Discrete Comput Geom (2014) 52:780–798
DOI 10.1007/s00454-014-9632-3

Fast and RIP-Optimal Transforms

Nir Ailon · Holger Rauhut

Received: 13 October 2013 / Revised: 27 July 2014 / Accepted: 18 August 2014 /
Published online: 4 September 2014

Abstract We study constructions of $k \times n$ matrices A that both (1) satisfy the restricted isometry property (RIP) at sparsity s with optimal parameters, and (2) are efficient in the sense that only $O(n \log n)$ operations are required to compute Ax given a vector x. Our construction is based on repeated application of independent transformations of the form DH, where H is a Hadamard or Fourier transform and D is a diagonal matrix with random $\{+1, -1\}$ elements on the diagonal, followed by any $k \times n$ matrix of orthonormal rows (e.g. selection of k coordinates). We provide guarantees (1) and (2) for a regime of parameters that is comparable with previous constructions, but using a construction that uses Fourier transforms and diagonal matrices *only*. Our main result can be interpreted as a rate of convergence to a random matrix of a random walk in the orthogonal group, in which each step is obtained by a Fourier transform H followed by a random sign change matrix D. After a few number of steps, the resulting matrix is random enough in the sense that any *arbitrary* selection of rows gives rise to an RIP matrix for, sparsity as high as slightly below $s = \sqrt{n}$, with high probability. The proof uses a bootstrapping technique that, roughly speaking, says that if a matrix A has some suboptimal RIP parameters, then the action of *two steps* in this random walk on this matrix has improved parameters. This idea is interesting in its own right, and may be used to strengthen other constructions.

N. Ailon
Technion Israel Institute of Technology, Haifa, Israel
e-mail: nailon@cs.technion.ac.il

H. Rauhut
Lehrstuhl C für Mathematik (Analysis), RWTH Aachen University,
Templergraben 55, 52056 Aachen, Germany
e-mail: rauhut@mathc.rwth-aachen.de

Keywords Restricted isometry · Johnson–Lindenstrauss transformations · Compressive sensing

1 Introduction

The theory of compressive sensing predicts that sparse vectors can be stably reconstructed from a small number of linear measurements via efficient reconstruction algorithms including ℓ_1-minimization [8,9]. The restricted isometry property (RIP) of the measurement matrix streamlines the analysis of various reconstruction algorithms [5,6,10,11]. All known matrices that satisfy the RIP in the optimal parameter regime (see below for details) are based on randomness. Well known examples include Gaussian and Bernoulli matrices where all entries are independent. Unfortunately, such matrices do not possess any structure and therefore no fast matrix-vector multiplication algorithm. The latter is important for speed-up of recovery algorithms. This article addresses constructions of matrices that satisfy the RIP in the optimal parameter regime and have fast matrix-vector multiplication algorithms. It provides an analysis of such matrices obtained after a few steps in a natural random walk in the orthogonal group.

A vector $x \in \mathbb{C}^n$ is said to be s-sparse if the number of nonzero entries of x is at most s. A matrix $A \in \mathbb{C}^{k\times n}$ satisfies the RIP with respect to parameters (s, δ) if, for all s-sparse vectors $x \in \mathbb{C}^n$,

$$(1-\delta)\|x\|_2 \le \|Ax\|_2 \le (1+\delta)\|x\|_2, \tag{1.1}$$

where $\|\cdot\|_2$ denotes the Euclidean norm.[1] If A satisfies the RIP with parameters $(2s, \delta^*)$ for a suitable $\delta^* < 1$ then a variety of recovery algorithms reconstruct an s-sparse vector exactly from $y = Ax$. Moreover, reconstruction is stable under passing from sparse to approximately sparse vectors and under adding noise on the measurements. The value of δ^* depends only on the reconstruction algorithm [4,6,10,11,24].

It is well known by now [3,5,17] that a Gaussian random matrix (having independent normal distributed entries of variance $1/m$) satisfies the RIP with parameters (s, δ) with probability at least $1 - e^{-c\delta^2 k}$ if

$$k \ge C\delta^{-2} s \ln(n/s),$$

where $c, C > 0$ are universal constants. Using lower bounds for Gelfand widths of ℓ_p-balls for $0 < p \le 1$, it can be shown that k must be at least $C_\delta s \log(n/s)$ for the RIP to hold [12]. It can further be shown [11, Theorem 6.8] that the constant (as a function of δ) satisfies $C_\delta \ge C\delta^{-2}$. Since we will always assume in this paper that $s \le Cn^{1/2}$, $\log(n/s)$ is equivalent to $\log(n)$ up to constants. Hence, we will say that a $k \times n$ matrix is *RIP-optimal* at s if it satisfies the RIP with (s, δ) for

[1] In much of the related literature, the definition of RIP uses *squared* Euclidean norms. The definition (1.1) is, however, more convenient for our purposes. Of course, both versions are equivalent up to a transformation of the parameter δ.

$$\delta = C\sqrt{\frac{s\log n}{k}}.$$

(The reader should keep in mind that for large s, RIP optimality should be defined to hold when k is at most $C\delta^{-2}s\log(n/s)$.)

The restricted isometry property is closely connected to Johnson–Lindenstrauss embeddings. We say that a random $k\times n$ matrix A satisfies the Johnson–Lindenstrauss property (JLP) with parameters (N,δ) if for any set $X\subseteq\mathbb{R}^n$ of cardinality at most N (1.1) holds uniformly for all $x\in X$ with constant probability. It is well known that a matrix of independently drawn subgaussian elements satisfies JLP if $k\geq C\delta^{-2}\log N$. Specializations of this fact to Gaussians and to Bernoulli random variables can be found in [1,13]. The general claim is obtainable by noting that the subgaussian property is the crux of these proofs. If A satisfies JLP with (N,δ) for $k\leq C\delta^{-2}\log N$, then we say that A is JLP-optimal.

The JLP and RIP properties are known to be almost equivalent, in a certain sense. One direction is stated as follows: A (random) $k\times n$ matrix satisfying JLP with (N,δ) satisfies RIP with $(C(\log N)/(\log n),2\delta)$ with constant probability [3,17]. This implies that we can always obtain, with high probability, an RIP-optimal matrix if we know how to draw a JLP-optimal matrix with $k=C\delta^{-2}s\log n$. The derivation of RIP from JLP is a specialization of JLP to a set X consisting of a ε-net of s sparse unit vectors, for $\varepsilon=0.1$ say, which has cardinality $N\leq(Cn)^s$.

The other direction is a remarkable recent result by Krahmer and Ward [14] implying that if A has RIP with $(s,\delta/4)$, then AD has JLP with (N,δ) as long as $N\leq 2^s$, where D is a diagonal matrix with independent random signs (±1) on the diagonal. Notice that from this result, RIP-optimality of A does not imply JLP-optimality of AD, because RIP-optimality implies that the embedding dimension of k is at least $C\delta^{-2}s\log n$, which suffers from an additional factor of $\log n$ compared to the JLP-optimality guarantee bound of $C\delta^{-2}\log N=C\delta^{-2}s$ (for $N=2^s$). From this observation we intuitively conclude the following stipulation:

RIP-optimality is weaker than JLP-optimality, and RIP-optimal constructions are easier to obtain.

One of the results of this paper, roughly speaking, confirms this by providing constructions of RIP-optimal matrices which are simpler than previously known constructions that relied on JLP optimality.

1.1 Known Constructions of RIP-Optimal and JLP-Optimal Matrices

No deterministic RIP-optimal matrix constructions are known. Deterministic constructions of RIP matrices are known only for a grossly suboptimal regime of parameters. See [14] for a nice survey of such constructions.

Of particular interest are RIP or JLP matrices A that are *efficient* in the sense that for any vector $x\in\mathbb{C}^n$, Ax can be computed in time $O(n\log n)$. Such constructions are known for JLP-optimal (and hence also RIP-optimal) matrices as long as $k\leq n^{1/2-\mu}$ for any arbitrarily small μ [2]. This is achieved with the transformation $BHD^{(1)}HD^{(2)}H\ldots HD^{(r)}$, where $D^{(i)}$ are independent random sign diagonal matrices, H are Hadamard transforms and B is a subsampled (and rescaled)

Hadamard transform, where the subset of sampled coordinates is related to a carefully constructed dual binary code, and r is at most C/μ. For larger k, the best efficient constructions satisfying RIP were for a long while due to Rudelson and Vershynin [22] with $k \geq Cs\log^4 n$, namely a factor of $\log^3 n$ away from optimal, see also [20]. It was more recently improved by Nelson et al. to only $\log^2 n$ factors away from optimal [18] (note also an improvement of Rudelson et al.'s result by Cheraghchi et al. [8]). The construction in [22] is a Fourier transform followed by a subsampling operator. Another family of RIP almost optimal matrices with a fast transform algorithm is that of partial random circulant matrices and time-frequency structured random matrices. The best known results in that vein have recently appeared in the work of Krahmer et al. [15], where RIP matrices of almost optimal embedding dimension $k = Cs\log^4 n$ are designed. These constructions improve on previous work in [21].

We should also mention that, in light of the result [14] coupled with [22] (and the more recent improved [18]), a method for obtaining JLP-optimal (and hence, RIP-optimal) efficient constructions for target dimension $k \approx \sqrt{n}/\operatorname{polylog}(n)$ could be obtained by applying a slightly sub-optimal transformation in time $O(n\log n)$ (paying an additional $\operatorname{polylog}(n)$ factor in target dimension), and then reducing the unwanted excess in dimensions by multiplying the result by a naïve JLP matrix, e.g. using a random Gaussian or Bernoulli matrix containing $O(n\log n)$ entries (due to our choice of s). Since a major open problem in this field is obtaining JLP (resp. RIP) optimality efficiently for *any* target dimension (resp. any sparsity), we argue that our approach provides an additional avenue for potentially achieving this goal and is hence important to explore.

1.2 Contribution

We pay particular attention to the aforementioned JLP matrix construction $BHD^{(1)}HD^{(2)}H\ldots HD^{(r)}$ from [2]. The combination HD of a Hadamard transform and a random diagonal sign matrix, repeatedly iterated, can be viewed as a random walk in the orthogonal group. The matrix B there is carefully chosen using an error correcting code, and is not easy to implement in practice.

We show that for the purpose of obtaining efficient RIP-optimality in the regime $s \leq n^{1/2-\mu}$ (the same regime studied in [2]), the matrix B can be replaced with any "reasonable" matrix. More precisely, for this regime, we show that the transformation $PD^{(1)}HD^{(2)}H\ldots D^{(r)}H$ is RIP-optimal and efficient, where the $D^{(i)}$'s are as above, P is an *arbitrary* deterministic matrix with properly normalized, pairwise orthogonal rows and r is at most C/μ. One could even set P to be a subsampling onto the set of first coordinates. No binary code designs are necessary.

Our main proof techniques involve concentration inequalities of vector valued Rademacher chaos of degree 2, together with a bootstrapping argument that, roughly speaking, shows that the RIP parameters of $ADHD'H$ are better than those of A.

We believe that the random walk in the orthogonal group induced by the random steps HD is interesting in its own right, and leave the question of studying stronger convergence rates for this walk to future work.

2 Notation and Main Results

Throughout, the letter C denotes a general global constant, whose value may change from appearance to appearance. The integer n denotes the ambient dimension, $k \leq n$ denotes the embedding dimension, and $\mathbb{C}^n$ denotes the n dimensional complex space with standard inner product. The usual ℓ_p-norms are denoted by $\|\cdot\|_p$, the spectral norm of a matrix A by $\|A\|$ and the Frobenius norm as $\|A\|_F = \sqrt{\text{trace}(A^*A)}$.

We let $H \in \mathbb{C}^{n\times n}$ denote a fixed matrix with the following properties:

(1) H is unitary,
(2) the maximal magnitude of an entry of H is $n^{-1/2}$,
(3) the transformation Hx given a vector $x \in \mathbb{C}^n$ can be computed in time $O(n \log n)$.

Note that any matrix satisfying 1. and 2. is usually called a *bounded orthogonal system*. Both the discrete Fourier matrix and the Walsh-Hadamard matrix are examples of such matrices. Note that the upper bound of $n^{-1/2}$ in Property 2. above could be replaced by $Kn^{-1/2}$ for any constant K, thus encompassing transformations such as the discrete cosine transform (with $K = \sqrt{2}$) with little effect on the guarantees. We have decided to concentrate on the case $K = 1$ for simplicity.

For any vector $z \in \mathbb{C}^n$, D_z denotes a diagonal matrix with the elements of z on the diagonal. For a subset Ω of $\{1, \ldots, n\}$, let $\text{ind}(\Omega) \in \mathbb{R}^n$ denote the vector with 1 at coordinates $i \in \Omega$ and 0 elsewhere. Then define $P_\Omega = D_{\text{ind}(\Omega)}$ and (letting $k = |\Omega|$) $R_\Omega \in \mathbb{R}^{k\times n}$ to be the map that restricts a vector in $\mathbb{R}^n$ to its entries in Ω, i.e., a subsampling operator.

Recall that a $k \times n$ matrix A has the RIP property with respect to parameters (s, δ) where s is an integer and $\delta > 0$, if for any s-sparse unit vector $x \in \mathbb{C}^n$,

$$1 - \delta \leq \|Ax\|_2 \leq 1 + \delta.$$

(Note that we allow $\delta > 1$, unlike typical definitions of RIP). For a fixed sparsity parameter s, we denote by $\delta_s(A)$ the infimum over all δ such that A has the RIP property with respect to parameters (s, δ). We say that A is RIP optimal at a given sparsity parameter s if

$$\delta_s(A) \leq C\sqrt{\frac{s \log n}{k}}. \tag{2.1}$$

A random vector ε with independent entries that take the values ± 1 with equal probability is called a Rademacher vector.

Our first main result provides a simple RIP-optimal matrix with a fast transform algorithm for small sparsities $s = \mathcal{O}(n^{1/3}/\log^{2/3} n)$.

Theorem 2.1 *Let $\varepsilon, \varepsilon' \in \{\pm 1\}^n$ be two independent Rademacher vectors, and let Ω be any subset of $\{1, \ldots, n\}$ of size k. The $k \times n$ random matrix $A = R_\Omega H D_\varepsilon H D_{\varepsilon'} H$ is RIP-optimal for the regime $k \leq \sqrt{n/s}$. More precisely, if*

$$\sqrt{n/s} \geq k \geq C\delta^{-2} s \log n, \tag{2.2}$$

then A satisfies the RIP (1.1) with probability at least $1 - e^{-C\delta^2 k}$. In particular, the conditions on k entail

$$s \leq \frac{C\delta^{4/3} n^{1/3}}{\log^{2/3} n}.$$

Clearly, $A = R_\Omega H D_\varepsilon H D_{\varepsilon'} H x$ can be computed in $\mathcal{O}(n \log n)$ time by assumption on H. Also note that (2.2) implies a restriction on δ for which the result applies, namely

$$\delta \geq c \frac{s^{3/4} \sqrt{\log n}}{n^{1/4}}. \tag{2.3}$$

Our second main result gives an RIP-optimal matrix construction with a fast transform for the enlarged parameter regime $s = O(\sqrt{n}/\log n)$ and $k = O(n/(s \log n))$.

Theorem 2.2 *Assume $s_0 \leq s \log n \leq k \leq \sqrt{n}$ and $\kappa = C\sqrt{\frac{sk \log n}{n}} < 1/2$, where s_0 is a global constant. Let A be an arbitrary $k \times n$ matrix satisfying $AA^* = \frac{n}{k} \mathrm{Id}_k$. Let $r = \lceil \frac{-\log(2\sqrt{n/k})}{\log \kappa} \rceil$, and let $\varepsilon_{(1)}, \ldots, \varepsilon_{(r)} \in \{\pm 1\}^n$ denote independent Rademacher vectors. Then the $k \times n$ matrix*

$$\hat{A} = A D_{\varepsilon_{(1)}} H D_{\varepsilon'_{(1)}} H D_{\varepsilon_{(2)}} H D_{\varepsilon'_{(2)}} H \ldots D_{\varepsilon_{(r+1)}} H D_{\varepsilon'_{(r+1)}} H \tag{2.4}$$

is RIP-optimal with probability at least 0.99, that is, (1.1) holds if $k \geq C\delta^{-2} s \log n$.

In particular, if we strengthen the constraints by requiring $s \leq n^{1/2-\mu}$ for some global $\mu > 0$, and that A is efficient, then $\hat{A}$ is also computationally efficient in the sense that $\hat{A}x$ can be computed in time $\mathcal{O}(n \log n)$.

(The second part of the theorem clearly follows the first, because if $s \leq n^{1/2-\mu}$ then $r = O(1/\mu)$.) The probability 0.99 in the above theorem is arbitrary and can be replaced by any different value in $(0, 1)$. This effects only the constant C. However, we remark that the present proof does not seem to give the optimal dependence of C in terms of the probability bound.

It is presently not clear whether the restrictions $s = \mathcal{O}(n^{1/3}/\ln^{2/3} n)$ and $s = \mathcal{O}(\sqrt{n}/\log n)$ in the above theorems can be removed. In any case, regimes of small s are of interest in many applications of compressive sensing. Nevertheless, we remark that also larger regimes of s may be important, and the extension of our main result to larger s remains an interesting open problem.

Section 3 is dedicated to proving Theorem 2.1, and Sect. 4 proves Theorem 2.2.

3 The Regime $s = \mathcal{O}(n^{1/3}/\log^{2/3} n)$

Our first randomized, computationally efficient, RIP-optimal construction involves three applications of H, two random sign diagonal matrices, and a choice of an *arbitrary* set of k coordinates. Fix $x \in U_s := \{x \in \mathbb{C}^n : \|x\|_2 = 1, x \text{ is } s\text{-sparse}\}$ and let

$\varepsilon, \varepsilon' \in \{\pm 1\}^n$ be two random sign vectors. Consider the following random variable, indexed by x,

$$\alpha(x) = \sqrt{\frac{n}{k}} \left\| P_\Omega H D_\varepsilon H D_{\varepsilon'} H x \right\|_2,$$

where k is the cardinality of Ω. It is not hard to see that $\mathbb{E}[\alpha(x)^2] = 1$. Indeed, denoting $\tilde{x} = H D_{\varepsilon'} H x$ and conditioning on a fixed value of ε', for any $i \in \{1, \ldots, n\}$,

$$\mathbb{E}_\varepsilon \| P_{\{i\}} H D_\varepsilon \tilde{x} \|_2^2 = \|\tilde{x}\|_2^2 / n = 1/n.$$

The random variable $\alpha(x)$ is the norm of a decoupled Rademacher chaos of degree 2. For the sake of notational convenience, we denote, for $i, j \in \{1, \ldots, n\}$,

$$\mathbf{x}_{ij} = \sqrt{\frac{n}{k}} P_\Omega H P_{\{i\}} H P_{\{j\}} H x, \tag{3.1}$$

so that we can conveniently write

$$\alpha(x) = \Big\| \sum_{i=1}^n \sum_{j=1}^n \varepsilon_i \varepsilon'_j \mathbf{x}_{ij} \Big\|_2.$$

By a seminal result of Talagrand [23] (Theorem 1.2), a Rademacher chaos concentrates around its median. We will exploit the following version.

Theorem 3.1 *With a double sequence $\mathbf{x}_{ij}$, $i, j = 1, \ldots, n$, of vectors in $\mathbb{C}^n$ and two independent Rademacher vectors $\varepsilon, \varepsilon' \in \{\pm 1\}^n$ let*

$$\alpha = \Big\| \sum_{i=1}^n \sum_{j=1}^n \varepsilon_i \varepsilon'_j \mathbf{x}_{ij} \Big\|_2.$$

Let M_α be a median of α. For $\mathbf{y} \in \mathbb{C}^n$ introduce the $n \times n$ matrix $B_\mathbf{y} = (\mathbf{y}^* \, \mathbf{x}_{ij})_{i,j=1}^n$ and the parameters

$$U = \sup_{\mathbf{y} \in \mathbb{C}^n, \|\mathbf{y}\|_2 \le 1} \|B_\mathbf{y}\| \tag{3.2}$$

$$V = \mathbb{E} \sup_{\mathbf{y} \in \mathbb{C}^n, \|\mathbf{y}\|_2 \le 1} (\|B_\mathbf{y} \varepsilon\|_2^2 + \|B_\mathbf{y}^* \, \varepsilon'\|_2^2)^{1/2}. \tag{3.3}$$

Then, for $t > 0$,

$$\Pr(|\alpha - M_\alpha| \ge t) \le 2 \exp\left(- C \min\left\{ \frac{t^2}{V^2}, \frac{t}{U} \right\} \right). \tag{3.4}$$

Proof With $A = (\mathbf{x}_{ij})_{i,j=1}^n$ and

$$S = \frac{1}{2}\begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix}, \qquad \tilde{\varepsilon} = \begin{pmatrix} \varepsilon \\ \varepsilon' \end{pmatrix}$$

we can rewrite the decoupled chaos as the coupled symmetric chaos

$$\tilde{\varepsilon}^* \ S\tilde{\varepsilon} = \sum_{i,j=1}^n \varepsilon_i \varepsilon'_j \mathbf{x}_{ij},$$

where matrix multiplication is extended in an obvious way to matrices with vector-valued entries. Observe that S has zero diagonal. Therefore, the claim follows from Theorem 1.2 in [23]. □

We bound the quantity $U = U(x)$ in (3.2), where the $\mathbf{x}_{ij}$ are defined by (3.1). Note that $\sum_{i,j} \alpha_i \beta_j \mathbf{y}^* \ \mathbf{x}_{ij} = \mathbf{y}^* \ P_\Omega H D_\alpha H D_\beta H x$, and hence

$$\begin{aligned}
U &= \sup_{\|\mathbf{y}\|_2, \|\alpha\|_2, \|\beta\|_2 \le 1} \sum_{i=1}^n \sum_{j=1}^n \bar{\alpha}_i \beta_j \mathbf{y}^* \ \mathbf{x}_{ij} \\
&= \sqrt{\frac{n}{k}} \sup_{\mathbf{y},\alpha,\beta} \mathbf{y}^* \ P_\Omega H D_{\alpha^*} \ H D_\beta H x \\
&= \sqrt{\frac{n}{k}} \sup_{\mathbf{y},\alpha,\beta} \alpha^* \ D_{\mathbf{y}^* \ P_\Omega H} H D_{Hx} \beta \\
&\le \sqrt{\frac{n}{k}} \sup_{\mathbf{y},\alpha,\beta} \|\alpha\|_2 \cdot \|D_{\mathbf{y}^* \ P_\Omega H}\| \cdot \|D_{Hx}\| \cdot \|\beta\|_2 \\
&\le \sqrt{\frac{n}{k}} \sup_{\mathbf{y},\beta} \|\mathbf{y}^* \ P_\Omega H\|_\infty \cdot \|Hx\|_\infty \cdot \|\beta\|_2 \\
&\le \sqrt{\frac{n}{k}} \sup_{\mathbf{y}} \|\mathbf{y}^* \ P_\Omega\|_1 n^{-1/2} \cdot \|x\|_1 n^{-1/2} \\
&\le \sqrt{\frac{n}{k}} k^{1/2} n^{-1/2} \|x\|_1 n^{-1/2} \\
&\le \sqrt{s/n}.
\end{aligned} \tag{3.5}$$

To bound V, we define a process $\nu(\mathbf{y})$ as

$$\nu(\mathbf{y}) = \sqrt{\|B_\mathbf{y} \varepsilon\|_2^2 + \|B_\mathbf{y}^* \ \varepsilon'\|_2^2} \tag{3.6}$$

so that $V = \mathbb{E} \sup_{\|\mathbf{y}\| \le 1} \nu(\mathbf{y})$. By the definition of the vectors $\mathbf{x}_{ij}$, it clearly suffices to take the supremum on vectors $\mathbf{y}$ supported on Ω. For any such $\mathbf{y}$, let $\mu_\mathbf{y} = \mathbb{E}\nu(\mathbf{y})$. Jensen's inequality yields

$$\mu_{\nu(\mathbf{y})} \le \sqrt{\mathbb{E}\nu_{\mathbf{y}}^2} \le \sqrt{\frac{2n}{k}} \|D_{\mathbf{y}^* P_\Omega H} H D_{Hx}\|_F$$
$$\le \sqrt{\frac{2n}{k}} \cdot \|H\|_\infty \cdot \|Hx\|_2 \cdot \|H P_\Omega \mathbf{y}\|_2.$$

By definition of the Frobenius norm, together with the fact that the matrix elements of H are bounded above by $n^{-1/2}$ in absolute value, we obtain

$$\mu_{\nu(\mathbf{y})} \le \|\mathbf{y}\|_2 \sqrt{2/k}. \tag{3.7}$$

We use a concentration bound for vector-valued Rademacher sums (tail inequality (1.9) in [16]) to notice that for any $\mathbf{y}$ and $t > 0$,

$$\Pr\big(|\nu(\mathbf{y}) - M_{\nu(\mathbf{y})}| > t\big) \le 4\exp\big(-Ct^2/\sigma_{\nu(\mathbf{y})}^2\big), \tag{3.8}$$

where $M_{\nu(\mathbf{y})}$ is a median of $\nu(\mathbf{y})$ and, with $A = \sqrt{n/k} P_\Omega$,

$$\begin{aligned}
\sigma_{\nu(\mathbf{y})} &= \sup_{\|\beta\|_2^2 + \|\gamma\|_2^2 \le 1} \left(\sum_{j=1}^n \Big| \sum_{i=1}^n \beta_i \mathbf{y}^* \mathbf{x}_{ij} + \gamma_i \mathbf{y}^* \mathbf{x}_{ji} \Big|^2 \right)^{1/2} \\
&\le \sup_{\|\beta\|_2, \|\gamma\|_2 \le 1} \|\mathbf{y}^* A D_\beta H D_{Hx}\|_2 + \|D_{\mathbf{y}^* A} H D_\gamma H x\|_2 \\
&\le \sup_{\substack{\|\beta\|_2, \|\gamma\|_2 \le 1 \\ \|\beta'\|_2, \|\gamma'\|_2 \le 1}} \mathbf{y}^* A D_\beta H D_{\beta'} H x + \mathbf{y}^* A D_\gamma H D_{\gamma'} H x \\
&\le 2 \sup_{\|\beta\|_2, \|\gamma\|_2 \le 1} \mathbf{y}^* A D_\beta H D_\gamma H x \\
&\le 2\sqrt{n/k} \|D_{\mathbf{y}^* P_\Omega H} H D_{Hx}\| \qquad (3.9) \\
&\le 2\|\mathbf{y}\|_2 \cdot \|x\|_1/\sqrt{n} \le 2\sqrt{s/n}. \qquad (3.10)
\end{aligned}$$

Upper bounding the expression (3.9) was done exactly as above when upper bounding U. Using (3.7) and the second part of Lemma 1, we conclude that

$$M_{\nu(\mathbf{y})} \le \mu_{\nu(\mathbf{y})} + C\sigma_{\mathbf{y}} \le \|\mathbf{y}\|_2 \sqrt{2/k} + C\sqrt{s/n}. \tag{3.11}$$

We will now bound V.To that end, we use a general epsilon-net argument. Given a subset T of a Euclidean space, we recall that a set $\mathcal{N} \subset T$ is called μ-separated if $\|\mathbf{y} - \mathbf{y}'\|_2 > \mu$ for all $\mathbf{y}, \mathbf{y}' \in \mathcal{N}$, $\mathbf{y} \ne \mathbf{y}'$. It is called maximally μ-separated if no additional vector can be added to $\mathcal{N}$ in a μ-separated position.

Lemma 3.2 *Let $\gamma : \mathbb{C}^m \mapsto \mathbb{R}^+$ be a seminorm, and let $\mathcal{N}$ denote a maximal μ-separated set of Euclidean unit vectors $\mathbf{y} \in \mathbb{C}^m$ for some $\mu < 1$. Let*

$$S = \sup_{\mathbf{y} \in \mathcal{N}} \gamma(\mathbf{y}) \quad I = \inf_{\mathbf{y} \in \mathcal{N}} \gamma(\mathbf{y}).$$

Then

$$\sup_{\|\mathbf{y}\|_2=1} \gamma(\mathbf{y}) \le \frac{1}{1-\mu} S \tag{3.12}$$

$$\inf_{\|\mathbf{y}\|_2=1} \gamma(\mathbf{y}) \ge I - \frac{\mu}{1-\mu} S. \tag{3.13}$$

In particular, if $\kappa := \sup_{\mathbf{y}\in\mathcal{N}} |\gamma(\mathbf{y}) - 1|$, *then*

$$\sup_{\|\mathbf{y}\|_2=1} |\gamma(\mathbf{y}) - 1| \le \frac{\kappa+\mu}{1-\mu}. \tag{3.14}$$

The proof of the bound (3.12) is contained in [25] (Lemma 5.3). The proof of (3.13) is similar and implicitly contained in [3] (inside the proof of Lemma 5.1).

We use the lemma by constructing a maximal $\eta = 0.1$-separated set $\mathcal{N}$ of $S_\Omega := \{\mathbf{y} \in \mathbb{C}^n : \|\mathbf{y}\|_2 = 1, \operatorname{supp} \mathbf{y} \subset \Omega\}$. Using a standard volumetric argument (see e.g. [20, Proposition 10.1])

$$\operatorname{card} \mathcal{N} \le (1 + 2/\eta)^{2k} = 21^{2k}. \tag{3.15}$$

We also notice that $\nu(\cdot)$ is a seminorm for any fixed $\varepsilon, \varepsilon'$. Using (3.12),

$$\sup_{\substack{\mathbf{y}:\mathbf{y}=P_\Omega \mathbf{y} \\ \|\mathbf{y}\|_2=1}} \nu(\mathbf{y}) \le \frac{1}{1-0.1} \sup_{\mathbf{y}'\in\mathcal{N}} \nu(\mathbf{y}').$$

Taking expectations yields

$$\mathbb{E} \sup_{\substack{\mathbf{y}:\mathbf{y}=P_\Omega \mathbf{y} \\ \|\mathbf{y}\|_2=1}} \nu(\mathbf{y}) \le 1.2 \mathbb{E} \sup_{\mathbf{y}'\in\mathcal{N}} \nu(\mathbf{y}'). \tag{3.16}$$

The expectation on the right hand side can now be bounded, in light of (3.8) and using Lemma 5.2 (with $\sigma_i' = 0$), as follows:

$$\mathbb{E} \sup_{\mathbf{y}'\in\mathcal{N}} \nu(\mathbf{y}') \le \sup_{\mathbf{y}} M_{\nu(\mathbf{y})} + \sup_{\mathbf{y}} \sigma_{\nu(\mathbf{y})} \sqrt{\log \operatorname{card} \mathcal{N}}. \tag{3.17}$$

Together with (3.11), (3.15), and (3.10), this implies

$$\mathbb{E} \sup_{\mathbf{y}'\in\mathcal{N}} \nu(\mathbf{y}') \le C\big(\sqrt{1/k} + \sqrt{s/n} + \sqrt{\frac{s}{n}k}\big). \tag{3.18}$$

If we now assume that $k \le \frac{1}{2}\sqrt{\frac{n}{s}}$, then we conclude from (3.16), (3.18) and (3.5) that

$$V = \mathbb{E} \sup_{\substack{\mathbf{y}:\mathbf{y}=P_\Omega \mathbf{y} \\ \|\mathbf{y}\|=1}} \nu(\mathbf{y}) \le C/\sqrt{k} \quad \text{and} \quad U \le 2/k.$$

Plugging these upper bounds into (3.4) we conclude that for all $0 < t \le 1$

$$\Pr(|\alpha(x) - M_{\alpha(x)}| \ge t) \le 2\exp\big(-Ct^2k\big)$$

(because $\min\{t^2/V^2, t/U\} = Ct^2/V^2$ for $0 < t \le 1$ and for the derived values of U, V). But we also know, using Lemma 1, that $|\sqrt{\mathbb{E}\alpha(x)^2} - M_{\alpha(x)}| \le C/\sqrt{k}$. Recalling that $\mathbb{E}\alpha(x)^2 = 1$ and combining, we conclude that for $t \le 1$,

$$\Pr(|\alpha(x) - 1| \ge t) \le C\exp\big\{-Ct^2k\big\}. \tag{3.19}$$

Now we fix a support set $T \subset \{1, \dots, N\}$ of size s and consider the complex unit sphere restricted to T, i.e., $S_T = \{x \in \mathbb{C}^n : \|x\|_2 = 1, \operatorname{supp} x \in T\}$. Let $\mathcal{N}_T$ be a maximal η-separated set of S_T, which has cardinality at most $(1 + 2/\eta)^{2s}$ by (3.15). By a union bound, we have

$$\begin{aligned}\Pr\Big(\max_{x\in\mathcal{N}_T} |\alpha(x) - 1| \ge t\Big) &\le (1 + 2/\eta)^{2s} Ce^{-Ct^2k} \\ &= C\exp\big(-Ct^2k + 2s\ln(1 + 2/\eta)\big).\end{aligned}$$

It follows from Lemma 3.2 that

$$\begin{aligned}&\Pr\Big(\max_{x\in U_s} |\alpha(x) - 1| \ge (t + \eta)/(1 - \eta)\Big) \\ &= \Pr\Big(\max_{\#T=s}\max_{x\in S_T} |\alpha(x) - 1| \ge (t + \eta)/(1 - \eta)\Big) \\ &\le \sum_{\#T=s} \Pr\Big(\max_{x\in\mathcal{N}_T} |\alpha(x) - 1| \ge t\Big) \le \binom{n}{s} C\exp\big(-Ct^2k + 2s\ln(1 + 2/\eta)\big) \\ &\le \exp\big(-Ct^2k + 2s\ln(1 + 2/\eta) + s\ln(en/s)\big).\end{aligned}$$

Choosing $\eta = \min\{t, 0.5\}$ we conclude that $\max_{x\in U_s} |\alpha(x) - 1| \le 4t$ with probability at least $1 - \varepsilon$ if

$$k \ge Ct^2(s(\ln(1 + 2/t) + \ln(en/s))) + \ln(C\varepsilon^{-1}).$$

Replacing t by $\delta/4$ and noting that (2.3) implies $\ln(1 + 8/\delta) \le c\ln(en/s)$ concludes the proof.

4 The Regime $s = \mathcal{O}(\sqrt{n}/\log n)$

We now use the idea developed in the previous section to bootstrap an efficient RIP-optimal construction for a larger regime.

Let A be a fixed $k \times n$ matrix with pairwise orthogonal rows of Euclidean length $\sqrt{n/k}$ each. Namely,

$$AA^* = \frac{n}{k}\,\mathrm{Id}_k\,. \tag{4.1}$$

The strategy will be to improve the RIP parameter $\delta_s(A)$ of A by replacing A with $\tilde{A} = AD_\varepsilon H D_{\varepsilon'} H$. To analyze the (random) RIP parameter $\delta_s(\tilde{A})$, fix an s-sparse unit vector $x \in \mathbb{C}^n$. Now define the random variable

$$\alpha(x) = \|\tilde{A}x\|_2.$$

As before, we note that $\alpha(x)$ is the norm of a decoupled Rademacher chaos of degree 2 in a k-dimensional Hilbert space, which can be conveniently written as $\alpha(x) = \|\sum_{i=1}^n \sum_{j=1}^n \varepsilon_i \varepsilon'_j \mathbf{x}_{ij}\|_2$, where

$$\mathbf{x}_{ij} = AHP_{\{i\}}HP_{\{j\}}Hx.$$

As in the previous discussion, we bound the invariants of interest U and V as defined in (3.2) and (3.3), respectively. We start by bounding U. By definition,

$$U \leq \sup_{\substack{\mathbf{y},\alpha,\beta \\ \|\mathbf{y}\|_2, \|\alpha\|_2, \|\beta\|_2 \leq 1}} \mathbf{y}^* AD_\alpha H D_\beta Hx. \tag{4.2}$$

Notice now that since x is s-sparse by assumption, we have $\|x\|_1 \leq \sqrt{s}$ and hence $\|Hx\|_\infty \leq \sqrt{s/n}$ so that

$$\|D_\beta Hx\|_2 \leq \sqrt{s/n} \quad \text{and} \quad \|D_\beta Hx\|_1 \leq 1.$$

The right hand side inequality is due to Cauchy–Schwarz. In turn, this implies

$$\|HD_\beta Hx\|_\infty \leq 1/\sqrt{n} \quad \text{and} \quad \|HD_\beta Hx\|_2 \leq \sqrt{s/n}.$$

Therefore,

$$\|D_\alpha HD_\beta Hx\|_2 \leq 1/\sqrt{n} \quad \text{and} \quad \|D_\alpha HD_\beta Hx\|_1 \leq \sqrt{s/n}. \tag{4.3}$$

Again, the right hand side inequality is due to Cauchy–Schwarz. We need the following simple lemma, see also [19, Lemma 3.1].

Lemma 4.1 *Let $w \in \mathbb{C}^n$ be such that $\|w\|_1 \leq \sqrt{s}\rho$ and $\|w\|_2 \leq \rho$ for some integer s and number $\rho > 0$. Then there exist $N = \lceil n/s \rceil$ vectors $w^{(1)}, \ldots, w^{(N)}$ such that $w^{(i)}$ is s-sparse for each i, $w = \sum_{i=1}^N w^{(i)}$ and $\sum_{i=1}^N \|w^{(i)}\|_2 \leq 2\rho$.*

Proof Assume wlog that the coordinates $w_1, \ldots, w_n$ of w are sorted so that $|w_1| \geq |w_2| \geq \cdots \geq |w_n|$. For $i = 1, \ldots, N$, let

$$w^{(i)} = (\underbrace{0, \ldots, 0}_{(i-1)s \text{ times}}, w_{(i-1)s+1}, \ldots, w_{is}, 0, \ldots, 0)^* \in \mathbb{C}^n$$

and $\alpha_i = \|w^{(i)}\|_\infty = |w_{(i-1)s+1}|$. Clearly $w = \sum_{i=1}^N w^{(i)}$, and we have:

$$\sum \|w^{(i)}\|_2 = \|w^{(1)}\|_2 + \sum_{i=2}^N \|w^{(i)}\|_2 \leq \rho + \sum_{i=2}^N \alpha_i \sqrt{s}. \tag{4.4}$$

But now notice that for all $i = 1, \ldots, N$, $\|w^{(i)}\|_1 \geq \alpha_{i+1} s$ (where we define $\alpha_{N+1} = 0$), hence we conclude, using the assumptions, that

$$\sum_{i=2}^{N} \alpha_i s \leq \sum_{i=1}^{N} \|w^{(i)}\|_1 = \|w\|_1 \leq \sqrt{s}\rho.$$

Therefore, $\sum_{i=1}^{N} \alpha_i \sqrt{s} \leq \rho$. Together with (4.4), this implies the lemma. □

Remark 4.2 Note that the technique of grouping together monotonically decreasing coordinates of a vector in blocks is rather standard in compressive sensing, see for example [6] or [14].

From (4.3) we conclude that $D_\alpha H D_\beta H x$ can be decomposed as $\sum_{i=1}^{N} w^{(i)}$ as in the lemma with $\rho = n^{-1/2}$. For brevity, we will henceforth use δ to denote $\delta_s(A)$. By definition, for each $i = 1, \ldots, N$, $\|A w^{(i)}\|_2 \leq \|w^{(i)}\|_2 (1 + \delta)$. By the lemma's premise, and using the triangle inequality, this implies

$$U \leq 2(1 + \delta)/\sqrt{n}. \tag{4.5}$$

Bounding V is done as follows. We define the process ν as

$$\nu(\mathbf{y}) = \left(\|B_{\mathbf{y}} \varepsilon\|_2^2 + \|B_{\mathbf{y}}^* \ \varepsilon'\|_2^2\right)^{1/2},$$

where $(B_{\mathbf{y}})_{ij} = \mathbf{y}^* \ \mathbf{x}_{ij}$, over the set $\{\mathbf{y} \in \mathbb{R}^k : \|\mathbf{y}\|_2 \leq 1\}$, so that $V = \mathbb{E} \sup_{\mathbf{y}} \nu(\mathbf{y})$. For any $\mathbf{y}$, $\nu(\mathbf{y})$ is a Rademacher sum in k-dimensional Hilbert space. Thus, we can use (3.8) to conclude that for all $\mathbf{y}$,

$$\Pr\left(\nu(\mathbf{y}) > M_{\nu(\mathbf{y})} + t\right) \leq 4 \exp(-t^2/8\sigma_{\nu(\mathbf{y})}^2),$$

where $M_{\nu(\mathbf{y})}$ is a median of $\nu(\mathbf{y})$ and

$$\begin{aligned}
\sigma_{\nu(\mathbf{y})} &= \sup_{\|\beta\|^2 + \|\gamma\|^2 \leq 1} \left(\sum_{j=1}^{n} \Big| \sum_{i=1}^{n} \beta_i \mathbf{y}^* \ \mathbf{x}_{ij} + \gamma_i \mathbf{y}^* \ \mathbf{x}_{ji} \Big|^2 \right)^{1/2} \\
&\leq \sup_{\|\beta\|_2, \|\gamma\|_2 \leq 1} \|\mathbf{y}^* \ A D_\beta H D_{Hx}\|_2 + \|D_{\mathbf{y}^* \ A} H D_\gamma H x\|_2 \\
&\leq \sup_{\substack{\|\beta\|_2, \|\gamma\|_2 \leq 1 \\ \|\beta'\|_2, \|\gamma'\|_2 \leq 1}} \mathbf{y}^* \ A D_\beta H D_{\beta'} H x + \mathbf{y}^* \ A D_\gamma H D_{\gamma'} H x \\
&\leq 2 \sup_{\|\beta\|_2, \|\gamma\|_2 \leq 1} \mathbf{y}^* \ A D_\beta H D_\gamma H x \qquad (4.6) \\
&\leq 4(1 + \delta)/\sqrt{n}. \qquad (4.7)
\end{aligned}$$

For the last inequality, notice that (4.6) is bounded by twice the RHS of (4.2), and recall the derivation of (4.5). Using the first part of Lemma 1, we conclude that for all $\mathbf{y}$ such that $\|\mathbf{y}\| = 1$,

$$M_{\nu(\mathbf{y})} \leq \mu_{\nu(\mathbf{y})} + C(1 + \delta)/\sqrt{n}, \tag{4.8}$$

where $\mu_{\nu(\mathbf{y})}$ is the expectation of $\nu(\mathbf{y})$. Jensen's inequality yields

$$\begin{aligned}\mu_{\nu(\mathbf{y})} &\le \|D_{H^*\,A^*\,\mathbf{y}} H D_{Hx}\|_F \le \|H^*\,A^*\,\mathbf{y}\|_2 \cdot \|Hx\|_2/\sqrt{n} = \|A^*\,\mathbf{y}\|_2 \cdot \|x\|_2/\sqrt{n} \\ &\le (\sqrt{n/k})/\sqrt{n} = k^{-1/2}.\end{aligned}$$

Again we notice that for any fixed $\varepsilon, \varepsilon'$, ν is a seminorm. As before, let $\mathcal{N}$ denote a maximal 0.1-separated set of Euclidean unit vectors in $\mathbb{C}^k$. Hence by (3.12) in Lemma 3.2, for any fixed $\varepsilon, \varepsilon'$,

$$\sup_{\|\mathbf{y}\|_2=1} \nu(\mathbf{y}) \le 1.2 \sup_{\mathbf{y}\in\mathcal{N}} \nu(\mathbf{y}).$$

Taking expectation on both sides and using Lemma 5.2 to bound the right hand side (recalling that the cardinality of $\mathcal{N}$ is at most 21^{2k}), we conclude

$$V = \mathbb{E} \sup_{\|\mathbf{y}\|_2=1} \nu(\mathbf{y}) \le \sup_{\|\mathbf{y}\|_2=1} M_{\nu(\mathbf{y})} + C\sqrt{k} \sup_{\|\nu(\mathbf{y})\|_2=1} \sigma_{\nu(\mathbf{y})}.$$

By (4.7) and by our bound on $M_{\nu(\mathbf{y})}$, we conclude

$$\begin{aligned}V &\le k^{-1/2} + C(1+\delta)/\sqrt{n} + C\sqrt{k}(1+\delta)/\sqrt{n} \\ &\le (k^{-1/2} + (1+\delta)C\sqrt{k/n}).\end{aligned} \tag{4.9}$$

From (3.4), (4.5), (4.9) we then conclude that for all $t > 0$,

$$\begin{aligned}&\Pr(|\alpha(x) - M_{\alpha(x)}| \ge t) \\ &\le 2\exp\left(-C\min\left\{\frac{t^2}{((1+\delta)\sqrt{k/n} + 1/\sqrt{k})^2}, \frac{t\sqrt{n}}{(1+\delta)}\right\}\right).\end{aligned} \tag{4.10}$$

Using the first part of Lemma 1 and recalling that $\mathbb{E}\alpha(x)^2 = 1$ this implies that

$$\left|M_{\alpha(x)} - 1\right| \le C((1+\delta)\sqrt{k/n} + 1/\sqrt{k}), \tag{4.11}$$

We now use the net-technique to pass to the supremum over all s-sparse unit vectors to provide an estimate of the restricted isometry constant. For each subset $T \subset \{1, \ldots, n\}$ of cardinality s we consider a maximal μ-separated net $\mathcal{N}_T$ of the unit sphere S_T of complex unit length vectors with support T where $\mu = 1/k$. By (3.15) and since $k < \sqrt{n/s}$ and $s \le \sqrt{n}$, the union $\mathcal{N} = \bigcup_{\#T=s} \mathcal{N}_T$ is bounded in size by

$$\#\mathcal{N} \le \binom{N}{s}(1+2/\eta)^{2s} \le (en/s)^s(1+2\sqrt{n/s})^{2s} \le \exp(Cs\log n). \tag{4.12}$$

Using Lemma 5.2, (4.10) and (4.11), we conclude that

$$\begin{aligned}\mathbb{E}\sup_{x\in\mathcal{N}}|\alpha(x)-1| &\le C((1+\delta)\sqrt{k/n}+1/\sqrt{k})+C\sqrt{s\log n}(1+\delta)\sqrt{k/n}\\ &\quad +C(s\log n)(1+\delta)/\sqrt{n}\\ &\le C\left((1+\delta)\left(\sqrt{\frac{sk\log n}{n}}+\frac{s\log n}{\sqrt{n}}\right)+k^{-1/2}\right). \end{aligned} \tag{4.13}$$

We will assume in what follows that

$$s\log n \le k, \tag{4.14}$$

so that (4.13) takes the simpler form

$$\mathbb{E}\sup_{x\in\mathcal{N}}|\alpha(x)-1| \le C\left((1+\delta)\sqrt{\frac{sk\log n}{n}}+k^{-1/2}\right). \tag{4.15}$$

Recalling that α is a seminorm and applying (3.14) in Lemma 3.2 we pass to the set of all s-sparse Euclidean unit normed vectors,

$$\begin{aligned}\mathbb{E}\sup_{\substack{\|x\|_2=1\\ \|x\|_0\le s}}|\alpha(x)-1| &= \mathbb{E}\max_{\#T=s}\sup_{x\in S_T}|\alpha(x)-1|\\ &\le C\left((1+\delta)\sqrt{\tfrac{sk\log n}{n}}+k^{-1/2}\right). \end{aligned} \tag{4.16}$$

4.1 A Bootstrapping Argument

Let δ' denote $\delta_s(\tilde{A})$. Assume henceforth that the parameters s, k satisfy

$$\kappa := C\sqrt{\frac{sk\log n}{n}} < 1/2. \tag{4.17}$$

Clearly (4.15) is a bound on $\mathbb{E}[\delta']$. With the new notation, we get

$$\mathbb{E}[1+\delta'] \le (1+\delta)\kappa + 1 + Ck^{-1/2}.$$

Denote $\tilde{A}$ by $A^{(1)}$ and A by $A^{(0)}$. Now consider inductively repeating the above process, obtaining (for $i \ge 2$) $A^{(i)}$ from $A^{(i-1)}$ by

$$A^{(i)} = A^{(i-1)} D_{\varepsilon_{(i)}} H D_{\varepsilon'_{(i)}} H,$$

where $\varepsilon_{(i)}, \varepsilon'_{(i)}$ are independent copies of $\varepsilon, \varepsilon'$. Let $\delta^{(i)}$ denote $\delta_s(A^{(i)})$. By independence and the principle of conditional expectation, we conclude that

$$\mathbb{E}[1+\delta^{(i)}] \le (1+\delta^{(0)})\kappa^i + \frac{1+Ck^{-1/2}}{(1-\kappa)} \le (1+\delta^{(0)})\kappa^i + 2(1+Ck^{-1/2}).$$

Assume in what follows that k is large enough so that $Ck^{-1/2} \le 1$. Then the last inequality conveniently implies $\mathbb{E}[1+\delta^{(i)}] \le (1+\delta^{(0)})\kappa^i + 4$. Recall by our definition of A that $\delta = \delta^{(0)}$ can be no more than $\sqrt{n/k}$. Let r be taken as

$$r := \left\lceil \frac{-\log(2\sqrt{n/k})}{\log \kappa} \right\rceil \tag{4.18}$$

so that $(1+\delta(0))\kappa^r \le (1+\sqrt{n/k})\kappa^r \le 2\sqrt{n/k}\kappa^r \le 1$ and hence

$$\mathbb{E}[1+\delta^{(r)}] \le 5.$$

Using Markov's inequality, this implies that with probability at least, say, 0.995

$$1+\delta^{(r)} \le 1000. \tag{4.19}$$

From now on assume that the event (4.19) holds. Now for an s-sparse unit vector x, let $x^{(r+1)} := A^{(r+1)}x$. The assumption $k \le \sqrt{n}$ is equivalent to $1/\sqrt{k} \ge \sqrt{k/n}$. Using this, and substituting a constant for $(1+\delta)$, (4.10) implies

$$\begin{aligned} &\Pr(|\|x^{(r+1)}\|_2 - M_{\|x^{(r+1)}\|_2}| \ge t) \\ &\le 2\exp\big(-C\min\big(\tfrac{t^2}{(k^{-1/2})^2}, t\sqrt{n}\big)\big). \end{aligned} \tag{4.20}$$

where $M_{\|x^{(r+1)}\|_2}$ is a median of $\|x^{(r+1)}\|_2$.

Once again we consider maximal μ-separated sets $\mathcal{N}_T$ of S_T with $\mu = 1/k$ for each $T \subset \{1,\dots,N\}$ of size s and form $\mathcal{N} = \bigcup_{\#T=s} \mathcal{N}_T$. The cardinality of $\mathcal{N}$ is at most $\exp\{Cs\log n\}$, see (4.12). We can now use a union bound over $\mathcal{N}$, to conclude that with probability at least 0.995,

$$\max_{x\in\mathcal{N}} \big|\|x^{(r+1)}\|_2 - M_{\|x^{(r+1)}\|_2}\big| \le C\max\left(\sqrt{\frac{s\log n}{k}}, \frac{s\log n}{\sqrt{n}}\right). \tag{4.21}$$

Using (4.11), this implies

$$\begin{aligned} \max_{x\in\mathcal{N}} \big|\|x^{(r+1)}\|_2 - 1\big| &\le C\left(\max\left(\sqrt{\frac{sk\log n}{n}}, \frac{s\log n}{\sqrt{n}}\right) + k^{-1/2}\right) \\ &\le C\left(\sqrt{\frac{sk\log n}{n}} + k^{-1/2}\right), \end{aligned} \tag{4.22}$$

where the last inequality used (4.14). As before, using (3.14) in Lemma 3.2, allows us to pass to the set of all s-sparse vectors:

$$\sup_{\substack{\|x\|=1 \\ \|x\|_0 \le s}} \left| \|x^{(r+1)}\|_2 - 1 \right| \le C\left(\sqrt{\frac{sk \log n}{n}} + k^{-1/2} \right). \tag{4.23}$$

Recalling the assumption $k \le \sqrt{n}$, this implies

$$\delta^{(r+1)} \le C\sqrt{\frac{s \log n}{k}}$$

and the proof of Theorem 2.2 is concluded.

Acknowledgments Nir Ailon acknowledges the support of a Marie Curie International Reintegration Grant PIRG07-GA-2010-268403, and a grant from the GIF, the German-Israeli Foundation for Scientific Research and Development. Work was done during his visit to the Hausdorff Center for Mathematics at the University of Bonn. Holger Rauhut acknowledges support by the Hausdorff Center for Mathematics at the University of Bonn and funding by the European Research Council through the Grant StG 258926.

Appendix: Properties of Mixed Gaussian and Exponential Processes

Let us collect some auxiliary results relating tails and expectations of certain random variables, which are required in the proofs of our main results.

Lemma 5.1 *Assume X is a random variable such that for some number M and for all $t \ge 0$,*

$$\Pr[|X - M| > t] \le C \exp\left\{ - \min\{t^2/\sigma_1^2, t/\sigma_2\}\right\},$$

for some $\sigma_1, \sigma_2 \ge 0$. Then

(1) $\left|(\mathbb{E}X^2)^{1/2} - |M|\right| \le C''\sqrt{\sigma_1^2 + \sigma_2^2}$
(2) $|\mathbb{E}X - M| \le C''\left|\sigma_1 + \sigma_2 t\right|$

for some constant C'' that depends only on C.

Proof For the first part, assume that $M \ge 0$ for the moment. By integrating and changing variables,

$$\begin{aligned}
\mathbb{E}(X - M)^2 &= \int_0^\infty \Pr[|X - M| \ge \sqrt{t}]dt \\
&\le C\int_0^\infty \exp\left\{ - t/\sigma_1^2\right\}dt + C\int_0^\infty \exp\left\{ - \sqrt{t}/\sigma_2\right\}dt \\
&= C\sigma_1^2 \int_0^\infty e^{-s}ds + 2C\sigma_2^2 \int_0^\infty e^{-s}sds \\
&\le (\sigma_1^2 + \sigma_2^2)C'
\end{aligned} \tag{5.1}$$

for some constant $C' > 0$. On the other hand, with $p^2 = \mathbb{E}X^2$ we have $\mathbb{E}(X - M)^2 = p^2 + M^2 - 2M\mathbb{E}X$ and $\mathbb{E}X \le \mathbb{E}|X| \le \sqrt{\mathbb{E}X^2} = p$. We hence conclude that $(p - M)^2 \le (\sigma_1^2 + \sigma_2^2)C'$. If $M < 0$ then we simply replace the random variable X with $-X$ and M with $-M$.

The second part is obtained in the same way by integrating to bound $\mathbb{E}|X - M|$. □

Lemma 5.2 *Assume that X_i, $i = 1, \dots, N$ are random variables such that for each i there exist numbers M_i and $\sigma_i, \sigma_i' \ge 0$ such that for all $t \ge 0$,*

$$\Pr[|X_i - M_i| > t] \le 2\exp\big\{-\min\{t^2/\sigma_i^2, t/\sigma_i'\}\big\}.$$

Then

$$\mathbb{E}\sup_i |X_i - M_i| \le C\big(\sqrt{\log N}\sup_i \sigma_i + \log N \sup_i \sigma_i'\big). \tag{5.2}$$

Note that the variables X_i are not required to be independent. The proof can be done by integration by parts, very similar to the derivation of (3.6) in [16].

References

1. Achlioptas, D.: Database-friendly random projections: Johnson–Lindenstrauss with binary coins. J. Comput. Syst. Sci. **66**(4), 671–687 (2003)
2. Ailon, N., Liberty, E.: Fast dimension reduction using Rademacher series on dual BCH codes. Discrete Comput. Geom. **42**(4), 615–630 (2009)
3. Baraniuk, R.G., Davenport, M., DeVore, R.A., Wakin, M.: A simple proof of the restricted isometry property for random matrices. Constr. Approx. **28**(3), 253–263 (2008)
4. Blumensath, T., Davies, M.: Iterative hard thresholding for compressed sensing. Indepen. Compon. Analysis Signal Sep. **27**(3), 265–274 (2009)
5. Candès, E.J., Tao, T.: Near optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory **52**(12), 5406–5425 (2006)
6. Candès, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
7. Cheraghchi, M., Guruswami, V., Velingker, A.: Restricted isometry of fourier matrices and list decodability of random linear codes. In: Symposium on Discrete Algorithms (SODA), pp. 432–442 (2013)
8. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
9. Fornasier, M., Rauhut, H.: Compressive sensing. In: Scherzer, O. (ed.) Handbook of Mathematical Methods in Imaging, pp. 187–228. Springer, Berlin (2011)
10. Foucart, S.: Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants. In: Proceedings of the 13th International Conference on Approximation Theory (2010).
11. Foucart, S., Rauhut, H.: A Mathematical Introduction to Compressive Sensing. Birkhäuser, Basel (2013)
12. Foucart, S., Pajor, A., Rauhut, H., Ullrich, T.: The Gelfand widths of ℓ_p-balls for $0 < p \le 1$. J. Complexity **26**, 629–640 (2010)
13. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, STOC '98, pp. 604–613 (1998).
14. Krahmer, F., Ward, R.: New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. SIAM J. Math. Analysis **43**(3), 1269–1281 (2011)
15. Krahmer, F., Mendelson, S., Rauhut, H.: Suprema of chaos processes and the restricted isometry property. http://arxiv.org/abs/1207.0235 (2012)
16. Ledoux, M., Talagrand, M.: Probability in Banach spaces: isoperimetry and processes. Series of modern surveys in mathematics. Springer-Verlag, Berlin (2010)
17. Mendelson, S., Pajor, A., Tomczak, N.: Jaegermann. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. Constr. Approx. **28**(3), 277–289 (2009)

18. Nelson, J., Price, E., Wootters, M.: New constructions of rip matrices with fast multiplication and fewer rows. http://arxiv.org/abs/1211.0986 (2012)
19. Plan Y., Vershynin, R. One-bit compressed sensing by linear programming. Commun. Pure Appl. Math. **66**(8), 1275–1297 (2013)
20. Rauhut, H.: Compressive sensing and structured random matrices. In: Fornasier, M. (ed.) Theoretical Foundations and Numerical Methods for Sparse Recovery. Radon Series on Computational and Applied Mathematics, vol. 9, pp. 1–92. deGruyter, Berlin (2010).
21. Rauhut, H., Romberg, J.K., Tropp, J.A.: Restricted isometries for partial random circulant matrices. Appl. Comput. Harmon. Anal. **32**(2), 242–254 (2012)
22. Rudelson, M., Vershynin, R.: On sparse reconstruction from Fourier and Gaussian measurements. Commun. Pure Appl. Math. **61**, 1025–1045 (2008)
23. Talagrand, M.: New concentration inequalities in product spaces. Invent. Math. **126**, 505–563 (1996)
24. Tropp, J.A., Needell, D.: CoSaMP: iterative signal recovery from incomplete and inaccurate samples. Appl. Comput. Harmon. Anal. **26**(3), 301–321 (2008)
25. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. In: Eldar, Y.C., Kutyniok, G. (eds.) Compressed Sensing: Theory and Applications, pp. 210–268. Cambridge University Press, Cambridge (2012)