New constructions of RIP matrices with fast multiplication and fewer rows

Jelani Nelson* Eric Price† Mary Wootters‡

November 5, 2012

Abstract

In compressed sensing, the *restricted isometry property* (RIP) is a sufficient condition for the efficient reconstruction of a nearly k-sparse vector $x \in \mathbb{C}^d$ from m linear measurements Φx. It is desirable for m to be small, and for Φ to support fast matrix-vector multiplication. In this work, we give a randomized construction of RIP matrices $\Phi \in \mathbb{C}^{m \times d}$, preserving the ℓ_2 norms of all k-sparse vectors with distortion $1+\varepsilon$, where the matrix-vector multiply Φx can be computed in nearly linear time. The number of rows m is on the order of $\varepsilon^{-2} k \log d \log^2(k \log d)$. Previous analyses of constructions of RIP matrices supporting fast matrix-vector multiplies, such as the sampled discrete Fourier matrix, required m to be larger by roughly a $\log k$ factor.

Supporting fast matrix-vector multiplication is useful for iterative recovery algorithms which repeatedly multiply by Φ or Φ^*. Furthermore, our construction, together with a connection between RIP matrices and the Johnson-Lindenstrauss lemma in [Krahmer-Ward, SIAM. J. Math. Anal. 2011], implies fast Johnson-Lindenstrauss embeddings with asymptotically fewer rows than previously known.

Our approach is a simple twist on previous constructions. Rather than choosing the rows for the embedding matrix to be rows sampled from some larger structured matrix (such as the discrete Fourier transform or a random circulant matrix), we instead choose each row of the embedding matrix to be a linear combination of a small number of rows of the original matrix, with random sign flips as coefficients. The main tool in our analysis is a recent bound for the supremum of certain types of Rademacher chaos processes in [Krahmer-Mendelson-Rauhut, arXiv abs/1207.0235].

1 Introduction

The goal of *compressed sensing* [12, 24] is to efficiently reconstruct sparse, high-dimensional signals from a small set of linear measurements. We say that a $x \in \mathbb{C}^d$ is *k-sparse* if $\|x\|_0 \le k$, where $\|x\|_0$ denotes the number of non-zero entries. The idea is that if x is guaranteed to be sparse or nearly sparse (that is, close to a sparse vector), then we should be able to recover it with far fewer than d measurements. Organizing the measurements as the rows of a matrix $\Phi \in \mathbb{C}^{m \times d}$, one wants an efficient algorithm $\mathcal{R}$ which approximately recovers a signal $x \in \mathbb{C}^d$ from the measurements Φx; that is, $\|\mathcal{R}(\Phi x) - x\|_2$ should be small. There are several goals in the design of Φ and $\mathcal{R}$. We would

*Institute for Advanced Study. minilek@ias.edu. Supported by NSF CCF-0832797 and NSF DMS-1128155.

†MIT. ecprice@mit.edu. Work supported by an NSF Graduate Research Fellowship and a Simons Fellowship

‡University of Michigan, Ann Arbor. wootters@umich.edu. Supported by NSF CCF-0743372 and NSF CCF-1161233.

like $m \ll d$ to be as small as possible, so that Φx can be interpreted as a compression of x. We also ask that the recovery algorithm $\mathcal{R}$ be efficient, and satisfy a reasonable recovery guarantee when x is close to a sparse vector.

The recovery guarantee most popular in the literature is the ℓ_2/ℓ_1 *guarantee*, which compares the error between x and the recovery $\mathcal{R}(\Phi x)$ to the error between x and the best k-sparse approximation of x. More precisely, to satisfy the ℓ_2/ℓ_1 guarantee there must exist a constant C such that for every x, $\mathcal{R}(\Phi x)$ satisfies

$$\|\mathcal{R}(\Phi x) - x\|_2 \leq \frac{C}{\sqrt{k}} \cdot \inf_{\substack{y \in \mathbb{C}^d \\ \|y\|_0 \leq k}} \|x - y\|_1. \tag{1}$$

The value of m and the pair $\Phi, \mathcal{R}$ can depend on d and k. Above, $\|\cdot\|_p$ denotes the ℓ_p norm $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$ and $\|x\|_0$ denotes the number of non-zero entries of x.

In this work, we will be concerned with a sufficient condition for the ℓ_2/ℓ_1 guarantee, known as the $(\varepsilon, 2k)$ *restricted isometry property*, or $(\varepsilon, 2k)$-*RIP*. We say that a matrix $\Phi \in \mathbb{C}^{m \times d}$ has the (ε, k)-RIP if

$$\forall x \in \mathbb{C}^d, \ \|x\|_0 \leq k \Rightarrow (1-\varepsilon)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\varepsilon)\|x\|_2^2. \tag{2}$$

It is known that if Φ satisfies the (ε, k)-RIP for $\varepsilon < \sqrt{2} - 1$, then Φ enables the ℓ_2/ℓ_1 guarantee for some constant C [11, 13, 14]. Furthermore, this guarantee is achievable by efficient methods such as solving a linear program [13, 17, 25].

In this work, we construct matrices Φ which satisfy the RIP with few rows, and which additionally support fast matrix-vector multiplication. The speed of the encoding time is important not just for encoding x as Φx, but also for the reconstruction of x. Aside from linear programming, there are several iterative algorithms for recovering x from Φx when Φ satisfies the RIP: for example Iterative Hard Thresholding [8], Gradient Descent with Sparsification [29], CoSaMP [44], Hard Thresholding Pursuit [27], Orthogonal Matching Pursuit [54], Stagewise OMP (StOMP) [26], and Regularized OMP (ROMP) [45, 46]. All these algorithms have running times essentially bounded by the number of iterations (which is usually logarithmic in d and an error parameter) times the running time required to perform a matrix-vector multiply with either Φ or Φ^*, and so it is important that this operation be fast.

If we do not require fast matrix-vector multiplication, it is known that RIP matrices exist with $m = \Theta(k \log(d/k))$. For example, any matrix with i.i.d. Gaussian or subgaussian entries suffices [7, 15, 42]. This is known to be optimal even for the ℓ_2/ℓ_1 recovery problem itself via a connection to Gelfand widths [30, 37] (see a discussion in [7, Section 3]), and is even required to obtain a weaker randomized guarantee [23]. However, for such matrices, naïve matrix-vector multiplication requires time $O(dm)$. Ideally, for the applications above, this would instead be nearly linear in d. This has caused a search for RIP matrices that support fast matrix-vector multiplication, leading to constructions that unfortunately require m to be larger than the optimal by several logarithmic factors. We discuss previous work in closing this gap, and our contribution, in more detail in Section 1.2 below.

1.1 Johnson-Lindenstrauss

The Johnson-Lindenstrauss (JL) lemma of [34] is related to the RIP, and, as we will see below, our constructions of RIP matrices will imply constructions of Johnson-Lindenstrauss transforms with fast embedding time. The JL lemma states that there is a way to embed N points in ℓ_2^d into a linear subspace of dimension approximately $\log N$, with very little distortion. [1]

Lemma 1. *For any $0 < \varepsilon < 1/2$ and any $x_1, \ldots, x_N \in \mathbb{R}^d$, there exists a linear map $A \in \mathbb{R}^{m \times d}$ for $m = O(\varepsilon^{-2} \log N)$ such that for all $1 \leq 1 < j \leq N$,*

$$(1-\varepsilon)\|x_i - x_j\|_2 \leq \|Ax_i - Ax_j\|_2 \leq (1+\varepsilon)\|x_i - x_j\|_2.$$

For any fixed set of vectors $x_1, \ldots, x_N$, we call a matrix A as in the lemma an *ε-JL matrix* for that set. It is known that there are sets of N vectors for which $m = \Omega((\varepsilon^{-2}/\log(1/\varepsilon))\log N)$ is required [5]. In fact, this bound holds for any, not necessarily linear, embedding into ℓ_2^m.

The JL lemma is a useful tool for speeding up solutions to several problems in high-dimensional computational geometry; see for example [32,55]. Often, one has an algorithm which is fast in terms of the number of points but slow as a function of dimension: a good strategy to approximate a solution quickly is to first reduce the input dimension via the JL lemma before running the algorithm. Recently dimensionality reduction via linear maps has also found applications in approximate numerical algebra problems such as linear regression and low-rank approximation [19, 20, 43, 47, 52], and for the k-means clustering problem [9]. Going back to our original problem, the JL lemma also implies the existence of (ε, k)-RIP matrices with $O(\varepsilon^{-2} k \log(d/k))$ rows [7].

Due to its algorithmic importance, it is of interest to obtain JL matrices which allow for fast embedding time, i.e. for which the matrix-vector product Ax can be computed quickly. Paralleling the situation with the RIP, if we do not require that A support fast matrix-vector multiplication, there are many constructions of dense matrices A which are JL matrices with high probability [1, 6, 22, 28, 33, 34, 41]. For example, we may take A to have i.i.d. Gaussian or subgaussian entries. However, for such A matrix-vector multiplication takes time $O(dm)$, where as before we would like it to be nearly linear in d. As with the RIP, if we require this embedding time, there is gap of several logarithmic factors between the upper and lower bounds on the target dimension m. We review previous work and state our contributions on this gap below.

1.2 Previous Work on Fast RIP/JL, and Our Contribution

Above, we saw the importance of constructing RIP and JL matrices which not only have few rows but also support fast matrix-vector multiplication. Below, we review previous work in this direction. We then state our contributions and improvements, which are summarized in Figure 1.

The best known construction of RIP matrices with fast multiplication come from either subsampled Fourier matrices (or related constructions) or from partial circulant matrices. Candès and Tao showed in [15] that a matrix whose rows are $m = O(k \log^6 d)$ random rows from the Fourier matrix satisfies the $(O(1), k)$-RIP with positive probability. The analysis of Rudelson and Vershynin [51] and an optimization of it by Cheraghchi, Guruswami, and Velingker [18] improved the number of rows required for the (ε, k)-RIP to $m = O(\varepsilon^{-2} k \log d \log^3 k)$. For circulant matrices, initial works required $m \gg k^{3/2}$ to obtain the (ε, k)-RIP [31, 50]; Krahmer, Mendelson and Rauhut [38] recently improved the number of rows required to $m = O(\varepsilon^{-2} k \log^2 d \log^2 k)$.

The first work on JL matrices with fast multiplication was by Ailon and Chazelle [2], which had $m = O(\varepsilon^{-2} \log N)$ rows and embedding time $O(d \log d + m^3)$. In certain applications N can

[1] The JL lemma is most commonly stated over $\mathbb{R}$, so we state it this way here. However, as in [39], all of our results extend to complex vectors and complex matrices.

be exponentially large in a parameter of interest, e.g. when one wants to preserve the geometry of an entire subspace for numerical linear algebra [19, 52] or k-means clustering [9], or the set of all sparse vectors in compressed sensing [7]. Thus, while the number of rows in this construction is optimal, for some applications it is important to improve the dependence on m in the running time. Ailon and Liberty [3] improved the running time to $O(d \log m + m^{2+\gamma})$ for any desired $\gamma > 0$ (with the same number of rows), and more recently the same authors gave a construction with $m = O(\varepsilon^{-4} \log N \log^4 d)$ supporting matrix-vector multiplies in time $O(d \log d)$ [4]. Krahmer and Ward [39] improved the target dimension to $m = O(\varepsilon^{-2} \log N \log^4 d)$.

This last improvement of [39] is actually a more general result. Specifically, they showed that, when the columns are multiplied by independent random signs, any $(O(\varepsilon), O(\log N))$-RIP matrix becomes an ε-JL matrix for a fixed set of N vectors with probability $1 - N^{-\Omega(1)}$. Since we saw above that sampling $O(\varepsilon^{-2} k \log d \log^3 k)$ rows from the discrete Fourier or Hadamard matrix satisfies (ε, k)-RIP with constant probability, conditioning on this event and applying the result of [39] implies a JL matrix with $m = O(\varepsilon^{-2} \log N \log d \log^3(\log N)) = O(\varepsilon^{-2} \log N \log^4 d)$ and embedding time $O(d \log d)$. We will use the same method to obtain fast JL matrices from our constructions of RIP matrices.

Another way to obtain JL matrices which support fast matrix-vector multiplication is to construct sparse JL matrices [10, 21, 35, 36, 56]. These constructions allow for very fast multiplication Ax when the vector x is itself sparse. However, these constructions have an $\Omega(\varepsilon)$ fraction of nonzero entries, and it is known that any JL transform with $O(\varepsilon^{-2} \log N)$ rows requires an $\Omega(\varepsilon / \log(1/\varepsilon))$ fraction of nonzero entries [48]. Thus, for constant ε and dense x, multiplication still requires time $\Theta(dm)$.

In this work we propose and analyze a new method for constructing RIP matrices that support fast matrix-vector multiplication. Loosely speaking, our method takes any “good” ensemble of RIP matrices, and produces an ensemble of RIP matrices with fewer rows by multiplying by a suitable hash matrix. We can apply our method to either subsampled Fourier matrices or partial circulant matrices to obtain our improved RIP matrices.

Our construction follows a natural intuition. For example, let A be the discrete Fourier matrix, and suppose that S is an $m \times d$ matrix with i.i.d. Rademacher entries, appropriately normalized. If $m = \Theta(\varepsilon^{-2} k \log(d/k))$, then SA satisfies the (ε, k)-RIP with high probability, because S has the RIP, and A is an isometry. Unfortunately, this construction has slow matrix-vector multiplication time. On the other hand, if S' is an extremely sparse random sign matrix, with only one non-zero per row, then $S'A$ is a subsampled Fourier matrix, supporting fast multiplication. Unfortunately, in order to show that $S'A$ satisfies the RIP with high probability, m must be increased by $\mathrm{polylog}(k)$ factors. This raises the question: can we get the best of both worlds? How sparse must the sign matrix S be to ensure RIP with few rows, and can it be sparse enough to maintain fast matrix-vector multiplication? In some sense, this question, and our results, connects the two lines of research—structured matrices and sparse matrices—on fast JL matrices mentioned above. Our results imply we can improve the number of rows over previous work by using such a sparse sign matrix with only $\mathrm{polylog}(d)$ non-zeroes per row.

Our Main Contribution: We give randomized constructions of (ε, k)-RIP matrices with $m = O(\varepsilon^{-2} k \log d \log^2(k \log d))$ and which support matrix-vector multiplication in time $O(d \log d) + m \cdot \log^{O(1)} d$. When combined with [39], we obtain a JL matrix with a number of rows $m = O(\varepsilon^{-2} \log N \log d \log^2((\log N) \log d)) = O(\varepsilon^{-2} \log N \log^3 d)$ and same embedding time. Thus for

Ensemble	# rows m needed for RIP	Matrix-vector multiplication time	Restrictions	Reference
Partial Fourier	$O(\varepsilon^{-2} k \log d \log^3 k)$	$O(d \log d)$		[18, 51]
Partial Circulant	$O(\varepsilon^{-2} k \log^2 d \log^2 k)$	$O(d \log m)$		[38]
Hash $\times$ Partial Fourier	$O(\varepsilon^{-2} k \log d \log^2 (k \log d))$	$O(d \log d) + m \operatorname{polylog} d$	$k \geq \log^{2.5} m$	this work
Hash $\times$ Partial Circulant	$O(\varepsilon^{-2} k \log d \log^2 (k \log d))$	$O(d \log m) + m \operatorname{polylog} d$	$k \geq \log^2 m$	this work

Figure 1: Table of results.

both RIP and JL, our constructions support fast matrix-vector multiply using the fewest rows known.

Our RIP and JL matrices maintain the $O(d \log d)$ running time of the sampled discrete Fourier matrix as long as $k < d/\operatorname{polylog} d$, and never have multiplication time larger than $d \cdot \log^{O(1)} d$ even for k as large as d. Our results are given in Figure 1.

We remark that the restrictions $k \geq \operatorname{polylog} m$ in Figure 1 can be eliminated as long as ε is not too small, because in this case it is already known how to obtain optimal RIP matrices with fast multiplication for small k. More precisely, the Fast Johnson-Lindenstrauss Transform of [2], combined with [7], give an (ε, k)-RIP matrix with $m = O(\varepsilon^{-2} k \log(d/k))$ rows that supports matrix-vector multiplies in time $O(d \log d)$ as long a $k \leq \varepsilon^{2/3} d^{1/3} / \operatorname{polylog} d$. Meanwhile, our restrictions in Figure 1 require $k \geq \operatorname{polylog} m$. Thus, the only case when neither our result nor the results of [2, 7] applies occurs when $\varepsilon < (\operatorname{polylog} d)/\sqrt{d}$. We note that when $\varepsilon < 1/\sqrt{d}$, it is unknown how to obtain any (ε, k)-RIP matrix with fewer than $d < 1/\varepsilon^2$ rows, and this is already trivially obtained by the identity matrix.

1.3 Notation and Preliminaries

We set some notation. We use $[n]$ to denote the set $\{1, \ldots, n\}$. We use $\|\cdot\|_2$ denote the ℓ_2 norm of a vector, and $\|\cdot\|$, $\|\cdot\|_F$ to denote the operator and Frobenius norms of a matrix, respectively. For a set $\mathcal{S}$ and a norm $\|\cdot\|_X$, $d_{\|\cdot\|_X}(\mathcal{S})$ denotes the diameter of $\mathcal{S}$ with respect to $\|\cdot\|_X$. The set of k-sparse vectors $x \in \mathbb{C}^d$ with $\|x\|_2 \leq 1$ is denoted T_k. In addition to $O(\cdot)$ notation, for two functions f, g, we use the shorthand $f \lesssim g$ (resp. $\gtrsim$) to indicate that $f \leq Cg$ (resp. $\geq$) for some absolute constant C. We use $f \approx g$ to mean $cf \leq g \leq Cf$ for some constants c, C. For clarity, we have made no attempt to optimize the values of the constants in our analyses.

Once we define the randomized construction of our RIP matrix Φ, we will control $|\|\Phi x\|_2^2 - \|x\|_2^2|$ uniformly over T_k, and thus will need some tools for controlling the supremum of a stochastic process on a compact set. For a metric space (T, d), the *δ-covering number* $\mathcal{N}(T, d, \delta)$ is the size of the smallest δ-net of T with respect to the metric d. One way to control a stochastic process on T is simply to union bound over a sufficiently fine net of T; a more powerful way to control stochastic processes, due to Talagrand, is through the γ_2 functional [53].

Definition 2. *For a metric space (T, d), an* admissible sequence *of T is a sequence of nets*

$A_1, A_2, \ldots$ *of* T *so that* $|A_n| \leq 2^{2^n}$*. Then*

$$\gamma_2(T, d) := \inf \sup_{t \in T} \sum_{n=1}^{\infty} 2^{n/2} d(A_n, t),$$

where the infimum is taken over all admissible sequences $\{A_n\}$*.*

Intuitively, $\gamma_2(T, d)$ measures how "clustered" T is with respect to d: if T is very clustered, then the union bound over nets above can be improved by a chaining argument. A similar idea is used in Dudley's integral inequality [40, Theorem 11.1], and indeed they are related (see [53], Section 1.2) by

$$\gamma_2(T, d) \lesssim \int_0^{\mathrm{diam}_d(T)} \sqrt{\log \mathcal{N}(T, d, u)}\, du. \tag{3}$$

It is this latter form that will be useful to us.

1.4 Organization

In Section 2 we define our construction and give an overview of our techniques. We also state our most general theorem, Theorem 6, which gives a recipe for turning a "good" ensemble of RIP matrices into an ensemble of RIP matrices with fewer rows. In Section 3, we apply Theorem 6 to obtain the results listed in Figure 1. Finally, we prove Theorem 6 in Sections 4 and 5.

2 Technical Overview

Our construction is actually a general method for turning any "good" RIP matrix with a suboptimal number of rows into an RIP matrix with fewer rows. Many previous constructions of RIP matrices involve beginning with an appropriately structured matrix (a DFT or Hadamard matrix, or a circulant matrix, for example), and keeping only a subset of the rows. In this work we propose a simple twist on this idea: each row of our new matrix is a linear combination of a small number of rows from the original matrix, with random sign flips as the coefficients. Formally, we define our construction as follows.

Let $\mathcal{A}_M$ be a distribution on $M \times d$ matrices, defined for all M, and fix parameters m and B. Define the injective function $h : [m] \times [B] \to [mB]$ as $h(b, i) = B(b-1) + i$ to partition $[mB]$ into m buckets of size B, so $h(b, i)$ denotes the i^{th} element in bucket b. We draw a matrix A from $\mathcal{A}_{mB}$, and then construct our $m \times d$ matrix $\Phi(A)$ by using h to hash the rows of A into m buckets of size B.

Definition 3 (Our construction)**.** *Let* $\mathcal{A}_M$ *be as above, and fix parameters* m *and* B*. Define a new distribution on* $m \times d$ *matrices by constructing a matrix* $\Phi \in \mathbb{C}^{m \times d}$ *as follows.*

1. *Draw* $A \sim \mathcal{A}_{mB}$*, and let* a_i *denote the rows of* A*.*
2. *For each* $(b, i) \in [m] \times [B]$*, choose a sign* $\sigma_{b,i} \in \{\pm 1\}$ *independently, uniformly at random.*
3. *For* $b = 1, \ldots, m$ *let*

$$\varphi_b = \sum_{i \in [B]} \sigma_{b,i} a_{h(b,i)},$$

and let $\Phi = \Phi(A, \sigma)$ *be the matrix with rows* φ_b*.*

We use A_b to denote the $B \times d$ matrix with rows $a_{h(b,i)}$ for $i \in [B]$.

Equivalently, Φ may be obtained by writing $\Phi = HA$, where $A \sim \mathcal{A}_{mB}$, and H is the $m \times mB$ random matrix with columns indexed by $(b,i) \in [m] \times [B]$, so that

$$H_{j,(b,i)} = \begin{cases} \sigma_{b,i} & b = j \\ 0 & b \neq j \end{cases}.$$

Note that there are two sources of randomness in the construction of Φ: there is the choice of $A \sim \mathcal{A}_{mB}$, and also the choice of the sign flips which determine the matrix H. Our RIP matrix will be the appropriately normalized matrix $\Phi/\sqrt{mB}$.

We consider two example distributions for $\mathcal{A}_M$. First, we consider a bounded orthogonal ensemble.

Definition 4 (Bounded orthogonal ensembles)**.** *Let $U \in \mathbb{C}^{d\times d}$ be any unitary matrix with $|U_{ij}| \leq 1$ for all entries U_{ij} of U. Let u_i denote the i^{th} row of U. A matrix $A \in \mathbb{C}^{M\times d}$ is drawn from the* bounded orthogonal ensemble *associated with U as follows. Select, independently and uniformly at random, a multi-set $\Omega = \{t_1, \ldots, t_M\}$ with $t_i \in [d]$. Then let $A \in \mathbb{C}^{M\times d}$ be the matrix with rows $u_{t_1}, \ldots, u_{t_M}$.*

Popular choices (and our choices) for U include the d-dimensional discrete Fourier transform (resulting in the *Fourier ensemble*), or the $d\times d$ Hadamard matrix, both of which support $O(d \log d)$ time matrix-vector multiplication.

The second family we consider is the partial circulant ensemble.

Definition 5 (Partial Circulant Ensemble)**.** *For $z \in \mathbb{C}^d$, the* circulant matrix *$H_z \in \mathbb{C}^{d\times d}$ is given by $H_z x = z * x$, where $*$ denotes convolution. Fix $\Omega \subset [d]$ of size M arbitrarily. A matrix A is drawn from the* partial circulant ensemble *as follows. Choose $\varepsilon \in \{\pm 1\}^d$ uniformly at random, and let A be the rows of H_ε indexed by Ω.*

As long as the original matrix ensemble $\mathcal{A}$ supports fast matrix-vector multiplication, so does the resulting matrix Φ. Indeed, writing $\Phi x = HAx$ as above, we observe that there are mB nonzero entries in H, so computing the product HAx takes time $O(mB)$, plus the time it takes to compute Ax. When A is drawn from the partial Fourier ensemble, Ax may be computed in time $O(d \log d)$ via the fast Fourier transform. We will choose $B = \mathrm{polylog}(d)$, and so Φx may be computed in time $O(d \log d + m \,\mathrm{polylog}\, d)$. When A is the partial circulant ensemble, Ax may be computed in time $d \log(mB)$ by breaking it up into $d/(mB)$ blocks, each of which is a $mB \times mB$ Toeplitz matrix supporting matrix-vector multiplication in time $O(mB \log(mB))$. Thus, in this case Φx may be computed in time $O(d \log(mB) + mB) = O(d \log m) + m \,\mathrm{polylog}\, d$.

Having established the "multiplication time" column of Figure 1, we turn to the more difficult task of establishing the bounds on m, the number of rows. We note that $\Phi/\sqrt{mB}$ has the (ε, k)-RIP if and only if

$$\sup_{x\in T_k} \left| \frac{1}{mB} \|\Phi x\|_2^2 - \|x\|_2^2 \right| \leq \varepsilon,$$

and so our goal will be to establish bounds on $\sup_{x\in T_k} \left| \|\Phi x\|_2^2 / (mB) - \|x\|_2^2 \right|$. We will show that if $\mathcal{A}$ satisfies certain properties, then in expectation this quantity is small. Specifically we require the following two conditions. First, we require a random matrix from $\mathcal{A}$ to have the RIP with a reasonable, though perhaps suboptimal, number of rows:

$$\mathbb{E}_{A\sim\mathcal{A}} \sup_{x\in T_k} \left| \frac{1}{M} \|Ax\|_2^2 - \|x\|_2^2 \right| \lesssim \sqrt{\frac{L}{M}} \tag{$\star$}$$

for some quantity L, for suitably large $M > M_0$.

Second, the matrices A_b whose rows are the rows of A indexed by $h(b,i)$ for $i \in [B]$ should be well-behaved. Define (†) to be the event that

$$\max_{b\in[m]} \sup_{x\in T_s} \|A_b x\|_2 \le \ell(s) \tag{†}$$

for some function $\ell(s)$ and all $s \le 2k$. We require that (†) happen with constant probability:

$$\mathbb{P}_{A\sim\mathcal{A}}\left[(\dagger) \text{ holds}\right] \ge 7/8. \tag{$\star\star$}$$

for some sufficiently small function ℓ.

As long as these two requirements on $\mathcal{A}$ are satisfied, and all matrices in the support of $\mathcal{A}$ have entries of bounded magnitude, the construction of Definition 3 yields a RIP matrix, with appropriate parameters. The following is our most general theorem.

Theorem 6. *Fix $\varepsilon \in (0,1)$, and fix integers m and B. Let $\mathcal{A} = \mathcal{A}_{mB}$ be a distribution on $mB \times d$ matrices so that $\|a_i\|_\infty \le 1$ almost surely for all rows a_i of $A \sim \mathcal{A}$. Suppose that ($\star$) holds with*

$$L \le mB\varepsilon^2,$$

and $M = mB > M_0$. Suppose further that ($\star\star$) holds, with

$$\ell(s) \le Q_1\sqrt{B} + Q_2\sqrt{s}$$

and that

$$B \ge \max\{Q_2^2 \log^2 m, Q_1^2 \log m \log k\}, \ \text{and } k \ge Q_1^2 \log^2 m.$$

Finally, suppose that $m > m_0$, for

$$m_0 = \frac{k \log d \log^2(Bk)}{\varepsilon^2}.$$

Let Φ be drawn from the distribution of Definition 3. Then

$$\sup_{x\in T_k} \left| \frac{1}{mB} \|\Phi x\|_2^2 - \|x\|_2^2 \right| \lesssim \varepsilon,$$

that is, $\frac{1}{\sqrt{mB}}\Phi$ satisfies the $(O(\varepsilon), k)$-RIP, with $3/4$ probability.

In Section 3, we will show how to use Theorem 6 to prove the results reported in Figure 1, but first we will outline the intuition of the proof of Theorem 6.

By construction, the expectation of $\|\Phi x\|_2^2$ over the sign flips σ is simply $\|Ax\|_2^2$, and ($\star$) guarantees that this expectation is under control, uniformly over $x \in T_k$. The trick is that A has mB rows, rather than m, and this provides slack to handle the fact that the guarantee ($\star$) is not optimal.

The problem is then to argue that for all $x \in T_k$, $\|\Phi x\|_2^2$ is close to its expectation. The proof of Theorem 6 proceeds in two steps. First, we condition on A and control the deviation

$$\mathop{\mathbb{E}}_{\sigma} \sup_{x \in T_k} \left| \|\Phi x\|_2^2 - \mathop{\mathbb{E}}_{\sigma} \|\Phi x\|_2^2 \right|. \tag{4}$$

Second, we take the expectation with respect to $A \sim \mathcal{A}_{mB}$.

In Theorem 11 we carry out the first step and bound the deviation (4) by Talagrand's γ_2 functional $\gamma_2(T_k, \|\cdot\|_X)$, where $\|x\|_X := \max_b \|A_b x\|_2$ is a norm which measures the contribution to $\|\Phi x\|_2$ of the worst bucket b of the partition function h. Our strategy is to write $\|\Phi x\|_2^2$ as $\|X(x)\sigma\|_2^2$, for an appropriate matrix $X(x)$ that depends on A. Finally we use a result of Krahmer, Mendelson, and Rauhut [38] to control the Rademacher chaos, obtaining an expression in terms of $\gamma_2(T_k, \|\cdot\|_X)$.

In the second step, we unfix A, and $\gamma_2(T_k, \|\cdot\|_X)$ becomes a random variable. In Theorem 12, we show that, as long as $(\star\star)$ holds, $\gamma_2(T_k, \|\cdot\|_X)$ is small with high probability over the choice of $A \sim \mathcal{A}_{mB}$. By (3), it is sufficient to bound the covering numbers $\mathcal{N}(T_k, \|\cdot\|_X, u)$. This is similar to [51], which must bound the same $\mathcal{N}(T_k, \|\cdot\|_X, u)$ but in a setting where $B = 1$. Both papers use Maurey's empirical method to relate the covering number to $\mathbb{E}[\max_b \|A_b g\|_2]$ for a Gaussian process g. But while [51] loses a $\sqrt{\log m}$ factor in a union bound over b, we only lose a constant factor as long as $B \geq \operatorname{polylog} d$. This difference is what gives our $\log k$ improvement in m. It is also the most technical piece of our proof, and is presented in Section 5.

Finally, we put all of the pieces together. As long as mB is large enough and the condition $(\star)$ holds, $\mathbb{E}_\sigma \|\Phi x\|_2 / \sqrt{mB}$ will be close to $\|x\|_2$ in expectation over A. At the same time as long as the condition $(\star\star)$ holds, the deviation (4) is small in expectation over $A \sim \mathcal{A}_{mB}$. Choosing B appropriately controls the restricted isometry constant of Φ, at the cost of slightly increasing the embedding time.

3 Main Results

Before we prove Theorem 6, let us show how we may use it to conclude the results in Figure 1. To do this, we must compute L and $\ell(s)$ from the conditions $(\star)$ and $(\star\star)$, when $\mathcal{A}$ is the Fourier ensemble (or any bounded orthogonal ensemble), and when $\mathcal{A}$ is the partial circulant ensemble.

3.1 Bounded orthogonal ensembles

Suppose $\mathcal{A}$ is a bounded orthogonal ensemble. The RIP analysis of [18, 51] shows

$$\mathop{\mathbb{E}}_{A \sim \mathcal{A}} \sup_{x \in T_k} \left| \frac{1}{M} \|Ax\|_2^2 - \|x\|_2^2 \right| \lesssim \sqrt{\frac{k \log^3 k \log d}{M}},$$

provided that $M \gtrsim k \log^3 k \log d$, so we may take $L \lesssim k \log^3 k \log d$. Further, the analysis of [51] (see Lemma 17) implies that

$$\mathbb{P}_{A \sim \mathcal{A}} \left[\exists s \in [2k] : \max_{b \in [m]} \sup_{x \in T_s} \|A_b x\|_2 \geq \ell(s) \right] \leq 2km \max_{s \in [2k]} \mathbb{P}_{A \sim \mathcal{A}} \left[\sup_{x \in T_s} \|A_1 x\|_2 \geq \ell(s) \right] \leq 1/8$$

when

$$\ell(s) \eqsim \log^{1/4}(m)\sqrt{B} + \log^{1/4}(m)\sqrt{s \log^2(k) \log(d) \log(B)}.$$

Thus, we may take $Q_1 \lesssim \log^{1/4} m$ and $Q_2 \lesssim \log^{1/4}(m)\log(k)\sqrt{\log(d)\log(B)} \lesssim \log^{2.5}(d)$ With these parameter settings, Theorem 6 implies the following theorem.

Theorem 7. *Let $\varepsilon \in (0,1)$. Let $\mathcal{A}$ be a bounded orthogonal ensemble (for example, the Fourier ensemble), and suppose that Φ is as in Definition 3. Further suppose $B \geq \log^{6.5} d$ and $k \geq \log^{2.5} m$. Then for some value*

$$m = O\left(\frac{k \log d \log^2(k \log d)}{\varepsilon^2}\right),$$

we have that

$$\sup_{x \in T_k} \left| \|\Phi x\|_2^2 - \|x\|_2^2 \right| \leq \varepsilon$$

with 3/4 probability.

3.2 Circulant Matrices

Suppose that $\mathcal{A}$ is the partial circulant ensemble. By the analysis in [38],

$$\mathop{\mathbb{E}}_{A \sim \mathcal{A}} \sup_{x \in T_k} \left| \frac{1}{M} \|Ax\|_2^2 - \|x\|_2^2 \right| \lesssim \sqrt{\frac{k \log^2 k \log^2 d}{M}},$$

for $M \gtrsim k \log^2 k \log^2 d$. Concentration also follows from the analysis in [38], as a corollary of Theorem 10 (see [38, Theorem 4.1]).

Lemma 8. *(Implicit in [38])*

$$\mathbb{P}_{A \sim \mathcal{A}} \left[\exists s \in [2k] : \max_{b \in [m]} \sup_{x \in T_s} \|A_b x\|_2 \geq \ell(s) \right] \leq \frac{1}{8}$$

when

$$\ell(s) \approx \sqrt{B} + \sqrt{s} \log k \log d.$$

Thus, we may take $Q_1 \lesssim 1$ and $Q_2 \lesssim \log k \log d$. Then Theorem 6 implies the following theorem.

Theorem 9. *Let $\varepsilon \in (0,1)$. Let $\mathcal{A}$ be the partial circulant ensemble, and suppose Φ is constructed as in Definition 3. Further suppose $B \geq \log^2 m \log^2 k \log^2 d$ and $k \geq \log^2 m$. Then, for some value*

$$m = O\left(\frac{k \log d \log^2(k \log d)}{\varepsilon^2}\right),$$

we have that, as long as $m < d/B$,

$$\sup_{x \in T_k} \left| \|\Phi x\|_2^2 - \|x\|_2^2 \right| \leq \varepsilon$$

with 3/4 probability.

We remark that the condition $m \leq d/B$ does not actually effect the results reported in Figure 1. Indeed, if $mB > d$, we may artificially increase d to $d' = mB$ by embedding T_k in $\mathbb{C}^{d'}$ by zero-padding. Applying Theorem 9 with $d = d'$ implies an RIP matrix with $O(\varepsilon^{-2} k \log d' \log^2(k \log d))$ rows and embedding time $O(d' \log d') + m \operatorname{polylog} d'$. Because $B = \operatorname{polylog} d$, we have $d' = d \operatorname{polylog}(d)$, and there is no asymptotic loss in m by extending d to d'. Further, in this parameter regime, $d' \log d' = mB \log d' = m \operatorname{polylog} d$.

4 Proof of Theorem 6

We will use the following theorem from [38].

Theorem 10. *[38, Theorem 1.4] Let $\mathcal{S} \subset \mathbb{C}^{m\times M}$ be a symmetric set of matrices, $\mathcal{S} = -\mathcal{S}$. Let $\sigma \in \{\pm 1\}^M$ uniformly at random. Then*

$$\mathbb{E} \sup_{X\in\mathcal{S}} \left| \|X\sigma\|_2^2 - \mathbb{E}\|X\sigma\|_2^2 \right| \lesssim \left(d_F(\mathcal{S})\gamma_2(\mathcal{S}, \|\cdot\|) + \gamma_2^2(\mathcal{S}, \|\cdot\|) \right) =: E'.$$

Furthermore, for all $t > 0$,

$$\mathbb{P}\left[\sup_{X\in\mathcal{S}} \left| \|X\sigma\|_2^2 - \mathbb{E}\|X\sigma\|_2^2 \right| > C_1 E' + t \right] \le 2\exp\left(-C_2 \min\left\{ \frac{t^2}{V^2}, \frac{t}{U} \right\} \right),$$

where C_1 and C_2 are constants,

$$V = d_{2\to 2}(\mathcal{S})(\gamma_2(\mathcal{S}, \|\cdot\|) + d_F(\mathcal{S})),$$

and

$$U = d_{2\to 2}^2(\mathcal{S}).$$

The first step in proving Theorem 6 is to bound the restricted isometry constant of Φ in terms of the γ_2 functional, removing the dependence on σ.

Theorem 11. *Suppose $\mathcal{A} = \mathcal{A}_M$ is a distribution on $M \times d$ matrices so that $(\star)$ holds, and let Φ be as in Definition 3. Then*

$$\mathbb{E} \sup_{x\in T_k} \left| \frac{1}{mB} \|\Phi x\|_2^2 - \|x\|_2^2 \right| \lesssim \frac{1}{mB} \left(\mathbb{E}_A \sup_{x\in T_k} \|Ax\|_2 \gamma_2(T_k, \|\cdot\|_X) + \mathbb{E}_A \gamma_2^2(T_k, \|\cdot\|_X) \right) + \sqrt{\frac{L}{mB}}. \tag{5}$$

where

$$\|x\|_X := \max_{b\in[m]} \|A_b x\|_2 .$$

Proof. Let $H(b) = \{ h(b,i) : i \in [B] \}$ be the multiset of indices of the rows of A in bucket b, and as above let A_b denote the $B \times d$ matrix whose rows are indexed by $H(b)$. Let $\sigma_b = \sum_{i=1}^B \sigma_{b,i} e_i$ denote the vector of sign flips associated with bucket b. Notice that, by construction, conditioning on $A \sim \mathcal{A}$, we have

$$\mathbb{E}_\sigma \|\Phi x\|_2^2 = \|Ax\|_2^2 , \tag{6}$$

and so

$$\begin{aligned}
&\mathbb{E} \sup_{x\in T_k} \left| \frac{1}{mB} \|\Phi x\|_2^2 - \|x\|^2 \right| \\
&\quad\le \mathbb{E}_A \left[\frac{1}{mB} \mathbb{E}_\sigma \sup_{x\in T_k} \left| \|\Phi x\|_2^2 - \mathbb{E}_\sigma \|\Phi x\|_2^2 \right| + \sup_{x\in T_k} \left| \frac{1}{mB} \mathbb{E}_\sigma \|\Phi x\|_2^2 - \|x\|_2^2 \right| \right] \\
&\quad= \frac{1}{mB} \mathbb{E}_A \mathbb{E}_\sigma \sup_{x\in T_k} \left| \|\Phi x\|_2^2 - \|Ax\|_2^2 \right| + \mathbb{E}_A \sup_{x\in T_k} \left| \frac{1}{mB} \|Ax\|_2^2 - \|x\|_2^2 \right| \\
&\quad\lesssim \frac{1}{mB} \mathbb{E}_A \mathbb{E}_\sigma \sup_{x\in T_k} \left| \|\Phi x\|_2^2 - \|Ax\|_2^2 \right| + \sqrt{\frac{L}{mB}},
\end{aligned} \tag{7}$$

where we have used ($\star$) in the last line and (6) in the penultimate line.

Condition on the choice of A until further notice, and consider the first term. We may write

$$E := \mathop{\mathbb{E}}_{\sigma} \sup_{x \in T_k} \left| \|\Phi x\|_2^2 - \mathop{\mathbb{E}}_{\sigma} \|\Phi x\|_2^2 \right| = \mathop{\mathbb{E}}_{\sigma} \sup_{x \in T_k} \left| \sum_b |\langle \sigma_b, A_b x \rangle|^2 - \mathop{\mathbb{E}}_{\sigma} \sum_b |\langle \sigma_b, A_b x \rangle|^2 \right|.$$

Now, we apply Theorem 10 to $\mathcal{S} = \{X(x) \in \mathbb{C}^{m \times mB} \mid x \in T_k\}$, where $X(x)$ is defined as follows:

$$X(x) = \begin{bmatrix} -(A_1 x)^* - & 0 & 0 & \cdots & 0 \\ 0 & -(A_2 x)^* - & 0 & \cdots & 0 \\ 0 & 0 & -(A_3 x)^* - & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & -(A_m x)^* - \end{bmatrix}.$$

Let σ be the vector in $\{-1,1\}^M$ defined as $(\sigma_1^*, \ldots, \sigma_m^*)^*$. By construction, $\|X(x)\sigma\|_2^2 = \sum_b |\langle \sigma_b, A_b x \rangle|^2$, and so by Theorem 10, it suffices to control $d_F(\mathcal{S})$ and $\gamma_2(\mathcal{S}, \|\cdot\|)$. The Frobenius norm of $X(x)$ is

$$\|X(x)\|_F^2 = \sum_{b \in [m]} \|A_b x\|_2^2 = \|Ax\|_2^2.$$

For the γ_2 term, notice that for any $x, y \in T_k$,

$$\|X(x) - X(y)\| = \max_{b \in [m]} \|A_b(x-y)\|_2 = \|x - y\|_X.$$

Thus, $\gamma_2(\mathcal{S}, \|\cdot\|) = \gamma_2(T_k, \|\cdot\|_X)$. Then Theorem 10 implies that

$$E \lesssim \max_{x \in T_k} \|Ax\|_2 \, \gamma_2(T_k, \|\cdot\|_X) + \gamma_2^2(T_k, \|\cdot\|_X).$$

Plugging this into (7), we conclude

$$\mathbb{E} \sup_{x \in T_k} \left| \frac{1}{mB} \|\Phi x\|_2^2 - \|x\|_2^2 \right| \lesssim \frac{1}{mB} \left(\mathop{\mathbb{E}}_{A} \sup_{x \in T_k} \|Ax\|_2 \, \gamma_2(T_k, \|\cdot\|_X) + \mathop{\mathbb{E}}_{A} \gamma_2^2(T_k, \|\cdot\|_X) \right) + \sqrt{\frac{L}{mB}}. \quad (8)$$

■

Theorem 11 leaves us with the task of controlling $\gamma_2(T_k, \|\cdot\|_X)$, which we do in the following theorem.

Theorem 12. *Suppose that A is a matrix such that (†) holds, with*

$$\ell(s) \le Q_1 \sqrt{B} + Q_2 \sqrt{s}.$$

Suppose further that $\|a_i\|_\infty \le 1$ for all i, and suppose that

$$B \ge \max\{Q_2^2 \log^2 m, Q_1^2 \log m \log k\}, \text{ and } k \ge Q_1^2 \log^2 m.$$

Then

$$\gamma_2(T_k, \|\cdot\|_X) \lesssim \sqrt{kB \log d} \cdot \log(Bk).$$

Proof. By (3),

$$\gamma_2(T_k, \|\cdot\|_X) \lesssim \int_{u=0}^{Q} \sqrt{\log \mathcal{N}(T_k, \|\cdot\|_X, u)}\, du, \tag{9}$$

where $Q = \sup_{x\in T_k} \|x\|_X$. Notice that we can bound

$$Q^2 = \sup_{x\in T_k} \max_b \|A_b x\|_2^2 = \sup_{x\in T_k} \max_b \sum_{i\in[B]} \left|\left\langle a_{h(b,i)}, x\right\rangle\right|^2 \le B \sup_{x\in T_k} \|x\|_1^2 \le Bk$$

using the fact that each entry of $a_{h(b,i)}$ has magnitude at most 1. We follow the approach of [51] and estimate the covering number using two nets, one for small u and one for large u.

For small u, we use a standard ℓ_2 net of B_2: we have

$$\|x\|_X \le Q\,\|x\|_2$$

so $\mathcal{N}(T_k, \|\cdot\|_X, u) \le \mathcal{N}(T_k, \|\cdot\|_2, u/Q)$. Observing that T_k is the union of $\binom{d}{k} = \left(\frac{d}{k}\right)^{O(k)}$ copies of B_2^k (the unit ℓ_2-ball of dimension k), we may cover T_k by covering cover each copy of B_2^k with a net of width u/Q. By a standard volume estimate [49, Eqn. (5.7)], the size of each such net is $(1+2Q/u)^k$, and so

$$\sqrt{\log \mathcal{N}(T_k, \|\cdot\|_X, u)} \lesssim \sqrt{k\log(d/k) + k\log(1+2Q/u)} \lesssim \sqrt{k\log(dQ/u)}.$$

For large u the situation is not as simple. We show in Lemma 13 that, as long as (†) holds,

$$\sqrt{\log \mathcal{N}(T_k, \|\cdot\|_X, u)} \lesssim \frac{\sqrt{kB\log d}}{u}.$$

We plug these bounds into (9) and integrate, using the first net for $u \in (0,1)$ and the second for $u > 1$. We find

$$\begin{aligned}
\int_{u=0}^{Q} \sqrt{\log \mathcal{N}(T_k, \max_b \|F_b \cdot\|, u)}\, du &\lesssim \int_{u=0}^{1} \sqrt{k\log(dQ/u)}\, du + \int_{u=1}^{Q} \frac{\sqrt{kB\log d}}{u}\, du \\
&\lesssim \sqrt{k\log(dQ)} + \sqrt{kB\log d}\log Q \\
&\lesssim \sqrt{kB\log d}\log Q \\
&\le \sqrt{kB\log d}\log(Bk)
\end{aligned}$$

as claimed. ■

It remains to put Theorem 11 and Theorem 12 together to prove Theorem 6.

Proof. (Proof of Theorem 6.) We need to show that

$$\Delta := \sup_{x\in T_k} \left|\frac{1}{mB}\|\Phi x\|_2^2 - \|x\|_2^2\right| \lesssim \varepsilon$$

with 3/4 probability. We have by (⋆⋆) that (†) holds with 7/8 probability over $\mathcal{A}$, and we will show that $\Delta \lesssim \varepsilon$ with 7/8 probability when A is drawn from the distribution $\mathcal{A}' = (\mathcal{A} \mid (\dagger) \text{ holds})$. Together, this will imply the conclusion of Theorem 6.

Note that as long as ($\star$) holds for $\mathcal{A}$, ($\star$) holds for $\mathcal{A}'$ as well. Indeed,

$$\mathop{\mathbb{E}}_{A\sim\mathcal{A}'} \sup_{x\in T_k} \left| \frac{1}{mB} \|Ax\|_2^2 - \|x\|_2^2 \right| \le \left(\frac{8}{7}\right) \mathop{\mathbb{E}}_{A\sim\mathcal{A}} \sup_{x\in T_k} \left| \frac{1}{mB} \|Ax\|_2^2 - \|x\|_2^2 \right| \lesssim \varepsilon,$$

so ($\star$) holds for $\mathcal{A}'$. For the rest of the proof, we consider $A \sim \mathcal{A}'$, so we have

$$\frac{1}{\sqrt{mB}} \mathop{\mathbb{E}}_{A} \sup_{x\in T_k} \|Ax\|_2 \le \sqrt{1+O(\varepsilon)} \lesssim 1.$$

Under the parameters of Theorem 6 and because (†) holds for all $A \sim \mathcal{A}'$, Theorem (12) implies

$$\gamma_2(T_k, \|\cdot\|_X) \le \sqrt{kB\log d} \cdot \log(Bk).$$

Then

$$\frac{1}{mB} \mathop{\mathbb{E}}_{A} \left[\sup_{x\in T_k} \|Ax\|_2 \cdot \gamma_2(T_k, \|\cdot\|_X) \right] \lesssim \frac{\sqrt{k\log(d)} \cdot \log(Bk)}{\sqrt{m}} \le \varepsilon.$$

Similarly,

$$\frac{1}{mB} \mathop{\mathbb{E}}_{A} \gamma_2^2(T_k, \|\cdot\|_X) \lesssim \frac{k\log(d)\log^2(Bk)}{m} \le \varepsilon^2.$$

By Theorem 11, and using the above bounds,

$$\begin{aligned}
\mathop{\mathbb{E}}_{A}[\Delta] &\lesssim \frac{1}{mB} \left(\mathop{\mathbb{E}}_{A} \sup_{x\in T_k} \|Ax\|_2 \, \gamma_2(T_k, \|\cdot\|_X) + \mathop{\mathbb{E}}_{A} \gamma_2^2(T_k, \|\cdot\|_X) \right) + \sqrt{\frac{L}{mB}} \\
&\lesssim \varepsilon + \varepsilon^2 + \varepsilon \\
&\lesssim \varepsilon.
\end{aligned}$$

Therefore by Markov's inequality, we have $\Delta \lesssim \varepsilon$ with arbitrarily high constant probability over $A \sim \mathcal{A}'$. In particular, we may adjust the constants so that $\Delta \lesssim \varepsilon$ with probability at least 7/8 over $A \sim \mathcal{A}'$, which was our goal. ■

5 Covering number bound

In this section, we prove the covering number lemma needed for the proof of Theorem 12. Recall the definition $\|x\|_X = \max_{b\in[m]} \|A_b x\|_2$, and that T_k is the set of k-sparse vectors in $\mathbb{C}^d$ with ℓ_2 norm at most 1.

Lemma 13. *Suppose that the conditions of Theorem 12 hold. Then*

$$\mathcal{N}(T_k/\sqrt{k}, \|\cdot\|_X, u) \le (2d+1)^{O(B/u^2)}.$$

We will prove this under the assumption that $x \in T_k$ is real, using only that (†) holds for $s \le k$ and that A has bounded entries. Then by Proposition 16 in the Appendix, we have $\mathcal{N}(T_k/\sqrt{k}, \|\cdot\|_X, u)$ over the complex numbers is less than $\mathcal{N}(T_{2k}/\sqrt{2k}, \|\cdot\|_{\tilde{X}}, u)$ over the reals, where $\|\cdot\|_{\tilde{X}}$ denotes a version of the $\|\cdot\|_X$ for a matrix $\tilde{A}$ of bounded entries that satisfies (†) for $s \leq 2k$. Adjusting the constants by a factor of 2 gives the final result.

As in [51], we use Maurey's empirical method (see [16]). Consider $x \in T_k/\sqrt{k}$, and choose a parameter s. For $i \in [s]$, define a random variable Z_i, so that $Z_i = e_j \mathrm{sign}(x_j)$ with probability $|x_j|$ for all $j \in [d]$, and 0 with probability $1 - \|x\|_1$. Notice that by the assumption that x is real, $\mathrm{sign}(x_j)$ is well defined. Further, because $T_k/\sqrt{k} \subset B_1$, this is a valid probability distribution. We want to show for every x that

$$\mathbb{E}\left\|x - \frac{1}{s}\sum Z_i\right\|_X \lesssim \sqrt{\frac{B}{s}}. \tag{10}$$

This would imply that the right hand side is at most u for $s \lesssim B/u^2$. If this holds, then the set of all possible $\frac{1}{s}\sum Z_i$ forms a u-covering. As there are only $2d+1$ choices for each Z_i, there are only $(2d+1)^s$ different vectors of the form $\frac{1}{s}\sum_{i=1}^s Z_i$. These form a u-covering, so Eq. (10) will imply

$$\mathcal{N}(T_k, \|\cdot\|_X, u) \leq (2d+1)^{O(B/u^2)}.$$

We now show Eq. (10). Draw a Gaussian vector $g \sim N(0, I_s)$, and define

$$\mathcal{G}(x) = \mathop{\mathbb{E}}_{g,Z}\left\|\sum Z_i g_i\right\|_X$$

By a standard symmetrization argument followed by a comparison principle (Lemma 6.3 and Eq. (4.8) in [40] respectively, or the proof of Lemma 3.9 in [51]),

$$\mathbb{E}\left\|x - \frac{1}{s}\sum Z_i\right\|_X \lesssim \frac{1}{s}\mathcal{G}(x),$$

so it suffices to bound $\mathcal{G}(x)$ by $O(\sqrt{Bs})$.

Let $L = \{i : |x_i| > \frac{\log m}{k}\}$ be the set of coordinates of x with "large" value in magnitude. Then

$$\mathcal{G}(x) \leq \mathcal{G}(x_L) + \mathcal{G}(x_{\overline{L}})$$

by partitioning the Z_i into those from L and those from $\overline{L}$ and applying the triangle inequality. Notice that x_L is "spiky" and $x_{\overline{L}}$ is "flat:" more precisely, we have

$$\|x_L\|_1 \leq \frac{1}{\log m} \qquad \text{and} \qquad \|x_{\overline{L}}\|_\infty \leq \frac{\log m}{k}, \tag{11}$$

using Cauchy-Schwarz to bound the ℓ_1 norm. To bound $\mathcal{G}(x_L)$ and $\mathcal{G}(x_{\overline{L}})$ we use the following lemma.

Lemma 14. *Suppose that* (†) *holds. Then the following inequalities hold for all x:*

$$\mathcal{G}(x) \lesssim \sqrt{Bs\,\|x\|_1 \log m} \tag{12}$$

$$\mathcal{G}(x) \lesssim \sqrt{Bs} + \sqrt{\log m}\left(Q_1\sqrt{B} + Q_2\sqrt{\min(k,s)}\right)\sqrt{s\|x\|_\infty + \log k} \tag{13}$$

Proof. Let $\mathbf{Z} \in \{-1, 0, 1\}^{d \times s}$ have columns Z_i, and $Z = \sum_i Z_i$. Then

$$\mathcal{G}(x) = \mathbb{E} \max_{b \in [m]} \|A_b \mathbf{Z} g\|_2$$

Consider $\|A_b \mathbf{Z} g\|_2$ for a single $b \in [m]$. This is a C-Lipschitz function of a Gaussian for $C = \|A_b \mathbf{Z}\|_{2 \to 2}$. Therefore [40, Eq. (1.4)],

$$\mathbb{P}_g[\|A_b \mathbf{Z} g\|_2 > \mathop{\mathbb{E}}_g \|A_b \mathbf{Z} g\|_2 + t \, \|A_b \mathbf{Z}\|_{2 \to 2}] < e^{-\Omega(t^2)}.$$

Hence by a standard computation for subgaussian random variables [40, Eq. (3.13)]),

$$\mathcal{G}(x) \lesssim \mathop{\mathbb{E}}_Z \max_{b \in [m]} \mathop{\mathbb{E}}_g \|A_b \mathbf{Z} g\|_2 + \sqrt{\log m} \, \|A_b \mathbf{Z}\|_{2 \to 2}.$$

Now,

$$\mathop{\mathbb{E}}_g \|A_b \mathbf{Z} g\|_2 \le \sqrt{\mathop{\mathbb{E}}_g \|A_b \mathbf{Z} g\|_2^2} = \|A_b \mathbf{Z}\|_F = \sqrt{B \, \|Z\|_1} \tag{14}$$

and

$$\mathop{\mathbb{E}}_Z \sqrt{B \, \|Z\|_1} \le \sqrt{B \mathop{\mathbb{E}}_Z \|Z\|_1} = \sqrt{Bs \, \|x\|_1} \le \sqrt{Bs}. \tag{15}$$

Thus

$$\mathcal{G}(x) \le \sqrt{Bs \, \|x\|_1} + O\left(\mathop{\mathbb{E}}_Z \max_{b \in [m]} \sqrt{\log m} \, \|A_b \mathbf{Z}\|_{2 \to 2} \right). \tag{16}$$

Thus it suffices to bound $\|A_b \mathbf{Z}\|_{2 \to 2}$ in terms of $\|x\|_1$ and $\|x\|_\infty$. First, we have

$$\|A_b \mathbf{Z}\|_{2 \to 2} \le \|A_b \mathbf{Z}\|_F$$

and so by Equations (14) and (15) we have

$$\mathcal{G}(x) \le \sqrt{Bs \, \|x\|_1 \log m},$$

as desired for Equation (12).

Second, we turn to Equation (13). For a matrix $A \in m \times d$ and a set $S \subset [d]$, let $A|_S$ denote the $m \times d$ matrix with all the columns not indexed by S set to zero. Then, we have

$$\|A_b \mathbf{Z}\|_{2 \to 2} \le \left\| A_b|_{\mathrm{supp}(Z)} \right\|_{2 \to 2} \|\mathbf{Z}\|_{2 \to 2} \le \max_{|S| \le \min(k,s)} \| A_b|_S \|_{2 \to 2} \, \|Z\|_\infty^{1/2}. \tag{17}$$

In the final step, we used the fact that for any matrix A, $\|A\|_{2 \to 2} \le \sqrt{\|A\|_{1 \to 1} \|A\|_{\infty \to \infty}}$ (see Lemma 15 in the Appendix). By the assumption (†) and the choice of ℓ,

$$\max_{b \in [m]} \sup_{x \in T_{\min(k,s)}} \|A_b x\|_2 \le Q_1 \sqrt{B} + Q_2 \sqrt{\min(k,s)},$$

so

$$\max_{b \in [m]} \|A_b \mathbf{Z}\|_{2 \to 2} \le \|Z\|_\infty^{1/2} \left(Q_1 \sqrt{B} + Q_2 \sqrt{\min(k,s)} \right).$$

Finally, we bound $\mathbb{E}_Z \|Z\|_\infty$. By a Chernoff bound, for any $j \in \mathrm{supp}(x)$, we have

$$\mathbb{P}\left[\left|\left(\sum Z_i\right)_j\right| > s\,|x_j| + t\right] \le e^{-\Omega(t)}.$$

Integrating, we have

$$\mathbb{E}\,\|Z\|_\infty \lesssim s\,\|x\|_\infty + \log k.$$

Thus

$$\mathop{\mathbb{E}}_{Z}\max_{b\in[m]} \|A_b\mathbf{Z}\|_{2\to 2} \lesssim (s\,\|x\|_\infty + \log k)^{1/2}\left(Q_1\sqrt{B} + Q_2\sqrt{\min(k,s)}\right).$$

Combining this with Equation (16) gives (13). ■

We return to the proof of Lemma 13. Recall that the goal was to bound

$$\mathcal{G}(x_L) + \mathcal{G}(x_{\overline{L}}) \lesssim \sqrt{Bs}.$$

By (11) and (12), $\mathcal{G}(x_L) \lesssim \sqrt{Bs}$. Furthermore,

$$\begin{aligned}
\mathcal{G}(x_{\overline{L}}) &\lesssim \sqrt{Bs} + \sqrt{\log m}\left(Q_1\sqrt{B} + Q_2\sqrt{\min(k,s)}\right)\sqrt{s\|x_{\overline{L}}\|_\infty + \log k} \\
&\le \sqrt{Bs} + \sqrt{\log m}\left(Q_1\sqrt{B} + Q_2\sqrt{\min(k,s)}\right)\left(\sqrt{\frac{s\log m}{k}} + \sqrt{\log k}\right) \\
&= \sqrt{Bs}\left(1 + Q_1\left(\frac{\log m}{\sqrt{k}} + \sqrt{\frac{\log m \log k}{s}}\right) + Q_2\left(\frac{\sqrt{\min(k,s)}\log m}{\sqrt{kB}} + \sqrt{\frac{\log m\log k}{B}\frac{\min(k,s)}{s}}\right)\right).
\end{aligned}$$

Since we have assumed $B \gtrsim Q_2^2\log^2 m$, the Q_2 term is bounded by a constant. Further, $k \gtrsim Q_1^2\log^2 m$, and $s \ge B \gtrsim Q_1^2 \log m \log k$, and so the Q_1 term is also constant. Thus, we conclude

$$\mathcal{G}(x) \le \mathcal{G}(x_L) + \mathcal{G}(x_{\overline{L}}) \lesssim \sqrt{Bs},$$

which was our goal.

6 Conclusion

In compressed sensing, it is of interest to obtain RIP matrices Φ supporting fast (i.e. nearly linear time) matrix-vector multiplication, with as few rows as possible. Not only does fast multiplication reduce the amount of time it takes to collect measurements, it also speeds up many iterative recovery algorithms, which are based on repeatedly multiplying by Φ or Φ^*. Similarly, because of applications in computational geometry, numerical linear algebra, and others, one wants to obtain JL matrices with few rows and fast matrix-vector multiplication. In this work, we have shown how to construct RIP matrices supporting fast matrix-vector multiplication, with fewer rows than was previously known. Combined with the work of [39], this also implies improved constructions of fast JL matrices.

Our work leaves the obvious open question of removing the two $O(\log(k\log d))$ factors separating our constructions from the lower bounds. It seems that both logarithmic factors come from the estimation (3). It would be interesting to see if they could be removed by more sophisticated chaining techniques such as majorizing measures.

Acknowledgments

We thank Piotr Indyk for suggesting the construction in this work and for asking us whether it yields any stronger guarantees for the restricted isometry property.

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 Comput. Geom.*, 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] R. I. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63(2):161–182, 2006.

[7] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28:253–263, 2008.

[8] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *J. Fourier Anal. Appl.*, 14:629–654, 2008.

[9] C. Boutsidis, A. Zouzias, M. W. Mahoney, and P. Drineas. Stochastic dimensionality reduction for K-means clustering. *CoRR*, abs/1110.2897, 2011.

[10] V. Braverman, R. Ostrovsky, and Y. Rabani. Rademacher chaos, random Eulerian graphs and the sparse Johnson-Lindenstrauss transform. *CoRR*, abs/1011.2590, 2010.

[11] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *C. R. Acad. Sci. Paris*, 346:589–592, 2008.

[12] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, (52):489–509, 2006.

[13] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8), 2006.

[14] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, 2005.

[15] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52:5406–5425, 2006.

[16] B. Carl. Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in Banach spaces. *Ann. Inst. Fourier (Grenoble)*, 35(3):79–118, 1985.

[17] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, 2001.

[18] M. Cheraghchi, V. Guruswami, and A. Velingker. Restricted isometry of Fourier matrices and list decodability of random linear codes. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), to appear*, 2013. full version at CoRR abs/1207.1140.

[19] K. L. Clarkson and D. P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41st ACM Symposium on Theory of Computing (STOC)*, pages 205–214, 2009.

[20] K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. *CoRR*, abs/1207.6365, 2012.

[21] A. Dasgupta, R. Kumar, and T. Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)*, pages 341–350, 2010.

[22] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003.

[23] K. Do Ba, P. Indyk, E. Price, and D. P. Woodruff. Lower bounds for sparse recovery. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1190–1197, 2010.

[24] D. L. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306, 2006.

[25] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inf. Theory*, 52(1), 2006.

[26] D. L. Donoho, Y. Tsaig, I. Drori, and J. Starck. Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit. *IEEE Trans. Inf. Theory*, 58:1094–1121, 2012.

[27] S. Foucart. Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM J. Numer. Anal.*, 49(6):2543–2563, 2011.

[28] P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *J. Comb. Theory. Ser. B*, 44(3):355–362, 1988.

[29] R. Garg and R. Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 337–344, 2009.

[30] A. Y. Garnaev and E. D. Gluskin. On the widths of the Euclidean ball. *Soviet Mathematics Doklady*, 30:200–203, 1984.

[31] J. Haupt, W. U. Bajwa, G. Raz, and R. Nowak. Toeplitz compressed sensing matrices with applications to sparse channel estimation. *IEEE Trans. Inform. Theory*, 56(11):5862–5875, 2010.

[32] P. Indyk. Algorithmic applications of low-distortion geometric embeddings. In *Proceedings of the 42nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 10–33, 2001.

[33] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th ACM Symposium on Theory of Computing (STOC)*, pages 604–613, 1998.

[34] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[35] D. M. Kane and J. Nelson. A derandomized sparse Johnson-Lindenstrauss transform. *CoRR*, abs/1006.3585, 2010.

[36] D. M. Kane and J. Nelson. Sparser Johnson-Lindenstrauss transforms. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1195–1206. SIAM, 2012.

[37] B. S. Kašin. The widths of certain finite-dimensional sets and classes of smooth functions. *Izv. Akad. Nauk SSSR Ser. Mat.*, 41(2):334–351, 478, 1977.

[38] F. Krahmer, S. Mendelson, and H. Rauhut. Suprema of chaos processes and the restricted isometry property. *arXiv*, abs/1207.0235, 2012.

[39] 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.

[40] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

[41] J. Matousek. On variants of the Johnson-Lindenstrauss lemma. *Random Struct. Algorithms*, 33(2):142–156, 2008.

[42] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4):1248–1282, 2007.

[43] X. Meng and M. W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. *CoRR*, abs/1210.3135, 2012.

[44] D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.*, 26:301–332, 2009.

[45] D. Needell and R. Vershynin. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations of Computational Mathematics*, 9(3):317–334, 2009.

[46] D. Needell and R. Vershynin. Signal recovery from inaccurate and incomplete measurements via regularized orthogonal matching pursuit. *IEEE Journal of Selected Topics in Signal Processing*, 4:310–316, 2010.

[47] J. Nelson and H. L. Nguyễn. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. Manuscript, 2012.

[48] J. Nelson and H. L. Nguyễn. Sparsity lower bounds for dimensionality-reducing maps. Manuscript, 2012.

[49] G. Pisier. *The volume of convex bodies and Banach space geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1989.

[50] H. Rauhut, J. Romberg, and J. A. Tropp. Restricted isometries for partial random circulant matrices. *Appl. and Comput. Harmon. Anal.*, 32(2):242–254, 2012.

[51] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

[52] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

[53] M. Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Verlag, 2005.

[54] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Info. Theory*, 53(12):4655–4666, 2007.

[55] S. Vempala. *The random projection method*, volume 65 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. American Mathematical Society, 2004.

[56] K. Q. Weinberger, A. Dasgupta, J. Langford, A. J. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 1113–1120, 2009.

Appendix

Lemma 15. *For any complex matrix A, $\|A\|_{2\to 2}^2 \le \|A\|_{1\to 1} \cdot \|A\|_{\infty\to\infty}$.*

Proof. First we consider the case of Hermitian A, then arbitrary A. For Hermitian A, let λ be the largest (in magnitude) eigenvalue of A and v be the associated eigenvector. We have

$$\|A\|_{1\to 1} \ge \frac{\|Av\|_1}{\|v\|_1} = \frac{\|\lambda v\|_1}{\|v\|_1} = |\lambda| = \|A\|_{2\to 2}.$$

For arbitrary A,

$$\|A\|_{2\to 2}^2 = \|A^*A\|_{2\to 2} \le \|A^*A\|_{1\to 1} \le \|A^*\|_{1\to 1} \cdot \|A\|_{1\to 1} = \|A\|_{\infty\to\infty} \cdot \|A\|_{1\to 1}$$

as desired. In the last inequality we used the fact that $\|\cdot\|_{\infty\to\infty}$ is equal to the largest ℓ_1 norm of any row, and $\|\cdot\|_{1\to 1}$ is equal to the largest ℓ_1 norm of any column. ■

Proposition 16. *Let $f : \mathbb{C}^d \to \mathbb{R}^{2d}$ act entrywise by replacing $a + b\mathbf{i}$ with (a, b). For any integer r, define $F : \mathbb{C}^{r\times d} \to \mathbb{R}^{2r\times 2d}$ to act entrywise by replacing an entry $a + b\mathbf{i}$ by the 2×2 matrix*

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix}.$$

Recall that $T_k \subset \mathbb{C}^d$ is the set of unit norm k-sparse complex vectors, and let $S_s \subset \mathbb{R}^s$ be the set of unit norm s-sparse real vectors. Recall that $\|\cdot\|_X$ is a norm on $\mathbb{C}^d$ given by $\|x\|_X = \max_b \|A_b x\|_2$, and let $\|\cdot\|_{\tilde{X}}$ be a norm on $\mathbb{R}^{2d}$ given by $\|x\|_{\tilde{X}} = \max_b \|F(A_b)x\|_2$. Then

1. *If (†) holds, then $\max_b \sup_{x\in S_s} \|F(A_b)x\|_2 \le \ell(s)$ for $s \le 2k$.*
2. *With $\|\cdot\|_{\tilde{X}}$ as above, we have*

$$\mathcal{N}(T_k, \|\cdot\|_X, u) \le \mathcal{N}(S_{2k}, \|\cdot\|_{\tilde{X}}, u).$$

Proof. By construction, we have $f(Ax) = F(A)f(x)$, and also $\|f(x)\|_2 = \|x\|_2$. Further, $f(T_k) \subset S_{2k}$ and $f^{-1}(S_s) \subset T_s$. Thus, item 1 follows because

$$\max_b \sup_{x\in S_s} \|F(A_b)x\|_2 \le \max_b \sup_{y\in T_s} \|F(A_b)f(y)\|_2 = \max_b \sup_{y\in T_s} \|A_b y\|_2 \le l(s)$$

Similarly, item 2 follows because for any $x, y \in T_k$,

$$\begin{aligned}
\|x - y\|_X &= \max_{b\in[m]} \|A_b(x - y)\|_2 \\
&= \max_{b\in[m]} \|F(A_b)f(x - y)\|_2 \\
&= \max_{b\in[m]} \|F(A_b)(f(x) - f(y))\|_2 \\
&= \|f(x) - f(y)\|_{\tilde{X}}.
\end{aligned}$$

Hence

$$N(T_k, \|\cdot\|_X, u) = N(f(T_k), \|\cdot\|_{\tilde{X}}, u) \le N(S_{2k}, \|\cdot\|_{\tilde{X}}, u).$$

■

Lemma 17. *Let $\mathcal{F}$ denote the $d \times d$ Fourier matrix. Let Ω with $|\Omega| = B$ be a random multiset with elements in $[d]$, and for $S \subseteq [d]$ let $\mathcal{F}_{\Omega\times S}$ denote the $|\Omega| \times |S|$ matrix whose rows are the rows of $\mathcal{F}$ in Ω, restricted to the columns in S. Then for any $t > 1$,*

$$\max_{|S|=k} \|\mathcal{F}_{\Omega\times S}\| \lesssim \sqrt{t(B + k\beta)}$$

with probability at least

$$1 - O\left(\exp\left(-\min\left\{t^2, t\beta\right\}\right)\right),$$

where

$$\beta = \log^2 k \log d \log B.$$

Proof. (Implicit in [51]). Let $X = \sup_{|S|=k} \left\| I_k - \frac{1}{B}\mathcal{F}^*_{\Omega\times S}\mathcal{F}_{\Omega\times S} \right\|$, where I_k is the $k \times k$ identity matrix. It is shown in [51] that

$$\mathop{\mathbb{E}}_{\Omega} X \lesssim \sqrt{\frac{k \log^2 k \log d \log B}{B}(\mathbb{E}\, X + 1)} =: \sqrt{\frac{k\beta}{B}(\mathbb{E}\, X + 1)}.$$

This implies that

$$\mathbb{E}\, X \leq 1 + \frac{O(k\beta)}{B} =: \alpha. \tag{18}$$

Indeed, whenever $x^2 \leq A(x+1)$, we have $x < A + 1$ or else we conclude $(A+1)^2 \leq A^2 + 2A$. Let α denote the right hand side of (18). We may plug this expectation into the proof of Theorem 3.9 in [51], and we obtain

$$\mathbb{P}\left[X > Ct\alpha\right] \leq 3\exp(-C' t\alpha B/k) + 2\exp(-t^2)$$

for constants C and C'. In the case $X \leq Ct\alpha$, we have

$$\max_{|S|=k} \|\mathcal{F}_{\Omega\times S}\| \leq \sqrt{B(1 + Ct\alpha)} \leq \sqrt{B} + \sqrt{BCt\alpha},$$

and so we conclude that

$$\max_{|S|=k} \|\mathcal{F}_{\Omega\times S}\| \leq \sqrt{B} + O\left(\sqrt{t(B + k\beta)}\right)$$

with probability at least

$$1 - 3\exp(-C' t(\beta + B/k)) - 2\exp(-t^2).$$

$\blacksquare$