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 $\ell_p$ 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 $\tilde{\Theta}(k^p)$ for all $p \in [1, \infty) \setminus \{2\}$, as opposed to $\tilde{\Theta}(k)$ 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.

1 Introduction

The main object of our interest is a matrix with Restricted Isometry Property for the $\ell_p$ 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 $\ell_p$ 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 $k \in [n] \equiv \{1, \ldots, n\}$, distortion $D > 1$, and the $\ell_p$ 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 $A$.)
11 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].

In the same paper [4] it is observed that the same construction works for any subgaussian random variables, such as random \( \mathcal{N}(0,1) \) Gaussians is \( O(k \log(n/k)/\varepsilon^2) \) rows of \( (k, 1 + \varepsilon) \)-RIP-1 with high probability. 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 Candès [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 \( \pm 1 \)'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.

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

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 Nguyêń 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 $p \in \{1, 1 + \frac{1}{\log k}, 2\}$, 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 $p \in \{1, 1 + \frac{1}{\log k}, 2\}$, 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 $m = \tilde{O}(k^p)$ rows. Here, we use the $\tilde{O}(\cdot)$-notation to hide factors that depend on $\varepsilon$, $p$, and are polynomial in $\log n$. More precisely, we show that a (scaled) random sparse $0/1$ matrix with $\tilde{O}(k^p)$ rows and column sparsity $\tilde{O}(k^{p-1})$ 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.
matrices based on the incoherence argument.\footnote{That is, a (scaled) random $m \times n$ binary matrix with $m = O(\varepsilon^{-2} k^2 \log(n/k))$ rows and sparsity $d = O(\varepsilon^{-1} \log(n/k))$ 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.}

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 $m \in (1, \infty) \setminus \{2\}$, 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 $m = \Omega(k^{p-1})$.\footnote{Also, the same argument gives the lower bound $\Omega(k^p)$ for binary RIP-$p$ matrices for every $p \in [1, \infty)$.} 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”\footnote{A similar singularity is known to exist for linear dimension reduction for arbitrary point sets with $\ell_2$-RIP matrices and sparse recovery. 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 $\ell_p/\ell_1$ guarantee requires $m = \tilde{\Theta}(k^p)$ measurements for every $p \in [1; \infty) \setminus \{2\}$, and requires $d = \Theta(k^{p-1})$ column sparsity for every $p \in [1, \infty)$. Our results together draw tradeoffs between the following three parameters in stable sparse recovery: \begin{itemize} \item $p$, the $\ell_p/\ell_1$ guarantee for the stable sparse recovery,\footnote{We note that the $\ell_p/\ell_1$ and the $\ell_q/\ell_1$ guarantees are incomparable. However, it is often more desirable to have larger $p$ in this $\ell_p/\ell_1$ guarantee to ensure a better recovery quality. This is because, if the noise vector $e = 0$, the $\ell_q/\ell_1$ guarantee (with $C_1 = O(k^{1+1/q})$) can be shown to be stronger than the $\ell_p/\ell_1$ 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 $p \neq q$.} \item $m$, the number of measurements needed for sketching, and \item $d$, the running time (per input coordinate) needed for sketching. \end{itemize} It was pointed out by an anonymous referee that for the noiseless case – that is, when the noise vector 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

\begin{itemize} \item for $p = 2$, the noiseless stable sparse recovery procedure

\begin{itemize} \item for $p \neq 2$, the noiseless stable sparse recovery procedure

\begin{itemize} \item for $p > 2$, the noiseless stable sparse recovery procedure

\begin{itemize} \item for $p < 2$, the noiseless stable sparse recovery procedure

\end{itemize} \end{itemize} \end{itemize}
with the $\ell_p/\ell_1$ guarantee using only $m = \tilde{O}(k^{2-2/p})$ 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 $m = \tilde{O}(k^p)$ rows and $n$ columns, independently for each column of $A$, we choose $d = \tilde{O}(k^{p-1})$ 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 $\Pr[S \leq 1000 \cdot E[S]]$ 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 $\Pr[X_i^{p-1} \geq t] \approx \exp(-t^{\frac{1}{p-1}})$, 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 $(\ell, d, \delta)$-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 $\tilde{O}(k)$ 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 $\tilde{O}(k^2)$ 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 $\ell = 2$. 

\(\text{SoCG’15}\)
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 \( \ell_\infty \)-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 \( n ( \frac{k}{m})^{2/p-1} \) if \( p \in (1, 2) \), or at most \( n ( \frac{k}{m})^{2/p-1} \) if \( p > 2 \). In the fourth step, we combine (a) (b) and (c) together by arguing about the relationships between the \( \ell_p \), \( \ell_\infty \) and \( \ell_2 \) 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 Nguyen [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
  
  \[
  m = \rho \left( \frac{p}{\varepsilon^2} \right) \cdot \frac{k^p}{\varepsilon} \cdot \log k^{-1} n \quad \text{and} \quad d = \rho \left( \frac{p}{\varepsilon} \right) \cdot \frac{k^{p-1}}{\varepsilon} \cdot \log k^{-1} n \leq m
  \]

  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

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

(3.1)

\begin{itemize}
  \item \textbf{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 $1 < p < 2$.
\end{itemize}

### 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 $\|Ax\|_p$ as

$$\|Ax\|_p^p = \frac{1}{d} \sum_{i=1}^m \left| \sum_{j \in [k]: i \in S_j} x_j \right|^p \leq \frac{1}{d} \sum_{i=1}^m \left| \{j' \in [k] \mid i \in S_{j'} \} \right|^{p-1} \cdot \sum_{j' \in [k]: i \in S_{j'}} |x_j|^p$$

$$= \frac{1}{d} \cdot \sum_{j=1}^k |S_j| \cdot \sum_{i \in S_j} \left| \{j' \in [k] \mid i \in S_{j'} \} \right|^{p-1},$$  

(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 $\left| \{j' \in [k] \mid i \in S_{j'} \} \right|$ 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

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

(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 $X_i \overset{\text{def}}{=} \left| \{j' \in [k] \mid i \in S_{j'} \} \right| - 1$. It is easy to see that $X_i$ is distributed as $\text{Bin}(k-1, d/m)$, 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 $\delta \overset{\text{def}}{=} dk/m$. 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

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

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

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

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

\end{itemize}
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.

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

Finally, we are ready to prove Theorem 3.2.

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

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

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

Recall that, owing to (3.2), the inequality (3.3) implies that \( \|Ax\|_p \leq (1 + \varepsilon)\|x\|_p \) 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 \( \|Ax\|_p \leq (1 + \varepsilon)\|x\|_p \) 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 \( \|Ax\|_p \geq (1 - \varepsilon)\|x\|_p \) also holds with probability at least 0.99. Overall we conclude that with probability at least 0.98, we have \( \|Ax\|_p \in (1 \pm \varepsilon)\|x\|_p \) for every \( k \)-sparse vector \( x \in \mathbb{R}^n. \)
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 $x \in \mathbb{R}^n$.

Note that, when $k \geq \varepsilon^{-\frac{\delta^2}{\log n}}$, the above bounds on $m$ and $k$ can be simplified as

$$m = O\left(\frac{k^p \cdot \log n}{\varepsilon^a}\right)$$

and

$$d = O\left(\frac{k^{p-1} \cdot \log n}{\varepsilon}\right).$$

Our proof of the above theorem is based on the existence of $(\ell, d, \delta)$ bipartite expanders (recall the definition of such expanders from Definition 2.1):

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

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 $\delta = \Theta\left(\frac{\log n}{d}\right)$ and $\ell = \Theta\left(\frac{\log n}{\delta}\right)$.

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 $\ell = \Theta\left(k^{2\gamma-}\right)$ and $\delta = \Theta\left(\min\left\{\frac{\varepsilon}{\ell^{2/\gamma}}, \frac{1}{\ell^{\gamma/p}}\right\}\right)$, 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 Nguyôn, 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**


**Restricted Isometry Property for General p-Norms**