On Deterministic Sketching and Streaming for Sparse Recovery and Norm Estimation

Jelani Nelson[1*], Huy L. Nguyễn[1**], and David P. Woodruff[2]

[1] Princeton University, USA,
{minilek,hlnguyen}@princeton.edu,
[2] IBM Almaden Research Center, San Jose, USA,
dpwoodru@us.ibm.com

Abstract. We study classic streaming and sparse recovery problems using *deterministic* linear sketches, including ℓ_1/ℓ_1 and ℓ_∞/ℓ_1 sparse recovery problems, norm estimation, and approximate inner product. We focus on devising a fixed matrix $A \in \mathbb{R}^{m\times n}$ and a deterministic recovery/estimation procedure which work for all possible input vectors simultaneously. We contribute several improved bounds for these problems.

- A proof that ℓ_∞/ℓ_1 sparse recovery and inner product estimation are equivalent, and that incoherent matrices can be used to solve both problems. Our upper bound for the number of measurements is $m = O(\varepsilon^{-2}\min\{\log n, (\log n/\log(1/\varepsilon))^2\})$. We can also obtain fast sketching and recovery algorithms by making use of the Fast Johnson-Lindenstrauss transform. Both our running times and number of measurements improve upon previous work. We can also obtain better error guarantees than previous work in terms of a smaller tail of the input vector.
- A new lower bound for the number of linear measurements required to solve ℓ_1/ℓ_1 sparse recovery. We show $\Omega(k/\varepsilon^2 + k\log(n/k)/\varepsilon)$ measurements are required to recover an x' with $\|x - x'\|_1 \le (1+\varepsilon)\|x_{tail(k)}\|_1$, where $x_{tail(k)}$ is x projected onto all but its largest k coordinates in magnitude.
- A tight bound of $m = \Theta(\varepsilon^{-2}\log(\varepsilon^2 n))$ on the number of measurements required to solve deterministic norm estimation, i.e., to recover $\|x\|_2 \pm \varepsilon\|x\|_1$.

For all the problems we study, tight bounds are already known for the randomized complexity from previous work, except in the case of ℓ_1/ℓ_1 sparse recovery, where a nearly tight bound is known. Our work thus aims to study the deterministic complexities of these problems.

1 Introduction

In this work we provide new results for the point query problem as well as several other related problems: approximate inner product, ℓ_1/ℓ_1 sparse recovery,

* Supported by NSF grant CCF-0832797.

** Supported in part by NSF grant CCF-0832797 and a Gordon Wu fellowship.

and deterministic norm estimation. For many of these problems efficient randomized sketching and streaming algorithms exist, and thus we are interested in understanding the *deterministic* complexities of these problems.

1.1 Applications

Here we give a motivating application of the point query problem; for a formal definition of the problem, see below. Consider k servers $S^1, \ldots, S^k$, each holding a database $D^1, \ldots, D^k$, respectively. The servers want to compute statistics of the union D of the k databases. For instance, the servers may want to know the frequency of a record or attribute-pair in D. It may be too expensive for the servers to communicate their individual databases to a centralized server, or to compute the frequency exactly. Hence, the servers wish to communicate a short summary or "sketch" of their databases to a centralized server, who can then combine the sketches to answer frequency queries about D.

We model the databases as vectors $x^i \in \mathbb{R}^n$. To compute a sketch of x^i, we compute Ax^i for some $A \in \mathbb{R}^{m \times n}$. Importantly, $m \ll n$, and so Ax^i is much easier to communicate than x^i. The servers compute $Ax^1, \ldots, Ax^k$, respectively, and transmit these to a centralized server. Since A is a linear map, the centralized server can compute Ax for $x = c_1 x^1 + \ldots c_k x^k$ for any real numbers $c_1, \ldots, c_k$. Notice that the c_i are allowed to be both positive and negative, which is crucial for estimating the frequency of record or attribute-pairs in the difference of two datasets, which allows for tracking which items have experienced a sudden growth or decline in frequency. This is useful in network anomaly detection [11, 17, 24, 32, 37], and also for transactional data [16]. It is also useful for maintaining the set of frequent items over a changing database relation [16].

Associated with A is an output algorithm Out which given Ax, outputs a vector x' for which $\|x' - x\|_\infty \leq \varepsilon \|x_{tail(k)}\|_1$ for some number k, where $x_{tail(k)}$ denotes the vector x with the top k entries in magnitude replaced with 0. Thus x' approximates x well on every coordinate. We call the pair (A, Out) a solution to the point query problem. Given such a matrix A and an output algorithm Out, the centralized server can obtain an approximation to the value of every entry in x, which depending on the application, could be the frequency of an attribute-pair. It can also, e.g., extract the maximum frequencies of x, which are useful for obtaining the most frequent items. The centralized server obtains an entire histogram of values of coordinates in x, which is a useful low-memory representation of x. Notice that the communication is mk words, as opposed to nk if the servers were to transmit $x^1, \ldots, x^n$.

Our correctness guarantees hold for all input vectors simultaneously using one fixed (A, Out) pair, and thus it is stronger and should be contrasted with the guarantee that the algorithm succeeds given Ax with high probability for some fixed input x. For example, for the point query problem, the latter guarantee is achieved by the CountMin sketch [15] or CountSketch [13]. One of the reasons the randomized guarantee is less useful is because of *adaptive* queries. That is, suppose the centralized server computes x' and transmits information about x' to $S^1, \ldots, S^k$. Since x' could depend on A, if the servers were to then use the same matrix A to compute sketches $Ay^1, \ldots, Ay^k$ for databases $y^1, \ldots, y^k$ which depend on x', then A need not succeed, since it is not guaranteed to be correct with high probability for inputs y^i which depend on A.

1.2 Notation and Problem Definitions

Throughout this work $[n]$ denotes $\{1, \ldots, n\}$. For q a prime power, $\mathbb{F}_q$ denotes the finite field of size q. For $x \in \mathbb{R}^n$ and $S \subseteq [n]$, x_S denotes the vector with $(x_S)_i = x_i$ for $i \in S$, and $(x_S)_i = 0$ for $i \notin S$. The notation x_{-i} is shorthand for $x_{[n]\backslash\{i\}}$. For a matrix $A \in \mathbb{R}^{m\times n}$ and integer $i \in [n]$, A_i denotes the ith column of A. For matrices A and vectors x, A^T and x^T denote their transposes. For $x \in \mathbb{R}^n$ and integer $k \le n$, we let $head(x, k) \subseteq [n]$ denote the set of k largest coordinates in x in absolute value, and $tail(x, k) = [n]\backslash head(x, k)$. We often use $x_{head(k)}$ to denote $x_{head(x,k)}$, and similarly for the tail. For real numbers $a, b, \varepsilon \ge 0$, we use the notation $a = (1 \pm \varepsilon)b$ to convey that $a \in [(1 - \varepsilon)b, (1 + \varepsilon)b]$. A collection of vectors $\{C_1, \ldots, C_n\} \in [q]^t$ is called a *code* with *alphabet size* q and *block length* t, and we define $\Delta(C_i, C_j) = |\{k : (C_i)_k \ne (C_j)_k\}|$. The *relative distance* of the code is $\max_{i\ne j} \Delta(C_i, C_j)/t$.

We now define the problems that we study in this work, which all involve some *error parameter* $0 < \varepsilon < 1/2$. We want to design a fixed $A \in \mathbb{R}^{m\times n}$ and deterministic algorithm *Out* for each problem satisfying the following.

Problem 1: In the *ℓ_∞/ℓ_1 recovery problem*, also called the *point query problem*, $\forall x \in \mathbb{R}^n$, $x' = Out(Ax)$ satisfies $\|x - x'\|_\infty \le \varepsilon\|x\|_1$. The pair (A, Out) furthermore satisfies the *k-tail guarantee* if actually $\|x - x'\|_\infty \le \varepsilon\|x_{tail(k)}\|_1$.

Problem 2: In the *inner product problem*, $\forall x, y \in \mathbb{R}^n$, $\alpha = Out(Ax, Ay)$ satisfies $|\alpha - \langle x, y\rangle| \le \varepsilon\|x\|_1\|y\|_1$.

Problem 3: In the *ℓ_1/ℓ_1 recovery problem with the k-tail guarantee*, $\forall x \in \mathbb{R}^n$, $x' = Out(Ax)$ satisfies $\|x - x'\|_1 \le (1 + \varepsilon)\|x_{tail(k)}\|_1$.

Problem 4: In the *ℓ_2 norm estimation problem*, $\forall x \in \mathbb{R}^n$, $\alpha = Out(Ax)$ satisfies $|\|x\|_2 - \alpha| \le \varepsilon\|x\|_1$.

We note that for the first, second, and fourth problems above, our errors are additive and not relative. This is because relative error is impossible to achieve with a sublinear number of measurements. If A is a fixed matrix with $m < n$, then it has some non-trivial kernel. Since for all the problems above an *Out* procedure would have to output 0 when $Ax = 0$ to achieve bounded relative approximation, such a procedure would fail on any input vector in the kernel which is not the 0 vector.

For Problem 2 one could also ask to achieve additive error $\varepsilon\|x\|_p\|y\|_p$ for $p > 1$. For $y = e_i$ for a standard unit vector e_i, this would mean approximating x_i up to additive error $\varepsilon\|x\|_p$. This is not possible unless $m = \Omega(n^{2-2/p})$ for $1 < p \le 2$ and $m = \Omega(n)$ for $p \ge 2$ [21]. For Problem 3, it is known that the analogous guarantee of returning x' for which $\|x - x'\|_2 \le \varepsilon\|x_{tail(k)}\|_2$ is not possible unless $m = \Omega(n)$ [14].

1.3 Our Contributions and Related Work

We study the four problems stated above, where we have the deterministic guarantee that a single pair (A, Out) provides the desired guarantee for all input vectors simultaneously. We first show that point query and inner product are equivalent up to changing ε by a constant factor. We then show that any "incoherent matrix" A can be used for these two problems to perform the linear measurements; that is, a matrix A whose columns have unit ℓ_2 norm and such that each pair of columns has dot product at most ε in magnitude. Such matrices can be obtained from the Johnson-Lindenstrauss (JL) lemma [29], almost pairwise independent sample spaces [7, 38], or error-correcting codes, and they play a prominent role in compressed sensing [18, 36] and mathematical approximation theory [25]. The connection between point query and codes was implicit in [22], though a suboptimal code was used, and the observation that the more general class of incoherent matrices suffices is novel. This connection allows us to show that $m = O(\varepsilon^{-2} \min\{\log n, (\log n/\log(1/\varepsilon))^2\})$ measurements suffice, and where Out and the construction of A are completely deterministic. Alon has shown that any incoherent matrix must have $m = \Omega(\varepsilon^{-2} \log n/\log(1/\varepsilon))$ [6]. Meanwhile the best known lower bound for point query is $m = \Omega(\varepsilon^{-2} + \varepsilon^{-1} \log(\varepsilon n))$ [19, 20, 27], and the previous best known upper bound was $m = O(\varepsilon^{-2} \log^2 n/(\log 1/\varepsilon + \log\log n))$ [22]. If the construction of A is allowed to be Las Vegas polynomial time, then we can use the Fast Johnson-Lindenstrauss transforms [2–4, 34] so that Ax can be computed quickly, e.g. in $O(n \log m)$ time as long as $m < n^{1/2-\gamma}$ [3], and with $m = O(\varepsilon^{-2} \log n)$. Our Out algorithm is equally fast. We also show that for point query, if we allow the measurement matrix A to be constructed by a polynomial Monte Carlo algorithm, then the $1/\varepsilon^2$-tail guarantee can be obtained essentially "for free", i.e. by keeping $m = O(\varepsilon^{-2} \log n)$. Previously the work [22] only showed how to obtain the $1/\varepsilon$-tail guarantee "for free" in this sense of not increasing m (though the m in [22] was worse). We note that for randomized algorithms which succeed with high probability for any given input, it suffices to take $m = O(\varepsilon^{-1} \log n)$ by using the CountMin data structure [15], and this is optimal [30] (the lower bound in [30] is stated for the so-called heavy hitters problem, but also applies to the ℓ_∞/ℓ_1 recovery problem).

For the ℓ_1/ℓ_1 sparse recovery problem with the k-tail guarantee, we show a lower bound of $m = \Omega(k \log(\varepsilon n/k)/\varepsilon + k/\varepsilon^2)$. The best upper bound is $O(k \log(n/k)/\varepsilon^2)$ [28]. Our lower bound implies a separation for the complexity of this problem in the case that one must simply pick a random (A, Out) pair which works for some given input x with high probability (i.e. not for all x simultaneously), since [39] showed an $m = O(k \log n \log^3(1/\varepsilon)/\sqrt{\varepsilon})$ upper bound in this case. The first summand of our lower bound uses techniques used in [9, 39]. The second summand uses a generalization of an argument of Gluskin [27], which was later rediscovered by Ganguly [20], which showed the lower bound $m = \Omega(1/\varepsilon^2)$ for point query.

Finally, we show how to devise an appropriate (A, Out) for ℓ_2 norm estimation with $m = O(\varepsilon^{-2} \log(\varepsilon^2 n))$, which is optimal. The construction of A is randomized but then works for all x with high probability. The proof takes A

according to known upper bounds on Gelfand widths, and the recovery procedure *Out* requires solving a simple convex program. As far as we are aware, this is the first work to investigate this problem in the deterministic setting. In the case that (A, Out) can be chosen randomly to work for any fixed x with high probability, one can use the AMS sketch [8] with $m = O(\varepsilon^{-2}\log(1/\delta))$ to succeed with probability $1-\delta$ and to obtain the better guarantee $\varepsilon\|x\|_2$. The AMS sketch can also be used for the inner product problem to obtain error guarantee $\varepsilon\|x\|_2\|y\|_2$ with the same m.

Due to space constraints, many of our proofs are omitted or abbreviated. Full proofs can be found in the full version.

2 Point Query and Inner Product Estimation

We first show that the problems of point query and inner product estimation are equivalent up to changing the error parameter ε by a constant factor.

Theorem 1. *Any solution (A, Out') to inner product estimation with error parameter ε yields a solution (A, Out) to the point query problem with error parameter ε. Also, a solution (A, Out) for point query with error ε yields a solution (A, Out') to inner product with error 12ε. The time complexities of Out and Out′ are equal up to* $\mathrm{poly}(n)$ *factors.*

Proof: Let (A, Out') be a solution to the inner product problem such that $Out'(Ax, Ay) = \langle x, y\rangle \pm \varepsilon\|x\|_1\|y\|_1$. Then given $x \in \mathbb{R}^n$, to solve the point query problem we return the vector with $Out(Ax)_i = Out'(Ax, Ae_i)$, and our guarantees are immediate.

Now let (A, Out) be a solution to the point query problem. Given $x, y \in \mathbb{R}^n$, let $x' = Out(Ax), y' = Out(Ay)$. Our estimate for $\langle x, y\rangle$ is $Out'(Ax, Ay) = \left\langle x'_{head(1/\varepsilon)}, y'_{head(1/\varepsilon)}\right\rangle$. Correctness is proven in the full version. ■

Since the two problems are equivalent up to changing ε by a constant factor, we focus on point query. We first have the following lemma, stating that any *incoherent matrix* A has a correct associated *Out* procedure (namely, multiplication by A^T). An incoherent matrix, is an $m \times n$ matrix A for which all columns A_i of A have unit ℓ_2 norm, and for all $i \neq j$ we have $|\langle A_i, A_j\rangle| \leq \varepsilon$.

Lemma 1. *Any incoherent matrix A with error parameter ε has an associated* $\mathrm{poly}(mn)$*-time deterministic recovery procedure Out for which (A, Out) is a solution to the point query problem. In fact, for any $x \in \mathbb{R}^n$, given Ax and $i \in [n]$, the output x'_i satisfies $|x'_i - x_i| \leq \varepsilon\|x_{-i}\|_1$.*

It is known that any incoherent matrix has $m = \Omega((\log n)/(\varepsilon^2 \log 1/\varepsilon))$ [6], and the JL lemma implies such matrices with $m = O((\log n)/\varepsilon^2)$ [29]. For example, there exist matrices in $\{-1/\sqrt{m}, 1/\sqrt{m}\}^{m\times n}$ satisfying this property [1], which can also be found in $\mathrm{poly}(n)$ time [41] (we note that [41] gives running time exponential in precision, but the proof holds if the precision is taken to be $O(\log(n/\varepsilon))$. It is also known that incoherent matrices can be obtained from almost pairwise independent sample spaces [7, 38] or error-correcting codes, and thus these tools can also be used to solve the point query problem. The connection to codes was already implicit in [22], though the code used in that work is suboptimal, as we will show soon. Below we elaborate on what bounds these tools provide for incoherent matrices, and thus the point query problem.

Incoherent matrices from JL: The upside of the connection to the JL lemma is that we can obtain incoherent matrices A such that Ax can be computed quickly, via the Fast Johnson-Lindenstrauss Transform introduced by Ailon and Chazelle [2] or related subsequent works. The JL lemma states the following.

Theorem 2 (JL lemma). *For any $x_1, \ldots, x_N \in \mathbb{R}^n$ and any $0 < \varepsilon < 1/2$, there exists $A \in \mathbb{R}^{m\times n}$ with $m = O(\varepsilon^{-2}\log N)$ such that for all $i, j \in [N]$ we have $\|Ax_i - Ax_j\|_2 = (1 \pm \varepsilon)\|x_i - x_j\|_2$.*

Consider the matrix A obtained from the JL lemma when the set of vectors is $\{0, e_1, \ldots, e_n\} \in \mathbb{R}^n$. Then columns A_i of A have ℓ_2 norm $1 \pm \varepsilon$, and furthermore for $i \neq j$ we have $|\langle A_i, A_j\rangle| = (\|A_i - A_j\|_2^2 - \|A\|_i^2 - \|A\|_j^2)/2 = ((1\pm\varepsilon)^2 2 - (1\pm\varepsilon) - (1\pm\varepsilon))/2 \le 2\varepsilon + \varepsilon^2/2$. By scaling each column to have ℓ_2 norm exactly 1, we still preserve that dot products between pairs of columns are $O(\varepsilon)$ in magnitude.

Incoherent matrices from almost pairwise independence: An *ε-almost pairwise independent sample space* a set $S \subseteq \{-1, 1\}^n$ satisfying the following. For any $i \neq j \in [n]$, the ℓ_1 distance between the uniform distribution over $\{-1, 1\}^2$ and the distribution of x_i, x_j when x is drawn uniformly at random from S is at most ε. A matrix whose rows are the elements of S, divided by a scale factor of $\sqrt{S}$, is incoherent. Details are in the full version, but we do not delve deeper since this approach does not improve upon the bounds via JL matrices.

Incoherent matrices from codes: Finally we explain the connection between incoherent matrices and codes. A connection to balanced binary codes was made in [6], and to arbitrary codes over larger alphabets without detail in a remark in [5]. Though not novel, we elaborate on this latter connection for the sake of completeness. Let $\mathcal{C} = \{C_1, \ldots, C_n\}$ be a code with alphabet size q, block length t, and relative distance $1-\varepsilon$. The fact that such a code gives rise to a matrix $A \in \mathbb{R}^{m\times n}$ for point query with error parameter ε was implicit in [22], but we make it explicit here. We let $m = qt$ and conceptually partition the rows of A arbitrarily into t sets each of size q. For the column A_i, let $(A_i)_{j,k}$ denote the entry of A_i in the kth coordinate of the jth block. We set $(A_i)_{j,k} = 1/\sqrt{t}$ if $(C_i)_j = k$, and $(A_i)_{j,k} = 0$ otherwise. Each column has exactly t non-zero entries of value $1/\sqrt{t}$, and thus has ℓ_2 norm 1. Furthermore, for $i \neq j$, $\langle A_i, A_j\rangle = (t - \Delta(C_i, C_j))/t \le \varepsilon$.

The work [22] instantiated the above with the following *Chinese remainder code* [35, 42, 44], which yielded $m = O(\varepsilon^{-2}\log^2 n/(\log 1/\varepsilon + \log\log n))$. We observe here that this bound is never optimal. A random code with $q = 2/\varepsilon$ and $t = O(\varepsilon^{-1}\log n)$ has the desired properties by applying the Chernoff bound on a pair of codewords, then a union bound over codewords (alternatively, such a code is promised by the Gilbert-Varshamov (GV) bound). If ε is sufficiently small, a Reed-Solomon code performs even better. That is, we take a finite field $\mathbb{F}_q$ for $q = \Theta(\varepsilon^{-1}\log n/(\log\log n + \log(1/\varepsilon)))$ and $q = t$, and each C_i corresponds to a distinct degree-d polynomial p_i over $\mathbb{F}_q$ for $d = \Theta(\log n/(\log\log n + \log(1/\varepsilon)))$ (note there are at least $q^d > n$ such polynomials). We set $(C_i)_j = p_i(j)$. The relative distance is as desired since $p_i - p_j$ has at most d roots over $\mathbb{F}_q$ and thus can be 0 at most $d \le \varepsilon t$ times. This yields $qt = O(\varepsilon^{-2}(\log n/(\log\log n + \log(1/\varepsilon))^2)$, which surpasses the GV bound for $\varepsilon < 2^{-\Omega(\sqrt{\log n})}$, and is always better than the Chinese remainder code. We note that this construction of a binary matrix based on Reed-Solomon codes is identical to one used by Kautz and Singleton in the different context of group testing [33].

Time	m	Details	Explicit?
$O((n\log n)/\varepsilon^2)$	$O(\varepsilon^{-2}\log n)$	$A \in \{-1/\sqrt{m}, 1/\sqrt{m}\}^{m\times n}$ [1, 41]	yes
$O((n\log n)/\varepsilon)$	$O(\varepsilon^{-2}\log n)$	sparse JL [31], GV code	no
$O(nd\log^2 d\log\log d/\varepsilon)$	$O(d^2/\varepsilon^2)$	Reed-Solomon code	yes
$O_\gamma(n\log m + m^{2+\gamma})$	$O(\varepsilon^{-2}\log n)$	FFT-based JL [3]	no
$O(n\log n)$	$O(\varepsilon^{-2}\log^5 n)$	FFT-based JL [4, 34]	no

Fig. 1. Implications for point query from JL matrices and codes. Time indicates the running time to compute Ax given x. In the case of Reed-Solomon, $d = O(\log n/(\log\log n + \log(1/\varepsilon)))$. We say the construction is "explicit" if A can be computed in deterministic time poly(n); otherwise we only provide a polynomial time Las Vegas algorithm to construct A.

In Figure 1 we elaborate on what known constructions of codes and JL matrices provide for us in terms of point query. In the case of running time for the Reed-Solomon construction, we use that degree-d polynomials can be evaluated on $d+1$ points in a total of $O(d\log^2 d\log\log d)$ field operations over $\mathbb{F}_q$ [43, Ch. 10]. In the case of [3], the constant $\gamma > 0$ can be chosen arbitrarily, and the constant in the big-Oh depends on $1/\gamma$. We note that except in the case of Reed-Solomon codes, the construction of A is randomized (though once A is generated, incoherence can be verified in polynomial time, thus providing a poly(n)-time Las Vegas algorithm).

Note that Lemma 1 did not just give us error $\varepsilon\|x\|_1$, but actually gave us $|x_i - x'_i| \le \varepsilon\|x_{-i}\|_1$, which is stronger. We now show that an even stronger guarantee is possible. We will show that in fact it is possible to obtain $\|x - x'\|_\infty \le \varepsilon\|x_{tail(1/\varepsilon^2)}\|_1$ while increasing m by only an additive $O(\varepsilon^{-2}\log(\varepsilon^2 n))$, which is less than our original m except potentially in the Reed-Solomon construction. The idea is to, in parallel, recover a good approximation of $x_{head(1/\varepsilon^2)}$ with error proportional to $\|x_{tail(1/\varepsilon^2)}\|_1$ via compressed sensing, then to subtract from Ax before running our recovery procedure. We now give details.

We in parallel run a k-*sparse recovery* algorithm which has the following guarantee: there is a pair (B, Out') such that for any $x \in \mathbb{R}^n$, we have that

$x' = Out'(Bx) \in \mathbb{R}^n$ satisfies $\|x' - x\|_2 \le O(1/\sqrt{k})\|x_{tail(k)}\|_1$. Such a matrix B can be taken to have the *restricted isometry property of order k* (k-RIP), i.e. that it preserves the ℓ_2 norm up to a small multiplicative constant factor for all k-sparse vectors in $\mathbb{R}^n$.[3] It is known [26] that any such x' also satisfies the guarantee that $\|x'_{head(k)} - x\|_1 \le O(1)\|x_{tail(k)}\|_1$, where $x'_{head(k)}$ is the vector which agrees with x' on the top k coordinates in magnitude and is 0 on the remaining coordinates. Moreover, it is also known [10] that if B satisfies the JL lemma for a particular set of $N = (en/k)^{O(k)}$ points in $\mathbb{R}^n$, then B will be k-RIP. The associated output procedure Out' takes Bx and outputs $\mathrm{argmin}_{z|Bx=Bz}\|z\|_1$ by solving a linear program [12]. All the JL matrices in Figure 1 provide this guarantee with $O(k\log(en/k))$ rows, except for the last row which satisfies k-RIP with $O(k\log(en/k)\log^2 k\log(k\log n))$ rows [40].

Theorem 3. *Let A be an incoherent matrix A with error parameter ε, and let B be k-RIP. Then there is an output procedure Out which for any $x \in \mathbb{R}^n$, given only Ax, Bx, outputs a vector x' with $\|x' - x\|_\infty \le \varepsilon\|x_{tail(k)}\|_1$.*

Proof: Given Bx, we first run the k-sparse recovery algorithm to obtain a vector y with $\|x - y\|_1 = O(1)\|x_{tail(k)}\|_1$. We then construct our output vector x' coordinate by coordinate. To construct x'_i, we replace y_i with 0, obtaining the vector z^i. Then we compute $A(x - z^i)$ and run the point query output procedure associated with A and index i. The guarantee is that the output w^i of the point query algorithm satisfies $|w^i_i - (x - z^i)_i| \le \varepsilon\|(x - z^i)_{-i}\|_1$, where

$$\|(x - z^i)_{-i}\|_1 = \|(x - y)_{-i}\|_1 \le \|x - y\|_1 = O(1)\|x_{tail(k)}\|_1,$$

and so $|(w^i + z^i)_i - x_i| = O(\varepsilon)\|x_{tail(k)}\|_1$. If we define our output vector by $x'_i = w^i_i + z^i_i$ and rescale ε by a constant factor, this proves the theorem. ■

By setting $k = 1/\varepsilon^2$ in Theorem 3 and stacking the rows of a k-RIP and incoherent matrix each with $O((\log n)/\varepsilon^2)$ rows, we obtain the following corollary.

Corollary 1. *There is an $m \times n$ matrix A and associated output procedure Out which for any $x \in \mathbb{R}^n$, given Ax, outputs a vector x' with $\|x' - x\|_\infty \le \varepsilon\|x_{tail(1/\varepsilon^2)}\|_1$. Here $m = O((\log n)/\varepsilon^2)$.*

It is also possible to obtain a tail-error guarantee for inner product.

Theorem 4. *Suppose $1/\varepsilon^2 < n/2$. There is an (A, Out) with $A \in \mathbb{R}^{m\times n}$ for $m = O(\varepsilon^{-2}\log n)$ such that for any $x, y \in \mathbb{R}^n$, $Out(Ax, Ay)$ gives an output which is $\langle x, y\rangle \pm \varepsilon(\|x\|_2\|y_{tail(1/\varepsilon^2)}\|_1 + \|x_{tail(1/\varepsilon^2)}\|_1\|y\|_2) + \varepsilon^2\|x_{tail(1/\varepsilon^2)}\|_1\|y_{tail(1/\varepsilon^2)}\|_1$.*

Here we state a lower bound for the point query problem. The proof can be found in the full version and follows from the works [20, 27] and volume arguments as used in [9].

[3] Unfortunately currently the only known constructions of k-RIP constructions with the values of m we discuss are Monte Carlo, forcing our algorithms in this section with the k-tail guarantee to only be Monte Carlo polynomial time when constructing the measurement matrix.

Theorem 5. *Let $0 < \varepsilon < \varepsilon_0$ for some universal constant $\varepsilon_0 < 1$. Suppose $1/\varepsilon^2 < n/2$, and A is an $m \times n$ matrix for which given Ax it is always possible to produce a vector x' such that $\|x - x'\|_\infty \le \varepsilon \|x_{tail(k)}\|_1$. Then $m = \Omega(k \log(n/k)/\log k + \varepsilon^{-2} + \varepsilon^{-1} \log n)$.*

3 Lower Bounds for ℓ_1/ℓ_1 recovery

Recall in the ℓ_1/ℓ_1-recovery problem, we would like to design a matrix $A \in \mathbb{R}^{m\times n}$ such that for any $x \in \mathbb{R}^n$, given Ax we can recover $x' \in \mathbb{R}^n$ such that $\|x - x'\|_1 \le (1+\varepsilon)\|x_{tail(k)}\|_1$. We now show two lower bounds.

Theorem 6. *Let $0 < \varepsilon < 1/\sqrt{8}$ be arbitrary, and k be an integer. Suppose $k/\varepsilon^2 < (n-1)/2$. Then any matrix $A \in \mathbb{R}^{m\times n}$ which allows ℓ_1/ℓ_1-recovery with the k-tail guarantee with error ε must have $m \ge \min\{n/2, (1/16)k/\varepsilon^2\}$.*

Proof: Without loss of generality we may assume that the rows of A are orthonormal. This is because first we can discard rows of A until the rows remaining form a basis for the rowspace of A. Call this new matrix with potentially fewer rows A'. Note that any dot products of rows of A with x that the recovery algorithm uses can be obtained by taking linear combinations of entries of $A'x$. Next, we can then find a matrix $T \in \mathbb{R}^{m\times m}$ so that TA' has orthonormal rows, and given $TA'x$ we can recover $A'x$ in post-processing by left-multiplication with T^{-1}. We henceforth assume that the rows of A are orthonormal. Since $A \cdot 0 = 0$, and our recovery procedure must in particular be accurate for $x = 0$, the recovery procedure must output $x' = 0$ for any $x \in ker(A)$. We consider $x = (I - A^T A)y$ for $y = \sum_{i=1}^k \sigma_i e_{\pi(i)}$. Here π is a random permutation on n elements, and $\sigma_1, \ldots, \sigma_k$ are independent and uniform random variables in $\{-1, 1\}$. Since $x \in ker(A)$, which follows since $AA^T = I$ by orthonormality of the rows of A, the recovery algorithm will output $x' = 0$. Nevertheless, in the full version we show that unless $m \ge \min\{n/2, (1/16)k/\varepsilon^2\}$, we will have $\|x\|_1 > (1+\varepsilon)\|x_{tail(k)}\|_1$ with positive probability so that by the probabilistic method there exists $x \in ker(A)$ for which $x' = 0$ is not a valid output. ∎

We now give another lower bound via a different approach. As in [9, 39], we use 2-party communication complexity to prove an $\Omega((k/\varepsilon)\log(\varepsilon n/k))$ bound on the number of rows of any ℓ_1/ℓ_1 sparse recovery scheme. The main difference from prior work is that we use deterministic communication complexity and a different communication problem.

We show how to use a pair (A, Out) with the property that for all vectors z, the output z' of $Out(Az)$ satisfies $\|z - z'\|_1 \le (1+\varepsilon)\|z_{tail(k)}\|_1$, to construct a correct protocol for the equality function on strings $x, y \in \{0,1\}^r$ for $r = \Theta((k/\varepsilon)\log n \log(\varepsilon n/k))$, where the communication is an $O(\log n)$ factor larger than the number of rows of A. We then show how this implies the number of rows of A is $\Omega((k/\varepsilon)\log(\varepsilon n/k))$. Details are in the full version.

Theorem 7. *Any matrix A which allows ℓ_1/ℓ_1-recovery with the k-tail guarantee with error ε satisfies $m = \Omega((k/\varepsilon)\log(\varepsilon n/k))$.*

4 Deterministic Norm Estimation and the Gelfand Width

Theorem 8. *For $1 \leq p < q \leq \infty$, let m be the minimum number such that there is an $n-m$ dimensional subspace S of $\mathbb{R}^n$ satisfying $\sup_{v \in S} \frac{\|v\|_q}{\|v\|_p} \leq \varepsilon$. Then there is an $m \times n$ matrix A and associated output procedure Out which for any $x \in \mathbb{R}^n$, given Ax, outputs an estimate of $\|v\|_q$ with additive error at most $\varepsilon\|v\|_p$. Moreover, any matrix A with fewer rows fails to perform this task.*

Proof: Consider a matrix A whose kernel is such a subspace. For any sketch z, we need to return a number in the range $[\|x\|_q - \varepsilon\|x\|_p, \|x\|_q + \varepsilon\|x\|_p]$ for any x satisfying $Ax = z$. Assume for contradiction that it is not possible. Then there exist x and y such that $Ax = Ay$ but $\|x\|_q - \varepsilon\|x\|_p > \|y\|_q + \varepsilon\|y\|_p$. However, since $x-y$ is in the kernel of A, $\|x\|_q - \|y\|_q \leq \|x-y\|_q \leq \varepsilon\|x-y\|_p \leq \varepsilon(\|x\|_p + \|y\|_p)$. Thus, we have a contradiction. This argument also shows that given the sketch z, the output procedure can return $\min_{x:Ax=z} \|x\|_q + \varepsilon\|x\|_p$. This is a convex optimization problem that can be solved in polynomial time using the ellipsoid algorithm; details are in the full version.

For the lower bound, consider a matrix A with fewer than m rows. Then in the kernel of A, there exists v such that $\|v\|_q > \varepsilon\|v\|_p$. Both v and the zero vector give the same sketch (a zero vector). However, by the stated requirement, we need to output 0 for the zero vector but some positive number for v. Thus, no matrix A with fewer than m rows can solve the problem. ■

The subspace S of highest dimension of $\mathbb{R}^n$ satisfying $\sup_{v \in S} \frac{\|v\|_q}{\|v\|_p} \leq \varepsilon$ is related to the *Gelfand width*, a well-studied notion in functional analysis. For $p < q$, the Gelfand width of order m of ℓ_p and ℓ_q unit balls in $\mathbb{R}^n$ is defined as the infimum over all subspaces $A \subseteq \mathbb{R}^n$ of codimension m of $\sup_{v \in A} \frac{\|v\|_q}{\|v\|_p}$. Using known bounds for the Gelfand width for $p = 1$ and $q = 2$ [19, 23], we obtain the following corollary.

Corollary 2. *Assume that $1/\varepsilon^2 < n/2$. There is an $m \times n$ matrix A and associated output procedure Out which for any $x \in \mathbb{R}^n$, given Ax, outputs an estimate e such that $\|x\|_2 - \varepsilon\|x\|_1 \leq e \leq \|x\|_2 + \varepsilon\|x\|_1$. Here $m = O(\varepsilon^{-2}\log(\varepsilon^2 n))$ and this bound for m is tight.*

Acknowledgments

We thank Raghu Meka for answering several questions about almost k-wise independent sample spaces. We thank an anonymous reviewer for pointing out the connection between incoherent matrices and ε-biased spaces.

References

1. D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

2. N. Ailon and B. Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009.
3. N. Ailon and E. Liberty. Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete & Computational Geometry*, 42(4):615–630, 2009.
4. N. Ailon and E. Liberty. Almost optimal unrestricted fast Johnson-Lindenstrauss transform. In *Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 185–191, 2011.
5. N. Alon. Problems and results in extremal combinatorics - I. *Discrete Mathematics*, 273(1-3):31–53, 2003.
6. N. Alon. Perturbed identity matrices have high rank: Proof and applications. *Combinatorics, Probability & Computing*, 18(1-2):3–15, 2009.
7. N. Alon, O. Goldreich, J. Håstad, and R. Peralta. Simple construction of almost k-wise independent random variables. *Rand. Struct. Alg.*, 3(3):289–304, 1992.
8. N. Alon, Y. Matias, and M. Szegedy. The Space Complexity of Approximating the Frequency Moments. *JCSS*, 58(1):137–147, 1999.
9. K. D. Ba, P. Indyk, E. Price, and D. P. Woodruff. Lower bounds for sparse recovery. In *SODA*, pages 1190–1197, 2010.
10. R. Baraniuk, M. A. Davenport, R. DeVore, and M. Wakin. A simple proof of the Restricted Isometry Property. *Constructive Approximation*, 28(3):253–263, 2008.
11. D. Barbará, N. Wu, and S. Jajodia. Detecting novel network intrusions using Bayes estimators. In *Proceedings of the 1st SIAM International Conference on Data Mining*, 2001.
12. E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Information Theory*, 52(2):489–509, 2006.
13. M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3–15, 2004.
14. A. Cohen, W. Dahmen, and R. A. DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, 22:211–231, 2009.
15. G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *J. Algorithms*, 55(1):58–75, 2005.
16. G. Cormode and S. Muthukrishnan. What's hot and what's not: tracking most frequent items dynamically. *ACM Trans. Database Syst.*, 30(1):249–278, 2005.
17. E. D. Demaine, A. López-Ortiz, and J. I. Munro. Frequency estimation of Internet packet streams with limited space. In *ESA*, pages 348–360, 2002.
18. D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Th.*, 47:2558–2567, 2001.
19. S. Foucart, A. Pajor, H. Rauhut, and T. Ullrich. The Gelfand widths of ℓ_p-balls for $0 < p \leq 1$. *Journal of Complexity*, 26(6):629–640, 2010.
20. S. Ganguly. Lower bounds on frequency estimation of data streams. In *CSR*, pages 204–215, 2008.
21. S. Ganguly. Deterministically estimating data stream frequencies. In *COCOA*, pages 301–312, 2009.
22. S. Ganguly and A. Majumder. CR-precis: A deterministic summary structure for update data streams. In *ESCAPE*, pages 48–59, 2007.
23. A. Y. Garnaev and E. D. Gluskin. On the widths of the Euclidean ball. *Soviet Mathematics Doklady*, 30:200–203, 1984.
24. A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. J. Strauss. Quicksand: Quick summary and analysis of network data. DIMACS Technical Report 2001-43, 2001.
25. A. C. Gilbert, S. Muthukrishnan, and M. Strauss. Approximation of functions over redundant dictionaries using coherence. In *SODA*, pages 243–252, 2003.

26. A. C. Gilbert, M. J. Strauss, J. A. Tropp, and R. Vershynin. One sketch for all: fast algorithms for compressed sensing. In *STOC*, pages 237–246, 2007.
27. E. D. Gluskin. On some finite-dimensional problems in the theory of widths. *Vestn. Leningr. Univ. Math.*, 14:163–170, 1982.
28. P. Indyk and M. Ružić. Near-optimal sparse recovery in the L_1 norm. In *FOCS*, pages 199–207, 2008.
29. W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.
30. H. Jowhari, M. Saglam, and G. Tardos. Tight bounds for L_p samplers, finding duplicates in streams, and related problems. In *PODS*, pages 49–58, 2011.
31. D. M. Kane and J. Nelson. Sparser Johnson-Lindenstrauss transforms. In *SODA*, pages 1195–1206, 2012.
32. R. M. Karp, S. Shenker, and C. H. Papadimitriou. A simple algorithm for finding frequent elements in streams and bags. *ACM Trans. Database Syst.*, 28:51–55, 2003.
33. W. H. Kautz and R. C. Singleton. Nonrandom binary superimposed codes. *IEEE Trans. Inf. Theory*, 10:363–377, 1964.
34. F. Krahmer and R. Ward. New and improved Johnson-Lindenstrauss embeddings via the Restricted Isometry Property. *SIAM J. Math. Anal.*, 43(3):1269–1281, 2011.
35. H. Krishna, B. Krishna, K.-Y. Lin, and J.-D. Sun. *Computational Number Theory and Digital Signal Processing: Fast Algorithms and Error Control Techniques.* CRC, Boca Raton, FL, 1994.
36. S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12):3397–3415, 1993.
37. J. Misra and D. Gries. Finding repeated elements. *Sci. Comput. Program.*, 2(2):143–152, 1982.
38. J. Naor and M. Naor. Small-bias probability spaces: Efficient constructions and applications. *SIAM J. Comput.*, 22(4):838–856, 1993.
39. E. Price and D. P. Woodruff. (1 + eps)-approximate sparse recovery. In *FOCS*, pages 295–304, 2011.
40. M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:1025–1045, 2008.
41. D. Sivakumar. Algorithmic derandomization via complexity theory. In *STOC*, pages 619–626, 2002.
42. M. A. Soderstrand, W. K. Jenkins, G. A. Jullien, and F. J. Taylor. *Residue Number System Arithmetic: Modern Applications in Digital Signal Processing.* IEEE Press, New York, 1986.
43. J. von zur Gathen and J. Gerhard. *Modern Computer Algebra.* Cambridge University Press, 1999.
44. R. W. Watson and C. W. Hastings. Self-checked computation using residue arithmetic. *Proc. IEEE*, 4(12):1920–1931, 1966.