on a fixed set of values for $(X)_{[n]\setminus W}$ and then use the fact that $I(A;B|C) = \mathbb{E}_c[I(A;B \mid C=c)]$ to arrive at the theorem statement.

Suppose that for all $i \in [n]\setminus W$, $X_i = x_i$. Then, the number of bits of information known about $(X)_W$ may be denoted $\tilde{b} = b((x)_{[n]\setminus W}) = |W|\log(n) - H((X)_W \mid (x)_{[n]\setminus W})$. Now, we may construct sequentially conditioned partitions only on the domains of $(X)_W$ and in the order $l_1, l_2, \ldots, l_{|W|}$. We will denote by J_W the conditioning over the partitions of the $(X)_W$ in the chosen order.

Let $W' \subseteq W$ be a set of indices which we shall choose later. Consider the mutual information between a set of random variables $(X)_{W'}$ and the measurements conditioned on the variables not in W'. Using the chain rule of mutual information:

$$\begin{aligned} &I((X)_{W'}; Y_1^m \mid (X)_{W \setminus W'}, (x)_{[n]\setminus W}, W') \\ &\qquad \leq I((X)_{W'}; Y_1^m \mid E_{J_W}, (X)_{W\setminus W'}, (x)_{[n]\setminus W}, W') + H(J_W \mid (x_{[n]\setminus W})) \end{aligned}$$

Using Lemma A.2, there exists a constant c_2 such that for all $i \in [|W|]$, $H(J_{l_i} \mid J_{l_1}^{l_{i-1}}, (x)_{[n]\setminus W}) \leq c_2(\log(n) - H(X_{l_1} \mid J_{l_1}^{l_{i-1}}, (x)_{[n]\setminus W}) + 1)$. Since conditioning only reduces entropy, we know that $H(J_{l_i} \mid J_{l_1}^{l_{i-1}}, (x)_{[n]\setminus W}) \leq c_2(\log(n) - H(X_{l_1} \mid X_{l_1}, \ldots, X_{l_{i-1}}, (x)_{[n]\setminus W}) + 1)$. So, $H(J_W \mid (x_{[n]\setminus W})) = \sum_{i\in[|W|]} H(J_{l_i} \mid J_{l_1}^{l_{i-1}}, (x)_{[n]\setminus W}) \leq c_2(\tilde{b} + k)$. Using the definition of conditional mutual information, and the fact that measurements are chosen independently,

$$\begin{aligned} &I((X)_{W'}; Y_1^m \mid (X)_{W\setminus W'}, (x)_{[n]\setminus W}, W') \\ &\qquad \leq \mathop{\mathbb{E}}_{(x)_{W\setminus W'}} \Big(\sum_{s=1}^{m} I((X)_{W'}; Y_s \mid E_{J_W}, (x)_{[n]\setminus W'}, W') \Big) + c_2(\tilde{b} + k) \end{aligned}$$

Applying the Data Processing Inequality to the first term, we get:

$$\begin{aligned} &I((X)_{W'}; Y_1^m \mid (X)_{W\setminus W'}, (x)_{[n]\setminus W}, W') \\ &\qquad \leq \mathop{\mathbb{E}}_{(x)_{W\setminus W'}} \Big(\sum_{s=1}^{m} I\big(\sum_{i\in W'} (v_s)_{(i-1)n+X_i}; Y_s \mid E_{J_W}, (x)_{[n]\setminus W'}, W' \big) \Big) + c_2(\tilde{b} + k) \end{aligned}$$

Observe that $Y_s = \sum_{i\in W'} (v_s)_{(i-1)n+X_i} + \sum_{i\in[n]\setminus W'} (v_s)_{(i-1)n+x_i} + N(0, \sigma^2)$. Since $(x)_{[n]\setminus W'}$ are conditioned out, we may subtract their contribution and we get:

$$\begin{aligned} &I\big(\sum_{i\in W'} (v_s)_{(i-1)n+X_i}; Y_s \mid E_{J_W}, (x)_{[n]\setminus W'}, W' \big) \\ &\qquad = I\big(\sum_{i\in W'} (v_s)_{(i-1)n+X_i}; \sum_{i\in W'} (v_s)_{(i-1)n+X_i} + \eta \mid E_{J_W}, (x)_{[n]\setminus W'}, W' \big) \end{aligned}$$

where $\eta \sim N(0, \sigma^2)$ is additive white gaussian noise. We may now use the Shannon-Hartley Theorem (Theorem A.1) on this quantity to bound the mutual information in terms of a variance term:

$$\begin{aligned} &I((X)_{W'}; Y_1^m \mid (X)_{W\setminus W'}, (x)_{[n]\setminus W}, W') \\ &\qquad \leq \mathop{\mathbb{E}}_{(x)_{W\setminus W'}} \sum_{s=1}^{m} \mathop{\mathbb{E}}_{j_W} \big[\log\big(1 + \frac{\mathbb{E}_{(X)_{W'}\mid E_{j_W}, (x)_{[n]\setminus W'}} (\sum_{i\in W'} [(v_s)_{(i-1)n+X_i}])^2}{\sigma^2} \big) \big] + c_2(\tilde{b} + k) \end{aligned}$$

Using Cauchy-Schwartz, then applying Jensen's inequality, and then using the convexity of log and the definition of Z_{i,J_W}:

$$\begin{aligned} &I((X)_{W'}; Y_1^m \mid (X)_{W\setminus W'}, (x)_{[n]\setminus W}, W') \\ &\qquad \leq \mathop{\mathbb{E}}_{j_W} \Big(\sum_{s=1}^{m} \log\big(1 + |W'| \frac{\sum_{i\in W'} \mathbb{E}_{X_i\mid E_{j_W}} [(v_s)^2_{(i-1)n+X_i}]}{\sigma^2} \big) \Big) + c_2(\tilde{b} + k) \\ &\qquad \leq m \mathop{\mathbb{E}}_{j_W} \Big(\log\big(1 + \frac{|W'| \sum_{i\in W'} \mathbb{E}_{X_i\mid E_{j_W}} [\sum_{s=1}^{m} (v_s)^2_{(i-1)n+X_i}]}{\sigma^2 \cdot m} \big) \Big) + c_2(\tilde{b} + k) \\ &\qquad = m \mathop{\mathbb{E}}_{j_W} \Big(\log\big(1 + \frac{|W'| \sum_{i\in W'} Z_{i,j_W}}{\sigma^2 \cdot m} \big) \Big) + c_2(\tilde{b} + k) \qquad (8) \end{aligned}$$

So, we need to set W' to be the set that contains indices in W with low values of Z_{i,j_W}. More precisely, for a fixed partition sequence j_w, we set $W' = \{i \in W \mid \log(1 + Z_{i,j_W}) < \log(1 + (\sum_{s=1}^{m} \left\|v_{s|i}\right\|_2^2 /n)) + (c_3/\alpha) \cdot ((\tilde{b}/k) + 1)\}$ where c_3 is a constant which will be set later. Suppose $M_{i_l,j_W} = n \cdot \max_{l \in T_{j_1^i}}(\Pr[X_{i_l} = x_{i_l} \mid E_{j_W}])$. We may use Lemma 2.3 on the indices in W since the indices in $[n] \setminus W$ has been fixed. So, there is a constant c_1 such that for all $i_l \in W$,

$$\mathop{\mathbb{E}}_{J_{i_1},\dots,J_{i_{|W|}}} \left[\log\left(1 + M_{i_l,J_W}\right)\right] \le c_1(\tilde{b}_{i_l} + 1)$$

where $\tilde{b}_{i_l} = \log(n) - H(X_{i_l} \mid X_{i_1}, \dots, X_{i_{l-1}}, (x)_{[n]\setminus W})$. Observe that $\sum \tilde{b}_{i_l} = |W| \log(n) - \sum H(X_{i_l} \mid X_{i_1}, \dots, X_{i_{l-1}}, (x)_{[n]\setminus W}) = |W| \log(n) - H(X_{i_1}, \dots, X_{i_{|W|}} \mid (x)_{[n]\setminus W}) = \tilde{b}$. Suppose I is distributed uniformly over W. Then using Jensen's inequality,

$$\begin{aligned}
\mathop{\mathbb{E}}_{I}\Big[\mathop{\mathbb{E}}_{J_W}\left[\log\left(1 + M_{I,J_W}\right)\right]\Big] &\le c_1 \mathop{\mathbb{E}}_{I}[\tilde{b}_I + 1] \\
&\le \frac{c_1(\tilde{b} + k)}{|W|} \\
&\le \frac{c_3(\tilde{b} + k)}{k}
\end{aligned}$$

where the third inequality follows because we are only considering W such that $|W| = ck$ for a constant fraction c and $c_3 = (c_1/c)$.

Now, since each $M_{I,J_W} \ge 0$, we may use Markov's inequality to show that:

$$\Pr_{(I,J_W)}\left[\log(1 + M_{I,J_W}) \ge \frac{c_3(\tilde{b} + k)}{\alpha k}\right] \le \alpha$$

Define $U = \{(i, j_W) \mid i \in W, \log(1 + M_{i,j_W}) < c_3(\tilde{b} + k)/\alpha k\}$ and for all $i \in W$, we may define $p_i^U = \Pr_{J_W}[(i, J_W) \notin U]$. Note that $E[|W \setminus W'|] \le \sum_{i\in W} p_i^U \le \alpha |W|$ and using Markov's inequality, we may say that $\Pr[|W \setminus W'| \ge \alpha |W| /\gamma] \le \gamma$. Plugging the definition of W' and $\sigma^2 = \Theta(k) = c'k$, into (8) gives

$$\begin{aligned}
&I((X)_{W'}; Y_1^m \mid (X)_{W\setminus W'}, (x)_{[n]\setminus W}, W') \\
&\qquad \le m \log\left(1 + \frac{|W'| \cdot \sum_{i\in W'} 2^{\log(1+\frac{1}{n}\sum_{s=1}^{m}\left\|v_{s|i}\right\|_2^2) + \frac{1}{\alpha}\left(c_3(\tilde{b}/k)+1\right)}}{c'mk}\right) + c_2(\tilde{b} + k) \\
&\qquad \le m \log\left(1 + \frac{|W'| \cdot 2^{\frac{1}{\alpha}\left(c_3(\tilde{b}/k)+1\right)} \sum_{i\in W'}\left(1 + \frac{1}{n}\sum_{s=1}^{m}\left\|v_{s|i}\right\|_2^2\right)}{c'mk}\right) + c_2(\tilde{b} + k) \qquad (9)
\end{aligned}$$

Since $\sum_{i\in[n]} \left\|v_{s|i}\right\|_2^2 = N$, we know that $\sum_{i\in W'}\left(1 + \frac{1}{n}\sum_{s=1}^{m}\left\|v_{s|i}\right\|_2^2\right) \le |W'| + \frac{Nm}{n} = |W'| + km$. Plugging this into (9), we get:

$$\begin{aligned}
&I((X)_{W'}; Y_1^m \mid (X)_{W\setminus W'}, (x)_{[n]\setminus W}, W') \\
&\qquad \le m \log\left(1 + \frac{|W'| \cdot 2^{\frac{c_3}{\alpha}(\tilde{b}/k+1)} \cdot (W' + km)}{c'mk}\right) + c_2(\tilde{b} + k) \\
&\qquad \le m \log(1 + \left|W'\right| /c') + m \log(1 + 2^{\frac{c_3}{\alpha}(\tilde{b}/k+1)}) + c_2(\tilde{b} + k) \\
&\qquad \le m \log(1 + k) + m \log(1 + c/c') + m \log(1 + 2^{\frac{c_3}{\alpha}(\tilde{b}/k+1)}) + c_2(\tilde{b} + k) \\
&\qquad \le 2m + m \log(k) + m \log(1 + c/c') + 2m + m \log(2^{\frac{c_3}{\alpha}(\tilde{b}/k+1)}) + c_2(\tilde{b} + k) \\
&\qquad \le m \log(k) + \frac{c_3 m\tilde{b}}{\alpha k} + \frac{c_4 m}{\alpha} + c_2(\tilde{b} + k)
\end{aligned}$$

where $c_4 = 4 + \log(1 + c/c')$ is a constant. So, with probability $1 - \gamma$ there exists a set $W' \subseteq W$ such that $|W'| \geq (1 - \alpha/\gamma)\,|W|$ and

$$I((X)_{W'}; Y_1^m \mid (X)_{W\setminus W'}, (x)_{[n]\setminus W}, W') \leq c_3 \frac{m\tilde{b}}{\alpha k} + m\log(k) + \frac{c_4 m}{\alpha} + c_2(\tilde{b} + k)$$

Now, taking the expectation of this term over $(x)_{[n]\setminus W}$, with probability $1 - \gamma$ there exists a set $W' \subseteq W$ such that $|W'| \geq (1 - \alpha/\gamma)\,|W|$ and

$$I((X)_{W'}; Y_1^m \mid (X)_{[n]\setminus W'}, W') \leq c_3 \frac{mb}{\alpha k} + m\log(k) + \frac{c_4 m}{\alpha} + c_2(b + k)$$

□

By applying Lemma 2.4 every round, we get the desired lower bound on m.

Theorem 2.5. *Any scheme using R adaptive rounds with $m_1, \dots, m_R$ measurements in each round and m total measurements has a set $W \subseteq [k], |W| \geq \Omega(k)$ such that with probability $\geq 3/4$*

$$I((X_i)_{i\in W}; Y_1, \dots, Y_m \mid (X_i)_{i\notin W}, W) \leq \left(\prod_{j=2}^{R} \big(2c_5 + \frac{32 c_6 R^2 m_j}{k}\big)\right) \max\{k\log(k), m_1\}$$

where c_5 and c_6 are constants. Consequently, for (k, C)-sparse recovery with $C = O(1)$, it must hold that

$$m \geq \frac{k}{C'R} \min\left\{ \Big(\log(N/k)\Big)^{1/R}, \left(\frac{\log(N/k)}{\log(k)}\right)^{1/(R-1)} \right\}$$

for some constant C'.

Proof. Let A^r be the measurement matrix in round r (which we may assume is deterministically chosen as a function of all the previous rounds). Since the first round is non-adaptive, we may use the Shannon-Hartley Theorem (as per [PW12]) to show that for $W_2 = [k]$,

$$I((X_i)_{i\in W_2}; Y_{1,1}, \dots, Y_{1,m_1} \mid (X_i)_{i\notin W_2}, W_2) \leq m_1$$

For each round r, by p_r we denote Bob's prior distribution at the beginning of that round. We also denote by $b^{(r)} = |W_r|\log(n) - H(X_{W_r} \mid X_{[n]\setminus W_r})$ the number of bits of information in the prior $(X_i)_{i\in W_r}$ conditioned on $(X_i)_{i\notin W_r}$.

Since the rows of A^r are deterministic given the observations in previous rounds, we may apply Lemma 2.4 with $\alpha = 1/(16R^2)$, $\gamma = 1/4R$, and with probability $(1-(1/4R))$ obtain a set $W_{r+1} \subseteq W_r$ such that $|W_{r+1}| \geq (1 - \frac{\alpha}{\gamma})\,|W_r|$ and:

$$I((X_i)_{i\in W_{r+1}}; Y^{r+1} \mid y^1, \dots, y^r, (X_i)_{i\notin W_{r+1}}, W_{r+1}) \leq c_3 \frac{m_{r+1} b_r}{\alpha k} + m_{r+1}\log(k) + \frac{c_4 m_{r+1}}{\alpha} + c_2(b_r + k)$$

Let us define $B_{r+1} = I((X_i)_{i\in W_{r+1}}; Y^{r+1}, \dots, Y^1 \mid (X_i)_{i\notin W_{r+1}}, W_{r+1})$. Using the chain rule of mutual information for $r > 1$

$$\begin{aligned} B_{r+1} &= I((X_i)_{i\in W_{r+1}}; Y^r, \dots, Y^1 \mid (X_i)_{i\notin W_{r+1}}) \\ &\quad + I((X_i)_{i\in W_{r+1}}; Y^{r+1} \mid Y^r, \dots, Y^1, (X_i)_{i\notin W_{r+1}}, W_{r+1}) \\ &\leq B_r + \mathop{\mathbb{E}}_{y^1,\dots,y^r}[I((X_i)_{i\in W_{r+1}}; Y^{r+1} \mid y^1, \dots, y^r, (X_i)_{i\notin W_{r+1}}, W_{r+1})] \end{aligned}$$

So,

$$\begin{aligned} B_{r+1} &\le B_r + c_3\frac{m_{r+1}B_r}{\alpha k} + m_{r+1}\log(k) + \frac{c_4 m_{r+1}}{\alpha} + c_2(B_r + k) \\ &\le \big(c_5 + \frac{c_3 m_{r+1}}{\alpha k}\big)B_r + m_{r+1}\log(k) + \frac{c_4 m_{r+1}}{\alpha} + c_2 k \end{aligned} \tag{10}$$

where $c_5 = c_2 + 1$. We know using the Shannon-Hartley Theorem that $B_1 \le m_1$. Further, we assume that $B_1 \ge k\log(k)$. While this weakens our lower bound, it allows us to make a cleaner inductive argument into Claim A.3. Plugging $\alpha = 1/16R^2$ in Claim A.3, we get:

$$B_R \le \Bigg(\prod_{j=2}^{R}\big(2c_5 + \frac{32c_6R^2m_j}{k}\big)\Bigg)\max\{k\log(k), m_1\}$$

It follows using the AM-GM inequality that:

$$B_R \le \max\Big\{k\cdot\Big(2c_5 + \frac{32c_6R\cdot m}{k}\Big)^R, k\log(k)\cdot\Big(2c_5 + \frac{32c_6R\cdot m}{k}\Big)^{(R-1)}\Big\}$$

So, after R rounds with probability $\ge 3/4$, we have a set W_R such that $|W_R| \ge (1-\frac{\alpha}{\gamma})^R k \ge e^{-4}k$ with $I((X_i)_{i\in W_R}; Y^R, \ldots, Y^1 \mid (X_i)_{i\notin W_R}, W_R)$ bounded as above. We may scale the variance of w (gaussian noise) by appropriate constants, so that for sparse recovery to succeed $k(1-\frac{1}{2e^4})$ indices must be fully recovered with probability $\ge 3/4$. So, for the set W_R it must hold that $I((X_i)_{i\in W_R}; Y^R, \ldots, Y^1 \mid (X_i)_{i\notin W_R}, W_R) \ge \frac{k}{2e^4}\log(N/k)$ and as a consequence, it must hold that:

$$\max\Big\{\Big(2c_5 + \frac{32c_6R\cdot m}{k}\Big)^R, \Big(2c_5 + \frac{32c_6R\cdot m}{k}\Big)^{(R-1)}k\log(k)\Big\} \ge \frac{k}{2e^4}\log(N/k)$$

If we simplify this and set $C' = 32c_6$, we get

$$m \ge \min\Bigg\{\frac{k}{C'R}\Bigg(\log(N/k)\Bigg)^{1/R}, \frac{k}{C'R}\Bigg(\frac{\log(N/k)}{\log(k)}\Bigg)^{1/(R-1)}\Bigg\}$$

□

If we restrict our sparsity parameter k to be $O(2^{(\log(N))^{1/R}})$ we observe that this lower bound is tight.

Corollary 2.6. *Let $C > 1$. Any (k, C)-sparse recovery scheme for vectors in $\mathbb{R}^N$ that uses R adaptive rounds and m total measurements with $k = O(2^{\log^{1/R} N})$ must satisfy*

$$m \ge \frac{k}{C'R}\Bigg(\log(N/k)\Bigg)^{1/R}$$

for some constant C'.

3 Upper Bound

In this section we present our algorithm for (k, C)-sparse recovery in R rounds. The main goal is to prove Theorem 3.9 which shows that Algorithm 3.2 achieves (k, C) sparse recovery using $O(k\log_C(n/k)^{1/R}\log^*(k)\cdot 2^R)$ measurements. Lemma 3.6 shows that in each round we lose a small amount of mass from the vector. Lemma 3.7 and Lemma 3.8 show that with a constant increase in the sparsity parameter from one round to the next, we can ensure that the "noise" carried over to the next round decreases by a factor.

3.1 Preliminaries

We start with a few definitions. Let x be an n-dimensional vector.

Definition 3.1. *Define*

$$H_k(x) = \underset{\substack{S \in [n] \\ |S| = k}}{\arg\max} \|x_S\|_2$$

to be the largest k coefficients in x.

Definition 3.2. *Define the "noise" or "error"*

$$\operatorname{Err}^2(x, k) = \left\| x_{\overline{H_k(x)}} \right\|_2^2$$

Definition 3.3. *Given a vector x, a recovered vector x^* satisfies (k, C)-sparse recovery under the ℓ_2/ℓ_2 guarantee if:*

$$\|x - x^*\|_2^2 \le C \operatorname{Err}^2(x, k)$$

Definition 3.4. *Given a hash function $h : [n] \to [D]$, a (D, h)-**gaussian hash projection** of a vector $x \in \mathbb{R}^n$ into $\mathbb{R}^D$ is given by $y \in \mathbb{R}^D$ such that $y_j = \sum_{i:h(i)=j} x_i \cdot g_i$ where $g_i \sim \mathcal{N}(0, 1)$ is i.i.d normal with variance 1 and mean 0.*

We denote by HighSNR-Recover(x, k, C, δ) a black-box algorithm which makes linear measurements on the input x and whose output achieves (k, C) sparse recovery with probability $1 - \delta$. The best known algorithm for achieving (k, C)-sparse recovery when $C \ge 1$ is the algorithm from [PW12]:

Theorem 3.5. *There exists an algorithm that takes $O\big(k \log^*(k) \log_C(n/k) \log(1/\delta)\big)$ linear measurements and outputs a k-sparse vector that achieves (k, C)-sparse recovery under the ℓ_2/ℓ_2 guarantee with success probability $1 - \delta$.*

3.2 Algorithm

```
procedure 1-RoundSNRBoost($x$, $n$, $D$, $C$, $k$, $\delta$) $\triangleright$ Recover most of the mass of heavy hitters while reducing noise by factor $D/k$
    For $i \in [n], h(i) \leftarrow [D]$
    For $i \in [n], t \in \{1, 2, 3\}$    $g_i^{(t)} \leftarrow \mathcal{N}(0, 1)$
    For $j \in [D], t \in \{1, 2, 3\}$ define $y_j^{(t)} = \sum_{i \in h^{-1}(j)} g_i^{(t)} x_i$
    For $t \in \{1, 2, 3\}$ , $U^{(t)} \leftarrow \text{supp}(\text{HighSNR-Recover}(y^{(t)}, k, C, \delta/3))$
    return $\cup_{j \in U^{(1)} \cup U^{(2)} \cup U^{(3)}} h^{-1}(j)$
end procedure
```

Algorithm 3.1: 1 round SNR-Boost

Lemma 3.6. *Let $x \in \mathbb{R}^n$, $D \ge k$, $C \ge 1$. Suppose $h : [n] \to [D]$ is drawn from a fully independent family of hash functions and $y^{(1)}$, $y^{(2)}$ and $y^{(3)}$ are independent (D, h)-gaussian hash projections of x. Then, if $\mathcal{A}$ is an algorithm that achieves (k, C) sparse recovery with probability $\ge 8/9$, and $U^{(t)} = \operatorname{supp}(\mathcal{A}(y^{(t)}))$ for $t \in \{1, 2, 3\}$,*

$$\mathbb{E}\left[\sum_{\substack{j \in [D] \\ j \notin U^{(1)} \cup U^{(2)} \cup U^{(3)}}} \left\| x_{h^{-1}(j)} \right\|_2^2 \,\middle|\, \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3 \right] \le 9C \operatorname{Err}^2(x, k)$$

```
procedure $R$-Round-K-SparseRec($x$, $k$, $C$, $R$ )
    $S_0 = [n]$
    $C_0 = C/8$
    for $r \leftarrow 1, \ldots, R-1$ do
        $k_r \leftarrow k5^{r-1}$
        $D_r \leftarrow k_r C_0^{5(\log_{C_0}(n))^{r/R}}$
        $C_r \leftarrow C_0^{(\log_{C_0}(n))^{(r-1)/R}}$
        $\delta_r \leftarrow 2^{-(r+3)}$
        $S_r \leftarrow$ 1-RoundSNRBoost($x_{S_{r-1}}, |S_{r-1}|, D_r, C_r, k_r, \delta_r$)
    end for
    return $\hat{x} \leftarrow$ HighSNR-Recover($x_{S_{R-1}}, 5k_{R-1}, C_0^{(\log_{C_0}(n))^{(R-1)/R}}, 2^{-(R+3)}$)
end procedure
```

Algorithm 3.2: R-Round-k-Sparse Recovery

where $\mathcal{E}^{(t)}$ represents the event that $\mathcal{A}(y^{(t)})$ successfully performs (k, C)-sparse recovery.

Proof. Let y be a (D, h)-gaussian hash projection of x. From the definition of $H_k(y)$, we know that for all S such that $|S| \leq k$, $\mathrm{Err}^2(y, k) = \sum_{j \in \overline{H_k(y)}} y_j^2 \leq \sum_{j \in \overline{S}} y_j^2$. If we choose $S = h(H_k(x))$, we get $\mathrm{Err}^2(y, k) \leq \sum_{j \in \overline{h(H_k(x))}} y_j^2$. Furthermore,

$$\begin{aligned}
\mathop{\mathbb{E}}_g[\mathrm{Err}^2(y,k)] &\leq \mathop{\mathbb{E}}_g[\sum_{j \in \overline{h(H_k(x))}} y_j^2] \\
&= \mathop{\mathbb{E}}_g[\sum_{j \in \overline{h(H_k(x))}} \Big(\sum_{i \in h^{-1}(j)} x_i \cdot g_i\Big)^2] \\
&= \sum_{j \in \overline{h(H_k(x))}} \sum_{i \in h^{-1}(j)} x_i^2 \\
&\leq \sum_{i \in \overline{H_k(x)}} x_i^2 = \mathrm{Err}^2(x, k)
\end{aligned}$$

where the second equality follows because $g_i \sim \mathcal{N}(0, 1)$ for all $i \in [n]$.

Let E_j be the indicator random variable for the event that $j \notin U$ where $U = \mathrm{supp}(\mathcal{A}(y))$. For a successful run of $\mathcal{A}$, the ℓ_2 mass of the unrecovered indices is bounded by:

$$\sum_{j \in [D]} E_j y_j^2 \leq C \,\mathrm{Err}^2(y, k)$$

Let $\mathcal{E}$ be the event that $\mathcal{A}(y)$ satisfies the (k, C)-sparse recovery guarantee for y. Then, if $\mathbb{I}(\mathcal{E})$ is the indicator random variable for the event $\mathcal{E}$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{g,\mathcal{A}}[\sum_{j \in [D]} E_j y_j^2 \mid \mathcal{E}] &\leq \mathop{\mathbb{E}}_{g,\mathcal{A}}[(\sum_{j \in [D]} E_j y_j^2)\mathbb{I}(\mathcal{E})] / \Pr_{g,\mathcal{A}}[\mathcal{E}] \\
&\leq \frac{9C}{8} \mathop{\mathbb{E}}_g[\mathrm{Err}^2(y,k)] \\
&\leq \frac{9C}{8} \mathrm{Err}^2(x, k) \qquad (11)
\end{aligned}$$

Let $q_j = \mathbb{E}_{g,\mathcal{A}}\left[E_j \mid \mathcal{E}\right]$ denote the probability $\big($over (D,h) projections and $\mathcal{A}\big)$ that $j \notin \operatorname{supp}(\mathcal{A}(y))$. Then for $j \in [D]$ and any $\theta > 0$,

$$\mathop{\mathbb{E}}_{g,\mathcal{A}}[E_j y_j^2 | \mathcal{E}] \geq \Pr\left[\Big(j \notin U\Big) \wedge \Big(|y_j| \geq (q_j/2)\theta\Big) \Bigm| \mathcal{E}\right] \cdot \theta^2$$

Observe that:

$$\Pr\left[\Big(j \notin U\Big) \wedge \Big(|y_j| \geq \theta\Big) \Bigm| \mathcal{E}\right] \geq 1 - \Pr\left[j \in U \Bigm| \mathcal{E}\right] - \Pr\left[|y_j| < \theta \Bigm| \mathcal{E}\right]$$

Since $y_j \sim \mathcal{N}(0, \theta^2)$ we may use the gaussian anti-concentration inequality i.e. $\Pr[|X| \leq \delta\theta] \leq \delta$ to get:

$$\Pr\left[\Big(j \notin U\Big) \wedge \Big(|y_j| \geq \theta\Big) \Bigm| \mathcal{E}\right] \geq 1 - (1 - q_j) - \frac{\theta}{\left\|x_{h^{-1}(j)}\right\|_2}$$

Setting $\theta = \frac{q_j}{2}\left\|x_{h^{-1}(j)}\right\|_2$:

$$\Pr\left[\Big(j \notin U\Big) \wedge \Big(|y_j| \geq \frac{q_j}{2}\left\|x_{h^{-1}(j)}\right\|_2\Big) \Bigm| \mathcal{E}\right] \geq q_j/2$$

and for all $j \in [D]$,

$$\mathop{\mathbb{E}}_{g,A}[E_j y_j^2 \mid \mathcal{E}] \geq \frac{q_j^3}{8}\left\|x_{h^{(-1)}(j)}\right\|_2^2. \tag{12}$$

Now, consider the $U^{(t)} = \operatorname{supp}(\mathcal{A}(y^{(t)}, k, C))$ for $t = 1, 2, 3$ where $y^{(1)}, y^{(2)}, y^{(3)}$ are independent (D,h) gaussian projections of x. Then,

$$\begin{aligned}
\mathbb{E}\Bigg[\sum_{\substack{j\in[D]:\\ j\notin U^{(1)}\cup U^{(2)}\cup U^{(3)}}} \left\|x_{h^{-1}(j)}\right\|_2^2 \Bigg| \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3\Bigg] &= \sum_{j\in[D]} \left\|x_{h^{-1}(j)}\right\|_2^2 \cdot \mathbb{E}[E_j^{(1)} \cdot E_j^{(2)} \cdot E_j^{(3)} \mid \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3] \\
&= \sum_{j\in[D]} \left\|x_{h^{-1}(j)}\right\|_2^2 \cdot \mathbb{E}[E_j^{(1)} \mid \mathcal{E}_1] \cdot \mathbb{E}[E_j^{(2)} \mid \mathcal{E}_2] \cdot \mathbb{E}[E_j^{(3)} \mid \mathcal{E}_3] \\
&= \sum_{j\in[D]} \left\|x_{h^{-1}(j)}\right\|_2^2 \cdot q_j^3
\end{aligned}$$

where the expectation is taken over $g^{(1)}, g^{(2)}, g^{(3)}, \mathcal{A}(y^{(1)}), \mathcal{A}(y^{(2)}), \mathcal{A}(y^{(3)})$. The second equality follows from the independence of $y^{(1)}, y^{(2)}, y^{(3)}$. So, using (11) and (12),

$$\begin{aligned}
\mathbb{E}\Bigg[\sum_{\substack{j\in[D]:\\ j\notin U^{(1)}\cup U^{(2)}\cup U^{(3)}}} \left\|x_{h^{-1}(j)}\right\|_2^2 \mid \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3\Bigg] &\leq \sum_{j\in[D]} \left\|x_{h^{-1}(j)}\right\|_2^2 \cdot q_j^3 \\
&\leq 8 \sum_{j\in[D]} \mathop{\mathbb{E}}_{g,\mathcal{A}}[E_j y_j^2 \mathbb{I}(\mathcal{E})] \\
&\leq 9C \operatorname{Err}^2(x, k).
\end{aligned}$$

□

Lemma 3.7. *Let $z \in \mathbb{R}^n$ and $h : [n] \to [D]$ be randomly chosen from a fully independent family of hash functions where $D \le n$. Then, with probability $1 - 2\delta$,*

$$\max_{l \in [D]} \Big[\sum_{i \in h^{-1}(l)} z_i^2 \Big] \le 4\Big(\frac{\|z\|_2^2}{D} + 5 \|z\|_\infty^2 \log\big(\frac{D \cdot \log(n/\delta)}{\delta}\big)\Big)$$

Proof. Let $\beta_j = \|z\|_\infty^2 \cdot 2^{-j}$ for all $j \in \mathbb{Z}$ and let $t = O(\log(n/\delta))$. Partition $[n]$ into $t+2$ sets: $R_j = \{i \in [n] \mid \beta_{j+1} \le z_i^2 \le \beta_j\}$ for all $0 \le j \le t$ and $R_{t+1} = \{i \in [n] \mid z_i^2 \le \beta_{t+1}\}$. Then, for a fixed R_j and $l \in [D]$ we may apply the Bernstein bounds(Theorem B.1) to get:

$$\Pr\Big[\big|R_j \cap h^{-1}(l)\big| \ge \frac{|R_j|}{D} + 4\log(1/\delta) + 4\sqrt{\frac{\log(1/\delta)\,|R_j|}{D}} \Big] \le \delta$$

Taking a union bound over all $R_0, \ldots, R_t$ and all $l \in [D]$:

$$\Pr\Big[\exists j \in [t], l \in [D] \;\Big|\; \big|R_j \cap h^{-1}(l)\big| \ge \frac{|R_j|}{D} + 4\log(\frac{D \cdot t}{\delta}) + 4\sqrt{\frac{\log(\frac{D \cdot t}{\delta})\,|R_j|}{D}} \Big] \le \delta$$

The ℓ_2 mass from $R_0, \ldots, R_t$ falling into any $j \in [D]$ is bounded by:

$$\begin{aligned}
\sum_{j=0}^{t} \beta_j \Big(\frac{|R_j|}{D} + 4\log(\frac{D \cdot t}{\delta}) + 4\sqrt{\frac{\log(\frac{D \cdot t}{\delta})\,|R_j|}{D}} \Big) &\le 2 \sum_{j=0}^{t} \beta_j \big(\frac{|R_j|}{D} + 4\log(\frac{D \cdot t}{\delta}) \big) \\
&\le 4\big(\frac{\|z\|_2^2}{D} + 4\log(\frac{D \cdot t}{\delta})\beta_0 \big) \\
&= 4\big(\frac{\|z\|_2^2}{D} + 4\log(\frac{D \cdot t}{\delta}) \|z\|_\infty^2 \big)
\end{aligned}$$

where the second inequality follows because $\sum_{j=0}^{t} |R_j| \beta_j \le 2 \sum_{i \in [n]} z_i^2 \le 2 \|z\|_2^2$.

Next, we bound contribution of R_{t+1} to the ℓ_2 mass hashed to each location. The total ℓ_2 mass in R_{t+1} is $\|z_{R_{t+1}}\|_2^2 \le \beta_{t+1} \cdot n$. So, the expected amount of ℓ_2 mass in a given location $l \in [D]$ is $\le n\beta_{t+1}/D$. Using Markov's inequality, with probability $1 - \delta$, we know that the ℓ_2 mass from R_{t+1} hashed to each location in $[D]$ is $\le n \cdot \|z\|_\infty^2 \cdot 2^{-(t+1)}/\delta \le \|z\|_\infty^2$. So,

$$\max_{l \in [D]} \Big[\sum_{i \in h^{-1}(l)} z_i^2 \Big] \le 4\Big(\frac{\|z\|_2^2}{D} + 5 \|z\|_\infty^2 \log\big(\frac{D \cdot \log(n/\delta)}{\delta}\big)\Big)$$

□

Lemma 3.8. *Let $z \in \mathbb{R}^n$, $k \le D \le n$ and $h : [n] \to [D]$ be randomly chosen from a fully independent family of hash functions. Then, with probability $1 - \delta$, for all $U \subseteq [D]$:*

$$\mathrm{Err}^2(z_{h^{-1}(U)}, |U| + k) \le \|z\|_2^2 \, \frac{|U|\, O(\log(n/\delta))}{\sqrt{kD\delta}}$$

Proof. Consider all indices in the set $J = \{i \in [n] \mid z_i^2 \ge \|z\|_2^2 / L\}$ where $L = \sqrt{kD\delta}$. Observe that the expected number of collisions among these elements under the hash function h is $\le \binom{L}{2}/D \le k\delta/2$. By Markov's inequality, the number of collisions is at most k with probability $1 - (\delta/2)$. So, with probability $1 - \delta/2$:

$$\forall U \subset [D], \big|J \cap h^{-1}(U)\big| \le |U| + k \tag{13}$$

Suppose, we restricted ourselves only to the indices in the set $\overline{J}$. Observe that $\left\|z_{\overline{J}}\right\|_2 \leq \|z\|_2$ and $\left\|z_{\overline{J}}\right\|_\infty^2 \leq \|z\|_2^2/L$. Applying Lemma 3.7, with probability $1-\delta/2$:

$$\max_{l\in[D]}\Bigg[\sum_{i\in\overline{J}:h(i)=l} z_i^2\Bigg] \leq 4\Bigg(\frac{\|z\|_2^2}{D} + \frac{5\,\|z\|_2^2}{L}\log\big(\frac{4D\log(4n/\delta)}{\delta}\big)\Bigg) = O\Bigg(\,\|z\|_2^2\,\frac{O(\log(n/\delta))}{L}\Bigg) \tag{14}$$

So, with probability $1-\delta$, both (13) and (14) hold. Hence,

$$\begin{aligned}\mathrm{Err}^2(z_{h^{-1}(U)}, |U|+k) &\leq |U|\cdot\big(\,\|z\|_2^2\,\frac{O(\log(n/\delta))}{L}\big)\\ &\leq \|z\|_2^2\,\frac{|U|\,O(\log(n/\delta))}{\sqrt{kD\delta}}\end{aligned}$$

□

Theorem 3.9. *Suppose there exists an algorithm that takes $O(k\log_C(n/k)\log(1/\delta)\cdot g(k))$ linear measurements of its input where $g(k)$ is a non-decreasing function in k and outputs a k sparse vector that achieves (k,C) sparse recovery with probability $(1-\delta)$. Then, for $R\leq\log\log(n/k)/2\log\log\log(n)$ and $C>16$, Algorithm 3.2 takes $O(k5^R(\log_C(n/k))^{1/R}\cdot g(5^Rk))$ linear measurements of $x\in\mathbb{R}^n$ over R adaptive rounds and outputs a vector that achieves (k,C) sparse recovery of x with probability $\geq\frac{3}{4}$.*

Proof. In this proof, we will achieve $(k,16C)$ sparse recovery for all $C>1$. We may rescale C to get the theorem statement. We define

$$\begin{aligned}\delta_r &= 2^{-(r+3)}\\ k_r &= k5^{r-1}\\ D_r &= k_rC^{5(\log_C(n/k))^{r/R}}\\ C_r &= C^{(\log_C(n/k))^{(r-1)/R}}\end{aligned}$$

for $r>0$ and $S_0=[n]$

In each round $r\in\{1,\ldots,R-1\}$, we use Algorithm 3.1 with these parameters to get a subset $S_r\subseteq S_{r-1}$. We sample a random hash function $h: S_{r-1}\to[D_r]$ and generate 3 independent (D_r,h)-gaussian hash projections $y^{(1)},y^{(2)},y^{(3)}$ of $x_{S_{r-1}}$ and perform HighSNR-Recover on each of them with parameters $(k_r,C_r,\delta_r/3)$. Let $U^{(1)},U^{(2)},U^{(3)}$ be supports of the recovered vectors. Since HighSNR-Recover generates k_r sparse output, $\left|U^{(1)}\right|,\left|U^{(2)}\right|,\left|U^{(3)}\right|\leq k_r$. Let $U_r=U^{(1)}\cup U^{(2)}\cup U^{(3)}$, and set $S_r=h^{-1}(U_r)\subseteq S_{r-1}$ to be the set of indices carried into the next round. So, if we set $z=x_{S_{r-1}\cap\overline{H_{k_{r-1}}(x_{S_{r-1}})}}$ and let $U=U_r$ in Lemma 3.8:

$$\begin{aligned}\mathrm{Err}^2(z_{h^{-1}(U_r)},|U_r|+k_{r-1}) &\leq \frac{\|z\|_2^2}{\sqrt{D_r\delta_r/k_rO(\log(n/\delta_r))}}\\ &\leq \frac{\|z\|_2^2}{2^{2(\log(n))^{r/R}}}\end{aligned}$$

where the second inequality follows because $\log(n)=o(C^{2(\log_C(n))^{1/R}})$ when $2^r\leq C^{2(\log_C(n))^{r/R}}$ and $R\leq\frac{\log\log(n)}{2\log\log\log(n)}$. Since $z=x_{S_{r-1}\cap\overline{H_{k_{r-1}}(x_{S_{r-1}})}}$, we have both $\|z\|_2^2=\left\|x_{S_{r-1}\cap\overline{H_{k_{r-1}}(x_{S_{r-1}})}}\right\|_2^2=$

$\mathrm{Err}^2(x_{S_{r-1},k_{r-1}})$ and $\mathrm{Err}^2(z_{h^{-1}(U)}, |U| + k_r) \geq \mathrm{Err}^2(x_{h^{-1}(U)}, |U| + k_{r-1} + k_{r-1})$. Since $|U| \leq 3k_{r-1}$ and $5k_{r-1} = k_r$, we conclude:

$$\mathrm{Err}^2(x_{S_r}, k_r) \leq \frac{\mathrm{Err}^2(x_{S_{r-1}}, k_{r-1})}{C^{2(\log_C(n))^{r/R}}}$$

If we successively apply Theorem 3.6 under the above parameters for rounds $1, \ldots, R-1$, then for any $r \in \{1, \ldots, R-1\}$

$$\begin{aligned}\mathbb{E}[\left\|x_{S_r} - x_{S_{r+1}}\right\|_2^2] &\leq C_r \,\mathrm{Err}^2(x_{S_r}, k_r)\\ &\leq \frac{C_r}{C^{2(\log_C(n))^{r/R}}} \,\mathrm{Err}^2(x_{S_{r-1}}, k_{r-1})\end{aligned}$$

Since $\mathrm{Err}^2(x_{S_{r-1}}, k_{r-1}) \leq \mathrm{Err}^2(x, k)$ and we have set $C_r = C^{(\log_C(n/k))^{(r-1)/R}}$,

$$\mathbb{E}[\left\|x_{S_r} - x_{S_{r+1}}\right\|_2^2] \leq \frac{1}{C^{(\log_C(n))^{r/R}}} \,\mathrm{Err}^2(x, k)$$

In the final round, we run HighSNR-Recover$(x_{S_{R-1}}, k_R, C_R)$ and find $\hat{x}$ such that $\left\|x_{S_{R-1}} - \hat{x}\right\|_2^2 \leq C_R \,\mathrm{Err}^2(x_{S_{R-1}}, k_R)$. So,

$$\begin{aligned}\mathbb{E}\left[\,\|x - \hat{x}\|_2^2\,\right] &\leq \sum_{r=1}^{R-1} \mathbb{E}\left[\,\left\|x_{S_{r-1}} - x_{S_r}\right\|_2^2\right] + \mathbb{E}\left[\,\left\|x_{S_{R-1}} - \hat{x}\right\|_2^2\right]\\ &\leq \sum_{r=1}^{R} C_r \,\mathrm{Err}^2(x_{S_r}, k_r)\\ &\leq C \,\mathrm{Err}^2(x_{S_1}, k_1) + \sum_{r=2}^{R} \frac{1}{C^{(\log_C(n))^{(r-1)/R}}} \,\mathrm{Err}^2(x, k)\\ &\leq 2C \,\mathrm{Err}^2(x, k)\end{aligned}$$

So, with probability $\geq 7/8$, after R rounds $\|x - \hat{x}\|_2^2 \leq 16C \,\mathrm{Err}^2(x, k)$. In each round, we use independently call HighSNR-Recover$(x_{S_{r-1}}, k_r, C_r)$ thrice with failure probability $\delta_r/3 = 2^{-(r+3)}/3$ and condition on them being successful. So, over R rounds all calls to HighSNR-Recover are successful with probability $\geq 1 - \sum_{r=1}^{R} \delta_r = 1 - \sum_{r=1}^{R} 2^{-(r+3)} = 7/8$.

The total number of measurements over R rounds is bounded by:

$$\begin{aligned}\sum_{r=1}^{R} 3k_r \log(3/\delta_r) \cdot g(5^r k) \cdot (\log_{C_{r-1}}(D_r/k)) &= \sum_{r=1}^{R} 3k_r \log(3/\delta_r) \cdot g(5^r k) \cdot (\log_C(n/k))^{1/R}\\ &\leq \sum_{r=1}^{R} 3k \cdot 5^r \cdot 2r \cdot g(5^r k)(\log_C(n/k))^{1/R}\\ &= O(5^R k(\log_C(n/k))^{1/R} \cdot g(5^R k))\end{aligned}$$

So, the output of Algorithm 3.2 achieves $(k, 16C)$ sparse recovery in R rounds with probability $\geq 3/4$ and uses $O(5^R k(\log_C(n/k))^{1/R} \cdot g(5^R k))$ measurements. If we rescale C by a factor of 16, we get the desired guarantee. □

As a consequence of Theorem 3.9 and Theorem 3.5, we get the following guarantee on our algorithm:

Corollary 3.10. *For $R \leq \frac{\log\log(n/k)}{\log\log\log(n)}$ and $C > 16$, Algorithm 3.2 takes $O(k5^R(\log_C(n/k))^{1/R} \cdot \log^*(5^R k))$ linear measurements of $x \in \mathbb{R}^n$ over R adaptive rounds and outputs a vector that achieves (k, C) sparse recovery of x with probability $\geq \frac{3}{4}$.*

References

[ACD13] Ery Arias-Castro, Emmanuel J. Candès, and Mark A. Davenport. On the fundamental limits of adaptive sensing. *IEEE Trans. Information Theory*, 59(1):472–481, 2013.

[BJKS04] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.*, 68(4):702–732, 2004.

[CCF02] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. *ICALP*, 2002.

[CHNR08] Rui M. Castro, Jarvis D. Haupt, Robert D. Nowak, and Gil M. Raz. Finding needles in noisy haystacks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, March 30 - April 4, 2008, Caesars Palace, Las Vegas, Nevada, USA*, pages 5133–5136, 2008.

[CM04] G. Cormode and S. Muthukrishnan. Improved data stream summaries: The count-min sketch and its applications. *LATIN*, 2004.

[CM06] G. Cormode and S. Muthukrishnan. Combinatorial algorithms for compressed sensing. *SIROCCO*, 2006.

[CRT06] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52:489–509, 2006.

[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.

[DIPW10] K. Do Ba, P. Indyk, E. Price, and D. Woodruff. Lower bounds for sparse recovery. *SODA*, 2010.

[ECG$^+$09] Yaniv Erlich, Kenneth Chang, Assaf Gordon, Roy Ronen, Oron Navon, Michelle Rooks, and Gregory J Hannon. Dna sudoku—harnessing high-throughput sequencing for multiplexed spec imen analysis. *Genome research*, 19(7):1243–1253, 2009.

[GLPS10] Anna C. Gilbert, Yi Li, Ely Porat, and Martin J. Strauss. Approximate sparse recovery: optimizing time and measurements. In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 475–484, 2010.

[HBCN12] Jarvis D. Haupt, Richard G. Baraniuk, Rui M. Castro, and Robert D. Nowak. Sequentially designed compressed sensing. In *IEEE Statistical Signal Processing Workshop, SSP 2012, Ann Arbor, MI, USA, August 5-8, 2012*, pages 401–404, 2012.

[HCN11] Jarvis D. Haupt, Rui M. Castro, and Robert D. Nowak. Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Trans. Information Theory*, 57(9):6222–6235, 2011.

[IPW11] Piotr Indyk, Eric Price, and David P. Woodruff. On the power of adaptivity in sparse recovery. In *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, pages 285–294, 2011.

[JXC08] Shihao Ji, Ya Xue, and Lawrence Carin. Bayesian compressive sensing. *IEEE Trans. Signal Processing*, 56(6):2346–2356, 2008.

[LDSP08] M. Lustig, D.L. Donoho, J.M. Santos, and J.M. Pauly. Compressed sensing mri. *Signal Processing Magazine, IEEE*, 25(2):72–82, 2008.

[LNW18] Yi Li, Vasileios Nakos, and David P. Woodruff. On low-risk heavy hitters and sparse recovery schemes. *APPROX*, 2018.

[MSW08] Dmitry M. Malioutov, Sujay Sanghavi, and Alan S. Willsky. Compressed sensing with sequential observations. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, March 30 - April 4, 2008, Caesars Palace, Las Vegas, Nevada, USA*, pages 3357–3360, 2008.

[NSWZ18] Vasileios Nakos, Xiaofei Shi, David P. Woodruff, and Hongyang Zhang. Improved algorithms for adaptive compressed sensing. In *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, July 9-13, 2018, Prague, Czech Republic*, pages 90:1–90:14, 2018.

[PW11] Eric Price and David P. Woodruff. (1 + eps)-approximate sparse recovery. In *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, pages 295–304, 2011.

[PW12] Eric Price and David P. Woodruff. Applications of the shannon-hartley theorem to data streams and sparse recovery. In *Proceedings of the 2012 IEEE International Symposium on Information Theory, ISIT 2012, Cambridge, MA, USA, July 1-6, 2012*, pages 2446–2450, 2012.

[PW13] Eric Price and David P. Woodruff. Lower bounds for adaptive sparse recovery. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 652–663, 2013.

A Appendix for Lower Bound

Theorem A.1 (Shannon-Hartley). *Let S be a random variable such that $\mathbb{E}[S^2] = \tau^2$. Consider the random variable $S + T$, where $T \sim \mathcal{N}(0, \sigma^2)$. Then*

$$I(S; S+T) \le \frac{1}{2} \lg \big(1 + \frac{\tau^2}{\sigma^2}\big).$$

Lemma A.2. *Consider a random variable $X \in [n]$ with probability distribution $p(l) = \Pr[X = l]$. Suppose $b = \lg(n) - H(X)$. Let $T_i = \{j \mid 2^i \le np(j) \le 2^{i+1}\}$ and $T_0 = \{j \mid np(j) \le 2\}$ and let $q_i = \sum_{j \in T_i} p(j)$. Then,* *(a)* $\sum_{i=0}^{\infty} iq_i \le b+1$

(b) $\sum_{i=0}^{\infty} q_i \lg(1+\frac{1}{q_i}) \le O(b+1)$

(c) if J *is the random variable that denotes the index of the partition containing* X*, then* $H(J) < O(b+1)$.

Proof.

$$\begin{aligned}\sum_{i=0}^{\infty} iq_i &= \sum_{i>0}\sum_{j\in T_i} \Pr[X=j]\cdot i \\ &\le \sum_{i>0}\sum_{j\in T_i} \Pr[X=j]\lg(n\Pr[X=j]) \\ &= b - \sum_{j\in T_0} \Pr[X=j]\lg(n\Pr[X=j]) \\ &= b - q_0\lg(nq_0/\left|T_0\right|) \\ &\le b + \left|T_0\right|/ne\end{aligned}$$

using convexity and minimizing $x\lg(ax)$ at $x = 1/ae$. Hence,

$$\sum_{i=0}^{\infty} iq_i \le b+1 \tag{15}$$

Next, consider $\sum_{i=0}^{\infty} q_i \lg(1+\frac{1}{q_i})$. When $q_i \le 1/2$, we have $\lg(1+\frac{1}{q_i}) \le 2\lg(\frac{1}{q_i})$. So,

$$\sum_{i=0}^{\infty} q_i \lg(1+\frac{1}{q_i}) \le 2\Big(\sum_{i|q_i\le 1/2} t_i\lg(1/t_i) + \sum_{i|q_i>1/2} 1\Big) \le 2\Big(H(J)+1\Big) \tag{16}$$

Now, in order to bound the entropy term, consider the partition $T_+ = \{i \mid q_i > 1/2^i\}$ and $T_- = \{i \mid q_i \le 1/2^i\}$. Then

$$\begin{aligned}H(J) &= \sum_i q_i\lg(\frac{1}{q_i}) \\ &\le \sum_{i\in T_+} iq_i + \sum_{i\in T_-} q_i\lg(\frac{1}{q_i}) \\ &\le b+1+\sum_{i\in T_-} q_i\lg(\frac{1}{q_i})\end{aligned}$$

Observe that $x\log(1/x)$ increases on $[0,1/e]$, so

$$\sum_{i\in T_-} q_i\lg(\frac{1}{q_i}) \le q_0\log(\frac{1}{q_0}) + q_1\lg(\frac{1}{q_1}) + \sum_{i\ge 2}\frac{1}{2^i}\lg(1/2^i) \le 2/e+3/2 < 3$$

Hence $H(J) < b+4$. So, in (16),

$$\sum_{i=0}^{\infty} q_i \lg(1+\frac{1}{q_i}) \le 2(b+5) \tag{17}$$

□

Claim A.3. *Let the sequence $B_1 \leq B_2 \leq B_3 \ldots,$, satisfy $B_1 \geq k\log(k)$,$B_1 \leq \max m_1, k\log(k)$ and for all $r \geq 1$,*

$$B_{r+1} \leq \big(c_5 + \frac{c_3 m_{r+1}}{\alpha k}\big) B_r + m_{r+1}\log(k) + \frac{c_4 m_{r+1}}{\alpha} + c_2 k$$

for constants $c_2, c_3, c_4, c_5 > 1$. Then, for all $r \geq 1$,

$$B_r \leq \left(\prod_{j=2}^{r+1} \big(2c_5 + \frac{2c_6 m_j}{k\alpha}\big)\right) \max\{k\log(k), m_1\}$$

where c_6 is a constant.

Proof. The base case holds because :

$$B_1 = \max m_1, k\log(k)$$

Now, assume that the claim holds for r, then:

$$\begin{aligned} B_{r+1} &\leq B_r\left(c_5 + \frac{c_3 m_{r+1}}{\alpha k}\right) + m_{r+1}\log(k) + \frac{c_4 m_{r+1}}{\alpha} + c_2 k \\ &= B_r\left(c_5 + \frac{c_3 \cdot m_{r+1}}{\alpha k}\right) + \frac{m_{r+1}}{k}(k\log(k)) + \frac{c_4 m_{r+1}}{\alpha} + c_2 k \\ &\leq 2B_r\left(c_5 + \frac{c_6 \cdot m_{r+1}}{\alpha k}\right) \\ &\leq \left(\prod_{j=2}^{r+1} \big(2c_5 + \frac{2c_6 m_j}{k\alpha}\big)\right) \max\{k\log(k), m_1\} \end{aligned}$$

where the third line follows because $B_r \geq B_1 \geq k\log(k)$ and $B_r \geq B_1 \geq m_1$ and $c_6 = \max(c_3, c_4{+}1)$ is a constant. □

B Appendix for Upper Bound

The following form of Bernstein's inequality is well known:

Theorem B.1 (Bernstein)**.** *Let $X_1, \ldots, X_n$ be i.i.d Bernoulli random variables with parameter p and $X = \sum_{i=1}^n X_i$. Then,*

$$\Pr[X \geq np + 4\log(1/\delta) + 4\sqrt{np\log(1/\delta)}] \leq \delta.$$