Restricted Isometry Property for General p-Norms*

Zeyuan Allen-Zhu, Rati Gelashvili, and Ilya Razenshteyn

MIT CSAIL, Cambridge, MA, USA
`{zeyuan,gelash,ilyaraz}@csail.mit.edu`

Abstract

The Restricted Isometry Property (RIP) is a fundamental property of a matrix which enables sparse recovery. Informally, an $m \times n$ matrix satisfies RIP of order k for the ℓ_p norm, if $\|Ax\|_p \approx \|x\|_p$ for every vector x with at most k non-zero coordinates.

For every $1 \leq p < \infty$ we obtain almost tight bounds on the minimum number of rows m necessary for the RIP property to hold. Prior to this work, only the cases $p = 1$, $1 + 1/\log k$, and 2 were studied. Interestingly, our results show that the case $p = 2$ is a "singularity" point: the optimal number of rows m is $\widetilde{\Theta}(k^p)$ for all $p \in [1, \infty) \setminus \{2\}$, as opposed to $\widetilde{\Theta}(k)$ for $k = 2$.

We also obtain almost tight bounds for the column sparsity of RIP matrices and discuss implications of our results for the Stable Sparse Recovery problem.

1998 ACM Subject Classification F.2 Analysis of Algorithms and Problem Complexity, G.3 Probability and Statistics

Keywords and phrases compressive sensing, dimension reduction, linear algebra, high-dimensional geometry

Digital Object Identifier 10.4230/LIPIcs.SOCG.2015.451

1 Introduction

The main object of our interest is a matrix with *Restricted Isometry Property for the ℓ_p norm* (RIP-p). Informally speaking, we are interested in a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ with $m \ll n$ that approximately preserves ℓ_p norms for *all* vectors that have only few non-zero coordinates.

More precisely, an $m \times n$ matrix $A \in \mathbb{R}^{m \times n}$ is said to have *(k, D)-RIP-p property* for sparsity $k \in [n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$, distortion $D > 1$, and the ℓ_p norm for $p \in [1, \infty)$, if for every vector $x \in \mathbb{R}^n$ with at most k non-zero coordinates one has

$$\|x\|_p \leq \|Ax\|_p \leq D \cdot \|x\|_p \ .$$

In this work we investigate the following question: given $p \in [1, \infty)$, $n \in \mathbb{N}$, $k \in [n]$, and $D > 1$,

What is the smallest $m \in \mathbb{N}$ so that there exists a (k, D)-RIP-p matrix $A \in \mathbb{R}^{m \times n}$?

Besides that, the following question arises naturally from the complexity of computing Ax:

What is the smallest column sparsity d for such a (k, D)-RIP-p matrix $A \in \mathbb{R}^{m \times n}$?

(Above, we denote by column sparsity the maximum number of non-zero entries in a column of A.)

* The full version of this paper can be found at `http://arxiv.org/abs/1407.2178` [2].

31st International Symposium on Computational Geometry (SoCG'15).
Editors: Lars Arge and János Pach; pp. 451–460
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1.1 Motivation

Why are RIP matrices important? RIP-2 matrices were introduced by Candès and Tao [7] for decoding a vector f from corrupted linear measurements $Bf + e$ under the assumption that the vector of errors e is sufficiently sparse (has only few non-zero entries). Later Candès, Romberg and Tao [6] used RIP-2 matrices to solve *the (Noisy) Stable Sparse Recovery* problem, which has since found numerous applications in areas such as compressive sensing of signals [6, 11], genetic data analysis [16], and data stream algorithms [19, 12].

The (noisy) stable sparse recovery problem is defined as follows. The input signal $x \in \mathbb{R}^n$ is assumed to be close to k-sparse, that is, to have most of the "mass" concentrated on k coordinates. The goal is to design a set of m linear measurements that can be represented as a single $m \times n$ matrix A such that, given a *noisy sketch* $y = Ax + e \in \mathbb{R}^m$, where $e \in \mathbb{R}^n$ is a noise vector, one can "approximately" recover x. Formally, the recovered vector $\widehat{x} \in \mathbb{R}^n$ is required to satisfy

$$\|x - \widehat{x}\|_p \leq C_1 \min_{k\text{-sparse } x^*} \|x - x^*\|_1 + C_2 \cdot \|e\|_p \tag{1.1}$$

for some $C_1, C_2 > 0$, $p \in [1, \infty)$, and $k \in [n]$.

(In order for (1.1) to be meaningful, we also require $\|A\|_p \leq 1$ – or equivalently, $\|Ax\|_p \leq \|x\|_p$ for all x – since otherwise, by scaling A up, the noise vector e will become negligible.)

We refer to (1.1) as the ℓ_p/ℓ_1 *guarantee*. The parameters of interest include: the number of measurements m, the column sparsity of the measurement matrix A, the approximation factors C_1, C_2 and the complexity of the recovery procedure.

Candès, Romberg and Tao [6] proved that if A is $(O(k), 1 + \varepsilon)$-RIP-2 for a sufficiently small $\varepsilon > 0$, then one can achieve the ℓ_2/ℓ_1 guarantee with $C_1 = O(k^{-1/2})$ and $C_2 = O(1)$ in polynomial time.

The $p = 1$ case was first studied by Berinde *et al.* [4]. They prove that if A is $(O(k), 1+\varepsilon)$-RIP-1 for a sufficiently small $\varepsilon > 0$ and has a certain additional property, then one can achieve the ℓ_1/ℓ_1 guarantee with $C_1 = O(1)$, $C_2 = O(1)$.

We note that *any* matrix A that allows the (noisy) stable sparse recovery with the ℓ_p/ℓ_1 guarantee *must have the* (k, C_2)*-RIP-p property.* For the proof see the full version.

Known constructions and limitations. Candès and Tao [7] proved that for every $\varepsilon > 0$, a matrix with $m = O(k \log(n/k)/\varepsilon^2)$ rows and n columns whose entries are sampled from i.i.d. Gaussians is $(k, 1 + \varepsilon)$-RIP-2 with high probability. Later, a simpler proof of the same result was discovered by Baraniuk *et al.* [3][1]. Berinde *et al.* [4] showed that a (scaled) *random sparse binary matrix* with $m = O(k \log(n/k)/\varepsilon^2)$ rows is $(k, 1 + \varepsilon)$-RIP-1 with high probability[2].

Since the number of measurements is very important in practice, it is natural to ask, how optimal is the dimension bound $m = O(k \log(n/k))$ that the above constructions achieve? The results of Do Ba *et al.* [10] and Candés [8] imply the lower bound $m = \Omega(k \log(n/k))$ for $(k, 1 + \varepsilon)$-RIP-p matrices for $p \in \{1, 2\}$, provided that $\varepsilon > 0$ is sufficiently small.

Another important parameter of a measurement matrix A is its *column sparsity*: the maximum number of non-zero entries in a single column of A. If A has column sparsity d, then we can perform multiplication $x \mapsto Ax$ in time $O(nd)$ as opposed to the naive $O(nm)$ bound. Moreover, for sparse matrices A, one can maintain the sketch $y = Ax$ very efficiently

[1] This proof has an advantage that it works for any subgaussian random variables, such as random ± 1's.
[2] In the same paper [4] it is observed that the same construction works for $p = 1 + 1/\log k$.

Table 1 Prior and new bounds on RIP-p matrices.

p	rows m	column sparsity d	references
1	$\Theta(k\log(n/k))$	$\Theta(\log(n/k))$	[4, 10, 20, 14]
$1+\frac{1}{\log k}$	$O(k\log(n/k))$	$O(\log(n/k))$	[4]
$(1,2)$	$\widetilde{\Theta}(k^p)$	$\widetilde{\Theta}(k^{p-1})$	this work
2	$\Theta(k\log(n/k))$	$\Theta(k\log(n/k))$	[7, 6, 8, 3, 10, 9, 23]
$(2,\infty)$	$\widetilde{\Theta}(k^p)$	$\widetilde{\Theta}(k^{p-1})$	this work

if we update x. Namely, if we set $x \leftarrow x + \alpha \cdot e_i$, where $\alpha \in \mathbb{R}$ and $e_i \in \mathbb{R}^n$ is a basis vector, then we can update y in time $O(d)$ instead of the naive bound $O(m)$.

The aforementioned constructions of RIP matrices exhibit very different behavior with respect to column sparsity. RIP-2 matrices obtained from random Gaussian matrices are obviously dense, whereas the construction of RIP-1 matrices of Berinde *et al.* [4] gives very small column sparsity $d = O(\log(n/k)/\varepsilon)$. It is known that in both cases the bounds on column sparsity are essentially tight.

Indeed, Nelson and Nguyễn showed [23] that any non-trivial column sparsity is impossible for RIP-2 matrices unless m is much larger than $O(k\log(n/k))$. Nachin showed [20] that any RIP-1 matrix with $O(k\log(n/k))$ rows must have column sparsity $\Omega(\log(n/k))$. Besides that, Indyk and Razenshteyn showed [14] that every RIP-1 matrix 'must be sparse': any RIP-1 matrix with $O(k\log(n/k))$ rows can be perturbed slightly and made $O(\log(n/k))$-sparse.

Another notable difference between RIP-1 and RIP-2 matrices is the following. The construction of Berinde *et al.* [4] provides RIP-1 matrices with non-negative entries, whereas Chandar proved [9] that any RIP-2 matrix with non-negative entries must have $m = \Omega(k^2)$ (and this was later improved to $m = \Omega(k^2\log(n/k))$ [23, 1]). In other words, negative signs are crucial in the construction of RIP-2 matrices but not for the RIP-1 case.

1.2 Our results

Motivated by these discrepancies between the optimal constructions for RIP-p matrices with $p \in \left\{1, 1+\frac{1}{\log k}, 2\right\}$, we initiate the study of RIP-p matrices for the general $p \in [1,\infty)$.

Having in mind that the upper bound $m = O(k\log(n/k))$ holds for RIP-p matrices with $p \in \left\{1, 1+\frac{1}{\log k}, 2\right\}$, it would be natural to conjecture that the same bound holds at least for every $p \in (1,2)$. As we will see, surprisingly, this conjecture is very far from being true.

Also, knowing that the column sparsity $d = O(k\log(n/k))$ can be obtained for $p = 2$ while $d = O(\log(n/k))$ can be obtained for $p = 1$, it is interesting to "interpolate" these two bounds.

Besides the mathematical interest, a more "applied" reason to study RIP-p matrices for the general p is to get new guarantees for the stable sparse recovery. Indeed, we obtain new results in this direction.

Our upper bounds. On the positive side, for all $\varepsilon > 0$ and all $p \in (1,\infty)$, we construct $(k, 1+\varepsilon)$-RIP-p matrices with $m = \widetilde{O}(k^p)$ rows. Here, we use the $\widetilde{O}(\cdot)$-notation to hide factors that depend on ε, p, and are polynomial in $\log n$. More precisely, we show that a (scaled) *random sparse* $0/1$ *matrix* with $\widetilde{O}(k^p)$ rows and column sparsity $\widetilde{O}(k^{p-1})$ has the desired RIP property with high probability.

This construction essentially matches that of Berinde *et al.* [4] when p approaches 1. At the same time, when $p = 2$, our result matches known constructions of non-negative RIP-2 matrices based on the incoherence argument.[3]

Our lower bounds. Surprisingly, we show that, despite our upper bounds being suboptimal for $p = 2$, the are essentially tight for every constant $p \in (1, \infty)$ except 2. Namely, they are optimal both in terms of the dimension m and the column sparsity d.

More formally, on the dimension side, for every $p \in (1, \infty) \setminus \{2\}$, distortion $D > 1$, and (k, D)-RIP-p matrix $A \in \mathbb{R}^{m\times n}$, we show that $m = \Omega(k^p)$, where $\Omega(\cdot)$ hides factors that depend on p and D. Note that, it is not hard to extend an argument of Chandar [9] and obtain a lower bound $m = \Omega(k^{p-1})$.[4] This additional factor k is exactly what makes our lower bound non-trivial and tight for $p \in (1, \infty) \setminus \{2\}$, and thus enables us to conclude that $p = 2$ is a "singularity".[5]

As for the column sparsity, we present a simple extension of the argument of Chandar [9] and prove that for every $p \in [1, \infty)$ any (k, D)-RIP-p matrix must have column sparsity $\Omega(k^{p-1})$.

RIP matrices and sparse recovery. We extend the result of Candès, Romberg and Tao [6] to show that, for every $p > 1$, RIP-p matrices allow the stable sparse recovery with the ℓ_p/ℓ_1 guarantee and approximation factors $C_1 = O\big(k^{-1+1/p}\big)$, $C_2 = O(1)$ in polynomial time. This extension is quite straightforward and seems to be folklore, but, to the best of our knowledge, it is not recorded anywhere.

On the other hand, for every $p \geq 1$, it is almost immediate that *any* matrix A that allows the stable sparse recovery with the ℓ_p/ℓ_1 guarantee – even if it works only for k-sparse signals – *must have the* (k, C_2)*-RIP-*p *property.* For the sake of completeness, we have included both the above proofs in the full version.

Implications to sparse recovery. Using the above equivalent relationship between the stable sparse recovery problem and the RIP-p matrices, we conclude that the stable sparse recovery with the ℓ_p/ℓ_1 guarantee requires $m = \widetilde{\Theta}(k^p)$ measurements for every $p \in [1; \infty) \setminus \{2\}$, and requires $d = \tilde{\Theta}(k^{p-1})$ column sparsity for every $p \in [1, \infty)$. Our results together draw tradeoffs between the following three parameters in stable sparse recovery:

- p, the ℓ_p/ℓ_1 guarantee for the stable sparse recovery,[6]
- m, the number of measurements needed for sketching, and
- d, the running time (per input coordinate) needed for sketching.

It was pointed out by an anonymous referee that for the *noiseless* case – that is, when the noise vector e is always zero – better upper bounds are possible. Using the result of Gilbert *et al.* [13], one can obtain, for every $p \geq 2$, the noiseless stable sparse recovery procedure

---

[3] That is, a (scaled) random $m \times n$ binary matrix with $m = O(\varepsilon^{-2}k^2 \log(n/k))$ rows and sparsity $d = O(\varepsilon^{-1}k \log(n/k))$ satisfies the $(k, 1+\varepsilon)$-RIP-2 property. This can be proved using for instance the incoherence argument from [24]: any incoherent matrix satisfies the RIP-2 property with certain parameters.

[4] Also, the same argument gives the lower bound $\Omega(k^p)$ for *binary* RIP-p matrices for every $p \in [1, \infty)$.

[5] A similar singularity is known to exist for linear dimension reduction for arbitrary point sets with respect to ℓ_p norms [18]; alas, tight bounds for that problem are not known.

[6] We note that the ℓ_p/ℓ_1 and the ℓ_q/ℓ_1 guarantees are incomparable. However, it is often more desirable to have larger p in this ℓ_p/ℓ_1 guarantee to ensure a better recovery quality. This is because, if the noise vector $e = 0$, the ℓ_q/ℓ_1 guarantee (with $C_1 = O(k^{-1+1/q})$) can be shown to be stronger than the ℓ_p/ℓ_1 one (with $C_1 = O(k^{-1+1/p})$) whenever $q > p$. However, when there is a noise term, the guarantee $\|x - \hat{x}\|_p \leq O(1) \cdot \|e\|_p$ is incomparable to $\|x - \hat{x}\|_q \leq O(1) \cdot \|e\|_q$ for $p \neq q$.

with the ℓ_p/ℓ_1 guarantee using only $m = \widetilde{O}(k^{2-2/p})$ measurements. Therefore, our results also imply a very large gap, both in terms of m and d, between the *noiseless* and the *noisy* stable sparse recovery problems.

2 Overview of the Proofs

2.1 Upper bounds

We construct RIP-p matrices as follows. Beginning with a zero matrix A with $m = \widetilde{O}(k^p)$ rows and n columns, independently for each column of A, we choose $d = \widetilde{O}(k^{p-1})$ out of m entries uniformly at random (without replacement), and assign the value $d^{-1/p}$ to those selected entries. For this construction, we have two very different analyses of its correctness: one works only for $p \geq 2$, and the other works only for $1 < p < 2$.

For $p \geq 2$, the most challenging part is to show that $\|Ax\|_p \leq (1+\varepsilon)\|x\|_p$ holds with high probability, for all k-sparse vectors x. We reduce this problem to a probabilistic question *similar in spirit* to the following "balls and bins" question. Consider n bins in which we throw n balls uniformly and independently. As a result, we get n numbers $X_1, X_2, \ldots, X_n$, where X_i is the number of balls falling into the i-th bin. We would like to upper bound the tail $\Pr[S \geq 1000 \cdot \mathrm{E}[S]]$ for the random variable $S = \sum_{i=1}^n X_i^{p-1}$. (Here, the constant 1000 can be replaced with any large enough one since we do not care about constant factors in this paper.) The first challenge is that X_i's are not independent. To deal with this issue we employ the notion of *negative association* of random variables introduced by Joag-Dev and Proschan [15]. The second problem is that the random variables X_i^{p-1} are heavy tailed: they have tails of the form $\Pr\left[X_i^{p-1} \geq t\right] \approx \exp(-t^{\frac{1}{p-1}})$, so the standard technique of bounding the moment-generating function does not work. Instead, we bound the high moments of S directly, which introduces certain technical challenges. Let us remark that sums of i.i.d. heavy-tailed variables were thoroughly studied by Nagaev [21, 22], but it seems that for the results in these papers the independence of summands is crucial.

One major reason the above approach fails to work for $1 < p < 2$ is that, in this range, even the best possible tail inequality for S is too weak for our purposes. Another challenge in this regime is that, to bound the "lower tail" of $\|Ax\|_p^p$ (that is, to prove that $\|Ax\|_p \geq (1-\varepsilon)\|x\|_p$ holds for all k-sparse x), the simple argument used for $p \geq 2$ no longer works. Our solution to both problems above is to instead build our RIP matrices based on the following general notion of bipartite expanders.

▶ **Definition 2.1.** Let $G = (U, V, E)$ with $|U| = n$, $|V| = m$ and $E \subseteq U \times V$ be a bipartite graph such that all vertices from U have the same degree d. We say that G is an (ℓ, d, δ)-*expander*, if for every $S \subseteq U$ with $|S| \leq \ell$ we have

$$\left|\{v \in V \mid \exists u \in S \ \ (u,v) \in E\}\right| \geq (1-\delta)d|S| \ .$$

It is known that random d-regular graphs are good expanders, and we can take the (scaled) adjacency matrix of such an expander and prove that it satisfies the desired RIP-p property for $1 < p < 2$. Our argument can be seen as a subtle interpolation between the argument from [4], which proves that (scaled) adjacency matrices of $(k, d, \Theta(\varepsilon))$-expanders (with $\widetilde{O}(k)$ rows) are $(k, 1+\varepsilon)$-RIP-1 and the one using incoherence argument,[7] which shows that $(2, d, \Theta(\varepsilon/k))$-expanders give $(k, 1+\varepsilon)$-RIP-2 matrices (with $\widetilde{O}(k^2)$ rows).

[7] It is known [24] that an incoherent matrix satisfies the RIP-2 property with certain parameters. At the same time, the notion of incoherence can be interpreted as expansion for $\ell = 2$.

2.2 Lower bounds

In the full version of our paper [2], we derive our dimension lower bound $m = \Omega(k^p)$ essentially from norm inequalities. The high-level idea can be described in four simple steps. Consider any (k, D)-RIP-p matrix $A \in \mathbb{R}^{n\times m}$, and assume that D is very close to 1 in this high-level description.

In the first three steps, we deduce from the RIP property that (a) the sum of the p-th powers of all entries in A is approximately n, (b) the largest entry in A (i.e., the vector ℓ_∞-norm of A) is essentially at most $k^{1/p-1}$, and (c) the sum of squares of all entries in A is at least $n\big(\frac{k}{m}\big)^{2/p-1}$ if $p \in (1, 2)$, or at most $n\big(\frac{k}{m}\big)^{2/p-1}$ if $p > 2$. In the fourth step, we combine (a) (b) and (c) together by arguing about the relationships between the ℓ_p, ℓ_∞ and ℓ_2 norms of entries of A, and prove the desired lower bound on m.

The sparsity lower bound $d = \Omega(k^{p-1})$ can be obtained via a simple extension of the argument of Chandar [9]. It is possible to extend the techniques of Nelson and Nguyễn [23] to obtain a slightly better sparsity lower bound. However, since we were unable to obtain *a tight* bound this way, we decided not to include it.

3 RIP Construction for $p \geq 2$

In this section, we construct $(k, 1 + \varepsilon)$-RIP-p matrices for $p \geq 2$ by proving the following theorem.

▶ **Definition 3.1.** We say that an $m \times n$ matrix A is a *random binary matrix with sparsity* $d \in [m]$, if A is generated by assigning $d^{-1/p}$ to d random entries per column (selected uniformly at random without replacement), and assigning 0 to the remaining entries.

▶ **Theorem 3.2.** *For all $n \in \mathbb{Z}_+$, $k \in [n]$, $\varepsilon \in (0, \frac{1}{2})$ and $p \in [2, \infty)$, there exist $m, d \in \mathbb{Z}_+$ with*

$$m = p^{O(p)} \cdot \frac{k^p}{\varepsilon^2} \cdot \log^{p-1} n \quad \textit{and} \quad d = p^{O(p)} \cdot \frac{k^{p-1}}{\varepsilon} \cdot \log^{p-1} n \leq m$$

such that, letting A be a random binary $m \times n$ matrix of sparsity d, with probability at least 98%, A satisfies $(1 - \varepsilon)\|x\|_p^p \leq \|Ax\|_p^p \leq (1 + \varepsilon)\|x\|_p^p$ for all k-sparse vectors $x \in \mathbb{R}^n$.

Our proof is divided into two steps: (1) the "lower-tail step", that is, with probability at least 0.99 we have $\|Ax\|_p^p \geq (1 - \varepsilon)\|x\|_p^p$ for all k-sparse x, and (2) the "upper-tail step", that is, with probability at least 0.99, we have $\|Ax\|_p^p \leq (1 + \varepsilon)\|x\|_p^p$.

For every $j \in [n]$, let us denote by $S_j \subseteq [m]$ the set of non-zero rows of the j-th column of A.

3.1 The Lower-Tail Step

To lower-tail step is very simple. It suffices to show that, with high probability, $|S_i \cap S_j|$ is small for every pair of different $i, j \in [n]$, which will then imply that if only k columns of A are considered, every S_i has to be almost disjoint from the union of the S_j of the $k - 1$ remaining columns. This can be summarized by the following claim, whose proof is deferred to the full version of this paper.

▶ **Claim 3.3.** *If $d \geq C\varepsilon^{-1}k \log n$ and $m \geq 2dk/\varepsilon$, where C is some large enough constant, then*

$$\Pr\left[\forall 1 \leq i < j \leq n \quad |S_i \cap S_j| \leq \frac{\varepsilon d}{k}\right] \geq 0.99 \ .$$

Now, to prove the lower tail, without loss of generality, let us assume that x is supported on $[k]$, the first k coordinates. For every $j \in [k]$, we denote by $S'_j = S_j \setminus \bigcup_{j' \in [k] \setminus \{j\}} S_{j'}$, the set of non-zero rows in column j that are not shared with the supports of other columns in $[k] \setminus \{j\}$. If the event in Claim 3.3 holds, then for every $j \in [k]$, we have $|S'_j| \geq (1-\varepsilon)d$. Thus, we can lower bound $\|Ax\|_p$ as

$$\|Ax\|_p^p = \frac{1}{d} \cdot \sum_{i=1}^{m} \Bigg| \sum_{j \in [k]: i \in S_j} x_j \Bigg|^p \geq \frac{1}{d} \cdot \sum_{i=1}^{m} \Bigg| \sum_{j \in [k]: i \in S'_j} x_j \Bigg|^p = \frac{1}{d} \cdot \sum_{j \in [k]} |S'_j| \cdot |x_j|^p \geq (1-\varepsilon)\|x\|_p^p \ . \tag{3.1}$$

▶ Remark. The above claim only works when $m = \Omega(k^2 \log n / \varepsilon^2)$, and therefore we cannot use it in for the case of $1 < p < 2$.

3.2 The Upper-Tail Step

Below we describe the framework of our proof for the upper-tail step, deferring all technical details to the full version of this paper.

Suppose again that x is supported on $[k]$. Then, we upper bound $\|Ax\|_p^p$ as

$$\begin{aligned}\|Ax\|_p^p = \frac{1}{d} \cdot \sum_{i=1}^{m} \Bigg| \sum_{j \in [k]: i \in S_j} x_j \Bigg|^p &\leq \frac{1}{d} \cdot \sum_{i=1}^{m} \big| \{j' \in [k] \mid i \in S_{j'}\} \big|^{p-1} \cdot \sum_{j \in [k]: i \in S_j} |x_j|^p \\ &= \frac{1}{d} \cdot \textstyle\sum_{j=1}^{k} |x_j|^p \cdot \sum_{i \in S_j} \big| \{j' \in [k] \mid i \in S_{j'}\} \big|^{p-1} \ , \end{aligned} \tag{3.2}$$

where the first inequality follows from the fact that $(a_1 + \cdots + a_N)^p \leq N^{p-1}(a_1^p + \cdots + a_N^p)$ for any sequence of N non-negative reals $a_1, \ldots, a_N$. Note that the quantity $\big| \{j' \in [k] \mid i \in S_{j'}\} \big| \in [k]$ captures the number of non-zeros of A in the i-th row and the first k columns. From now on, in order to prove the desired upper tail, it suffices to show that, with high probability

$$\forall j \in [k], \quad \textstyle\sum_{i \in S_j} \big| \{j' \in [k] \mid i \in S_{j'}\} \big|^{p-1} \leq (1+\varepsilon)d \ . \tag{3.3}$$

To prove this, let us fix some $j^* \in [k]$ and upper bound the probability that (3.3) holds for $j = j^*$, and then take a union bound over the choices of j^*. Without loss of generality, assume that $S_{j^*} = \{1, 2, \ldots, d\}$, consisting of the first d rows. For every $i \in S_{j^*}$, define a random variable $X_i \stackrel{\text{def}}{=} \big| \{j' \in [k] \mid i \in S_{j'}\} \big| - 1$. It is easy to see that X_i is distributed as $\mathrm{Bin}(k-1, d/m)$, the binomial distribution that is the sum of $k-1$ i.i.d. random 0/1 variables, each being 1 with probability d/m. For notational simplicity, let us define $\delta \stackrel{\text{def}}{=} dk/m$. We will later choose $\delta < \varepsilon$ to be very small. Our goal in (3.3) can now be reformulated as follows: upper bound the probability

$$\Pr\Big[\textstyle\sum_{i=1}^{d} ((X_i+1)^{p-1} - 1) > \varepsilon d \Big] \ .$$

We begin with a lemma showing an upper bound on the moments of each $Y_i \stackrel{\text{def}}{=} (X_i + 1)^{p-1} - 1$.

▶ **Lemma 3.4.** *There exists a constant $C \geq 1$ such that, if X is drawn from the binomial distribution $\mathrm{Bin}(k-1, \delta/k)$ for some $\delta < 1/(2e^2)$, and $p \geq 2$, then for any real $\ell \geq 1$,*

$$\mathbb{E}[((X+1)^{p-1} - 1)^{\ell}] \leq C \cdot \delta (\ell(p-1)+1)^{\ell(p-1)+1} \ .$$

Next, we note that although the random variables X_i's are dependent, they can be verified to be *negatively associated*, a notion introduced by Joag-Dev and Proschan [15]. This theory allows us to conclude the following bound on the moments.

▶ **Lemma 3.5.** *Let $\widetilde{X}_1, \dots, \widetilde{X}_d$ be d random variables, each drawn* independently *from* $\mathrm{Bin}(k-1, \delta/k)$. *Then, for every integer $t \geq 1$ we have*

$$\mathbb{E}\left[\left(\sum_{i=1}^d ((X_i+1)^{p-1}-1)\right)^t\right] \leq \mathbb{E}\left[\left(\sum_{i=1}^d ((\widetilde{X}_i+1)^{p-1}-1)\right)^t\right] .$$

Now, using the moments of random variables $Y_i = (X_i+1)^{p-1}-1$ from Lemma 3.4, as well as Lemma 3.5, we can compute the tail bound of the sum $\sum_{i=1}^d Y_i$. Our proof of the following Lemma uses the result of Latała [17].

▶ **Lemma 3.6.** *There exists constants $C \geq 1$ such that, whenever $\delta \leq \varepsilon/p^{Cp}$ and $d \geq p^{Cp}/\varepsilon$, we have*

$$\Pr\left[\sum_{i=1}^d ((X_i+1)^{p-1}-1) > \varepsilon d\right] \leq e^{-\Omega(\frac{(\varepsilon d)^{1/(p-1)}}{p})} .$$

Finally, we are ready to prove Theorem 3.2.

Proof of Theorem 3.2. We can choose $d = \Theta(p)^{p-1} \cdot \frac{k^{p-1}}{\varepsilon} \cdot \log^{p-1} n$ so that $e^{-\Omega(\frac{(\varepsilon d)^{1/(p-1)}}{p})} < \frac{1}{100} \frac{1}{k\binom{n}{k}}$. Since our choice of $m = \frac{dkp^{\Theta(p)}}{\varepsilon}$ ensures that $\delta = dk/m \leq \varepsilon/p^{Cp}$, and our choice of d ensures $d \geq p^{Cp}/\varepsilon$, we can apply Lemma 3.6 and conclude that with probability at least $1 - \frac{1}{100}\frac{1}{k\binom{n}{k}}$ one has

$$\sum_{i \in S_{j^*}} \left| \{j' \in [k] \mid i \in S_{j'}\} \right|^{p-1} = \sum_{i=1}^d (X_i+1)^{p-1} \leq (1+\varepsilon)d .$$

Therefore, by applying the union bound over all $j^* \in [k]$, we conclude that with probability at least $1 - \frac{1}{100}\frac{1}{\binom{n}{k}}$, the desired inequality (3.3) is satisfied for all $j \in [k]$.

Recall that, owing to (3.2), the inequality (3.3) implies that $\|Ax\|_p^p \leq (1+\varepsilon)\|x\|_p^p$ for every $x \in \mathbb{R}^n$ that is supported on the *first* k coordinates. By another union bound over the choices of all possible $\binom{n}{k}$ subsets of $[n]$, we conclude that with probability at least 0.99, we have $\|Ax\|_p^p \leq (1+\varepsilon)\|x\|_p^p$ for all k-sparse vectors x.

On the other hand, since our choice of d and m satisfies the assumptions $d \geq \Omega(k \log n/\varepsilon)$ and $m \geq 2dk/\varepsilon$ in Claim 3.3, the lower tail $\|Ax\|_p^p \geq (1-\varepsilon)\|x\|_p^p$ also holds with probability at least 0.99. Overall we conclude that with probability at least 0.98, we have $\|Ax\|_p^p \in (1 \pm \varepsilon)\|x\|_p^p$ for every k-sparse vector $x \in \mathbb{R}^n$. ◀

4 RIP Construction for $1 < p < 2$

In this section, we construct $(k, 1+\varepsilon)$-RIP-p matrices for $1 < p < 2$ by proving the following theorem.

We assume that $1 + \tau \leq p \leq 2 - \tau$ for some $\tau > 0$, and whenever we write $O_\tau(\cdot)$, we assume that some factor that depends on τ is hidden. (For instance, factors of $p/(1-p)$ may be hidden.)

▶ **Theorem 4.1.** *For every $n \in \mathbb{Z}_+$, $k \in [n]$, $0 < \varepsilon < 1/2$ and $1 + \tau \leq p \leq 2 - \tau$, there exist $m, d \in \mathbb{Z}_+$ with*

$$m = O_\tau\left(k^p \frac{\log n}{\varepsilon^2} + k^{4-2/p-p} \frac{\log n}{\varepsilon^{2/(p-1)}}\right) \quad \textit{and} \quad d = O_\tau\left(\frac{k^{p-1} \cdot \log n}{\varepsilon} + \frac{k^{(p-1)/p} \cdot \log n}{\varepsilon^{1/(p-1)}}\right)$$

such that, letting A be a random binary $m \times n$ matrix of sparsity d, with probability at least 98%, A satisfies $(1-\varepsilon)\|x\|_p^p \le \|Ax\|_p^p \le (1+\varepsilon)\|x\|_p^p$ for all k-sparse vectors $x \in \mathbb{R}^n$.

Note that, when $k \ge \varepsilon^{-\frac{p(2-p)}{(p-1)^3}}$, the above bounds on m and k can be simplified as

$$m = O_\tau\Big(\frac{k^p \cdot \log n}{\varepsilon^2}\Big) \quad \text{and} \quad d = O_\tau\Big(\frac{k^{p-1} \cdot \log n}{\varepsilon}\Big) \ .$$

Our proof of the above theorem is based on the existence of (ℓ, d, δ) bipartite expanders (recall the definition of such expanders from Definition 2.1):

▶ **Lemma 4.2** ([5, Lemma 3.10]). *For every $\delta \in (0, \frac{1}{2})$, and $\ell \in [n]$, there exist (ℓ, d, δ)-expanders with $d = O\big(\frac{\log n}{\delta}\big)$ and $m = O(dl/\delta) = O\big(\frac{\ell \log n}{\delta^2}\big)$.*

In fact, the proof of Lemma 4.2 implies a simple probabilistic construction of such expanders: with probability at least 98%, a random binary matrix A of sparsity d is the adjacency matrix of a $(2\ell, d, \delta)$-expander scaled by $d^{-1/p}$, for $\delta = \Theta(\frac{\log n}{d})$ and $\ell = \Theta(\frac{\delta m}{d})$.

In the full version of this paper [2] we argue that, when A is the (scaled) adjacency matrix of a $(2\ell, d, \delta)$-expander, for parameters choices $\ell = \Theta_\tau(k^{2-p})$ and $\delta = \Theta_\tau\big(\min\big\{\frac{\varepsilon}{k^{p-1}}, \frac{\varepsilon^{1/(p-1)}}{k^{(p-1)/p}}\big\}\big)$, it satisfies that $\|Ax\|_p^p = 1 \pm \varepsilon$. This proof is very technical, but we have included a high-level description of its idea in the full version of this paper.

It is perhaps interesting to be noted that, our construction confirms our description in the introduction: it interpolates between the expander construction of RIP-1 matrices from [4] that uses $\ell = k$, and the construction of RIP-2 matrices using incoherence argument that essentially corresponds to $\ell = 2$.

Acknowledgments. We thank Piotr Indyk for encouraging us to work on this project and for many valuable conversations. We are grateful to Piotr Indyk and Ronitt Rubinfeld for teaching "Sublinear Algorithms", where parts of this work appeared as a final project. We thank Artūrs Bačkurs, Chinmay Hegde, Gautam Kamath, Sepideh Mahabadi, Jelani Nelson, Huy Nguyễn, Eric Price and Ludwig Schmidt for useful conversations and feedback. Thanks to Leonid Boytsov for pointing us to [21, 22]. We are grateful to anonymous referees for pointing out some relevant literature. The first author is partly supported by a Simons Graduate Student Award under grant no. 284059.

References

1 Zeyuan Allen-Zhu, Rati Gelashvili, Silvio Micali, and Nir Shavit. Johnson-Lindenstrauss Compression with Neuroscience-Based Constraints. *ArXiv e-prints*, abs/1411.5383, November 2014. Also appeared in the Proceedings of the National Academy of Sciences of the USA, vol 111, no 47.

2 Zeyuan Allen-Zhu, Rati Gelashvili, and Ilya Razenshteyn. Restricted Isometry Property for General p-Norms. *ArXiv e-prints*, abs/1407.2178v3, February 2015.

3 Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

4 Radu Berinde, Anna C. Gilbert, Piotr Indyk, Howard Karloff, and Martin J. Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. In *Proceedings of the 46th Annual Allerton Conference on Communication, Control, and Computing (Allerton 2008)*, pages 798–805, 2008.

5 Harry Buhrman, Peter Bro Miltersen, Jaikumar Radhakrishnan, and Srinivasan Venkatesh. Are bitvectors optimal? *SIAM Journal on Computing*, 31(6):1723–1744, 2002.

6 Emmanuel Candès, Justin Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

7 Emmanuel Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

8 Emmanuel J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9–10):589–592, 2008.

9 Venkat B. Chandar. *Sparse Graph Codes for Compression, Sensing, and Secrecy.* PhD thesis, Massachusetts Institute of Technology, 2010.

10 Khanh Do Ba, Piotr Indyk, Eric Price, and David P. Woodruff. Lower bounds for sparse recovery. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'10)*, pages 1190–1197, 2010.

11 David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

12 Anna C. Gilbert and Piotr Indyk. Sparse recovery using sparse matrices. *Proceedings of IEEE*, 98(6):937–947, 2010.

13 Anna C. Gilbert, Martin J. Strauss, Joel A. Tropp, and Roman Vershynin. One sketch for all: fast algorithms for compressed sensing. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC 2007)*, pages 237–246, 2007.

14 Piotr Indyk and Ilya Razenshteyn. On model-based RIP-1 matrices. In *Proceedings of the 40th International Colloquium on Automata, Languages, and Programming (ICALP'13)*, pages 564–575, 2013.

15 Kumar Joag-Dev and Frank Proschan. Negative association of random variables with applications. *Annals of Statistics*, 11(1):286–295, 1983.

16 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.

17 Rafał Latała. Estimation of moments of sums of independent real random variables. *Annals of Probability*, 25(3):1502–1513, 1997.

18 James R. Lee, Manor Mendel, and Assaf Naor. Metric structures in L_1: dimension, snowflakes, and average distortion. *European Journal of Combinatorics*, 26(8):1180–1190, 2005.

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

20 Mergen Nachin. Lower bounds on the column sparsity of sparse recovery matrices. undergraduate thesis, MIT, 2010.

21 A.V. Nagaev. Integral limit theorems taking large deviations into account when Cramér's condition does not hold. I. *Theory of Probability and Its Applications*, 14(1):51–64, 1969.

22 A.V. Nagaev. Integral limit theorems taking large deviations into account when Cramér's condition does not hold. II. *Theory of Probability and Its Applications*, 14(2):193–208, 1969.

23 Jelani Nelson and Huy L. Nguyễn. Sparsity lower bounds for dimensionality reducing maps. In *Proceedings of the 45th ACM Symposium on the Theory of Computing (STOC'13)*, pages 101–110, 2013.

24 Holger Rauhut. Compressive sensing and structured random matrices. *Theoretical foundations and numerical methods for sparse recovery*, 9:1–92, 2010.