Compressive Sensing and Structured Random Matrices

Holger Rauhut

Abstract. These notes give a mathematical introduction to compressive sensing focusing on recovery using ℓ_1-minimization and structured random matrices. An emphasis is put on techniques for proving probabilistic estimates for condition numbers of structured random matrices. Estimates of this type are key to providing conditions that ensure exact or approximate recovery of sparse vectors using ℓ_1-minimization.

Keywords. compressive sensing, ℓ_1-minimization, basis pursuit, structured random matrices, condition numbers, random partial Fourier matrix, partial random circulant matrix, Khintchine inequalities, bounded orthogonal systems.

AMS classification. 15A12, 15A60, 15B02, 15B19, 15B52, 42A05, 42A61, 46B09, 46B10, 60B20, 60G50, 90C05, 90C25, 90C90, 94A12, 94A20.

H. R. acknowledges support by the Hausdorff Center for Mathematics and by the WWTF project SPORTS (MA 07-004).
Version of June 12, 2011.

1 Introduction

Compressive sensing is a recent theory that predicts that sparse vectors in high dimensions can be recovered from what was previously believed to be incomplete information. The seminal papers by E. Candès, J. Romberg and T. Tao [19, 23] and by D. Donoho [38] have caught significant attention and have triggered enormous research activities after their appearance. These notes make an attempt to introduce to some mathematical aspects of this vastly growing field. In particular, we focus on ℓ_1-minimization as recovery method and on structured random measurement matrices such as the random partial Fourier matrix and partial random circulant matrices. We put emphasis on methods for showing probabilistic condition number estimates for structured random matrices. Among the main tools are scalar and noncommutative Khintchine inequalities. It should be noted that modified parts of these notes together with much more material will appear in a monograph on compressive sensing [55] that is currently under preparation by the author and Simon Foucart.

The main motivation for compressive sensing is that many real-world signals can be well-approximated by sparse ones, that is, they can be approximated by an expansion in terms of a suitable basis, which has only a few non-vanishing terms. This is the key why many (lossy) compression techniques such as JPEG or MP3 work so well. To obtain a compressed representation one computes the coefficients in the basis (for instance a wavelet basis) and then keeps only the largest coefficients. Only these will be stored while the rest of them will be put to zero when recovering the compressed signal.

When complete information on the signal or image is available this is certainly a valid strategy. However, when the signal has to be acquired first with a somewhat costly, difficult, or time-consuming measurement process, this seems to be a waste of resources: First one spends huge efforts to collect complete information on the signal and then one throws away most of the coefficients to obtain its compressed version. One might ask whether there is a more clever way of obtaining somewhat more directly the compressed version of the signal. It is not obvious at first sight how to do this: measuring directly the large coefficients is impossible since one usually does not know *a-priori*, which of them are actually the large ones. Nevertheless, compressive sensing provides a way of obtaining the compressed version of a signal using only a small number of linear and non-adaptive measurements. Even more surprisingly, compressive sensing predicts that recovering the signal from its undersampled measurements can be done with computationally efficient methods, for instance convex optimization, more precisely, ℓ_1-minimization.

Of course, arbitrary undersampled linear measurements – described by the so-called measurement matrix – will not succeed in recovering sparse vectors. By now, necessary and sufficient conditions are known for the matrix to recover sparse vectors using ℓ_1-minimization: the null space property and the restricted isometry property. Basically, the restricted isometry property requires that all column submatrices of the measurement matrix of a certain size are well-conditioned. It turns out to be quite difficult to check this condition for deterministic matrices – at least when one aims to work with the minimal amount of measurements. Indeed, the seminal papers [19, 38] obtained their breakthrough by actually using random matrices. While the use of random matrices in sparse signal processing was rather uncommon before the advent of compressive sensing, we note that they were used quite successfully already much earlier, for instance in the very related problem from Banach space geometry of estimating Gelfand widths of ℓ_1^N-balls [54, 57, 74].

Introducing randomness allows to show optimal (or at least near-optimal) conditions on the number of measurements in terms of the sparsity that allow recovery of sparse vectors using ℓ_1-minimization. To this end, often Gaussian or Bernoulli matrices are used, that is, random matrices with stochastically independent entries having a standard normal or Bernoulli distribution.

Applications, however, often do not allow the use of "completely" random matrices, but put certain physical constraints on the measurement process and limit the

amount of randomness that can be used. For instance, when sampling a trigonometric polynomial having sparse coefficients one might only have the freedom to choose the sampling points at random. This leads then to a structured random measurement matrix, more precisely, a random partial Fourier type matrix. Indeed, such type of matrices were already investigated in the initial papers [19, 23] on compressive sensing. These notes will give an introduction on recovery results for ℓ_1-minimization that can be obtained using such structured random matrices. A focus is put on methods for probabilistic estimates of condition numbers such as the noncommutative Khintchine inequalities and Dudley's inequality.

Although we will not cover specific applications in these notes, let us mention that compressive sensing may be applied in imaging [44, 109], A/D conversion [133], radar [69, 49] and wireless communication [126, 95], to name a few.

These notes contain some improvements and generalizations of existing results, that have not yet appeared elsewhere in the literature. In particular, we generalize from random sampling of sparse trigonometric polynomials to random sampling of functions having sparse expansions in terms of bounded orthonormal systems. The probability estimate for the so-called restricted isometry constants for the corresponding matrix is slightly improved. Further, also the sparse recovery result for partial random circulant and Toeplitz matrices presented below is an improvement over the one in [105].

These lecture notes only require basic knowledge of analysis, linear algebra and probability theory, as well as some basic facts about vector and matrix norms.

2 Recovery via ℓ_1-minimization

2.1 Preliminaries and Notation

Let us first introduce some notation. For a vector $\mathbf{x} = (x_1, \dots, x_N) \in \mathbb{C}^N$, the usual p-norm is denoted

$$\|\mathbf{x}\|_p := \left(\sum_{\ell=1}^{N} |x_\ell|^p \right)^{1/p}, \quad 1 \leq p < \infty,$$
$$\|\mathbf{x}\|_\infty := \max_{\ell \in [N]} |x_\ell|,$$

where $[N] := \{1, 2, \dots, N\}$. For a matrix $A = (a_{jk}) \in \mathbb{C}^{m \times N}$ we denote $A^* = (\overline{a_{kj}})$ its conjugate transpose. The operator norm of a matrix from ℓ_p into ℓ_p is defined as

$$\|A\|_{p \to p} := \max_{\|\mathbf{x}\|_p = 1} \|A\mathbf{x}\|_p.$$

For the cases $p = 1, 2, \infty$ an explicit expression for the operator norm of A is given by

$$\|A\|_{1\to 1} = \max_{k\in[N]} \sum_{j=1}^{m} |a_{jk}|, \tag{2.1}$$

$$\|A\|_{\infty\to\infty} = \max_{j\in[m]} \sum_{k=1}^{N} |a_{jk}|,$$

$$\|A\|_{2\to 2} = \sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^*A)},$$

where $\sigma_{\max}(A)$ denotes the largest singular value of A and $\lambda_{\max}(A^*A) \geq 0$ is the largest eigenvalue of A^*A. Clearly, $\|A\|_{1\to 1} = \|A^*\|_{\infty\to\infty}$. It follows from the Riesz-Thorin interpolation theorem [118, 7] that

$$\|A\|_{2\to 2} \leq \max\{\|A\|_{1\to 1}, \|A\|_{\infty\to\infty}\}. \tag{2.2}$$

The above inequality is sometimes called the Schur test, and it can also be derived using Hölder's inequality, see for instance [64]; or alternatively using Gershgorin's disc theorem [8, 71, 135]. In particular, if $A = A^*$ is hermitian, then

$$\|A\|_{2\to 2} \leq \|A\|_{1\to 1}. \tag{2.3}$$

All eigenvalues of a hermitian matrix $A = A^* \in \mathbb{C}^{n\times n}$ are contained in

$$\{\langle A\mathbf{x}, \mathbf{x}\rangle : \mathbf{x} \in \mathbb{C}^n, \|\mathbf{x}\|_2 = 1\} \subset \mathbb{R}.$$

In particular, for hermitian $A = A^*$,

$$\|A\|_{2\to 2} = \sup_{\|\mathbf{x}\|_2=1} |\langle A\mathbf{x}, \mathbf{x}\rangle|. \tag{2.4}$$

For real scalars $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and vectors $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{C}^m$ the matrix $\sum_{j=1}^n \alpha_j \mathbf{z}_j \mathbf{z}_j^*$ is hermitian and we have

$$\|\sum_{j=1}^{n} \alpha_j \mathbf{z}_j \mathbf{z}_j^*\|_{2\to 2} = \sup_{\|\mathbf{x}\|_2=1} \left|\left\langle \sum_{j=1}^{n} \alpha_j \mathbf{z}_j \mathbf{z}_j^* \mathbf{x}, \mathbf{x}\right\rangle\right| = \sup_{\|\mathbf{x}\|_2=1} \left|\sum_{j=1}^{n} \alpha_j |\langle \mathbf{z}_j, \mathbf{x}\rangle|^2\right|$$

$$\leq \max_{k\in[n]} |\alpha_k| \sup_{\|\mathbf{x}\|_2=1} \sum_{j=1}^{n} |\langle \mathbf{z}_j, \mathbf{x}\rangle|^2 = \|\sum_{j=1}^{n} \mathbf{z}_j \mathbf{z}_j^*\|_{2\to 2} \max_{k\in[n]} |\alpha_k|. \tag{2.5}$$

Also the Frobenius norm will be of importance. For a matrix $A = (a_{jk})$ it is defined as

$$\|A\|_F := \sqrt{\sum_{j,k} |a_{jk}|^2} = \sqrt{\mathrm{Tr}(A^*A)},$$

where Tr denotes the trace. The Frobenius norm is induced by the inner product $\langle A, B\rangle_F = \mathrm{Tr}(B^*A)$. The Cauchy Schwarz inequality for the trace states that

$$|\langle A, B\rangle_F| = |\,\mathrm{Tr}(B^*A)| \leq \|A\|_F \|B\|_F\,. \tag{2.6}$$

The null space of a matrix $A \in \mathbb{C}^{m\times N}$ is denoted by $\ker A = \{\mathbf{x} \in \mathbb{C}^N, A\mathbf{x} = 0\}$. We usually write $\mathbf{a}_\ell \in \mathbb{C}^m$, $\ell = 1, \ldots, N$, for the columns of a matrix $A \in \mathbb{C}^{m\times N}$. The column submatrix of A consisting of the columns indexed by S will be written $A_S = (\mathbf{a}_j)_{j\in S}$. If $S \subset [N]$, then for $\mathbf{x} \in \mathbb{C}^N$ we denote by $\mathbf{x}_S \in \mathbb{C}^N$ the vector that coincides with $\mathbf{x}$ on S and is set to zero on $S^c = [N] \setminus S$. Similarly, $\mathbf{x}^S \in \mathbb{C}^S$ denotes the vector $\mathbf{x}$ restricted to the entries in S. The support of a vector is defined as $\mathrm{supp}\,\mathbf{x} = \{\ell, x_\ell \neq 0\}$. We write Id for the identity matrix. The complement of a set $S \subset [N]$ is denoted $S^c = [N] \setminus S$, while $|S|$ is its cardinality.

If $A \in \mathbb{C}^{m\times n}$, $m \geq n$, is of full rank (i.e. injective), then its Moore-Penrose pseudo-inverse is given by

$$A^\dagger = (A^*A)^{-1}A^*. \tag{2.7}$$

In this case, it satisfies $A^\dagger A = \mathrm{Id} \in \mathbb{C}^{n\times n}$. We refer to [8, 71, 59] for more information on the pseudo inverse.

All the constants appearing in this note – usually denoted by C or D – are universal, which means that they do not depend on any of the involved quantities.

2.2 Sparse Recovery

Let $\mathbf{x} \in \mathbb{C}^N$ be a (high-dimensional) vector that we will sometimes call signal. It is called s-sparse if

$$\|\mathbf{x}\|_0 := |\mathrm{supp}\,\mathbf{x}| \leq s. \tag{2.8}$$

The quantity $\|\cdot\|_0$ is often called ℓ_0-norm although it is actually not a norm, not even a quasi-norm.

In practice it is generally not realistic that a signal $\mathbf{x}$ is exactly s-sparse, but rather that its error of best s-term approximation $\sigma_s(\mathbf{x})_p$ is small,

$$\sigma_s(\mathbf{x})_p := \inf\{\|\mathbf{x} - \mathbf{z}\|_p, \mathbf{z} \text{ is } s\text{-sparse}\}. \tag{2.9}$$

(This is the standard notation in the literature, and we hope that no confusion with the singular values of a matrix will arise.)

Taking linear measurements of $\mathbf{x}$ is modeled as the application of a measurement matrix $A \in \mathbb{C}^{m\times N}$,

$$\mathbf{y} = A\mathbf{x}. \tag{2.10}$$

The vector $\mathbf{y} \in \mathbb{C}^m$ is called the measurement vector. We are interested in the case of undersampled measurements, that is, $m \ll N$. Reconstructing $\mathbf{x}$ amounts to solving (2.10). By basic linear algebra, this system of equations has infinitely many solutions (at least if A has full rank). Hence, it seems impossible at first sight to guess the

correct $\mathbf{x}$ among these solutions. If, however, we impose the additional requirement (2.8) that $\mathbf{x}$ is s-sparse, the situation changes, as we will see. Intuitively, it is natural to search then for the solution with smallest support, that is, to solve the ℓ_0-minimization problem

$$\min_{\mathbf{z}\in\mathbb{C}^N} \|\mathbf{z}\|_0 \quad \text{subject to} \quad A\mathbf{z} = \mathbf{y}. \tag{2.11}$$

The hope is that the solution $\mathbf{x}^\#$ of this optimization problem coincides with $\mathbf{x}$. Indeed, rather easy recovery conditions on $A \in \mathbb{C}^{m\times N}$ and on the sparsity s can be shown, see for instance [28]. There exist matrices $A \in \mathbb{C}^{m\times N}$ such that $2s \leq m$ suffices to always ensure recovery; choose the columns of A in general position.

Unfortunately, the combinatorial optimization problem (2.11) is NP hard in general [35, 88]. In other words, an algorithm that solves (2.11) for any matrix A and any vector $\mathbf{y}$ must be intractable (unless maybe the famous Millenium problem $P = NP$ is solved in the affirmative, on which we will not rely here). Therefore, (2.11) is completely impractical for applications and tractable alternatives have to be found. Essentially two approaches have mainly been pursued: greedy algorithms and convex relaxation. We will concentrate here on the latter and refer the reader to the literature [40, 58, 78, 90, 91, 103, 131, 127] for further information concerning greedy methods.

The ℓ_1-minimization problem

$$\min_{\mathbf{z}\in\mathbb{C}^N} \|\mathbf{z}\|_1 \quad \text{subject to} \quad A\mathbf{z} = \mathbf{y} \tag{2.12}$$

can be understood as convex relaxation of (2.11). Sometimes (2.12) is also referred to as basis pursuit [25]. In contrast to (2.11), the ℓ_1-minimization problem can be solved with efficient convex optimization methods. In the real-valued case (2.12) can be rewritten as a linear program and can be solved with linear programming techniques, while in the complex-valued case (2.12) is equivalent to a second order cone program (SOCP), for which also efficient solvers exist [15]. We refer the interested reader to [32, 33, 34, 43, 47, 76] for further efficient algorithms for ℓ_1-minimization.

Of course, our hope is that the solution of (2.12) coincides with the one of (2.11). One purpose of these notes is to provide an understanding under which conditions this is actually guaranteed.

2.3 Null Space Property and Restricted Isometry Property

In this section we present conditions on the matrix A that ensure exact reconstruction of all s-sparse vectors using ℓ_1-minimization. Our first notion is the so-called null space property.

Definition 2.1. A matrix $A \in \mathbb{C}^{m\times N}$ satisfies the null space property of order s if for all subsets $S \subset [N]$ with $|S| = s$ it holds

$$\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{S^c}\|_1 \quad \text{for all } \mathbf{v} \in \ker A \setminus \{0\}. \tag{2.13}$$

Remark 2.2. We deal here with the complex case. For real-valued matrices one might restrict the kernel to the real-valued vectors and define an obvious real-valued analogue of the null space property above. However, it is not obvious that the real and the complex null space property are the same for real-valued matrices. Nevertheless this fact can be shown [52].

Based on this notion we have the following recovery result concerning ℓ_1-minimization.

Theorem 2.3. *Let $A \in \mathbb{C}^{m\times N}$. Then every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is the unique solution of the ℓ_1-minimization problem* (2.12) *with $\mathbf{y} = A\mathbf{x}$ if and only if A satisfies the null space property of order s.*

Proof. Assume first that every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $A\mathbf{z} = A\mathbf{x}$. Then, in particular, for any $\mathbf{v} \in \ker A\backslash\{0\}$ and any $S \subset [N]$ with $|S| = s$, the s-sparse vector $\mathbf{v}_S$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $A\mathbf{z} = A\mathbf{v}_S$. Observe that $A(-\mathbf{v}_{S^c}) = A\mathbf{v}_S$ and $-\mathbf{v}_{S^c} \neq \mathbf{v}_S$, because $A(\mathbf{v}_{S^c} + \mathbf{v}_S) = A\mathbf{v} = 0$ and because $\mathbf{v} \neq 0$. Therefore we must have $\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{S^c}\|_1$. This establishes the null space property.

For the converse, let us assume that the null space property of order s holds. Then, given an s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ and a vector $\mathbf{z} \in \mathbb{C}^N$, $\mathbf{z} \neq \mathbf{x}$, satisfying $A\mathbf{z} = A\mathbf{x}$, we consider $\mathbf{v} := \mathbf{x} - \mathbf{z} \in \ker A \setminus \{0\}$ and $S := \mathrm{supp}(\mathbf{x})$. In view of the null space property we obtain

$$\begin{aligned}\|\mathbf{x}\|_1 &\leq \|\mathbf{x} - \mathbf{z}_S\|_1 + \|\mathbf{z}_S\|_1 = \|\mathbf{x}_S - \mathbf{z}_S\| + \|\mathbf{z}_S\|_1 = \|\mathbf{v}_S\|_1 + \|\mathbf{z}_S\|_1 \\ &< \|\mathbf{v}_{S^c}\|_1 + \|\mathbf{z}_S\|_1 = \| - \mathbf{z}_{S^c}\|_1 + \|\mathbf{z}_S\|_1 = \|\mathbf{z}\|_1.\end{aligned}$$

This establishes the required minimality of $\|\mathbf{x}\|_1$. □

This theorem seems to have first appeared explicitly in [60], although it was used implicitly already in [41, 48, 97]. The term null space property was coined by A. Cohen, W. Dahmen, and R. DeVore in [28]. One may obtain also a stable version of the above theorem by passing from sparse vectors to compressible ones, for which $\sigma_s(\mathbf{x})_1$ is small. Then the condition (2.13) has to be strengthened to $\|\mathbf{v}_S\|_1 < \gamma\|\mathbf{v}_{S^c}\|_1$ for some $\gamma \in (0, 1)$.

The null space property is usually somewhat difficult to show directly. Instead, the so called restricted isometry property [22], which was introduced by E. Candès and T. Tao in [23] under the term uniform uncertainty principle (UUP), has become very popular in compressive sensing.

Definition 2.4. The restricted isometry constant δ_s of a matrix $A \in \mathbb{C}^{m\times N}$ is defined as the smallest δ_s such that

$$(1 - \delta_s)\|\mathbf{x}\|_2^2 \leq \|A\mathbf{x}\|_2^2 \leq (1 + \delta_s)\|\mathbf{x}\|_2^2 \tag{2.14}$$

for all s-sparse $\mathbf{x} \in \mathbb{C}^N$.

We say that a matrix A satisfies the restricted isometry property (RIP) if δ_s is small for reasonably large s (whatever "small" and "reasonably large" might mean in a concrete situation).

Before relating the restricted isometry property with the null space property let us first provide some simple properties of the restricted isometry constants.

Proposition 2.5. *Let $A \in \mathbb{C}^{m\times N}$ with isometry constants δ_s.*

(a) The restricted isometry constants are ordered, $\delta_1 \leq \delta_2 \leq \delta_3 \leq \cdots$.

(b) It holds

$$\begin{aligned}\delta_s &= \max_{S\subset[N],|S|\leq s} \|A_S^* A_S - \mathrm{Id}\|_{2\to 2} \\ &= \sup_{\mathbf{x}\in T_s} |\langle (A^*A - \mathrm{Id})\mathbf{x}, \mathbf{x}\rangle|,\end{aligned}$$

where $T_s = \{\mathbf{x} \in \mathbb{C}^N, \|\mathbf{x}\|_2 = 1, \|\mathbf{x}\|_0 \leq s\}$.

(c) Let $\mathbf{u}, \mathbf{v} \subset \mathbb{C}^N$ with disjoint supports, $\operatorname{supp}\mathbf{u} \cap \operatorname{supp}\mathbf{v} = \emptyset$. Let $s = |\operatorname{supp}\mathbf{u}| + |\operatorname{supp}\mathbf{v}|$. Then

$$|\langle A\mathbf{u}, A\mathbf{v}\rangle| \leq \delta_s \|\mathbf{u}\|_2 \|\mathbf{v}\|_2.$$

Proof. Since an s-sparse vector is also $s+1$-sparse the statement (a) is immediate.

The definition (2.14) is equivalent to

$$\big|\|A\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \leq \delta_s \|\mathbf{x}\|_2^2 \quad \text{for all } S \subset [N], |S| \leq s, \text{ for all } \mathbf{x} \in \mathbb{C}^N, \operatorname{supp}\mathbf{x} \subset S.$$

The term on the left hand side can be rewritten as $|\langle (A^*A - \mathrm{Id})\mathbf{x}, \mathbf{x}\rangle|$. Taking the supremum over all $\mathbf{x} \in \mathbb{C}^N$ with $\operatorname{supp}\mathbf{x} \subset S$ and unit norm $\|\mathbf{x}\|_2 = 1$ yields the operator norm $\|A_S^* A_S - \mathrm{Id}\|_{2\to 2}$ by (2.4). Taking also the maximum over all subsets S of cardinality at most s completes the proof of (b).

For (c) we denote $S = \operatorname{supp}\mathbf{u}, \Xi = \operatorname{supp}\mathbf{v}$ and let $\widetilde{\mathbf{u}}, \widetilde{\mathbf{v}}$ denote the vectors $\mathbf{u}, \mathbf{v}$ restricted to their supports. Then we write

$$\begin{aligned}\langle A\mathbf{u}, A\mathbf{v}\rangle &= \widetilde{\mathbf{u}}^* A_S^* A_\Xi \widetilde{\mathbf{v}} = (\widetilde{\mathbf{u}}^*, 0_\Xi^*) A_{S\cup\Xi}^* A_{S\cup\Xi} (0_S^*, \widetilde{\mathbf{v}}^*)^* \\ &= (\widetilde{\mathbf{u}}^*, 0_\Xi^*)(A_{S\cup\Xi}^* A_{S\cup\Xi} - \mathrm{Id})(0_S^*, \widetilde{\mathbf{v}}^*)^*,\end{aligned}$$

where 0_S is the zero-vector on the indices in S. Therefore, one may estimate

$$|\langle A\mathbf{u}, A\mathbf{v}\rangle| \leq \|A_{S\cup\Xi}^* A_{S\cup\Xi} - \mathrm{Id}\|_{2\to 2} \|\mathbf{u}\|_2 \|\mathbf{v}\|_2.$$

Applying part (b) completes the proof. □

Part (b) shows that the restricted isometry property requires in particular that all column submatrices of A of size s are well-conditioned. Indeed, all eigenvalues of $A_S^* A_S$ should be contained in the interval $[1-\delta_s, 1+\delta_s]$, which bounds the condition number of $A_S^* A_S$ by $\frac{1+\delta_s}{1-\delta_s}$ and therefore the one of A_S by $\sqrt{\frac{1+\delta_s}{1-\delta_s}}$.

The restricted isometry property implies the null space property as stated in the next theorem.

Theorem 2.6. *Suppose the restricted isometry constants δ_{2s} of a matrix $A \in \mathbb{C}^{m\times N}$ satisfies*

$$\delta_{2s} < \frac{1}{3}, \tag{2.15}$$

then the null space property of order s is satisfied. In particular, every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is recovered by ℓ_1-minimization.

Proof. Let $\mathbf{v} \in \ker A$ be given. It is enough to consider an index set S_0 of s largest modulus entries of the vector $\mathbf{v}$. We partition the complement of S_0 as $S_0^c = S_1 \cup S_2 \cup \ldots$, where S_1 is an index set of s largest absolute entries of $\mathbf{v}$ in $[N] \setminus S_0$, S_2 is an index set of s largest absolute entries of $\mathbf{v}$ in $[N] \setminus (S_0 \cup S_1)$ etc. In view of $\mathbf{v} \in \ker A$, we have $A(\mathbf{v}_{S_0}) = -A(\mathbf{v}_{S_1} + \mathbf{v}_{S_2} + \cdots)$, so that

$$\begin{aligned} \|\mathbf{v}_{S_0}\|_2^2 &\leq \frac{1}{1-\delta_{2s}} \|A(\mathbf{v}_{S_0})\|_2^2 = \frac{1}{1-\delta_{2s}} \langle A(\mathbf{v}_{S_0}), A(-\mathbf{v}_{S_1}) + A(-\mathbf{v}_{S_2}) + \cdots \rangle \\ &= \frac{1}{1-\delta_{2s}} \sum_{k\geq 1} \langle A(\mathbf{v}_{S_0}), A(-\mathbf{v}_{S_k}) \rangle. \end{aligned} \tag{2.16}$$

Proposition 2.5(c) yields then

$$\langle A(\mathbf{v}_{S_0}), A(-\mathbf{v}_{S_k}) \rangle \leq \delta_{2s} \|\mathbf{v}_{S_0}\|_2 \|\mathbf{v}_{S_k}\|_2. \tag{2.17}$$

Substituting (2.17) into (2.16) and dividing by $\|\mathbf{v}_{S_0}\|_2$ gives

$$\|\mathbf{v}_{S_0}\|_2 \leq \frac{\delta_{2s}}{1-\delta_{2s}} \sum_{k\geq 1} \|\mathbf{v}_{S_k}\|_2.$$

Since the s entries of $\mathbf{v}_{S_k}$ do not exceed the s entries of $\mathbf{v}_{S_{k-1}}$ for $k \geq 1$, we have

$$|v_j| \leq \frac{1}{s} \sum_{\ell \in S_{k-1}} |v_\ell| \quad \text{for all } j \in S_k$$

and therefore

$$\|\mathbf{v}_{S_k}\|_2 = \left(\sum_{j\in S_k} |v_j|^2 \right)^{1/2} \leq \frac{1}{\sqrt{s}} \|\mathbf{v}_{S_{k-1}}\|_1.$$

We obtain by the Cauchy–Schwarz inequality

$$\|\mathbf{v}_{S_0}\|_1 \leq \sqrt{s}\|\mathbf{v}_{S_0}\|_2 \leq \frac{\delta_{2s}}{1-\delta_{2s}} \sum_{k\geq 1} \|\mathbf{v}_{S_{k-1}}\|_1 \leq \frac{\delta_{2s}}{1-\delta_{2s}}(\|\mathbf{v}_{S_0}\|_1 + \|\mathbf{v}_{S^c}\|_1) \quad (2.18)$$

as announced. Since $\frac{\delta_{2s}}{1-\delta_{2s}} < 1/2$ by assumption, the null space property follows. □

The restricted isometry property also implies stable recovery by ℓ_1-minimization for vectors that can be well-approximated by sparse ones, and it further implies robustness under noise on the measurements. This fact was first noted in [23, 21]. The sufficient condition on the restricted isometry constants was successively improved in [18, 28, 53, 51]. We present without proof the so far best known result [51, 55] concerning recovery using a noise aware variant of ℓ_1-minimization.

Theorem 2.7. *Assume that the restricted isometry constant δ_{2s} of the matrix $A \in \mathbb{C}^{m\times N}$ satisfies*

$$\delta_{2s} < \frac{3}{4+\sqrt{6}} \approx 0.465. \quad (2.19)$$

Then the following holds for all $\mathbf{x} \in \mathbb{C}^N$. Let noisy measurements $\mathbf{y} = A\mathbf{x} + \mathbf{e}$ be given with $\|\mathbf{e}\|_2 \leq \eta$. Let $\mathbf{x}^\#$ be a solution of

$$\min_{\mathbf{z}\in\mathbb{C}^N} \|\mathbf{z}\|_1 \quad \text{subject to } \|A\mathbf{z} - \mathbf{y}\|_2 \leq \eta. \quad (2.20)$$

Then

$$\|\mathbf{x} - \mathbf{x}^\#\|_2 \leq c\eta + d\frac{\sigma_s(\mathbf{x})_1}{\sqrt{s}}$$

for some constants $c, d > 0$ that depend only on δ_{2s}.

Note that the previous theorem ensures exact recovery of s-sparse signals using ℓ_1-minimization (2.12) under condition (2.19) in the noise-free case $\eta = 0$.

In contrast to the null space property, the restricted isometry property is not necessary for sparse recovery by ℓ_1-minimization. Indeed, the null space property of A is invariant under multiplication from the left with an invertible matrix $U \in \mathbb{C}^{m\times m}$ as this does not change the null space, while the restricted isometry property is certainly not invariant (simply take a matrix U with large condition number).

We will soon see examples of measurement matrices with small restricted isometry constants.

2.4 Recovery of Individual Vectors

We will later need also a condition ensuring sparse recovery which not only depends on the matrix A but also on the sparse vector $\mathbf{x} \in \mathbb{C}^N$ to be recovered. The following theorem is due to J.J. Fuchs [56] in the real-valued case and was extended to the

complex-valued case by J. Tropp [128]. Its statement requires introducing the sign vector $\mathrm{sgn}(\mathbf{x}) \in \mathbb{C}^N$ having entries

$$\mathrm{sgn}(\mathbf{x})_j := \begin{cases} \dfrac{x_j}{|x_j|} & \text{if } x_j \neq 0, \\ 0 & \text{if } x_j = 0, \end{cases} \qquad j \in [N].$$

Theorem 2.8. *Let $A \in \mathbb{C}^{m\times N}$ and $\mathbf{x} \in \mathbb{C}^N$ with $S := \mathrm{supp}(\mathbf{x})$. Assume that A_S is injective and that there exists a vector $\mathbf{h} \in \mathbb{C}^m$ such that*

$$\begin{aligned} A_S^* \mathbf{h} &= \mathrm{sgn}(x^S), \\ |(A^*\mathbf{h})_\ell| &< 1, \quad \ell \in [N] \setminus S. \end{aligned} \tag{2.21}$$

Then $\mathbf{x}$ is the unique solution to the ℓ_1-minimization problem (2.12) *with $\mathbf{y} = A\mathbf{x}$.*

Proof. Let $\mathbf{h} \in \mathbb{C}^m$ be the vector with the described property. We have

$$\|\mathbf{x}\|_1 = \langle A^*\mathbf{h}, \mathbf{x}\rangle = \langle \mathbf{h}, A\mathbf{x}\rangle.$$

Thus, for $\mathbf{z} \in \mathbb{C}^N$, $\mathbf{z} \neq \mathbf{x}$, such that $A\mathbf{z} = \mathbf{y}$, we derive

$$\begin{aligned} \|\mathbf{x}\|_1 &= \langle \mathbf{h}, A\mathbf{z}\rangle = \langle A^*\mathbf{h}, \mathbf{z}\rangle = \langle A^*\mathbf{h}, \mathbf{z}_S\rangle + \langle A^*\mathbf{h}, \mathbf{z}_{S^c}\rangle \\ &\leq \|(A^*\mathbf{h})_S\|_\infty \|\mathbf{z}_S\|_1 + \|(A^*\mathbf{h})_{S^c}\|_\infty \|\mathbf{z}_{S^c}\|_1 < \|\mathbf{z}_S\|_1 + \|\mathbf{z}_{S^c}\|_1 = \|\mathbf{z}\|_1. \end{aligned}$$

The strict inequality follows from $\|\mathbf{z}_{S^c}\|_1 > 0$, which holds because otherwise the vector $\mathbf{z}$ would be supported on S and the equality $A\mathbf{z} = A\mathbf{x}$ would then be in contradiction with the injectivity of A_S. We have therefore shown that the vector $\mathbf{x}$ is the unique minimizer of $\|\mathbf{z}\|_1$ subject to $A\mathbf{z} = \mathbf{y}$, as desired. □

The above result makes clear that the success of sparse recovery by ℓ_1-minimization only depends on the support set S and on the sign pattern of the non-zero coefficients of $\mathbf{x}$.

Choosing the vector $h = \left(A_S^\dagger\right)^* \mathrm{sgn}(\mathbf{x}^S)$ leads to the following corollary, which will become a key tool later on.

Corollary 2.9. *Let $A \in \mathbb{C}^{m\times N}$ and $\mathbf{x} \in \mathbb{C}^N$ with $S := \mathrm{supp}(\mathbf{x})$. If the matrix A_S is injective and if*

$$|\langle A_S^\dagger \mathbf{a}_\ell, \mathrm{sgn}(\mathbf{x}^S)\rangle| < 1 \quad \text{for all } \ell \in [N] \setminus S, \tag{2.22}$$

then the vector $\mathbf{x}$ is the unique solution to the ℓ_1-mininimization problem (2.12) *with $\mathbf{y} = A\mathbf{x}$.*

Proof. The vector $h = \left(A_S^\dagger\right)^* \mathrm{sgn}(\mathbf{x}^S)$ satisfies $A_S^* h = A_S^* A_S (A_S^* A_S)^{-1} \mathrm{sgn}(\mathbf{x}^S) = \mathrm{sgn}(\mathbf{x}^S)$, and the condition (2.22) translates into (2.21). Hence, the statement follows from Theorem 2.8. □

2.5 Coherence

A classical way to measure the quality of a measurement matrix A with normalized columns, $\|\mathbf{a}_j\|_2 = 1, j \in [N]$, is the coherence [39, 40, 60, 61, 127], defined by

$$\mu := \max_{j \neq k} |\langle \mathbf{a}_j, \mathbf{a}_k \rangle|.$$

If the coherence is small than the columns of A are almost mutually orthogonal. A small coherence is desired in order to have good sparse recovery properties.

A refinement of the coherence is the 1-coherence function or Babel function, defined by

$$\mu_1(s) := \max_{\ell \in [N]} \max_{\substack{S \subset [N] \setminus \{\ell\} \\ |S| \leq s}} \sum_{j \in S} |\langle \mathbf{a}_j, \mathbf{a}_\ell \rangle| \leq s\mu.$$

The following proposition lists simple properties of μ and μ_1 and relates the coherence to the restricted isometry constants.

Proposition 2.10. *Let $A \in \mathbb{C}^{m \times N}$ with unit norm columns, coherence μ, 1-coherence function $\mu_1(s)$ and restricted isometry constants δ_s. Then*

(a) $\mu = \delta_2$,

(b) $\mu_1(s) = \max_{S \subset [N], |S| \leq s+1} \|A_S^* A_S - \mathrm{Id}\|_{1 \to 1}$,

(c) $\delta_s \leq \mu_1(s-1) \leq (s-1)\mu$.

Proof. (a) If $S = \{j, \ell\}$ has cardinality two then

$$A_S^* A_S - \mathrm{Id} = \begin{pmatrix} 0 & \langle \mathbf{a}_j, \mathbf{a}_\ell \rangle \\ \langle \mathbf{a}_\ell, \mathbf{a}_j \rangle & 0 \end{pmatrix},$$

by the normalization $\|\mathbf{a}_j\|_2 = \|\mathbf{a}_\ell\|_2 = 1$. The operator norm of this matrix equals $|\langle \mathbf{a}_j, \mathbf{a}_\ell \rangle|$. Taking the maximum over all two element subsets S shows that $\delta_2 = \mu$ by Proposition 2.5(b).

(b) Again by normalization, the matrix $A_S^* A_S - \mathrm{Id}$ has zeros on the diagonal. The explicit expression (2.1) for the operator norm on ℓ_1 then yields

$$\|A_S^* A_S - \mathrm{Id}\|_{1 \to 1} = \max_{j \in S} \sum_{k \in S \setminus \{j\}} |\langle \mathbf{a}_j, \mathbf{a}_k \rangle|.$$

Taking also the maximum over all $S \subset [N]$ with $|S| \leq s+1$ gives

$$\max_{S \subset [N], |S| \leq s+1} \|A_S^* A_S - \mathrm{Id}\|_{1 \to 1} = \max_{S \subset [N], |S| \leq s+1} \max_{j \in S} \sum_{k \in S \setminus \{j\}} |\langle \mathbf{a}_j, \mathbf{a}_k \rangle|$$

$$= \max_{j \in [N]} \max_{S \subset [N] \setminus \{j\}, |S| \leq s} \sum_{k \in S} |\langle \mathbf{a}_j, \mathbf{a}_k \rangle| = \mu_1(s),$$

which establishes (b).

For (c) observe that by Proposition 2.5(b) and inequality (2.3) for hermitian matrices

$$\delta_s = \max_{S\subset[N],|S|\leq s} \|A_S^* A_S - \mathrm{Id}\|_{2\to 2} \leq \max_{S\subset[N],|S|\leq s} \|A_S^* A_S - \mathrm{Id}\|_{1\to 1} = \mu_1(s-1)$$
$$\leq (s-1)\mu \tag{2.23}$$

by part (b). □

In combination with Theorem 2.6 (or Theorem 2.7) we see that $s - 1 \leq 1/(3\mu)$ or $\mu_1(s-1) \leq 1/3$ implies exact recovery (and also stable recovery) of all s-sparse vectors by ℓ_1-minimization. We note that the slightly weaker sufficient conditions

$$\mu_1(s-1) + \mu_1(s) < 1 \tag{2.24}$$

or $(2s-1)\mu < 1$ ensuring recovery by ℓ_1-minimization can be shown by working directly with the coherence or the 1-coherence function [39, 60, 127, 129] instead of the restricted isometry constants. It is worth noting that (2.24) also implies recovery by the greedy algorithm (orthogonal) matching pursuit [127, 62].

A simple example of a matrix $A \in \mathbb{C}^{m\times 2m}$ with small coherence is a concatenation of the identity with a Fourier matrix $F \in \mathbb{C}^{m\times m}$, i.e., $A = (\mathrm{Id}|F)$, where the entries of F are given by

$$F_{j,k} = \frac{1}{\sqrt{m}} e^{2\pi i jk/m}.$$

It is well known that F is unitary and it is easy to see that $\mu = \frac{1}{\sqrt{m}}$ and $\mu_1(s) = \frac{s}{\sqrt{m}}$ for $s = 1, \ldots, m-1$. It follows that

$$\delta_s \leq \frac{s-1}{\sqrt{m}}. \tag{2.25}$$

Hence, if

$$s < \frac{\sqrt{m}}{6} + 1 \tag{2.26}$$

then recovery by ℓ_1-minimization is ensured. There exist also matrices with many more columns still having coherence on the order $1/\sqrt{m}$. Indeed, [2, 121] give examples of matrices $A \in \mathbb{C}^{m\times m^2}$ satisfying

$$\mu = \frac{1}{\sqrt{m}}$$

(and one can also check that $\mu_1(s) = \frac{s}{\sqrt{m}}$ for $s = 1, \ldots, m-1$ for those matrices).

The drawback of these results is that the sparsity s is required to be tiny compared to the number m of measurements in (2.26). Or in other words, the number m of samples (measurements) required to recover an s-sparse vector scales quadratically in

s. As we will see, there exists (random) matrices for which the quadratic scaling can be improved to a much better linear scaling (up to log-factors). However, such results cannot be obtained by analyzing the coherence or the 1-coherence function as follows from the lower bounds in the next theorem.

Theorem 2.11. *Let $A \in \mathbb{C}^{m\times N}$ with normalized columns, coherence μ and 1-coherence function $\mu_1(s)$. Then*

(a) $\mu \geq \sqrt{\frac{N-m}{m(N-1)}}$,

(b) $\mu_1(s) \geq s\sqrt{\frac{N-m}{m(N-1)}}$ *whenever* $s \leq \sqrt{N-1}$.

The inequality in part (a) is also called Welch bound and can be found in [121, 111]. The proof of Part (b) is contained in [119]. Note that the case $s > \sqrt{N-1}$ is of minor importance to us, since then $\mu_1(s) > \sqrt{N-1}\sqrt{\frac{N-m}{m(N-1)}} = \sqrt{\frac{N}{m}-1}$ which will be larger than 1 provided $N \geq 2m$. The latter will be the case in all situations where compressive sensing is interesting. Then Proposition 2.10 implies only that $\delta_s \leq 1$, which does not allow any conclusion concerning ℓ_1-minimization.

For large enough N — say $N \geq 2m$ — the above lower bound for the coherence scales like $\frac{1}{\sqrt{m}}$, while the one for $\mu_1(s)$ scales like $\frac{s}{\sqrt{m}}$. Hence, those bounds explain to some extent why it is difficult to obtain significantly better recovery bounds than (2.26) for deterministic matrices. Indeed, the estimate (2.3) – or Gershgorin's theorem [8, 71, 135] that is often applied in the sparse approximation literature [127, 37] – which is used to establish Propostion 2.10(c), seems to be the optimal estimate one may obtain by taking into account only the absolute values of the Gramian matrix A^*A. In particular, it is *not* possible to improve on (2.25) by using Gershgorin's disc theorem, or by using Riesz-Thorin interpolation between $\|\cdot\|_{1\to 1}$ and $\|\cdot\|_{\infty\to\infty}$ (Schur's test).

Hence, to overcome the 'quadratic bottleneck' (2.25) or (2.26), that is, $m \geq Cs^2$, one should take into account cancellations that result from the signs of the entries of the Gramian A^*A. This task seems to be rather difficult, however, for deterministic matrices. The major breakthrough for beating the "quadratic bottleneck" was obtained using random matrices [19, 23, 38]. The problem of exploiting cancellations in the Gramian matrix is handled much easier with probabilistic methods than with deterministic techniques. And indeed, it is presently still an open problem to come up with deterministic matrices offering the same performance guarantees for sparse recovery as the ones for random matrices we will see below.

2.6 Restricted Isometry Property of Gaussian and Bernoulli Random Matrices

By now, many papers deal with Gaussian or Bernoulli random matrices in connection with sparse recovery, or more generally, subgaussian random matrices, [5, 23, 38, 42, 87, 114, 116]. The entries of a random Bernoulli matrix take the value $+\frac{1}{\sqrt{m}}$ or

$-\frac{1}{\sqrt{m}}$ with equal probability, while the entries of a Gaussian matrix are independent and follow a normal distribution with expectation 0 and variance $1/m$. With high probability such random matrices satisfy the restricted isometry property with a (near) optimal order in s, and therefore allow sparse recovery using ℓ_1-minimization.

Theorem 2.12. *Let $A \in \mathbb{R}^{m\times N}$ be a Gaussian or Bernoulli random matrix. Let $\epsilon, \delta \in (0,1)$ and assume*

$$m \geq C\delta^{-2}(s\ln(N/s) + \ln(\varepsilon^{-1})) \tag{2.27}$$

for a universal constant $C > 0$. Then with probability at least $1-\varepsilon$ the restricted isometry constant of A satisfies $\delta_s \leq \delta$.

There are by now several proofs of this result. In [5] a particularly nice and simple proof is given, which, however, yields an additional $\log(\delta^{-1})$-term. It shows in connection with Theorem 2.6 that with probability at least $1-\varepsilon$ all s-sparse vectors $\mathbf{x} \in \mathbb{C}^N$ can be recovered from $\mathbf{y} = A\mathbf{x}$ using ℓ_1-minimization (2.12) provided

$$m \geq C'(s\ln(N/s) + \ln(\varepsilon^{-1})). \tag{2.28}$$

Moreover, Theorem 2.7 predicts also stable and robust recovery under this condition. Note that choosing $\varepsilon = \exp(-cm)$ with $c = 1/(2C')$, we obtain that recovery by ℓ_1-minimization is successful with probability at least $1 - e^{-cm}$ provided

$$m \geq 2C's\ln(N/s). \tag{2.29}$$

This is the statement usually found in the literature.

The important point in the bound (2.29) is that the number of required samples only scales linearly in s up to the logarithmic factor $\ln(N/s)$ – in contrast to the quadratic scaling in the relation $m \geq 36(s-1)^2$ deduced from (2.26). Moreover, the ambient dimension N enters only very mildly into (2.29), and if N is large and s is rather small then m can be chosen significantly smaller than N and still allow for recovery by ℓ_1-minimization. In particular, an s-sparse $\mathbf{x}$ can be reconstructed exactly although at first sight the available information seems highly incomplete.

Let us note that (2.29) is optimal as can be shown by using lower bounds for Gelfand widths of the ℓ_1^N ball [54, 57]. In particular, the factor $\ln(N/s)$ cannot be improved.

3 Structured Random Matrices

While Gaussian and Bernoulli matrices ensure sparse recovery via ℓ_1-minimization with the optimal bound (2.28) on the number of measurements, they are of somewhat limited use in applications for several reasons. Often the design of the measurement matrix is subject to physical or other constraints of the application, or it is actually given to us without having the freedom to design anything, and therefore it is often

not justifiable that the matrix follows a Gaussian or Bernoulli distribution. Moreover, Gaussian or other unstructured matrices have the disadvantage that no fast matrix multiplication is available, which may speed up recovery algorithms significantly, so that large scale problems are not practicable with Gaussian or Bernoulli matrices. Even storing an unstructured matrix may be difficult.

From a computational and an application oriented view point it is desirable to have measurement matrices with structure. Since it is hard to rigorously prove good recovery conditions for deterministic matrices as outlined above, we will nevertheless allow randomness to come into play. This leads to the study of structured random matrices.

We will consider basically two types of structured random matrices. The larger part of these notes will be devoted to the recovery of randomly sampled functions that have a sparse expansion in terms of an orthonormal system $\{\psi_j, j = 1, \dots, N\}$ with uniformly bounded L^∞-norm, $\sup_{j\in[N]} \|\psi_j\|_\infty = \sup_{j\in[N]} \sup_x |\phi_j(x)| \leq K$. The corresponding measurement matrix has entries $(\psi_j(t_\ell))_{\ell,j}$, where the t_ℓ are random sampling points. So the structure is determined by the function system ψ_j, while the randomness comes from the sampling locations.

The random partial Fourier matrix, which consists of randomly chosen rows of the discrete Fourier matrix can be seen as a special case of this setup and was studied already in the very first papers on compressive sensing [19, 23]. It is important to note that in this case the fast Fourier transform (FFT) algorithm can be used to compute a fast application of a partial Fourier matrix in $\mathcal{O}(N \log(N))$ operations [30, 59, 137] – to be compared with the usual $\mathcal{O}(mN)$ operations for a matrix vector multiply with an $m \times N$ matrix. Commonly, $m \geq Cs \log(N)$ in compressive sensing, so that an $\mathcal{O}(N \log(N))$ matrix multiply implies a substantial complexity gain.

The second type of structured random matrices we will study are partial random circulant and Toeplitz matrices. They arise in applications where convolutions are involved. Since circulant and Toeplitz matrices can be applied efficiently using again the FFT, they are also of interest for computationally efficient sparse recovery.

Other types of structured random matrices, that will not be discussed here in detail, are the following.

- **Random Gabor System.** On $\mathbb{C}^m$ a time-shift or translation is the circular shift operator $(T_k g)_j = g_{j-k \mod m}$, while a frequency shift is the modulation operator $(M_\ell g)_j = e^{2\pi i \ell j/m} g_j$. Now fix a vector g and construct a matrix $A = A_g \in \mathbb{C}^{m\times m^2}$ by selecting its columns as the time-frequency shifts $M_\ell T_k g \in \mathbb{C}^m$, $\ell, k \in [m]$. Here the entries of g are chosen independently and uniformly at random from the torus $\{\mathbf{z} \in \mathbb{C}, |\mathbf{z}| = 1\}$. Then $A = A_g$ is a structured random matrix called a random Gabor system. Corresponding sparse recovery results can be found in [95, 96].

- **Random Demodulator.** This type of random matrix is motivated by analog to digital conversion. We refer to [133] for details.

3.1 Nonuniform versus Uniform Recovery

Showing recovery results for ℓ_1-minimization in connection with structured random matrices is more delicate than for unstructured Gaussian matrices. Nevertheless, we will try to get as close to the recovery condition (2.28) as possible. We will not be able to obtain precisely this condition, but we will only suffer from a slightly larger log-term. Our recovery bounds will have the form

$$m \geq Cs \log^\alpha(N/\varepsilon)$$

(or similar) for some $\alpha \geq 1$, where $\varepsilon \in (0,1)$ corresponds to the probability of failure. In particular, the important linear scaling of m in s up to log-factors is retained.

We will pursue different strategies in order to come up with rigorous recovery results. In particular, we distinguish between uniform and nonuniform recovery guarantees. A uniform recovery guaranty means that once the random matrix is chosen, then with high probability all sparse signals can be recovered. A nonuniform recovery result states only that each fixed sparse signal can be recovered with high probability using a random draw of the matrix. In particular, such weaker results allow in principle that the small exceptional set of matrices for which recovery may fail is dependent on the signal, in contrast to a uniform statement. Clearly, uniform recovery implies nonuniform recovery, but the converse is not true.

It is usually easier to obtain nonuniform recovery results for structured random matrices, and the provable bounds on the maximal allowed sparsity (or on the minimal number of measurements) are usually slightly worse for uniform recovery.

Uniform recovery is clearly guaranteed once we prove that the restricted isometry property of a random matrix holds with high probability. Indeed, the corresponding Theorems 2.6 or 2.7 are purely deterministic and guarantee recovery of all s-sparse signals once the restricted isometry constant δ_{2s} of the measurement matrix is small enough.

In order to obtain nonuniform recovery results we will use the recovery condition for individual vectors, Corollary 2.9. If the signal is fixed then also its support is fixed, and hence, applying Corollary 2.9 means in the end that only a weaker property than the restricted isometry property has to be checked for the random matrix. In order to simplify arguments even further we can also choose the signs of the non-zero coefficients of the sparse vector at random.

4 Random Sampling in Bounded Orthonormal Systems

An important class of structured random matrices is connected with random sampling of functions in certain finite dimensional function spaces. We require an orthonormal basis of functions which are uniformly bounded in the L^∞-norm. The most prominent example consists of the trigonometric system [19, 102, 104, 78]. In a discrete setup, the resulting matrix is a random partial Fourier matrix, which actually was the first

structured random matrix investigated in connection with compressive sensing [19, 23, 116].

4.1 Bounded Orthonormal Systems

Let $\mathcal{D} \subset \mathbb{R}^d$ be endowed with a probability measure ν. Further, let $\psi_1, \dots, \psi_N$ be an orthonormal system of complex-valued functions on $\mathcal{D}$, that is, for $j, k \in [N]$,

$$\int_{\mathcal{D}} \psi_j(t)\overline{\psi_k(t)} d\nu(t) = \delta_{j,k} = \begin{cases} 0 & \text{if } j \neq k, \\ 1 & \text{if } j = k. \end{cases} \tag{4.1}$$

The orthonormal system will be assumed to be uniformly bounded in L^∞,

$$\|\psi_j\|_\infty = \sup_{t \in \mathcal{D}} |\psi_j(t)| \leq K \quad \text{for all } j \in [N]. \tag{4.2}$$

The smallest value that the constant K can take is $K = 1$. Indeed,

$$1 = \int_{\mathcal{D}} |\psi_j(t)|^2 d\nu(x) \leq \sup_{t \in \mathcal{D}} |\psi_j(t)|^2 \int_{\mathcal{D}} d\nu(t) = K^2.$$

In the extreme case $K = 1$ we necessarily have $|\psi_j(t)| = 1$ for ν-almost all $t \in \mathcal{D}$.

Remark 4.1. (a) Note that *some* bound K can be found for most reasonable sets of functions ψ_j, $j \in [N]$. The crucial point of the boundedness condition (4.2) is that $K = \sup_{j \in [N]} \|\psi_j\|_\infty$ should ideally be independent of N, or at least depend only mildly on N, such as $K \leq C \ln^\alpha(N)$ for some $\alpha > 0$. Such a condition excludes for instance that the functions ψ_j are very localized in small regions of $\mathcal{D}$.

Expressed differently, the quotients $\|\psi_j\|_\infty / \|\psi_j\|_2$ should be uniformly bounded in j (in case that the functions ψ_j are not yet normalized); or at least grow only very slowly.

(b) It is not essential that $\mathcal{D}$ is a (measurable) subset of $\mathbb{R}^d$. This assumption was only made for convenience. In fact, $\mathcal{D}$ can be any measure space endowed with a probability measure ν.

We consider functions of the form

$$f(t) = \sum_{k=1}^{N} x_k \psi_k(t), \quad t \in \mathcal{D} \tag{4.3}$$

with coefficients $x_1, \dots, x_N \in \mathbb{C}$.

Let $t_1, \dots, t_m \in \mathcal{D}$ be some points and suppose we are given the sample values

$$y_\ell = f(t_\ell) = \sum_{k=1}^{N} x_k \psi_k(t_\ell) \ , \quad \ell = 1, \dots, m.$$

Introducing the sampling matrix $A \in \mathbb{C}^{m \times N}$ with entries

$$A_{\ell,k} = \psi_k(t_\ell) \ , \quad \ell = 1, \dots, m, \ k = 1, \dots, N, \tag{4.4}$$

the vector $\mathbf{y} = (y_1, \dots, y_m)^T$ of sample values (measurements) can be written in the form

$$\mathbf{y} = A\mathbf{x}, \tag{4.5}$$

where $\mathbf{x}$ is the vector of coefficients in (4.3).

Our task is to reconstruct the polynomial f — or equivalently its vector $\mathbf{x}$ of coefficients — from the vector of samples $\mathbf{y}$. We wish to perform this task with as few samples as possible. Without further knowledge this is clearly impossible if $m < N$. As common in compressive sensing we therefore assume sparsity.

A polynomial f of the form (4.3) is called s-sparse if its coefficient vector $\mathbf{x}$ is s-sparse. The problem of recovering an s-sparse polynomial from m sample values reduces then to solving (4.5) with a sparsity constraint, where A is the matrix in (4.4). We consider ℓ_1-minimization for this task.

Now we introduce randomness. We assume to this end that the sampling points $t_1, \dots, t_m$ are selected independently at random according to the probability measure ν. This means in particular that $\mathbb{P}(t_\ell \in B) = \nu(B)$, $\ell = 1, \dots, m$, for a measurable subset $B \subset \mathcal{D}$. The matrix A in (4.4) becomes then a structured random matrix.

Let us give examples of bounded orthonormal systems.

(i) **Trigonometric Polynomials.** Let $\mathcal{D} = [0, 1]$ and for $k \in \mathbb{Z}$ set

$$\psi_k(t) = e^{2\pi i k t}, \quad t \in [0, 1].$$

The probability measure ν is taken to be the Lebesgue measure on $[0, 1]$. Then for all $j, k \in \mathbb{Z}$,

$$\int_0^1 \psi_k(t)\overline{\psi_j(t)}dt = \delta_{j,k}. \tag{4.6}$$

The constant in (4.2) is clearly $K = 1$. For a subset $\Gamma \subset \mathbb{Z}$ of size N we then consider the trigonometric polynomials of the form

$$f(t) = \sum_{k \in \Gamma} x_k \psi_k(t) = \sum_{k \in \Gamma} x_k e^{2\pi i k t}.$$

A common choice is $\Gamma = \{-q, -q+1, \dots, q-1, q\}$ resulting in trigonometric polynomials of degree at most q (then $N = 2q+1$). We emphasize, however, that an arbitrary choice of $\Gamma \subset \mathbb{Z}$ of size $|\Gamma| = N$ is possible. Introducing sparsity on the coefficient vector $\mathbf{x} \in \mathbb{C}^N$ then leads to the notion of s-sparse trigonometric polynomials.

The sampling points $t_1, \dots, t_m$ will be chosen independently and uniformly at random from $[0, 1]$. The entries of the associated structured random matrix A are given by

$$A_{\ell,k} = e^{2\pi i k t_\ell} \ , \quad \ell = 1, \dots, m \ , \quad k \in \Gamma, \tag{4.7}$$

Such A is a Fourier type matrix, sometimes also called a nonequispaced Fourier matrix.

This example extends to multivariate trigonometric polynomials on $[0,1]^d$, $d \in \mathbb{N}$. Indeed, the monomials $\psi_{\mathbf{k}}(t) = e^{2\pi i \langle \mathbf{k}, t\rangle}$, $\mathbf{k} \in \mathbb{Z}^d$, $t \in [0,1]^d$, form an orthonormal system. For readers familiar with abstract harmonic analysis we mention that this example can be further generalized to characters of a compact commutative group. The corresponding measure will be the Haar measure of the group [50, 117].

The matrix A in (4.7) has a fast (approximate) matrix multiplication algorithm, called the non-equispaced fast Fourier transform (NFFT) [46, 101]. Similarly to the FFT, it has complexity $\mathcal{O}(N \log(N))$.

(ii) **Real Trigonometric Polynomials.** Instead of the complex exponentials above we may also take the real functions

$$\begin{aligned} \psi_{2k}(t) &= \sqrt{2}\cos(2\pi k t), \quad k \in \mathbb{N}_0, \quad \psi_0(t) = 1, \\ \psi_{2k+1}(t) &= \sqrt{2}\sin(2\pi k t), \quad k \in \mathbb{N}. \end{aligned} \tag{4.8}$$

They also form an orthonormal system on $[0,1]$ with respect to the Lebesgue measure and the constant in (4.2) is $K = \sqrt{2}$. The samples $t_1, \dots, t_m$ are chosen again according to the uniform distribution on $[0,1]$.

(iii) **Discrete Orthonormal Systems.** Let $U = (U_{tk}) \in \mathbb{C}^{N\times N}$ be a unitary matrix. The normalized columns $\sqrt{N}\mathbf{u}_k \in \mathbb{C}^N$, $k \in [N]$, then form an orthonormal system with respect to the discrete uniform probability measure on $[N]$, $\nu(B) = |B|/N$ for $B \subset [N]$; written out, this means

$$\frac{1}{N}\sum_{t=1}^{N} \sqrt{N}\mathbf{u}_k(t)\overline{\sqrt{N}\mathbf{u}_\ell(t)} = \langle \mathbf{u}_k, \mathbf{u}_\ell\rangle = \delta_{k,\ell}\,, \quad k,\ell \in [N].$$

Here, $\mathbf{u}_k(t) = U_{tk}$ denotes the tth entry of the kth column of U. The boundedness condition (4.2) requires that the normalized entries of U are bounded, i.e.,

$$\sqrt{N}\max_{k,t\in[N]} |U_{tk}| = \max_{k,t\in[N]} |\sqrt{N}\mathbf{u}_k(t)| \le K. \tag{4.9}$$

Choosing the points $t_1, \dots, t_m$ independently and uniformly at random from $[N]$ corresponds then to creating the random matrix A by selecting its rows independently and uniformly at random from the rows of $\sqrt{N}U$, that is,

$$A = \sqrt{N} R_T U\,,$$

where $R_T : \mathbb{C}^N \to \mathbb{C}^m$ denotes the random subsampling operator

$$(R_T \mathbf{z})_\ell = \mathbf{z}_{t_\ell}, \ell = 1, \dots, m. \tag{4.10}$$

Compressive sensing in this context yields the situation that only a small portion of the entries of $\tilde{\mathbf{y}} = \sqrt{N}U\mathbf{x} \in \mathbb{C}^N$ are observed of a sparse vector $\mathbf{x} \in \mathbb{C}^N$. In other words, $\mathbf{y} = R_T\tilde{\mathbf{y}} \in \mathbb{C}^m$, and we wish to recover $\mathbf{x}$ from the undersampled $\mathbf{y}$.

Note that it may happen with non-zero probability that a row of $\sqrt{N}U$ is selected more than once because the probability measure is discrete in this example. Hence, A is allowed to have repeated rows. One can avoid this effect by passing to a different probability model where the subset $\{t_1, \dots, t_m\} \subset [N]$ is selected uniformly at random among all subsets of $[N]$ of cardinality m. This probability model requires a slightly different analysis than the model described above, and we refer to [19, 23, 20, 55, 116, 130] for more information. The difference between the two models, however, is very slight in practice and the corresponding recovery results are almost the same.

(iv) **Partial Discrete Fourier Transform.** Our next example uses the discrete Fourier matrix $F \in \mathbb{C}^{N\times N}$ with entries

$$F_{\ell,k} = \frac{1}{\sqrt{N}} e^{2\pi i \ell k/N}, \quad \ell, k = 1, \dots, N. \tag{4.11}$$

It is well-known (and easy to see) that F is unitary. The constant in (4.2) or (4.9) is clearly $K = 1$. The result $\hat{\mathbf{x}} = F\mathbf{x}$ of a applying F to a vector is called the Fourier transform of $\mathbf{x}$. Applying the setup of the previous example to this sitatuation results in the problem of reconstructing a sparse vector $\mathbf{x}$ from m random entries of its Fourier transform $\hat{\mathbf{x}}$, that are independent and uniformly distributed on $\mathbb{Z}_N := \{\frac{k}{N}, k = 1, \dots, N\}$. The resulting matrix A is called random partial Fourier matrix. Such a matrix can also be seen as a special case of the non-equispaced Fourier type matrix in (4.7) with the points t_ℓ being chosen from the grid $\mathbb{Z}_N$ instead of from the whole interval $[0, 1]$. Note that the discrete Fourier matrix in (4.11) can also be extended to higher dimensions, i.e., to grids $\mathbb{Z}_N^d$ for $d \in \mathbb{N}$.

A crucial point for applications is that the Fourier transform has a fast algorithm for matrix-vector multiplication, the so called fast Fourier transform (FFT) [30, 137]. It computes the Fourier transform of a vector $\mathbf{x} \in \mathbb{C}^N$ in complexity $\mathcal{O}(N \log(N))$.

(v) **Incoherent Bases.** Let $V, W \in \mathbb{C}^{N\times N}$ be two unitary matrices. Their columns $(\mathbf{v}_\ell)_{\ell=1}^N$ and $(\mathbf{w}_\ell)_{\ell=1}^N$ form two orthonormal bases of $\mathbb{C}^N$. Assume that a vector $\mathbf{z} \in \mathbb{C}^N$ is sparse with respect to the basis $(\mathbf{v}_\ell)$ rather than the canonical basis, that is, $\mathbf{z} = V\mathbf{x}$ for a sparse vector $\mathbf{x}$. Further, assume that $\mathbf{z}$ is sampled with respect to the basis $(\mathbf{w}_\ell)$, i.e., we obtain measurements

$$y_k = \langle z, \mathbf{w}_{t_k} \rangle, \quad k = 1, \dots, m$$

with $T := \{t_1, \dots, t_m\} \subset [N]$. In matrix vector form this can be written as

$$\mathbf{y} = R_T W^* \mathbf{z} = R_T W^* V \mathbf{x},$$

where R_T is again the random sampling operator (4.10). Defining the unitary matrix $U = W^*V \in \mathbb{C}^{N\times N}$ we are back to the situation of the third example. The condition (4.9) now reads

$$\sqrt{N} \max_{\ell,k\in[N]} |\langle \mathbf{v}_\ell, \mathbf{w}_k\rangle| \leq K. \tag{4.12}$$

The quantity on the left hand side (without the $\sqrt{N}$) is known as the mutual coherence of the bases $(\mathbf{v}_\ell), (\mathbf{w}_\ell)$, and they are called incoherent if K can be chosen small. The two previous examples also fall into this setting by choosing one of the bases as the canonical basis, $W = \mathrm{Id} \in \mathbb{C}^N$. The Fourier basis and the canonical basis are actually maximally incoherent, since then $K = 1$.

(vi) **Haar-Wavelets and Noiselets.** This example is a special case of the previous one, which is potentially useful for image processing applications. It is convenient to start with a continuous description of Haar-wavelets and noiselets [29], and then pass to the discrete setup via sampling. The Haar scaling function on $\mathbb{R}$ is defined as the characteristic function of the interval $[0, 1)$,

$$\phi(x) = \chi_{[0,1)}(x) = \begin{cases} 1 & \text{if } x \in [0,1), \\ 0 & \text{otherwise.} \end{cases} \tag{4.13}$$

The Haar wavelet is then defined as

$$\psi(x) = \phi(2x) - \phi(2x-1) = \begin{cases} 1 & \text{if } x \in [0,1/2), \\ -1 & \text{if } x \in [1/2,1), \\ 0 & \text{otherwise.} \end{cases} \tag{4.14}$$

Further, denote

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad \phi_k(x) = \phi(x-k), \quad x \in \mathbb{R}, j \in \mathbb{Z}, k \in \mathbb{Z}. \tag{4.15}$$

It is well-known [138] (and can easily be seen) that, for $n \in \mathbb{N}$, the Haar-wavelet system

$$\mathbf{\Psi}_n := \{\phi_k, k \in \mathbb{Z}\} \cup \{\psi_{j,k},\, k = 0, \dots, 2^j - 1, j = 0, \dots, n-1\} \tag{4.16}$$

forms an orthonormal basis of

$$V_n = \{f \in L^2([0,1]) : f \text{ is constant on } [k2^{-n}, (k+1)2^{-n}), k = 0, \dots, 2^n - 1\}.$$

Now let $N = 2^n$ for some $n \in \mathbb{N}$. Since the functions $\psi_{j,k}$, $j \leq n-1$, are constant on intervals of the form $[2^{-n}k, 2^{-n}(k+1))$ we conclude that the vectors $\widetilde{\phi}, \widetilde{\psi}^{(j,k)} \in \mathbb{C}^N$, $j = 0, \dots, n-1, k = 0, \dots, 2^j - 1$, with entries

$$\widetilde{\phi}_t = 2^{-n/2}\phi(t/N), \quad t = 0, \dots, N-1$$

$$\widetilde{\psi}_t^{(j,k)} = 2^{-n/2}\psi_{j,k}(t/N), \quad t = 0, \dots, N-1$$

form an orthonormal basis of $\mathbb{C}^N$. We collect these vectors as the columns of a unitary matrix $\Psi \in \mathbb{C}^{N\times N}$.

Next we introduce the noiselet system on $[0,1]$. Let $g_1 = \phi = \chi_{[0,1)}$ be the Haar scaling function and define, for $r \geq 1$, recursively the complex-valued functions

$$g_{2r}(x) = (1-i)g_r(2x) + (1+i)g_r(2x-1),$$
$$g_{2r+1}(x) = (1+i)g_r(2x) + (1-i)g_r(2x-1).$$

It is shown in [29] that the functions $\{2^{-n/2}g_r, r = 2^n, \dots, 2^{n+1}-1\}$ form an orthonormal basis of V_n. The key property for us consists in the fact that they are maximally incoherent with respect to the Haar basis. Indeed, Lemma 10 in [29] states that

$$\left| \int_0^1 g_r(x)\psi_{j,k}(x)dx \right| = 1 \quad \text{provided } r \geq 2^j - 1, \quad 0 \leq k \leq 2^j - 1. \quad (4.17)$$

For the discrete noiselet basis on $\mathbb{C}^N$, $N = 2^n$, we take the vectors

$$\widetilde{g}_t^{(r)} = 2^{-n} g_{N+r}(t/N), \quad r = 0, \dots, N-1, \quad t = 0, \dots, N-1.$$

Again, since the functions $g_{N+r}, r = 0, \dots, N-1$, are constant on intervals of the form $[2^{-n}k, 2^{-n}(k+1))$ it follows that the vectors $\widetilde{g}^{(r)}$, $r = 0, \dots, N-1$, form an orthonormal basis of $\mathbb{C}^N$. We collect these as columns into a unitary matrix $G \in \mathbb{C}^{N\times N}$. Due to (4.17) the unitary matrix $U = G^*\Psi \in \mathbb{C}^{N\times N}$ satisfies (4.9) with $K = 1$ – or in other words, the incoherence condition (4.12) for the Haar basis and the noiselet basis holds with the minimal constant $K = 1$.

Due to the their recursive definition, both the Haar wavelet transform and the noiselet transform, that is, the application of Ψ and G and their adjoints, come with a fast algorithm that computes a matrix vector multiply in $\mathcal{O}(N\log(N))$ time.

As a simple signal model, images or other types of signals are sparse in the Haar wavelet basis. The described setup corresponds to randomly sampling such functions with respect to noiselets. For more information on wavelets we refer to [27, 31, 83, 138].

(vii) **Legendre polynomials**. The Legendre polynomials P_j are a system of orthogonal polynomials, where P_j is a polynomial of precise degree j, and orthonormality is with respect to the normalized Lebesgue measure $dx/2$ on $[-1,1]$. Their supremum norm is given by $\|P_j\|_\infty = \sqrt{2j+1}$, so considering the polynomials $P_j, j = 0, \dots, N-1$, yields the constant $K = \sqrt{2N-1}$. Unfortunately, K grows therefore rather quickly with N. This problem can be avoided with a trick. One takes sampling points with respect to the "Chebyshev" measure $d\nu(x) = \pi^{-1}(1-x^2)^{-1/2}dx$ and uses a preconditioned measurement matrix. We refer to [106] for details.

Figure 1 shows an example of exact recovery of a 10-sparse vector in dimension 300 from 30 Fourier samples (example (iv) above) using ℓ_1-minimization. For comparison the reconstruction via ℓ_2-minimization is also shown.

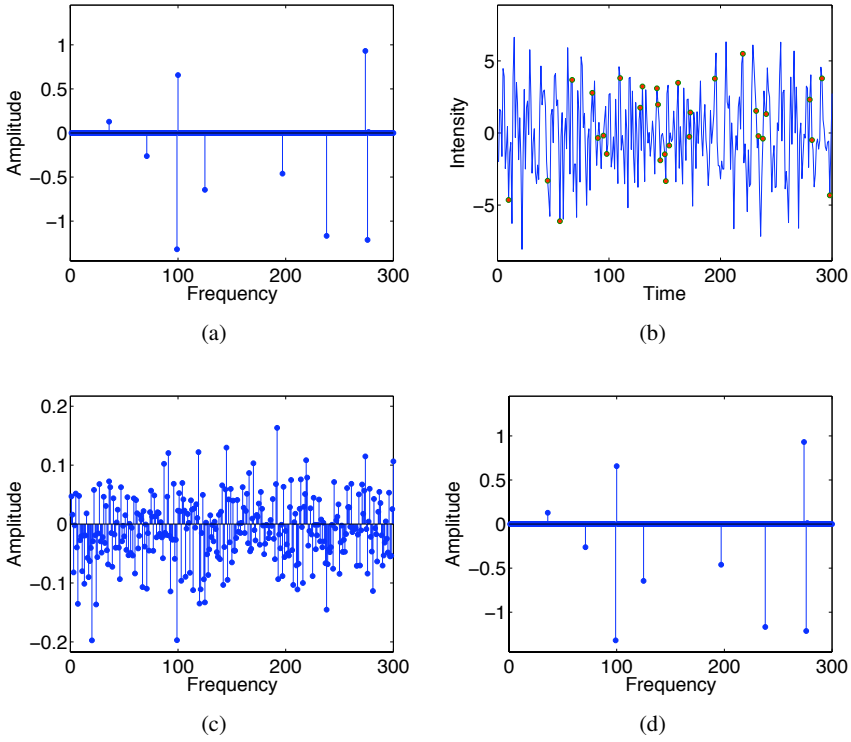

Figure 1. (a) 10-sparse Fourier spectrum, (b) time domain signal of length 300 with 30 samples, (c) reconstruction via ℓ_2-minimization, (d) exact reconstruction via ℓ_1-minimization

4.2 Nonuniform Recovery

We start with a nonuniform recovery result that additionally assumes that the signs of the non-zero entries of the signal $\mathbf{x}$ are chosen at random.

Theorem 4.2. *Let $S \subset [N]$ be of cardinality $|S| = s$ and let $\boldsymbol{\epsilon} = (\epsilon_\ell)_{\ell \in S} \in \mathbb{C}^s$ be a sequence of independent random variables that take the values ± 1 with equal probability. Alternatively, the ϵ_ℓ may be uniformly distributed on the torus $\{\mathbf{z} \in \mathbb{C}, |\mathbf{z}| = 1\}$. Let $\mathbf{x}$ be an s-sparse vector with support S and $\operatorname{sgn}(\mathbf{x}^S) = \boldsymbol{\epsilon}$.*

Let $A \in \mathbb{C}^{m \times N}$ be the sampling matrix (4.4) *associated to an orthonormal system that satisfies the boundedness condition* (4.2) *for some constant $K \geq 1$. Assume*

that the random sampling points $t_1, \dots, t_m$ are chosen independently and distributed according to the orthogonalization measure ν. Assume that

$$m \geq CK^2 s \ln^2(6N/\varepsilon), \tag{4.18}$$

where $C \approx 26.25$. Set $\mathbf{y} = A\mathbf{x}$. Then with probability at least $1 - \varepsilon$ the vector $\mathbf{x}$ is the unique solution to the ℓ_1-minimization problem (2.12)*.*

The proof will be contained in Chapter 7. With more effort (which we will not do here), the exponent 2 at the log-term in (4.18) can be replaced by 1. More precisely, one may obtain also the following sufficient recovery condition [55]

$$m \geq C_1 K^2 \max\{s, C_2 \ln(6N/\varepsilon)\} \ln(6N/\varepsilon) \tag{4.19}$$

with (reasonable) constants $C_1, C_2 > 0$. In the special case of a discrete orthonormal system (see example (3) in the previous section), a version of Theorem 4.2 with recovery condition (4.19) was shown in [20] under a slightly different probability model.

The constants provided in (4.18) and (4.19) are likely not optimal. Numerical experiments suggest much better values. In the special case of the Fourier matrix (examples (1) and (4) in the previous section) indeed slightly better constants are available [65, 102, 55]. However, we note that condition (4.19) gives an estimate that is valid for any possible support set S of size $|S| \leq s$. Clearly, it is impossible to test all such subsets numerically. So only limited conclusions on the constants in (4.19) and (4.18) can be drawn from numerical experiments.

In case of random sampling in the Fourier system (examples (1) and (4) in the previous section) the assumption of randomness in the sign pattern of the non-zero entries of $\mathbf{x}$ can be removed [19, 102].

Theorem 4.3. *Let $\mathbf{x} \in \mathbb{C}^N$ be s-sparse. Assume A is the random Fourier type matrix* (4.7) *or the random partial Fourier matrix of example (4) above. If*

$$m \geq Cs \log(N/\varepsilon)$$

then $\mathbf{x}$ is the unique solution of the ℓ_1-minimization problem (2.12) *with probability at least $1 - \varepsilon$.*

The techniques of the proof of this theorem [19, 102] heavily use the algebraic structure of the Fourier system and do not easily extend to general bounded orthonormal systems. In fact, the general case is still open.

4.3 Uniform Recovery

Our main theorem concerning the recovery of sparse polynomials in bounded orthonormal systems from random samples reads as follows.

Theorem 4.4. *Let $A \in \mathbb{C}^{m\times N}$ be the sampling matrix* (4.4) *associated to an orthonormal system that satisfies the boundedness condition* (4.2) *for some constant $K \geq 1$. Assume that the random sampling points $t_1, \dots, t_m$ are chosen independently at random according to the orthogonalization measure ν. Suppose*

$$\frac{m}{\ln(m)} \geq CK^2 s \ln^2(s) \ln(N), \tag{4.20}$$

$$m \geq DK^2 s \ln(\varepsilon^{-1}), \tag{4.21}$$

where $C, D > 0$ are some universal constants. Then with probability at least $1-\varepsilon$ every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is recovered from the samples

$$\mathbf{y} = A\mathbf{x} = \left(\sum_{j=1}^{N} x_j \phi_j(t_\ell) \right)_{\ell=1}^{m}$$

by ℓ_1-minimization (2.12).

Moreover, with probability at least $1-\varepsilon$ the following holds for every $\mathbf{x} \in \mathbb{C}^N$. Let noisy samples $\mathbf{y} = A\mathbf{x} + \mathbf{e}$ with

$$\|\mathbf{e}\|_2 = \sqrt{\sum_{\ell=1}^{m} |e_\ell|^2} \leq \eta\sqrt{m}$$

be given and let $\mathbf{x}^*$ be the solution of the ℓ_1-minimization problem (2.20)*, where η is replaced by $\eta\sqrt{m}$. Then*

$$\|\mathbf{x} - \mathbf{x}^*\|_2 \leq c\frac{\sigma_s(\mathbf{x})_1}{\sqrt{s}} + d\eta$$

for suitable constants $c, d > 0$.

This result is proven in Chapter 8 by estimating the restricted isometry constants δ_s of A. Thereby, also explicit constants are provided, see Theorem 8.4. The reader will notice that its proof is considerably more involved than the one of Theorem 4.2.

Remark 4.5. We may choose ε such that there is equality in (4.21). Then condition (4.20) implies recovery with probability at least

$$1 - N^{-\gamma \ln(m) \ln^2(s)}$$

where $\gamma = C/D$. A condition that is easier to remember is derived by noting that $s \leq N$ and $m \leq N$ (otherwise, we are not in the range of interest for compressive sensing). Indeed,

$$m \geq CK^2 s \ln^4(N) \tag{4.22}$$

implies recovery by ℓ_1-minimization with probability at least $1 - N^{-\gamma \ln^3(N)}$.

E. Candès and T. Tao [23] obtained the sufficient condition (4.22) in case of the random partial Fourier matrix with an exponent 6 instead of 4 at the $\ln(N)$ term. M. Rudelson and R. Vershynin [116] improved this to an exponent 5 at the $\ln(N)$-term; or alternatively to an exponent 4 for constant probability ε, see also Theorem 8.1 below. The condition (4.22) with exponent 4 and super-polynomially decreasing failure probability $N^{-\gamma \ln(N)^3}$ is presently the best known result. (In the Fourier case this is already contained in the proof of the main result in [104], but the author did not realize at that time that this was actually a slight improvement over the estimate of Rudelson and Vershynin in [116].) Our proof in Chapter 8 follows the ideas of Rudelson and Vershynin in [116] with some modifications and the mentioned improvements.

5 Partial Random Circulant Matrices

This section will be devoted to a different type of structured random matrices, which are important in applications such as wireless communications and radar, see [4, 68, 110]. We will study partial random circulant matrices and partial random Toeplitz matrices. Presently, there are less recovery results available than for the structured random matrices in the preceding section. In particular, a good estimate for the restricted isometry constants is still under investigation at the time of writing. (The estimates in [4, 68] only provide a quadratic scaling of the number of measurements in terms of the sparsity, similar to (2.26).) Therefore, we will only be able to present a nonuniform recovery result in the spirit of Theorem 4.2, which is a slight improvement of the main result in [105]. We believe that the mathematical approach to its proof should be of interest on its own.

We consider the following measurement matrices. For $\mathbf{b} = (b_0, b_1, \dots, b_{N-1}) \in \mathbb{C}^N$ we define its associated circulant matrix $\Phi = \Phi(\mathbf{b}) \in \mathbb{C}^{N \times N}$ by setting

$$\Phi_{k,j} = b_{j-k \mod N}, \quad k, j = 1, \dots, N.$$

Note that the application of Φ to a vector is the convolution,

$$(\Phi \mathbf{x})_j = (\mathbf{x} * \widetilde{\mathbf{b}})_j = \sum_{\ell=1}^{N} x_\ell \widetilde{b}_{j-\ell \mod N},$$

where $\widetilde{b}_j = b_{N-j}$. Similarly, for a vector $\mathbf{c} = (c_{-N+1}, c_{-N+2}, \dots, c_{N-1})$ its associated Toeplitz matrix $\Psi = \Psi(\mathbf{c}) \in \mathbb{C}^{N \times N}$ has entries $\Psi_{k,j} = c_{j-k}$, $k, j = 1, \dots, N$.

Now we choose an arbitrary subset $\Theta \subset [N]$ of cardinality $m < N$ and let the partial circulant matrix $\Phi^\Theta = \Phi^\Theta(\mathbf{b}) \in \mathbb{C}^{m \times N}$ be the submatrix of Φ consisting of the rows indexed by Θ. The partial Toeplitz matrix $\Psi^\Theta = \Psi^\Theta(\mathbf{c}) \in \mathbb{C}^{m \times N}$ is defined similarly. For the purpose of this exposition we will choose the vectors $\mathbf{b}$ and $\mathbf{c}$ as Rademacher sequences, that is, the entries of $\mathbf{b}$ and $\mathbf{c}$ are independent random variables that take the value ± 1 with equal probability. Standard Gaussian vectors or

Steinhaus sequences (independent random variables that are uniformly distributed on the complex torus) are possible as well.

It is important from a computational viewpoint that circulant matrices can be diagonalized using the discrete Fourier transform [59]. Therefore, there is a fast matrix vector multiplication algorithm for partial circulant matrices of complexity $\mathcal{O}(N \log(N))$ that uses the FFT. Since Toeplitz matrices can be seen as submatrices of circulant matrices [59], this remark applies to partial Toeplitz matrices as well.

Of particular interest is the case $N = mL$ with $L \in \mathbb{N}$ and $\Theta = \{L, 2L, \dots, mL\}$. Then the application of $\Phi^\Theta(\mathbf{b})$ and $\Psi^\Theta(\mathbf{c})$ corresponds to (periodic or non-periodic) convolution with the sequence $\mathbf{b}$ (or $\mathbf{c}$, respectively) followed by a downsampling by a factor of L. This setting was studied numerically in [132] by J. Tropp et al. (using orthogonal matching pursuit instead of ℓ_1-minimization). Also of interest is the case $\Theta = [m]$ which was investigated in [4, 68].

Since Toeplitz matrices can be embedded into circulant matrices as just mentioned, we will deal only with the latter in the following. The result below (including its proof) holds without a difference (and even with the same constants) for Toeplitz matrices as well. Similarly to Theorem 4.2 we deal with nonuniform recovery, where additionally the signs $x_j/|x_j|$ of the non-zero coefficients of the vector x to be recovered are chosen at random.

Theorem 5.1. *Let $\Theta \subset [N]$ be an arbitrary (deterministic) set of cardinality m. Let $\mathbf{x} \in \mathbb{C}^N$ be s-sparse such that the signs of its non-zero entries form a Rademacher or Steinhaus sequence. Choose $\mathbf{b} = \boldsymbol{\epsilon} \in \mathbb{R}^N$ to be a Rademacher sequence. Let $\mathbf{y} = \Phi^\Theta(\boldsymbol{\epsilon})\mathbf{x} \in \mathbb{C}^m$. Then*

$$m \geq 57 s \ln^2(17N^2/\varepsilon) \tag{5.1}$$

implies that with probability at least $1 - \varepsilon$ the vector $\mathbf{x}$ is the unique solution to the ℓ_1-minimization problem (2.12).

The proof of Theorem 5.1 will be presented in Chapter 9. We note that the exponent 2 of the log-term in (5.1) is a slight improvement over an exponent 3 present in the main result of [105]. The constant 57 is very likely not optimal. With the much more technical (and combinatorial) approach of [19, 102, 95] we expect that the randomness in the signs can be removed and the exponent 2 at the log-factor can be improved to 1.

6 Tools from Probability Theory

The proofs of the results presented in the two previous chapters will require tools from probability theory that might not be part of an introductory course on probability. This chapter collects the necessary background. We will only assume familiarity of the reader with basic probability theory that can be found in most textbooks on the subject, see for instance [63, 112].

In the following we discuss the relation of moments and tail estimates, symmetrization, decoupling, and scalar and noncommutative Khintchine inequalities. The latter represent actually a very powerful tool that presently does not seem to be widely acknowledged. Furthermore, we present Dudley's inequality on the expectation of the supremum of a subgaussian process. Much more material of a similar flavor can be found in the monographs [36, 73, 79, 80, 125, 134].

6.1 Basics on Probability

In this section we recall some important facts from basic probability theory. Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, where Σ denotes a σ-algebra on the sample space Ω and $\mathbb{P}$ a probability measure on (Ω, Σ). The probability of an event $B \in \Sigma$ is denoted by

$$\mathbb{P}(B) = \int_B d\mathbb{P}(\omega) = \int_\Omega I_B(\omega) d\mathbb{P}(\omega),$$

where the characteristic function $I_B(\omega)$ takes the value 1 if $\omega \in B$ and 0 otherwise. The union bound (or Bonferroni's inequality, or Boole's inequality) states that for a collection of events $B_\ell \in \Sigma$, $\ell = 1, \dots, n$, we have

$$\mathbb{P}\left(\bigcup_{\ell=1}^n B_\ell\right) \leq \sum_{\ell=1}^n \mathbb{P}(B_\ell). \tag{6.1}$$

We assume knowledge of basic facts on random variables. The expectation or mean of a random variable X is denoted by

$$\mathbb{E}X = \int_\Omega X(\omega) d\mathbb{P}(\omega).$$

The quantities $\mathbb{E}|X|^p$, $0 < p < \infty$, are called (absolute) moments. For $1 \leq p < \infty$, $(\mathbb{E}|\mathbf{X}|^p)^{1/p}$ defines a norm on the $L^p(\Omega, \mathbb{P})$-space of all p-integrable random variables, in particular, the triangle inequality

$$(\mathbb{E}|X+Y|^p)^{1/p} \leq (\mathbb{E}|X|^p)^{1/p} + (\mathbb{E}|Y|^p)^{1/p} \tag{6.2}$$

holds for $X, Y \in L^p(\Omega, \mathbb{P}) = \{X \text{ measurable }, \mathbb{E}|X|^p < \infty\}$.

Let $p, q \geq 1$ with $1/p + 1/q = 1$, Hölder's inequality states that $|\mathbb{E}XY| \leq (\mathbb{E}|X|^p)^{1/p} (\mathbb{E}|Y|^q)^{1/q}$ for random variables X, Y. The special case $p = q = 2$ is the Cauchy-Schwarz inequality. It follows from Hölder's inequality that for all $0 < p \leq q < \infty$,

$$(\mathbb{E}|X|^p)^{1/p} \leq (\mathbb{E}|X|^q)^{1/q}. \tag{6.3}$$

Absolute moments can be computed by means of the following formula.

Proposition 6.1. *The absolute moments of a random variable X can be expressed as*

$$\mathbb{E}|X|^p = p\int_0^\infty \mathbb{P}(|X| \geq t)\, t^{p-1} dt, \quad p > 0.$$

Proof. Recall that $I_{\{|X|^p \geq x\}}$ is the random variable that takes the value 1 on the event $|X|^p \geq x$ and 0 otherwise. Using Fubini's theorem we derive

$$\begin{aligned}\mathbb{E}|X|^p &= \int_\Omega |X|^p d\mathbb{P} = \int_\Omega \int_0^{|X|^p} dx d\mathbb{P} = \int_\Omega \int_0^\infty I_{\{|X|^p \geq x\}} dx d\mathbb{P} \\ &= \int_0^\infty \int_\Omega I_{\{|X|^p \geq x\}} d\mathbb{P} dx = \int_0^\infty \mathbb{P}(|X|^p \geq x) dx \\ &= p\int_0^\infty \mathbb{P}(|X|^p \geq t^p) t^{p-1} dt = p\int_0^\infty \mathbb{P}(|X| \geq t) t^{p-1} dt,\end{aligned}$$

where we also applied a change of variables. □

The function $t \mapsto \mathbb{P}(|X| \geq t)$ is called the tail of X. The Markov inequality is a simple way of estimating a tail.

Theorem 6.2. *(Markov inequality) Let X be a random variable. Then*

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}|X|}{t} \quad \textit{for all } t > 0.$$

Proof. Note that $\mathbb{P}(|X| \geq t) = \mathbb{E}I_{\{|X| \geq t\}}$ and $tI_{\{|X| \geq t\}} \leq |X|$. Hence, $t\mathbb{P}(|X| \geq t) = \mathbb{E}tI_{\{|X| \geq t\}} \leq \mathbb{E}|X|$ and the proof is complete. □

A random vector $\mathbf{X} = (X_1, \ldots, X_n)^T \in \mathbb{R}^n$ is a collection of n random variables X_ℓ on a common probability space. Its expectation is the vector

$$\mathbb{E}\mathbf{X} = (\mathbb{E}X_1, \ldots, \mathbb{E}X_n)^T \in \mathbb{R}^n.$$

A complex random vector $\mathbf{Z} = \mathbf{X} + i\mathbf{Y} \in \mathbb{C}^n$ is a special case of a $2n$-dimensional real random vector $(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{2n}$.

A collection of random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_N \in \mathbb{C}^n$ is called (stochastically) independent if for all measurable subsets $B_1, \ldots, B_N \subset \mathbb{C}^n$,

$$\mathbb{P}(\mathbf{X}_1 \in B_1, \mathbf{X}_2 \in B_2, \ldots, \mathbf{X}_N \in B_N) = \mathbb{P}(\mathbf{X}_1 \in B_1)\mathbb{P}(\mathbf{X}_2 \in B_2) \cdots \mathbb{P}(\mathbf{X}_N \in B_N).$$

Functions of independent random vectors are again independent. A random vector $\mathbf{X}'$ in $\mathbb{C}^n$ will be called an independent copy of $\mathbf{X}$ if $\mathbf{X}$ and $\mathbf{X}'$ are independent and have the same distribution, that is, $\mathbb{P}(\mathbf{X} \in B) = \mathbb{P}(\mathbf{X}' \in B)$ for all measurable $B \subset \mathbb{C}^n$.

Jensen's inequality reads as follows.

Theorem 6.3. *(Jensen's inequality) Let $f : \mathbb{C}^n \to \mathbb{R}$ be a convex function, and let $\mathbf{X} \in \mathbb{C}^n$ be a random vector. Then*

$$f(\mathbb{E}\mathbf{X}) \leq \mathbb{E} f(\mathbf{X}) \, . \tag{6.4}$$

Finally, we state the Borel-Cantelli lemma.

Lemma 6.4. *(Borel-Cantelli) Let $A_1, A_2, \ldots \in \Sigma$ be events and let*

$$A^* = \limsup_{n\to\infty} A_n = \cap_{n=1}^{\infty} \cup_{m=n}^{\infty} A_m.$$

If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$ then $\mathbb{P}(A^*) = 0$.

Proof. Since $A^* \subset \bigcup_{m=n}^{\infty} A_m$ for all n, it holds $\mathbb{P}(A^*) \leq \sum_{m=n}^{\infty} \mathbb{P}(A_m) \to 0$ as $n \to \infty$ whenever $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$. □

6.2 Moments and Tails

It will be very crucial for us that tails of random variables can be estimated by means of their moments. The next statement is rather simple but very effective, see also [130].

Proposition 6.5. *Suppose Z is a random variable satisfying*

$$(\mathbb{E}|Z|^p)^{1/p} \leq \alpha \beta^{1/p} p^{1/\gamma} \quad \text{for all } p \geq p_0$$

for some constants $\alpha, \beta, \gamma, p_0 > 0$. Then

$$\mathbb{P}(|Z| \geq e^{1/\gamma} \alpha u) \leq \beta e^{-u^\gamma/\gamma}$$

for all $u \geq p_0^{1/\gamma}$.

Proof. By Markov's inequality, Theorem 6.2, we obtain for an arbitrary $\kappa > 0$

$$\mathbb{P}(|Z| \geq e^{\kappa} \alpha u) \leq \frac{\mathbb{E}|Z|^p}{(e^\kappa \alpha u)^p} \leq \beta \left(\frac{\alpha p^{1/\gamma}}{e^\kappa \alpha u} \right)^p .$$

Choose $p = u^\gamma$ and the optimal value $\kappa = 1/\gamma$ to obtain the claim. □

Also a converse of the above proposition can be shown [55, 80]. Important special cases are $\gamma = 1, 2$. In particular, if $(\mathbb{E}|Z|^p)^{1/p} \leq \alpha \beta^{1/p} \sqrt{p}$ for all $p \geq 2$ then Z satisfies the subgaussian tail estimate.

$$\mathbb{P}(|Z| \geq e^{1/2} \alpha u) \leq \beta e^{-u^2/2} \quad \text{for all } u \geq \sqrt{2}. \tag{6.5}$$

For random variables satisfying a subgaussian tail estimate, the following useful estimate of the expectation of their maximum can be shown [80].

Lemma 6.6. *Let $X_1, \dots, X_M$ be random variables satisfying*

$$\mathbb{P}(|X_\ell| \geq u) \leq \beta e^{-u^2/2} \quad \text{for } u \geq \sqrt{2}, \quad \ell = 1, \dots, M,$$

for some $\beta \geq 1$. Then

$$\mathbb{E} \max_{\ell=1,\dots,M} |X_\ell| \leq C_\beta \sqrt{\ln(4\beta M)}$$

with $C_\beta \leq \sqrt{2} + \frac{1}{4\sqrt{2}\ln(4\beta)}$.

Proof. According to Proposition 6.1 we have, for some $\alpha \geq \sqrt{2}$,

$$\begin{aligned}
\mathbb{E} \max_{\ell=1,\dots,M} |X_\ell| &= \int_0^\infty \mathbb{P}\left(\max_{\ell=1,\dots,M} |X_\ell| > u \right) du \\
&\leq \int_0^\alpha 1 du + \int_\alpha^\infty \mathbb{P}\left(\max_{\ell=1,\dots,M} |X_\ell| > u \right) du \leq \alpha + \int_\alpha^\infty \sum_{\ell=1}^M \mathbb{P}(|X_\ell| > u) du \\
&\leq \alpha + M\beta \int_\alpha^\infty e^{-u^2/2} du.
\end{aligned}$$

In the second line we have applied the union bound. Using Proposition 10.2 in the Appendix we obtain

$$\mathbb{E} \max_{\ell=1,\dots,M} |X_\ell| \leq \alpha + \frac{M\beta}{\alpha} e^{-\alpha^2/2}.$$

Now we choose $\alpha = \sqrt{2\ln(4\beta M)} \geq \sqrt{2\ln(4)} \geq \sqrt{2}$. This yields

$$\begin{aligned}
\mathbb{E} \max_{\ell=1,\dots,M} |X_\ell| &\leq \sqrt{2\ln(4\beta M)} + \frac{1}{4\sqrt{2\ln(4\beta M)}} \\
&= \left(\sqrt{2} + \frac{1}{4\sqrt{2}\ln(4\beta M)} \right) \sqrt{\ln(4\beta M)} \leq C_\beta \sqrt{\ln(4\beta M)}
\end{aligned}$$

by our choice of C_β. The proof is completed. □

6.3 Rademacher Sums and Symmetrization

A Rademacher variable is presumably the simplest random variable. It takes the values $+1$ or -1, each with probability $1/2$. A sequence $\boldsymbol{\epsilon}$ of independent Rademacher variables $\epsilon_j, j = 1, \dots, M$, is called a Rademacher sequence. The technique of symmetrization leads to so called Rademacher sums $\sum_{j=1}^M \epsilon_j x_j$ where the x_j are scalars, vectors or matrices. Although quite simple, symmetrization is very powerful because there are nice estimates for Rademacher sums available – the so called Khintchine inequalities to be treated later on.

A random vector $\mathbf{X}$ is called symmetric, if $\mathbf{X}$ and $-\mathbf{X}$ have the same distribution. In this case $\mathbf{X}$ and $\epsilon\mathbf{X}$, where ϵ is a Rademacher variable independent of $\mathbf{X}$, have the same distribution as well.

The following lemma, see also [80, 36], is the key to symmetrization.

Lemma 6.7. *(Symmetrization) Assume that $\boldsymbol{\xi} = (\xi_j)_{j=1}^M$ is a sequence of independent random vectors in $\mathbb{C}^n$ equipped with a (semi-)norm $\|\cdot\|$, having expectations $\mathbf{x}_j = \mathbb{E}\xi_j$. Then for $1 \leq p < \infty$*

$$\left(\mathbb{E}\|\sum_{j=1}^M (\xi_j - \mathbf{x}_j)\|^p\right)^{1/p} \leq 2\left(\mathbb{E}\|\sum_{j=1}^M \epsilon_j \xi_j\|^p\right)^{1/p}, \tag{6.6}$$

where $\boldsymbol{\epsilon} = (\epsilon_j)_{j=1}^N$ is a Rademacher sequence independent of $\boldsymbol{\xi}$.

Proof. Let $\boldsymbol{\xi}' = (\xi_1', \dots, \xi_M')$ denote an independent copy of the sequence of random vectors $(\xi_1, \dots, \xi_M)$. Since $\mathbb{E}\xi_j' = \mathbf{x}_j$ an application of Jensen's inequality (6.4) yields

$$E := \mathbb{E}\|\sum_{j=1}^M (\xi_j - \mathbf{x}_j)\|^p = \mathbb{E}\|\sum_{j=1}^M (\xi_j - \mathbb{E}\xi_j')\|^p \leq \mathbb{E}\|\sum_{j=1}^M (\xi_j - \xi_j')\|^p.$$

Now observe that $(\xi_j - \xi_j')_{j=1}^M$ is a vector of independent symmetric random variables; hence, it has the same distribution as $(\epsilon_j(\xi_j - \xi_j'))_{j=1}^M$. The triangle inequality gives

$$\begin{aligned} E^{1/p} &\leq (\mathbb{E}\|\sum_{j=1}^M \epsilon_j(\xi_j - \xi_j')\|^p)^{1/p} \leq (\mathbb{E}\|\sum_{j=1}^M \epsilon_j \xi_j\|^p)^{1/p} + (\mathbb{E}\|\sum_{j=1}^M \epsilon_j \xi_j'\|^p)^{1/p} \\ &= 2(\mathbb{E}\|\sum_{j=1}^M \epsilon_j \xi_j\|^p)^{1/p}. \end{aligned}$$

The last equality is due to the fact that $\boldsymbol{\xi}'$ is an independent copy of $\boldsymbol{\xi}$. □

Note that this lemma holds also in infinite-dimensional spaces [80]. Since it is rather technical to introduce random vectors in infinite dimensions we stated the lemma only for the finite-dimensional case. Further, also a converse inequality to (6.6) can be shown [36, 80].

6.4 Scalar Khintchine Inequalities

Khintchine inequalities provide estimates of the moments of Rademacher and related sums. In this section we present the scalar Khintchine inequalities, while in the next section we concentrate on the noncommutative (matrix-valued) Khintchine inequalities.

Theorem 6.8. *(Khintchine's inequality) Let* $\mathbf{b} \in \mathbb{C}^M$ *and* $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)$ *be a Rademacher sequence. Then, for all* $n \in \mathbb{N}$*,*

$$\mathbb{E}|\sum_{j=1}^{M} \epsilon_j b_j|^{2n} \leq \frac{(2n)!}{2^n n!} \|\mathbf{b}\|_2^{2n}. \tag{6.7}$$

Proof. First assume that the b_j are real-valued. Expanding the expectation on the left hand side of (6.7) with the multinomial theorem, which states that

$$\left(\sum_{j=1}^{M} x_j\right)^n = \sum_{\substack{k_1+\cdots+k_M=n \\ k_i \in \{0,1,\dots n\}}} \frac{n!}{k_1! \cdots k_M!} x_1^{k_1} \cdots x_M^{k_M}, \tag{6.8}$$

yields

$$\begin{aligned} E &:= \mathbb{E}|\sum_{j=1}^{M} \epsilon_j b_j|^{2n} \\ &= \sum_{\substack{j_1+\cdots+j_M=n \\ j_i \in \{0,1,\dots,n\}}} \frac{(2n)!}{(2j_1)! \cdots (2j_M)!} |b_1|^{2j_1} \cdots |b_M|^{2j_M} \mathbb{E}\epsilon_1^{2j_1} \cdots \mathbb{E}\epsilon_M^{2j_M} \\ &= \sum_{\substack{j_1+\cdots+j_M=n \\ j_i \in \{0,1,\dots,n\}}} \frac{(2n)!}{(2j_1)! \cdots (2j_M)!} |b_1|^{2j_1} \cdots |b_M|^{2j_M}. \end{aligned}$$

Hereby we used the independence of the ϵ_j and the fact that $\mathbb{E}\epsilon_j^k = 0$ if k is an odd integer. For integers satisfying $j_1 + \cdots + j_M = n$ it holds

$$2^n j_1! \cdots j_M! = 2^{j_1} j_1! \cdots 2^{j_M} j_M! \leq (2j_1)! \cdots (2j_M)!.$$

This implies

$$\begin{aligned} E &\leq \frac{(2n)!}{2^n n!} \sum_{\substack{j_1+\cdots+j_M=n \\ j_i \in \{0,1,\dots,n\}}} \frac{n!}{j_1! \cdots j_n!} |b_1|^{2j_1} \cdots |b_M|^{2j_M} \\ &= \frac{(2n)!}{2^n n!} \left(\sum_{j=1}^{M} |b_j|^2\right)^n = \frac{(2n)!}{2^n n!} \|\mathbf{b}\|_2^{2n}. \end{aligned}$$

The general complex case is derived by splitting into real and imaginary parts as follows

$$\left(\mathbb{E}|\sum_{j=1}^{M} \epsilon_j (\mathrm{Re}(b_j) + i\,\mathrm{Im}(b_j))|^{2n}\right)^{1/2n}$$

$$= \left(\mathbb{E} \left(|\sum_{j=1}^M \epsilon_j \operatorname{Re}(b_j)|^2 + |\sum_{j=1}^M \epsilon_j \operatorname{Im}(b_j)|^2 \right)^n \right)^{1/2n}$$

$$\leq \left(\left(\mathbb{E}|\sum_{j=1}^M \epsilon_j \operatorname{Re}(b_j)|^{2n} \right)^{1/n} + \left(\mathbb{E}|\sum_{j=1}^M \epsilon_j \operatorname{Im}(b_j)|^{2n} \right)^{1/n} \right)^{1/2}$$

$$\leq \left(\left(\frac{(2n)!}{2^n n!} \right)^{1/n} (\|\operatorname{Re}(\mathbf{b})\|_2^2 + \|\operatorname{Im}(\mathbf{b})\|_2^2) \right)^{1/2} = \left(\frac{(2n)!}{2^n n!} \right)^{1/2n} \|\mathbf{b}\|_2.$$

This concludes the proof. □

Except that we allowed the coefficient vector $\mathbf{b}$ to be complex valued, the above formulation and the proof is due to Khintchine [75]. Using the central limit theorem, one can show that the constants in (6.7) are optimal. Based on Theorem 6.8 we can also estimate the general absolute pth moment of a Rademacher sum.

Corollary 6.9. *(Khintchine's inequality) Let $\mathbf{b} \in \mathbb{C}^M$ and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)$ be a Rademacher sequence. Then, for all $p \geq 2$,*

$$\left(\mathbb{E}|\sum_{j=1}^M \epsilon_j b_j|^p \right)^{1/p} \leq 2^{3/(4p)} e^{-1/2} \sqrt{p} \|\mathbf{b}\|_2. \tag{6.9}$$

Proof. Without loss of generality we assume that $\|\mathbf{b}\|_2 = 1$. Stirling's formula for the factorial,

$$n! = \sqrt{2\pi n}\, n^n e^{-n} e^{\lambda_n}, \tag{6.10}$$

where $\frac{1}{12n+1} \leq \lambda_n \leq \frac{1}{12n}$, gives

$$\frac{(2n)!}{2^n n!} = \frac{\sqrt{2\pi 2n}(2n/e)^{2n} e^{\lambda_{2n}}}{2^n \sqrt{2\pi n}(n/e)^n e^{\lambda_n}} \leq \sqrt{2}\,(2/e)^n n^n. \tag{6.11}$$

An application of Hölder's inequality yields for $\theta \in [0,1]$ and an arbitrary random variable Z,

$$\mathbb{E}|Z|^{2n+2\theta} = \mathbb{E}[|Z|^{(1-\theta)2n} |Z|^{\theta(2n+2)}] \leq (\mathbb{E}|Z|^{2n})^{1-\theta} (\mathbb{E}|Z|^{2n+2})^{\theta}. \tag{6.12}$$

Combine the two estimates above and the Khintchine inequality (6.7) to get

$$\begin{aligned}
\mathbb{E}|\sum_{j=1}^M \epsilon_j b_j|^{2n+2\theta} &\leq (\mathbb{E}|\sum_{j=1}^M \epsilon_j b_j|^{2n})^{1-\theta} (\mathbb{E}|\sum_{j=1}^M \epsilon_j b_j|^{2n+2})^{\theta} \\
&\leq (\sqrt{2}(2/e)^n n^n)^{1-\theta} (\sqrt{2}(2/e)^{n+1}(n+1)^{n+1})^{\theta} \\
&= \sqrt{2}(2/e)^{n+\theta} n^{n(1-\theta)} (n+1)^{\theta(n+1)}
\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{2}(2/e)^{n+\theta}(n^{1-\theta}(n+1)^{\theta})^{n+\theta}\left(\frac{n+1}{n}\right)^{\theta(1-\theta)} \\
&\leq \sqrt{2}(2/e)^{n+\theta}(n+\theta)^{n+\theta}\left(\frac{n+1}{n}\right)^{\theta(1-\theta)} \\
&\leq 2^{3/4}(2/e)^{n+\theta}(n+\theta)^{n+\theta}. \qquad (6.13)
\end{aligned}
$$

In the second line from below the inequality of the geometric and arithmetic mean was applied. The last step used that $(n+1)/n \leq 2$ and $\theta(1-\theta) \leq 1/4$. Replacing $n+\theta$ by $p/2$ completes the proof. □

The optimal constants $C_p = \left(2^{\frac{p-1}{2}}\frac{\Gamma(p/2)}{\Gamma(3/2)}\right)^{1/p}$, $p \geq 2$, instead of $2^{3/(4p)}e^{-1/2}\sqrt{p}$ for Khintchine's inequality (6.9) are actually slightly better than the ones computed above, but deriving these requires much more effort [67, 89].

Combining Corollary 6.9 with Proposition 6.5 yields the following special case of Hoeffding's inequality [70], also known as Chernoff's bound [26].

Corollary 6.10. *(Hoeffding's inequality for Rademacher sums) Let* $\mathbf{b} = (b_1, \ldots, b_M) \in \mathbb{C}^M$ *and* $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_M)$ *be a Rademacher sequence. Then, for* $u \geq \sqrt{2}$*,*

$$
\mathbb{P}\left(|\sum_{j=1}^{M}\epsilon_j b_j| \geq \|\mathbf{b}\|_2 u\right) \leq 2^{3/4}\exp(-u^2/2). \qquad (6.14)
$$

For completeness we also give the standard version and proof of Hoeffding's inequality for (real) Rademacher sums.

Proposition 6.11. *(Hoeffding's inequality for Rademacher sums) Let* $\mathbf{b} = (b_1, \ldots, b_M) \in \mathbb{R}^M$ *and* $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_M)$ *be a Rademacher sequence. Then, for* $u > 0$*,*

$$
\mathbb{P}\left(\sum_{j=1}^{M}\epsilon_j b_j \geq \|\mathbf{b}\|_2 u\right) \leq \exp(-u^2/2) \qquad (6.15)
$$

and consequently,

$$
\mathbb{P}\left(|\sum_{j=1}^{M}\epsilon_j b_j| \geq \|\mathbf{b}\|_2 u\right) \leq 2\exp(-u^2/2). \qquad (6.16)
$$

Proof. Without loss of generality we may assume $\|\mathbf{b}\|_2 = 1$. By Markov's inequality

(Theorem 6.2) and independence we have, for $\lambda > 0$,

$$\mathbb{P}\left(\sum_{j=1}^M \epsilon_j b_j \geq u\right) = \mathbb{P}\left(\exp(\lambda \sum_{j=1}^M \epsilon_j b_j) \geq e^{\lambda u}\right) \leq e^{-\lambda u}\mathbb{E}[\exp(\lambda \sum_{j=1}^M \epsilon_j b_j)]$$
$$= e^{-\lambda u}\prod_{j=1}^M \mathbb{E}[\exp(\epsilon_j \lambda b_j)].$$

Note that, for $s \in \mathbb{R}$,

$$\mathbb{E}[\exp(\epsilon_j s)] = \frac{1}{2}(e^{-s} + e^s) = \frac{1}{2}\left(\sum_{k=0}^\infty \frac{(-s)^k}{k!} + \sum_{k=0}^\infty \frac{s^k}{k!}\right) = \sum_{k=0}^\infty \frac{s^{2k}}{(2k)!}$$
$$\leq \sum_{k=0}^\infty \frac{s^{2k}}{2^k k!} = e^{s^2/2}.$$

This yields

$$\mathbb{P}\left(\sum_{j=1}^M \epsilon_j b_j \geq u\right) \leq e^{-\lambda u}\prod_{j=1}^M e^{\lambda^2 b_j^2/2} = e^{-\lambda u + \lambda^2 \|\mathbf{b}\|_2^2/2}.$$

Choosing $\lambda = u$ and recalling that $\|\mathbf{b}\|_2 = 1$ yields (6.15). Finally,

$$\mathbb{P}\left(|\sum_{j=1}^M \epsilon_j b_j| \geq \|\mathbf{b}\|_2 u\right) = \mathbb{P}\left(\sum_{j=1}^M \epsilon_j b_j \geq \|\mathbf{b}\|_2 u\right) + \mathbb{P}\left(\sum_{j=1}^M \epsilon_j b_j \leq -\|\mathbf{b}\|_2 u\right)$$
$$\leq 2e^{-u^2/2},$$

since $-\epsilon_j$ has the same distribution as ϵ_j. □

As mentioned earlier, a complex random variable which is uniformly distributed on the torus $\mathbb{T} = \{z \in \mathbb{C}, |z| = 1\}$ is called a Steinhaus variable. A sequence $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)$ of independent Steinhaus variables is called a Steinhaus sequence. There is also a version of Khintchine's inequality for Steinhaus sequences.

Theorem 6.12. *(Khintchine's inequality for Steinhaus sequences) Let* $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_M)$ *be a Steinhaus sequence and* $\mathbf{b} = (b_1, \ldots, b_M) \in \mathbb{C}^M$. *Then*

$$\mathbb{E}|\sum_{j=1}^M \epsilon_j b_j|^{2n} \leq n!\|\mathbf{b}\|_2^{2n} \quad \text{for all } n \in \mathbb{N}.$$

Proof. Expand the moment of the Steinhaus sum using the multinomial theorem (6.8),

$$\begin{aligned}
\mathbb{E}|\sum_{j=1}^{M}\epsilon_j b_j|^{2n} &= \mathbb{E}\left(\sum_{j=1}^{M}\epsilon_j b_j\right)^n\left(\sum_{k=1}^{M}\overline{\epsilon_k b_k}\right)^n \\
&= \mathbb{E}\left(\sum_{\substack{j_1+\cdots+j_M=n\\ j_\ell\geq 0}}\frac{n!}{j_1!\cdots j_M!}b_1^{j_1}\cdots b_M^{j_M}\epsilon_1^{j_1}\cdots\epsilon_M^{j_M}\right) \\
&\quad\times\left(\sum_{\substack{k_1+\cdots k_M=n\\ k_\ell\geq 0}}\frac{n!}{k_1!\cdots k_M!}\overline{b_1^{k_1}\cdots b_M^{k_M}\epsilon_1^{k_1}\cdots\epsilon_M^{k_M}}\right) \\
&= \sum_{\substack{j_1+\cdots+j_M=n\\ k_1+\cdots k_M=n\\ j_\ell,k_\ell\geq 0}}\frac{n!}{j_1!\cdots j_M!}\frac{n!}{k_1!\cdots k_M!}b_1^{j_1}\overline{b_1^{k_1}}\cdots b_M^{j_M}\overline{b_M^{k_M}}\,\mathbb{E}[\epsilon_1^{j_1}\overline{\epsilon_1^{k_1}}\cdots\epsilon_M^{j_M}\overline{\epsilon_M^{k_M}}].
\end{aligned}$$

Since the ϵ_j are independent and uniformly distributed on the torus it holds

$$\mathbb{E}[\epsilon_1^{j_1}\overline{\epsilon_1^{k_1}}\cdots\epsilon_M^{j_M}\overline{\epsilon_M^{k_M}}] = \mathbb{E}[\epsilon_1^{j_1-k_1}]\cdots\mathbb{E}[\epsilon_M^{j_M-k_M}] = \delta_{j_1,k_1}\cdots\delta_{j_M,k_M}.$$

This yields

$$\begin{aligned}
\mathbb{E}|\sum_{j=1}^{M}\epsilon_j b_j|^{2n} &= \sum_{\substack{k_1+\cdots k_M=n\\ k_\ell\geq 0}}\left(\frac{n!}{k_1!\cdots k_M!}\right)^2|b_1|^{2k_1}\cdots|b_M|^{2k_M} \\
&\leq n!\sum_{\substack{k_1+\cdots k_M=n\\ k_\ell\geq 0}}\frac{n!}{k_1!\cdots k_M!}|b_1|^{2k_1}\cdots|b_M|^{2k_M} \\
&= n!\left(\sum_{j=1}^{M}|b_j|^2\right)^{2n},
\end{aligned}$$

where the multinomial theorem (6.8) was applied once more in the last step. □

The above moment estimate leads to a Hoeffding type inequality for Steinhaus sums.

Corollary 6.13. *(Hoeffding's inequality for Steinhaus sequences) Let* $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_M)$ *be a Steinhaus sequence,* $\mathbf{b} = (b_1, \ldots, b_M) \in \mathbb{C}^M$ *and* $0 < \lambda < 1$. *Then*

$$\mathbb{P}(|\sum_{j=1}^{M}\epsilon_j b_j| \geq u\|\mathbf{b}\|_2) \leq \frac{1}{1-\lambda}e^{-\lambda u^2} \quad \text{for all } u \geq 0. \tag{6.17}$$

In particular, using the optimal choice $\lambda = 1 - u^{-2}$,

$$\mathbb{P}(|\sum_{j=1}^{M} \epsilon_j b_j| \geq u\|\mathbf{b}\|_2) \leq \exp(-u^2 + \log(u^2) + 1) \quad \text{for all } u \geq 1. \tag{6.18}$$

Note that the argument of the exponential in (6.18) is always negative for $u > 1$.

Proof. Without loss of generality assume $\|\mathbf{b}\|_2 = 1$. Markov's inequality gives

$$\begin{aligned}
&\mathbb{P}(|\sum_{j=1}^{M} \epsilon_j b_j| \geq u) = \mathbb{P}(\exp(\lambda|\sum_{j=1}^{M} \epsilon_j b_j|^2) \geq \exp(\lambda u^2)) \\
&\leq \mathbb{E}[\exp(\lambda|\sum_{j=1}^{M} \epsilon_j b_j|^2)] \exp(-\lambda u^2) = \exp(-\lambda u^2) \sum_{n=0}^{\infty} \frac{\lambda^n \mathbb{E}|\sum_{j=1}^{M} \epsilon_j b_j|^{2n}}{n!} \\
&\leq \exp(-\lambda u^2) \sum_{n=0}^{\infty} \lambda^n = \frac{1}{1-\lambda} e^{-\lambda u^2}.
\end{aligned}$$

In the second line the Taylor expansion of the exponential function was used together with Fubini's theorem in order to interchange the expectation and the series. In the third line Theorem 6.12 was applied. □

For more information and extensions of scalar Khintchine inequalities we refer the interested reader to [94, 93].

6.5 Noncommutative Khintchine Inequalities

The scalar Khintchine inequalities above can be generalized to the case where the coefficients are matrices. Combined with symmetrization the resulting noncommutative Khintchine inequalities are a very powerful tool in the theory of random matrices. Schatten class norms have to be introduced to formulate them.

For a matrix A we let $\sigma(A) = (\sigma_1(A), \ldots, \sigma_n(A))$ be its sequence of singular values. Then the Schatten p-norm is defined as

$$\|A\|_{S_p} := \|\sigma(A)\|_p, \quad 1 \leq p \leq \infty. \tag{6.19}$$

It is actually nontrivial to show the triangle inequality for Schatten p-norms. We refer the interested reader to [8, 72, 120].

The hermitian matrix AA^* can be diagonalized using a unitary matrix U,

$$AA^* = U^* D^2 U$$

where $D = \operatorname{diag}(\sigma_1(A), \dots, \sigma_n(A))$ (possibly filled up with zeros). As the trace is cyclic, that is $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$, and since $UU^* = \mathrm{Id}$, we get for $n \in \mathbb{N}$

$$\begin{aligned} \|A\|_{S_{2n}}^{2n} &= \|\sigma(A)\|_{2n}^{2n} = \operatorname{Tr}(D^{2n}) = \operatorname{Tr}(D^{2n}UU^*) = \operatorname{Tr}(U^*D^{2n}U) \\ &= \operatorname{Tr}((U^*D^2U)^n) = \operatorname{Tr}((AA^*)^n). \end{aligned} \tag{6.20}$$

As a special case, the Frobenius norm is the Schatten 2-norm, $\|A\|_F = \|A\|_{S_2}$. The operator norm is also a Schatten norm,

$$\|A\|_{2\to 2} = \sigma_1(A) = \|\sigma(A)\|_\infty = \|A\|_{S_\infty}.$$

By the analogous property of the vector p-norm we have $\|A\|_{S_q} \le \|A\|_{S_p}$ for $q \ge p$. In particular, the following estimate will be very useful,

$$\|A\|_{2\to 2} \le \|A\|_{S_p} \quad \text{for all } 1 \le p \le \infty\ . \tag{6.21}$$

If A has rank r then it follows from the corresponding property of ℓ_p-norms that

$$\|A\|_{S_p} \le r^{1/p}\|A\|_{2\to 2}. \tag{6.22}$$

Let us now state the noncommutative Khintchine inequality for matrix-valued Rademacher sums, which was first formulated by F. Lust–Piquard [82] with unspecified constants. The optimal constants were provided by A. Buchholz in [16, 17], although it is not obvious at first sight that the results of his paper [16] allow to deduce our next theorem, see also [130]. The proof follows the ideas of Buchholz [16].

Theorem 6.14. *Let $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)$ be a Rademacher sequence, and let B_j, $j = 1, \dots, M$, be complex matrices of the same dimension. Choose $n \in \mathbb{N}$. Then*

$$\begin{aligned} &\mathbb{E}\|\sum_{j=1}^M \epsilon_j B_j\|_{S_{2n}}^{2n} \\ &\le \frac{(2n)!}{2^n n!} \max\left\{ \left\| \left(\sum_{j=1}^M B_j B_j^*\right)^{1/2} \right\|_{S_{2n}}^{2n}, \left\| \left(\sum_{j=1}^M B_j^* B_j\right)^{1/2} \right\|_{S_{2n}}^{2n} \right\}. \end{aligned} \tag{6.23}$$

Note that the matrices $B_jB_j^*$ and $B_j^*B_j$ in (6.23) are self-adjoint and positive, so that the square-roots in (6.23) are well-defined.

In order to prove the noncommutative Khintchine inequalities we need to introduce the notion of pairings.

Definition 6.15.

(a) A pairing is a partition of the set $[2n]$ into two-element subsets, called blocks. The set $\mathbb{P}_{2n}$ denotes the set of all pairings of $[2n]$.

(b) The canonical pairing $\mathbb{1} = \{D_1, \dots, D_n\}$ has blocks $D_j = \{2j-1, 2j\}$.

(c) Let $\pi = \{D_1, \dots, D_n\}$ be a pairing. Then its cyclic shift $T\pi$ is the pairing with blocks TD_ℓ, where $T\{j,k\} = \{j+1, k+1\}$ with addition understood modulo $2n$.

(d) The "symmetrized" pairing $\overleftarrow{\pi}$ contains all blocks $\{j,k\}$ of a pairing π satisfying $j,k \le n$ and in addition the "reflected" blocks $\{2n+1-j, 2n+1-k\}$. The blocks of π with both elements being larger than n are omitted in $\overleftarrow{\pi}$ and the blocks $\{j,k\}$ with $j \le n$ and $k > n$ are replaced by the "symmetric" block $\{j, 2n+1-j\}$.

(e) Similarly, the pairing $\overrightarrow{\pi}$ contains all blocks $\{j,k\}$ of the pairing π satisfying $j,k > n$ and in addition the "reflected" blocks $\{2n+1-j, 2n+1-k\}$. The blocks of π with both elements smaller than $n+1$ are omitted in $\overrightarrow{\pi}$ and the blocks $\{j,k\}$ with $j \le n$ and $k > n$ are replaced by the "symmetric" block $\{2n+1-k, k\}$.

Let $\mathcal{B} = (B_1, \dots, B_M)$ be a sequence of matrices of the same dimension and $\pi = \{D_1, \dots, D_n\} \in \mathbb{P}_{2n}$. We define the mapping $\alpha = \alpha_\pi : [2n] \to [n]$ such that $\alpha(j) = \ell$ iff $j \in D_\ell$. Using this notation we introduce

$$\pi(\mathcal{B}) = \sum_{k_1,\dots,k_n=1}^{M} B_{k_{\alpha(1)}} B_{k_{\alpha(2)}}^* B_{k_{\alpha(3)}} B_{k_{\alpha(4)}}^* \cdots B_{k_{\alpha(2n-1)}} B_{k_{\alpha(2n)}}^*. \tag{6.24}$$

Note that $\pi(\mathcal{B})$ is independent of the chosen numbering of $D_1, \dots, D_n$. The following lemma will be the key to the proof of the noncommutative Khintchine inequalities.

Lemma 6.16. *Let $\pi \in \mathbb{P}_{2n}$ and $\mathcal{B} = (B_1, \dots, B_M)$ a sequence of complex matrices of the same dimension. Then there is $\gamma \ge 1/(4n)$ and non-negative numbers $p_\rho = p_\rho(\pi), \rho \in \mathbb{P}_{2n}$, satisfying $\gamma + \sum_{\rho \in \mathbb{P}_{2n}} p_\rho = 1$, such that*

$$\begin{aligned} &|\operatorname{Tr} \pi(\mathcal{B})| \\ &\le \max\left\{ \left\| \left(\sum_{k=1}^M B_k B_k^* \right)^{1/2} \right\|_{S_{2n}}^{2n}, \left\| \left(\sum_{k=1}^M B_k^* B_k \right)^{1/2} \right\|_{S_{2n}}^{2n} \right\}^{\gamma} \prod_{\rho \in \mathbb{P}_{2n}} |\operatorname{Tr} \rho(\mathcal{B})|^{p_\rho}. \end{aligned} \tag{6.25}$$

Proof. First observe that

$$\mathbb{1}(\mathcal{B}) = \sum_{k_1,\dots,k_n=1}^{N} \prod_{j=1}^{n} B_{k_j} B_{k_j}^* = \left(\sum_{k=1}^M B_k B_k^* \right)^n.$$

Since the matrix inside the bracket is self-adjoint and positive semi-definite we can take its square root and (6.20) yields

$$\operatorname{Tr} \mathbb{1}(\mathcal{B}) = \left\| \left(\sum_{k=1}^M B_k B_k^* \right)^{1/2} \right\|_{S_{2n}}^{2n}. \tag{6.26}$$

Since the trace is cyclic we similarly obtain

$$\operatorname{Tr} T\mathbb{1}(\mathcal{B}) = \left\| \left(\sum_{k=1}^{M} B_k^* B_k \right)^{1/2} \right\|_{S_{2n}}^{2n}. \tag{6.27}$$

The idea of the proof is to successively provide estimates of $|\operatorname{Tr}\pi(\mathcal{B})|$ in terms of traces of operators $\rho(\mathcal{B})$ which become more and more 'similar' to $\mathbb{1}(\mathcal{B})$ or $T\mathbb{1}(\mathcal{B})$.

Let $t \in \{0, 1, \dots, n\}$ be the maximal number such that, for some p, $\{p, p+1\}, \{p+2, p+3\}, \dots, \{p+2t-2, p+2t-1\}$ are blocks of the partition π. If $t = n$ then $\pi = \mathbb{1}$ or $\pi = T\mathbb{1}$ and we are done. We postpone the case $t = 0$ to later and assume $t \in [n-1]$.

By cyclicity of the trace, it holds $\operatorname{Tr}\pi(\mathcal{B}) = \operatorname{Tr}(T^{n-p-2t+1}\pi)(\mathcal{B})$ if $n-p$ is odd and $\operatorname{Tr}\pi(\mathcal{B}) = \operatorname{Tr}(T^{n-p-2t+1}\pi)(\mathcal{B}^*)$ if $n-p$ is even, where $\mathcal{B}^* = (B_1^*, \dots, B_M^*)$. Note that the blocks $\{n-2t+1, n-2t+2\}, \{n-2t+3, n-2t+4\}, \dots, \{n-1, n\}$ (with addition modulo $2n$) are part of the partition $T^{n-p-2t+1}\pi$. Assume n even and p odd for the moment. Denote the blocks of $T^{n-p-2t+1}\pi$ by $D_1, \dots, D_n$ and let $\alpha = \alpha_\pi : [2n] \to [n]$ be the mapping defined by $\alpha(j) = \ell$ iff $j \in D_\ell$. Divide $[n]$ into three sets L, R, U. The subset L (resp. R) contains the indices ℓ, for which both elements of D_ℓ are in $\{1, \dots, n\}$ (resp. $\{n+1, \dots, 2n\}$), while U contains the remaining indices for which the blocks have elements in both $\{1, \dots, n\}$ and $\{n+1, \dots, 2n\}$.

The Cauchy Schwarz inequality for the trace (2.6) and for the usual Euclidean inner product yields

$$\begin{aligned}
&|\operatorname{Tr}\pi(\mathcal{B})| = |\operatorname{Tr}(T^{n-p-2t+1}\pi)(\mathcal{B})| \\
&= \left| \sum_{k_i \in [M], i \in U} \operatorname{Tr}\left(\left(\sum_{k_i \in [M], i \in L} B_{k_{\alpha(1)}} \cdots B_{k_{\alpha(n)}}^* \right) \left(\sum_{k_i \in [M], i \in R} B_{k_{\alpha(n+1)}} \cdots B_{k_{\alpha(2n)}}^* \right) \right) \right| \\
&\le \sum_{k_i, i \in U} \sqrt{\operatorname{Tr}\left(\left(\sum_{k_i, i \in L} B_{k_{\alpha(1)}} \cdots B_{k_{\alpha(n)}}^* \right) \left(\sum_{k_i, i \in L} B_{k_{\alpha(n)}} \cdots B_{k_{\alpha(1)}}^* \right) \right)} \\
&\qquad \times \sqrt{\operatorname{Tr}\left(\left(\sum_{k_i, i \in R} B_{k_{\alpha(n+1)}} \cdots B_{k_{\alpha(2n)}}^* \right) \left(\sum_{k_i, i \in R} B_{k_{\alpha(2n)}} \cdots B_{k_{\alpha(n+1)}}^* \right) \right)} \\
&\le \sqrt{\sum_{k_i, i \in U} \operatorname{Tr}\left(\left(\sum_{k_i, i \in L} B_{k_{\alpha(1)}} \cdots B_{k_{\alpha(n)}}^* \right) \left(\sum_{k_i, i \in L} B_{k_{\alpha(n)}} \cdots B_{k_{\alpha(1)}}^* \right) \right)} \\
&\qquad \times \sqrt{\sum_{k_i, i \in U} \operatorname{Tr}\left(\left(\sum_{k_i, i \in R} B_{k_{\alpha(n+1)}} \cdots B_{k_{\alpha(2n)}}^* \right) \left(\sum_{k_i, i \in R} B_{k_{\alpha(2n)}} \cdots B_{k_{\alpha(n+1)}}^* \right) \right)}
\end{aligned}$$

$$= |\operatorname{Tr}(\overleftarrow{T^{n-p-2t+1}\pi})(\mathcal{B})|^{1/2} |\operatorname{Tr}\rho(\mathcal{B})|^{1/2} \tag{6.28}$$

with $\rho = \overrightarrow{T^{n-p-2t+1}\pi} \in \mathbb{P}_{2n}$. If $t \geq n/2$ then $\overleftarrow{T^{n-p-2t+1}\pi}$ equals $\mathbb{1}$ or $T\mathbb{1}$ and we are done. If $t < n/2$ then $\overleftarrow{T^{n-p-2t+1}\pi}$ contains the blocks $\{n-2t+1, n-2t+2\}, \ldots, \{n-1, n\}, \{n+1, n+2\}, \ldots, \{n+2t-1, n+2t\}$. Apply the same estimates with $t' = 2t$ as above to $T^{-2t}(\overleftarrow{T^{n-p-2t+1}\pi})$ to obtain

$$|\operatorname{Tr}\pi(\mathcal{B})| \leq |\operatorname{Tr}(\overleftarrow{T^{-2t}(\overleftarrow{T^{n-p-2t+1}\pi})}(\mathcal{B})|^{1/4} |\operatorname{Tr}\tilde{\rho}(\mathcal{B})|^{1/4} |\operatorname{Tr}\rho(\mathcal{B})|^{1/2}$$

for suitable $\tilde{\rho}, \rho \in \mathbb{P}_{2n}$. Similarly, as above if $t \geq n/4$ then $\overleftarrow{T^{-2t}(\overleftarrow{T^{n-p-2t}\pi})}$ equals $\mathbb{1}$ or $T\mathbb{1}$ and we are done. If $t < n/4$ then we continue in this way, and after at most $\lceil \log_2(n) \rceil$ estimation steps of the form (6.28) inequality (6.25) is obtained with $\gamma \geq 1/2^{\lceil \log_2(n) \rceil} \geq 1/(2n)$.

If initially $t = 0$, then we apply the above method to $T^q\pi$ where q was chosen such that $\{n, p\}$ for some $p > n$ is a block of $T^q\pi$. Using the same estimates as in (6.28) yields $|\operatorname{Tr}\pi(\mathcal{B})| \leq |\operatorname{Tr}\tilde{\pi}(\mathcal{B})|^{1/2} |\operatorname{Tr}\rho(\mathcal{B})|^{1/2}$ for some partition ρ, where $\tilde{\pi}$ contains the block $\{n, n+1\}$. Then invoke the above method to obtain (6.25) with $\gamma \geq 1/(4n)$.

If n is odd and p even, then an obvious modification of the chain of inequalities (6.28) applies. If $n - p$ is even then $\mathcal{B}^*$ instead of $\mathcal{B}$ appears after the first equality in (6.28). Noting that $\operatorname{Tr}\mathbb{1}(\mathcal{B}^*) = \operatorname{Tr}T\mathbb{1}(\mathcal{B})$ and $\operatorname{Tr}T\mathbb{1}(\mathcal{B}^*) = \operatorname{Tr}\mathbb{1}(\mathcal{B})$ by cyclicity concludes the proof. □

Corollary 6.17. *Under the same assumptions as in Lemma 6.16, for all $\pi \in \mathbb{P}_{2n}$,*

$$|\operatorname{Tr}\pi(\mathcal{B})| \leq \max\left\{ \left\| \left(\sum_{k=1}^{M} B_k B_k^* \right)^{1/2} \right\|_{S_{2n}}^{2n}, \left\| \left(\sum_{k=1}^{M} B_k^* B_k \right)^{1/2} \right\|_{S_{2n}}^{2n} \right\}. \tag{6.29}$$

Proof. Denote the right hand side of (6.29) by D. The constant γ in Lemma 6.16 may be chosen the same for all partitions $\pi \in \mathbb{P}_{2n}$, for instance $\gamma = \gamma_1 = 1/(4n)$. Indeed, if γ is initially larger, then by (6.26) and (6.27) simply move some weight from D^γ to $|\operatorname{Tr}\mathbb{1}(\mathcal{B})|^{p_{\mathbb{1}}(\pi)}$ or to $|\operatorname{Tr}T\mathbb{1}(\mathcal{B})|^{p_{T\mathbb{1}}(\pi)}$, whichever term is larger.

Apply Lemma 6.16 to itself to obtain

$$\begin{aligned}
|\operatorname{Tr}\pi(\mathcal{B})| &\leq D^{\gamma_1} \prod_{\kappa \in \mathbb{P}_{2n}} |\operatorname{Tr}\kappa(\mathcal{B})|^{p_\kappa(\pi)} \\
&\leq D^{\gamma_1} \prod_{\kappa \in \mathbb{P}_{2n}} D^{\gamma_1 p_\kappa(\pi)} \prod_{\rho \in \mathbb{P}_{2n}} |\operatorname{Tr}\rho(\mathcal{B})|^{p_\kappa(\pi) p_\rho(\kappa)} \\
&= D^{\gamma_1 + \gamma_1(1-\gamma_1)} \prod_{\rho \in \mathbb{P}_{2n}} |\operatorname{Tr}\rho(\mathcal{B})|^{\sum_{\kappa \in \mathbb{P}_{2n}} p_\rho(\kappa) p_\kappa(\pi)}.
\end{aligned}$$

This yields (6.25) with new constants

$$\gamma_2 = \gamma_1 + \gamma_1(1-\gamma_1), \quad p_\rho^{(2)}(\pi) = \sum_{\kappa\in\mathbb{P}_{2n}} p_\rho(\kappa)p_\kappa(\pi).$$

Since $\gamma_1 = 1/(4n)$, in particular, $0 < \gamma_1 < 1$, the new constant γ_2 is larger than γ_1. Iterating this process yields increasingly larger constants γ_ℓ defined recursively by

$$\gamma_{\ell+1} = \gamma_\ell + \gamma_\ell(1-\gamma_\ell).$$

Elementary calculus shows that $\lim_{\ell\to\infty}\gamma_\ell = 1$. Since the corresponding constants $p_\rho^{(\ell)}(\pi)$ satisfy $\gamma_\ell + \sum_{\rho\in\mathbb{P}_{2n}} p_\rho^{(\ell)}(\pi) = 1$ for all ℓ one concludes $\lim_{\ell\to\infty} p_\rho^{(\ell)}(\pi) = 0$ for all $\rho \in \mathbb{P}_{2n}$. This completes the proof. □

Now we are in the position to complete the proof of the noncommutative Khintchine inequalities.

Proof of Theorem 6.14. By (6.20)

$$\begin{aligned} E := \mathbb{E}\|\sum_{k=1}^M \epsilon_k B_k\|_{S_{2n}}^{2n} &= \mathbb{E}\,\mathrm{Tr}\left(\left(\sum_{k=1}^M \epsilon_k B_k \sum_{j=1}^M \epsilon_j B_j^*\right)^n\right) \\ &= \sum_{k_1,\ldots,k_{2n}=1}^M \mathbb{E}[\epsilon_{k_1}\cdots\epsilon_{k_{2n}}]\,\mathrm{Tr}(B_{k_1}B_{k_2}^*B_{k_3}\cdots B_{k_{2n}}^*). \end{aligned}$$

Observe that $\mathbb{E}[\epsilon_{k_1}\cdots\epsilon_{k_{2n}}] = 1$ if and only if each $j \in [2n]$ can be paired with an $\ell \in [2n]$ such that $k_j = k_\ell$ and $\mathbb{E}[\epsilon_{k_1}\cdots\epsilon_{k_{2n}}] = 0$ otherwise. Therefore, denoting $\mathcal{B} = (B_1,\ldots,B_M)$, Corollary 6.17 yields (recall also the definition in (6.24))

$$\begin{aligned} E &= \sum_{\pi\in\mathbb{P}_{2n}} \mathrm{Tr}\,\pi(\mathcal{B}) \\ &\le |\mathbb{P}_{2n}| \max\left\{ \left\| \left(\sum_{k=1}^M B_k B_k^*\right)^{1/2} \right\|_{S_{2n}}^{2n}, \left\| \left(\sum_{k=1}^M B_k^* B_k\right)^{1/2} \right\|_{S_{2n}}^{2n} \right\}. \end{aligned}$$

Elementary considerations show that the number $|\mathbb{P}_{2n}|$ of pairings of a set with $2n$ elements equals $\frac{(2n)!}{2^n n!}$. □

The noncommutative Khintchine inequalities may be extended to general $p \ge 2$, similarly to Corollary 6.9 in the scalar case, see also [130]. For our purposes the present version will be sufficient.

6.6 Rudelson's Lemma

Rudelson's lemma [113] is a very useful estimate for the operator norm of a Rademacher sum of rank one matrices. The statement below is slightly different from the formulation in [113], but allows to draw the same conclusions, and makes constants explicit. Its proof is a nice application of the noncommutative Khintchine inequality.

Lemma 6.18. *Let $A \in \mathbb{C}^{m\times M}$ of rank r with columns $\mathbf{a}_1, \dots, \mathbf{a}_M$. Let $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)$ be a Rademacher sequence. Then, for $2 \le p < \infty$,*

$$\left(\mathbb{E}\|\sum_{j=1}^{M} \epsilon_j \mathbf{a}_j \mathbf{a}_j^*\|_{2\to 2}^p\right)^{1/p} \le 2^{3/(4p)} r^{1/p} \sqrt{p} e^{-1/2} \|A\|_{2\to 2} \max_{j=1,\dots,M} \|\mathbf{a}_j\|_2. \quad (6.30)$$

Proof. Write $p = 2n + 2\theta$ with $n \in N$ and $\theta \in [0, 1)$. Denote $C_n = \left(\frac{(2n)!}{2^n n!}\right)^{1/(2n)}$. Note that $(\mathbf{a}_j\mathbf{a}_j^*)^*(\mathbf{a}_j\mathbf{a}_j^*) = (\mathbf{a}_j\mathbf{a}_j^*)(\mathbf{a}_j\mathbf{a}_j^*)^* = \|\mathbf{a}_j\|_2^2 \mathbf{a}_j\mathbf{a}_j^*$. Therefore, the noncommutative Khintchine inequality (6.23) yields

$$\begin{aligned} E := \left(\mathbb{E}\|\sum_j \epsilon_j \mathbf{a}_j\mathbf{a}_j^*\|_{2\to 2}^{2n}\right)^{1/(2n)} &\le \left(\mathbb{E}\|\sum_j \epsilon_j \mathbf{a}_j\mathbf{a}_j^*\|_{S_{2n}}^{2n}\right)^{1/(2n)} \\ &\le C_n \left\| \left(\sum_j \|\mathbf{a}_j\|_2^2 \mathbf{a}_j\mathbf{a}_j^*\right)^{1/2} \right\|_{S_{2n}} \end{aligned}$$

The operator $\sum_j \|\mathbf{a}_j\|_2^2 \mathbf{a}_j\mathbf{a}_j^*$ has rank at most r. The estimate (6.22) of the Schatten norm by the operator norm together with (2.5) gives therefore

$$\begin{aligned} E &\le C_n r^{1/(2n)} \left\| \left(\sum_j \|\mathbf{a}_j\|_2^2 \mathbf{a}_j\mathbf{a}_j^*\right)^{1/2} \right\|_{2\to 2} \\ &\le C_n r^{1/(2n)} \|\sum_j \mathbf{a}_j\mathbf{a}_j^*\|_{2\to 2}^{1/2} \max_{k=1,\dots,M} \|\mathbf{a}_k\|_2. \end{aligned}$$

Observing that $\|\sum_j \mathbf{a}_j\mathbf{a}_j^*\|_{2\to 2}^{1/2} = \|AA^*\|_{2\to 2}^{1/2} = \|A\|_{2\to 2}$ yields

$$E \le C_n r^{1/(2n)} \|A\|_{2\to 2} \max_{j=1,\dots,M} \|\mathbf{a}_j\|_2.$$

With the estimate (6.21) of the operator norm by the Schatten norm together with

(6.12) we obtain

$$\begin{aligned}\mathbb{E}\|\sum_{j=1}^{M}\epsilon_j\mathbf{a}_j\mathbf{a}_j^*\|_{2\to 2}^{2n+2\theta} &\le (\mathbb{E}\|\sum_{j}\epsilon_j\mathbf{a}_j\mathbf{a}_j^*\|_{S_{2n}}^{2n})^{1-\theta}(\mathbb{E}\|\sum_{j}\epsilon_j\mathbf{a}_j\mathbf{a}_j^*\|_{S_{2n+2}}^{2n+2})^{\theta}\\ &\le \left(\frac{(2n)!}{2^n n!}\right)^{1-\theta}\left(\frac{(2n+2)!}{2^{n+1}(n+1)!}\right)^{\theta} r\left(\|A\|_{2\to 2}\max_{j=1,\dots,M}\|\mathbf{a}_j\|_2\right)^{2n+2\theta}\\ &\le 2^{3/4}(2/e)^{n+\theta}(n+\theta)^{n+\theta} r\left(\|A\|_{2\to 2}\max_{j=1,\dots,M}\|\mathbf{a}_j\|_2\right)^{2n+2\theta}.\end{aligned}$$

Hereby, we applied (6.11) and the same chain of inequalities as in (6.13). Substituting $p/2 = n+\theta$ completes the proof. □

Proposition 6.5 leads to the following statement.

Corollary 6.19. *Let $A \in \mathbb{C}^{m\times M}$ of rank r with columns $\mathbf{a}_1,\dots,\mathbf{a}_M$. Let $\boldsymbol{\epsilon} \in \mathbb{R}^M$ be a Rademacher sequence. Then for all $u \ge \sqrt{2}$*

$$\mathbb{P}\left(\|\sum_{j=1}^{M}\epsilon_j\mathbf{a}_j\mathbf{a}_j^*\|_{2\to 2} \ge u\|A\|_{2\to 2}\max_{j=1,\dots,M}\|\mathbf{a}_j\|_2\right) \le 2^{3/4} r e^{-u^2/2}. \tag{6.31}$$

The formulation of Rudelson's lemma which is most commonly used follows then from an application of Lemma 6.6 (where the "maximum" is taken only over one random variable) after estimating $2^{3/4} < 2$.

Corollary 6.20. *Let $A \in \mathbb{C}^{m\times M}$ of rank r with columns $\mathbf{a}_1,\dots,\mathbf{a}_M$. Let $\boldsymbol{\epsilon} \in \mathbb{R}^M$ be a Rademacher sequence. Then*

$$\mathbb{E}\|\sum_{j=1}^{M}\epsilon_j\mathbf{a}_j\mathbf{a}_j^*\|_{2\to 2} \le C\sqrt{\ln(8r)}\|A\|_{2\to 2}\max_{j=1,\dots,M}\|\mathbf{a}_j\|_2$$

with $C \le \sqrt{2} + \frac{1}{4\sqrt{2\ln(8)}} \approx 1.499 < 1.5$.

6.7 Decoupling

Decoupling is a technique that reduces stochastic dependencies in certain sums of random variables, called chaos variables. A typical example is a sum of the form

$$\sum_{j\neq\ell}\epsilon_j\epsilon_\ell\mathbf{x}_{j\ell}$$

where $\mathbf{x}_{j\ell}$ are some vectors and $\boldsymbol{\epsilon} = (\epsilon_j)$ is a Rademacher series. Such a sum is called Rademacher chaos of order 2. The following statement, taken from [13], provides a way of "decoupling" the sum. Many more results concerning decoupling can be found in the monograph [36].

Lemma 6.21. *Let $\boldsymbol{\xi} = (\xi_1, \dots, \xi_M)$ be a sequence of independent random variables with $\mathbb{E}\xi_j = 0$ for all $j = 1, \dots, M$. Let $B_{j,k}$, $j, k = 1, \dots, M$, be a double sequence of elements in a vector space with norm $\|\cdot\|$, where $B_{j,j} = 0$ for all $j = 1, \dots, M$. Then for $1 \leq p < \infty$*

$$\mathbb{E}\left\|\sum_{j,k=1}^{M} \xi_j \xi_k B_{j,k}\right\|^p \leq 4^p \mathbb{E}\left\|\sum_{j,k=1}^{M} \xi_j \xi_k' B_{j,k}\right\|^p, \tag{6.32}$$

where $\boldsymbol{\xi}'$ denotes an independent copy of $\boldsymbol{\xi}$.

Proof. Introduce a sequence $\delta = (\delta_j)_{j=1}^M$, of independent random variables δ_j taking only the values 0 and 1 with probability $1/2$. Then for $j \neq k$

$$\mathbb{E}\delta_j(1 - \delta_k) = 1/4.$$

Since $B_{j,j} = 0$ this gives

$$\begin{aligned} E := \mathbb{E}\left\|\sum_{j,k=1}^{M} \xi_j \xi_k B_{j,k}\right\|^p &= 4^p\, \mathbb{E}_\xi \left\|\sum_{j,k=1}^{M} \mathbb{E}_\delta[\delta_j(1-\delta_k)]\xi_j \xi_k B_{j,k}\right\|^p \\ &\leq 4^p\, \mathbb{E}\left\|\sum_{j,k=1}^{M} \delta_j(1-\delta_k)\xi_j \xi_k B_{j,k}\right\|^p, \end{aligned}$$

where Jensen's inequality was applied in the last step. Now let

$$\sigma(\delta) := \{j = 1, \dots, M : \delta_j = 1\}.$$

Then, by Fubini's theorem,

$$E \leq 4^p\, \mathbb{E}_\delta \mathbb{E}_\xi \left\|\sum_{j \in \sigma(\delta)} \sum_{k \notin \sigma(\delta)} \xi_j \xi_k B_{j,k}\right\|^p.$$

For a fixed δ the sequences $(\xi_j)_{j \in \sigma(\delta)}$ and $(\xi_k)_{k \notin \sigma(\delta)}$ are independent, and hence,

$$E \leq 4^p\, \mathbb{E}_\delta \mathbb{E}_\xi \mathbb{E}_{\xi'} \left\|\sum_{j \in \sigma(\delta)} \sum_{k \notin \sigma(\delta)} \xi_j \xi_k' B_{j,k}\right\|^p.$$

This implies the existence of a δ_0, and hence a $\sigma = \sigma(\delta_0)$ such that

$$E \leq 4^p\, \mathbb{E}_\xi \mathbb{E}_{\xi'} \left\|\sum_{j \in \sigma} \sum_{k \notin \sigma} \xi_j \xi_k' B_{j,k}\right\|^p.$$

Since $\mathbb{E}\xi_j = \mathbb{E}\xi'_j = 0$, an application of Jensen's inequality yields

$$E \le 4^p \, \mathbb{E} \left\| \sum_{j\in\sigma} \left(\sum_{k\notin\sigma} \xi_j \xi'_k B_{j,k} + \sum_{k\in\sigma} \xi_j \mathbb{E}[\xi'_k] B_{j,k} \right) + \sum_{j\notin\sigma} \mathbb{E}[\xi_j] \sum_{k=1}^{M} \xi'_k B_{j,k} \right\|^p$$
$$\le 4^p \, \mathbb{E} \left\| \sum_{j=1}^{M} \sum_{k=1}^{M} \xi_j \xi'_k B_{j,k} \right\|^p ,$$

and the proof is completed. □

We note that the mean-zero assumption $\mathbb{E}\xi_j = 0$ may be removed by introducing a larger constant 8 instead of 4, see Theorem 3.1.1 in [36] and its proof. The sum $\sum_{j,k} \xi_j \xi'_k B_{j,k}$ on the right hand side of (6.32) is called a decoupled chaos.

6.8 Noncommutative Khintchine Inequalities for Decoupled Rademacher Chaos

The previous section showed the usefulness of studying decoupled chaoses. Next we state the noncommutative Khintchine inequality for decoupled Rademacher chaos [105], see also [100, p. 111] for a slightly more general inequality (without explicit constants). A scalar version can be found, for instance, in [86].

Theorem 6.22. *Let $B_{j,k} \in \mathbb{C}^{r\times t}$, $j,k = 1,\dots,M$, be complex matrices of the same dimension. Let $\boldsymbol{\epsilon}, \boldsymbol{\epsilon}'$ be independent Rademacher sequences. Then, for $n \in \mathbb{N}$,*

$$\left[\mathbb{E} \left\| \sum_{j,k=1}^{M} \epsilon_j \epsilon'_k B_{j,k} \right\|_{S_{2n}}^{2n} \right]^{1/2n} \le 2^{1/(2n)} \left(\frac{(2n)!}{2^n n!} \right)^{1/n} \tag{6.33}$$
$$\times \max \left\{ \left\| \left(\sum_{j,k=1}^{M} B_{j,k} B_{j,k}^* \right)^{1/2} \right\|_{S_{2n}}, \left\| \left(\sum_{j,k=1}^{M} B_{j,k}^* B_{j,k} \right)^{1/2} \right\|_{S_{2n}}, \|F\|_{S_{2n}}, \|\widetilde{F}\|_{S_{2n}} \right\},$$

where F, $\widetilde{F}$ are the block matrices $F = (B_{j,k})_{j,k=1}^M$ and $\widetilde{F} = (B_{j,k}^*)_{j,k=1}^M$.

We note that the factor $2^{1/(2n)}$ may be removed with a more technical proof that uses the same strategy as the proof of the (ordinary) noncommutative Khintchine inequality (6.25) above. Our proof below rather proceeds by applying (6.25) twice. Taking scalars instead of matrices $B_{j,k}$ results in a scalar Khintchine inequality for decoupled Rademacher chaos. In the scalar case the first two terms in the maximum in (6.33) coincide and the third one is always dominated by the first.

Proof of Theorem 6.22. Denote $C_n = \frac{(2n)!}{2^n n!}$. Fubini's theorem and an application of the noncommutative Khintchine inequality (6.23) yields

$$E := \mathbb{E} \left\| \sum_{j,k=1}^{M} \epsilon_j \epsilon'_k B_{j,k} \right\|_{S_{2n}}^{2n} \tag{6.34}$$

$$\leq C_n \mathbb{E}_\epsilon \max \left\{ \left\| \left(\sum_{k=1}^{M} H_k(\epsilon)^* H_k(\epsilon) \right)^{1/2} \right\|_{S_{2n}}^{2n}, \left\| \left(\sum_{k=1}^{M} H_k(\epsilon) H_k(\epsilon)^* \right)^{1/2} \right\|_{S_{2n}}^{2n} \right\},$$

where $H_k(\epsilon) := \sum_{j=1}^{N} \epsilon_j B_{j,k}$. We define

$$\widehat{B}_{j,k} = (0|\dots|0|B_{j,k}|0|\dots|0) \in \mathbb{C}^{r \times tM},$$

and similarly

$$\widetilde{B}_{j,k} = \big(0|\dots|0|B_{j,k}^*|0|\dots|0\big)^* \in \mathbb{C}^{rM \times t},$$

where in both cases the non-zero block $B_{j,k}$ is the kth one. Then

$$\widehat{B}_{j,k} \widehat{B}_{j',k'}^* = \begin{cases} 0 & \text{if } k \neq k', \\ B_{j,k} B_{j',k}^* & \text{if } k = k', \end{cases} \tag{6.35}$$

$$\widetilde{B}_{j,k}^* \widetilde{B}_{j',k'} = \begin{cases} 0 & \text{if } k \neq k', \\ B_{j,k}^* B_{j',k} & \text{if } k = k'. \end{cases}$$

Since the singular values obey $\sigma_k(A) = \sigma_k((AA^*)^{1/2})$, the Schatten class norm satisfies $\|A\|_{S_{2n}} = \|(AA^*)^{1/2}\|_{S_{2n}}$. This allows us to verify that

$$\left\| \sum_{j=1}^{M} \epsilon_j \sum_{k=1}^{M} \widehat{B}_{j,k} \right\|_{S_{2n}} = \left\| \left(\sum_{j,j'} \epsilon_j \epsilon_{j'} \sum_{k,k'} \widehat{B}_{j,k} \widehat{B}_{j',k'}^* \right)^{1/2} \right\|_{S_{2n}}$$

$$= \left\| \left(\sum_{j,j'} \epsilon_j \epsilon_{j'} \sum_{k} B_{j,k} B_{j',k}^* \right)^{1/2} \right\|_{S_{2n}} = \left\| \left(\sum_{k} H_k(\epsilon) H_k(\epsilon)^* \right)^{1/2} \right\|_{S_{2n}}.$$

Similarly, we also get

$$\left\| \left(\sum_{k} H_k(\epsilon)^* H_k(\epsilon) \right)^{1/2} \right\|_{S_{2n}} = \left\| \sum_{j=1}^{M} \epsilon_j \sum_{k=1}^{M} \widetilde{B}_{j,k} \right\|_{S_{2n}}.$$

Plugging the above expressions into (6.34) we can further estimate

$$E \leq C_n \left(\mathbb{E} \left\| \sum_{j=1}^{M} \epsilon_j \sum_{k=1}^{M} \widehat{B}_{j,k} \right\|_{S_{2n}}^{2n} + \mathbb{E} \left\| \sum_{j=1}^{M} \epsilon_j \sum_{k=1}^{M} \widetilde{B}_{j,k} \right\|_{S_{2n}}^{2n} \right).$$

Using Khintchine's inequality (6.23) once more we obtain

$$\begin{aligned} E_1 &:= \mathbb{E} \left\| \sum_{j=1}^{M} \epsilon_j \sum_{k=1}^{M} \widehat{B}_{j,k} \right\|_{S_{2n}}^{2n} \\ &\leq C_n \max \left\{ \left\| \left(\sum_j \widetilde{H}_j \widetilde{H}_j^* \right)^{1/2} \right\|_{S_{2n}}^{2n}, \left\| \left(\sum_j \widetilde{H}_j^* \widetilde{H}_j \right)^{1/2} \right\|_{S_{2n}}^{2n} \right\}, \end{aligned}$$

where $\widetilde{H}_j = \sum_{k=1}^{M} \widehat{B}_{j,k}$. Using (6.35) we see that

$$\sum_j \widetilde{H}_j \widetilde{H}_j^* = \sum_{k,j} B_{j,k} B_{j,k}^*.$$

Furthermore, noting that

$$F = \begin{pmatrix} B_{1,1} & B_{1,2} & \dots & B_{1,M} \\ B_{2,1} & B_{2,2} & \dots & B_{2,M} \\ \vdots & \vdots & \vdots & \vdots \\ B_{M,1} & B_{M,2} & \dots & B_{M,M} \end{pmatrix} = \begin{pmatrix} \widetilde{H}_1 \\ \widetilde{H}_2 \\ \vdots \\ \widetilde{H}_M \end{pmatrix},$$

we have

$$\|(\sum_j \widetilde{H}_j^* \widetilde{H}_j)^{1/2}\|_{S_{2n}}^{2n} = \|(F^* F)^{1/2}\|_{S_{2n}}^{2n} = \|F\|_{S_{2n}}^{2n}.$$

Hence,

$$E_1 \leq C_n \max \left\{ \|(\sum_{j,k=1}^{M} B_{j,k} B_{j,k}^*)^{1/2}\|_{S_{2n}}^{2n}, \|F\|_{S_{2n}}^{2n} \right\}.$$

As $\widetilde{B}_{j,k}$ differs from $\widehat{B}_{j,k}$ only by interchanging $B_{j,k}$ with $B_{j,k}^*$ we obtain similarly

$$E_2 := \mathbb{E} \| \sum_{j=1}^{M} \epsilon_j \sum_{k=1}^{M} \widetilde{B}_{j,k} \|_{S_{2n}}^{2n} \leq C_n \max \left\{ \| \sum_{j,k=1}^{M} B_{j,k}^* B_{j,k})^{1/2} \|_{S_{2n}}^{2n}, \|\widetilde{F}\|_{S_{2n}}^{2n} \right\}.$$

Finally,

$$\begin{aligned} E &\leq C_n(E_1 + E_2) \\ &\leq 2 \cdot C_n^2 \max \left\{ \left\| \left(\sum_{j,k=1}^M B_{j,k}^* B_{j,k} \right)^{1/2} \right\|_{S_{2n}}^{2n}, \left\| \left(\sum_{j,k=1}^M B_{j,k} B_{j,k}^* \right)^{1/2} \right\|_{S_{2n}}^{2n}, \right. \\ &\qquad \left. \|F\|_{S_{2n}}^{2n}, \|\widetilde{F}\|_{S_{2n}}^{2n} \right\}. \end{aligned}$$

This concludes the proof. □

6.9 Dudley's Inequality

A stochastic process is a collection X_t, $t \in \hat{T}$, of complex-valued random variables indexed by some set $\hat{T}$. We are interested in bounding the moments of its supremum. In order to avoid measurability issues (in general, the supremum of an uncountable number of random variables might not be measurable any more) we define, for a subset $T \subset \hat{T}$, the so called lattice supremum as

$$\mathbb{E} \sup_{t \in T} |X_t| := \sup\{\mathbb{E} \sup_{t \in F} |X_t|, F \subset T, F \text{ finite}\}. \tag{6.36}$$

Note that for a countable set T, where no measurability problems can arise, $\mathbb{E}(\sup_{t \in T} |X_t|)$ equals the right hand side above. Dudley's inequality, which was originally formulated and shown in [45] for the expectation, bounds the moments $\mathbb{E} \sup_{t \in T} |X_t|^p$ from above by a geometric quantity involving the covering numbers of T.

We endow $\hat{T}$ with the pseudometric

$$d(s,t) = \left(\mathbb{E}|X_t - X_s|^2\right)^{1/2}. \tag{6.37}$$

Recall that in contrast to a metric a pseudometric does not need to separate points, i.e., $d(s,t) = 0$ does not necessarily imply $s = t$. We assume that the increments of the process satisfy,

$$\mathbb{P}(|X_t - X_s| \geq u d(t,s)) \leq 2 \exp\left(-u^2/2\right), \quad t, s \in \hat{T}, \quad u > 0. \tag{6.38}$$

We will later apply Dudley's inequality for the special case of Rademacher processes of the form

$$X_t = \sum_{j=1}^M \epsilon_j x_j(t), \quad t \in \hat{T}, \tag{6.39}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_M)$ is a Rademacher sequence and the $x_j : \hat{T} \to \mathbb{C}$ are some deterministic functions. Observe that

$$d(t,s)^2 = \mathbb{E}|X_t - X_s|^2 = \mathbb{E}|\sum_{j=1}^{M} \epsilon_j (x_j(t) - x_j(s))|^2$$

$$= \sum_{j=1}^{M} (x_j(t) - x_j(s))^2 = \|\mathbf{x}(t) - \mathbf{x}(s)\|_2^2,$$

where $\mathbf{x}(t)$ denotes the vector with components $x_j(t), j = 1, \ldots, M$. Therefore, we define the (pseudo-)metric

$$d(s,t) = \left(\mathbb{E}|X_t - X_s|^2\right)^{1/2} = \|\mathbf{x}(t) - \mathbf{x}(s)\|_2. \tag{6.40}$$

Hoeffding's inequality (Proposition 6.11) shows that the Rademacher process (6.39) satisfies (6.38). Although we will need Dudley's inequality only for Rademacher processes here, we note that the original formulation was for Gaussian processes, see also [3, 79, 80, 99, 125].

For a subset $T \subset \hat{T}$, the covering number $N(T, d, \varepsilon)$ is defined as the smallest integer N such that there exists a subset $E \subset \hat{T}$ with cardinality $|E| = N$ satisfying

$$T \subset \bigcup_{t \in E} B_d(t, \varepsilon),$$

where $B_d(t, \varepsilon) = \{s \in \hat{T}, d(t,s) \leq \varepsilon\}$. In words, T can be covered with N balls of radius ε in the metric d. Note that some authors additionally require that $E \subset T$. For us $E \subset \hat{T}$ will be sufficient. Denote the diameter of T in the metric d by

$$\Delta(T) := \sup_{s,t \in T} d(s,t).$$

With these concepts at hand our version of Dudley's inequality reads as follows.

Theorem 6.23. *Let $X_t, t \in \hat{T}$, be a complex-valued process indexed by a pseudometric space $(\hat{T}, d)$ with pseudometric defined by* (6.37) *which satisfies* (6.38)*. Then, for a subset $T \subset \hat{T}$ and any point $t_0 \in T$ it holds*

$$\mathbb{E} \sup_{t \in T} |X_t - X_{t_0}| \leq C_1 \int_0^{\Delta(T)} \sqrt{\ln(N(T,d,u))} du + D_1 \Delta(T) \tag{6.41}$$

with constants $C_1 = 16.51$ and $D_1 = 4.424$. Furthermore, for $p \geq 2$,

$$\left(\mathbb{E} \sup_{t \in T} |X_t - X_{t_0}|^p\right)^{1/p} \leq \beta^{1/p} \sqrt{p} \left(C \int_0^{\Delta(T)} \sqrt{\ln(N(T,d,u))} du + D\Delta(T)\right) \tag{6.42}$$

with constants $C = 14.372$, $D = 5.818$ and $\beta = 6.028$.

We note that the estimate (6.42) also holds for $1 \leq p \leq 2$ with possibly slightly different constants (this can be seen, for instance, via interpolation between $p = 1$ and $p = 2$). Further, the theorem and its proof easily extend to Banach space valued processes satisfying $\mathbb{P}(\|X_t - X_s\| > ud(t,s)) \leq 2e^{-u^2/2}$. Inequality (6.42) for the increments of the process can be used in the following way to bound the supremum,

$$\left(\mathbb{E}\sup_{t\in T}|X_t|^p\right)^{1/p} \leq \inf_{t_0\in T}\left(\left(\mathbb{E}\sup_{t\in T}|X_t - X_{t_0}|^p\right)^{1/p} + (\mathbb{E}|X_{t_0}|^p)^{1/p}\right)$$
$$\leq \beta^{1/p}\sqrt{p}\left(\int_0^{\Delta(T)}\sqrt{\log(N(T,d,u))}du + D\Delta(T)\right) + \inf_{t_0\in T}(\mathbb{E}|X_{t_0}|^p)^{1/p}.$$

The second term is usually easy to estimate. Further, note that for a centered real-valued process, that is, $\mathbb{E}X_t = 0$ for all $t \in \hat{T}$, we have

$$\mathbb{E}\sup_{t\in T} X_t = \mathbb{E}\sup_{t\in T}(X_t - X_{t_0}) \leq \mathbb{E}\sup_{t\in T}|X_t - X_{t_0}|. \tag{6.43}$$

For completeness we also state the usual version of Dudley's inequality.

Corollary 6.24. *Let $X_t, t \in T$, be a real-valued centered process indexed by a pseudometric space (T,d) such that* (6.38) *holds. Then*

$$\mathbb{E}\sup_{t\in T} X_t \leq 30\int_0^{\Delta(T)}\sqrt{\ln(N(T,d,u))}du. \tag{6.44}$$

Proof. As in the proof of Theorem 6.23 below, we may assume without loss of generality that $\Delta(T) = 1$. Then it follows that $N(T,d,u) \geq 2$ for all $u < 1/2$. Indeed, if $N(T,d,u) = 1$ for some $u < 1/2$ then, for any $\epsilon > 0$, there would be two points of distance at least $1-\epsilon$ that are covered by one ball of radius u, a contradiction to the triangle inequality. Therefore,

$$\int_0^{\Delta(T)}\sqrt{\ln(N(T,d,u))}du \geq \int_0^{1/2}\sqrt{\ln(2)}du = \frac{\sqrt{\ln 2}}{2}\Delta(T).$$

Therefore, (6.44) follows from (6.41) and (6.43) and the estimate

$$C_1 + \frac{2D_1}{\sqrt{\ln 2}} < 30.$$

□

Generalizations of Dudley's inequality are contained in [80, 125]; in particular, extensions to generic chaining inequalities, or bounds of suprema of random processes by means of majorizing measure conditions.

Proof of Theorem 6.23. Without loss of generality we may assume that the right hand sides of (6.41) and (6.42) are finite and non-vanishing. Otherwise, the statement becomes trivial. In particular, $0 < \Delta(T) < \infty$ and $N(T, d, u) < \infty$ for all $u > 0$. By eventually passing to a rescaled process $X'_t = X_t/\Delta(T)$ we may assume $\Delta(T) = 1$.

Now let $b > 1$ to be specified later. According to the definition of the covering numbers, there exist finite subsets $E_j \subset \hat{T}$, $j \in \mathbb{N}\setminus\{0\}$, of cardinality $|E_j| = N(T, d, b^{-j})$ such that

$$T \subset \bigcup_{t\in E_j} B_d(t, b^{-j}).$$

For each $t \in T$ and $j \in \mathbb{N} \setminus \{0\}$ we can therefore define $\pi_j(t) \in E_j$ such that

$$d(t, \pi_j(t)) \le b^{-j}.$$

Further set $\pi_0(t) = t_0$. Then by the triangle inequality

$$d(\pi_j(t), \pi_{j-1}(t)) \le d(\pi_j(t), t) + d(\pi_{j-1}(t), t) \le (1 + b) \cdot b^{-j} \quad \text{for all } j \ge 2$$

and $d(\pi_1(t), \pi_0(t)) \le \Delta(T) = 1$. Therefore,

$$d(\pi_j(t), \pi_{j-1}(t)) \le (1 + b) \cdot b^{-j}, \quad \text{for all } j \ge 1. \tag{6.45}$$

Now we claim the chaining identity

$$X_t - X_{t_0} = \sum_{j=1}^{\infty} (X_{\pi_j(t)} - X_{\pi_{j-1}(t)}) \quad \text{almost surely.} \tag{6.46}$$

Indeed, by (6.38) we have

$$\begin{aligned}
&\mathbb{P}(|X_{\pi_j(t)} - X_{\pi_{j-1}(t)}| \ge b^{-j/2}) \\
&\le \mathbb{P}\left(|X_{\pi_j(t)} - X_{\pi_{j-1}(t)}| \ge \frac{b^{j/2}}{1+b} d(\pi_j(t), \pi_{j-1}(t))\right) \le 2\exp\left(-\frac{1}{2(1+b)^2} b^j\right).
\end{aligned}$$

This implies that $\sum_{j=1}^{\infty} \mathbb{P}(|X_{\pi_j(t)} - X_{\pi_{j-1}(t)}| \ge b^{-j/2}) < \infty$. It follows from the Borel Cantelli lemma (Lemma 6.4) that the event that there exists an increasing sequence $j_\ell, \ell = 1, 2, \ldots$ of integers with $j_\ell \to \infty$ as $\ell \to \infty$ such that $|X_{\pi_{j_\ell}(t)} - X_{\pi_{j_\ell - 1}(t)}| \ge b^{-j/2}$ has zero probability. In conclusion, $|X_{\pi_j(t)} - X_{\pi_{j-1}(t)}| < b^{-j/2}$ for all $j \ge j_0$ and some j_0 holds almost surely. Consequently, the series on the right hand side of (6.46) converges almost surely. Furthermore,

$$\begin{aligned}
&\mathbb{E}\left|X_t - X_{t_0} - \sum_{j=1}^{J} (X_{\pi_j(t)} - X_{\pi_{j-1}(t)})\right|^2 = \mathbb{E}|X_t - X_{\pi_J(t)}|^2 \\
&= d(t, \pi_J(t))^2 \to 0 \quad (J \to \infty)
\end{aligned}$$

by definition (6.37) of the metric d and construction of the $\pi_j(t)$. The chaining identity (6.46) follows.

Now let F be a finite subset of T. Let $a_j > 0$, $j > 0$, be numbers to be determined later. For brevity of notation we write $N(T, d, b^{-j}) = N(b^{-j})$. Then

$$\begin{aligned}
&\mathbb{P}\left(\max_{t\in F}|X_t - X_{t_0}| > u\sum_{j=1}^{\infty} a_j\right)\\
&\leq \mathbb{P}\left(\max_{t\in F}\sum_{j=1}^{\infty}|X_{\pi_j(t)} - X_{\pi_{j-1}(t)}| > u\sum_{j=1}^{\infty} a_j\right)\\
&\leq \sum_{j=1}^{\infty}\mathbb{P}\left(\max_{t\in F}|X_{\pi_j(t)} - X_{\pi_{j-1}(t)}| > ua_j\right)\\
&\leq \sum_{j=1}^{\infty} N(b^{-j})N(b^{-(j-1)})\max_{t\in F}\mathbb{P}\left(|X_{\pi_j(t)} - X_{\pi_{j-1}(t)}| \geq ua_j\right)\\
&\leq \sum_{j=1}^{\infty} N(b^{-j})N(b^{-(j-1)})\max_{t\in F}\mathbb{P}\left(|X_{\pi_j(t)} - X_{\pi_{j-1}(t)}| \geq \frac{ua_j d(\pi_j(t), \pi_{j-1}(t))}{(1+b)\cdot b^{-j}}\right)\\
&\leq 2\sum_{j=1}^{\infty} N(b^{-j})N(b^{-(j-1)})\exp\left(-\frac{u^2(b^j a_j)^2}{2(1+b)^2}\right). \qquad (6.47)
\end{aligned}$$

Hereby we used that the number of possible increments $|X_{\pi_j(t)} - X_{\pi_{j-1}(t)}|$ is bounded by the product $N(b^{-j})N(b^{-(j-1)})$ of the cardinalities of the sets E_j and E_{j-1}. Further, we have applied (6.45) and (6.38). Now for a number $\alpha > 0$ to be determined later, we choose

$$a_j = \sqrt{2}\alpha^{-1}(1+b)\cdot b^{-j}\sqrt{\ln(b^j N(b^{-j})N(b^{-(j-1)}))}, \quad j \geq 1.$$

Continuing the chain of inequalities (6.47) yields, for $u \geq \alpha$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{t\in F}|X_t - X_{t_0}| > u\sum_{j=1}^{\infty} a_j\right)\\
&\leq 2\sum_{j=1}^{\infty} N(b^{-j})N(b^{-(j-1)})\left(b^j N(b^{-j})N(b^{-(j-1)})\right)^{-u^2/\alpha^2}\\
&\leq 2\sum_{j=1}^{\infty} b^{-ju^2/\alpha^2} \leq 2b^{-u^2/\alpha^2}\sum_{j=0}^{\infty} b^{-j} = \frac{2b}{b-1}b^{-u^2/\alpha^2}.
\end{aligned}$$

Using $N(b^{-(j-1)}) \leq N(b^{-j})$ we further obtain

$$\Theta := \sum_{j=1}^{\infty} a_j \leq \sqrt{2}\alpha^{-1}(b+1)\sum_{j=1}^{\infty} b^{-j}\sqrt{j\ln(b) + 2\ln(N(b^{-j}))}$$
$$\leq \sqrt{2}\alpha^{-1}(b+1)\sum_{j=1}^{\infty} b^{-j}\sqrt{j\ln(b)} + 2\alpha^{-1}(b+1)\sum_{j=1}^{\infty} b^{-j}\sqrt{\ln(N(b^{-j}))}. \quad (6.48)$$

By comparing sums and integrals, the second sum in (6.48) is upperbounded by

$$\sum_{j=1}^{\infty} b^{-j}\sqrt{\ln(N(b^{-j}))} = \frac{b}{b-1}\sum_{j=1}^{\infty}\sqrt{\ln(N(b^{-j}))}\int_{b^{-(j+1)}}^{b^{-j}} du$$
$$\leq \frac{b}{b-1}\int_0^{b^{-1}} \sqrt{\ln(N(T,d,u))}du \leq \frac{b}{b-1}\int_0^{\Delta(T)} \sqrt{\ln(N(T,d,u))}du,$$

where we additionally used that $N(T,d,b^{-j}) \leq N(T,d,u)$ for all $u \in [b^{-(j+1)}, b^{-j}]$. Plugging into (6.48) shows that

$$\Theta \leq C(b,\alpha)\Delta(T) + \frac{2b(b+1)}{\alpha(b-1)}\int_0^{\Delta(T)} \sqrt{\ln(N(T,d,u))}du \quad (6.49)$$

with

$$C(b,\alpha) := \sqrt{2}\alpha^{-1}(b+1)\sqrt{\ln(b)}\sum_{j=1}^{\infty} b^{-j}\sqrt{j}, \quad (6.50)$$

while

$$\mathbb{P}(\max_{t\in F}|X_t - X_{t_0}| > u\Theta) \leq \frac{2b}{b-1}b^{-u^2/\alpha^2}, \quad u \geq \alpha.$$

Using that any probability is bounded by 1, Proposition 6.1 yields for the moments

$$\mathbb{E}\sup_{t\in F}|X_t - X_{t_0}|^p = p\int_0^{\infty}\mathbb{P}(\sup_{t\in F}|X_t - X_{t_0}| \geq v)v^{p-1}dv$$
$$= p\Theta^p\int_0^{\infty}\mathbb{P}(\sup_{t\in F}|X_t - X_{t_0}| \geq u\Theta)u^{p-1}du$$
$$\leq p\Theta^p\left(\int_0^{\alpha} u^{p-1}du + \frac{2b}{b-1}\int_{\alpha}^{\infty} b^{-u^2/\alpha^2}u^{p-1}du\right)$$
$$= p\Theta^p\left(\frac{\alpha^p}{p} + \frac{2b}{b-1}\int_{\alpha}^{\infty} b^{-u^2/\alpha^2}u^{p-1}du\right).$$

Taking the supremum over all finite subsets $F \subset T$ yields

$$\left(\mathbb{E}\sup_{t\in T}|X_t - X_{t_0}|^p\right)^{1/p} \leq K_1(p,b,\alpha)\int_0^{\Delta(T)}\sqrt{\ln(N(T,d,u))}du$$
$$+ K_2(p,b,\alpha)\Delta(T),$$

where

$$K_1(p,b,\alpha) = p^{1/p}\frac{2b(b+1)}{\alpha(b-1)}\left(\frac{\alpha^p}{p} + \frac{2b}{b-1}\int_\alpha^\infty b^{-u^2/\alpha^2}u^{p-1}du\right)^{1/p},$$

and

$$K_2(p,b,\alpha) = p^{1/p}C(b,\alpha)\left(\frac{\alpha^p}{p} + \frac{2b}{b-1}\int_\alpha^\infty b^{-u^2/\alpha^2}u^{p-1}du\right)^{1/p}.$$

(Readers who do not care about the values of the constants may be satisfied at this point.) We choose $\alpha = \sqrt{2\ln(b)}$. Consider first $p = 1$. Lemma 10.2 in the Appendix yields

$$\int_\alpha^\infty b^{-u^2/\alpha^2}du = \int_{\sqrt{2\ln(b)}}^\infty e^{-u^2/2}du \leq \frac{1}{b\sqrt{2\ln(b)}}.$$

Hence,

$$\hat{K}_1(b) := K_1(1,b,\sqrt{2\ln(b)}) \leq \frac{2b(b+1)}{b-1} + \frac{2b(b+1)}{\ln(b)(b-1)^2}.$$

In order to estimate K_2 we note that, for $x < 1$,

$$\sum_{j=1}^\infty x^j\sqrt{j} \leq \sum_{j=1}^\infty x^j j = x\frac{d}{dx}(\sum_{j=1}^\infty x^j) = \frac{x}{(1-x)^2}.$$

Therefore,

$$C(b,\alpha) \leq \sqrt{2\ln(b)}\alpha^{-1}\frac{b(b+1)}{(b-1)^2}, \tag{6.51}$$

and

$$\hat{K}_2(b) := K_2(1,b,\sqrt{2\ln(b)}) \leq \frac{b(b+1)}{(b-1)^2}\sqrt{2\ln(b)} + \frac{b(b+1)}{(b-1)^3\ln(b)}.$$

The choice $b = 3.8$ yields $\hat{K}_1(3.8) \leq 16.51 = C_1$ and $\hat{K}_2(3.8) \leq 4.424 = D_1$. This yields the claim for $p = 1$.

Now assume $p \geq 2$. We use the Gamma function $\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt$ and the inequality $\Gamma(x) \leq \frac{x^{x-1/2}}{e^{x-1}}$, for $x \geq 1$, see [81], to estimate (recall $\alpha = \sqrt{2\ln(b)}$)

$$\int_\alpha^\infty b^{-u^2/\alpha^2}u^{p-1}du = \int_{\sqrt{2\ln(b)}}^\infty e^{-u^2/2}u^{p-1}du \leq \int_0^\infty e^{-u^2/2}u^{p-1}du$$
$$= 2^{p/2-1}\int_0^\infty e^{-t}t^{p/2-1}dt = 2^{p/2-1}\Gamma(p/2) \leq (2/e)^{p/2-1}(p/2)^{p/2-1/2}$$
$$= \frac{e}{\sqrt{2p}}(p/e)^{p/2}.$$

This yields, recalling that $p \geq 2$ and using that $p^{1/(2p)} \leq e^{1/(2e)}$,

$$\begin{aligned}
&\hat{K}_1(p) := K_1(p, b, \sqrt{2\ln b}) \\
&\leq p^{1/p} \frac{2b(b+1)}{\sqrt{2\ln b}(b-1)} \left(\frac{(2\ln b)^{p/2}}{p} + \frac{2b}{b-1} \frac{e}{\sqrt{2p}} (p/e)^{p/2} \right)^{1/p} \\
&\leq \frac{2b(b+1)}{b-1} + p^{1/(2p)} \frac{2b(b+1)}{\sqrt{2\ln b}(b-1)} \left(\frac{\sqrt{2}eb}{b-1} \right)^{1/p} e^{-1/2}\sqrt{p} \\
&\leq \frac{\sqrt{2}b(b+1)}{b-1} \left(\frac{\sqrt{2}eb}{b-1} \right)^{1/p} \sqrt{p} + \frac{2e^{1/(2e)}e^{-1/2}b(b+1)}{\sqrt{2\ln b}(b-1)} \left(\frac{\sqrt{2}eb}{b-1} \right)^{1/p} \sqrt{p} \\
&= \sqrt{2}b(b+1) \left(\frac{1}{b-1} + \frac{e^{1/(2e)-1/2}}{\sqrt{\ln b}(b-1)} \right) \left(\frac{\sqrt{2}eb}{b-1} \right)^{1/p} \sqrt{p}.
\end{aligned}$$

Using (6.51) we estimate similarly

$$\begin{aligned}
&\hat{K}_2(p) := K_2(p, b, \sqrt{2\ln b}) \\
&\leq p^{1/p} \frac{b(b+1)}{(b-1)^2} \left(\frac{(2\ln b)^{p/2}}{p} + \frac{2b}{b-1} \frac{e}{\sqrt{2p}} (p/e)^{p/2} \right)^{1/p} \\
&\leq \frac{b(b+1)\sqrt{2\ln b}}{(b-1)^2} + p^{1/(2p)} e^{-1/2} \frac{b(b+1)}{(b-1)^2} \left(\frac{\sqrt{2}eb}{b-1} \right)^{1/p} \sqrt{p} \\
&\leq \left(\frac{b(b+1)\sqrt{\ln b}}{(b-1)^2} + \frac{e^{1/(2e)-1/2}b(b+1)}{(b-1)^2} \right) \left(\frac{\sqrt{2}eb}{b-1} \right)^{1/p} \sqrt{p}.
\end{aligned}$$

In conclusion, inequality (6.42) holds with

$$\begin{aligned}
&\beta = \frac{\sqrt{2}eb}{b-1}, \quad C = \sqrt{2}b(b+1) \left(\frac{1}{b-1} + \frac{e^{1/(2e)-1/2}}{\sqrt{\ln b}(b-1)} \right), \\
&D = \frac{b(b+1)\sqrt{\ln b}}{(b-1)^2} + \frac{e^{1/(2e)-1/2}b(b+1)}{(b-1)^2}.
\end{aligned}$$

Now we choose $b = 2.76$ to obtain $\beta = 6.028$, $C = 14.372$ and $D = 5.818$. This completes the proof. □

6.10 Deviation Inequalities for Suprema of Empirical Processes

The strong probability estimate of Theorem 4.4 depends on a deviation inequality for suprema of empirical processes that we present in this section. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_M$ be

independent random vectors in $\mathbb{C}^n$ and let $\mathcal{F}$ be a countable collection of functions from $\mathbb{C}^n$ into $\mathbb{R}$. We are interested in the random variable

$$Z = \sup_{f \in \mathcal{F}} \sum_{\ell=1}^{M} f(\mathbf{Y}_\ell), \tag{6.52}$$

that is, the supremum of an empirical process. The next theorem estimates the probability that Z deviates much from its mean.

Theorem 6.25. *Let $\mathcal{F}$ be a countable set of functions $f : \mathbb{C}^n \to \mathbb{R}$. Let $\mathbf{Y}_1, \dots, \mathbf{Y}_M$ be independent copies of a random vector $\mathbf{Y}$ on $\mathbb{C}^n$ such that $\mathbb{E} f(\mathbf{Y}) = 0$ for all $f \in \mathcal{F}$, and assume $f(\mathbf{Y}) \leq 1$ almost surely. Let Z be the random variable defined in* (6.52) *and $\mathbb{E}Z$ its expectation. Let $\sigma^2 > 0$ such that $\mathbb{E}\left[f(\mathbf{Y})^2\right] \leq \sigma^2$ for all $f \in \mathcal{F}$. Set $v_M = M\sigma^2 + 2\mathbb{E}Z$. Then, for all $t > 0$,*

$$\mathbb{P}(Z \geq \mathbb{E}Z + t) \leq \exp(-v_M h(t/v_M)) \leq \exp\left(-\frac{t^2}{2v_M + 2t/3}\right), \tag{6.53}$$

where $h(t) = (1+t)\ln(1+t) - t$.

This theorem, in particular, the left-hand inequality (6.53), is taken from [14]. The second inequality in (6.53) follows from $h(t) \geq \frac{t^2}{2+2t/3}$ for all $t > 0$. If $\mathcal{F}$ consists only of a single function f, then Theorem 6.25 reduces to the ordinary Bernstein or Bennett inequality [6, 134]. Hence, (6.53) can be viewed as a far reaching generalization of these inequalities.

The proof of (6.53), which uses the concept of entropy, is beyond the scope of these notes. We refer the interested reader to [14]. Deviation inequalities for suprema of empirical processes were already investigated in the 1980ies by P. Massart and others, see e.g. [84, 1]. M. Talagrand obtained major breakthroughs in [122, 123], in particular, he obtained also a concentration inequality of the following type: Let $\mathbf{Y}_1, \dots, \mathbf{Y}_M$ be independent random vectors and $|f(\mathbf{Y}_\ell)| \leq 1$ almost surely for all $f \in \mathcal{F}$ and all $\ell = 1, \dots, M$. Then

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq 3\exp\left(-\frac{t}{C}\log\left(1 + \frac{t}{M\sigma^2}\right)\right), \tag{6.54}$$

where $C > 0$ is a universal constant. The constants in the deviation and concentration inequalities were successfully improved in [85, 107, 108, 14, 77]. Extensions of deviation and concentration inequalities can be found in [10, 9, 79].

7 Proof of Nonuniform Recovery Result for Bounded Orthonormal Systems

In this section we prove Theorem 4.2.

7.1 Nonuniform Recovery with Coefficients of Random Signs

In order to obtain nonuniform recovery results we use the recovery condition for individual vectors, Corollary 2.9. In order to simplify arguments we also choose the signs of the non-zero coefficients of the sparse vector at random. A general recovery result reads as follows.

Proposition 7.1. *Let $A = (\mathbf{a}_1, \dots, \mathbf{a}_N) \in \mathbb{C}^{m\times N}$ and let $S \subset [N]$ of size $|S| = s$. Assume A_S is injective and*

$$\|A_S^\dagger \mathbf{a}_\ell\|_2 \leq \alpha < 1/\sqrt{2} \quad \text{for all } \ell \notin S, \tag{7.1}$$

where $A^\dagger$ is the Moore-Penrose pseudo-inverse of A_S. Let $\boldsymbol{\epsilon} = (\epsilon_j)_{j\in S} \in \mathbb{C}^s$ be a (random) Rademacher or Steinhaus sequence. Then with probability at least

$$1 - 2^{3/4}(N-s)\exp(-\alpha^{-2}/2)$$

every vector $\mathbf{x} \in \mathbb{C}^N$ with support S and $\operatorname{sgn}(\mathbf{x}^S) = \boldsymbol{\epsilon}$ is the unique solution to the ℓ_1-minimization problem (2.12)*.*

Proof. In the Rademacher case the union bound and Hoeffding's inequality, Corollary 6.10, yield

$$\begin{aligned}\mathbb{P}(\max_{\ell\notin S} |\langle A_S^\dagger \mathbf{a}_\ell, \operatorname{sgn}(\mathbf{x}_S)\rangle| \geq 1) &\leq \sum_{\ell\notin S} \mathbb{P}\left(|\langle A_S^\dagger \mathbf{a}_\ell, \operatorname{sgn}(\mathbf{x}_S)\rangle| \geq \|A_S^\dagger \mathbf{a}_\ell\|_2 \alpha^{-1}\right) \\ &\leq (N-s) 2^{3/4} \exp(-\alpha^{-2}/2).\end{aligned}$$

In the Steinhaus case we even obtain a better estimate from Corollary 6.13. An application of Corollary 2.9 finishes the proof. □

In view of the previous proposition it is enough to show that $\|A_S^\dagger \mathbf{a}_\ell\|_2$ is small. The next statement indicates a way how to pursue this task.

Proposition 7.2. *Let $A \in \mathbb{C}^{m\times N}$ with coherence μ and let $S \subset [N]$ of size s. Assume that*

$$\|A_S^* A_S - \mathrm{Id}\|_{2\to 2} \leq \delta \tag{7.2}$$

for some $\delta \in (0,1)$. Then

$$\|A_S^\dagger \mathbf{a}_\ell\|_2 \leq \frac{\sqrt{s}\mu}{1-\delta} \quad \text{for all } \ell \notin S.$$

Proof. By definition of the operator norm

$$\|A_S^\dagger \mathbf{a}_\ell\|_2 = \|(A_S^* A_S)^{-1} A_S^* \mathbf{a}_\ell\|_2 \leq \|(A_S^* A_S)^{-1}\|_{2\to 2} \|A_S^* \mathbf{a}_\ell\|_2. \tag{7.3}$$

The Neumann series yields

$$\|(A_S^* A_S)^{-1}\|_{2\to 2} = \|\sum_{k=0}^{\infty} (\mathrm{Id} - A_S^* A_S)^k\|_{2\to 2} \leq \sum_{k=0}^{\infty} \|\mathrm{Id} - A_S^* A_S\|_{2\to 2}^k$$

$$\leq \sum_{k=0}^{\infty} \delta^k = \frac{1}{1-\delta}$$

by the geometric series formula. The second term in (7.3) can be estimated using the coherence,

$$\|A_S^* \mathbf{a}_\ell\|_2 = \sqrt{\sum_{j\in S} |\langle \mathbf{a}_\ell, \mathbf{a}_j\rangle|^2} \leq \sqrt{s}\mu.$$

Combining the two estimates completes the proof. □

We note that in contrast to the usual definition of coherence, we do not require the columns of A to be normalized in the previous statement. Condition (7.2) is a different way of saying that the eigenvalues of $A_S^* A_S$ are contained in $[1-\delta, 1+\delta]$, or that the singular values of A_S are contained in $[\sqrt{1-\delta}, \sqrt{1+\delta}]$.

7.2 Condition Number Estimate for Column Submatrices

Let us return now to the situation of Theorem 4.4. Proposition 7.2 requires to provide an estimate on the coherence of A and on $\|A_S^* A_S - \mathrm{Id}\|_{2\to 2}$. The latter corresponds to a probabilistic condition number estimate of a column submatrix of the structured random matrix A of the form (4.4). The estimate of the coherence will follow as a corollary. (Note, however, that the coherence alone might be estimated with simpler tools, see for instance [78].) The main theorem of this section reads as follows.

Theorem 7.3. *Let $A \in \mathbb{C}^{m\times N}$ be the sampling matrix* (4.4) *associated to an orthonormal system that satisfies the boundedness condition* (4.2) *for some constant $K \geq 1$. Let $S \subset [N]$ be of cardinality $|S| = s \geq 2$. Assume that the random sampling points $t_1, \ldots, t_m$ are chosen independently according to the orthogonalization measure ν. Let $\delta \in (0, 1/2]$. Then with probability at least*

$$1 - 2^{3/4} s \exp\left(-\frac{m\delta^2}{\widetilde{C} K^2 s}\right), \tag{7.4}$$

where $\widetilde{C} = 9 + \sqrt{17} \approx 13.12$, the normalized matrix $\widetilde{A} = \frac{1}{\sqrt{m}} A$ satisfies

$$\|\widetilde{A}_S^* \widetilde{A}_S - \mathrm{Id}\|_{2\to 2} \leq \delta.$$

We note that the theorem also holds for $1/2 \leq \delta < 1$ with a slightly larger constant $\widetilde{C}$.

Proof. Denote by $\mathbf{X}_\ell = (\overline{\psi_j}(t_\ell))_{j\in S} \in \mathbb{C}^s$ a column vector of A_S^*. By independence of the t_ℓ these are i.i.d. random vectors. Their 2-norm is bounded by

$$\|\mathbf{X}_\ell\|_2 = \sqrt{\sum_{j\in S} |\psi_j(t_\ell)|^2} \leq K\sqrt{s}. \tag{7.5}$$

Furthermore,

$$\mathbb{E}\left(\mathbf{X}_\ell \mathbf{X}_\ell^*\right)_{j,k} = \mathbb{E}\left[\psi_k(t_\ell)\overline{\psi_j(t_\ell)}\right] = \int_{\mathcal{D}} \psi_j(t)\overline{\psi_k(t)} d\nu(t) = \delta_{j,k}, j,k \in S,$$

or in other words, $\mathbb{E}\mathbf{X}_\ell\mathbf{X}_\ell^* = \mathrm{Id}$. Using symmetrization, Lemma 6.7, we estimate, for $p \geq 2$,

$$\begin{aligned} E_p := \mathbb{E}\|\widetilde{A}_S^*\widetilde{A}_S - \mathrm{Id}\|_{2\to 2}^p &= \mathbb{E}\|\frac{1}{m}\sum_{\ell=1}^m \left(\mathbf{X}_\ell\mathbf{X}_\ell^* - \mathbb{E}\mathbf{X}_\ell\mathbf{X}_\ell^*\right)\|_{2\to 2}^p \\ &\leq \left(\frac{2}{m}\right)^p \mathbb{E}\|\sum_{\ell=1}^m \epsilon_\ell \mathbf{X}_\ell\mathbf{X}_\ell^*\|_{2\to 2}^p, \end{aligned}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)$ is a Rademacher sequence, independent of $\mathbf{X}_1, \ldots, \mathbf{X}_m$. Now, we are in the position to apply Rudelson's lemma 6.18. To this end we note that A_S has rank at most s. Using Fubini's theorem and applying Rudelson's lemma conditional on $(\mathbf{X}_1, \ldots, \mathbf{X}_m)$ yields

$$\begin{aligned} E_p &\leq \left(\frac{2}{m}\right)^p 2^{3/4} s p^{p/2} e^{-p/2} \mathbb{E}\left[\|A_S\|_{2\to 2}^p \max_{\ell\in[m]} \|\mathbf{X}_\ell\|_2^p\right] \\ &\leq \left(\frac{2}{\sqrt{m}}\right)^p 2^{3/4} s p^{p/2} e^{-p/2} \sqrt{\mathbb{E}\|\widetilde{A}_S^*\widetilde{A}_S\|_{2\to 2}^p \mathbb{E}\left[\max_{\ell\in[m]} \|\mathbf{X}_\ell\|_2^{2p}\right]}. \end{aligned} \tag{7.6}$$

In the last step we applied the Cauchy Schwarz inequality. Using the bound (7.5), which holds for all realizations of $\mathbf{X}_1, \ldots, \mathbf{X}_m$, inserting the identity Id into the operator norm and applying the triangle inequality yields

$$\begin{aligned} E_p &\leq \left(\frac{2}{\sqrt{m}}\right)^p 2^{3/4} s p^{p/2} e^{-p/2} s^{p/2} K^p \sqrt{\mathbb{E}\left[(\|\widetilde{A}_S^*\widetilde{A}_S - \mathrm{Id}\|_{2\to 2} + 1)^p\right]} \\ &\leq \left(2K\sqrt{\frac{s}{m}}\right)^p 2^{3/4} s e^{-p/2} p^{p/2} \left((\mathbb{E}\|\widetilde{A}_S^*\widetilde{A}_S - \mathrm{Id}\|_{2\to 2}^p)^{1/p} + 1\right)^{p/2}. \end{aligned}$$

Denoting

$$D_{p,m,s} = 2K\sqrt{\frac{s}{m}} 2^{3/(4p)} s^{1/p} e^{-1/2}\sqrt{p}$$

we have deduced

$$E_p^{1/p} \leq D_{p,m,s}\sqrt{E_p^{1/p} + 1}.$$

Squaring this inequality and completing the squares yields

$$(E_p^{1/p} - D_{p,m,s}^2/2)^2 \leq D_{p,m,s}^2 + D_{p,m,s}^4/4,$$

which gives

$$E_p^{1/p} \leq \sqrt{D_{p,m,s}^2 + D_{p,m,s}^4/4} + D_{p,m,s}^2/2 \tag{7.7}$$

Assuming $D_{p,m,s} \leq 1/2$ this yields

$$E_p^{1/p} \leq \sqrt{1 + \frac{1}{16}} D_{p,m,s} + \frac{1}{4} D_{p,m,s} = \kappa D_{p,m,s} \tag{7.8}$$

with $\kappa = \frac{\sqrt{17}+1}{4}$. Hence,

$$\left(\mathbb{E} \min\{1/2, \|\widetilde{A}_S^* \widetilde{A}_S - \mathrm{Id}\|_{2\to 2}^p\}\right)^{1/p} \leq \min\{1/2, (\mathbb{E}\|\widetilde{A}_S^* \widetilde{A}_S - \mathrm{Id}\|_{2\to 2}^p)^{1/p})\}$$
$$\leq \kappa D_{p,M,s}.$$

It follows from Proposition 6.5 that for $u \geq \sqrt{2}$,

$$\mathbb{P}\left(\min\{1/2, \|\widetilde{A}_S^* \widetilde{A}_S - \mathrm{Id}\|_{2\to 2}\} \geq 2\kappa K \sqrt{\frac{s}{m}} u\right) \leq 2^{3/4} s \exp(-u^2/2),$$

hence, for $2\kappa K \sqrt{\frac{2s}{m}} \leq \delta \leq 1/2$

$$\mathbb{P}(\|A_S^* A_S - \mathrm{Id}\|_{2\to 2} \geq \delta) \leq 2^{3/4} s \exp\left(-\frac{m\delta^2}{8\kappa^2 K^2 s}\right). \tag{7.9}$$

The right hand side in (7.9) is less than ε provided

$$m \geq \frac{\widetilde{C} K^2 s}{\delta^2} \ln(2^{3/4} s/\varepsilon) \tag{7.10}$$

with $\widetilde{C} = 8\kappa^2 = (\sqrt{17}+1)^2/2 = 9 + \sqrt{17} \approx 13.12$. In order to have a non-trivial statement we must have $\varepsilon < 1$. In fact, for $s \geq 2$ condition (7.10) then implies that $\delta \geq 2\kappa K \sqrt{2s/m}$. We conclude that (7.9) holds trivially also for $0 < \delta < 2\kappa K \sqrt{2s/m}$, which finishes the proof. □

The above proof followed ideas contained in [115, 136]. Similar techniques were used in [92]. We remark that in the special case of the trigonometric system (examples (1) and (4) in Section (4.1)), the constant 13.12 in (7.4) can be essentially improved to $3e \approx 8.15$ (see [65] for the precise statement) by exploiting the algebraic structure of the Fourier system [102, 65]. Indeed, one may estimate $\mathbb{E}\|\frac{1}{m} A_S^* A_S - \mathrm{Id}\|_{S_{2n}}^{2n} = \mathbb{E}\,\mathrm{Tr}\left((\frac{1}{m} A_S A_S - \mathrm{Id})^n\right)$ directly in this case, i.e., without Rudelson's lemma or the

Khintchine inequality. This approach, however, is more technical and uses elements from combinatorics.

Note furthermore that the conclusion of the theorem can be reformulated as follows: If for $\varepsilon \in (0,1), \delta \in (0,1/2]$ condition (7.10) holds, then with probability at least $1-\varepsilon$ the normalized matrix $\widetilde{A} = \frac{1}{\sqrt{m}} A$ satisfies

$$\|\widetilde{A}_S^* \widetilde{A}_S - \mathrm{Id}\|_{2\to 2} \leq \delta.$$

The above proof also indicates how the boundedness condition (4.2) may be weakened. Indeed, the term $\mathbb{E} \max_{\ell \in [m]} \|\mathbf{X}_\ell\|_2^{2p}$ in (7.6) was estimated by $K^{2p} s^p$ using the boundedness condition (4.2). Instead, we might actually impose also finite moment conditions of the form

$$\sup_{j \in [N]} \int_{\mathcal{D}} |\psi_j(t)|^p d\nu(t) \leq K_p, \quad 2 \leq p < \infty.$$

A suitable growth condition on the constants K_p should then still allow a probabilistic estimate of $\|\widetilde{A}_S^* \widetilde{A}_S - \mathrm{Id}\|_{2\to 2}$ – possibly with a worse probability decay than in (7.4). Details remain to be worked out.

Let us now turn to the probabilistic estimate of the coherence of the matrix A in (4.4)

Corollary 7.4. *Let $A \in \mathbb{C}^{m \times N}$ be the sampling matrix* (4.4) *associated to an orthonormal system that satisfies the boundedness condition* (4.2) *for some constant $K \geq 1$. Then the coherence of the renormalized matrix $\widetilde{A} = \frac{1}{\sqrt{m}} A$ satisfies*

$$\mu \leq \sqrt{\frac{2\widetilde{C} K^2 \ln(2^{3/4} N^2/\varepsilon)}{m}}$$

with probability at least $1-\varepsilon$ – provided the right hand side is at most $1/2$. The constant is the same as in the previous statement, $\widetilde{C} = 9 + \sqrt{17} \approx 13.12$.

Proof. Let $S = \{j,k\}$ be a two element set. Then the matrix $\widetilde{A}_S^* \widetilde{A}_S - \mathrm{Id}$ contains $\langle \widetilde{\mathbf{a}}_j, \widetilde{\mathbf{a}}_k \rangle$ as a matrix entry. Since the absolute value of any entry of a matrix is bounded by the operator norm of the matrix on ℓ_2, we have

$$|\langle \widetilde{\mathbf{a}}_j, \widetilde{\mathbf{a}}_k \rangle| \leq \|\widetilde{A}_S^* \widetilde{A}_S - \mathrm{Id}\|_{2\to 2}.$$

By Theorem 7.3 the probability that the operator norm on the right is *not* bounded by $\delta \in (0, 1/2]$ is at most

$$2^{3/4} \cdot 2 \exp\left(-\frac{m\delta^2}{\widetilde{C} K^2 \cdot 2}\right).$$

Taking the union bound over all $N(N-1)/2 \leq N^2/2$ two element sets $S \subset [N]$ shows that

$$\mathbb{P}(\mu \geq \delta) \leq 2^{3/4} N^2 \exp\left(-\frac{m\delta^2}{2\widetilde{C}K^2}\right).$$

Requiring that the right hand side is at most ε leads to the desired conclusion. □

7.3 Finishing the proof

Now we complete the proof of Theorem 4.2. Set $\alpha = \frac{\sqrt{st}}{1-\delta}$ for some $t, \delta \in (0, 1/2]$ to be chosen later. Let μ be the coherence of $\widetilde{A} = \frac{1}{\sqrt{m}} A$. By Proposition 7.1 and Proposition 7.2 the probability that recovery by ℓ_1-minimization fails is bounded from above by

$$\begin{aligned} &2^{3/4}(N-s)e^{-\alpha^{-2}/2} + \mathbb{P}\left(\max_{\ell \in [N]\setminus S} \|\widetilde{A}_S^\dagger \widetilde{\mathbf{a}}_\ell\|_2 \geq \alpha\right) \\ &\leq 2^{3/4}(N-s)e^{-\alpha^{-2}/2} + \mathbb{P}(\|\widetilde{A}_S^* \widetilde{A}_S - \mathrm{Id}\|_{2\to 2} > \delta) + \mathbb{P}(\mu > t). \end{aligned} \tag{7.11}$$

By Theorem 7.3 we have $\mathbb{P}(\|\widetilde{A}_S^* \widetilde{A}_S - \mathrm{Id}\|_{2\to 2} > \delta) \leq \varepsilon$ provided

$$m \geq \frac{\widetilde{C}K^2}{\delta^2} s \ln(2^{3/4} s/\varepsilon), \tag{7.12}$$

while Corollary 7.4 asserts that $\mathbb{P}(\mu > t) \leq \varepsilon$ provided

$$m \geq \frac{2\widetilde{C}K^2}{t^2} \ln(2^{3/4} N^2/\varepsilon). \tag{7.13}$$

Set $t = \delta\sqrt{\frac{2}{s}}$. Then (7.13) implies (7.12), and $\alpha = \frac{\delta\sqrt{2}}{1-\delta}$. The first term in (7.11) is then bounded by ε if

$$\delta^{-2} = 4 \ln(2^{3/4} N/\varepsilon).$$

Plugging this into the definition of t and then into (7.13) we find that recovery by ℓ_1-minimization fails with probability at most 3ε provided

$$\begin{aligned} m &\geq \widetilde{C}K^2 s \ln(2^{3/4} N/\varepsilon) \ln(2^{3/4} N^2/\varepsilon) \\ &= \widetilde{C}K^2 s \ln(2^{3/4} N/\varepsilon) \left(\ln(N) + \ln(2^{3/4} N/\varepsilon)\right). \end{aligned}$$

Replacing ε by $\varepsilon/3$, this is satisfied if (4.18) holds with $C = 2\widetilde{C}$. □

8 Proof of Uniform Recovery Result for Bounded Orthonormal Systems

In this chapter we first prove the theorem below concerning the restricted isometry constants δ_s of $\widetilde{A} = \frac{1}{\sqrt{m}}A$, associated to random sampling in bounded orthogonal system, see (4.4). Rudelson and Vershynin have shown an analog result for discrete orthonormal systems in [116]. Later in Section 8.6 we strengthen Theorem 8.1 to Theorem 8.4, which ultimately shows Theorem 4.4.

Theorem 8.1. *Let $A \in \mathbb{C}^{m\times N}$ be the sampling matrix* (4.4) *associated to an orthonormal system that satisfies the boundedness condition* (4.2) *for some constant $K \geq 1$. Assume that the random sampling points $t_1, \dots, t_m$ are chosen independently at random according to the orthogonalization measure ν. Suppose, for some $\varepsilon \in (0,1), \delta \in (0, 1/2]$, that*

$$\frac{m}{\ln(10m)} \geq DK^2\delta^{-2}s\ln^2(100s)\ln(4N)\ln(7\varepsilon^{-1}) \tag{8.1}$$

where the constant $D \leq 243\,150$, then with probability at least $1-\varepsilon$ the restricted isometry constant of the renormalized matrix $\frac{1}{\sqrt{m}}A$ satisfies $\delta_s \leq \delta$.

8.1 Start of Proof

We use the characterization of the restricted isometry constants in Proposition 2.5(b),

$$\delta_s = \max_{S\subset N, |S|\leq s} \|\widetilde{A}_S^*\widetilde{A}_S - \mathrm{Id}\|_{2\to 2}.$$

Let us introduce the set

$$D_{s,N}^2 := \{\mathbf{z}\in\mathbb{C}^N, \|\mathbf{z}\|_2 \leq 1, \|\mathbf{z}\|_0 \leq s\} = \bigcup_{S\subset[N],|S|=s} \mathcal{B}_S^2,$$

where $\mathcal{B}_S^2 = \{\mathbf{z}\in\mathbb{C}^N, \|\mathbf{z}\|_2 \leq 1, \operatorname{supp}\mathbf{z}\subset S\}$. The quantity

$$\|\!|B|\!\|_s := \sup_{\mathbf{z}\in D_{s,N}^2} |\langle B\mathbf{z},\mathbf{z}\rangle|$$

defines a norm on self-adjoint matrices $B = B^* \in \mathbb{C}^{N\times N}$ (a semi-norm on all of $\mathbb{C}^{N\times N}$), and

$$\delta_s = \|\!|\widetilde{A}^*\widetilde{A} - \mathrm{Id}|\!\|_s.$$

Let $\mathbf{X}_\ell = \left(\overline{\psi_j(t_\ell)}\right)_{j=1}^N \in \mathbb{C}^N$ be the random column vector associated to the sampling point t_ℓ, $\ell \in [m]$. Then $\mathbf{X}_\ell^*$ is a row of A. Observe that $\mathbb{E}\mathbf{X}_\ell\mathbf{X}_\ell^* = \mathrm{Id}$ by the orthogonality relation 4.1. We can express the restricted isometry constant of $\widetilde{A}$ as

$$\delta_s = \|\!|\frac{1}{m}\sum_{\ell=1}^m \mathbf{X}_\ell\mathbf{X}_\ell^* - \mathrm{Id}|\!\|_s = \frac{1}{m}\|\!|\sum_{\ell=1}^m (\mathbf{X}_\ell\mathbf{X}_\ell^* - \mathbb{E}\mathbf{X}_\ell\mathbf{X}_\ell^*)|\!\|_s. \tag{8.2}$$

Let us first consider the moments of δ_s. Using symmetrization (Lemma 6.7) we estimate, for $p \geq 1$,

$$\left(\mathbb{E}\|\sum_{\ell=1}^m (\mathbf{X}_\ell \mathbf{X}_\ell^* - \mathbb{E}\mathbf{X}_\ell \mathbf{X}_\ell^*)\|_s^p\right)^{1/p} \leq 2\left(\mathbb{E}\|\sum_{\ell=1}^m \epsilon_\ell \mathbf{X}_\ell \mathbf{X}_\ell^*\|_s^p\right)^{1/p}. \tag{8.3}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_m)$ is a Rademacher sequence, which is independent of the random sampling points t_ℓ, $\ell \in [m]$.

8.2 The Crucial Lemma

The following lemma, which heavily relies on Dudley's inequality, is key to the estimate of the moments in (8.3).

Lemma 8.2. *Let $\mathbf{x}_1, \dots, \mathbf{x}_m$ be vectors in $\mathbb{C}^N$ with $\|\mathbf{x}_\ell\|_\infty \leq K$ for $\ell \in [m]$ and assume $s \leq m$. Then,*

$$\mathbb{E}\|\sum_{\ell=1}^m \epsilon_\ell \mathbf{x}_\ell \mathbf{x}_\ell^*\|_s \leq \tilde{C}_1 K\sqrt{s}\ln(100s)\sqrt{\ln(4N)\ln(10m)}\sqrt{\|\sum_{\ell=1}^m \mathbf{x}_\ell \mathbf{x}_\ell^*\|_s} \tag{8.4}$$

where $\tilde{C}_1 = 94.81$. Furthermore, for $p \geq 2$,

$$\left(\mathbb{E}\|\sum_{\ell=1}^m \epsilon_\ell \mathbf{x}_\ell \mathbf{x}_\ell^*\|_s^p\right)^{1/p}$$
$$\leq \beta^{1/p}\tilde{C}_2\sqrt{p}K\sqrt{s}\ln(100s)\sqrt{\ln(4N)\ln(10m)}\sqrt{\|\sum_{\ell=1}^m \mathbf{x}_\ell \mathbf{x}_\ell^*\|_s}, \tag{8.5}$$

where $\tilde{C}_2 \approx 82.56$ and $\beta = 6.028$ is the constant in Dudley's inequality (6.42).

PROOF. Observe that

$$E_p := \left(\mathbb{E}\|\sum_{\ell=1}^m \epsilon_\ell \mathbf{x}_\ell \mathbf{x}_\ell^*\|_s^p\right)^{1/p} = \left(\mathbb{E}\sup_{\mathbf{u}\in D_{s,N}^2}\left|\sum_{\ell=1}^m \epsilon_\ell |\langle \mathbf{x}_\ell, \mathbf{u}\rangle|^2\right|^p\right)^{1/p}.$$

This is moment of the supremum of a Rademacher process, $X_\mathbf{u} = \sum_{\ell=1} \epsilon_\ell |\langle \mathbf{x}_\ell, \mathbf{u}\rangle|^2$, which has associated pseudo-metric

$$d(\mathbf{u}, \mathbf{v}) = \left(\mathbb{E}|X_\mathbf{u} - X_\mathbf{v}|^2\right)^{1/2} = \sqrt{\sum_{\ell=1}^m \left(|\langle \mathbf{x}_\ell, \mathbf{u}\rangle|^2 - |\langle \mathbf{x}_\ell, \mathbf{v}\rangle|^2\right)^2},$$

see also (6.40). Then, for $\mathbf{u}, \mathbf{v} \in D_{s,N}^2$ we can estimate

$$\begin{aligned} d(\mathbf{u},\mathbf{v}) &= \left(\sum_{\ell=1}^{m} \left(|\langle \mathbf{x}_\ell, \mathbf{u}\rangle| - |\langle \mathbf{x}_\ell, \mathbf{v}\rangle| \right)^2 \left(|\langle \mathbf{x}_\ell, \mathbf{u}\rangle| + |\langle \mathbf{x}_\ell, \mathbf{v}\rangle| \right)^2 \right)^{1/2} \\ &\leq \max_{\ell\in[m]} \left| |\langle \mathbf{x}_\ell, \mathbf{u}\rangle| - |\langle \mathbf{x}_\ell, \mathbf{v}\rangle| \right| \sup_{\mathbf{u},\mathbf{v}\in D_{s,N}^2} \sqrt{\sum_{\ell=1}^{m} (|\langle \mathbf{x}_\ell, \mathbf{u}\rangle| + |\langle \mathbf{x}_\ell, \mathbf{v}\rangle|)^2} \\ &\leq 2R \max_{\ell\in[m]} |\langle \mathbf{x}_\ell, \mathbf{u}-\mathbf{v}\rangle| \ , \end{aligned}$$

where

$$R = \sup_{\mathbf{u}\in D_{s,N}^2} \sqrt{\sum_{\ell=1}^{m} |\langle \mathbf{x}_\ell, \mathbf{u}\rangle|^2} = \sqrt{\left\| \sum_{\ell=1}^{m} \mathbf{x}_\ell \mathbf{x}_\ell^* \right\|_s} \ .$$

We further introduce the auxiliary seminorm

$$\|\mathbf{u}\|_X := \max_{\ell\in[m]} |\langle \mathbf{x}_\ell, \mathbf{u}\rangle| \ , \mathbf{u} \in \mathbb{C}^N \ . \tag{8.6}$$

We derived that the rescaled process $X_{\mathbf{u}}/(2R)$ satisfies

$$\left(\mathbb{E}|X_{\mathbf{u}}/(2R) - X_{\mathbf{v}}/(2R)|^2\right)^{1/2} \leq \|\mathbf{u}-\mathbf{v}\|_X .$$

It follows from Dudley's inequality for Rademacher process, Theorem 6.23 with $t_0 = 0$, that

$$E_p \leq 2\beta^{1/p} R \left(C \int_0^{\Delta(D_{s,N}^2)} \sqrt{\ln(N(D_{s,N}^2, \|\cdot\|_X, t))}\, dt + D\Delta(D_{s,N}^2) \right) \ . \tag{8.7}$$

By the Cauchy-Schwarz inequality, for $\mathbf{u} \in D_{s,N}^2$,

$$\|\mathbf{u}\|_X = \max_{\ell\in[m]} |\langle \mathbf{x}_\ell, \mathbf{u}\rangle| \leq \|\mathbf{u}\|_1 \max_{\ell\in[m]} \|\mathbf{x}_\ell\|_\infty \leq K\sqrt{s}\|\mathbf{u}\|_2 \leq K\sqrt{s} \ . \tag{8.8}$$

Therefore, the diameter $\Delta(D_{s,N}^2)$ in the $\|\cdot\|_X$-norm satisfies

$$\Delta(D_{s,N}^2) = \Delta(D_{s,N}^2, \|\cdot\|_X) \leq 2K\sqrt{s} \ . \tag{8.9}$$

Our next task is to estimate the covering numbers $N(D_{s,N}^2, \|\cdot\|_X, t)$. We will do this in two different ways. One estimate will be good for small values of t and the other one for large values of t. For small values, we introduce the norm

$$\|\mathbf{z}\|_1^* := \sum_{j=1}^{N} (|\operatorname{Re}(z_j)| + |\operatorname{Im}(z_j)|) \ , \mathbf{z} \in \mathbb{C}^N \ ,$$

which is the usual ℓ_1-norm after identification of $\mathbb{C}^N$ with $\mathbb{R}^{2N}$. Then we have the embedding

$$D_{s,N}^2 \subset \sqrt{2s} B_{\|\cdot\|_1^*}^N, \quad \text{where } B_{\|\cdot\|_1^*}^N = \{x \in \mathbb{C}^N, \|x\|_1^* \leq 1\}.$$

The next lemma provides an estimate of the covering numbers of an arbitrary subset of $B_{\|\cdot\|_1^*}^N$.

8.3 Covering Number Estimate

The next lemma provides an estimate of the covering numbers of an arbitrary subset of $B_{\|\cdot\|_1^*}^N$.

Lemma 8.3. *Let U be a subset of $B_{\|\cdot\|_1^*}^N$ and $0 < t < \sqrt{2}K$. Then*

$$\sqrt{\ln(N(U, \|\cdot\|_X, t))} \leq 3K\sqrt{\ln(10m)\ln(4N)}t^{-1}.$$

Proof. Fix $\mathbf{x} \in U$. The idea is to approximate $\mathbf{x}$ by a finite set of very sparse vectors. In order to find a vector $\mathbf{z}$ from this finite set that is close to $\mathbf{x}$ we use the so called empirical method of Maurey. To this end we define a random vector $\mathbf{Z}$ that takes the value $\operatorname{sgn}(\operatorname{Re}(x_j))\mathbf{e}_j$ with probability $|\operatorname{Re}(x_j)|$, the value $i\operatorname{sgn}(\operatorname{Im}(x_j))\mathbf{e}_j$ with probability $|\operatorname{Im}(x_j)|$ for $j = 1, \ldots, N$, and the zero vector 0 with probability $1 - \|x\|_1^*$. Here, $\mathbf{e}_j$ denotes the jth canonical unit vector, $(\mathbf{e}_j)_k = \delta_{j,k}$. Since $\|\mathbf{x}\|_1^* \leq 1$ this is a valid probability distribution. Note that

$$\mathbb{E}\mathbf{Z} = \sum_{j=1}^N \operatorname{sgn}(\operatorname{Re}(x_j))|\operatorname{Re}(x_j)|\mathbf{e}_j + i\sum_{j=1}^N \operatorname{sgn}(\operatorname{Im}(x_j))|\operatorname{Im}(x_j)|\mathbf{e}_j = \mathbf{x}.$$

Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$ be independent copies of $\mathbf{Z}$, where M is a number to be determined later. We attempt to approximate $\mathbf{x}$ with the M-sparse vector

$$\mathbf{z} = \frac{1}{M}\sum_{k=1}^M \mathbf{Z}_k.$$

We estimate the expected distance of $\mathbf{z}$ to $\mathbf{x}$ in $\|\cdot\|_X$ by first using symmetrization (Lemma 6.7),

$$\mathbb{E}\|\mathbf{z} - \mathbf{x}\|_X = \mathbb{E}\|\frac{1}{M}\sum_{k=1}^M (\mathbf{Z}_k - \mathbb{E}\mathbf{Z}_k)\|_X \leq \frac{2}{M}\mathbb{E}\|\sum_{k=1}^M \epsilon_k \mathbf{Z}_k\|_X$$

$$= \frac{2}{M}\mathbb{E}\max_{\ell\in[m]}\left|\sum_{k=1}^M \epsilon_k \langle \mathbf{x}_\ell, \mathbf{Z}_k\rangle\right|,$$

where $\boldsymbol{\epsilon}$ is a Rademacher sequence, which is independent of $(\mathbf{Z}_1, \dots, \mathbf{Z}_M)$. Now we fix a realization of $(\mathbf{Z}_1, \dots, \mathbf{Z}_M)$ and consider only expectation and probability with respect to $\boldsymbol{\epsilon}$ for the moment (that is, conditional on $(\mathbf{Z}_1, \dots, \mathbf{Z}_M)$). Since $\|\mathbf{x}_\ell\|_\infty \le K$ and $\mathbf{Z}_k$ has only a single non-zero component of magnitude 1, we have $|\langle \mathbf{x}_\ell, \mathbf{Z}_k \rangle| \le K$. It follows that

$$\|(\langle \mathbf{x}_\ell, \mathbf{Z}_k \rangle)_{k=1}^M\|_2 \le \sqrt{M} K, \quad \ell \in [m].$$

Using Hoeffding's inequality (Proposition 6.11) we obtain

$$\mathbb{P}_{\boldsymbol{\epsilon}} \left(|\sum_{k=1}^M \epsilon_k \langle \mathbf{x}_\ell, \mathbf{Z}_k \rangle| > K\sqrt{M} u \right) \le \mathbb{P}_{\boldsymbol{\epsilon}} \left(|\sum_{k=1}^M \epsilon_k \langle \mathbf{x}_\ell, \mathbf{Z}_k \rangle| > \|(\langle \mathbf{x}_\ell, \mathbf{Z}_k \rangle)_{k=1}^M\|_2 u \right)$$
$$\le 2e^{-u^2/2}, \quad \text{for all } u > 0, \ell \in [m].$$

Lemma 6.6 yields

$$\mathbb{E} \max_{\ell \in [m]} |\sum_{k=1}^M \epsilon_k \langle \mathbf{x}_\ell, \mathbf{Z}_k \rangle| \le CK\sqrt{M}\sqrt{\ln(8m)} \tag{8.10}$$

with $C = \sqrt{2} + \frac{1}{4\sqrt{2}\ln(8)} \approx 1.499 < 1.5$. By Fubini's theorem we finally obtain

$$\mathbb{E}\|\mathbf{z} - \mathbf{x}\|_X \le \frac{2}{M} \mathbb{E}_{\mathbf{Z}} \mathbb{E}_{\boldsymbol{\epsilon}} \max_{\ell \in [m]} \left| \sum_{k=1}^M \epsilon_k \langle \mathbf{x}_\ell, \mathbf{Z}_k \rangle \right| \le \frac{3K}{\sqrt{M}} \sqrt{\ln(8m)} \, .$$

This implies that there exists a vector of the form

$$\mathbf{z} = \frac{1}{M} \sum_{k=1}^M \mathbf{z}_k, \tag{8.11}$$

where each $\mathbf{z}_k$ is one of the vectors in $\{\pm \mathbf{e}_j, \pm i\mathbf{e}_j : j \in [N]\}$, such that

$$\|\mathbf{z} - \mathbf{x}\|_X \le \frac{3K}{\sqrt{M}} \sqrt{\ln(8m)}. \tag{8.12}$$

(Note that $\mathbf{z}$ has sparsity at most M.) In particular,

$$\|\mathbf{z} - \mathbf{x}\|_X \le t \tag{8.13}$$

provided

$$\frac{3K}{\sqrt{M}} \sqrt{\ln(8m)} \le t. \tag{8.14}$$

Each $\mathbf{z}_k$ takes $4N + 1$ values, so that $\mathbf{z}$ can take at most $(4N + 1)^M$ values. (Actually, it takes strictly less than $(4N)^M$ values, since if some $\mathbf{e}_j$ appears more than once in

the sum, then it always appears with the same sign.) For each $\mathbf{x} \in U$ we can therefore find a vector $\mathbf{z}$ of the form (8.11) such that $\|\mathbf{x} - \mathbf{z}\|_X \leq t$. The choice

$$M = \left\lfloor \frac{9K^2}{t^2} \ln(10m) \right\rfloor$$

satisfies (8.14). Indeed, then

$$\begin{aligned} M &\geq \frac{9K^2}{t^2} \ln(10m) - 1 = \frac{9K^2}{t^2} \ln(8m) + \frac{9K^2 \ln(10/8)}{t^2} - 1 \\ &\geq \frac{9K^2}{t^2} \ln(8m) + \frac{9 \ln(10/8)}{2} - 1 \geq \frac{9K^2}{t^2} \ln(8m) \end{aligned}$$

since $t \leq \sqrt{2}K$ and $\frac{9\ln(10/8)}{2} > 1$. Therefore, (8.14) is satisfied. We deduce that the covering numbers can be estimated by

$$\begin{aligned} &\sqrt{\ln(N(U, \|\cdot\|_X, t))} \leq \sqrt{\ln((4N)^M)} \leq \sqrt{\left\lfloor \frac{9K^2}{t^2} \ln(10m) \right\rfloor \ln(4N)} \\ &\leq 3K \sqrt{\ln(10m) \ln(4N)} t^{-1}, \end{aligned}$$

This completes the proof of the lemma. □

8.4 Finishing the Proof of the Crucial Lemma

The estimate of the covering number in the lemma of the previous section will be good for larger values of t. For small values of t we use a volumetric argument. To this end, we observe that

$$D_{s,N}^2 \subset \sqrt{s} D_{s,N}^1 := \{\mathbf{x} \in \mathbb{C}^N, \|\mathbf{x}\|_1 \leq 1, \|\mathbf{x}\|_0 \leq s\} = \bigcup_{|S|=s} \mathcal{B}_1^S \,,$$

where $\mathcal{B}_1^S = \{\mathbf{x} \in \mathbb{C}^N, \|\mathbf{x}\|_1 \leq 1, \operatorname{supp} \mathbf{x} \subset S\}$. The estimate (8.8) reveals that $\|\mathbf{u}\|_X \leq K\|\mathbf{u}\|_1$, so that

$$\mathcal{B}_1^S \subset K\mathcal{B}_X^S = \{\mathbf{x} \in \mathbb{C}^N, \|\mathbf{x}\|_X \leq K, \operatorname{supp} \mathbf{x} \subset S\} \,.$$

It follows from Proposition 10.1 after identifying $\mathbb{C}^s$ with $\mathbb{R}^{2s}$ that

$$N(\mathcal{B}_1^S, \|\cdot\|_X, t) \leq N(\mathcal{B}_X^S, \|\cdot\|_X, t/K) \leq (1 + 2K/t)^{2s} \,.$$

There are

$$\binom{N}{s} = \frac{N(N-1)\cdots(N-s+1)}{s!} \leq \frac{N^s}{s!} = \frac{s^s}{s!} \frac{N^s}{s^s} \leq e^s \frac{N^s}{s^s} = \left(\frac{eN}{s}\right)^s$$

subsets of $[N]$ of cardinality s. Hence, by subadditivity of the covering numbers we obtain

$$N(D_{s,N}^2, \|\cdot\|_X, t) \leq \sum_{|S|=s} N(\mathcal{B}_S^1, \|\cdot\|_X, t/\sqrt{s}) \leq (eN/s)^{2s}(1+2K\sqrt{s}/t)^{2s} .$$

Together with Lemma 8.3, and noting that $D_{s,N}^2 \subset \sqrt{2s}B_{\|\cdot\|_1^*}^N$, we get the two bounds

$$\begin{aligned}
\sqrt{\ln(N(D_{s,N}^2, \|\cdot\|_X, t))} &\leq 3K\sqrt{2s}\sqrt{\ln(10m)\ln(4N)}t^{-1}, \quad 0 < t \leq 2K\sqrt{s}, \\
\sqrt{\ln(N(D_{s,N}^2, \|\cdot\|_X, t))} &\leq \sqrt{2s}\sqrt{\ln(eN/s) + \ln(1+2K\sqrt{s}/t)} \\
&\leq \sqrt{2s}\left(\sqrt{\ln(eN/s)} + \sqrt{\ln(1+2K\sqrt{s}/t)}\right), \quad t > 0 .
\end{aligned}$$

Next we combine these inequalities to estimate the "Dudley integral". We obtain, for $\kappa > 0$, noting also that $\Delta(D_{s,N}^2) \leq 2K\sqrt{s}$ by (8.9),

$$\begin{aligned}
I &:= \int_0^{\Delta(D_{s,N}^2)} \sqrt{\ln(N(D_{s,N}^2, \|\cdot\|_X, t)}dt \\
&\leq \sqrt{2s}\int_0^{\kappa} \sqrt{\ln(eN/s)} + \sqrt{\ln\left(1+2K\sqrt{s}t^{-1}\right)}dt \\
&\quad + 3K\sqrt{2s}\sqrt{\ln(10m)\ln(4N)}\int_{\kappa}^{2K\sqrt{s}} t^{-1}dt \\
&\leq \kappa\sqrt{2s}\sqrt{\ln(eN/s)} + 2\sqrt{2}Ks\int_0^{\kappa/(2K\sqrt{s})} \sqrt{\ln(1+u^{-1})}du \\
&\quad + 3K\sqrt{2s}\sqrt{\ln(10m)\ln(4N)}\ln(2K\sqrt{s}/\kappa) \\
&\leq \kappa\sqrt{2s}\left(\sqrt{\ln(eN/s)} + \sqrt{\ln(e(1+2K\sqrt{s}/\kappa))}\right) \\
&\quad + 3K\sqrt{2s}\sqrt{\ln(10m)\ln(4N)}\ln(2K\sqrt{s}/\kappa). \qquad (8.15)
\end{aligned}$$

In the last step we have applied Lemma 10.3. The choice $\kappa = K/5$ yields

$$\begin{aligned}
I &\leq \frac{K}{5}\sqrt{2s}\left(\sqrt{\ln(eN/s)} + \sqrt{\ln\left(e\left(1+10\sqrt{s}\right)\right)}\right) \\
&\quad + 3K\sqrt{2s}\sqrt{\ln(10m)\ln(4N)}\ln\left(\sqrt{100s}\right) \\
&\leq C_0K\sqrt{s}\sqrt{\ln(10m)\ln(4N)}\ln(100s) .
\end{aligned}$$

where $C_0 = \frac{\sqrt{2}}{5} + \frac{1}{\sqrt{\ln(100)}} + \frac{3}{\sqrt{2}} \approx 2.87$. Hereby, we applied the inequality

$$\begin{aligned}\sqrt{\ln(e(1+20\sqrt{s}))} &\leq \sqrt{\ln(100s)/2 + \ln(11e/10)} \\ &\leq \frac{1}{\sqrt{2\ln(100)}} \ln(100s) \sqrt{\ln(11e/10)} \\ &\leq \frac{1}{\sqrt{2\ln(100)}} \ln(100s) \sqrt{\ln(10m)\ln(4N)}.\end{aligned}$$

Plugging the above estimate and (8.9) into (8.7) yields

$$\begin{aligned}E_p &\leq 2\beta^{1/p}\sqrt{s}\left(C_0 C K \sqrt{\ln(10m)\ln(4N)}\ln(100s) + 2DK\right) R \\ &\leq \beta^{1/p}\widetilde{C}_2\sqrt{s}\sqrt{\ln(10m)\ln(4N)}\ln(100s)\, R,\end{aligned}$$

where, for $p \geq 2$, (and $N, m \geq 2$),

$$\widetilde{C} = \tilde{C}_2 = 2C_0 C + \frac{4D}{C_0 C\sqrt{\ln(20)\ln(8)}\ln(100)} \approx 82.56.$$

For the case $p = 1$ we can use the slight better constants C_1 and D_1 in Dudley's inequality (6.41) to obtain

$$\widetilde{C} = \tilde{C}_1 = 2C_0 C_1 + \frac{4D_1}{C_1 C\sqrt{\ln(20)\ln(8)}\ln(100)} \approx 94.81.$$

The proof of Lemma 8.2 is completed.

8.5 Completing the Proof of Theorem 8.1

We proceed similarly as in Section 7.2. Denote, for $p \geq 2$,

$$E_p := (\mathbb{E}\delta_s^p)^{1/p} = \left(\mathbb{E}\|\frac{1}{m}\sum_{\ell=1}^m \mathbf{X}_\ell \mathbf{X}_\ell^* - \mathrm{Id}\|_s^p\right)^{1/p}.$$

Then (8.3) together with Lemma (8.2) yields

$$\begin{aligned}E_p^p &\leq \left(\frac{2D_{N,m,s,p}}{\sqrt{m}}\right)^p \mathbb{E}\|\frac{1}{m}\sum_{\ell=1}^m \mathbf{X}_\ell \mathbf{X}_\ell^*\|_s^{p/2} \\ &\leq \left(\frac{2D_{N,m,s,p}}{\sqrt{m}}\right)^p \mathbb{E}\left(\|\frac{1}{m}\sum_{\ell=1}^m \mathbf{X}_\ell \mathbf{X}_\ell^* - \mathrm{Id}\|_s + 1\right)^{p/2}, \qquad (8.16)\end{aligned}$$

where

$$D_{N,m,s,p} = \beta^{1/p}\tilde{C}_2\sqrt{p}K\sqrt{s}\ln(100s)\sqrt{\ln(4N)\ln(10m)}$$

Using the triangle inequality we conclude that

$$E_p \le \frac{2D_{N,m,s,p}}{\sqrt{m}}\sqrt{E_p+1}.$$

Proceeding in the same way as in Section 7.2, see (7.8), and setting $\kappa = \frac{\sqrt{17}+1}{4}$ yields

$$\begin{aligned}(\mathbb{E}\min\{1/2,\delta_s\}^p)^{1/p} &\le \frac{2\kappa D_{N,m,s,p}}{\sqrt{m}} \\ &= \beta^{1/p}2\kappa\tilde{C}_2\sqrt{p}\sqrt{\frac{s}{m}}\sqrt{s}\ln(100s)\sqrt{\ln(4N)\ln(10m)}, \quad p\ge 2.\end{aligned} \tag{8.17}$$

Proposition (6.5) shows that for all $u \ge 2$,

$$\mathbb{P}\left(\min\{1/2,\delta_s\} \ge 2\kappa e^{1/2}\tilde{C}_2\sqrt{\frac{s}{m}}\ln(100s)\sqrt{\ln(4N)\ln(10m)}\ u\right) < 7e^{-u^2/2},$$

where we used that $\beta < 7$. Expressed differently, $\delta_s \le \delta \le 1/2$ with probability at least $1-\varepsilon$ provided

$$m \ge D\delta^{-2}s\ln^2(100s)\ln(4N)\ln(10m)\ln(7\varepsilon^{-1})$$

with $D = 2(2\kappa e^{1/2}\tilde{C}_2)^2 \approx 243\,150$.

8.6 Strengthening the Probability Estimate

In this section we slightly improve on Theorem 8.1. The next theorem immediately implies Theorem 4.4 by noting Theorems 2.6 and 2.7. Its proof uses the deviation inequality of Section 6.10.

Theorem 8.4. *Let A be the random sampling matrix* (4.4) *associated to random sampling in a bounded orthonormal system obeying* (4.2) *with some constant $K \ge 1$. Let $\varepsilon \in (0,1), \delta \in (0,1/2]$. If*

$$\begin{aligned}\frac{m}{\ln(10m)} &\ge C\delta^{-2}K^2 s\ln^2(100s)\ln(4N), \\ m &\ge D\delta^{-2}K^2 s\ln(\varepsilon^{-1}),\end{aligned} \tag{8.18}$$

then with probability at least $1-\varepsilon$ the restricted isometry constant δ_s of $\frac{1}{\sqrt{m}}A$ satisfies $\delta_s \le \delta$. The constants satisfy $C \le 50\,963$ and $D \le 456$.

Proof. Set $E = \mathbb{E}\delta_s$. Using Lemma 8.2 for $p=1$ and proceeding similarly as in the preceding section we obtain

$$E \le \frac{2D_{N,m,s,1}}{\sqrt{m}}\sqrt{E+1} = G_{N,m,s}\sqrt{E+1}$$

with

$$G_{N,m,s} = C' \sqrt{\frac{s}{m}} \ln(100s) \sqrt{\ln(10m)\ln(4N)}$$

and $C' = 2\tilde{C}_1$. It follows from (7.7) that, if

$$G_{N,m,s} \leq \sigma\delta, \qquad \text{with } \sigma := 0.84 \tag{8.19}$$

for $\delta \leq 1/2$, then

$$E \leq \mathbb{E}\delta_s < 8\delta/9.$$

It remains to show that δ_s does not deviate much from its expectation with high probability. To this end we use the deviation inequality of Theorem 6.25. By definition of the norm $|||\cdot|||_s$ we can write

$$\begin{aligned} m\delta_s &= |||\sum_{\ell=1}^m (\mathbf{X}_\ell \mathbf{X}_\ell^* - \mathrm{Id})|||_s = \sup_{S\subset[N],|S|\leq s} \|\sum_{\ell=1}^m (\mathbf{X}_\ell^S (\mathbf{X}_\ell^S)^* - \mathrm{Id}_S)\|_{2\to 2} \\ &= \sup_{(z,w)\in Q_{s,N}^2} \mathrm{Re}\left(\langle \sum_{\ell=1}^m (\mathbf{X}_\ell \mathbf{X}_\ell^* - \mathrm{Id})\mathbf{z}, \mathbf{w}\rangle \right) \\ &= \sup_{(z,w)\in Q_{s,N}^{2,*}} \sum_{\ell=1}^m \mathrm{Re}\left(\left\langle \sum_{\ell=1}^m (\mathbf{X}_\ell \mathbf{X}_\ell^* - \mathrm{Id})\mathbf{z}, \mathbf{w} \right\rangle \right), \end{aligned}$$

where $\mathbf{X}_\ell^S$ denotes the vector $\mathbf{X}_\ell$ restricted to the entries in S, and

$$Q_{s,N}^2 = \bigcup_{S\subset[N],|S|\leq s} Q_{S,N},$$

where $Q_{S,N} = \{(\mathbf{z},\mathbf{w}) : \mathbf{z},\mathbf{w} \in \mathbb{C}^N, \|\mathbf{z}\|_2 = \|\mathbf{w}\|_2 = 1, \operatorname{supp}\mathbf{z}, \operatorname{supp}\mathbf{w} \subset S\}$. Further, let $Q_{s,N}^{2,*}$ denote a dense countable subset of $Q_{s,N}^2$. Introducing $f_{\mathbf{z},\mathbf{w}}(\mathbf{X}) = \mathrm{Re}(\langle(\mathbf{X}\mathbf{X}^* - \mathrm{Id})\mathbf{z},\mathbf{w}\rangle)$ we therefore have

$$m^{-1}\delta_s = \sup_{(\mathbf{z},\mathbf{w})\in Q_{s,N}^{2,*}} \sum_{\ell=1}^m f_{\mathbf{z},\mathbf{w}}(\mathbf{X}_\ell).$$

Since $\mathbb{E}\mathbf{X}\mathbf{X}^* = \mathrm{Id}$ it follows that $f_{\mathbf{z},\mathbf{w}}(\mathbf{X}) = 0$. Let us check the boundedness of $f_{\mathbf{z},\mathbf{w}}$ for $(\mathbf{z},\mathbf{w}) \in Q_{S,N}$ with $|S| \leq s$,

$$\begin{aligned} |f_{\mathbf{z},\mathbf{w}}(\mathbf{X})| &\leq |\langle(\mathbf{X}\mathbf{X}^* - \mathrm{Id})\mathbf{z},\mathbf{w}\rangle| \leq \|\mathbf{z}\|_2\|\mathbf{w}\|_2 \|\mathbf{X}^S(\mathbf{X}^S)^* - \mathrm{Id}_S\|_{2\to 2} \\ &\leq \|\mathbf{X}^S(\mathbf{X}^S)^* - \mathrm{Id}_S\|_{1\to 1} = \max_{j\in S} \sum_{k\in S} |\psi_j(t)\overline{\psi_k(t)} - \delta_{j,k}| \\ &\leq sK^2 \end{aligned}$$

by the boundedness condition (4.2). Hereby, we used that the operator norm on ℓ_2 is bounded by the one on ℓ_1 for self-adjoint matrices, see (2.3) as well as the explicit expression (2.1) for $\|\cdot\|_{1\to 1}$. For the variance term σ^2 we estimate

$$\begin{aligned}
&\mathbb{E}|f_{\mathbf{z},\mathbf{w}}(\mathbf{X}_\ell)|^2 \le \mathbb{E}|\langle(\mathbf{X}\mathbf{X}^* - \mathrm{Id})\mathbf{z},\mathbf{w}\rangle|^2 = \mathbb{E}\mathbf{w}^*(\mathbf{X}\mathbf{X} - \mathrm{Id})\mathbf{z}((\mathbf{X}\mathbf{X}^* - \mathrm{Id})\mathbf{z})^*\mathbf{w}\\
&\le \|\mathbf{w}\|_2^2\mathbb{E}\|(\mathbf{X}\mathbf{X} - \mathrm{Id})\mathbf{z}((\mathbf{X}\mathbf{X}^* - \mathrm{Id})\mathbf{z})^*\|_{2\to 2} = \mathbb{E}\|(\mathbf{X}\mathbf{X}^* - \mathrm{Id})\mathbf{z}\|_2^2\\
&= \mathbb{E}\left[\|\mathbf{X}\|_2^2|\langle\mathbf{X},\mathbf{z}\rangle|^2\right] - 2\mathbb{E}|\langle\mathbf{X},\mathbf{z}\rangle|^2 + 1.
\end{aligned}$$

Hereby we used that $\|\mathbf{u}\mathbf{u}^*\|_{2\to 2} = \|\mathbf{u}\|_2^2$. Observe that $\|X\|_2 \le \sqrt{s}K$ by the Cauchy Schwarz inequality and the boundedness condition (4.2). Furthermore,

$$\mathbb{E}|\langle\mathbf{X},z\rangle|^2 = \sum_{j,k\in S} z_j\overline{z_k}\mathbb{E}[\psi_k(t)\overline{\psi_j(t)}] = \|\mathbf{z}\|_2^2 = 1$$

by orthogonality (4.1). Hence,

$$\begin{aligned}
\mathbb{E}|f_{\mathbf{z},\mathbf{w}}(\mathbf{X}_\ell)|^2 &\le \mathbb{E}\left[\|X\|_2^2|\langle\mathbf{X},\mathbf{z}\rangle|^2\right] - 2\mathbb{E}|\langle\mathbf{X},\mathbf{z}\rangle|^2 + 1 \le (sK^2-2)\mathbb{E}|\langle\mathbf{X},\mathbf{z}\rangle|^2 + 1\\
&= sK^2 - 1 < sK^2.
\end{aligned}$$

Now we are prepared to apply Theorem 6.25. Under condition (8.19) it gives

$$\begin{aligned}
&\mathbb{P}(\delta_s \ge \delta) \le \mathbb{P}(\delta_s \ge \mathbb{E}\delta_s + \delta/9)\\
&= \mathbb{P}(\|\!|\sum_{\ell=1}^m(\mathbf{X}_\ell\mathbf{X}_\ell^* - \mathrm{Id})\|\!|_s \ge \mathbb{E}\|\!|\sum_{\ell=1}^m(\mathbf{X}_\ell\mathbf{X}_\ell^* - \mathrm{Id})\|\!|_s + \delta m/9)\\
&\le \exp\left(-\frac{(\delta m/9)^2}{2msK^2 + 4(8\delta/9)m + 2(\delta m/9)/3}\right)\\
&= \exp\left(-\frac{m\delta^2}{sK^2}\frac{1}{9^2(2 + 4\frac{8\delta}{9sK^2} + \frac{2\delta}{3\cdot 9sK^2})}\right) \le \exp\left(-\frac{m\delta^2}{sK^2}\frac{1}{162 + 9\cdot 32 + 6}\right)\\
&= \exp\left(-\frac{m\delta^2}{456sK^2}\right).
\end{aligned}$$

In the third line, it was used a second time that $\mathbb{E}\|\!|\sum_{\ell=1}^m(\mathbf{X}_\ell\mathbf{X}_\ell^* - \mathrm{Id}\|\!|_s = m\mathbb{E}\delta_s \le m8\delta/9$. Also, note that $\delta/(sK^2) < 1$. It follows that $\delta_s \le \delta$ with probability at least $1-\varepsilon$ provided

$$m \ge 456\,\delta^{-2}K^2 s\ln(\varepsilon^{-1}).$$

Taking also (8.19) into account, we proved that $\delta_s \le \delta$ with probability at least $1-\varepsilon$ provided that m satisfies the two conditions

$$\begin{aligned}
\frac{m}{\ln(10m)} &\ge C\delta^{-2}K^2 s\ln^2(100s)\ln(4N),\\
m &\ge 456\,\delta^{-2}K^2 s\ln(\varepsilon^{-1}).
\end{aligned}$$

with $C = \sigma^{-2}(C')^2 = 4\sigma^{-2}\tilde{C}_1^2 < 50\,963$. □

8.7 Notes

The estimates of the restricted isometry constants are somewhat related to the Λ_1-set problem [11, 12], where one aims at selecting a subset of characters (or bounded orthonormal functions), such that all their linear combinations have comparable L_1 and L_2-norms, up to a logarithmic factor, see [124, 66]. The paper [66] considers also the more involved problem of providing a Kashin splitting of a set of bounded orthonormal functions. It is interesting to note that the analysis in [66] also uses the norm $\|\cdot\|_X$ introduced in (8.6).

9 Proof of Recovery Theorem for Partial Circulant Matrices

The proof of Theorem 5.1 is based on Proposition 7.2, which requires to estimate the coherence of $A = \frac{1}{\sqrt{m}}\Phi^{\Theta}(\mathbf{b})$ and to provide a probabilistic estimate of $\|A_S^* A_S - \mathrm{Id}\|_{2\to 2}$, where $S = \mathrm{supp}(\mathbf{x})$. We start with the coherence estimate.

9.1 Coherence

The proof of the following coherence bound uses similar ideas as the one of Theorem 5.1 in [96].

Proposition 9.1. *Let μ be the coherence of the partial random circulant matrix $A = \frac{1}{\sqrt{m}}\Phi^{\Theta}(\boldsymbol{\epsilon}) \in \mathbb{R}^{m\times N}$, where $\boldsymbol{\epsilon}$ is a Rademacher sequence and $\Theta \subset [N]$ has cardinality m. For convenience assume that m is divisible by three. Then with probability at least $1-\varepsilon$ the coherence satisfies*

$$\mu \leq \sqrt{\frac{6\log(3N^2/\varepsilon)}{m}}. \tag{9.1}$$

Proof. The inner product between two different columns $\mathbf{a}_\ell$, $\mathbf{a}_k$, $\ell \neq k$, of A can be written

$$\langle \mathbf{a}_\ell, \mathbf{a}_k\rangle = \frac{1}{m}\sum_{j\in\Theta} \epsilon_{\ell-j}\epsilon_{k-j},$$

where here and in the following $\ell - j$ and $k-j$ is understood modulo N. The random variables $\tilde{\epsilon}_j = \epsilon_{\ell-j}\epsilon_{k-j}$, $j\in\Theta$, are again Rademacher variables by independence of the ϵ_j and since $\ell \neq k$. We would like to apply Hoeffding's inequality, but unfortunately the $\tilde{\epsilon}_j$, $j\in\Theta$, are not independent in general. Nevertheless, we claim that we can always partition the $\tilde{\epsilon}_j$, $j\in\Theta$ into three sets $\Theta_1, \Theta_2, \Theta_3$ of cardinality $m/3$, such that for each Θ_i the corresponding family $\{\tilde{\epsilon}_j, j\in\Theta_i\}$ forms a sequence of independent Rademacher variables. To this end consider the sets $G_j = \{\ell - j, k-j\}$, $j\in\Theta$, and let

$$H = \{j\in\Theta : \exists j' \in \Theta \text{ such that } G_j \cap G_{j'} \neq \emptyset\}.$$

The random variables $\tilde{\epsilon}_j$, $j \in H$, are not independent. In order to construct the desired splitting into three sets, consider the graph with vertices $j \in \Theta$. The graph contains an edge between j and j' if and only if $G_j \cap G_{j'} \neq \emptyset$. Since any $r \in \Theta$ can be contained in at most two sets G_j (once as $\ell - j$ and once as $k - j$), this graph has degree at most 2. The independence problems are caused by the connected components of the graph. In order to start with the construction of the three sets $\Theta_1, \Theta_2, \Theta_3$ we choose a connected component of the graph, and then one of its endpoints (that is, a vertex that is only connected to one other vertex). If the connected component is a cycle then we choose an arbitrary vertex as starting point. We then move along the connected component, and add the starting vertex j to Θ_1, the second vertex to Θ_2, the third to Θ_3, the fourth to Θ_1 etc. If the connected component is actually a cycle, and if the last vertex was to be added to Θ_1, then we add it to Θ_2 instead. It is easily seen that after dealing with the first connected component of the graph by this process, the random variables $\{\tilde{\epsilon}_j, j \in \Theta_i\}$, are independent for each $i = 1, 2, 3$. All random variables $\tilde{\epsilon}_j$ with j not being a member of the first connected component are independent of the already treated ones. So we may repeat this process with the next connected component in the same way, starting now with a Θ_i satisfying $|\Theta_i| \leq |\Theta_j|$, $j \in \{1, 2, 3\} \setminus \{i\}$. After going through all connected components in this way, we add each element of $\Theta \setminus H$ arbitrarily to one of the Θ_i, such that at the end $|\Theta_i| = m/3$ for $i = 1, 2, 3$. By construction, the random variables $\tilde{\epsilon}_j$, $j \in \Theta_i$, are independent for each $i = 1, 2, 3$. By the triangle inequality, the union bound, and Hoeffding's inequality (6.16) we obtain

$$\begin{aligned}
\mathbb{P}(|\langle \mathbf{a}_\ell, \mathbf{a}_k \rangle| \geq u) &\leq \sum_{i=1}^{3} \mathbb{P}\left(\frac{1}{m} |\sum_{j \in \Theta_i} \tilde{\epsilon}_j| \geq u/3 \right) \\
&= \sum_{i=1}^{3} \mathbb{P}\left(|\sum_{j \in \Theta_i} \tilde{\epsilon}_j| \geq \sqrt{|\Theta_i|} \frac{um}{3\sqrt{|\Theta_i|}} \right) \leq 2 \sum_{i=1}^{3} \exp\left(-\frac{u^2 m^2}{18|\Theta_i|} \right) \\
&\leq 6 \exp\left(-\frac{u^2 m}{6} \right). \qquad (9.2)
\end{aligned}$$

Taking the union bound over all $N(N-1)/2$ possible pairs $\{\ell, k\} \subset [N]$ we get the coherence bound

$$\mathbb{P}(\mu \geq u) \leq 3N(N-1)e^{-u^2 m/6}.$$

This implies that the coherence satisfies $\mu \leq u$ with probability at least $1 - \varepsilon$ provided

$$m \geq \frac{6}{u^2} \ln(3N^2/\varepsilon).$$

Yet another reformulation is the statement of the proposition. □

Note that (9.1) is a slight improvement with respect to Proposition III.2 in [105]. It implies a non-optimal estimate for the restricted isometry constants of A.

Corollary 9.2. *The restricted isometry constant of the renormalized partial random circulant matrix $A \in \mathbb{R}^{m\times N}$ (with m divisible by three) satisfies $\delta_s \leq \delta$ with probability exceeding $1-\varepsilon$ provided*

$$m \geq 6\delta^{-2}s^2 \ln(3N^2/\varepsilon).$$

Proof. Combine Proposition 9.1 with Proposition 2.10(c). □

9.2 Conditioning of Submatrices

Our key estimate, which will be presented next, is mainly based on the noncommutative Khintchine inequality for Rademacher chaos, Theorem 6.22.

Theorem 9.3. *Let $\Theta, S \subset [N]$ with $|\Theta| = m$ and $|S| = s \in \mathbb{N}$. Let $\boldsymbol{\epsilon} \in \mathbb{R}^N$ be a Rademacher sequence. Denote $A = \frac{1}{\sqrt{m}}\Phi^\Theta(\boldsymbol{\epsilon})$ and assume, for $\varepsilon \in (0, 1/2], \delta \in (0,1)$,*

$$m \geq 16\delta^{-2}s \ln^2(2^{5/2}s^2/\varepsilon), \tag{9.3}$$

Then with probability at least $1-\varepsilon$ it holds $\|A_S^*A_S - \mathrm{Id}\|_{2\to 2} \leq \delta$.

Proof. Let us denote $H_S = A_S^*A_S - \mathrm{Id}_S$. We introduce the elementary shift operators on $\mathbb{R}^N$,

$$(T_j\mathbf{x})_\ell = x_{\ell-j \mod N}, \quad j = 1, \dots, N.$$

Further, denote by $R_\Theta : \mathbb{C}^N \to \mathbb{C}^\Theta$ the operator that restricts a vector to the indices in Θ. Then we can write

$$\Phi^\Theta(\boldsymbol{\epsilon}) = R_\Theta \sum_{j=1}^N \epsilon_j T_j. \tag{9.4}$$

We introduce $R_S^* : \mathbb{C}^S \to \mathbb{C}^N$ to be the extension operator that fills up a vector in $\mathbb{C}^S$ with zeros outside S. Observe that

$$\begin{aligned} A_S^*A_S &= \frac{1}{m}\sum_{j=1}^N \epsilon_j R_S T_j^* R_\Theta^* \sum_{k=1}^N \epsilon_k R_\Theta T_k R_S^* \\ &= \frac{1}{m}\sum_{\substack{j,k=1\\ j\neq k}}^N \epsilon_j\epsilon_k R_S T_j^* P_\Theta T_k R_S^* + \frac{1}{m} R_S \left(\sum_{j=1}^N T_j^* P_\Theta T_j\right) R_S^*, \end{aligned}$$

where $P_\Theta = R_\Theta^* R_\Theta$ denotes the projection operator which cancels all components of a vector outside Θ. It is straightforward to check that

$$\sum_{j=1}^N T_j^* P_\Theta T_j = m\mathrm{Id}_N, \tag{9.5}$$

where Id_N is the identity on $\mathbb{C}^N$. Since $R_S R_S^* = \mathrm{Id}_S$ we obtain

$$H_S = \frac{1}{m} \sum_{j \neq k} \epsilon_j \epsilon_k R_S T_j^* P_\Theta T_k R_S^* = \frac{1}{m} \sum_{j \neq k} \epsilon_j \epsilon_k B_{j,k}$$

with $B_{j,k} = R_S T_j^* P_\Theta T_k R_S^*$. Our goal is to apply the noncommutative Khintchine inequality for decoupled Rademacher chaos, Theorem 6.22. To this end we first observe that by (9.5)

$$\sum_{j=1}^N B_{j,k}^* B_{j,\ell} = R_S T_k^* P_\Theta \left(\sum_{j=1}^N T_j P_S T_j^* \right) P_\Theta T_\ell R_S^* = s R_S T_k^* P_\Theta T_\ell R_S^*.$$

Using (9.5) once more this yields

$$\sum_{j,k=1}^N B_{j,k}^* B_{j,k} = s R_S \left(\sum_{k=1}^N T_k^* P_\Theta T_k \right) R_S^* = sm R_S R_S^* = sm \mathrm{Id}_S.$$

Since the entries of all matrices $B_{j,k}$ are non-negative we get

$$\begin{aligned} \|(\sum_{j \neq k} B_{j,k}^* B_{j,k})^{1/2}\|_{S_{2n}}^{2n} &= \mathrm{Tr} \left(\sum_{j \neq k} B_{j,k}^* B_{j,k} \right)^n \\ &\leq \mathrm{Tr} \left(\sum_{j,k} B_{j,k}^* B_{j,k} \right)^n = \mathrm{Tr} \, (sm \mathrm{Id}_S)^n = s^{n+1} m^n. \end{aligned}$$

Furthermore, since $B_{j,k}^* = B_{k,j}$ we have $\sum_{j \neq k} B_{j,k}^* B_{j,k} = \sum_{j \neq k} B_{j,k} B_{j,k}^*$. Let F denote the block matrix $F = (\tilde{B}_{j,k})_{j,k}$ where $\tilde{B}_{j,k} = B_{j,k}$ if $j \neq k$ and $\tilde{B}_{j,j} = 0$. Expressing the product $(F^* F)^n$ as multiple sums over the block-components $\tilde{B}_{j,k}$ and applying the trace yields

$$\begin{aligned} &\|F\|_{S_{2n}}^{2n} = \mathrm{Tr} \left[(F^* F)^n \right] \\ &= \mathrm{Tr} \left[\sum_{\substack{j_1, j_2, \ldots, j_n = 1 \\ k_1, k_2, \ldots, k_n = 1}}^N \tilde{B}_{j_1,k_1}^* \tilde{B}_{j_1,k_2} \tilde{B}_{j_2,k_2}^* \tilde{B}_{j_2,k_3} \cdots \tilde{B}_{j_n,k_n}^* \tilde{B}_{j_n,k_1} \right] \\ &\leq \mathrm{Tr} \sum_{k_1, \ldots, k_n = 1}^N \left[\sum_{j_1=1}^N B_{j_1,k_1}^* B_{j_1,k_2} \cdots \sum_{j_n=1}^N B_{j_n,k_n}^* B_{j_n,k_1} \right] \\ &= s^n \, \mathrm{Tr} \sum_{k_1, \ldots, k_n = 1}^N \left[R_S T_{k_1}^* P_\Theta T_{k_2} R_S^* R_S T_{k_2}^* P_\Theta T_{k_3} R_S^* \cdots R_S T_{k_n}^* P_\Theta T_{k_1} R_S^* \right], \end{aligned}$$

where we applied also (9.5) once more. In the inequality step we used again that the entries of all matrices are non-negative. Using the cyclicity of the trace and applying (9.5) another time, together with the fact that $T_k = T^*_{-k \mod N}$, gives

$$\begin{aligned}\|F\|_{S_{2n}}^{2n} &\leq s^n \operatorname{Tr}\left[\sum_{k_1=1}^{N} T_{k_1} P_S T_{k_1}^* P_\Theta \sum_{k_2=1}^{N} T_{k_2} P_S T_{k_2}^* P_\Theta \cdots \sum_{k_n=1}^{N} T_{k_n} P_S T_{k_n}^* P_\Theta\right] \\ &= s^{2n} \operatorname{Tr}[P_\Theta] = m s^{2n}.\end{aligned}$$

Next, let $\widetilde{F}$ denote the block matrix $\widetilde{F} = (\tilde{B}^*_{j,k})_{j,k}$. Similarly as above we get

$$\begin{aligned}\|\widetilde{F}\|_{S_{2n}}^{2n} &= \operatorname{Tr}\left[\sum_{\substack{j_1,j_2,\ldots,j_n=1 \\ k_1,k_2,\ldots,k_n=1}}^{N} \tilde{B}_{j_1,k_1} \tilde{B}^*_{j_1,k_2} \tilde{B}_{j_2,k_2} \tilde{B}^*_{j_2,k_3} \cdots \tilde{B}_{j_n,k_n} \tilde{B}^*_{j_n,k_1}\right] \\ &\leq \operatorname{Tr}\left[\sum_{\substack{j_1,j_2,\ldots,j_n=1 \\ k_1,k_2,\ldots,k_n=1}}^{N} R_S T^*_{j_1} P_\Theta T_{k_1} P_S T^*_{k_2} P_\Theta T_{j_1} P_S \cdots P_S T^*_{j_n} P_\Theta T_{k_n} P_S T^*_{k_1} P_\Theta T_{j_n} R^*_S\right].\end{aligned}$$

Using that $T_k^* P_\Theta = P_{\Theta - k} T_k$ and $T_j T_k^* = T_k^* T_j$ we further obtain

$$\begin{aligned}\|\widetilde{F}\|_{S_{2n}}^{2n} &\leq \operatorname{Tr}\left[\sum_{\substack{j_1,j_2,\ldots,j_n=1 \\ k_1,k_2,\ldots,k_n=1}}^{N} R_S T_{k_1} T^*_{j_1} P_{\Theta-k_1} P_S P_{\Theta-k_2} T_{j_1} T^*_{k_2} P_S \right. \\ &\qquad \left. \cdots P_S T_{k_n} T^*_{j_n} P_{\Theta-k_n} P_S T_{\Theta-k_1} T_{j_n} T^*_{k_1} R^*_S\right] \\ &= \operatorname{Tr}\left[\sum_{k_1=1}^{N} |(\Theta - k_1) \cap S \cap (\Theta - k_2)| T^*_{k_1} P_S T_{k_1} \right. \\ &\qquad \left. \cdots \sum_{k_n=1}^{N} |(\Theta - k_n) \cap S \cap (\Theta - k_1)| T^*_{k_n} P_S T_{k_n}\right].\end{aligned}$$

In the last step we have used that $P_{\Theta-k_1} P_S P_{\Theta-k_2} = P_{(\Theta-k_1)\cap S \cap (\Theta-k_2)}$ together with (9.5), and in addition the cyclicity of the trace. Clearly,

$$|(\Theta - k_1) \cap S \cap (\Theta - k_2)| \leq |(\Theta - k_1) \cap S| \leq |S| = s,$$

and furthermore, $|(\Theta - k_1) \cap S|$ is non-zero if and only if $k_1 \in \Theta - S$. This implies that

$$\sum_{k_1=1}^{N} |(\Theta - k_1) \cap S \cap (\Theta - k_2)| T_{k_1}^* P_S T_{k_1} \leq \sum_{k_1=1}^{N} |(\Theta - k_1) \cap S| P_{S-k_1} \leq s^2 P_{S+S-\Theta},$$

where the inequalities are understood entrywise. Combining the previous estimates yields

$$\|\widetilde{F}\|_{S_{2n}}^{2n} \leq s^{2n} \operatorname{Tr}[P_{S+S-\Theta}] = s^{2n} |S + S - \Theta| \leq s^{2n} s^2 m.$$

Since by assumption (9.3) $s \leq m$ it follows that

$$\max \left\{ \|(\sum_{j \neq k} B_{j,k}^* B_{j,k})^{1/2}\|_{S_{2n}}^{2n}, \|(\sum_{j \neq k} B_{j,k} B_{j,k}^*)^{1/2}\|_{S_{2n}}^{2n}, \|F\|_{S_{2n}}^{2n}, \|\widetilde{F}\|_{S_{2n}}^{2n} \right\}$$
$$\leq m^n s^{n+2}.$$

Using $\|H_S\|_{2\to 2} = \|H_S\|_{S_\infty} \leq \|H_S\|_{S_p}$ and applying the decoupling Lemma 6.21 and the Khintchine inequality in Theorem 6.22 we obtain for an integer n

$$\mathbb{E}\|H_S\|_{2\to 2}^{2n} = \mathbb{E}\|A_S^* A_S - \mathrm{Id}_S\|_{2\to 2}^{2n} \leq \mathbb{E}\|A_S^* A_S - \mathrm{Id}_S\|_{S_{2n}}^{2n}$$
$$= \frac{1}{m^{2n}} \mathbb{E}\| \sum_{j \neq k} \epsilon_j \epsilon_k B_{j,k}\|_{S_{2n}}^{2n} \leq \frac{4^{2n}}{m^{2n}} \mathbb{E}\| \sum_{j \neq k} \epsilon_j \epsilon_k' B_{j,k}\|_{S_{2n}}^{2n} \leq 2 \cdot 4^{2n} \left(\frac{(2n)!}{2^n n!} \right)^2 \frac{s^{n+2}}{m^n}.$$

Here $\boldsymbol{\epsilon}'$ denotes a Rademacher sequence, inpendent of $\boldsymbol{\epsilon}$. Let $p = 2n + 2\theta = (1 - \theta)2n + \theta(2n + 2)$ with $\theta \in [0, 1]$. Applying Hölder's inequality, see also (6.12), and the series of inequalities in (6.13) yields

$$\mathbb{E}\|H_S\|_{2\to 2}^{2n+2\theta} \leq (\mathbb{E}\|H_S\|_{2\to 2}^{2n})^{1-\theta} (\mathbb{E}\|H_S\|_{2\to 2}^{2n+2})^{\theta}$$
$$\leq 2 \cdot 4^{2n+2\theta} \left(\left(\frac{(2n)!}{2^n n!} \right)^{1-\theta} \left(\frac{(2(n+1))!}{2^{n+1}(n+1)!} \right)^{\theta} \right)^2 \frac{s^{n+\theta+2}}{m^{n+\theta}}$$
$$\leq 2 \cdot 2^{3/2} 4^{2n+2\theta} (2/e)^{2n+2\theta} (n+\theta)^{2n+2\theta} \frac{s^{n+\theta+2}}{m^{n+\theta}}.$$

In other words, for $p \geq 2$,

$$\left(\mathbb{E}\|H_S\|_{2\to 2}^p\right)^{1/p} \leq 4e^{-1} \sqrt{\frac{s}{m}} (2^{5/2} s^2)^{1/p} p.$$

An application of Proposition 6.5 yields

$$\mathbb{P}\left(\|H_S\|_{2\to 2} \geq 4\sqrt{\frac{s}{m}} u \right) \leq 2^{5/2} s^2 e^{-u} \tag{9.6}$$

for all $u \geq 2$. Note that $s \geq 1$ implies $2^{5/2} s^2 e^{-u} \geq 1/2$ for $u < 2$. Therefore, setting the right hand side equal $\varepsilon \leq 1/2$ yields $u \geq 2$. In particular, $\|H_S\| \leq \delta$ with probability at least $1 - \varepsilon$ provided (9.3) holds true. □

9.3 Completing the Proof

Let us now complete the proof of Theorem 5.1. Set

$$\alpha = \frac{\sqrt{s}t}{1-\delta} \tag{9.7}$$

for some $t, \delta \in (0,1)$ to be chosen later such that $\alpha < 1/\sqrt{2}$. According to Propositions 7.1 and 7.2, the probability that recovery fails is bounded from above by

$$2^{3/4}(N-s)e^{-\alpha^{-2}/2} + \mathbb{P}(\|A_S^* A_S - \mathrm{Id}\|_{2\to 2} > \delta) + \mathbb{P}(\mu > t). \tag{9.8}$$

By Theorem 9.3 we have $\mathbb{P}(\|A_S^* A_S - \mathrm{Id}\|_{2\to 2} > \delta) \leq \varepsilon/3$ provided

$$m \geq 16\delta^{-2} s \ln^2(3 \cdot 2^{5/2} s^2/\varepsilon), \tag{9.9}$$

and Proposition 9.1 yields $\mathbb{P}(\mu > t) \leq \varepsilon/3$ if

$$m \geq 6t^{-2} \ln(9N^2/\varepsilon). \tag{9.10}$$

The first term of (9.8) equals $\varepsilon/3$ for

$$\alpha = \frac{1}{\sqrt{2\ln(2^{3/4} \cdot 3(N-s)/\varepsilon)}} < \frac{1}{\sqrt{2}}.$$

Solving for t in (9.7) gives

$$t = \frac{1-\delta}{\sqrt{2s\ln(2^{3/4} \cdot 3(N-s)/\varepsilon)}},$$

and plugging into (9.10) yields the condition

$$m \geq \frac{12s}{(1-\delta)^2} \ln(9N^2/\varepsilon) \ln(2^{3/4} \cdot 3(N-s)/\varepsilon). \tag{9.11}$$

Choose $\delta = 8/15$. Then (5.1) implies both (9.9) and (9.11). □

10 Appendix

Here we show some lemmas that are needed in some of the proofs.

10.1 Covering Numbers for the Unit Ball

Proposition 10.1. *Let $\|\cdot\|$ be some semi-norm on $\mathbb{R}^n$ and let U be a subset of the unit ball $B = \{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\| \leq 1\}$. Then the covering numbers satisfy, for $t > 0$,*

$$N(U, \|\cdot\|, t) \leq \left(1 + \frac{2}{t}\right)^n. \tag{10.1}$$

Proof. If $\|\cdot\|$ fails to be a norm, we consider the quotient space $X = \mathbb{R}^n/\mathcal{N}$ where $\mathcal{N} = \{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\| = 0\}$ is the kernel of $\|\cdot\|$. Then $\|\cdot\|$ is a norm on X, and the latter is isomorphic to $\mathbb{R}^{n-d}$, where d is the dimension of $\mathcal{N}$. Hence, we may assume without loss of generality that $\|\cdot\|$ is actually a norm.

Let $\{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subset U$ be a maximal t packing of U, that is, a maximal set satisfying $d(\mathbf{x}_i, \mathbf{x}_j) > t$ for all $i \neq j$. Then the balls $B(\mathbf{x}_\ell, t/2) = \{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}-\mathbf{x}_\ell\| \leq t/2\}$ do not intersect and they are contained in the scaled unit ball $(1+t/2)B$. By comparing volumes (that is, Lebesgue measures) of the involved balls we get

$$\operatorname{vol}\left(\bigcup_{\ell=1}^{N} B(\mathbf{x}_\ell, t/2)\right) = N \operatorname{vol}\left((t/2)B\right) \leq \operatorname{vol}\left((1+t/2)B\right).$$

(Note that $\operatorname{vol}(B) < \infty$ since $\|\cdot\|$ is a norm.) On $\mathbb{R}^n$ the volume satisfies $\operatorname{vol}(tB) = t^n \operatorname{vol}(B)$, hence, $N(t/2)^n \operatorname{vol}(B) \leq (1+t/2)^n \operatorname{vol}(B)$ or $N \leq (1+2/t)^n$. To conclude the proof, observe that the balls $B(\mathbf{x}_\ell, t)$, $\ell = 1, \dots, N$ form a covering of U. Indeed, if there were an $\mathbf{x} \in U$ that is not covered, then $d(\mathbf{x}_\ell, \mathbf{x}) > t$, so that $\mathbf{x}$ could be added to the packing. But this is a contradiction to the maximality of the packing. □

10.2 Integral Estimates

This section contains estimates for two integrals.

Lemma 10.2. *For $u > 0$ it holds*

$$\int_u^\infty e^{-t^2/2} dt \leq \min\left\{\sqrt{\frac{\pi}{2}}, \frac{1}{u}\right\} \exp(-u^2/2).$$

Proof. A change of variables yields

$$\int_u^\infty e^{-t^2/2} dt = \int_0^\infty e^{-\frac{(t+u)^2}{2}} dt = e^{-u^2/2} \int_0^\infty e^{-tu} e^{-t^2/2} dt.$$

On the one hand, using that $e^{-tu} \leq 1$ for $t, u \geq 0$, we get

$$\int_u^\infty e^{-t^2/2} dt \leq e^{-u^2/2} \int_0^\infty e^{-t^2/2} dt = \sqrt{\frac{\pi}{2}} e^{-u^2/2}.$$

On the other hand, using that $e^{-t^2} \leq 1$ for $t \geq 0$ yields

$$\int_u^\infty e^{-t^2/2} dt \leq e^{-u^2/2} \int_0^\infty e^{-tu} dt = \frac{1}{u} e^{-u^2/2}. \tag{10.2}$$

This shows the desired estimate. □

Lemma 10.3. *For $\alpha > 0$ it holds*

$$\int_0^\alpha \sqrt{\ln(1+t^{-1})}dt \leq \alpha\sqrt{\ln(e(1+\alpha^{-1}))}. \tag{10.3}$$

Proof. First apply the Cauchy-Schwarz inequality to obtain

$$\int_0^\alpha \sqrt{\ln(1+t^{-1})}dt \leq \sqrt{\int_0^\alpha 1dt \int_0^\alpha \ln(1+t^{-1})dt}.$$

A change of variables and integration by parts yields

$$\begin{aligned}
&\int_0^\alpha \ln(1+t^{-1})dt = \int_{\alpha^{-1}}^\infty u^{-2}\ln(1+u)du \\
&= -u^{-1}\ln(1+u)\big|_{\alpha^{-1}}^\infty + \int_{\alpha^{-1}}^\infty u^{-1}\frac{1}{1+u}du \leq \alpha\ln(1+\alpha^{-1}) + \int_{\alpha^{-1}}^\infty \frac{1}{u^2}du \\
&= \alpha\ln(1+\alpha^{-1}) + \alpha.
\end{aligned}$$

Combining the above estimates concludes the proof. □

Acknowledgments. I would like to thank Massimo Fornasier for organizing the summer school "Theoretical Foundations and Numerical Methods for Sparse Recovery" and for inviting me to present this course. The time in Linz was very enjoyable and fruitful. Also I would like to thank Simon Foucart for the joint adventure of writing the monograph [55], which influenced very much these notes, and for sharing his insights and ideas on compressive sensing. Further, I greatly acknowledge RICAM and the START grant "Sparse Approximation and Optimization in High Dimensions" for hosting the summer school. I would further like to thank several people for nice and interesting discussions on the subject: Joel Tropp, Roman Vershynin, Rachel Ward, Stefan Kunis, Ingrid Daubechies, Karlheinz Gröchenig, Ron DeVore, Thomas Strohmer, Emmanuel Candès, Justin Romberg, Götz Pfander, Jared Tanner, Karin Schnass, Rémi Gribonval, Pierre Vandergheynst, Tino Ullrich, Albert Cohen, Alain Pajor. Also, I thank the following people for identifying errors and providing comments on previous versions of these notes: Jan Vybíral, Jan Haskovec, Silvia Gandy, Felix Krahmer, Ulas Ayaz, Tino Ullrich, Deanna Needell, Rachel Ward, Thomas Strohmer, Dirk Lorenz, Pasc Gavruta, Jan-Olov Strömberg, and Mike Wakin. My work on this topic started when being a PostDoc at NuHAG in Vienna. I would like to thank Hans Feichtinger and the whole NuHAG group for providing a very nice and productive research environment. I enjoyed my time there very much. Also I am very grateful to the Hausdorff Center for Mathematics (funded by the DFG) and to the Institute for Numerical Simulation at the University of Bonn for providing excellent working conditions and financial support. Last but not least, my greatest thanks go to Daniela and to our little children Niels and Paulina for making my math-free time so enjoyable.

Bibliography

[1] K. Alexander, Probability inequalities for empirical processes and a law of the iterated logarithm, *Ann. Probab.* 12 (1984), 1041–1067.

[2] W.O. Alltop, Complex sequences with low periodic correlations, *IEEE Trans. Inform. Theory* 26 (1980), 350–354.

[3] J.-M. Azaïs and M. Wschebor, *Level Sets and Extrema of Random Processes and Fields*, John Wiley & Sons Inc., 2009.

[4] W. Bajwa, J. Haupt, G. Raz, S.J. Wright and R. Nowak, Toeplitz-structured compressed sensing matrices., 2007, IEEE Workshop SSP.

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

[6] G. Bennett, Probability inequalities for the sum of independent random variables., *J. Amer. Statist. Assoc.* 57 (1962), 33–45.

[7] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Springer, 1976.

[8] R. Bhatia, *Matrix Analysis*, Graduate Texts in Mathematics 169, Springer-Verlag, New York, 1997.

[9] S. Boucheron, O. Bousquet, G. Lugosi and P. Massart, Moment inequalities for functions of independent random variables, *Ann. Probab.* 33 (2005), 514–560.

[10] S. Boucheron, G. Lugosi and P. Massart, Concentration inequalities using the entropy method, *Ann. Probab.* 31 (2003), 1583–1614.

[11] J. Bourgain, Bounded orthogonal systems and the $\Lambda(p)$-set problem, *Acta Math.* 162 (1989), 227–245.

[12] ———, *Λ_p-sets in analysis: results, problems and related aspects*, Handbook of the Geometry of Banach Spaces, Vol I, North-Holland, 2001, pp. 195–232.

[13] J. Bourgain and L. Tzafriri, Invertibility of 'large' submatrices with applications to the geometry of Banach spaces and harmonic analysis, *Israel J. Math.* 57 (1987), 137–224.

[14] O. Bousquet, *Concentration inequalities for sub-additive functions using the entropy method*, Stochastic Inequalities and Applications, Progr. Probab. 56, Birkhäuser, Basel, 2003, pp. 213–247.

[15] S. Boyd and L. Vandenberghe, *Convex Optimization.*, Cambridge Univ. Press, 2004.

[16] A. Buchholz, Operator Khintchine inequality in non-commutative probability, *Math. Ann.* 319 (2001), 1–16.

[17] ———, Optimal constants in Khintchine type inequalities for fermions, Rademachers and q-Gaussian operators, *Bull. Pol. Acad. Sci. Math.* 53 (2005), 315–321.

[18] E.J. Candès, The restricted isometry property and its implications for compressed sensing, *C. R. Acad. Sci. Paris S'er. I Math.* 346 (2008), 589–592.

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

[20] E.J. Candès and J. Romberg, Sparsity and incoherence in compressive sampling, *Inverse Problems* 23 (2007), 969–985.

[21] E.J. Candès, J. Romberg and T. Tao, Stable signal recovery from incomplete and inaccurate measurements, *Comm. Pure Appl. Math.* 59 (2006), 1207–1223.

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

[23] _____, Near optimal signal recovery from random projections: universal encoding strategies?, *IEEE Trans. Inform. Theory* 52 (2006), 5406–5425.

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

[25] S. S. Chen, D.L. Donoho and M. A. Saunders, Atomic decomposition by Basis Pursuit, *SIAM J. Sci. Comput.* 20 (1999), 33–61.

[26] H. Chernoff, A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations, *Ann. Math. Statist.* 23 (1952), 493–507.

[27] A. Cohen, *Numerical Analysis of Wavelet Methods*, North-Holland, 2003.

[28] A. Cohen, W. Dahmen and R. DeVore, Compressed sensing and best k-term approximation, *J. Amer. Math. Soc.* 22 (2009), 211–231.

[29] R. Coifman, F. Geshwind and Y. Meyer, Noiselets, *Appl. Comput. Harmon. Anal.* 10 (2001), 27–44.

[30] J. Cooley and J. Tukey, An algorithm for the machine calculation of complex Fourier series, *Math. Comp.* 19 (1965), 297–301.

[31] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics 61, SIAM, Society for Industrial and Applied Mathematics, 1992.

[32] I. Daubechies, M. Defrise and C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Comm. Pure Appl. Math.* 57 (2004), 1413–1457.

[33] I. Daubechies, R. DeVore, M. Fornasier and S. Güntürk, Iteratively re-weighted least squares minimization for sparse recovery, *Comm. Pure Appl. Math.* 63 (2010), 1–38.

[34] I. Daubechies, M. Fornasier and I. Loris, Accelerated projected gradient methods for linear inverse problems with sparsity constraints, *J. Fourier Anal. Appl.* 14 (2008), 764–792.

[35] G. Davis, S. Mallat and M. Avellaneda, Adaptive greedy approximations, *Constr. Approx.* 13 (1997), 57–98.

[36] V. de la Peña and E. Giné, *Decoupling. From Dependence to Independence*, Probability and its Applications (New York), Springer-Verlag, New York, 1999.

[37] R.A. DeVore, Deterministic constructions of compressed sensing matrices, *J. Complexity* 23 (2007), 918–925.

[38] D.L. Donoho, Compressed sensing, *IEEE Trans. Inform. Theory* 52 (2006), 1289–1306.

[39] D.L. Donoho and M. Elad, Optimally sparse representations in general (non-orthogonal) dictionaries via ell^1 minimization, *Proc. Nat. Acad. Sci.* 100 (2002), 2197–2202.

[40] D.L. Donoho, M. Elad and V.N. Temlyakov, Stable recovery of sparse overcomplete representations in the presence of noise, *IEEE Trans. Inform. Theory* 52 (2006), 6–18.

[41] D.L. Donoho and X. Huo, Uncertainty principles and ideal atomic decompositions, *IEEE Trans. Inform. Theory* 47 (2001), 2845–2862.

[42] D.L. Donoho and J. Tanner, Counting faces of randomly-projected polytopes when the projection radically lowers dimension, *J. Amer. Math. Soc.* 22 (2009), 1–53.

[43] D.L. Donoho and Y. Tsaig, Fast solution of l1-norm minimization problems when the solution may be sparse, *IEEE Trans. Inform. Theory* 54 (2008), 4789–4812.

[44] M.F. Duarte, M.A. Davenport, D. Takhar, J.N. Laska, S. Ting, K.F. Kelly and R.G. Baraniuk, Single-Pixel Imaging via Compressive Sampling, *IEEE Signal Processing Magazine* 25 (2008), 83–91.

[45] R.M. Dudley, The sizes of compact subsets of Hilbert space and continuity of Gaussian processes, *J. Functional Analysis* 1 (1967), 290–330.

[46] A. Dutt and V. Rokhlin, Fast Fourier transforms for nonequispaced data, *SIAM J. Sci. Comput.* 14 (1993), 1368 – 1393.

[47] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, Least angle regression, *Ann. Statist.* 32 (2004), 407–499.

[48] M. Elad and A.M. Bruckstein, A generalized uncertainty principle and sparse representation in pairs of bases., *IEEE Trans. Inform. Theory* 48 (2002), 2558–2567.

[49] A. Fannjiang, P. Yan and T. Strohmer, Compressed Remote Sensing of Sparse Objects, *preprint* (2009).

[50] G.B. Folland, *A Course in Abstract Harmonic Analysis*, CRC Press, 1995.

[51] S. Foucart, A note on ensuring sparse recovery via ℓ_1-minimization, *preprint* (2009).

[52] S. Foucart and R. Gribonval, Real vs. complex null space properties for sparse vector recovery, *preprint* (2009).

[53] S. Foucart and M. Lai, Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$, *Appl. Comput. Harmon. Anal.* 26 (2009), 395–407.

[54] S. Foucart, A. Pajor, H. Rauhut and T. Ullrich, The Gelfand widths of ℓ_p-balls for $0 < p \leq 1$, *preprint* (2010).

[55] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Appl. Numer. Harmon. Anal., Birkhäuser, Boston, in preparation.

[56] J. J. Fuchs, On sparse representations in arbitrary redundant bases, *IEEE Trans. Inform. Theory* 50 (2004), 1341–1344.

[57] A.Y. Garnaev and E.D. Gluskin, On widths of the Euclidean ball, *Sov. Math., Dokl.* 30 (1984), 200–204.

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

[59] G. Golub and C.F. van Loan, *Matrix Computations*, 3rd ed, The Johns Hopkins University Press, 1996.

[60] R. Gribonval and M. Nielsen, Sparse representations in unions of bases, *IEEE Trans. Inform. Theory* 49 (2003), 3320–3325.

[61] _____, Highly sparse representations from dictionaries are unique and independent of the sparseness measure, *Appl. Comput. Harmon. Anal.* 22 (2007), 335–355.

[62] R. Gribonval and P. Vandergheynst, On the exponential convergence of matching pursuits in quasi-incoherent dictionaries, *IEEE Trans. Inform. Theory* 52 (2006), 255–261.

[63] G. Grimmett and D. Stirzaker, *Probability and random processes*, Third ed, Oxford University Press, New York, 2001.

[64] K. Gröchenig, *Foundations of Time-Frequency Analysis*, Appl. Numer. Harmon. Anal., Birkhäuser Boston, 2001.

[65] K. Gröchenig, B. Pötscher and H. Rauhut, Learning trigonometric polynomials from random samples and exponential inequalities for eigenvalues of random matrices, *preprint* (2007).

[66] O. Guédon, S. Mendelson, A. Pajor and N. Tomczak Jaegermann, Majorizing measures and proportional subsets of bounded orthonormal systems, *Rev. Mat. Iberoam.* 24 (2008), 1075–1095.

[67] U. Haagerup, The best constants in the Khintchine inequality, *Studia Math.* 70 (1981), 231–283 (1982).

[68] J. Haupt, W. Bajwa, G. Raz and R. Nowak, Toeplitz compressed sensing matrices with applications to sparse channel estimation, *preprint* (2008).

[69] M. Herman and T. Strohmer, High-resolution radar via compressed sensing, *IEEE Trans. Signal Process.* 57 (2009), 2275–2284.

[70] W. Hoeffding, Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.* 58 (1963), 13–30.

[71] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, 1990.

[72] _____, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1994, Corrected reprint of the 1991 original.

[73] W. B. Johnson and J. Lindenstrauss (eds.), *Handbook of the Geometry of Banach Spaces Vol I*, North-Holland Publishing Co., Amsterdam, 2001.

[74] B.S. Kashin, Diameters of some finite-dimensional sets and classes of smooth functions., *Math. USSR, Izv.* 11 (1977), 317–333.

[75] A. Khintchine, Über dyadische Brüche, *Math. Z.* 18 (1923), 109–116.

[76] S. Kim, K. Koh, M. Lustig, S. Boyd and D. Gorinevsky, A method for large-scale l1-regularized least squares problems with applications in signal processing and statistics, *IEEE J. Sel. Top. Signal Proces.* 4 (2007), 606–617.

[77] T. Klein and E. Rio, Concentration around the mean for maxima of empirical processes, *Ann. Probab.* 33 (2005), 1060–1077.

[78] S. Kunis and H. Rauhut, Random sampling of sparse trigonometric polynomials II - orthogonal matching pursuit versus basis pursuit, *Found. Comput. Math.* 8 (2008), 737–763.

[79] M. Ledoux, *The Concentration of Measure Phenomenon*, AMS, 2001.

[80] M. Ledoux and M. Talagrand, *Probability in Banach Spaces.*, Springer-Verlag, Berlin, Heidelberg, NewYork, 1991.

[81] X. Li and C.-P. Chen, Inequalities for the Gamma function, *JIPAM. J. Inequal. Pure Appl. Math.* 8 (2007), Article 28, 3 pp. (electronic).

[82] F. Lust-Piquard, Inégalites de Khintchine dans $C_p(1 < p < \infty)$, *C. R. Acad. Sci. Paris S'er. I Math.* 303 (1986), 289–292.

[83] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1998.

[84] P. Massart, Rates of convergence in the central limit theorem for empirical processes, *Ann. Inst. H. Poincar'e Probab. Statist.* 22 (1986), 381–423.

[85] ———, About the constants in Talagrand's concentration inequalities for empirical processes, *Ann. Probab.* 28 (2000), 863–884.

[86] T. McConnell and M. Taqqu, Decoupling inequalities for multilinear forms in independent symmetric random variables, *Annals Prob.* 11 (1986), 943–951,.

[87] S. Mendelson, A. Pajor and N. Tomczak Jaegermann, Uniform uncertainty principle for Bernoulli and subgaussian ensembles, *Constr. Approx.* 28 (2009), 277–289.

[88] B. K. Natarajan, Sparse approximate solutions to linear systems., *SIAM J. Comput.* 24 (1995), 227–234.

[89] F. Nazarov and A. Podkorytov, *Ball, Haagerup, and distribution functions*, Complex Analysis, Operators, and related Topics, Oper. Theory Adv. Appl. 113, Birkhäuser, Basel, 2000, pp. 247–267.

[90] D. Needell and R. Vershynin, Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit, *Found. Comput. Math.* 9 (2009), 317–334.

[91] ———, Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit, *IEEE J. Sel. Topics Sig. Process.* (to appear).

[92] A. Pajor and S. Mendelson, On singular values of matrices with independent rows, *Bernoulli* 12 (2006), 761–773.

[93] G. Peškir, Best constants in Kahane-Khintchine inequalities for complex Steinhaus functions, *Proc. Amer. Math. Soc.* 123 (1995), 3101–3111.

[94] G. Peškir and A. N. Shiryaev, The Khintchine inequalities and martingale expanding sphere of their action, *Russian Math. Surveys* 50 (1995), 849–904.

[95] G. Pfander and H. Rauhut, Sparsity in time-frequency representations, *J. Fourier Anal. Appl.* 16 (2010), 233–260.

[96] G.E. Pfander, H. Rauhut and J. Tanner, Identification of matrices having a sparse representation, *IEEE Trans. Signal Process.* 56 (2008), 5376–5388.

[97] A. Pinkus, *On L^1-Approximation*, Cambridge Tracts in Mathematics 93, Cambridge University Press, Cambridge, 1989.

[98] G. Pisier, *Remarques sur un résultat non publié de B. Maurey*, Seminar on Functional Analysis, 1980–1981, École Polytech., 1981, pp. Exp. No. V, 13.

[99] ———, *The Volume of Convex Bodies and Banach Space Geometry*, Cambridge Tracts in Mathematics 94, Cambridge University Press, Cambridge, 1989.

[100] ———, Non-commutative vector valued L_p-spaces and completely p-summing maps, *Ast'erisque* (1998).

[101] D. Potts, G. Steidl and M. Tasche, *Fast Fourier transforms for nonequispaced data: A tutorial*, Modern Sampling Theory: Mathematics and Applications (J.J. Benedetto and P.J.S.G. Ferreira, eds.), Birkhäuser, 2001, pp. 247 – 270.

[102] H. Rauhut, Random sampling of sparse trigonometric polynomials, *Appl. Comput. Harmon. Anal.* 22 (2007), 16–42.

[103] ———, On the impossibility of uniform sparse reconstruction using greedy methods, *Sampl. Theory Signal Image Process.* 7 (2008), 197–215.

[104] ———, Stability results for random sampling of sparse trigonometric polynomials, *IEEE Trans. Information Theory* 54 (2008), 5661–5670.

[105] ———, Circulant and Toeplitz matrices in compressed sensing, in: *Proc. SPARS'09*, Saint-Malo, France, 2009.

[106] H. Rauhut and R. Ward, Sparse Legendre expansions via ℓ_1-minimization, *preprint* (2010).

[107] E. Rio, Inégalités de concentration pour les processus empiriques de classes de parties, *Probab. Theory Related Fields* 119 (2001), 163–175.

[108] ———, Une inégalité de Bennett pour les maxima de processus empiriques, *Ann. Inst. H. Poincar'e Probab. Statist.* 38 (2002), 1053–1057, En l'honneur de J. Bretagnolle, D. Dacunha-Castelle, I. Ibragimov.

[109] J. Romberg, Imaging via Compressive Sampling, *IEEE Signal Process. Magazine* 25 (2008), 14–20.

[110] ———, Compressive sensing by random convolution, *SIAM J. Imaging Sci.* 2 (2009), 1098–1128.

[111] M. Rosenfeld, *In praise of the Gram matrix*, The Mathematics of Paul Erdős, II, Algorithms Combin. 14, Springer, 1997, pp. 318–323.

[112] S. Ross, *Introduction to Probability Models*, Nineth ed, Academic Press, 2006.

[113] M. Rudelson, Random vectors in the isotropic position, *J. Funct. Anal.* 164 (1999), 60–72.

[114] M. Rudelson and R. Vershynin, Geometric approach to error-correcting codes and reconstruction of signals, *Internat. Math. Res. Notices* (2005), 4019–4041.

[115] ———, Sampling from large matrices: an approach through geometric functional analysis, *J. ACM* 54 (2007), Art. 21, 19 pp. (electronic).

[116] _____, On sparse reconstruction from Fourier and Gaussian measurements, *Comm. Pure Appl. Math.* 61 (2008), 1025–1045.

[117] W. Rudin, *Fourier Analysis on Groups*, Interscience Publishers, 1962.

[118] _____, *Functional Analysis*, McGraw-Hill Book Company, 1973.

[119] K. Schnass and Pierre Vandergheynst, Dictionary preconditioning for greedy algorithms, *IEEE Trans. Signal Process.* 56 (2008), 1994–2002.

[120] B. Simon, *Trace Ideals and their Applications.*, Cambridge University Press, Cambridge, 1979.

[121] T. Strohmer and R.W. jun. Heath, Grassmannian frames with applications to coding and communication., *Appl. Comput. Harmon. Anal.* 14 (2003), 257–275.

[122] M. Talagrand, Isoperimetry and integrability of the sum of independent Banach-space valued random variables, *Ann. Probab.* 17 (1989), 1546–1570.

[123] _____, New concentration inequalities in product spaces, *Invent. Math.* 126 (1996), 505–563.

[124] _____, Selecting a proportion of characters, *Israel J. Math.* 108 (1998), 173–191.

[125] _____, *The Generic Chaining*, Springer Monographs in Mathematics, Springer-Verlag, Berlin, 2005.

[126] Georg Tauböck, Franz Hlawatsch, Daniel Eiwen and Holger Rauhut, Compressive Estimation of Doubly Selective Channels in Multicarrier Systems: Leakage Effects and Sparsity-Enhancing Processing, *IEEE J. Sel. Top. Sig. Process.* 4 (2010), 255–271.

[127] J.A. Tropp, Greed is good: Algorithmic results for sparse approximation, *IEEE Trans. Inform. Theory* 50 (2004), 2231–2242.

[128] _____, Recovery of short, complex linear combinations via l_1 minimization, *IEEE Trans. Inform. Theory* 51 (2005), 1568–1570.

[129] _____, Just relax: Convex programming methods for identifying sparse signals in noise, *IEEE Trans. Inform. Theory* 51 (2006), 1030–1051.

[130] _____, On the conditioning of random subdictionaries, *Appl. Comput. Harmon. Anal.* 25 (2008), 1–24.

[131] J.A. Tropp and D. Needell, CoSaMP: Iterative signal recovery from incomplete and inaccurate samples, *Appl. Comput. Harmon. Anal.* 26 (2008), 301–321.

[132] J.A. Tropp, M. Wakin, M. Duarte, D. Baron and R.G. Baraniuk, Random filters for compressive sampling and reconstruction, *Proc. 2006 IEEE Int. Conf. Acoustics, Speech, and Signal Processing* 3 (2006), 872–875.

[133] Joel A. Tropp, Jason N. Laska, Marco F. Duarte, Justin K. Romberg and Richard G. Baraniuk, Beyond Nyquist: Efficient sampling of sparse bandlimited signals, *IEEE Trans. Inform. Theory* 56 (2010), 520 –544.

[134] A. W. Van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*, Springer-Verlag, 1996.

[135] R. Varga, *Gershgorin and his Circles*, Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2004.

[136] R. Vershynin, Frame expansions with erasures: an approach through the non-commutative operator theory, *Appl. Comput. Harmon. Anal.* 18 (2005), 167–176.

[137] J.S. Walker, *Fast Fourier Transforms*, CRC Press, 1991.

[138] P. Wojtaszczyk, *A Mathematical Introduction to Wavelets*, Cambridge University Press, 1997.

Author information

Holger Rauhut, Hausdorff Center for Mathematics & Institute for Numerical Simulation, University of Bonn, Endenicher Allee 60, 53115 Bonn, Germany.
E-mail: rauhut@hcm.uni-bonn.de