Lower Bounds on the Column Sparsity of Compressed Sensing Matrices

Mergen Nachin (mergen@mit.edu)
Supervisor: Professor Piotr Indyk (indyk@mit.edu)

Abstract

It was recently shown that the following algorithms [BGI+08], [BI09] can approximately recover n-dimensional signal x from its *sketch* Ax, where the sketch length is $O(k \log(n/k))$ and the column sparsity of A is $O(\log(n/k))$. Our main goal in this report is to show that this column sparsity bound is tight when A is an $m \times n$ matrix with $m = \Theta(k \log(n/k))$.

1 Introduction

How can we maintain a succinct representation of a signal and still be able to efficiently reconstruct the signal to a specified precision? In the past few years, this question has developed a whole new area of research called *Compressive Sensing*[RCS] which has been attracting many different communities, including theoretical computer science, applied mathematics, image processing and digital signal processing. For any signal x of length n, the *sketch* is defined to be Ax where the length of Ax is much shorter than the original signal x. For a carefully chosen matrix A, the low dimensional vector Ax contains plenty of useful information about x. Obviously, one cannot hope to find an *exact* solution to an under-determined system of equations. But instead, we wish to find a signal $\hat{x}$ such that approximation error $||x - \hat{x}||_p$ is small.

Linear sketches may be motivated by the following two examples. The first one comes from the *digital signal processing* [WLD+06], [LKM+06] area. One wants to sense a signal x and traditional approaches first capture the

entire signal and then process it. But the new approach senses the signal in a way that it approximately computes a dot product with a pre-specified measurement vector at unit cost. As one can see, our linear sketch model can very well capture this scenario. The next example comes from the *data streaming* [Ind07], [Mut03] area. One often wants to support linear updates of the high dimensional signal x and still maintain a low dimensional sketch: this can be easily done, since $A(cx+y) = cAx+Ay$. For incremental updates, it is often crucial that the matrix A is *sparse*, i.e. contains a few nonzero elements *per each column*. The time needed to update the sketch Ax under $x' = x + e_i$ is proportional to the number of nonzero elements on ith column, since $Ax' = Ax + Ae_i$.

Formally, the (p, q, k)-**stable sparse recovery problem** is defined as follows: Given the sketch Ax, we wish to find a vector $\hat{x}$ such that the approximation error $||x - \hat{x}||_p$ satisfies:

$$||x - \hat{x}||_p \leq C \cdot Err_k^q(x) \tag{1}$$

$Err_k^q(x)$ is defined to be the smallest ℓ_q approximation error $||x - \acute{x}||_q$ where $\acute{x}$ ranges over all k-*sparse* vectors (i.e, that have at most k non-zero entries). It is easy to see that $||x - \acute{x}||_q$ is minimized when $\acute{x}$ consists of the k largest (in magnitude) coefficients of x.

It was recently shown [BIPW10] that any algorithm that solves the stable sparse recovery problem must require the sketch length to be $\Omega(k \log(n/k))$ when the length of the original signal x is n. In fact, this bound is tight; it is known [BGI+08], [BI09] that there exist matrices A and associated recovery algorithm that solves the problem with sketch length $O(k \log(n/k))$ and column sparsity (i.e. number of nonzero elements per each column) $O(\log(n/k))$. For a survey of sparse recovery using sparse matrices, see [GI10].

Our goal in this paper is to study the column sparsity of a matrix when it has the optimal dimension, that is when the sketch length is $\Theta(k \log(n/k))$. Roughly speaking, **our main result is the following**: any deterministic algorithm that solves the stable spares recovery problem must require the matrix A to have at least $\Omega(\log(n/k))$ nonzero entries per each column provided that A is an $m \times n$ matrix, where $m = \Theta(k \log(n/k))$. As mentioned earlier, *column sparsity* plays an important role in data stream model; the time needed for an incremental update is proportional to column sparsity. In the next subsection, we will formally give definitions and precisely state the results.

1.1 Formal definitions and results

Definition 1. We say a matrix A has *column sparsity* s if the number of non-zero entries is at most s per each column.

Definition 2. We define a C-approximate *deterministic* ℓ_1/ℓ_1 recovery algorithm $\mathbf{D}$ to be a pair $(A, \mathcal{A})$ where A is an $m \times n$ matrix and $\mathcal{A}$ is a deterministic algorithm that, for any x, maps Ax to some $\hat{x}$ that satisfies Equation (1) for $p = q = 1$.

Definition 3. We define a C-approximate *randomized* ℓ_1/ℓ_1 recovery algorithm $\mathbf{R}$ to be a pair $(A, \mathcal{A})$ where A is an $m \times n$ matrix chosen from some probability distribution and $\mathcal{A}$ is a deterministic algorithm that, for any x, maps (Ax, A) to some $\hat{x}$ that satisfies Equation (1) with probability at least $3/4$ for $p = q = 1$.

Theorem 1. *Let $(A, \mathcal{A})$ be a C-approximate deterministic ℓ_1/ℓ_1 recovery algorithm, where the matrix A is a real $m \times n$ with $m = \Theta(k \log(n/k))$, and A has column sparsity s. Then $s = \Omega(\log(n/k)$.*

Definition 4. We define a *deterministic* k-sparse *exact* recovery algorithm $\mathbf{D}$ to be a pair $(A, \mathcal{A})$ where A is an $m \times n$ matrix and $\mathcal{A}$ is a deterministic algorithm that exactly recovers x from Ax, for all x with at most k nonzero entries. The *randomized* algorithm is also defined analogously.

At this point, a reader might wonder why we introduced the definition 4. As mentioned earlier, for any q, $||x - \acute{x}||_q$ is minimized when $\acute{x}$ consists of the k largest (in magnitude) coefficients of x, and therefore, $Err_k^q(x)$ is 0 when x has at most k nonzero entries. Hence, any C-approximate ℓ_1/ℓ_1 recovery algorithm is also an *exact* k-sparse recovery algorithm. We prove the following stronger results for zero-one matrices:

Theorem 2. *Let $(A, \mathcal{A})$ be a deterministic exact recovery algorithm, where the matrix A is in $\{0,1\}^{m \times n}$ with $m = \Theta(k \log(n/k))$, and A has column sparsity s. Then $s = \Omega(\log(n/k)$.*

We also have the same bound for the randomized case.

Theorem 3. *Let $(A, \mathcal{A})$ be a randomized exact recovery algorithm, where the matrix A is in $\{0,1\}^{m \times n}$ with $m = \Theta(k \log(n/k))$, and A has column sparsity s. Then $s = \Omega(\log(n/k)$.*

1.2 Our techniques

To prove Theorem 1, we use a volume argument similar to [BIPW10]. Consider a large set of k-sparse vectors Y with minimum ℓ_1 distance $\Omega(1)$. If the column sparsity of A is bounded then the image of any k-sparse vector under A is also sparse. Therefore, by averaging argument, there is a relatively 'large' subset $Z \subset Y$ and relatively 'small' subset $I \subset [m]$ such that the support of Ax is contained in I for all $x \in Z$. Moreover, it can be seen that for any two elements $x, y \in Z$, the balls of radius $\Theta(1)$ around x and y, as well as their images, must be disjoint. Since those images lie in low dimensional subspace, where the dimension depends on s, we were able to give an inequality with dependence on s based on their volumes.

To prove Theorem 2, the main idea is to use counting argument. Any two distinct *k-sparse* vectors must map to different elements, otherwise a deterministic recovery algorithm can't distinguish them, and therefore can't recover their preimages. There are $\binom{n}{k}$ distinct *k-sparse* zero-one vectors, and we show that if most entries of A are zero then there will two vectors that maps to a same element. For the randomized case, we use the same counting argument along with the Yao's principle.

1.3 Related work

To best of knowledge, the only prior research related to column sparsity were done in [Cha08]. This paper studies matrices with *restricted isometry property* or RIP.

A matrix A satisfies $RIP(p)$ if ℓ_p norm of x is approximately preserved under the linear transformation Ax, for any *k-sparse* vector x. It was shown that a linear program can approximately recover the signal if the matrix A had the property [CRT06](when $p = 2$), [BGI+08](when $p = 1$). The paper [Cha08] proved that if A is a $m \times n$ matrix which satisfies $RIP(2)$ where $m = \Theta(k \log(n/k))$, then column sparsity of A is at least $\Omega(\frac{n}{k \log(n/k)})$.

For a survey of sparse recovery using sparse matrices, see [GI10].

2 Properties of compressed sensing matrices

Consider $m \times n$ matrix A, where $m < n$, for which there exists *deterministic* k-sparse *exact* recovery algorithm. First of all, note that, as A is a linear

map from $\mathbb{R}^m$ to $\mathbb{R}^n$, where $m < n$, there are two vectors x and y such that $Ax = Ay$. But on the other hand, for any *k-sparse* vector x, the deterministic algorithm must be able to exactly recover x from Ax. Therefore,

Lemma 1. *Let (A, D) be a deterministic k-sparse exact recovery algorithm. If x and y are k-*sparse *vectors then $Ax \neq Ay$.*

Proof. Otherwise, the deterministic will not be able to distinguish its preimage. □

We will also prove a similar impossibility result for C-approximate deterministic recovery algorithms.

Lemma 2. *Let (A, D) be a C-approximate deterministic ℓ_1/ℓ_1 recovery algorithm. Let x and y be k-*sparse *vectors. If z_1 and z_2 are real vectors with ℓ_1 norm less than $||x - y||_1/(2C + 2)$ then $A(x + z_1) \neq A(y + z_2)$.*

Proof (adapted from the proof of Theorem 3.1 in [BIPW10]) . Suppose, for the sake of contradiction, $A(x + z_1) = A(y + z_2)$.

Let w be the result of running D on $A(x + z_1)$. Then by the property of the algorithm, we have

$$||x + z_1 - w||_1 \leq C \cdot Err_k^1(x + z_1) \leq C\,||z_1||_1$$

On the other hand, we have $||x - w||_1 - ||z_1||_1 \leq ||x + z_1 - w||_1$, and therefore we get,

$$||x - w||_1 \leq (1 + C)\,||z_1||_1 \tag{2}$$

Similarly $||y - w||_2 \leq (1 + C)\,||z_2||_1$, so

$$||x - y||_1 \leq ||x - w||_1 + ||y - w||_1 \leq (1 + C)\,||z_1||_1 + (1 + C)\,||z_2||_1 \tag{3}$$

which is a contradiction when $||z_1||_1 < ||x - y||_1/(2C + 2)$ and $||z_2||_1 < ||x - y||_1/(2C + 2)$. □

3 Deterministic lower bound for general matrices

In this section, we will prove Theorem 1.

Lemma 3. *[Gilbert Varshamov] For any $q, k \in \mathbb{Z}^+$, $\epsilon \in \mathbb{R}^+$ with $\epsilon < 1 - 1/q$, there exists a set $Y \subset \{0,1\}^{qk}$ of binary vectors with exactly k ones such that Y has minimum Hamming distance $2\epsilon k$ and*

$$\log |Y| > (1 - H_q(\epsilon))k \log q$$

where H_q is the q-ary entropy function $H_q(x) = -x \log_q \frac{x}{q-1} - (1-x)\log_q(1-x)$.

See the appendix of [BIPW10] for proof.

Corollary 1. *There exists a set $X \subset \{0, \frac{1}{k}\}^n$ of k-sparse vectors with minimum ℓ_1 distance 1 such that $\log |X| = \Omega(k \log(n/k))$*

Proof. Let Y be the maximal set of k-sparse n-dimensional binary vectors with Hamming distance k. Take $X = \{\frac{1}{k}y : y \in Y\}$. By Lemma 3, $\log |X| = \log |Y| = \Omega(k \log(n/k))$. □

Proof of Theorem 1. Take a set X as defined in Corollary 1 and consider any vector $x \in X$. Since each column of A has at most s nonzero entries and x has exactly k nonzero entries, Ax has at most sk nonzero entries. To put in other way, for any $x \in X$, there exists a sequence $1 \leq a_1 < a_2 \ldots < a_{sk} \leq m$ such that Ax lies in the span of $< e_{a_1}, \ldots, e_{a_{sk}} >$ where e_i is defined to be i-th standard basis element. By averaging argument, there exists a set $Z \subset X$ of size $\frac{|X|}{\binom{m}{sk}}$ such that the images of elements of Z all lie in a same sk-dimensional subspace W. Let U be the maximal subspace such that image of U under A is contained in W. We denote $B_1^U(r)$ be the ℓ_1 ball in U of radius r.

Consider the following $|Z|$ balls: $\{x + B_1^U(r) | x \in Z\}$ where $r = 1/(2C+3)$. Note that, these balls all lie within $B_1^U(1+r)$ because $||x||_1 = 1$ for all $x \in Z$. Moreover, since $r = 1/(2C+3) < 1/(2C+2) \leq ||x-y||_1/(2C+2)$ for any $x, y \in Z$, by Lemma 2 the images of all those $|Z|$ balls are disjoint. The volume argument goes as follows.

Let $S = AB_1^U(1)$ be the image ℓ_1 ball in U of radius 1. This is a polytope of dimension sk with some volume V. The image of $B_1^U(r)$ is a linearly scaled S with volume $r^{sk}V$, and similarly the image of $B_1^U(1+r)$ has volume $(1+r)^{sk}V$. Hence, we get the following inequality $|Z|r^{sk} \leq (1+r)^{sk}$. Since

$|Z| = \frac{|X|}{\binom{m}{sk}}$ we get,

$$\frac{|X|}{\binom{m}{sk}} V r^{sk} \leq (1+r)^{sk} V$$

$$|X| \leq \binom{m}{sk} \left(\frac{1+r}{r}\right)^{sk}$$

$$|X| \leq \left(\frac{c'm}{sk}\right)^{sk} \text{ where } c' = e(1+r)/r$$

Without loss of generality, let's suppose $m = k\log(n/k)$. By Corollary 1, we have that $c''k\log(n/k) \leq \log|X|$ for some constant c''. Therefore,the following inequality holds:

$$c''k\log(n/k) \leq sk\log\left(c'\frac{\log(n/k)}{s}\right)$$

which is equivalent to

$$c''u \leq \log(c'u)$$

where $u = u(s) = \frac{\log(n/k)}{s}$. Note that, $c''u$ grows faster than $\log(c'u)$, and therefore, u is bounded by some constant T. Hence, we get $s \geq \frac{1}{T}\log(n/k)$. □

4 Deterministic lower bound for 0-1 matrices

Our main goal in this section is to prove Theorem 2.

Lemma 4. *The number of distinct vectors $v \in \mathbb{N}^m$ such that $||v||_1 \leq t$ is $\binom{m+t}{t}$.*

Proof. First of all, let's consider the case when $v \in \mathbb{N}^m$ and $||v||_1 = t$. It is the same as counting the number of multisets of cardinality t, with the elements taken from $\{1, 2, \ldots, m\}$. To see this, consider the following bijective map: if an element i is in the multiset with repetition k, then take $v_i = k$, and the inverse map is defined in the obvious way. From the basic combinatorial theory, the number of such multisets is $\binom{m+t-1}{t}$.

Next, we define the following bijective map between vectors $v \in \mathbb{N}^m$ with $||v||_1 \leq t$, and vectors $v \in \mathbb{N}^{m+1}$ with $||v||_1 = t$. Suppose we have a vector $v \in \mathbb{N}^m$ with $||v||_1 \leq t$, then we can add $(m+1)$-the entry with value $t - ||v||_1$. Now, suppose we have a vector $v \in \mathbb{N}^{m+1}$ with $||v||_1 = t$, then we can just delete the $(m+1)$-th entry. Therefore, the number of distinct vectors $v \in \mathbb{N}^m$ such that $||v||_1 \leq t$ is $\binom{m+t}{t}$. □

Proof of Theorem 2. Without loss of generality, let's suppose $m = k \log(n/k)$. For the sake of contradiction, suppose $s \leq B \log(n/k)$ (B will be determined later).

Consider $X = \{x \in \{0,1\}^n : x \text{ is k-sparse}\}$. Note that, since each column of A has ℓ_1 norm at most s and Ax is a sum of k columns, $||Ax||_1 \leq sk$ for $x \in X$. Moreover, since $A \in \{0,1\}^{m \times n}$, the entries of Ax are nonnegative integers.

The size of X is $\binom{n}{k}$, and by Lemma 1 the images of points in X must be disjoint. Therefore, by Lemma 4 the following inequality holds:

$$\binom{n}{k} \leq \binom{m + sk}{sk} \tag{4}$$

Recall the following inequality on binomial coefficients.

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k \tag{5}$$

Applying inequality (5) to (4) we get,

$$\begin{aligned}
\left(\frac{n}{k}\right)^k &\leq \left(\frac{e(k \log(n/k) + Bk \log(n/k))}{Bk \log(n/k)}\right)^{Bk \log(n/k)} \\
\left(\frac{n}{k}\right) &\leq \left(\frac{e(1+B)}{B}\right)^{B \log(n/k)} \\
\log\left(\frac{n}{k}\right) &\leq B \log\left(\frac{n}{k}\right) \log\left(\frac{e(1+B)}{B}\right) \\
1 &\leq B \log\left(\frac{e(1+B)}{B}\right)
\end{aligned}$$

which is impossible when B is, say, $1/10$. □

5 Randomized lower bound for 0-1 matrices

Our main goal in this section is to prove Theorem 3. A natural approach is to use Yao's principle; we find a distribution of hard inputs, and show that any deterministic algorithm is likely to fail on that distribution. Motivated by our deterministic bound, we take the input distribution to be uniform distribution over *k-sparse* vectors. The reader is suggested to read the proof of the deterministic case first.

Proof of Theorem 3. We define X to be $X = X_1 \cup X_2$ where

$$X_1 = \{x | x \in \{0,1\}^n \text{and } x \text{ is } k\text{-sparse}\}$$

and

$$X_2 = \{x | x \in \{0,2\}^n \text{and } x \text{ is } k\text{-sparse}\}$$

Let $\mathcal{I}$ be a uniform distribution over X. Note that $|X| = 2\binom{n}{k}$.

On one hand, by our assumption, there exists a distribution $\mathcal{R}$ of recovery algorithms, and by definition, the following must hold:

$$\text{For any } k\text{-}\textit{sparse} \text{ vector } x,\ P_{R\in\mathcal{R}}[R \text{ recovers } x] \geq 3/4 \tag{6}$$

On the other hand, by applying Yao's principle we get that,

$$\text{There exists deterministic algorithm D, } P_{x\in\mathcal{I}}[D \text{ recovers } x] \geq 3/4 \tag{7}$$

For any deterministic algorithm D,

$$P_{x\in\mathcal{I}}[D \text{ recovers } x] = \frac{\text{number of recoverable } x\text{'s in X}}{|X|} \tag{8}$$

As D is a deterministic algorithm, the number of recoverable x's in X cannot exceed the number of distinct images of points in X. Note that, since $||Ax||_1 \leq 2sk$ for $x \in X$, by Lemma 4 we have that, number of distinct images of points in X is at most $\binom{m+2sk}{2sk}$. Therefore, from (7) and (8) we get,

$$\frac{3}{4} \leq P_{x\in\mathcal{I}}[D \text{ recovers } x] \leq \frac{\binom{m+2sk}{2sk}}{2\binom{n}{k}}$$

which implies

$$6\binom{n}{k} \leq 4\binom{m+2sk}{2sk}$$

which implies

$$\binom{n}{k} \leq \binom{m+2sk}{2sk}$$

The same calculations as in the proof of Theorem 2 will yield that $s > B\log(n/k)$ where $B = 1/20$. □

6 Further discussions

We pose the following conjecture.

Conjecture 1. Let $(A, \mathcal{A})$ be a C-approximate *randomized* ℓ_1/ℓ_1 recovery algorithm, where the matrix A is a real $m \times n$ with $m = \Theta(k\log(n/k))$, and A has column sparsity s. Then $s = \Omega(\log(n/k)$.

The packing argument as in the proof of Theorem 1 doesn't seem to extend trivially for the randomized case. One approach might involve with communication complexity as in [BIPW10].

At this point, we have a good understanding about the sketch length and the column sparsity (at least for the deterministic case) of recovery matrices. For further interesting research, one might study about their tradeoffs.

We would like remark on matrices with *restricted isometry property*. As we have mentioned earlier, one could recover the signal if the matrix A had $RIP(1)$ [CRT06] or $RIP(2)$ [BGI+08], and therefore, our result implies that any matrix with $RIP(1)$ or $RIP(2)$ must have column sparsity at least $\Omega(\log(n/k))$ provided that it has $\Theta(k\log(n/k))$ rows.

7 Acknowledgements

I am very much privileged to have Prof. Piotr Indyk as my undergraduate thesis supervisor. I owe him much gratitude for his support, effort, kindness, and enlightening guidance and discussions throughout the Spring 2010 semester.

I would also like to thank my good friend Gabriel Bujokas for many insightful discussions.

References

[BGI+08] Radu Berinde, Anna C. Gilbert, Piotr Indyk, Howard J. Karloff, and Martin J. Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. *Allerton*, 2008.

[BI09] Radu Berinde and Piotr Indyk. Sequential sparse matching pursuit. *Allerton*, 2009.

[BIPW10] Khanh Do Ba, Piotr Indyk, Eric Price, and David P. Woodruff. Lower bounds for sparse recovery. *SODA*, 2010.

[Cha08] Venkat Chandar. A negative result concerning explicit matrices with the restricted isometry property. *Preprint*, 2008.

[CRT06] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[GI10] Anna Gilbert and Piotr Indyk. Sparse recovery using sparse matrices. In *Proceedings of IEEE*, 2010.

[Ind07] Piotr Indyk. Sketching, streaming and sublinear-space algorithms. *Graduate course notes available at http://stellar.mit.edu/S/course/6/fa07/6.895/*, 2007.

[LKM+06] J. N. Laska, S. Kirolos, Y. Massoud, R. Baraniuk, Anna Gilbert, Mark Iwen, and Martin Strauss. Random sampling for analog-to-information conversion of wideband signals. In *In Proceedings of the IEEE Dallas Circuits and Systems Workshop (DCAS)*, 2006.

[Mut03] S. Muthukrishnan. Data streams: algorithms and applications. In *SODA*, pages 413–413, 2003.

[RCS] RCSG. Rice compressive sensing group. *http://dsp.rice.edu/cs*.

[WLD+06] Michael B. Wakin, Jason N. Laska, Marco F. Duarte, Dror Baron, Shriram Sarvotham, Dharmpal Takhar, Kevin F. Kelly, and Richard G. Baraniuk. An architecture for compressive imaging. In *ICIP*, pages 1273–1276, 2006.