Improved Algorithms for Adaptive Compressed Sensing

Vasileios Nakos
Harvard University, Cambridge, USA
vasileiosnakos@g.harvard.edu

Xiaofei Shi
Carnegie Mellon University, Pittsburgh, USA
xiaofeis@andrew.cmu.edu

David P. Woodruff
Carnegie Mellon University, Pittsburgh, USA
dwoodruf@cs.cmu.edu

Hongyang Zhang
Carnegie Mellon University, Pittsburgh, USA
hongyanz@cs.cmu.edu

Abstract

In the problem of adaptive compressed sensing, one wants to estimate an approximately k-sparse vector $x \in \mathbb{R}^n$ from m linear measurements $A_1x, A_2x, \ldots, A_mx$, where A_i can be chosen based on the outcomes $A_1x, \ldots, A_{i-1}x$ of previous measurements. The goal is to output a vector $\hat{x}$ for which

$$\|x - \hat{x}\|_p \leq C \cdot \min_{k\text{-sparse } x'} \|x - x'\|_q,$$

with probability at least $2/3$, where $C > 0$ is an approximation factor. Indyk, Price and Woodruff (FOCS'11) gave an algorithm for $p = q = 2$ for $C = 1+\epsilon$ with $\mathcal{O}((k/\epsilon)\text{loglog}(n/k))$ measurements and $\mathcal{O}(\log^*(k)\text{loglog}(n))$ rounds of adaptivity. We first improve their bounds, obtaining a scheme with $\mathcal{O}(k \cdot \text{loglog}(n/k) + (k/\epsilon) \cdot \text{loglog}(1/\epsilon))$ measurements and $\mathcal{O}(\log^*(k)\text{loglog}(n))$ rounds, as well as a scheme with $\mathcal{O}((k/\epsilon) \cdot \text{loglog}(n \log(n/k)))$ measurements and an optimal $\mathcal{O}(\text{loglog}(n))$ rounds. We then provide novel adaptive compressed sensing schemes with improved bounds for (p, p) for every $0 < p < 2$. We show that the improvement from $O(k \log(n/k))$ measurements to $O(k \log\log(n/k))$ measurements in the adaptive setting can persist with a better ϵ-dependence for other values of p and q. For example, when $(p, q) = (1, 1)$, we obtain $O(\frac{k}{\sqrt{\epsilon}} \cdot \log\log n \log^3(\frac{1}{\epsilon}))$ measurements. We obtain nearly matching lower bounds, showing our algorithms are close to optimal. Along the way, we also obtain the first nearly-optimal bounds for (p, p) schemes for every $0 < p < 2$ even in the non-adaptive setting.

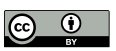
2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms

Keywords and phrases Compressed Sensing, Adaptivity, High-Dimensional Vectors

Digital Object Identifier 10.4230/LIPIcs.ICALP.2018.90

Related Version A full version of the paper is available at `https://arxiv.org/pdf/1804.09673.pdf`.

Funding This work was partially supported by NSF grant IIS-144741.



45th International Colloquium on Automata, Languages, and Programming (ICALP 2018).
Editors: Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella;
Article No. 90; pp. 90:1–90:14

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Compressed sensing, also known as sparse recovery, is a central object of study in data stream algorithms, with applications to monitoring network traffic [7], analysis of genetic data [19, 12], and many other domains [16]. The problem can be stated as recovering an underlying signal $x \in \mathbb{R}^n$ from *measurements* $A_1 \cdot x, ..., A_m \cdot x$ with the C-approximate ℓ_p/ℓ_q recovery guarantee being

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

where the A_i are drawn from a distribution and $m \ll n$. The focus of this work is on *adaptive compressed sensing*, in which the measurements are chosen in rounds, and the choice of measurement in each round depends on the outcome of the measurements in previous rounds.

Adaptive compressed sensing has been studied in a number of different works [11, 4, 8, 9, 14, 1, 10, 18] in theoretical computer science, machine learning, image processing, and many other domains [10, 18, 2]. In theoretical computer science and machine learning, adaptive compressed sensing serves as an important tool to obtain sublinear algorithms for active learning in both time and space [10, 5, 18, 2]. In image processing, the study of adaptive compressed sensing has led to compressed acquisition of sequential images with various applications in celestial navigation and attitude determination [6].

Despite a large amount of works on adaptive compressed sensing, the power of adaptivity remains a long-standing open problem. Indyk, Price, and Woodruff [10] were the first to show that without any assumptions on the signal x, one can obtain a number m of measurements which is a $\log(n)/\log\log(n)$ factor smaller than what can be achieved in the non-adaptive setting. Specifically, for $p = q = 2$ and $C = 1 + \epsilon$, they show that $m = \mathcal{O}(\frac{k}{\epsilon}\log\log(n))$ measurements suffice to achieve guarantee (1), whereas it is known that any non-adaptive scheme requires $k = \Omega(\frac{k}{\epsilon}\log(\frac{n}{k}))$ measurements, provided $\epsilon > \sqrt{\frac{k\log n}{n}}$ (Theorem 4.4 of [17], see also [3]). Improving the sample complexity as much as possible is desired, as it might correspond to, e.g., the amount of radiation a hospital patient is exposed to, or the amont of time a patient must be present for diagnosis.

The ℓ_1/ℓ_1 problem was studied in [17], for which perhaps surprisingly, a better dependence on ϵ was obtained than is possible for ℓ_2/ℓ_2 schemes. Still, the power of adaptivity for the ℓ_1/ℓ_1 recovery problem over its non-adaptive counterpart has remained unclear. An $O(\frac{k}{\sqrt{\epsilon}}\log n\log^3(\frac{1}{\epsilon}))$ non-adaptive bound was shown in [17], while an adaptive lower bound of $\Omega(\frac{k}{\sqrt{\epsilon}}/\log\frac{k}{\sqrt{\epsilon}})$ was shown in [18]. Recently several works [20, 15] have looked at other values of p and q, even those for which $0 < p, q < 1$, which do not correspond to normed spaces. The power of adaptivity for such error measures is also unknown.

1.1 Our Results

Our work studies the problem of adaptive compressed sensing by providing affirmative answers to the above-mentioned open questions. We improve over the best known results for $p = q = 2$, and then provide novel adaptive compressed sensing guarantees for $0 < p = q < 2$ for every p and q. See Table 1 for a comparison of results.

For ℓ_1/ℓ_1, we design an adaptive algorithm which requires only $\mathcal{O}(\frac{k}{\sqrt{\epsilon}}\text{loglog}(n)\log^{\frac{5}{2}}(\frac{1}{\epsilon}))$ measurements for the ℓ_1/ℓ_1 problem. More generally, we study the ℓ_p/ℓ_p problem for $0 < p < 2$. One of our main theorems is the following.

▶ **Theorem 1** (ℓ_p/ℓ_p Recovery Upper Bound). *Let $x \in \mathbb{R}^n$ and $0 < p < 2$. There exists a randomized algorithm that performs $\mathcal{O}(\frac{k}{\epsilon^{p/2}}\text{loglog}(n)\,\text{poly}(\log(\frac{1}{\epsilon})))$ adaptive linear measurements*

Table 1 The sample complexity of adaptive compressed sensing. Results without any citation given correspond to our new results.

C, Guarantees	Upper Bounds	Rounds	Lower Bounds
$1+\epsilon$, ℓ_1/ℓ_1	$\mathcal{O}(\frac{k}{\sqrt{\epsilon}}\text{loglog}(n)\log^{\frac{5}{2}}(\frac{1}{\epsilon}))$	$\mathcal{O}(\text{loglog}(n))$	$\Omega(\frac{k}{\sqrt{\epsilon}\log(k/\sqrt{\epsilon})})$ [18]
$1+\epsilon$, ℓ_p/ℓ_p	$\mathcal{O}(\frac{k}{\epsilon^{p/2}}\text{loglog}(n)\,\text{poly}(\log(\frac{1}{\epsilon})))$	$\mathcal{O}(\text{loglog}(n))$	$\Omega(\frac{k}{\epsilon^{p/2}}\frac{1}{\log^2(k/\epsilon)})$
$\sqrt{\frac{1}{k}}$, ℓ_∞/ℓ_2	$\mathcal{O}(k\text{loglog}(n)+k\log(k))$	$\mathcal{O}(\text{loglog}(n))$	-
$1+\epsilon, \ell_2/\ell_2$	$\mathcal{O}(\frac{k}{\epsilon}\text{loglog}(\frac{n\epsilon}{k}))$ [10] $\mathcal{O}(k\text{loglog}(\frac{n}{k})+\frac{k}{\epsilon}\text{loglog}(\frac{1}{\epsilon}))$ $\mathcal{O}(\frac{k}{\epsilon}\text{loglog}(\frac{n\log(n\epsilon)}{k}))$	$\mathcal{O}(\log^*(k)\text{loglog}(\frac{n\epsilon}{k}))$ [10] $\mathcal{O}(\log^*(k)\text{loglog}(\frac{n}{k}))$ $\mathcal{O}(\text{loglog}(n\log(\frac{n\epsilon}{k}))$	$\Omega(\frac{k}{\epsilon}+\text{loglog}(n))$ [18]

on x in $\mathcal{O}(\text{loglog}(n))$ rounds, and with probability $2/3$, returns a vector $\hat{x}\in\mathbb{R}^n$ such that $\|x-\hat{x}\|_p \le (1+\epsilon)\|x_{-k}\|_p$.

Theorem 1 improves the previous sample complexity upper bound for the case of $C=1+\epsilon$ and $p=q=1$ from $\mathcal{O}(\frac{k}{\sqrt{\epsilon}}\log(n)\log^3(\frac{1}{\epsilon}))$ to $\mathcal{O}(\frac{k}{\sqrt{\epsilon}}\text{loglog}(n)\log^{\frac{5}{2}}(\frac{1}{\epsilon}))$. Compared with the non-adaptive $(1+\epsilon)$-approximate ℓ_1/ℓ_1 upper bound of $\mathcal{O}(\frac{k}{\sqrt{\epsilon}}\log(n)\log^3(\frac{1}{\epsilon}))$, we show that adaptivity exponentially improves the sample complexity w.r.t. the dependence on n over non-adaptive algorithms while retaining the improved dependence on ϵ of non-adaptive algorithms. Furthermore, Theorem 1 extends the working range of adaptive compressed sensing from $p=1$ to general values of $p\in(0,2)$.

We also state a complementary lower bound to formalize the hardness of the above problem.

▶ **Theorem 2** (ℓ_p/ℓ_p Recovery Lower Bound). *Fix $0<p<2$, any $(1+\epsilon)$-approximate ℓ_p/ℓ_p recovery scheme with sufficiently small constant failure probability must make $\Omega(\frac{k}{\epsilon^{p/2}}/\log^2(\frac{k}{\epsilon}))$ measurements.*

Theorem 2 shows that our upper bound in Theorem 1 is tight up to the $\log(k/\epsilon)$ factor.

We also study the case when $p\neq q$. In particular, we focus on the case when $p=\infty, q=2$ and $C=\sqrt{\frac{1}{k}}$, as in the following theorem.

▶ **Theorem 3** (ℓ_∞/ℓ_2 Recovery Upper Bound). *Let $x\in\mathbb{R}^n$. There exists a randomized algorithm that performs $\mathcal{O}(k\log(k)+k\text{loglog}(n))$ linear measurements on x in $\mathcal{O}(\text{loglog}(n))$ rounds, and with probability $1-1/\text{poly}(k)$ returns a vector $\hat{x}$ such that $\|x-\hat{x}\|_\infty^2\le\frac{1}{k}\|x_{-k}\|_2^2$, where $x_{-k}\in\mathbb{R}^n$ is the vector with the largest $n-k$ coordinates (in the sense of absolute value) being zeroed out.*

We also provide an improved result for $(1+\epsilon)$-approximate ℓ_2/ℓ_2 problems.

▶ **Theorem 4** (ℓ_2/ℓ_2 Sparse Recovery Upper Bounds). *Let $x\in\mathbb{R}^n$. There exists a randomized algorithm that*

- *uses $\mathcal{O}(\frac{k}{\epsilon}\text{loglog}(\frac{1}{\epsilon})+k\text{loglog}(\frac{n}{k}))$ linear measurements on x in $\mathcal{O}(\text{loglog}(\frac{n}{k})\cdot\log^*(k))$ rounds;*
- *uses $\mathcal{O}(\frac{k}{\epsilon}\text{loglog}(\frac{n\log(n\epsilon)}{k}))$ linear measurements on x in $\mathcal{O}(\text{loglog}(\epsilon n\log(\frac{n}{k})))$ rounds;*

and with constant probability returns a vector $\hat{x}$ such that $\|x-\hat{x}\|_2\le(1+\epsilon)\|x_{-k}\|_2$.

Previously the best known tradeoff was $\mathcal{O}(\frac{k}{\epsilon}\text{loglog}(\frac{n\epsilon}{k}))$ samples and $\mathcal{O}(\log^*(k)\text{loglog}(\frac{n\epsilon}{k}))$ rounds for $(1+\epsilon)$-approximation for the ℓ_2/ℓ_2 problem [10]. Our result improves both the sample complexity (the first result) and the number of rounds (the second result). We summarize our results in Table 1.

1.2 Our Techniques

ℓ_∞/ℓ_2 Sparse Recovery. Our ℓ_∞/ℓ_2 sparse recovery scheme hashes every $i \in [n]$ to poly(k) buckets, and then proceeds by finding all the buckets that have ℓ_2 mass at least $\Omega(\frac{1}{\sqrt{k}}\|x_{-\Omega(k)}\|_2)$. We then find a set of buckets that contain all heavy coordinates, which are isolated from each other due to hashing. Then, we run a 1-sparse recovery in each bucket in parallel in order to find all the heavy coordinate. However, since we have $\mathcal{O}(k)$ buckets, we cannot afford to take a union bound over all one-sparse recovery routines called. Instead, we show that most buckets succeed and hence we can substract from x the elements returned, and then run a standard CountSketch algorithm to recover everything else. This algorithm obtains an optimal $\mathcal{O}(\text{loglog}(n))$ number of rounds and $\mathcal{O}(k\log(k) + k\text{loglog}(n))$ number of measurements, while succeeding with probability at least $1 - 1/\text{poly}(k)$.

We proceed by showing an algorithm for ℓ_2/ℓ_2 sparse recovery with $\mathcal{O}(\frac{k}{\epsilon}\text{loglog}(n))$ measurements and $\mathcal{O}(\text{loglog}(n))$ rounds. This will be important for our more general ℓ_p/ℓ_p scheme, saving a $\log^*(k)$ factor from the number of rounds, achieving optimality with respect to this quantity. For this scheme, we utilize the ℓ_∞/ℓ_2 scheme we just developed, observing that for small $k < \mathcal{O}(\log(n))$, the measurement complexity is $\mathcal{O}(k\text{loglog}(n))$. The algorithm hashes to $k/(\epsilon\log(n))$ buckets, and in each bucket runs ℓ_∞/ℓ_2 with sparsity k/ϵ. The ℓ_∞/ℓ_2 algorithm in each bucket succeeds with probability $1 - 1/\text{polylog}(n))$; this fact allows us to argue that all but a $1/\text{polylog}(n)$ fraction of the buckets will succeed, and hence we can recover all but a $k/\text{polylog}(n))$ fraction of the heavy coordinates. The next step is to subtract these coordinates from our initial vector, and then run a standard ℓ_2/ℓ_2 algorithm with decreased sparsity.

ℓ_p/ℓ_p Sparse Recovery. Our ℓ_p/ℓ_p scheme, $0 < p < 2$, is based on carefully invoking several ℓ_2/ℓ_2 schemes with different parameters. We focus our discussion on $p = 1$, then mention extensions to general p. A main difficulty of adapting the ℓ_1/ℓ_1 scheme of [17] is that it relies upon an ℓ_∞/ℓ_2 scheme, and all known schemes, including ours, have at least a $k\log k$ dependence on the number of measurements, which is too large for our overall goal.

A key insight in [17] for ℓ_1/ℓ_1 is that since the output does not need to be exactly k-sparse, one can compensate for mistakes on approximating the top k entries of x by accurately outputting enough smaller entries. For example, if $k = 1$, consider two possible signals $x = (1, \epsilon, \ldots, \epsilon)$ and $x' = (1 + \epsilon, \epsilon, \ldots, \epsilon)$, where ϵ occurs $1/\epsilon$ times in both x and x'. One can show, using known lower bound techniques, that distinguishing x from x' requires $\Omega(1/\epsilon)$ measurements. Moreover, $x_1 = (1, 0, \ldots, 0)$ and $x'_1 = (1 + \epsilon, 0, \ldots, 0)$, and any 1-sparse approximation to x or x' must therefore distinguish x from x', and so requires $\Omega(1/\epsilon)$ measurements. An important insight though, is that if one does not require the output signal y to be 1-sparse, then one can output $(1, \epsilon, 0, \ldots, 0)$ in both cases, without actually distinguishing which case one is in!

As another example, suppose that $x = (1, \epsilon, \ldots, \epsilon)$ and $x' = (1 + \epsilon^c, \epsilon, \ldots, \epsilon)$ for some $0 < c < 1$. In this case, one can show that one needs $\Omega(1/\epsilon^c)$ measurements to distinguish x and x', and as before, to output an exactly 1-sparse signal providing a $(1 + \epsilon)$-approximation requires $\tilde{\Theta}(1/\epsilon^c)$ measurements. In this case if one outputs a signal y with $y_1 = 1$, one cannot simply find a single other coordinate ϵ to "make up" for the poor approximation on the first coordinate. However, if one were to output $1/\epsilon^{1-c}$ coordinates each of value ϵ, then the ϵ^c "mass" lost by poorly approximating the first coordinate would be compensated for by outputting $\epsilon \cdot 1/\epsilon^{1-c} = \epsilon^c$ mass on these remaining coordinates. It is not clear how to find such remaining coordinates though, since they are much smaller; however, if one randomly subsamples an ϵ^c fraction of coordinates, then roughly $1/\epsilon^{1-c}$ of the coordinates of value ϵ

survive and these could all be found with a number of measurements proportional to $1/\epsilon^{1-c}$. Balancing the two measurement complexities of $1/\epsilon^c$ and $1/\epsilon^{1-c}$ at $c = 1/2$ gives roughly the optimal $1/\epsilon^{1/2}$ dependence on ϵ in the number of measurements.

To extend this to the adaptive case, a recurring theme of the above examples is that the top k, while they need to be found, they do not need to be approximated very accurately. Indeed, they do need to be found, if, e.g., the top k entries of x were equal to an arbitrarily large value and the remaining entries were much smaller. We accomplish this by running an ℓ_2/ℓ_2 scheme with parameters $k' = \Theta(k)$ and $\epsilon' = \Theta(\sqrt{\epsilon})$, as well as an ℓ_2/ℓ_2 scheme with parameters $k' = \Theta(k/\sqrt{\epsilon})$ and $\epsilon' = \Theta(1)$ (up to logarithmic factors in $1/\epsilon$). Another theme is that the mass in the smaller coordinates we find to compensate for our poor approximation in the larger coordinates also does not need to be approximated very well, and we find this mass by subsampling many times and running an ℓ_2/ℓ_2 scheme with parameters $k' = \Theta(1)$ and $\epsilon' = \Theta(1)$. This technique is surprisingly general, and does not require the underlying error measure we are approximating to be a norm. It just uses scale-invariance and how its rate of growth compares to that of the ℓ_2-norm.

ℓ_2/ℓ_2 Sparse Recovery. Our last algorithm, which concerns ℓ_2/ℓ_2 sparse recovery, achieves $\mathcal{O}(k\text{loglog}(n) + \frac{k}{\epsilon}\text{loglog}(1/\epsilon))$ measurements, showing that ϵ does not need to multiply loglog(n). The key insight lies in first solving the 1-sparse recovery task with $\mathcal{O}(\text{loglog}(n) + \frac{1}{\epsilon}\text{loglog}(1/\epsilon))$ measurements, and then extending this to the general case. To achieve this, we hash to polylog$(1/\epsilon)$ buckets, then solve ℓ_2/ℓ_2 with constant sparsity on a new vector, where coordinate j equals the ℓ_2 norm of the jth bucket; this steps requires only $\mathcal{O}(\frac{1}{\epsilon}\text{loglog}(1/\epsilon))$ measurements. Now, we can run standard 1-sparse recovery in each of these buckets returned. Extending this idea to the general case follows by plugging this sub-routine in the iterative algorithm of [10], while ensuring that sub-sampling does not increase the number of measurements. For that we also need to sub-sample at a slower rate, slower roughly by a factor of ϵ.

Notation: For a vector $x \in \mathbb{R}^n$, we define $H_k(x)$ to be the set of its largest k coordinates in absolute value. For a set S, denote by x_S the vector with every coordinate $i \notin S$ being zeroed out. We also define $x_{-k} = x_{[n]\setminus H_k(x)}$ and $H_{k,\epsilon}(x) = \{i \in [n] : |x_i| \geq \frac{\epsilon}{k}\|x_{-k}\|_2^2\}$, where $[n]$ represents the set $\{1, 2, ..., n\}$. For a set S, let $|S|$ be the cardinality of S.

Due to space constraints, we defer the proof of Theorem 2 to the full version[1].

2 Adaptive ℓ_p/ℓ_p Recovery

This section is devoted to proving Theorem 1. Our algorithm for ℓ_p/ℓ_p recovery is in Algorithm 1.

Let $f = \epsilon^{p/2}$, $r = 2/(p\log(1/f))$ and $q = \max\{p - \frac{1}{2}, 0\} = (p - \frac{1}{2})^+$. We will invoke the following ℓ_2/ℓ_2 oracle frequently throughout the paper.

▶ **Oracle 1** (AdaptiveSparseRecovery$_{\ell_p/\ell_q}(x, k, \epsilon)$). *The oracle is fed with (x, k, ϵ) as input parameters, and outputs a set of coordinates $i \in [n]$ of size $\mathcal{O}(k)$ which corresponds to the support of vector $\hat{x}$, where $\hat{x}$ can be any vector for which $\|x-\hat{x}\|_p \leq (1+\epsilon)\min_{\mathcal{O}(k)\text{-sparse } x'} \|x-x'\|_q$.*

[1] see https://arxiv.org/pdf/1804.09673.pdf

Algorithm 1 Adaptive ℓ_p/ℓ_p Recovery.

1. $A \leftarrow \text{AdaptiveSparseRecovery}_{\ell_2/\ell_2}(x, 2k/f, 1/10)$.
2. $B \leftarrow \text{AdaptiveSparseRecovery}_{\ell_2/\ell_2}(x, 4k, f/r^2)$.
3. $S \leftarrow A \cup B$.
4. For $j = 1 : r$
5. Uniformly sample the entries of x with probability $2^{-j}f/k$ for $k/(2f(r+1)^q)$ times.
6. Run the adaptive $\text{AdaptiveSparseRecovery}_{\ell_2/\ell_2}(x, 2, 1/(4(r+1))^{\frac{2}{p}})$ algorithm on each of the $k/(2f(r+1)^q)$ subsamples to obtain sets $A_{j,1}, A_{j,2}, \ldots, A_{j,k/(2f(r+1)^q)}$.
7. Let $S_j \leftarrow \cup_{t=1}^{k/(2f(r+1)^q)} A_{j,t} \setminus \cup_{t=0}^{j-1} S_t$.
8. End For
9. Request the entries of x with coordinates $S_0, ..., S_r$.
Output: $\hat{x} = x_{S_0 \cup \cdots \cup S_r}$.

Existing algorithms can be applied to construct Oracle 1 for the ℓ_2/ℓ_2 case, such as [10]. Without loss of generality, we assume that the coordinates of x are ranked in decreasing value, i.e., $x_1 \geq x_2 \geq \cdots \geq x_n$.

▶ **Lemma 5.** *Suppose we subsample x with probability p and let y be the subsampled vector formed from x. Then with failure probability $e^{-\Omega(k)}$, $\|y_{-2k}\|_2 \leq \sqrt{2p}\left\|x_{-k/p}\right\|_2$.*

Proof. Let T be the set of coordinates in the subsample. Then $\mathbb{E}\left[\left|T \cap \left[\frac{3k}{2p}\right]\right|\right] = \frac{3k}{2}$. So by the Chernoff bound, $\Pr\left[\left|T \cap \left[\frac{3k}{2p}\right]\right| > 2k\right] \leq e^{-\Omega(k)}$. Thus $\left|T \cap \left[\frac{3k}{2p}\right]\right| \leq 2k$ holds with high probability. Let $Y_i = x_i^2$ if $i \in T$ $Y_i = 0$ if $i \in [n] \setminus T$. Then $\mathbb{E}\left[\sum_{i > \frac{3k}{2p}} Y_i\right] = p\left\|x_{-\frac{3k}{2p}}\right\|_2^2 \leq p\left\|x_{-k/p}\right\|_2^2$. Notice that there are at least $\frac{k}{2p}$ elements in $x_{-k/p}$ with absolute value larger than $\left|x_{\frac{3k}{2p}}\right|$. Thus for $i > \frac{3k}{2p}$, $Y_i \leq \left|x_{\frac{3k}{2p}}\right|^2 \leq \frac{2p}{k}\left\|x_{-k/p}\right\|_2^2$. Again by a Chernoff bound, $\Pr\left[\sum_{i > \frac{3k}{2p}} Y_i \geq \frac{4p}{3}\left\|x_{-k/p}\right\|_2^2\right] \leq e^{-\Omega(k)}$. Conditioned on the latter event not happening, $\|y_{-2k}\|_2^2 \leq \sum_{i > \frac{3k}{2p}} Y_i \leq \frac{4p}{3}\left\|x_{-k/p}\right\|_2^2 \leq 2p\left\|x_{-k/p}\right\|_2^2$. By a union bound, with failure probability $e^{-\Omega(k)}$, we have $\|y_{-2k}\|_2 \leq \sqrt{2p}\left\|x_{-k/p}\right\|_2$. ◀

▶ **Lemma 6.** *Let $\hat{x}$ be the output of the ℓ_2/ℓ_2 scheme on x with parameters $(k, \epsilon/2)$. Then with small constant failure probability, $\left\|x_{[k]}\right\|_p^p - \|\hat{x}\|_p^p \leq k^{1-\frac{p}{2}}\epsilon^{\frac{p}{2}}\|x_{-k}\|_2^p$.*

Proof. Notice that with small constant failure probability, the ℓ_2/ℓ_2 guarantee holds and we have

$$\left\|x_{[k]}\right\|_2^2 - \|\hat{x}\|_2^2 = \|x - \hat{x}\|_2^2 - \|x_{-k}\|_2^2 \leq (1+\epsilon)\|x_{-k}\|_2^2 - \|x_{-k}\|_2^2 = \epsilon\|x_{-k}\|_2^2.$$

Let $S \subset [n]$ be such that $x_S = \hat{x}$, and define $y = x_{[k]\setminus S}$, $z = x_{S\setminus[k]}$. Then if $\|y\|_p^p \leq k^{1-\frac{p}{2}}\epsilon^{\frac{p}{2}}\|x_{-k}\|_2^p$ we are done. Otherwise, let $1 \leq k' \leq k$ denote the size of $[k] \setminus S$, and define $c = \|y\|_2/\sqrt{k'}$.

$$\begin{aligned}\left\|x_{[k]}\right\|_p^p - \|\hat{x}\|_p^p = \|y\|_p^p - \|z\|_p^p &\leq k'^{1-\frac{p}{2}}\|y\|_2^p - \|z\|_p^p = \frac{\|y\|_2^2}{c^{2-p}} - \|z\|_p^p \\ &\leq \frac{\|y\|_2^2 - \|z\|_2^2}{c^{2-p}} = \frac{\left\|x_{[k]}\right\|_2^2 - \|\hat{x}\|_2^2}{c^{2-p}} \leq \frac{\epsilon\|x_{-k}\|_2^2}{c^{2-p}}.\end{aligned}$$

Since $c \geq \frac{\|y\|_p}{k'^{\frac{1}{p}}} \geq \frac{\|y\|_p}{k^{\frac{1}{p}}} \geq \sqrt{\frac{\epsilon}{k}} \|x_{-k}\|_2$, we have $\left\|x_{[k]}\right\|_p^p - \|\hat{x}\|_p^p \leq k^{\frac{2-p}{2}} \epsilon^{1-\frac{2-p}{2}} \|x_{-k}\|_2^{2-(2-p)} = k^{1-\frac{p}{2}} \epsilon^{\frac{p}{2}} \|x_{-k}\|_2^p$. ◀

▶ **Theorem 7.** *Fix $0 < p < 2$. For $x \in \mathbb{R}^n$, there exists a $(1+\epsilon)$-approximation algorithm that performs $\mathcal{O}(\frac{k}{\epsilon^{p/2}} \text{loglog}(n) \log^{\frac{2}{p}+1-(p-\frac{1}{2})^+}(\frac{1}{\epsilon}))$ adaptive linear measurements in $\mathcal{O}(\text{loglog}(n))$ rounds, and with probability at least $2/3$, we can find a vector $\hat{x} \in \mathbb{R}^n$ such that*

$$\|x - \hat{x}\|_p \leq (1+\epsilon) \|x_{-k}\|_p . \tag{2}$$

Proof. The algorithm is stated in Algorithm 1. We first consider the difference $\left\|x_{[k]}\right\|_p^p - \|x_{S_0}\|_p^p$.

Let $i^*(0)$ be the smallest integer such that for any $l > i^*(0)$, $|x_l| \leq \|x_{-2k/f}\|_2/\sqrt{k}$.

Case 1. $i^*(0) > 4k$

Then for all $k < j \leq 4k$, we have $|x_j| > \|x_{-2k/f}\|_2/\sqrt{k}$. Hence x_{S_0} must contain at least $1/2$ of these indices; if not, the total squared loss is at least $1/2 \cdot 3k \|x_{-2k/f}\|_2^2/k \geq (3/2)\|x_{-2k/f}\|_2^2$, a contradiction to $\epsilon' = 1/10$. It follows that $\|x_{S_0 \cap \{k+1,\ldots,4k\}}\|_p^p \geq \frac{3}{2}k \left[\frac{\|x_{-2k/f}\|_2}{\sqrt{k}}\right]^p = \frac{3}{2}k^{1-\frac{p}{2}} \|x_{-2k/f}\|_2^p$. On the other hand, $\left\|x_{[k]}\right\|_p^p - \|x_{S_0}\|_p^p$ is at most $1.1k^{1-\frac{p}{2}} \|x_{-2k/f}\|_2^p$, since by the ℓ_2/ℓ_2 guarantee

$$\|x_{[k]}\|_p^p - \|x_{S_0 \cap [k]}\|_p^p \leq k^{1-\frac{p}{2}} \|x_{[k]} - x_{S_0 \cap [k]}\|_2^p \leq k^{1-\frac{p}{2}} \|x - x_{S_0}\|_2^p \leq \frac{11}{10} k^{1-\frac{p}{2}} \|x_{-2k/f}\|_2^p.$$

It follows that

$$\begin{aligned} \|x_{[k]}\|_p^p - \|x_{S_0}\|_p^p &= \|x_{[k]}\|_p^p - \|x_{S_0 \cap [k]}\|_p^p - \|x_{S_0 \cap \{k+1,\ldots,4k\}}\|_p^p \\ &\leq \frac{11}{10} k^{1-\frac{p}{2}} \|x_{-2k/f}\|_2^p - \frac{3}{2} k^{1-\frac{p}{2}} \|x_{-2k/f}\|_2^p \leq 0. \end{aligned}$$

Case 2. $i^*(0) \leq 4k$, and $\sum_{j=i^*(0)+1}^{2k/f} x_j^2 \geq 4\|x_{-2k/f}\|_2^2$.

We claim that x_{S_0} must contain at least a $5/8$ fraction of coordinates in $\{i^*(0)+1, \ldots, 2k/f\}$; if not, then the cost for missing at least a $3/8$ fraction of the ℓ_2-norm of $x_{\{i^*(0)+1,\ldots,2k/f\}}$ will be at least $(3/2)\|x_{-2k/f}\|_2^2$, contradicting the ℓ_2/ℓ_2 guarantee. Since all coordinates x_j's for $j > i^*(0)$ have value at most $\|x_{-2k/f}\|_2/\sqrt{k}$, it follows that the p-norm of coordinates corresponding to $\{i^*(0)+1, \ldots, 2k/f\} \cap S_0$ is at least $\left\|x_{\{i^*(0)+1,\ldots,2k/f\} \cap S_0}\right\|_p^p \geq \frac{5}{2} k^{\frac{2-p}{2}} \frac{\|x_{-2k/f}\|_2^2}{\|x_{-2k/f}\|_2^{2-p}} = \frac{5}{2} k^{1-\frac{p}{2}} \|x_{-2k/f}\|_2^p$. Then

$$\begin{aligned} \|x_{[k]}\|_p^p - \|x_{S_0}\|_p^p &\leq \frac{11}{10} k^{1-\frac{p}{2}} \|x_{-2k/f}\|_2^p + k \left(\frac{\|x_{-2k/f}\|_2}{\sqrt{k}}\right)^p - \|x_{\{i^*(0)+1,\ldots,2k/f\} \cap S_0}\|_p^p \\ &\leq \frac{21}{10} k^{1-\frac{p}{2}} \|x_{-2k/f}\|_2^p - \frac{5}{2} k^{1-\frac{p}{2}} \|x_{-2k/f}\|_2^p \leq 0. \end{aligned}$$

Case 3. $i^*(0) \leq 4k$, and $\sum_{j=i^*(0)+1}^{2k/f} x_j^2 \leq 4\|x_{-2k/f}\|_2^2$.

With a little abuse of notation, let x_{S_0} denote the output of the ℓ_2/ℓ_2 with parameters $(4k, f/r^2)$. Notice that there are at most $8k$ non-zero elements in x_{S_0}, and $\|x_{-4k}\|_2^2 \leq \|x_{-i^*(0)}\|_2^2 = \sum_{j=i^*(0)+1}^{2k/f} x_j^2 + \|x_{-2k/f}\|_2^2 \leq 5\|x_{-2k/f}\|_2^2$. By Lemma 6, we have $\left\|x_{[k]}\right\|_p^p - \|x_{S_0}\|_p^p \leq \left\|x_{[4k]}\right\|_p^p - \|x_{S_0}\|_p^p \leq (4k)^{1-\frac{p}{2}} \frac{f^{\frac{p}{2}}}{r^p} \|x_{-4k}\|_2^p \leq \mathcal{O}\left(\frac{1}{r^p}\right) k^{1-\frac{p}{2}} f^{\frac{p}{2}} \|x_{-2k/f}\|_2^p$. According to the above three cases, we conclude that $\|x_{[k]}\|_p^p - \|x_{S_0}\|_p^p \leq \mathcal{O}\left(\frac{1}{r^p}\right) k^{1-\frac{p}{2}} f^{\frac{p}{2}} \|x_{-2k/f}\|_2^p$. Thus with failure probability at most $1/6$,

$$\|x - \hat{x}\|_p^p - \|x_{-k}\|_p^p = \|x_{[k]}\|_p^p - \sum_{j=0}^{r} \|x_{S_j}\|_p^p \leq \mathcal{O}\left(\frac{1}{r^p}\right) k^{1-\frac{p}{2}} f^{\frac{p}{2}} \|x_{-2k/f}\|_2^p - \sum_{j=1}^{r} \left\|x_{S_j}\right\|_p^p . \tag{3}$$

In order to convert the first term on the right hand side of (3) to a term related to the ℓ_p norm (which is a semi-norm if $0 < p < 1$), we need the following inequalities: for every u and s, by splitting into chunks of size s, we have

$$s^{1-\frac{p}{2}} \|u_{-2s}\|_2^p \le \|u_{-s}\|_p^p, \qquad \text{and} \qquad \left\|u_{\overline{[s]}\cap[2s]}\right\|_2 \le \sqrt{s}\,|u_s|\,.$$

Define $c = (r+1)^{\min\{p,1\}}$. This gives us that, for $0 < p < 2$ $\frac{1}{(r+1)^p} k^{1-\frac{p}{2}} f^{\frac{p}{2}} \left\|x_{-2k/f}\right\|_2^p \le \frac{k^{1-\frac{p}{2}} f^{\frac{p}{2}}}{c} \left\|x_{-2k/f^{1+\frac{2}{p}}}\right\|_2^p + \frac{k^{1-\frac{p}{2}} f^{\frac{p}{2}}}{c} \sum_{j=1}^{r} \left\|x_{\overline{[2^j k/f]}\cap[2^{j+1}k/f]}\right\|_2^p \le \frac{f^{(1-\frac{p}{2})(1+\frac{2}{p})+\frac{p}{2}}}{c} \left\|x_{-k/f^{1+\frac{2}{p}}}\right\|_p^p + \frac{1}{c}\sum_{j=1}^{r} k 2^{pj/2} \left|x_{2^j k/f}\right|^p$. Therefore,

$$\begin{aligned}\|\hat{x}-x\|_p^p - \|x_{-k}\|_p^p &\le \mathcal{O}\left(\frac{1}{c}\right) f^{\frac{2}{p}} \left\|x_{-k/f^{1+\frac{2}{p}}}\right\|_p^p + \sum_{j=1}^{r} \mathcal{O}\left(\frac{1}{c}\right) k 2^{pj/2} |x_{2^j k/f}|^p - \sum_{j=1}^{r} \|x_{S_j}\|_p^p \\ &\le \mathcal{O}\left(\frac{1}{c}\right) f^{\frac{2}{p}} \left\|x_{-k/f}\right\|_p^p + \sum_{j=1}^{r} \mathcal{O}\left(\frac{1}{c}\right) k 2^{pj/2} |x_{2^j k/f}|^p - \sum_{j=1}^{r} \|x_{S_j}\|_p^p. \end{aligned} \tag{4}$$

Let $y = x_T$ denote an independent subsample of x with probability $f/(2^j k)$, and $\hat{y}$ be the output of the ℓ_2/ℓ_2 algorithm with parameter $\mathrm{s}(2, 1/(4(r+1))^{\frac{2}{p}})$. Notice that $|S_j| \le 2k/(r+1)f$ by the adaptive ℓ_2/ℓ_2 guarantee. Define $Q = [2^j k/f] \setminus (S_0 \cup \cdots \cup S_{j-1})$. There are at least $2^j k/(2f)$ elements in Q, and every element in Q has absolute value at least $\left|x_{2^j k/f}\right|$. In each subsample, notice that $\mathbb{E}[|T \cap Q|] = \frac{1}{2}$. Thus with sufficiently small constant failure probability there exists at least 1 element in y with absolute value at least $|x_{2^j k/f}|$. On the other hand, by Lemma 6 and Lemma 5,

$$\left\|y_{[1]}\right\|_p^p - \|\hat{y}\|_p^p \le \left\|y_{[2]}\right\|_p^p - \|\hat{y}\|_p^p \le \frac{2^{1-\frac{p}{2}}}{4(r+1)} \|y_{-2}\|_2^p \le \frac{1}{2(r+1)} \left(\frac{f}{2^j k}\right)^{\frac{p}{2}} \left\|x_{-2^j k/f}\right\|_2^p, \tag{5}$$

with sufficiently small constant failure probability given by the union bound. For the $k/(2f(r+1)^q)$ independent copies of subsamples, by a Chernoff bound, a $1/4$ fraction of them will have the largest absolute value in Q and (5) will also hold, with the overall failure probability being $e^{-\Omega(k/(fr^q))}$. Therefore, since $k/f > 2^{pj/2}k$,

$$\begin{aligned}\left\|x_{S_j}\right\|_p^p &\ge \frac{2^{pj/2}k}{8(r+1)^q} \left[\left|x_{2^j k/f}\right|^p - \frac{1}{2(r+1)} \left(\frac{f}{2^j k}\right)^{\frac{p}{2}} \left\|x_{-2^j k/f}\right\|_2^p\right] \\ &\ge \frac{2^{pj/2}k}{8(r+1)^q} \left|x_{2^j k/f}\right|^p - \frac{k^{1-\frac{p}{2}} f^{\frac{p}{2}}}{16(r+1)^{q+1}} \left\|x_{-2k/f}\right\|_2^p, \end{aligned}$$

and by the fact that $0 < q < p < 2$,

$$\begin{aligned}\|x-\hat{x}\|_p^p - \|x_{-k}\|_p^p &\le \mathcal{O}(\frac{1}{r^p}) k^{1-\frac{p}{2}} f^{\frac{p}{2}} \|x_{-2k/f}\|_2^p - \sum_{j=1}^{r} \left\|x_{S_j}\right\|_p^p \\ &\le \left[\mathcal{O}\left(\frac{1}{r^p}\right) + \frac{r}{16(r+1)^{q+1}}\right] k^{1-\frac{p}{2}} f^{\frac{p}{2}} \|x_{-2k/f}\|_2^p - \sum_{j=1}^{r} \frac{2^{pj/2}k}{8(r+1)^q} \left|x_{2^j k/f}\right|^p \\ &\le \mathcal{O}\left(\frac{1}{c}\right) f^{\frac{2}{p}} \left\|x_{-k/f}\right\|_p^p + \left[\mathcal{O}\left(\frac{1}{c}\right) + \frac{1}{16(r+1)^q} - \frac{1}{8(r+1)^q}\right] \sum_{j=1}^{r} k 2^{pj/2} \left|x_{2^j k/f}\right|^p \\ &\le f^{\frac{2}{p}} \left\|x_{-k/f}\right\|_p^p \le \epsilon \|x_{-k}\|_p^p\,. \end{aligned}$$

The total number of measurements will be at most

$$\mathcal{O}\left(\frac{k}{f}\mathrm{loglog}(n) + \frac{4kr^2}{f}\mathrm{loglog}(n) + \frac{kr}{2fr^q} r^{\frac{2}{p}} \mathrm{loglog}(n)\right) = \mathcal{O}\left(\frac{k}{\epsilon^{\frac{p}{2}}} \mathrm{loglog}(n) \log^{\frac{2}{p}+1-(p-\frac{1}{2})^+}\left(\frac{1}{\epsilon}\right)\right),$$

while the total failure probability given by the union bound is $1/6 + e^{-\Omega(k/(fr^q))} < 1/3$, which completes the proof. ◀

3 ℓ_∞/ℓ_2 Adaptive Sparse Recovery

In this section, we will prove Theorem 3. Our algorithm first approximates $\|x_{-k}\|_2$. The goal is to compute a value V which is not much smaller than $\frac{1}{k}\|x_{-k}\|_2^2$, and also at least $\Omega(\frac{1}{k})\|x_{-\Omega(k)}\|_2^2$. This value will be used to filter out coordinates that are not large enough, while ensuring that heavy coordinates are included. We need the following lemma, which for example can be found in Section 4 of [13].

▶ **Lemma 8.** *Using* $\log(1/\delta)$ *non-adaptive measurements we can find with probability* $1-\delta$ *a value* V *such that* $\frac{1}{C_1 k}\|x_{-C_2 k}\|_2^2 \le V \le \frac{1}{k}\|x_{-k}\|_2^2$, *where* C_1, C_2 *are absolute constants larger than* 1.

We use the aforementioned lemma with $\Theta(\log k)$ measuremenents to obtain such a value V with probability $1-1/\text{poly}(k)$. Now let c be an absolute constant and let $g : [n] \to [k^c]$ be a random hash function. Then, with probability at least $1 - \frac{1}{\text{poly}(k)}$ we have that for every $i, j \in H_k(x)$, $g(i) \neq g(j)$. By running PartitionCountSketch $(x, 2C_1 k, \{g^{-1}(1), g^{-1}(2), \ldots, g^{-1}(k^c)\}$, we get back an estimate w_j for every $j \in [k^c]$; here C_1 is an absolute constant. Let γ' be an absolute constant to be chosen later. We set $S = \{j \in [k^c] : w_j^2 \ge \gamma' V\}$ and $T = \bigcup_{j \in S} g^{-1}(j)$. We prove the following lemma.

▶ **Lemma 9.** *Let* C' *be an absolute constant. With probability at least* $1 - 1/\text{poly}(k)$ *the following holds.*

1. $|S| = \mathcal{O}(k)$.
2. *Every* $j \in [k^c]$ *such that there exists* $i \in H_k(x) \cap g^{-1}(j)$, *will be present in* S.
3. *For every* $j \in S$, *there exists exactly one coordinate* $i \in g^{-1}(j)$ *with* $x_i^2 \ge \frac{1}{C' k}\|x_{-C_2 k}\|_2^2$.
4. *For every* $j \in S$, $\|x_{g^{-1}(j) \setminus H_k(x)}\|_2^2 \le \frac{1}{k^2}\|x_{-k}\|_2^2$.

Proof. Let C_0 be an absolute constant larger than 1. Note that with probability $1 - C_0^2 \cdot k^{6-c}$, all $i \in H_{C_0 k^3}(x)$ (and, hence, also in $H_{C_0 k^3, 1/k^3}(x)$) are isolated under g. Fix $j \in [k^c]$ and, for $i \in [n]$, define the random variable $Y_i = 1_{g(x_i)=j} x_i^2$. Now observe that

$$\mathbb{E}\left[\sum_{i \in g^{-1}(j) \setminus H_{C_0 k^3, 1/k^3}(x)} Y_i\right] = \frac{1}{k^c}\|x_{-C_0 k^3}\|_2^2.$$

Applying Bernstein's inequality to the variables Y_i with

$$K = \frac{1}{C_0 k^3}\|x_{-C_0 k^3}\|_2^2, \qquad \text{and} \qquad \sigma^2 < \frac{1}{k^{c+3}}\|x_{-C_0 k^3}\|_2^4,$$

we have that

$$\Pr\left[\sum_{i \in g^{-1}(j) \setminus H_{C_0 k^3, 1/k^3}(x)} x_i^2 \ge 1/k^2 \|x_{-C_0 k^2}\|_2^2\right] \le e^{-k},$$

where c is an absolute constant. This allows us to conclude that the above statement holds for all different k^c possible values j, by a union-bound. We now prove the bullets one by one. We remind the reader that PartitionCountSketch aproximates the value of every $\|x_{g^{-1}(j)}\|_2^2$ with a multiplicate error in $[1-\gamma, 1+\gamma]$ and additive error $\frac{1}{C_0 k}\|x_{-k}\|_2^2$.

1. Since there are at most $\frac{1}{\gamma'(1+\gamma)}C_2k + C_2k$ indices j with $(1+\gamma)\|x_{g^{-1}(j)}\|_2^2 \geq \frac{\gamma'}{k}\|x_{-k}\|_2^2 \geq \gamma' V$, the algorithm can output at most $\mathcal{O}(k)$ indices.
2. The estimate for such a j will be at least $(1-\gamma)\frac{1}{k}\|x_{-k}\|_2^2 - \frac{1}{2C_1k}\|x_{-C_2k}\|_2^2 \geq \gamma' V$, for some suitable choice of γ'. This implies that j will be included in S.
3. Because of the guarantee for V and the guarantee of PartitionCountSketch, we have that all j that are in S satisfy $(1+\gamma)\|x_{g^{-1}(j)}\|_2^2 + \frac{1}{k}\|x_{-2C_1k}\|_2^2 \geq \frac{\gamma'}{k}\|x_{-C_2k}\|_2^2$, and since

$$\sum_{i \in g^{-1}(j) \setminus H_{C_0k^3}(x)} x_i^2 \leq \frac{1}{k^2}\|x_{-k}\|_2^2,$$

   this implies that there exists $i \in H_{C_0k^3}(x) \cap g^{-1}(j)$. But since all $i \in H_{C_0k^3}(x)$ are perfectly hashed under g, this implies that this i should satisfy $x_i^2 \geq \frac{1}{C'k}\|x_{-C_2k}\|_2^2$, from which the claim follows.
4. Because elements in $H_{C_0k^3}(x)$ are perfectly hashed, we have that

$$\|x_{g^{-1}(j)\setminus H_k(x)}\|_2^2 = \|x_{g^{-1}(j)\setminus H_{C_0k^3}}(x)\|_2^2 \leq \frac{1}{k^2}\|x_{-k}\|_2^2$$

   for C_0 large enough. ◀

Given S, we proceed in the following way. For every $j \in S$, we run the algorithm guaranteed by Lemma 15 from the full version [2] to obtain an index i_j, using $\mathcal{O}(k\text{loglog}n)$ measurements. Then we observe directly x_{i_j} using another $\mathcal{O}(k)$ measurements, and form vector $z = x - x_{\{i_j\}_{j\in S}}$. We need the following lemma.

▶ **Lemma 10.** *With probability* $1 - 1/\text{poly}(k)$, $|H_k(x) \setminus \{i_j\}_{j\in S}| \leq \frac{k}{\log^2 n}$.

Proof. Let us consider the calls to the 1-sparse recovery routine in j for which there exists $i \in H_k(x) \cap g^{-1}(j)$. Since the 1-sparse recovery routine succeeds with probability $1 - 1/\text{poly}(\log n)$, then the probability that we have more than $\frac{k}{\log^2 n}$ calls that fail, is

$$\binom{k}{\frac{k}{\log^2 n}}\left(\frac{1}{\text{poly}(\log n)}\right)^{k/\log^2 n} \leq \frac{1}{\text{poly}(k)}.$$

This gives the proof of the lemma. ◀

For the last step of our algorithm, we run PartitionCountSketch$(z_T, k/\log(n), [n])$ to estimate the entries of z. We then find the coordinates with the largest $2k$ estimates, and observe them directly. Since

$$\frac{\log n}{k}\|(z_T)_{-k/\log n}\|_2^2 \leq \frac{\log n}{k} \cdot \frac{1}{k^2}\|x_{-k}\|_2^2 = \frac{\log n}{k^3}\|x_{-k}\|_2^2,$$

every coordinate will be estimated up to additive error $\frac{\log n}{k^3}\|x_{-k}\|_2^2$, which shows that every coordinate in $T \cap H_{k,1/k}(x)$ will be included in the top $2k$ coordinates. Putting everything together, we obtain the desired result.

[2] see https://arxiv.org/pdf/1804.09673.pdf

4 ℓ_2/ℓ_2 Adaptive Sparse Recovery in Optimal Rounds

In this section, we give an algorithm for ℓ_2/ℓ_2 compressed sensing using $\mathcal{O}(\text{loglog}n)$ rounds, instead of $\mathcal{O}(\log^* k \cdot \text{loglog}n)$ rounds. Specifically, we prove the first bullet of Theorem 4. We call this algorithm AdaptiveSparseRecovery$_{\ell_\infty/\ell_2}$.

We proceed with the design and the analysis of the algorithm. We note that for $k/\epsilon = \mathcal{O}(\log^5 n)$[3], ℓ_∞/ℓ_2 gives already the desired result. So, we focus on the case of $k/\epsilon = \Omega(\log^5 n)$. We pick a hash function $h : [n] \to [B]$, where $B = ck/(\epsilon \log n)$ for some constant c large enough. The following follows by an application of Bernstein's Inequality and the Chernoff Bound, similarly to ℓ_∞/ℓ_2.

▶ **Lemma 11.** *With probability* $1 - 1/\text{poly}(n)$*, the following holds:*

$$\forall j \in [B] : |H_{k/\epsilon}(x) \cap h^{-1}(j)| \le \log n, \qquad \text{and} \qquad \left| \sum_{i \in h^{-1}(j) \setminus H_{k/\epsilon}(x)} x_i^2 \right| \le \frac{\epsilon}{k} \|x_{-k}\|_2^2.$$

We now run the ℓ_∞/ℓ_2 algorithm for the previous section on vectors $x_{h^{-1}(1)}, x_{h^{-1}(2)}, \ldots, x_{h^{-1}(B)}$ with sparsity parameter $\mathcal{O}(\log n)$, to obtain vectors $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_B$. The number of rounds is $\mathcal{O}(\text{loglog}(n))$, since we can run the algorithm in every bucket in parallel. By the definition of the ℓ_∞/ℓ_2 algorithm, one can see that $|\text{supp}(\hat{x}_j)| \le \mathcal{O}(\log n)$. We set $S = \cup_{j \in B} |\text{supp}(x_j)|$, and observe that $|S| = ck/(\epsilon \log n) \cdot \mathcal{O}(\log n) = \mathcal{O}(k/\epsilon)$. The number of measurements equals $ck/(\epsilon \log n) \cdot \mathcal{O}(\log n \cdot \text{loglog}(n \log(n/k))) = \mathcal{O}((k/\epsilon) \cdot \text{loglog}(n \log(n/k)))$.

▶ **Lemma 12.** *With probability* $1 - 1/\text{poly}(n)$*, we have that* $|S \setminus H_{k/\epsilon}(x)| \le \frac{k}{\epsilon \log^2 n}$.

Proof. Since every call to ℓ_∞/ℓ_2 fails with probability $1/\text{poly}(\log n)$, the probability that we have more than a $\frac{1}{\log n}$ fraction of the calls that fail is at most

$$\binom{B}{B/\log^2 n} \left(\frac{1}{\log n}\right)^{B/\log n} \le (e \log^2 n)^{\log n} (\log n)^{-B/\log n} \le \frac{1}{\text{poly}(n)}.$$

This implies that S will contain all but at most $B/\log^2 n \cdot \log n = k/(\epsilon \log^2 n)$ coordinates $i \in H_k(x)$. ◀

We now observe x_S directly and form the vector $z = x - x_S$, for which $\|z_{-k/(\epsilon \log^2 n)}\|_2 \le \|x_{-k/\epsilon}\|_2$. We now run a standard ℓ_2/ℓ_2 algorithm that fails with probability $1/\text{poly}(n)$ to obtain a vector $\hat{z}$ that approximates z (for example PartitionCountSketch$(z, k/(\epsilon \log^2 n), [n])$ suffices). We then output $\hat{z} + x_S$, for which $\|\hat{z} + x_S - x\|_2 = \|\hat{z} - z\| \le (1+\epsilon)\|z_{-k/(\epsilon \log n)}\|_2 \le (1+\epsilon)\|x_{-k}\|_2$. The number of measurements of this step is $\mathcal{O}(\frac{1}{\epsilon} \frac{k}{\log^2 n} \cdot \log n) = o(\frac{k}{\epsilon})$. The total number of rounds is clearly $\mathcal{O}(\text{loglog}(n \log(\frac{n\epsilon}{k})))$.

5 ℓ_2/ℓ_2 with Improved Dependence on ϵ

In this section, we prove the second part of Theorem 4. We first need an improved algorithm for the 1-sparse recovery problem.

▶ **Lemma 13.** *Let* $x \in \mathbb{R}^n$*. There exists an algorithm* ImprovedOneSparseRecovery*, that uses* $\mathcal{O}(\text{loglog}n + \frac{1}{\epsilon}\text{loglog}(\frac{1}{\epsilon}))$ *measurements in* $\mathcal{O}(\text{loglog}(n))$ *rounds, and finds with sufficiently small constant probability an* $\mathcal{O}(1)$*-sparse vector* $\hat{x}$ *such that* $\|\hat{x} - x\|_2 \le (1 + \epsilon)\|x_{-1}\|_2$.

[3] the constant 5 is arbitrary

Proof. We pick a hash function $h : [n] \to [B]$, where $B = \lceil 1/\epsilon^h \rceil$ for a sufficiently large constant h. Observe that all elements of $H_{\sqrt{B}}(x)$ are perfectly hashed under h with constant probability, and, $\forall j \in [B]$, $\mathbb{E}\left[\left\|x_{h^{-1}(j)\setminus H_{\sqrt{B}}}(x)\right\|_2\right] \leq 1/B\|x_{-\sqrt{B}}\|_2$. As in the previous sections, invoking Bernstein's inequality we can get that with probability $1 - 1/\text{poly}(B)$, $\forall j \in [B]$, $\left\|x_{h^{-1}(j)\setminus H_{\sqrt{B}}(x)\|_2}\right\|_2^2 \leq \frac{c\log B}{B}\|x_{-\sqrt{B}}\|_2^2$, where c is some absolute constant, and the exponent in the failure probability is a function of c.

We now define the vector $z \in \mathbb{R}^B$, the j-th coordinate of which equals $z_j = \sum_{i\in h^{-1}(j)} \sigma_{i,j}x_i$. We shall invoke Khintchine inequality to obtain $\forall j$,

$$\Pr\left[\left|\sum_{i\in h^{-1}(j)\setminus H_{\sqrt{B}}(x)} \sigma_{i,j}x_i\right|^2 > \frac{c'}{\epsilon}\left\|x_{h^{-1}(j)\setminus H_{\sqrt{B}}(x)}\right\|_2^2\right] \leq e^{-\Omega(1/\epsilon^2)},$$

for some absolute constant c'. This allows us to take a union-bound over all $B = \lceil 1/\epsilon^h \rceil$ entries of z to conclude that there exists an absolute constant ζ such that $\forall j \in [B]$, $\left|\sum_{i\in h^{-1}(j)\setminus H_{\sqrt{B}}(x)} \sigma_{i,j}x_i\right|^2 \leq \frac{c'}{\epsilon}\|x_{h^{-1}(j)\setminus H_{\sqrt{B}}(x)}\|_2^2 < \zeta\epsilon\|x_{-1}\|_2^2$, by setting h large enough. Now, for every coordinate $j \in [B]$ for which $h^{-1}(j) \cap H_{1,\epsilon}(x) = i^*$ or some $i^* \in [n]$, we have that $|z_j| \geq \left||x_{i^*}| - \sqrt{\frac{c\log B}{B}\cdot\frac{c'}{\epsilon}}\|x_{-\sqrt{B}}\|_2\right| \geq (1-\zeta)\sqrt{\epsilon}\|x_{-1}\|_2$, whereas for every $j \in [B]$ such that $h^{-1}(j)\cap H_{1,\epsilon\zeta}(x) = \emptyset$ it holds that $|z_j| \leq 2\zeta\sqrt{\epsilon}\|x_{-1}\|_2$. We note that $H_{1,\epsilon}(x) \subset H_{\sqrt{B}}(x)$, and hence all elements of $H_{1,\epsilon}(x)$ are also perfectly hashed under h. Moreover, observe that $\mathbb{E}\|z_{-1}\|_2^2 \leq \|x_{-1}\|_2^2$, and hence by Markov's inequality, we have that $\|z_{-1}\|_2^2 \leq 10\|x_{-1}\|_2^2$ holds with probability $9/10$. We run the ℓ_2/ℓ_2 algorithm of Theorem 4 for vector z with the sparsity being set to 1, and obtain vector $\hat{z}$. We then set $S = \text{supp}(\hat{z})$. We now define $w = (|z_1|, |z_2|, \ldots)$, for which $\|w_{-1}\|_2 = \|z_{-1}\|_2$. Clearly, $\|z - z_S\|_2^2 \leq \|z - \hat{z}\|_2^2 \leq (1+\epsilon)\|z_{-1}\|_2^2 = (1+\epsilon)\|w_{-1}\|_2^2$. So $\|w - w_S\|_2^2 = \|z - z_S\|_2^2 \leq (1+\epsilon)\|w_{-1}\|_2^2$. We now prove that $\left\|x - x_{\cup_{j\in S}h^{-1}(j)}\right\|_2 \leq (1+\mathcal{O}(\epsilon))\|x_{-1}\|_2$. Let i^* be the largest coordinate in magnitude of x, and $j^* = h(i^*)$. If $j^* \in S$, then it follows easily that $\|x - x_{\cup_{j\in S}h^{-1}(j)}\|_2 \leq \|x_{-1}\|_2$. Otherwise, since $\sum_{j\neq j^*} w_j^2 = \|w_{-1}\|_2^2$, and $\sum_{j\notin S} w_j^2 \leq (1+\epsilon)\|w_{-1}\|_2^2$, it must be the case that $\left|w_{j^*}^2 - \|w_S\|_2^2\right| \leq \epsilon\|w_{-1}\|_2^2 \leq 10\epsilon\|x_{-1}\|_2^2$. The above inequality, translates to $\sum_{i\in h^{-1}(j^*)} x_i^2 \leq |S|\zeta\epsilon\|x_{-1}\|_2^2 + \zeta\epsilon\|x_{-1}\|_2^2 + 10\epsilon\|x_{-1}\|_2^2 + \sum_{j\in S}\sum_{i\in h^{-1}(j)} x_j^2 = \mathcal{O}(\epsilon)\|x_{-1}\|_2^2 + \sum_{j\in S}\sum_{i\in h^{-1}(j)} x_j^2$. This gives $\left\|x - x_{\cup_{j\in S}h^{-1}(j)}\right\|_2 = \sum_{i\in h^{-1}(j^*)} x_i^2 + \sum_{j\notin S\cup\{j^*\}}\sum_{i\in h^{-1}(j)} x_i^2 \leq \mathcal{O}(\epsilon)\|x_{-1}\|_2^2 + \mathcal{O}(1)\zeta\epsilon\|x_{-1}\|_2^2 + \sum_{j\in S}\sum_{i\in h^{-1}(j)} x_j^2 + \sum_{j\notin S\cup\{j^*\}}\sum_{i\in h^{-1}(j)} x_i^2 + \leq (1+\mathcal{O}(\epsilon))\|x_{-1}\|_2^2$.

Given S, we run the 1-sparse recovery routine on vectors x_j for $j \in S$, with a total of $\mathcal{O}(\text{loglog}n)$ measurements and $\mathcal{O}(\text{loglog}n)$ rounds. We then output $\{x_{i_j}\}_{j\in S}$. Let i_j be the index returned for $j \in S$ by the 1-sparse recovery routine. Since we have a constant number of calls to the 1-sparse recovery routine (because S is of constant size), all our 1-sparse recovery routines will succeed. We now have that $\|x - x_{\cup_{j\in S} i_j}\|_2 \leq \|x_{\bar{S}}\|_2 + \sum_{j\in S}\|x_{h^{-1}(j)} - x_{i_j}\|_2 \leq \|x_{\bar{S}}\|_2 + \sum_{j\in S}(1+\epsilon)\|x_{h^{-1}(j)\setminus H_1(x)}\|_1 \leq (1+\mathcal{O}(\epsilon))\|x_{-1}\|_2$. Rescaling ϵ, we get the desired result. ◀

The algorithm for general k is similar to [10], apart from the fact that we subsample at a slower rate, and also use our new 1-sparse recovery algorithm as a building block. In the algorithm below, R_r is the universe we are restricting our attention on at the rth round. Moreover, J is the set of coordinates that we have detected so far. We are now ready to prove Theorem 4.

Proof. The number of measurements is bounded in the exact same way as in Theorem 3.7 from [10].

Algorithm 2 Adaptive ℓ_2/ℓ_2 Sparse Recovery.

1. $R_0 \leftarrow [n]$.
2. $x_0 \leftarrow \vec{0}$.
3. $\delta_0 \leftarrow \delta/2, \epsilon_0 \leftarrow \epsilon/e, f_0 \leftarrow 1/32, k_0 \leftarrow k$.
4. $J \leftarrow \emptyset$.
5. **For** $r = 0$ to $\mathcal{O}(\log^* k)$ **do**
6. **For** $t = 0$ to $\Theta(k_r \log(1/(\delta_r f_r)))$ **do**
7. $S_t \leftarrow \textsc{Subsample}(x - x^{(r)}, R_r, 1/(C_0 k_r))$.
8. $J \leftarrow J \cup \textsc{ImprovedOneSparseRecovery}((x - x^{(r)})_{S_t})$.
9. **End For**
10. $R_{r+1} \leftarrow [n] \setminus J$.
11. $\delta_{r+1} \leftarrow \delta_r/8$.
12. $\epsilon_{r+1} \leftarrow \epsilon_r/2$.
13. $f_{r+1} \leftarrow 1/2^{1/(4^{i+r} f_r)}$.
14. $k_{r+1} \leftarrow f_r k_r$.
15. $R_{r+1} \leftarrow [n] \setminus J$.
16. **End For**
17. $\hat{x} \leftarrow x^{(r+1)}$.
18. Return $\hat{x}$.

We fix a round r and $i \in H_{k_r,\epsilon_r}(x^{(r)})$. Then the call to $\textsc{Subsample}(R_r, 1/(C_0 k_r))$ yields

$$\Pr\left[|H_{k_r,\epsilon_r}(x - x^{(r)}) \cap S_t| = \{i\}\right] \geq \frac{1}{C_0 k_r}, \quad \mathbb{E}\left[\|x_{S_t \setminus H_{k_r,\epsilon_i}(x^{(r)})}\|_2^2\right] = \frac{1}{C_0 k_r}\|x_{-k_r}\|_2^2.$$

Setting C_0 to be large enough and combining Markov's inequality with the guarantee of Lemma 13, we get that the probability that the call to $\textsc{ImprovedOneSparseRecovery}(x_{S_t})$ returns i is $\Theta(1/k_r)$. Because we repeat $k_r \log(1/(f_r\delta_r))$, the probability that i or a set S_i of size $\mathcal{O}(1)$ such that $\|x_{\{i\}} - x_{S_i}\|_2 \leq \epsilon_i \|x_{-k_r}\|_2^2$, is not added in J is at most $(1 - 1/k_r)^{k_r \log(1/(f_r\delta_r))} = f_r\delta_r$.

Given the above claim, the number of measurements is $\mathcal{O}((k\text{loglog}n + k/\epsilon\text{loglog}(1/\epsilon)\log(1/\delta))$ and the analysis of the iterative loop proceeds almost identically to Theorem 3.7 of [10]. ◀

References

1 Akram Aldroubi, Haichao Wang, and Kourosh Zarringhalam. Sequential adaptive compressed sampling via Huffman codes. *arXiv preprint arXiv:0810.4916*, 2008.
2 Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Annual Conference on Learning Theory*, pages 152–192, 2016.
3 Khanh Do Ba, Piotr Indyk, Eric Price, and David P. Woodruff. Lower bounds for sparse recovery. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1190–1197, 2010.
4 Rui M. Castro, Jarvis Haupt, Robert Nowak, and Gil M. Raz. Finding needles in noisy haystacks. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5133–5136, 2008.
5 Anna C. Gilbert, Yi Li, Ely Porat, and Martin J. Strauss. Approximate sparse recovery: optimizing time and measurements. *SIAM Journal on Computing*, 41(2):436–453, 2012.
6 Rishi Gupta, Piotr Indyk, Eric Price, and Yaron Rachlin. Compressive sensing with local geometric features. *International Journal of Computational Geometry & Applications*, 22(04):365–390, 2012.

7 Jarvis Haupt, Waheed U Bajwa, Michael Rabbat, and Robert Nowak. Compressed sensing for networked data. *IEEE Signal Processing Magazine*, 25(2):92–101, 2008.

8 Jarvis Haupt, Robert Nowak, and Rui Castro. Adaptive sensing for sparse signal recovery. In *Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop*, pages 702–707, 2009.

9 Jarvis D. Haupt, Richard G. Baraniuk, Rui M. Castro, and Robert D. Nowak. Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements. In *Asilomar Conference on Signals, Systems and Computers*, pages 1551–1555, 2009.

10 Piotr Indyk, Eric Price, and David P. Woodruff. On the power of adaptivity in sparse recovery. In *Annual IEEE Symposium on Foundations of Computer Science*, pages 285–294, 2011.

11 Shihao Ji, Ya Xue, and Lawrence Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346–2356, 2008.

12 Raghunandan M. Kainkaryam, Angela Bruex, Anna C. Gilbert, John Schiefelbein, and Peter J. Woolf. poolmc: Smart pooling of mrna samples in microarray experiments. *BMC Bioinformatics*, 11:299, 2010.

13 Yi Li and Vasileios Nakos. Sublinear-time algorithms for compressive phase retrieval. *arXiv preprint arXiv:1709.02917*, 2017.

14 Dmitry M. Malioutov, Sujay Sanghavi, and Alan S. Willsky. Compressed sensing with sequential observations. In *International Conference on Acoustics, Speech and Signal Processing*, pages 3357–3360, 2008.

15 Tom Morgan and Jelani Nelson. A note on reductions between compressed sensing guarantees. *CoRR*, abs/1606.00757, 2016.

16 Shanmugavelayutham Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2):117–236, 2005.

17 Eric Price and David P. Woodruff. (1+eps)-approximate sparse recovery. In *IEEE Symposium on Foundations of Computer Science*, pages 295–304, 2011.

18 Eric Price and David P. Woodruff. Lower bounds for adaptive sparse recovery. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 652–663, 2013.

19 Noam Shental, Amnon Amir, and Or Zuk. Rare-allele detection using compressed se(que)nsing. *CoRR*, abs/0909.0400, 2009.

20 Tasuku Soma and Yuichi Yoshida. Non-convex compressed sensing with the sum-of-squares method. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 570–579, 2016.