Technische Universität München
TUM School of Computation, Information and Technology

# Representing and Recovering Structured High-Dimensional Data: Fast Dimension Reduction, Recovery Guarantees, and Neural Network Representation

Stefan Julian Bernhard Bamberger

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Christian Kühn

Prüfer*innen der Dissertation:

1. Prof. Dr. Felix Krahmer
2. Assistant Prof. Dr. Michael Kapralov
3. Assoc. Prof. Dr. Rachel Ward

Die Dissertation wurde am 30.05.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 10.03.2023 angenommen.

## Abstract

High-dimensional signals with a known structure and specifically their reconstruction from incomplete measurements in significantly lower dimensions have been the central object of study in the field of compressed sensing. The most commonly studied example of this is the sparse recovery problem, in which the signal vectors are sparse, i.e., have a limited number of non-zero entries, and the measurements in a lower dimension are obtained by applying a linear transformation. A crucial tool in this field is the restricted isometry property (RIP), which states that a matrix approximately preserves the norms of all sparse vectors. This property yields fundamental guarantees for the successful recovery of sparse signals. This work studies four selected topics from compressed sensing and related fields.

The first part deals with Johnson-Lindenstrauss embeddings, functions that reduce the dimension of data points while preserving their pairwise distances. An established construction for such embeddings is based on randomizing matrices that satisfy the RIP. In our work, we adapt this method to a class of embeddings with a structure that allows a particularly fast application to data points that are Kronecker products of multiple smaller vectors.

In the second part, we prove results about higher-order random tensors relevant to the third part and establish more general versions of them.

As the third topic, we investigate to what extent the sparse recovery problem is solvable with neural networks. We specifically restrict our analysis to networks that are invariant under positive scaling since we know that the considered problem also has this structure. In this context, we also study the related question to what extent neural networks can approximate continuous, positive scale-invariant functions in general.

In the fourth part, we study the RIP of random partial Fourier transforms under a modified sparsity model that limits the number of non-zero entries in multiple blocks separately rather than in the entire vector. For this problem, which is motivated by the sparsity structure of natural images in a wavelet basis, we can improve the required embedding dimension by logarithmic factors in the sparsity, compared to the best previous result.

## Zusammenfassung

Hochdimensionale Signale mit einer bekannten Struktur sowie deren Rekonstruktion aus unvollständigen Messungen in wesentlich niedrigeren Dimensionen sind der zentrale Forschungsgegenstand in der Theorie des Compressed Sensing. Das am meisten untersuchte solche Szenario ist das Rekonstruktionsproblem für dünnbesetzte Vektoren, in dem die Signalvektoren höchstens eine vorgegebene Anzahl von null verschiedener Einträge haben können. Von großer Bedeutung ist in diesem Zusammenhang die als Restricted Isometry Property (RIP) bekannte Eigenschaft bestimmter Matrizen, die Normen aller dünnbesetzten Vektoren annähernd zu erhalten. Aus dieser Eingenschaft folgen wichtige Garantien für die erfolgreiche Rekonstruktion dünnbesetzter Vektoren. Diese Arbeit untersucht vier spezielle Themen aus dem Compressed Sensing und verwandten Gebieten.

Im ersten Teil untersuchen wir Johnson-Lindenstrauss-Einbettungen. Diese reduzieren die Dimension eine Menge von Datenpunkten unter Erhaltung der paarweisen Abstände. Eine verbreitete Methode zur Konstruktion solcher Einbettungen basiert auf einer Randomisierung von Matrizen, welche die RIP erfüllen. In dieser Arbeit entwickeln wir eine Variante dieses Ansatzes für eine Klasse von Einbettungen, die eine besonders schnelle Transformation für Vektoren mit Kronecker-Struktur ermöglicht. Dies bedeutet, dass jeder Datenvektor das Kronecker-Produkt mehrerer kleinerer Vektoren ist.

Im zweiten Teil beweisen wir mehrere Ergebnisse über Zufallstensoren höherer Ordnung. Spezielle Fälle dieser Ergebnisse werden zur vollständigen Behandlung des Themas aus dem ersten Teil benötigt.

Als drittes Thema untersuchen wir, inwiefern das Rekonstruktionsproblem für dünnbesetzte Vektoren mit neuronalen Netzen gelöst werden kann. Dabei beschränken wir uns auf Lösungen, die invariant gegenüber positiver Skalierung sind, da auch das zu lösende Problem diese Eigenschaft besitzt. In diesem Zusammenhang untersuchen wir auch, unter welchen Umständen neuronale Netze im Allgemeinen stetige Funktionen approximieren können, die invariant unter positiver Skalierung sind.

Im vierten Teil untersuchen wir die RIP von zufälligen partiellen Fouriertransformationen für eine Klasse dünnbesetzter Vektoren, für die im Gegensatz zur sonst üblichen Definition die Anzahl der von null verschiedenen Einträge nicht nur für den gesamten Vektor beschränkt ist, sondern für mehrere Blöcke im Vektor einzeln. Im Vergleich zum besten vorherigen Ergebnis verbessern wie die benötigte Einbettungsdimension um logarithmische Faktoren in der Anzahl der von null verschiedener Einträge.

## Acknowledgements

First of all, I am deeply indebted to my adviser Felix Krahmer for his consistent support throughout this project, ranging from help in finding suitable topics and contributing mathematical ideas to providing helpful suggestions to enhance this work. I am also grateful for being a part of his research group throughout my doctorate. I would like to thank all the other colleagues in the groups of applied numerical analysis and optimization and data analysis.

I would also like to thank Holger Boche for the valuable mathematical discussions and for being part of his research group over a large part of the duration of this project. Therefore I would like to thank my former colleagues at the chair of theoretical information technology.

Furthermore, I also appreciate the support from Reinhard Heckel and Rachel Ward, including advice, discussions, and feedback on my work. The collaboration with them made the respective parts of this thesis, about neural networks and Kronecker-structured objects, possible.

Next, I would like to express my gratitude to the examination committee for their engagement in the final review process of this thesis. This includes the leader of the examination committee Blanka Horvath, the already mentioned readers, Felix Krahmer and Rachel Ward, and the third reader Michael Kapralov.

Finally, I also appreciate the personal support from the people close to me. This especially concerns my family, my mother and brother, and, in silent memory, my father, who unfortunately cannot see my graduation anymore. And in particular, I would like to show my appreciation to my girlfriend for her backing and respect for the large amounts of time I invested into this thesis.

# Contents

# Introduction: Compressed Sensing and Content Overview

The field of *compressed sensing* (also known as *compressive sensing*) originated in the works [CRT06; Don06] and has grown into a widely studied theory about the reconstruction of signals from few measurements. In particular, it is about situations in which structural assumptions on the signal enable its reconstruction, which would not have been possible with standard linear algebra. The most prominent example of such a structural assumption is *sparsity*, which is also studied in [CRT06]. We assume that there is a signal $x \in \mathbb{R}^N$ that is known to be $s$-sparse, i.e., at most $s$ of its entries are non-zero. Then we observe measurements

$$y = Ax \in \mathbb{R}^m \tag{0.1}$$

for a matrix $A \in \mathbb{R}^{m \times N}$, where $m \ll N$. The question is, under what circumstances the signal $x$ can be recovered from the measurements $y$. From a linear algebra perspective, since $m < N$, the matrix $A$ has a non-trivial kernel, and therefore there is a whole affine subspace of dimension $\geq N - m$ of vectors $z \in \mathbb{R}^N$ that all satisfy $Az = y = Ax$. However, it is possible that within this subspace of possible solutions, the original signal $x$ is the only vector that is $s$-sparse. In this case, $x$ can be uniquely identified as the only $s$-sparse vector satisfying $Ax = y$, and therefore, the sparsity as an additional structural constraint on the signal makes this problem solvable.

Indeed, [CRT06] shows that this *sparse recovery problem* has a unique solution for a certain class of matrices $A$ with a number of measurements $m \ll N$. With follow-up works on related topics, this has evolved into a wide theory about reconstructing signals with structural assumptions. The most important aspects of compressed sensing can be found in the textbooks [FR13] and [EK12].

Even though sparsity is not the only possible structural assumption on the signal considered in the field, it has become one of the most important and widely studied ones. One reason for this is the typical sparsity of natural images in the wavelet basis [OSL00]. Compressed sensing has found important applications in various types of image reconstruction, for which [SK18] provides an overview. This includes magnetic resonance imaging (MRI) [Lus+08] and the single-pixel camera, which can reconstruct an image using only measurements from one single brightness sensor, and radar systems [HS09].

One main research topic in compressed sensing is the question of what matrices $A$ in (0.1) allow a unique recovery of all $s$-sparse signals $x$ from the measurements $Ax$. One is interested in properties of the matrix that can guarantee this unique recovery and what matrices fulfill them. Specific attention has been given to matrices $A$ that are relevant for certain practical applications. Moreover, the goal is usually to show the success of sparse recovery for a number $m$ of rows of $A$ that is as small as possible. As a related question, also lower bounds on $m$ have been shown that are required for matrices to enable successful sparse recovery with certain beneficial properties. The most important aspects of these questions that are relevant for this work are summarized in Section 0.2.

Another important and extensively studied aspect of the theory of compressed sensing is the question of how the original signal can be computed from the measurements efficiently if the unique recovery is known to be possible. More details about this topic are discussed in Section 0.3.

For both of the aforementioned topics, we are interested in stable and robust approaches. This means that the reconstruction should still work (up to a small error) for small perturbations of the signal (i.e., if the signal is not exactly $s$-sparse) and small perturbations of the measurements. So we should be able to approximately recover $x$ that is approximately $s$-sparse from

$$y = Ax + e \tag{0.2}$$

where $e \in \mathbb{R}^m$ is a *noise* vector with a small norm. Both, conditions on the matrix $A$ and recovery algorithms have been developed to make this possible.

Although deterministic constructions also exist, the best known matrices $A$ suitable for sparse recovery are random constructions. For this reason, high-dimensional probability theory is closely related to the field of compressed sensing. We discuss more details about this topic in Section 0.5.

Techniques developed for signal recovery in compressed sensing have also found important applications in dimension reduction in the form of Johnson-Lindenstrauss embeddings. These connections are discussed in Section 0.4.

First we summarize some important notation that will be used in all parts of the thesis. Then in each of the following four Sections 0.2 to 0.5, we give a short introduction to one topic from compressed sensing and related fields. Each of these topics is related to one of the four sections in the main part of this thesis. Therefore, we introduce each topic with a subsequent summary of the corresponding section in the main part.

The sections in the main part do not depend on each other, except for Subsection 2.5, which technically belongs to the topic of Section 1 but requires tools and notation from Section 2 and is therefore treated afterwards.

Most concepts used in this thesis can be considered in complex numbers and will be introduced in this introduction accordingly. However, to simplify the presentation, we restrict our results to real numbers in Sections 1, 2, 3. In Section 4, we explicitly present the results in complex numbers since the discrete Fourier transform as the main application generally has complex values.

Most results from this thesis have been submitted for publication elsewhere previously. Section 0.7 lists the details of this.

## 0.1 Notation

Throughout this thesis, we make use of the following notation.

- $[N] = \{1, \dots, N\}$ is the set of integers from 1 to $N$ for any $N \in \mathbb{Z}_{\geq 0}$.
- For any finite set $S$, we denote $|S|$ for the number of its elements.
- For a subset $S \subset M$, we denote $S^c := M \backslash S$ for its complement.
- $Id_N \in \mathbb{R}^{N \times N}$ is the identity matrix of size $N \times N$.
- $\sigma_1(A) \geq \dots \geq \sigma_r(A)$ are the singular values of the matrix $A \in \mathbb{C}^{m \times N}$ of rank $r$.
- $A^\dagger \in \mathbb{C}^{N \times m}$ is the Moore-Penrose pseudoinverse of $A \in \mathbb{C}^{m \times N}$, i.e., if $\operatorname{rank}(A) = r$ and $U\Sigma V^*$ is a singular value decomposition with $U \in \mathbb{C}^{m \times r}$, $\Sigma \in \mathbb{R}^{r \times r}$, and $V \in \mathbb{C}^{N \times r}$, then $A^\dagger = V\Sigma^\dagger U^*$ where $\Sigma^\dagger_{k,k} = \frac{1}{\Sigma_{k,k}}$ for $k \in [r]$.
- The *support* of $x \in \mathbb{C}^N$ is the set of indices with non-zero entries
$$\operatorname{supp}(x) := \{k \in [N] \,|\, x_k \neq 0\}.$$
- For $x \in \mathbb{C}^N$ and some set $S \subset [N]$, we define $x_S \in \mathbb{C}^{|S|}$ to be the vector of all entries with indices in $S$.
- For $p \geq 1$ and a vector $x \in \mathbb{C}^N$, we define the $\ell_p$ norm
$$\|x\|_p := \left( \sum_{k=1}^{N} |x_k|^p \right)^{\frac{1}{p}}$$

and the $\ell_\infty$ norm

$$\|x\|_\infty := \max_{k\in[N]} |x_k|.$$

- The $\ell_0$ "norm" of $x \in \mathbb{C}^N$ (which is not an actual norm) is
$$\|x\|_0 := |\operatorname{supp}(x)|.$$

- We call $x \in \mathbb{C}^N$ an $s$-sparse vector if $\|x\|_0 \leq s$. The set of $s$-sparse vectors is
$$\Sigma_s := \{x \in \mathbb{C}^N \,\big|\, \|x\|_0 \leq s\}$$
or the corresponding subset of $\mathbb{R}^N$ depending on the context.

- For a real number $p \geq 1$ and a random variable $X$ with values in $\mathbb{C}$, we define the $L_p$ norm
$$\|X\|_{L_p} := (\mathbb{E}|X|^p)^{\frac{1}{p}}$$
and the $L_\infty$ norm as the minimal $K \geq 0$ such that
$$|X| \leq K \qquad \text{almost surely.}$$

- The unit sphere is
$$S^{N-1} := \{x \in \mathbb{R}^N \,\big|\, \|x\|_2 = 1\}$$
or the corresponding subset of $\mathbb{C}^N$ depending on the context.

- For a metric space $(T, d)$ (on $\mathbb{R}^N$ or $\mathbb{C}^N$ we consider the $\|\cdot\|_2$ norm unless noted otherwise), the open ball with radius $r \geq 0$ and center $x_0 \in T$ is
$$B_r(x_0) := \{x \in T \,\big|\, d(x, x_0) < r\}$$
and the closed ball
$$\bar{B}_r(x_0) := \{x \in T \,\big|\, d(x, x_0) \leq r\}.$$

- $X \sim N(\mu, \sigma^2)$ if the random variable $X$ follows a normal distribution with mean $\mu$ and variance $\sigma^2$.

- $g \sim N(\mu, \Sigma)$ if the random vector $g$ in $\mathbb{R}^N$ follows a multivariate normal distribution with mean $\mu \in \mathbb{R}^N$ and covariance matrix $\Sigma \in \mathbb{R}^{N\times N}$.

- For functions $f, g : S \to [0, \infty)$ on some set $S$, we denote $f(x) \gtrsim g(x)$ if there is a constant $C > 0$ such that $f(x) \geq Cg(x)$ for all $x \in S$. $f(x) \lesssim g(x)$ is defined analogously and we write $f(x) \sim g(x)$ if both relations holds.

- $2^M$ refers to the power set of the set $M$.

- For $A \in \mathbb{C}^{m_1\times n_1}$ and $B \in \mathbb{C}^{m_2\times n_2}$, the *Kronecker product* is
$$A \otimes B = \begin{pmatrix} A_{1,1}B & \dots & A_{1,n_1}B \\ \vdots & \ddots & \vdots \\ A_{m_1,1}B & \dots & A_{m_1,n_1}B \end{pmatrix} \in \mathbb{C}^{m_1m_2\times n_1n_2}$$
and the analogous definition holds for vectors.

- For $x \in \mathbb{R}$, we denote the floor function $\lfloor x \rfloor := n$ for the unique $n \in \mathbb{Z}$ such that $n \leq x < n + 1$ and the ceiling function $\lceil x \rceil := n$ for the unique $n \in \mathbb{Z}$ such that $n - 1 < x \leq n$.

- We denote $\log : (0, \infty) \to \mathbb{R}$ for the natural logarithm (base $e$) and $\log_a : (0, \infty) \to \mathbb{R}$ for the logarithm to base $a$ for any $a > 1$.

## 0.2 The Restricted Isometry Property and Section 4

### 0.2.1 The Restricted Isometry Property

From (0.1) it is apparent that for any two different $s$-sparse vectors $x^{(1)}$, $x^{(2)} \in \mathbb{R}^N$ with the same measurements $Ax^{(1)} = Ax^{(2)}$, it holds that $A(x^{(1)} - x^{(2)}) = 0$. $x^{(1)} - x^{(2)}$ is a $2s$-sparse vector and it is clear that any $2s$-sparse vector can be written in this form. Therefore, the uniqueness of every $s$-sparse vector in its measurements under $A$ is equivalent to the property that $Ax \neq 0$ for any $2s$-sparse $x \in \mathbb{R}^N \backslash \{0\}$, i.e., that $\ker(A)$ does not contain any non-zero $2s$-sparse vector.

A stronger property that implies this and in addition makes efficient recovery with noise in the sense of (0.2) possible is given by the following definition.

**Definition 0.1** ($\ell_q$-Robust Null Space Property, Definition 4.21 of [FR13])**.** *Let $q \geq 1$. $A \in \mathbb{C}^{m\times N}$ satisfies the $\ell_q$-robust null space property of order $s$ (with respect to the norm $\|\cdot\|$) with constants $0 < \rho < 1$ and $\tau > 0$ if for all $S \subset [N]$ with at most $s$ elements,*

$$\|v_S\|_q \leq \frac{\rho}{s^{1-1/q}} \|v_{S^c}\|_1 + \tau \|Av\|$$

*holds for all $v \in \mathbb{C}^N$, where $v_S \in \mathbb{C}^{|S|}$ is the restriction of $v$ to the indices in $S$ and $S^c$ is the complement of $S$.*

For any parameters $0 < \rho < 1$, $\tau > 0$, this implies that $Ax \neq 0$ for any $2s$-sparse $x \in \mathbb{C}$ and therefore that the measurements of any $s$-sparse vector are unique. Otherwise take a $2s$-sparse $x \in \mathbb{C}^N \backslash \{0\}$ such that $Ax = 0$ and define $S \subset [N]$, $|S| = s$ to be the set of indices of the $s$ entries of $x$ with largest absolute value. Using the Hölder inequality, we obtain

$$\|v_{S^c}\|_1 \leq \|v_S\|_1 = \sum_{j\in S} |v_j| \cdot 1 \leq \left( \sum_{j\in S} |v_j|^q \right)^{\frac{1}{q}} \cdot \left( \sum_{j\in S} 1 \right)^{1-\frac{1}{q}} = \|v_S\|_q \cdot s^{1-1/q}$$

and therefore

$$\frac{\rho}{s^{1-1/q}} \|v_{S^c}\|_1 + \tau\|Av\| = \frac{\rho}{s^{1-1/q}} \|v_{S^c}\|_1 < \frac{\|v_{S^c}\|_1}{s^{1-1/q}} \leq \|v_S\|_q,$$

such that the $\ell_q$-robust null space property cannot hold. Beyond this, this property also ensures that a stable, robust and algorithmically efficient method for sparse recovery is possible as explained in the next Section 0.3.

Even though the $\ell_q$-robust null space property is enough to guarantee solutions for the sparse recovery problem, the stronger *restricted isometry property* has played a crucial role in compressed sensing. On the one hand, for many different matrices it can be proven more conveniently. On the other hand, it also has other applications beyond the sparse recovery problem such as the Johnson-Lindenstrauss embeddings discussed in Section 0.4. It states that a matrix $A$ approximately preserves the norms of all $s$-sparse vectors.

**Definition 0.2** (Restricted Isometry Property, Definition 6.1 in [FR13])**.** *For $A \in \mathbb{C}^{m\times N}$, the $s$-th restricted isometry constant $\delta_s = \delta_s(A)$ is the smallest $\delta \geq 0$ such that*

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2 \tag{0.3}$$

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

*The matrix $A$ satisfies the $(s, \delta)$-restricted isometry property (RIP) if $\delta_s \leq \delta$.*

So $A$ is said to satisfy the $(s, \delta)$-RIP if (0.3) holds for all $s$-sparse $x \in \mathbb{C}^N$. By Theorem 6.13 in [FR13], the $(2s, \delta)$-RIP for $\delta < 0.62$ implies the $\ell_2$-robust null space property with constants $\rho$ and $\tau$ that only depend on $\delta$.

As mentioned before, the best known constructions for matrices satisfying the RIP are random matrices that satisfy the RIP with a probability that is close to 1. Specifically, for any $\eta \in (0,1)$, a *Gausian* matrix, i.e, $A \in \mathbb{R}^{m\times N}$ with independent, normally distributed entries $A_{j,k} \sim N(0, \frac{1}{m})$ satisfies the $(s,\delta)$-RIP with probability $1 - \eta$ if $m \geq C_\eta \delta^{-2} s \log(\frac{N}{s})$, where $C_\eta > 0$ is a constant that only depends on $\eta$. This is a special case of a result for a class of matrices whose rows are independent subgaussian random vectors shown in Theorem 9.6 of [FR13]. Subgaussian random vectors will be discussed in Section 0.5.

On the other hand, with a technique known as *Gelfand widths* (see Chapter 10 of [FR13]), lower bounds on $m$ can be shown, proving that for any matrix that has the $(2s,\delta)$-RIP for a $\delta < 0.62$, the number of rows must be $m \geq cs\log(\frac{eN}{s})$ for a constant $c > 0$ (Corollary 10.8 in [FR13]). In this sense, the number of rows required for the Gaussian matrices to satisfy the RIP is known to be optimal up to constant factors.

There have also been approaches to construct deterministic matrices that satisfy the RIP, for example [Ban+13] lists several techniques. However, most approaches such as [DeV07] require the matrix to have $m \gtrsim s^2(\log N)^\alpha$ rows for a constant $\alpha \geq 1$. Therefore in contrast to the random matrices above, this scales with $s^2$ instead of $s$ which is far above the known lower bound for the RIP. The best known result [Bou+11] showed that it is possible with an exponent smaller than 2 but still requires $m \gtrsim s^{2-\epsilon}$ for a very small constant $\epsilon > 0$.

Between the fully random Gaussian matrices and deterministic ones, *structured* random matrices have been analyzed that are constructed from a small number of random parameters, compared to the Gaussian matrices above in which every single entry is an independent random variable. One particular example is given by subsampled Fourier matrices whose $m$ rows are randomly drawn from the discrete Fourier transform $F \in \mathbb{C}^N$ (see Section 0.6 below). Because of the dependencies between their entries, these subsampled Fourier matrices are more difficult to analyze. However, as shown in [HR16], they satisfy the $(s,\delta)$-RIP with high probability for $m \gtrsim \delta^{-2}(\log(\delta^{-1}))^2 s\log(N)(\log(s/\delta))^2$. This bound on $m$ has the optimal scaling in $s$ and is only off the lower bound by logarithmic factors. These matrices are relevant for certain applications and also provide the advantage that there is a fast algorithm for computing the matrix-vector product.

### 0.2.2 Summary of Section 4

We discuss more about this topic of subsampled Fourier matrices in **Section 4**. There we consider a version of the RIP for a sparsity model in which not only the total number of non-zero entries in the vector $x$ is bounded. Instead, the entries of $x$ are divided into $r$ blocks and within each block $k$, there is a maximal number $s_k$ of non-zero entries. So the regular sparsity is a special case of this for $r = 1$. The RIP of subsampled Fourier matrices for this sparsity model has already been shown in [LA19]. In Section 4 we show an improved variant of their main result that has a weaker requirement on the number of rows $m$.

## 0.3 Approaches for Sparse Recovery and Section 3

### 0.3.1 Established Approaches for Sparse Recovery

Given a guarantee that the measurements of all $s$-sparse signals are unique as discussed in the previous section, a natural question is how the unique signal $x$ can be reconstructed from the measurements $y = Ax$ algorithmically. Picking the sparsest $z$ such that $Az = y$, i.e., taking the solution of

$$\min \|z\|_0 \qquad\qquad \text{s.t. } Az = y$$

would recover $x$ but solving this minimization problem has been shown to be NP-hard in general [Nat95].

To enable more efficient solutions, the $\|\cdot\|_0$ function has been replaced by the $\ell_1$ norm. This is still guaranteed to recover the signal if $A$ satisfies the $\ell_q$-robust null space property (Definition 0.1). More specifically, in order to also allow for noisy measurements (0.2) for $\|e\|_2 \leq \eta$, the following problem known as *quadratically constrained basis pursuit* has been studied (Section 3.1 in [FR13])

$$\min \|z\|_1 \qquad\qquad \text{s.t. } \|Az - y\|_2 \leq \eta. \tag{0.4}$$

An according recovery guarantee is given in the following theorem.

**Theorem 0.3** (Theorem 4.22 in [FR13])**.** *Let $A \in \mathbb{C}^{m\times N}$ fulfill the $\ell_2$-robust null space property of order $s$ with constants $0 < \rho < 1$ and $\tau > 0$. Any solution $\hat{x}$ of (0.4) with $\|\cdot\| = \|\cdot\|_2$, $y = Ax + e$ for a signal vector $x \in \mathbb{C}^N$ and an error vector $\|e\|_2 \leq \eta$ is close to $x$ in the sense that*

$$\|x - \hat{x}\|_p \leq \frac{C}{s^{1-1/p}}\sigma_s(x)_1 + Ds^{1/p-1/2}\eta$$

*for $1 \leq p \leq 2$, constants $C, D > 0$ that only depend on $\rho$ and $\tau$, and*

$$\sigma_s(x)_1 := \inf_{\tilde{x}\in\Sigma_s} \|x - \tilde{x}\|_1.$$

An alternative version to (0.4) is the *basis pursuit denoising* ((3.2) in [FR13])

$$\min \lambda\|z\|_1 + \|Az - y\|_2^2. \tag{0.5}$$

As shown in Proposition 3.2 of [FR13], the problems (0.4) and (0.5) are closely related in the sense that any unique solution of one of them also minimizes the other one for a suitable parameter.

Unlike for $\ell_0$ minimization, there are efficient algorithms to solve these minimization problems. For example, (0.4) for $\eta = 0$ in real numbers can be written as a linear program and also for the general version, known efficient techniques from convex optimization can be used (see Chapter 15 in [FR13]). Moreover, there is also a variety of algorithms that specifically solve (0.4) or (0.5) or that iteratively approximate the signal $x$ directly from the measurements with a matrix that satisfies the $\ell_q$-robust null space property or the RIP. Chapters 3 and 15 of [FR13] introduce some of them.

### 0.3.2 Neural Networks and Summary of Section 3

In contrast to these established reconstruction methods in compressed sensing, in **Section 3**, we consider neural networks to reconstruct vectors from their measurements. Specifically, feed-forward neural networks define functions $f : \mathbb{R}^m \to \mathbb{R}^n$ of the type

$$f(x) = W_{d+1}\sigma\left(W_d\sigma\left(\dots W_2\sigma(W_1x + b_1) + b_2 \dots\right) + b_d\right) + b_{d+1},$$

where $W_1 \in \mathbb{R}^{k_1\times m}, W_2 \in \mathbb{R}^{k_2\times k_1}, \dots, W_{d+1} \in \mathbb{R}^{n\times k_d}$ are *weight matrices*, $b_1 \in \mathbb{R}^{k_1}, \dots, b_d \in \mathbb{R}^{k_d}, b_{d+1} \in \mathbb{R}^n$ are *bias vectors*, $d$ is the number of hidden layers or *depth* and $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function that is applied to vectors component-wise. A very common choice for $\sigma$ is the rectified linear unit (ReLU) activation function defined by $\mathrm{ReLU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$

Such networks have empirically been proven to be successful for many applications in image recovery and processing such as the reconstruction of images from incomplete measurements [Jin+17] or noise reduction or completion of missing parts in images [XXC12]. On the theoretical side, the established *universal approximation theorem* can guarantee that any continuous function on a compact domain can be approximated by a network with only one hidden layer with an arbitrarily small maximal deviation.

In this work, our goal is to analyze their theoretical performance on the simple model of sparse recovery. We are interested in a neural network function $f$ with ReLU activation function corresponding to a measurement matrix $A$ (that satisfies the RIP), such that $f(Ax) = x$ (or only $f(Ax) \approx x$) for all sparse vectors. Moreover, we are also interested in robustness results like Theorem 0.3 for these methods. Rather than explicit constructions for the networks, we investigate under what circumstances such networks exist at all and specifically which minimal depth $d$ they need to have.

We take one more special aspect into account for this. Since we know for any $s$-sparse signal $x$ and $\lambda \geq 0$, if $y = Ax$ is the measurement vector produced by $x$, then $\lambda y$ is produced by $\lambda x$. Therefore, we restrict our analysis to neural networks that by design satisfy the condition $f(\lambda y) = \lambda f(y)$ for $\lambda \geq 0$. We call functions satisfying this *positive homogeneous.* With such networks, we can ensure that the reconstruction will eventually work for all $s$-sparse vectors and not only those with a bounded norm, like it might be an issue for approaches based on the usual universal approximation theorem. For the ReLU activation function, we will show the following contrast under these circumstances. Sparse recovery cannot be performed at all with one hidden layer but it can be performed with an arbitrarily small error and in a robust way with two hidden layers.

As we will see, this is closely related to the problem of approximating general continuous positive homogeneous functions with neural networks. We investigate, under which circumstances such functions can be approximated with arbitrary precision. A version of the universal approximation theorem has already been established for positive homogeneous functions. We connect to this and show that having two layers is necessary for this and that the ReLU function, up to some modifications of it, is the only activation function that generate a class of positive homogeneous networks that can approximate any continuous positive homogeneous function.

We also show that the solutions of optimization problems like (0.4) and (0.5) can be approximated with neural networks.

Furthermore, we show that the sparse recovery problem can be solved exactly (the other approaches always allow an error even though it can be made arbitrarily small) by a ReLU network with $\lceil \log_2(s) + 1 \rceil$ layers. For the case $s = 1$, we also give an explicit construction for such a network with a relatively small width.

## 0.4 Johnson-Lindenstrauss Embeddings and Section 1

### 0.4.1 Johnson-Lindenstrauss Embeddings

Johnson-Lindenstrauss embeddings, first introduced in [JL84], were developed as functions that map a finite set of vectors from a high-dimensional space into a space of significantly lower dimension in such a way that that the pairwise distances within this finite set are preserved. This preserves the structure of the data that can be used for applications like large matrix multiplication [Sar06], clustering [FB03], and dictionary learning [ST20] whose performance can be improved due to the reduced dimension.

While in their first occurrence [JL84] they were realized by random rotations (i.e., linear maps), subsequent works have shown that various other kinds of randomized linear maps can be used [Ach01; DG03; AV06]. In this spirit, a Johnson-Lindenstrauss can be defined in the following way that has been used in Definition 1 of [BK21].

**Definition 0.4** (Johnson-Lindenstrauss Embedding). *Let $A \in \mathbb{R}^{m \times N}$ be a random matrix and $p \in \mathbb{Z}_{\geq 1}$, $\epsilon, \eta \in (0,1)$. We say that $A$ is a $(p, \epsilon, \eta)$-Johnson Lindenstrauss embedding (JLE) if for any set $E \subset \mathbb{R}^N$ with $|E| = p$, with probability $\geq 1 - \eta$, the inequality*

$$(1-\epsilon)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\epsilon)\|x\|_2^2 \tag{0.6}$$

*holds for all $x \in E$ simultaneously.*

For some purposes, also the *distributional* Johnson-Lindenstrauss property is considered for random matrices, meaning that $\mathbb{P}(\big|\|Ax\|_2^2 - \|x\|_2^2\big| > \epsilon\|x\|_2^2) \leq \eta'$ holds for all $x \in \mathbb{R}^N$. In this case, by a union bound, the norms of $p$ points are preserved simultaneously with probability $\geq 1 - \eta$ if $\eta' \leq \frac{\eta}{p}$. More details are described in Definition 1.1 and the following remark.

These embeddings are considered for the case that $m \ll N$, i.e., the map significantly reduces the dimension. In this case, it is clear that the matrix $A$ in the above dimension cannot be deterministic. Otherwise, we can take a non-zero vector $x_0$ of the kernel into $E$ for which $Ax_0 = 0$ and therefore (0.6) would be violated. The best constructions allow an embedding dimension of $m \sim \epsilon^{-2}\log(p)$, which has also been shown to be optimal [LN17].

A connection between Johnson-Lindenstrauss embeddings and compressed sensing is based on the fact that the inequality required for Johnson-Lindenstrauss embeddings (0.6) and for the restricted isometry property (0.3) is the same. The difference is that in the former case, it is required to hold on any finite set of vectors of given size, in the latter case it is required to hold for the one set of $s$-sparse vectors (which is infinite). Nevertheless, [Bar+08] established a relation between the properties showing that a JLE for $p \geq \exp(Cs\log(N/s))$ for a constant $C > 0$ satisfies the restricted isometry property with a high probability.

On the other hand, [KW11] establishes a converse of this and shows that also every matrix satisfying the restricted isometry property can be turned into a Johnson-Lindenstrauss embedding by multiplying all its columns by random signs. In this sense, these concepts are equivalent.

Because of the particular motivation to speed up algorithms by reducing the size of their input data with Johnson-Lindenstrauss embeddings, fast Johnson-Lindenstrauss embeddings that can be applied to individual vectors with an efficient algorithm have been studied [AC06; AL08]. The above connection between Johnson-Lindenstrauss [KW11] provides a new method to construct fast Johnson-Lindenstrauss embeddings from RIP matrices that are known to have a fast transform such as subsampled Fourier matrices.

### 0.4.2 Summary of Section 1

In **Section 1** we connect to this and investigate fast Johnson-Lindenstrauss embeddings with an additional requirement. For certain applications, we are interested in data vectors that are Kronecker products $x = x^{(1)} \otimes \cdots \otimes x^{(d)}$ of $d$ vectors. A Kronecker fast Johnson-Lindenstrauss embedding should offer an efficient algorithm for the matrix-vector product with such vectors. Especially, the construction discussed there does not require the Kronecker product $x$ to be computed explicitly as this would be required for standard matrix multiplication. Improving some other recent works on this field, in Section 1 we generalize the approach from [KW11] to constructions that can be Kronecker efficient. The proof and result of [KW11] is a special case of this if the number of factors in the Kronecker product is $d = 1$.

## 0.5 High-Dimensional Probability and Section 2

### 0.5.1 High-Dimensional Probability

As mentioned at the beginning, because of the importance of random matrices in compressed sensing, it is closely related to a field that is often referred to as *high-dimensional probability*, which deals with random vectors and matrices in usually high-dimensional spaces. Notable works that cover this topic in depth include the textbook [Ver18], the technical report [Van14], but also in the compressed sensing literature Chapters 7 and 8 in [FR13] and Chapter 5 in [EK12]. In this section, we summarize the most important terms that are used in the subsequent sections.

An important concept for data processing with random matrices are *subgaussian* random variables. Their characterization is, vaguely speaking, that their fluctuations are at most as big as the ones of a normally distributed variable. This can be stated precisely in the following four equivalent ways.

**Proposition 0.5** (Proposition 2.5.2 in [Ver18])**.** *Let $X$ be a random variable in $\mathbb{R}$. Then the following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.*

- *Tails:*

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2/K_1^2) \qquad \text{for all } t \geq 0.$$

- *Moments:*

$$\|X\|_{L_p} = (\mathbb{E}|X|^p)^{1/p} \leq K_2\sqrt{p} \qquad \text{for all } p \geq 1.$$

- *Moment generating function (MGF) of $X^2$:*

$$\mathbb{E}\exp(\lambda^2 X^2) \leq \exp(K_3^2\lambda^2) \qquad \text{for all } |\lambda| \leq \frac{1}{K_3}.$$

- *MGF of $X^2$ bounded at some point:*

$$\mathbb{E}\exp(X^2/K_4^2) \leq 2.$$

*Moreover, if $\mathbb{E} = 0$, then the above properties are also equivalent to*

- *MGF of $X$:*

$$\mathbb{E}\exp(\lambda X) \leq \exp(K_5^2\lambda^2) \qquad \text{for all } \lambda \in \mathbb{R}.$$

We call those variables that satisfy these equivalent condition *subgaussian* and the corresponding $K_i$ the *subgaussian norm.* Any of the four conditions can be used to define this term. We pick the following one.

**Definition 0.6** (Definition 2.5.6 in [Ver18])**.** *Let $X \in \mathbb{R}$ be a random variable. We define the subgaussian norm*

$$\|X\|_{\psi_2} := \inf\{K \geq 0 \,|\, \mathbb{E}\exp(X^2/K^2) \leq 2\}.$$

*$X$ is called subgaussian if $\|X\|_{\psi_2} < \infty$.*

One can show that this is actually a norm for the subgaussian random variables (see Section 2.5.6 in [Ver18]). Analogously, the $\|\cdot\|_{\psi_2}$ norm can also be defined a complex valued random variable $X \in \mathbb{C}$ as the $\psi_2$-norm of $|X|$.

Example 2.5.8 in [Ver18] lists three examples of subgaussian variables: A Gaussian variable $X \sim N(0, \sigma^2)$ has $\|X\|_{\psi_2} \leq C\sigma$ for a constant $C$, a Bernoulli variable $X \in \{1, -1\}$ with $\mathbb{E}X = 0$ has $\|X\|_{\psi_2} = \frac{1}{\sqrt{\log 2}}$. If $X$ is bounded almost surely, then $\|X\|_{\psi_2} \leq C\|X\|_{L_\infty}$ for a constant $C$.

Similar to the subgaussian variables that arise from a comparison of the tails to a Gaussian distribution, there are also *subexponential* variables that arise from a comparison to the tails of an exponential distribution and have an analogous characterization to the above proposition (see Proposition 2.7.1 in [Ver18]) and a subexponential norm $\|X\|_{\psi_1} = \inf\{K \geq 0 \,|\, \mathbb{E}\exp(X/K) \leq 2\}$. An example is $X^2$ for $X \sim N(0,1)$. More generally, the concept of *strong domination* discussed in **Section 2.3.1** allows to draw conclusions from the comparison of general tails of random variables.

One important property of subgaussian random variables is the following relation about the $\|\cdot\|_{\psi_2}$ norm of a sum of independent subgaussian variables.

**Proposition 0.7** (Proposition 2.6.1 in [Ver18])**.** *Assume that the random variables $X_1, \dots, X_N$ in $\mathbb{C}$ are independent, mean zero, and subgaussian. Then the sum $\sum_{k=1}^N X_k$ is also subgaussian and has the norm*

$$\left\| \sum_{k=1}^N X_k \right\|_{\psi_2}^2 \leq C \sum_{k=1}^N \|X_k\|_{\psi_2}^2,$$

*where $C$ is an absolute constant.*

The cited Proposition 2.6.1 in [Ver18] only considers the real-valued case but the statement for complex numbers follows from applying the real-valued version to the real and imaginary part of the sum.

Besides one-dimensional random variables, the subgaussian property can also be generalized to random vectors in arbitrary dimension as in the following definition for the real-valued case.

**Definition 0.8** (Definition 3.4.1 in [Ver18])**.** *We say that a random vector $X \in \mathbb{R}^N$ is subgaussian if $\langle X, x\rangle$ is a subgaussian random variable in $\mathbb{R}$ for each $x \in \mathbb{R}^N$ and*

$$\|X\|_{\psi_2} := \sup_{x \in S^{N-1}} \|\langle X, x\rangle\|_{\psi_2}.$$

Moreover, we say that a random vector $X \in \mathbb{R}^N$ is *isotropic* if $\mathbb{E}XX^T = Id_N$ (Definition 3.2.1 in [Ver18]).

Examples of vectors that are both, subgaussian and isotropic, are a multivariate normally distributed $g \sim N(0, Id_N)$ or a *Rademacher vector* $\xi \in \{\pm 1\}^N$ whose entries are independent variables with $\mathbb{P}(\xi_k = -1) = \mathbb{P}(\xi_k = 1) = \frac{1}{2}$ for each $k \in [N]$.

An important application of Rademacher vectors is given with the *symmetrization* technique that can be used to turn a sum of mean zero variables into a sum of symmetric random variables.

**Lemma 0.9** (Symmetrization, Lemma 6.4.2 in [Ver18])**.** *Let $(V, \|\cdot\|)$ be a normed space. Assume that $X_1, \dots, X_N$ are independent, mean zero random vectors in $V$. Then*

$$\frac{1}{2}\mathbb{E}\left\| \sum_{k=1}^N \xi_k X_k \right\| \leq \mathbb{E}\left\| \sum_{k=1}^N X_k \right\| \leq 2\mathbb{E}\left\| \sum_{k=1}^N \xi_k X_k \right\|,$$

*where $\xi \in \{\pm 1\}^N$ is a Rademacher vector that is independent of $X_1, \dots, X_N$.*

This lemma can be used in various scenarios. In this work, it is used at the beginning of the main proof of **Section 4**. It also occurs in **Section 2** in the generalized form of Lemma 2.15.

Another useful tool from this field is the decoupling technique that can be applied to a double sum of random variables.

**Theorem 0.10** (Decoupling, Theorem 8.11 in [FR13])**.** *Let $X_1, \dots, X_n$ be independent, mean $0$ random variables, $A \in \mathbb{R}^{N \times N}$, and $F : \mathbb{R} \to \mathbb{R}$ a convex function. Then*

$$\mathbb{E}F\left( \sum_{\substack{j,k=1 \\ j \neq k}}^N A_{j,k} X_j X_k \right) \leq \mathbb{E}F\left( 4 \sum_{j,k=1}^N A_{j,k} X_j \bar{X}_k \right),$$

*where $(\bar{X}_1, \dots, \bar{X}_N)$ is an independent copy of $(X_1, \dots, X_N)$.*

This decoupling technique is a main ingredient for the Hanson-Wright inequality that can be used to control the tails of a double sum that is also known as *chaos*. Versions for multiple different classes of distributions have been shown. The following one concerns subgaussian variables.

**Theorem 0.11** (Hanson-Wright Inequality, Theorem 1.1 from [RV13])**.** *Let $A \in \mathbb{R}^{N \times N}$. Let $X \in \mathbb{R}^N$ be a random vector with independent entries such that $\mathbb{E}X = 0$ and $\|X\|_{\psi_2} \leq K$. Then for every $t \geq 0$,*

$$\mathbb{P}(|X^T AX - \mathbb{E}X^T AX| > t) \leq 2 \exp\left[-c \min\left\{\frac{t^2}{K^4 \|A\|_F^2}, \frac{t}{K^2 \|A\|_{2\to 2}}\right\}\right],$$

*where $\|A\|_F$ is the Frobenius and $\|A\|_{2\to 2}$ the spectral norm of $A$.*

*Covering numbers* are another concept that is commonly used in the context of high-dimensional probability.

**Definition 0.12** (Covering numbers, Definition 4.2.2 in [Ver18])**.** *Let $(T, d)$ be a metric space. For $u > 0$, the covering number $\mathcal{N}(T, d, u)$ is the smallest number of closed balls with centers in $T$ and radii $u$ whose union contains $T$. If $(V, \|\cdot\|)$ is a normed space, we denote $\mathcal{N}(V, \|\cdot\|, u)$ for the covering number with respect to the metric that is induced by the norm.*

Such numbers will be of particular importance for controlling suprema of stochastic processes on an infinite index set by finite approximations in **Section 1**. A well-known standard estimate is the one for the Euclidean unit ball

**Lemma 0.13** (Covering numbers of the Euclidean ball, Corollary 4.2.13 in [Ver18])**.** *The following inequalities hold for the covering numbers of the Euclidean unit ball $\bar{B}_1(0) = \{x \in \mathbb{R}^n \mid \|x\|_2 \leq 1\}$ for any $u > 0$,*

$$\left(\frac{1}{u}\right)^n \leq \mathcal{N}(\bar{B}_1(0), \|\cdot\|_2, u) \leq \left(1 + \frac{2}{u}\right)^n.$$

### 0.5.2 Summary of Section 2

The main result of **Section 2** is a generalized version of the above Hanson-Wright inequality for subgaussian variables with a sum that unlike $X^T AX = \sum_{j,k=1}^N A_{j,k} X_j X_k$ ranges over $2d$ indices instead of only two, i.e.,

$$\sum_{i_1,\dots,i_{2d}=1}^{N} A_{i_1,\dots,i_d,i_{d+1},\dots,i_{2d}} X_{i_1}^{(1)} \dots X_{i_d}^{(d)} X_{i_{d+1}}^{(1)} \dots X_{i_{2d}}^{(d)},$$

where the coefficients $A_{i_1,\dots,i_d,i_{d+1},\dots,i_{2d}}$ depend on index tuples of length $2d$ and $X^{(1)}, \dots, X^{(d)}$ are vectors with independent subgaussian entries. On this way, we also establish a generalized version of the decoupling Theorem 0.10.

Instead of the coefficient matrix $A$, we have a coefficient *array* whose entries $A_{i_1,\dots i_{2d}}$ are indexed by tuples of length $2d$. We also develop a special notation to deal with such long index tuples which is introduced in **Section 2.1.4**. Even though **Section 1** provides a self-contained proof for a special case of its main result that shows the most important ideas behind the proof, the general case requires the notation and results of **Section 2** and is therefore postponed to **Section 2.5**.

### 0.6 Some Relevant Orthonormal Bases

Important constructions for matrices with the RIP and also Johnson-Lindenstrauss embeddings are based on the discrete Fourier transform (DFT). Details about the DFT and related transformations and algorithms can be found in [RKH10]. The normalized DFT is the linear map described by the normalized DFT or Fourier matrix $F \in \mathbb{C}^{N\times N}$ whose entries are

$$F_{j,k} = \frac{1}{\sqrt{N}} e^{-\frac{2\pi i(j-1)(k-1)}{N}}.$$

Beside its relevance for practical applications such as MRI ([LDP07]), this matrix has the following interesting mathematical properties.

- $F$ is unitary, i.e., $F^*F = Id_N$.
- All entries of $F$ have the same value $|F_{j,k}| = \frac{1}{\sqrt{N}}$.
- There is a fast algorithm, the fast Fourier transform (FFT, [RKH10]) that can compute the matrix vector product $Fx$ for any vector $x \in \mathbb{C}^N$ in $\mathcal{O}(N \log N)$ operations.

Moreover, there is another important matrix that satisfies the same properties but only has real-valued entries. The Hadamard matrix $H_n \in \mathbb{R}^{N\times N}$ for a dimension $N = 2^n$ that is a power of 2, can also be seen as a multidimensional variant of a Fourier transform [Kun79], exists only for dimensions $N$ that are powers of 2 and is recursively defined as

$$H_0 = (1) \qquad\qquad H_{n+1} = \frac{1}{\sqrt{2}} \begin{pmatrix} H_n & H_n \\ H_n & -H_n \end{pmatrix}.$$

One can also show that this is unitary with entries of absolute value $\frac{1}{\sqrt{N}}$ and the fast Walsh-Hadamard transform (see also [RKH10]) can compute the matrix-vector product in $\mathcal{O}(N \log N)$ operations.

In **Section 4**, such matrices will play an important role for the construction of matrices with a modified RIP. In **Section 1**, our main construction of fast Johnson-Lindenstrauss embeddings will be based on Hadamard matrices.

The Haar wavelets form another class of orthonormal bases that are relevant in this work. General wavelets are given by different shifts and scales of a base function called mother wavelet. Using them, general wavelet transforms can be defined, which exist for discrete or continuous coefficients and for signals on a discrete or a continuous domain. Details about this wide-ranging topic can be found in the textbook [Dau92]. Its relevance for compressed sensing is constituted by the typical sparsity of natural images in wavelet coefficient representations discussed in [OSL00], which concerns the discrete wavelet transformation for discrete signals. Important classes of wavelet transforms can be shown to define orthonormal bases. The Haar wavelets, as a particular example for this, exist for a dimension $N = 2^r$ that is a power of 2, and are defined, for example, in Section II of [AHR16], as $\phi_0 \in \mathbb{R}^N$ and $\phi_{j,p} \in \mathbb{R}^N$ for $j = 0, \dots, r-1$, $p = 0, \dots, 2^j - 1$ with entries

$$\phi_0(t) = 2^{-r}$$
$$\phi_{j,p}(t) = \begin{cases} 2^{\frac{j-r}{2}} & \text{for } 2^{r-j}p \leq t < 2^{r-j}(p+\frac{1}{2}) \\ -2^{\frac{j-r}{2}} & \text{for } 2^{r-j}(p+\frac{1}{2}) \leq t < 2^{r-j}(p+1) \\ 0 & \text{otherwise.} \end{cases}$$

for $1 \leq t \leq N$. This defines $2^r = N$ vectors that form an orthonormal basis. For $\phi_{j,p}$, $j$ defines the scale and $p$ the shift of the wavelet.

Wavelet coefficients of natural images are usually sparse in this basis but there is even an additional structure. Typically, they are sparser for fine scales (i.e., $\phi_{j,k}$ for large $j$) than for coarse ones. This phenomenon, previously discussed in [Adc+17], is a key motivation for the result shown in **Section 4**, which particularly considers the case of Haar wavelets.

## 0.7 Previous Publications

Some results of this thesis have previously been submitted for publication individually. This concerns the following submissions.

- "Johnson-Lindenstrauss Embeddings with Kronecker Structure" by authors Stefan Bamberger, Felix Krahmer, and Rachel Ward, submitted to *SIAM Journal on Matrix Analysis and Applications*, publisher: Society for Industrial and Applied Mathematics
  Copyright ©by SIAM. Unauthorized reproduction of this article is prohibited.
  A preprint of this work is available at `https://arxiv.org/abs/2106.13349`, [BKW21a].
  The content of this submission mostly agrees with Sections 1.1 to 1.6 and Section 2.5.

- "The Hanson-Wright Inequality for Random Tensors" by authors Stefan Bamberger, Felix Krahmer, and Rachel Ward, submitted to *Sampling Theory, Signal Processing, and Data Analysis*, publisher: Springer Nature Switzerland AG
  A preprint of this work is available at `https://arxiv.org/abs/2106.13345`, [BKW21b].
  The content of this submission mostly agrees with Sections 2.1 to 2.4.

The usage in the author's thesis is permitted by each of the publishers' agreements.

# 1 Johnson-Lindenstrauss Embeddings with Kronecker Structure

This section, along with Section 2.5 and parts of Section 2.1.4, shares major similarities with the article “Johnson-Lindenstrauss Embeddings with Kronecker Structure” by authors Stefan Bamberger, Felix Krahmer, and Rachel Ward, that was submitted to *SIAM Journal on Matrix Analysis and Applications*. A preprint of this work is available at `https://arxiv.org/abs/2106.13349`, [BKW21a].

## 1.1 Introduction

As discussed in Section 0.4, Johnson-Lindenstrauss embeddings provide a random embedding of finitely many points into a lower-dimensional vector space while preserving the structure of these points, i.e. their pairwise Euclidean distances, which has a wide range of applications.

In particular, the technique of *sketching* – for which [Woo14] provides a detailed overview – uses dimension reduction transforms such as JL embeddings to reduce the complexity of problems in numerical linear algebra. For example, instead of solving the classical linear regression problem $\min_x \|Ax - b\|_2^2$, one can apply a Johnson-Lindenstrauss embedding $\Phi$ to $b$ and the columns of $A$ which leads to a smaller-dimensional problem $\min_x \|\Phi Ax - \Phi b\|_2^2$ which can often be solved more efficiently. The Johnson-Lindenstrauss assumption is a simple sufficient condition under which the solution of the reduced problem is guaranteed to yield a good approximation to the original problem [Sar06].

In response to the driving application of improving algorithmic complexity of sketched linear algebra problems at massive scale, a line of research on *fast* Johnson-Lindenstrauss embeddings emerged, concerning the construction and analysis of random matrices $\Phi$ with the Johnson-Lindenstrauss property and which also have structure allowing for fast matrix-vector multiplication This analysis was initiated with the fast JL transform introduced in [AC06], in the form of a randomly row-subsampled discrete Hadamard matrix with randomized column signs. This construction was later improved and refined in [Vyb11; AL13], and ultimately sharpened to the best-known embedding power in [KW11] by establishing a near-equivalence between the Johnson-Lindenstrauss embedding property and a deterministic restricted isometry property [KW11]. Recently, this line of work found new energy following the work [BBK18] which proposed the use of a row-subsampled discrete Hadamard matrix with column signs randomized according to a Kronecker-structured Rademacher vector, and conjectured that such an embedding satisfies the Johnson-Lindenstrauss property. The Kronecker structure allows for even faster matrix-vector multiplication when applied to data points with Kronecker structure themselves, as arise naturally when dealing with multidimensional data arrays (see, for example, applications to kernel methods with polynomial sketching [Ahl+20a], and solving least squares problems with tensor structure [JKW20; Iwe+21]). Indeed, suppose we want to embed a data point $x = x^{(1)} \otimes \cdots \otimes x^{(d)} \in \mathbb{R}^{n^d} = \mathbb{R}^N$ which is a Kronecker product of $d$ data vectors, each of dimension $n$. If the embedding matrix $\Phi \in \mathbb{R}^{m \times N}$ itself has Kronecker structure $\Phi = \Phi^{(1)} \otimes \cdots \otimes \Phi^{(d)}$ where the dimensions of the factors of $\Phi$ correspond to the factor dimensions of $x$, then the matrix-vector multiplication $\Phi x$ can be factored as $\Phi x = (\Phi^{(1)} x^{(1)}) \otimes \cdots \otimes (\Phi^{(d)} x^{(d)})$, and *can be computed factor by factor, without constructing $x$ explicitly.* Because the Kronecker product of discrete Hadamard matrices is itself a discrete Hadamard matrix, embedding matrices in the form of discrete Hadamard matrices with Kronecker-structured random column signs fall within this framework, and it is natural to study the embedding power of such transforms.

In this paper, we study Johnson-Lindenstrauss embeddings of the type $\Phi D_\xi$ where $\Phi$ satisfies the restricted isometry property and $\xi$ is a Kronecker product of $d$ Rademacher vectors. Motivated by their relation to tensor subspace embeddings, such embeddings have been analyzed in a number of works [Ahl+20a; MB20; Iwe+21; JKW20]. In particular Lemma 4.11 of [Ahl+20a] shows that an embedding dimension of $\Theta((\log p)^{d+1})$ is sufficient for embedding $p$ points simultaneously in the case that $\Phi$ is a subsampled Hadamard transform and Theorem

2.1 of [JKW20] establishes that $\Theta((\log p)^{2d-1})$ is sufficient for general $\Phi$. Generalizing an approach from [KW11] on near-equivalence between the Johnson-Lindenstrauss property and the restricted isometry property to higher-degree tensor embeddings, we show that in fact an embedding dimension of $\Theta((\log p)^d)$ is both sufficient and necessary up to logarithmic factors in the dimension.

### 1.1.1 Background and prior work

Recall the distributional version of the JL Lemma: for any $\epsilon > 0$ and $\eta < 1/2$ and positive integer $N$, there exists a distribution over $\mathbb{R}^{m\times N}$ such that for a fixed unit-length vector $x \in \mathbb{R}^N$ and for a random matrix $\Phi$ drawn from this distribution with $m = O(\epsilon^{-2}\log(1/\eta))$,

$$\mathbb{P}(\big|\|\Phi x\|_2^2 - 1\big| > \epsilon) < \eta. \tag{1.1}$$

The dependence $m = O(\epsilon^{-2}\log(1/\eta))$, as achieved by (properly normalized) random matrices with independent and identically distributed subgaussian entries [DG03], is tight, as shown recently in [LN17] improving on a previous (nearly-tight) lower bound [Alo03].

For a given $\Phi : \mathbb{R}^N \to \mathbb{R}^m$ generated as such, computing the matrix-vector product $x \to \Phi x$ has time complexity $O(mN)$. The fast Johnson-Lindenstrauss as introduced in [AC06] and improved in [AL13; KW11], is constructed by randomly flipping the column signs of a random subset of $m$ rows from the $N \times N$ Discrete Fourier (or Discrete Hadamard) Transform. Exploiting the FFT algorithm, the fast JLT computes a matrix-vector product in time $O(N\log(N))$. The trade-off for this time savings is that the fast JLT has reduced embedding power $m = O(\epsilon^{-2}\log(1/\eta)\log^3(N))$.

More recently, the Kronecker fast JL transform (KFJLT) was proposed in [BBK18], to further improve the algorithmic complexity of the fast JL embedding in applications to Kronecker-structured data.

Such a construction has found applications as a key ingredient of the oblivious sketching procedure [Ahl+20a], a multiscale construction for dimension reduction applicable for subspace embeddings and approximate matrix multiplication. A central idea of this construction is the repeated application of the Kronecker FJLT of order $d = 2$.

The KFJLT of order $d$ acts on a Kronecker-structured vector $x = x^{(1)} \otimes \cdots \otimes x^{(d)} \in \mathbb{R}^{n_1 \dots n_d} = \mathbb{R}^N$ as follows: For fixed diagonal matrices with i.i.d. Rademacher random variables $D_{(1)}, \dots, D_{(d)}$ of dimensions $n_1, \dots, n_d$ respectively, and for a random subset $\Omega \subset [N]$ of size $|\Omega| = m$:

1. Randomly flip signs of the entries in each vector factor according to $x^{(k)} \to D_{(k)}x^{(k)} =: z^{(k)}$;

2. Compute the DFTs of each factor $y^{(k)} = H_{n_k}z^{(k)}$, where $H_n$ is the $n \times n$ DFT matrix (normalized to be a unitary transform).

3. Compress $y = y^{(1)} \otimes \cdots \otimes y^{(d)} \in \mathbb{C}^N$ to $y_\Omega \in \mathbb{C}^m$, where $y_\Omega$ consists of the entries in $y$ restricted to the subset $S$

4. Rescale $y_\Omega$ by $\sqrt{N/m}$.

The Kronecker JL transform extends to a well-defined linear map for any input $x \in \mathbb{R}^N$, taking the form of a matrix which can be expressed as the product of three matrix types:

$$\sqrt{\frac{N}{m}} \cdot P_\Omega \cdot H \cdot D_\xi \in \mathbb{R}^{m\times N} \tag{1.2}$$

where $D_\xi$ is the $N \times N$ diagonal matrix with diagonal vector $\xi = \xi_1 \otimes \xi_2 \otimes \cdots \otimes \xi_d$ the Kronecker product of $n_1, \dots, n_d$-dimensional Rademacher vectors, $H = H_{n_1} \otimes \cdots \otimes H_{n_d}$ is the Kronecker product of orthonormal DFTs (or, more generally, of bounded orthogonal matrices, including

DFTs, Hadamard, etc .... ), and $P_\Omega : \mathbb{R}^N \to \mathbb{R}^m$ denotes the projection matrix onto the coordinate subset $\Omega$. Our results hold for more general constructions of Kronecker products of matrix factors satisfying the restricted isometry property, after randomizing their column signs.

In the special case $d = 1$, the Kronecker FJLT reduces to the standard FJLT as considered in [KW11]. However, when the input vector has Kronecker structure so that the mapping can be applied separately to the matrix-vector factors, the complexity of computing a KJLT transform matrix-vector product improves to $O(n_1 \log(n_1) + \cdots + n_d \log(n_d) + md)$. The price that is paid is that the embedding power (that is, the minimal scaling of the embedding dimension in $m$ necessary for the distributional JL (1.1)) is weakened by the reduced randomness in $\xi$. For a numerical demonstration of the suboptimal scaling, we refer the reader to [JKW20]. A general theoretical lower bound was, to our knowledge, not available before this paper; lower bounds for related but somewhat different constructions were shown in [Ahl+20a].

At the same time, a number of works have investigated sufficient conditions on the embedding dimension to ensure that the map given by (1.2) satisfies the distributional JL property (1.1). The papers [Ahl+20a; MB20] show, among other results, that for (1.2) based on the Hadamard transform, a sufficient condition is given by $m = C_d \cdot \frac{1}{\epsilon^2} \log(1/\eta)^{d+1}$, up to logarithmic factors in $\log(1/\eta), 1/\epsilon$, and $N$. While the analysis in [MB20] is restricted to vectors with a Kronecker structure, the generalization of [Ahl+20a] applies to arbitrary vectors.

On the other hand, the paper [JKW20] used the near-equivalence between JL embedding and restricted isometry property from [KW11] to provide the sufficient condition $m = C_d \frac{1}{\epsilon^2}(\log(1/\eta))^{2d-1}$ for any subsampled bounded orthonormal transform, thus including but not limited to constructions based on Hadamard transform.

To put these two results into perspective, we remind the reader that the tensor degree $d$ is typically small – recall that the oblivious sketching procedure of [Ahl+20a] only uses the case $d = 2$, where the two conditions basically agree. Hence also for our results, we will pay special attention to optimizing the dependence for small values of $d$.

Compared to the corresponding result stated in [Ahl+20a], our main theorem is – like [JKW20] – not restricted to the specific construction (1.2) but can be applied to arbitrary $\Phi D_\xi \in \mathbb{R}^{m \times N}$ where $\Phi$ satisfies an RIP of sufficient order. Beside (1.2), a different possible application of this with a fast transformation of Kronecker structured vectors is given by $\Phi D_\xi \in \mathbb{R}^{m \times N}$ where

$$\Phi = \frac{1}{\sqrt{m}} P_\Omega (A_{v^{(1)}} \otimes \cdots \otimes A_{v^{(d)}}) \tag{1.3}$$

where $P_\Omega$ and $D_\xi$ are like in (1.2), $v^{(1)}, \ldots, v^{(d)}$ are independent Rademacher vectors and $A_{v^{(j)}} \in \mathbb{R}^{n_j}$ represents the circular convolution by the vector $v^{(j)}$, i.e., $A_{v^{(j)}} z = v^{(j)} * z$. For $d = 1$, [HPX19] shows this $\Phi$ to have the $(s, \delta)$-RIP if $m \gtrsim \delta^{-2} (\log(1/\delta))^2 s (\log(s/\delta))^2 \log N$, using the main result of [HR16] and the fact that $A_{v^{(1)}}$ has bounded entries and that $\frac{1}{\sqrt{N}} A_{v^{(1)}}$ satisfies the RIP for $s \lesssim \delta^2 n / (\log n)^4$.

For arbitrary $d$, both of the aforementioned two observations about $A_{v^{(1)}}$ can be transferred from the single factors to the Kronecker product of the $A_{v^{(j)}}$ (losing a factor $\leq 2^d$ in the RIP constant). In this way, the proof for the RIP can be adapted to the matrix $\Phi$ in (1.3).

### 1.1.2 Contributions of this work

In this work, we improve the existing bounds on the embedding dimension for the general Kronecker FJLT to $m = C_d \frac{1}{\epsilon^2} (\log(1/\eta))^d$, up to logarithmic factors in $\log(1/\eta)$, $1/\epsilon$, and in $N$, improving the results in [Ahl+20a] by a factor of $\log(1/\eta)$. In particular, for the case of $d = 2$ at the core of the oblivious sketching procedure [Ahl+20a], our results improve the scaling of the embedding dimension in $\log(\frac{1}{\eta})$ from cubic to quadratic.

We additionally prove that this embedding result is optimal in the $\eta$ dependence by providing a lower bound of $m = \Theta((\log(1/\eta))^d)$ in Section 1.3. We achieve the optimal bounds by generalizing the near-equivalence between the JL property and the restricted isometry property of [KW11] to higher-order tensors, in a sharper way than what was shown in [JKW20], by carefully using a higher-dimensional analog of the Hanson-Wright inequality for random tensors.

We state our main results in Section 1.2. In Section 1.3, we provide lower bounds to prove the optimality of our main results. The proof of the general result introduces an advanced index notation for higher order tensors. This notation is introduced for the next part of the thesis in Section 2.1.4. To simplify the presentation, we first prove the special case $d = 2$ in Section 1.5. This proof already contains the most important ideas. This makes it easier to follow the general proof which uses the notation and results from Section 2 and is therefore postponed to Section 2.5. Then we conclude by discussing the implications of our work in Section 1.6.

### 1.1.3 Related work

Tensor Johnson-Lindenstrauss constructions have become a recent topic of study, even beyond the concrete construction of (1.2).

Tensor JL embeddings based on sparse matrix structure have been studied in the context of vectors with Kronecker structure, based on the count sketch technique [CCF02], which has been extended to the tensorized version known as tensor sketch in [PP13]. Applications to problems including subspace embeddings and approximate matrix multiplication are presented in [ANW14]. However, these methods have a worse dependence on the failure probability compared to Kronecker FJLT.

The paper [Iwe+21] derived fast tensor embeddings for subspaces. Their embedding consists of the Kronecker FJLT from [JKW20] and a subsequent vector JL embedding. Thus, with our work we can improve the intermediate dimension and thus the number of random bits and the time complexity for the application to low rank tensors (for arbitrary tensors, the time complexity is dominated by the number of entries in the entire tensor). [MB20] uses the Kronecker FJLT for embeddings of subspaces consisting of low-rank tensors which can also be improved with the FJLT result of our work. The paper [Sun+21] proposed tensor random projections as matrices whose rows are i.i.d. Kronecker products of independent Gaussian vectors, and proved embedding properties for such constructions for Kronecker products of order $d = 2$. The paper [Ahl+20a] extended the analysis beyond $d = 2$, and [CJ20] further refined and extended these results in the context of sketching constrained least squares problems.

### 1.1.4 Notation

We make use of the essential notation introduced in the introduction. Especially, we denote $H \in \mathbb{R}^{N\times N}$ and $H_k \in \mathbb{R}^{2^k\times 2^k}$ for the Hadamard introduced in Section 0.6.

A random vector with independent entries which are $\pm 1$ with probability $\frac{1}{2}$ each, is called a Rademacher vector.

## 1.2 Main Result

**Definition 1.1.** *For $\epsilon, \eta > 0$, a random matrix $A \in \mathbb{R}^{m\times N}$ satisfies the $(\epsilon, \eta)$ distributional Johnson-Lindenstrauss property if for all $x \in \mathbb{R}^N$ with $\|x\|_2 = 1$,*

$$\mathbb{P}\left(\left|\|Ax\|_2^2 - 1\right| > \epsilon\|x\|_2^2\right) \leq \eta.$$

**Remark 1.2.** *If $A \in \mathbb{R}^{m\times N}$ has the $(\epsilon, \frac{\tilde{\eta}}{p(p-1)})$ distributional Johnson-Lindenstrauss property, then for any set $E \subset \mathbb{R}^N$ with $|E| = p$ elements, by a union bound we obtain*

$$\begin{aligned}&\mathbb{P}\left(\exists x, y \in E : \left|\|Ax - Ay\|_2^2 - \|x-y\|_2^2\right| > \epsilon\|x-y\|_2^2\right)\\ &\leq \sum_{\substack{x,y\in E\\ x\neq y}} \mathbb{P}\left(\left|\|A\frac{x-y}{\|x-y\|_2}\|_2^2 - 1\right| > \epsilon\right) \leq |E|(|E|-1)\cdot\frac{\tilde{\eta}}{p(p-1)} = \tilde{\eta}.\end{aligned}$$

*So with a probability of at least $1 - \tilde{\eta}$, it holds that*

$$\forall x, y \in E : \left| \|Ax - Ay\|_2^2 - \|x - y\|_2^2 \right| \leq \epsilon \|x - y\|_2^2.$$

*Then $A$ preserves all pairwise distances in the set $E$ up to a factor of $1 \pm \epsilon$.*

**Theorem 1.3.** *For $d \geq 1$, let $n_1, \dots, n_d$ be dimensions such that $N = n_1 \dots n_d$. Let $0 < \epsilon, \eta < 1$ and $\Phi \in \mathbb{R}^{m \times N}$ be a matrix satisfying the $(4s^d, \delta)$-RIP for $s \geq \log \frac{1}{\eta}$ and $\delta \leq C(d)\epsilon$ where $C(d)$ is a constant that only depends on $d$.*

*Let $\xi^{(1)} \in \{\pm 1\}^{n_1}, \dots, \xi^{(d)} \in \{\pm 1\}^{n_d}$ be independent Rademacher vectors and $\xi := \xi^{(1)} \otimes \dots \otimes \xi^{(d)} \in \mathbb{R}^N$. Define $A := \Phi D_\xi \in \mathbb{R}^{m \times N}$ where $D_\xi$ is a diagonal matrix with the entries of $\xi$ on its diagonal.*

*Then $A$ satisfies the $(\epsilon, \eta)$ distributional Johnson-Lindenstrauss property.*

Applying the result by Haviv and Regev about the RIP of subsampled bounded orthonormal matrices [HR16] to the case of a subsampled Kronecker product of $d$ Hadamard matrices, we obtain the following corollary.

**Corollary 1.4.** *Let $n_1, \dots, n_d, N, \epsilon, \eta, \xi$ be as in Theorem 1.3, $\nu \in (0,1)$, $\Phi = \sqrt{\frac{N}{m}} P_\Omega F \in \mathbb{R}^{m \times N}$ where $P_\Omega \in \mathbb{R}^{m \times N}$ represents uniform independent subsampling of rows with replacement and $F \in \mathbb{R}^{N \times N}$ is a unitary matrix with entries bounded by $\frac{D}{\sqrt{N}}$ in absolute value.*

*If $N \geq \frac{1}{(\nu)^{C_1 d \log\log(\frac{1}{\nu})}}$ and*

$$m \geq C(d) D^2 \epsilon^{-2} \left( \log \frac{1}{\eta} \right)^d \left( \log \frac{C(d)}{\epsilon} \right)^2 \log N \left( \log \frac{C(d) \log \frac{1}{\eta}}{\epsilon} \right)^2,$$

*then with probability $\geq 1 - \nu$ (with respect to $P_\Omega$), we obtain a matrix $\Phi$ such that $\Phi D_\xi$ satisfies the $(\epsilon, \eta)$ distributional Johnson-Lindenstrauss property (with respect to the probability in $\xi$).*

*$C_1$ is an absolute constant and $C(d)$ only depends on $d$.*

**Remark 1.5.** *For norm preservation of $p$ points simultaneously through a union bound, an $(\epsilon, \frac{c}{p})$ distributional Johnson-Lindenstrauss property is required for a constant $c \in (0,1)$. The RIP is a property that holds uniformly for all sparse vectors such that in Corollary 1.4, no union bound over the probability in $P_\Omega$ is required and $\nu$ can be chosen to be constant and especially independent of $p$.*

*So even though the lower bound on $N$ in Corollary 1.4 implies $\log N \gtrsim \log \frac{1}{\nu}$ in the formula for the lower bond on $m$, the dependence of $m$ on $p$ will only be $m \gtrsim (\log p)^d$.*

## 1.3 Lower Bounds

The goal of this section is to show that our results, especially Corollary 1.4 that we obtain for Hadamard matrices, are optimal with respect to the probability $\eta$. To do this, we apply the Tensor randomized subsampled Hadamard transform to a set of $p$ points. By a union bound and Corollary 1.4, this randomized transform simultaneously preserves the norms of $p$ vectors simultaneously with probability $1 - \nu$ if

$$m \geq C(d) \epsilon^{-2} \left( \log \frac{p}{\nu} \right)^d \left( \log \frac{C(d)}{\epsilon} \right)^2 \log N \left( \log \frac{C(d) \log \frac{p}{\nu}}{\epsilon} \right)^2$$

and

$$N \geq \frac{1}{\nu^{C_1 d \log\log \frac{p}{\nu}}}.$$

We will prove that the dependence $m \gtrsim (\log p)^d$ on $p$ (neglecting double logarithmic factors) is optimal.

Regard the Hadamard transform $H = \mathbb{R}^{N\times N}$ as the Fourier transform on $\mathbb{F}_2^n$ where $N = 2^n$. That is,

$$H_{jk} = (H_n)_{jk} = \frac{1}{\sqrt{N}}(-1)^{\langle j-1,k-1\rangle_b},$$

where $\langle a, b\rangle_b$ denotes the inner product of the binary representations of $a$ and $b$.

Our approach is based on the special behavior of the Hadamard matrix on indicator vectors of subspaces of $\mathbb{F}_2^n$. This principle has been used before to show lower bounds for the restricted isometry property of subsampled Hadamard matrices in [Bla+19] and was then adapted to Johnson-Lindenstrauss embeddings in [BK21].

Denote $\mathbb{G}_{n,r}$ for the set of all $r$-dimensional subspaces of $\mathbb{F}_2^n$. For any subset $M \subset \mathbb{F}_2^n$ we write $\mathbb{1}_M \in \mathbb{R}^N$ for the indicator vector of $M$ normalized such that $\|\mathbb{1}_M\|_2 = 1$. With this notation, it holds for any $V \in \mathbb{G}_{n,r}$ that (see Lemma II.1 in [Bla+19])

$$H\mathbb{1}_V = \mathbb{1}_{V^\perp}$$

Let $P_\Omega \in \mathbb{R}^{m\times N}$ be the matrix representing subsampling $m$ out of $N$ entries independently and uniformly with replacement and rescaling by $\sqrt{\frac{N}{m}}$.

Let $N = N_1 \cdot \dots \cdot N_d$, $N_j = 2^{n_j}$ for $1 \le j \le d$. Consider the matrix

$$A = P_\Omega F D_\xi$$

where $\xi = \xi^{(1)} \otimes \dots \otimes \xi^{(d)}$ is a Kronecker product of $d$ Rademacher vectors, $\xi^{(j)} \in \{\pm 1\}^{N_j}$ and $F \in \mathbb{R}^{N\times N}$ is a bounded orthonormal matrix.

Let $2 \le r \le \min\{n_1, \dots, n_d\}$ and $s = 2^r$. For each $1 \le j \le d$ consider a subspace $V_j \in \mathbb{G}_{n_j,r}$. By taking $F := H_{n_1} \otimes \dots \otimes H_{n_d}$ and $x := \mathbb{1}_{V_1} \otimes \dots \otimes \mathbb{1}_{V_d}$, we obtain

$$y = Fx = (H_{n_1}\mathbb{1}_{V_1}) \otimes \dots \otimes (H_{n_d}\mathbb{1}_{V_d}) = \mathbb{1}_{V_1^\perp} \otimes \dots \otimes \mathbb{1}_{V_d^\perp}.$$

The vector $y$ has $\frac{N_1}{s} \cdot \dots \cdot \frac{N_d}{s} = \frac{N}{s^d}$ entries of size $\sqrt{\frac{s^d}{N}}$. In subsampling with replacement, each selected entry is $\sqrt{\frac{s^d}{N}}$ with probability $\frac{1}{s^d}$ and 0 with probability $1 - \frac{1}{s^d}$. Then

$$\mathbb{P}(P_\Omega y = 0) = (1 - \frac{1}{s^d})^m \ge \exp(-\frac{2m}{s^d}).$$

Now consider the set $E := \{(D_{\hat{\xi}^{(1)}}\mathbb{1}_{V_1}) \otimes \dots \otimes (D_{\hat{\xi}^{(d)}}\mathbb{1}_{V_d}) | \hat{\xi}^{(1)} \in \{\pm 1\}^{N_1}, \dots, \hat{\xi}^{(d)} \in \{\pm 1\}^{N_d}\}$. Corresponding to each factor, there are $2^s$ sign patterns such that $p := |E| \le 2^{ds}$.

With respect to the matrix $A = P_\Omega F D_\xi$, we note that for any value of the random vector $\xi$, there exists $\hat{x} \in E$ such that $D_\xi \hat{x} = x$. Then $A\hat{x} = P_\Omega F D_\xi \hat{x} = P_\Omega y$. We obtain $A\hat{x} = 0$ with probability $\ge$

$$\exp\left(-\frac{2m}{s^d}\right) = \exp\left(-2m\left(\frac{d\log 2}{\log p}\right)^d\right) \tag{1.4}$$

with respect to the randomness in $P_\Omega$.

Altogether, with the probability (1.4),

$$\sup_{x\in E} |\|Ax\|_2 - 1| \ge 1,$$

i.e., the Johnson-Lindenstrauss condition is violated.

To achieve that (1.4) is $\le \nu$, we need that $m \ge \frac{1}{2}(\log\frac{1}{\nu})(\frac{\log p}{d\log 2})^d$.

## 1.4 Notation and Required Tools

For the proof of Theorem 1.3 in the case $d = 2$, we use general order $d'$ arrays $\mathbf{A} \in \mathbb{R}^{[n]^{d'}}$ whose entries $A_{i_1,\dots,i_{d'}}$ are indexed by $d'$ indices $i_1, \dots, i_{d'} \in [n]$ or a tuple $i \in [n]^{d'}$ of $d'$ entries. For $d' = 1$, these are vectors, for $d' = 2$, these are matrices. For $d' = 0$, the index is just one empty tuple and thus we can identify the arrays with real numbers. We define the Frobenius norm $\|\mathbf{A}\|_F = \left(\sum_{i\in[n]^{d'}} A_i^2\right)^{\frac{1}{2}}$. For the $d = 2$ case we will need arrays of order up to 4.

For indices $i \in [n]^4$, for any subset $S \subset [4]$, we denote $i_S$ for the tuple of entries $i_l$ for $l \in S$. In this sense, we use the expression "'for all $i_S$"' in the sense of "'for all choices of all $i_l$, $l \in S$"' and write $\sum_{i_S\in[n]^{|S|}}$ and $\max_{i_S\in[n]^{|S|}}$ for the sum or maximum ranging over all choices of the $i_l$ ($l \in S$) in $[n]$. For an array $\mathbf{A}$ of order $|S|$, we denote $A_{i_S}$ for the entry indexed by the tuple $i_S$.

Using this notation of arrays and indices, we can define the following norms which have been used to bound Gaussian chaos of arbitrary order in [Lat06]. Let $\mathbf{B} \in \mathbb{R}^{[n]^{d'}}$ and denote $S(d', \kappa)$ for the set of all partitions of $[d']$ into $\kappa$ nonempty disjoint sets. Then for each $(I_1, \dots, I_\kappa) \in S(d', \kappa)$, we define

$$\|\mathbf{B}\|_{I_1,\dots,I_\kappa} := \sup\left\{ \sum_{i\in[n]^{d'}} B_i \alpha^{(1)}_{i_{I_1}} \dots \alpha^{(\kappa)}_{i_{I_\kappa}} \,\middle|\, \boldsymbol{\alpha}^{(1)} \in \mathbb{R}^{[n]^{|I_1|}}, \dots, \boldsymbol{\alpha}^{(\kappa)} \in \mathbb{R}^{[n]^{|I_\kappa|}}, \right.$$
$$\left. \|\boldsymbol{\alpha}^{(1)}\|_F = \dots = \|\boldsymbol{\alpha}^{(\kappa)}\|_F = 1 \right\}.$$

The following statement shows that joining some of the partition sets cannot decrease the corresponding norm.

**Lemma 1.6.** *Let $\mathbf{B} \in \mathbb{R}^{[n]^{d'}}$ be an array and $I_1, \dots, I_\kappa$ and $\bar{I}_1, \dots, \bar{I}_{\bar{\kappa}}$ two partitions of $[d']$ into non-empty disjoint sets such that every $\bar{I}_j$ ($1 \le j \le \bar{\kappa}$) is a union of at least one of the sets $I_1, \dots, I_\kappa$. Then*

$$\|\mathbf{B}\|_{I_1,\dots,I_\kappa} \le \|\mathbf{B}\|_{\bar{I}_1,\dots,\bar{I}_{\bar{\kappa}}}.$$

The proof of this is postponed to Subsection 1.5.5.

The following Theorem 1.7 is shown for Gaussian random vectors by Latala in [Lat06]. As explained in Section 1 of [AW15], Theorem 1.4 in [AW15] generalizes the upper bound of Latala's result to subgaussian vectors which is the following statement.

**Theorem 1.7.** *Let $\mathbf{n} \in \mathbb{N}^d$, $\mathbf{B} \in \mathbb{R}^{\mathbf{n}}$, $p \ge 2$.*

*Let $S(\kappa, d)$ denote the set of partitions of $[d]$ into $\kappa$ nonempty disjoint subsets. Define*

$$m_p(\mathbf{B}) := \sum_{\kappa=1}^{d} p^{\kappa/2} \sum_{(I_1,\dots,I_\kappa)\in S(\kappa,d)} \|\mathbf{B}\|_{I_1,\dots,I_\kappa}.$$

*Consider random vectors $X^{(1)} \in \mathbb{R}^{n_1}, \dots, X^{(d)} \in \mathbb{R}^{n_d}$ with independent, mean 0, variance 1 entries with subgaussian norm bounded by $L \ge 1$. Then*

$$\left\| \sum_{i_1\in[n_1],\dots,i_d\in[n_d]} B_{i_1,\dots,i_d} X^{(1)}_{i_1} \dots X^{(d)}_{i_d} \right\|_{L_p} \le C(d) L^d m_p(\mathbf{B}),$$

*where $C(d) > 0$ is a constant that only depends on $d$.*

We also make use of the restricted isometry through the following lemma. Especially, this lemma will be used in Subsection 1.5.5. For a more general overview of the restricted isometry property and similar tools, see Chapter 6 in [FR13].

**Lemma 1.8.** *Let $\Phi \in \mathbb{R}^{m\times N}$ have the $(2s, \delta)$-RIP. Then for any $S, T \subset [N]$ of size $|S| = |T| = s$, the submatrix $B = (\Phi^*\Phi - Id_N)_{S,T}$ satisfies $\|B\|_{2\to 2} \leq \delta$.*

*Proof.* Let $x, y \in \mathbb{R}^N$ such that $\operatorname{supp}(x) = S$, $\operatorname{supp}(y) = T$ and $\|x\|_2 = \|y\|_2 = 1$. Then by the polarization identity and the RIP

$$\begin{aligned} |x_S^* B y_T| &= |x^*\Phi^*\Phi y - x^* y| \\ &= \frac{1}{4}\left|\|\Phi(x+y)\|_2^2 - \|\Phi(x-y)\|_2^2 - \|x+y\|_2^2 + \|x-y\|_2^2\right| \\ &\leq \frac{\delta}{4}\left(\|x+y\|_2^2 + \|x-y\|_2^2\right) = \frac{\delta}{4}\left(2\|x\|_2^2 + 2\|y\|_2^2\right) = \delta. \end{aligned}$$

□

## 1.5 Proof of the Main Theorem for $d = 2$

In this section, we prove the main Theorem 1.3 for the special case $d = 2$. Considering this case will make the presentation of the proof significantly easier compared to the general proof while the most important ideas are still covered. Furthermore, in this section, compared to the general proof in Section 2.5, we restrict the dimensions $n_1, n_2$ of the Rademacher vectors $\xi^{(1)}, \xi^{(2)}$ to be the same $n_1 = n_2 = n$. This assumption does not lead to any essential restrictions in the proof but further simplifies the notation.

A crucial part of the proof is contained in the Lemmas 1.12 and 1.17 which are shown in the general form. These two lemmas will also be used in the general proof in Section 2.5.

### 1.5.1 Overview

This subsection gives a short overview of the proof contained in Subsections 1.5.2 to 1.5.5.

We start by taking an arbitrary signal vector $x \in \mathbb{R}^{n^2}$ which we can assume to satisfy $\|x\|_2 = 1$. Then we arrange the vector $x$ to a matrix $\mathbf{x} \in \mathbb{R}^{n\times n}$ and accordingly the matrix $B := \Phi^*\Phi - Id_N$ to an array $\mathbf{B} \in \mathbb{R}^{n\times n\times n\times n}$ of order 4 in such a way that

$$\|\Phi D_{\xi^{(1)}\otimes\xi^{(2)}} x\|_2^2 - \|x\|_2^2 = \sum_{i\in[n]^4} B_{i_1,\ldots,i_4} x_{i_1,i_2} x_{i_3,i_4} \xi_{i_1}^{(1)} \xi_{i_3}^{(1)} \xi_{i_2}^{(2)} \xi_{i_4}^{(2)}.$$

This expression cannot be controlled by Theorem 1.7 directly since each of the vectors $\xi^{(1)}$, $\xi^{(2)}$ appears twice in each term of the sum.

However, a decoupling technique which is shown in Subsection 1.5.2 shows that it is enough to bound the decoupled chaos

$$\|\Phi D_{\xi^{(1)}\otimes\xi^{(2)}} x\|_2^2 - \|x\|_2^2 = \sum_{i\in[n]^4} B_{i_1,\ldots,i_4} x_{i_1,i_2} x_{i_3,i_4} \xi_{i_1}^{(1)} \xi_{i_3}^{(3)} \xi_{i_2}^{(2)} \xi_{i_4}^{(4)} \tag{1.5}$$

for all $\mathbf{x}$, where $\xi^{(3)}$ and $\xi^{(4)}$ are new independent Rademacher vectors.

This decoupled chaos could be controlled with Theorem 1.7, however, the resulting bound – which would also hold for Gaussian vectors instead of the Rademacher vectors $\xi^{(j)}$ – is not strong enough to prove the theorem.

In order to use special properties of Rademacher vectors in (1.5), we first split up $\mathbf{x}$ as a sum

$$\mathbf{x} = \mathbf{x}^{(\emptyset)} + \mathbf{x}^{(\{1\})} + \mathbf{x}^{(\{2\})} + \mathbf{x}^{(\{1,2\})}$$

of four matrices with disjoint supports. All the entries of $\mathbf{x}$ will be distributed to these four matrices depending on their absolute value. Using the decomposition of $\mathbf{x}$, we can write (1.5) as the sum over

$$\sum_{i\in[n]^4} B_{i_1,\ldots,i_4} x_{i_1,i_2}^{(S)} x_{i_3,i_4}^{(T)} \xi_{i_1}^{(1)} \xi_{i_3}^{(3)} \xi_{i_2}^{(2)} \xi_{i_4}^{(4)}$$

for all $S, T \subset [2]$. For each such expression, now we condition on some of the $\xi^{(l)}$ ($l \in [4]$), bound their entries by $\pm 1$ (which is specific to Rademacher vectors) and then we regard the expression as a chaos in terms of the other $\xi^{(l)}$ which can then be controlled using Theorem 1.7. Then for all choices of $S, T \subset [2]$, we need to control the coefficient array of the corresponding chaos in terms of all its $\|\cdot\|_{I_1,\ldots,I_\kappa}$ norms. Before we derive the bounds on these norms in Subsection 1.5.4, we complete the proof of the main Theorem 1.3 by bounding the $L_p$ norms of $\|\Phi D_{\xi^{(1)}\otimes\xi^{(2)}} x\|_2^2 - \|x\|_2^2$ and therefore $\mathbb{P}\left(\left|\|\Phi D_{\xi^{(1)}\otimes\xi^{(2)}} x\|_2^2 - \|x\|_2^2\right| > \epsilon\right)$.

Subsection 1.5.4 is entirely devoted to bounding the $\|\cdot\|_{I_1,\ldots,I_\kappa}$ norms. Afterwards, Subsection 1.5.5 shows some lemmas that have been used in the previous parts by generalizing techniques used in [KW11].

**Remark 1.9.** *In [Ahl+20a], Lemma 4.11 (TensorSRHT) provides a Johnson-Lindenstrauss result which is similar to our Corollary 1.4 restricted to Hadamard matrices. The proof of this lemma can be found in the extended version [Ahl+20b]. Their proof uses general moment bounds for sums of independent mean* 0 *variables to control the probability in the subsampling $P_\Omega$ while conditioning on the random sign vector $\xi$. In contrast, our approach conditions on the RIP of $P_\Omega H$ and then shows the Johnson-Lindenstrauss property by controlling the probability in $\xi$. This gives an advantage for the case that the Johnson-Lindenstrauss property is shown for p vectors simultaneously. For our approach, once $P_\Omega H$ has the RIP, this holds for all s-sparse vectors uniformly. Then we only need to show the Johnson-Lindenstrauss property by a union bound with respect to the probability in $\xi$ but not with respect to $P_\Omega$. The advantage of this is that in this case the dependence of the embedding dimension in [Ahl+20a] is* $(\log p)^{d+1}$ *(up to smaller logarithmic factors) while our result only requires* $(\log p)^d$ *which the example in Section 1.3 proves to be optimal.*

*On the other hand, our approach makes controlling the probability in $\xi$ more intricate. In [Ahl+20a], Lemma 4.9 provides a result similar to the one by Latala [Lat06] with a better dependence on d but all $\|\cdot\|_{I_1,\ldots,I_\kappa}$ bounded by the Frobenius norm. This suffices to control $\xi^T(D_{H_j}x)$ sufficiently for arbitrary $x \in \mathbb{R}^N$ where $H_j$ is the j-th row of the Hadamard matrix. The latter is required in [Ahl+20a]. In our case, we need to control $\xi^T D_x \Phi^T \Phi D_x \xi$ for which we make use of the RIP of $\Phi$ and control all the $\|\cdot\|_{I_1,\ldots,I_\kappa}$ norms separately. We will discuss more aspects of the relation of our work to [Ahl+20a] in Section 1.6.*

### 1.5.2 Decoupling for $d = 2$

Decoupling is a commonly used technique to relate a chaos of the type $\sum_{j,k=1}^n X_{j,k}\xi_j\xi_k$ with the same vector $(\xi_1, \ldots, \xi_n)$ occurring twice to a decoupled chaos $\sum_{j,k=1}^n X_{j,k}\xi_j\xi'_k$ containing an independent copy $\xi'$ of $\xi$. The statement can be found in numerous textbooks such as [FR13].

**Theorem 1.10** (Theorem 8.11 in [FR13])**.** *Let $\xi_1, \ldots, \xi_n$ be independent, mean* 0 *random variables, $X \in \mathbb{R}^{n\times n}$, and $F : \mathbb{R} \to \mathbb{R}$ a convex function. Then*

$$\mathbb{E}F\left(\sum_{\substack{j,k=1\\ j\neq k}}^n X_{j,k}\xi_j\xi_k\right) \le \mathbb{E}F\left(4\sum_{j,k=1}^n X_{j,k}\xi_j\bar{\xi}_k\right),$$

*where $(\bar{\xi}_1, \ldots, \bar{\xi}_n)$ is an independent copy of $(\xi_1, \ldots, \xi_n)$.*

Note that by taking $F(x) = |x|^p$ for $p \ge 1$ and then taking the $p$-th root, the conclusion can be written as

$$\left\|\sum_{\substack{j,k=1\\ j\neq k}}^n X_{j,k}\xi_j\xi_k\right\|_{L_p} \le 4\left\|\sum_{j,k=1}^n X_{j,k}\xi_j\bar{\xi}_k\right\|_{L_p}. \tag{1.6}$$

Some higher order versions of this decoupling theorem have been developed [Kwa87; AG93]. In the general $d$ case, $\|\Phi D_\xi x\|_2^2 - \|x\|_2^2$ leads to the particular situation of $d$ independent vectors $\xi^{(j)}$, each occurring twice for which we will use Theorem 2.5 developed in [BKW21b] for the general case. To provide a self-contained proof with a simplified notation, we prove the following special case for $d = 2$ by repeated application of Theorem 1.10.

**Lemma 1.11.** *Let $\mathbf{B} \in \mathbb{R}^{n\times n\times n\times n}$ and $\gamma \geq 0$. Let $\xi^{(1)}, \dots, \xi^{(4)} \in \{\pm 1\}^n$ be independent Rademacher vectors. Assume that for all $\mathbf{x} \in \mathbb{R}^{n\times n}$,*

$$\left\| \sum_{i_1,\dots,i_4 \in [n]} B_{i_1,\dots,i_4} \xi^{(1)}_{i_1} \xi^{(2)}_{i_2} \xi^{(3)}_{i_3} \xi^{(4)}_{i_4} x_{i_1,i_2} x_{i_3,i_4} \right\|_{L_p} \leq \gamma \|\mathbf{x}\|_F^2.$$

*Then for all $\mathbf{x} \in \mathbb{R}^{n\times n}$,*

$$\left\| \sum_{i_1,\dots,i_4 \in [n]} B_{i_1,\dots,i_4} \xi^{(1)}_{i_1} \xi^{(2)}_{i_2} \xi^{(1)}_{i_3} \xi^{(2)}_{i_4} x_{i_1,i_2} x_{i_3,i_4} \right\|_{L_p} \leq 25\gamma \|\mathbf{x}\|_F^2.$$

*Proof.* By repeatedly separating diagonal entries and applying the decoupling Theorem 1.10, we obtain (in all sums each index ranges over $[n]$)

$$\begin{aligned}
&\left\| \sum_{i_2,i_4} \left( \sum_{i_1,i_3} B_{i_1,\dots,i_4} \xi^{(1)}_{i_1} \xi^{(1)}_{i_3} x_{i_1,i_2} x_{i_3,i_4} \right) \xi^{(2)}_{i_2} \xi^{(2)}_{i_4} \right\|_{L_p} \\
&\leq \left\| \sum_{\substack{i_2,i_4 \\ i_2 \neq i_4}} \left( \sum_{i_1,i_3} B_{i_1,\dots,i_4} \xi^{(1)}_{i_1} \xi^{(1)}_{i_3} x_{i_1,i_2} x_{i_3,i_4} \right) \xi^{(2)}_{i_2} \xi^{(2)}_{i_4} \right\|_{L_p} \\
&+ \left\| \sum_{i_2} \left( \sum_{i_1,i_3} B_{i_1,i_2,i_3,i_2} \xi^{(1)}_{i_1} \xi^{(1)}_{i_3} x_{i_1,i_2} x_{i_3,i_2} \right) \right\|_{L_p} \\
&\leq 4 \left\| \sum_{i_2,i_4} \left( \sum_{i_1,i_3} B_{i_1,\dots,i_4} \xi^{(1)}_{i_1} \xi^{(1)}_{i_3} x_{i_1,i_2} x_{i_3,i_4} \right) \xi^{(2)}_{i_2} \xi^{(4)}_{i_4} \right\|_{L_p} \\
&+ \left\| \sum_{i_1,i_3} \left( \sum_{i_2} B_{i_1,i_2,i_3,i_2} x_{i_1,i_2} x_{i_3,i_2} \right) \xi^{(1)}_{i_1} \xi^{(1)}_{i_3} \right\|_{L_p} \\
&\leq 4 \left\| \sum_{i_1,i_3} \left( \sum_{i_2,i_4} B_{i_1,\dots,i_4} \xi^{(2)}_{i_2} \xi^{(4)}_{i_4} x_{i_1,i_2} x_{i_3,i_4} \right) \xi^{(1)}_{i_1} \xi^{(1)}_{i_3} \right\|_{L_p} \\
&+ \left\| \sum_{\substack{i_1,i_3 \\ i_1 \neq i_3}} \left( \sum_{i_2} B_{i_1,i_2,i_3,i_2} x_{i_1,i_2} x_{i_3,i_2} \right) \xi^{(1)}_{i_1} \xi^{(1)}_{i_3} \right\|_{L_p} + \left\| \sum_{i_1,i_2} B_{i_1,i_2,i_1,i_2} x_{i_1,i_2} x_{i_1,i_2} \right\|_{L_p} \\
&\leq 4 \left\| \sum_{\substack{i_1,i_3 \\ i_1 \neq i_3}} \left( \sum_{i_2,i_4} B_{i_1,\dots,i_4} \xi^{(2)}_{i_2} \xi^{(4)}_{i_4} x_{i_1,i_2} x_{i_3,i_4} \right) \xi^{(1)}_{i_1} \xi^{(1)}_{i_3} \right\|_{L_p} \\
&+ 4 \left\| \sum_{i_1} \left( \sum_{i_2,i_4} B_{i_1,i_2,i_1,i_4} \xi^{(2)}_{i_2} \xi^{(4)}_{i_4} x_{i_1,i_2} x_{i_1,i_4} \right) \right\|_{L_p}
\end{aligned}$$

$$\begin{aligned}
&+4\left\|\sum_{i_1,i_3}\left(\sum_{i_2} B_{i_1,i_2,i_3,i_2} x_{i_1,i_2} x_{i_3,i_2}\right)\xi^{(1)}_{i_1}\xi^{(3)}_{i_3}\right\|_{L_p} + \left\|\sum_{i_1,i_2} B_{i_1,i_2,i_1,i_2} x_{i_1,i_2} x_{i_1,i_2}\right\|_{L_p}\\
&\le 16\left\|\sum_{i_1,i_2,i_3,i_4} B_{i_1,\dots,i_4}\xi^{(1)}_{i_1}\xi^{(2)}_{i_2}\xi^{(3)}_{i_3}\xi^{(4)}_{i_4} x_{i_1,i_2}x_{i_3,i_4}\right\|_{L_p}\\
&+4\left\|\sum_{i_1,i_2,i_4} B_{i_1,i_2,i_1,i_4}\xi^{(2)}_{i_2}\xi^{(4)}_{i_4}x_{i_1,i_2}x_{i_1,i_4}\right\|_{L_p}\\
&+4\left\|\sum_{i_1,i_2,i_3} B_{i_1,i_2,i_3,i_2}\xi^{(1)}_{i_1}\xi^{(3)}_{i_3}x_{i_1,i_2}x_{i_3,i_2}\right\|_{L_p} + \left\|\sum_{i_1,i_2} B_{i_1,i_2,i_1,i_2}x_{i_1,i_2}x_{i_1,i_2}\right\|_{L_p}.
\end{aligned}\tag{1.7}$$

For $i'_1 \in [n]$, define $\mathbf{x}^{(i'_1)} \in \mathbb{R}^{n\times n}$ by

$$x^{(i'_1)}_{i_1,i_2} = \begin{cases} x_{i_1,i_2} & \text{if } i_1 = i'_1\\ 0 & \text{otherwise.}\end{cases}$$

Then all the $x^{(i'_1)}$ have disjoint supports and

$$\mathbf{x} = \sum_{i'_1\in[n]} \mathbf{x}^{(i'_1)}.$$

Then

$$\begin{aligned}
&\left\|\sum_{i_1,i_2,i_4} B_{i_1,i_2,i_1,i_4}\xi^{(2)}_{i_2}\xi^{(4)}_{i_4}x_{i_1,i_2}x_{i_1,i_4}\right\|_{L_p} = \left\|\sum_{i_1,i_2,i_4} B_{i_1,i_2,i_1,i_4}\xi^{(2)}_{i_2}\xi^{(4)}_{i_4}x^{(i_1)}_{i_1,i_2}x^{(i_1)}_{i_1,i_4}\right\|_{L_p}\\
&\le \sum_{i_1}\left\|\sum_{i_2,i_4} B_{i_1,i_2,i_1,i_4}\xi^{(2)}_{i_2}\xi^{(4)}_{i_4}x^{(i_1)}_{i_1,i_2}x^{(i_1)}_{i_1,i_4}\right\|_{L_p}\\
&= \sum_{i_1}\left\|\sum_{i'_1,i_2,i_3,i_4} B_{i'_1,i_2,i_3,i_4}\xi^{(1)}_{i'_1}\xi^{(2)}_{i_2}\xi^{(3)}_{i_3}\xi^{(4)}_{i_4}x^{(i_1)}_{i'_1,i_2}x^{(i_1)}_{i_3,i_4}\right\|_{L_p} \le \sum_{i_1}\gamma\|\mathbf{x}^{(i_1)}\|_F^2 = \gamma\|\mathbf{x}\|_F^2.
\end{aligned}$$

where in the third step we used that $x^{(i_1)}_{i'_1,i_2}x^{(i_1)}_{i_3,i_4} = 0$ except for the one term $i'_1 = i_3 = i_1$ in which $\xi^{(1)}_{i'_1}\xi^{(3)}_{i_3} = \pm 1$.

In (1.7), the first term on the right hand side can be bounded by $16\gamma\|\mathbf{x}\|_F^2$ by assumption, the above argument shows that the second term is $\le 4\gamma\|\mathbf{x}\|_F^2$, and the same bound follows for the third term analogously. For the fourth term we can define arrays $\mathbf{x}^{(i'_1,i'_2)} \in \mathbb{R}^{n\times n}$ for any $i'_1, i'_2 \in [n]$ by

$$x^{(i'_1,i'_2)}_{i_1,i_2} = \begin{cases} x_{i_1,i_2} & \text{if } i'_1 = i_1 \text{ and } i'_2 = i_2\\ 0 & \text{otherwise.}\end{cases}$$

Then we can do an analogous argument and also bound the fourth term by $\gamma\|\mathbf{x}\|_F^2$. So altogether it follows that

$$\left\|\sum_{i_1,i_2,i_3,i_4\in[n]} B_{i_1,\dots,i_4}\xi^{(1)}_{i_1}\xi^{(2)}_{i_2}\xi^{(1)}_{i_3}\xi^{(2)}_{i_4}x_{i_1,i_2}x_{i_3,i_4}\right\|_{L_p} \le 25\gamma.$$

□

### 1.5.3 Proof of Theorem 1.3 for $d = 2$

For the $d = 2$ case, the signal vector $x \in \mathbb{R}^{n^2}$ is entry-wise multiplied by the Kronecker product $\xi^{(1)} \otimes \xi^{(2)}$ of Rademacher vectors $\xi^{(1)}, \xi^{(2)} \in \mathbb{R}^n$. The Kronecker product $\xi^{(1)} \otimes \xi^{(2)} \in \mathbb{R}^{n^2}$ is a rearrangement of the entries of the matrix $\xi^{(1)}(\xi^{(2)})^* \in \mathbb{R}^{n \times n}$. We define $\mathcal{I}_2 : [n] \times [n] \to [n^2]$ to be the bijective map that maps pairs of row/column index of $\xi^{(1)}(\xi^{(2)})^* \in \mathbb{R}^{n\times n}$ to the corresponding index of $\xi^{(1)} \otimes \xi^{(2)} \in \mathbb{R}^{n^2}$, i.e., for all $j, k \in [n]$,

$$\left(\xi^{(1)}(\xi^{(2)})^*\right)_{j,k} = \xi^{(1)}_j \xi^{(2)}_k = \left(\xi^{(1)} \otimes \xi^{(2)}\right)_{\mathcal{I}_2(j,k)}. \tag{1.8}$$

Now we rearrange the vector $x$ in the same way to a matrix $\mathbf{x} \in \mathbb{R}^{n\times n}$. Then the entry-wise multiplication of $\xi^{(1)} \otimes \xi^{(2)}$ and $x$ corresponds to entry-wise multiplication of $\xi^{(1)}(\xi^{(2)})^*$ and $\mathbf{x}$. We consider the matrix $B := \Phi^*\Phi - Id_N$ and rearrange its rows and columns each in the same way we rearranged $x$ to $\mathbf{x}$ to obtain the array $\mathbf{B} \in \mathbb{R}^{n\times n\times n\times n}$ of order 4 (i.e., $B_{i_1,i_2,i_3,i_4} = B_{\mathcal{I}_2(i_1,i_2),\mathcal{I}_2(i_3,i_4)}$). Then

$$\begin{aligned}\|\Phi D_{\xi^{(1)}\otimes\xi^{(2)}} x\|_2^2 - \|x\|_2^2 &= x^* D_{\xi^{(1)}\otimes\xi^{(2)}}(\Phi^*\Phi - Id_N) D_{\xi^{(1)}\otimes\xi^{(2)}} x \\ &= \sum_{i,j\in[n]} B_{i,j} (\xi^{(1)} \otimes \xi^{(2)})_i (\xi^{(1)} \otimes \xi^{(2)})_j x_i x_j \\ &= \sum_{i_1,\ldots,i_4\in[n]} B_{i_1,\ldots,i_4} \xi^{(1)}_{i_1}\xi^{(2)}_{i_2}\xi^{(1)}_{i_3}\xi^{(2)}_{i_4} x_{i_1,i_2} x_{i_3,i_4}.\end{aligned}$$

Our goal is to bound

$$\left\| \|\Phi D_{\xi^{(1)}\otimes\xi^{(2)}} x\|_2^2 - \|x\|_2^2 \right\|_{L_p} = \left\| \sum_{i_1,\ldots,i_4\in[n]} B_{i_1,\ldots,i_4} \xi^{(1)}_{i_1}\xi^{(2)}_{i_2}\xi^{(1)}_{i_3}\xi^{(2)}_{i_4} x_{i_1,i_2} x_{i_3,i_4} \right\|_{L_p}$$

for all $\mathbf{x} \in \mathbb{R}^{n\times n}$. By Lemma 1.11, it is sufficient to bound

$$\left\| \sum_{i_1,\ldots,i_4\in[n]} B_{i_1,\ldots,i_4} \xi^{(1)}_{i_1}\xi^{(2)}_{i_2}\xi^{(3)}_{i_3}\xi^{(4)}_{i_4} x_{i_1,i_2} x_{i_3,i_4} \right\|_{L_p}$$

for all $\mathbf{x} \in \mathbb{R}^{n\times n}$ and by homogeneity it is enough to do this for all such $\mathbf{x}$ satisfying $\|\mathbf{x}\|_F = 1$.

So consider any $\mathbf{x} \in \mathbb{R}^{n\times n}$ with $\|\mathbf{x}\|_F = 1$. We split the matrix $\mathbf{x}$ up into the sum

$$\mathbf{x} = \mathbf{x}^{(\emptyset)} + \mathbf{x}^{(\{1\})} + \mathbf{x}^{(\{2\})} + \mathbf{x}^{(\{1,2\})}$$

of four matrices with disjoint support where

- $\mathbf{x}^{(\{1,2\})}$ contains the $s^2$ largest entries of $\mathbf{x}$.
- $\mathbf{x}^{(\{2\})}$ contains the largest $s$ of the remaining entries of every row.
- $\mathbf{x}^{(\{1\})}$ contains the largest $s$ of the now remaining entries of every column.
- $\mathbf{x}^{(\emptyset)}$ contains all the entries that are still remaining.

In all cases, "largest“ refers to the corresponding entries with the largest absolute value. We pick one such choice even if it is not unique.

By the definition these matrices, in every row $i_1 \in [n]$, the $s$ largest entries of $(x_{i_1,i_2})_{i_2\in[n]}$ are not contained in $\mathbf{x}^{(\emptyset)}$. This implies that for each $i_1 \in [n]$ there are at least $s$ indices $i_2 \in [n]$ such that $|x_{i_1,i_2}| \geq \max_{i_2'\in[n]} |x^{(\emptyset)}_{i_1,i_2'}|$. This implies that for each $i_1 \in [n]$, $\max_{i_2\in[n]} (x^{(\emptyset)}_{i_1,i_2})^2 \leq \frac{1}{s}\sum_{i_2\in[n]} x^2_{i_1,i_2}$.

With analogous arguments and $\|\mathbf{x}\|_F = 1$, we can obtain all of the following statements.

$$\begin{aligned} \max_{(i_1,i_2)\in[n]^2} (x^{(\emptyset)}_{i_1,i_2})^2 &\le \frac{1}{s^2} \\ \max_{(i_1,i_2)\in[n]^2} (x^{(\{1\})}_{i_1,i_2})^2 &\le \frac{1}{s^2} \\ \max_{(i_1,i_2)\in[n]^2} (x^{(\{2\})}_{i_1,i_2})^2 &\le \frac{1}{s^2} \\ \text{For each } i_1 \in [n], \quad \max_{i_2\in[n]} (x^{(\emptyset)}_{i_1,i_2})^2 &\le \frac{1}{s} \sum_{i_2\in[n]} x^2_{i_1,i_2} \\ \text{For each } i_2 \in [n], \quad \max_{i_1\in[n]} (x^{(\emptyset)}_{i_1,i_2})^2 &\le \frac{1}{s} \sum_{i_1\in[n]} x^2_{i_1,i_2}. \end{aligned} \tag{1.9}$$

The definitions of the $\mathbf{x}^{(S)}$ directly imply that

$$\begin{aligned} x^{(\{1,2\})}_{i_1,i_2} &\neq 0 \text{ for at most } s^2 \text{ pairs } (i_1,i_2) \in [n]^2 \\ \text{For every } i_2 \in [n],\ x^{(\{1\})}_{i_1,i_2} &\neq 0 \text{ for at most } s \text{ indices } i_1 \in [n] \\ \text{For every } i_1 \in [n],\ x^{(\{2\})}_{i_1,i_2} &\neq 0 \text{ for at most } s \text{ indices } i_2 \in [n]. \end{aligned} \tag{1.10}$$

We have split up $\mathbf{x}$ into $\mathbf{x} = \sum_{S\subset[2]} \mathbf{x}^{(S)}$. Thus we obtain

$$\sum_{i_1,\ldots,i_4\in[n]} B_{i_1,\ldots,i_4} \xi^{(1)}_{i_1} \xi^{(2)}_{i_2} \xi^{(3)}_{i_3} \xi^{(4)}_{i_4} x_{i_1,i_2} x_{i_3,i_4} = \sum_{S,T\subset[2]} X^{(S,T)}$$

where

$$\begin{aligned} X^{(S,T)} &:= \sum_{i_1,\ldots,i_4\in[n]} B_{i_1,\ldots,i_4} \xi^{(1)}_{i_1} \xi^{(2)}_{i_2} \xi^{(3)}_{i_3} \xi^{(4)}_{i_4} x^{(S)}_{i_1,i_2} x^{(T)}_{i_3,i_4} \\ &= \sum_{i_1,\ldots,i_4\in[n]} B_{i_1,\ldots,i_4} x^{(S)}_{i_1,i_2} x^{(T)}_{i_3,i_4} \prod_{l=1}^{4} \xi^{(l)}_{i_l} = \sum_{i_{U^c}\in[n]^{|U^c|}} B^{(S,T)}_{i_{U^c}} \prod_{l\in U^c} \xi^{(l)}_{i_l} \end{aligned}$$

where $U = S \cup (T+2) \subset [4]$ and

$$B^{(S,T)}_{i_{U^c}} = \sum_{i_U\in[n]^{|U|}} B_{i_1,\ldots,i_4} x^{(S)}_{i_1,i_2} x^{(T)}_{i_3,i_4} \prod_{l\in U} \xi^{(l)}_{i_l}$$

such that $\mathbf{B}^{(S,T)} \in \mathbb{R}^{[n]^{|U^c|}}$.

For example $S = \{1\}$, $T = \{2\}$, then $U = \{2,3\}$ and

$$X^{(\{1\},\{2\})} = \sum_{i_2,i_3\in[n]} B^{(\{1\},\{2\})}_{i_2,i_3} \xi^{(2)}_{i_2} \xi^{(3)}_{i_3},$$

where

$$B^{(\{1\},\{2\})}_{i_2,i_3} = \sum_{i_1,i_4\in[n]} B_{i_1,\ldots,i_4} x^{(\{1\})}_{i_1,i_2} x^{(2)}_{i_3,i_4} \xi^{(1)}_{i_1} \xi^{(4)}_{i_4}.$$

The next step is to condition on $(\xi^{(l)})_{l\in U}$ and bound the $L_p$ norm of $X^{(S,T)}$ only with respect to $(\xi^{(l)})_{l\in U^c}$, in such a way that the bound is deterministic. Then the same bound also holds for the total $L_p$ norm because if $\mathbb{E}\left[|X^{(S,T)}|^p \,\middle|\, (\xi^{(l)})_{l\in U}\right] \le C$ for deterministic $C$, then

$$\mathbb{E}\left[|X^{(S,T)}|^p\right] = \mathbb{E}\left[\mathbb{E}\left[|X^{(S,T)}|^p \,\middle|\, (\xi^{(l)})_{l\in U}\right]\right] \le \mathbb{E}[C] = C.$$

In terms of $(\xi^{(l)})_{l\in U^c}$, $X^{(S,T)}$ is a chaos of order $|U^c|$. If $(S,T)\neq([2],[2])$, then $U^c\neq\emptyset$ and by Theorem 1.7,

$$\left\|X^{(S,T)}\right\|_{L_p}\leq Cm_p(\mathbf{B}^{(S,T)}), \tag{1.11}$$

where

$$m_p(\mathbf{B}^{(S,T)}):=\sum_{\kappa=1}^{|U^c|}p^{\kappa/2}\sum_{(I_1,\dots,I_\kappa)\in S(\kappa,|U^c|)}\|\mathbf{B}^{(S,T)}\|_{I_1,\dots,I_\kappa},$$

so our goal is to bound $\|\mathbf{B}^{(S,T)}\|_{I_1,\dots,I_\kappa}$ for all $S,T\subset[2]$ and $I_1,\dots,I_\kappa\in S(\kappa,4-|S|-|T|)$. We can bijectively map $[4-|S|-|T|]=[|U^c|]$ to $U^c$ such that instead of considering all partitions $I_1,\dots,I_\kappa$ of $[4-|S|-|T|]$, we consider the partitions of $U^c$ for the norms. This makes the notation easier such that we can write

$$\begin{aligned}\|\mathbf{B}^{(S,T)}\|_{I_1,\dots,I_\kappa}&=\sup_{\substack{\boldsymbol{\alpha}^{(1)}\in\mathbb{R}^{[n]^{|I_1|}},\dots,\boldsymbol{\alpha}^{(\kappa)}\in\mathbb{R}^{[d]^{|I_\kappa|}}\\ \|\boldsymbol{\alpha}^{(1)}\|_F=\dots=\|\boldsymbol{\alpha}^{(\kappa)}\|_F=1}}\sum_{i_{U^c}\in[n]^{|U^c|}}B^{(S,T)}_{i_{U^c}}\alpha^{(1)}_{i_{I_1}}\dots\alpha^{(\kappa)}_{i_{I_\kappa}}\\ &=\sup_{\substack{\boldsymbol{\alpha}^{(1)}\in\mathbb{R}^{[n]^{|I_1|}},\dots,\boldsymbol{\alpha}^{(\kappa)}\in\mathbb{R}^{[d]^{|I_\kappa|}}\\ \|\boldsymbol{\alpha}^{(1)}\|_F=\dots=\|\boldsymbol{\alpha}^{(\kappa)}\|_F=1}}\sum_{i\in[n]^4}B_{i_1,\dots,i_4}x^{(S)}_{i_1,i_2}x^{(T)}_{i_3,i_4}\left(\prod_{l\in U}\xi^{(l)}_{i_l}\right)\alpha^{(1)}_{i_{I_1}}\dots\alpha^{(\kappa)}_{i_{I_\kappa}}.\end{aligned} \tag{1.12}$$

We will show

$$\|\mathbf{B}^{(S,T)}\|_{I_1,\dots,I_\kappa}\leq C\frac{\delta}{s^{\frac{\kappa}{2}}}, \tag{1.13}$$

for all $S,T\subset[2]$, $(S,T)\neq([2],[2])$ and all partitions $I_1,\dots,I_\kappa$ of $\{1,2,3,4\}\backslash(S\cup(T+2))$ where $C>0$ is a constant and $\delta$ comes from the RIP assumption of $\Phi$. The case $(S,T)=([2],[2])$ will be considered separately.

However, we postpone the proof of (1.13) separately to Subsection 1.5.4 and continue to complete the proof of Theorem 1.3 here under the assumption that it holds for all $S,T$, $(S,T)\neq([2],[2])$.

From (1.13) for all $S,T$, $(S,T)\neq([2],[2])$, we can conclude using (1.11) that for all such $S$, $T$,

$$\left\|X^{(S,T)}\right\|_{L_p}\leq C_1\sum_{\kappa=1}^{4-|S|-|T|}p^{\kappa/2}\frac{\delta}{s^{\kappa/2}},$$

where $C_1>0$ is a constant.

For the remaining case $S=T=\{1,2\}$, we observe that $\mathbf{x}^{(\{1,2\})}$ is $s^2$-sparse and thus

$$\begin{aligned}|X^{([2],[2])}|&=\left|\sum_{i_1,\dots,i_4\in[n]}B_{i_1,\dots,i_4}x^{(S)}_{i_1,i_2}x^{(T)}_{i_3,i_4}\prod_{l=1}^{4}\xi^{(l)}_{i_l}\right|\\ &=\left|\left(\operatorname{vec}(\mathbf{x}^{(\{1,2\})})\circ(\xi^{(1)}\otimes\xi^{(2)})\right)^*B\left(\operatorname{vec}(\mathbf{x}^{(\{1,2\})})\circ(\xi^{(1)}\otimes\xi^{(2)})\right)\right|\\ &\leq\delta\|\mathbf{x}^{(\{1,2\})}\|_F^2\leq\delta\end{aligned}$$

by Lemma 1.8 where $\circ$ denotes the entry-wise product and $\operatorname{vec}(\mathbf{x}^{(\{1,2\})})\in\mathbb{R}^{n^2}$ is the vectorized rearrangement of the matrix $\mathbf{x}^{(\{1,2\})}$ (i.e. $\operatorname{vec}(\mathbf{x}^{(\{1,2\})})_{\mathcal{I}_2(i_1,i_2)}=x^{(\{1,2\})}_{i_1,i_2}$).

We obtain that

$$\left\|\sum_{i_1,\dots,i_4\in[n]}B_{i_1,\dots,i_4}\xi^{(1)}_{i_1}\xi^{(2)}_{i_2}\xi^{(3)}_{i_3}\xi^{(4)}_{i_4}x_{i_1,i_2}x_{i_3,i_4}\right\|_{L_p}\leq\sum_{S,T\subset[2]}\|X^{(S,T)}\|_{L_p}$$

$$\leq \delta + \sum_{\substack{S,T \subset [2] \\ (S,T) \neq ([2],[2])}} \|X^{(S,T)}\|_{L_p} \leq \delta + 4^2 C_1 \sum_{\kappa=1}^{4} p^{\kappa/2} \frac{\delta}{s^{\kappa/2}} \leq C_2 \delta \sum_{\kappa=0}^{4} \left(\frac{p}{s}\right)^{\frac{\kappa}{2}},$$

where $C_2 = 4^2 C_1$, assuming $C_2 \geq 1$.

As we have shown using the decoupling Lemma 1.11, the fact that the above inequality holds for all $\mathbf{x} \in \mathbb{R}^{n \times n}$ with $\|\mathbf{x}\|_F = 1$ is enough to show that for all $x \in \mathbb{R}^{n^2}$, $\|x\|_2 = 1$,

$$\left\| \|\Phi D_{\xi^{(1)} \otimes \xi^{(2)}} x\|_2^2 - \|x\|_2^2 \right\|_{L_p} \leq C_3 \delta \sum_{\kappa=0}^{4} \left(\frac{p}{s}\right)^{\frac{\kappa}{2}},$$

where $C_3 = 25 C_2$.

We assume $\delta \leq \frac{\epsilon}{5eC_3}$ and obtain for the particular choice $p = s \geq 2$,

$$\left\| \|\Phi D_{\xi^{(1)} \otimes \xi^{(2)}} x\|_2^2 - \|x\|_2^2 \right\|_{L_s} \leq C_3 \delta \sum_{\kappa=0}^{4} 1 = 5 C_3 \delta.$$

With Markov's inequality, we obtain

$$\begin{aligned} \mathbb{P}\left( \left| \|\Phi D_{\xi^{(1)} \otimes \xi^{(2)}} x\|_2^2 - \|x\|_2^2 \right| > \epsilon \right) &\leq \left( \frac{\left\| \|\Phi D_{\xi^{(1)} \otimes \xi^{(2)}} x\|_2^2 - \|x\|_2^2 \right\|_{L_s}}{\epsilon} \right)^s \\ &\leq \left( \frac{5 C_3 \delta}{5 e C_3 \delta} \right)^s = e^{-s}. \end{aligned}$$

By assumption, $s \geq \log \frac{1}{\eta}$ such that this probability is $\leq \eta$ which completes the proof of Theorem 1.3.

### 1.5.4 Bounding the Tensor Norms

This entire subsection is devoted to the proof of (1.13) for all required cases.

Take sets $S, T \subset \{1,2\}$, such that $(S,T) \neq (\{1,2\},\{1,2\})$ and any partition $I_1, \dots, I_\kappa$ of $[4] \backslash (S \cup (T+2))$ into non-empty disjoint sets.

We define $\bar{I} \subset \{1,2\}$ to be the union of all sets among $I_1, \dots, I_\kappa$ that are contained in $\{1,2\}$, $\bar{I}' \subset \{3,4\}$ the union of all $I_1, \dots, I_\kappa$ that are contained in $\{3,4\}$ and $\bar{\bar{J}} \subset [4]$ the union of all other sets of the partition. Furthermore, define $\bar{J} := \bar{\bar{J}} \cap \{1,2\}$ and $\bar{J}' = \bar{\bar{J}} \cap \{3,4\}$, such that $S \cup \bar{I} \cup \bar{J} = \{1,2\}$ and $(T+2) \cup \bar{I}' \cup \bar{J}' = \{3,4\}$. Then $\bar{I}, \bar{I}', \bar{\bar{J}}$ is again a partition of $[4] \backslash (S \cup (T+2))$. However, these three sets might not all be non-empty. Since joining some of the partition sets does not increase the $\|\cdot\|_{I_1,\dots,I_\kappa}$ norm (Lemma 1.6), we obtain

$$\|\mathbf{B}^{(S,T)}\|_{I_1,\dots,I_\kappa} \leq \|\mathbf{B}^{(S,T)}\|_{\bar{I},\bar{I}',\bar{\bar{J}}},$$

where we denote $\|\cdot\|_{\bar{I},\bar{I}',\bar{\bar{J}}}$ for the norm corresponding to the partition obtained by restricting $\bar{I}, \bar{I}', \bar{\bar{J}}$ to the non-empty sets among them.

So we need to show

$$\|\mathbf{B}^{(S,T)}\|_{\bar{I},\bar{I}',\bar{\bar{J}}} \leq C \frac{\delta}{s^{\frac{\kappa}{2}}} \tag{1.14}$$

for a constant $C > 0$ in all cases. Note however, that this inequality still contains the cardinality $\kappa$ of the original partition $I_1, \dots, I_\kappa$.

In general, we have according to (1.12),

$$\|\mathbf{B}^{(S,T)}\|_{\bar{I},\bar{I}',\bar{\bar{J}}}$$

$$= \sup_{\substack{\boldsymbol{\alpha}^{(1)} \in \mathbb{R}^{[n]^{|\bar{I}|}} \boldsymbol{\alpha}^{(2)} \in \mathbb{R}^{[n]^{|\bar{I}'|}}, \\ \boldsymbol{\alpha}^{(3)} \in \mathbb{R}^{[n]^{|\bar{\bar{J}}|}} \\ \|\boldsymbol{\alpha}^{(1)}\|_F = \|\boldsymbol{\alpha}^{(2)}\|_F \\ = \|\boldsymbol{\alpha}^{(3)}\|_F = 1}} \sum_{i \in [n]^4} B_{i_1,\dots,i_4} x^{(S)}_{i_1,i_2} x^{(T)}_{i_3,i_4} \alpha^{(1)}_{i_{\bar{I}}} \alpha^{(2)}_{i_{\bar{I}'}} \alpha^{(3)}_{i_{\bar{\bar{J}}}} \prod_{l \in S \cup (T+2)} \xi^{(l)}_{i_l}. \tag{1.15}$$

Note that it can happen that, for example, $\bar{I} = \emptyset$. In this case we take the supremum over the array $\boldsymbol{\alpha}^{(1)} \in \mathbb{R}^{[n]^0}$ of order 0 with Frobenius norm 1. We regard this as a real number with absolute value 1 which leads to the same effect in the expression as dropping $\bar{I}$ and considering $\|\cdot\|_{\bar{I}',\bar{\bar{J}}}$ instead. The same holds if some of $\bar{I}', \bar{\bar{J}}$ are empty. Because of $(S,T) \neq ([2],[2])$, not all of them can be empty.

For example, if $S = \{1\}$, $\bar{I} = \emptyset$, $\bar{\bar{J}} = \{2,4\}$, $\bar{I}' = \{3\}$ and $T = \emptyset$, then

$$\|\mathbf{B}^{(S,T)}\|_{\bar{I},\bar{I}',\bar{\bar{J}}} = \sup_{\substack{\alpha^{(1)} \in \mathbb{R}, \alpha^{(2)} \in \mathbb{R}^n \\ \alpha^{(3)} \in \mathbb{R}^{n \times n} \\ |\alpha^{(1)}| = \|\alpha^{(2)}\|_2 \\ = \|\alpha^{(3)}\|_F = 1}} \sum_{i \in [n]^4} B_{i_1,\dots,i_4} x^{(\{1\})}_{i_1,i_2} \xi^{(1)}_{i_1} \alpha^{(1)} x^{(\emptyset)}_{i_3,i_4} \alpha^{(2)}_{i_3} \alpha^{(3)}_{i_2,i_4}.$$

Our goal is to bound the expressions (1.15) in all cases. However, depending on the choices of $S, T, \bar{I}, \bar{I}', \bar{\bar{J}}$, this expression has a different shape. Since already in the case $d = 2$ for example, there are in total 41 possibilities to choose these sets, we develop a unified approach to handle all these expressions. For this, we first observe that $S, \bar{I}, \bar{J}, T+2, \bar{I}', \bar{J}'$ is a partition of $\{1,2,3,4\}$ in which some sets are empty and each set has $\leq 2$ elements. In this respect, we split up the entire index tuple $(i_1, i_2, i_3, i_4)$ of the sum in (1.15) into the (partly empty) tuples $i_{\bar{J}}$, $i_{\bar{I}}$, $i_S$, $i_{\bar{J}'}$, $i_{\bar{I}'}$, $i_{T+2}$. These tuples will then be mapped to integers $j$, $k$, $l$, $j'$, $k'$, $l'$, respectively. Specifically, if $\bar{J} = \emptyset$, then $i_{\bar{J}}$ will be empty and thus there is just one possible value for it and we define $j = 1$. If $|\bar{J}| = 1$, i.e., $\bar{J} = \{r\}$ for some $r \in [4]$, then $i_{\bar{J}} = i_r$ and we define $j = i_r$. If $|\bar{J}| = 2$, $\bar{J} = \{r_1, r_2\}$, $r_1 < r_2$, then $i_{\bar{J}} = (i_{r_1}, i_{r_2})$ and we map these tuples of two indices in $[n]$ to one integer in $[n^2]$ using the function $\mathcal{I}_2 : [n]^2 \to [n^2]$ (see the explanation before (1.8) for the definition of $\mathcal{I}_2$). In this way, the set of all $j$ obtained in this way, is always $[n^{|\bar{J}|}]$. We do the same for the other sets besides $\bar{J}$ to obtain the other indices besides $j$.

After this rearrangement of the indices, the factor $x^{(S)}_{i_1,i_2} \prod_{r \in S} \xi^{(r)}_{i_r}$ in (1.15) that previously depended on the indices $i_1, i_2$, will now depend on $j, k, l$ (where for $d = 2$, at least one of the three indices can only take the value 1). Therefore, we will rearrange these $x^{(S)}_{i_1,i_2} \prod_{r \in S} \xi^{(r)}_{i_r}$ into an array $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ whose entries $X_{i,j,l}$ depend on three indices. In a similar way, we will rearrange the other factors in (1.15) in terms of the new indices and then obtain the following arrays.

$$\begin{aligned} x^{(S)}_{i_1,i_2} \prod_{r \in S} \xi^{(r)}_{i_r} &\to X_{j,k,l} & \qquad x^{(T)}_{i_3,i_4} \prod_{r \in T+2} \xi^{(r)}_{i_r} &\to Y_{j',k',l'} \\ x_{i_1,i_2} &\to \bar{X}_{j,k,l} & x_{i_1,i_2} &\to \bar{Y}_{j',k',l'} \\ \alpha^{(1)}_{i_{\bar{I}}} &\to \alpha_k & \alpha^{(2)}_{i_{\bar{I}'}} &\to \alpha'_{k'} \\ \alpha^{(3)}_{i_{\bar{\bar{J}}}} &\to \Gamma_{j,j'}. \end{aligned} \tag{1.16}$$

The precise definition of all these objects will be given later. However, this overview is given to demonstrate that with such a rearrangement of indices and entries, we will be able to rewrite the sum in (1.15) in the form

$$\sum_{\substack{(j,j') \in [n_1] \times [n_1'] \\ (k,k') \in [n_2] \times [n_2'] \\ (l,l') \in [n_3] \times [n_3']}} B_{\mathcal{I}(j,k,l), \mathcal{I}'(j',k',l')} X_{j,k,l} \alpha_k Y_{j',k',l'} \alpha'_{k'} \Gamma_{j,j'},$$

where $\mathcal{I}, \mathcal{I}'$ are certain bijections that map triples $(j,k,l)$ to row or column indices of the matrix $B \in \mathbb{R}^{n^2 \times n^2}$.

Expressions of this type can then be controlled using a unified approach which is given in the following lemma. Note that properties (c) to (f) will be consequences of (1.9) and (1.10).

**Lemma 1.12.** *For each $r \in \{1,2,3\}$, let $n_r$, $s_r$, $n_r'$, $s_r'$ be positive integers such that*

$$N := n_1 n_2 n_3 = n_1' n_2' n_3', \quad s := s_1 s_2 s_3 = s_1' s_2' s_3'.$$

*Let $\Phi \in \mathbb{R}^{m \times N}$ be a matrix that satisfies the $(4s, \delta)$-RIP and $B := \Phi^* \Phi - Id_n$.*

*Let $\mathbf{X}, \bar{\mathbf{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ be arrays such that:*

*(a) For each $(j,k,l) \in [n_1] \times [n_2] \times [n_3]$, $|X_{j,k,l}| \leq |\bar{X}_{j,k,l}|$.*

*(b) $\|\bar{\mathbf{X}}\|_F = 1$.*

*(c)* $\displaystyle\max_{j \in [n_1], k \in [n_2]} \sum_{l \in [n_3]} X_{j,k,l}^2 \leq \frac{1}{s_1 s_2}$.

*(d)* $\displaystyle\max_{k \in [n_2]} \sum_{l \in [n_3]} X_{j,k,l}^2 \leq \frac{1}{s_2} \sum_{k \in [n_2], l \in [n_3]} \bar{X}_{j,k,l}^2$ *for each $j \in [n_1]$.*

*(e)* $\displaystyle\max_{j \in [n_1]} \sum_{l \in [n_3]} X_{j,k,l}^2 \leq \frac{1}{s_1} \sum_{j \in [n_1], l \in [n_3]} \bar{X}_{j,k,l}^2$ *for each $k \in [n_2]$.*

*(f) For every $j \in [n_1]$, $k \in [n_2]$, there are at most $s_3$ indices $l \in [n_3]$ such that $X_{j,k,l} \neq 0$.*

*Assume that $\mathbf{Y}, \bar{\mathbf{Y}} \in \mathbb{R}^{n_1' \times n_2' \times n_3'}$ satisfy the analogous conditions with the numbers $s_1, s_2, s_3$ replaced by $s_1', s_2', s_3'$.*

*Let $\alpha \in \mathbb{R}^{n_2}$, $\alpha' \in \mathbb{R}^{n_2'}$, $\Gamma \in \mathbb{R}^{n_1 \times n_1'}$ satisfy $\|\alpha\|_2 = \|\alpha'\|_2 = \|\Gamma\|_F = 1$.*

*Let $\mathcal{I} : [n_1] \times [n_2] \times [n_3] \to [N]$ and $\mathcal{I}' : [n_1'] \times [n_2'] \times [n_3'] \to [N]$ be bijections that map tuples $(j,k,l)$ to row/column indices of the matrix $B$.*

*Then*

$$\sum_{\substack{(j,j') \in [n_1] \times [n_1'] \\ (k,k') \in [n_2] \times [n_2'] \\ (l,l') \in [n_3] \times [n_3']}} B_{\mathcal{I}(j,k,l), \mathcal{I}'(j',k',l')} X_{j,k,l} \alpha_k Y_{j',k',l'} \alpha_{k'}' \Gamma_{j,j'} \leq 4 \frac{\delta}{(s_1 s_1')^{\frac{1}{4}} (s_2 s_2')^{\frac{1}{2}}}. \tag{1.17}$$

Our next step is to give the precise definition of the aforementioned arrays outlined in (1.16) and then show that with those, we can rewrite (1.15) in the form of Lemma 1.12 where all the requirements are fulfilled.

For the precise definition of these new arrays, we first define the dimensions

$$\begin{array}{lll}
n_1 = n^{|\bar{J}|}, & n_2 = n^{|\bar{I}|}, & n_3 = n^{|S|} \\
s_1 = s^{|\bar{J}|}, & s_2 = s^{|\bar{I}|}, & s_3 = s^{|S|} \\
n_1' = n^{|\bar{J}'|}, & n_2' = n^{|\bar{I}'|}, & n_3' = n^{|T|} \\
s_1' = s^{|\bar{J}'|}, & s_2' = s^{|\bar{I}'|}, & s_3' = s^{|T|}.
\end{array}$$

Then $n_1 n_2 n_3 = n^{|S \cup \bar{I} \cup \bar{J}|} = n^{|\{1,2\}|} = n^2 = N$ and in the same way $n_1' n_2' n_3' = N$ and $s_1 s_2 s_3 = s_1' s_2' s_3' = s^2$. In this respect, our original assumption that $\Phi \in \mathbb{R}^{m \times N}$ has the $(4s^2, \delta)$-RIP then ensures the corresponding requirement of Lemma 1.12.

For the next definition, we use the following special notation: Let $i \in [n]^d$ and $M \subset [d]$ with $|M| \leq 2$. If $|M| = 2$, i.e. $M = \{m_1, m_2\}$ with $m_1 < m_2$, then $(i_M) = \mathcal{I}_2(i_{m_1}, i_{m_2})$. If

$|M| = 1$, i.e. $M = \{m_1\}$, then $(i_M) = i_{m_1}$ and if $M = \emptyset$, then always $(i_M) = 1$. Note that this $(\cdot)$ operation is a bijection between the tuples in $[n]^{|M|}$ and the integers in $[n^{|M|}]$ (assuming $[n]^0$ contains exactly one empty tuple). With this notation, we define $X, \bar{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ such that for all $i = (i_1, i_2) \in [n]^2$,

$$X_{(i_{\bar{J}}),(i_{\bar{I}}),(i_S)} = x^{(S)}_{i_1,i_2} \prod_{l \in S} \xi^{(l)}_{i_l}, \qquad \bar{X}_{(i_{\bar{J}}),(i_{\bar{I}}),(i_S)} = x_{i_1,i_2}.$$

This defines all entries of these arrays. Analogously, we define $Y, \bar{Y} \in \mathbb{R}^{n'_1 \times n'_2 \times n'_3}$ such that for all $(i_3, i_4) \in [n]^2$,

$$Y_{(i_{\bar{J}'}),(i_{\bar{I}'}),(i_{T+2})} = x^{(T)}_{i_3,i_4} \prod_{l \in T+2} \xi^{(l)}_{i_l}, \qquad \bar{Y}_{(i_{\bar{J}'}),(i_{\bar{I}'}),(i_{T+2})} = x_{i_3,i_4}.$$

Furthermore, $\alpha \in \mathbb{R}^{n_2}$, $\alpha' \in \mathbb{R}^{n'_2}$, $\Gamma \in \mathbb{R}^{n_1 \times n'_1}$ such that for all $i \in [n]^4$,

$$\alpha_{(i_{\bar{I}})} = \alpha^{(1)}_{i_{\bar{I}}}, \qquad \alpha'_{(i_{\bar{I}'})} = \alpha^{(2)}_{i_{\bar{I}'}}, \qquad \Gamma_{(i_{\bar{J}}),(i_{\bar{J}'})} = \alpha^{(3)}_{i_{\bar{\bar{J}}}}.$$

Finally, we define the maps $\mathcal{I} : [n_1] \times [n_2] \times [n_3] \to [N]$ and $\mathcal{I}' : [n'_1] \times [n'_2] \times [n'_3] \to [N]$ such that for all $i \in [n]^4$,

$$\mathcal{I}((i_{\bar{J}}), (i_{\bar{I}}), (i_S)) = \mathcal{I}_2(i_1, i_2), \qquad \mathcal{I}'((i_{\bar{J}'}), (i_{\bar{I}'}), (i_{T+2})) = \mathcal{I}_2(i_3, i_4).$$

Because the $(\cdot)$ operation and $\mathcal{I}_2$ are bijective, also $\mathcal{I}$ and $\mathcal{I}'$ are bijections.

Considering that for all $i \in [n]^4$, $B_{i_1,\dots,i_4} = B_{\mathcal{I}_2(i_1,i_2),\mathcal{I}_2(i_3,i_4)}$, we obtain

$$\begin{aligned} &\sum_{i \in [n]^4} B_{i_1,\dots,i_4} x^{(S)}_{i_1,i_2} x^{(T)}_{i_3,i_4} \alpha^{(1)}_{i_{\bar{I}}} \alpha^{(2)}_{i_{\bar{I}'}} \alpha^{(3)}_{i_{\bar{\bar{J}}}} \prod_{l \in S \cup (T+2)} \xi^{(l)}_{i_l} \\ &= \sum_{i \in [n]^4} B_{\substack{\mathcal{I}((i_{\bar{J}}),(i_{\bar{I}}),(i_S)), \\ \mathcal{I}'((i_{\bar{J}'}),(i_{\bar{I}'}),(i_{T+2}))}} X_{(i_{\bar{J}}),(i_{\bar{I}}),(i_S)} Y_{(i_{\bar{J}'}),(i_{\bar{I}'}),(i_{T+2})} \alpha_{(i_{\bar{I}})} \alpha'_{(i_{\bar{I}'})} \Gamma_{(i_{\bar{J}}),(i_{\bar{J}'})} \\ &= \sum_{\substack{(j,j') \in [n_1] \times [n'_1] \\ (k,k') \in [n_2] \times [n'_2] \\ (l,l') \in [n_3] \times [n'_3]}} B_{\mathcal{I}(j,k,l),\mathcal{I}'(j',k',l')} X_{j,k,l} \alpha_k Y_{j',k',l'} \alpha'_{k'} \Gamma_{j,j'}, \end{aligned} \tag{1.18}$$

where in the first step we substituted the definitions of all these arrays and in the second step we used that summing over all $i \in [n]^4$ is the same as summing over all possible values of

$$((i_{\bar{J}}), (i_{\bar{I}}), (i_S), (i_{\bar{J}'}), (i_{\bar{I}'}), (i_{T+2})).$$

Now in order to apply Lemma 1.12 to (1.18), we need to check that all the remaining assumptions are fulfilled. We start by checking (a) to (f) for $\mathbf{X}, \bar{\mathbf{X}}$. (a) follows directly from the definitions of $\mathbf{x}$ and $\mathbf{x}^{(S)}$ and the fact that $|\xi^{(l)}_{i_l}| = 1$ for all $l \in [4]$, $i_l \in [n]$. Since $\bar{\mathbf{X}}$ is a rearrangement of the entries of $\mathbf{x}$, $\|\bar{\mathbf{X}}\|_F = \|\mathbf{x}\|_F = 1$, implying (b). To show (c), we apply the inequalities (1.9) depending on $S$ and obtain the following four cases.

- $S = \emptyset$: Then
$$\max_{j \in [n_1], k \in [n_2]} \sum_{l \in [n_3]} X^2_{j,k,l} = \max_{i_1,i_2 \in [n]} (x^{(\emptyset)}_{i_1,i_2})^2 \leq \frac{1}{s^2} = \frac{1}{s_1 s_2}.$$

- $S = \{1\}$:
$$\begin{aligned} \max_{j \in [n_1], k \in [n_2]} \sum_{l \in [n_3]} X^2_{j,k,l} &= \max_{i_2 \in [n]} \sum_{i_1 \in [n]} (x^{(\{1\})}_{i_1,i_2} \xi^{(1)}_{i_1})^2 \leq \sum_{i_1 \in [n]} \max_{i_2 \in [n]} (x^{(\{1\})}_{i_1,i_2})^2 \\ &\leq \frac{1}{s} \sum_{i_1,i_2 \in [n]} x^2_{i_1,i_2} = \frac{1}{s_1 s_2}. \end{aligned}$$

- $S = \{2\}$:

$$\max_{j\in[n_1],k\in[n_2]} \sum_{l\in[n_3]} X_{j,k,l}^2 = \max_{i_1\in[n]} \sum_{i_2\in[n]} (x_{i_1,i_2}^{(\{2\})}\xi_{i_2}^{(2)})^2 \leq \sum_{i_2\in[n]} \max_{i_1\in[n]} (x_{i_1,i_2}^{(\{2\})})^2$$
$$\leq \frac{1}{s} \sum_{i_1,i_2\in[n]} x_{i_1,i_2}^2 = \frac{1}{s_1 s_2}.$$

- $S = \{1,2\}$:

$$\max_{j\in[n_1],k\in[n_2]} \sum_{l\in[n_3]} X_{j,k,l}^2 = \sum_{i_1,i_2\in[n]} (x_{i_1,i_2}^{(\{1,2\})}\xi_{i_1}^{(1)}\xi_{i_2}^{(2)})^2 = \|\mathbf{x}^{(\{1,2\})}\|_F^2 \leq 1 = \frac{1}{s_1 s_2}.$$

Condition (d) follows directly if $\bar{I} = \emptyset$ since then $n_2 = s_2 = 1$ and then

$$\begin{aligned} & \max_{k\in[n_2]} \sum_{l\in[n_3]} X_{j,k,l}^2 \leq \frac{1}{s_2} \sum_{k\in[n_2],l\in[n_3]} \bar{X}_{j,k,l}^2 \\ \Leftrightarrow \quad & \sum_{l\in[n_3]} X_{j,1,l}^2 \leq \sum_{l\in[n_3]} \bar{X}_{j,1,l}^2, \end{aligned}$$

and the latter follows from (a) which we have already shown. If $\bar{J} = \emptyset$, then $n_1 = s_1 = 1$ and (d) is equivalent to

$$\begin{aligned} & \max_{j\in[n_1],k\in[n_2]} \sum_{l\in[n_3]} X_{j,k,l}^2 \leq \frac{1}{s_2} \sum_{j\in[n_1],k\in[n_2],l\in[n_3]} \bar{X}_{j,k,l}^2 \\ \Leftrightarrow \quad & \max_{j\in[n_1],k\in[n_2]} \sum_{l\in[n_3]} X_{j,k,l}^2 \leq \frac{1}{s_2}. \end{aligned}$$

This is (c) which we have already shown. If both $\bar{I}$ and $\bar{J}$ are $\neq \emptyset$, then there are precisely the two cases $\bar{I} = \{1\}, \bar{J} = \{2\}$ and $\bar{I} = \{2\}, \bar{J} = \{1\}$, implying $S = \emptyset$. In the first case, (d) is equivalent to the statement that for all $i_2 \in [n]$,

$$\max_{i_1\in[n]} (x_{i_1,i_2}^{(\emptyset)})^2 \leq \frac{1}{s_2} \sum_{i_1,i_2\in[n]} x_{i_1,i_2}^2.$$

This follows from (1.9). The other case $\bar{I} = \{2\}, \bar{J} = \{1\}$ follows analogously, completing the proof of (d). The property (e) for a certain choice of $S, \bar{I}, \bar{J}$ is equivalent to the corresponding case of (d) where $\bar{I}$ and $\bar{J}$ are exchanged, so (e) always holds.

For (f), we can see that it trivially holds if $S = \emptyset$, i.e., $n_3 = s_3 = 1$. For all other choices of $S$, it directly follows from (1.10).

Analogously, it also follows that these properties (a) to (f) hold for $Y, \bar{Y}$.

By the definitions, also

$$\|\alpha\|_2 = \|\boldsymbol{\alpha}^{(1)}\|_F = \|\alpha'\|_2 = \|\boldsymbol{\alpha}^{(2)}\|_F = \|\Gamma\|_F = \|\boldsymbol{\alpha}^{(3)}\|_F = 1,$$

which completes the check of all assumptions.

Thus, we can apply Lemma 1.12 to (1.18) such that we obtain

$$\|\mathbf{B}^{(S,T)}\|_{I_1,\ldots,I_\kappa} \leq 4\frac{\delta}{(s_1 s_1')^{\frac{1}{4}}(s_2 s_2')^{\frac{1}{2}}} = 4\frac{\delta}{s^{\frac{1}{4}|\bar{\bar{J}}|+\frac{1}{2}(|\bar{I}|+|\bar{I}'|)}} \leq 4\frac{\delta}{s^{\frac{\kappa}{2}}},$$

where in the last step we used that $\frac{1}{4}|\bar{\bar{J}}| + \frac{1}{2}(|\bar{I}| + |\bar{I}'|) \geq \frac{\kappa}{2}$. We will show this fact as Lemma 1.17 in Subsection 1.5.5.

This completes the proof of (1.13) for all $S, T \subset [2]$ with $(S,T) \neq (\{1,2\},\{1,2\})$ and all partitions with the constant $C = 4$.

### 1.5.5 Proof of the Lemmas

The goal of this subsection is to give all remaining proofs of the lemmas that were given previously in Section 1.5. First, we give the proof of Lemma 1.6 from Section 1.4. Then the main part will be about proving Lemma 1.12 which we used in the previous Subsection 1.5.4. On this way, we will establish the technical auxiliary Lemmas 1.13 and 1.16 and make use of a certain class of partitions introduced in Definition 1.14.

*Proof of Lemma 1.6.* It is sufficient to show the case

$$\|\mathbf{B}\|_{I_1,\dots,I_\kappa} \leq \|\mathbf{B}\|_{I_1,\dots,I_{\kappa-2},I_{\kappa-1}\cup I_\kappa},$$

i.e., we join the last two partition sets. Successively applying this and reordering the partition sets then yields the result.

To show the aforementioned case, consider any $\boldsymbol{\alpha}^{(1)} \in \mathbb{R}^{[n]^{|I_1|}}, \dots, \boldsymbol{\alpha}^{(\kappa)} \in \mathbb{R}^{[n]^{|I_\kappa|}}$ with $\|\boldsymbol{\alpha}^{(1)}\|_F = \dots = \|\boldsymbol{\alpha}^{(\kappa)}\|_F = 1$. Denote $m := |I_{\kappa-1}|$ and $\bar{m} := |I_\kappa|$ and define $\tilde{\boldsymbol{\alpha}}^{(\kappa-1)} \in \mathbb{R}^{[n]^{m+\bar{m}}}$ such that for all $i \in [n]^{d'}$,

$$\tilde{\alpha}^{(\kappa-1)}_{i_{I_{\kappa-1}\cup I_\kappa}} = \alpha^{(\kappa-1)}_{i_{I_{\kappa-1}}} \cdot \alpha^{(\kappa)}_{i_{I_\kappa}}.$$

Then

$$\begin{aligned}\|\tilde{\boldsymbol{\alpha}}^{(\kappa-1)}\|_F^2 &= \sum_{i_{I_{\kappa-1}\cup I_\kappa} \in [n]^{m+\bar{m}}} (\tilde{\alpha}^{(\kappa-1)}_{i_{I_{\kappa-1}\cup I_\kappa}})^2 = \sum_{i_{I_{\kappa-1}}\in[n]^m} \sum_{i_{I_\kappa}\in[n]^{\bar{m}}} (\alpha^{(\kappa-1)}_{i_{I_{\kappa-1}}} \cdot \alpha^{(\kappa)}_{i_{I_\kappa}})^2 \\ &= \sum_{i_{I_{\kappa-1}}\in[n]^m} (\alpha^{(\kappa-1)}_{i_{I_{\kappa-1}}})^2 \cdot \sum_{i_{I_\kappa}\in[n]^{\bar{m}}} (\alpha^{(\kappa)}_{i_{I_\kappa}})^2 = \|\boldsymbol{\alpha}^{(\kappa-1)}\|_F^2 \cdot \|\boldsymbol{\alpha}^{(\kappa)}\|_F^2 = 1\end{aligned}$$

and therefore

$$\sum_{i\in[n]^{d'}} B_i \alpha^{(1)}_{i_{I_1}} \dots \alpha^{(\kappa)}_{i_{I_\kappa}} = \sum_{i\in[n]^{d'}} B_i \alpha^{(1)}_{i_{I_1}} \dots \alpha^{(\kappa-2)}_{i_{I_{\kappa-2}}} \tilde{\alpha}^{(\kappa-1)}_{i_{I_{\kappa-1}\cup I_\kappa}} \leq \|\mathbf{B}\|_{I_1,\dots,I_{\kappa-2},I_{\kappa-1}\cup I_\kappa}.$$

Taking the supremum over the $\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(\kappa)}$ shows

$$\|\mathbf{B}\|_{I_1,\dots,I_\kappa} \leq \|\mathbf{B}\|_{I_1,\dots,I_{\kappa-2},I_{\kappa-1}\cup I_\kappa}.$$

□

**Lemma 1.13.** *Let $n, R, s_1, s_2$ and $n', R', s_1', s_2'$ be positive integers such that $s := s_1 s_2 = s_1' s_2'$ and $\Phi \in \mathbb{R}^{m\times N}$ has the $(4s,\delta)$-RIP, $B := \Phi^*\Phi - Id_N$.*

*Consider vectors $x^{(j,K)}, y^{(j',K')} \in \mathbb{R}^N$ for $(j,k) \in [n]\times[R]$ and $(j',k') \in [n']\times[R']$ such that all $x^{(j,K)}$ are $s_2$-sparse with disjoint supports and all $y^{(j',K')}$ are $s_2'$-sparse with disjoint supports.*

*For each $K \in [R]$, let $b_K(1), \dots, b_K(R_1)$ be a partition of $[n]$ into sets of size $\leq s_1$ each. Analogously, for each $K' \in [R']$, let $b'_{K'}(1), \dots, b'_{K'}(R_1')$ be a partition of $[n']$ into sets of size $\leq s_2'$.*

*Then*

$$\begin{aligned}&\sum_{(j,j')\in[n]\times[n']} \left( \sum_{(K,K')\in[R]\times[R']} (x^{(j,K)})^* B y^{(j',K')} \right)^2 \\ \leq{}& \delta^2 \left[ \sum_{(J,K,\bar{K})\in[R_1]\times[R]^2} \sqrt{\sum_{j\in b_K(J)} \|x^{(j,K)}\|_2^2 \|x^{(j,\bar{K})}\|_2^2} \right] \\ &\cdot \left[ \sum_{(J',K',\bar{K}')\in[R_1']\times[R']^2} \sqrt{\sum_{j'\in b_{K'}(J')} \|y^{(j',K')}\|_2^2 \|y^{(j',\bar{K}')}\|_2^2} \right].\end{aligned}$$

*Proof.* We can bound the desired expression by rearranging terms and block-wise summation.

$$
\begin{aligned}
&\sum_{(j,j')\in[n]\times[n']}\left(\sum_{(K,K')\in[R]\times[R']}(x^{(j,K)})^*By^{(j',K')}\right)^2\\
=&\sum_{(j,j')\in[n]\times[n']}\ \sum_{(K,\bar{K},K',\bar{K}')\in[R]^2\times[R']^2}(x^{(j,K)})^*By^{(j',K')}(y^{(j',\bar{K}')})^*B^*x^{(j,\bar{K})}\\
=&\sum_{\substack{(J,J')\in[R_1]\times[R_1']\\(K,\bar{K})\in[R]^2\\(K',\bar{K}')\in[R']^2}}\ \sum_{j\in b_K(J)}(x^{(j,K)})^*B\left(\sum_{j'\in b'_{K'}(J')}y^{(j',K')}(y^{(j',\bar{K}')})^*\right)B^*x^{(j,\bar{K})}\\
=&\sum_{\substack{(J,J')\in[R_1]\times[R_1']\\(K,\bar{K})\in[R]^2\\(K',\bar{K}')\in[R']^2}}\left\langle\sum_{j\in b_K(J)}x^{(j,K)}(x^{(j,\bar{K})})^*,B\left(\sum_{j'\in b'_{K'}(J')}y^{(j',K')}(y^{(j',\bar{K}')})^*\right)B^*\right\rangle_F.
\end{aligned}
$$

In the third step we used that for $x,y\in\mathbb{R}^N$, $A\in\mathbb{R}^{N\times N}$, $x^*Ay=\operatorname{tr}(x^*Ay)=\operatorname{tr}(yx^*A)=\langle xy^*,A\rangle_F$.

Note that every $x^{(j,K)}$ is $s_2$-sparse and $|b_K(J)|\leq s_1$. Thus, the number of nonzero rows and the number of nonzero columns of the matrix $\sum_{j\in b_K(J)}x^{(j,K)}(x^{(j,\bar{K})})^*$ can be at most $s=s_1s_2$ each. The same holds for $\sum_{j'\in b'_{K'}(J')}y^{(j',K')}(y^{(j',\bar{K}')})^*$. So for the above expression, we can restrict $B$ to a submatrix of $2s$ rows and $2s$ columns which has an operator norm $\leq\delta$ by the RIP assumption (Lemma 1.8). Using that $\langle A,BCB^*\rangle_F\leq\|A\|_F\|BCB^*\|_F\leq\|A\|_F\|B\|_{2\to2}^2\|C\|_F$, we can bound the expression by

$$
\begin{aligned}
&\delta^2\sum_{\substack{(J,J')\in[R_1]\times[R_1']\\(K,\bar{K})\in[R]^2\\(K',\bar{K}')\in[R']^2}}\left\|\sum_{j\in b_K(J)}x^{(j,K)}(x^{(j,\bar{K})})^*\right\|_F\left\|\sum_{j'\in b'_{K'}(J')}y^{(j',K')}(y^{(j',\bar{K}')})^*\right\|_F\\
\leq&\delta^2\sum_{\substack{(J,J')\in[R_1]\times[R_1']\\(K,\bar{K})\in[R]^2\\(K',\bar{K}')\in[R']^2}}\sqrt{\sum_{j\in b_K(J)}\|x^{(j,K)}\|_2^2\|x^{(j,\bar{K})}\|_2^2\sum_{j'\in b'_{K'}(J')}\|y^{(j',K')}\|_2^2\|y^{(j',\bar{K}')}\|_2^2}\\
=&\delta^2\left[\sum_{\substack{J\in[R_1]\\(K,\bar{K})\in[R]^2}}\sqrt{\sum_{j\in b_K(J)}\|x^{(j,K)}\|_2^2\|x^{(j,\bar{K})}\|_2^2}\right]\\
&\cdot\left[\sum_{\substack{J'\in[R_1']\\(K',\bar{K}')\in[R']^2}}\sqrt{\sum_{j'\in b'_{K'}(J')}\|y^{(j',K')}\|_2^2\|y^{(j',\bar{K}')}\|_2^2}\right],
\end{aligned}
$$

where in the first step we used that the $x^{(j,K)}(x^{(j,\bar{K})})^*$ have disjoint supports.

□

A central argument used for the case $d=1$ in the previous paper [KW11] is the strategy to divide the signal vector $x\in\mathbb{R}^n$ into blocks of size $s$ by descending absolute value of its entries.

The following Definition 1.14 and Lemma 1.15 generalize this idea. However, it is not required to know the proof in [KW11] for the subsequent statements.

We consider an arbitrary finite indexed family $(x_i)_{i\in I}$. For example, this could be a vector for $I = [n]$ or a matrix for $I = [n] \times [n]$. Now take $b(1)$ to be the set of the $s$ indices $i \in I$ with the largest $|x_i|$, then $b(2)$ the set of the $s$ remaining indices $i \in I$ with the largest $|x_i|$ and so on. This leads to the definition of the following partitions.

**Definition 1.14.** *Let $I$ be a finite set, $(x_i)_{i\in I}$ an indexed family with values in $\mathbb{R}$, and $s$ a positive integer. We define $\mathcal{P}((x_i)_{i\in I}, s) = (b(1), \dots, b(R))$ for $R = \lceil \frac{|I|}{s} \rceil$ to be a partition of $I$ such that $|b(1)| = \dots = |b(R-1)| = s$ and for all $J = 1, \dots, R-1$ and all $i_1 \in b(J)$, $|x_{i_1}| \geq \max_{i_2 \in b(J+1)} |x_{i_2}|$. By ordering the values $x_i$ by their absolute value, such a partition can always be constructed.*

Then the main use of these partitions can be summarized in the following simple lemma.

**Lemma 1.15.** *Let $I$ be a finite set, $(x_i)_{i\in I}$ an indexed family with values in $\mathbb{R}$, $(b(1), \dots, b(R)) = \mathcal{P}((x_i)_{i\in I}, s)$. Then*

$$\sum_{J=2}^{R} \max_{i\in b(J)} |x_i| \leq \frac{1}{s} \sum_{i\in I} |x_i|$$

*and hence*

$$\sum_{J=1}^{R} \max_{i\in b(J)} |x_i| \leq \max_{i\in I} |x_i| + \frac{1}{s} \sum_{i\in I} |x_i|.$$

*Proof.* For each $J = 2, \dots, R$, $|b(J-1)| = s$ and for each $i_1 \in b(J-1)$, $|x_{i_1}| \geq \max_{i_2\in b(J)} |x_{i_2}|$, so $\max_{i_2\in b(J)} |x_{i_2}| \leq \frac{1}{s} \sum_{i_1\in b(J-1)} |x_i|$ and

$$\sum_{J=2}^{R} \max_{i\in b(J)} |x_i| \leq \sum_{J=2}^{R} \frac{1}{s} \sum_{i\in b(J-1)} |x_i| \leq \frac{1}{s} \sum_{i\in I} |x_i|.$$

Then

$$\sum_{J=1}^{R} \max_{i\in b(J)} |x_i| = \max_{i\in b(1)} |x_i| + \sum_{J=2}^{R} \max_{i\in b(J)} |x_i| \leq \max_{i\in I} |x_i| + \frac{1}{s} \sum_{i\in I} |x_i|.$$

□

Using the partitions from Definition 1.14, we can establish the following lemma which will eventually be used to control the expressions occurring on the right hand side in Lemma 1.13.

**Lemma 1.16.** *Let $\mathbf{X}, \bar{\mathbf{X}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ be arrays that satisfy the conditions (a) to (e) from Lemma 1.12 and $\alpha \in \mathbb{R}^{n_2}$ such that $\|\alpha\|_2 = 1$.*

*Consider the partitions*

- $(b(1), \dots, b(R)) = \mathcal{P}\left(\left(\sum_{j\in[n_1], l\in[n_3]} \bar{X}_{j,k,l}^2\right)_{k\in[n_2]}, s_2\right)$,
- $(b_K(1), \dots, b_K(R_1)) = \mathcal{P}\left(\left(\sum_{k\in b(K), l\in[n_3]} (X_{j,k,l}\alpha_k)^2\right)_{j\in[n_1]}, s_1\right)$

  *for each $K = 1, \dots, R$.*

*Then*

$$\sum_{\substack{J\in[R]\\K,\bar{K}\in[R_1]}}\sqrt{\sum_{j\in b_K(J)}\left(\sum_{\substack{k\in b(K)\\l\in[n_3]}}(X_{j,k,l}\alpha_k)^2\right)\left(\sum_{\substack{k\in b(\bar{K})\\l\in[n_3]}}(X_{j,k,l}\alpha_k)^2\right)}\le\frac{4}{\sqrt{s_1}s_2}.$$

*Proof.* We use the Hölder inequality (for $\ell_\infty$ and $\ell_1$ norm) on $\sum_{j\in b_K(J)}$ and then the Cauchy-Schwarz inequality on $\sum_{J\in[R]}$ to obtain

$$\begin{aligned}
&\sum_{\substack{J\in[R]\\K,\bar{K}\in[R_1]}}\sqrt{\sum_{j\in b_K(J)}\left(\sum_{\substack{k\in b(K)\\l\in[n_3]}}(X_{j,k,l}\alpha_k)^2\right)\left(\sum_{\substack{k\in b(\bar{K})\\l\in[n_3]}}(X_{j,k,l}\alpha_k)^2\right)}\\
&\le\sum_{J\in[R],K,\bar{K}\in[R_1]}\sqrt{\left(\max_{j\in b_K(J)}\sum_{\substack{k\in b(K)\\l\in[n_3]}}(X_{j,k,l}\alpha_k)^2\right)\left(\sum_{j\in b_K(J)}\sum_{\substack{k\in b(\bar{K})\\l\in[n_3]}}(X_{j,k,l}\alpha_k)^2\right)}\\
&\le\sum_{K,\bar{K}\in[R_1]}\sqrt{\left(\sum_{J\in[R]}\max_{j\in b_K(J)}\sum_{\substack{k\in b(K)\\l\in[n_3]}}(X_{j,k,l}\alpha_k)^2\right)\left(\sum_{J\in[R]}\sum_{j\in b_K(J)}\sum_{\substack{k\in b(\bar{K})\\l\in[n_3]}}(X_{j,k,l}\alpha_k)^2\right)}\\
&=\left[\sum_{K\in[R_1]}\sqrt{\sum_{J\in[R]}\max_{j\in b_K(J)}\sum_{\substack{k\in b(K)\\l\in[n_3]}}(X_{j,k,l}\alpha_k)^2}\right]\left[\sum_{\bar{K}\in[R_1]}\sqrt{\sum_{k\in b(\bar{K})}\sum_{\substack{j\in[n_1]\\l\in[n_3]}}(X_{j,k,l}\alpha_k)^2}\right]\\
&=:(I)\cdot(II).
\end{aligned}\tag{1.19}$$

Now we apply Lemma 1.15 on the partitions $b_K$, then again the Hölder inequality ($\ell_\infty$ and $\ell_1$) on $\sum_{k\in b(K)}$ to obtain,

$$\begin{aligned}
&\sum_{J\in[R]}\max_{j\in b_K(J)}\sum_{\substack{k\in b(K)\\l\in[n_3]}}(X_{j,k,l}\alpha_k)^2\\
&\le\max_{j\in[n_1]}\sum_{\substack{k\in b(K)\\l\in[n_3]}}(X_{j,k,l}\alpha_k)^2+\frac{1}{s_1}\sum_{\substack{k\in b(K)\\j\in[n_1],l\in[n_3]}}(X_{j,k,l}\alpha_k)^2\\
&\le\max_{k\in b(K)}\max_{j\in[n_1]}\sum_{l\in[n_3]}X_{j,k,l}^2\cdot\sum_{k\in b(K)}\alpha_k^2+\frac{1}{s_1}\max_{k\in b(K)}\sum_{j\in[n_1],l\in[n_3]}X_{j,k,l}^2\cdot\sum_{k\in b(K)}\alpha_k^2.
\end{aligned}$$

For $K=1$, we can use the assumptions of this lemma to bound this by

$$\begin{aligned}
&\overset{\|a\|_2=1}{\le}\max_{j\in[n_1],k\in[n_2]}\sum_{l\in[n_3]}X_{j,k,l}^2\cdot 1+\frac{1}{s_1}\max_{k\in[n_2]}\sum_{j\in[n_1],l\in[n_3]}X_{j,k,l}^2\cdot 1\\
&\overset{(c)}{\le}\frac{1}{s_1s_2}+\frac{1}{s_1}\sum_{j\in[n_1]}\max_{k\in[n_2]}\sum_{l\in[n_3]}X_{j,k,l}^2
\end{aligned}$$

$$\begin{aligned}
&\overset{(d)}{\leq} \frac{1}{s_1 s_2} + \frac{1}{s_1 s_2} \sum_{j\in[n_1],k\in[n_2],l\in[n_3]} \bar{X}_{j,k,l}^2 \\
&\overset{(b)}{\leq} \frac{2}{s_1 s_2}.
\end{aligned}$$

On the other hand, for $K \geq 2$, we use (e) and (a) to obtain the bound

$$\frac{2}{s_1} \max_{k\in b(K)} \sum_{j\in[n_1],l\in[n_3]} \bar{X}_{j,k,l}^2 \cdot \sum_{k\in b(K)} \alpha_k^2.$$

Altogether, we obtain with the above results, the Cauchy-Schwarz inequality for $\sum_{K\in[R]\setminus\{1\}}$, Lemma 1.15 for the partition $b$ and the assumptions of this lemma,

$$\begin{aligned}
(I) &\leq \sqrt{\frac{2}{s_1 s_2}} + \sum_{K\in[R]\setminus\{1\}} \sqrt{\sum_{J\in[R]} \max_{j\in b_K(J)} \sum_{\substack{k\in b(K)\\ l\in[n_3]}} (X_{j,k,l}\alpha_k)^2} \\
&\leq \sqrt{\frac{2}{s_1 s_2}} + \sum_{K\in[R]\setminus\{1\}} \sqrt{\left(\frac{2}{s_1} \max_{k\in b(K)} \sum_{j\in[n_1],l\in[n_3]} \bar{X}_{j,k,l}^2\right)\left(\sum_{k\in b(K)} \alpha_k^2\right)} \\
&\overset{\text{C.-S.}}{\leq} \sqrt{\frac{2}{s_1 s_2}} + \sqrt{\frac{2}{s_1}\left(\sum_{K\in[R]\setminus\{1\}} \max_{k\in b(K)} \sum_{j\in[n_1],l\in[n_3]} \bar{X}_{j,k,l}^2\right)\left(\sum_{K\in[R]\setminus\{1\}} \sum_{k\in b(K)} \alpha_k^2\right)} \\
&\overset{\|\alpha\|_2=1}{\leq} \sqrt{\frac{2}{s_1 s_2}} + \sqrt{\frac{2}{s_1} \sum_{K\in[R]\setminus\{1\}} \max_{k\in b(K)} \sum_{j\in[n_1],l\in[n_3]} \bar{X}_{j,k,l}^2} \\
&\overset{\text{Lem.1.15}}{\leq} \sqrt{\frac{2}{s_1 s_2}} + \sqrt{\frac{2}{s_1 s_2} \sum_{j\in[n_1],k\in[n_2],l\in[n_3]} \bar{X}_{j,k,l}^2} \\
&\overset{(b)}{=} \frac{2\sqrt{2}}{\sqrt{s_1 s_2}}.
\end{aligned}$$

Furthermore, for the other factor $(II)$, we use the Hölder inequality for $\sum_{k\in b(\bar{K})}$, the Cauchy-Schwarz inequality for $\sum_{\bar{K}\in[R_1]}$, Lemma 1.15 for the partition $b$ and the assumptions of this lemma to obtain

$$\begin{aligned}
(II) &= \sum_{\bar{K}\in[R_1]} \sqrt{\sum_{k\in b(\bar{K})} \sum_{\substack{j\in[n_1]\\ l\in[n_3]}} (X_{j,k,l}\alpha_k)^2} \\
&\overset{\text{Hölder}}{\leq} \sum_{\bar{K}\in[R_1]} \sqrt{\left(\max_{k\in b(\bar{K})} \sum_{j\in[n_1],l\in[n_3]} X_{j,k,l}^2\right)\left(\sum_{k\in b(\bar{K})} \alpha_k^2\right)} \\
&\overset{\text{C.-S.}}{\leq} \sqrt{\left(\sum_{\bar{K}\in[R_1]} \max_{k\in b(\bar{K})} \sum_{j\in[n_1],l\in[n_3]} X_{j,k,l}^2\right)\left(\sum_{\bar{K}\in[R_1]} \sum_{k\in b(\bar{K})} \alpha_k^2\right)} \\
&\overset{\|\alpha\|_2=1}{=} \sqrt{\sum_{\bar{K}\in[R_1]} \max_{k\in b(\bar{K})} \sum_{j\in[n_1],l\in[n_3]} X_{j,k,l}^2} \\
&\overset{(a)}{\leq} \sqrt{\max_{k\in b(1)} \sum_{j\in[n_1],l\in[n_3]} X_{j,k,l}^2 + \sum_{\bar{K}\in[R_1]\setminus\{1\}} \max_{k\in b(\bar{K})} \sum_{j\in[n_1],l\in[n_3]} \bar{X}_{j,k,l}^2}
\end{aligned}$$

$$\begin{aligned}
&\overset{\text{Lem.1.15}}{\leq} \sqrt{\max_{k\in[n_2]} \sum_{j\in[n_1],l\in[n_3]} X_{j,k,l}^2 + \frac{1}{s_2} \sum_{k\in[n_2]} \sum_{j\in[n_1],l\in[n_3]} \bar{X}_{j,k,l}^2} \\
&\leq \sqrt{\sum_{j\in[n_1]} \max_{k\in[n_2]} \sum_{l\in[n_3]} X_{j,k,l}^2 + \frac{1}{s_2} \sum_{k\in[n_2]} \sum_{j\in[n_1],l\in[n_3]} \bar{X}_{j,k,l}^2} \\
&\overset{(d),(b)}{\leq} \sqrt{\frac{2}{s_2}}.
\end{aligned}$$

Substituting the bounds for $(I)$ and $(II)$ into (1.19) yields the claim of the lemma. □

Now we have established all required tools and can prove Lemma 1.12.

*Proof of Lemma 1.12.* We need to show the inequality (1.17). Let $\beta$ be the left hand side of the inequality (1.17). Recall that the corresponding expression contains the matrix $B$, the order 3 arrays $\mathbf{X}, \mathbf{Y}$, vectors $\alpha, \alpha'$, the matrix $\Gamma$ and bijective functions $\mathcal{I}$, $\mathcal{I}'$ that map index triples (corresponding to $\mathbf{X}, \mathbf{Y}$) to row/column indices of $B$.

We apply the Cauchy-Schwarz inequality to the sum $\sum_{(j,j')\in[n_1]\times[n_1']}$ and observe that

$$\begin{aligned}
\beta^2 &= \left( \sum_{\substack{(j,j')\in[n_1]\times[n_1'] \\ (k,k')\in[n_2]\times[n_2'] \\ (l,l')\in[n_3]\times[n_3']}} B_{\mathcal{I}(j,k,l),\mathcal{I}'(j',k',l')} X_{j,k,l} \alpha_k Y_{j',k',l'} \alpha'_{k'} \Gamma_{j,j'} \right)^2 \\
&\leq \sum_{(j,j')\in[n_1]} \left( \sum_{\substack{(k,k')\in[n_2]\times[n_2'] \\ (l,l')\in[n_3]\times[n_3']}} B_{\mathcal{I}(j,k,l),\mathcal{I}'(j',k',l')} X_{j,k,l} \alpha_k Y_{j',k',l'} \alpha'_{k'} \right)^2 \cdot \sum_{(j,j')\in[n_1]} \Gamma_{j,j'}^2 \\
&= \sum_{(j,j')\in[n_1]} \left( \sum_{\substack{(k,k')\in[n_2]\times[n_2'] \\ (l,l')\in[n_3]\times[n_3']}} B_{\mathcal{I}(j,k,l),\mathcal{I}'(j',k',l')} X_{j,k,l} \alpha_k Y_{j',k',l'} \alpha'_{k'} \right)^2
\end{aligned}$$

Now we choose partitions

$$(b(1),\ldots,b(R)) = \mathcal{P}\left( \left( \sum_{j\in[n_1],l\in[n_3]} \bar{X}_{j,k,l}^2 \right)_{k\in[n_2]}, s_2 \right)$$

and

$$(b'(1),\ldots,b'(R')) = \mathcal{P}\left( \left( \sum_{j'\in[n_1'],l'\in[n_3']} \bar{Y}_{j',k',l'}^2 \right)_{k'\in[n_2']}, s_2' \right).$$

Using those, we can further conclude

$$\beta^2 \leq$$

$$\sum_{(j,j')\in[n_1]} \left( \sum_{K\in[R],K'\in[R']} \sum_{\substack{(k,k')\in b(K)\times b'(K') \\ (l,l')\in[n_3]\times[n_3']}} B_{\mathcal{I}(j,k,l),\mathcal{I}'(j',k',l')} X_{j,k,l} \alpha_k Y_{j',k',l'} \alpha'_{k'} \right)^2.$$

Now for each $j \in [n_1]$ and $K \in [R]$, we define the vector $x^{(j,K)} \in \mathbb{R}^N$ by

$$x^{(j,K)}_{\mathcal{I}(\bar{j},k,l)} = \begin{cases} X_{j,k,l}\alpha_k & \text{if } \bar{j} = j \text{ and } k \in b(K) \\ 0 & \text{otherwise.} \end{cases}$$

for all $(\bar{j}, k, l) \in [n_1] \times [n_2] \times [n_3]$. Analogously, we define $y^{(j',K')} \in \mathbb{R}^N$ for $j' \in [n_1']$, $K' \in [R']$ by

$$y^{(j',K')}_{\mathcal{I}'(\bar{j}',k',l')} = \begin{cases} Y_{j',k',l'}\alpha'_{k'} & \text{if } \bar{j}' = j' \text{ and } k' \in b'(K') \\ 0 & \text{otherwise.} \end{cases}$$

With these vectors, we obtain

$$\begin{aligned}\beta^2 &\leq \sum_{(j,j')\in[n_1]} \left( \sum_{K\in[R],K'\in[R']} \sum_{r,r'\in[N]} B_{r,r'} x_r^{(j,K)} y_{r'}^{(j',K')} \right)^2 \\ &= \sum_{(j,j')\in[n_1]} \left( \sum_{K\in[R],K'\in[R']} (x^{(j,K)})^* B y^{(j',K')} \right)^2 \end{aligned} \tag{1.20}$$

For every $K \in [R]$, $|b(K)| \leq s_2$ and for every $j, k$, $X_{j,k,l}$ is only non-zero for $\leq s_3$ values of $l$. The $b(K)$ are disjoint and consequently, all the vectors $x^{(j,K)}$ are $s_2 s_3$-sparse with disjoint supports. Analogously, also the $y^{(j',K')}$ are $s_2' s_3'$-sparse with disjoint supports. Furthermore, for each $K \in [R]$, $K' \in [R']$, we choose the partitions

$$(b_K(1), \ldots, b_K(R_1)) = \mathcal{P}\left( \left( \sum_{k\in b(K), l\in[n_3]} (X_{j,k,l}\alpha_k)^2 \right)_{j\in[n_1]}, s_1 \right)$$

$$(b'_{K'}(1), \ldots, b'_{K'}(R_1')) = \mathcal{P}\left( \left( \sum_{k\in b(K), l\in[n_3]} (X_{j,k,l}\alpha_k)^2 \right)_{j\in[n_1]}, s_1 \right)$$

Then the requirements are fulfilled to bound (1.20) using Lemma 1.13 such that

$$\begin{aligned}\beta^2 \leq \delta^2 &\left[ \sum_{(J,K,\bar{K})\in[R_1]\times[R]^2} \sqrt{\sum_{j\in b_K(J)} \|x^{(j,K)}\|_2^2 \|x^{(j,\bar{K})}\|_2^2} \right] \\ &\cdot \left[ \sum_{(J',K',\bar{K}')\in[R_1']\times[R']^2} \sqrt{\sum_{j'\in b'_{K'}(J')} \|y^{(j',K')}\|_2^2 \|y^{(j',\bar{K}')}\|_2^2} \right] \\ &=: \delta^2 \cdot (I) \cdot (II). \end{aligned}$$

$$\begin{aligned}(I) &= \sum_{\substack{J\in[R_1] \\ K,\bar{K}\in[R]}} \sqrt{\sum_{j\in b_K(J)} \left( \sum_{r\in[N]} (x_r^{(j,K)})^2 \cdot \sum_{r\in[N]} (x_r^{(j,\bar{K})})^2 \right)} \\ &= \sum_{\substack{J\in[R_1] \\ K,\bar{K}\in[R]}} \sqrt{\sum_{j\in b_K(J)} \left( \sum_{\substack{k\in b(K) \\ l\in[n_3]}} (X_{j,k,l}\alpha_k)^2 \cdot \sum_{\substack{k\in b(\bar{K}) \\ l\in[n_3]}} (X_{j,k,l}\alpha_k)^2 \right)} \end{aligned}$$

The requirements (a) to (e) are assumed to hold in this lemma. Also $\|\alpha\|_2 = 1$ and the choices of the partitions $b$ and $b_K$ also match with Lemma 1.16. So we can use Lemma 1.16 to conclude

$$(I) \le \frac{4}{\sqrt{s_1 s_2}}.$$

We can perform the same argument with $(II)$ and use Lemma 1.16 to show

$$(II) \le \frac{4}{\sqrt{s_1' s_2'}}.$$

Combining these bounds, we obtain

$$\beta^2 \le 16 \frac{\delta}{(s_1 s_1')^{\frac{1}{2}} s_2 s_2'}.$$

□

Besides the previous lemma, also the following Lemma 1.17 is used for the proof in Subsection 1.5.4. Like the other results in the current subsection, it is not restricted to the special case $d = 2$ and will be used in the same form in the general proof in Section 2.5.

**Lemma 1.17.** *Let $d \ge 1$ be an integer and $I_1, \dots, I_\kappa \subset [2d]$ pairwise disjoint, non-empty sets. Define*

$$\bar{I} = \bigcup_{j \in [\kappa] : I_j \subset [d]} I_j, \quad \bar{I}' = \bigcup_{j \in [\kappa] : I_j \subset ([2d] \backslash [d])} I_j$$
$$\bar{\bar{J}} = (I_1 \cup \dots \cup I_\kappa) \backslash (\bar{I} \cup \bar{I}'),$$

*i.e., $\bar{I}$ is the union of those sets among $I_1, \dots, I_\kappa$ that are contained in $[d]$, $\bar{I}'$ the union of the sets contained in $\{d+1, \dots, 2d\}$ and $\bar{\bar{J}}$ the union of all the other sets. Then*

$$\frac{1}{4}|\bar{\bar{J}}| + \frac{1}{2}(|\bar{I}| + |\bar{I}'|) \ge \frac{\kappa}{2}.$$

*Proof of Lemma 1.17.* Define $I_0 := I_1 \cup \dots \cup I_\kappa$.

Then $|\bar{\bar{J}}| + |\bar{I}| + |\bar{I}'| = |I_0|$. If $\kappa \le \frac{|I_0|}{2}$,

$$\frac{1}{4}|\bar{\bar{J}}| + \frac{1}{2}(|\bar{I}| + |\bar{I}'|) \ge \frac{1}{4}(|\bar{\bar{J}}| + |\bar{I}| + |\bar{I}'|) = \frac{|I_0|}{4} \ge \frac{\kappa}{2}.$$

Now assume that $\kappa > \frac{|I_0|}{2}$. Let $\kappa' \le \kappa$ be the number of indices $l \in [\kappa]$ such that $|I_l| = 1$. All other sets $I_l$ must contain at least two elements and the total number of elements is $|I_0| = \sum_{l \in [\kappa]} |I_l| \ge \kappa' + 2(\kappa - \kappa') = 2\kappa - \kappa'$. This implies that $\kappa' \ge 2\kappa - |I_0|$. Every one-element set $I_l$ is completely contained in either $[d]$ or $[2d] \backslash [d]$ and thus $I_l \subset \bar{I}$ or $I_l \subset \bar{I}'$. So $|\bar{I}| + |\bar{I}'| \ge \kappa' \ge 2\kappa - |I_0|$ and we obtain

$$\frac{1}{4}|\bar{\bar{J}}| + \frac{1}{2}(|\bar{I}| + |\bar{I}'|) = \frac{1}{4}(|\bar{\bar{J}}| + |\bar{I}| + |\bar{I}'|) + \frac{1}{4}(|\bar{I}| + |\bar{I}'|)$$
$$\ge \frac{1}{4}|I_0| + \frac{1}{4} \cdot (2\kappa - |I_0|) = \frac{\kappa}{2}.$$

□

## 1.6 Conclusions and Implications for Oblivious Sketching

Our approach provides a sharp generalization of the near equivalence between Johnson-Lindenstrauss property and restricted isometry property from [KW11]; the special case $d = 1$ in our work recovers the result of [KW11]. We prove the Johnson-Lindenstrauss property without any assumption on the vectors it is applied to, i.e., it is not necessary for them to have Kronecker structure. As Section 1.3 shows, Corollary 1.4 is optimal with respect to the dependence on the probability $\eta$ even for vectors with Kronecker structure, implying that even for this case, the dependence on the required sparsity level $s$ on $\eta$ in Theorem 1.3 is optimal.

With this provably optimal $\eta$ dependence, Corollary 1.4 also provides an improvement compared to Lemma 4.11 in [Ahl+20a]. In that work, the construction $P_\Omega H D_\xi$ as in Corollary 1.4 is introduced as TensorSRHT and is used as one element of a more extensive fast embedding for vectors with Kronecker structure which allows for a computational complexity that is only polynomial in the degree $d$. This embedding is based on a tree structure. Starting from a vector $x = x^{(1)} \otimes \cdots \otimes x^{(d)}$ with Kronecker structure, first a sparse Johnson-Lindenstrauss transform (OSNAP) is applied to each $x^{(j)}$ from $n$ to $m \geq m_1 = \Theta(\epsilon^{-2} \log \frac{1}{\eta})$ dimensions (Lemma 4.8 in [Ahl+20a]). Subsequently the TensorSRHT is applied to $\frac{d}{2}$ pairs of these vectors, reducing the corresponding Kronecker products of two factors separately. In this way, the result is a reduced Kronecker product of $\frac{d}{2}$ factors. This reduction is applied successively until only a single factor remains at the end. In each level, the TensorSRHT acts as an embedding $\mathbb{R}^{m^2} \to \mathbb{R}^m$ for a suitable $m \geq m_1$. As such, the dimension is reduced from $m^d$ to $m$ after the application of OSNAP.

Observe that this construction uses the setup of Corollary 1.4 for the case $d = 2$ and $N = m^2$. Choose

$$m := \left\lceil C\epsilon^{-2} \left(\log \frac{1}{\epsilon}\right)^2 \left(\log \frac{1}{\eta}\right)^2 \left(\log \frac{\log \frac{1}{\eta}}{\epsilon}\right)^3 \right\rceil .$$

Then for sufficiently large $C$, $m \geq m_1$ such that OSNAP provides a suitable embedding $\mathbb{R}^{m^2} \to \mathbb{R}^m$. Also, as required by the aforementioned construction, after choosing the RIP matrix with constant success probability, Corollary 1.4 provides an embedding $\mathbb{R}^{m^2} \to \mathbb{R}^m$ satisfying the $(\epsilon, \eta)$-distributional Johnson-Lindenstrauss property since the required embedding dimension is

$$\begin{aligned} m' &= C'\epsilon^{-2} \left(\log \frac{1}{\epsilon}\right)^2 \left(\log \frac{1}{\eta}\right)^2 \left(\log(m^2)\right) \left(\log \frac{\log \frac{1}{\eta}}{\epsilon}\right)^2 \\ &\leq 2\tilde{C}'(\log C)\epsilon^{-2} \left(\log \frac{1}{\epsilon}\right)^2 \left(\log \frac{1}{\eta}\right)^2 \left(\log \frac{\log \frac{1}{\eta}}{\epsilon}\right)^3 \end{aligned}$$

which is $\leq m$ for sufficiently large $C$. So omitting $\log \frac{1}{\epsilon}$ and $\log \log \frac{1}{\eta}$ factors, our result requires an embedding dimension $m$ of $\Omega\left(\epsilon^{-2} \left(\log \frac{1}{\eta}\right)^2\right)$ compared to the dimension $\Omega\left(\epsilon^{-2} \left(\log \frac{1}{\eta}\right)^3\right)$ in [Ahl+20a]. Thus, our result leads to both an improved embedding power and, consequently, an improved computational complexity of the tensor computation procedure.

# 2 The Hanson-Wright Inequality for Random Tensors

This section, except for part 2.5, shares major similarities with the article "The Hanson-Wright Inequality for Random Tensors" by authors Stefan Bamberger, Felix Krahmer, and Rachel Ward, that was submitted to *Sampling Theory, Signal Processing, and Data Analysis.* A preprint of this work is available at `https://arxiv.org/abs/2106.13345`, [BKW21b].

The corresponding source for part 2.5 is mentioned at the beginning of Section 1.

## 2.1 Introduction

### 2.1.1 Background and Studied Objects

Given a matrix $A \in \mathbb{R}^{n \times n}$ and a random vector $X \in \mathbb{R}^n$, the Hanson-Wright inequality provides a tail bound for the chaos $X^T AX - \mathbb{E}X^T AX$. In the original work [HW71], $X$ was assumed to have independent subgaussian entries whose distributions are symmetric about 0.

This result has been improved and adapted to various settings in a number of works. For example, the version shown in the introduction (Theorem 0.11), which is cited from [RV13], holds for vectors with general subgaussian entries without the symmetry assumption of the distribution:

**Theorem 2.1** (Theorem 1.1 from [RV13])**.** *Let $A \in \mathbb{R}^{n \times n}$. Let $X \in \mathbb{R}^n$ be a random vector with independent entries such that $\mathbb{E}X = 0$ and such that $X$ has a subgaussian norm of at most $K$. Then for every $t \geq 0$,*

$$\mathbb{P}(|X^T AX - \mathbb{E}X^T AX| > t) \leq 2 \exp \left[ -c \min \left\{ \frac{t^2}{K^4 \|A\|_F^2}, \frac{t}{K^2 \|A\|_{2\to 2}} \right\} \right]$$

*where $\|A\|_F$ is the Frobenius and $\|A\|_{2\to 2}$ the spectral norm of $A$.*

Today, the Hanson-Wright inequality is an important probabilistic tool and can be found in various textbooks covering the basics of signal processing and probability theory, such as [FR13] and [Ver18]. It has found numerous applications, in particular it has been a key ingredient for the construction of fast Johnson-Lindenstrauss embeddings [KW11].

For subgaussian $X \in \mathbb{R}^n$, linear expressions $\sum_{k=1}^n a_k X_k$ can be controlled by Hoeffding's inequality, while quadratic (order 2) expressions $X^T AX = \sum_{j,k=1}^n A_{j,k} X_j X_k$ can be controlled by the Hanson-Wright inequality. Thus, it is natural to wonder to what extent such control extends to a higher-order subgaussian chaos of the form

$$\sum_{i_1,\dots,i_d} A_{i_1,\dots,i_d} X_{i_1} \dots X_{i_d}. \tag{2.1}$$

Expressions of this type for subgaussian vectors have been considered in [AW15] where they are controlled using specific tensor norms of the arrays of all expected partial derivatives of certain degree with respect to the entries in $X$.

In contrast, for *independent* random vectors $X^{(1)}, \dots, X^{(d)}$, the decoupled chaos

$$\sum_{i_1,i_2,\dots,i_d=1}^{n} A_{i_1,\dots,i_d} X_{i_1}^{(1)} \dots X_{i_d}^{(d)}, \tag{2.2}$$

can be controlled with simpler bounds and has been considered in multiple previous works for numerous different distributions of the random vectors [Lat06; AL12; KL15].

In the course of adapting fast Johnson-Lindenstrauss embeddings to data with Kronecker structure as introduced in [BBK18] (see also [Ahl+20a; JKW20]), one encounters expressions of the form $(X^{(1)} \otimes \cdots \otimes X^{(d)})^T A(X^{(1)} \otimes \cdots \otimes X^{(d)})$ which are somewhat intermediate between (2.1) and (2.2), as they can be expanded as

$$\sum_{i_1,\dots,i_{2d}=1}^{n} A_{i_1,\dots,i_d,i_{d+1},\dots,i_{2d}} X^{(1)}_{i_1} \dots X^{(d)}_{i_d} X^{(1)}_{i_{d+1}} \dots X^{(d)}_{i_{2d}}. \tag{2.3}$$

Vershynin [Ver20] recently studied embeddings of random tensors which requires controlling

$$\|B(X^{(1)} \otimes \cdots \otimes X^{(d)})\|_2 \tag{2.4}$$

for a matrix $B$. This is of the form (2.3) where $A$ is a rearrangement of $B^T B$.

Even though (2.3) can be cast as a specific case of (2.1) for which [AW15] provides optimal bounds, these bounds are not straightforward to use in this specific situation since they are given in terms of partial derivatives and not in terms of the coefficients $A_{i_1,\dots,i_{2d}}$.

The main results of this paper provide moment estimates for the semi-decoupled chaos process (2.3) that are easier to use as they are explicitly given in terms of the coefficients $A_{i_1,\dots,i_{2d}}$. Our bounds imply improved estimates for (2.4) and lay the foundations for the order-optimal analysis of fast Kronecker-structured Johnson-Lindenstrauss embeddings from Section 1 in the case of arbitrary order $d$ which will be completed at the end of this section. We nevertheless expect that our results should find broader use beyond these specific applications.

### 2.1.2 Previous Work

For the case where $X^{(1)}, \dots, X^{(d)}$ are independent *Gaussian* vectors, the concentration of (2.2) has been studied in [Lat06] which provides upper and lower moment bounds which match up to a constant factor depending only on the order $d$. We will obtain our main results for subgaussian vectors by careful reduction to the Gaussian bounds.

Higher order chaos expressions have also been studied for distributions beyond Gaussian. Specifically, [Bou+05], Section 9, considers (2.1) for the case of Rademacher vectors. However, the bounds are more intricate than in [Lat06] and the coefficient array $\mathbf{A} = (A_{i_1,\dots,i_d})_{i_1,\dots,i_d=1}^n$ must satisfy a symmetry condition and be diagonal-free, i.e., $A_{i_1,\dots,i_d} = 0$ if any two of the indices $i_1, \dots, i_d$ coincide.

Upper and lower bounds on the moments of (2.2) are shown in [AL12] and [KL15] for the case of symmetric random variables with logarithmically concave and convex tails, meaning that for a random variable $X \in \mathbb{R}$, the function $t \mapsto -\log \mathbb{P}(|X| \geq t)$ is convex or concave, respectively. However, for general subgaussian random variables, neither of these has to be the case. In addition, these works only consider the decoupled chaos (2.2) and provide a decoupling inequality to control (2.1) for diagonal-free $\mathbf{A}$.

Upper moment bounds for general polynomials of independent subgaussian random variables are provided in [AW15]. Similar to our work, the authors utilize the decoupling techniques of [AG93]. Since (2.3) is a polynomial in the entries of $X^{(1)}, \dots, X^{(d)}$, it can also be controlled using the results from [AW15]. Because the aforementioned work also shows that these moment bounds are tight for the case of Gaussian vectors, one of the main results (Theorem 2.9) of our work can also be shown using their results. However, their result bounds the corresponding $L_p$ norms in terms of norms of the array of all $d' \leq 2d$ expected partial derivatives, meaning that significant additional work would be required to relate these derivatives to the expressions in Theorem 2.9. We believe, that our approach is not much longer but more insightful. In addition, it provides the decoupling result Theorem 2.11 which will be of independent interest.

More work on related topics include [Mel16; Mel19] where upper and lower bounds for the case of random variables satisfying the moment condition $\|X\|_{2p} \leq \alpha \|X\|_p$ are considered for the case of positive variables of order 2. The recent work [GSS21] provides similar bounds to [AW15] for distributions of bounded $\psi_\alpha$ norm for $\alpha \in (0, 1]$ (or $\alpha \in (0, 2]$ for some fo their results), such as subexponential distributions. Like in [AW15], their bounds are given in terms of partial derivatives, not directly in terms of the coefficients.

The decoupling technique used in many proofs of the standard Hanson-Wright inequality relates $X^TAX$ to $X^TA\bar{X}$ where $\bar{X}$ is an independent copy of $X$. This approach was first introduced in [MT86], already in a general higher-dimensional form. The general idea is to upper bound convex functions (e.g. moments) of (2.1) by the corresponding expressions of (2.2), up to a constant. Beside independent, symmetrically distributed entries of the random vectors, the result also requires the coefficient array to be symmetric and diagonal free.

The subsequent work [Kwa87] has also shown the reverse decoupling bound, up to constant factors, proving that through (2.2), one can also provide lower bounds on the moments of (2.1) with the same assumptions on the coefficient array. However, in some applications it can be interesting to consider non-diagonal-free coefficient arrays. For example, in the scenario of $\|B(X^{(1)} \otimes \cdots \otimes X^{(d)})\|_2^2$, the coefficient array $B^TB$ cannot be expected to fulfill the diagonal-free condition in general. The work in [AG93] lifts the restriction of a diagonal-free coefficient array and bounds the tails of slight modifications of (2.2) and (2.1) by each other up to certain constants in the case of Gaussian random variables.

The concentration of the norm (2.4) has recently been studied for the subgaussian case in [Ver20]. It is shown that

$$\mathbb{P}\left(\left|\|B(X^{(1)} \otimes \cdots \otimes X^{(d)})\|_2 - \|B\|_F\right| > t\right) \leq 2\exp\left(-\frac{ct^2}{dn^{d-1}\|B\|_{2\to 2}^2}\right) \tag{2.5}$$

for an absolute constant $c$ and for $0 \leq t \leq 2n^{\frac{d}{2}}\|B\|_{2\to 2}$. This bound suggests that techniques like the chaos moment bounds in [Lat06] could be applied to this problem, which is what we do in this work and leads to Theorem 2.13 below.

### 2.1.3 Overview of our Contribution

The goal of this work is to provide upper and lower bounds for the moments of the deviation of (2.3) from its expectation for vectors with independent subgaussian entries (Theorem 2.9 below). Key steps of the proof include a decoupling inequality for expressions of the form (2.3), Theorem 2.11, and a comparison to Gaussian random vectors. Finally, based on our results for (2.3), we provide a concentration inequality for (2.4) as stated in Theorem 2.13 which extends previous results of [Ver20].

Possible applications of such results include recent developments in norm-preserving maps for vectors with tensor structure in the context of machine learning methods using the kernel trick [BBK18; Ahl+20a; JKW20].

### 2.1.4 Notation

Our results on $X^TAX$ where $X$ is a Kronecker product of $d$ random vectors will depend crucially on the structure of the coefficient matrix $A$ rearranged as a higher-order (specifically order $2d$) array. As such, we must establish sophisticated notation for such arrays and their indices.

Consider a vector of dimensions $\mathbf{n} = (n_1, n_2, \ldots, n_d)$ and a subset $I \subset [d]$. We call a function $\mathbf{i} : I \to \mathbb{N}$ a partial index of order $d$ on $I$ if for all $l \in I$, $\mathbf{i}_l := \mathbf{i}(l) \in [n_l]$. Assume there is exactly one such function if $I = \emptyset$. If $I = [d]$, then $\mathbf{i}$ is called an index of order $d$. We denote the set of all partial indices of order $d$ on $I$ as $\mathbf{J^n}(I)$; the set of all indices of order $d$ is denoted by $\mathbf{J^n} := \mathbf{J^n}([d])$. $\mathbf{J^n}$ can be identified with $[n_1] \times \cdots \times [n_d]$.

A function $\mathbf{B} : \mathbf{J^n} \to \mathbb{R}$ is called an array of order $d$. Because of the aforementioned identification, we also write $\mathbf{B} \in \mathbb{R}^{n_1 \times \cdots \times n_d} =: \mathbb{R}^{\mathbf{n}}$. For $I \subset [d]$, we define $\mathbb{R}^{\mathbf{n}}(I)$ to be the set of partial arrays $\mathbf{B} : \mathbf{J^n}(I) \to \mathbb{R}$. For $I = [d]$, this is just the aforementioned array definition.

We denote

$$\|\mathbf{B}\|_2 := \left[\sum_{\mathbf{i}\in\mathbf{J}^{\mathbf{n}}(I)} B_{\mathbf{i}}^2\right]^{\frac{1}{2}}$$

for the Frobenius norm of the (partial) array where $B_{\mathbf{i}} := \mathbf{B}(\mathbf{i})$ are its entries.

For disjoint sets $I$, $J \subset [d]$ and corresponding partial indices $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(I)$, $\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(J)$, define the partial index $\mathbf{i}\dot{\times}\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(I \cup J)$ by

$$(\mathbf{i}\dot{\times}\mathbf{j})_l = \begin{cases} \mathbf{i}_l & \text{if } l \in I \\ \mathbf{j}_l & \text{if } l \in J. \end{cases} \tag{2.6}$$

We will often work with arrays of order $2d$ whose dimensions along the first $d$ axes are the same as the dimensions along the remaining $d$ ones. We use the notation

$$\mathbf{n}^{\times 2} = (n_1, \ldots, n_d, n_1, \ldots, n_d)$$

for the dimensions of such arrays.

For sets $I \subset [2d]$, $J \subset [d]$ such that $I \cap (J + d) = \emptyset$ and for corresponding partial indices $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(I)$, $\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(J)$, define the partial index $\mathbf{i}\dot{+}\mathbf{j} \in \mathbf{J}^{\mathbf{n}^{\times 2}}(I \cup (J + d))$ by

$$(\mathbf{i}\dot{+}\mathbf{j})_l = \begin{cases} \mathbf{i}_l & \text{if } l \in I \\ \mathbf{j}_{l-d} & \text{if } l \in J + d. \end{cases} \tag{2.7}$$

For $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(I)$ and $J \subset I$, define $\mathbf{i}_J \in \mathbf{J}^{\mathbf{n}}(J)$ to be the restriction of $\mathbf{i}$ to $J$, i.e., $(\mathbf{i}_J)_l = \mathbf{i}_l$ for all $l \in J$.

**Remark 2.2.** *To see how the $\dot{+}$ and $\dot{\times}$ operators work, consider the following example for $d = 3$ and $n_1 = n_2 = n_3 = 10$. Take the subset $\{1, 3\} \subset [3]$ and partial indices $\mathbf{i}, \mathbf{j} \in \mathbf{J}^{\mathbf{n}}(\{1, 3\})$ such that $\mathbf{i}_1 = 2, \mathbf{i}_3 = 4$, and $\mathbf{j}_1 = 5, \mathbf{j}_3 = 8$. Then $\mathbf{i}$ and $\mathbf{j}$ each define indices along the axes 1 and 3 but not along axis 2.*

*We can join $\mathbf{i}$ and $\mathbf{j}$ using the $\dot{+}$ operator to one index $\mathbf{i}\dot{+}\mathbf{j} =: \mathbf{k} \in \mathbf{J}^{\mathbf{n}^{\times 2}}(\{1, 3, 4, 6\})$. Then $\mathbf{k}_1 = \mathbf{i}_1 = 2$, $\mathbf{k}_3 = \mathbf{i}_3 = 4$ and the entries of $\mathbf{j}$ get shifted by $d = 3$ such that $\mathbf{k}_4 = \mathbf{j}_1 = 5$, $\mathbf{k}_6 = \mathbf{j}_3 = 8$.*

*Now to extend $\mathbf{k} \in \mathbf{J}^{\mathbf{n}^{\times 2}}(\{1, 3, 4, 6\})$ to the remaining axes 2 and 5, we can join it with another partial index $\mathbf{l} \in \mathbf{J}^{\mathbf{n}^{\times 2}}(\{2, 5\})$ using the $\dot{\times}$ operator. Assume $\mathbf{l}_2 = 3$ and $\mathbf{l}_5 = 7$, then we obtain the total index $\mathbf{m} := \mathbf{k}\dot{\times}\mathbf{l} \in \mathbf{J}^{\mathbf{n}^{\times 2}}$ for which*

$$\mathbf{m}_1 = \mathbf{k}_1 = 2,\ \mathbf{m}_2 = \mathbf{l}_2 = 3,\ \mathbf{m}_3 = \mathbf{k}_3 = 4,\ \mathbf{m}_4 = \mathbf{k}_4 = 5,\ \mathbf{m}_5 = \mathbf{l}_5 = 7,\ \mathbf{m}_6 = \mathbf{k}_6 = 8,$$

*such that $\mathbf{m}$ corresponds to the 6-tuple $(2, 3, 4, 5, 7, 8)$. Note that in contrast to the $\dot{+}$ operator, $\dot{\times}$ does not shift the entries of the second index.*

The following function establishes a relation between array indices and indices of the rearrangement of the array as a vector.

**Definition 2.3.** *For a dimension vector $\mathbf{n} = (n_1, n_2, \ldots, n_d)$, a subset $I = \{j_1, \ldots, j_{|I|}\} \subset [d]$ for $j_1 < \cdots < j_{|I|}$ and $N := \prod_{l\in I} n_l$, define the function $\mathcal{I}_I^{\mathbf{n}} : \mathbf{J}^{\mathbf{n}}(I) \to [N]$ by*

$$\mathcal{I}_I^{\mathbf{n}}(\mathbf{j}) = \sum_{l=1}^{|I|} (\mathbf{i}_{j_l} - 1) \prod_{l'=1}^{l-1} n_{j_{l'}} + 1.$$

*which defines a bijection. Its inverse is called $\hat{\mathcal{I}}_I^{\mathbf{n}} : [N] \to \mathbf{J}^{\mathbf{n}}(I)$. We define $\mathcal{I}^{\mathbf{n}} := \mathcal{I}_{[d]}^{\mathbf{n}}$.*

**Definition 2.4.** *For an array* $\mathbf{a} \in \mathbb{R}^{\mathbf{n}}$*, the vectorization* $\operatorname{vec}(\mathbf{a}) \in \mathbb{R}^N$ *is defined such that for all* $\mathbf{j} \in \mathbf{J}^{\mathbf{n}}$*,* $(\operatorname{vec}(\mathbf{a}))_{\mathcal{I}^{\mathbf{n}}(\mathbf{j})} = \mathbf{a}_{\mathbf{j}}$*.*

**Definition 2.5.** *Let* $I \subset J \subset [d]$*. For a partial index* $\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(J)$*, define the restriction* $\mathbf{j}_I \in \mathbf{J}^{\mathbf{n}}(I)$ *such that for all* $l \in I$*,* $(\mathbf{j}_I)_l = \mathbf{j}_l$*.*

As suggested by the explanations above, our convention is to use bold letters for higher order arrays (e.g., $\mathbf{A}$) while their entries are denoted in non-bold letters (e.g., $A_{\mathbf{i}}$). For some of our results, we will convert matrices into higher-order arrays by rearranging their entries. In these cases, we will denote the matrices in non-bold letters and use the same letter in bold for the array, e.g., $A$ and $\mathbf{A}$. For the entries, it will be clear from the indices which object is being referred to. Besides that, we will also always use bold letters for array indices (e.g., $\mathbf{i}$), for vectors of array dimensions (e.g. $\mathbf{n}$), and for the set $\mathbf{J}^{\mathbf{n}}$.

### 2.1.5 Previous Relevant Results

Since our result is based on the bounds given by Latala in [Lat06], we also consider the following norms which are also used in that result. In our notation, the norms of interest are stated as follows.

**Definition 2.6.** *For* $\mathbf{n} \in \mathbb{N}^d$ *and an array* $\mathbf{B} \in \mathbb{R}^{\mathbf{n}}$*, we define the following norms for any partition* $I_1, \ldots, I_\kappa$ *of* $[d]$*.*

$$\|\mathbf{B}\|_{I_1,\ldots,I_\kappa} := \sup_{\substack{\alpha^{(1)} \in \mathbb{R}^{\mathbf{n}}(I_1),\ldots,\alpha^{(\kappa)} \in \mathbb{R}^{\mathbf{n}}(I_\kappa), \\ \|\alpha^{(1)}\|_2 = \cdots = \|\alpha^{(\kappa)}\|_2 = 1}} \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i}} \alpha^{(1)}_{\mathbf{i}_{I_1}} \ldots \alpha^{(\kappa)}_{\mathbf{i}_{I_\kappa}}.$$

For example, when $d = 2$, the array $\mathbf{B}$ is a matrix and $\|\cdot\|_{\{1,2\}}$ coincides with the Frobenius and $\|\cdot\|_{\{1\},\{2\}}$ with the spectral norm. Latala [Lat06] proved the following upper and lower moment bounds for a decoupled Gaussian chaos of arbitrary order. Even though it is only shown for $p \geq 2$ in [Lat06], it holds for all $p \geq 1$ as explained in Remark 2.8 below.

**Theorem 2.7** (Theorem 1 in [Lat06]). *Let* $\mathbf{n} \in \mathbb{N}^d$*,* $\mathbf{B} \in \mathbb{R}^{\mathbf{n}}$*,* $p \geq 1$*.*

*Let* $S(d, \kappa)$ *denote the set of partitions of* $[d]$ *into* $\kappa$ *nonempty disjoint subsets. Define*

$$m_p(\mathbf{B}) := \sum_{\kappa=1}^{d} p^{\kappa/2} \sum_{(I_1,\ldots,I_\kappa) \in S(d,\kappa)} \|\mathbf{B}\|_{I_1,\ldots,I_\kappa}. \tag{2.8}$$

*Consider independent Gaussian random vectors* $g^{(1)} \sim N(0, Id_{\mathbf{n}_1}), \ldots, g^{(d)} \sim N(0, Id_{\mathbf{n}_d})$*. Then*

$$\frac{1}{C(d)} m_p(\mathbf{B}) \leq \left\| \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i}} \prod_{l \in [d]} g^{(l)}_{\mathbf{i}_l} \right\|_{L_p} \leq C(d) m_p(\mathbf{B}),$$

*where* $C(d) > 0$ *is a constant that only depends on* $d$*.*

**Remark 2.8.** *Theorem 1 in [Lat06] only shows this statement for* $p \geq 2$*. However, by a small adjustment, we can see that it also holds for* $1 \leq p \leq 2$ *with a possibly different* $C(d)$*. Let* $X := \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i}} \prod_{l \in [d]} g^{(l)}_{\mathbf{i}_l}$*. For the upper bound we have for* $1 \leq p \leq 2$*,*

$$\|X\|_{L_p} \leq \|X\|_{L_2} \leq C(d) m_2(\mathbf{B}) \leq 2^{\frac{d}{2}} C(d) m_p(\mathbf{B}).$$

*For the lower bound, we consider the recent work [ALM21] about a generalized Gaussian chaos with values in an arbitrary Banach space. Theorem 2.1 in their work states the lower bound*

$$\frac{1}{C(d)}\sum_{J\subset[d]}\sum_{\mathcal{P}\in\mathcal{P}(J)} p^{|\mathcal{P}|/2}|||\mathbf{B}|||_{\mathcal{P}} \leq \|X\|_{L_p}, \tag{2.9}$$

*for all $p \geq 1$, where $\mathcal{P}(J)$ is defined as the set of all partitions of $J$ (into non-empty, pairwise disjoint sets) and $|||\mathbf{B}|||_{\mathcal{P}}$, defined in (2.2) of [ALM21], is a non-negative expression that coincides with our definition of $\|\mathbf{B}\|_{I_1,\ldots,I_\kappa}$ if $\mathcal{P}=(I_1,\ldots,I_\kappa)$ is a partition of the entire set $[d]$. Therefore we can restrict the sum over $J$ in* (2.9) *to the term $J=[d]$ and obtain*

$$\frac{1}{C(d)}m_p(\mathbf{B}) = \frac{1}{C(d)}\sum_{\mathcal{P}\in\mathcal{P}([d])} p^{|\mathcal{P}|/2}|||\mathbf{B}|||_{\mathcal{P}} \leq \|X\|_{L_p}.$$

## 2.2 Main Results

The main contribution of our work is the following result which gives a generalization of the Hanson-Wright inequality (Theorem 2.1) in terms of upper and lower moment bounds. Note that the operators $\dot{\times}$ and $\dot{+}$ are defined in (2.6) and (2.7).

**Theorem 2.9.** *For $d \geq 1$, let $\mathbf{n}=(n_1,\ldots,n_d)$ be a vector of dimensions, and let $N = n_1 \ldots n_d$.*

*Let $A \in \mathbb{R}^{N\times N}$ and $X^{(1)} \in \mathbb{R}^{n_1},\ldots,X^{(d)} \in \mathbb{R}^{n_d}$ be random vectors with independent, mean* 0*, variance* 1 *entries with subgaussian norms bounded by $L \geq 1$. Define $X := X^{(1)}\otimes\cdots\otimes X^{(d)}$. There exists a constant $C(d)$, depending only on $d$, such that for all $p \geq 1$,*

$$\left\|X^TAX - \mathbb{E}X^TAX\right\|_{L_p} \leq C(d)m_p.$$

*The numbers $m_p$ are defined as follows. By rearranging its entries, regard $A$ as an array $\mathbf{A} \in \mathbb{R}^{\mathbf{n}^{\times 2}}$ of order $2d$ such that*

$$X^TAX = \sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}} A_{\mathbf{i}\dot{+}\mathbf{i}'}\prod_{l\in[d]} X^{(l)}_{\mathbf{i}_l}X^{(l)}_{\mathbf{i}'_l}.$$

*For any $I \subset [d]$ and for $I^c = [d]\backslash I$, define $\mathbf{A}^{(I)} \in \mathbb{R}^{\mathbf{n}^{\times 2}}(I^c \cup (I^c+d))$ by*

$$A^{(I)}_{\mathbf{i}\dot{+}\mathbf{i}'} = \sum_{\mathbf{k}\in\mathbf{J}^{\mathbf{n}}(I)} A_{(\mathbf{i}\dot{\times}\mathbf{k})\dot{+}(\mathbf{i}'\dot{\times}\mathbf{k})} \tag{2.10}$$

*for all $\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}(I^c)$.*

*For $T \subset [2d]$ and $1 \leq \kappa \leq 2d$, denote by $S(T,\kappa)$ the set of partitions of $T$ into $\kappa$ sets. Then for any $p \geq 1$, define*

$$m_p := L^{2d}\sum_{\kappa=1}^{2d} p^{\frac{\kappa}{2}} \sum_{\substack{I\subset[d]\\ I\neq[d]}} \sum_{(I_1,\ldots,I_\kappa)\in S((I^c)\cup(I^c+d),\kappa)} \|\mathbf{A}^{(I)}\|_{I_1,\ldots,I_\kappa}.$$

*If in addition, $X^{(1)} \sim N(0, Id_{n_1}),\ldots,X^{(d)} \sim N(0, Id_{n_d})$ are normally distributed (i.e. $L$ is constant), and $\mathbf{A}$ satisfies the symmetry condition that for all $l \in [d]$ and any $\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}([d]\backslash\{l\})$, $\mathbf{j},\mathbf{j}' \in \mathbf{J}^{\mathbf{n}}(\{l\})$,*

$$A_{(\mathbf{i}\dot{\times}\mathbf{j})\dot{+}(\mathbf{i}'\dot{\times}\mathbf{j}')} = A_{(\mathbf{i}\dot{\times}\mathbf{j}')\dot{+}(\mathbf{i}'\dot{\times}\mathbf{j})}, \tag{2.11}$$

*then also the lower bound*

$$\tilde{C}(d)m_p \leq \left\|X^TAX - \mathbb{E}X^TAX\right\|_{L_p}$$

*holds for all $p \geq 1$. Here, $\tilde{C}(d) > 0$ only depends on $d$.*

Note that these upper bounds can directly be converted to tail bounds in the style of Theorems 2.1 or 2.13 using Lemma 2.22. After introducing the required tools, the proof of Theorem 2.9 will be split up into two parts. We will prove the upper bound in Subsection 2.3.2 and then the lower bound in Subsection 2.3.3.

**Remark 2.10.** *The symmetry condition required for the lower bound is not satisfied for all matrices. However, for any matrix $A$, we can find a matrix $\tilde{A}$ satisfying the symmetry condition and such that $X^T A X = X^T \tilde{A} X$ always holds. To do this, in the array notation we can define $\tilde{\mathbf{A}}$ by transposing $\mathbf{A}$ along all possible sets of axes and then taking the mean $\tilde{A}_{\mathbf{i} \dot{+} \mathbf{i}'} = \frac{1}{2^d} \sum_{I \subset [d]} A_{(\mathbf{i}_{I^c} \dot{\times} \mathbf{i}'_I) \dot{+} (\mathbf{i}_I \dot{\times} \mathbf{i}'_{I^c})}$ for any $\mathbf{i}, \mathbf{i}' \in \mathbf{J}^{\mathbf{n}}$. This is a generalization of taking $\tilde{A} = \frac{1}{2}(A + A^T)$ for $d = 1$. Note however, that $\tilde{A}$ might have significantly smaller norms than $A$ which is why the lower moment bounds in Theorem 2.9 might not hold for $A$ directly.*

A central part of our argument is the following specialized decoupling result for expressions as in (2.3) which might be of independent interest.

**Theorem 2.11.** *Let $\mathbf{n} = (n_1, \ldots, n_d) \in \mathbb{N}^d$, $\mathbf{A} \in \mathbb{R}^{\mathbf{n}^{\times 2}}$, $X^{(1)} \in \mathbb{R}^{n_1}, \ldots, X^{(d)} \in \mathbb{R}^{n_d}$ random vectors with independent mean $0$, variance $1$ entries and $\bar{X}^{(1)}, \ldots, \bar{X}^{(d)}$ corresponding independent copies. Then for all $p \geq 1$,*

$$\left\| \sum_{\mathbf{i}, \mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} A_{\mathbf{i} \dot{+} \mathbf{i}'} \prod_{l \in [d]} X^{(l)}_{\mathbf{i}_l} X^{(l)}_{\mathbf{i}'_l} - \mathbb{E} \sum_{\mathbf{i}, \mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} A_{\mathbf{i} \dot{+} \mathbf{i}'} \prod_{l \in [d]} X^{(l)}_{\mathbf{i}_l} X^{(l)}_{\mathbf{i}'_l} \right\|_{L_p}$$

$$\leq \sum_{\substack{I, J \subset [d]: \\ J \subset I,\, I \setminus J \neq [d]}} 4^{d - |I|} \left\| \sum_{\substack{\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(J) \\ \mathbf{j} \in \mathbf{J}^{\mathbf{n}}(I \setminus J) \\ \mathbf{k}, \mathbf{k}' \in \mathbf{J}^{\mathbf{n}}(I^c)}} A_{\substack{(\mathbf{i} \dot{\times} \mathbf{j} \dot{\times} \mathbf{k}) \\ \dot{+} (\mathbf{i} \dot{\times} \mathbf{j} \dot{\times} \mathbf{k}')}} \prod_{l \in J} \left[ (X^{(l)}_{\mathbf{i}_l})^2 - 1 \right] \prod_{l \in I^c} X^{(l)}_{\mathbf{k}_l} \bar{X}^{(l)}_{\mathbf{k}'_l} \right\|_{L_p}$$

**Remark 2.12.** *Consider the special case in Theorem 2.11 of $X^{(1)}, \ldots, X^{(d)}$ being Rademacher vectors, i.e., having independent entries that are $\pm 1$ with a probability of $\frac{1}{2}$ each. Then any squared entry is $1$ almost surely. This implies that the factor $\prod_{l \in J} \left[ (X^{(l)}_{\mathbf{i}_l})^2 - 1 \right]$ is $0$ unless $J = \emptyset$. So on the right hand side of the inequality in Theorem 2.11, only the terms with $J = \emptyset$ need to be considered.*

Theorem 2.9 also leads to the following new tail bound for $\|A(X^{(1)} \otimes \cdots \otimes X^{(d)})\|_2$. Note that it contains the deviation of the non-squared norm. This improves upon the previous result by Vershynin [Ver20] as described in (2.5), up to the constant $C(d)$. By comparison, our result provides a strictly stronger bound for matrices with smaller Frobenius norm and holds for all $t \geq 0$.

**Theorem 2.13.** *Let $B \in \mathbb{R}^{n_0 \times n^d}$ be a matrix, $X^{(1)}, \ldots, X^{(d)} \in \mathbb{R}^n$ independent random vectors with independent, mean $0$, variance $1$ entries with subgaussian norm bounded by $L \geq 1$, and let $X := X^{(1)} \otimes \cdots \otimes X^{(d)} \in \mathbb{R}^{n^d}$. Then for a constant $C(d)$ depending only on $d$ and for any $t > 0$,*

$$\begin{aligned} &\mathbb{P}\left( \left| \|BX\|_2 - \|B\|_F \right| > t \right) \\ &\leq \begin{cases} e^2 \exp\left( -C(d) \frac{t^2}{n^{d-1} \|B\|_{2\to 2}^2} \right) & \text{\textit{if} } t \leq n^{\frac{d}{2}} \|B\|_{2\to 2} \\ e^2 \exp\left( -C(d) \left( \frac{t}{\|B\|_{2\to 2}} \right)^{\frac{2}{d}} \right) & \text{\textit{if} } t \geq n^{\frac{d}{2}} \|B\|_{2\to 2} \\ e^2 \exp\left( -C(d) \frac{t^2}{n^{\frac{d-1}{2}} \|B\|_F^2} \right) & \text{\textit{if} } n^{\frac{d-1}{4}} \|B\|_{2\to 2} \leq t \leq n^{\frac{d-1}{4}} \|B\|_F. \end{cases} \end{aligned}$$

Note that the third interval intersects the first two intervals. In any interval of intersection, both bounds hold. For slightly more complicated but provably optimal moment bounds, we refer the reader to Corollary 2.32.

**Remark 2.14.** *In addition to extending the previous result in* (2.5) *from [Ver20] to all* $t \geq 0$*, our result provides a strict improvement of that result for matrices with stable rank* $(\|B\|_F / \|B\|_{2\to 2})^2$ *in* $(1, n^{\frac{d-1}{2}})$.

*As an example, consider a square matrix* $B \in \mathbb{R}^{n^d \times n^d}$ *of full rank with mildly exponentially decreasing singular values* $\sigma_j = e^{-\frac{1}{2}n^{-\frac{d}{4}}(j-1)}$ *for* $1 \leq j \leq n^d$*. Then* $\|B\|_{2\to 2} = \sigma_1 = 1$ *and one can check, using* $e^{-x} \geq 1 - x$ *for all* $x \in \mathbb{R}$ *and* $e^{-x} \leq 1 - \frac{x}{2}$ *for* $x \in [0,1]$*, that*

$$\|B\|_F^2 = \frac{1 - e^{-n^{\frac{3}{4}d}}}{1 - e^{-n^{-\frac{d}{4}}}} \in \left[\frac{1}{2}n^{\frac{d}{4}}, 2n^{\frac{d}{4}}\right]$$

*So the stable rank is* $\in [\frac{1}{2}n^{\frac{d}{4}}, 2n^{\frac{d}{4}}]$*. Indeed, for at least the (for large enough* $n$ *non-empty) interval* $n^{\frac{1}{4}d-\frac{1}{4}} \leq t \leq \frac{1}{2}n^{\frac{3}{8}d-\frac{1}{4}}$*, the third line in Theorem 2.13 provides a probability bound* $\leq e^2 \exp\left(-C(d)\frac{t^2}{2n^{\frac{3}{4}d-\frac{1}{2}}}\right)$ *while the first line only provides a bound of* $e^2 \exp\left(-C(d)\frac{t^2}{n^{d-1}}\right)$*, i.e., there is an improvement for* $d \geq 3$.

## 2.3 Main Proofs

### 2.3.1 Preliminaries

The classical symmetrization lemma for normed spaces, cited as Lemma 0.9, can be extended to increasing convex functions of norms as the following result from [Peñ92] shows.

**Lemma 2.15** (Special case of Lemma A1 in [Peñ92])**.** *Let* $X_1, \dots, X_n$ *be independent, mean* 0 *real-valued random variables and* $p \geq 1$*. Let* $\xi_1, \dots, \xi_n$ *be independent Rademacher variables that are independent of* $X_1, \dots, X_n$*. Then*

$$\frac{1}{2^p}\mathbb{E}\left|\sum_{k=1}^n \xi_k X_k\right|^p \leq \mathbb{E}\left|\sum_{k=1}^n X_k\right|^p \leq 2^p\mathbb{E}\left|\sum_{k=1}^n \xi_k X_k\right|^p$$

The decoupling theorem for quadratic forms relates double sums $\sum_{j,k=1}^n A_{j,k}X_jX_k$ over random variables $(X_j)_{j\in[n]}$ to a "decoupled" expression $\sum_{j,k=1}^n A_{j,k}X_j\bar{X}_k$ where the $\bar{X}_k$ are independent copies of the $X_k$. Different versions have been used in probability for a long time and we refer to Section 3.6 in [PG99] for an overview of their history. The following version for convex functions, from Theorem 8.11 in the textbook [FR13], is an adaption of Proposition 1.9 in [BT87] for norms in Banach spaces.

**Theorem 2.16.** *Let* $A \in \mathbb{R}^{n\times n}$ *be a matrix,* $X \in \mathbb{R}^n$ *a vector with independent mean* 0 *entries, and* $\bar{X}$ *and independent copy of* $X$*. Let* $F : \mathbb{R} \to \mathbb{R}$ *be a convex function. Then*

$$\mathbb{E}F\left(\sum_{\substack{j,k=1\\ j\neq k}}^n A_{jk}X_jX_k\right) \leq \mathbb{E}F\left(4\sum_{j,k=1}^n A_{jk}X_j\bar{X}_k\right)$$

Also the following elementary result will be used.

**Lemma 2.17.** *Let* $T$ *be a finite set. Then*

$$\sum_{S\subset T}(-1)^{|S|} = \begin{cases} 1 & \text{if } T = \emptyset \\ 0 & \text{otherwise.}\end{cases}$$

*Proof.* By grouping all $S \subset T$ of the same size and applying the binomial theorem,

$$\sum_{S\subset T}(-1)^{|S|} = \sum_{k=0}^{|T|}\sum_{\substack{S\subset T\\|S|=k}}(-1)^{|S|} = \sum_{k=0}^{|T|}\binom{|T|}{k}(-1)^k \cdot 1^{|T|-k} = (-1+1)^{|T|}$$
$$= \begin{cases}1 & \text{if } T=\emptyset\\ 0 & \text{otherwise.}\end{cases}$$

□

Although this is a very elementary statement and consequence of the binomial theorem, we are not aware of any previous usages of precisely this identity. One somewhat similar tool is given by Mazur-Orlicz formula ((11) in [MO34]), which has also been used in a problem related to decoupling inequalities in [PM95]. It is stated as

$$(-1)^k \sum_{\epsilon_1,\dots,\epsilon_k=0}^{1}(-1)^{-(\epsilon_1+\dots+\epsilon_k)}\epsilon_1^{v_1}\dots\epsilon_k^{v_k} = (1-0^{v_1})\dots(1-0^{v_k}).$$

With $v_1 = \dots = v_k = 0$, this becomes

$$(-1)^k \sum_{\epsilon_1,\dots,\epsilon_k=0}^{1}(-1)^{-(\epsilon_1+\dots+\epsilon_k)} = 0^k.$$

For $k = |T| > 0$, the $\{0,1\}$-tuples $(\epsilon_1,\dots,\epsilon_k)$ can be identified with the subsets $S \subset T$ such that $|S| = \epsilon_1 + \dots + \epsilon_k$ and then this identity implies Lemma 2.17 for $T \neq \emptyset$.

For the norms in Definition 2.6, we need the following property about restricting arrays to some diagonal entries. This can be obtained directly from a repeated application of Lemma 5.2 in [AW15] (where $K = \{l, l+d\}$ for each $l \in I$). Here again, we use the notation of $\dot{\times}$ and $\dot{+}$ from (2.6) and (2.7).

**Lemma 2.18.** *Let $\mathbf{A} \in \mathbb{R}^{\mathbf{n}^{\dot{\times}2}}$, $I \subset [d]$ and define $\mathbf{A}^{[I]} \in \mathbb{R}^{\mathbf{n}^{\dot{\times}2}}$ by*

$$A^{[I]}_{\mathbf{i}\dot{+}\mathbf{i}'} := \begin{cases}A_{\mathbf{i}\dot{+}\mathbf{i}'} & \textit{if } \forall l \in I : \mathbf{i}_l = \mathbf{i}'_l\\ 0 & \textit{otherwise.}\end{cases}$$

*for all $\mathbf{i}, \mathbf{i}' \in \mathbf{J}^{\mathbf{n}}$. Then for any partition $I_1,\dots,I_\kappa$ of $[2d]$, we have*

$$\|\mathbf{A}^{[I]}\|_{I_1,\dots,I_\kappa} \leq \|\mathbf{A}\|_{I_1,\dots,I_\kappa}.$$

For comparisons between functions of subgaussian and of Gaussian variables, we will use the concept of strong domination of random variables. See, e.g., [KW92] for the following definition and further explanations.

**Definition 2.19** (Definition 3.2.1 in [KW92])**.** *Let $X, Y \in \mathbb{R}$ be random variables and $\kappa, \lambda > 0$. We say that $X$ is $(\kappa,\lambda)$-strongly dominated by $Y$ ($X \prec_{(\kappa,\lambda)} Y$) if for every $t > 0$,*

$$\mathbb{P}(|X| > t) \leq \kappa\mathbb{P}(\lambda|Y| > t).$$

It can be shown that linear combinations of independent, strongly dominated random variables are again strongly dominated which in turn implies the following statement about expectations of convex functions of these linear combinations.

**Theorem 2.20** (Corollary 3.2.1 in [KW92]). *Let $X_1, \dots, X_n, Y_1, \dots, Y_n \in \mathbb{R}$ be independent symmetric random variables and $a_1, \dots, a_n \in \mathbb{R}$ fixed coefficients such that $X_i \prec_{(\kappa,\lambda)} Y_i$. Then for any nondecreasing $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$,*

$$\mathbb{E}\varphi\left(\left|\sum_{i=1}^n a_i X_i\right|\right) \le 2\lceil\kappa\rceil \mathbb{E}\varphi\left(\lceil\kappa\rceil\lambda\left|\sum_{i=1}^n a_i Y_i\right|\right).$$

Statements similar to the following lemma have been used in multiple works to establish a relation between $|\|Ax\|_2 - a|$ and $\left|\|Ax\|_2^2 - a^2\right|$, for example in the proof of Lemma 5.36 in [Ver12]. For completeness, we state it as a separate result with its proof here.

**Lemma 2.21.** *For real numbers $a, b \ge 0$, $b \ne 0$, it holds that*

$$\frac{1}{3}\min\left\{\frac{|a^2-b^2|}{b}, \sqrt{|a^2-b^2|}\right\} \le |a-b| \le \min\left\{\frac{|a^2-b^2|}{b}, \sqrt{|a^2-b^2|}\right\}.$$

*Proof.* We obtain

$$|a-b| = \frac{|a^2-b^2|}{|a+b|} \le \frac{|a^2-b^2|}{b},$$

and since $a, b \ge 0$, i.e., $|a-b| \le |a| + |b| = |a+b|$, it follows that $|a-b|^2 \le |a-b||a+b| = |a^2-b^2|$, proving the second inequality.

For the first inequality, first assume the case $a \le 2b$. Then $a + b \le 3b$ such that

$$\frac{1}{3}\frac{|a^2-b^2|}{b} \le \frac{|a^2-b^2|}{a+b} = |a-b|.$$

In the case that $a \ge 2b$, i.e., $a - b \ge b \ge 0$, we obtain

$$\begin{aligned}\frac{1}{3}\sqrt{|a^2-b^2|} &\le \frac{1}{3}\sqrt{|a+b||a-b|} \le \frac{1}{3}\sqrt{(|a-b|+2b)|a-b|} \\ &\le \frac{1}{3}\sqrt{(|a-b|+2|a-b|)|a-b|} = \frac{1}{\sqrt{3}}|a-b| \le |a-b|.\end{aligned}$$

□

Relations between moments and tail bounds have also been well-known in the field. For an overview see, e.g., Chapter 7.3 in [FR13]. In this spirit, we state and prove the following small tool for the case of mixed tails which we encounter in this work.

**Lemma 2.22** (Moments and tail bounds). *Let $T$ be a finite set and $X$ an $\mathbb{R}$ valued random variable such that for all $p \ge p_0 \ge 0$,*

$$\|X\|_{L_p} \le \sum_{k=1}^d \min_{l \in T} p^{e_{k,l}} \gamma_{k,l}$$

*for values $\gamma_{k,l} > 0$, $e_{k,l} > 0$.*

*Then for all $t > 0$,*

$$\mathbb{P}(|X| > t) \le e^{p_0} \exp\left(-\min_{k\in[d]} \max_{l \in T}\left(\frac{t}{ed\gamma_{k,l}}\right)^{\frac{1}{e_{k,l}}}\right).$$

*Proof.* Fix any $u > 0$. For any $k \in [d]$, define $l'(k) := \operatorname{argmax}_{l\in T}\left(\frac{u}{\gamma_{k,l}}\right)^{\frac{1}{e_{k,l}}}$, then choose $k' := \operatorname{argmin}_{k\in[d]}\left(\frac{u}{\gamma_{k,l'(k)}}\right)^{\frac{1}{e_{k,l'(k)}}}$, and $p := \left(\frac{u}{\gamma_{k',l'(k')}}\right)^{\frac{1}{e_{k',l'(k')}}}$, such that $p = \min_{k\in[d]}\max_{l\in T}\left(\frac{u}{\gamma_{k,l}}\right)^{\frac{1}{e_{k,l}}}$.

If $p < p_0$, then $\mathbb{P}(|X| > edu) \leq 1 = e^{p_0} \exp(-p_0) \leq e^{p_0} \exp(-p)$.

If $p \geq p_0$, then by the choice of $p$,

$$\begin{aligned}\|X\|_{L_p} &\leq \sum_{k=1}^{d} \min_{l\in T} p^{e_{k,l}} \gamma_{k,l} \leq \sum_{k=1}^{d} \min_{l\in T} \left[ \left( \frac{u}{\gamma_{k',l'(k')}} \right)^{\frac{1}{e_{k',l'(k')}}} \right]^{e_{k,l}} \gamma_{k,l} \\ &\leq \sum_{k=1}^{d} \left[ \left( \frac{u}{\gamma_{k',l'(k')}} \right)^{\frac{1}{e_{k',l'(k')}}} \right]^{e_{k,l'(k)}} \gamma_{k,l'(k)} \\ &\leq \sum_{k=1}^{d} \left[ \left( \frac{u}{\gamma_{k,l'(k)}} \right)^{\frac{1}{e_{k,l'(k)}}} \right]^{e_{k,l'(k)}} \gamma_{k,l'(k)} \leq \sum_{k=1}^{d} u = du.\end{aligned}$$

So by Markov's inequality,

$$\mathbb{P}(|X| > edu) \leq \mathbb{P}(|X|^p > (edu)^p) \leq \frac{\mathbb{E}|X|^p}{(edu)^p} = \left( \frac{\|X\|_{L_p}}{edu} \right)^p \leq e^{-p}.$$

In all cases, we obtain

$$\mathbb{P}(|X| > edu) \leq e^{p_0} e^{-p} = e^{p_0} \exp\left( -\min_{k\in[d]} \max_{l\in T} \left( \frac{u}{\gamma_{k,l}} \right)^{\frac{1}{e_{k,l}}} \right).$$

The result follows by taking $u := \frac{t}{ed}$. □

### 2.3.2 Proof of the Upper Bound

**Required tools:**

**Lemma 2.23.** *There is an absolute constant $C$ such that the following holds. Let $X \in \mathbb{R}^n$ be random with mean 0 and $\|X\|_{\psi_2} \leq L$. Take a Gaussian vector $g \sim N(0, Id_n)$ and $a \in \mathbb{R}^n$. Then for all $p \geq 1$,*

$$\mathbb{E}\left|\sum_{k=1}^{n} a_k X_k\right|^p \leq (CL)^p \mathbb{E}\left|\sum_{k=1}^{n} a_k g_k\right|^p.$$

*Proof.* By the assumption on $X$, $\sum_{k=1}^{n} a_k X_k = \langle a, X\rangle$ is mean 0 with $\|\langle a, X\rangle\|_{\psi_2} \leq L\|a\|_2$, implying that for any $p \geq 1$,

$$\mathbb{E}|\langle a, X\rangle|^p \leq (C_1 L\|a\|_2)^p p^{\frac{p}{2}}.$$

On the other hand, $\langle a, g\rangle \sim N(0, \|a\|_2^2)$, so by the known absolute moments of the normal distribution and Stirling's approximation,

$$\begin{aligned}\mathbb{E}|\langle a, g\rangle|^p =&\|a\|_2^p \cdot \frac{2^{\frac{p}{2}}}{\sqrt{\pi}} \Gamma\left(\frac{p+1}{2}\right) \geq \|a\|_2^p \frac{2^{\frac{p}{2}}}{\sqrt{\pi}} \sqrt{2\pi} \left(\frac{p+1}{2}\right)^{\frac{p}{2}} \exp(-\frac{p+1}{2}) \\ \geq& 2^{\frac{p}{2}} \|a\|_2^p \sqrt{\frac{2}{e}} \left(\frac{p}{2e}\right)^{\frac{p}{2}} \geq \sqrt{\frac{2}{e}} \left(\frac{1}{e}\right)^{\frac{p}{2}} \|a\|_2^p p^{\frac{p}{2}} \geq \left(\frac{2}{e^2}\right)^{\frac{p}{2}} \|a\|_2^p p^{\frac{p}{2}},\end{aligned}$$

implying that $\mathbb{E}|\langle a, X\rangle|^p \leq \left(\frac{C_1 e}{\sqrt{2}} L\right)^p \mathbb{E}|\langle a, g\rangle|^p$. □

In order to control arbitrary chaoses, we will derive a similar result as Lemma 2.23 for squared subgaussian and Gaussian variables. To achieve this, we make use of strong domination. The following theorem states that this can be used to compare squared subgaussian and Gaussian variables.

**Lemma 2.24.** *There exist absolute constants $\kappa, \lambda > 0$ such that the following holds. Let $X$ be a random variable with $\mathbb{E}X^2 = 1$ and $\|X\|_{\psi_2} \leq L$, $L \geq 1$ and $g \sim N(0,1)$. Let $\xi$, $\xi' \in \{\pm 1\}$ be Rademacher variables that are independent of $X$ and $g$. Then $\xi(X^2-1) \prec_{(\kappa, \lambda L^2)} \xi'(g^2-1)$ in the sense of Definition 2.19.*

*Proof.* For any $t > 0$,

$$\mathbb{P}\left(|\xi(X^2-1)| > t\right) = \mathbb{P}\left(X^2 - 1 > t\right) + \mathbb{P}\left(-(X^2-1) > t\right)$$

For a constant $c \geq 1$, the first term can be bounded by

$$\mathbb{P}\left(X^2 - 1 > t\right) = \mathbb{P}\left(|X| > \sqrt{1+t}\right) \leq \exp\left(1 - \frac{1+t}{c^2L^2}\right) \leq e \cdot e^{-\frac{t}{c^2L^2}}.$$

The second term is 0 if $t \geq 1$ since $-(X^2-1) \leq 1$. For $t \leq 1$, $e^{-\frac{t}{c^2L^2}} \geq e^{-\frac{1}{c^2L^2}} \geq e^{-1}$. Then it holds that $\mathbb{P}(-(X^2-1) > t) \leq 1 \leq e \cdot e^{-\frac{t}{c^2L^2}}$, and altogether we obtain

$$\mathbb{P}\left(|\xi(X^2-1)| > t\right) \leq 2e \cdot e^{-\frac{t}{c^2L^2}}.$$

On the other hand, for any $\lambda > 0$,

$$\begin{aligned}\mathbb{P}\left(\lambda L^2 |\xi'(g^2-1)| > t\right) &\geq \mathbb{P}\left(g^2 - 1 > \frac{t}{\lambda L^2}\right) = \mathbb{P}\left(|g| > \sqrt{1 + \frac{t}{\lambda L^2}}\right) \\ &= \mathbb{P}\left(|g| \geq \sqrt{1 + \frac{t}{\lambda L^2}}\right).\end{aligned}$$

To bound this, we use the following properties of the normal distribution: (see Proposition 7.5 in [FR13])

$$\mathbb{P}(|g| \geq u) \geq \sqrt{\frac{2}{\pi}} \frac{1}{u}\left(1 - \frac{1}{u^2}\right) e^{-\frac{u^2}{2}}, \qquad \mathbb{P}(|g| \geq u) \geq \left(1 - \sqrt{\frac{2}{\pi}} u\right) e^{-\frac{u^2}{2}}. \tag{2.12}$$

For $0 < u \leq \frac{1}{4}$, the second inequality in (2.12) yields

$$\mathbb{P}\left(|g| \geq \sqrt{1+u}\right) \geq \frac{1}{10} e^{-\frac{1+u}{2}} \geq \frac{1}{10} e^{-\frac{1}{2}} \cdot e^{-u} \geq \frac{1}{17} e^{-u}.$$

For $u \geq \frac{1}{4}$, the first inequality in (2.12) gives $\mathbb{P}\left(|g| \geq \sqrt{1+u}\right) \geq \frac{1}{5}\sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{1+u}} e^{-\frac{1+u}{2}}$. Using that $\frac{1}{\sqrt{1+u}} \geq e^{-\frac{1}{2}u}$ for all $u > 0$, we obtain for $u \geq \frac{1}{4}$,

$$\mathbb{P}\left(|g| \geq \sqrt{1+u}\right) \geq \frac{1}{5}\sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}u} \exp\left(-\frac{1+u}{2}\right) = \frac{1}{5}\sqrt{\frac{2}{\pi}} \exp\left(-\frac{1}{2} - u\right) \geq \frac{1}{11} e^{-u}.$$

So for any $u > 0$, $\mathbb{P}(|g| > \sqrt{1+u}) \geq \frac{1}{17} e^{-u}$. By choosing $\lambda = c^2$ and combining,

$$\mathbb{P}\left(|\xi(X^2-1)| > t\right) \leq 2e \cdot e^{-\frac{t}{\lambda L^2}} \leq 93 \cdot \frac{1}{17} e^{-\frac{t}{\lambda L^2}} \leq 93 \mathbb{P}\left(\lambda L^2 |\xi'(g^2-1)| > t\right).$$

□

**Theorem 2.25.** *There is an absolute constant $C > 0$ such that the following holds. Let $X \in \mathbb{R}^n$ have independent entries that have mean 0 and variance 1 and are subgaussian with $\psi_2$ norm $\leq L$ for an $L \geq 1$. Take a Gaussian vector $g \sim N(0, Id_n)$ and $a \in \mathbb{R}^n$. Then*

$$\mathbb{E}\left|\sum_{k=1}^n a_k (X_k^2 - 1)\right|^p \leq (CL^2)^p \mathbb{E}\left|\sum_{k=1}^n a_k (g_k^2 - 1)\right|^p.$$

*Proof.* Consider independent Rademacher variables $\xi_1, \dots, \xi_n, \bar{\xi}_1, \dots, \bar{\xi}_n \in \{\pm 1\}^n$ that are also independent of $X$ and $g$. By the symmetrization Lemma 2.15, it holds that

$$\mathbb{E}\left|\sum_{k=1}^n a_k(X_k^2-1)\right|^p \le 2^p\mathbb{E}\left|\sum_{k=1}^n a_k\xi_k(X_k^2-1)\right|^p$$
$$\mathbb{E}\left|\sum_{k=1}^n a_k\bar{\xi}_k(g_k^2-1)\right|^p \le 2^p\mathbb{E}\left|\sum_{k=1}^n a_k(g_k^2-1)\right|^p. \tag{2.13}$$

Using that $\xi_k(X^2-1) \prec_{(\kappa,\lambda L^2)} \bar{\xi}_k(g^2-1)$ by Lemma 2.24 and that $|\cdot|^p$ is a convex nondecreasing function $\mathbb{R}^+ \to \mathbb{R}^+$, Theorem 2.20 implies that there is a constant $\tilde{C} > 0$ such that

$$\mathbb{E}\left|\sum_{k=1}^n a_k\xi_k(X_k^2-1)\right|^p \le (\tilde{C}L^2)^p\mathbb{E}\left|\sum_{k=1}^n a_k\bar{\xi}_k(g_k^2-1)\right|^p.$$

□

The next theorem is an important tool for the proof of our decoupling result (Theorem 2.11). Its purpose is to rearrange a chaos in such a way that – under some changes – the quadratic factor that occur (here $(X_{\mathbf{i}_l}^{(l)})^2$) are replaced by corresponding mean 0 factors of the type $\left[(X_{\mathbf{i}_l}^{(l)})^2-1\right]$ that also occur in Theorem 2.11.

Rearranging the terms with this theorem enables an iterative application of the standard decoupling Theorem 2.16 in the proof of Theorem 2.11. Furthermore, the factors $\left[(X_{\mathbf{i}_l}^{(l)})^2-1\right]$ are 0 in the Rademacher case (Remark 2.12). In the general case, after the comparison with Gaussians, they will be turned into a product of two independent factors with the subsequent Lemma 2.27 in the proof of Theorem 2.9.

Note that this is a purely arithmetic rearrangement of the chaos. We do not require or take any randomness of the $X^{(l)}$ into account.

**Theorem 2.26.** *Let* $\mathbf{n} \in \mathbb{N}^d$, $\mathbf{A} \in \mathbb{R}^{\mathbf{n}}$, $X^{(1)} \in \mathbb{R}^{n_1}, \dots, X^{(d)} \in \mathbb{R}^{n_d}$, $I \subset [d]$. *Then*

$$\sum_{\mathbf{i}\in\mathbf{J^n}} A_{\mathbf{i}} \prod_{l\in[d]} (X_{\mathbf{i}_l}^{(l)})^2 = \sum_{I\subset[d]} \sum_{\mathbf{i}\in\mathbf{J^n}([d]\backslash I)} A_{\mathbf{i}}^{(I)} \prod_{l\in[d]\backslash I} \left[(X_{\mathbf{i}_l}^{(l)})^2-1\right]$$

*where for any* $\mathbf{i} \in \mathbf{J^n}([d]\backslash I)$,

$$A_{\mathbf{i}}^{(I)} = \sum_{\mathbf{j}\in\mathbf{J^n}(I)} A_{\mathbf{i}\dot{\times}\mathbf{j}}.$$

*Proof.* Observing that for any $I \subset [d]$, $\mathbf{i} \in \mathbf{J^n}(I)$,

$$\prod_{l\in[d]\backslash I} \left[(X_{\mathbf{i}_l}^{(l)})^2-1\right] = \sum_{I'\subset[d]\backslash I} (-1)^{|[d]\backslash(I\cup I')|} \prod_{l\in I'} (X_{\mathbf{i}_l}^{(l)})^2,$$

we obtain

$$\begin{aligned}
&\sum_{\substack{I\subset[d]\\ \mathbf{i}\in\mathbf{J^n}([d]\backslash I)}} A_{\mathbf{i}}^{(I)} \prod_{l\in[d]\backslash I} \left[(X_{\mathbf{i}_l}^{(l)})^2-1\right]\\
&= \sum_{\substack{I\subset[d]\\ \mathbf{i}\in\mathbf{J^n}([d]\backslash I)\\ \mathbf{j}\in\mathbf{J^n}(I)}} A_{\mathbf{i}\dot{\times}\mathbf{j}} \sum_{I'\subset[d]\backslash I} (-1)^{|[d]\backslash(I\cup I')|} \prod_{l\in I'} (X_{\mathbf{i}_l}^{(l)})^2
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\substack{I\subset[d]\\ I'\subset[d]\setminus I}} (-1)^{|[d]\setminus(I\cup I')|} \sum_{\substack{\mathbf{i}\in\mathbf{J^n}([d]\setminus I)\\ \mathbf{j}\in\mathbf{J^n}(I)}} A_{\mathbf{i}\dot{\times}\mathbf{j}} \prod_{l\in I'} (X^{(l)}_{\mathbf{i}_l})^2 \\
&= \sum_{\substack{I'\subset[d]\\ I\subset[d]\setminus I'}} (-1)^{|[d]\setminus(I\cup I')|} \sum_{\mathbf{i}\in\mathbf{J^n}} A_{\mathbf{i}} \prod_{l\in I'} (X^{(l)}_{\mathbf{i}_l})^2 \\
&= \sum_{I'\subset[d]} \left[ \left( \sum_{I\subset[d]\setminus I'} (-1)^{|([d]\setminus I')\setminus I|} \right) \cdot \left( \sum_{\mathbf{i}\in\mathbf{J^n}} A_{\mathbf{i}} \prod_{l\in I'} (X^{(l)}_{\mathbf{i}_l})^2 \right) \right].
\end{aligned}
$$

This implies the claim using Lemma 2.17. □

A key to the proof of the upper moment bound in our main result (Theorem 2.9) is the decoupling technique of Theorem 2.11. With the above auxiliary results, we can give the proof of it here.

*Proof of Theorem 2.11.*

$$
\begin{aligned}
b &:= \sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J^n}} A_{\mathbf{i}\dot{+}\mathbf{i}'} \prod_{l\in[d]} X^{(l)}_{\mathbf{i}_l} X^{(l)}_{\mathbf{i}'_l} \\
&= \sum_{I\subset[d]} \sum_{\substack{\mathbf{i}\in\mathbf{J^n}(I)\\ \mathbf{j},\mathbf{j}'\in\mathbf{J^n}(I^c)\\ \forall l\in I^c:\mathbf{j}_l\neq\mathbf{j}'_l}} A_{(\mathbf{i}\dot{\times}\mathbf{j})\dot{+}(\mathbf{i}\dot{\times}\mathbf{j}')} \prod_{l\in I} (X^{(l)}_{\mathbf{i}_l})^2 \prod_{l\in I^c} X^{(l)}_{\mathbf{j}_l} X^{(l)}_{\mathbf{j}'_l}
\end{aligned}
$$

since each summand $\mathbf{i},\mathbf{i}'$ is precisely considered in the sum for $I = \{l\in[d] : \mathbf{i}_l = \mathbf{i}'_l\}$ and no other $I$.

Now applying Theorem 2.26 yields

$$
\begin{aligned}
b &= \sum_{I\subset[d]} \sum_{\mathbf{i}\in\mathbf{J^n}(I)} \left( \sum_{\substack{\mathbf{j},\mathbf{j}'\in\mathbf{J^n}(I^c)\\ \forall l\in I^c:\mathbf{j}_l\neq\mathbf{j}'_l}} A_{(\mathbf{i}\dot{\times}\mathbf{j})\dot{+}(\mathbf{i}\dot{\times}\mathbf{j}')} \prod_{l\in I^c} X^{(l)}_{\mathbf{j}_l} X^{(l)}_{\mathbf{j}'_l} \right) \prod_{l\in I} (X^{(l)}_{\mathbf{i}_l})^2 \\
&= \sum_{\substack{I,J\subset[d]:\\ J\subset I}} \sum_{\substack{\mathbf{i}\in\mathbf{J^n}(J)\\ \mathbf{k}\in\mathbf{J^n}(I\setminus J)}} \left( \sum_{\substack{\mathbf{j},\mathbf{j}'\in\mathbf{J^n}(I^c)\\ \forall l\in I^c:\mathbf{j}_l\neq\mathbf{j}'_l}} A_{\substack{(\mathbf{i}\dot{\times}\mathbf{j}\dot{\times}\mathbf{k})\\ \dot{+}(\mathbf{i}\dot{\times}\mathbf{j}'\dot{\times}\mathbf{k})}} \prod_{l\in I^c} X^{(l)}_{\mathbf{j}_l} X^{(l)}_{\mathbf{j}'_l} \right) \prod_{l\in J} \left[ (X^{(l)}_{\mathbf{i}_l})^2 - 1 \right] \\
&= \sum_{\substack{I,J\subset[d]:\\ J\subset I}} \sum_{\substack{\mathbf{i}\in\mathbf{J^n}(J)\\ \mathbf{k}\in\mathbf{J^n}(I\setminus J)\\ \mathbf{j},\mathbf{j}'\in\mathbf{J^n}(I^c)\\ \forall l\in I^c:\mathbf{j}_l\neq\mathbf{j}'_l}} A_{\substack{(\mathbf{i}\dot{\times}\mathbf{j}\dot{\times}\mathbf{k})\\ \dot{+}(\mathbf{i}\dot{\times}\mathbf{j}'\dot{\times}\mathbf{k})}} \prod_{l\in I^c} X^{(l)}_{\mathbf{j}_l} X^{(l)}_{\mathbf{j}'_l} \prod_{l\in J} \left[ (X^{(l)}_{\mathbf{i}_l})^2 - 1 \right] \\
&=: \sum_{\substack{I,J\subset[d]:\\ J\subset I}} S_{I,J}.
\end{aligned}
$$

Because of

$$
S_{[d],\emptyset} = \sum_{\mathbf{k}\in\mathbf{J^n}} A_{\mathbf{k}\dot{+}\mathbf{k}} = \mathbb{E} \sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J^n}} A_{\mathbf{i}\dot{+}\mathbf{i}'} \prod_{l\in[d]} X^{(l)}_{\mathbf{i}_l} X^{(l)}_{\mathbf{i}'_l}
$$

and the triangle inequality, we obtain

$$
\|b - \mathbb{E}b\|_{L_p} \leq \sum_{\substack{I,J\subset[d]:\\ J\subset I, I\setminus J\neq\emptyset}} \|S_{I,J}\|_{L_p}. \tag{2.14}
$$

For any fixed $l_0 \in I^c$, we obtain that $\|S_{I,J}\|_{L_p} =$

$$\left\| \sum_{\substack{\bar{\mathbf{j}},\bar{\mathbf{j}}' \in \mathbf{J^n}(\{l_0\}) \\ \bar{\mathbf{j}}_{l_0} \neq \bar{\mathbf{j}}'_{l_0}}} \left( \sum_{\substack{\mathbf{i} \in \mathbf{J^n}(J) \\ \mathbf{k} \in \mathbf{J^n}(I \backslash J) \\ \mathbf{j},\mathbf{j}' \in \mathbf{J^n}(I^c \backslash \{l_0\}) \\ \forall l \in I^c : \mathbf{j}_l \neq \mathbf{j}'_l}} A_{\substack{(\mathbf{i} \dot{\times} \mathbf{j} \dot{\times} \bar{\mathbf{j}} \dot{\times} \mathbf{k}) \\ \dot{+} (\mathbf{i} \dot{\times} \mathbf{j}' \dot{\times} \bar{\mathbf{j}}' \dot{\times} \mathbf{k})}} \prod_{l \in I^c} X_{\mathbf{j}_l}^{(l)} X_{\mathbf{j}'_l}^{(l)} \prod_{l \in J} \left[ (X_{\mathbf{i}_l}^{(l)})^2 - 1 \right] \right) X_{\bar{\mathbf{j}}_{l_0}}^{(l_0)} X_{\bar{\mathbf{j}}'_{l_0}}^{(l_0)} \right\|_{L_p} .$$

We can apply the decoupling Theorem 2.16 to this for the convex function $|\cdot|^p$ and the expectation conditioned on all variables except $X^{(l_0)}$. This leads to $\|S_{I,J}\|_{L_p} \leq$

$$4 \left\| \sum_{\bar{\mathbf{j}},\bar{\mathbf{j}}' \in \mathbf{J^n}(\{l_0\})} \left( \sum_{\substack{\mathbf{i} \in \mathbf{J^n}(J) \\ \mathbf{k} \in \mathbf{J^n}(I \backslash J) \\ \mathbf{j},\mathbf{j}' \in \mathbf{J^n}(I^c \backslash \{l_0\}) \\ \forall l \in I^c : \mathbf{j}_l \neq \mathbf{j}'_l}} A_{\substack{(\mathbf{i} \dot{\times} \mathbf{j} \dot{\times} \bar{\mathbf{j}} \dot{\times} \mathbf{k}) \\ \dot{+} (\mathbf{i} \dot{\times} \mathbf{j}' \dot{\times} \bar{\mathbf{j}}' \dot{\times} \mathbf{k})}} \prod_{l \in I^c} X_{\mathbf{j}_l}^{(l)} X_{\mathbf{j}'_l}^{(l)} \prod_{l \in J} \left[ (X_{\mathbf{i}_l}^{(l)})^2 - 1 \right] \right) X_{\bar{\mathbf{j}}_{l_0}}^{(l_0)} \bar{X}_{\bar{\mathbf{j}}'_{l_0}}^{(l_0)} \right\|_{L_p} .$$

Repeating this procedure iteratively for all other $l \in I^c$, we obtain

$$\|S_{I,J}\|_{L_p} \leq 4^{d-|I|} \left\| \sum_{\substack{\mathbf{i} \in \mathbf{J^n}(J) \\ \mathbf{k} \in \mathbf{J^n}(I \backslash J) \\ \mathbf{j},\mathbf{j}' \in \mathbf{J^n}(I^c)}} A_{(\mathbf{i} \dot{\times} \mathbf{j} \dot{\times} \mathbf{k}) \dot{+} (\mathbf{i} \dot{\times} \mathbf{j}' \dot{\times} \mathbf{k})} \prod_{l \in I^c} X_{\mathbf{j}_l}^{(l)} \bar{X}_{\mathbf{j}'_l}^{(l)} \prod_{l \in J} \left[ (X_{\mathbf{i}_l}^{(l)})^2 - 1 \right] \right\|_{L_p} .$$

Substituting this into (2.14) completes the proof. □

The works in [Kwa87] and [AG93] have investigated polynomials with higher powers of Gaussian variables. Since in our scenario, we only have two occurrences of every vector, thus we can repeatedly apply their result for the case of two coinciding indices. Considering that $H_2(x) = x^2 - 1$ is the Hermite polynomial of degree 2 and leading coefficient 1, equation (2.9) in [AG93] in our setup can be written as follows. Note that as suggested there, the case $p \geq 1$ can also be shown using Jensen's inequality which can be used to show this inequality with coefficient 2.

**Lemma 2.27.** *Let $a \in \mathbb{R}^n$, $g, \bar{g} \sim N(0, Id_n)$, $p \geq 1$. Then*

$$\left\| \sum_{k=1}^n a_k (g_k^2 - 1) \right\|_{L_p} \leq 2 \left\| \sum_{k=1}^n a_k g_k \bar{g}_k \right\|_{L_p} .$$

Combining the previous lemmas, now we can prove the upper bound in the main Theorem 2.9.

**Proof of Theorem 2.9, upper bound:**

**Step 1: Decoupling**

Let $\alpha := \|X^T A X - \mathbb{E} X^T A X\|_{L_p}$. By Theorem 2.11, $\alpha \leq$

$$\sum_{\substack{J \subset I \subset [d] \\ I \backslash J \neq [d]}} 4^{d-|I|} \left\| \sum_{\substack{\mathbf{i} \in \mathbf{J^n}(J) \\ \mathbf{k} \in \mathbf{J^n}(I \backslash J) \\ \mathbf{j},\mathbf{j}' \in \mathbf{J^n}(I^c)}} A_{(\mathbf{i} \dot{\times} \mathbf{j} \dot{\times} \mathbf{k}) \dot{+} (\mathbf{i} \dot{\times} \mathbf{j}' \dot{\times} \mathbf{k})} \prod_{l \in I^c} X_{\mathbf{j}_l}^{(l)} \bar{X}_{\mathbf{j}'_l}^{(l)} \prod_{l \in J} \left[ (X_{\mathbf{i}_l}^{(l)})^2 - 1 \right] \right\|_{L_p} . \qquad (2.15)$$

**Step 2: Replacing the subgaussian factors by Gaussians**
In (2.15), we can repeatedly apply Lemma 2.23 to replace all the linear subgaussian factors by Gaussian ones. Afterwards, Theorem 2.25 allows the same for the quadratic terms. Together, this yields that $\alpha \leq$

$$\sum_{\substack{J\subset I\subset[d]\\ I\backslash\neq[d]}} (CL)^{|I^c|+|J|} \left\| \sum_{\substack{\mathbf{i}\in\mathbf{J^n}(J)\\ \mathbf{k}\in\mathbf{J^n}(I\backslash J)\\ \mathbf{j},\mathbf{j}'\in\mathbf{J^n}(I^c)}} A_{\substack{(\mathbf{i}\dot\times\mathbf{j}\dot\times\mathbf{k})\\ \dot+(\mathbf{i}\dot\times\mathbf{j}'\dot\times\mathbf{k})}} \prod_{l\in I^c} g_{\mathbf{j}_l}^{(l)} \bar{g}_{\mathbf{j}'_l}^{(l)} \prod_{l\in J} \left[(g_{\mathbf{i}_l}^{(l)})^2 - 1\right] \right\|_{L_p} . \tag{2.16}$$

**Step 3: Decoupling of squared Gaussians** In an analogous fashion as in step 2, we can successively replace all the factors $\left[(g_{\mathbf{i}_l}^{(l)})^2 - 1\right]$ in (2.16) by $g_{\mathbf{i}_l}^{(l)} \bar{g}_{\mathbf{i}_l}^{(l)}$ using Lemma 2.27. This leads to

$$\begin{aligned}
\alpha &\leq \sum_{\substack{J\subset I\subset[d]\\ I\backslash J\neq[d]}} (CL)^{|I^c|+|J|} \left\| \sum_{\substack{\mathbf{i}\in\mathbf{J^n}(J)\\ \mathbf{k}\in\mathbf{J^n}(I\backslash J)\\ \mathbf{j},\mathbf{j}'\in\mathbf{J^n}(I^c)}} A_{\substack{(\mathbf{i}\dot\times\mathbf{j}\dot\times\mathbf{k})\\ \dot+(\mathbf{i}\dot\times\mathbf{j}'\dot\times\mathbf{k})}} \prod_{l\in I^c} g_{\mathbf{j}_l}^{(l)} \bar{g}_{\mathbf{j}'_l}^{(l)} \prod_{l\in J} g_{\mathbf{i}_l}^{(l)} \bar{g}_{\mathbf{i}_l}^{(l)} \right\|_{L_p} \\
&= \sum_{\substack{J\subset I\subset[d]\\ I\backslash J\neq[d]}} (CL)^{|I^c|+|J|} \left\| \sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J^n}(I^c\cup J)} A^{(I,J)}_{\mathbf{i}\dot+\mathbf{i}'} \prod_{l\in I^c\cup J} g_{\mathbf{j}_l}^{(l)} \bar{g}_{\mathbf{j}'_l}^{(l)} \right\|_{L_p} .
\end{aligned}$$

where for all $\mathbf{i},\mathbf{i}' \in \mathbf{J^n}(J\cup I^c)$,

$$A^{(I,J)}_{\mathbf{i}\dot+\mathbf{i}'} = \begin{cases} \sum_{\mathbf{k}\in\mathbf{J^n}(I\backslash J)} A_{(\mathbf{i}\dot\times\mathbf{k})\dot+(\mathbf{i}'\dot\times\mathbf{k})} & \text{if } \forall l \in J : \mathbf{i}_l = \mathbf{i}'_l \\ 0 & \text{otherwise.} \end{cases} \tag{2.17}$$

**Step 4: Completing the proof** Then Theorem 2.7 yields that

$$\left\| \sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J^n}(J\cup I^c)} A^{(I,J)}_{\mathbf{i}\dot+\mathbf{i}'} \prod_{l\in I^c\cup J} g_{\mathbf{i}_l}^{(l)} \bar{g}_{\mathbf{i}'_l}^{(l)} \right\|_{L_p} \leq \tilde{m}_p^{(I,J)}$$

where for $S((J\cup I^c)\cup((J\cup I^c)+d),\kappa)$ being the set of all partitions of $(J\cup I^c)\cup((J\cup I^c)+d)$ into $\kappa$ sets,

$$\tilde{m}_p^{(I,J)} := \sum_{\kappa=1}^{d} p^{\kappa/2} \sum_{(I_1,\ldots,I_\kappa)\in S((J\cup I^c)\cup((J\cup I^c)+d),\kappa)} \|\mathbf{A}^{(I,J)}\|_{I_1,\ldots,I_\kappa}.$$

By Lemma 2.18, $\|\mathbf{A}^{(I,J)}\|_{I_1,\ldots,I_\kappa} \leq \|\mathbf{A}^{(I)}\|_{I_1,\ldots,I_\kappa}$ where $\mathbf{A}^{(I)} = \mathbf{A}^{(I,\emptyset)}$ as given in the statement of Theorem 2.9. Together with this, the upper bound in Theorem 2.9 follows.

### 2.3.3 Proof of the Lower Bound

**Required tools:**
In this section, we will prove the lower bound in Theorem 2.9. Unlike the upper bound, we will only prove this for the case of Gaussian vectors. Indeed, for arbitrary subgaussian distributions, the lower bound fails to hold as the following simple example for the case $d = 1$

shows: Consider the identity matrix $Id_n$ and a Rademacher vector $\xi \in \{\pm 1\}^n$. Then the object of interest in Theorem 2.9 is $\xi^T Id_n \xi - \mathbb{E}[\xi^T Id_n \xi] = 0$ even though the moment bounds $m_p$ would be $> 0$.

We follow the approach of reversing all steps in the proof of the upper bound, without the Gaussian comparison steps. This is why also the two decoupling steps before and after the Gaussian comparison can be performed together.

As mentioned before, Gaussian decoupling, with upper as well as lower bounds, has been studied in [AG93] where central ideas of [Kwa87] have been used. [AG93] provides a decoupling inequality for Gaussian chaos with an arbitrary number of coinciding indices. Similarly to Lemma 2.27, we can adapt the result of Equation (2.9) in [AG93] to our situation as follows.

**Lemma 2.28.** *Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix, $g, \bar{g} \sim N(0, Id_n)$ be independent, and $p \geq 1$.*

$$\left\| \sum_{j,k \in [n]} A_{j,k} g_j \bar{g}_k \right\|_{L_p} \leq \left\| \sum_{j,k \in [n]} A_{j,k} (g_j g_k - \mathbb{1}_{j=k}) \right\|_{L_p} .$$

To generalize this to cases of multiple axes, we iteratively apply Lemma 2.28 to obtain the following corollary.

**Corollary 2.29.** *Let $\mathbf{n} \in \mathbb{N}^d$, $\mathbf{A} \in \mathbb{R}^{\mathbf{n}^{\times 2}}$ such that $\mathbf{A}$ satisfies the symmetry condition that for all $l \in [d]$ and any $\mathbf{i}, \mathbf{i}' \in \mathbf{J}^{\mathbf{n}}([d] \backslash \{l\})$, $\mathbf{j}, \mathbf{j}' \in \mathbf{J}^{\mathbf{n}}(\{l\})$,*

$$A_{(\mathbf{i} \dot{\times} \mathbf{j}) \dot{+} (\mathbf{i}' \dot{\times} \mathbf{j}')} = A_{(\mathbf{i} \dot{\times} \mathbf{j}') \dot{+} (\mathbf{i}' \dot{\times} \mathbf{j})} \tag{2.18}$$

*Let $g^{(1)}, \bar{g}^{(1)} \sim N(0, Id_{\mathbf{n}_1}), \ldots, g^{(d)}, \bar{g}^{(d)} \sim N(0, Id_{\mathbf{n}_d})$ be independent. Then for any set $I \subset [d]$, $p \geq 1$,*

$$\begin{aligned} &\left\| \sum_{\substack{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}(I) \\ \mathbf{j} \in \mathbf{J}^{\mathbf{n}}(I^c)}} A_{(\mathbf{i} \dot{\times} \mathbf{j}) \dot{+} (\mathbf{i}' \dot{\times} \mathbf{j})} \prod_{l \in I} g^{(l)}_{\mathbf{i}_l} \bar{g}^{(l)}_{\mathbf{i}'_l} \right\|_{L_p} \\ \leq &\left\| \sum_{\substack{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}(I) \\ \mathbf{j} \in \mathbf{J}^{\mathbf{n}}(I^c)}} A_{(\mathbf{i} \dot{\times} \mathbf{j}) \dot{+} (\mathbf{i}' \dot{\times} \mathbf{j})} \prod_{l \in I} \left[ g^{(l)}_{\mathbf{i}_l} g^{(l)}_{\mathbf{i}'_l} - \mathbb{1}_{\mathbf{i}_l = \mathbf{i}'_l} \right] \right\|_{L_p} \end{aligned}$$

Independently of the Gaussian decoupling approach, the following two lemmas provide a tool to reverse the application of the rearrangement result Theorem 2.26 in the proof of the upper bound.

**Lemma 2.30.** *Let $\mathbf{A} \in \mathbb{R}^{\mathbf{n}^{\times 2}}$ be an array of order $2d$ and $X^{(1)} \in \mathbb{R}^{n_1}, \ldots X^{(d)} \in \mathbb{R}^{n_d}$ vectors. Then*

$$\begin{aligned} &\sum_{I \subset [d]} \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}(I)} \sum_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(I^c)} A_{(\mathbf{i} \dot{\times} \mathbf{j}) \dot{+} (\mathbf{i}' \dot{\times} \mathbf{j})} \prod_{l \in I} \left[ X^{(l)}_{\mathbf{i}_l} X^{(l)}_{\mathbf{i}'_l} - \mathbb{1}_{\mathbf{i}_l = \mathbf{i}'_l} \right] \\ &= \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} A_{\mathbf{i} \dot{+} \mathbf{i}'} \prod_{l \in [d]} X^{(l)}_{\mathbf{i}_l} X^{(l)}_{\mathbf{i}'_l}. \end{aligned}$$

*Proof.* Note that

$$\prod_{l \in I} \left[ X^{(l)}_{\mathbf{i}_l} X^{(l)}_{\mathbf{i}'_l} - \mathbb{1}_{\mathbf{i}_l = \mathbf{i}'_l} \right] = \sum_{J \subset I} (-\mathbb{1}_{\mathbf{i}_l = \mathbf{i}'_l})^{|I \backslash J|} \prod_{l \in J} X^{(l)}_{\mathbf{i}_l} X^{(l)}_{\mathbf{i}'_l}.$$

Using this, we obtain

$$\begin{aligned}\alpha :=&\sum_{I\subset[d]}\sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}(I)}\sum_{\mathbf{j}\in\mathbf{J}^{\mathbf{n}}(I^c)} A_{(\mathbf{i}\dot{\times}\mathbf{j})\dot{+}(\mathbf{i}'\dot{\times}\mathbf{j})}\prod_{l\in I}\left[X^{(l)}_{\mathbf{i}_l}X^{(l)}_{\mathbf{i}'_l}-\mathbb{1}_{\mathbf{i}_l=\mathbf{i}'_l}\right]\\ =&\sum_{I\subset[d]}\sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}(I)}\sum_{\mathbf{j}\in\mathbf{J}^{\mathbf{n}}(I^c)} A_{(\mathbf{i}\dot{\times}\mathbf{j})\dot{+}(\mathbf{i}'\dot{\times}\mathbf{j})}\sum_{J\subset I}\prod_{l\in I\backslash J}(-\mathbb{1}_{\mathbf{i}_l=\mathbf{i}'_l})\prod_{l\in J}X^{(l)}_{\mathbf{i}_l}X^{(l)}_{\mathbf{i}'_l}\end{aligned}$$

Observing that

$$\prod_{l\in I\backslash J}(-\mathbb{1}_{\mathbf{i}_l=\mathbf{i}'_l})=\begin{cases}(-1)^{|I\backslash J|} & \text{if } \forall j\in I\backslash J:\mathbf{i}_l=\mathbf{i}'_l\\ 0 & \text{otherwise,}\end{cases}$$

we can conclude

$$\begin{aligned}\alpha =&\sum_{I\subset[d]}\sum_{J\subset I}\sum_{\substack{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}(J)\\ \mathbf{k}\in\mathbf{J}^{\mathbf{n}}(I\backslash J)}}\sum_{\mathbf{j}\in\mathbf{J}^{\mathbf{n}}(I^c)} A_{(\mathbf{i}\dot{\times}\mathbf{j}\dot{\times}\mathbf{k})\dot{+}(\mathbf{i}'\dot{\times}\mathbf{j}\dot{\times}\mathbf{k})}(-1)^{|I\backslash J|}\prod_{l\in J}X^{(l)}_{\mathbf{i}_l}X^{(l)}_{\mathbf{i}'_l}\\ =&\sum_{J\subset[d]}\sum_{I\supset J}(-1)^{|I\backslash J|}\sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}(J)}\sum_{\mathbf{j}\in\mathbf{J}^{\mathbf{n}}(J^c)} A_{(\mathbf{i}\dot{\times}\mathbf{j})\dot{+}(\mathbf{i}'\dot{\times}\mathbf{j})}\prod_{l\in J}X^{(l)}_{\mathbf{i}_l}X^{(l)}_{\mathbf{i}'_l}\end{aligned}$$

Lemma 2.17 yields

$$\sum_{I\supset J}(-1)^{|I\backslash J|}=\sum_{I'\subset[d]\backslash J}(-1)^{|I'|}=\begin{cases}1 & \text{if } J=[d]\\ 0 & \text{otherwise,}\end{cases}$$

such that

$$\alpha=\sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}}A_{\mathbf{i}\dot{+}\mathbf{i}'}\prod_{l\in[d]}X^{(l)}_{\mathbf{i}_l}X^{(l)}_{\mathbf{i}'_l}.$$

□

**Lemma 2.31.** *Let $\mathbf{A}\in\mathbb{R}^{\mathbf{n}^{\times 2}}$ be an array of order $2d$ and $X^{(1)}\in\mathbb{R}^{n_1},\dots X^{(d)}\in\mathbb{R}^{n_d}$ independent random vectors with mean $0$, variance $1$ entries. Then for any subset $\emptyset\neq I\subset[d]$, $p\geq 1$,*

$$\begin{aligned}&\left\|\sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}(I)}\sum_{\mathbf{j}\in\mathbf{J}^{\mathbf{n}}(I^c)} A_{(\mathbf{i}\dot{\times}\mathbf{j})\dot{\times}(\mathbf{i}'\dot{\times}\mathbf{j})}\prod_{l\in I}\left[X^{(l)}_{\mathbf{i}_l}X^{(l)}_{\mathbf{i}'_l}-\mathbb{1}_{\mathbf{i}_l=\mathbf{i}'_l}\right]\right\|_{L_p}\\ \leq& C(|I|)\left\|\sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}}A_{\mathbf{i}\dot{+}\mathbf{i}'}\prod_{l\in[d]}X^{(l)}_{\mathbf{i}_l}X^{(l)}_{\mathbf{i}'_l}-\mathbb{E}\sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}}A_{\mathbf{i}\dot{+}\mathbf{i}'}\prod_{l\in[d]}X^{(l)}_{\mathbf{i}_l}X^{(l)}_{\mathbf{i}'_l}\right\|_{L_p},\end{aligned}\tag{2.19}$$

*where $C(|I|)$ is a constant only depending on $|I|$.*

*Proof.* By the assumptions on the vectors $X^{(l)}$,

$$E:=\mathbb{E}\sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}}A_{\mathbf{i}\dot{+}\mathbf{i}'}\prod_{l\in[d]}X^{(l)}_{\mathbf{i}_l}X^{(l)}_{\mathbf{i}'_l}=\sum_{\mathbf{i}\in\mathbf{J}^{\mathbf{n}}}A_{\mathbf{i}\dot{+}\mathbf{i}}.$$

Since this is exactly the term for $I=\emptyset$ in Lemma 2.30, we obtain for the term on the right hand side of (2.19),

$$b:=\sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}}A_{\mathbf{i}\dot{+}\mathbf{i}'}\prod_{l\in[d]}X^{(l)}_{\mathbf{i}_l}X^{(l)}_{\mathbf{i}'_l}-E$$

$$= \sum_{\substack{\emptyset \neq J \subset [d] \\ \mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}(J) \\ \mathbf{j} \in \mathbf{J}^{\mathbf{n}}(J^c)}} A_{(\mathbf{i} \dot{\times} \mathbf{j}) \dot{+} (\mathbf{i}' \dot{\times} \mathbf{j})} \prod_{l \in J} \left[ X^{(l)}_{\mathbf{i}_l} X^{(l)}_{\mathbf{i}'_l} - \mathbb{1}_{\mathbf{i}_l = \mathbf{i}'_l} \right] =: \sum_{\substack{J \subset [d] \\ J \neq \emptyset}} S_J.$$

Using these terms, we need to show that $\|S_I\|_{L_p} \leq C(|I|)\|b\|_{L_p}$ for all $\emptyset \neq I \subset [d]$.

Now we prove this by induction over $|I|$. First assume $I = \{l_0\}$. For any $J \neq \emptyset, I$, there exists an $l \in J \backslash I$ and then

$$\mathbb{E} \left[ \prod_{l \in J} \left[ X^{(l)}_{\mathbf{i}_l} X^{(l)}_{\mathbf{i}'_l} - \mathbb{1}_{\mathbf{i}_l = \mathbf{i}'_l} \right] \middle| X^{(l_0)} \right] = 0$$

since there is at least one factor whose conditional expectation is 0.

We conclude

$$\begin{aligned} \mathbb{E}\,|S_I|^p =& \mathbb{E} \left| S_I + \mathbb{E} \left[ \sum_{J \subset [d] : J \neq \emptyset, I} S_J \middle| X^{(l_0)} \right] \right|^p \\ =& \mathbb{E} \left| \mathbb{E} \left[ \sum_{J \subset [d] : J \neq \emptyset} S_J \middle| X^{(l_0)} \right] \right|^p \leq \mathbb{E}|b|^p, \end{aligned}$$

where we used Jensen's inequality on the conditional expectation in the last step.

Now assume that we have already shown (2.19) for all $\emptyset \neq I' \subset [d]$ with $|I'| < |I|$.

For all $J \subset [d]$ such that $J \neq \emptyset, I$, one of the following holds.

- $J \backslash I = \emptyset$, i.e., $J \subset I$: Because $J \neq I$, $|J| < |I|$, so by induction
$$\|S_J\|_{L_p} \leq C(|J|)\|b\|_{L_p}. \tag{2.20}$$
- $J \backslash I \neq \emptyset$. Since there is an $l' \in J \backslash I$,
$$\mathbb{E} \left[ \prod_{l \in J} \left[ X^{(l)}_{\mathbf{i}_l} X^{(l)}_{\mathbf{i}'_l} - \mathbb{1}_{\mathbf{i}_l = \mathbf{i}'_l} \right] \middle| (X^{(l)})_{l \in I} \right] = 0. \tag{2.21}$$

The triangle inequality yields together with (2.20), that $\|S_I\|_{L_p} \leq$

$$\left\| S_I + \sum_{\substack{J \subset I \\ J \neq \emptyset, I}} S_J \right\|_{L_p} + \sum_{\substack{J \subset I \\ J \neq \emptyset, I}} \|S_J\|_{L_p} \leq \left\| S_I + \sum_{\substack{J \subset I \\ J \neq \emptyset, I}} S_J \right\|_{L_p} + \left[ \sum_{J \subset I, J \neq \emptyset, I} C(|J|) \right] \|b\|_{L_p}.$$

The first term on the right hand side can be controlled with (2.21) and Jensen's inequality,

$$\begin{aligned} \mathbb{E} \left| S_I + \sum_{J \subset I : J \neq \emptyset, I} S_J \right|^p =& \mathbb{E} \left| S_I + \sum_{J \subset I : J \neq \emptyset, I} S_J + \mathbb{E} \left[ \sum_{J \subset [d] : J \backslash I \neq \emptyset} S_J \middle| (X^{(l)})_{l \in I} \right] \right|^p \\ =& \mathbb{E} \left| \mathbb{E} \left[ \sum_{J \subset [d] : J \neq \emptyset} S_J \middle| (X^{(l)})_{l \in I} \right] \right|^p \leq \mathbb{E}|b|^p. \end{aligned}$$

So altogether $\|S_I\|_{L_p} \leq C(|I|)\|b\|_{L_p}$ where $C(|I|) := \sum_{J \subset I : J \neq \emptyset, I} C(|J|) + 1$ depends only on $|I|$. □

Now we introduced all the necessary tools and can prove the lower bound of the main result, Theorem 2.9.

**Proof of Theorem 2.9, lower bound:**

For any $J \subset I \subset [d]$, define the array $\mathbf{A}^{(I,J)}$ as in the proof of the upper bound (2.17) and

$$\alpha^{(I,J)} := \left\| \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}(J \cup I^c)} A^{(I,J)}_{\mathbf{i} \dot{+} \mathbf{i}'} \prod_{l \in I^c \cup J} g^{(l)}_{\mathbf{i}_l} \bar{g}^{(l)}_{\mathbf{i}'_l} \right\|_{L_p} \tag{2.22}$$

**Step 1: Adding off-diagonal terms**

Define independent Rademacher vectors $(\xi^{(l)})_{l \in J}$ that are also independent of the vectors $g^{(1)}, \dots g^{(d)}$, $\bar{g}^{(1)}, \dots, \bar{g}^{(d)}$.

Noting that $\mathbb{E}_\xi[\xi^{(l)}_{\mathbf{i}_l} \xi^{(l)}_{\mathbf{i}'_l}] = \mathbb{1}_{\mathbf{i}_l = \mathbf{i}'_l}$, we obtain

$$\begin{aligned}
&\mathbb{E}_\xi \left[ \sum_{\substack{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}(J) \\ \mathbf{k} \in \mathbf{J}^{\mathbf{n}}(I \setminus J) \\ \mathbf{j},\mathbf{j}' \in \mathbf{J}^{\mathbf{n}}(I^c)}} A_{(\mathbf{i} \dot{\times} \mathbf{j} \dot{\times} \mathbf{k}) \dot{+} (\mathbf{i}' \dot{\times} \mathbf{j}' \dot{\times} \mathbf{k})} \prod_{l \in I^c} g^{(l)}_{\mathbf{j}_l} \bar{g}^{(l)}_{\mathbf{j}'_l} \prod_{l \in J} (\xi^{(l)}_{\mathbf{i}_l} g^{(l)}_{\mathbf{i}_l})(\xi^{(l)}_{\mathbf{i}'_l} \bar{g}^{(l)}_{\mathbf{i}'_l}) \right] \\
&= \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}(J \cup I^c)} A^{(I,J)}_{\mathbf{i} \dot{+} \mathbf{i}'} \prod_{l \in I^c \cup J} g^{(l)}_{\mathbf{i}_l} \bar{g}^{(l)}_{\mathbf{i}'_l}
\end{aligned}$$

Substituting into (2.22) and applying Jensen's inequality and Fubini's theorem yields

$$\begin{aligned}
&(\alpha^{(I,J)})^p \\
&= \mathbb{E}_{g,\bar{g}} \left| \mathbb{E}_\xi \sum_{\substack{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}(J) \\ \mathbf{j},\mathbf{j}' \in \mathbf{J}^{\mathbf{n}}(I^c)}} \sum_{\mathbf{k} \in \mathbf{J}^{\mathbf{n}}(I \setminus J)} A_{(\mathbf{i} \dot{\times} \mathbf{j} \dot{\times} \mathbf{k}) \dot{+} (\mathbf{i}' \dot{\times} \mathbf{j}' \dot{\times} \mathbf{k})} \prod_{l \in I^c} g^{(l)}_{\mathbf{j}_l} \bar{g}^{(l)}_{\mathbf{j}'_l} \prod_{l \in J} (\xi^{(l)}_{\mathbf{i}_l} g^{(l)}_{\mathbf{i}_l})(\xi^{(l)}_{\mathbf{i}'_l} \bar{g}^{(l)}_{\mathbf{i}'_l}) \right|^p \\
&\leq \mathbb{E}_\xi \mathbb{E}_{g,\bar{g}} \left| \sum_{\substack{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}(J) \\ \mathbf{j},\mathbf{j}' \in \mathbf{J}^{\mathbf{n}}(I^c)}} \sum_{\mathbf{k} \in \mathbf{J}^{\mathbf{n}}(I \setminus J)} A_{(\mathbf{i} \dot{\times} \mathbf{j} \dot{\times} \mathbf{k}) \dot{+} (\mathbf{i}' \dot{\times} \mathbf{j}' \dot{\times} \mathbf{k})} \prod_{l \in I^c} g^{(l)}_{\mathbf{j}_l} \bar{g}^{(l)}_{\mathbf{j}'_l} \prod_{l \in J} (\xi^{(l)}_{\mathbf{i}_l} g^{(l)}_{\mathbf{i}_l})(\xi^{(l)}_{\mathbf{i}'_l} \bar{g}^{(l)}_{\mathbf{i}'_l}) \right|^p
\end{aligned}$$

By the symmetry of the normal distribution, conditioned on $(\xi^{(l)})_{l \in J}$, $(\xi^{(l)}_{\mathbf{i}_l} g^{(l)}_{\mathbf{i}_l}, \xi^{(l)}_{\mathbf{i}'_l} \bar{g}^{(l)}_{\mathbf{i}'_l})$ and $(g^{(l)}_{\mathbf{i}_l}, \bar{g}^{(l)}_{\mathbf{i}'_l})$ have the same distribution. So we can conclude

$$\alpha^{(I,J)} \leq \left\| \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}(J \cup I^c)} \sum_{\mathbf{k} \in \mathbf{J}^{\mathbf{n}}(I \setminus J)} A_{(\mathbf{i} \dot{\times} \mathbf{k}) \dot{+} (\mathbf{i}' \dot{\times} \mathbf{k})} \prod_{l \in J \cup I^c} g^{(l)}_{\mathbf{i}_l} \bar{g}^{(l)}_{\mathbf{i}'_l} \right\|_{L_p}.$$

**Step 2: Inverse Gaussian decoupling**

For every $J \subset I \subset [d]$, we obtain then by the symmetry of $\mathbf{A}$ and Corollary 2.29,

$$\alpha^{(I,J)} \leq \left\| \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}(J \cup I^c)} \sum_{\mathbf{k} \in \mathbf{J}^{\mathbf{n}}(I \setminus J)} A_{(\mathbf{i} \dot{\times} \mathbf{k}) \dot{+} (\mathbf{i}' \dot{\times} \mathbf{k})} \prod_{l \in J \cup I^c} \left[ g^{(l)}_{\mathbf{i}_l} g^{(l)}_{\mathbf{i}'_l} - \mathbb{1}_{\mathbf{i}_l = \mathbf{i}'_l} \right] \right\|_{L_p}.$$

**Step 3: Removing the mean subtractions in every factor**

Since $I \backslash J \neq [d]$, $J \cup I^c \neq \emptyset$ and Lemma 2.31 provides that $\alpha^{(I,J)} \leq$

$$C_1(|J \cup I^c|) \left\| \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} A_{\mathbf{i} \dot{+} \mathbf{i}'} \prod_{l \in [d]} g^{(l)}_{\mathbf{i}_l} g^{(l)}_{\mathbf{i}'_l} - \mathbb{E} \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} A_{\mathbf{i} \dot{+} \mathbf{i}'} \prod_{l \in [d]} g^{(l)}_{\mathbf{i}_l} g^{(l)}_{\mathbf{i}'_l} \right\|_{L_p}.$$

Adding this up over all $J \subset I \subset [d]$, $I\backslash J \neq [d]$ yields

$$\begin{aligned} &\sum_{\substack{J\subset I\subset[d]\\ I\backslash J\neq[d]}} \alpha^{(I,J)} \\ &\leq C(d) \left\| \sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}} A_{\mathbf{i}+\mathbf{i}'} \prod_{l\in[d]} g^{(l)}_{\mathbf{i}_l} g^{(l)}_{\mathbf{i}'_l} - \mathbb{E} \sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}} A_{\mathbf{i}+\mathbf{i}'} \prod_{l\in[d]} g^{(l)}_{\mathbf{i}_l} g^{(l)}_{\mathbf{i}'_l} \right\|_{L_p} \end{aligned} \tag{2.23}$$

where $C(d) := \sum_{J\subset I\subset[d]: I\backslash J\neq[d]} C_1(|J\cup I^c|)$ depends only on $d$.

**Step 4: Completing the proof**

Restricting the left hand side in (2.23) to the terms in which $J=\emptyset$. The remaining terms $\alpha^{(I,\emptyset)}$ only contain the arrays $\mathbf{A}^{(I,\emptyset)}$ which are equal to the $\mathbf{A}^{(I)}$ from the theorem statement. Subsequently, we can bound the $\alpha^{(I,\emptyset)}$ from below using Theorem 2.7 (similarly to the upper bound) to obtain the lower bound in Theorem 2.9.

### 2.3.4 Concentration of $\|BX\|_2$

In this section, we apply our main results to the concentration of $\|BX\|_2$ where $X = X^{(1)} \otimes \cdots \otimes X^{(d)}$ is a Kronecker product of independent vectors with subgaussian entries. The following statement is a direct consequence from Theorem 2.9 and Lemma 2.21.

**Corollary 2.32.** *Let $B \in \mathbb{R}^{n_0\times N}$ be a matrix where $N = n_1 \dots n_d$ and $X := X^{(1)} \otimes\cdots\otimes X^{(d)} \in \mathbb{R}^N$ a random vector as in Theorem 2.9.*

*Let $\mathbf{A} \in \mathbb{R}^{\mathbf{n}^{\times 2}}$ be the rearrangement of the matrix $A = B^*B$ as an array with $2d$ axes. For any $I \subset [d]$, define the array $\mathbf{A}^{(I)}$ as in (2.10).*

*For $T \subset [2d]$, $1 \leq \kappa \leq 2d$, denote $S(T,\kappa)$ for the set of partitions of $T$ into $\kappa$ sets and $I^c = [d]\backslash I$. Define for any $p \geq 1$ and any $\kappa \in [2d]$,*

$$m_{p,\kappa} := \sum_{\substack{I\subset[d]\\ I\neq[d]}} \sum_{(I_1,\dots,I_\kappa)\in S((I^c)\cup(I^c+d),\kappa)} \|\mathbf{A}^{(I)}\|_{I_1,\dots,I_\kappa}$$

$$m_p := L^{2d} \sum_{\kappa=1}^{2d} \min\left\{ p^{\frac{\kappa}{2}} \frac{m_{p,\kappa}}{\|B\|_F}, p^{\frac{\kappa}{4}} \sqrt{m_{p,\kappa}} \right\}$$

*Then there is a constant $C(d) > 0$, depending only on $d$, such that for all $p \geq 1$,*

$$\|\|BX\|_2 - \|B\|_F\|_{L_p} \leq C(d) m_p.$$

*If in addition, $X^{(1)} \sim N(0, Id_{n_1}), \dots, X^{(d)} \sim N(0, Id_{n_d})$ are normally distributed (i.e., $L$ is constant) and $\mathbf{A}$ satisfies the symmetry condition (2.11), then also the lower bound*

$$\tilde{C}(d) m_p \leq \|\|BX\|_2 - \|B\|_F\|_{L_p}$$

*holds for all $p \geq 1$. Above, $\tilde{C}(d) > 0$ that depends only on $d$.*

**Lemma 2.33.** *Let $\mathbf{B} \in \mathbb{R}^{n_1\times\cdots\times n_d}$. Assume that $I_1, \dots, I_\kappa$ is a partition of $[d]$. Let $\bar{I}_\kappa \cup \bar{I}_{\kappa+1} = I_\kappa$ be a partition into two subsets. Then*

$$\begin{aligned} \|\mathbf{B}\|_{I_1,\dots,I_{\kappa-1},\bar{I}_\kappa,\bar{I}_{\kappa+1}} &\leq \|\mathbf{B}\|_{I_1,\dots,I_\kappa} \\ &\leq \sqrt{\min\left\{ \prod_{l\in\bar{I}_\kappa} n_l, \prod_{l\in\bar{I}_{\kappa+1}} n_l \right\}} \|\mathbf{B}\|_{I_1,\dots,I_{\kappa-1},\bar{I}_\kappa,\bar{I}_{\kappa+1}}. \end{aligned}$$

*Proof.* Take arrays $\alpha^{(1)} \in \mathbb{R}^{\mathbf{n}}(I_1), \ldots, \alpha^{(\kappa-1)} \in \mathbb{R}^{\mathbf{n}}(I_{\kappa-1}), \bar{\alpha}^{(\kappa)} \in \mathbb{R}^{\mathbf{n}}(\bar{I}_\kappa), \bar{\alpha}^{(\kappa+1)} \in \mathbb{R}^{\mathbf{n}}(\bar{I}_{\kappa+1})$, with Frobenius norm 1 each, such that $\|\mathbf{B}\|_{I_1,\ldots,I_{\kappa-1},\bar{I}_\kappa,\bar{I}_{\kappa+1}} = \sum_{\mathbf{i}\in\mathbf{J}^{\mathbf{n}}} B_{\mathbf{i}} \alpha^{(1)}_{\mathbf{i}_{I_1}} \ldots \alpha^{(\kappa-1)}_{\mathbf{i}_{I_{\kappa-1}}} \bar{\alpha}^{(\kappa)}_{\mathbf{i}_{\bar{I}_\kappa}} \bar{\alpha}^{(\kappa+1)}_{\mathbf{i}_{\bar{I}_{\kappa+1}}}$. Now define $\alpha^{(\kappa)} \in \mathbb{R}^{\mathbf{n}}(I_\kappa)$ by $\alpha^{(\kappa)}_{\mathbf{i}} = \bar{\alpha}^{(\kappa)}_{\mathbf{i}_{\bar{I}_\kappa}} \bar{\alpha}^{(\kappa+1)}_{\mathbf{i}_{\bar{I}_{\kappa+1}}}$ for every $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(I_\kappa)$. Then $\|\alpha^{(\kappa)}\|_2 = 1$ and by the definition of $\|\cdot\|_{I_1,\ldots,I_\kappa}$ as the supremum over $\alpha^{(1)}, \ldots, \alpha^{(\kappa)}$, we obtain

$$\|\mathbf{B}\|_{I_1,\ldots,I_{\kappa-1},\bar{I}_\kappa,\bar{I}_{\kappa+1}} = \sum_{\mathbf{i}\in\mathbf{J}^{\mathbf{n}}} B_{\mathbf{i}} \alpha^{(1)}_{\mathbf{i}_{I_1}} \ldots \alpha^{(\kappa)}_{\mathbf{i}_{I_\kappa}} \leq \|\mathbf{B}\|_{I_1,\ldots,I_\kappa},$$

which proves the first inequality.

To prove the second inequality, take arrays $\alpha^{(1)} \in \mathbb{R}^{\mathbf{n}}(I_1), \ldots, \alpha^{(\kappa)} \in \mathbb{R}^{\mathbf{n}}(I_\kappa)$ such that

$$\|\mathbf{B}\|_{I_1,\ldots,I_\kappa} = \sum_{\mathbf{i}\in\mathbf{J}^{\mathbf{n}}} B_{\mathbf{i}} \alpha^{(1)}_{\mathbf{i}_{I_1}} \ldots \alpha^{(\kappa)}_{\mathbf{i}_{I_\kappa}}.$$

Now define $\tilde{\mathbf{B}} \in \mathbb{R}^{\mathbf{n}}(I_\kappa)$ such that for all $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(\bar{I}_\kappa), \mathbf{j} \in \mathbf{J}^{\mathbf{n}}(\bar{I}_{\kappa+1})$,

$$\tilde{\mathbf{B}}_{\mathbf{i}\dot{\times}\mathbf{j}} = \sum_{\mathbf{k}\in\mathbf{J}^{\mathbf{n}}([d]\setminus I_\kappa)} B_{\mathbf{i}\dot{\times}\mathbf{j}\dot{\times}\mathbf{k}} \alpha^{(1)}_{\mathbf{k}_{I_1}} \ldots \alpha^{(\kappa-1)}_{\mathbf{k}_{I_{\kappa-1}}}.$$

For $N_1 := \prod_{l\in\bar{I}_\kappa} n_l$ and $N_2 := \prod_{l\in\bar{I}_{\kappa+1}} n_l$, we can interpret $\tilde{\mathbf{B}}$ as a matrix $\tilde{B} \in \mathbb{R}^{N_1\times N_2}$ with rows indexed by $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(\bar{I}_\kappa)$ and columns indexed by $\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(\bar{I}_{\kappa+1})$.

Then

$$\|\tilde{B}\|_F = \sup_{\beta\in\mathbb{R}^{\mathbf{n}}(I_\kappa),\|\beta\|_2=1} \sum_{\mathbf{i}\in\mathbf{J}^{\mathbf{n}}(I_\kappa)} \tilde{B}_{\mathbf{i}}\beta_{\mathbf{i}},$$

$$\|\tilde{B}\|_{2\to2} = \sup_{\substack{\beta^{(1)}\in\mathbb{R}^{\mathbf{n}}(\bar{I}_\kappa),\beta^{(2)}\in\mathbb{R}^{\mathbf{n}}(\bar{I}_{\kappa+1}),\\ \|\beta^{(1)}\|_2=\|\beta^{(2)}\|_2=1}} \sum_{\substack{\mathbf{i}\in\mathbf{J}^{\mathbf{n}}(\bar{I}_\kappa)\\ \mathbf{j}\in\mathbf{J}^{\mathbf{n}}(\bar{I}_{\kappa+1})}} \tilde{B}_{\mathbf{i}\dot{\times}\mathbf{j}}\beta^{(1)}_{\mathbf{i}}\beta^{(2)}_{\mathbf{j}},$$

such that

$$\begin{aligned}\|\tilde{B}\|_F &= \sup_{\beta\in\mathbb{R}^{\mathbf{n}}(I_\kappa),\|\beta\|_2=1} \sum_{\mathbf{i}\in\mathbf{J}^{\mathbf{n}}(I_\kappa)} \sum_{\mathbf{k}\in\mathbf{J}^{\mathbf{n}}([d]\setminus I_\kappa)} B_{\mathbf{i}\dot{\times}\mathbf{k}} \alpha^{(1)}_{\mathbf{k}_{I_1}} \ldots \alpha^{(\kappa-1)}_{\mathbf{k}_{I_{\kappa-1}}} \beta_{\mathbf{i}} \\ &= \sup_{\beta\in\mathbb{R}^{\mathbf{n}}(I_\kappa),\|\beta\|_2=1} \sum_{\mathbf{i}\in\mathbf{J}^{\mathbf{n}}} B_{\mathbf{i}} \alpha^{(1)}_{\mathbf{i}_{I_1}} \ldots \alpha^{(\kappa-1)}_{\mathbf{i}_{I_{\kappa-1}}} \beta_{\mathbf{i}_{I_\kappa}},\end{aligned}$$

where by definition the maximum is attained at $\beta = \alpha^{(\kappa)}$, implying

$$\|\tilde{B}\|_F = \|\mathbf{B}\|_{I_1,\ldots,I_\kappa}. \tag{2.24}$$

For the spectral norm, we obtain from the definition of $\|\cdot\|_{I_1,\ldots,I_{\kappa-1},\bar{I}_\kappa,\bar{I}_{\kappa+1}}$,

$$\begin{aligned}&\|\tilde{B}\|_{2\to2} \\ &= \sup_{\substack{\beta^{(1)}\in\mathbb{R}^{\mathbf{n}}(\bar{I}_\kappa),\beta^{(2)}\in\mathbb{R}^{\mathbf{n}}(\bar{I}_{\kappa+1}),\\ \|\beta^{(1)}\|_2=\|\beta^{(2)}\|_2=1}} \sum_{\substack{\mathbf{i}\in\mathbf{J}^{\mathbf{n}}(\bar{I}_\kappa)\\ \mathbf{j}\in\mathbf{J}^{\mathbf{n}}(\bar{I}_{\kappa+1})}} \sum_{\mathbf{k}\in\mathbf{J}^{\mathbf{n}}([d]\setminus I_\kappa)} B_{\mathbf{i}\dot{\times}\mathbf{j}\dot{\times}\mathbf{k}} \alpha^{(1)}_{\mathbf{k}_{I_1}} \ldots \alpha^{(\kappa-1)}_{\mathbf{k}_{I_{\kappa-1}}} \beta^{(1)}_{\mathbf{i}} \beta^{(2)}_{\mathbf{j}} \\ &= \sup_{\substack{\beta^{(1)}\in\mathbb{R}^{\mathbf{n}}(\bar{I}_\kappa),\beta^{(2)}\in\mathbb{R}^{\mathbf{n}}(\bar{I}_{\kappa+1}),\\ \|\beta^{(1)}\|_2=\|\beta^{(2)}\|_2=1}} \sum_{\mathbf{i}\in\mathbf{J}^{\mathbf{n}}} B_{\mathbf{i}} \alpha^{(1)}_{\mathbf{i}} \ldots \alpha^{(\kappa-1)}_{\mathbf{i}_{I_{\kappa-1}}} \beta^{(1)}_{\mathbf{i}_{\bar{I}_\kappa}} \beta^{(2)}_{\mathbf{j}_{\bar{i}_{\kappa+1}}} \\ &\leq \|\mathbf{B}\|_{I_1,\ldots,I_{\kappa-1},\bar{I}_\kappa,\bar{I}_{\kappa+1}}.\end{aligned} \tag{2.25}$$

The second inequality now follows from (2.24), (2.25) and the general property of matrices that

$$\|\tilde{B}\|_F \leq \sqrt{\operatorname{rank}(\tilde{B})}\|\tilde{B}\|_{2\to2} \leq \sqrt{\min\{N_1,N_2\}}\|\tilde{B}\|_{2\to2}.$$

□

**Lemma 2.34.** *Let $\mathbf{A} \in \mathbb{R}^{\mathbf{n}^{\times 2}}$, $I \subset [d]$. Define $\mathbf{A}^{(I)}$ as in* (2.10).

*Let $I_1, \ldots, I_\kappa$ be a partition of $([d] \backslash I) \cup (d + ([d] \backslash I))$. Let $I_{\kappa+1}, \ldots, I_{\kappa+|I|}$ be the sets $\{j, j+d\}$ for every $j \in I$. Then $I_1, \ldots, I_{\kappa+|I|}$ is a paritition of $[2d]$ and*

$$\|\mathbf{A}^{(I)}\|_{I_1,\ldots,I_\kappa} \leq \sqrt{\prod_{l \in I} n_l} \|\mathbf{A}\|_{I_1,\ldots,I_{\kappa+|I|}}$$

*Proof.* Take $\alpha^{(1)} \in \mathbb{R}^{\mathbf{n}^{\times 2}}(I_1), \ldots, \alpha^{(\kappa)} \in \mathbb{R}^{\mathbf{n}^{\times 2}}(I_\kappa)$, all having a Frobenius norm of 1, such that

$$\begin{aligned}
\|\mathbf{A}^{(I)}\|_{I_1,\ldots,I_\kappa} &= \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}^{\times 2}}(I^c \cup (I^c + d))} A^{(I)}_{\mathbf{i}} \alpha^{(1)}_{\mathbf{i}_{I_1}} \ldots \alpha^{(\kappa)}_{\mathbf{i}_{I_\kappa}} \\
&= \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}^{\times 2}}(I^c \cup (I^c + d))} \sum_{\mathbf{k} \in \mathbf{J}^{\mathbf{n}}(I)} A_{\mathbf{i} \dot{\times} (\mathbf{k} \dot{+} \mathbf{k})} \alpha^{(1)}_{\mathbf{i}_{I_1}} \ldots \alpha^{(\kappa)}_{\mathbf{i}_{I_\kappa}} \\
&= \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}^{\times 2}}} A_{\mathbf{i}} \alpha^{(1)}_{\mathbf{i}_{I_1}} \ldots \alpha^{(\kappa)}_{\mathbf{i}_{I_\kappa}} \mathbb{1}_{\forall l \in I : \mathbf{i}_l = \mathbf{i}_{l+d}}.
\end{aligned} \tag{2.26}$$

Now define $\alpha^{(\kappa+1)} \in \mathbb{R}^{\mathbf{n}^{\times 2}}(\{j_1, j_1 + d\}), \ldots, \alpha^{(\kappa+|I|)} \in \mathbb{R}^{\mathbf{n}^{\times 2}}(\{j_{|I|}, j_{|I|} + d\})$ (where $I = \{j_1, \ldots, j_{|I|}\}$) such that for all $r \in [|I|]$ and $\mathbf{i} \in \mathbf{J}^{\mathbf{n}^{\times 2}}(\{j_r, j_r + d\})$,

$$\alpha^{(\kappa+r)}_{\mathbf{i}} = \begin{cases} \frac{1}{\sqrt{n_{j_r}}} & \text{if } \mathbf{i}_{j_r} = \mathbf{i}_{j_r+d} \\ 0 & \text{otherwise.} \end{cases}$$

Then for $\mathbf{i} \in \mathbf{J}^{\mathbf{n}^{\times 2}}(I \cup (I + d))$

$$\alpha^{(\kappa+1)}_{\mathbf{i}_{I_{\kappa+1}}} \ldots \alpha^{(\kappa+|I|)}_{\mathbf{i}_{I_{\kappa+|I|}}} = \frac{1}{\sqrt{\prod_{l \in I} n_l}} \mathbb{1}_{\forall l \in I : \mathbf{i}_l = \mathbf{i}_{l+d}}$$

Substituting this into (2.26) yields

$$\begin{aligned}
\|\mathbf{A}^{(I)}\|_{I_1,\ldots,I_\kappa} &= \sqrt{\prod_{l \in I} n_l} \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}^{\times 2}}} A_{\mathbf{i}} \alpha^{(1)}_{\mathbf{i}_{I_1}} \ldots \alpha^{(\kappa)}_{\mathbf{i}_{I_\kappa}} \alpha^{(\kappa+1)}_{\mathbf{i}_{I_{\kappa+1}}} \ldots \alpha^{(\kappa+|I|)}_{\mathbf{i}_{I_{\kappa+|I|}}} \\
&\leq \sqrt{\prod_{l \in I} n_l} \|\mathbf{A}\|_{I_1,\ldots,I_{\kappa+|I|}}
\end{aligned}$$

□

Using the aforementioned results, we can give the proof of Theorem 2.13 about $\|B(X^{(1)} \otimes \cdots \otimes X^{(d)})\|_2$ in which we find suitable bounds for all the tensor norms of $B^*B$ in terms of $\|B\|_{2\to 2}$ and $\|B\|_F$.

*Proof of Theorem 2.13.* Let $A := B^*B \in \mathbb{R}^{n^d \times n^d}$ and $\mathbf{A} \in \mathbb{R}^{\mathbf{n}^{\times 2}}$ be the corresponding array of order $2d$ obtained by rearranging $A$ for $\mathbf{n} = (n, \ldots, n)$. Note that here the dimensions along all axes are equal. For $I \subset [2d]$, define $\mathbf{A}^{(I)}$ as in Corollary 2.32.

**Step 1: Showing the norm inequalities**

$$\|\mathbf{A}^{(I)}\|_{I_1,\ldots,I_\kappa} \leq n^{\frac{|I|}{2}} \|A\|_F \qquad \|\mathbf{A}^{(I)}\|_{I_1,\ldots,I_\kappa} \leq n^{d - \frac{\kappa}{2}} \|A\|_{2\to 2}. \tag{2.27}$$

In both cases, we start by extending $I_1, \ldots, I_\kappa$ to $I_1, \ldots, I_{\kappa+|I|}$ as in Lemma 2.34, obtaining

$$\|\mathbf{A}^{(I)}\|_{I_1,\ldots,I_\kappa} \leq n^{\frac{|I|}{2}} \|\mathbf{A}\|_{I_1,\ldots,I_{\kappa+|I|}} \tag{2.28}$$

Then the first inequality of (2.27) follows by repeatedly joining all the sets $I_1, \dots, I_{\kappa+|I|}$ in the sense of Lemma 2.33 (first inequality) yielding $\|\mathbf{A}\|_{I_1,\dots,I_{\kappa+|I|}} \leq \|\mathbf{A}\|_{[2d]} = \|A\|_F$.

For the second inequality in (2.27), we distinguish two cases. First assume that $\kappa \leq d - |I|$. Then $|I| \leq d - \kappa$. Since $A$ is a matrix in $\mathbb{R}^{n^d \times n^d}$, $\|A\|_{2\to 2} \leq n^{\frac{d}{2}} \|A\|_F$ and with the first inequality in (2.27), we obtain

$$\|\mathbf{A}^{(I)}\|_{I_1,\dots,I_\kappa} \leq n^{\frac{|I|}{2}} n^{\frac{d}{2}} \|A\|_{2\to 2} \leq n^{\frac{d-\kappa}{2}} n^{\frac{d}{2}} \|A\|_{2\to 2} = n^{d-\frac{\kappa}{2}} \|A\|_{2\to 2}.$$

In the other case that $\kappa > d - |I|$, denote $\kappa'$ for the number of sets among $I_1, \dots, I_\kappa$ that only contain one element. Since each of the other sets must contain at least two elements, this leads to the inequality

$$\kappa' + 2(\kappa - \kappa') \leq |I_1 \cup \dots \cup I_\kappa| \qquad \Rightarrow 2\kappa - \kappa' \quad \leq 2(d - |I|) \qquad \Rightarrow \kappa' \geq 2(\kappa - d + |I|).$$

This implies that among $I_1, \dots, I_\kappa$, there must be at least $\kappa - d + |I|$ sets with exactly one element that are all contained in $[d]$ or all contained in $[2d]\backslash[d]$. Without loss of generality, we can assume that these are $I_1, \dots, I_{\kappa-d+|I|}$. Now take the unions $\bar{I}_1 := I_1 \cup \dots \cup I_{\kappa-d+|I|}$ and $\bar{I}_2 := I_{\kappa-d+|I|+1} \cup \dots \cup I_{\kappa+|I|}$. With (2.28) and the first inequality of Lemma 2.33, we obtain

$$\|\mathbf{A}^{(I)}\|_{I_1,\dots,I_\kappa} \leq n^{\frac{|I|}{2}} \|\mathbf{A}\|_{\bar{I}_1,\bar{I}_2}.$$

Now split up $\bar{I}_2$ into $\bar{I}_{2,1} := \bar{I}_2 \cap [d]$ and $\bar{I}_{2,2} := \bar{I}_2 \cap ([2d]\backslash[d])$. If neither $\bar{I}_{2,1}$ nor $\bar{I}_{2,2}$ is empty, then with the second inequality of Lemma 2.33, we obtain

$$\begin{aligned}\|\mathbf{A}^{(I)}\|_{I_1,\dots,I_\kappa} \leq & n^{\frac{|I|}{2}} n^{\frac{1}{2}\min\{|\bar{I}_{2,1}|,|\bar{I}_{2,2}|\}} \|\mathbf{A}\|_{\bar{I}_1,\bar{I}_{2,1},\bar{I}_{2,2}} \\ \leq & n^{\frac{|I|}{2}+\frac{1}{2}\min\{|\bar{I}_{2,1}|,|\bar{I}_{2,2}|\}} \|\mathbf{A}\|_{[d],([2d]\backslash[d])},\end{aligned}$$

where in the last step we used the first inequality in Lemma 2.33 with the fact that $\bar{I}_1 \cup \bar{I}_{2,1} \cup \bar{I}_{2,2} = [2d]$ and each of these three sets is contained in either $[d]$ or $[2d]\backslash[d]$. Note that the inequality between the first and the third term still holds in the case that $\bar{I}_{2,1}$ or $\bar{I}_{2,2}$ is empty and thus Lemma 8.4 cannot be applied in the first step.

Now assume $\bar{I}_1 \subset [d]$ (otherwise $\bar{I}_1 \subset [2d]\backslash[d]$ and the proof works analogously). Then $\bar{I}_1 \cup \bar{I}_{2,1} = [d]$ and $\bar{I}_{2,1} = [2d]\backslash[d]$. So $\min\{|\bar{I}_{2,1}|, |\bar{I}_{2,2}|\} = |\bar{I}_{2,1}| = d - |\bar{I}_1| = d - (\kappa - d + |I|) = 2d - \kappa - |I|$. This implies

$$\|\mathbf{A}^{(I)}\|_{I_1,\dots,I_\kappa} \leq n^{\frac{|I|}{2}+\frac{1}{2}(2d-\kappa-|I|)} \|\mathbf{A}\|_{[d],([2d]\backslash[d])} = n^{d-\frac{\kappa}{2}} \|A\|_{2\to 2}.$$

This completes the proof of (2.27).

**Step 2: Moment and tail bounds**

Now, use Corollary 2.32 and its notation of $m_{p,\kappa}$ and $m_p$. The number of terms in the sum of the definition of $m_{p,\kappa}$ only depends on $d$. This fact together with (2.27) leads to

$$\begin{aligned} m_{p,\kappa} \leq & C_1(d) \max_{I \subset [d], I \neq [d]} n^{\frac{|I|}{2}} \|A\|_F = C_1(d) n^{\frac{d-1}{2}} \|A\|_F \leq C_1(d) n^{\frac{d-1}{2}} \|B\|_2 \|B\|_F. \\ m_{p,\kappa} \leq & C_1(d) n^{d-\frac{\kappa}{2}} \|A\|_{2\to 2} = C_1(d) n^{d-\frac{\kappa}{2}} \|B\|_{2\to 2}^2, \end{aligned}$$

where $C_1(d)$ is a constant depending only on $d$. Furthermore, we obtain

$$\begin{aligned} m_p \leq & C_1(d) L^{2d} \\ & \cdot \sum_{\kappa=1}^{2d} \min\left\{ p^{\frac{\kappa}{2}} n^{\frac{d-1}{2}} \|B\|_{2\to 2}, p^{\frac{\kappa}{2}} n^{d-\frac{\kappa}{2}} \frac{\|B\|_{2\to 2}^2}{\|B\|_F}, \right. \end{aligned}$$

$$p^{\frac{\kappa}{4}} n^{\frac{d-1}{4}} \sqrt{\|B\|_{2\to 2}\|B\|_F}, p^{\frac{\kappa}{4}} n^{\frac{d}{2}-\frac{\kappa}{4}} \|B\|_{2\to 2}\Big\}.$$

Since this is an upper bound on the $L_p$ norm of $\|BX\|_2 - \|B\|_F$, Lemma 2.22 implies

$$\mathbb{P}\left(|\|BX\|_2 - \|B\|_F| > t\right) \leq e^2 \exp\left(-C_2(d) \min_{\kappa\in[2d]} \beta_\kappa\right)$$

where

$$\begin{aligned}\beta_\kappa := \max\Bigg\{ &\left(\frac{t}{n^{\frac{d-1}{2}}\|B\|_{2\to 2}}\right)^{\frac{2}{\kappa}}, \left(\frac{t\|B\|_F}{n^{d-\frac{\kappa}{2}}\|B\|_{2\to 2}^2}\right)^{\frac{2}{\kappa}}, \\ &\left(\frac{t}{n^{\frac{d-1}{4}}\sqrt{\|B\|_{2\to 2}\|B\|_F}}\right)^{\frac{4}{\kappa}}, \left(\frac{t}{n^{\frac{d}{2}-\frac{\kappa}{4}}\|B\|_{2\to 2}}\right)^{\frac{4}{\kappa}}\Bigg\}. \end{aligned} \tag{2.29}$$

Now, for each of multiple different ranges of $t$, we select one of the four terms in (2.29).

**Step 3: Bound for $t \leq n^{\frac{d}{2}}\|B\|_{2\to 2}$**

For $\kappa = 1$, we obtain using the first term in (2.29), $\beta_1 \geq \left(t/(n^{\frac{d-1}{2}}\|B\|_{2\to 2})\right)^2$.

For $\kappa \geq 2$, we can use the fourth term in (2.29) to show the same bound because

$$\begin{aligned}\beta_\kappa \geq &\left(\frac{t}{n^{\frac{d}{2}-\frac{\kappa}{4}}\|B\|_{2\to 2}}\right)^{\frac{4}{\kappa}} = n\left(\frac{t}{n^{\frac{d}{2}}\|B\|_{2\to 2}}\right)^{\frac{4}{\kappa}} \\ \geq & n\left(\frac{t}{n^{\frac{d}{2}}\|B\|_{2\to 2}}\right)^{2} = \frac{t^2}{n^{d-1}\|B\|_{2\to 2}^2}.\end{aligned}$$

This implies that

$$\mathbb{P}\left(|\|BX\|_2 - \|B\|_F| > t\right) \leq e^2 \exp\left(-C_2(d)\frac{t^2}{n^{d-1}\|B\|_{2\to 2}^2}\right).$$

**Step 5: Bound for $t \geq n^{\frac{d}{2}}\|B\|_{2\to 2}$**

For all $\kappa \in [2d]$, using the fourth term in (2.29) yields

$$\begin{aligned}\beta_\kappa \geq &\left(\frac{t}{n^{\frac{d}{2}-\frac{\kappa}{4}}\|B\|_{2\to 2}}\right)^{\frac{4}{\kappa}} = n\left(\frac{t}{n^{\frac{d}{2}}\|B\|_{2\to 2}}\right)^{\frac{4}{\kappa}} \\ \geq & n\left(\frac{t}{n^{\frac{d}{2}}\|B\|_{2\to 2}}\right)^{\frac{4}{2d}} = \left(\frac{t}{\|B\|_{2\to 2}}\right)^{\frac{2}{d}},\end{aligned}$$

such that

$$\mathbb{P}\left(|\|BX\|_2 - \|B\|_F| > t\right) \leq e^2 \exp\left(-C_2(d)\left(\frac{t}{\|B\|_{2\to 2}}\right)^{\frac{2}{d}}\right).$$

**Step 6: Bound for $n^{\frac{d-1}{4}}\|B\|_{2\to 2} \leq t \leq n^{\frac{d-1}{4}}\|B\|_F$**

Using the third term in (2.29), we obtain that

$$\beta_\kappa \geq \left(\frac{t^2}{n^{\frac{d-1}{2}}\|B\|_{2\to 2}\|B\|_F}\right)^{\frac{2}{\kappa}} \geq \left(\frac{tn^{\frac{d-1}{4}}\|B\|_{2\to 2}}{n^{\frac{d-1}{2}}\|B\|_{2\to 2}\|B\|_F}\right)^{\frac{2}{\kappa}}$$

$$= \left( \frac{t}{n^{\frac{d-1}{4}} \|B\|_F} \right)^{\frac{2}{\kappa}} \geq \frac{t^2}{n^{\frac{d-1}{2}} \|B\|_F^2},$$

implying

$$\mathbb{P} \left( |\|BX\|_2 - \|B\|_F| > t \right) \leq e^2 \exp \left( -C_2(d) \frac{t^2}{n^{\frac{d-1}{2}} \|B\|_F^2} \right).$$

□

## 2.4 Discussion

In total, for a chaos of the type

$$\sum_{i_1,\dots,i_{2d}=1}^{n} A_{i_1,\dots,i_d,i_{d+1},\dots,i_{2d}} X_{i_1}^{(1)} \dots X_{i_d}^{(d)} X_{i_{d+1}}^{(1)} \dots X_{i_{2d}}^{(d)},$$

we have shown moment bounds that are tight (up to dependence on $d$) for the Gaussian case. Along with this, we have also shown a specific decoupling inequality for the above expression and improved moment and tail bounds for $\|B(X^{(1)} \otimes \cdots \otimes X^{(d)})\|_2$.

One particular application of this is the arbitrary order case of the Johnson-Lindenstrauss embeddings studied in Section 1. In Section 2.5, we complete this proof by making use of the decoupling Theorem 2.11, more specifically the Rademacher case (Remark 2.12) in which the terms significantly simplify.

## 2.5 Proof of the General Case of Section 1

### 2.5.1 Overview

In this Section 2.5, we prove the general version of Theorem 2.9 for arbitrary $d \geq 1$. This proof is a generalization of the one in Section 1.5 and follows an analogous outline.

Presenting the general case for arbitrary order arrays yields some notational difficulties which is why we present this proof using the special notation for indices and arrays which we introduced in Section 2.1.4. We will also use the decoupling Theorem 2.11. Using this particular notation in the subsequent Section 2.5.2, we will adjust Theorem 2.11 to this situation and show Theorem 2.35 that generalizes Lemma 1.11 that we used for the case $d = 2$. Then Section 2.5.3 presents the main part of the proof analogously to Section 1.5.3 for the case $d = 2$. We will generalize the distribution of the entries $\mathbf{x} = \mathbf{x}^{(\emptyset)} + \mathbf{x}^{(\{1\})} + \mathbf{x}^{(\{2\})} + \mathbf{x}^{(\{1,2\})}$ from the simplified case to a general $\mathbf{x} = \sum_{S \subset [d]} \mathbf{x}^{(S)}$ where the sum runs over all subsets $S \subset [d]$. Again, using a general tensor norm bound ((2.35) which is analogous to (1.13)), we can complete the proof. In the last Section 2.5.4, we complete the proof by showing the remaining inequality (2.35) for which we use the same Lemma 1.12 as for the previous simplified case.

### 2.5.2 Decoupling

In this first step, we establish a generalization of Lemma 1.11 by using the decoupling Theorem 2.11 along with Remark 2.12 and by taking into account that we will consider higher order chaos expressions for arrays $\mathbf{A} \in \mathbb{R}^{\mathbf{n}^{\times 2}}$ such that

$$A_{\mathbf{i}+\mathbf{i}'} = B_{\mathbf{i}+\mathbf{i}'} x_{\mathbf{i}} x_{\mathbf{i}'}$$

for all $\mathbf{i}, \mathbf{i}' \in \mathbf{J}^{\mathbf{n}}$, where $\mathbf{B} \in \mathbb{R}^{\mathbf{n}^{\times 2}}$ and $\mathbf{x} \in \mathbb{R}^{\mathbf{n}}$. Specifically, we need to obtain norm bounds that hold for one particular choice of $\mathbf{B}$ but all $\mathbf{x} \in \mathbb{R}^{\mathbf{n}}$. This leads to the following statement.

**Theorem 2.35.** *Let $\mathbf{B} \in \mathbb{R}^{\mathbf{n}^{\times 2}}$ and $\xi^{(1)} \in \{\pm 1\}^{n_1}, \ldots, \xi^{(d)} \in \{\pm 1\}^{n_d}$ be independent Rademacher vectors. Define $\boldsymbol{\xi} \in \mathbb{R}^{\mathbf{n}}$ by*

$$\xi_{\mathbf{i}} = \prod_{l=1}^{d} \xi_{\mathbf{i}_l}^{(l)}.$$

*and let $\bar{\boldsymbol{\xi}}$ be an independent copy of $\boldsymbol{\xi}$. Let $p \geq 1$.*

*Assume that for all $\mathbf{x} \in \mathbb{R}^{\mathbf{n}}$, $\|\mathbf{x}\|_2 = 1$, it holds that*

$$\left\| \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i} \dot{+} \mathbf{i}'} x_{\mathbf{i}} x_{\mathbf{i}'} \xi_{\mathbf{i}} \bar{\xi}_{\mathbf{i}'} \right\|_{L_p} \leq \gamma_p.$$

*Then also for all $\mathbf{x} \in \mathbb{R}^{\mathbf{n}}$, $\|\mathbf{x}\|_2 = 1$, it holds that*

$$\left\| \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i} \dot{+} \mathbf{i}'} x_{\mathbf{i}} x_{\mathbf{i}'} \xi_{\mathbf{i}} \xi_{\mathbf{i}'} \right\|_{L_p} \leq 5^d \gamma_p.$$

*Proof.* By Theorem Theorem 2.11 and Remark 2.12,

$$\left\| \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i} \dot{+} \mathbf{i}'} x_{\mathbf{i}} x_{\mathbf{i}'} \xi_{\mathbf{i}} \xi_{\mathbf{i}'} - \mathbb{E} \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i} \dot{+} \mathbf{i}'} x_{\mathbf{i}} x_{\mathbf{i}'} \xi_{\mathbf{i}} \xi_{\mathbf{i}'} \right\|_{L_p} \leq \sum_{\substack{I \subset [d] \\ I \neq [d]}} 4^{d-|I|} \|\bar{S}_I\|_{L_p} \tag{2.30}$$

for

$$\bar{S}_I := \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(I)} \sum_{\mathbf{j},\mathbf{j}' \in \mathbf{J}^{\mathbf{n}}(I^c)} B_{(\mathbf{i} \dot{\times} \mathbf{j}) \dot{+} (\mathbf{i} \dot{\times} \mathbf{j}')} x_{\mathbf{i} \dot{\times} \mathbf{j}} x_{\mathbf{i} \dot{\times} \mathbf{j}'} \prod_{l \in I^c} \xi_{\mathbf{j}_l}^{(l)} \bar{\xi}_{\mathbf{j}'_l}^{(l)}.$$

Furthermore, by the definition of Rademacher vectors, for any $\mathbf{i}, \mathbf{i}' \in \mathbf{J}^{\mathbf{n}}$,

$$\mathbb{E}[\xi_{\mathbf{i}} \xi_{\mathbf{i}'}] = \begin{cases} 1 & \text{if } \mathbf{i} = \mathbf{i}' \\ 0 & \text{otherwise,} \end{cases}$$

such that

$$\mathbb{E} \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i} \dot{+} \mathbf{i}'} x_{\mathbf{i}} x_{\mathbf{i}'} \xi_{\mathbf{i}} \xi_{\mathbf{i}'} = \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i} \dot{+} \mathbf{i}} x_{\mathbf{i}} x_{\mathbf{i}} = \bar{S}_{[d]}.$$

So by the triangle inequality and (2.30),

$$\left\| \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i} \dot{+} \mathbf{i}'} x_{\mathbf{i}} x_{\mathbf{i}'} \xi_{\mathbf{i}} \xi_{\mathbf{i}'} \right\|_{L_p} \leq \sum_{I \subset [d]} 4^{d-|I|} \|\bar{S}_I\|_{L_p}. \tag{2.31}$$

Now fix $I \subset [d]$ and for each $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(I)$, define

$$x_{\mathbf{i}' \dot{\times} \mathbf{j}}^{(\mathbf{i})} = \begin{cases} x_{\mathbf{i} \dot{\times} \mathbf{j}} & \text{if } \mathbf{i}' = \mathbf{i} \\ 0 & \text{otherwise.} \end{cases}$$

for $\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}(I)$, $\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(I^c)$.

This gives us

$$\begin{aligned} \bar{S}_I &= \sum_{\bar{\mathbf{i}} \in \mathbf{J}^{\mathbf{n}}(I)} \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}(I)} \sum_{\mathbf{j},\mathbf{j}' \in \mathbf{J}^{\mathbf{n}}(I^c)} B_{(\mathbf{i} \dot{\times} \mathbf{j}) \dot{+} (\mathbf{i} \dot{\times} \mathbf{j}')} x_{\mathbf{i} \dot{\times} \mathbf{j}}^{(\bar{\mathbf{i}})} x_{\mathbf{i}' \dot{\times} \mathbf{j}'}^{(\bar{\mathbf{i}})} \prod_{l \in I^c} \xi_{\mathbf{j}_l}^{(l)} \bar{\xi}_{\mathbf{j}'_l}^{(l)} \\ &= \sum_{\bar{\mathbf{i}} \in \mathbf{J}^{\mathbf{n}}(I)} \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i} \dot{+} \mathbf{i}'} x_{\mathbf{i}}^{(\bar{\mathbf{i}})} x_{\mathbf{i}'}^{(\bar{\mathbf{i}})} \prod_{l \in I^c} \xi_{\mathbf{i}_l}^{(l)} \bar{\xi}_{\mathbf{i}'_l}^{(l)} \end{aligned}$$

Note that each summand is non-zero only if $\mathbf{i}_I = \mathbf{i}'_I = \bar{\mathbf{i}}$ and in that case

$$\prod_{l\in I^c} \xi^{(l)}_{\mathbf{i}_l}\bar{\xi}^{(l)}_{\mathbf{i}'_l} \cdot 1 = \prod_{l\in I^c} \xi^{(l)}_{\mathbf{i}_l}\bar{\xi}^{(l)}_{\mathbf{i}'_l} \cdot \left(\prod_{l\in I} \xi^{(l)}_{\bar{\mathbf{i}}_l}\bar{\xi}^{(l)}_{\bar{\mathbf{i}}_l}\right)^2 = \xi_{\mathbf{i}}\bar{\xi}_{\mathbf{i}'} \cdot \prod_{l\in I} \xi^{(l)}_{\bar{\mathbf{i}}_l}\bar{\xi}^{(l)}_{\bar{\mathbf{i}}_l}.$$

So this yields

$$\bar{S}_I = \sum_{\bar{\mathbf{i}}\in\mathbf{J}^{\mathbf{n}}(I)} \left(\prod_{l\in I} \xi^{(l)}_{\bar{\mathbf{i}}_l}\bar{\xi}^{(l)}_{\bar{\mathbf{i}}_l}\right) \sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}} B_{\mathbf{i}\dot{+}\mathbf{i}'} x^{(\bar{\mathbf{i}})}_{\mathbf{i}} x^{(\bar{\mathbf{i}})}_{\mathbf{i}'} \xi_{\mathbf{i}}\bar{\xi}_{\mathbf{i}'}.$$

Because all the Rademacher variables are ±1, it holds that $\left|\prod_{l\in I} \xi^{(l)}_{\bar{\mathbf{i}}_l}\bar{\xi}^{(l)}_{\bar{\mathbf{i}}_l}\right| = 1$. Using this and the triangle inequality, we obtain

$$\begin{aligned} \|\bar{S}_I\|_{L_p} &\leq \sum_{\bar{\mathbf{i}}\in\mathbf{J}^{\mathbf{n}}(I)} \left\| \left|\prod_{l\in I} \xi^{(l)}_{\bar{\mathbf{i}}_l}\bar{\xi}^{(l)}_{\bar{\mathbf{i}}_l}\right| \left| \sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}} B_{\mathbf{i}\dot{+}\mathbf{i}'} x^{(\bar{\mathbf{i}})}_{\mathbf{i}} x^{(\bar{\mathbf{i}})}_{\mathbf{i}'} \xi_{\mathbf{i}}\xi_{\mathbf{i}'} \right| \right\|_{L_p} \\ &= \sum_{\bar{\mathbf{i}}\in\mathbf{J}^{\mathbf{n}}(I)} \left\| \sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}} B_{\mathbf{i}\dot{+}\mathbf{i}'} x^{(\bar{\mathbf{i}})}_{\mathbf{i}} x^{(\bar{\mathbf{i}})}_{\mathbf{i}'} \xi_{\mathbf{i}}\xi_{\mathbf{i}'} \right\|_{L_p} =: \sum_{\bar{\mathbf{i}}\in\mathbf{J}^{\mathbf{n}}(I)} \|\bar{S}_{I,\bar{\mathbf{i}}}\|_{L_p}. \end{aligned}$$

If $\|\mathbf{x}^{(\bar{\mathbf{i}})}\|_2 = 0$, then $\|\bar{S}_{I,\bar{\mathbf{i}}}\|_{L_p} = 0 = \|\mathbf{x}^{(\bar{\mathbf{i}})}\|_2^2 \cdot \gamma_p$. Otherwise, the array with entries $\frac{x^{(\bar{\mathbf{i}})}_{\mathbf{i}}}{\|\mathbf{x}^{(\bar{\mathbf{i}})}\|_2}$ has a $\|\cdot\|_2$ norm of 1 and by the assumption of the theorem

$$\|\bar{S}_{I,\bar{\mathbf{i}}}\|_{L_p} \leq \|\mathbf{x}^{(\bar{\mathbf{i}})}\|_2^2\gamma_p$$

which then holds in all cases and this implies

$$\|\bar{S}_I\|_{L_p} \leq \sum_{\bar{\mathbf{i}}\in\mathbf{J}^{\mathbf{n}}(I)} \|\mathbf{x}^{(\bar{\mathbf{i}})}\|_2^2\gamma_p = \|\mathbf{x}\|_2^2\gamma_p = \gamma_p.$$

Substituting into (2.31) yields

$$\left\| \sum_{\mathbf{i},\mathbf{i}'\in\mathbf{J}^{\mathbf{n}}} B_{\mathbf{i}\dot{+}\mathbf{i}'} x_{\mathbf{i}} x_{\mathbf{i}'} \xi_{\mathbf{i}}\xi_{\mathbf{i}'} \right\|_{L_p} \leq \sum_{I\subset[d]} 4^{d-|I|}\gamma_p = \sum_{k=0}^{d} \sum_{\substack{I\subset[d]\\|I|=k}} 4^{d-k}\gamma_p.$$

Noting that there are precisely $\binom{d}{k}$ sets $I \subset [d]$ with $|I| = k$, such that this is

$$\gamma_p \sum_{k=0}^{d} \binom{d}{k} 4^{d-k} \cdot 1^k = \gamma_p(4+1)^d = 5^d\gamma_p,$$

which is the desired upper bound. □

### 2.5.3 Proof of Theorem 2.9

Since every vector $x \in \mathbb{R}^N$ can be rearranged to an array in $\mathbb{R}^{\mathbf{n}}$, it is sufficient to prove

$$\mathbb{P}\left(\left|\|A\operatorname{vec}(\mathbf{x})\|_2^2 - \|\mathbf{x}\|_2^2\right| > \epsilon\right) \leq \eta$$

for any $\mathbf{x} \in \mathbb{R}^{\mathbf{n}}$ with $\|\mathbf{x}\|_2 = 1$. So take an arbitrary such $\mathbf{x}$.

**Splitting up x into $\mathbf{x}^{(S)}$:**

For every subset $S \subset [d]$, define the set $\mathbf{K}(S) \subset \mathbf{J}^{\mathbf{n}}$ of indices in the following way: For each $\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(S^c)$, choose $s^{|S|}$ indices $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(S)$ with the largest $|x_{\mathbf{i} \dot{\times} \mathbf{j}}|$ and $\mathbf{K}(S)$ is the set of all $\mathbf{i} \dot{\times} \mathbf{j}$ obtained in this way.

Now for every index $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}$, choose $S(\mathbf{i})$ to be a set $S \subset [d]$ of largest cardinality such that $\mathbf{i} \in \mathbf{K}(S)$. Since $\mathbf{K}(\emptyset) = \mathbf{J}^{\mathbf{n}}$, such an $S$ always exists.

Then for any set $S \subset [d]$, define $\mathbf{x}^{(S)} \in \mathbb{R}^{\mathbf{n}}$ such that for each $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}$,

$$x_{\mathbf{i}}^{(S)} := \begin{cases} x_{\mathbf{i}} & \text{if } S(\mathbf{i}) = S \\ 0 & \text{otherwise.} \end{cases}$$

Since for every index $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}$, we chose exactly one $S(\mathbf{i})$,

$$\mathbf{x} = \sum_{S \subset [d]} \mathbf{x}^{(S)}.$$

A direct consequence from these definitions is the following lemma.

**Lemma 2.36.** *Let $S \subset [d]$. For any index $\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(S^c)$, there can be at most $s^{|S|}$ different indices $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(S)$ such that*

$$x_{\mathbf{i} \dot{\times} \mathbf{j}}^{(S)} \neq 0.$$

*Proof.* Fix $\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(S^c)$.

If $x_{\mathbf{i} \dot{\times} \mathbf{j}}^{(S)} \neq 0$ holds for $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(S)$, then $S(\mathbf{i} \dot{\times} \mathbf{j}) = S$, implying that $\mathbf{i} \dot{\times} \mathbf{j} \in \mathbf{K}(S)$. By definition of $\mathbf{K}(S)$ however, this can only be the case for $s^{|S|}$ different indices $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(S)$. □

**Lemma 2.37.** *For any $S, T \subset [d]$ with $|S| < |T|$ and any $\mathbf{k} \in \mathbf{J}^{\mathbf{n}}(T^c)$,*

$$\max_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)} |x_{\mathbf{j} \dot{\times} \mathbf{k}}^{(S)}|^2 \leq \frac{1}{s^{|T|}} \sum_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)} |x_{\mathbf{j} \dot{\times} \mathbf{k}}|^2.$$

*Proof.* Let $S, T \subset [d]$, $|S| < |T|$ and $\mathbf{k} \in \mathbf{J}^{\mathbf{n}}(T^c)$. Choose $\mathbf{j}_0 \in \mathbf{J}^{\mathbf{n}}(T)$ such that $|x_{\mathbf{j}_0 \dot{\times} \mathbf{k}}^{(S)}|$ is maximal.

If $|x_{\mathbf{j}_0 \dot{\times} \mathbf{k}}^{(S)}| = 0$, then the claim is fulfilled. Otherwise we know that $S(\mathbf{j}_0 \dot{\times} \mathbf{k}) = S$. Especially, this implies that $\mathbf{j}_0 \dot{\times} \mathbf{k} \notin \mathbf{K}(T)$ since $|T| > |S|$. By the definition of $\mathbf{K}(T)$, there is a set $\bar{\mathbf{J}} \subset \mathbf{J}^{\mathbf{n}}(T)$ of $s^{|T|}$ indices such that for all $\mathbf{j} \in \bar{\mathbf{J}}$,

$$|x_{\mathbf{j} \dot{\times} \mathbf{k}}| \geq |x_{\mathbf{j}_0 \dot{\times} \mathbf{k}}|.$$

Assuming that $|x_{\mathbf{j}_0 \dot{\times} \mathbf{k}}^{(S)}|^2 > \frac{1}{s^{|T|}} \sum_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)} |x_{\mathbf{j} \dot{\times} \mathbf{k}}|^2$ implies

$$\begin{aligned} \sum_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)} |x_{\mathbf{j} \dot{\times} \mathbf{k}}|^2 \geq & \sum_{\mathbf{j} \in \bar{\mathbf{J}}} |x_{\mathbf{j} \dot{\times} \mathbf{k}}|^2 \geq \sum_{\mathbf{j} \in \bar{\mathbf{J}}} |x_{\mathbf{j}_0 \dot{\times} \mathbf{k}}|^2 \\ = & |\bar{\mathbf{J}}| |x_{\mathbf{j}_0 \dot{\times} \mathbf{k}}|^2 \geq s^{|T|} |x_{\mathbf{j}_0 \dot{\times} \mathbf{k}}^{(S)}|^2 \\ > & s^{|T|} \frac{1}{s^{|T|}} \sum_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)} |x_{\mathbf{j} \dot{\times} \mathbf{k}}|^2 = \sum_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)} |x_{\mathbf{j} \dot{\times} \mathbf{k}}|^2. \end{aligned}$$

This is a contradiction which completes the proof. □

**Lemma 2.38.** *Let $S, T \subset [d]$ and $S \cap T = \emptyset$. Then for any index $\mathbf{k} \in \mathbf{J}^{\mathbf{n}}([d] \backslash (S \cup T))$,*

$$\max_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)} \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(S)} |x_{\mathbf{i} \dot{\times} \mathbf{j} \dot{\times} \mathbf{k}}^{(S)}|^2 \leq \frac{1}{s^{|T|}} \sum_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)} \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(S)} |x_{\mathbf{i} \dot{\times} \mathbf{j} \dot{\times} \mathbf{k}}|^2.$$

*Proof.* If $T = \emptyset$, then $s^{|T|} = 1$ and $\mathbf{J}^{\mathbf{n}}(T)$ has exactly one element such that the claim holds since $|x^{(S)}_{\mathbf{i}\dot{\times}\mathbf{j}\dot{\times}\mathbf{k}}| \leq |x_{\mathbf{i}\dot{\times}\mathbf{j}\dot{\times}\mathbf{k}}|$ for any indices $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(S)$, $\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)$, $\mathbf{k} \in \mathbf{J}^{\mathbf{n}}([d]\backslash(S \cup T))$. So we can assume that $T \neq \emptyset$.

Then we can apply Lemma 2.37 to the sets $S$ and $S \cup T$. Since $S$ and $T$ are disjoint and $T \neq \emptyset$, $|S \cup T| > |S|$. The lemma yields that for any $\mathbf{k} \in \mathbf{J}^{\mathbf{n}}([d]\backslash(S \cup T))$,

$$\max_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(S \cup T)} |x^{(S)}_{\mathbf{j}\dot{\times}\mathbf{k}}|^2 \leq \frac{1}{s^{|S \cup T|}} \sum_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(S \cup T)} |x_{\mathbf{j}\dot{\times}\mathbf{k}}|^2.$$

We can rewrite this as

$$\max_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(S)} \max_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)} |x^{(S)}_{\mathbf{i}\dot{\times}\mathbf{j}\dot{\times}\mathbf{k}}|^2 \leq \frac{1}{s^{|S|+|T|}} \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(S)} \sum_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)} |x_{\mathbf{i}\dot{\times}\mathbf{j}\dot{\times}\mathbf{k}}|^2. \tag{2.32}$$

By Lemma 2.36, for fixed $\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)$ and $\mathbf{k} \in \mathbf{J}^{\mathbf{n}}([d]\backslash(S \cup T))$, there are at most $2^{|S|}$ indices $\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(S)$ such that $x^{(S)}_{\mathbf{i}\dot{\times}\mathbf{j}\dot{\times}\mathbf{k}} \neq 0$. Thus we obtain

$$\max_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)} \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(S)} |x^{(S)}_{\mathbf{i}\dot{\times}\mathbf{j}\dot{\times}\mathbf{k}}|^2 \leq \max_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)} s^{|S|} \max_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(S)} |x^{(S)}_{\mathbf{i}\dot{\times}\mathbf{j}\dot{\times}\mathbf{k}}|^2$$

Combining this with (2.32) yields

$$\begin{aligned} \max_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)} \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(S)} |x^{(S)}_{\mathbf{i}\dot{\times}\mathbf{j}\dot{\times}\mathbf{k}}|^2 &\leq \frac{s^{|S|}}{s^{|S|+|T|}} \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(S)} \sum_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)} |x_{\mathbf{i}\dot{\times}\mathbf{j}\dot{\times}\mathbf{k}}|^2 \\ &= \frac{1}{s^{|T|}} \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}}(S)} \sum_{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(T)} |x_{\mathbf{i}\dot{\times}\mathbf{j}\dot{\times}\mathbf{k}}|^2. \end{aligned}$$

□

Let $N = n_1 n_2 \dots n_d$ and assume that the matrix $\Phi \in \mathbb{R}^{m \times N}$ has the $(4s^d, \delta)$-RIP. We regard the matrix $\Phi^*\Phi - Id_N \in \mathbb{R}^{N \times N}$ as an array $\mathbf{B}$ of order $2d$ with dimensions $\mathbf{n}^{\times 2} = (n_1, \dots, n_d, n_1, \dots, n_d)$ such that for all $\mathbf{i}, \mathbf{i}' \in \mathbf{J}^{\mathbf{n}}$,

$$B_{\mathbf{i}\dot{+}\mathbf{i}'} = \sum_{k=1}^{m} \Phi_{k,\mathcal{I}^{\mathbf{n}}(\mathbf{i})} \Phi_{k,\mathcal{I}^{\mathbf{n}}(\mathbf{i}')} - \mathbb{1}_{\mathbf{i}=\mathbf{i}'}. \tag{2.33}$$

Then for any arrays $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathbf{n}}$,

$$\langle \Phi \operatorname{vec}(\mathbf{x}), \Phi \operatorname{vec}(\mathbf{y}) \rangle - \langle \operatorname{vec}(\mathbf{x}), \operatorname{vec}(\mathbf{y}) \rangle = \sum_{\mathbf{i}, \mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i}\dot{+}\mathbf{i}'} x_{\mathbf{i}} y_{\mathbf{i}'}$$

Let $\boldsymbol{\xi} \in \mathbb{R}^{\mathbf{n}}$ be the Rademacher tensor of order $d$, i.e., for $\mathbf{j} \in \mathbf{J}^{\mathbf{n}}$,

$$\xi_{\mathbf{j}} = \xi^{(1)}_{\mathbf{j}_1} \dots \xi^{(d)}_{\mathbf{j}_d}$$

where $\xi^{(1)}, \dots, \xi^{(d)}$ are the independent Rademacher vectors from the assumption of Theorem 2.9. Let $\bar{\xi}^{(1)}, \dots, \bar{\xi}^{(d)}$ and $\bar{\boldsymbol{\xi}}$ be corresponding independent copies.

Consider the norm deviation represented by the chaos

$$\tilde{X} := \langle A \operatorname{vec}(\mathbf{x}), A \operatorname{vec}(\mathbf{x}) \rangle - \langle \operatorname{vec}(\mathbf{x}), \operatorname{vec}(\mathbf{x}) \rangle = \sum_{\mathbf{i}, \mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i}\dot{+}\mathbf{i}'} x_{\mathbf{i}} x_{\mathbf{i}'} \xi_{\mathbf{i}} \xi_{\mathbf{i}'}$$

and the corresponding decoupled chaos

$$X := \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i} \dotplus \mathbf{i}'} x_{\mathbf{i}} x_{\mathbf{i}'} \xi_{\mathbf{i}} \bar{\xi}_{\mathbf{i}'}$$

Our goal is to bound the moments of $|X|$ which will also lead to bounds on the moments of $|\tilde{X}| = \big| \|A \operatorname{vec}(\mathbf{x})\|_2^2 - \|\mathbf{x}\|_2^2 \big|$ by the application of the decoupling Theorem 2.35. This in turn, will lead to the proof of Theorem 2.9.

Fix $S, T \subset [d]$. Define the sums

$$\begin{aligned} \tilde{X}^{(S,T)} &:= \langle A \operatorname{vec}(\mathbf{x}^{(S)}), A \operatorname{vec}(\mathbf{x}^{(T)}) \rangle - \langle \operatorname{vec}(\mathbf{x}^{(S)}), \operatorname{vec}(\mathbf{x}^{(T)}) \rangle \\ &= \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i} \dotplus \mathbf{i}'} \xi_{\mathbf{i}} \xi_{\mathbf{i}'} x_{\mathbf{i}}^{(S)} x_{\mathbf{i}'}^{(T)} \end{aligned}$$

and their decoupled counterparts

$$X^{(S,T)} := \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J}^{\mathbf{n}}} B_{\mathbf{i} \dotplus \mathbf{i}'} \xi_{\mathbf{i}} \bar{\xi}_{\mathbf{i}'} x_{\mathbf{i}}^{(S)} x_{\mathbf{i}'}^{(T)}$$

such that

$$X = \sum_{S,T \subset [d]} X^{(S,T)} \quad \text{and} \quad \tilde{X} = \sum_{S,T \subset [d]} \tilde{X}^{(S,T)}.$$

We obtain

$$X^{(S,T)} := \sum_{\substack{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(S^c), \mathbf{k} \in \mathbf{J}^{\mathbf{n}}(S) \\ \mathbf{j}' \in \mathbf{J}^{\mathbf{n}}(T^c), \mathbf{k}' \in \mathbf{J}^{\mathbf{n}}(T)}} B_{(\mathbf{j} \dot{\times} \mathbf{k}) \dotplus (\mathbf{j}' \dot{\times} \mathbf{k}')} \xi_{\mathbf{j} \dot{\times} \mathbf{k}} \bar{\xi}_{\mathbf{j}' \dot{\times} \mathbf{k}'} x_{\mathbf{j} \dot{\times} \mathbf{k}}^{(S)} x_{\mathbf{j}' \dot{\times} \mathbf{k}'}^{(T)}$$

Now for any set $\tilde{I} \subset [d]$, define the array $\boldsymbol{\xi}^{(\tilde{I})} \in \mathbb{R}^{\mathbf{n}}(\tilde{I})$ by

$$\xi_{\mathbf{j}}^{(\tilde{I})} = \prod_{l \in \tilde{I}} \xi_{\mathbf{j}_l}^{(l)}$$

for any $\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(\tilde{I})$ and analogously for $\bar{\boldsymbol{\xi}}$.

With this, we obtain

$$X^{(S,T)} = \sum_{\substack{\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(S^c), \mathbf{k} \in \mathbf{J}^{\mathbf{n}}(S) \\ \mathbf{j}' \in \mathbf{J}^{\mathbf{n}}(T^c), \mathbf{k}' \in \mathbf{J}^{\mathbf{n}}(T)}} B_{(\mathbf{j} \dot{\times} \mathbf{k}) \dotplus (\mathbf{j}' \dot{\times} \mathbf{k}')} \xi_{\mathbf{k}}^{(S)} \xi_{\mathbf{j}}^{(S^c)} \bar{\xi}_{\mathbf{k}'}^{(T)} \bar{\xi}_{\mathbf{j}'}^{(T^c)} x_{\mathbf{j} \dot{\times} \mathbf{k}}^{(S)} x_{\mathbf{j}' \dot{\times} \mathbf{k}'}^{(T)}. \tag{2.34}$$

Note that $\boldsymbol{\xi}^{(S)}$, $\boldsymbol{\xi}^{(S^c)}$, $\bar{\boldsymbol{\xi}}^{(T)}$, $\bar{\boldsymbol{\xi}}^{(T^c)}$ are independent. Condition on $\boldsymbol{\xi}^{(S)}$ and $\bar{\boldsymbol{\xi}}^{(T)}$ and treat (2.34) as a chaos of order $2d - |S| - |T|$ (depending on $\boldsymbol{\xi}^{(S^c)}$ and $\bar{\boldsymbol{\xi}}^{(T^c)}$) with corresponding index array $\mathbf{B}^{(S,T)}$ given by

$$B_{\mathbf{j} \dot{\times} \mathbf{j}'}^{(S,T)} = \sum_{\mathbf{k} \in \mathbf{J}^{\mathbf{n}}(S), \mathbf{k}' \in \mathbf{J}^{\mathbf{n}}(T)} B_{(\mathbf{j} \dot{\times} \mathbf{k}) \dotplus (\mathbf{j}' \dot{\times} \mathbf{k}')} \xi_{\mathbf{k}}^{(S)} \bar{\xi}_{\mathbf{k}'}^{(T)} x_{\mathbf{j} \dot{\times} \mathbf{k}}^{(S)} x_{\mathbf{j}' \dot{\times} \mathbf{k}'}^{(T)}$$

for $\mathbf{j} \in \mathbf{J}^{\mathbf{n}}(S^c)$, $\mathbf{j}' \in \mathbf{J}^{\mathbf{n}}(T^c)$. If $(S, T) \neq ([d], [d])$, then the chaos order is $\geq 1$ and we can apply the chaos norm bound Theorem 1.7 for which we need to control the tensor norms $\|\mathbf{B}^{(S,T)}\|_{I_1,\dots,I_\kappa}$. Note that any array with varying dimensions along the axes can be extended by 0 entries to an array with equal dimensions along all axes.

Let $1 \leq \kappa \leq 2d - |S| - |T|$ and $I_1, \dots, I_\kappa$ be a partition of the set $[2d] \backslash (S \cup (T + d))$. Let $\boldsymbol{\alpha}^{(l)} \in \mathbb{R}^{\mathbf{n}^{\times 2}}(I_l)$ and $\|\boldsymbol{\alpha}^{(l)}\|_2 = 1$ for $1 \leq l \leq \kappa$. In this notation, we obtain for the norm,

$$\|\mathbf{B}^{(S,T)}\|_{I_1,\dots,I_\kappa} = \sup_{\boldsymbol{\alpha}^{(1)},\dots,\boldsymbol{\alpha}^{(\kappa)}} \sum_{\mathbf{i} \in \mathbf{J}^{\mathbf{n}^{\times 2}}(I_1 \cup \dots \cup I_\kappa)} B_{\mathbf{i}}^{(S,T)} \alpha_{\mathbf{i}_{I_1}}^{(1)} \dots \alpha_{\mathbf{i}_{I_\kappa}}^{(\kappa)}$$

where the supremum is taken over all possible choices of the aforementioned arrays $\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(\kappa)}$.

We will show that for all $S, T \subset [d]$, $(S,T) \neq ([d],[d])$ and all partitions $I_1, \ldots, I_\kappa$ of $[2d] \backslash (S \cup (T+d))$ into non-empty disjoint sets,

$$\|\mathbf{B}^{(S,T)}\|_{I_1,\ldots,I_\kappa} \leq 4 \frac{\delta}{s^{\frac{\kappa}{2}}}. \tag{2.35}$$

The proof of this inequality is postponed to Subsection 2.5.4. At this point we complete the proof using that (2.35) holds.

With the moment bound (Theorem 1.7), this implies that for $(S,T) \neq ([d],[d])$ and $p \geq 2$,

$$\|X^{(S,T)}\|_{L_p} \leq C_1(d) \sum_{\kappa=1}^{2d} p^{\frac{\kappa}{2}} \frac{\delta}{s^{\frac{\kappa}{2}}}$$

with $C_1(d)$ depending only on $d$. For the remaining case $(S,T) = ([d],[d])$, observe that according to Lemma 2.36, there are only $s^d$ indices $\mathbf{i} \in \mathbf{J^n}$ such that $x_{\mathbf{i}}^{([d])} \neq 0$, i.e., $\mathbf{x}^{([d])}$ is $s^d$-sparse. We obtain

$$\begin{aligned} X^{([d],[d])} &= \sum_{\mathbf{k},\mathbf{k'} \in \mathbf{J^n}} B_{\mathbf{k} \dotplus \mathbf{k'}} \xi_{\mathbf{k}} \bar{\xi}_{\mathbf{k'}} x_{\mathbf{k}}^{([d])} x_{\mathbf{k'}}^{([d])} \\ &= \left\langle \operatorname{vec}(\mathbf{x}^{([d])}) \circ (\xi^{(1)} \otimes \cdots \otimes \xi^{(d)}), (\Phi^*\Phi - Id_N)(\operatorname{vec}(\mathbf{x}^{([d])}) \circ (\bar{\xi}^{(1)} \otimes \cdots \otimes \bar{\xi}^{(d)})) \right\rangle, \end{aligned}$$

where $\circ$ denotes the element-wise product. Since $\operatorname{vec}(\mathbf{x}^{([d])})$ is an $s^d$-sparse vector with norm $\|\operatorname{vec}(\mathbf{x}^{([d])})\|_2 \leq 1$, we can use Lemma 1.8 with the RIP of $\Phi$ to bound

$$|X^{([d],[d])}| \leq \delta.$$

So altogether, we obtain for a $C_2(d)$ depending only on $d$,

$$\|X\|_{L_p} \leq \sum_{S,T \subset [d]} \|X^{(S,T)}\|_{L_p} \leq C_2(d) \sum_{\kappa=0}^{2d} p^{\frac{\kappa}{2}} \frac{\delta}{s^{\frac{\kappa}{2}}}.$$

This moment bound holds for all $\|\mathbf{x}\|_2 = 1$ so by Theorem 2.35, we also obtain

$$\|\tilde{X}\|_{L_p} \leq C_3(d) \sum_{\kappa=0}^{2d} p^{\frac{\kappa}{2}} \frac{\delta}{s^{\frac{\kappa}{2}}},$$

where $C_3(d) > 0$ only depends on $d$. Using the particular choice $p = s \geq 2$, we obtain,

$$\|\tilde{X}\|_{L_s} \leq C_3(d) \sum_{\kappa=0}^{2d} s^{\frac{\kappa}{2}} \frac{\delta}{s^{\frac{\kappa}{2}}} = \delta(2d+1)C_3(d).$$

Finally, applying Markov's inequality implies for $\delta \leq \frac{1}{(2d+1)eC_3(d)}\epsilon$,

$$\begin{aligned} \mathbb{P}\left(\left|\|A\operatorname{vec}(\mathbf{x})\|_2^2 - \|\mathbf{x}\|_2^2\right| > \epsilon\right) = \mathbb{P}\left(|X| > \epsilon\right) &\leq \left(\frac{\|X\|_{L_s}}{\epsilon}\right)^s \\ &\leq \left(\frac{\delta(2d+1)C_3(d)}{\epsilon}\right)^s \leq e^{-s} \leq \eta \end{aligned}$$

for $s \geq \log \frac{1}{\eta}$.

### 2.5.4 Bounding the Tensor Norms

In this subsection, we complete the proof by showing the remaining inequality (2.35) for all $(S,T) \neq ([d],[d])$ and partitions $I_1, \dots, I_\kappa$ of $[2d]\backslash(S \cup (T+d))$ into non-empty disjoint sets.

The partition sets $I_1, \dots, I_\kappa$ can contain elements of $[d]$ and of $[2d]\backslash[d]$. Analogously to Subsection 1.5.4, we separate them by whether they intersect only $[d]$, only $[2d]\backslash[d]$ or both of them. In this sense, we define

$$\bar{I} = \bigcup_{j \in [\kappa]: I_j \subset [d]} I_j, \qquad \bar{I}' = \bigcup_{j \in [\kappa]: I_j \subset ([2d]\backslash[d])} I_j$$

$$\bar{\bar{J}} = (I_1 \cup \dots \cup I_\kappa)\backslash(\bar{I} \cup \bar{I}'), \qquad \bar{J} = \bar{\bar{J}} \cap [d], \qquad \bar{J}' = \bar{\bar{J}} \cap ([2d]\backslash[d]).$$

Joining partition sets does not increase the corresponding partition norm by Lemma 1.6. Note that this lemma can also applied to arrays which do not have the same dimension $n$ along all the axes since any array can be turned into one of those by extending it by zero entries. Thus we obtain,

$$\begin{aligned}\|\mathbf{B}^{(S,T)}\|_{I_1,\dots,I_\kappa} &\leq \|\mathbf{B}^{(S,T)}\|_{\bar{I},\bar{I}',\bar{\bar{J}}} \\ &= \sup_{\boldsymbol{\alpha}^{(1)},\boldsymbol{\alpha}^{(2)},\boldsymbol{\alpha}^{(3)}} \sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J^n}} B_{\mathbf{i},\mathbf{i}'} x^{(S)}_{\mathbf{i}} x^{(T)}_{\mathbf{i}'} \xi^{(S)}_{\mathbf{i}_S} \bar{\xi}^{(T)}_{\mathbf{i}'_T} \alpha^{(1)}_{(\mathbf{i}\dot{+}\mathbf{i}')_{\bar{I}}} \alpha^{(2)}_{(\mathbf{i}\dot{+}\mathbf{i}')_{\bar{I}'}} \alpha^{(3)}_{(\mathbf{i}\dot{+}\mathbf{i}')_{\bar{\bar{J}}}},\end{aligned}$$

where the supremum is taken over all $\boldsymbol{\alpha}^{(1)} \in \mathbb{R}^{\mathbf{n}^{\times 2}}(\bar{I})$, $\boldsymbol{\alpha}^{(2)} \in \mathbb{R}^{\mathbf{n}^{\times 2}}(\bar{I}')$, $\boldsymbol{\alpha}^{(3)} \in \mathbb{R}^{\mathbf{n}^{\times 2}}(\bar{\bar{J}})$ with $\|\boldsymbol{\alpha}^{(1)}\|_2 = \|\boldsymbol{\alpha}^{(2)}\|_2 = \|\boldsymbol{\alpha}^{(3)}\|_2 = 1$.

$\bar{J}, \bar{I}, S$ forms a partition of $[d]$ and thus in the above expression for the tensor norm, every index $\mathbf{i} \in \mathbf{J^n}$ in the sum has a unique decomposition $\mathbf{i} = \mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}$ where $\mathbf{j} \in \mathbf{J^n}(\bar{J})$, $\mathbf{k} \in \mathbf{J^n}(\bar{I})$ and $\mathbf{l} \in \mathbf{J^n}(S)$. We can do an analogous decomposition of $\mathbf{i}' \in \mathbf{J^n}$ corresponding to the partition $\bar{J}' - d, \bar{I}' - d, T$ of $[d]$ and then rewrite the sum $\sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J^n}}$ as a sum over six partial indices which leads to

$$\begin{aligned}&\sum_{\mathbf{i},\mathbf{i}' \in \mathbf{J^n}} B_{\mathbf{i},\mathbf{i}'} x^{(S)}_{\mathbf{i}} x^{(T)}_{\mathbf{i}'} \xi^{(S)}_{\mathbf{i}_S} \bar{\xi}^{(T)}_{\mathbf{i}'_T} \alpha^{(1)}_{(\mathbf{i}\dot{+}\mathbf{i}')_{\bar{I}}} \alpha^{(2)}_{(\mathbf{i}\dot{+}\mathbf{i}')_{\bar{I}'}} \alpha^{(3)}_{(\mathbf{i}\dot{+}\mathbf{i}')_{\bar{\bar{J}}}} \\ &= \sum_{\substack{\mathbf{j}\in\mathbf{J^n}(\bar{J}),\mathbf{j}'\in\mathbf{J^n}(\bar{J}'-d) \\ \mathbf{k}\in\mathbf{J^n}(\bar{I}),\mathbf{k}'\in\mathbf{J^n}(\bar{I}'-d) \\ \mathbf{l}\in\mathbf{J^n}(S),\mathbf{l}'\in\mathbf{J^n}(T)}} B_{(\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l})\dot{+}(\mathbf{j}'\dot{\times}\mathbf{k}'\dot{\times}\mathbf{l}')} x^{(S)}_{\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}} x^{(T)}_{\mathbf{j}'\dot{\times}\mathbf{k}'\dot{\times}\mathbf{l}'} \xi^{(S)}_{\mathbf{l}} \bar{\xi}^{(T)}_{\mathbf{l}'} \alpha^{(1)}_{\mathbf{k}} \alpha^{(2)}_{\mathbf{k}'} \alpha^{(3)}_{\mathbf{j}\dot{+}\mathbf{j}'} \qquad (2.36)\end{aligned}$$

In order to apply Lemma 1.12 to this expression, we will perform the following rearrangements:

$$\begin{aligned}x^{(S)}_{\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}} \xi^{(S)}_{\mathbf{l}} &\to X_{j,k,l} & x^{(T)}_{\mathbf{j}'\dot{\times}\mathbf{k}'\dot{\times}\mathbf{l}'} \xi^{(T)}_{\mathbf{l}'} &\to Y_{j',k',l'} & & \\ x_{\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}} &\to \bar{X}_{j,k,l} & x_{\mathbf{j}'\dot{\times}\mathbf{k}'\dot{\times}\mathbf{l}'} &\to \bar{Y}_{j',k',l'} & & \\ \alpha^{(1)}_{\mathbf{k}} &\to \alpha_k & \alpha^{(2)}_{\mathbf{k}'} &\to \alpha'_{k'} & \alpha^{(3)}_{\mathbf{j}\dot{+}\mathbf{j}'} &\to \Gamma_{j,j'}.\end{aligned}$$

For the precise definitions of the above arrays, we first define the dimensions

$$\bar{n}_1 := \prod_{r \in \bar{J}} n_r \qquad \bar{n}_2 := \prod_{r \in \bar{I}} n_r \qquad \bar{n}_3 := \prod_{r \in S} n_r$$

$$\bar{n}'_1 := \prod_{r \in \bar{J}'-d} n_r \qquad \bar{n}'_2 := \prod_{r \in \bar{I}'-d} n_r \qquad \bar{n}'_3 := \prod_{r \in T} n_r$$

and

$$s_1 := s^{|\bar{J}|} \qquad s_2 := s^{|\bar{I}|} \qquad s_3 := s^{|S|}$$

$$s_1' := s^{|\bar{J}'|} \qquad s_2' := s^{|\bar{I}'|} \qquad s_3' := s^{|T|}.$$

Since $S, \bar{I}, \bar{J}$ and $T, \bar{I}' - d, \bar{J}' - d$ are both partitions of $[d]$, we obtain for these dimensions

$$\bar{n}_1\bar{n}_2\bar{n}_3 = \bar{n}_1'\bar{n}_2'\bar{n}_3' = \prod_{r\in[d]} n_r = N \qquad s_1s_2s_3 = s_1's_2's_3' = s^d,$$

and the matrix $\Phi \in \mathbb{R}^{m\times N}$ satisfies the $(4s^d, \delta)$-RIP which complies with the assumption of Lemma 1.12.

With these dimensions, also $|\mathbf{J^n}(\bar{J})| = \bar{n}_1$ and we can take a bijective $\mathcal{I}_{\bar{J}} : \mathbf{J^n}(\bar{J}) \to [\bar{n}_1]$ that maps any array index $\mathbf{j} \in \mathbf{J^n}(\bar{J})$ to an integer $j \in [\bar{n}_1]$. Analogously, we define all the bijective maps

$$\begin{aligned} &\mathcal{I}_{\bar{J}} : \mathbf{J^n}(\bar{J}) \to [\bar{n}_1], && \mathcal{I}_{\bar{I}} : \mathbf{J^n}(\bar{I}) \to [\bar{n}_2], && \mathcal{I}_S : \mathbf{J^n}(S) \to [\bar{n}_3] \\ &\mathcal{I}_{\bar{J}'} : \mathbf{J^n}(\bar{J}' - d) \to [\bar{n}_1'], && \mathcal{I}_{\bar{I}'} : \mathbf{J^n}(\bar{I}' - d) \to [\bar{n}_2'], && \mathcal{I}_T : \mathbf{J^n}(T) \to [\bar{n}_3'] \end{aligned}$$

and then define arrays $\mathbf{X}, \bar{\mathbf{X}} \in \mathbb{R}^{\bar{n}_1\times\bar{n}_2\times\bar{n}_3}$ such that

$$X_{\mathcal{I}_{\bar{J}}(\mathbf{j}),\mathcal{I}_{\bar{I}}(\mathbf{k}),\mathcal{I}_S(\mathbf{l})} = x^{(S)}_{\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}}\xi^{(S)}_{\mathbf{l}} \qquad \bar{X}_{\mathcal{I}_{\bar{J}}(\mathbf{j}),\mathcal{I}_{\bar{I}}(\mathbf{k}),\mathcal{I}_S(\mathbf{l})} = x_{\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}}$$

for all $\mathbf{j} \in \mathbf{J^n}(\bar{J})$, $\mathbf{k} \in \mathbf{J^n}(\bar{I})$, $\mathbf{l} \in \mathbf{J^n}(S)$. Analogously, we define $\mathbf{Y}, \bar{\mathbf{Y}} \in \mathbb{R}^{\bar{n}_1'\times\bar{n}_2'\times\bar{n}_3'}$ such that

$$Y_{\mathcal{I}_{\bar{J}'}(\mathbf{j}'),\mathcal{I}_{\bar{I}'}(\mathbf{k}'),\mathcal{I}_T(\mathbf{l}')} = x^{(T)}_{\mathbf{j}'\dot{\times}\mathbf{k}'\dot{\times}\mathbf{l}'}\xi^{(T)}_{\mathbf{l}'} \qquad \bar{Y}_{\mathcal{I}_{\bar{J}'}(\mathbf{j}'),\mathcal{I}_{\bar{I}'}(\mathbf{k}'),\mathcal{I}_T(\mathbf{l}')} = x_{\mathbf{j}'\dot{\times}\mathbf{k}'\dot{\times}\mathbf{l}'}$$

for all $\mathbf{j}' \in \mathbf{J^n}(\bar{J}' - d)$, $\mathbf{k}' \in \mathbf{J^n}(\bar{I}' - d)$, $\mathbf{l}' \in \mathbf{J^n}(T)$.

Furthermore, we define vectors $\alpha \in \mathbb{R}^{\bar{n}_2}, \alpha' \in \mathbb{R}^{\bar{n}_2'}$ and a matrix $\Gamma \in \mathbb{R}^{\bar{n}_1\times\bar{n}_1'}$ such that

$$\alpha_{\mathcal{I}_{\bar{I}}(\mathbf{k})} = \alpha^{(1)}_{\mathbf{k}} \qquad \alpha_{\mathcal{I}_{\bar{I}'}(\mathbf{k}')} = \alpha^{(2)}_{\mathbf{k}'} \qquad \Gamma_{\mathcal{I}_{\bar{J}}(\mathbf{j}),\mathcal{I}_{\bar{J}'}(\mathbf{j}')} = \alpha^{(3)}_{\mathbf{j}\dot{+}\mathbf{j}'}.$$

We define one more function $\mathcal{I} : [\bar{n}_1] \times [\bar{n}_2] \times [\bar{n}_3] \to [N]$ such that

$$\mathcal{I}(\mathcal{I}_{\bar{J}}(\mathbf{j}), \mathcal{I}_{\bar{I}}(\mathbf{k}), \mathcal{I}_S(\mathbf{l})) = \mathcal{I}^{\mathbf{n}}(\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}).$$

for all $\mathbf{j} \in \mathbf{J^n}(\bar{J})$, $\mathbf{k} \in \mathbf{J^n}(\bar{I})$, $\mathbf{l} \in \mathbf{J^n}(S)$. Recall that $\mathcal{I}^{\mathbf{n}} : \mathbf{J^n} \to [N]$ is the bijection that maps row/column indices of the matrix $\Phi^*\Phi - Id_N$ to corresponding indices for the rearranged order $d$ array $\mathbf{B}$ (see Definition 2.3 and (2.33)), i.e., for all $\mathbf{i}, \mathbf{i}' \in \mathbf{J^n}$, $B_{\mathbf{i}\dot{+}\mathbf{i}} = (\Phi^*\Phi - Id_N)_{\mathcal{I}^{\mathbf{n}}(\mathbf{i}),\mathcal{I}^{\mathbf{n}}(\mathbf{i}')}$. Since every $\mathbf{i} \in \mathbf{J^n}$ has a unique representation as $\mathbf{i} = \mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}$ with $\mathbf{j} \in \mathbf{J^n}(\bar{J})$, $\mathbf{k} \in \mathbf{J^n}(\bar{I})$, $\mathbf{l} \in \mathbf{J^n}(S)$ and the functions $\mathcal{I}_{\bar{J}}$, $\mathcal{I}_{\bar{I}}$, $\mathcal{I}_S$ are bijective, also $\mathcal{I}$ is bijective. Analogously, we define $\mathcal{I}' : [\bar{n}_1'] \times [\bar{n}_2'] \times [\bar{n}_3'] \to [N]$ such that

$$\mathcal{I}(\mathcal{I}_{\bar{J}'}(\mathbf{j}'), \mathcal{I}_{\bar{I}'}(\mathbf{k}'), \mathcal{I}_T(\mathbf{l}')) = \mathcal{I}^{\mathbf{n}}(\mathbf{j}'\dot{\times}\mathbf{k}'\dot{\times}\mathbf{l}').$$

for all $\mathbf{j}' \in \mathbf{J^n}(\bar{J}' - d)$, $\mathbf{k}' \in \mathbf{J^n}(\bar{I} - d)$, $\mathbf{l}' \in \mathbf{J^n}(T)$, and also this function is bijective.

Using all the aforementioned definition and the matrix $B = \Phi^*\Phi - Id_N$, we can rewrite (2.36) as

$$\begin{aligned} &\sum_{\substack{\mathbf{j}\in\mathbf{J^n}(\bar{J}),\mathbf{j}'\in\mathbf{J^n}(\bar{J}'-d)\\ \mathbf{k}\in\mathbf{J^n}(\bar{I}),\mathbf{k}'\in\mathbf{J^n}(\bar{I}'-d)\\ \mathbf{l}\in\mathbf{J^n}(S),\mathbf{l}'\in\mathbf{J^n}(T)}} B_{(\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l})\dot{+}(\mathbf{j}'\dot{\times}\mathbf{k}'\dot{\times}\mathbf{l}')} x^{(S)}_{\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}} x^{(T)}_{\mathbf{j}'\dot{\times}\mathbf{k}'\dot{\times}\mathbf{l}'} \xi^{(S)}_{\mathbf{l}} \bar{\xi}^{(T)}_{\mathbf{l}'} \alpha^{(1)}_{\mathbf{k}} \alpha^{(2)}_{\mathbf{k}'} \alpha^{(3)}_{\mathbf{j}\dot{+}\mathbf{j}'} \\ &= \sum_{\substack{(j,j')\in[n_1]\times[n_1']\\ (k,k')\in[n_2]\times[n_2']\\ (l,l')\in[n_3]\times[n_3']}} B_{\mathcal{I}(j,k,l),\mathcal{I}'(j',k',l')} X_{j,k,l}\alpha_k Y_{j',k',l'}\alpha'_{k'}\Gamma_{j,j'}. \end{aligned} \tag{2.37}$$

Before we can apply Lemma 1.12 to this, we need to check that all the remaining requirements are fulfilled. We check the following conditions for $\mathbf{X}, \bar{\mathbf{X}}$.

(a) follows directly from the definitions of $\mathbf{X}$ and $\bar{\mathbf{X}}$ and the fact that all entries of $\boldsymbol{\xi}^{(S)}$ have absolute value 1.

(b) By the definition of $\bar{\mathbf{X}}$, $\|\bar{\mathbf{X}}\|_F = \|\mathbf{x}\|_F = 1$.

(c) We use Lemma 2.38 for the disjoint sets $S$ and $\bar{J} \cup \bar{I}$ which implies

$$\max_{j\in[\bar{n}_1],k\in[\bar{n}_2]} \sum_{l\in[\bar{n}_3]} X_{j,k,l}^2 = \max_{\mathbf{j}\in\mathbf{J^n}(\mathbf{J}),\mathbf{k}\in\mathbf{J^n}(\bar{I})} \sum_{\mathbf{l}\in\mathbf{J^n}(S)} (x^{(S)}_{\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}}\xi^{(S)}_{\mathbf{l}})^2$$
$$= \max_{\mathbf{i}\in\mathbf{J^n}(\bar{J}\cup\bar{I})} \sum_{\mathbf{l}\in\mathbf{J^n}(S)} (x^{(S)}_{\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}})^2 \leq \frac{1}{s^{|\bar{J}\cup\bar{I}|}} \sum_{\mathbf{i}\in\mathbf{J^n}} (x_{\mathbf{i}})^2 = \frac{1}{s_1 s_2}.$$

(d) By Lemma 2.38 for the disjoint set $S$ and $\bar{I}$, for each $\mathbf{j} \in \mathbf{J^n}(\bar{J})$,

$$\max_{\mathbf{k}\in\mathbf{J^n}(\bar{I})} \sum_{\mathbf{l}\in\mathbf{J^n}(S)} (x^{(S)}_{\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}}\xi^{(S)}_{\mathbf{l}})^2 = \max_{\mathbf{k}\in\mathbf{J^n}(\bar{I})} \sum_{\mathbf{l}\in\mathbf{J^n}(S)} (x^{(S)}_{\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}})^2$$
$$\leq \frac{1}{s_2} \sum_{\mathbf{k}\in\mathbf{J^n}(\bar{I}),\mathbf{l}\in\mathbf{J^n}(S)} (x_{\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}})^2.$$

By the definition of $\mathbf{X}$ and $\bar{\mathbf{X}}$, this translates to the condition (d).

(e) Analogous to the previous case (d), the condition follows from applying Lemma 2.38 for $S$ and $\bar{J}$, which shows for all $\mathbf{k} \in \mathbf{J^n}(\bar{I})$,

$$\max_{\mathbf{j}\in\mathbf{J^n}(\bar{J})} \sum_{\mathbf{l}\in\mathbf{J^n}(S)} (x^{(S)}_{\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}}\xi^{(S)}_{\mathbf{l}})^2 \leq \frac{1}{s_1} \sum_{\mathbf{j}\in\mathbf{J^n}(\bar{J}),\mathbf{l}\in\mathbf{J^n}(S)} (x_{\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}})^2.$$

(f) By Lemma 2.36, for any $\mathbf{j} \in \mathbf{J^n}(\bar{J})$, $\mathbf{k} \in \mathbf{J^n}(\bar{I})$, there are at most $s^{|S|} = s_3$ different indices $\mathbf{l} \in \mathbf{J^n}(S)$ such that $x^{(S)}_{\mathbf{j}\dot{\times}\mathbf{k}\dot{\times}\mathbf{l}} \neq 0$. By the definition of $\mathbf{X}$, this implies the property (f).

For the arrays $\mathbf{Y}, \bar{\mathbf{Y}}$, the properties (a) to (f) follow analogously. Furthermore, the corresponding definitions directly yield

$$\|\alpha\|_2 = \|\boldsymbol{\alpha}^{(1)}\|_2 = 1 \qquad \|\alpha'\|_2 = \|\boldsymbol{\alpha}^{(2)}\|_2 = 1 \qquad \|\Gamma\|_F = \|\boldsymbol{\alpha}^{(3)}\|_2 = 1.$$

This completes the proof that all the assumptions of Lemma 1.12 are fulfilled and we can apply this lemma to bound (2.37) and therefore also (2.36) and $\|\mathbf{B}^{(S,T)}\|_{I_1,\dots,I_\kappa}$ by

$$\|\mathbf{B}^{(S,T)}\|_{I_1,\dots,I_\kappa} \leq 4\frac{\delta}{(s_1 s_1')^{\frac{1}{4}}(s_2 s_2')^{\frac{1}{2}}} = 4\frac{\delta}{s^{\frac{1}{4}|\bar{\bar{J}}|+\frac{1}{2}(|\bar{I}|+|\bar{I}'|)}} \leq 4\frac{\delta}{s^{\frac{\kappa}{2}}},$$

where we used Lemma 1.17 in the last step. This completes the proof of (2.35).

# 3 Scale-Invariant Neural Networks for Inverse Problems

## 3.1 Introduction

In this part of the thesis, we study the performance of neural networks on the sparse recovery problem as opposed to usual methods that are based on convex optimization. As a related problem, we also study the approximation of functions that are invariant under positive scaling with neural networks.

Recently, neural networks often outperform the classical sparsity-based methods for a variety of signal and image reconstruction problems. Specifically, neural networks achieve state-of-the-art results in tasks such as denoising [Zha+17] and reconstructing imaget from few and noisy examples [Zbo+18]. However, contrary to optimization-based methods for which a rich literature on performance guarantees exists [FR13], many underlying theoretical questions are still open for neural network-based signal reconstruction,

In this work, we consider the question of whether a sparse signal can be provably recovered with a neural network. Given measurements $y = Ax$ for a sparse signal $x$, we study neural networks $f$ that recover $x$ from a coarse approximation given by $A^T y$, specifically networks that obey $x = f(A^T y)$ or at least ensure that the reconstruction error $\|x - f(A^T y)\|$ is small.

Note that in the first step, a neural network applies a linear transformation to its input. Thus, given a neural network $f$, we can define another network $f_2$ such that $f_2(y) = f(A^T y)$. Considering in addition that $(A^\dagger)^T A^T A = A$ (with $A^\dagger$ being the Moore-Penrose pseudoinverse), we can see that the problems of finding a neural network $f$ such that $f(A^T y)$ is (approximately) $x$ and finding $f_2$ such that $f_2(y)$ is (approximately) $x$ are equivalent. This is why we consider the latter case in the results of this work.

This work aims to investigate under what circumstances ReLU networks can approximately solve the sparse recovery problem. In particular, we are interested in the number of layers that such a network requires.

Moreover, we are interested in solutions that make use of the positive homogeneous structure of the problem. That is, we want the recovery network function $f$ to satisfy $f(\lambda y) = \lambda f(y)$ for all $\lambda \geq 0$ because we also know that if $x$ has measurements $y = Ax$, then the measurements $\lambda y$ will be obtained from the signal $\lambda x$. In this way, no prior knowledge of the size of the signal $x$ or training for different sizes is required. Such a network can recover every $s$-sparse vector and therefore works on an unbounded domain. Therefore investigating the number of required layers goes beyond the usual universal approximation theorem ([Pin99], see Theorem 3.1 below), which can guarantee arbitrarily precise approximations but only on a compact domain and without incorporating the positive homogeneous structure of the problem into the network.

Taking knowledge about a function into account for the design of the network to approximate it is a strategy that can significantly improve reliability and training effort. For this reason, a large number of works have studied this strategy over a long period of time for different types of functions [WS96; Dug+09; KSO21; Chi+19]. Specifically, [Tan+20] considers the aforementioned class of functions satisfying $\tilde{f}(\lambda y) = \lambda \tilde{f}(y)$ for all $\lambda \in [0, \infty)$. We call these functions positive homogeneous.

Furthermore, positive homogeneity can also have applications for other problems. For example in image denoising, rescaling the brightness of a picture might not change the underlying procedure and if the corresponding network is designed to be positive homogeneous, different brightness levels do not need to be learned separately.

### 3.1.1 Contributions of this work

We first show that with one hidden layer, it is not possible to even approximately recover 1-sparse vectors. Secondly, we show that two hidden layers are sufficient to recover sparse vectors with arbitrary sparsity levels $s$ and to arbitrary precision. Furthermore, we also show a robustness guarantee for these networks establishing that the resulting network reconstruction function can reconstruct vectors that are only approximately $s$-sparse and from corrupted measurements.

We show the positive result for two hidden layers for a more general class of inverse problems. In general, instead of the set of sparse vectors, we can have any subset $U \in \mathbb{R}^n$ that is positive homogeneous (i.e., $\lambda u \in U$ for any $\lambda \in [0, \infty)$ and $u \in U$), and instead of a linear map $A$ we can have a positive homogeneous function $g$ that satisfies certain requirements. Besides sparse recovery, using this generalized result we also show that the low-rank matrix recovery and the phase retrieval problem can be solved using ReLU networks with two hidden layers.

Since for problems like sparse recovery or phase retrieval, there are solution methods based on optimization problems, we also show a method how these can be translated to a ReLU network with two hidden layers. The central argument is that there exists a continuous solution function of the optimization problem. We show this using a generalization of the continuity concept to functions with multiple values.

Furthermore, we also gain more insights about the general approximation of continuous, positive homogeneous functions. Specifically, [Tan+20] already shows that with the ReLU activation function, the unbiased networks with two hidden layers represent a class of functions such that (i) all these functions are positive homogeneous and (ii) they can approximate every continuous positive homogeneous function to arbitrary precision. We complement their result by showing that (up to certain modifications of itself), the ReLU function is the only activation function such that the unbiased networks satisfy these two conditions (i) and (ii). We establish this in a theorem which is similar to the classical universal approximation theorem of neural networks. Furthermore, using the negative results about sparse recovery, we also prove that this universal approximation property fails to hold for just one hidden layer such that the assumption of two hidden layers is actually necessary.

In Section 3.2, we present our main reslts in two parts. One part is about solving inverse problems and the other one about universal approximation of positive homogeneous functions. Then Section 3.3 contains the proof of the main result about universal approximation. In Section 3.4, we prove the main results regarding inverse problems and show some other applications of them. Section 3.5 then shows that ReLU networks can be used to solve inverse problems in the way optimization-based methods do. In Section 3.8, we discuss implications and relations to other work.

### 3.1.2 Previous work on universal approximation of NNs

To solve the aforementioned signal recovery problem, we need to compute the function that maps measurements $y = Ax$ to their original signals $x$. Compressed sensing guarantees the well-definedness of this function and the question is if, how, and how well this function can be approximated by certain classes of neural networks.

The general question of how well certain functions can be approximated has been a central question in the research of neural networks for a long time. Cybenko [Cyb89] showed that neural networks with only one hidden layer and any bounded measurable sigmoidal activation function can approximate any continuous function on the $n$-dimensional unit cube to arbitrary precision if the width of the network is sufficiently large. This result has been known as the *universal approximation theorem* and has been extended several times. For example, Leshno et al. [Les+93] generalized it to the case of any non-polynomial activation function. [Pin99] even proved for a large class of functions that this approximation property is equivalent to the function being non-polynomial.

**Theorem 3.1** (Universal Approximation Theorem, [Pin99])**.** *Let $n \geq 1$ be a dimension and $\sigma : \mathbb{R} \to \mathbb{R}$ continuous. Then the following are equivalent.*

*(a) For any compact $K \subset \mathbb{R}^n$, any continuous $f : K \to \mathbb{R}$ and any $\delta > 0$, there exists a network with one hidden layer and activation function $\sigma$, representing $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$, such*

*that for all $x \in K$,*

$$|\tilde{f}(x) - f(x)| \leq \delta.$$

*(b) $\sigma$ is not a polynomial.*

As mentioned previously, taking known properties of the approximated functions into account for the network design has been studied in multiple previous works. To mention some concrete examples, [Dug+09] considers functions satisfying a certain monotonicity and convexity condition, [KSO21] considers functions that are invariant under certain permutations of their input variables, and [Tan+20] considers the positive homogeneous functions.

All these works construct a class of networks and show that, on the one hand, all these networks represent functions of the particular class, and on the other hand, every function of the respective class can be approximated by one of these networks. The latter part corresponds to the implication (b) ⇒ (a) in the universal approximation Theorem 3.1 for the class of general continuous functions on compact domains.

In particular Tang et al. [Tan+20] show these things for positive homogeneous functions. In this work, we extend their result to an equivalence statement similar to Theorem 3.1 which will be Theorem 3.7. We also show that their requirement of having at least two hidden layers is required.

### 3.1.3 Previous work on sparse recovery with NNs

A popular approach for sparse recovery of a vector $x$ from a linear measurement $y = Ax$ is the basis pursuit denoising (0.5). As a convex optimization problem, it can be solved by proximal methods such ahs the *iterative shrinkage thresholding algorithm* (ISTA). ISTA is initialized at some $x^0$ and iterates for $\ell = 1, 2, \dots$

$$x^{\ell+1} = \eta_{\lambda/L}\left(x^\ell - \frac{1}{L}A^T(Ax^\ell - y)\right), \tag{3.1}$$

where $\eta_z$ is the soft-thresholding function, i.e., $\eta_z(t) = \text{sign}(t)(|t| - z)$ if $|t| \geq z$ and $\eta_z = 0$ otherwise. A sequence of $d$ iterations can be regarded as a recurrent neural network of depth $d$. Based on this, a number of works starting with [GL10] studied unrolled algorithms which unrolls $d$ ISTA iterations as

$$x^{\ell+1} = \eta_{\lambda/L}\left(W_1^\ell x^\ell + W_2^\ell y\right). \tag{3.2}$$

This is a feed-forward neural network of depth $d$ and $W_1^\ell$ and $W_2^\ell$ are weight matrices that are typically learned based on data. In all layers, $\eta_{\lambda/L}$ is used as an activation function. Chen at al. [Che+18] (Thm. 2) established that there exist choices of weights such that an $s$-sparse signal $x$ with entries bounded by $|x_i| \leq B$, and with $s$ sufficiently small can be approximated as

$$\|x^d - x\|_2 \leq sBe^{-cd},$$

where $c$ is a constant depending on the matrix $A$ and mildly on the sparsity $s$ of the signal. This result requires at least $s^2 \leq m$, where $m$ is the number of measurements, as it works with the incoherence of the matrix $A$, and requires the sparsity to be sufficiently small relative to the incoherence (see [Che+18, Appendix B, Step 3]. This result establishes that there is a relatively shallow neural network that can approximate sparse signals well.

So with those approaches, depth $d = \mathcal{O}(\log s)$ is sufficient to approximate the signal $x^*$ with the output of the network. In contrast, the goal of this work is to determine the exact number of layers that is necessary and sufficient to solve the sparse recovery problem. Furthermore, other than the aforementioned unrolling approach, which works for signals whose entries are bounded by $B$, we consider networks that can solve the sparse recovery problem on the entire (unbounded) set of possible signals. We prove that for these requirements and with ReLU activation function, one hidden layer is not sufficient to solve the problem, but two hidden layers are, even for a stable solution. However, our results do not yield a construction for these networks and also don't specify their width.

### 3.1.4 Robustness

An important aspect of solutions to the sparse recovery problem is how sensitive they are to noisy measurements $y = Ax + e$. For a robust recovery method $f : \mathbb{R}^m \to \mathbb{R}^n$, we expect $\|f(Ax+e) - f(Ax)\|_2$ to be small for small $\|e\|_2$. With minimization-based methods such as the quadratically-constrained basis pursuit (3.14), robust recovery has been proven to be successful (see Theorem 0.3).

For neural networks on inverse problems, the question of robustness is currently studied under various aspects. In [GMM22], an empirical analysis is conducted suggesting that neural networks can provide robust solutions to specifically chosen problems similar to sparse recovery and image reconstruction in a similar or even better way compared to optimization-based methods.

In contrast to this, [Got+20] provides a theoretical analysis of certain situations that necessarily lead to robustness issues for neural networks on inverse problems. Specifically, they show that in certain situations, neural networks applied to inverse problems necessarily have large local Lipschitz constants.

We also review our work in the context of the aforementioned results. This provide a possible interpretation of this seeming contradiction. We show robustness similar to the one for the minimization problem (0.4) but still the local Lipschitz constants of our solution might be very large. This is due to possible large gradients $\frac{\|f(Ax+e)-f(Ax)\|_2}{\|e\|_2}$ for very small error levels $\|e\|_2$. This is discussed in detail in Section 3.8.

### 3.1.5 Notation

We consider neural networks with the rectified linear unit activation function ReLU : $\mathbb{R} \to \mathbb{R}$, defined by $\mathrm{ReLU}(x) = \max\{x, 0\}$. We also use the shorter notation $\phi :=$ ReLU.

A lot of statements in this work concern feedforward neural networks for which we use the notation and terminology introduced in Section 0.3.2.

Furthermore, we work with positive homogeneous sets and functions according to the following definition.

**Definition 3.2.** *We define a set $U \subset \mathbb{R}^n$ to be positive homogeneous if for all $\lambda \in [0, \infty)$ and all $x \in U$, also $\lambda x \in U$.*

*If $U \subset \mathbb{R}^n$ is a positive homogeneous set, we define a function $f : U \to \mathbb{R}^m$ to be positive homogeneous if for all $\lambda \in [0, \infty)$ and all $x \in U$, $f(\lambda x) = \lambda f(x)$.*

## 3.2 Main Results

### 3.2.1 Inverse Problems

**Sparse Recovery:**

The main results of this work regarding sparse recovery are the following two Corollaries 3.3 and 3.4 which are consequences of the slightly more general Theorems 3.5 and 3.6 below respectively.

Corollary 3.3 below states that a ReLU network with one hidden layer cannot recover all sparse vectors from any $m \ll n$ linear measurements, not even approximately and for 1-sparse vectors.

**Corollary 3.3** (Impossibility result for one hidden layer)**.** *Let $A \in \mathbb{R}^{m\times n}$, $m \leq n$, and $f : \mathbb{R}^m \to \mathbb{R}^n$ be a function represented by a* ReLU *network with one hidden layer. Then, for any width and any choice of the network parameters,*

$$\sup_{x\in\Sigma_1\setminus\{0\}} \frac{\|x - f(Ax)\|_2}{\|x\|_2} \geq \sqrt{1 - \frac{m}{n}}.$$

Note that in usual recovery problems, $m \ll n$ such that the lower bound for the relative error is close to 1. Thus, the reconstruction function is guaranteed to make a large error for reconstructing at least one 1-sparse signal.

In strong contrast, Corollary 3.4 states that for a ReLU network with two hidden layers, recovery of all $s$-sparse vectors is possible to arbitrary precision and in a stable way for not exactly sparse signals or corrupted measurements.

**Corollary 3.4.** *Let $A \in \mathbb{R}^{m\times n}$ be a matrix satisfying the $(2s, \delta)$-RIP for a $\delta \in (0, 1)$. Then for each $\delta' \in (0, 1)$, there exists a function $\tilde{f} : \mathbb{R}^m \to \mathbb{R}^n$, represented by an unbiased* ReLU *network with two hidden layers such that for all $x \in \mathbb{R}^n$, $e \in \mathbb{R}^m$,*

$$\|\tilde{f}(Ax + e) - x\|_2 \leq \delta'\|x\|_2 + C\sigma_s(x)_1 + D\|e\|_2,$$

*where $C = 1 + 2\frac{1+\delta}{1-\delta}$, $D = \frac{3}{1-\delta}$, and*

$$\sigma_s(x)_1 = \inf_{x'\in\Sigma_s} \|x - x'\|_1.$$

**General Form:**

For each of the above statements, we actually prove a more general version. The following theorem is not restricted to sparse recovery but applicable to a wider range of inverse problems which, as we will show in Section 3.4, also includes low-rank matrix recovery and phase retrieval.

**Theorem 3.5.** *Let $U \subset \mathbb{R}^n$ be positive homogeneous. Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be a positive homogeneous function such that*

$$\inf_{\substack{x^{(1)},x^{(2)}\in U\\ x^{(1)}\neq x^{(2)}}} \frac{\|g(x^{(1)}) - g(x^{(2)})\|_2}{\|x^{(1)} - x^{(2)}\|_2} =: \tau > 0 \qquad \sup_{\substack{x^{(1)},x^{(2)}\in\mathbb{R}^n\\ x^{(1)}\neq x^{(2)}}} \frac{\|g(x^{(1)}) - g(x^{(2)})\|_2}{\|x^{(1)} - x^{(2)}\|_{II}} =: \rho < \infty \tag{3.3}$$

*where $\|\cdot\|_{II}$ is a norm on $\mathbb{R}^n$ with $\|\cdot\|_2 \leq \|\cdot\|_{II}$.*

*Let $\delta \in (0, 1)$. Then there exists a function $\tilde{f} : \mathbb{R}^m \to \mathbb{R}^n$, represented by a* ReLU *network with two hidden layers, such that for all $x \in \mathbb{R}^n$, $e \in \mathbb{R}^m$,*

$$\|\tilde{f}(g(x) + e) - x\|_2 \leq \delta\|x\|_2 + Cd_{II}(x, U) + D\|e\|_2,$$

*where $C = 1 + \frac{2\rho}{\tau}$ and $D = \frac{3}{\tau}$ only depend on $\tau$ and $\rho$ and*

$$d_{II}(x, U) := \inf_{x'\in U} \|x - x'\|_{II}.$$

Furthermore, the following result generalizes Corollary 3.3 to general unions of subspaces instead of the set of sparse vectors.

**Theorem 3.6.** *Let $A \in \mathbb{R}^{m\times n}$, $m \leq n$, and $f : \mathbb{R}^m \to \mathbb{R}^n$ be a function represented by a* ReLU *network with one hidden layer.*

*Let $x_1, \ldots, x_{\tilde{n}} \in \mathbb{R}^n$ be vectors with $\|\cdot\|_2$ norm 1 and $X := (x_1\, x_2\, \ldots\, x_{\tilde{n}}) \in \mathbb{R}^{n\times\tilde{n}}$. Let $U = \bigcup_{k=1}^{\tilde{n}} \operatorname{span}(x_k)$. Then*

$$\sup_{x\in U\setminus\{0\}} \frac{\|f(Ax) - x\|_2}{\|x\|_2} \geq \sqrt{\frac{1}{\tilde{n}} \sum_{k=m+1}^{\tilde{n}} (\sigma_k(X))^2}.$$

Note that an essential requirement for the lower bounds, Theorem 3.6 and thus Corollary 3.3, to hold is that we consider the approximation errors on the entire (unbounded) domain $U$ or $\Sigma_1$ respectively. The purpose of this is that we are interested in incorporating the positive homogenous structure of the problem into the solution. That is, the recovery function $\tilde{f}$ should by design satisfy $\tilde{f}(\lambda y) = \lambda \tilde{f}(y)$ for all $\lambda \geq 0$ because we know that the measurements $\lambda Ax$ are produced by the signal $\lambda x$. If $\tilde{f}$ is positive homogeneous and $\|\tilde{f}(Ax) - x\|_2 \leq \delta \|x\|_2$ holds for all $x \in \bar{B}_r(0) \cap U$ for some radius $r > 0$, then this is also true on the entire positive homogeneous set $U$. So Theorem 3.6 and Corollary 3.3 can also be interpreted in the sense that there is no *positive homogeneous* network function $\tilde{f}$ that provides a good recovery for all vectors in a (possibly bounded) neighborhood of 0.

So the key question of this work is when positive homogeneous functions can be approximated (to arbitrary precision) with positive homogeneous networks. As mentioned above, for this case it is irrelevant if the domain is only a neighborhood of 0 or the entire space. In the following subsection, Theorems 3.7 and 3.8 provide the answer that it is possible with any number of layers $d \geq 2$ but not with $d = 1$. In addition, Theorem 3.7 also classifies the possible activation functions for this.

### 3.2.2 Universal Approximation of Positive Homogeneous Functions

We show the following statement about the universal approximation of positive homogeneous functions. It can be seen as an analogous version of the equivalence in Theorem 3.1 for positive homogeneous functions. The essential proof step for the direction (b) ⇒ (a) has already been established by Tang et al. [Tan+20]. We extend this to an equivalence statement, showing that the activation functions described in (b) are actually the only ones for which the unbiased networks represent a class of functions which are all positive homogeneous and also powerful enough to approximate any other continuous positive homogeneous function.

**Theorem 3.7.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous function and $d \geq 2$ an integer. Then the following two statements are equivalent.*

*(a)*

- *For every non-empty, closed, positive homogeneous $U \subset \mathbb{R}^n$, every continuous, positive homogeneous function $f : U \to \mathbb{R}$, and every $\delta > 0$, there exists a function $\tilde{f} : U \to \mathbb{R}$ that can be represented by a neural network with $d$ hidden layers and activation function $\sigma$, such that for all $x \in U$,*

$$|\tilde{f}(x) - f(x)| \leq \delta \|x\|_2$$

***and***

- *every unbiased neural network with $d$ hidden layers and activation function $\sigma$ represents a positive homogeneous function.*

*(b) There are $\alpha, \beta \in \mathbb{R}$, $|\alpha| \neq |\beta|$ such that*

$$\sigma(x) = \alpha \, \mathrm{ReLU}(x) + \beta \, \mathrm{ReLU}(-x)$$

*for all $x \in \mathbb{R}$.*

*In case these statements hold, the network representing $\tilde{f}$ in (a) can be chosen to be unbiased.*

In addition, we also complement Theorem 3.7 by the following consequence of Corollary 3.3 which proves that Theorem 3.7 fails to hold if we consider networks with only one hidden layer. Therefore, its assumption $d \geq 2$ is actually necessary.

**Theorem 3.8.** *Let $m \geq 2$, $n \geq 1$ be integers. There exists a continuous, positive homogeneous function $f : \mathbb{R}^m \to \mathbb{R}^n$ such that for each $\tilde{f} : \mathbb{R}^m \to \mathbb{R}^n$ that is represented by a* ReLU *network with one hidden layer,*

$$\sup_{x \in \mathbb{R}^m \setminus \{0\}} \frac{\|\tilde{f}(x) - f(x)\|_2}{\|x\|_2} \geq \begin{cases} \sqrt{1 - \frac{2}{n}} & \text{if } n > 4 \\ \sqrt{\frac{n}{8}} & \text{if } n \leq 4. \end{cases}$$

In particular, the case $n = 1$ in Theorem 3.8 shows that contrary to higher depths $d \geq 2$, the first part of (a) does not hold for the ReLU activation function and any $\delta < \sqrt{\frac{1}{8}}$ if $d = 1$.

### 3.3 Universal Approximation

This section is devoted to the proof of Theorem 3.7. The following lemmma will be used to establish the implication (a) ⇒ (b).

**Lemma 3.9.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous function. Let $k \in \mathbb{Z}_{\geq 1}$. Assume that for all $\gamma \in \mathbb{R}^k$, the function $\sigma_\gamma : \mathbb{R} \to \mathbb{R}$ defined by*

$$\sigma_\gamma(x) = \gamma_k \sigma(\gamma_{k-1}\sigma(\gamma_{k-2}\sigma(\dots\sigma(\gamma_1\sigma(x))\dots))),$$

*with $k$ applications of $\sigma$, is positive homogeneous.*

*Then there exist $\alpha, \beta \in \mathbb{R}$ such that for all $x \in \mathbb{R}$,*

$$\sigma(x) = \alpha\phi(x) + \beta\phi(-x). \tag{3.4}$$

*Proof.* If $k = 1$, then choose $\gamma = 1 \in \mathbb{R}$ such that $\sigma_\gamma = \sigma(x)$. By the positive homogeneity, we obtain that for $x \geq 0$,

$$\sigma(x) = \sigma_\gamma(x \cdot 1) = x\sigma(1) = \sigma(1)\phi(x)$$

and for $x \leq 0$,

$$\sigma(x) = \sigma_\gamma(|x| \cdot (-1)) = |x|\sigma(-1) = \sigma(-1)\phi(-x),$$

showing the representation (3.4).

Now we assume that $k \geq 2$. First, we take $\gamma = (1, 0, \dots, 0)^T$. Then for all $x \in \mathbb{R}$, $\sigma_\gamma(x) = 1 \cdot \sigma(0 \cdot \sigma(\dots)) = \sigma(0)$. Since $\sigma_\gamma$ is positive homogeneous, $\sigma(0) = \sigma_\gamma(1) = \frac{1}{2}\sigma_\gamma(2) = \frac{1}{2}\sigma(0)$, so we know that $\sigma(0) = 0$.

If $\sigma(x) = 0$ for all $x \geq 0$, then the representation (3.4) holds for all $x \geq 0$. Otherwise there exists a $y_0 > 0$ such that $\sigma(y_0) \neq 0$. Then we choose $\gamma_0 = \frac{y_0}{\sigma(y_0)} \in \mathbb{R}$ and $\gamma = (\gamma_0, \dots, \gamma_0)^T \in \mathbb{R}^k$. We define $\tilde{\sigma} : \mathbb{R} \to \mathbb{R}$ by $\tilde{\sigma}(x) = \gamma_0\sigma(x)$. Then $\sigma_\gamma = \tilde{\sigma} \circ \dots \circ \tilde{\sigma}$ with $k$ applications.

By the choice of $\gamma_0$, $\tilde{\sigma}(y_0) = y_0$ such that also $\sigma_\gamma(y_0) = y_0$. By the assumption that $\sigma_\gamma$ is positive homogeneous and the previously shown case $k = 1$, we obtain that there are $\eta, \tau \in \mathbb{R}$ such that for all $x \in \mathbb{R}$,

$$\sigma_\gamma(x) = \eta\phi(x) + \tau\phi(-x) \tag{3.5}$$

and $\sigma_\gamma(y_0) = y_0$ implies that $\eta = 1$, i.e., $\sigma_\gamma(x) = x$ for all $x \geq 0$.

So we know that $\sigma_\gamma$ is injective on the interval $[0, \infty)$ and then the same must hold for $\tilde{\sigma}$. As a continuous, $\mathbb{R}$-valued, injective function in the interval $[0, \infty)$, $\tilde{\sigma}$ must be either strictly increasing or strictly decreasing. We have already shown $\tilde{\sigma}(0) = 0$ and $\tilde{\sigma}(y_0) = y_0$ where $y_0 > 0$, so it must be strictly increasing. This also implies that $\tilde{\sigma}(x) \geq 0$ for all $x \geq 0$.

Now assume that $\tilde{\sigma}(x) = x$ does not hold for all $x \in [0, \infty)$. Then we can find an $x_0 > 0$ such that $\tilde{\sigma}(x_0) \neq x_0$. Starting from this $x_0$, we construct a sequence $(x_j)$ in $[0, \infty)$ by defining $x_{j+1} = \tilde{\sigma}(x_j)$ for $j = 0, 1, 2, \dots$. We observe the following for all $j \in \mathbb{Z}_{\geq 0}$.

- If $x_{j+1} > x_j$, then by monotonicity also $x_{j+2} = \tilde{\sigma}(x_{j+1}) > \tilde{\sigma}(x_j) = x_{j+1}$.

- If $x_{j+1} < x_j$, then by monotonicity also $x_{j+2} = \tilde{\sigma}(x_{j+1}) < \tilde{\sigma}(x_j) = x_{j+1}$.

By the choice of $x_0$, we have $x_1 > x_0$ or $x_1 < x_0$ and thus successively either $x_0 < x_1 < x_2 < \dots$ or $x_0 > x_1 > x_2 > \dots$, respectively. In any case, $x_0 \neq x_k$. However, by the definition of the sequence we obtain $x_k = \sigma_\gamma(x_0) = x_0$, which is a contradiction. This completes the proof that $\tilde{\sigma}(x) = x$, i.e., $\sigma(x) = \frac{1}{\gamma_0}x$ holds for all $x \geq 0$.

To complete the proof also for all $x \leq 0$, we consider the function $\bar{\sigma} : \mathbb{R} \to \mathbb{R}$, $\bar{\sigma}(x) = -\sigma(-x)$. For this function we obtain that $\bar{\sigma}_\gamma(x) = -\sigma_\gamma(-x)$. The latter function is then also positive homogeneous such that by the previous proof there is a $\beta \in \mathbb{R}$ such that for all $x \geq 0$, $\bar{\sigma}(x) = \beta x$ and thus for all $x \leq 0$, $\sigma(x) = -\bar{\sigma}(-x) = -\beta(-x) = \beta x$. □

The above statement implies the following corollary.

**Corollary 3.10.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous function and $d \in \mathbb{Z}_{\geq 1}$. If all unbiased neural networks with $d$ hidden layers and activation function $\sigma$ represent positive homogeneous functions, then there are $\alpha, \beta \in \mathbb{R}$ such that for all $x \in R$,*

$$\sigma(x) = \alpha\phi(x) + \beta\phi(-x).$$

*Proof.* All the functions $\sigma_\gamma$ for $\gamma \in \mathbb{R}^d$ from Lemma 3.9 are represented by unbiased networks with $d$ hidden layers with one neuron and weights $\gamma_j$ in each of them and activation function $\sigma$. □

The next statement will be the core of the proof of (b) ⇒ (a) and has already been shown in almost the same form in [Tan+20].

**Theorem 3.11.** *Let $f : U \to \mathbb{R}$ be a positive homogeneous, continuous function on a positive homogeneous domain $U \subset \mathbb{R}^m$ and $\epsilon > 0$. Then there exists an unbiased* ReLU *network with two hidden layers, representing the function $\tilde{f} : U \to \mathbb{R}$, such that for all $x \in U$,*

$$|\tilde{f}(x) - f(x)| \leq \epsilon \|x\|_2.$$

*Proof.* The previous work [Tan+20] shows a similar statement in its supplement in Theorem B.2.2. For completeness of the presentation, we repeat their argument here and adapt it to the situation of Theorem 3.11.

First, we restrict $f$ to the set $U \cap B_1$ where $B_1 = \{x \in \mathbb{R}^m \,|\, \|x\|_1 = 1\}$ is the $\ell_1$ unit ball. Since $U$ is assumed to be closed, this domain $U \cap B_1$ is compact. Therefore, we can apply the universal approximation theorem (Theorem 3.1) to obtain a network function $\tilde{g} : U \cap B_1 \to \mathbb{R}$, $\tilde{g}(x) = W_2\phi(W_1 x + b_1) + b_2$ where $W_1 \in \mathbb{R}^{k \times m}$, $W_2 \in \mathbb{R}^{1 \times k}$, $b_1 \in \mathbb{R}^k$, $b_2 \in \mathbb{R}$ such that for all $x \in U \cap B_1$,

$$|\tilde{g}(x) - f(x)| \leq \frac{1}{\sqrt{n}}\epsilon.$$

Now we define $\tilde{f} : U \to \mathbb{R}$ by $\tilde{f}(0) = 0$ and for $x \in U \backslash \{0\}$

$$\tilde{f}(x) = \|x\|_1 \tilde{g}(\frac{x}{\|x\|_1}) = \|x\|_1(W_2\phi(W_1\frac{x}{\|x\|_1} + b_1) + b_2) = W_2\phi(W_1 x + \|x\|_1 b_1) + \|x\|_1 b_2.$$

Now to show that $\tilde{f}$ can be represented by a ReLU network with two hidden layers, we only need to represent the $\|\cdot\|_1$ function with one hidden layer which is done by

$$\|x\|_1 = 1_{2m}^T \operatorname{ReLU}\left(\begin{pmatrix} Id_m \\ -Id_m \end{pmatrix} x\right),$$

where $1_{2m} \in \mathbb{R}^{2m}$ is the vector whose all entries are 1. Substituting this into the above expression for $\tilde{f}$ yields

$$\tilde{f}(x) = \begin{pmatrix} W_2 & b_2 \end{pmatrix} \phi\left[\begin{pmatrix} W_1 + b_1 1_m^T & -W_1 + b_1 1_m^T \\ 1_m^T & 1_m^T \end{pmatrix} \phi\left(\begin{pmatrix} Id_m \\ -Id_m \end{pmatrix} x\right)\right],$$

such that $\tilde{f}$ can be represented by an unbiased ReLU network with two hidden layers and is therefore also positive homogeneous.

Furthermore, for all $x \in U\backslash\{0\}$, $\frac{x}{\|x\|_1} \in U \cap B_1$ such that $|\tilde{g}(\frac{x}{\|x\|_1}) - f(\frac{x}{\|x\|_1})| \leq \frac{\epsilon}{\sqrt{n}}$ and then

$$|\tilde{f}(x) - f(x)| = \|x\|_1 \left|\tilde{f}(\frac{x}{\|x\|_1}) - f(\frac{x}{\|x\|_1})\right| = \|x\|_1 \left|\tilde{g}(\frac{x}{\|x\|_1}) - f(\frac{x}{\|x\|_1})\right| \leq \frac{\epsilon}{\sqrt{n}}\|x\|_1 \leq \epsilon\|x\|_2.$$

□

Now we have established all the requirements to prove the specialized universal approximation theorem.

*Proof of Theorem 3.7.* First we show the implication (b) $\Rightarrow$ (a). So let $f : U \to \mathbb{R}$ be continuous, positive homogeneous and $\delta > 0$. By Theorem 3.11, there exists an unbiased ReLU network with 2 hidden layers, representing $\tilde{f} : U \to \mathbb{R}$, such that for all $x \in U$, $|\tilde{f}(x) - f(x)| \leq \delta\|x\|_2$.

The one-layer ReLU network function $\mathbb{R}^{n_1} \to \mathbb{R}^{n_1}$

$$x \mapsto (Id_{n_1} \quad -Id_{n_1})\phi\left(\begin{pmatrix} Id_{n_1} \\ -Id_{n_1}\end{pmatrix} x\right) = \phi(x) - \phi(-x) = x,$$

is the identity. Thus, we can add $d-2$ such identity layers to the 2 layer network representing $\tilde{f}$ without changing the represented function. In this way, $\tilde{f}$ can be represented by an unbiased ReLU network with $d$ hidden layers.

Since $|\alpha| \neq |\beta|$, we can define $\gamma_1 := \frac{\alpha}{\alpha^2-\beta^2}$ and $\gamma_2 := \frac{\beta}{\alpha^2-\beta^2}$ such that

$$\gamma_1\sigma(x) - \gamma_2\sigma(-x) = \frac{\alpha}{\alpha^2-\beta^2}(\alpha\phi(x) + \beta\phi(-x)) - \frac{\beta}{\alpha^2-\beta^2}(\alpha\phi(-x) + \beta\phi(x)) = \phi(x).$$

Each of the hidden layers in the network for $\tilde{f}$ performs a function $f_j : \mathbb{R}^{n_1} \to \mathbb{R}^{n_3}$,

$$f_j(x) = A\phi(Bx)$$

for matrices $A \in \mathbb{R}^{n_3 \times n_2}$, $B \in \mathbb{R}^{n_2 \times n_1}$. Then

$$\begin{aligned}(\gamma_1 A \quad -\gamma_2 A)\sigma\left(\begin{pmatrix} B \\ -B\end{pmatrix} x\right) &= \gamma_1 A\sigma(Bx) - \gamma_2 A\sigma(-Bx) = A(\gamma_1\sigma(Bx) - \gamma_2\sigma(-Bx)) \\ &= A\phi(Bx) = f_j(x)\end{aligned}$$

performs the same operation as one layer with activation function $\sigma$. So we can replace all ReLU layers by suitable layers with activation function $\sigma$ and eventually obtain an unbiased network with $d$ hidden layers and activation function $\sigma$ that represents $\tilde{f}$.

Now we show the other implication (a) $\Rightarrow$ (b). The second statement of (a) together with Corollary 3.10 implies that there are $\alpha, \beta \in \mathbb{R}$ such that for all $x \in \mathbb{R}$,

$$\sigma(x) = \alpha\phi(x) + \beta\phi(-x).$$

It remains to show that $|\alpha| \neq |\beta|$. If this is not the case, then $\alpha = \beta$ or $\alpha = -\beta$. In the first case, $\alpha = -\beta$, then $\sigma(x) = \alpha(\phi(x) - \phi(-x)) = \alpha x$ is linear such that also all corresponding network functions $\tilde{f}$ must be affine linear. For example, the function $x \mapsto |x|$ cannot be approximated by such functions to arbitrary precision even though it is positive homogeneous.

In the other case, $\sigma(x) = \alpha(\phi(x) + \phi(-x)) = \alpha|x|$. Consider the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \sigma(ax + b)$ for $a, b \in \mathbb{R}$. For this function, it holds that

$$\lim_{x\to\infty} \frac{f(x)}{|x|} = \lim_{x\to-\infty} \frac{f(x)}{|x|} \in \mathbb{R}. \tag{3.6}$$

Let $f^{(1)} : \mathbb{R} \to \mathbb{R}^k$, $f^{(1)} : x \mapsto \sigma(W_1 x + b_1)$ for $W_1 \in \mathbb{R}^{k_1 \times 1}$, $b_1 \in \mathbb{R}^{k_1}$ be the first layer of a network with activation function $\sigma$ (with domain $\mathbb{R}^1$). Then the condition (3.6) holds for each component of $f^{(1)}$.

Any linear combination of functions that fulfill (3.6) satisfies (3.6) again. In addition, if $f : \mathbb{R} \to \mathbb{R}$ satisfies (3.6) with limit $\tau$, then for $\bar{f} : \mathbb{R} \to \mathbb{R}$, $\bar{f}(x) = |f(x) + c|$, $c \in \mathbb{R}$, it holds that

$$\lim_{x \to \pm\infty} \frac{\bar{f}(x)}{|x|} = \lim_{x \to \pm\infty} \left| \frac{f(x)}{|x|} + \frac{c}{|x|} \right| = |\tau + 0| \in \mathbb{R},$$

such that (3.6) also holds for $\bar{f}$. In total, we can successively conclude that for any neural network with activation function $\sigma$, any component of any layer, as a function $\mathbb{R} \to \mathbb{R}$, satisfies (3.6). Therefore also all functions $f : \mathbb{R} \to \mathbb{R}$ that are represented by a network with activation function $\sigma$ of any depth and width must satisfy (3.6). Clearly, $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x$ is a positive homogeneous function that cannot be approximated by these functions.

So in any case that $|\alpha| = |\beta|$, this contradicts the first part of (a). □

## 3.4 Inverse Problems

### 3.4.1 General Statement

Our goal in this part is the proof of the general inverse problem Theorem 3.5. One ingredient for the proof is Kirszbraun's theorem which is known in functional analysis and measure theory and allows us to extend a Lipschitz continuous function from a subset of $\mathbb{R}^m$ to the entire space.

**Theorem 3.12** (*Kirszbraun's theorem*, Theorem 2.10.43 in [Fed96]). *Let $U \subset \mathbb{R}^n$ and $f : U \to \mathbb{R}^m$ be a Lipschitz continuous function with Lipschitz constant $L$. Then there exists an extension $g : \mathbb{R}^n \to \mathbb{R}^m$ of $f$ with Lipschitz constant $L$.*

Considering positive homogeneous functions however, Kirszbraun's theorem cannot guarantee that this extension will be positive homogeneous again. However, in the following lemma we show that we can circumvent this by first extending the function to the entire space, then restricting it to the unit sphere and then extend it as a positive homogeneous function again. In this way, the Lipschitz constant will increase by a factor of at most 2.

**Lemma 3.13.** *Let $U \subset \mathbb{R}^n$ be non-empty, positive homogeneous and $f : U \to \mathbb{R}^m$ a function that is positive homogeneous and Lipschitz continuous with constant $L$.*

*Then there is an extension $\tilde{f} : \mathbb{R}^n \to \mathbb{R}^m$ of $f$ which is positive homogeneous and Lipschitz continuous with constant $2L$.*

*Proof.* By the positive homogeneity, $0 \in U$. So we can restrict $f$ to $(S^{n-1} \cap U) \cup \{0\}$ where it is still Lipschitz continuous with constant $L$. By Theorem 3.12, we can extend this function to $f : S^{n-1} \cup \{0\} \to \mathbb{R}^m$ on the entire set $S^{n-1} \cup \{0\}$ (by extending to the entire $\mathbb{R}^m$ and then restricting it again) such that it still has Lipschitz constant $L$.

Now we define $\tilde{f} : \mathbb{R}^n \to \mathbb{R}^m$ by $\tilde{f}(x) = \|x\|_2 f(\frac{x}{\|x\|_2})$ for $x \neq 0$ and $\tilde{f}(0) = f(0) = 0$.

Clearly, $\tilde{f}$ is positive homogeneous and an extension of $f$ and we will show that it is Lipschitz continuous with constant $2L$. For this, consider two different points $x, y \in \mathbb{R}^n$. We can assume $\|x\|_2 \leq \|y\|_2$ and thus $y \neq 0$.

Then it holds that

$$\|x - y\|_2 \geq \left\| x - \frac{\|x\|_2}{\|y\|_2} y \right\|_2$$

because

$$\|x - y\|_2^2 \geq \left\| x - \frac{\|x\|_2}{\|y\|_2} y \right\|_2^2$$

$$
\begin{aligned}
\Leftrightarrow \quad & \|x\|_2^2 + \|y\|_2^2 - 2\langle x, y\rangle \geq \|x\|_2^2 + \|x\|_2^2 - 2\frac{\|x\|_2}{\|y\|_2}\langle x, y\rangle \\
\Leftrightarrow \quad & \|y\|_2^2 - \|x\|_2^2 \geq 2\left(1 - \frac{\|x\|_2}{\|y\|_2}\right)\langle x, y\rangle \\
\Leftarrow \quad & \|y\|_2^2 - \|x\|_2^2 \geq 2\left(1 - \frac{\|x\|_2}{\|y\|_2}\right)\|x\|_2\|y\|_2 \\
\Leftrightarrow \quad & \|y\|_2^2 - \|x\|_2^2 \geq 2(\|x\|_2\|y\|_2 - \|x\|_2^2) \\
\Leftrightarrow \quad & \|y\|_2^2 + \|x\|_2^2 - 2(\|x\|_2\|y\|_2) \geq 0 \\
\Leftrightarrow \quad & (\|y\|_2 - \|x\|_2)^2 \geq 0,
\end{aligned}
$$

which is always fulfilled.

If also $x \neq 0$, then we can conclude.

$$
\begin{aligned}
\frac{\|\tilde{f}(x) - \tilde{f}(y)\|_2}{\|x - y\|_2} &\leq \frac{\|\tilde{f}(x) - \tilde{f}(\frac{\|x\|_2}{\|y\|_2}y)\|_2}{\|x - y\|_2} + \frac{\|\tilde{f}(y) - \tilde{f}(\frac{\|x\|_2}{\|y\|_2}y)\|_2}{\|x - y\|_2} \\
&\leq \frac{\|\tilde{f}(x) - \tilde{f}(\frac{\|x\|_2}{\|y\|_2}y)\|_2}{\left\|x - \frac{\|x\|_2}{\|y\|_2}y\right\|_2} + \frac{|\|y\|_2 - \|x\|_2|\,\|\tilde{f}(\frac{y}{\|y\|_2})\|_2}{\|x - y\|_2} \\
&\leq \frac{\|f(\frac{x}{\|x\|_2}) - f(\frac{y}{\|y\|_2})\|_2}{\left\|\frac{x}{\|x\|_2} - \frac{y}{\|y\|_2}\right\|_2} + \frac{\|f(\frac{y}{\|y\|_2}) - f(0)\|_2}{\left\|\frac{y}{\|y\|_2} - 0\right\|_2} \leq 2L,
\end{aligned}
$$

and if otherwise $x = 0$,

$$
\frac{\|\tilde{f}(x) - \tilde{f}(y)\|_2}{\|x - y\|_2} = \frac{\|f(0) - f(\frac{y}{\|y\|_2})\|_2}{\left\|0 - \frac{y}{\|y\|_2}\right\|_2} \leq L,
$$

which completes the proof. □

Now we can use this lemma in the proof for our generalized main theorem for inverse problems.

*Proof of Theorem 3.5.* If there are two $x^{(1)}, x^{(2)} \in U$ with $g(x^{(1)}) = g(x^{(2)})$, then by assumption on $g$, $0 = \|g(x^{(1)}) - g(x^{(2)})\|_2 \geq \tau\|x^{(1)} - x^{(2)}\|_2$. Since $\tau > 0$, this implies that $x^{(1)} = x^{(2)}$. Therefore, $g : U \to g(U)$ is bijective and has an inverse function $g^{-1} = f_0 : g(U) \to U$.

We obtain

$$
\begin{aligned}
\sup_{\substack{y^{(1)}, y^{(2)} \in g(U) \\ y^{(1)} \neq y^{(2)}}} \frac{\|f_0(y^{(1)}) - f_0(y^{(2)})\|_2}{\|y^{(1)} - y^{(2)}\|_2} &= \sup_{\substack{x^{(1)}, x^{(2)} \in U \\ x^{(1)} \neq x^{(2)}}} \frac{\|x^{(1)} - x^{(2)}\|_2}{\|g(x^{(1)}) - g(x^{(2)})\|_2} \\
&= \left(\inf_{\substack{x^{(1)}, x^{(2)} \in U \\ x^{(1)} \neq x^{(2)}}} \frac{\|g(x^{(1)}) - g(x^{(2)})\|_2}{\|x^{(1)} - x^{(2)}\|_2}\right)^{-1} = \frac{1}{\tau}
\end{aligned}
$$

So $f_0 : g(U) \to U$ is Lipschitz continuous with Lipschitz constant $\frac{1}{\tau}$. By Lemma 3.13, there exists a positive homogeneous extension $f : \mathbb{R}^m \to \mathbb{R}^n$ with Lipschitz constant $\frac{2}{\tau}$.

For any $x \in \mathbb{R}^n$, $e \in \mathbb{R}^m$ and any $\epsilon > 0$, there exists an $x' \in U$ such that $\|x - x'\|_{II} \leq d_{II}(x, U) + \epsilon$ and then

$$
\begin{aligned}
\|f(g(x) + e) - x\|_2 \leq \|f(g(x')) - x'\|_2 &+ \|f(g(x)) - f(g(x'))\|_2 \\
&+ \|f(g(x) + e) - f(g(x))\|_2 + \|x - x'\|_2
\end{aligned}
$$

$$\begin{aligned}
&\leq 0 + \frac{2}{\tau}\|g(x) - g(x')\|_2 + \frac{2}{\tau}\|e\|_2 + \|x - x'\|_2 \\
&\leq \frac{2\rho}{\tau}\|x - x'\|_{II} + \|x - x'\|_2 + \frac{2}{\tau}\|e\|_2 \\
&\leq \left(1 + \frac{2\rho}{\tau}\right)\|x - x'\|_{II} + \frac{2}{\tau}\|e\|_2 \\
&\leq \left(1 + \frac{2\rho}{\tau}\right)(d_{II}(x, U) + \epsilon) + \frac{2}{\tau}\|e\|_2,
\end{aligned}$$

where we used that $\|\cdot\|_2 \leq \|\cdot\|_{II}$. Since this holds for all $\epsilon > 0$, we must have

$$\|f(g(x) + e) - x\|_2 \leq \left(1 + \frac{2\rho}{\tau}\right) d_{II}(x, U) + \frac{2}{\tau}\|e\|_2.$$

By equivalence of norms, there exists a number $M > 0$ (that possibly depends on the dimension $n$), such that $\|\cdot\|_{II} \leq M\|\cdot\|_2$. $f$ is positive homogeneous and continuous, so by Theorem 3.7, for each component $f_j$ of $f$, there exists an unbiased ReLU network with 2 hidden layers that approximates $f_j$ up to a relative error of $\frac{1}{\sqrt{n}}\min\{\frac{1}{\tau}, \frac{\delta}{\rho M}\} > 0$. Combining these into one network, we obtain an unbiased ReLU network with 2 hidden layers representing $\tilde{f} : \mathbb{R}^m \to \mathbb{R}^n$ such that for all $y \in \mathbb{R}^m$,

$$\|f(y) - \tilde{f}(y)\|_2 \leq \min\left\{\frac{1}{\tau}, \frac{\delta}{\rho M}\right\}\|y\|_2.$$

Then for all $x \in \mathbb{R}^n$, $e \in \mathbb{R}^m$,

$$\begin{aligned}
\|\tilde{f}(g(x) + e) - x\|_2 &\leq \|\tilde{f}(g(x) + e) - f(g(x) + e)\|_2 + \|f(g(x) + e) - x\|_2 \\
&\leq \min\left\{\frac{1}{\tau}, \frac{\delta}{\rho M}\right\}\|g(x) + e\|_2 + \|f(g(x) + e) - x\|_2 \\
&\leq \frac{\delta}{\rho M}\rho\|x\|_{II} + \frac{1}{\tau}\|e\|_2 + \|f(g(x) + e) - x\|_2 \\
&\leq \delta\|x\|_2 + \left(1 + \frac{2\rho}{\tau}\right) d_{II}(x, U) + \frac{3}{\tau}\|e\|_2,
\end{aligned}$$

where we used $\|x\|_{II} \leq M\|x\|_2$. □

### 3.4.2 Restricted Isometries

The next theorem is an application of the general Theorem 3.5 for the case of a linear measurement function that satisfies a restricted isometry property on the signal set $U$. This includes the usual restricted isometry property for sparse vectors but also other generalizations like the one for low-rank matrices in [CCG15].

**Theorem 3.14.** *Consider norms $\|\cdot\|_I$ on $\mathbb{R}^m$ and $\|\cdot\|_{II}$ on $\mathbb{R}^n$.*

*Let $U \subset \mathbb{R}^n$ be a positive homogeneous subset and $A \in \mathbb{R}^{m\times n}$ a linear map such that there are $\delta^{lb} \in (0, 1), \delta^{ub} \in (0, \infty)$ such that for all $x^{(1)}, x^{(2)} \in U$,*

$$(1 - \delta^{lb})\|x^{(1)} - x^{(2)}\|_2 \leq \|Ax^{(1)} - Ax^{(2)}\|_I \leq (1 + \delta^{ub})\|x^{(1)} - x^{(2)}\|_2. \tag{3.7}$$

*Furthermore, assume that $\|\cdot\|_I \leq \alpha\|\cdot\|_2$ on $\mathbb{R}^m$ for an $\alpha \geq 1$, $\|\cdot\|_2 \leq \|\cdot\|_{II}$ on $\mathbb{R}^n$, and each $x \in \mathbb{R}^n$ can be decomposed as $x = x^{(1)} + \cdots + x^{(M)}$ where $x^{(1)}, \ldots, x^{(M)} \in U$ and $\|x^{(1)}\|_2 + \cdots + \|x^{(M)}\|_2 \leq \|x\|_{II}$.*

*Then, for any $\delta' > 0$, there exists an unbiased* ReLU *network with two hidden layers that represents a function $\tilde{f} : \mathbb{R}^m \to \mathbb{R}^n$ such that for any $x \in \mathbb{R}^n$, $e \in \mathbb{R}^m$,*

$$\|f(Ax + e) - x\|_2 \leq \delta'\|x\|_2 + C\alpha d_{II}(x, U) + D\alpha\|e\|_2.$$

*where* $C = 1 + 2\frac{1+\delta^{ub}}{1-\delta^{lb}}$ *and* $D = \frac{3}{1-\delta^{lb}}$ *and*

$$d_{II}(x, U) := \inf_{x' \in U} \|x - x'\|_{II}.$$

*Proof.* By the assumption on $g$,

$$\tau := \inf_{\substack{x^{(1)}, x^{(2)} \in U \\ x^{(1)} \neq x^{(2)}}} \frac{\|Ax^{(1)} - Ax^{(2)}\|_2}{\|x^{(1)} - x^{(2)}\|_2} \geq \frac{1}{\alpha} \inf_{\substack{x^{(1)}, x^{(2)} \in U \\ x^{(1)} \neq x^{(2)}}} \frac{\|Ax^{(1)} - Ax^{(2)}\|_I}{\|x^{(1)} - x^{(2)}\|_2} \geq \frac{1}{\alpha}(1 - \delta^{lb}) > 0,$$

such that the first assumption of Theorem 3.5 is fulfilled.

For any $x^{(1)}, x^{(2)} \in \mathbb{R}^n$, define $z := x^{(1)} - x^{(2)}$. By assumption, there is a decomposition $z = z^{(1)} + \cdots + z^{(M)}$ such that $z^{(1)}, \ldots, z^{(M)} \in U$ and $\|z^{(1)}\|_2 + \cdots + \|z^{(M)}\|_2 \leq \|z\|_{II}$. Then

$$\|Ax^{(1)} - Ax^{(2)}\|_2 = \|Az\|_2 \leq \sum_{j=1}^{M} \|Az^{(j)}\|_2 \leq (1 + \delta^{ub}) \sum_{j=1}^{M} \|z^{(j)}\|_2 \leq (1 + \delta^{ub})\|x^{(1)} - x^{(2)}\|_{II},$$

and therefore

$$\rho := \sup_{\substack{x^{(1)}, x^{(2)} \in \mathbb{R}^n \\ x^{(1)} \neq x^{(2)}}} \frac{\|Ax^{(1)} - Ax^{(2)}\|_2}{\|x^{(1)} - x^{(2)}\|_{II}} \leq 1 + \delta^{ub}.$$

So we can apply Theorem 3.5 and obtain that for any $\delta' > 0$, there exists a function $\tilde{f} : \mathbb{R}^m \to \mathbb{R}^n$ such that for any $x \in \mathbb{R}^n$ and $e \in \mathbb{R}^m$,

$$\begin{aligned} \|f(Ax + e) - x\|_2 &\leq \delta'\|x\|_2 + \left(1 + \frac{2\rho}{\tau}\right) d_{II}(x, U) + \frac{3}{\tau}\|e\|_2 \\ &\leq \delta'\|x\|_2 + C\alpha d_{II}(x, U) + D\alpha\|e\|_2 \end{aligned}$$

where $C = 1 + 2\frac{1+\delta^{ub}}{1-\delta^{lb}}$ and $D = \frac{3}{1-\delta^{lb}}$.

□

A first immediate consequence from the above theorem is the main result about sparse recovery for matrices with the restricted isometry property.

*Proof of Corollary 3.4.* If $A$ satisfies the $(s, \delta)$-restricted isometry property (for sparse vectors), then (3.7) is fulfilled for $\delta^{lb} = \delta^{ub} = \delta$, $U = \Sigma_s$ and $\|\cdot\|_I = \|\cdot\|_2$. Furthermore, we choose $\|\cdot\|_{II} = \|\cdot\|_1$ such that any $x \in \mathbb{R}^n$ can be decomposed as $x = \sum_{j=1}^n x_j e_j$ where the $e_j \in \mathbb{R}^n$ are the canonical basis vectors. Then clearly, each $x_j e_j \in U$ and $\sum_{j=1}^n \|x_j e_j\|_2 = \sum_{j=1}^n |x_j| = \|x\|_1 = \|x\|_{II}$. Then Theorem 3.14 implies Corollary 3.4. □

Another application of Theorem 3.14 is low-rank matrix recovery. Besides sparse vectors, the inequality (3.7) has also been studied for linear operators on low-rank matrices. Using these results, we can prove the following consequence of Theorem 3.14. It involves the nuclear norm $\|X\|_*$ of a matrix which is defined as the sum of its singular values $\|X\|_* := \sum_{k=1}^{\operatorname{rank}(X)} \sigma_k(X)$.

**Corollary 3.15.** *There are universal constants* $C, D, C_3, c_3 > 0$ *such that the following holds. Let* $A \in \mathbb{R}^{m\times n}$ *have i.i.d. subgaussian entries* $A_{j,k}$ *satisfying*

$$\mathbb{E}[A_{j,k}] = 0 \qquad \mathbb{E}[A_{j,k}^2] = 1 \qquad \mathbb{E}[A_{j,k}^4] > 1.$$

*Define the operator* $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ *such that for all* $X \in \mathbb{R}^{n\times n}$,

$$(\mathcal{A}(X))_j = \sum_{k,l=1}^{n} A_{j,k}A_{j,l}X_{k,l}.$$

*Let $1 \leq r \leq n$ be an integer and $m \geq c_4 nr$. Then with probability $\geq 1 - C_3 e^{-c_3 m}$, the following holds: For any $\delta' > 0$, there exists function $\tilde{f} : \mathbb{R}^m \to \mathbb{R}^{n \times n}$, represented by an unbiased* ReLU *network with* 2 *hidden layers, such that for all $X \in \mathbb{R}^{n \times n}$ and $e \in \mathbb{R}^m$,*

$$\|\tilde{f}(\mathcal{A}(X) + e) - X\|_F \leq \delta' \|X\|_F + C\sqrt{m} d_*(X, U_r) + \frac{D}{\sqrt{m}} \|e\|_2,$$

*where $U_r \subset \mathbb{R}^{n \times n}$ is the set of rank $\leq r$ matrices and $d_*$ denotes the distance in $\| \cdot \|_*$ (nuclear norm).*

*Proof.* According to Corollary 1 in [CCG15], with probability $\geq 1 - C_3 e^{-c_3 m}$, $\mathcal{A}$ satisfies the RIP for low-rank matrices in the sense that for all $X \in \mathbb{R}^{n \times n}$ of rank $\leq 2r$,

$$(1 - \delta^{lb}) \|X\|_F \leq \frac{1}{m} \|\mathcal{A}(X)\|_1 \leq (1 + \delta^{ub}) \|X\|_F$$

for universal constants $\delta^{lb} \in (0, 1)$ and $\delta^{ub} > 0$. Therefore, for any $X^{(1)}, X^{(2)} \in \mathbb{R}^{n \times n}$ of rank $\leq r$,

$$(1 - \delta^{lb}) \|X^{(1)} - X^{(2)}\|_F \leq \frac{1}{m} \|\mathcal{A}(X^{(1)} - X^{(2)})\|_1 \leq (1 + \delta^{ub}) \|X^{(1)} - X^{(2)}\|_F,$$

such that (3.7) is fulfilled for the $\| \cdot \|_F$ norm which corresponds to the $\| \cdot \|_2$ norm of the vectorized matrices and $\| \cdot \|_I = \| \cdot \|_1$ on $\mathbb{R}^m$. Then $\| \cdot \|_I \leq \alpha \| \cdot \|_2$ for $\alpha = \sqrt{m}$. Furthermore, define $U \subset \mathbb{R}^{n \times n}$ to be the set of rank $\leq r$ matrices. Then $U$ is positive homogeneous. Define $\| \cdot \|_{II}$ to be the nuclear norm $\| \cdot \|_*$. Then any matrix $X \in \mathbb{R}^{n \times n}$ has a singular value decomposition $\sum_{j=1}^n \sigma_j u_j v_j^*$ with singular values $\sigma_1, \ldots, \sigma_n$ and orthonormal $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ in $\mathbb{R}^n$. Then every $\sigma_j u_j v_j^*$ is in $U$ and

$$\sum_{j=1}^n \|\sigma_j u_j v_j^*\|_F = \sum_{j=1}^n \sigma_j = \|X\|_* = \|X\|_{II}.$$

Then by Theorem 3.14, for each $\delta' > 0$, there exists a $\hat{\tilde{f}} : \mathbb{R}^m \to \mathbb{R}^n$, represented by a ReLU network with two hidden layers, such that for all $X \in \mathbb{R}^{n \times n}$ and $e \in \mathbb{R}^m$,

$$\|\hat{\tilde{f}}(\frac{1}{m} \mathcal{A}(X) + \frac{e}{m}) - X\|_F \leq \delta' \|X\|_F + C\sqrt{m} d_*(X, U_r) + D\sqrt{m} \|\frac{e}{m}\|_2$$

and thus if we define $\tilde{f}(y) = \hat{\tilde{f}}(\frac{1}{m} y)$, which can also be represented by a ReLU network with two hidden layers,

$$\|\tilde{f}(\mathcal{A}(X) + e) - X\|_F \leq \delta' \|X\|_F + C\sqrt{m} d_*(X, U_r) + \frac{D}{\sqrt{m}} \|e\|_2$$

for $C = 1 + 2 \frac{1 + \delta^{ub}}{1 - \delta^{lb}}$ and $D = \frac{3}{1 - \delta^{lb}}$. □

**Remark 3.16.** *Corollary 3.15 has the error dependence $\frac{1}{\sqrt{m}} \|e\|_2$. This is worse or equal to the dependence $\frac{1}{m} \|e\|_1$ in Theorem 1 of [CCG15]. This is caused by the upper bound $\| \cdot \|_1 \leq \sqrt{m} \| \cdot \|_2$ which is needed because we apply Kirszbraun's theorem for the $\ell_2$ norm. This could be improved if Kirszbraun's theorem also holds for functions on a domain with the $\ell_1$ norm. The same holds for the additional $\sqrt{m}$ factor in the dependence on $d_*(X, U_r)$.*

**Remark 3.17.**
- *In Corollary 3.15, if $r = 1$, the operator $\mathcal{A}$ applied to rank* 1 *matrices of the type $xx^*$ for $x \in \mathbb{R}^n$, yields*

$$(\mathcal{A}(xx^*))_j = \sum_{k,l=1}^n A_{j,k} A_{j,l} x_k x_l = \left| \sum_{k=1}^n A_{j,k} x_k \right|^2 = |(Ax)_j|^2,$$

*so $\mathcal{A}(xx^*) = |Ax|^2$, where $|Ax|^2$ contains the squared absolute values of the entries of $Ax$. So the* ReLU *network function $\tilde{f}$ can (approximately) reconstruct $xx^*$ (and therefore indirectly $x$) from $|Ax|^2$, which is the widely studied phase retrieval problem.*

- *So Corollary 3.15 enables us to solve the phase retrieval problem in the sense that from $|Ax|^2$ we can calculate $xx^*$ using an end-to-end network. One might wonder whether it is also possible to calculate the vector $x$ from $|Ax|^2$ or $|Ax|$ directly. However, the problem is that this $x$ is not unique since for any $|\lambda| = 1$ (thus $\lambda = \pm 1$ in $\mathbb{R}$), $|Ax| = |A\lambda x|$. It is not even possible to define a continuous function $f : \mathbb{R}^m \to \mathbb{R}^n$ such that for each $y$, $|Af(y)| = y$ if $A$ enables a unique solution of the phase retrieval problem up to global phase. To see this, define $g(x) = |Ax|$ which is continuous. If $f$ is continuous, then also $f \circ g$. Let $e_1$ be the first canonical basis vector. If the phase retrieval problem for $A$ is uniquely solvable, then we need to have $f(g(e_1)) = \pm e_1$. Without loss of generality, we can assume $f(g(e_1)) = +e_1$. Now consider $f \circ g$ along the connected line $t \mapsto (1-t)e_1 + te_2$ ($t \in [0, 1]$). By the uniqueness of the solution up to global phase, we must have $(f \circ g)((1-t)e_1 + te_2) = \pm((1-t)e_1 + te_2)$ for each $t \in [0, 1]$. Since we assume that the $\pm 1$ sign is $+1$ for $t = 0$ and always $(1-t)e_1 + te_2 \neq 0$, by continuity we know that $(f \circ g)((1-t)e_1 + te_2) = +((1-t)e_1 + te_2)$ for all $t \in [0, 1]$. Especially, $(f \circ g)(e_2) = e_2$. Analogously, new we can consider $f \circ g$ along the connected line from $e_2$ to $-e_1$ and conclude that $(f \circ g)(-e_1) = -e_1$. However, $g(-e_1) = |A(-e_1)| = g(e_1)$ and therefore also $-e_1 = (f \circ g)(-e_1) = (f \circ g)(e_1)$. This contradicts the previous assumption that $(f \circ g)(e_1) = +e_1$.*

### 3.4.3 Lower Bounds

In this subsection, we prove the lower bounds for sparse recovery (Corollary 3.3, Theorem 3.6) and subsequently also for universal approximation (Theorem 3.8).

*Proof of Theorem 3.6.* Let $f = W_2(W_1x + b_1) + b_2$ be the network function with $W_1 \in \mathbb{R}^{k\times m}$, $W_2 \in \mathbb{R}^{m\times k}$, $b_1 \in \mathbb{R}^k$, $b_2 \in \mathbb{R}^n$.

We first show that it is sufficient to prove the statement for networks with zero biases ($b_1 = 0$ and $b_2 = 0$), because if we scale the signal $x$ with a sufficiently large constant, the biases become irrelevant.

To see more formally that that we can set $b_1 = 0$ and $b_2 = 0$, recall that we denote $\phi$ for the ReLU function. Note that for any numbers $\lambda, a \in \mathbb{R}$ and $\lambda \geq 0$, $\phi(\lambda a) = \lambda\phi(a)$. Note that by this observation and the continuity of $\phi$ and $\|\cdot\|_2$,

$$\begin{aligned}
\sup_{x\in U\setminus\{0\}} \frac{\|x - f(Ax)\|_2}{\|x\|_2} &= \sup_{x\in U\setminus\{0\}} \sup_{\lambda>0} \frac{\|\lambda x - f(A\lambda x)\|_2}{\|\lambda x\|_2} \\
&= \sup_{x\in U\setminus\{0\}} \sup_{\lambda>0} \frac{\|\lambda x - W_2\phi(W_1A\lambda x + b_1) - b_2\|_2}{\|\lambda x\|_2} \\
&= \sup_{x\in U\setminus\{0\}} \sup_{\lambda>0} \frac{\|x - W_2\phi(W_1Ax + \frac{b_1}{\lambda}) - \frac{b_2}{\lambda}\|_2}{\|x\|_2} \\
&\geq \sup_{x\in U\setminus\{0\}} \lim_{\lambda\to\infty} \frac{\|x - W_2\phi(W_1Ax + \frac{b_1}{\lambda}) - \frac{b_2}{\lambda}\|_2}{\|x\|_2} \\
&= \sup_{x\in U\setminus\{0\}} \frac{\|x - W_2\phi(W_1Ax)\|_2}{\|x\|_2}.
\end{aligned}$$

So it is sufficient to prove the statement for a network $f$ with no biases (i.e., with $b_1, b_2 = 0$).

$f$ is defined on $\mathbb{R}^m$ and for a matrix $M \in \mathbb{R}^{m\times m'}$ we define $f(M) \in \mathbb{R}^{n\times m'}$ to be the columnwise application of $f$ on $M$. Then for the matrix $X \in \mathbb{R}^{n\times\tilde{n}}$ whose columns are the vectors

$x_1, \dots, x_{\tilde{n}}$, we obtain

$$\|f(AX) - X\|_F^2 = \sum_{k=1}^{\tilde{n}} \|f(Ax_k) - x_k\|_2^2$$

as the sum of the squared deviations. Now we find a lower bound for $\|f(AX) - X\|_F$.

We observe that

$$f(AX) - f(-AX) = W_2\phi(W_1AX) - W_2\phi(-W_1AX) = W_2\left[\phi(W_1AX) - \phi(-W_1AX)\right].$$

For any $x \in \mathbb{R}$, $\phi(x) - \phi(-x) = x$ such that

$$f(AX) - f(-AX) = W_2W_1AX.$$

Since $A \in \mathbb{R}^{m\times n}$, it has rank $\leq m$ and therefore also $f(AX)-f(-AX)$ has rank $\leq m$. Our goal is to bound $\|f(AX)-X\|_F$ or $\|f(-AX)-(-X)\|_F$ for which we bound $\|f(AX)-f(-AX)-2X\|_F$. By the Eckart-Young-Mirsky theorem, the best rank $m$ approximation of $2X$ in Frobenius norm can be obtained by truncating its sigular value decomposition after the largest $m$ singular values and therefore, for any rank $\leq m$ matrix $M \in \mathbb{R}^{n\times\tilde{n}}$,

$$\|M - 2X\|_F^2 \geq \sum_{k=m+1}^{\tilde{n}} (\sigma_k(2X))^2.$$

Note that this even holds if $X$ itself has rank $\leq m$, in which case $\sigma_k(X) = 0$ for all $m+1 \leq k \leq \tilde{n}$. Since $f(AX) - f(-Ax)$ has rank $\leq m$,

$$\begin{aligned}2\alpha := 2\sqrt{\sum_{k=m+1}^{\tilde{n}} (\sigma_k(X))^2} &\leq \|f(AX) - f(-AX) - 2X\|_F \\ &\leq \|f(AX) - X\|_F + \|f(-AX) - (-X)\|_F \leq 2\max\left\{\|f(AX) - X\|_F, \|f(-AX) - (-X)\|_F\right\}.\end{aligned}$$

So one of these norms on the right hand side is $\geq \alpha$. W.l.o.g. we assume that it is the first one. Then

$$\sum_{k=m+1}^{\tilde{n}} (\sigma_k(X))^2 \leq \|f(AX) - X\|_F^2 = \sum_{k=1}^{\tilde{n}} \|f(Ax_k) - x_k\|_2^2 \leq \tilde{n} \max_{k\in[\tilde{n}]} \|f(Ax_k) - x_k\|_2^2.$$

So we can conclude

$$\sup_{x\in U\setminus\{0\}} \frac{\|f(Ax) - x\|_2}{\|x\|_2} \geq \max_{k\in[\tilde{n}]} \|f(Ax_k) - x_k\|_2 \geq \sqrt{\frac{1}{\tilde{n}} \sum_{k=m+1}^{\tilde{n}} (\sigma_k(X))^2}.$$

□

*Proof of Corollary 3.3.* Corollary 3.3 follows from Theorem 3.6 by choosing $x_1 = e_1, \dots, x_n = e_n$. Then $X = (x_1 \dots x_n) = Id_n$ and $U = \Sigma_1$. The lower bound simplifies to

$$\sqrt{\frac{1}{n} \sum_{k=m+1}^{n} (\sigma_k(Id_n))^2} = \sqrt{\frac{1}{n}(n - m)} = \sqrt{1 - \frac{m}{n}}.$$

□

Using the lower bound for the specific case of sparse recovery, we can also show a lower bound for the general approximation of continuous, positive homogeneous functions.

*Proof of Theorem 3.8.* Let $w_1, \dots, w_n \in \mathbb{R}$ be pairwise distinct. Define the matrix

$$A := \begin{pmatrix} \frac{1}{\sqrt{1+w_1^2}} & \cdots & \frac{1}{\sqrt{1+w_n^2}} \\ \frac{w_1}{\sqrt{1+w_1^2}} & \cdots & \frac{w_n}{\sqrt{1+w_n^2}} \end{pmatrix} \in \mathbb{R}^{2\times n}.$$

All the columns of $A$ have an $\ell_2$-norm of 1. Furthermore, if there exists a 2-sparse $x \in \mathbb{R}^n$ such that $Ax = 0$, then there is a $2 \times 2$-subdeterminant which is 0, i.e., for some $k, l \in [n]$,

$$0 = \det \begin{pmatrix} \frac{1}{\sqrt{1+w_k^2}} & \frac{1}{\sqrt{1+w_l^2}} \\ \frac{w_k}{\sqrt{1+w_k^2}} & \frac{w_l}{\sqrt{1+w_l^2}} \end{pmatrix} = \frac{w_l - w_k}{\sqrt{1+w_k^2}\sqrt{1+w_l^2}}$$

and thus $w_l = w_k$, contradicting the assumption that the numbers are pairwise distinct.

So $Ax \neq 0$ must hold for all 2-sparse $x \in \mathbb{R}^n$. Furthermore, $A$ is injective on $\Sigma_1$. So there exists an inverse map $f : A\Sigma_1 \to \Sigma_1$. The set $\Sigma_1 \cap S^{n-1}$ is compact and thus there exists

$$\tau := \min_{x\in\Sigma_2\cap S^{n-1}} \|Ax\|_2 = \min_{x\in\Sigma_2\setminus\{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \min_{\substack{x,y\in\Sigma_1 \\ x\neq y}} \frac{\|Ax - Ay\|_2}{\|x-y\|_2}$$

and by the previous observation, $\tau > 0$. So for all $x, y \in \Sigma_1$, $x \neq y$, $\|x-y\|_2 \leq \frac{1}{\tau}\|Ax - Ay\|_2$ such that the inverse map $f$ is Lipschitz continuous with Lipschitz constant $\frac{1}{\tau}$.

Furthermore, since $A$ as a function is positive homogeneous, also its restricted inversion $f$ must be positive homogeneous. By Lemma 3.13, there exists a positive homogeneous extension $f : \mathbb{R}^2 \to \mathbb{R}^n$ on the entire space with Lipschitz constant $\frac{2}{\tau}$.

Now let $\tilde{f} : \mathbb{R}^2 \to \mathbb{R}^n$ be any function that can be represented by a ReLU network with one hidden layer. By Corollary 3.3,

$$\sup_{x\in\Sigma_1\setminus\{0\}} \frac{\|x - \tilde{f}(Ax)\|_2}{\|x\|_2} \geq \sqrt{1 - \frac{2}{n}}.$$

Now for each $x \in \Sigma_1$, by the definition of $f$, $f(Ax) = x$ and since $A$ has normalized columns, $\|Ax\|_2 = \|x\|_2$. Therefore we can conclude

$$\begin{aligned} \sqrt{1-\frac{2}{n}} &\leq \sup_{x\in\Sigma_1\setminus\{0\}} \frac{\|x - \tilde{f}(Ax)\|_2}{\|x\|_2} = \sup_{x\in\Sigma_1\setminus\{0\}} \frac{\|f(Ax) - \tilde{f}(Ax)\|_2}{\|Ax\|_2} \\ &= \sup_{y\in A\Sigma_1\setminus\{0\}} \frac{\|f(y) - \tilde{f}(y)\|_2}{\|y\|_2} \leq \sup_{y\in\mathbb{R}^2\setminus\{0\}} \frac{\|\tilde{f}(y) - f(y)\|_2}{\|y\|_2}. \end{aligned}$$

$f$ is positive homogeneous and Lipschitz continuous, thus also continuous. We can expand $f : \mathbb{R}^2 \to \mathbb{R}^n$ to a function $\bar{f} : \mathbb{R}^m \to \mathbb{R}^n$ by setting $\bar{f}(y_1, \dots, y_m) = f(y_1, y_2)$. In this way, $\bar{f}$ is still positive homogeneous and continuous. For any $\tilde{\tilde{f}} : \mathbb{R}^m \to \mathbb{R}^n$ that is represented by a ReLU network with one hidden layers, also $\tilde{f} : \mathbb{R}^2 \to \mathbb{R}^n$, $\tilde{f}(y) = \tilde{\tilde{f}}(y_1, y_2, 0, \dots, 0)$ can be represented by a ReLU network with one hidden layer such that

$$\begin{aligned} \sup_{\tilde{y}\in\mathbb{R}^m\setminus\{0\}} \frac{\|\tilde{\tilde{f}}(\tilde{y}) - \bar{f}(\tilde{y})\|_2}{\|\tilde{y}\|_2} &\geq \sup_{y\in\mathbb{R}^2\setminus\{0\}} \frac{\|\tilde{\tilde{f}}(y_1, y_2, 0, \dots, 0) - \bar{f}(y_1, y_2, 0, \dots, 0)\|_2}{\|y\|_2} \\ &= \sup_{y\in\mathbb{R}^2\setminus\{0\}} \frac{\|\tilde{f}(y) - f(y)\|_2}{\|y\|_2} \geq \sqrt{1-\frac{2}{n}}. \end{aligned} \tag{3.8}$$

Now take an $n' \leq n$ and let $\bar{f} : \mathbb{R}^m \to \mathbb{R}^n$ as above such that (3.8) holds for all ReLU networks $\tilde{\tilde{f}}$ with one hidden layer. For each subset $S \subset [n]$ of size $|S| = n'$, we can define a function $\bar{f}_S : \mathbb{R}^m \to \mathbb{R}^n$ such that for all $y \in \mathbb{R}^m$, $j \in [n]$,

$$(\bar{f}_S(y))_j := \begin{cases} (\bar{f}(y))_j & \text{if } j \in S \\ 0 & \text{otherwise.} \end{cases}$$

Now assume that every continuous, positive homogeneous function $\mathbb{R}^m \to \mathbb{R}^{n'}$ can be approximated by a one-layer ReLU network up to relative precision $n'(\frac{1}{n} - \frac{2}{n^2}) - \epsilon$ for an $\epsilon > 0$. Then for each $S \subset [n]$, $|S| = n'$, there exists a ReLU network function $\tilde{\tilde{f}}_{(S)} : \mathbb{R}^m \to \mathbb{R}^n$ such that $(\tilde{\tilde{f}}_{(S)})_j = 0$ for $j \in [n] \backslash S$ and for all $y \in \mathbb{R}^m$,

$$\frac{\|\tilde{\tilde{f}}_{(S)}(y) - \bar{f}_S(y)\|_2^2}{\|y\|_2^2} \le n' \left( \frac{1}{n} - \frac{2}{n^2} \right) - \epsilon.$$

Since every $j \in [n]$ is contained in exactly $\binom{n-1}{n'-1}$ subsets $S \subset [n]$ of cardinality $|S| = n'$, for each $y \in \mathbb{R}^m$, $j \in [n]$,

$$\bar{f}_j(y) = \frac{1}{\binom{n-1}{n'-1}} \sum_{\substack{S \subset [n] \\ \text{s.t. } |S| = n'}} (\bar{f}_S(y))_j.$$

We define the ReLU network function $\tilde{\tilde{f}} : \mathbb{R}^m \to \mathbb{R}^n$ with one hidden layer by

$$\tilde{\tilde{f}}(y) = \frac{1}{\binom{n-1}{n'-1}} \sum_{\substack{S \subset [n] \\ \text{s.t. } |S| = n'}} \tilde{\tilde{f}}_{(S)}(y).$$

Then by (3.8), we have

$$\sup_{y \in \mathbb{R}^m \backslash \{0\}} \frac{\|\tilde{\tilde{f}}(y) - \bar{f}(y)\|_2^2}{\|y\|_2^2} \ge 1 - \frac{2}{n}$$

On the other hand, for all $y \in \mathbb{R}^m \backslash \{0\}$,

$$\begin{aligned}
\frac{\|\tilde{\tilde{f}}(y) - \bar{f}(y)\|_2}{\|y\|_2} = \sum_{j \in [n]} \frac{|\tilde{\tilde{f}}_j(y) - \bar{f}_j(y)|^2}{\|y\|_2} &\le \frac{1}{\binom{n-1}{n'-1}} \sum_{j \in [n]} \sum_{\substack{S \subset [n] \\ \text{s.t. } |S| = n'}} \frac{|(\tilde{\tilde{f}}_{(S)}(y))_j - (\bar{f}_S(y))_j|^2}{\|y\|_2} \\
&\le \frac{1}{\binom{n-1}{n'-1}} \sum_{\substack{S \subset [n] \\ \text{s.t. } |S| = n'}} \sum_{j \in S} \frac{|(\tilde{\tilde{f}}_{(S)}(y))_j - (\bar{f}_S(y))_j|^2}{\|y\|_2} \\
&= \frac{1}{\binom{n-1}{n'-1}} \sum_{\substack{S \subset [n] \\ \text{s.t. } |S| = n'}} \frac{\|\tilde{\tilde{f}}_{(S)}(y) - \bar{f}_S(y)\|_2^2}{\|y\|_2} \\
&\le \frac{\binom{n}{n'}}{\binom{n-1}{n'-1}} \left[ n' \left( \frac{1}{n} - \frac{2}{n^2} \right) - \epsilon \right] = \frac{n}{n'} \left[ n' \left( \frac{1}{n} - \frac{2}{n^2} \right) - \epsilon \right] = 1 - \frac{2}{n} - \epsilon \frac{n}{n'} < 1 - \frac{2}{n}.
\end{aligned}$$

This is a contradiction. So we can concluded that for $n' \le n$, there exists a function $f : \mathbb{R}^m \to \mathbb{R}^{n'}$ such that for all one-layer ReLU networks $\tilde{f} : \mathbb{R}^m \to \mathbb{R}^{n'}$,

$$\sup_{y \in \mathbb{R}^m \backslash \{0\}} \frac{\|\tilde{f}(y) - f(y)\|_2^2}{\|y\|_2^2} \ge n' \left( \frac{1}{n} - \frac{2}{n^2} \right).$$

The second factor $\left( \frac{1}{n} - \frac{2}{n^2} \right)$ becomes maximal for $n = 4$. For $n' \le 4$, we choose $n = 4$ and otherwise $n = n'$, which proves the bound from the theorem statement.

□

## 3.5 Networks from Optimization Based Approaches

Other than the neural network approaches of this work, classical compressed sensing studies optimization based methods to solve this problem, see Chapter 4 in [FR13] for an overview. In contrast to the previous result in this work, which is independent of this, in this section we solve the sparse recovery problem by approximating the solution of an optimization problem with a neural network. In particular, we recall the $\ell_1$ minimization problem (0.4)

$$\min \|z\|_1 \qquad\qquad \text{s.t. } \|Az - y\|_2 \leq \eta, \tag{3.9}$$

for an $\eta \geq 0$, which can be shown to give a stable reconstruction of sparse $x$ from their measurements $y = Ax$ for suitable measurement matrices $A$ (Theorem 0.3).

To show that (3.9) can be solved using a ReLU network, we need to show that the function that maps vectors $y$ to the corresponding minimizer in (3.9) is continuous. However, it is not clear that for each $y$, (3.9) has a unique solution. In fact, previous works such as [ZYY16] have shown uniqueness under certain circumstances but this might not be the case in general.

This leads to the concept of multifunctions. Unlike a usual function $f : X \to Y$ that maps each $x \in X$ to exactly one $f(x) \in Y$, a multifunction maps each $x \in X$ to a subset of $Y$ while usually the empty set is excluded. Therefore, a multifunction can also be seen as a function $F : X \to 2^Y \backslash \{\emptyset\}$ with values in the power set $2^Y$ of $Y$. Using this concept, we can always describe the map from the vector $y$ to the set of minimizers of (3.9) as a multifunction if the minimization problem is feasible. We will show that this multifunction satisfies a generalization of continuity and that eventually there is a continuous selection function that maps every $y$ to an approximation of *one* of the solutions of (3.9).

There exists an extensive theory about multifunctions and generalizations of well-known concepts of functions to them. This has been known as *set-valued analysis* or *multivalued analysis* and textbooks such as [AF09] and [HP97] can provide a detailed summary of this. In the following presentation of the most important concepts, we mostly use the notation and terminology of [HP97].

**Definition 3.18.** *Let $X, Y$ be sets. A multifunction $F : X \to 2^Y \backslash \{\emptyset\}$ is a function that maps from $M$ to the power set $2^Y$ of $Y$ without the empty set.*

One important tool which we need in this section is the generalization of continuity to multifunctions. The following properties have also been known under the terms "upper/lower *hemi*continuous" in the literature.

**Definition 3.19** (Definition 2.3 / Remark 2.4 in [HP97])**.** *Let $F : X \to 2^Y \backslash \{\emptyset\}$ be a multifunction between Hausdorff topological spaces $X$ and $Y$. For $x_0 \in X$, we say that*

- *$F$ is upper semicontinuous at $x_0$ if for all open sets $V \subset Y$ with $F(x_0) \subset V$, there exists a neighborhood $U$ of $x_0$ such that for all $x \in U$, $F(x) \subset V$,*
- *$F$ is lower semicontinuous at $x_0$ if for all open sets $V \subset Y$ with $F(x_0) \cap V \neq \emptyset$, there exists a neighborhood of $x_0$ such that for all $x \in U$, $F(x) \cap V \neq \emptyset$,*
- *$F$ is continuous at $x_0$ if $F$ is upper semicontinuous at $x_0$ and lower semicontinuous at $x_0$.*

*We say that $F$ is (upper/lower semi-)continuous if it is (upper/lower semi-)continuous at all points $x_0 \in X$.*

For single-valued functions, i.e., multifunctions $F$ such that $|F(x)| = 1$ for all $x$ in the domain, all these terms coincide to the usual term of continuity of functions. A simple standard example for a multivalued function that is upper but not lower semicontinuous is given by $F_1 : \mathbb{R} \to 2^{\mathbb{R}} \backslash \{\emptyset\}$, $F_1(x) = \{1\}$ if $x \neq 0$ and $F_1(0) = [0, 1]$. A lower but not upper semicontinuous function is given by $F_2 : \mathbb{R} \to 2^{\mathbb{R}} \backslash \{\emptyset\}$, $F_2(x) = [0, 1]$ if $x \neq 0$ and $F_2(0) = \{0\}$. Example 2.8

in [HP97] gives more details and explanations about this. So in general, none of these two properties implies the other.

In optimization problems such as (3.9), the feasible region is a multifunction of the parameters (here $y$). Berge's maximum theorem states that if this feasible region is a continuous multifunction and the objective function is a continuous function, then the multifunction mapping the parameters to the set of optimal solutions is upper semicontinuous.

**Theorem 3.20** (Berge's maximum theorem, Theorem 3.4 in [HP97])**.** *Let $u : X \times Y \to \mathbb{R}$ be a continuous function and $F : Y \to 2^X \backslash \{\emptyset\}$ a continuous multifunction with compact values. Consider the optimization problem*

$$\max_x u(x,y) \qquad\qquad s.t.\ x \in F(y).$$

*Let $S : Y \to 2^X \backslash \{\emptyset\}$ be the multifunction mapping each $y \in Y$ to the optimizers and $v : Y \to \mathbb{R}$ be the function mapping each $y$ to the optimal value.*

*Then $S$ is upper continuous with compact values and $v$ is continuous.*

For a multifunction $F : X \to 2^Y \backslash \{\emptyset\}$, a *selection* is defined as a single-valued function $f : X \to Y$ such that $f(x) \in F(x)$ for all $x \in X$. A particular question that has been studied is when there exists a continuous selection of $F$. Michael's selection theorem (Theorem 4.6 in [HP97]) states that lower semicontinuity of $F$ is enough for this. However, the above Theorem 3.20 can only guarantee upper semicontinuity which is not sufficient for a continuous selection (also see [HP97]). Nevertheless, upper semicontinuous multifunctions still allow approximate selections with arbitrarily small perturbations in the argument and in the function value. In the following theorem, we use the notation $F(M)$ with a multifunction $F : X \to 2^Y$ and a subset $M \subset X$ for $F(M) := \bigcup_{x \in M} F(x)$.

**Theorem 3.21** (Theorem 4.42 in [HP97])**.** *Let $X$ be a metric space, $Y$ a Banach space, $W \subset X$ open, $K \subset W$ compact, $F : \bar{W} \to 2^Y \backslash \{\emptyset\}$ an upper semicontinuous multifunction with convex values, then for every $\epsilon > 0$, there is an open neighborhood $G_\epsilon$ of $K$ and a locally Lipschitz function $f_\epsilon : G_\epsilon \to \mathrm{conv} F(K)$ with finite dimensional range such that for every $x \in G_\epsilon$, $f_\epsilon(x) \in F(K \cap B_\epsilon(x)) + B_\epsilon(0)$.*

Now in order to apply the aforementioned results to the particular problem of approximating the solution of (3.9) with a continuous function, the first step is to show that the feasible region of the problem is described by a continuous multifunction. We show this for a slight restriction of the feasible region which is compact but this will not change the eventual minimizer set.

**Lemma 3.22.** *Let $A \in \mathbb{R}^{k \times n}$ ($k \leq n$), $B \in \mathbb{R}^{k \times m}$, and $\eta \in [0, \infty)$. Let $\| \cdot \|$ be a norm on $\mathbb{R}^k$ and $g : \mathbb{R}^m \to [1, \infty)$ a continuous function.*

*Define the multifunctions $F_1, F_2, F : \mathbb{R}^m \to 2^{\mathbb{R}^n} \backslash \{\emptyset\}$ by*

$$\begin{aligned} F_1(y) &= \{x \in \mathbb{R}^n \mid \|Ax + By\| \leq \eta\}, \\ F_2(y) &= \{x \in \mathbb{R}^n \mid \|P_{\ker(A)} x\|_2 \leq g(y)\}, \\ F(y) &= F_1(y) \cap F_2(y), \end{aligned}$$

*where $P_{\ker(A)}$ is the orthogonal projection onto the kernel of $A$.*

*Furthermore, assume that*

- *$g(y) \geq \|A^\dagger B y\|_2$ for all $y \in \mathbb{R}^m$*
- *$B(\mathbb{R}^m) \subset A(\mathbb{R}^n)$*

*Then $F$ is well-defined (i.e., $F(y) \neq \emptyset$ for all $y \in \mathbb{R}^m$), continuous and has compact values.*

*Proof.* The condition that the range of $B$ is contained in the range of $A$ implies that

$$AA^\dagger B = P_{A(\mathbb{R}^n)}B = B,$$

where $P_{A(\mathbb{R}^n)}$ is the orthogonal projection onto the range of $A$.

Furthermore, by the equivalence of all norms on $\mathbb{R}^k$, there exists a constant $C > 0$ such that $\|z\|_2 \leq C\|z\|$ holds for all $z \in \mathbb{R}^k$.

**Step 1: $F$ is well-defined:** Since $AA^\dagger B = B$, for each $y \in \mathbb{R}^m$, $\|A(-A^\dagger By) + By\| = 0$, such that $-A^\dagger By \in F_1(y)$ and by assumption $\| - A^\dagger By\|_2 \leq g(y)$ such that also $-A^\dagger By \in F_2(y)$ and thus $F(y) \neq \emptyset$.

**Step 2: $F$ has compact values:** For each $y$, $F_1(y)$ and $F_2(y)$ are closed such that $F(y)$ is closed. Furthermore, note that $A^\dagger A = P_{\ker(A)^\perp}$. So for any $y \in \mathbb{R}^m$, $x \in F(y)$, we obtain

$$\begin{aligned}\|x\|_2 &\leq \|P_{\ker(A)}x\|_2 + \|A^\dagger Ax\|_2 \leq g(y) + \|A^\dagger(Ax + By)\|_2 + \|A^\dagger By\|_2 \\ &\leq g(y) + \|A^\dagger\|_{2\to2}\|Ax + By\|_2 + \|A^\dagger By\|_2 \\ &\leq g(y) + \|A^\dagger\|_{2\to2}C\eta + \|A^\dagger By\|_2.\end{aligned}$$

The right hand side does not depend on $x$ and therefore, $F(y)$ is also bounded and thus compact.

**Step 3: $F$ is lower semicontinuous**

Let $y_0 \in \mathbb{R}^m$ and $V \subset \mathbb{R}^n$ open such that there exists an $x_0 \in F(y_0) \cap V$. Since $V$ is open, there exists a radius $\epsilon > 0$ such that $B_\epsilon(x_0) \subset V$.

Since $g$ is continuous, there is a $\tilde{\delta} > 0$ such that for all $y \in B_{\tilde{\delta}}(y_0)$, $|g(y) - g(y_0)| < \frac{\epsilon}{4}$. Now choose $\delta := \min\{\tilde{\delta}, \frac{\epsilon}{4\|A^\dagger B\|_{2\to2}}\} > 0$ if $A^\dagger B \neq 0$ and $\delta = \tilde{\delta}$ otherwise. Let $y \in B_\delta(y_0)$. We define the number $\lambda \in (0, 1]$ by

$$\lambda := \begin{cases} 1 & \text{if } \|P_{\ker(A)}x_0\|_2 \leq \frac{\epsilon}{4} \\ \frac{\epsilon}{4\|P_{\ker(A)}x_0\|_2} & \text{otherwise,} \end{cases}$$

such that we always have $1 - \lambda \geq 0$ and $\lambda\|P_{\ker(A)}x_0\|_2 \leq \frac{\epsilon}{4}$. Then we define

$$x := x_0 + A^\dagger B(y_0 - y) - \lambda P_{\ker(A)}x_0. \tag{3.10}$$

We observe $\|Ax + By\| = \|Ax_0 + AA^\dagger B(y_0 - y) - 0 + By\| = \|Ax_0 + By\| \leq \eta$, so $x \in F_1(y)$. Furthermore,

$$\begin{aligned}\|P_{\ker(A)}x\|_2 &= \|(1 - \lambda)P_{\ker(A)}x_0\|_2 \\ &= \|P_{\ker(A)}x_0\|_2 - \lambda\|P_{\ker(A)}x_0\|_2 \begin{cases} = 0 \leq g(y) & \text{if } \lambda = 1 \\ \leq g(y_0) - \frac{\epsilon}{4} < g(y) & \text{otherwise,} \end{cases}\end{aligned}$$

showing that $x \in F_2(y)$, i.e., $x \in F(y)$. We also obtain

$$\|x - x_0\|_2 = \|A^\dagger B(y_0 - y) - \lambda P_{\ker(A)}x_0\|_2 \leq \|A^\dagger B\|_{2\to2}\delta + \lambda\|P_{\ker(A)}x_0\|_2 \leq \frac{\epsilon}{4} + \frac{\epsilon}{4} < \epsilon.$$

This implies $x \in B_\epsilon(x_0) \subset V$. Therefore, $F(y) \cap V \neq \emptyset$ for any $y \in B_\delta(y_0)$. This shows that $F$ is lower semicontinuous.

**Step 4: $F$ is upper semicontinuous**

Let $y_0 \in \mathbb{R}^m$ and take an open set $V \subset \mathbb{R}^n$ such that $F(y_0) \subset V$.

For points $y \in \mathbb{R}^m$, define the distance $d(y, F(y_0)) = \min_{y' \in F(y_0)} \|y - y'\|_2$. Since $F(y_0)$ is compact, this minimum always exists.

Assume that for each integer $k \geq 1$, there is a $y_k \in \mathbb{R}^m \backslash V$ such that $d(y_k, F(y_0)) \leq \frac{1}{k}$. Then $(y_k)$ forms a sequence in the compact set $F(y_0) + \bar{B}_1(0)$. Therefore, it has a convergent subsequence $(y_{k_l})$ with limit $\bar{y}$. By continuity, $d(\bar{y}, F(y_0)) = 0$ such that by compactness $\bar{y} \in$

$F(y_0)$. On the other hand, $\mathbb{R}^m \backslash V$ is closed such that $\bar{y} \in \mathbb{R}^m \backslash V$. This contradicts the assumption that $F(y_0) \subset V$. Therefore, there is a radius $\epsilon > 0$ such that all $y \in \mathbb{R}^m$ with $d(y, F(y_0)) < \epsilon$ belong to $V$, i.e., $F(y_0) + B_\epsilon(0) \subset V$.

The rest of the argument is similar to the proof of lower semicontinuity. Since $g$ is continuous, there is a $\tilde{\delta} > 0$ such that for all $y \in B_{\tilde{\delta}}(y_0)$, $|g(y) - g(y_0)| < \frac{\epsilon}{4}$. Now choose $\delta := \min\{\tilde{\delta}, \frac{\epsilon}{4\|A^\dagger B\|_{2\to 2}}\} > 0$ if $A^\dagger B \neq 0$ and $\delta = \tilde{\delta}$ otherwise. Let $y \in B_\delta(y_0)$ and take any $x \in F(y)$. We define the number $\lambda \in (0, 1]$ by

$$\lambda := \begin{cases} 1 & \text{if } \|P_{\ker(A)} x\|_2 \leq \frac{\epsilon}{4} \\ \frac{\epsilon}{4\|P_{\ker(A)} x\|_2} & \text{otherwise,} \end{cases}$$

such that we always have $1 - \lambda \geq 0$ and $\lambda \|P_{\ker(A)} x\|_2 \leq \frac{\epsilon}{4}$. Then we define

$$\bar{x} := x + A^\dagger B(y - y_0) - \lambda P_{\ker(A)} x.$$

Note that this definition is analogous to (3.10) in the proof of the lower semicontinuity. Therefore, we can follow the same subsequent steps and prove that $\bar{x} \in F(y_0)$ and $\|\bar{x} - x\|_2 < \epsilon$.

This implies $x \in F(y_0) + B_\epsilon(0) \subset V$. Since this holds for any $x \in F(y)$, $F(y) \subset V$ for any $y \in B_\delta(y_0)$, proving that $F$ is upper semicontinuous. □

Now, using the continuity of the feasible region from above, we can use the tools from multivalued analysis to show that for a certain class of minimization problems, there exists a continuous function whose values are approximate optimal solutions.

**Lemma 3.23.** *Take $\tilde{A} \in \mathbb{R}^{k\times n}$, $\tilde{B} \in \mathbb{R}^{k\times m}$, $\eta \in [0, \infty)$, a norm $\|\cdot\|$ on $\mathbb{R}^k$, and a continuous function $u : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. Consider the optimization problem*

$$\min_{z\in\mathbb{R}^n} u(z, y) \qquad \text{s.t. } \|\tilde{A}z + \tilde{B}y\| \leq \eta, \tag{3.11}$$

*where $\tilde{B}(\mathbb{R}^m) \subset \tilde{A}(\mathbb{R}^n)$. Furthermore, assume that there exists a continuous function $\bar{g} : \mathbb{R}^m \to \mathbb{R}$ and a coefficient $\alpha \in (0, \infty)$, such that for all $y \in \mathbb{R}^m$ and feasible $z \in \mathbb{R}^n$,*

$$\|z\|_2 \leq \alpha u(z, y) + \bar{g}(y). \tag{3.12}$$

*Let $\|\cdot\|_I$ be a norm on $\mathbb{R}^m$.*

*For each $\epsilon > 0$ and each compact $V \subset \mathbb{R}^m$, there is a function $\tilde{f} : V \to \mathbb{R}^n$, represented by a* ReLU *network with one hidden layer, such that for all $y \in V$, there is a $\tilde{y} \in V$ and a solution $\tilde{x} \in \mathbb{R}^n$ of* (3.11) *for $\tilde{y}$, such that $\|y - \tilde{y}\|_2 < \epsilon$ and $\|\tilde{f}(y) - \tilde{x}\|_I < \epsilon$.*

*Proof.* First we define the continuous function $g(y) := \bar{g}(y) + \alpha u(-\tilde{A}^\dagger \tilde{B} y, y)$ and consider the corresponding multifunction $F$ defined in Lemma 3.22 (with the matrices $\tilde{A}$ and $\tilde{B}$). Then every minimizer of

$$\min u(z, y) \qquad \text{s.t. } z \in F(y) \tag{3.13}$$

also minimizes (3.11).

Assume that this is not the case and there is a minimizer $\tilde{x}$ of (3.13) that does not minimize (3.11). Then there is an optimal solution $\hat{x}$ to (3.11) (because of $\tilde{B}(\mathbb{R}^m) \subset \tilde{A}(\mathbb{R}^n)$, it is always feasible) with $u(\hat{x}, y) < u(\tilde{x}, y)$. So $\hat{x}$ cannot be feasible for (3.13), i.e., $\hat{x} \notin F(y)$. On the other hand, $-\tilde{A}^\dagger \tilde{B} y$ is feasible for (3.13). This implies

$$\|P_{\ker(\tilde{A})} \hat{x}\|_2 \leq \|\hat{x}\|_2 \leq \alpha u(\hat{x}, y) + \bar{g}(y) \leq \alpha u(\tilde{x}, y) + \bar{g}(y) \leq \alpha u(-\tilde{A}^\dagger \tilde{B} y, y) + \bar{g}(y) = g(y)$$

and therefore $\hat{x}$ is also feasible for (3.13), which contradicts the above observation.

Now by Lemma 3.22, $F$ is continuous with compact values such that by Theorem 3.20, the solution multifunction of (3.13), $S : \mathbb{R}^m \to 2^{\mathbb{R}^n}\backslash\{\emptyset\}$ is upper semicontinuous with compact values and the optimal value function $v : \mathbb{R}^m \to \mathbb{R}$ is continuous.

We apply the approximate selection Theorem 3.21. As a domain, consider the metric space $V$ endowed with the $\|\cdot\|_2$ norm. Then $V$ is an open and compact subset of itself. The space $\mathbb{R}^n$ endowed with the $\|\cdot\|_I$ norm is finite-dimensional and therefore a Banach space. We have shown that $S : \mathbb{R}^m \to 2^{\mathbb{R}^n}\backslash\{\emptyset\}$ is upper semicontinuous with convex values and this remains the case if we restrict $S$ to the metric space $V$ (whose topology is the subspace topology of $\mathbb{R}^m$). Therefore by Theorem 3.21, for every $\epsilon > 0$, there exists a continuous function $f : V \to \mathbb{R}^n$ such that for every $y \in V$, $f(y) \in S(B_{\epsilon/2}(y)) + B_{\epsilon/2}(0)$. This means that there exists $\tilde{y} \in V$ and $\tilde{x} \in S(\tilde{y})$ such that $\|y - \tilde{y}\|_2 < \frac{\epsilon}{2}$ and $\|f(y) - \tilde{x}\|_I < \frac{\epsilon}{2}$.

By the universal approximation theorem for compact sets (Theorem 3.1 for each coordinate of $f$), for each $\epsilon$, there exists a ReLU network with one hidden layer that represents $\tilde{f} : V \to \mathbb{R}^n$, such that for all $y \in V$, $\|\tilde{f}(y) - f(y)\|_I < \frac{\epsilon}{2}$. Then for all $y \in V$, there exists a $\tilde{y} \in V$ and $\tilde{x} \in S(\tilde{y})$ such that $\|y - \tilde{y}\|_2 < \frac{\epsilon}{2} < \epsilon$ and $\|\tilde{f}(y) - \tilde{x}\|_I \leq \|\tilde{f}(y) - f(y)\|_I + \|f(y) - \tilde{x}\|_I < \epsilon$. □

**Remark 3.24.** *Lemma 3.23 can be applied to the following optimization problems that have been used to solve the sparse recovery problem, i.e., recovering $x$ from $y = Ax + e$. Some of these approaches have already been mentioned in Section 0.3. An overview with a detailed explanation of the following techniques can be found in Section 3.1 in [FR13].*

- *Quadratically constrained basis pursuit:*

$$\min_{z\in\mathbb{R}^n} \|z\|_1 \qquad \text{s.t. } \|Az - y\|_2 \leq \eta \tag{3.14}$$

  *for $\|e\|_2 \leq \eta$. Here $u(z, y) = \|z\|_1$ is continuous, $\tilde{A} = A$, and $\tilde{B} = Id_m$. So Lemma 3.23 can be applied if* rank$(A) = m$. *If this is not the case and* rank$(A) = m' < m$, *we can replace $A$ by $PA \in \mathbb{R}^{m'\times n}$ where $P \in \mathbb{R}^{m'\times m}$ is a bijective and orthogonal map from $A(\mathbb{R}^n)$ to $\mathbb{R}^{m'}$. Then $PA$ satisfies the same RIP as $A$.*

  *Note that also the condition* (3.12) *is fulfilled since for all feasible $z$, $\|z\|_2 \leq \|z\|_1 = 1 \cdot u(z, y) + 0$.*

- *Basis pursuit denoising:*

$$\min_{z\in\mathbb{R}^n} \lambda\|z\|_1 + \|Az - y\|_2^2 \tag{3.15}$$

  *for a parameter $\lambda > 0$. For each feasible $z$, $\|z\|_2 \leq \frac{1}{\lambda}u(z, y)$ such that* (3.12) *is fulfilled for $\alpha = \frac{1}{\lambda}$. Again the objective function is continuous and we can apply Lemma 3.23 for $\tilde{A} = 0 \in \mathbb{R}^{1\times n}$ and $\tilde{B} = 0 \in \mathbb{R}^{1\times m}$.*

- *LASSO:*

$$\min_{z\in\mathbb{R}^n} \|Az - y\|_2 \qquad \text{s.t. } \|z\|_1 \leq \tau \tag{3.16}$$

  *for a parameter $\tau \geq 0$.*(3.12) *is fulfilled since for all feasible $z$, $\|z\|_2 \leq \|z\|_1 \leq u(z, y) + \tau = u(z, y) + \bar{g}(y)$ for the continuous function $\bar{g}(y) = \tau$. Lemma 3.23 can be applied again for $\tilde{A} = Id_n$ and $\tilde{B} = 0 \in \mathbb{R}^{n\times m}$.*

- *Dantzig selector:*

$$\min_{z\in\mathbb{R}^n} \|z\|_1 \qquad \text{s.t. } \|A^*(Az - y)\|_\infty \leq \eta \tag{3.17}$$

  *for a parameter $\tau \geq 0$ for $\|A^*e\|_\infty \leq \eta$. Here $\tilde{A} = A^*A$, $\tilde{B} = A^*$ and therefore $\tilde{A}(\mathbb{R}^n) = A^*(\mathbb{R}^m) = \tilde{B}(\mathbb{R}^m)$, so Lemma 3.23 can be applied again.* (3.12) *is fulfilled for the same reason as in* (3.14).

**Remark 3.25.** *Lemma 3.23 provides an approximate selection of solutions of the optimization problem* (3.11)*, i.e.,* $\tilde{f}(y)$ *is close to an optimal solution of* (3.11) *for a parameter that is close to* $y$*. To show this, we used the approximate selection Theorem 3.21.*

*One might wonder whether there exists a continuous exact selection, i.e., a continuous function* $f : \mathbb{R}^m \to \mathbb{R}^n$ *such that for each* $y \in \mathbb{R}^m$*,* $f(y)$ *is exactly a solution of* (3.11) *for the parameter* $y$*. Indeed, in the field of multivalued analysis, Michael's selection theorem (Theorem 4.6 in [HP97]) can guarantee the existence of a continuous selection of a multifunction. It requires this multifunction (i.e. the multifunction of solutions of* (3.11) *in our application) to be lower semicontinuous. However, Berge's maximum theorem (Theorem 3.20) can only guarantee upper semicontinuity for the solution function.*

*Indeed, the following example shows that actually not in all cases in which Lemma 3.23 can be applied, an exact continuous selection exists. Consider the continuous function* $u : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$

$$u(z,y) = \left\| \begin{pmatrix} y_1 & 0 \\ 0 & y_2 \end{pmatrix} z \right\|_1 + \max\{2, \|z\|_2\}$$

*and the minimization problem*

$$\min_{z \in \mathbb{R}^2} u(z,y) \qquad \text{s.t. } \left| \begin{pmatrix} 1 & 1 \end{pmatrix} z - \begin{pmatrix} 1 & 0 \end{pmatrix} y \right| \leq 0, \tag{3.18}$$

*which satisfies the requirements of Lemma 3.23 (including* (3.12)*).*

*We are interested in the case* $y \in \{1\} \times (0,2)$*. The* $\max\{2, \|z\|_2\}$ *condition is only required to ensure* (3.12) *but it will not change the optimal solutions in this case. To see this, consider the minimization problem without the* $\max\{2, \|z\|_2\}$*, i.e., with* $u(z,y)$ *replaced by* $\tilde{u}(z,y) = u(z,y) - 2\max\{2, \|z\|_2\}$ *for* $y \in \{1\} \times (0,2)$*. This becomes*

$$\min_{z \in \mathbb{R}^2} |z_1| + |y_2 z_2| \qquad \text{s.t. } z_1 + z_2 = 1,$$

*which is equivalent to*

$$\min_{z_1 \in \mathbb{R}} |z_1| + |y_2(1 - z_1)|.$$

*We obtain the following optimal values and sets of all optimal solutions depending on* $y_2$*:*

- $y_2 \in (0,1)$*: minimum* $y_2$ *at* $z_1 = 0$
- $y_2 = 1$*: minimum* $y_2 = 1$ *at* $z_1 \in [0,1]$
- $y_2 \in (1,2)$*: minimum* $1$ *at* $z_1 = 1$*.*

*All the solutions have* $z_1 \in [0,1]$ *and therefore* $z_2 \in [0,1]$*,* $\|z\|_2 \leq \sqrt{2} < 2$*. This shows that adding* $\max\{2, \|z\|_2\}$ *to the objective function will not change the set of minimizers and the solution sets from above are also the solution sets of* (3.18)*.*

*So if there is a continuous function* $f : \mathbb{R}^2 \to \mathbb{R}^2$*, such that for every* $y \in \mathbb{R}^2$*,* $f(y)$ *is an optimal solution of* (3.18)*, we would need to have*

$$\begin{aligned} (f(1,y_2))_1 &= 0 && \text{for all } y_2 \in (0,1) \\ (f(1,y_2))_1 &= 1 && \text{for all } y_2 \in (1,2). \end{aligned}$$

*However, in this way* $f$ *cannot be continuous at the point* $(1,1)$*.*

*Nevertheless, there might still be exact continous selections for some of the most important applications listed in Remark 3.24. The work in [Bri+18] considers problem* (3.15) *and shows that there is an optimal solution that continuously depends on the parameter* $\lambda$*. With similar techniques, it might also be possible to show that there is an optimal solution that continuously depends on* $y$*. However, the above counterexample shows that this is not always possible in the generalized setting of Lemma 3.23. Furthermore, since we approximate the solution functions using the universal approximation theorem with an arbitrary but positive precision* $\delta > 0$*, having an exact selection would not lead to any essential improvement anyway.*

The above Lemma 3.23 concerns compact domains (and therefore only one hidden layer). The following Theorem turns this into a positive homogeneous version that enables results similar to the previous parts of this work.

**Theorem 3.26.** *Let $A \in \mathbb{R}^{m\times n}$, $U \subset \mathbb{R}^n$ positive homogeneous. Consider matrices $\tilde{A} \in \mathbb{R}^{k\times n}$, $\tilde{B} \in \mathbb{R}^{k\times m}$ with $\tilde{B}(\mathbb{R}^m) \subset \tilde{A}(\mathbb{R}^n)$, a continuous $u : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $\eta \in [0,1]$ and a norm $\|\cdot\|$ on $\mathbb{R}^k$. We define the minimization problem*

$$\min_{z\in\mathbb{R}^n} u(z,y) \qquad\qquad s.t.\ \|\tilde{A}z + \tilde{B}y\| \leq \eta. \tag{3.19}$$

*Furthermore, assume that there exists a continuous function $\bar{g} : \mathbb{R}^m \to \mathbb{R}$ and a coefficient $\alpha \in (0,\infty)$, such that for all $y \in \mathbb{R}^m$ and feasible $z \in \mathbb{R}^n$,*

$$\|z\|_2 \leq \alpha u(z,y) + \bar{g}(y).$$

*Let $\|\cdot\|_I$ be a norm on $\mathbb{R}^n$ and $\|\cdot\|_{II}$ a norm on $\mathbb{R}^m$.*

*Assume that for each $x \in \mathbb{R}^n$ and $e \in \mathbb{R}^m$, $\|e\|_{II} \leq \eta$, any optimal solution $\hat{x}$ of (3.19) for $y = Ax + e$ satisfies*

$$\|\hat{x} - x\|_I \leq v(x,\eta),$$

*where $v : \mathbb{R}^n \times \mathbb{R} \to [0,\infty)$ satisfies $v(\lambda x, \lambda\eta) = \lambda v(x,\eta)$ for all $\lambda \geq 0$, $x \in \mathbb{R}^n$, $\eta \in \mathbb{R}$ and $\eta \mapsto v(x,\eta)$ is increasing for each $x$. Assume that this also holds for $\eta = 0$.*

*Then for each $\delta > 0$, there exists a function $\tilde{f} : \mathbb{R}^m \to \mathbb{R}^n$, represented by a* ReLU *network with two hidden layers, such that for all $x \in \mathbb{R}^n$, $e \in \mathbb{R}^m$, $\|e\|_{II} \leq \frac{\eta}{3}\|Ax\|_{II}$, $y = Ax + e$,*

$$\|\tilde{f}(y) - x\|_I \leq \delta\|x\|_2 + v(x, \frac{4}{3}\eta\|Ax\|_{II}).$$

*Proof.* Consider the unit sphere of the $\|\cdot\|_{II}$ norm

$$S_{II} := \{x \in \mathbb{R}^m \,\big|\, \|x\|_{II} = 1\} \subset \mathbb{R}^m.$$

$S_{II}$ is a compact set and therefore we can apply Lemma 3.23 to obtain that for each $\epsilon > 0$, there is a continuous function $f : S_{II} \to \mathbb{R}^n$ such that for all $y \in S_{II}$, there exists a $\tilde{y} \in S_{II}$ and a solution $\tilde{x} \in \mathbb{R}^n$ of (3.19) for $\tilde{y}$ such that $\|y - \tilde{y}\|_2 < \epsilon$ and $\|\tilde{f}(y) - \tilde{x}\|_I < \epsilon$.

$f$ is defined on $S_{II}$ such that we can extend it to a positive homogeneous, continuous function $f : \mathbb{R}^m \to \mathbb{R}^n$ on the entire space. Now take any $x \in \mathbb{R}^n$ and $e \in \mathbb{R}^m$ with $\|e\|_{II} \leq \frac{\eta}{3}\|Ax\|_{II}$. Let $y = Ax + e$. Then

$$\frac{2}{3}\|Ax\|_{II} \leq (1 - \frac{\eta}{3})\|Ax\|_{II} \leq \|y\|_{II} \leq (1 + \frac{\eta}{2})\|Ax\|_{II} \leq \frac{4}{3}\|Ax\|_{II}. \tag{3.20}$$

Assume $y \neq 0$ for now. Define

$$\bar{x} = \frac{x}{\|y\|_{II}} \qquad\qquad \bar{e} = \frac{e}{\|y\|_{II}} \qquad\qquad \bar{y} = A\bar{x} + \bar{e} = \frac{y}{\|y\|_{II}},$$

such that $\bar{y} = A\bar{x} + \bar{e}$. So $\bar{y} \in S_{II}$ and thus by the previous observation, there is a $\bar{y}' \in S_{II}$ and an optimal solution $\bar{x}' \in \mathbb{R}^n$ of (3.19) for $\bar{y}'$ such that $\|\bar{y}' - \bar{y}\|_2 < \epsilon$ and $\|f(\bar{y}) - \bar{x}'\|_I < \epsilon$.

There is a constant $C > 0$ such that $\|w\|_{II} \leq C\|w\|_2$ for all $w \in \mathbb{R}^m$. We can choose $\epsilon \leq \frac{\eta}{2C}$.

Define $\bar{e}' := \bar{e} + \bar{y}' - \bar{y}$. Then $\bar{y}' = A\bar{x} + \bar{e}'$ and

$$\|\bar{e}'\|_{II} \leq \|\bar{e}\|_{II} + C\|\bar{y}' - \bar{y}\|_2 \leq \frac{\|e\|_{II}}{\|y\|_{II}} + C\epsilon \leq \frac{\frac{\eta}{3}\|Ax\|_{II}}{\frac{2}{3}\|Ax\|_{II}} + C\epsilon \leq \frac{\eta}{2} + \frac{\eta}{2} = \eta.$$

So since $\bar{x}'$ is an optimal solution of (3.19) for $\bar{y}'$,

$$\|\bar{x}' - \bar{x}\|_I \le v(\bar{x}, \eta).$$

Since $\|f(\bar{y}) - \bar{x}'\|_I < \epsilon$,

$$\|f(\bar{y}) - \bar{x}\|_I \le \|f(\bar{y}) - \bar{x}'\|_I + \|\bar{x}' - \bar{x}\|_I \le \epsilon + v(\bar{x}, \eta).$$

Now recall that $x = \|y\|_{II}\bar{x}$ and we defined $f$ by a positive homogeneous extension such that in general,

$$\begin{aligned}\|f(y) - x\|_I = \|y\|_{II}\|f(\bar{y}) - \bar{x}\|_I &\le \epsilon\|y\|_{II} + v(\frac{x}{\|y\|_{II}}, \eta)\|y\|_{II} = \epsilon\|y\|_{II} + v(x, \eta\|y\|_{II})\\ &\le \epsilon\frac{4}{3}\|Ax\|_{II} + v(x, \frac{4}{3}\eta\|Ax\|_{II}) \le \epsilon\frac{4}{3}C\|A\|_{2\to 2}\|x\|_2 + v(x, \frac{4}{3}\eta\|Ax\|_{II})\\ &\le \frac{\delta}{2}\|x\|_2 + v(x, \frac{4}{3}\eta\|Ax\|_{II}), \qquad (3.21)\end{aligned}$$

where the last step follows by choosing $\epsilon \le \frac{3\delta}{8C\|A\|_{2\to 2}}$.

It still remains to show (3.21) for the case that $y = 0$. Then by (3.20), also $Ax = 0$ and therefore $e = 0$. By the assumption of the theorem, in this case $z = 0$ is feasible and thus optimal in (3.19) for $\eta = 0$ and so

$$\|x - 0\|_I \le v(x, \eta) = v(x, 0).$$

Since we defined $f$ as a positive homogeneous extension, $f(y) = 0$ such that $\|f(y) - x\|_I = \|0 - x\|_I$ and (3.21) also holds for $y = 0$.

Now since $f$ is a continuous, positive homogeneous function, by Theorem 3.7 (applied to each component, together with equivalence of all norms), for each $\epsilon' > 0$, there is $\tilde{f} : \mathbb{R}^m \to \mathbb{R}^n$, represented by an unbiased ReLU network with two hidden layers, such that for all $y \in \mathbb{R}^m$,

$$\|\tilde{f}(y) - f(y)\|_I \le \epsilon'\|y\|_{II}.$$

Then $\tilde{f}(0) = 0 = f(0)$ and for $y \ne 0$, by combining everything, we obtain for all $\epsilon' > 0$,

$$\begin{aligned}\|\tilde{f}(y) - x\|_I \le \|\tilde{f}(y) - f(y)\|_I + \|f(y) - x\|_I &\le \epsilon'\|y\|_{II} + \frac{\delta}{2}\|x\|_2 + v(x, \frac{4}{3}\eta\|Ax\|_{II})\\ &\le \frac{4}{3}\epsilon'\|Ax\|_{II} + \frac{\delta}{2}\|x\|_2 + v(x, \frac{4}{3}\eta\|Ax\|_{II}) \le \delta\|x\|_2 + v(x, \frac{4}{3}\eta\|Ax\|_{II})\end{aligned}$$

by choosing $\epsilon' \le \frac{3\delta}{8C\|A\|_{2\to 2}}$.

□

With Theorem 3.26, we can construct a positive homogeneous network to solve an inverse problem that is known to be solved by a minimization problem. In particular, $\ell_1$ minimization has been studied for sparse recovery (see Chapter 4 of [FR13]). Applying Theorem 3.26 to the quadratically constrained basis pursuit (3.14), we obtain the following corollary.

**Corollary 3.27.** *Let $A \in \mathbb{R}^{m\times n}$ be a matrix of rank $m$, satisfying the $(2s, \delta)$-restricted isometry property for a $\delta < 0.7$ and $\eta \in [0, \frac{1}{3}]$. Then for each $\delta' > 0$, there exists a function $\tilde{f} : \mathbb{R}^m \to \mathbb{R}^n$, represented by an unbiased* ReLU *network with two hidden layers, such that for all $x \in \mathbb{R}^n$, $e \in \mathbb{R}^m$ with $\|e\|_2 \le \eta\|Ax\|_2$, $p \in [1, 2]$,*

$$\|\tilde{f}(Ax + e) - x\|_p \le \delta'\|x\|_2 + \frac{C}{s^{1-1/p}}\sigma_s(x)_1 + Ds^{1/p-1/2}\eta\|Ax\|_2,$$

*where $\sigma_s(x)_1 := \inf_{x' \in \Sigma_s}\|x - x'\|_1$ and $C$, $D$ only depend on $\delta$.*

*In particular, for $p = 1, 2$, we obtain*

$$\begin{aligned}\|\tilde{f}(Ax+e)-x\|_1 &\le \delta'\|x\|_2 + C\sigma_s(x)_1 + D\sqrt{s}\eta\|Ax\|_2\\ \|\tilde{f}(Ax+e)-x\|_2 &\le \delta'\|x\|_2 + \frac{C}{\sqrt{s}}\sigma_s(x)_1 + D\eta\|Ax\|_2.\end{aligned}$$

*Proof.* By Theorem 6.13 in [FR13], $A$ satisfies the $\ell_2$-robust null space property which in turn implies by Theorem 4.22 in [FR13] that the solution $\hat{x}$ of (3.14) always satisfies

$$\|\hat{x}-x\|_p \le \frac{C}{s^{1-1/p}}\sigma_s(x)_1 + Ds^{1/p-1/2}\eta$$

for $p \in [1, 2]$.

Then the result follows from Theorem 3.26 with $v(x,\eta) = \frac{C}{s^{1-1/p}}\sigma_s(x)_1 + Ds^{1/p-1/2}\eta$. □

As shown in Proposition 3.2 in [FR13], the basis pursuit denoising (3.15) and LASSO (3.16) are as powerful as (3.14) since a solution of one of them can be shown to also optimize the other ones for suitable parameters.

**Corollary 3.28.** *Let $A \in \mathbb{R}^{m\times n}$ be a matrix of rank $m$ that satisfies the $(2s,\delta)$-restricted isometry property for a $\delta < \frac{1}{3}$ and $\eta \in [0, \frac{1}{3}]$. Then for each $\delta' > 0$, there exists a function $\tilde{f} : \mathbb{R}^m \to \mathbb{R}^n$, represented by a* ReLU *network with two hidden layers, such that for all $x \in \mathbb{R}^n$, $e \in \mathbb{R}^m$ with $\|A^Te\|_\infty \le \eta\|A^TAx\|_\infty$,*

$$\|\tilde{f}(Ax+e)-x\|_2 \le \delta'\|x\|_2 + \frac{C}{\sqrt{s}}\sigma_s(x)_1 + D\eta\|A^TAx\|_\infty.$$

*Proof.* Analogously to Corollary 3.27, this is a consequence of Theorem 3.26, this time applied to the Dantzig selector (3.17).

As the only essential difference to Corollary 3.27, we need to ensure that $y \mapsto \|A^Ty\|_\infty$ is a norm on $\mathbb{R}^m$. Clearly it fulfills all properties except the positive definiteness. The latter one is fulfilled if $A^Ty \ne 0$ for all $y \ne 0$ which is equivalent to $\dim(\ker(A^T)) = 0$. This is fulfilled since $\dim(\ker(A^T)) = \dim((A(\mathbb{R}^n))^\perp) = m - \operatorname{rank}(A) = 0$. □

**Remark 3.29.**

- *Compared to the original minimization result, in Corollary 3.27 (and analogously Corollary 3.28), the upper bound on the error, $\frac{\eta}{3}\|Ax\|_2$, now depends on $\|Ax\|_2$. This arises from making the solution positive homogeneous. However, the term that contributes to the deviation of the result is still equal to the maximal error up to constant factors.*

- *The condition* $\operatorname{rank}(A) = m$ *is satisfied for most interesting matrices, for example for Gaussian ones with probability* 1*. If it is still not the case, we can replace $A$ by $PA$ for an orthogonal projection $P \in \mathbb{R}^{\operatorname{rank}(A)\times m}$ without changing its RIP.*

- *Compared to Corollary 3.4, Corollary 3.27 provides deviation bounds in other norms and for $p = 2$ a better dependence on $\sigma_s(x)_1$. However, it requires an explicit bound on $\|e\|_2$ which influences the result while in Corollary 3.4 there is one network that works for all possible error levels.*

- *Another approach that allows for robust sparse recovery without a previously known bound on $\|e\|_2$ is given in [Woj10] by a basis pursuit* (3.14) *with $\eta = 0$. However, to make this work, the measurement matrix $A$ must satisfy an additional condition beside the RIP which is known as the quotient property with respect to a norm $\|\cdot\|$. Then the reconstruction error depends on $\|e\|$. This additional property holds with respect to the norm $\|\cdot\|_2$ for example for Gaussian matrices but not for Bernoulli matrices as shown in Section 11.3 in [FR13].*

## 3.6 Approximation on Polytopes

In this section, we discuss topics related to the exact representation of the sparse recovery problem with ReLU networks. Even though for exact measurements, Corollary 3.4 states that $\|\tilde{f}(Ax) - x\|_2$ can be made arbitrarily small, i.e, $\leq \delta\|x\|_2$ for any $\delta > 0$, we cannot conclude an exact representation from it in the sense that $\tilde{f}(Ax) = x$ can be achieved for all signals $x$.

In the simple case of functions $\mathbb{R} \to \mathbb{R}$, it can be seen from the properties of the ReLU function, that the functions that are exactly represented by ReLU networks are the continuous piecewise linear (CPWL) functions. We call $f : \mathbb{R} \to \mathbb{R}$ CPWL if there are finitely many numbers $x_1 < x_2 < \cdots < x_n$ for some $n$ such that $f$ is affine linear on each interval $(-\infty, x_1], [x_1, x_2], \ldots, [x_{n-1}, x_n], [x_n, \infty)$ (which implies continuity since each of the interval boundary points are contained in both neighboring intervals).

This observation has been generalized to functions $f : \mathbb{R}^m \to \mathbb{R}$ on a higher-dimensional domain using higher-dimensional polyhedra. In this section, we introduce the most important terms related to these polyhedra and establish some tools to make this result applicable to the sparse recovery problem. In this way, we can prove the following result.

**Theorem 3.30.** *Let $A \in \mathbb{R}^{m \times n}$ be a matrix such that $Ax \neq 0$ for all $x \in \Sigma_{2s} \setminus \{0\}$.*

*Then there exists a function $f : \mathbb{R}^m \to \mathbb{R}^n$, represented by an unbiased* ReLU *network with $\lceil \log_2(s) + 1 \rceil$ hidden layers such that for all $x \in \Sigma_s$,*

$$f(Ax) = x.$$

### 3.6.1 General Terms Related to Polyhedra

The textbook [Zie12] covers various topics related to polyhedra and polytopes. We mostly use their notation and repeat the most important aspects here.

First, like in Section 0 and 1 of [Zie12], we use the following definitions for a subset $X \subset \mathbb{R}^n$.

- The *affine hull* of $X$ is

$$\operatorname{aff}(X) := \left\{ \sum_{j=1}^{k} \lambda_j x_j \,\middle|\, k \in \mathbb{Z}_{\geq 1},\, x_1, \ldots, x_k \in X,\, \lambda_1, \ldots, \lambda_k \in \mathbb{R},\, \sum_{j=1}^{k} \lambda_j = 1 \right\}$$

  which is the smallest (by set inclusion) affine subspace of $\mathbb{R}^n$ that contains $X$.

- The *convex hull* of $X$ is

$$\operatorname{conv}(X) := \left\{ \sum_{j=1}^{k} \lambda_j x_j \,\middle|\, k \in \mathbb{Z}_{\geq 1},\, x_1, \ldots, x_k \in X,\, \lambda_1, \ldots, \lambda_k \in [0, \infty),\, \sum_{j=1}^{k} \lambda_j = 1 \right\}$$

  which is the smallest (by set inclusion) convex subset of $\mathbb{R}^n$ that contains $X$.

- The *conic hull* of $X$ is

$$\operatorname{cone}(X) := \left\{ \sum_{j=1}^{k} \lambda_j x_j \,\middle|\, k \in \mathbb{Z}_{\geq 1},\, x_1, \ldots, x_k \in X,\, \lambda_1, \ldots, \lambda_k \in [0, \infty) \right\}.$$

Furthermore, we also need the following precise definitions of polyhedra and polytopes.

**Definition 3.31** (Definition 0.1 in [Zie12]).

- *A polyhedron $P \subset \mathbb{R}^n$ is an intersection of finitely many closed halfspaces, i.e.*

$$P = \{x \in \mathbb{R}^n \,|\, a_j^T x \leq b_j \text{ for all } 1 \leq j \leq k\}$$

  *for some $a_1, \ldots, a_k \in \mathbb{R}^n$ and $b_1, \ldots, b_k \in \mathbb{R}$.*

- *An $\mathcal{H}$-polytope $P \subset \mathbb{R}^n$ is a polyhedron that is bounded in the sense that there is no ray $\{x + \lambda y \,|\, \lambda \geq 0\} \subset P$ for any $x \in \mathbb{R}^n$, $y \in \mathbb{R}^n \backslash \{0\}$.*
- *A $\mathcal{V}$-polytope $P \subset \mathbb{R}^n$ is the convex hull of any finite set in $\mathbb{R}^n$.*
- *The dimension of a polyhedron $P$ is defined as the dimension of its affine hull*

$$\dim(P) := \dim(\operatorname{aff}(P)).$$

A fundamental result in the field of polytopes covered in [Zie12] is that the terms $\mathcal{H}$- and $\mathcal{V}$-polytope are equivalent.

**Theorem 3.32** (Theorem 1.1 in [Zie12])**.** *A subset $P \subset \mathbb{R}^n$ is an $\mathcal{H}$-polytope if and only if it is a $\mathcal{V}$-polytope.*

Therefore, hereafter we can just refer to both concepts as *polytopes.*

### 3.6.2 Polyhedra and ReLU networks

After having already explained the idea of continuous piecewise linear functions in one dimension, with the help of polyhedra we can generalize this to functions on higher-dimensional domains.

**Definition 3.33** (Continuous piecewise linear functions $\mathbb{R}^n \to \mathbb{R}$ (CPWL), Definition 3 in [Aro+18])**.** *$f : \mathbb{R}^n \to \mathbb{R}$ is continuous piecewise linear (CPWL) function if there are finitely many polyhedra covering $\mathbb{R}^n$ such that $f$ is affine linear on each of these polyhedra.*

As shown in [He+20], this class is equivalent to the class of functions that can be represented by ReLU functions. Moreover, the following result also states the number of layers that are sufficient for the representation.

**Theorem 3.34** (Representation of CPWL functions by ReLU networks, [He+20])**.** *Any* ReLU *network represents a CPWL function.*

*On the other hand, any CPWL function $\mathbb{R}^n \to \mathbb{R}$ can be represented by a* ReLU *network with $\leq \lceil \log_2(n + 1) \rceil$ hidden layers.*

In order to apply this theorem directly to the sparse recovery problem, we would need to show that each function $f_j : \mathbb{R}^m \to \mathbb{R}$ that recovers $x_j$ from $Ax$, is CPWL. We know that for each support $S \subset [N]$, $|S| = s$, with $\Sigma_S := \{x \in \mathbb{R}^n \,\big|\, \operatorname{supp}(x) \subset S\}$, $f$ is linear on $A\Sigma_S$ and $A\Sigma_S$ is a polyhedron. However, these polyhedra do not cover the entire space $\mathbb{R}^m$. [Ovc02] characterizes CPWL functions as an expression of minima and maxima of affine linear functions on a convex domain. This could be used to extend a CPWL function from a convex domain to the entire space. However the set $A\Sigma_s$ of images of sparse vectors is not convex and $f$ needs to be extended to the space between the images of the $s$-sparse supports in some way.

We pursue the following approach. We establish multiple tools in Section 3.6.3 that can be used to show that whenever a function is represented by one separate ReLU network on each polyhedron, then we can join them into one network which we can also ensure to be unbiased under certain circumstances. For technical reasons, we need the function to be 0 on the intersections between the polyhedra. In our application, this intersection would consist of the images $A\Sigma_{s-1}$ of the $(s-1)$-sparse vectors. Therefore in the end, we first start representing $f$ on $\Sigma_1$, then in the next step represent the difference between this function and $f$ on $\Sigma_2$ and so on until we reach $\Sigma_s$.

### 3.6.3 Auxiliary Results

The following lemma states that a function represented by a biased ReLU network on an affine subspace that does not include 0 can be represented on the same affine subspace by an unbiased network of the same depth.

**Lemma 3.35** (Bias elimination)**.** *Let $m, n \in \mathbb{N}$ with $n > m$ and $U \in \mathbb{R}^{n\times m}$ have orthonormal columns, i.e., $U^TU = Id_m$. Take $v \in \mathbb{R}^n$ such that $UU^Tv \neq v$. Consider a function $f : \mathbb{R}^m \to \mathbb{R}$ represented by $f(x) = W_{d+1}\phi(\dots W_2\phi(W_1x + b_1) + b_2 \dots) + b_{d+1}$, i.e., a* ReLU *network with $d$ hidden layers and biases. Assume dimensions $W_1 \in \mathbb{R}^{k_1\times m}, W_2 \in \mathbb{R}^{k_2\times k_1}, \dots, W_{d+1} \in \mathbb{R}^{1\times k_d}$.*

*Then there exist matrices $\tilde{W}_1 \in \mathbb{R}^{(k_1+1)\times n}, \tilde{W}_2 \in \mathbb{R}^{(k_2+1)\times(k_1+1)}, \dots, \tilde{W}_{d+1} \in \mathbb{R}^{1\times(k_d+1)}$ such that for $\tilde{f}(y) := \tilde{W}_{d+1}\phi\left(\dots \tilde{W}_2\phi(\tilde{W}_1y)\right)$, we obtain that for all $x \in \mathbb{R}^m$, $\tilde{f}(Ux+v) = f(Ux+v)$.*

*Proof of Lemma 3.35.* We can extend the orthonormal columns of $U$ to an orthonormal basis of $\mathbb{R}^n$ and define the matrix $\tilde{U} \in \mathbb{R}^{n\times(n-m)}$ whose columns are precisely the remaining basis entries. Then by orthogonality, $\tilde{U}^TU = 0$.

We define $\tilde{W}_1 := \begin{pmatrix} W_1 + b_1\tilde{v}^T \\ \tilde{v}^T \end{pmatrix} \in \mathbb{R}^{(k_1+1)\times n}$ where $\tilde{v} := \frac{\tilde{U}\tilde{U}^Tv}{\|\tilde{U}^Tv\|_2^2}$. Note that by assumption $UU^Tv \neq v$, so $\tilde{U}^Tv \neq 0$. For $j = 2, \dots, d$, we define $\tilde{W}_j \in \mathbb{R}^{(k_j+1)\times(k_{j-1}+1)}$ by

$$\tilde{W}_j := \begin{pmatrix} W_j & b_j \\ 0 & 1 \end{pmatrix}$$

and finally $\tilde{W}_{d+1} = (W_{d+1} \quad b_{d+1}) \in \mathbb{R}^{1\times(k_d+1)}$.

Then for any $x \in \mathbb{R}^m$,

$$\tilde{v}^T(Ux+v) = \frac{1}{\|\tilde{U}^Tv\|_2^2}\left(v^T\tilde{U}(\tilde{U}^TU)x + v^T\tilde{U}\tilde{U}^Tv\right) = \frac{(\tilde{U}^Tv)^T(\tilde{U}^Tv)}{\|\tilde{U}^Tv\|_2^2} = 1$$

such that

$$\tilde{W}_1(Ux+v) = \begin{pmatrix} W_1(Ux+v) + b_1\tilde{v}^T(Ux+v) \\ \tilde{v}^T(Ux+v) \end{pmatrix} = \begin{pmatrix} W_1(Ux+v) + b_1 \\ 1 \end{pmatrix}.$$

Then by induction for all $j = 2, \dots, d$,

$$\tilde{W}_j\phi\left(\dots \tilde{W}_2\phi(\tilde{W}_1(Ux+v))\dots\right) = \begin{pmatrix} W_j\phi(\dots W_2\phi(W_1(Ux+v) + b_1) + b_2 \dots) + b_j \\ 1 \end{pmatrix},$$

and defining

$$\tilde{f}(y) := \tilde{W}_{d+1}\phi\left(\dots \tilde{W}_2\phi(\tilde{W}_1y)\dots\right),$$

we obtain that $\tilde{f}(Ux+v) = f(Ux+v)$ holds for all $x \in \mathbb{R}^m$.

□

Now we establish that for a polyhedron defined by inequalities $a_j^Tx \leq b_j$, the distance of a point $x$ from the polyhedron is the maximal $\phi(a_j^Tx - b_j)$ up to a constant where $\phi$ is the ReLU function.

**Lemma 3.36.** *Let $a_1, \dots, a_k \in S^{n-1}$, $b_1, \dots, b_k \in \mathbb{R}$ ($k \geq 1$) and define the non-empty polyhedron*

$$P := \left\{x \in \mathbb{R}^n \,\middle|\, \forall j \in [k] : a_j^Tx \leq b_j\right\}$$

*Then there is a number $C(P) > 0$, depending on the polyhedron, such that for all $x \in \mathbb{R}^n$,*

$$\max_{j\in[k]} \phi(a_j^Tx - b_j) \leq d(x, P) \leq C(P) \max_{j\in[k]} \phi(a_j^Tx - b_j),$$

*where*

$$d(x, P) = \inf_{y\in P} \|x - y\|_2.$$

*Proof.* We can assume that $x \notin P$, otherwise the statement follows immediately.

For the first inequality, consider a $j_0 \in [k]$ such that $\phi(a_{j_0}^T x - b_{j_0})$ becomes maximal. Consider the half-space $H_{j_0} := \{x \in \mathbb{R}^n \,|\, a_{j_0}^T x - b_{j_0} \leq 0\}$. Then since $\|a_{j_0}\|_2 = 1$, $d(x, H_{j_0}) = \phi(a_{j_0}^T x - b_{j_0})$. Then the first inequality follows from the fact that $P \subset H_{j_0}$ and thus $d(x, H_{j_0}) \leq d(x, P)$.

Now we prove the second inequality. Denote $A \in \mathbb{R}^{k \times n}$ for the matrix with rows $a_j^T$ $(1 \leq j \leq k)$ and $b \in \mathbb{R}^n$ for the vector with entries $b_j$ $(1 \leq j \leq k)$.

Fix $x \in \mathbb{R}^n$. Define

$$\bar{x} := \operatorname{argmin}_{y \in P} \|x - y\|_2,$$

which exists since $P$ is closed and is unique since $P$ is convex. Furthermore, define the set

$$J := \{j \in [k] \,|\, a_j^T \bar{x} = b_j\} \subset [k]$$

and the polytope

$$P_J := \{x \in \mathbb{R}^n \,|\, \forall j \in J : a_j^T x \leq b_j\}$$

with the corresponding minimal distance point

$$\hat{x} := \operatorname{argmin}_{y \in P_J} \|x - y\|_2.$$

Now we show that $\hat{x} = \bar{x}$: Assume that $\hat{x} \neq \bar{x}$. Since $\hat{x}, \bar{x} \in P_J$ and since the $\ell_2$-distance of a point to a closed convex set has a unique minimizer, $\|x - \hat{x}\|_2 < \|x - \bar{x}\|_2$.

For any $\lambda \in [0, 1]$, $x_\lambda := \lambda \hat{x} + (1 - \lambda)\bar{x} \in P_J$. By definition of $J$, the strict inequalities $a_j^T \bar{x} < b_j$ are fulfilled for all $j \in J^c$. Note that $\bar{x} = x_\lambda$ for $\lambda = 0$. So by continuity, for a small enough $\lambda_1 \in (0, 1)$, these strict inequalities are still fulfilled for $x_{\lambda_1}$. Since in addition $x_{\lambda_1} \in P_J$, this implies that $x_{\lambda_1} \in P$. However, $\|x - x_{\lambda_1}\|_2 \leq \lambda \|x - \hat{x}\|_2 + (1 - \lambda)\|x - \bar{x}\|_2 < \|x - \bar{x}\|_2$. This contradicts the definition of $\bar{x}$ and thus $\hat{x} = \bar{x}$.

We can draw the following conclusions,

$$x \notin P \Rightarrow x \neq \bar{x} \Rightarrow x \notin P_J \Rightarrow J \neq \emptyset.$$

Define $J^{(0)} := J$ and $\bar{x}^{(0)} = \bar{x}$. We repeat the following steps for $l = 0, 1, \ldots$.

1. It holds that $J^{(l)} \neq \emptyset$, $x \notin P_{J^{(l)}}$, and $\bar{x}^{(l)}$ minimizes $\|\bar{x}^{(l)} - x\|_2$ in $P_{J^{(l)}}$.

2. Let $A_{J^{(l)}} \in \mathbb{R}^{|J^{(l)}| \times n}$ be the matrix with rows $a_j^T$ for $j \in J^{(l)}$ and correspondingly $b_{J^{(l)}} \in \mathbb{R}^{|J^{(l)}|}$. With the Moore-Penrose pseudoinverse $A_{J^{(l)}}^\dagger$, $y = A_{J^{(l)}}^\dagger v$ is the minimizer of $\|A_{J^{(l)}} y - v\|_2$ with minimal $\|y\|_2$. So $\bar{x}^{(l)} - x = A_{J^{(l)}}^\dagger (b_{J^{(l)}} - A_{J^{(l)}} x)$ since $\bar{x}^{(l)}$ is the minimizer of $\|A_{J^{(l)}}(\bar{x}^{(l)} - x) - (b_{J^{(l)}} - A_{J^{(l)}} x)\|_2 = \|A_{J^{(l)}} \bar{x}^{(l)} - b_{J^{(l)}}\|_2$ (which is 0 iff $a_j^T \bar{x}^{(l)} = b_j$ for all $j \in J^{(l)}$) with minimal $\|\bar{x}^{(l)} - x\|_2$.

3. This implies that

$$\|\bar{x}^{(l)} - x\|_2 \leq \|A_{J^{(l)}}^\dagger\|_{2\to 2} \|A_{J^{(l)}} x - b_{J^{(l)}}\|_2 \leq \sqrt{k} \|A_{J^{(l)}}^\dagger\|_{2\to 2} \|A_{J^{(l)}} x - b_{J^{(l)}}\|_\infty. \tag{3.22}$$

4. We pick a $j_l \in J^{(l)}$ such that $|a_{j_l}^T x - b_{j_l}|$ becomes maximal. Define $\bar{J}^{(l+1)} := J^{(l)} \backslash \{j_l\}$ and $\bar{x}^{(l+1)} := \operatorname{argmin}_{y \in P_{\bar{J}^{(l+1)}}} \|x - y\|_2$. In addition, $J^{(l+1)} := \{j \in \bar{J}^{(l+1)} \,|\, a_j^T \bar{x}^{(l+1)} = b_j\}$. Analogously to the beginning of this proof, then $\bar{x}^{(l+1)} = \operatorname{argmin}_{y \in P_{J^{(l+1)}}} \|x - y\|_2$.

   Now we distinguish the following three cases.

   (a) $a_{j_l}^T x - b_{j_l} \geq 0$.
   In this case, the right hand side of (3.22) can be bounded by $\phi(a_{j_l}^T x - b_{j_l})$. We stop the iteration here. (The previously defined $J^{(l+1)}$ is irrelevant in this case).

(b) $a_{j_l}^T x - b_{j_l} < 0$ and $|a_{j_l}^T x - b_{j_l}| \le \|\bar{x}^{(l+1)} - x\|_2$.
Then by (3.22), $\|\bar{x}^{(l)} - x\|_2 \le \sqrt{k}\|A_{J^{(l)}}^\dagger\|_{2\to 2}\|\bar{x}^{(l+1)} - x\|_2$.
Since $x \notin P_{J^{(l)}}$, there exists a $j \in J^{(l)}$ such that $a_j^T x - b_j > 0$. Then this $j \neq j_l$ is also contained in $\bar{J}^{(l+1)}$. So $\bar{x}^{(l+1)} \neq x$, and thus $J^{(l+1)} \neq \emptyset$ and $x \notin P_{J^{(l+1)}}$. We continue with the next iteration.

(c) $a_{j_l}^T x - b_{j_l} < 0$ and $|a_{j_l}^T x - b_{j_l}| > \|\bar{x}^{(l+1)} - x\|_2$.
Then, using that $\|a_{j_l}\|_2 = 1$, we obtain

$$\begin{aligned} a_{j_l}^T \bar{x}^{(l+1)} - b_{j_l} &= a_{j_l}^T x - b_{j_l} + a_{j_l}^T(\bar{x}^{(l+1)} - x) \le a_{j_l}^T x - b_{j_l} + \|a_{j_l}\|_2\|\bar{x}^{(l+1)} - x\|_2 \\ &< a_{j_l}^T x - b_{j_l} + |a_{j_l}^T x - b_{j_l}| \le 0. \end{aligned}$$

Together with $\bar{x}^{(l+1)} \in P_{\bar{J}^{(l+1)}}$, this implies that $\bar{x}^{(l+1)} \in P_{J^{(l)}}$. In addition, since $\bar{x}^{(l)}$ and $\bar{x}^{(l+1)}$ both are points of minimal distance and $P_{J^{(l)}} \subset P_{J^{(l+1)}}$, $\|\bar{x}^{(l+1)} - x\|_2 \le \|\bar{x}^{(l)} - x\|_2$.
$\bar{x}^{(l)}$ is the unique minimizer of $y \mapsto \|x - y\|_2$ in $P_{J^{(l)}}$, so we obtain that $\bar{x}^{(l+1)} = \bar{x}^{(l)}$. However, $a_{j_l}^T \bar{x}^{(l+1)} - b_{j_l} < 0$ and $a_{j_l}^T \bar{x}^{(l)} - b_{j_l} = 0$. So this third case leads to a contradiction and cannot occur.

In every step $l$, the iteration either terminates or reduces the size of $J^{(l)}$ by at least 1, while $|J^{(l)}|$ remains $\ge 1$. So it terminates after the iteration with index $L \le |J^{(0)}| \le k$ and then with

$$D := \max\left\{1, \sqrt{k} \max_{J\subset[k],|J|\ge 1} \|A_J^\dagger\|_{2\to 2}\right\},$$

we obtain that

$$d(x,P) = \|\bar{x}^{(0)} - x\|_2 \le D^L\|\bar{x}^{(L)} - x\|_2 \le D^L\phi(a_{j_L}^T x - b_{j_L}) \le D^k \max_{j\in[k]} \phi(a_j^T x - b_j).$$

□

In the next step we bound the distance of a point to the intersection $P \cap Q$ of two polyhedra in terms of the distance to one of them.

**Lemma 3.37.** *Let $P = \{x \in \mathbb{R}^n \,|\, \forall j \in [k] : a_j^T x \le b_j\}$ and $Q = \{x \in \mathbb{R}^n \,|\, \forall j \in [\tilde{k}] : \tilde{a}_j^T x \le \tilde{b}_j\}$ be non-empty polyhedra in $\mathbb{R}^n$ such that $P \cap Q \neq \emptyset$. There is a number $c(P,Q) > 0$, depending on the polyhedra, such that for all $x \in Q$,*

$$d(x, P\cap Q) \le c(P,Q)d(x,P).$$

*Proof.* Note that $P \cap Q = \{y \in \mathbb{R}^n \,|\, \forall j \in [k] : a_j^T y \le b_j \text{ and } \forall j \in [\tilde{k}] : \tilde{a}_j^T y \le \tilde{b}_j\}$ such that by Lemma 3.36, for all $x \in \mathbb{R}^n$,

$$d(x, P\cap Q) \le C(P\cap Q)\max\left\{\max_{j\in[k]}\phi(a_j^T x - b_j), \max_{j\in[\tilde{k}]}\phi(\tilde{a}_j^T x - \tilde{b}_j)\right\}.$$

If $x \in Q$, then $\phi(\tilde{a}_j^T - \tilde{b}_j) = 0$ for all $j \in [\tilde{k}]$ such that then

$$d(x, P\cap Q) \le C(P\cap Q)\max_{j\in[k]}\phi(a_j^T x - b_j).$$

On the other hand, Lemma 3.36 also states that for all $x \in \mathbb{R}^n$,

$$d(x,P) \ge \max_{j\in[k]}\phi(a_j^T x - b_j).$$

Together these two inequalities imply that for all $x \in Q$,

$$d(x, P\cap Q) \le C(P\cap Q)d(x,P).$$

□

The next theorem concerns a collection of polytopes that satisfy certain conditions, including that 0 is non of their affine hulls. It states that a function that can be represented by a ReLU network on each of these polytopes and that is 0 on their intersections, can also be represented by one ReLU network on the entire union of the polytopes. In addition, we can ensure this network to be positive homogeneous.

**Lemma 3.38.** *Let $\mathcal{P}$ be a finite set of non-empty polytopes in $\mathbb{R}^m$ such that:*

- *For all $P \in \mathcal{P}$, $0 \notin \operatorname{aff}(P)$.*
- *For any $P, Q \in \mathcal{P}$, $\operatorname{cone}(P) \cap \operatorname{cone}(Q) = \operatorname{cone}(P \cap Q) \cup \{0\}$.*

*Take a function $f : \mathbb{R}^m \to \mathbb{R}$ such that:*

- *For each $P \in \mathcal{P}$, there exists a function $f_P : \mathbb{R}^m \to \mathbb{R}$, represented by a (biased)* ReLU *network with $d$ hidden layers, such that $f_P(x) = f(x)$ for all $x \in P$.*
- *Let $R := \bigcup_{\substack{P,Q \in \mathcal{P} \\ P \neq Q}} (P \cap Q)$ be the set of all points in $\mathbb{R}^m$ that lie in more than one polytope in $\mathcal{P}$. Then $f(x) = 0$ for all $x \in R$.*

*Then there exists a function $\tilde{f} : \mathbb{R}^m \to \mathbb{R}$, represented by an* ***unbiased*** ReLU *network with $d + 1$ hidden layers such that $\tilde{f}(x) = f(x)$ for all $x \in \bigcup_{P \in \mathcal{P}} P$.*

*Proof.* Fix one $P \in \mathcal{P}$. We will show that there is a function $\tilde{f}_P : \mathbb{R}^m \to \mathbb{R}$ such that $\tilde{f}_P(x) = f_P(x)$ for all $x \in P$ and that for all other $Q \in \mathcal{P} \backslash \{P\}$, $\tilde{f}_P(x) = 0$ holds for all $x \in Q$.

Since $0 \notin \operatorname{aff}(P)$, we can use Lemma 3.35 to eliminate the bias in the network of $f_P$ without changing the function on $\operatorname{aff}(P) \supset P$. Thus, we can assume that the network representing $f_P$ is unbiased.

The condition $0 \notin \operatorname{aff}(P)$ also implies that $\dim(\operatorname{aff}(P)) \leq m - 1$, such that $P$ is contained in an affine hyperplane $\{x \in \mathbb{R}^m \mid a^T x = b_0\}$ for $a \in \mathbb{R}^m$ with $\|a\|_2 = 1$ and $b_0 \in \mathbb{R} \backslash \{0\}$.

Since $P$ is a convex polygon, there exist $A \in \mathbb{R}^{k \times m}$ and $b \in \mathbb{R}^k$ such that the set of solutions $x \in \mathbb{R}^m$ to

$$\begin{aligned} a^T x &= b_0 \\ Ax &\leq b \end{aligned}$$

(with element-wise $\leq$ in the second inequality) is precisely $P$. Since $b_0 \neq 0$, we can add a multiple of $a^T x = b_0$ to every row in $Ax \leq b$ such that the right hand side becomes 0. So without loss of generality we can assume that $b = 0$ and that $P$ is the set of solutions $x$ to

$$\begin{aligned} a^T x &= b_0 \\ Ax &\leq 0. \end{aligned}$$

Then $\operatorname{cone}(P)$ is the same as the polytope defined by the inequalities $Ax \leq 0$. We denote $a_1^T, \dots, a_k^T$ for the rows of the matrix $A$.

We observe the following facts:

- As a ReLU network function, $f_P$ is Lipschitz continuous with a Lipschitz constant $L_P > 0$.
- By Lemma 3.36, there exists a constant $C(P) > 0$ such that for all $x \in \mathbb{R}^m$, $d(x, \operatorname{cone}(P)) \leq C(P) \max_{j \in [k]} \phi(a_j^T x)$.
- $f_P$ is represented by an unbiased ReLU network and thus invariant under positive scaling, i.e., $f_P(\lambda x) = \lambda f_P(x)$ for all $\lambda \geq 0$ and $x \in \mathbb{R}^m$. Thus for each other $Q \in \mathcal{P} \backslash \{P\}$, the condition $f_P(x) = 0$ does not only hold for all $x \in P \cap Q$, but even for all $x \in \operatorname{cone}(P \cap Q)$.

We take $D(P)$ to be the maximum of all $c(\text{cone}(P), \text{cone}(Q))$ in Lemma 3.38 for $Q \in \mathcal{P}\backslash\{P\}$ such that $P \cap Q \neq \emptyset$. Then for each such cone$(Q)$ and all $x \in Q$,

$$d(x, \text{cone}(P) \cap \text{cone}(Q)) \leq D(P)d(x, \text{cone}(P)).$$

Let $\bar{Q} := \bigcup_{Q\in\mathcal{P}:P\cap Q=\emptyset}$. Since $\mathcal{P}$ is finite, $\bar{Q}$ is compact and disjoint from $P$. $\bar{Q}$ is even disjoint from cone$(P)$. Otherwise there is a $Q \in \mathcal{P}$ with $P \cap Q = \emptyset$ such that a non-zero vector (since $0 \notin Q$) is contained in $\text{cone}(P) \cap Q \subset \text{cone}(P) \cap \text{cone}(Q) = \text{cone}(P \cap Q) \cup \{0\} = \{0\}$, which is a contradiction. Since cone$(P)$ is convex and closed, for each $x \in \bar{Q}$, we can find a unique closest point $P_{\text{cone}(P)}(x) \in \text{cone}(P)$. Now define $g_P : \mathbb{R}^m \to \mathbb{R}$ by

$$g_P(x) = \left( L_P D(P) + \frac{\max_{y\in\bar{Q}} |f_P(P_{\text{cone}(P)}(y))|}{\min_{y\in\bar{Q}} \|y - P_{\text{cone}(P)}(y)\|_2} + L_P \right) C(P) \sum_{j=1}^{m} \phi(a_j^T x). \tag{3.23}$$

Since $\bar{Q}$ is compact and disjoint from cone$(P)$, all the minima and maxima in this expression exist and the denominator is $> 0$. Also, $g_P(x) \geq 0$ for all $x \in \mathbb{R}^m$.

The function $g_P$ has the following properties.

- For all $x \in P$,
$$g_P(x) = 0. \tag{3.24}$$

- For all $x \in Q$ for any $Q \in \mathcal{P}\backslash\{P\}$,
$$g_P(x) \geq |f_P(x)|. \tag{3.25}$$

(3.24) follows directly from the fact that $Ax \leq 0$ holds for all $x \in P$. To show the second property (3.25), we distinguish two cases.

- 1st case: $P \cap Q \neq \emptyset$.

  Then $P \cap Q$ is a polytope. Let $\bar{x}$ be the point in $\text{cone}(P) \cap \text{cone}(Q)$ with minimal distance to $x$. Since by assumption $\text{cone}(P) \cap \text{cone}(Q) = \text{cone}(P \cap Q) \cup \{0\}$, by the previous observation, $\bar{x}$ satisfies $f_P(\bar{x}) = 0$. Thus we obtain

$$\begin{aligned} |f_P(x)| &\leq |f_P(\bar{x})| + |f_P(\bar{x}) - f_P(x)| \leq 0 + L_P\|\bar{x} - x\|_2 = L_P d(x, \text{cone}(P) \cap \text{cone}(Q)) \\ &\leq L_P D(P) d(x, \text{cone}(P)) \leq L_P D(P) C(P) \max_{j\in[k]} \phi(a_j^T x) \leq g_P(x), \end{aligned}$$

  such that (3.25) holds.

- 2nd case: $P \cap Q = \emptyset$.

  Take the point $v = P_{\text{cone}(P)}(x)$ in cone$(P)$ with minimal distance from $x$. Then

$$\begin{aligned} |f_P(x)| &\leq |f_P(v)| + |f_P(x) - f_P(v)| \leq \max_{y\in\bar{Q}} |f_P(P_{\text{cone}(P)}(y))| + L_P d(x, \text{cone}(P)) \\ &\leq \left( \frac{\max_{y\in\bar{Q}} |f_P(P_{\text{cone}(P)}(y))|}{\min_{y\in\bar{Q}} \|y - P_{\text{cone}(P)}(y)\|_2} + L_P \right) d(x, \text{cone}(P)) \\ &\leq \left( \frac{\max_{y\in\bar{Q}} |f_P(P_{\text{cone}(P)}(y))|}{\min_{y\in\bar{Q}} \|y - P_{\text{cone}(P)}(y)\|_2} + L_P \right) C(P) \max_{j\in[k]} \phi(a_j^T x) \\ &\leq g_P(x), \end{aligned}$$

  showing (3.25).

From the definition (3.23), we can see that $g_P$ is represented by an unbiased ReLU network with one hidden layer (the factors outside of the sum are constant and do not depend on $x$). This network can then also be changed to $d$ layers.

Now we define $\tilde{f}_P : \mathbb{R}^m \to \mathbb{R}$ by

$$\tilde{f}_P(x) = \phi(f_P(x) - g_P(x)) - \phi(-f_P(x) - g_P(x))$$

for all $x \in \mathbb{R}^m$. This can then be represented by a ReLU network with $d+1$ hidden layers.

From the properties (3.24) and (3.25) of $g_P$, we can conclude that for $x \in P$,

$$\tilde{f}_P(x) = \phi(f_P(x) - 0) - \phi(-f_P(x) - 0) = f_P(x).$$

For any other $Q \in \mathcal{P}\backslash\{P\}$ and $x \in Q$, because of (3.25), $f_P(x) - g_P(x) \leq 0$ and $-f_P(x) - g_P(x) \leq 0$ such that $\tilde{f}_P(x) = 0$.

Now we can define $\tilde{f} : \mathbb{R}^m \to \mathbb{R}$ by

$$\tilde{f}(x) = \sum_{P \in \mathcal{P}} \tilde{f}_P(x).$$

By the representation of the $\tilde{f}_P$, also $\tilde{f}$ can be represented by an unbiased ReLU network with $d+1$ hidden layers. For any $x \in \bigcup_{P\in\mathcal{P}} P$, consider one $P' \in \mathcal{P}$ such that $x \in P'$. Then $\tilde{f}_{P'}(x) = f(x)$ and for any other $Q \in \mathcal{P}\backslash\{P'\}$, it holds that $\tilde{f}_Q(x) = 0$. Note that this is true even for the case that also $x \in Q$ ($f$ is zero on the intersections). So we obtain

$$\tilde{f}(x) = \tilde{f}_{P'}(x) + \sum_{Q \in \mathcal{P}\backslash\{P'\}} \tilde{f}_Q(x) = f(x).$$

□

In the next step we take a certain sequence of sets of polytopes. In our application $\mathcal{P}^{(r)}$ will be the set of polytopes that form the intersection of the images $A\Sigma_{s-r}$ of the $(s-r)$-sparse vectors under $A$ with the image of the $\ell_1$ unit sphere (which itself consists of polytopes) under $A$. These sets of polytopes have to fulfill a number of technical conditions, for example that the intersection of two elements in $\mathcal{P}^{(r)}$ is contained in an element of the next set $\mathcal{P}^{(r+1)}$. We will show that they are fulfilled in the subsequent Lemma 3.40.

Then the following Lemma states that any function that can be represented by a separate ReLU network on each of these polytopes, can also be represented on their entire union with just one network. We do not require the function to be zero on the intersections anymore.

**Lemma 3.39.** *Let $\mathcal{P}, \mathcal{P}^{(0)}, \mathcal{P}^{(1)}, \ldots, \mathcal{P}^{(R)}$ be finite sets of non-empty polytopes in $\mathbb{R}^m$ such that:*

*(a) $\mathcal{P}^{(0)} = \mathcal{P}$ and $\mathcal{P}^{(R)} = \emptyset$.*

*(b) For all $P \in \mathcal{P}$, $0 \notin \operatorname{aff}(P)$ and $\dim(P) \leq m'$.*

*(c) For $r = 0, \ldots, R-1$, for all $P, Q \in \mathcal{P}^{(r)}$, $\operatorname{cone}(P) \cap \operatorname{cone}(Q) = \operatorname{cone}(P \cap Q) \cup \{0\}$.*

*(d) For $r = 0, \ldots, R-1$, for all $P, Q \in \mathcal{P}^{(r)}$ with $P \neq Q$ and $P \cap Q \neq \emptyset$, there exists a $\tilde{Q} \in \mathcal{P}^{(r+1)}$ such that $P \cap Q \subset \tilde{Q}$.*

*(e) For $r = 0, \ldots, R-1$, for all $P \in \mathcal{P}^{(r+1)}$, there is a $Q \in \mathcal{P}^{(r)}$ such that $P \subset Q$.*

*Consider a function $f : \bigcup_{P\in\mathcal{P}} P \to \mathbb{R}$ such that for each $P \in \mathcal{P}$, there exists a* ReLU *network with $\leq \lceil \log_2(m'+1) \rceil$ hidden layers, representing $f_P : \mathbb{R}^m \to \mathbb{R}$, such that for all $x \in P$,*

$$f_P(x) = f(x). \tag{3.26}$$

*Then there exists an unbiased* ReLU *network with $\leq \lceil \log_2(m'+1) + 1 \rceil$ hidden layers, representing $\tilde{f} : \mathbb{R}^m \to \mathbb{R}$, such that for all $x \in \bigcup_{P\in\mathcal{P}} P$,*

$$\tilde{f}(x) = f(x).$$

*Proof.* Since $\mathcal{P}^{(R)} = \emptyset$, for all $P, Q \in \mathcal{P}^{(R-1)}$ with $P \neq Q$, $P \cap Q = \emptyset$.

With the polytopes in $\mathcal{P}^{(R-1)}$ and the function $f$, which is also defined on their union, we observe that the prerequisites of Lemma 3.38 are fulfilled:

The polytopes in $\mathcal{P}^{(R-1)}$ are subsets of the ones in $\mathcal{P}$ such that also for all $P \in \mathcal{P}^{(R-1)}$, $0 \notin \operatorname{aff}(P)$. The condition $\operatorname{cone}(P \cap Q) \cup \{0\} = \operatorname{cone}(P) \cap \operatorname{cone}(Q)$ also holds for all $P, Q \in \mathcal{P}^{(R-1)}$. For any $P \in \mathcal{P}^{(R-1)}$, there exists a $P' \in \mathcal{P}$ such that $P \subset P'$. By assumption, there is an $f_{P'} : \mathbb{R}^m \to \mathbb{R}$, represented by a ReLU network with $\leq \lceil \log_2(m'+1) \rceil$ hidden layers, such that for all $x \in P'$, $f_{P'}(x) = f(x)$. Furthermore, since all intersections of polytopes in $\mathcal{P}^{(R-1)}$ are empty, the condition that $f$ is 0 on the intersections, trivially holds.

So all requirements of Lemma 3.38 are fulfilled and there is a function $\tilde{f}_{(R-1)}$, represented by an unbiased network with $\leq \lceil \log_2(m'+1) + 1 \rceil$ hidden layers such that for all $x \in \bigcup_{P \in \mathcal{P}^{(R-1)}} P$,

$$\tilde{f}_{(R-1)}(x) = f(x).$$

Now we iteratively apply Lemma 3.38 again for $r = R-2, R-3, \ldots, 0$. We assume that there is a function $\tilde{f}_{(r+1)} : \mathbb{R}^m \to \mathbb{R}$, represented by an unbiased ReLU network with $\leq \lceil \log_2(m'+1) + 1 \rceil$ layers, such that for all $x \in \bigcup_{P \in \mathcal{P}^{(r+1)}} P$,

$$\tilde{f}_{(r+1)}(x) = f(x). \tag{3.27}$$

We define $g_{(r)} : \mathbb{R}^m \to \mathbb{R}$ by $g_{(r)}(x) = \tilde{f}_{(r+1)}(x) - f(x)$. Now we check the conditions to apply Lemma 3.38 to represent the function $g_{(r)}$ on the polytopes in $\mathcal{P}^{(r)}$. Clearly, the conditions $0 \notin P$ and $\operatorname{cone}(P \cap Q) \cup \{0\} = \operatorname{cone}(P) \cap \operatorname{cone}(Q)$ are again fulfilled for all $P, Q \in \mathcal{P}^{(r)}$. Now take any $P \in \mathcal{P}^{(r)}$. Then there is a $P' \in \mathcal{P}$ such that $P \subset P'$ and a corresponding $f_{P'}$, represented by a ReLU network with $\leq \lceil \log_2(m'+1) \rceil$ hidden layers, such that for all $x \in P'$, $f_{P'}(x) = f(x)$.

Then $g_{(r,P)} := \tilde{f}_{(r+1)} - f_{P'}$ can be represented by a ReLU network with $\leq \lceil \log_2(m'+1) + 1 \rceil$ hidden layers. With $m'' := \dim(P) \leq m'$, there is a bijective affine transformation $h : \mathbb{R}^{m''} \to \operatorname{aff}(P)$. Then $h^{-1} : \operatorname{aff}(P) \to \mathbb{R}^{m''}$ is also affine linear and $g_{(r,P)} \circ h : \mathbb{R}^{m''} \to \mathbb{R}$ can be represented by a ReLU network. With the classification Theorem 3.34, we can conclude that $g_{(r,P)} \circ h$ is a CPWL function on $\mathbb{R}^{m''}$ and can therefore also be represented by a ReLU network with $\leq \lceil \log_2(m'+1) \rceil$ layers. $g_{(r,P)} \circ h \circ h^{-1} : \operatorname{aff}(P) \to \mathbb{R}$ agrees with $g_{(r,P)}$ on its domain and because $h \circ h^{-1}$ is an affine transformation, it can also be represented by a ReLU network with $\leq \lceil \log_2(m'+1) \rceil$ layers. Let $\tilde{g}_{(r,P)} : \mathbb{R}^m \to \mathbb{R}$ be the function represented by this network (on the entire space $\mathbb{R}^m$). Then $\tilde{g}_{(r,P)}$ and $g_{(r,P)}$ agree on $P$ and we have for all $x \in P$,

$$\tilde{g}_{(r,P)}(x) = \tilde{f}_{(r+1)}(x) - f(x) = g_{(r)}(x),$$

showing the first requirement of Lemma 3.38 on the function $g_{(r)}$.

For the second requirement on $g_{(r)}$, consider $x \in P \cap Q$ where $P, Q \in \mathcal{P}^{(r)}$, $P \neq Q$. Then there is a $P' \in \mathcal{P}^{(r+1)}$ such that $P \cap Q \subset P'$. Especially, $x$ is contained in the set $\bigcup_{Q' \in \mathcal{P}^{(r+1)}} Q'$, for which we have shown $g_{(r)}(x) = \tilde{f}_{(r+1)}(x) - f(x) = 0$.

This completes the check of the prerequisites for Lemma 3.38 such that we can conclude that there exists $\tilde{g}_{(r)} : \mathbb{R}^m \to \mathbb{R}$, represented by an unbiased ReLU network with $\leq \lceil \log_2(m'+1) + 1 \rceil$ hidden layers, such that for all $x \in \bigcup_{P \in \mathcal{P}^{(r)}} P$, $\tilde{g}_{(r)}(x) = g_{(r)}(x)$. Defining $\tilde{f}_{(r)}(x) := \tilde{f}_{(r+1)}(x) - \tilde{g}_{(r)}(x)$, this function can also be represented by an unbiased ReLU network with $\leq \lceil \log_2(m'+1) + 1 \rceil$ hidden layers, we obtain that for all $x \in \bigcup_{P \in \mathcal{P}^{(r)}} P$,

$$\tilde{f}_{(r)}(x) = \tilde{f}_{(r+1)}(x) - \tilde{g}_{(r)}(x) = g_{(r)}(x) - \tilde{g}_{(r)}(x) + f(x) = f(x),$$

which matches with with the induction hypothesis (3.27) for the next step.

So by induction, it follows that there is a function $\tilde{f} = \tilde{f}_{(0)}$ on $\mathbb{R}^m \to \mathbb{R}$, represented by a ReLU network with $\leq \lceil \log_2(m'+1) + 1 \rceil$ layers, such that for all $x \in \bigcup_{P \in \mathcal{P}} P$,

$$\tilde{f}(x) = f(x),$$

which completes the proof of the lemma. □

In the next step, we show that the conditions on the polytopes of the previous lemma are fulfilled. The $\mathcal{P}^{(r)}$ below consist of the polytopes whose union is the image under $A$ of the intersection of $\Sigma_{s-r}$ with the $\ell_1$ unit sphere.

**Lemma 3.40.** *Let $a_1, \dots, a_n \in \mathbb{R}^m$ such that any $M \subset \{a_1, \dots, a_n\}$ with $|M| \leq 2s$ is linearly independent.*

*For each $\tau \in \{0, \pm 1\}^n$, define*

$$\Gamma_\tau := \operatorname{conv}(\{\tau_j a_j \,|\, 1 \leq j \leq n \text{ and } \tau_j \neq 0\}).$$

*and then*

$$\mathcal{P}^{(r)} = \{\Gamma_\tau \,|\, \|\tau\|_0 = s - r\}$$

*for $0 \leq r \leq s-1$ and $\mathcal{P}^{(s)} = \emptyset$.*

*Then $\mathcal{P} := \mathcal{P}^{(0)}, \mathcal{P}^{(1)}, \dots, \mathcal{P}^{(s)}$ are finite sets of non-empty polytopes satisfying the conditions (a) to (e) from Lemma 3.39 with $m' = s-1$.*

*Proof.* Clearly, all $\mathcal{P}^{(r)}$ are finite. Since every element of every $\mathcal{P}^{(r)}$ is a convex hull of finitely many and more than 0 points, it is a non-empty polytope (Theorem 3.32).

(a) follows directly from the definitions.

Proof of (b): Consider $\tau \in \{0, \pm 1\}^n$ with $\|\tau\|_0 \leq s$ and $\tau \neq 0$. If $0 \in \operatorname{aff}(\Gamma_\tau)$, then $0 \in \operatorname{aff}(\{\tau_j a_j \,|\, 1 \leq j \leq n \text{ and } \tau_j \neq 0\})$. So for $S = \{j \in [n] \,|\, \tau_j \neq 0\}$, there exist coefficients $(\lambda_j)_{j \in S}$ such that $\sum_{j \in S} \lambda_j \tau_j a_j = 0$ and $\sum_{j \in S} \lambda_j = 1$. Since $|S| \leq 2s$, by assumption the $(a_j)_{j \in S}$ are linearly independent. This implies $\lambda_j = 0$ for all $j \in S$, contradicting the fact that the sum of the $\lambda_j$ is 1.

Moreover, any $\Gamma_\tau$ is the convex hull of $\leq s$ points and therefore $\dim(\Gamma_\tau) \leq s-1$.

Proof of (c): Consider any $\tau, \tau' \in \{0, \pm 1\}^n$ with $0 < \|\tau\|_0, \|\tau'\|_0 \leq s$. It always holds that $\Gamma_\tau \cap \Gamma_{\tau'} \subset \operatorname{cone}(\Gamma_\tau) \cap \operatorname{cone}(\Gamma_{\tau'})$ and thus

$$\operatorname{cone}(\Gamma_\tau \cap \Gamma_{\tau'}) \cup \{0\} \subset \operatorname{cone}(\Gamma_\tau) \cap \operatorname{cone}(\Gamma_{\tau'}).$$

On the other hand, consider any $x \in \operatorname{cone}(\Gamma_\tau) \cap \operatorname{cone}(\Gamma_{\tau'})$. Define $S = \operatorname{supp}(\tau)$, $S' := \operatorname{supp}(\tau')$ and $\bar{S} := S \cup S'$. By the definition of $\operatorname{cone}(\Gamma_\tau)$ and $\operatorname{cone}(\Gamma_{\tau'})$ there are coefficients $(\mu_j)_{j \in S}$, $(\mu'_j)_{j \in S'} \geq 0$ such that

$$\sum_{j \in S} \mu_j \tau_j a_j = x = \sum_{j \in S'} \mu'_j \tau'_j a_j.$$

These are two representations of $x$ as a linear combination of $(a_j)_{j \in \bar{S}}$ which is linearly independent since $|\bar{S}| \leq 2s$. So the coefficients must be equal and since all $\mu_j$, $\mu'_j$ are $\geq 0$, $\mu_j = 0$ for all $j \notin \hat{S}$ where $\hat{S} = \{j \in [n] \,|\, \tau_j = \tau'_j \neq 0\}$. If $\mu_j = 0$ for all $j \in \hat{S}$, then $x = 0$ and the other inclusion follows directly. Otherwise

$$\frac{1}{\sum_{l \in \hat{S}} \mu_l} x = \sum_{j \in \hat{S}} \frac{\mu_j}{\sum_{l \in \hat{S}} \mu_l} \tau_j a_j$$

is a convex combination of $(\tau_j a_j)_{j \in \hat{S}}$ and thus in $\Gamma_\tau \cap \Gamma_{\tau'}$ such that $x \in \operatorname{cone}(\Gamma_\tau \cap \Gamma_{\tau'})$.

Proof of (d): Consider any $\tau, \tau' \in \{0, \pm 1\}^n$ with $\|\tau\|_0 = \|\tau'\|_0 = s-r$ and $\tau \neq \tau'$, $\Gamma_\tau \cap \Gamma_{\tau'} \neq \emptyset$. Note that since this intersection is non-empty, we cannot have $r = s-1$. Again, we define $S = \operatorname{supp}(\tau)$, $S' = \operatorname{supp}(\tau')$ and $\bar{S} = S \cup S'$. Any $x \in \Gamma_\tau \cap \Gamma_{\tau'}$ can then be represented as a convex combination

$$\sum_{j \in S} \lambda_j \tau_j a_j = x = \sum_{j \in S'} \lambda'_j \tau'_j a_j,$$

where $\lambda_j, \lambda'_j \geq 0$ for all $j$ and $\sum_{j \in S} \lambda_j = \sum_{j \in S'} \lambda'_j = 1$. Also here, $|\bar{S}| \leq 2s$ shows that $\bar{S}$ is independent and thus the coefficients must be equal such that $\lambda_j = 0$ and $\lambda'_j = 0$ whenever $j \notin \hat{S}$ for $\hat{S} = \{j \in [n] \,|\, \tau_j = \tau'_j \neq 0\}$. So if we define $\tilde{\tau} \in \{0, \pm 1\}^n$ such that $\tilde{\tau}_j = \tau_j$ for $j \in \hat{S}$ and $\tilde{\tau}_j = 0$ otherwise, then $x \in \Gamma_{\tilde{\tau}}$. Since $\hat{S} \subset S \cap S'$ and $\tau \neq \tau'$, we must have $\|\tilde{\tau}\|_0 = |\hat{S}| \leq s - r - 1$. Modifying $\tilde{\tau}$ to $\tilde{\tau}' \in \{0, \pm 1\}^n$ by changing some of its entries from 0 to 1, we can achieve that $\|\tilde{\tau}'\|_0 = s - r - 1$ and then $\Gamma_{\tilde{\tau}} \subset \Gamma_{\tilde{\tau}'}$ such that $x \in \Gamma_{\tilde{\tau}'}$ and $\Gamma_{\tilde{\tau}'} \in \mathcal{P}^{(r+1)}$.

Proof of (e): For any $\tau \in \{0, \pm 1\}^n$, $\|\tau\|_0 = s - r - 1 < n$, we can construct $\tau' \in \{0, \pm 1\}^n$ with $\|\tau'\|_0 = s - r$ by changing one of the entries from 0 to 1. Then $\Gamma_\tau \subset \Gamma_{\tau'}$ and $\Gamma_{\tau'} \in \mathcal{P}^{(r)}$. □

### 3.6.4 Proof of Theorem 3.30

Now we can combine all the previous tools to prove the main result about exact recovery.

*Proof of Theorem 3.30.* Let $a_1, \ldots, a_n$ be the columns of $A$. Then any subset of $\{a_1, \ldots, a_n\}$ of size $\leq 2s$ is linearly independent. Otherwise there would be an $x \in \Sigma_{2s} \backslash \{0\}$ such that $Ax = 0$.

For $a_1, \ldots, a_n$ and $s$, by Lemma 3.40, the sets $\mathcal{P}, \mathcal{P}^{(0)}, \ldots, \mathcal{P}^{(s)}$ satisfy (a) to (e) from Lemma 3.39 with $m' = s - 1$.

The map $A$ is injective on $\Sigma_s$, otherwise there would be $x \in \Sigma_{2s} \backslash \{0\}$ such that $Ax = 0$. So we can define an inverse map $f_0 : A\Sigma_s \to \Sigma_s$. For each $\tau \in \{0, \pm 1\}^n$, $1 \leq \|\tau\|_0 \leq s$, $A$ maps the subspace $\operatorname{span}(\{\tau_j e_j \,|\, j \in [n], \tau_j \neq 0\})$ bijectively to the subspace $\operatorname{span}(\{\tau_j a_j \,|\, j \in [n], \tau_j \neq 0\})$. So the inverse map $f_0$ is again linear on $\operatorname{span}(\{\tau_j a_j \,|\, j \in [n], \tau_j \neq 0\})$ and thus also on the subset $\Gamma_\tau$. So for each $P \in \mathcal{P}^{(r)}$, $r = 0, \ldots, s$, $f_0$ restricted to $P$ is linear and can thus be represented exactly by a ReLU network with one hidden layer, i.e., the condition (3.26) holds.

Altogether, we can apply Lemma 3.39 to $f_0$, such that there exists $f : \mathbb{R}^m \to \mathbb{R}^n$, represented by an unbiased ReLU network with $\lceil \log(s) + 1 \rceil$ hidden layers such that for all $y \in \bigcup_{P \in \mathcal{P}} P$,

$$f(y) = f_0(y).$$

Now for all $x \in \Sigma_s \backslash \{0\}$, $\frac{x}{\|x\|_1} \in \operatorname{conv}(\{\operatorname{sign}(x_j) e_j \,|\, \operatorname{sign}(x_j) \neq 0\})$, so $A\frac{x}{\|x\|_1} \in \Gamma_{\operatorname{sign}(x)}$ and then by the positive homogeneity of $f$,

$$f(Ax) = \|x\|_1 f(A\frac{x}{\|x\|_1}) = \|x\|_1 f_0(A\frac{x}{\|x\|_1}) = \|x\|_1 \cdot \frac{x}{\|x\|_1} = x,$$

and $f(Ax) = x$ for $x = 0$ follows directly by positive homogeneity.

□

## 3.7 Construction of Small Networks for $s = 1$

Although Corollary 3.4 states that we can solve the sparse recovery problem with a network of small depth, because of using the universal approximation theorem, it does not state anything about the width of the network. With the following result, we establish a first step towards investigating the width required for successful sparse recovery. We show that for 1-sparse vectors, the problem can be solved with a network with two hidden layers of comparably small widths $\lceil 5 \log(n) \rceil$ and $2n$.

**Theorem 3.41.** *Let $n > 2$ and $A \in \mathbb{R}^{m \times n}$, $m \geq 2$ such that $Ax \neq 0$ for all $x \in \Sigma_2$. Then for $k_1 = \lceil 5 \log(n) \rceil$, $k_2 = 2n$, there exist matrices $W_3 \in \mathbb{R}^{n \times k_2}$, $W_2 \in \mathbb{R}^{k_2 \times k_1}$, $W_1 \in \mathbb{R}^{k_1 \times n}$ such that the neural network function $f(y) = W_3 \operatorname{ReLU}(W_2 \operatorname{ReLU}(W_1 y))$ satisfies $f(Ax) = x$ for every $x \in \Sigma_1$.*

*Proof of Theorem 3.41.* Consider the matrix

$$B = \operatorname{ReLU}(W_1 A \begin{bmatrix} Id_n & -Id_n \end{bmatrix}) \in \mathbb{R}^{k_1 \times 2n}.$$

We first show that i) there is a choice of $W_1$ so that no two columns of $B$ are parallel. Next, we use this fact to show that ii) there is a choice of $W_2 \in \mathbb{R}^{2n \times k_1}$ so that $\operatorname{ReLU}(W_2 B) \in \mathbb{R}^{2n \times 2n}$ is invertible. Finally, define $W_3 = \begin{bmatrix} Id_n & -Id_n \end{bmatrix} (\operatorname{ReLU}(W_2 B))^{-1}$. It follows that

$$W_3 \operatorname{ReLU}(W_2 \operatorname{ReLU}(W_1 A \begin{bmatrix} Id_n & -Id_n \end{bmatrix})) = W_3 \operatorname{ReLU}(W_2 B) = \begin{bmatrix} Id_n & -Id_n \end{bmatrix},$$

establishing that for any $j \in [n]$,

$$f(Ae_j) = W_3 \operatorname{ReLU}(W_2 \operatorname{ReLU}(W_1(Ae_j))) = W_3 \operatorname{ReLU}(W_2 B) e_j = e_j,$$

and similarly $f(A(-e_j)) = -e_j$. Since the ReLU function is positive homogeneous, this implies that $f(\lambda v) = \lambda f(v)$ holds for every vector $v \in \mathbb{R}^n$ and thus $f(Ax) = x$ is satisfied for any 1-sparse $x \in \mathbb{R}^n$, which concludes the proof. It remains to establish the facts i and ii.

**i) There is a choice of the matrix $W_1$ so that no two columns of the matrix $B$ are parallel:** From the assumption that $Ax \neq 0$ for all $x \in \Sigma_2$, it follows that there are no two columns $a \neq b$ of $A$ such that $a$ and $b$ are parallel, and $A$ does not have a zero-column.

Define $\tilde{W}_1 \in \mathbb{R}^{k_1 \times n}$ to have independent $N(0,1)$ entries. For $J \subset \{1, \dots, k_1\}$, define $R_J \in \mathbb{R}^{|J| \times k_1}$ to be the restriction operator to the entries in $J$. For any vectors $a, b \in \mathbb{R}^n$ such that $a \nparallel b$, consider the probability

$$\begin{aligned} p_{a,b} &:= \mathbb{P}\left(\exists i, j \in [k_1] : i \neq j \text{ and } R_{\{i,j\}} \tilde{W}_1 a \parallel R_{\{i,j\}} \tilde{W}_1 b\right) \\ &\leq \mathbb{P}\left(\exists i, j \in [k_1] : i \neq j \text{ and } \det \begin{bmatrix} \langle (\tilde{W}_1)_i, a \rangle & \langle (\tilde{W}_1)_i, b \rangle \\ \langle (\tilde{W}_1)_j, a \rangle & \langle (\tilde{W}_1)_j, b \rangle \end{bmatrix} = 0 \right) \\ &\leq \sum_{\substack{i,j \in [k_1] \\ i \neq j}} \mathbb{P}\left(\langle (\tilde{W}_1)_i, a \rangle \langle (\tilde{W}_1)_j, b \rangle - \langle (\tilde{W}_1)_i, b \rangle \langle (\tilde{W}_1)_j, a \rangle = 0\right) \\ &= \sum_{\substack{i,j \in [k_1] \\ i \neq j}} \mathbb{P}\left(\langle (\tilde{W}_1)_i, a \langle (\tilde{W}_1)_j, b \rangle - b \langle (\tilde{W}_1)_j, a \rangle \rangle = 0\right) \\ &= \sum_{\substack{i,j \in [k_1] \\ i \neq j}} \mathbb{P}\left(a \langle (\tilde{W}_1)_j, b \rangle - b \langle (\tilde{W}_1)_j, a \rangle = 0\right), \end{aligned}$$

where we used that the rows $(\tilde{W}_1)_i$ and $(\tilde{W}_1)_j$ are independent and for any fixed vector $v \in \mathbb{R}^n$ and $g \sim N(0, Id_n)$, $\mathbb{P}(\langle g, v \rangle = 0) = 0$ if $v \neq 0$. Since $a \langle (\tilde{W}_1)_j, b \rangle - b \langle (\tilde{W}_1)_j, a \rangle = 0$ can only hold if $a \parallel b$, we obtain $p_{a,b} = 0$.

Let $V$ be the set of columns of the matrix $[A, -A]$. For any $v \in V$, since $v \neq 0$, and any index $j \in [k_1]$, $(\tilde{W}_1 v)_j$ is a Gaussian variable with mean 0 and positive variance, so $\mathbb{P}((\phi(\tilde{W}_1 v))_j \neq 0) = \mathbb{P}((\tilde{W}_1 v)_j > 0) = \frac{1}{2}$. Since the entries of $\tilde{W}_1 v$ are independent,

$$\begin{aligned} \mathbb{P}(\|\phi(\tilde{W}_1 v)\|_0 \leq 1) &= \mathbb{P}(\|\phi(\tilde{W}_1 v)\|_0 = 0) + \mathbb{P}(\|\phi(\tilde{W}_1 v)\|_0 = 1) \\ &= (1/2)^{k_1} + (1/2)^{k_1} n \\ &= 2^{-k_1}(n+1). \end{aligned}$$

Now we obtain

$$\begin{aligned} \mathbb{P}(\exists v \in V : \|\phi(\tilde{W}_1 v)\|_0 \leq 1) &\leq \sum_{v \in V} \mathbb{P}(\|\phi(\tilde{W}_1 v)\|_0 \leq 1) \\ &\leq |V| \max_{v \in V} \mathbb{P}(\|\phi(\tilde{W}_1 v)\|_0 \leq 1) \\ &\leq 2n(n+1) 2^{-k_1} \end{aligned}$$

$$\begin{aligned}&\leq 2n(n+1)2^{-5\log n} = 2n(n+1)n^{-5\log 2}\\ &\leq 2n(n+1)n^{-3} = \frac{2}{n} + \frac{2}{n^2} < 1\end{aligned}$$

since $n \geq 3$.

This implies that $\mathbb{P}(\forall v \in V : \|\phi(\tilde{W}_1 v)\|_0 \geq 2) > 0$, i.e., we can choose one realization $W_1$ of $\tilde{W}_1$ such that $\|\phi(W_1 v)\|_0 \geq 2$ for all columns of $[A^T A, -A^T A]$ and simultaneously for any distinct indices $i, j \in [k_1]$ and any non-parralel $a, b \in V$, $R_{\{i,j\}} W_1 a \not\parallel R_{\{i,j\}} W_1 b$.

Now assume that there are two parralel columns of $B$. Then there are two different $a, b \in V$ such that $\phi(W_1 a) = \lambda \phi(W_1 b)$ for a $\lambda \in \mathbb{R}$. Both $\phi(W_1 a)$ and $\phi(W_1 b)$ only have non-negative entries, so $\lambda \geq 0$. Since $\|\phi(W_1 a)\|_0 \geq 2$, $\phi(W_1 a) \neq 0$, thus $\lambda > 0$. By the choice of $W_1$, we can also pick two distinct indices $i, j \in [k_1]$ such that all entries in $R_{\{i,j\}}\phi(W_1 b)$ are $> 0$ and because of $\lambda > 0$ also the entries of $R_{\{i,j\}}\phi(W_1 a)$ are $> 0$.

This implies $R_{\{i,j\}}\phi(W_1 v) = R_{\{i,j\}} W_1 v$ for $v = a, b$ and thus $R_{\{i,j\}} W_1 a = \lambda R_{\{i,j\}} W_1 b$, i.e., $R_{\{i,j\}} W_1 a \parallel R_{\{i,j\}} W_1 b$. This however, can only happen if $a \parallel b$, as observed above. By the definition of $V$, this is only possible for $b = -a$. Since $\phi(W_1 a)_j \neq 0$ (i.e. $(W_1 a)_j > 0$) excludes $\phi(W_1(-a))_j \neq 0$, the vectors $\phi(W_1 a)$ and $\phi(W_1(-a))$ have disjoint supports, contradicting the assumption that they are non-zero and parallel.

So $B$ cannot have any two parralel columns.

**ii) There is a choice of $W_2 \in \mathbb{R}^{2n \times k_1}$ so that $\mathrm{ReLU}(W_2 B)$ is invertible:** Now let $g \sim N(0, Id_{k_1})$ and define the symmetric matrix $M := \mathbb{E}[\phi(B^T g)\phi(B^T g)^T] \in \mathbb{R}^{2n \times 2n}$. We show that $M$ and later also its random approximation $\phi(W_2 B)^T \phi(W_2 B)$ have rank $2n$, which establishes that the matrix $\mathrm{ReLU}(W_2 B)$ is invertible, as desired. A related scenario has been analyzed in [Du+19]. The following proof adapts those methods to our setup.

**ii)-1:** We first show that $M$ has rank $2n$. Towards this goal, assume for contradiction that $\mathrm{rank}(M) < 2n$. Then there exists a vector $u \in \mathbb{R}^{2n} \backslash \{0\}$ such that

$$0 = u^T M u = \mathbb{E}[u^T \phi(B^T g)\phi(B^T g)^T u] = \mathbb{E}[|\phi(B^T g)^T u|^2].$$

Since $|\phi(B^T g)^T u|^2 \geq 0$, this implies that $\phi(B^T g)^T u = 0$ holds almost surely. Since the Lebesgue measure $\lambda$ is absolutely continuous with respect to the probability measure of the standard normal distribution on $\mathbb{R}^{k_1}$, the equation $\phi(B^T x)^T u = 0$ has to hold $\lambda$-almost everywhere in $x$ and by continuity this implies $\phi(B^T x)^T u = 0$ for all $x \in \mathbb{R}^{k_1}$.

Let $b_1, \ldots, b_{2n}$ be the columns of the matrix $B$. Also, for $j \in [2n]$ let $D_j := \{x \in \mathbb{R}^{k_1} : \langle x, b_j \rangle = 0\}$. Since there are no parallel columns of $B$, $D_j \not\subset \bigcup_{l \neq j} D_l$ (see [Du+19], Lemma A.1). We know that $\sum_{j=1}^{2n} u_j \phi(\langle x, b_j \rangle) = 0$ holds for all $x \in \mathbb{R}^{k_1}$. Now fix a particular index $j_0 \in [2n]$. Choose $x \in D_{j_0} \backslash \bigcup_{j \neq j_0} D_j$. For all $j \neq j_0$, $\langle x, b_j \rangle \neq 0$ and by continuity, for a sufficiently small $\epsilon > 0$, for all $j \neq j_0$, $\mathrm{sign}(\langle x \pm \epsilon b_{j_0}, b_j \rangle) = \mathrm{sign}(\langle x, b_j \rangle)$. Because of $0 = \sum_{j=1}^{2n} u_j \phi(\langle x, b_j \rangle) = \sum_{j=1}^{2n} u_j \phi(\langle x + \epsilon b_{j_0}, b_j \rangle) = \sum_{j=1}^{2n} u_j \phi(\langle x - \epsilon b_{j_0}, b_j \rangle)$, we have

$$0 = \sum_{j=1}^{2n} u_j \left[\phi(\langle x + \epsilon b_{j_0}, b_j \rangle) + \phi(\langle x - \epsilon b_{j_0}, b_j \rangle) - 2\phi(\langle x, b_j \rangle)\right]. \tag{3.28}$$

For each $j \in [2n]$, one of the following holds

- $j \neq j_0$ and $\langle x, b_j \rangle > 0$: Then by assumption on $\epsilon$, also $\langle x \pm \epsilon b_{j_0}, b_j \rangle > 0$ and then

$$\begin{aligned}\phi(\langle x + \epsilon b_{j_0}, b_j \rangle) + \phi(\langle x - \epsilon b_{j_0}, b_j \rangle) - 2\phi(\langle x, b_j \rangle) &= \langle x + \epsilon b_{j_0}, b_j \rangle + \langle x - \epsilon b_{j_0}, b_j \rangle - 2\langle x, b_j \rangle\\ &= 0.\end{aligned}$$

- $j \neq j_0$ and $\langle x, b_j \rangle < 0$: Then by assumption on $\epsilon$, also $\langle x \pm \epsilon b_{j_0}, b_j \rangle < 0$ and then
$$\phi(\langle x + \epsilon b_{j_0}, b_j \rangle) + \phi(\langle x - \epsilon b_{j_0}, b_j \rangle) - 2\phi(\langle x, b_j \rangle) = 0 + 0 - 0 = 0.$$

- $j = j_0$: Then $\langle x \pm \epsilon b_{j_0}, b_j \rangle = \pm\epsilon \|b_{j_0}\|_2^2$, i.e.
$$\phi(\langle x + \epsilon b_{j_0}, b_j \rangle) + \phi(\langle x - \epsilon b_{j_0}, b_j \rangle) - 2\phi(\langle x, b_j \rangle) = \epsilon \|b_{j_0}\|_2^2 + 0 - 0 = \epsilon \|b_{j_0}\|_2^2.$$

Substituting into (3.28) yields $0 = u_{j_0}\epsilon\|b_{j_0}\|_2^2$, so $u_{j_0} = 0$. Since this argument holds for all $j_0 \in [2n]$, we obtain $u = 0$, contradicting the assumption that the expectation matrix $M$ does not have the maximal rank $2n$. So $\operatorname{rank}(M) = 2n$.

**ii)-2:** Now let $\tilde{W}_2 \in \mathbb{R}^{k_2' \times k_1}$ be a matrix with independent $N(0,1)$ entries. Define
$$\tilde{X} := \phi(\tilde{W}_2 B)^T \phi(\tilde{W}_2 B)$$
such that for $(j,l) \in [2n]^2$, and $(\tilde{W}_2)_r$ being the $r$-th row of $\tilde{W}_2$,
$$\tilde{X}_{jl} - \mathbb{E}[\tilde{X}_{jl}] = \sum_{r \in [k_2']} \left[ \phi((\tilde{W}_2)_r b_j)\phi((\tilde{W}_2)_r b_l) - M_{jl} \right] =: \sum_{r \in [k_2']} \tilde{X}_{jl}^{(r)}.$$

We show that $\tilde{X}$ has full rank $2n$ with high probabiliy, and therefore there exists a realization of weights $W_2$ so that the matrix $\phi(W_2 B)^T$ has full rank.

For $p \geq 2$,
$$\begin{aligned}
\mathbb{E}|\tilde{X}_{jl}^{(r)}|^p &= \mathbb{E}|\phi((\tilde{W}_2)_r b_j)\phi((\tilde{W}_2)_r b_l) - M_{jl}|^p \\
&\leq 2^p \left[ \mathbb{E}|\phi((\tilde{W}_2)_r b_j)\phi((\tilde{W}_2)_r b_l)|^p + \mathbb{E}|M_{jl}|^p \right] \\
&= 2^p \left[ \mathbb{E}|\phi((\tilde{W}_2)_r b_j)\phi((\tilde{W}_2)_r b_l)|^p + [\mathbb{E}|\phi((\tilde{W}_2)_r b_j)\phi((\tilde{W}_2)_r b_l)|]^p \right] \\
&\leq 2^p \left[ 2\mathbb{E}|\phi((\tilde{W}_2)_r b_j)\phi((\tilde{W}_2)_r b_l)|^p \right] \\
&\leq 2^p \left[ \mathbb{E}|\phi((\tilde{W}_2)_r b_j)^2 + \phi((\tilde{W}_2)_r b_l)^2|^p \right] \\
&\leq 4^p \left[ \mathbb{E}|\phi((\tilde{W}_2)_r b_j)|^{2p} + \mathbb{E}|\phi((\tilde{W}_2)_r b_l)|^{2p} \right]
\end{aligned}$$

Considering that for any $a \in \mathbb{R}^n$ and $g \sim N(0, Id_n)$,
$$\mathbb{E}|\phi(\langle g, a \rangle)|^{2p} \leq \mathbb{E}|\langle g, a \rangle|^{2p} = \frac{(2p)!}{2^p p!}\|a\|_2^{2p} \leq \frac{2^{2p}(p!)^2}{2^p p!}\|a\|_2^{2p} = 2^p p! \|a\|_2^{2p},$$
we obtain
$$\mathbb{E}|\tilde{X}_{jl}^{(r)}|^p \leq 8^p p! (\|b_j\|_2^{2p} + \|b_l\|_2^{2p}) \leq p! R^{p-2}\sigma^2/2$$
for $R = 8\max_{l \in [2n]} \|b_l\|_2^2$, $\sigma^2 = 256 \max_{l \in [2n]} \|b_l\|_2^4$.

Since the $\tilde{X}_{jl}^{(r)}$ are mean 0, by Bernstein's inequality ([FR13], Theorem 7.30), we obtain
$$\mathbb{P}\left( \frac{1}{k_2'} \left| \tilde{X}_{jl} - \mathbb{E}[\tilde{X}_{jl}] \right| \geq t \right) \leq 2\exp\left( -\frac{(k_2')^2 t^2/2}{k_2'\sigma^2 + Rtk_2'} \right) = \exp\left( -\frac{k_2' t^2/2}{\sigma^2 + Rt} \right)$$

Note that for any fixed $t > 0$, the right hand side converges to 0 if $k_2' \to \infty$.

Let $\lambda_0 > 0$ be the smallest eigenvalue of $M$. Choose $t := \frac{\lambda_0}{4n}$ and pick a $k_2'$ such that for all $(j,l) \in [2n]$,
$$\mathbb{P}\left( \frac{1}{k_2'} \left| \tilde{X}_{jl} - \mathbb{E}[\tilde{X}_{jl}] \right| \geq t \right) < \frac{1}{8n^2}.$$

Then

$$\mathbb{P}\left(\exists(j,l) \in [2n]^2 : \frac{1}{k_2'}\left|\tilde{X}_{jl} - \mathbb{E}[\tilde{X}_{jl}]\right| \geq t\right) < |[2n]|^2 \times \frac{1}{8n^2} = \frac{1}{2}.$$

In the event that $\frac{1}{k_2'}\left|\tilde{X}_{jl} - \mathbb{E}[\tilde{X}_{jl}]\right| < t$, which holds with probability $> \frac{1}{2}$, we have

$$\|\frac{1}{k_2'}\tilde{X} - M\|^2 \leq \|\frac{1}{k_2'}\tilde{X} - M\|_F^2 = \sum_{(j,l)\in[2n]^2}\left|\frac{1}{k_2'}\tilde{X}_{jl} - M_{jl}\right| <$$
$$4n^2 \cdot t^2 = 4n^2\left(\frac{\lambda_0}{4n}\right)^2 = \frac{\lambda_0^2}{4},$$

implying that

$$\lambda_{\min}(\frac{1}{k_2'}\tilde{X}) \geq \lambda_{\min}(M) - \|\frac{1}{k_2'}\tilde{X} - M\| \geq \lambda_0 - \frac{\lambda_0}{2} = \frac{\lambda_0}{2} > 0.$$

Then the matrix $\tilde{X} \in \mathbb{R}^{2n\times 2n}$ has rank $2n$ and thus by the definition of $\tilde{X}$, also the matrix $\phi(\tilde{W}_2B) \in \mathbb{R}^{k_2'\times 2n}$.

Since this holds with positive probability, there is a realization $\bar{W}_2 \in \mathbb{R}^{k_2'\times k_1}$ of the random matrix $\tilde{W}_2$, such that $\phi(\bar{W}_2B) \in \mathbb{R}^{k_2'\times 2n}$ has rank $2n$. Then $k_2' \geq 2n$ and we can select $2n$ rows of $\phi(\bar{W}_2B)$ such that the resulting matrix still has rank $2n$. We define $W_2 \in \mathbb{R}^{2n\times 2n}$ by taking these rows of the matrix $\bar{W}_2$ such that the matrix $\phi(W_2B)$ is an invertible square matrix.

□

### 3.8 Discussion

In this work, we have shown that ReLU networks with one hidden layer cannot even solve the sparse recovery problem for 1-sparse vectors while in contrast with two hidden layers, they are capable of approximating this problem to an arbitrary precision and for arbitrary sparsity levels. The latter result can also be generalized to a larger class of inverse problems.

A key assumption for these results is that we look at networks that take the positive homogeneous structure of the problem into account. This ensures the reconstruction to work for all possible signals without any bound on their norm.

This also improved our understanding of how continuous positive homogeneous functions can be approximated with neural networks in general. We have seen that the ReLU function plays a unqiue role in their approximation and that the general approximation necessarily requires two layers.

Despite showing that a good solution for the respective inverse problems is possible with rather shallow networks, our results of this work do not provide a statement about the width and efficiency of such networks. The previous Theorem 3.41 accomplishes a first step towards this but is still limited for the case $s = 1$. Possibly, future research could use width-limited versions of the universal approximation theorem to investigate this question. For example, [Tan+20] shows such a statement for positive homogeneous networks (Theorem 2 in the supplement) which is based on Theorem 1 in [Lu+17]. However, these results do not specify the depth of the network. Furthermore, future research could also search for guarantees regarding the training of the networks to solve inverse problems.

Our main theorems also address the robustness of our solution already mentioned at the beginning in Section 3.1.4. We can obtain similar guarantees to minimization-based approaches. This agrees with the empirical observation in [GMM22] that states that neural networks provide a similar robustness to total variation minimization which is related to $\ell_1$ minimization for sparse recovery. Note however, that in our work, we only study the existence of the networks but not training them.

The robustness seemingly contradicts the analysis of [Got+20] which analyzes certain scenarios in which problems with the robustness of neural network for inverse problems occur. In particular, they show (Theorem 3.1 in [Got+20]) that instabilities have to occur if one tries to recover signals whose difference is close to the kernel of the measurement matrix, i.e., $x$ and $x'$ such that $\|x-x'\|$ is large compared to $\|Ax-Ax'\|$. Avoiding this situation is referred to as *kernel awareness* (discussed in Section 4.2 in [Got+20]) which for sparse recovery can be achieved if $\|x\|_2 \leq \gamma\|Ax\|_2$ for a constant $\gamma > 0$ and all $2s$-sparse vectors. In fact, (3.3) in Theorem 3.5 is exactly such a kernel awareness condition which ensures that the considered problems are well-behaved in this respect. This condition is ensured by the $\ell_2$-robust null space property and therefore also the restricted isometry property which are assumed in many compressed sensing scenarios.

However, by a theoretical comparison of neural networks to an optimization approach similar to (3.14) (Theorem 6.3 in [Got+20]), they show that even for a problem related to sparse recovery, neural networks necessarily have a significantly larger local Lipschitz constant in some cases.

In the proof of Theorem 3.5 of our work, we first considered an extended inversion function $f$ which we prove to be Lipschitz continuous. Then we approximate $f$ by a ReLU network $\tilde{f}$ such that $\|\tilde{f}(y) - f(y)\|_2 \leq \delta'\|y\|_2$. And indeed, even though $f$ is Lipschitz continuous and $\delta'$ can be arbitrarily small, we cannot conclude anything about the local Lipschitz constants of $\tilde{f}$ based on this method. Specifically, in the sparse recovery case, Corollary 3.4 states that for an exactly sparse signal $x$ with $\|x\|_2 = 1$,

$$\|\tilde{f}(Ax+e) - x\|_2 \leq \delta' + D\|e\|_2.$$

Therefore, we obtain gradients

$$\frac{\|\tilde{f}(Ax+e) - \tilde{f}(Ax)\|_2}{\|e\|_2} \leq \frac{2\delta'}{\|e\|_2} + D.$$

For very small $\|e\|_2$, specifically $\|e\|_2 \lesssim \delta'$, this becomes very large and therefore it becomes clear that our method cannot provide a bound to control the local Lipschitz constant of $\tilde{f}$. However, these large gradients only occur for very small $\|e\|_2$ and if specifically $\|e\|_2 \geq \delta'$ (recall that $\delta'$ can be chosen arbitrarily small), then the above gradient is bounded by

$$\frac{\|\tilde{f}(Ax+e) - \tilde{f}(Ax)\|_2}{\|e\|_2} \leq 2 + D,$$

i.e., a constant. So to summarize, the networks provided by our method might actually have very large local Lipschitz constants. However, these are only relevant for very small deviations and in this way, robust recovery as in Theorem 3.5 is still possible.

The results in Section 3.5 also show that for a large class of minimization problems, neural networks can achieve the same robustness with respect to perturbations of size $\|e\|_2 \gtrsim \delta'$ even though the local Lipschitz constant might be significantly larger. Nevertheless, we can chose the $\delta'$ arbitrarily close to 0.

With respect to the sparse recovery problem, Corollaries 3.3 and 3.4 show that exactly two hidden layers are the smallest possible depth for approximate recovery. For exact recovery, Theorem 3.30 shows that $\lceil \log(s-1) + 2 \rceil$ hidden layers are sufficient but it is still an open question to what extent this is optimal. This is also related to the question whether the network depth in the CPWL representation Theorem 3.34. As stated in the conclusion section of [He+20], it is known to be optimal for $n = 2, 3$ (when $\lceil \log_2(n+1) \rceil = 2$) but also an open problem for larger $n$.

# 4 Improved Recovery Guarantees for the Sparsity in Levels Class

## 4.1 Introduction

Recall from the introduction Section 0.2 that the $(s, \delta)$-restricted isometry property (RIP) of a matrix $A \in \mathbb{C}^{m \times N}$ given in Definition 0.2 states that

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2$$

holds for all $s$-sparse $x \in \mathbb{C}^N$, and that one of the key insights is that this property guarantees that there is always a unique solution to the sparse recovery problem, which can be computed efficiently. Besides this, the construction of Johnson-Lindenstrauss embeddings as discussed in [KW11] and Section 0.4 provides another application of the RIP in a different context.

As we also discussed in Section 0.2, one particular class of matrices satisfying the RIP are the Gaussian matrices with embedding dimension $m \gtrsim \delta^{-2} s \log\left(\frac{N}{s}\right)$ and it has been shown that this dependence of $m$ is optimal up to constant factors.

We mentioned in the introduction of this thesis that one of the key motivations to apply compressed sensing with sparsity as a structural constraint is that natural images are typically sparse in a wavelet basis, which is discussed in [OSL00].

A more refined view on the coefficient representation of images, as performed in [Adc+17], shows that the wavelet coefficients are not uniformly distributed across all the scales. Instead there are typically more non-zero coefficients at the coarser scales of the wavelet basis and fewer ones at the finer scales. This motivates the definition of the sparsity in levels model. It partitions the entries of the vectors into $r$ fixed blocks and within each block $k \in [r]$, there is a maximal number $s_k$ of possible non-zero entries. For the imaging applications, these blocks would correspond to the scales of a wavelet basis. This is a generalization of the usual sparsity model, which is given by the special case $r = 1$.

Also for this refined model, one can define a restricted isometry property, the *restricted isometry property in levels* (RIPL), which is, for example, again fulfilled by Gaussian matrices.

### 4.1.1 Subsampled Bounded Orthonormal Systems

Even though the theory of the RIP is fully understood for Gaussian matrices, the property has also been studied extensively for other ones. One particular case is given by matrices whose rows are randomly selected from the DFT matrix $F \in \mathbb{C}^{N \times N}$ or the Hadamard matrix $H$ defined in Section 0.6. A partial Fourier matrix $\sqrt{\frac{N}{m}} P_\Omega F \in \mathbb{C}^{m \times N}$ consists of $m$ randomly selected (usually uniformly and independently) and rescaled rows of $F$ and the same can be done for $H$. In general, we can even consider a larger class of randomly subsampled matrices $\sqrt{\frac{N}{m}} P_\Omega U$, where $U \in \mathbb{C}^{N \times N}$ is unitary and bounded in the sense that all its entries satisfy $|U_{j,k}| \leq \frac{L}{\sqrt{N}}$ for a constant $L$. These are known as *bounded orthonormal systems* (BOSs). Most previous works regarding subsampled Fourier matrices also apply to this class.

Subsampled bounded orthonormal systems and specifically subsampled Fourier matrices have been studied for two main advantages over the subgaussian matrices. First, in most applications the measurement matrix cannot be chosen arbitrarily but they are determined by the application. Contrary to subgaussian matrices, measurements that are random Fourier coefficients can be found in multiple applications of compressed sensing including magnetic resonance imaging (MRI) [LDP07; Lus+08] or reducing the mutual inference of different radar systems [Che+22]. The second aspect is the computational complexity. For matrices like the ones with independent subgaussian entries, the best algorithm to compute the matrix-vector product $Ax$ is usually the standard algorithm, which requires $\mathcal{O}(mN)$ operations. As also explained in Section 0.6, there are fast algorithms to compute the matrix-vector product $Fx$ or $Hx$ in only $\mathcal{O}(N \log N)$

operations such that also $P_\Omega Fx$ can be computed in this complexity. This fast matrix-vector multiplication can improve the performance of sparse recovery algorithms such as CoSaMP [NT09]. RIP matrices are known to allow for constructions of Johnson-Lindenstrauss embeddings using [KW11]. In this way, subsampled bounded orthonormal systems can be used to construct Johnson-Lindenstrauss embeddings with a fast application to individual points. Section 1 of this work connects to this topic.

Because of the aforementioned applications, it has been an important research question what number $m$ of rows are sufficient or required for a subsampled BOS $\sqrt{\frac{N}{m}}P_\Omega U \in \mathbb{C}^{m\times N}$ to have the $(s,\delta)$-RIP with a high probability. In fact, there have been multiple results that have successively improved each other from 2006 until 2021. First, Candes and Tao [CT06] showed that for subsampled Fourier matrices, $m \gtrsim C_\delta s(\log N)^6$ is sufficient for $C_\delta > 0$ depending only on $\delta$. Rudelson and Vershynin [RV08] improved this to $m \gtrsim \delta^{-2}s\log(N)\log(m)(\log s)^2$, again for subsampled Fourier matrices. Although these two results are shown for Fourier matrices, their proofs also work for discrete BOSs. Subsequently, this bound was generalized to general, not necessarily discrete bounded orthonoral systems in [Rau10]. The bound was improved to $m \gtrsim \delta^{-2}s\log(N)(\log s)^3$ by Cheraghchi, Guruswami, and Velingker [CGV13] for the Fourier case (with a proof that can also be applied to discrete BOSs). Bourgain [Bou14] then showed that $m \gtrsim C_\delta s(\log N)^2\log s$ is sufficient for $C_\delta$ depending on $\delta$ (for discrete BOSs). Chkifa et al. [Chk+18] improved this to $m \gtrsim \delta^{-6}s\log(N)(\log\frac{s}{\delta})^2$ (for general BOSs). A further improvement by Haviv and Regev [HR16] proves the RIP for $m \gtrsim \delta^{-2}(\log\frac{1}{\delta})^2 s\log(N)(\log\frac{s}{\delta})^2$ (for discrete BOSs). The latest improvement to $m \gtrsim \delta^{-2}s\log(N)(\log\frac{s}{\delta})^2$ was achieved by Brugiapaglia et al. [Bru+21] (for general BOSs).

On the other hand, beyond the known lower bound for RIP matrices in general ($m \gtrsim \delta^{-2}s\log\frac{N}{s}$), [BLM18] shows that for subsampled Fourier matrices $m \gtrsim C_\delta s\log(eN)$ rows are necessary. Furthermore, [Bla+19] shows that for any $\delta \in (0,1)$, $m \gtrsim s\log(\frac{N}{s})\log(s)$ is necessary for the $(s,\delta)$-RIP of a subsampled Hadamard matrix to hold with high probability. Therefore, also no RIP result for arbitrary discrete BOSs can work for a smaller $m$.

Especially the number of required logarithmic factors in $N$ and $s$ (besides the dependence on $\delta$) in the bound has been subject to extensive study as it has been improved from 6 in [CT06] to 3 in [HR16]. On the other hand, [Bla+19] shows that 2 such logarithmic factors are necessary. The remaining gap is still an open problem.

### 4.1.2 Sparsity in Levels and Multilevel Sampling

Measurements of the type $P_\Omega Ux$ for sparse $s$ and unitary $U$ generally represent a scenario in which signals are sparse in one orthonormal basis and randomly sampled in another orthonormal basis. That is, a signal has a representation $V_1x$ for a sparse vector $x \in \mathbb{C}^{N\times N}$ and unitary $V_1 \in \mathbb{C}^{N\times N}$ and samples are taken from $V_2^*V_1x$ for another unitary $V_2 \in \mathbb{C}^{N\times N}$. Then $U := V_2^*V_1$ is unitary and the measurements are $P_\Omega Ux$ for sparse $x$. We call $\mu := \max_{j,k\in[N]}|U_{j,k}|^2$ the *coherence* of the orthonormal bases $V_1$ and $V_2$ or just the coherence of $U$. The above results about subsampled BOSs require all entries of $U$ to be small, specifically $\mu \leq \frac{L^2}{N}$ for a constant $L$. In fact, the RIP can still be guaranteed for $m \gtrsim L^2\delta^{-2}s\cdot\operatorname{polylog}(N)$ if $L$ is larger. This dependence of the sampling complexity on the coherence, first for the case of non-uniform recovery, was first observed in [CR07].

In the case of the Fourier basis as $V_2$ and the canonical basis $V_1$, the smallest possible coherence $\mu = \frac{1}{N}$ is achieved since all entries of $U$ have absolute value $\frac{1}{\sqrt{N}}$. However, for other pairs of orthonormal bases such as the Fourier basis and the Haar wavelet basis, situations arise in which most entries of $U$ are small but few ones are very large. This leads to a large coherence $\mu$ and therefore large requirements on $m$ by the above results. To address this problem, different techniques have been developed to improve the previous results on bounded orthonormal systems for this situation. For example [KW13] adjusts the sampling probabilities of the rows of $U$

according to the largest entry in the corresponding row. In this way, rows of $U$ that contain large entries are sampled with a higher probability. Adcock et al. [Adc+17] propose another modified sampling model in which the rows are divided into a number $r$ of blocks and within each block, a certain number of entries are sampled, which is a similar strategy and called *multilevel random sampling*.

The approach from [Adc+17] in addition also considers the refined signal model of *sparsity in levels* as already motivated above. In this sparsity model, not the number of non-zero entries in an entire vector is bounded, e.g. by $s$, but the indices $[N]$ of $x$ are divided into $r$ blocks and in the $k$-th block the number of non-zero entries is bounded by $s_k$. This refined model also allows for entry sizes that vary between different columns as compared to different rows. Since rearranging the columns of $U$ does not change the property that $U$ is unitary, we can assume that each of the $r$ blocks consists of consecutive indices. This leads to the following definition of sparsity in levels. We mostly adapt the notation used in [Adc+17] and [LA19].

**Definition 4.1** (Sparsity in levels, Definition 3.3 in [Adc+17], Definition 2.6 in [LA19])**.** *Let $r \in \mathbb{Z}_{\geq 1}$, $\mathbf{M} = (M_1, \dots, M_r)$ with integers $1 \leq M_1 < \cdots < M_r = N$, $\mathbf{s} = (s_1, \dots, s_r)$, where $s_k \leq M_k - M_{k-1}$ for $k = 1, \dots, r$ and $M_0 = 0$.*

*Moreover, for each $k = 1, \dots, r$ we define the set*

$$\mathcal{M}_k := \{M_{k-1} + 1, \dots, M_k\} \subset [N].$$

*We say that $x \in \mathbb{C}^N$ is $(\mathbf{s}, \mathbf{M})$-sparse if for all $k = 1, \dots, r$,*

$$|\operatorname{supp}(x) \cap \mathcal{M}_k| \leq s_k.$$

*We denote $\Sigma_{\mathbf{s},\mathbf{M}} \subset \mathbb{C}^N$ for the set of all $(\mathbf{s}, \mathbf{M})$-sparse vectors.*

*We call $\mathbf{M}$ sparsity levels and $\mathbf{s}$ local sparsities.*

*Furthermore, for $x \in \Sigma_{\mathbf{s},\mathbf{M}}$, we define*

$$S_k := \operatorname{supp}(x) \cap \mathcal{M}_k$$

*for $k = 1, \dots, r$.*

So the block distribution of the sparsity in levels model yields a partition of the columns of $U \in \mathbb{C}^{N \times N}$ into $r$ blocks. Analogously, for the multilevel random subsampling we also partition the rows of $U$ into $r$ blocks. Then from each of these blocks, a certain number $m_k$ of samples are taken. Furthermore, for the first $r_0$ blocks (where $0 \leq r_0 \leq r$), we simply take all entries.

**Definition 4.2** (Multilevel random subsampling, Definition 3.2 in [Adc+17], Definition 2.7 in [LA19])**.** *Let $r \in \mathbb{Z}_{\geq 1}$, $0 \leq r_0 \leq r$, $\mathbf{N} = (N_1, \dots, N_r)$ with integers $1 \leq N_1 < \cdots < N_r = N$, $\mathbf{m} = (m_1, \dots, m_r)$, where $m_k \leq N_k - N_{k-1}$ for $k = 1, \dots, r$ and $N_0 = 0$.*

*Moreover, for each $k = 1, \dots, r$ we define the set*

$$\mathcal{N}_k := \{N_{k-1} + 1, \dots, N_k\} \subset [N].$$

*Assume that for each $k = 1, \dots, r_0$, $m_k = |\mathcal{N}_k|$ and $\Omega_k = \mathcal{N}_k$. For each $k = r_0 + 1, \dots, r$, let $t_{k,1}, \dots, t_{k,m_k}$ be random variables that are chosen uniformly and independently with replacement from the set $\mathcal{N}_k$ and $\Omega_k = \{t_{k,1}, \dots, t_{k,m_k}\}$ (as a multiset).*

*Then we call $\Omega = \Omega_{\mathbf{N},\mathbf{m}} = \Omega_1 \cup \cdots \cup \Omega_r$ an $(\mathbf{N}, \mathbf{m})$-multilevel subsampling scheme.*

Now that we have divided both, the rows and the columns of $U$, into blocks, we generalize the term of coherence of a unitary matrix to the *local coherence* by considering the largest entry in each of the resulting submatrices instead of the entire matrix.

**Definition 4.3** (Local coherence in levels, Definition 2.8 in [LA19])**.** *Let* $\mathbf{N} = (N_1, \dots, N_r)$ *be sampling and* $\mathbf{M} = (M_1, \dots, M_r)$ *be sparsity levels. The* $(l,k)$*-th local coherence of a unitary matrix* $U \in \mathbb{C}^{N \times N}$ *is*

$$\mu_{l,k} = \max\{|U_{j,j'}|^2 \,|\, j \in \mathcal{N}_l,\, j' \in \mathcal{M}_k\}.$$

Analogously to the classical RIP, [BH17] introduced a restricted isometry property for the sparsity in levels model from Definition 4.1.

**Definition 4.4** (RIP in levels (RIPL), Definition 3.5 in [BH17], Definition 2.12 in [LA19])**.** *Let* $\mathbf{M} = (M_1, \dots, M_r)$ *be sparsity levels and* $\mathbf{s} = (s_1, \dots, s_r)$ *be local sparsities. the* $\mathbf{s}$*-th restricted isometry constant in levels (RICL)* $\delta_{\mathbf{s},\mathbf{M}}$ *of a matrix* $A \in \mathbb{C}^{m \times N}$ *is the smallest* $\delta \geq 0$ *such that*

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2$$

*holds for all* $x \in \Sigma_{\mathbf{s},\mathbf{M}}$*.*

*If* $0 \leq \delta_{\mathbf{s},\mathbf{M}} < 1$*, we say that the matrix* $A$ *satisfies the restricted isometry property in levels (RIPL) of order* $(\mathbf{s}, \mathbf{M})$*.*

Like the subsampling matrix $P_\Omega$ for the usual subsampled bounded orthonormal systems, we define subsampling operations as matrices $P_{\Omega_k} \in \mathbb{R}^{m_k \times N}$ for a multilevel subsampling scheme as in Definition 4.2 like they are used in [LA19].

**Definition 4.5.** *Let* $\Omega = \Omega_{\mathbf{N},\mathbf{m}} = \Omega_1 \cup \dots \cup \Omega_r$ *be an* $(\mathbf{N}, \mathbf{m})$*-multilevel subsampling scheme with* $\Omega_k = \{t_{k,1}, \dots, t_{k,m_k}\}$ *for* $k = 1, \dots, r$*.*

*For each* $k \in [r]$*, we define* $P_{\Omega_k} \in \mathbb{R}^{m_k \times N}$ *such that*

$$(P_{\Omega_k})_j x = x_{t_{k,j}}$$

*for all* $j \in [m_k]$*,* $x \in \mathbb{C}^N$*, i.e.,* $P_{\Omega_k}$ *is a subsampling to the entries in* $\Omega_k$*.*

*We use the notation*

$$\tilde{\mathcal{N}_k} = \{\sum_{k'=1}^{k-1} m_{k'} + 1, \dots, \sum_{k'=1}^{k} m_{k'}\} \subset [m]$$

*for each* $k \in [r]$*, where* $m = m_1 + \dots + m_r$*.*

Using these definitions, [LA19] generalizes the subsampled bounded orthonormal systems with the following construction

$$A = \begin{pmatrix} \frac{1}{\sqrt{p_1}} P_{\Omega_1} U \\ \frac{1}{\sqrt{p_2}} P_{\Omega_2} U \\ \vdots \\ \frac{1}{\sqrt{p_r}} P_{\Omega_r} U \end{pmatrix} \in \mathbb{C}^{m \times N}, \tag{4.1}$$

where $p_l = \frac{m_l}{|\mathcal{N}_l|}$ for $l = 1, \dots, r$ and $m = m_1 + \dots + m_r$.

### 4.1.3 Fourier Sampling and Haar Wavelet Sparsity

As the main motivation of the RIPL result, [LA19] considers the above situation of two orthonormal bases for the Fourier basis and the discrete Haar wavelet basis. Important properties of this combination have already been studied in [AHR16].

Recall from Section 0.6 that in [AHR16], the Haar wavelet basis vectors are defined for $N = 2^r$ as $\phi_0 \in \mathbb{R}^N$ and $\phi_{j,p} \in \mathbb{R}^N$ for $j = 0, \dots, r-1$, $p = 0, \dots, 2^j - 1$ with entries

$$\phi_0(t) = 2^{-r}$$

$$\phi_{j,p}(t) = \begin{cases} 2^{\frac{j-r}{2}} & \text{for } 2^{r-j}p \leq t < 2^{r-j}(p+\frac{1}{2}) \\ -2^{\frac{j-r}{2}} & \text{for } 2^{r-j}(p+\frac{1}{2}) \leq t < 2^{r-j}(p+1) \\ 0 & \text{otherwise.} \end{cases}$$

for $1 \leq t \leq N$. We put these vectors into a matrix $\Phi \in \mathbb{R}^{N\times N}$ in such a way that for $M_0 = 0$ and $M_j = 2^j$ for $j = 1, \dots, r$, the columns 1 and 2 contain $\phi_0$ and $\phi_{0,0}$ and for each $j = 1, \dots, r-1$, the columns $M_j + 1, \dots, M_{j+1}$ contain the $\phi_{j,p}$ for all possible $p$. This defines the matrix $\Phi \in \mathbb{R}^{N\times N}$ and also the sparsity levels $\mathbf{M} = (M_1, \dots, M_r)$.

Section II in [AHR16] also defines a rearrangement $\tilde{F}$ of the rows of the DFT matrix $F$ in such a way, that there are good bounds on the local coherence of the matrix $U = \tilde{F}^*\Phi$ with the aforementioned $\Phi$. For this, consider the Fourier vectors $f_j \in \mathbb{C}^N$ for $j = -\frac{N}{2}+1, \dots, \frac{N}{2}$ ($N = 2^r$ is even) with entries

$$f_j(t) = \frac{1}{\sqrt{N}} e^{-\frac{2\pi i j t}{N}}$$

for $1 \leq t \leq N$. Then we define $N_0 = 0$, $N_k = 2^k$ for $k = 1, \dots, r$ and $\tilde{F}^{N\times N}$ in such a way that columns 1 and 2 of $\tilde{F}$ contain $f_0, f_1$ and for $k = 1, \dots, r-1$, columns $N_k + 1, \dots, N_{k+1}$ of $\tilde{F}$ contain the vectors $f_j$ for $j \in \{-2^k+1, \dots, -2^{k-1}\} \cup \{2^{k-1}+1, \dots, 2^k\}$. This defines $\tilde{F} \in \mathbb{C}^{N\times N}$ and the sampling levels $\mathbf{N} = (N_1, \dots, N_r)$.

The important result Lemma 1 of [AHR16] states that for the unitary matrix $U = \tilde{F}^*\Phi \in \mathbb{C}^{N\times N}$, the sparsity levels $\mathbf{M}$ and the sampling levels $\mathbf{N}$, the local coherences (Definition 4.3) satisfy

$$\mu_{l,k} \lesssim 2^{-l} \cdot 2^{-|l-k|} \tag{4.2}$$

for all $l, k \in [r]$.

## 4.2 Previous Work

The authors of [LA19] first observe (Footnote 1 in Section 3.2, consequence of Corollary 5.4 in [Dir16]) that an $m \times N$ Gaussian random matrix satisfies the RIPL (Definition 4.4) with probability $\geq 1 - \eta$ for

$$m \gtrsim \delta^{-2} \left( \sum_{k=1}^{r} s_k \log\left(\frac{e|\mathcal{M}_k|}{s_k}\right) + \log(\eta^{-1}) \right).$$

However, the main result of [LA19] is to show the RIPL for subsampled bounded orthonormal systems with multilevel sampling as in the matrix given in (4.1). This is done in the following theorem.

**Theorem 4.6** (Theorem 3.2 in [LA19])**.** *Let $U \in \mathbb{C}^{N\times N}$ be unitary, $r \in \mathbb{Z}_{\geq 1}$ and $0 < \eta, \delta < 1$ and $0 \leq r_0 \leq r$ an integer. Let $\Omega = \Omega_{\mathbf{N},\mathbf{m}}$ be an $(\mathbf{N}, \mathbf{m})$-multilevel subsampling scheme, $\mathbf{M}$ sparsity levels and $\mathbf{s}$ local sparsities. Suppose that blocks $l = 1, \dots, r_0$ are fully sampled, i.e., $m_l = |\mathcal{N}_l|$ and*

$$m_l \geq C\delta^{-2}|\mathcal{N}_l| \left( \sum_{k=1}^{r} \mu_{l,k} s_k \right) \left( r \log(2\tilde{m}) \log(2N)(\log(2s))^2 + \log\left(\frac{1}{\eta}\right) \right)$$

*for $l = r_0 + 1, \dots, r$, where $\tilde{m} = m_{r_0+1} + \dots + m_r$ and $C > 0$ is an absolute constant. Then with probability at least $1 - \eta$, the matrix (4.1) satisfies the RIPL of order $(\mathbf{s}, \mathbf{M})$ with constant $\delta_{\mathbf{s},\mathbf{M}} \leq \delta$.*

Finally, [LA19] also provides the following corollary for the application to Fourier sampling and Haar sparsity described in Section 4.1.3. This follows from applying the local coherence bound (4.2) to Theorem 4.6.

**Corollary 4.7** (RIPL for Fourier/Haar system, Corollary 3.4 in [LA19])**.** *Let $N = 2^r$ and $U \in \mathbb{C}^{N\times N}$ and $\mathbf{M}, \mathbf{N}$ the Fourier/Haar matrix with sparsity and sampling levels defined in Section 4.1.3. Assume that the first $0 \leq r_0 \leq r$ blocks are fully sampled, i.e., $m_l = |\mathcal{N}_l|$ for $l = 1, \dots, r_0$ and for the other blocks samples are drawn according to a corresponding multilevel sampling scheme with*

$$m_l \geq C\delta^{-2}\left(s_l + \sum_{\substack{k=r_0+1\\k\neq l}}^{r} 2^{-|l-k|}s_k\right) \cdot (\log(2\tilde{m})(\log(2N))^2(\log(2s))^2 + \log(\eta^{-1}))$$

*for $l = r_0+1, \dots, r$, $\tilde{m} = m_{r_0+1} + \dots + m_r$. If $N_{r_0} \leq s_{r_0+1}$ then with probability at least $1-\eta$, the matrix (4.1) satisfies the RIPL with constant $\delta_{\mathbf{s},\mathbf{M}} \leq \delta$ where $\mathbf{s} = (s_1, \dots, s_r)$ and $s_k = |\mathcal{N}_k|$ for $k = 1, \dots, r_0$.*

## 4.3 Main Result

The goal of our work is to show the following improved version of Theorem 4.6 that has fewer logarithmic factors in the required number of rows in terms of $s$.

**Theorem 4.8.** *Let $U \in \mathbb{C}^{N\times N}$ $(N \geq 2)$ be unitary, $r \in \mathbb{Z}_{\geq 1}$ and $0 < \eta < 1$, $0 < \delta < \frac{1}{2}$ and $0 \leq r_0 \leq r$ an integer. Let $\Omega = \Omega_{\mathbf{N},\mathbf{m}}$ be an $(\mathbf{N},\mathbf{m})$-multilevel subsampling scheme, $\mathbf{M}$ sparsity levels and $\mathbf{s}$ local sparsities with $s \geq 4$. Suppose that blocks $l = 1, \dots, r_0$ are fully sampled, i.e., $m_l = |\mathcal{N}_l|$ and*

$$m_l \geq C\delta^{-2}|\mathcal{N}_l|\left(\sum_{k=1}^{r}\mu_{l,k}s_k\right)\left(\log(N)\log(\tilde{\rho})\left(\log(s)\log(1/\delta)+r\right)+\log\left(\frac{1}{\eta}\right)\right) \tag{4.3}$$

*for $l = r_0+1, \dots, r$, $C > 0$ is an absolute constant, and*

$$\tilde{\rho} := \min\left\{4\tilde{m}, \frac{s^4}{\delta}\sum_{l=r_0+1}^{r}|\mathcal{N}_l|\sum_{k=1}^{r}\mu_{l,k}s_k\right\},$$

*where $\tilde{m} = m_{r_0+1} + \dots + m_r$. Then with probability at least $1-\eta$, the matrix (4.1) satisfies the RIPL of order $(\mathbf{s},\mathbf{M})$ with constant $\delta_{\mathbf{s},\mathbf{M}} \leq \delta$.*

So for fixed $\delta$, this is an improvement over Theorem 4.6 by one $\log s$ factor and in addition we have a $\log(s) + r$ factor instead of $\log(s)\cdot r$. The latter aspect especially yields an improvement if the number $r$ of layers scales logarithmically in $s$ or $N$. This is the case for example in the important application of the Fourier/Haar basis in Corollary [LA19]. We can improve this corollary with our Theorem 4.8 and obtain the following result, using the bounds (4.2).

**Corollary 4.9.** *Let $N = 2^r$ and $U \in \mathbb{C}^{N\times N}$ and $\mathbf{M}, \mathbf{N}$ the Fourier/Haar matrix with sparsity and sampling levels defined in Section 4.1.3. Assume that the first $0 \leq r_0 \leq r$ blocks are fully sampled, i.e., $m_l = |\mathcal{N}_l|$ for $l = 1, \dots, r_0$ and for the other blocks samples are drawn according to a corresponding multilevel sampling scheme with*

$$m_l \geq C\delta^{-2}\left(s_l + \sum_{\substack{k=r_0+1\\k\neq l}}^{r} 2^{-|l-k|}s_k\right) \cdot (\log(4\tilde{m})(\log(N))^2\log(1/\delta) + \log(\eta^{-1}))$$

*for $l = r_0 + 1, \ldots, r$, $\tilde{m} = m_{r_0+1} + \cdots + m_r$. If $N_{r_0} \leq s_{r_0+1}$ then with probability at least $1 - \eta$, the matrix* (4.1) *satisfies the RIPL with constant $\delta_{\mathbf{s},\mathbf{M}} \leq \delta$ where $\mathbf{s} = (s_1, \ldots, s_r)$ and $s_k = |\mathcal{N}_k|$ for $k = 1, \ldots, r_0$.*

Now for a fixed $\delta$, this result has 3 logarithmic factors while the original one, Corollary 4.7 has 5.

## 4.4 Proof of the Main Result

This section is devoted to the proof of Theorem 4.8. The proof is an adaption of the technique by Adcock and Li in 4.8 with significant changes that involve ideas from other proofs of the RIP for bounded orthonormal systems, especially [HR16].

Section 4.4.1 introduced important general tools required for the proof, Section 4.4.2 explains the ideas from previous works on BOSs that are used in the proof, Section 4.4.3 provides an outline of the proof steps and Section 4.5 contains the proof itself.

### 4.4.1 Required Tools

Most previous proofs of the RIP for subsampled BOSs rely on established techniques for controlling the suprema of stochastic processes. In general, a stochastic process $(X_t)_{t\in T}$ is an indexed family of random variables over the same probability space. An extensive theory has been developed to control expressions like

$$\sup_{t\in T} X_t$$

for a real-valued stochastic process. The textbook [Tal14] provides a profound overview of this theory, of which we introduce the most important required aspects hereafter.

The above supremum is related to the RIPL problem since we can write

$$\delta_{\mathbf{s},\mathbf{M}} = \sup_{x\in\Sigma_{\mathbf{s},\mathbf{M}}\cap S^{N-1}} \left| \|Ax\|_2^2 - 1 \right| \tag{4.4}$$

as the supremum of a stochastic process indexed by the set $\Sigma_{\mathbf{s},\mathbf{M}} \cap S^{N-1} =: D_{\mathbf{s},\mathbf{M}}$.

One particular object of interest is the expectation of such a supremum $\mathbb{E}\sup_{t\in T} X_t$. Here we encounter the first problem that for a general (uncountable) index set $T$, the supremum might not even be a random variable, i.e., it might not be measurable. Therefore, the definition of this expectation is modified to

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

In the case that $T$ is countable however, $\sup_{t\in T} X_t$ is always measurable and the above definition is consistent with the expectation of it (see Section 8.6 in [FR13]). For the process (4.4), the supremum is equal to the supremum over a countable dense subset of $\Sigma_{\mathbf{s},\mathbf{M}} \cap S^{N-1}$ such that also here, the supremum $\delta_{\mathbf{s},\mathbf{M}}$ is actually a random variable, such that we can not only consider its expectation but also define events involving it.

A particular class of processes that have been studied in this theory are the *subgaussian* processes. They are defined for an index set $T$ that defines a metric space $(T, d)$ and they require the condition (1.4) in [Tal14],

$$\forall u > 0: \quad \mathbb{P}(|X_s - X_t| \geq u) \leq 2\exp\left(-\frac{u^2}{2d(s,t)^2}\right). \tag{4.5}$$

Examples of processes that satisfy this requirement are given for a set $T \subset \mathbb{R}^N$ by the Gaussian process

$$X_t = \langle g, t\rangle \tag{4.6}$$

where $g \sim N(0, Id_N)$ follows a multivariate normal distribution and for the Bernoulli process

$$X_t = \langle \xi, t \rangle \tag{4.7}$$

where $\xi \in \{\pm 1\}^N$ is a Rademacher vector, i.e. its entries are independent with value $\pm 1$ with probability $\frac{1}{2}$ each. Both these processes satisfy the above condition with the metric $d(s,t) = \|s-t\|_2$. Even though (4.4) is not subgaussian, we will relate it to a Bernoulli process that satisfies the subgaussian condition.

A central part of the theory about stochastic processes discussed in [Tal14] is the technique of *generic chaining*, which is applicable to subgaussian processes. Its key idea is that we approximate the supremum of $(X_t)$ in the entire index set $T$ by the maximum on finite subsets $(T_n)_{n\geq 0}$ of $T$ with increasing size. For each $n$, the probability that the values of $(X_t)$ on the points in $T_n$ and the points in $T_{n+1}$ deviate too much, is controlled using (4.4) and a union bound.

More precisely, we define a sequence $(T_n)_{n\geq 0}$ of subsets of $T$ to be an *admissible sequence* if

- $|T_0| = 1$
- $|T_n| \leq 2^{2^n}$

and define the $\gamma_2$ functional

$$\gamma_2(T,d) := \inf_{(T_n)_{n\geq 0}} \sup_{t\in T} \sum_{n=0}^{\infty} 2^{\frac{n}{2}} d(t, T_n) \tag{4.8}$$

where the $\inf_{(T_n)_{n\geq 0}}$ is taken over all admissible sequence $(T_n)_{n\geq 0}$.

Then the main result about generic chaining ((2.32) in [Tal14]) is that there is a universal constant $C > 0$ such that for all subgaussian processes that are centered, i.e., $\mathbb{E}X_t = 0$ for all $t \in T$,

$$\mathbb{E} \sup_{t\in T} |X_t| \leq C\gamma_2(T,d). \tag{4.9}$$

It has been shown that the generic chaining bound (4.9) is sharp up to constant factors for Gaussian processes as in (4.6). This is know as the *majorizing measure theorem* (Theorem 2.4.1 in [Tal14]). However, for Bernoulli processes as in (4.7) it is generally not sharp. As a particular example, for $T = \{x \in \mathbb{R}^N \,|\, \|x\|_1 = 1\}$, one can show that $\mathbb{E}\sup_{x\in T} |\langle g, x\rangle| \sim \log(N)$ is significantly larger than $\mathbb{E}\sup_{x\in T} |\langle \xi, x\rangle| = 1$. In contrast to Gaussian processes, for Bernoulli processes also the bound

$$\sup_{t\in T} |\langle \xi, t\rangle| \leq \sup_{t\in T} \|\xi\|_\infty \|t\|_1 = \sup_{t\in T} \|t\|_1 \tag{4.10}$$

holds, which follows from Hölder's inequality.

We can combine the generic chaining bound (4.9) and the Bernoulli bound (4.10) in the sense that we can split up $T \subset T^{(1)} + T^{(2)}$ for two subsets $T^{(1)}, T^{(2)} \subset \mathbb{R}^N$ and then apply (4.9) on $T^{(1)}$ and (4.10) on $T^{(2)}$ to obtain

$$\sup_{t\in T} |\langle \xi, t\rangle| \leq C\gamma_2(T^{(1)}, d) + \sup_{t\in T^{(2)}} \|t\|_1, \tag{4.11}$$

for a constant $C > 0$ as described in Proposition 5.1.4 in [Tal14]. The sharpness of this bound up to constant factors has been known as the *Bernoulli conjecture*, which was proven to be true in [BL14].

Even though (4.9) provides a bound that is provably sharp in the Gaussian case, it is not clear how an optimal admissible sequence $(T_n)_{n\geq 0}$ in (4.8) can be found. One possible way is choosing them as a covering such that $\sup_{t\in T} d(t, T_n)$ is as small as possible. With this method, the $\gamma_2$ functional can be bounded by an integral over covering numbers as in (2.38) in [Tal14],

$$\gamma_2(T, d) \lesssim \int_0^\infty \sqrt{\log \mathcal{N}(T, d, u)}\mathrm{d}u. \tag{4.12}$$

However, this bound is possibly not sharp even for Gaussian processes, see Exercise 2.2.15 in [Tal14].

Finally, we establish the following short lemma that will be used in our proof to connect the sum over $2^{\frac{n}{2}} d(t, T_n)$ in the $\gamma_2$ functional and a corresponding sum over $2^{\frac{n}{2}} d(t_n, t_{n-1})$, where the $t_{n'} \in T_{n'}$ usually approximate $t$.

**Lemma 4.10.** *Let $(X, d)$ be a metric space. For each $n \in \{n_1, n_1 + 2, \ldots, n_2\}$, let $T_n \subset T$ and $t_n \in T_n$. Then for $t \in T$,*

$$\sum_{n=n_1}^{n_2} 2^{\frac{n}{2}} d(t, T_n) \le 4 \left( 2^{\frac{n_2}{2}} d(t, t_{n_2}) + \sum_{n=n_1+1}^{n_2} 2^{\frac{n}{2}} d(t_n, t_{n-1}) \right).$$

*Proof.* First we obtain for each $n \in \{n_1, n_1 + 2, \ldots, n_2\}$ by the triangle inequality,

$$d(t, T_n) \le d(t, t_n) \le d(t, t_{n_2}) + \sum_{n'=n+1}^{n_2} d(t_{n'}, t_{n'-1}).$$

Then substituting this on the left hand side leads to

$$\begin{aligned}\sum_{n=n_1}^{n_2} 2^{\frac{n}{2}} d(t, T_n) &\le \sum_{n=n_1}^{n_2} 2^{\frac{n}{2}} d(t, t_{n_2}) + \sum_{n=n_1}^{n_2} 2^{\frac{n}{2}} \sum_{n'=n+1}^{n_2} d(t_{n'}, t_{n'-1}) \\ &\le 4 \cdot 2^{\frac{n_2}{2}} d(t, t_{n_2}) + \sum_{n'=n_1}^{n_2} \sum_{n=n_1}^{n'-1} 2^{\frac{n}{2}} d(t_{n'}, t_{n'-1}) \\ &\le 4 \cdot 2^{\frac{n_2}{2}} d(t, t_{n_2}) + 4 \sum_{n'=n_1}^{n_2} 2^{\frac{n'}{2}} d(t_{n'}, t_{n'-1}),\end{aligned}$$

which shows the claim while we used that

$$\sum_{n=n_1}^{n_2} 2^{\frac{n}{2}} \le \sum_{n=0}^{n_2} (\sqrt{2})^n = \frac{(\sqrt{2})^{n_2+1} - 1}{\sqrt{2} - 1} \le \frac{\sqrt{2}}{\sqrt{2} - 1} 2^{\frac{n_2}{2}} \le 4 \cdot 2^{\frac{n_2}{2}}.$$

□

In some cases, instead of constructing an admissible sequence for $T$ directly, it can be easier to construct a sequence $(T_n)_{n\ge 0}$ of sets that satisfy the size condition but are not subsets of $T$ but some bigger metric space that contains $T$. The following lemma states that we can still bound the $\gamma_2$ functional using such a sequence, only losing a factor of 2, even if $T$ is not necessarily closed.

**Lemma 4.11.** *Let $(\bar{T}, d)$ be a metric space and $\emptyset \ne T \subset \bar{T}$. Assume that $(T_n)_{n\ge 0}$ is an admissible sequence in $\bar{T}$, i.e., all $T_n \subset \bar{T}$, $|T_0| = 1$ and for all $n \ge 1$, $|T_n| \le 2^{2^n}$ (but not necessarily $T_n \subset T$). Then*

$$\gamma_2(T, d) \le 2 \sup_{t\in T} \sum_{n=0}^{\infty} 2^{\frac{n}{2}} d(t, T_n).$$

*Proof.* We construct an admissible sequence $(T'_n)_{n\geq 0}$ of subsets of $T$ from $(T_n)_{n\geq 0}$. For this, fix $\epsilon > 0$. For each $n \geq 0$ and each $t_0 \in T_n$, take one $t'_0 \in T$ such that $d(t_0, t'_0) \leq d(t_0, T) + 2^{-\frac{3}{2}n}\epsilon$. Define $T'_n \subset T$ to be the set of all $t'_0$ obtained in this way. Then $|T'_n| \leq |T_n|$ and for each $t \in T$, there exists a $t_0 \in T_n$ such that $d(t, t_0)$ is minimal, i.e., $d(t, t_0) = d(t, T_n)$ and with the corresponding $t'_0$ defined above

$$\begin{aligned} d(t, T'_n) &\leq d(t, t'_0) \leq d(t, t_0) + d(t_0, t'_0) \leq d(t, T_n) + d(t_0, T) + 2^{-\frac{3}{2}n}\epsilon \\ &\leq d(t, T_n) + d(t_0, t) + 2^{-\frac{3}{2}n}\epsilon \leq 2d(t, T_n) + 2^{-\frac{3}{2}n}\epsilon. \end{aligned}$$

So $(T'_n)_{n\geq 0}$ is an admissible sequence in $T$ and therefore

$$\begin{aligned} \gamma_2(T, d) &\leq \sup_{t\in T} \sum_{n=0}^{\infty} 2^{\frac{n}{2}} d(t, T'_n) \leq 2 \sup_{t\in T} \sum_{n=0}^{\infty} 2^{\frac{n}{2}} d(t, T_n) + \sum_{n=0}^{\infty} 2^{\frac{n}{2}} \cdot 2^{-\frac{3}{2}n}\epsilon \\ &\leq 2 \sup_{t\in T} \sum_{n=0}^{\infty} 2^{\frac{n}{2}} d(t, T_n) + 2\epsilon. \end{aligned}$$

Then the claim follows from the fact that this holds for all $\epsilon > 0$. □

Like it has been done in [RV08] and [LA19] before, in order to prove the RIPL, we will first control $\mathbb{E}\delta_{\mathbf{s},\mathbf{M}}$ and subsequently bound the probability that the RICL deviates too much from it expectation. For the latter part, we will use the following concentration result given as Theorem 8.42 in [FR13].

**Theorem 4.12.** *Let $\mathcal{F}$ be a countable set of functions $F : \mathbb{C}^n \to \mathbb{R}$. Let $Y_1, \ldots, Y_M$ be independent random vectors in $\mathbb{C}^n$ such that $\mathbb{E}F(Y_l) = 0$ and $F(Y_l) \leq K$ almost surely for all $l \in [M]$ and for all $F \in \mathcal{F}$ for some constant $K > 0$. Introduce*

$$Z = \sup_{F\in\mathcal{F}} \sum_{l=1}^{M} F(Y_l).$$

*Let $\sigma_l^2 > 0$ such that $\mathbb{E}[F(Y_l)^2] \leq \sigma_l^2$ for all $F \in \mathcal{F}$ and $l \in [M]$. Then, for all $t > 0$,*

$$\mathbb{P}(Z \geq \mathbb{E}Z + t) \leq \exp\left(-\frac{t^2/2}{\sigma^2 + 2K\mathbb{E}Z + tK/3}\right)$$

*where $\sigma^2 = \sum_{l=1}^{M} \sigma_l^2$.*

### 4.4.2 Previous Proofs

In this section, we review the main ideas of previous proofs for the RIP of bounded orthonormal systems since our proof in Section 4.5 combines multiples ones of them.

*Rudelson and Vershynin* [RV08] provide a proof for the RIP of subsampled BOSs that is also contained as Theorem 12.31 in the textbook [FR13] in a simplified form. Their main task is bounding the expectation $\mathbb{E}\delta_s$ of the restricted isometry constant and then the concentration can be concluded with Theorem 4.12.

As the first step, with a suitable application of the symmetrization Lemma 0.9, they can show that

$$\mathbb{E}\delta_s = \mathbb{E} \sup_{x\in\Sigma_s\cap S^{N-1}} \left|\|Ax\|_2^2 - 1\right| \leq C\mathbb{E} \sup_{y\in T} |\langle \xi, y\rangle|, \tag{4.13}$$

with a Rademacher vector $\xi \in \{\pm 1\}^m$ that is independent of $A$ and the index set $T := \{|Ax|^2 \,|\, x \in \Sigma_s \cap S^{N-1}\}$ that consists of the measurements $Ax$ with entry-wise squared absolute value. So the index set $T$ as well as $\xi$ are random and they are independent. So we can condition on $T$ and first control the expectation with respect to $\xi$. Then the problem boils down to controlling the expected supremum of a Bernoulli process of the type (4.7), for which the techniques introduced in Section 4.4.1 are available. As mentioned there, this process is subgaussian with the metric

$$d(|Ax|^2, |Az|^2) = \left(\sum_{j=1}^{m}(|(Ax)_j|^2 - |(Az)_j|^2)^2\right)^{\frac{1}{2}},$$

which for $x, z \in \Sigma_s \cap S^{N-1}$ is bounded in [RV08] by

$$\begin{aligned} d(|Ax|^2, |Az|^2) &= \left(\sum_{j=1}^{m}(|(Ax)_j| - |(Az)_j|)^2(|(Ax)_j| + |(Az)_j|)^2\right)^{\frac{1}{2}} \\ &\le \left(\sum_{j=1}^{m} 2|(A(x-z))_j|^2(|(Ax)_j|^2 + |(Az)_j|^2)\right)^{\frac{1}{2}} \\ &\le \left(2\max_{j\in[m]}|(A(x-z))_j|^2 \cdot \sum_{j=1}^{m} 2(|(Ax)_j|^2 + |(Az)_j|^2)\right)^{\frac{1}{2}} \\ &= 2\|A(x-z)\|_\infty \cdot \sqrt{\|Ax\|_2^2 + \|Ax\|_2^2} \\ &\le 2\sqrt{2}\|A(x-z)\|_\infty \cdot \sqrt{1+\delta_s}. \end{aligned} \tag{4.14}$$

Then [RV08] bounds the process (4.13) using the generic chaining (4.9) in the form of Dudley's inequality (4.12). This requires a bound on the covering numbers $\mathcal{N}(T, d, u)$. Bound (4.14) shows that with an additional factor $\sqrt{1+\delta_s}$ in the result, it is enough to bound the covering numbers $\mathcal{N}(\Sigma_s \cap S^{N-1}, \tilde{d}, u)$ for the modified metric

$$\tilde{d}(x, z) = \|A(x-z)\|_\infty$$

on $\Sigma_s \cap S^{N-1}$. Their strategy is to bound these covering numbers in two different ways. There are $\binom{N}{s}$ possible choices of a support of size $s$ in $[N]$. For small distances $u$, the covering of $\Sigma_s \cap S^{N-1}$ is obtained as a union of separate coverings for each possible support.

For larger distances $u$, a more involved approach known as *Maurey's empirical method* is applied. This method has been used before in [Car85] and pursues the following strategy. Each real-valued $s$-sparse vector $x \in S^{N-1}$ has $\|x\|_1 \le \sqrt{s}$ and therefore, we can define a random variable $Z \in \{0, \pm\sqrt{s}e_1, \dots, \pm\sqrt{s}e_N\}$ where $e_1, \dots, e_N$ are the canonical basis vectors, such that $Z = \sqrt{s}\,\mathrm{sign}(x_j)e_j$ with probability $\frac{|x_j|}{\sqrt{s}}$ for each $j \in [N]$ and $Z = 0$ with probability $1 - \frac{\|x\|_1}{\sqrt{s}}$. This is a valid probability distribution with $\mathbb{E}Z = x$. Now we define independent copies of $Z_1, \dots, Z_M$ of $Z$ and form $\bar{Z} := \frac{1}{M}\sum_{k=1}^{M} Z_k$. Now for given $u$, one can use standard concentration inequalities to show that $\tilde{d}(\bar{Z}, x) \le u$ with positive probability if $M$ is sufficiently large. Therefore, the set of all possible values of $\bar{Z}$ is a covering of $(\Sigma_s \cap S^{N-1}, \tilde{d})$ with distance $u$. On the other hand, by the definition of $\bar{Z}$, its total number of possible different values is $\le (2N+1)^M$. In this way, we can conclude that $\mathcal{N}(\Sigma_s \cap S^{N-1}, \tilde{d}, u) \le (2N+1)^M$ where $M$ depends on $u$.

With all of this, [RV08] shows a bound

$$\mathbb{E}\left[\sup_{y\in T}|\langle \xi, y\rangle| \,\middle|\, A\right] \le \alpha\sqrt{1+\delta_s}$$

and therefore by the law of total probability and Jensen's inequality

$$\mathbb{E}\delta_s \lesssim \mathbb{E}\left[\mathbb{E}\left[\sup_{y\in T}|\langle \xi, y\rangle| \,\middle|\, A\right]\right] \le \alpha\mathbb{E}\left[\sqrt{1+\delta_s}\right] \le \alpha\sqrt{1+\mathbb{E}\delta_s}$$

for an $\alpha$ that depends on $m$, $N$, $s$. This is an inequality that can be solved for $\mathbb{E}\delta_s$. Finally, Theorem 4.12 is applied to control the deviation of $\delta_s$ from its expectation.

The proof by *Haviv and Regev* [HR16] essentially improves the aforementioned result by Rudelson and Vershynin by one logarithmic factor in $s$. They follow an approach that is similar to generic chaining but not precisely written in the terminology introduced in Section 4.4.1. However, as shown in the work of *Brugiapaglia et al.* [Bru+21], their approach can be cast into the framework of symmetrization and generic chaining established in [RV08]. In this way, the proof bounds the supremum of the same Bernoulli process. However, some aspects are performed differently. Firstly, in contrast to [RV08], which only uses the generic chaining bound (4.9), the approach [HR16; Bru+21] uses the combined bound for Bernoulli processes (4.11), which is known to be stronger in certain cases. Secondly, they do not use Dudley's inequality but instead construct an admissible sequence for the generic chaining bound (4.9), (4.8) directly whose elements can approximate each $|Ax|^2$. These approximations are also obtained using Maurey's empirical method described above. However, in the construction of $T_{n_1+n}$ for some $n_1$, Maurey's empirical method is only used to approximate those entries of $|Ax|^2$ of approximate size $\ge 2^{-n}\frac{s}{m}$ (all entries of $|Ax|^2$ are $\le \frac{s}{m}$) and the remaining ones are just approximated by 0. Then also (4.14) is replaced by an improved bound that takes the sizes of the individual entries into account. After a certain number of steps that are controlled using the generic chaining method, this produces an approximation $w \in \mathbb{R}^m$ of $|Ax|^2$ such that $\left||(Ax)_j|^2 - w_j\right| \le \delta|(Ax)_j|^2$ holds for all $j$ except for some of them that are too few or have a too small $|(Ax)_j|^2$ to influence the result significantly. So in the end for the Bernoulli bound (4.11), we obtain an $\ell_1$ deviation

$$\left\||Ax|^2 - w\right\|_1 \le \sum_{j=1}^{m}\delta|(Ax)_j|^2 = \delta\|Ax\|_2^2 \le \delta\sqrt{1+\delta_s}.$$

In the end, all the possible values of $w$ are defined to form the $T^{(1)}$ in (4.11) and all possible values of $|Ax|^2 - w$ form $T^{(2)}$. Then (4.11) is used to bound the expected supremum of the Bernoulli process. The remaining part, including controlling the additional factor $\sqrt{1+\delta_s}$, solving for $\mathbb{E}\delta_s$ and the concentration are done in a similar way as in the proof of Rudelson/Vershynin [RV08] described above.

The work by *Adcock and Li* [LA19] shows the RIPL instead of the regular RIP. Their proof is an adaption of [RV08] to sparsity in levels and multilevel random sampling. Therefore, the overall procedure is similar. They also use Dudley's inequality and bound covering numbers for each support separately for small $u$ and use Maurey's empirical method for larger $u$. The most important difference is that in their application of Maurey's empirical method, they approximate each sparsity level of $x$ separately, so for each of the $r$ sparsity levels, there is a certain number $M$ of independent copies of the aforementioned random variable $Z$ such that the total number of possible values of all of them is $(2N+1)^{rM}$. This is where the additional factor $r$ in the final result of Theorem 4.6 arises.

### 4.4.3 Proof Outline

Our proof combines ideas from different approaches described in Section 4.4.2 along with some novel improvements. The procedure can be split into the following steps.

1. **Symmetrization and Bernoulli process**
   This first step is the same as in [LA19; RV08]. With the symmetrization technique (Lemma 0.9), the expectation is bounded by the expected supremum of a Bernoulli process over the index set $T := \{|Ax|^2 \,|\, x \in D_{\mathbf{s},\mathbf{M}}\}$ for $D_{\mathbf{s},\mathbf{M}} := \Sigma_{\mathbf{s},\mathbf{M}} \cap S^{N-1}$.

2. **Approximations with Maurey's empirical method**
   Like the other previously mentioned proofs, we use Maurey's empirical method to approximate $|Ax|^2$ with different accuracy levels. Like in [HR16; Bru+21] (and other than [RV08; LA19]), we do not directly use the resulting approximations for the generic chaining yet but apply some modifications in step 3 beforehand. For the approximation $\bar{Z}_n$, we show that $\left|(A\bar{Z}_n)_j - (Ax)_j\right|$ is sufficiently small enough for all $j \in [m]$ except a limited number. The set of entries for which the deviation is too large in any of the approximation levels in the generic chaining will be called $J \subset [m]$. We show that $|J|$ is small and accordingly bound the influence of these indices in the remaining parts of the proof.

   Moreover, we also approximate each sparsity level $k \in [r]$ of $x$ separately. However, unlike in [LA19], we do not choose the number of independent copies for Maurey's empirical method equal for all sparsity level but instead let it scale with the $\ell_2$ norm $\|x_{S_k}\|_2$ of $x$ in the corresponding sparsity block. This improves the dependence on $r$ in the final result (such that we have a factor $r + \log s$ instead of $r \log s$).

3. **Construction of the admissible sequence for $n \leq n_1 + \lceil \log_2(s) \rceil$**
   We will define $n_1 := n_0 + \lceil \log_2 \log_2 N + \log_2 \log \rho \rceil$ for a constant integer $n_0 \geq 0$ to be specified later. First, for $0 \leq n \leq n_1 + \lceil \log_2 r \rceil$, we define $T_n = \{0\}$.

   For $n_1 + \lceil \log_2 r \rceil \leq n \leq n_1 + \lceil \log_2 s \rceil$, we construct vectors $w^{(n-n_1)}(x) \in \mathbb{R}^m$, depending on $x \in D_{\mathbf{s},\mathbf{M}}$ that approximate $|Ax|^2$ on each level of the generic chaining. At each level $n$, the entries of a certain size are approximated similarly to the proof in [HR16; Bru+21] described in the previous section. The other entries of $w^{(n-n_1)}(x)$ will be set to 0. Larger entries of $|Ax|^2$ are approximated for smaller $n$ than smaller entries. The set of all these $w^{(n-n_1)}(x)$ across all $x \in D_{\mathbf{s},\mathbf{M}}$ will be $T_n$ and satisfy $|T_n| \leq 2^{2^n}$.

4. **Construction of $T^{(1)}$ and $T^{(2)}$ for (4.11) and $\ell_1$ bound**
   The sets $T_0, \ldots, T_{n_1+\lceil \log_2(s) \rceil}$ from the previous step are used to approximate $|Ax|^2 \in \mathbb{R}^m$. We construct $T^{(1)}$ in such a way that for each such $|Ax|^2$, there is a $z \in T^{(1)}$ based on the following idea. If entry $j$ of $|Ax|^2$ is approximated sufficiently well in the last level defined above, i.e., $w^{(\lceil \log_2(s) \rceil)} \in T_{n_1+\lceil \log_2(s) \rceil}$, then $z_j = w_j^{(\lceil \log_2(s) \rceil)}$ and otherwise $z_j = |(Ax)_j|^2$. In the first case, in the subsequent approximations defined in step 6, the entry $j$ will remain constant as $w_j^{(\lceil \log_2(s) \rceil)}$ and the deviation $|z_j - |(Ax)_j||$ will be controlled with the $\ell_1$ part in (4.11). In the other case, $|(Ax)_j|^2$ will be approximated with increasing precision in the remaining part of the admissible sequence such that the entry $j$ is controlled in the generic chaining part of (4.11). Then $T^{(2)}$ contains the differences $z - |Ax|^2$ and can be bounded in $\ell_1$ norm.

   This particular strategy is new and differs from the approaches [HR16; Bru+21]. In their proof, the generic chaining method described in step 3 is continued until all entries are approximated sufficiently well for the remaining distance to be bounded in the $\ell_1$ norm. In this way, $T^{(1)}$ is the last element of the partial admissible sequence and thus finite. However, this requires $\log(L\frac{s}{\delta})$ steps (for the classical RIP case $r = 1$, with $L$ being the constant such that all entries of U are bounded by $\frac{L}{\sqrt{N}}$). Our approach only requires $\log s$ steps.

   Therefore, our approach yields the slight advantage of having a $\log s$ factor in the final result instead of $\log(L\frac{s}{\delta})$.

5. **Application of Lemma 4.10 to the first part of the admissible sequence**
   For the first part of the admissible sequence, we controlled an expression of the type $\sum_n 2^{\frac{n}{2}} d(t_n, t_{n-1})$. Using Lemma 4.10, we turn this into a bound on $\sum_n 2^{\frac{n}{2}} d(t, T_n)$, which will later be combined with the corresponding sum for the remaining admissible sequence to bound $\gamma_2(T^{(1)}, d)$.

6. **Construction of the admissible sequence for $n > n_1 + \lceil \log_2(s) \rceil$**
   We complete the admissible sequence $(T_n)_n$ for the remaining larger $n$ by approximating the vectors $x \in D_{\mathbf{s},\mathbf{M}}$ on each single support separately. This has been done in [RV08] and [LA19]. Note that although [HR16; Bru+21] do not do an equivalent step, this part is required here because of the difference discussed in step 4.

7. **Combination of the bounds to control $\mathbb{E}\delta_{\mathbf{s},\mathbf{M}}$**
   We use the combined generic chaining and $\ell_1$ bound (4.11) to bound $\mathbb{E}\delta_{\mathbf{s},\mathbf{M}}$ in way that is similar to [Bru+21].

8. **Concentration of $\delta_{\mathbf{s},\mathbf{M}}$**
   We use Theorem 4.12 to control the deviation of $\delta_{\mathbf{s},\mathbf{M}}$ from its expectation analogously to [RV08; LA19].

## 4.5 Proof of Theorem 4.8

*Proof.* We quickly recall the notation of the matrix and Definitions 4.1 to 4.3.

- The matrix is

$$A = \begin{pmatrix} \frac{1}{\sqrt{p_1}} P_{\Omega_1} U \\ \frac{1}{\sqrt{p_2}} P_{\Omega_2} U \\ \vdots \\ \frac{1}{\sqrt{p_r}} P_{\Omega_r} U \end{pmatrix} \in \mathbb{C}^{m \times N}.$$

- $\mathcal{N}_1, \dots, \mathcal{N}_r \subset [N]$ partition of rows of $U$ for multilevel sampling, $\mathcal{N}_l = \{N_{l-1}+1, \dots, N_l\}$.
- $\tilde{\mathcal{N}_1}, \dots, \tilde{\mathcal{N}_r} \subset [m]$ partition of the rows of $A$ into samples from different blocks. ($|\tilde{\mathcal{N}_l}| = m_l$)
- $\mathcal{M}_1, \dots, \mathcal{M}_r \subset [N]$ blocks for sparsity
- $S_1, \dots, S_r \subset [N]$ supports of $x_{\mathcal{M}_1}, \dots, x_{\mathcal{M}_r}$, $|S_k| \le s_k$ $(1 \le k \le r)$.
- $m_l = |\mathcal{N}_l|$ for $l = 1, \dots, r_0$

**Step 1: Symmetrization and Bernoulli process**
This first step follows the same argument as the proof in [LA19]. First we define

$$D_{\mathbf{s},\mathbf{M}} := \Sigma_{\mathbf{s},\mathbf{M}} \cap S^{N-1},$$

such that

$$\delta_{\mathbf{s},\mathbf{M}} = \sup_{x \in D_{\mathbf{s},\mathbf{M}}} \left| \|Ax\|_2^2 - 1 \right| = \sup_{x \in D_{\mathbf{s},\mathbf{M}}} |\langle x, (A^*A - Id_N)x \rangle| =: |||A^*A - Id_N|||_{\mathbf{s},\mathbf{M}},$$

which is how we define the norm $|||\cdot|||_{\mathbf{s},\mathbf{M}}$ on $\mathbb{C}^{N \times N}$.

For each $j \in [m]$, we define $A_j \in \mathbb{C}^N$ as the adjoint vector of the $j$-th row of $A$ (such that $A_j^* x = \langle A_j, x \rangle = (Ax)_j$ for $x \in \mathbb{C}$, $j \in [m]$). Define $U_j$ $(j \in [N])$ analogously for the matrix $U$. Then for each $1 \le l \le r_0$ (i.e., the blocks of complete sampling),

$$\sum_{j \in \tilde{\mathcal{N}_l}} \mathbb{E} A_j A_j^* = \frac{1}{p_l} \sum_{j \in \tilde{\mathcal{N}_l}} U_j U_j^* = \sum_{j \in \mathcal{N}_l} U_j U_j^*.$$

For each of the other blocks $r_0 + 1 \leq l \leq r$,

$$\sum_{j\in\tilde{\mathcal{N}}_l} \mathbb{E}A_jA_j^* = \sum_{j\in\tilde{\mathcal{N}}_l} \frac{1}{|\mathcal{N}_l|} \sum_{j'\in\mathcal{N}_l} \frac{1}{p_l} U_{j'}U_{j'}^* = \frac{m_l}{|\mathcal{N}_l|p_l} \sum_{j'\in\mathcal{N}_l} U_{j'}U_{j'}^* = \sum_{j\in\mathcal{N}_l} U_jU_j^*.$$

So altogether

$$\sum_{j\in[m]} \mathbb{E}A_jA_j^* = \sum_{j\in[N]} U_jU_j^* = U^*U = Id_N.$$

Then

$$\begin{aligned}\mathbb{E}\delta_{\mathbf{s},\mathbf{M}} = \left|\!\left|\!\left| A^*A - Id_N \right|\!\right|\!\right|_{\mathbf{s},\mathbf{M}} &= \left|\!\left|\!\left| \sum_{j\in[m]} A_jA_j^* - \sum_{j\in[m]} \mathbb{E}A_jA_j^* \right|\!\right|\!\right|_{\mathbf{s},\mathbf{M}} \\ = \left|\!\left|\!\left| \sum_{l=1}^{r} \sum_{j\in\tilde{\mathcal{N}}_l} (A_jA_j^* - \mathbb{E}A_jA_j^*) \right|\!\right|\!\right|_{\mathbf{s},\mathbf{M}} &= \left|\!\left|\!\left| \sum_{l=r_0+1}^{r} \sum_{j\in\tilde{\mathcal{N}}_l} (A_jA_j^* - \mathbb{E}A_jA_j^*) \right|\!\right|\!\right|_{\mathbf{s},\mathbf{M}}.\end{aligned}$$

Using the symmetrization Lemma 0.9, we obtain

$$\mathbb{E}\delta_{\mathbf{s},\mathbf{M}} \leq 2\mathbb{E} \left|\!\left|\!\left| \sum_{l=r_0+1}^{r} \sum_{j\in\tilde{\mathcal{N}}_l} \xi_j A_jA_j^* \right|\!\right|\!\right|_{\mathbf{s},\mathbf{M}} \tag{4.15}$$

where $\xi_j$ for all $j \in \tilde{\mathcal{N}}_l$, $l \in \{r_0 + 1, \ldots, r\}$ are independent Rademacher variables that are independent of $A$.

Now we condition on the $A_j$, and first consider the expectation $\mathbb{E}_\xi$ with respect to the $\xi_j$. By definition of $|\!|\!|\cdot|\!|\!|_{\mathbf{s},\mathbf{M}}$,

$$\mathbb{E}_\xi \left|\!\left|\!\left| \sum_{l=r_0+1}^{r} \sum_{j\in\tilde{\mathcal{N}}_l} \xi_j A_jA_j^* \right|\!\right|\!\right|_{\mathbf{s},\mathbf{M}} = \mathbb{E}_\xi \sup_{x\in D_{\mathbf{s},\mathbf{M}}} \left| \sum_{l=r_0+1}^{r} \sum_{j\in\tilde{\mathcal{N}}_l} \xi_j |(Ax)_j|^2 \right|.$$

This is the expected supremum of the Bernoulli process defined by

$$X_y = \sum_{l=r_0+1}^{r} \sum_{j\in\tilde{\mathcal{N}}_l} \xi_j y_j \qquad\qquad y \in T$$

for the index set

$$T := \{|Ax|^2 \,|\, x \in D_{\mathbf{s},\mathbf{M}}\},$$

where $|Ax|^2$ consists of all squared absolute values of the entries of $Ax$. As a Bernoulli process it is subgaussian with the Euclidean metric with respect to the entries indexed by

$$\tilde{\mathcal{N}} := \bigcup_{l=r_0+1}^{r} \tilde{\mathcal{N}}_l, \tag{4.16}$$

i.e., $d(y, \bar{y}) = \left(\sum_{j\in\tilde{\mathcal{N}}} (y_j - \bar{y}_j)^2\right)^{\frac{1}{2}}$ for $y, \bar{y} \in T$. Note however, that this might not be a metric on $T$ since two different vectors in $T$ might only differ in entries in $[m]\backslash\tilde{\mathcal{N}}$. But it is a metric on the projections on $\tilde{\mathcal{N}}$,

$$T' = \{y_{\tilde{\mathcal{N}}} \,|\, y \in T\} \subset \mathbb{R}^{|\tilde{\mathcal{N}}|}.$$

To simplify the presentation, we will still consider the index set $T$ and an admissible sequence on $T$ but implicitly project everything onto the entries $\tilde{\mathcal{N}}$ such that $d$ becomes an actual metric instead of a pseudometric.

Then for any $u, v \in D_{\mathbf{s},\mathbf{M}}$,

$$d(|Au|^2, |Av|^2) = \left( \sum_{l=r_0+1}^{r} \sum_{j \in \tilde{\mathcal{N}}_l} (|(Au)_j|^2 - |(Av)_j|^2)^2 \right)^{\frac{1}{2}}.$$

**Step 2: Approximations with Maurey's empirical method**

We need the following definitions for $l = 1, 2, \ldots, r$,

$$\gamma_l := \sum_{k=1}^{r} \frac{\mu_{l,k} s_k}{p_l} \qquad\qquad \gamma := \max_{l \in \{r_0+1,\ldots,r\}} \gamma_l$$

$$\rho := \min \left\{ 4\tilde{m}, 4\delta^{-1} rs(\lceil \log_2(s) \rceil + 2) \cdot \sum_{l=r_0+1}^{r} m_l \gamma_l \right\} \leq \min \left\{ 4\tilde{m}, \frac{s^4}{\delta} \sum_{l=r_0+1}^{r} m_l \gamma_l \right\} \tag{4.17}$$

where $\tilde{m} = m_{r_0+1}, \ldots, m_r$.

Choose $\alpha_1, \ldots, \alpha_r \in \{0, 1, 2, \ldots, \lceil \log_2(r) \rceil\}$ such that for each $k = 1, \ldots, r$,

$$2^{-\alpha_k} \geq \|x_{S_k}\|_2^2 \geq 2^{-\alpha_k - 1} \qquad \text{or} \qquad 2^{-\alpha_k} = 2^{-\lceil \log_2(r) \rceil} \geq \|x_{S_k}\|_2^2. \tag{4.18}$$

Define $\tilde{n} := \lceil \log_2(r) \rceil$, for $n = \tilde{n}+1, \tilde{n}+2, \ldots, \lceil \log_2(s) \rceil + 2$ and $R_{n,k} = \lceil c_1 \log \rho \rceil \cdot 2^{n-\alpha_k}$ for an absolute constant $c_1 > 0$ that will be specified later. Since $n \geq \lceil \log_2(r) \rceil \geq \alpha_k$, $R_{n,k}$ is always an integer.

We define the independent random variables $Z^{(k)}$ for $k \in [r]$ such that for all $j \in S_k$,

$$Z^{(k)} = 2^{-\frac{\alpha_k}{2}} \sqrt{2s_k}\, \mathrm{sign}(\mathrm{Re}(x_j)) e_j \qquad \text{with probaility } 2^{\frac{\alpha_k}{2}} \frac{|\mathrm{Re}(x_j)|}{\sqrt{2s_k}}$$

$$Z^{(k)} = 2^{-\frac{\alpha_k}{2}} \sqrt{2s_k}\, \mathrm{sign}(\mathrm{Im}(x_j)) e_j \qquad \text{with probaility } 2^{\frac{\alpha_k}{2}} \frac{|\mathrm{Im}(x_j)|}{\sqrt{2s_k}}$$

and

$$Z^{(k)} = 0 \qquad \text{with probability } 1 - \frac{2^{\frac{\alpha_k}{2}}}{\sqrt{2s_k}} \sum_{j \in S_k} (|\mathrm{Re}(x_j)| + |\mathrm{Im}(x_j)|).$$

That this is a valid probability distribution since

$$\frac{2^{\frac{\alpha_k}{2}}}{\sqrt{2s_k}} \sum_{j \in S_k} (|\mathrm{Re}(x_j)| + |\mathrm{Im}(x_j)|) \leq \frac{2^{\frac{\alpha_k}{2}}}{\sqrt{2s_k}} \sum_{j \in S_k} \sqrt{2}|x_j| = \frac{2^{\frac{\alpha_k}{2}}}{\sqrt{s_k}} \|x_{S_k}\|_1 \leq 2^{\frac{\alpha_k}{2}} \|x_{S_k}\|_2 \leq 1$$

and furthermore we obtain for all $j \in [N]$,

$$(\mathbb{E}Z^{(k)})_j = \begin{cases} x_j & \text{if } j \in S_k \\ 0 & \text{otherwise.} \end{cases}$$

There are $\lceil \log_2(r) \rceil + 1$ different possible values of $\alpha_k$. Therefore the total number of values that $Z^{(k)}$ can attain across all $x \in D_{\mathbf{s},\mathbf{M}}$ is bounded by

$$4|\mathcal{M}_k|(\lceil \log_2(r) \rceil + 1) + 1 \leq 4Nr + 1 \leq 5N^2 \leq N^5$$

for $N \geq 2$.

We define $Z_1^{(k)}, \ldots, Z_{R_{n,k}}^{(k)}$ to be independent copies of $Z^{(k)}$ and

$$\bar{Z}_n = \sum_{k \in [r]} \frac{1}{R_{n,k}} \sum_{q=1}^{R_{n,k}} Z_q^{(k)}.$$

Then

$$\mathbb{E}\bar{Z}_n = x.$$

$\bar{Z}_n$ depends on $\alpha_1, \ldots, \alpha_r$. We take all choices of $\alpha_1, \ldots, \alpha_r \in \{0, 1, \ldots, \lceil \log_2(r) \rceil\}$ for which

$$\sum_{k=1}^{r} 2^{-\alpha_k} \leq 3. \tag{4.19}$$

We define $V_n$ to be the set of all possible values of $\bar{Z}_n$ for all of the above choices of $\alpha_1, \ldots, \alpha_r$.

For each of these choices of $(\alpha_1, \ldots, \alpha_r)$, each $Z_q^{(k)}$ in the definition of $\bar{Z}_n$ can have at most $N^5$ different values such that the number of possible values of $\bar{Z}_n$ for the choice of $(\alpha_1, \ldots, \alpha_r)$ is bounded by

$$\prod_{k \in [r]} (N^5)^{R_{n,k}} = N^{5 \sum_{k \in [r]} R_{n,k}} = N^{5 \sum_{k \in [r]} (\lceil c_1 \log \rho \rceil \cdot 2^n \cdot 2^{-\alpha_k})} \leq N^{15 \lceil c_1 \log \rho \rceil \cdot 2^n}. \tag{4.20}$$

The number of different choices of $\alpha_1, \ldots, \alpha_r$ is

$$(\lceil \log_2(r) \rceil + 1)^r \leq r^r \leq N^{2^n}.$$

So in total

$$|V_n| \leq N^{2^n} \cdot N^{15 \lceil c_1 \log \rho \rceil \cdot 2^n} \leq N^{16 \lceil c_1 \log \rho \rceil \cdot 2^n}. \tag{4.21}$$

For each $l = r_0 + 1, \ldots, r$, $i = 1, \ldots, m_l$,

$$|(AZ_q^{(k)})_j| \leq \|A_j\|_\infty \|Z_q^{(k)}\|_1 \leq \sqrt{\frac{2\mu_{l,k} s_k 2^{-\alpha_k}}{p_l}}.$$

So

$$\left\| (AZ_q^{(k)})_j - \mathbb{E}(AZ_q^{(k)})_j \right\|_{\psi_2} \leq c_2 \sqrt{\frac{\mu_{l,k} s_k 2^{-\alpha_k}}{p_l}}$$

for an absolute constant $c_2 > 0$.

Now we can use Proposition 0.7 about sums of centered subgaussian variables,

$$\begin{aligned}
\left\| (A\bar{Z}_n)_j - (Ax)_j \right\|_{\psi_2} &= \left\| \sum_{k \in [r]} \frac{1}{R_{n,k}} \sum_{q=1}^{R_{n,k}} ((AZ_q^{(k)})_j - \mathbb{E}(AZ_q^{(k)})_j) \right\|_{\psi_2} \\
&\leq c_3 \sqrt{\sum_{k \in [r]} \frac{1}{R_{n,k}^2} \sum_{q=1}^{R_{n,k}} \frac{\mu_{l,k} s_k 2^{-\alpha_k}}{p_l}} = c_3 \sqrt{\sum_{k \in [r]} \frac{\mu_{l,k} s_k 2^{-\alpha_k}}{R_{n,k} p_l}} \\
&\leq c_3 \sqrt{\max_{l' \in \{r_0+1, \ldots, r\}} \sum_{k \in [r]} \frac{\mu_{l',k} s_k}{\lceil c_1 \log \rho \rceil \cdot 2^n p_{l'}}}
\end{aligned}$$

$$\leq 2^{-\frac{n}{2}} \frac{c_3}{\sqrt{c_1 \log \rho}} \sqrt{\max_{l' \in \{r_0+1,\ldots,r\}} \sum_{k \in [r]} \frac{\mu_{l',k} s_k}{p_{l'}}} = 2^{-\frac{n}{2}} \frac{c_3}{\sqrt{c_1 \log \rho}} \sqrt{\gamma}$$

for an absolute constant $c_3 > 0$, such that

$$\mathbb{P}\left(\left|(A\bar{Z}_n)_j - (Ax)_j\right| > 2^{-\frac{n}{2}}\sqrt{\gamma}\right) \leq 2\exp\left(-\frac{c_4 c_1 \log \rho}{c_3^2}\right) \leq \frac{2}{\rho},$$

where $c_4 > 0$ is a constant and in the last step, we choose $c_1 = \frac{c_3^2}{c_4}$.

For each $l = r_0 + 1, \ldots, r$, $j \in \tilde{\mathcal{N}}_l$, define the indicator variable $\chi_j \in \{0, 1\}$ such that $\chi_j = 1$ if

$$\left|(A\bar{Z}_n)_j - (Ax)_j\right| > 2^{-\frac{n}{2}}\sqrt{\gamma} \tag{4.22}$$

and $\chi_j = 0$ otherwise.

The above probability bound shows

$$\mathbb{E}\chi_j \leq \frac{2}{\rho}.$$

For each $l = r_0 + 1, \ldots, r$, the sum $\sum_{j \in \tilde{\mathcal{N}}_l} \chi_j$ is the number of indices $j \in \tilde{\mathcal{N}}_l$ such that (4.22) holds for $j$ and we know

$$\mathbb{E} \sum_{j \in \tilde{\mathcal{N}}_l} \chi_j \leq \frac{2m_l}{\rho}.$$

First assume that the second part in the minimum of definition (4.17) is the smaller one, i.e., $\rho = 4\delta^{-1} rs(\lceil \log_2(s) \rceil + 2) \cdot \sum_{l=r_0+1}^{r} m_l \gamma_l$. Then we apply a union bound and Markov's inequality to show

$$\mathbb{P}\left(\exists l \in \{r_0+1, \ldots, r\} : \sum_{j \in \tilde{\mathcal{N}}_l} \chi_j \geq \frac{4rm_l}{\rho}\right) \leq \sum_{l=r_0+1}^{r} \mathbb{P}\left(\sum_{j \in \tilde{\mathcal{N}}_l} \chi_j \geq \frac{4rm_l}{\rho}\right)$$
$$\leq \sum_{l=r_0+1}^{r} \frac{\mathbb{E}\sum_{j \in \tilde{\mathcal{N}}_l} \chi_j}{\frac{4rm_l}{\rho}} \leq \sum_{l=r_0+1}^{r} \frac{1}{2r} \leq \frac{1}{2}.$$

So

$$\mathbb{P}\left(\forall l \in \{r_0+1, \ldots, r\} : \sum_{j \in \tilde{\mathcal{N}}_l} \chi_j < \frac{4rm_l}{\rho}\right) \geq \frac{1}{2} > 0$$

and there is one realization

$$\tilde{\pi}_n(x) \tag{4.23}$$

of $\bar{Z}_n$ such that for each $l = r_0 + 1, \ldots, r$, there are at most $\frac{4rm_l}{\rho}$ indices $j \in \tilde{\mathcal{N}}_l$ such that (4.22) holds. We define $\tilde{J}_{n,l} \subset \tilde{\mathcal{N}}_l$ to be the set of indices $j \in \tilde{\mathcal{N}}_l$ for which this is the case such that $|\tilde{J}_{n,l}| \leq \frac{4rm_l}{\rho}$. Furthermore, we define $J_l := \bigcup_{n=\tilde{n}+1}^{\lceil \log_2(s) \rceil + 2} \tilde{J}_{n,l}$ as the set of indices for which (4.22) holds for any $n$. Then $|J_l| \leq \frac{4rm_l(\lceil \log_2(s) \rceil + 2)}{\rho}$. We also define

$$J := \bigcup_{l=r_0+1}^{r} J_l.$$

The bound on $|J_l|$ and the definition of $\rho$ also imply the inequality

$$\sum_{l=r_0+1}^{r} |J_l|\gamma_l \leq \frac{4r(\lceil \log_2(s)\rceil + 2)}{\rho} \sum_{l=r_0+1}^{r} m_l\gamma_l = \frac{\sum_{l=r_0+1}^{r} m_l\gamma_l}{\delta^{-1}s \cdot \sum_{l'=r_0+1}^{r} m_{l'}\gamma_{l'}} = \frac{\delta}{s}. \tag{4.24}$$

If on the other hand $\rho = 4\tilde{m}$, then the expected number of indices $j \in \tilde{\mathcal{N}}$ in total (for all $l$) for which (4.22) holds, is

$$\mathbb{E}\sum_{j\in\tilde{\mathcal{N}}} \chi_j \leq \sum_{l=r_0+1}^{r} \frac{2m_l}{\rho} \leq \frac{1}{2} < 1.$$

Therefore, there is always a realization of $\bar{Z}_n$ such that (4.22) does not hold for any $j \in \tilde{\mathcal{N}}$ and we define this as $\tilde{\pi}_n(x) \in V_n$. Then consistent with the other case, we define $J_l = \emptyset$ for all $l \in \{r_0+1, \dots, r\}$ and $J = \emptyset$ such that (4.24) still holds.

In order to show that $\tilde{\pi}_n(x) \in V_n$, we only need to show that the choice of the $\alpha_k$ (4.18) satisfies the requirement (4.19) of the definition of $V_n$. To to this, note that for each $k \in [r]$, $2^{-\alpha_k} \leq 2\|x_{S_k}\|_2^2$ or $2^{-\alpha_k} \leq 2^{-\log_2(r)} = \frac{1}{r}$. So altogether

$$\sum_{k=1}^{r} 2^{-\alpha_k} \leq \sum_{k=1}^{r} \max\{2\|x_{S_k}\|_2^2, \frac{1}{r}\} \leq 2\sum_{k=1}^{r} \|x_{S_k}\|_2^2 + \sum_{k=1}^{r}\frac{1}{r} = 2\|x\|_2^2 + 1 \leq 3,$$

which shows that $z_n \in V_n$.

Moreover, for the random variable $\bar{Z}_n$ it holds with probability 1 that for all $l \in \{r_0+1, \dots, r\}$ and $j \in \tilde{\mathcal{N}}_l$,

$$\begin{aligned}|(A\bar{Z}_n)_j| &\leq \sum_{k=1}^{r}\sum_{k'\in S_k} |A_{j,k'}||(\bar{Z}_n)_{k'}| \leq \sum_{k=1}^{r}\sqrt{\frac{\mu_{l,k}}{p_l}}\|(\bar{Z}_n)_{S_k}\|_1 \leq \sum_{k=1}^{r}\sqrt{\frac{\mu_{l,k}}{p_l}}\frac{1}{R_{n,k}}\sum_{q=1}^{R_{n,k}}\|(Z_q^{(k)})_{S_k}\|_1 \\ &\leq \sum_{k=1}^{r}\sqrt{\frac{\mu_{l,k}}{p_l}}2^{-\frac{\alpha_k}{2}}\sqrt{2s_k} \leq \sqrt{2}\sqrt{\sum_{k=1}^{r}\frac{\mu_{l,k}s_k}{p_l}} \cdot \sqrt{\sum_{k=1}^{r}2^{-\alpha_k}} \leq \sqrt{6\gamma_l}\end{aligned}$$

and therefore

$$|(A\tilde{\pi}_n(x))_j| \leq \sqrt{6\gamma_l}. \tag{4.25}$$

**Step 3: Construction of the admissible sequence for $n \leq n_1 + \lceil\log_2(s)\rceil$**

In this step we define the first part of the admissible sequence for the index set $T$.

Note that we always have $r \leq s$. For $n = \lceil\log_2(r)\rceil + 1, \dots, \lceil\log_2(s)\rceil$, we define sets

$$I_n := \left\{ j \in \tilde{\mathcal{N}} \,\middle|\, |(A\tilde{\pi}_{n+2}(x))_j| \geq 2^{-\frac{n}{2}}\sqrt{\gamma} \right\} \setminus \left( \bigcup_{n'=\lceil\log_2(r)\rceil+1}^{n-1} I_{n'} \right). \tag{4.26}$$

of indices that depend on $x$ where $\tilde{\pi}_{n+2}(x)$ is the one defined in (4.23) and $\tilde{\mathcal{N}}$ is defined in (4.16). The intention of these sets $I_n$ is that it contains those $j \in \tilde{\mathcal{N}}$ for which $|(Ax)_j|^2$ is approximately in the range between $2^{-n}\gamma$ and $2^{-n+1}\gamma$. In order to limit the number of possible choices for the set $I_n$ across all $x \in D_{\mathbf{s},\mathbf{M}}$, we define it based on the approximation $|(A\tilde{\pi}_{n+2}(x))_j|$ instead of $|(Ax)_j|$.

Based on the sets $I_n$ from above, we choose $n_\delta := \lceil\log_2(\frac{1}{\delta^2})\rceil = \lceil 2\log_2(\frac{1}{\delta})\rceil$ and define

$$\tilde{I}_n^{(1)} := \bigcup_{n'=\max\{1,n-n_\delta\}}^{n} I_{n'} \qquad \tilde{I}_n^{(2)} := \bigcup_{n'=1}^{n-n_\delta-1} I_{n'} \qquad \tilde{I}_n^{(3)} := \tilde{\mathcal{N}}\backslash(\tilde{I}_n^{(1)} \cup \tilde{I}_n^{(2)}).$$

Each index $j \in \tilde{\mathcal{N}}$ is contained in exactly one of these sets.

Now we define the vectors $w^{(n)} = w^{(n)}(x) \in \mathbb{R}^m$ such that for $n = 0, \dots, \lceil \log_2(r) \rceil$,

$$w^{(n)} = 0$$

and for $n = \lceil \log_2(r) \rceil + 1, \dots, \lceil \log_2(s) \rceil$, $j \in [m]$,

$$w_j^{(n)} = \begin{cases} |(A\tilde{\pi}_{n+2}(x))_j|^2 & \text{if } j \in \tilde{I}_n^{(1)} \\ |(A\tilde{\pi}_{n'+n_\delta+2}(x))_j|^2 & \text{if } j \in \tilde{I}_n^{(2)} \text{ and } j \in I_{n'} \\ 0 & \text{if } j \in \tilde{I}_n^{(3)} \text{ or } j \in [m] \backslash \tilde{\mathcal{N}}. \end{cases} \tag{4.27}$$

We define $n_1 := n_0 + \lceil \log_2 \log_2(N) + \log_2 \log(\rho) \rceil$ for a constant integer $n_0 \geq 0$ that will be chosen later. Furthermore, we start defining an admissible sequence by setting

$$T_0 = T_1 = T_2 = \dots = T_{n_1} = \{0\} \subset \mathbb{R}^m$$

and for $n = 1, \dots, \lceil \log_2(s) \rceil$,

$$T_{n_1+n} = \{w^{(n)}(x) \,|\, x \in D_{\mathbf{s},\mathbf{M}}\}. \tag{4.28}$$

Note that $I_{\lceil \log_2(r) \rceil+1}, \dots, I_n$ and therefore also $w^{(n)}$ for $n \geq \lceil \log_2(r) \rceil + 1$ are completely determined by the approximations

$$\tilde{\pi}_{\lceil \log_2(r) \rceil+3}(x) \in V_{\lceil \log_2(r) \rceil+3}, \ \dots, \ \tilde{\pi}_{n+2}(x) \in V_{n+2}.$$

So we can use the bound (4.21) on the $|V_n|$ to bound the number of possible different values of $w^{(n)}(x)$ such that for $\lceil \log_2(r) \rceil + 1 \leq n \leq \lceil \log_2(s) \rceil$,

$$\begin{aligned} |T_{n_1+n}| &\leq \prod_{n'=\lceil \log_2(r) \rceil+3}^{n+2} |V_{n'}| \leq \prod_{n'=0}^{n+2} N^{16\lceil c_1 \log \rho \rceil \cdot 2^{n'}} = N^{16\lceil c_1 \log \rho \rceil \cdot \sum_{n'=0}^{n+2} 2^{n'}} \leq N^{\tilde{c}_1 \log(\rho) \cdot 2^n} \\ &= 2^{\tilde{c}_1 \log_2(N) \log(\rho) \cdot 2^n} \leq 2^{2^{n_1+n}} \end{aligned}$$

for a constant $\tilde{c}_1 > 0$ and $n_0 \geq \log_2(\tilde{c}_1)$. Then

$$|T_n| \leq 2^{2^n} \tag{4.29}$$

holds for all $1 \leq n \leq n_1 + \lceil \log_2(s) \rceil$.

Now we observe that for any $l \in [r]$ and $j' \in \mathcal{N}_l$,

$$\begin{aligned} |(Ux)_{j'}| &= \left| \sum_{k=1}^{r} \sum_{k' \in \mathcal{M}_k} U_{j',k'} x_{k'} \right| \leq \sum_{k=1}^{r} \left( \max_{k' \in \mathcal{M}_k} |U_{j',k'}| \cdot \sum_{k' \in \mathcal{M}_k} |x_{k'}| \right) \leq \sum_{k=1}^{r} \left( \sqrt{\mu_{l,k}} \cdot \|x_{S_k}\|_1 \right) \\ &\leq \sum_{k=1}^{r} \left( \sqrt{\mu_{l,k} s_k} \cdot \|x_{S_k}\|_2 \right) \leq \sqrt{\sum_{k=1}^{r} \mu_{l,k} s_k} \cdot \sqrt{\sum_{k=1}^{r} \|x_{S_k}\|_2^2} = \sqrt{\sum_{k=1}^{r} \mu_{l,k} s_k} \end{aligned} \tag{4.30}$$

and therefore for any $j \in \tilde{\mathcal{N}}_l$,

$$|(Ax)_j| \leq \sqrt{\sum_{k=1}^{r} \frac{\mu_{l,k} s_k}{p_l}} = \sqrt{\gamma_l} \leq \sqrt{\gamma}, \tag{4.31}$$

where the last part only holds for $l = r_0 + 1, \dots, r$.

For any $j \in I_{n'} \backslash J$ for $n' \geq \lceil \log_2(r) \rceil + 1$, we have $|(A\tilde{\pi}_{n'+2}(x))_j| \geq 2^{-\frac{n'}{2}}\sqrt{\gamma}$ and (because of $j \notin J$) $|(A\tilde{\pi}_{n'+2}(x))_j - (Ax)_j| \leq 2^{-\frac{n'+2}{2}}\sqrt{\gamma} = \frac{1}{2} \cdot 2^{-\frac{n'}{2}}\sqrt{\gamma}$. Hence,

$$|(Ax)_j| \geq |(A\tilde{\pi}_{n'+2}(x))_j| - |(A\tilde{\pi}_{n'+2}(x))_j - (Ax)_j| \geq \frac{1}{2} \cdot 2^{-\frac{n'}{2}}\sqrt{\gamma}.$$

If in addition $n' \geq \lceil \log_2(r) \rceil + 2$ holds, then we know that $j \notin I_{n'-1}$ and therefore $|(A\tilde{\pi}_{n'+1}(x))_j| < 2^{-\frac{n'-1}{2}}\sqrt{\gamma}$. Together with $|(A\tilde{\pi}_{n'+1}(x))_j - (Ax)_j| \leq 2^{-\frac{n'+1}{2}}\sqrt{\gamma} = \frac{1}{2} \cdot 2^{-\frac{n'-1}{2}}\sqrt{\gamma}$, this yields

$$|(Ax)_j| \leq |(A\tilde{\pi}_{n'+1}(x))_j| + |(A\tilde{\pi}_{n'+1}(x))_j - (Ax)_j| \leq \frac{3}{2} \cdot 2^{-\frac{n'-1}{2}}\sqrt{\gamma} \leq \frac{5}{2} \cdot 2^{-\frac{n'}{2}}\sqrt{\gamma}.$$

So we can summarize that for $j \in I_{n'} \backslash J$ ($n' \geq \lceil \log_2(r) \rceil + 1$)

$$\frac{1}{2} \cdot 2^{-\frac{n'}{2}}\sqrt{\gamma} \leq |(Ax)_j| \leq \frac{5}{2} \cdot 2^{-\frac{n'}{2}}\sqrt{\gamma}, \tag{4.32}$$

where the second bound only holds if $n' \geq \lceil \log_2(r) \rceil + 2$. In addition, for any $n \geq n'$,

$$\begin{aligned} |(A\tilde{\pi}_{n+2}(x))_j| &\leq |(Ax)_j| + |(A\tilde{\pi}_{n+2}(x))_j - (Ax)_j| \leq |(Ax)_j| + \frac{1}{2} \cdot 2^{-\frac{n}{2}}\sqrt{\gamma} \\ &\leq |(Ax)_j| + \frac{1}{2} \cdot 2^{-\frac{n'}{2}}\sqrt{\gamma} \leq 2 \cdot |(Ax)_j|. \end{aligned} \tag{4.33}$$

Now we assume $j \in I_{n'} \backslash J$ again for $\lceil \log_2(r) \rceil + 1 \leq n' \leq \lceil \log_2(s) \rceil$. In the next step, we bound $(w_j^{(n)} - w_j^{(n-1)})^2$ for all possible cases.

- $n \leq n' - 1$:

$$(w_j^{(n)} - w_j^{(n-1)})^2 = (0 - 0)^2 = 0.$$

- $n = n'$:

$$\begin{aligned} (w_j^{(n)} - w_j^{(n-1)})^2 &= (|(A\tilde{\pi}_{n+2}(x))_j|^2 - 0)^2 = |(A\tilde{\pi}_{n+2}(x))_j|^4 \\ &\leq 2^4 |(Ax)_j|^4 \leq 100 \cdot 2^{-n}\gamma |(Ax)_j|^2 \end{aligned}$$

- $n' + 1 \leq n \leq n' + n_\delta$:

$$\begin{aligned} (w_j^{(n)} - w_j^{(n-1)})^2 &= \left(|(A\tilde{\pi}_{n+2}(x))_j|^2 - |(A\tilde{\pi}_{n+1}(x))_j|^2\right)^2 \\ &= \left((|(A\tilde{\pi}_{n+2}(x))_j|^2 - |(Ax)_j|^2) + (|(Ax)_j|^2 - |(A\tilde{\pi}_{n+1}(x))_j|^2)\right)^2 \\ &= 2\left((|(A\tilde{\pi}_{n+2}(x))_j|^2 - |(Ax)_j|^2)^2 + (|(Ax)_j|^2 - |(A\tilde{\pi}_{n+1}(x))_j|^2)^2\right) \end{aligned}$$

  We can bound the second factor by

$$\begin{aligned} \left(|(A\tilde{\pi}_{n+1}(x))_j|^2 - |(Ax)_j|^2\right)^2 &= (|(A\tilde{\pi}_{n+1}(x))_j| - |(Ax)_j|)^2 (|(A\tilde{\pi}_{n+1}(x))_j| + |(Ax)_j|)^2 \\ &\leq |(A\tilde{\pi}_{n+1}(x))_j - (Ax)_j|^2 \cdot (3|(Ax)_j|)^2 \\ &\leq 9 \cdot (2^{-\frac{n+1}{2}}\sqrt{\gamma})^2 |(Ax)_j|^2 \\ &\leq \frac{9}{2} 2^{-n}\gamma |(Ax)_j|^2. \end{aligned}$$

  The same bound also follows for the other term $(|(Ax)_j|^2 - |(A\tilde{\pi}_{n+2}(x))_j|^2)^2$, such that in total we obtain

$$(w_j^{(n)} - w_j^{(n-1)})^2 \leq 18 \cdot 2^{-n}\gamma |(Ax)_j|^2$$

- $n \geq n' + n_\delta + 1$:

$$(w_j^{(n)} - w_j^{(n-1)})^2 = (|(A\tilde{\pi}_{n'+n_\delta+2}(x))_j|^2 - |(A\tilde{\pi}_{n'+n_\delta+2}(x))_j|^2)^2 = 0.$$

  Note that this especially also holds for the case $n = n' + n_\delta + 1$.

So if $j \notin J$ and $j \in I_{n'}$ for some $n' \leq \lceil \log_2(s) \rceil$, then we have shown that for any $\lceil \log_2(r) \rceil + 2 \leq n \leq \lceil \log_2(s) \rceil$,

$$(w_j^{(n)} - w_j^{(n-1)})^2 \begin{cases} \leq 100 \cdot 2^{-n}\gamma |(Ax)_j|^2 & \text{if } n' \leq n \leq n' + n_\delta \\ = 0 & \text{otherwise.} \end{cases}$$

If $j \notin I_{n'}$ for any $n' \leq \lceil \log_2(s) \rceil$, then $(w_j^{(n)} - w_j^{(n-1)})^2 = 0$ for all $n$.

All the other indices $j \in \tilde{\mathcal{N}}$ are in $J = \bigcup_{l=r_0+1}^{r} J_l$. From inequality (4.25), it follows that $|w_j^{(n)}|, |w_j^{(n-1)}| \leq 6\gamma_l$. Therefore we obtain

$$\begin{aligned} \sum_{j \in J} 2^n (w_j^{(n)} - w_j^{(n-1)})^2 &= \sum_{l=r_0+1}^{r} \sum_{j \in J_l} 2^n (w_j^{(n)} - w_j^{(n-1)})^2 \leq \sum_{l=r_0+1}^{r} |J_l| \cdot 2^n (12\gamma_l)^2 \\ &\leq 12^2 \cdot 2^n \cdot \max_{l \in \{r_0+1,\ldots,r\}} \gamma_l \cdot \sum_{l=r_0+1}^{r} |J_l|\gamma_l \leq 12^2 \cdot 2^n \gamma \cdot \frac{1}{s}, \end{aligned}$$

where we used (4.24) and $\delta \leq 1$ in the last step.

Towards bounding the contribution of the approximations $w^{(n)}$ in the $\gamma_2$ functional, we obtain with the above estimates,

$$\sum_{n=\lceil \log_2(r) \rceil+2}^{\lceil \log_2(s) \rceil} 2^{\frac{n}{2}} d(w^{(n)}, w^{(n-1)}) = \sum_{n=\lceil \log_2(r) \rceil+2}^{\lceil \log_2(s) \rceil} \left( \sum_{j \in \tilde{\mathcal{N}}} 2^n (w_j^{(n)} - w_j^{(n-1)})^2 \right)^{\frac{1}{2}} \tag{4.34}$$

$$\begin{aligned} &\leq \sum_{n=\lceil \log_2(r) \rceil+2}^{\lceil \log_2(s) \rceil} \left( \sum_{n'=\lceil \log_2(r) \rceil+1}^{\lceil \log_2(s) \rceil} \sum_{j \in I_{n'} \setminus J} 2^n (w_j^{(n)} - w_j^{(n-1)})^2 \right)^{\frac{1}{2}} \\ &\quad + \sum_{n=\lceil \log_2(r) \rceil+2}^{\lceil \log_2(s) \rceil} \left( \sum_{j \in J} 2^n (w_j^{(n)} - w_j^{(n-1)})^2 \right)^{\frac{1}{2}} \\ &\leq \sum_{n=\lceil \log_2(r) \rceil+2}^{\lceil \log_2(s) \rceil} \left( \sum_{n'=\lceil \log_2(r) \rceil+1}^{\lceil \log_2(s) \rceil} \sum_{j \in I_{n'}} 100\gamma |(Ax)_j|^2 \mathbb{1}_{n' \leq n \leq n'+n_\delta} \right)^{\frac{1}{2}} + \sum_{n=\lceil \log_2(r) \rceil+2}^{\lceil \log_2(s) \rceil} 12 \cdot 2^{\frac{n}{2}} \sqrt{\frac{\gamma}{s}} \\ &\leq \sqrt{100\gamma} \sum_{n=\lceil \log_2(r) \rceil+2}^{\lceil \log_2(s) \rceil} \left( \sum_{n'=\lceil \log_2(r) \rceil+1}^{\lceil \log_2(s) \rceil} \|(Ax)_{I_{n'}}\|_2^2 \mathbb{1}_{n' \leq n \leq n'+n_\delta} \right)^{\frac{1}{2}} + \tilde{C}_2\sqrt{\gamma} \qquad (4.35) \\ &\leq \sqrt{100\gamma \lceil \log_2(s) \rceil} \left( \sum_{n'=\lceil \log_2(r) \rceil+1}^{\lceil \log_2(s) \rceil} \|(Ax)_{I_{n'}}\|_2^2 \sum_{n=\lceil \log_2(r) \rceil+2}^{\lceil \log_2(s) \rceil} \mathbb{1}_{n' \leq n \leq n'+n_\delta} \right)^{\frac{1}{2}} + \tilde{C}_2\sqrt{\gamma} \qquad (4.36) \\ &\leq \sqrt{100\gamma \lceil \log_2(s) \rceil} \left( \sum_{n'=\lceil \log_2(r) \rceil+1}^{\lceil \log_2(s) \rceil} \|(Ax)_{I_{n'}}\|_2^2 (n_\delta + 1) \right)^{\frac{1}{2}} + \tilde{C}_2\sqrt{\gamma} \\ &\leq \sqrt{100\gamma \lceil \log_2(s) \rceil (n_\delta + 1)} \|Ax\|_2 + \tilde{C}_2\sqrt{\gamma} \end{aligned}$$

$$\leq C_2\sqrt{\gamma \log_2(s)\log_2(\frac{1}{\delta}))\cdot\left(1+\|Ax\|_2^2\right)}.$$

Here $C_2 > 0$ is a constant, $\mathbb{1}_{n'\leq n\leq n'+n_\delta}$ is 1 for $n' \leq n \leq n'+n_\delta$ and 0 otherwise, and in the step leading from (4.35) to (4.36), we applied the Cauchy-Schwarz inequality to the sum over $n$ in the sense that $\sum_n a_n^{\frac{1}{2}} = \sum_n (a_n^{\frac{1}{2}}\cdot 1) \leq \sqrt{\sum_n a_n}\cdot\sqrt{\sum_n 1} \leq \sqrt{\sum_n a_n}\cdot\sqrt{\lceil\log_2(s)\rceil}$.

Furthermore,

$$\sum_{n=1}^{\lceil\log_2(r)\rceil} 2^{\frac{n}{2}} d(w^{(n)}, w^{(n-1)}) = 0.$$

For the one missing term $n = \lceil\log_2(r)\rceil + 1$, first assume $j \in I_n\backslash J$. Then combining (4.25) and (4.33) yields

$$2^n(w_j^{(n)} - w_j^{(n-1)})^2 = 2^n(w_j^{(n)})^2 = 2^n|(A\tilde{\pi}_{n+2}(x))_j|^4 \leq 2^n\cdot 6\gamma|(Ax)_j|^2 \leq 100r\gamma|(Ax)_j|^2.$$

On the other hand, if $j \notin J$ and $j \notin I_n$, then $w_j^{(n)} - w_j^{(n-1)} = 0$. And also for $n = \lceil\log_2(r)\rceil + 1$, we obtain that

$$\sum_{j\in J} 2^n(w_j^{(n)} - w_j^{(n-1)})^2 \leq 16\gamma,$$

such that altogether for $n = \lceil\log_2(r)\rceil + 1$ and some constant $C_3 > 0$,

$$2^{\frac{n}{2}}\left(\sum_{l=r_0+1}^{r}\sum_{j\in\tilde{\mathcal{N}}_l}(w_j^{(n)} - w_j^{(n-1)})^2\right)^{\frac{1}{2}} \leq C_3\sqrt{\gamma\left(1 + r\sum_{l=r_0+1}^{r}\sum_{j\in\tilde{\mathcal{N}}_l}|(Ax)_j|^2\right)}.$$

So in total for a constant $C_4 > 0$,

$$\sum_{n=1}^{\lceil\log_2(s)\rceil} 2^{\frac{n}{2}} d(w^{(n)}, w^{(n-1)}) \leq C_4\sqrt{\gamma\log_2(s)\log_2(\frac{1}{\delta})) + \gamma\left(\log_2(s)\log_2(\frac{1}{\delta})) + r\right)\|Ax\|_2^2} \quad (4.37)$$

**Step 4: Construction of $T^{(1)}$ and $T^{(2)}$ for (4.11) and $\ell_1$ bound**

For each $x \in D_{\mathbf{s},\mathbf{M}}$, define $f(x) \in \mathbb{R}^m$ such that if $j \in I_{n'}$ for an $\lceil\log_2(r)\rceil + 1 \leq n' \leq \lceil\log_2(s)\rceil - n_\delta$ (recall that $n_\delta = \lceil\log_2(\frac{1}{\delta^2})\rceil$)

$$(f(x))_j = w_j^{(\lceil\log_2(s)\rceil)}$$

and otherwise, if there is no such $n'$,

$$(f(x))_j = |(Ax)_j|^2.$$

Then for each $j \in I_{n'}\backslash J$ (with $\lceil\log_2(r)\rceil + 1 \leq n' \leq \lceil\log_2(s)\rceil - n_\delta$),

$$\begin{aligned}\left|(f(x))_j - |(Ax)_j|^2\right| &= \left|w_j^{(\lceil\log_2(s)\rceil)} - |(Ax)_j|^2\right| = \left||(A\tilde{\pi}_{n'+n_\delta+2}(x))_j|^2 - |(Ax)_j|^2\right| \\ &\leq |(A\tilde{\pi}_{n'+n_\delta+2}(x))_j - (Ax)_j|\left(|(A\tilde{\pi}_{n'+n_\delta+2}(x))_j| + |(Ax)_j|\right) \\ &\leq 2^{-\frac{n'+n_\delta+2}{2}}\sqrt{\gamma}\cdot 3|(Ax)_j| \leq \frac{\delta}{2}\cdot 2^{-\frac{n'}{2}}\sqrt{\gamma}\cdot 3|(Ax)_j| \\ &\leq 3\delta|(Ax)_j|^2,\end{aligned}$$

where we used (4.33) and (4.32).

Furthermore, by (4.25) and (4.31), we have $|(f(x))_j| \leq 6\gamma_l$ and $|(Ax)_j|^2 \leq \gamma_l$ in any case for $l \in \{r_0+1,\ldots,r\}$, $j \in \tilde{\mathcal{N}_l}$, which implies

$$\sum_{j\in J} \left|(f(x))_j - |(Ax)_j|^2\right| = \sum_{l=r_0+1}^{r} \sum_{j\in J_l} \left|(f(x))_j - |(Ax)_j|^2\right| \leq 7 \sum_{l=r_0+1}^{r} |J_l|\gamma_l \leq 7\frac{\delta}{s} \leq 7\delta,$$

using (4.24). Then we obtain the $\ell_1$ norm bound

$$\begin{aligned} \left\| |Ax|^2 - f(x)\right\|_1 &= \sum_{n'=\lceil \log_2(r)\rceil+1}^{\lceil \log_2(s)\rceil - n_\delta} \sum_{j\in I_{n'}\setminus J} \left|(f(x))_j - |(Ax)_j|^2\right| + \sum_{j\in J}\left|(f(x))_j - |(Ax)_j|^2\right| \\ &\leq \delta\left(3\sum_{l=r_0+1}^{r}\sum_{j\in\tilde{\mathcal{N}_l}} |(Ax)_j|^2 + 7\right) \leq \delta\left(3\|Ax\|_2^2 + 7\right). \end{aligned} \tag{4.38}$$

Define

$$\tilde{\delta} := \delta\left(3 \sup_{x\in D_{\mathbf{s},\mathbf{M}}} \|Ax\|_2^2 + 7\right). \tag{4.39}$$

and the sets

$$\begin{aligned} T^{(1)} &:= \{f(x) \,|\, x \in D_{\mathbf{s},\mathbf{M}}\} \\ T^{(2)} &:= \{y \in \mathbb{R}^m \,|\, \|y\|_1 \leq \tilde{\delta}\}. \end{aligned}$$

Then (4.38) shows that

$$T = \{|Ax|^2 \,|\, x \in D_{\mathbf{s},\mathbf{M}}\}$$

satisfies

$$T \subset T^{(1)} + T^{(2)}.$$

We have already defined $T_n$ for $0 \leq n \leq n_1 + \lceil \log_2(s)\rceil$ and will extend this to an admissible sequence $(T_n)_{n\geq 0}$ to control $\gamma_2(T^{(1)}, d)$, while $T^{(2)}$ is controlled with its $\ell_1$ norm bound such that (4.11) can be applied.

**Step 5: Application of Lemma 4.10 to the first part of the admissible sequence**

If $j \in I_{n'}$ for any $\lceil \log_2(r)\rceil + 1 \leq n' \leq \lceil \log_2(s)\rceil - n_\delta$, then $(f(x))_j - w_j^{(\lceil \log_2(s)\rceil)} = 0$.

If $j \in I_{n'}\setminus J$ for $\lceil \log_2(s)\rceil - n_\delta + 1 \leq n' \leq \lceil \log_2(s)\rceil$, then

$$\begin{aligned} \left|(f(x))_j - w_j^{(\lceil \log_2(s)\rceil)}\right| &= \left||(Ax)_j|^2 - |(A\tilde{\pi}_{\lceil \log_2(s)\rceil+2}(x))_j|^2\right| \\ &\leq \left|(Ax)_j - (A\tilde{\pi}_{\lceil \log_2(s)\rceil+2}(x))_j\right| \left(|(Ax)_j| + |(A\tilde{\pi}_{\lceil \log_2(s)\rceil+2}(x))_j|\right) \\ &\leq 2^{-\frac{\lceil \log_2(s)\rceil+2}{2}}\sqrt{\gamma}\cdot 3|(Ax)_j| \leq \frac{3}{2}\sqrt{\frac{\gamma}{s}}|(Ax)_j|. \end{aligned}$$

If $j \notin I_{n'}$ for all $\lceil \log_2(r)\rceil + 1 \leq n' \leq \lceil \log_2(s)\rceil$ and $j \notin J$, then by the definition of the $I_{n'}$, $|(A\tilde{\pi}_{\lceil \log_2(s)\rceil+2}(x))_j| < 2^{-\frac{\lceil \log_2(s)\rceil}{2}}\sqrt{\gamma}$ and then

$$\begin{aligned} |(Ax)_j| &\leq |(A\tilde{\pi}_{\lceil \log_2(s)\rceil+2}(x))_j| + |(Ax)_j - (A\tilde{\pi}_{\lceil \log_2(s)\rceil+2}(x))_j| \\ &\leq 2^{-\frac{\lceil \log_2(s)\rceil}{2}}\sqrt{\gamma} + 2^{-\frac{\lceil \log_2(s)\rceil+2}{2}}\sqrt{\gamma} \leq 2\sqrt{\frac{\gamma}{s}}, \end{aligned}$$

such that

$$\left|(f(x))_j - w_j^{(\lceil \log_2(s) \rceil)}\right| = |(f(x))_j| = |(Ax)_j|^2 \leq 2\sqrt{\frac{\gamma}{s}}\,|(Ax)_j|\,.$$

For the remaining entries $j \in J$,

$$\begin{aligned}\sum_{j\in J}\left|(f(x))_j - w_j^{(\lceil \log_2(s) \rceil)}\right|^2 &\leq \sum_{l=r_0+1}^{r}\sum_{j\in J_l}\left|(f(x))_j - w_j^{(\lceil \log_2(s) \rceil)}\right|^2 \leq \sum_{l=r_0+1}^{r}|J_l|(12\gamma_l)^2\\ &\leq 12^2\gamma\sum_{l=r_0+1}|J_l|\gamma_l \leq 12^2\gamma\frac{\delta}{s} \leq 12^2\frac{\gamma}{s}.\end{aligned}$$

So in total,

$$\begin{aligned}d(f(x), w^{(\lceil \log_2(s) \rceil)})^2 &= \sum_{l=1}^{r}\sum_{j\in\tilde{\mathcal{N}}_l}\left|(f(x))_j - w_j^{(\lceil \log_2(s) \rceil)}\right|^2\\ &= \sum_{j\in J}\left|(f(x))_j - w_j^{(\lceil \log_2(s) \rceil)}\right|^2 + \sum_{l=1}^{r}\sum_{j\in\tilde{\mathcal{N}}_l\setminus J}\left|(f(x))_j - w_j^{(\lceil \log_2(s) \rceil)}\right|^2\\ &\leq \frac{\gamma}{s}\left(12^2 + 4\sum_{l=1}^{r}\sum_{j\in\tilde{\mathcal{N}}_l}|(Ax)_j|^2\right) \leq \frac{\gamma}{s}\left(12^2 + 4\|Ax\|_2^2\right). \qquad (4.40)\end{aligned}$$

Recall the definition of the sets $T_n$ for $0 \leq n \leq n_1 + \lceil \log_2(s) \rceil$ given at (4.28) that contain 0 or all the vectors $w^{(n-n_1)}(x)$. We always have $w^{(0)}(x) = 0$. For $x \in D_{\mathbf{s},\mathbf{M}}$, $f(x) \in T^{(1)}$ and

$$0 \in T_0,\, \ldots,\, 0 = w^{(0)} \in T_{n_1},\, w^{(1)} \in T_{n_1+1},\, w^{(2)} \in T_{n_1+2},\, \ldots,\, w^{(\lceil \log_2(s) \rceil)} \in T_{n_1+\lceil \log_2(s) \rceil},$$

we can apply Lemma 4.10 to obtain

$$\begin{aligned}&\sum_{n=0}^{n_1+\lceil \log_2(s) \rceil} 2^{\frac{n}{2}} d(f(x), T_n)\\ &\leq 4\left(2^{\frac{n_1+\lceil \log_2(s) \rceil}{2}} d(f(x), w^{(\lceil \log_2(s) \rceil)}) + \sum_{n=0}^{n_1} 2^{\frac{n}{2}}\cdot 0 + \sum_{n=n_1+1}^{n_1+\lceil \log_2(s) \rceil} 2^{\frac{n}{2}} d(w^{(n-n_1)}, w^{(n-n_1-1)})\right)\\ &\leq 4\cdot 2^{\frac{n_1}{2}}\left(\sqrt{s}d(f(x), w^{(\lceil \log_2(s) \rceil)}) + \sum_{n=1}^{\lceil \log_2(s) \rceil} 2^{\frac{n}{2}} d(w^{(n)}, w^{(n-1)})\right)\end{aligned}$$

Using (4.37), 4.40 and the definition of $n_1$, we can conclude further

$$\begin{aligned}&\sum_{n=0}^{n_1+\lceil \log_2(s) \rceil} 2^{\frac{n}{2}} d(f(x), T_n)\\ &\leq 4\cdot 2^{\frac{n_1}{2}}\left(\sqrt{\gamma\left(12^2 + 4\|Ax\|_2^2\right)} + C_4\sqrt{\gamma\log_2(s)\log_2(\frac{1}{\delta}) + \gamma\left(\log_2(s)\log_2(\frac{1}{\delta}) + r\right)\|Ax\|_2^2}\right)\\ &\leq C_5\cdot 2^{\frac{n_0}{2}}\sqrt{\gamma\log_2(N)\log(\rho)}\cdot\sqrt{\log_2(s)\log_2(\frac{1}{\delta}) + \left(\log_2(s)\log_2(\frac{1}{\delta}) + r\right)\|Ax\|_2^2} \qquad (4.41)\end{aligned}$$

**Step 6: Construction of the admissible sequence for $n > n_1 + \lceil \log_2(s) \rceil$**

To complete the remaining part of the admissible sequence $(T_n)$, we follow a similar approach as in [LA19] for small distances with the difference that instead of using Dudley's inequality, we construct the admissible sequence directly.

For any $x, y \in D_{\mathbf{s},\mathbf{M}}$, $k \in [r]$, $(x-y)_{\mathcal{M}_k}$ is $2s_k$-sparse. Therefore for any $l \in \{r_0+1, \dots, r\}$, $j \in \tilde{\mathcal{N}_l}$,

$$\begin{aligned}
|(A(x-y))_j| &\le \sum_{k=1}^{r} \sum_{k' \in \mathcal{M}_k} |A_{j,k'}||(x-y)_{k'}| \le \sum_{k=1}^{r} \sqrt{\frac{\mu_{l,k}}{p_l}} \|(x-y)_{\mathcal{M}_k}\|_1 \\
&\le \sum_{k=1}^{r} \sqrt{\frac{2\mu_{l,k}s_k}{p_l}} \|(x-y)_{\mathcal{M}_k}\|_2 \le \sqrt{\sum_{k=1}^{r} \frac{2\mu_{l,k}s_k}{p_l}} \cdot \sqrt{\sum_{k=1}^{r} \|(x-y)_{\mathcal{M}_k}\|_2^2} \\
&\le \sqrt{2\gamma_l}\|x-y\|_2 \le \sqrt{2\gamma}\|x-y\|_2
\end{aligned}$$

and then

$$\begin{aligned}
d(|Ax|^2, |Ay|^2) &= \sqrt{\sum_{j \in \tilde{\mathcal{N}}} \left(|(Ax)_j|^2 - |(Ay)_j|^2\right)^2} \le \sqrt{\sum_{j \in \tilde{\mathcal{N}}} |(Ax)_j - (Ay)_j|^2 \left(|(Ax)_j| + |(Ay)_j|\right)^2} \\
&\le \max_{j \in \tilde{\mathcal{N}}} |(A(x-y))_j| \cdot \sqrt{\sum_{j \in \tilde{\mathcal{N}}} \left(|(Ax)_j| + |(Ay)_j|\right)^2} \\
&\le \sqrt{2\gamma}\|x-y\|_2 \cdot [\|Ax\|_2 + \|Ay\|_2] \le 2\sqrt{2\gamma}\|x-y\|_2 \cdot \sqrt{\max_{z \in D_{\mathbf{s},\mathbf{M}}} \|Az\|_2^2} \\
&\le 2\sqrt{2\gamma\beta}\|x-y\|_2 \qquad (4.42)
\end{aligned}$$

$$\beta := \max_{z \in D_{\mathbf{s},\mathbf{M}}} \|Az\|_2^2.$$

For each $S \subset [N]$, define the unit ball with support $S$,

$$B_S := \{x \in \mathbb{C}^N \mid \|x\|_2 = 1 \text{ and } \operatorname{supp}(x) \subset S\}$$

For $|S| = s$, $B_S$ (in $\mathbb{C}$) is isometric to the $\ell_2$ unit sphere in $\mathbb{R}^{2s}$ and by the standard covering number estimates (Lemma 0.13),

$$\mathcal{N}(B_S, \|\cdot\|_2, u) \le \left(1 + \frac{2}{u}\right)^{2s}.$$

Each $x \in D_{\mathbf{s},\mathbf{M}}$ is $s$-sparse and therefore

$$D_{\mathbf{s},\mathbf{M}} \subset \bigcup_{\substack{S \subset [N] \\ |S|=s}} B_S.$$

Since there are $\binom{N}{s}$ choices of $S \subset [N]$, $|S| = s$, for each $u > 0$, there exists a set $\mathcal{T}'(u) \subset \mathbb{C}^N$ of cardinality $|\mathcal{T}'(u)| \le \binom{N}{s}(1 + \frac{2}{u})^{2s}$ such that for each $x \in D_{\mathbf{s},\mathbf{M}}$, there exists a $y \in \mathcal{T}'(u)$ with $\|x - y\|_2 \le u$.

For each integer $n \ge n_1 + \lceil \log_2(s) \rceil + 1$, define

$$\bar{T}_n := \{|Ax|^2 \in \mathbb{R}^m \mid x \in \mathcal{T}'(2^{-(n-n_0)})\}.$$

Then by (4.42) for each $y \in T$, there exists a $y' \in \bar{T}_n$ such that $d(y, y') \le 2^{-(n-n_0)} \cdot 2\sqrt{2\gamma\beta}$. Furthermore

$$|\bar{T}_n| \le \binom{N}{s} \left(1 + 2 \cdot 2^{n-n_0}\right)^{2s} \le \binom{N}{s} \left(2^{n-n_0+2}\right)^{2s} \le \binom{N}{s} \cdot 2^{2s(n-n_0+2)}.$$

Now we define the admissible sequence elements $T_n$ for $n \geq n_1 + \lceil \log_2(s) \rceil + 1$. Take any $y \in \bar{T}_n$ and any tuple $(w^{(n')})_{n'=\lceil \log_2(r)\rceil+1}^{\lceil \log_2(s)\rceil}$ of the vectors defined in (4.27) along with the corresponding sets $(I_{n'})_{n'=\lceil \log_2(r)\rceil+1}^{\lceil \log_2(s)\rceil}$ defined in (4.26). Then we define $z \in \mathbb{R}^m$ by

$$z_j = \begin{cases} w_j^{(\lceil \log_2(s)\rceil)} & \text{if } j \in I_{n'} \text{ for } \lceil \log_2(r)\rceil + 1 \leq n' \leq \lceil \log_2(s)\rceil - n_\delta \\ y_j & \text{otherwise.} \end{cases}$$

Note that this is well defined since the $I_{n'}$ are disjoint.

Define $T_n \subset \mathbb{R}^m$ to be the set of all $z$ obtained in this way. To bound $|T_n|$, note that by their definitions, $(w^{(n')})_{n'=\lceil \log_2(r)\rceil+1}^{\lceil \log_2(s)\rceil}$ and $(I_{n'})_{n'=\lceil \log_2(r)\rceil+1}^{\lceil \log_2(s)\rceil}$ are uniquely determined by

$$\tilde{\pi}_{\lceil \log_2(r)\rceil+3}(x) \in V_{\lceil \log_2(r)\rceil+3}, \dots, \tilde{\pi}_{\lceil \log_2(s)\rceil+2}(x) \in V_{\lceil \log_2(s)\rceil+2}$$

from (4.23). Taking into account the bound (4.21) on the $|V_{n'}|$ and that there are $|\bar{T}_n|$ choices for $y$, we can bound the number of possible $z$ by

$$\begin{aligned} |T_n| &\leq |\bar{T}_n| \cdot \prod_{n'=\lceil \log_2(r)\rceil+3}^{\lceil \log_2(s)\rceil+2} |V_{n'}| \leq \binom{N}{s} \cdot 2^{2s(n-n_0+2)} \cdot \prod_{n'=\lceil \log_2(r)\rceil+3}^{\lceil \log_2(s)\rceil+2} N^{16\lceil c_1 \log \rho\rceil \cdot 2^{n'}} \\ &\leq \binom{N}{s} \cdot 2^{4s(n-n_0)} \cdot N^{16\lceil c_1 \log \rho\rceil \cdot \sum_{n'=\lceil \log_2(r)\rceil+3}^{\lceil \log_2(s)\rceil+2} 2^{n'}} \leq \binom{N}{s} \cdot 2^{4s(n-n_0)} \cdot N^{\tilde{c}_2 s \log(\rho)} \end{aligned}$$

for a constant $\tilde{c}_2 > 0$ and $n \geq n_0 + 2$.

So with the standard bound $\binom{N}{s} \leq \left(\frac{eN}{s}\right)^s$,

$$\begin{aligned} \log_2 |T_n| &\leq s \log_2(\frac{eN}{s}) + 4s(n-n_0) + \tilde{c}_2 s \log_2(N) \log(\rho) \\ &\leq 4s(n-n_0) + \tilde{c}_3 s \log_2(N) \log(\rho) \end{aligned} \tag{4.43}$$

for a constant $\tilde{c}_3 > 0$.

For $n = n_0 + \lceil \log_2 \log_2(N) + \log_2 \log(\rho)\rceil + \lceil \log_2(s)\rceil$, we obtain using that $\log_2(x) \leq x$ for all $x > 0$,

$$\begin{aligned} s(n-n_0) &= s\lceil \log_2 \log_2(N) + \log_2 \log(\rho)\rceil + s\lceil \log_2(s)\rceil \\ &\leq s\lceil \log_2(N) \log(\rho)\rceil + s\lceil \log_2(N)\rceil \\ &\leq 2s\lceil \log_2(N) \log(\rho)\rceil \leq 2 \cdot 2^{n-n_0}. \end{aligned}$$

Since $x \mapsto \frac{2^x}{x}$ is strictly increasing for $x \geq 2$, $s(n-n_0) \leq 2 \cdot 2^{n-n_0}$ also holds for any $n \geq n_0 + \lceil \log_2 \log_2(N) + \log_2 \log(\rho)\rceil + \lceil \log_2(s)\rceil$. Then we obtain with (4.43) for all such $n$,

$$\log_2 |T_n| \leq 8 \cdot 2^{n-n_0} + \tilde{c}_3 \cdot 2^{n-n_0} \leq 2^{-n_0}(8 + \tilde{c}_3) \cdot 2^n \leq 2^n$$

for choosing $n_0 \geq \log_2(8 + \tilde{c}_3)$ (recall that $n_0$ can be chosen as a sufficiently large constant). So for all $n \geq n_0 + \lceil \log_2 \log_2(N) + \log_2 \log(\rho)\rceil + \lceil \log_2(s)\rceil$,

$$|T_n| \leq 2^{2^n}.$$

Again fix an $n \geq n_0 + \lceil \log_2 \log_2(N) + \log_2 \log(\rho)\rceil + \lceil \log_2(s)\rceil$. For each $f(x) \in T^{(1)}$ ($x \in D_{\mathbf{s},\mathbf{M}}$), we consider $y \in \bar{T}_n$ such that $d(y, |Ax|^2) \leq 2^{-(n-n_0)} \cdot 2\sqrt{2\gamma\beta}$ and the vectors $(w^{(n')}(x))_{n'}$ defined in (4.27). Associated to this $y$ and these $(w^{(n')})_{n'}$ there is one element $z \in T_n$. We obtain

$$d(f(x), z)^2 = \sum_{l=r_0+1}^{r} \sum_{j \in \tilde{\mathcal{N}}_l} ((f(x))_j - z_j)^2$$

$$
\begin{aligned}
&\le \sum_{l=r_0+1}^{r} \sum_{j\in\tilde{\mathcal{N}_l}} \left(|(Ax)_j|^2 - y_j\right)^2 + \sum_{n'=\lceil\log_2(r)\rceil+1}^{\lceil\log_2(s)\rceil-n_\delta} \sum_{j\in I_{n'}} (w_j^{(\lceil\log_2(s)\rceil)} - w_j^{(\lceil\log_2(s)\rceil)})^2 \\
&= d(y,|Ax|^2)^2 \le 2^{-2(n-n_0)} \cdot 8\gamma\beta.
\end{aligned}
$$

This yields for a constant $C_5 > 0$,

$$
\begin{aligned}
\sum_{n=n_1+\lceil\log_2(s)\rceil+1}^{\infty} 2^{\frac{n}{2}} d(f(x),T_n) &\le \sum_{n=n_1+\lceil\log_2(s)\rceil+1}^{\infty} 2^{\frac{n}{2}} \cdot 2\sqrt{2\gamma\beta}\cdot 2^{-(n-n_0)} \\
&\le 2\cdot 2^{n_0}\sqrt{2\gamma\beta}\sum_{n=0}^{\infty} 2^{-\frac{n}{2}} \le 2^{n_0}\cdot C_5\sqrt{\gamma\beta}. \qquad (4.44)
\end{aligned}
$$

And together with (4.41),

$$
\sum_{n=0}^{\infty} 2^{\frac{n}{2}} d(f(x),T_n) \le C_6\sqrt{\gamma\log_2(N)\cdot\log(\rho)}\sqrt{\log_2(s)\log_2(\frac{1}{\delta}) + \left(\log_2(s)\log_2(\frac{1}{\delta}) + r\right)\beta} \qquad (4.45)
$$

for a constant $C_6 > 0$ and all $x \in D_{\mathbf{s},\mathbf{M}}$.

**Step 7: Combination of the bounds to control $\mathbb{E}\delta_{\mathbf{s},\mathbf{M}}$**

We have shown that the sets $T_n$ for $n \ge 0$ satisfy the size condition for an admissible sequence but they are not necessarily subsets of $T^{(1)}$. However, by Lemma 4.11, this is still sufficient to prove

$$
\begin{aligned}
\gamma_2(T^{(1)},d) &\le 2\sup_{x\in D_{\mathbf{s},\mathbf{M}}}\sum_{n=0}^{\infty} 2^{\frac{n}{2}} d(f(x),T_n) \\
&\le 2C_6\sqrt{\gamma\log_2(N)\cdot\log(\rho)}\sqrt{\log_2(s)\log_2(\frac{1}{\delta}) + \left(\log_2(s)\log_2(\frac{1}{\delta}) + r\right)\beta} \\
&\le 4C_6\sqrt{\gamma\log_2(N)\cdot\log(\rho)\left(\log_2(s)\log_2(\frac{1}{\delta}) + r\right)\beta} \qquad (4.46)
\end{aligned}
$$

with (4.45) in the second step.

So with the Bernoulli bound (4.11), we can control the expectation of the Bernoulli process

$$
\begin{aligned}
\mathbb{E}_\xi\left\|\left|\sum_{l=r_0+1}^{r}\sum_{j\in\tilde{\mathcal{N}_l}} \xi_j A_j A_j^*\right\|\right\|_{\mathbf{s},\mathbf{M}} &\le C\gamma_2(T^{(1)},d) + \sup_{y\in T^{(2)}}\|y\|_1 \\
&\le C\gamma_2(T^{(1)},d) + \delta\left(3\sup_{x\in D_{\mathbf{s},\mathbf{M}}}\|Ax\|_2^2 + 7\right) \\
&\le C\gamma_2(T^{(1)},d) + \delta\left(3\delta_{\mathbf{s},\mathbf{M}} + 10\right),
\end{aligned}
$$

using (4.39). By forming the expectation on both sides, we obtain with (4.15),

$$
\begin{aligned}
&& \mathbb{E}\delta_{\mathbf{s},\mathbf{M}} &\le 2C\mathbb{E}\gamma_2(T^{(1)},d) + 2\delta(3\mathbb{E}\delta_{\mathbf{s},\mathbf{M}} + 10) \\
\Rightarrow && (1-6\delta)\mathbb{E}\delta_{\mathbf{s},\mathbf{M}} &\le 2C\mathbb{E}\gamma_2(T^{(1)},d) + 20\delta.
\end{aligned}
$$

If $\delta \le \frac{1}{12}$, i.e., $1-6\delta \ge \frac{1}{2}$,

$$
\mathbb{E}\delta_{\mathbf{s},\mathbf{M}} \le 4C\mathbb{E}\gamma_2(T^{(1)},d) + 40\delta. \qquad (4.47)
$$

This implies

$$(\mathbb{E}\delta_{\mathbf{s},\mathbf{M}})^2 \leq 2\left[(4C\mathbb{E}\gamma_2(T^{(1)},d))^2 + (40\delta)^2\right].$$

With (4.46) and $\beta \leq 1 + \delta_{\mathbf{s},\mathbf{M}}$, we obtain that

$$(\mathbb{E}\delta_{\mathbf{s},\mathbf{M}})^2 \leq C_7\left(\gamma\log_2(N)\cdot\log(\rho)\left(\log_2(s)\log_2(\frac{1}{\delta}) + r\right)\right)(1 + \mathbb{E}\delta_{\mathbf{s},\mathbf{M}}) + C_8\delta$$

for constants $C_7, C_8 > 0$.

Defining

$$E := \mathbb{E}\delta_{\mathbf{s},\mathbf{M}} \qquad\qquad D := \sqrt{C_7\gamma\log_2(N)\log(\rho)\left(\log_2(s)\log_2(\frac{1}{\delta})) + r\right)} \geq 0,$$

the above inequality becomes.

$$E^2 \leq D^2(E+1) + C_8\delta,$$

which implies for $E \geq 0$,

$$\mathbb{E}\delta_{\mathbf{s},\mathbf{M}} = E \leq \frac{D^2}{2} + \sqrt{\frac{D^4}{4} + D^2 + C_8\delta} \leq \frac{D^2}{2} + \frac{D^2}{2} + D + C_8\delta = D^2 + D + C_8\delta.$$

If for all $l \in \{r_0+1,\dots,r\}$,

$$m_l \geq C_7\delta^{-2}(N_l - N_{l-1})\left(\sum_{k=1}^{r}\mu_{l,k}s_k\right)\log_2(N)\log(\rho)\left(\log_2(s)\log_2(\frac{1}{\delta})) + r\right), \tag{4.48}$$

then

$$\gamma = \max_{l\in\{r_0+1,\dots,r\}}\sum_{k=1}^{r}\frac{\mu_{l,k}s_k}{m_l/(N_l - N_{l-1})} \leq \delta^2\left(C_7\log_2(N)\log(\rho)\left(\log_2(s)\log_2(\frac{1}{\delta})) + r\right)\right)^{-1}, \tag{4.49}$$

so

$$D \leq \delta$$

and therefore

$$\mathbb{E}\delta_{\mathbf{s},\mathbf{M}} \leq \delta + \delta^2 + C_8\delta \leq (C_8+2)\delta.$$

If (4.48) holds with $\delta$ replaced by $\delta' = \frac{\delta}{2(C_8+2)}$ for a $\delta \in (0, \frac{1}{2}]$, then this only increases the lower bound by at most a constant factor and we obtain

$$\mathbb{E}\delta_{\mathbf{s},\mathbf{M}} \leq (C_8+2)\delta' = \frac{\delta}{2}.$$

The condition $\delta' \leq \frac{1}{12}$ required for (4.47) is also fulfilled since we can assume that $2(C_8+2) \geq 6$.

**Step 8: Concentration of $\delta_{\mathbf{s},\mathbf{M}}$**

For the remaining part of the proof, we need to control the deviation of $\delta_{\mathbf{s},\mathbf{M}}$ from its expectation. To do this, we mostly follow the approach from [LA19] and [FR13] with slight modifications. We include the entire proof for completeness.

For each $l \in \{r_0+1, \dots, r\}$, $j \in \tilde{\mathcal{N}}_l$, define the random vector $Y_j \in \mathbb{C}^{N+1}$ as

$$Y_j = \begin{pmatrix} A_j \\ l \end{pmatrix}.$$

Since all the $A_j$ are independent, also all the $Y_j$ are independent.

Take a countable dense subset $\tilde{D}_{\mathbf{s},\mathbf{M}}$ of $D_{\mathbf{s},\mathbf{M}}$ and for each $x \in \tilde{D}_{\mathbf{s},\mathbf{M}}$ define $F_x : \mathbb{C}^{N+1} \to \mathbb{R}$ such that for all $l \in [r]$ and $v \in \mathbb{C}^{N+1}$,

$$F_x(v) = \begin{cases} |\langle v_{\{1,\dots,N\}}, x\rangle|^2 - \frac{\|(Ux)_{\mathcal{N}_l}\|_2^2}{m_l} & \text{if } v_{N+1} = l \\ 0 & \text{otherwise.} \end{cases}$$

Then for each $l \in \{r_0+1, \dots, r\}$, $j \in \tilde{\mathcal{N}}_l$,

$$F_x(Y_j) = |(Ax)_j|^2 - \frac{\|(Ux)_{\mathcal{N}_l}\|_2^2}{m_l}.$$

Furthermore,

$$\mathbb{E}F_x(Y_j) = \sum_{j' \in \mathcal{N}_l} \frac{1}{|\mathcal{N}_l|} \cdot \frac{1}{p_l} |U_{j'}^* x|^2 - \frac{\|(Ux)_{\mathcal{N}_l}\|_2^2}{m_l} = \frac{\sum_{j' \in \mathcal{N}_l} |(Ux)_{j'}|^2}{m_l} - \frac{\|(Ux)_{\mathcal{N}_l}\|_2^2}{m_l} = 0.$$

The blocks $l = 1, \dots, r_0$ are fully sampled and therefore

$$\begin{aligned} \sum_{l=1}^{r_0} \sum_{j \in \tilde{\mathcal{N}}_l} \left[ |(Ax)_j|^2 - \frac{\|(Ux)_{\mathcal{N}_l}\|_2^2}{m_l} \right] &= \sum_{l=1}^{r_0} \|(Ax)_{\tilde{\mathcal{N}}_l}\|_2^2 - \sum_{l=1}^{r_0} \|(Ux)_{\mathcal{N}_l}\|_2^2 \\ &= \sum_{l=1}^{r_0} \|(Ux)_{\mathcal{N}_l}\|_2^2 - \sum_{l=1}^{r_0} \|(Ux)_{\mathcal{N}_l}\|_2^2 = 0. \end{aligned}$$

such that

$$\begin{aligned} \sum_{j \in \tilde{\mathcal{N}}} F_x(Y_j) &= \sum_{l=r_0+1}^{r} \sum_{j \in \tilde{\mathcal{N}}_l} \left[ |(Ax)_j|^2 - \frac{\|(Ux)_{\mathcal{N}_l}\|_2^2}{m_l} \right] = \sum_{l=1}^{r} \sum_{j \in \tilde{\mathcal{N}}_l} \left[ |(Ax)_j|^2 - \frac{\|(Ux)_{\mathcal{N}_l}\|_2^2}{m_l} \right] \\ &= \sum_{j \in [m]} |(Ax)_j|^2 - \sum_{l=1}^{r} \sum_{j \in \tilde{\mathcal{N}}_l} \frac{\|(Ux)_{\mathcal{N}_l}\|_2^2}{m_l} = \|Ax\|_2^2 - \sum_{l=1}^{r} \|(Ux)_{\mathcal{N}_l}\|_2^2 = \|Ax\|_2^2 - 1. \end{aligned}$$

Now we define

$$\delta_{\mathbf{s},\mathbf{M},+} := \sup_{x \in D_{\mathbf{s},\mathbf{M}}} \left[ \|Ax\|_2^2 - 1 \right] \qquad \delta_{\mathbf{s},\mathbf{M},-} := \sup_{x \in D_{\mathbf{s},\mathbf{M}}} \left[ 1 - \|Ax\|_2^2 \right]$$

Since $\tilde{D}_{\mathbf{s},\mathbf{M}}$ is dense in $D_{\mathbf{s},\mathbf{M}}$ and the supremum is taken over a continuous function, $\sup_{x \in D_{\mathbf{s},\mathbf{M}}}$ can be replaced by $\sup_{x \in \tilde{D}_{\mathbf{s},\mathbf{M}}}$ and then

$$\sup_{F \in \mathcal{F}} \sum_{j \in \tilde{\mathcal{N}}} F(Y_j) = \sup_{x \in \tilde{D}_{\mathbf{s},\mathbf{M}}} \sum_{j \in \tilde{\mathcal{N}}} F_x(Y_j) = \sup_{x \in D_{\mathbf{s},\mathbf{M}}} \left[ \|Ax\|_2^2 - 1 \right] = \delta_{\mathbf{s},\mathbf{M},+}$$

and analogously

$$\sup_{F \in \mathcal{F}} \sum_{j \in \tilde{\mathcal{N}}} -F(Y_j) = \delta_{\mathbf{s},\mathbf{M},-}.$$

Then

$$\delta_{\mathbf{s},\mathbf{M}} = \sup_{x \in D_{\mathbf{s},\mathbf{M}}} \left| \|Ax\|_2^2 - 1 \right| = \max\{\delta_{\mathbf{s},\mathbf{M},+}, \delta_{\mathbf{s},\mathbf{M},-}\}.$$

In (4.30) and (4.31), we have seen that for all $l \in \{r_0+1, \dots, r\}$, $j \in \mathcal{N}_l$,

$$|(Ux)_j| \le \sqrt{\sum_{k=1}^r \mu_{l,k} s_k} \qquad\qquad |(Ax)_j| \le \sqrt{\gamma_l},$$

such that

$$\frac{\|(Ux)_{\mathcal{N}_l}\|_2^2}{m_l} \le \frac{|\mathcal{N}_l| \sum_{k=1}^r \mu_{l,k} s_k}{m_l} = \gamma_l.$$

and therefore

$$|F_x(Y_j)| \le \max_{l \in \{r_0+1,\dots,r\}} (2\gamma_l) = 2\gamma =: K.$$

We also obtain

$$\begin{aligned}\mathbb{E}|F_x(Y_j)|^2 &= \mathbb{E}|(Ax)_j|^4 - \frac{2\|(Ux)_{\mathcal{N}_l}\|_2^2}{m_l}\mathbb{E}|(Ax)_j|^2 + \frac{\|(Ux)_{\mathcal{N}_l}\|_2^4}{m_l^2} \\ &\le \gamma_l \mathbb{E}|(Ax)_j|^2 - \frac{\|(Ux)_{\mathcal{N}_l}\|_2^4}{m_l^2} \le \frac{\gamma_l}{m_l}\|(Ux)_{\mathcal{N}_l}\|_2^2 =: \sigma_l^2\end{aligned}$$

and therefore

$$\sigma^2 := \sum_{l=r_0+1}^r \sum_{j \in \tilde{\mathcal{N}}_l} \sigma_l^2 = \sum_{l=r_0+1}^r \gamma_l \|(Ux)_{\mathcal{N}_l}\|_2^2 \le \gamma \sum_{l=r_0+1}^r \|(Ux)_{\mathcal{N}_l}\|_2^2 \le \gamma.$$

So we can apply Theorem 4.12 which gives us

$$\begin{aligned}\mathbb{P}(\delta_{\mathbf{s},\mathbf{M},+} > \delta) &\le \mathbb{P}(\delta_{\mathbf{s},\mathbf{M},+} > \mathbb{E}\delta_{\mathbf{s},\mathbf{M}} + \frac{\delta}{2}) \le \mathbb{P}(\delta_{\mathbf{s},\mathbf{M},+} > \mathbb{E}\delta_{\mathbf{s},\mathbf{M},+} + \frac{\delta}{2}) \\ &\le \exp\left(-\frac{(\delta/2)^2}{\sigma^2 + 2K\mathbb{E}\delta_{\mathbf{s},\mathbf{M},+} + \frac{\delta}{2}\cdot K/3}\right) \\ &\le \exp\left(-\frac{(\delta/2)^2}{\gamma + 2\gamma\mathbb{E}\delta_{\mathbf{s},\mathbf{M},+} + \frac{\delta}{2}\cdot 2\gamma/3}\right).\end{aligned}$$

Since $\mathbb{E}\delta_{\mathbf{s},\mathbf{M},+} \le \mathbb{E}\delta_{\mathbf{s},\mathbf{M}} \le \delta/2 \le 1/2$, this can be bounded by

$$\exp\left(-\frac{\delta^2}{12\gamma}\right)$$

We can perform the same estimate for $\delta_{\mathbf{s},\mathbf{M},-}$ and conclude

$$\mathbb{P}(\delta_{\mathbf{s},\mathbf{M},-} > \delta) \le \exp\left(-\frac{\delta^2}{12\gamma}\right)$$

With the choice (4.48) of the $m_l$, we can bound $\gamma$ as in (4.49) such that we can conclude

$$\mathbb{P}(\delta_{\mathbf{s},\mathbf{M}} > \delta) \le 2\exp\left(-\frac{\delta^2}{12\gamma}\right).$$

If for all $l \in \{r_0+1, \dots, r\}$,

$$m_l \geq 12\delta^{-2}(N_l - N_{l-1}) \left( \sum_{k=1}^{r} \mu_{l,k} s_k \right) \log\left(\frac{2}{\eta}\right),$$

then $\gamma \leq \delta^2 \left(12 \log(2/\eta)\right)^{-1}$ and

$$\mathbb{P}(\delta_{\mathbf{s},\mathbf{M}} > \delta) \leq 2 \cdot \frac{\eta}{2} = \eta, \tag{4.50}$$

which is what we needed to show.

The prerequisite (4.3) of the theorem ensures that (4.48) (with modified constants due to changing the $\delta$) and also (4.50) are fulfilled. □

### 4.6 Discussion

We have shown that the proof idea of Haviv and Regev [HR16] can be adapted to the sparsity in levels and multilevel sampling scenario from [LA19] and therefore obtained an improvement for this by one $\log s$ factor for constant $\delta$.

Furthermore, by adjusting Maurey's empirical method in the proof to vectors whose sparsity blocks have different $\ell_2$ norms, we were also able to improve the dependence of the result on $r$ in the sense that our result contains a $\log(s) + r$ factor instead of $\log(s) \cdot r$. In this way, if $r \lesssim \log(N)$, like it is the case for the important Fourier and Haar basis, we still obtain a bound with three logarithmic factors that is independent of $r$. The original Fourier/Haar corollary in [LA19] had five logarithmic factors such that our result even improves this by $(\log s)^2$.

Adcock and Li [LA19] conjecture in their work that the factor $r$ in Theorem 4.6 is an artifact of the proof and that one can prove a corresponding bound without the $r$ dependence. We have proven this to be true for the case that $r \lesssim \log(N)$, but it is still open whether it also holds for larger $r$.

The $r$ dependence of our results originates from choosing the elements of the admissible sequence $T_{n_1+1}, \dots, T_{n_1+\lceil \log_2(r) \rceil} = \{0\}$ in the proof. We needed this because our subsequent construction of the sets $T_{n_1+n}$ requires $n \geq \lceil \log_2(r) \rceil$ for the definition of $R_{n,k}$ (after equation (4.18), the number of samples for Maurey's empirical method in block $k$) to be integer. We conjecture that the $r$ dependence can be removed completely by choosing more precise approximations on the part $T_{n_1+1}, \dots, T_{n_1+\lceil \log_2(r) \rceil}$ of the admissible sequence, possibly by applying a version of Maurey's empirical method with a suitable probability distribution.

For the classical subsampled Fourier matrix case, which corresponds to $r = 1$ and $\mu_{1,1} = \frac{1}{N}$, we obtain that the $(s, \delta)$-RIP holds with high probability

$$m \gtrsim \delta^{-2} s \left( \log\left(\frac{1}{\delta}\right) \log(N) \log\left(\frac{s}{\delta}\right) \log(s) \right). \tag{4.51}$$

As explained in step 4 of the proof outline Section 4.4.3, compared to [HR16] and [Bru+21], we can improve one $\log \frac{s}{\delta}$ factor (or $\log(L\frac{s}{\delta})$ if the entries of $U$ are bounded by $|U_{j,k}| \leq \frac{L}{\sqrt{N}}$) to $\log s$. In Theorem 4.8, this means that, unlike it would have been the case for directly adapting the other methods, the $\log s$ factor in the result neither depends on $\delta$ nor on the $\mu_{l,k}$ (like the $\log \tilde{\rho}$ factor does).

An important related open question is how (also for the RIPL, but especially for the classical RIP) the remaining gap between the known lower bound [Bla+19] with two logarithmic factors in $N$ and $s$ and the known guarantees with three logarithmic factors in $N$ and $s$ can be closed, i.e., if the RIP can also be guaranteed with a requirement of the type $m \gtrsim s \log(N) \log(s)$ for a constant $\delta$.

Although our method cannot answer this question, we can gain some further insights by tracking how the vectors in the counterexample of [Bla+19] are controlled in our proof and which step is not optimal. In detail, [Bla+19] is based on the Hadamard transform $H \in \mathbb{R}^{2^n \times 2^n}$ as the Fourier transform on the group (and vector space) $\mathbb{F}_2^n$, where $\mathbb{F}_2$ is the finite field with two elements. Based on subspaces of this finite vector space, they construct a set of $2^{c \log(N/s) \log(s)}$ vectors $x \in \mathbb{R}^N$ such that $x_j = \frac{1}{\sqrt{s}}$ for $s$ indices $j$ and $x_j = 0$ for all other $j$, and such that $(Hx)_j = \sqrt{\frac{s}{N}}$ for $\frac{N}{s}$ entries $j$ and $(Hx)_j = 0$ for all other $j$. For all these vectors we will have $(Ax)_j = \sqrt{\frac{s}{m}}$ for approximately $\frac{m}{s}$ indices $j \in [m]$ and $(Ax)_j = 0$ for all other $j \in [m]$.

Now we consider how these vectors are treated in the proof of Theorem 4.8 for $r = 1$, $r_0 = 0$, $\mu_{1,1} = \frac{1}{N}$. In this case we can check that $\gamma = \frac{s}{m}$ in (4.17). Up to small random changes that do not essentially influence the result, each set $I_n$ in (4.26) is constructed such that it contains the indices $j \in [m]$ for which $(Ax)_j \sim 2^{-\frac{n}{2}} \sqrt{\gamma} = 2^{-\frac{n}{2}} \sqrt{\frac{s}{m}}$. So for the vectors $x$ from the counterexample described above, we would have (up to minor deviations) that $I_1$ contains all the non-zero entries of $x$, i.e., $|I_1| = \frac{m}{s}$ and $I_n = \emptyset$ for all larger $n$.

Now, one can check that the expression controlled in formula (4.34) is later on multiplied by $C\sqrt{\log(N)\log(\rho)} \sim \sqrt{\log(N)\log(s/\delta)}$ and that this is what dominates the bound on $\mathbb{E}\delta_{\mathbf{s},\mathbf{M}}$. For the configuration of the sets $I_n$ described above, we can check that the application of the Cauchy-Schwarz inequality from (4.35) to (4.36) is not sharp and creates an additional $\log_2(s)$ factor that in the end also appears in the final bound (4.51). This Cauchy-Schwarz step would optimal in the case that $\|(Ax)_{I_n}\|_2$ is equal for all $1 \leq n \leq \lceil \log_2(s) \rceil$. One would need to investigate to what extent this is possible at all for bounded orthonormal matrices and sparse vectors, and if and how such vectors and similar ones can be controlled more efficiently in the generic chaining.

# Notation and Abbreviations

| Notation | Description | Pages |
|---|---|---|
| $2^M$ | power set of a set $M$ | 8 |
| $A^\dagger$ | Moore-Penrose pseudoinverse of $A$ | 7 |
| $B_r(x_0)$ | Open ball with radius $r$ and center $x_0$ | 8 |
| $Id_N$ | Identity matrix in $\mathbb{R}^N$ | 7 |
| $N(\mu, \Sigma)$ | Multivariate normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$ | 8 |
| $N(\mu, \sigma^2)$ | Normal distribution with mean $\mu$ and variance $\sigma^2$ | 8 |
| $S^c$ | Complement of the set $S$ | 7 |
| $S^{N-1}$ | $\ell_2$ unit sphere in $\mathbb{C}^N$ (or $\mathbb{R}^N$) | 8 |
| $[N]$ | $\{1, 2, \dots, N\}$ | 7 |
| $\Sigma_s$ | Set of $s$-sparse vectors in $\mathbb{C}^N$ (or $\mathbb{R}^N$) | 8 |
| $\bar{B}_r(x_0)$ | Closed ball with radius $r$ and center $x_0$ | 8 |
| $\gamma_2(T, d)$ | $\gamma_2$ functional | 132 |
| $\lceil \cdot \rceil$ | ceiling function | 8 |
| $\lesssim$, $\gtrsim$, $\sim$ | Bounded from above, below, or both up to constant factor | 8 |
| $\lfloor \cdot \rfloor$ | floor function | 8 |
| $\mathcal{N}(T, d, u)$ | covering number | 16 |
| $\mathrm{aff}(\cdot)$ | affine hull | 109 |
| $\mathrm{cone}(\cdot)$ | conic hull | 109 |
| $\mathrm{conv}(\cdot)$ | convex hull | 109 |
| $\otimes$ | Kronecker product | 8 |
| ReLU, $\phi$ | rectified linear unit | 85 |
| $\sigma_k(A)$ | $k$-th singular value of $A$ | 7 |
| $\sigma_s(x)_1$ | $\inf_{\tilde{x} \in \Sigma_s} \|x - \tilde{x}\|_1$ | 11 |
| $\mathrm{supp}(x)$ | Support of $x$ | 7 |
| $\|\cdot\|_0$ | $\ell_0$ "norm" of vectors, number of non-zero entries | 8 |
| $\|\cdot\|_p$ | $\ell_p$ norm of vectors, $1 \leq p \leq \infty$ | 7 |
| $\|\cdot\|_{L_p}$ | $L_p$ norm of random variable, $1 \leq p \leq \infty$s | 8 |
| $\|\cdot\|_{\psi_1}$ | subexponential norm | 15 |
| $\|\cdot\|_{\psi_2}$ | subgaussian norm (of random variable or vector) | 14, 15 |
| $x_S$ | Restriction of $x$ to the indices in $S$ | 7 |
| $\|S\|$ | Cardinality of the finite set $S$ | 7 |
| BOS | bounded orthonormal system | 125 |
| CPWL | continuous piecewise linear | 110 |
| JLE | Johnson-Lindenstrauss embedding | 13 |
| RIP | restricted isometry property | 10 |

## References

[AC06] Nir Ailon and Bernard Chazelle. "Approximate Nearest Neighbors and the Fast Johnson-Lindenstrauss Transform". In: *STOC '06 Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing* (2006), pp. 557–563.

[Ach01] Dimitris Achlioptas. "Database-friendly Random Projections". In: *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems.* 2001, pp. 274–281.

[Adc+17] Ben Adcock et al. "Breaking the coherence barrier: A new theory for compressed sensing". In: *Forum of Mathematics, Sigma.* Vol. 5. Cambridge University Press. 2017.

[AF09] Jean-Pierre Aubin and Hélène Frankowska. *Set-Valued Analysis.* Modern Birkhäuser Classics. Springer, Boston, 2009. ISBN: 978-0-8176-3478-0.

[AG93] Miguel A. Arcones and Evarist Giné. "On decoupling, series expansions, and tail behavior of chaos processes". In: *Journal of Theoretical Probability* 6.1 (1993), pp. 101–122.

[Ahl+20a] Thomas D. Ahle et al. "Oblivious Sketching of High-Degree Polynomial Kernels". In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms.* SIAM. 2020, pp. 141–160.

[Ahl+20b] Thomas D. Ahle et al. *Oblivious Sketching of High-Degree Polynomial Kernels.* 2020. arXiv: 1909.01410 [cs.DS].

[AHR16] Ben Adcock, Anders C. Hansen, and Bogdan Roman. "A Note on Compressed Sensing of Structured Sparse Wavelet Coefficients From Subsampled Fourier Measurements". In: *IEEE Signal Processing Letters* 23.5 (2016), pp. 732–736.

[AL08] Nir. Ailon and Edo. Liberty. "Fast Dimension Reduction Using Rademacher Series on Dual BCH Codes". In: *SODA '08: Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms* (2008), pp. 1–9.

[AL12] Radosław Adamczak and Rafał Latała. "Tail and moment estimates for chaoses generated by symmetric random variables with logarithmically concave tails". In: *Annales de l'I.H.P. Probabilités et statistiques* 48.4 (2012), pp. 1103–1136.

[AL13] Nir Ailon and Edo Liberty. "An Almost Optimal Unrestricted Fast Johnson-Lindenstrauss Transform". In: *ACM Trans. Algorithms* 9.3 (June 2013). ISSN: 1549-6325.

[ALM21] Radosław Adamczak, Rafał Latała, and Rafał Meller. "Moments of Gaussian chaoses in Banach spaces". In: *Electronic Journal of Probability* 26 (2021), pp. 1–36.

[Alo03] Noga Alon. "Problems and results in extremal combinatorics". In: *Discrete Math* 273 (2003), pp. 31–53.

[ANW14] Haim Avron, Huy Nguyen, and David Woodruff. "Subspace Embeddings for the Polynomial Kernel". In: *Advances in Neural Information Processing Systems.* Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc., 2014.

[Aro+18] Raman Arora et al. "Understanding Deep Neural Networks with Rectified Linear Units". In: *International Conference on Learning Representations.* 2018.

[AV06] Rosa I. Arriaga and Santosh Vempala. "An algorithmic theory of learning: Robust concepts and random projection". In: *Machine Learning* 63.2 (2006), pp. 161–182.

[AW15] Radosław Adamczak and Paweł Wolff. "Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order". In: *Probability Theory and Related Fields* 162.3 (2015), pp. 531–586.

[Ban+13] Afonso S. Bandeira et al. "The Road to Deterministic Matrices with the Restricted Isometry Property". In: *Journal of Fourier Analysis and Applications* 19.6 (2013), pp. 1123–1149.

[Bar+08] Richard Baraniuk et al. "A Simple Proof of the Restricted Isometry Property for Random Matrices". In: *Constructive Approximation* 28.3 (2008), pp. 253–263.

[BBK18] Casey Battaglino, Grey Ballard, and Tamara G. Kolda. "A Practical Randomized CP Tensor Decomposition". In: *SIAM Journal on Matrix Analysis and Applications* 39.2 (2018), pp. 876–901.

[BH17] Alexander Bastounis and Anders C. Hansen. "On the Absence of Uniform Recovery in Many Real-World Applications of Compressed Sensing and the Restricted Isometry Property and Nullspace Property in Levels". In: *SIAM Journal on Imaging Sciences* 10.1 (2017), pp. 335–371.

[BK21] Stefan Bamberger and Felix Krahmer. "Optimal fast Johnson–Lindenstrauss embeddings for large data sets". In: *Sampling Theory, Signal Processing, and Data Analysis* 19.1 (2021), pp. 1–23.

[BKW21a] Stefan Bamberger, Felix Krahmer, and Rachel Ward. *Johnson-Lindenstrauss embeddings with Kronecker structure.* 2021. arXiv: `2106.13349`.

[BKW21b] Stefan Bamberger, Felix Krahmer, and Rachel Ward. *The Hanson-Wright Inequality for Random Tensors.* 2021. arXiv: `2106.13345 [math.PR]`.

[BL14] Witold Bednorz and Rafał Latała. "On the boundedness of Bernoulli processes". In: *Annals of Mathematics* 180.3 (2014), pp. 1167–1203.

[Bla+19] Jaroslaw Blasiok et al. "An Improved Lower Bound for Sparse Reconstruction from Subsampled Hadamard Matrices". In: *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS).* IEEE. 2019, pp. 1564–1567.

[BLM18] Afonso S. Bandeira, Megan E. Lewis, and Dustin G. Mixon. "Discrete uncertainty principles and sparse signal processing". In: *Journal of Fourier Analysis and Applications* 24.4 (2018), pp. 935–956.

[Bou+05] Stéphane Boucheron et al. "Moment inequalities for functions of independent random variables". In: *The Annals of Probability* 33.2 (2005), pp. 514–560.

[Bou+11] Jean Bourgain et al. "Explicit constructions of RIP matrices and related problems". In: *Duke Mathematical Journal* 159.1 (2011), pp. 145–185.

[Bou14] Jean Bourgain. "An Improved Estimate in the Restricted Isometry Problem". In: *Geometric Aspects of Functional Analysis.* Springer, 2014, pp. 65–70.

[Bri+18] Björn Bringmann et al. "The homotopy method revisited: Computing solution paths of $\ell_1$-regularized problems". In: *Mathematics of Computation* 87.313 (2018), pp. 2343–2364.

[Bru+21] Simone Brugiapaglia et al. "Sparse recovery in bounded Riesz systems with applications to numerical methods for PDEs". In: *Applied and Computational Harmonic Analysis* 53 (2021), pp. 231–269.

[BT87] Jean Bourgain and Lior Tzafriri. "Invertibility of 'large' submatrices with applications to the geometry of Banach spaces and harmonic analysis". In: *Israel Journal of Mathematics* 57.2 (1987), pp. 137–224.

[Car85] Bernd Carl. "Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in Banach spaces". In: *Annales de l'institut Fourier.* Vol. 35. 3. 1985, pp. 79–118.

[CCF02] Moses Charikar, Kevin Chen, and Martin Farach-Colton. “Finding Frequent Items in Data Streams”. In: *International Colloquium on Automata, Languages, and Programming.* Springer. 2002, pp. 693–703.

[CCG15] Yuxin Chen, Yuejie Chi, and Andrea J Goldsmith. “Exact and Stable Covariance Estimation From Quadratic Sampling via Convex Programming”. In: *IEEE Transactions on Information Theory* 61.7 (2015), pp. 4034–4059.

[CGV13] Mahdi Cheraghchi, Venkatesan Guruswami, and Ameya Velingker. “Restricted Isometry of Fourier Matrices and List Decodability of Random Linear Codes”. In: *SIAM Journal on Computing* 42.5 (2013), pp. 1888–1914.

[Che+18] Xiaohan Chen et al. “Theoretical Linear Convergence of Unfolded ISTA and Its Practical Weights and Thresholds”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems.* NIPS’18. Curran Associates Inc., 2018, pp. 9079–9089.

[Che+22] Shengyi Chen et al. “Iterative 2D sparse signal reconstruction with masked residual updates for automotive radar interference mitigation”. In: *EURASIP Journal on Advances in Signal Processing* 2022.1 (2022), pp. 1–25.

[Chi+19] Benjamin Chidester et al. “Rotation equivariant and invariant neural networks for microscopy image analysis”. In: *Bioinformatics* 35.14 (July 2019), pp. i530–i537.

[Chk+18] Abdellah Chkifa et al. “Polynomial approximation via compressed sensing of high-dimensional functions on lower sets”. In: *Mathematics of Computation* 87.311 (2018), pp. 1415–1450.

[CJ20] Ke Chen and Ruhui Jin. *Nearly sharp structured sketching for constrained optimization.* 2020. arXiv: `2010.09791`.

[CR07] Emmanuel Candes and Justin Romberg. “Sparsity and incoherence in compressive sampling”. In: *Inverse problems* 23.3 (2007), p. 969.

[CRT06] E J. Candès, J. Romberg, and T Tao. “Stable signal recovery from incomplete and inaccurate measurements”. In: *Comm. Pure Appl. Math.* 59.8 (2006), pp. 1207–1223.

[CT06] Emmanuel J. Candes and Terence Tao. “Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?” In: *IEEE Transactions on Information Theory* 52.12 (2006), pp. 5406–5425.

[Cyb89] G. Cybenko. “Approximation by Superpositions of a Sigmoidal Function”. In: *Mathematics of Control, Signals, and Systems* 2 (1989), pp. 303–314.

[Dau92] Ingrid Daubechies. *Ten Lectures on Wavelets.* SIAM, 1992.

[DeV07] Ronald A DeVore. “Deterministic constructions of compressed sensing matrices”. In: *Journal of Complexity* 23.4-6 (2007), pp. 918–925.

[DG03] Sanjoy Dasgupta and Anupam Gupta. “An elementary proof of a theorem of Johnson and Lindenstrauss”. In: *Random Structures and Algorithms* 22 (2003), pp. 60–65.

[Dir16] Sjoerd Dirksen. “Dimensionality Reduction with Subgaussian Matrices: A Unified Theory”. In: *Foundations of Computational Mathematics* 16.5 (2016), pp. 1367–1396.

[Don06] D. L. Donoho. “Compressed sensing”. In: *IEEE Transactions on Information Theory* 52.4 (2006), pp. 1289–1306.

[Du+19] Simon S. Du et al. “Gradient Descent Provably Optimizes Over-parameterized Neural Networks”. In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.* 2019.

[Dug+09] Charles Dugas et al. “Incorporating Functional Knowledge in Neural Networks.” In: *Journal of Machine Learning Research* 10.6 (2009).

[EK12] Yonina C. Eldar and Gitta Kutyniok. *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.

[FB03] Xiaoli Z Fern and Carla E Brodley. “Random projection for high dimensional data clustering: A cluster ensemble approach”. In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 2003, pp. 186–193.

[Fed96] Herbert Federer. *Geometric Measure Theory*. Classics in Mathematics. Springer, Berlin, Heidelberg, 1996. ISBN: 978-3-642-62010-2.

[FR13] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013. ISBN: 978-0-8176-4947-0.

[GL10] Karol Gregor and Yann LeCun. “Learning Fast Approximations of Sparse Coding”. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML’10. USA: Omnipress, 2010, pp. 399–406. ISBN: 978-1-60558-907-7.

[GMM22] Martin Genzel, Jan Macdonald, and Maximilian Marz. “Solving Inverse Problems With Deep Neural Networks – Robustness Included?” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[Got+20] Nina M. Gottschling et al. *The troublesome kernel: why deep learning for inverse problems is typically unstable*. 2020. arXiv: 2001.01258.

[GSS21] Friedrich Götze, Holger Sambale, and Arthur Sinulis. “Concentration inequalities for polynomials in $\alpha$-sub-exponential random variables”. In: *Electronic Journal of Probability* 26 (2021), pp. 1–22.

[He+20] Juncai He et al. “Relu deep neural networks and linear finite elements”. In: *Journal of Computational Mathematics* 38.3 (2020), pp. 502–527. ISSN: 0254-9409.

[HP97] Shouchuan Hu and Nikolaos S. Papageorgiou. *Handbook of Multivalued Analysis. Volume I: Theory*. Mathematics and Its Applications. Springer US, 1997. ISBN: 978-0-7923-4682-1.

[HPX19] Meng Huang, Yuxuan Pang, and Zhiqiang Xu. “Improved bounds for the RIP of Subsampled Circulant Matrices”. In: *Sampling Theory, Signal Processing, and Data Analysis* 18 (Jan. 2019), pp. 1–8.

[HR16] Ishay Haviv and Oded Regev. “The Restricted Isometry Property of Subsampled Fourier Matrices”. In: *SODA ’16 Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms* (2016), pp. 288–297.

[HS09] Matthew Herman and Thomas Strohmer. “High-Resolution Radar via Compressed Sensing”. In: *IEEE Transactions on Signal Processing* 57.6 (2009), pp. 2275–2284.

[HW71] D. L. Hanson and F. T. Wright. “A Bound on Tail Probabilities for Quadratic Forms in Independent Random Variables”. In: *The Annals of Mathematical Statistics* 42.3 (June 1971), pp. 1079–1083.

[Iwe+21] Mark A Iwen et al. “Lower Memory Oblivious (Tensor) Subspace Embeddings with Fewer Random Bits: Modewise Methods for Least Squares”. In: *SIAM Journal on Matrix Analysis and Applications* 42.1 (2021), pp. 376–416.

[Jin+17] Kyong Hwan Jin et al. “Deep Convolutional Neural Network for Inverse Problems in Imaging”. In: *IEEE Transactions on Image Processing* 26.9 (2017), pp. 4509–4522.

[JKW20] Ruhui Jin, Tamara G Kolda, and Rachel Ward. "Faster Johnson–Lindenstrauss transforms via Kronecker products". In: *Information and Inference: A Journal of the IMA* (Oct. 2020). iaaa028. ISSN: 2049-8772.

[JL84] W.B. Johnson and J. Lindenstrauss. "Extensions of Lipschitz Mappings into a Hilbert Space". In: *Contemporary Mathematics* 26 (1984), pp. 189–206.

[KL15] Konrad Kolesko and Rafał Latała. "Moment estimates for chaoses generated by symmetric random variables with logarithmically convex tails". In: *Statistics & Probability Letters* 107 (2015), pp. 210–214.

[KSO21] Piotr Kicki, Piotr Skrzypczyński, and Mete Ozay. "A New Approach to Design Symmetry Invariant Neural Networks". In: *2021 International Joint Conference on Neural Networks (IJCNN).* 2021, pp. 1–8.

[Kun79] Henry O. Kunz. "On the equivalence between one-dimensional discrete Walsh-Hadamard and multidimensional discrete Fourier transforms". In: *IEEE Transactions on Computers* 28.03 (1979), pp. 267–268.

[KW11] Felix Krahmer and Rachel Ward. "New and Improved Johnson-Lindenstrauss Embeddings via the Restricted Isometry Property". In: *SIAM Journal on Mathematical Analysis* 43.3 (2011), pp. 1269–1281.

[KW13] Felix Krahmer and Rachel Ward. "Stable and Robust Sampling Strategies for Compressive Imaging". In: *IEEE Transactions on Image Processing* 23.2 (2013), pp. 612–622.

[KW92] Stanislaw Kwapien and Wojbar A. Woyczynski. *Random Series and Stochastic Integrals: Single and Multiple.* Boston: Birkhäuser, 1992.

[Kwa87] Stanislaw Kwapien. "Decoupling Inequalities for Polynomial Chaos". In: *The Annals of Probability* 15.3 (1987), pp. 1062–1071.

[LA19] Chen Li and Ben Adcock. "Compressed sensing with local structure: uniform recovery guarantees for the sparsity in levels class". In: *Applied and Computational Harmonic Analysis* 46.3 (2019), pp. 453–477.

[Lat06] Rafał Latała. "Estimates of Moments and Tails of Gaussian Chaoses". In: *The Annals of Probability* 34.6 (2006), pp. 2315–2331. ISSN: 00911798.

[LDP07] Michael Lustig, David Donoho, and John M Pauly. "Sparse MRI: The application of compressed sensing for rapid MR imaging". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 58.6 (2007), pp. 1182–1195.

[Les+93] Moshe Leshno et al. "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function". In: *Neural Networks* 6.6 (1993), pp. 861–867. ISSN: 0893-6080.

[LN17] Kasper Green Larsen and Jelani Nelson. "Optimality of the Johnson-Lindenstrauss lemma". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS).* IEEE. 2017, pp. 633–638.

[Lu+17] Zhou Lu et al. "The Expressive Power of Neural Networks: A View from the Width". In: *Advances in Neural Information Processing Systems* 30 (2017).

[Lus+08] Michael Lustig et al. "Compressed sensing MRI". In: *IEEE signal processing magazine* 25.2 (2008), pp. 72–82.

[MB20] Osman Asif Malik and Stephen Becker. "Guarantees for the Kronecker fast Johnson-Lindenstrauss transform using a coherence and sampling argument". In: *Linear Algebra and its Applications* 602 (2020), pp. 120–137.

[Mel16] Rafał Meller. "Two-sided moment estimates for a class of nonnegative chaoses". In: *Statistics & Probability Letters* 119 (2016), pp. 213–219.

[Mel19] Rafał Meller. "Tail and moment estimates for a class of random chaoses of order two". In: *Studia Mathematica* 249 (2019), pp. 1–32.

[MO34] Stanisław Mazur and Władysław Orlicz. "Grundlegende Eigenschaften der polynomischen Operationen. Erste Mitteilung". In: *Studia Mathematica* 1.5 (1934), pp. 50–68.

[MT86] Terry R. McConnell and Murad S. Taqqu. "Decoupling Inequalities for Multilinear Forms in Independent Symmetric Random Variables". In: *The Annals of Probability* 14.3 (1986), pp. 943–954.

[Nat95] Balas Kausik Natarajan. "Sparse Approximate Solutions to Linear Systems". In: *SIAM Journal on Computing* 24.2 (1995), pp. 227–234.

[NT09] Deanna Needell and Joel A Tropp. "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples". In: *Applied and Computational Harmonic Analysis* 26.3 (2009), pp. 301–321.

[OSL00] Bruno A Olshausen, Phil Sallee, and Michael S Lewicki. "Learning sparse wavelet codes for natural images". In: *Wavelet Applications in Signal and Image Processing VIII*. Vol. 4119. International Society for Optics and Photonics. 2000, pp. 200–207.

[Ovc02] Sergei Ovchinnikov. "Max-Min Representation of Piecewise Linear Functions". In: *Contributions to Algebra and Geometry* 43.1 (2002), pp. 297–302.

[Peñ92] Victor de la Peña. "Decoupling and Khintchine's Inequalities for $U$-Statistics". In: *The Annals of Probability* 20.4 (1992), pp. 1877–1892.

[PG99] Victor de la Peña and Evarist Giné. *Decoupling. From Dependence to Independence.* Probability and its Applications. Springer, New York, NY, 1999.

[Pin99] Allan Pinkus. "Approximation theory of the MLP model in neural networks". In: *Acta Numerica* 8 (1999), pp. 143–195.

[PM95] Victor de la Peña and Stephen J Montgomery-Smith. "Decoupling inequalities for the tail probabilities of multivariate U-statistics". In: *The Annals of Probability* (1995), pp. 806–816.

[PP13] Ninh Pham and Rasmus Pagh. "Fast and scalable polynomial kernels via explicit feature maps". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013, pp. 239–247.

[Rau10] Holger Rauhut. "Compressive Sensing and Structured Random Matrices". In: *Theoretical Foundations and Numerical Methods for Sparse Recovery*. Ed. by Massimo Fornasier. De Gruyter, 2010, pp. 1–92. DOI: `doi:10.1515/9783110226157.1`. URL: `https://doi.org/10.1515/9783110226157.1`.

[RKH10] Kamisetty Ramamohan Rao, Do Nyeon Kim, and Jae Jeong Hwang. *Fast Fourier Transform - Algorithms and Applications*. Springer, 2010. ISBN: 978-1-4020-6628-3.

[RV08] Mark Rudelson and Roman Vershynin. "On sparse reconstruction from Fourier and Gaussian measurements". In: *Communications on Pure and Applied Mathematics* 61 (2008), pp. 1025–1045.

[RV13] Mark Rudelson and Roman Vershynin. "Hanson-Wright inequality and sub-gaussian concentration". In: *Electronic Communications in Probability* 18 (2013), 9 pp.

[Sar06] Tamas Sarlos. "Improved approximation algorithms for large matrices via random projections". In: *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. IEEE. 2006, pp. 143–152.

[SK18] K Sreekala and E Krishna Kumar. "Compressed Sensing in Imaging and Reconstruction – An Insight Review". In: *International Conference on Intelligent Systems Design and Applications.* Springer. 2018, pp. 779–791.

[ST20] Karin Schnass and Flavio Teixeira. "Compressed Dictionary Learning". In: *Journal of Fourier Analysis and Applications* 26 (Mar. 2020).

[Sun+21] Yiming Sun et al. *Tensor random projection for low memory dimension reduction.* 2021. arXiv: `2105.00105`.

[Tal14] Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes.* Vol. 60. Springer, 2014.

[Tan+20] Hao Tang et al. "Towards scale-invariant graph-related problem solving by iterative homogeneous gnns". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 15811–15822.

[Van14] Ramon Van Handel. *Probability in High Dimension.* Tech. rep. Princeton University, 2014.

[Ver12] Roman Vershynin. "Introduction to the non-asymptotic analysis of random matrices". In: *Compressed Sensing: Theory and Applications.* Ed. by Yonina C. Eldar and Gitta Kutyniok. Cambridge: Cambridge University Press, 2012, pp. 210–268.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 2018.

[Ver20] Roman Vershynin. "Concentration inequalities for random tensors". In: *Bernoulli* 26.4 (2020), pp. 3139–3162.

[Vyb11] Jan Vybíral. "A variant of the Johnson-Lindenstrauss lemma for circulant matrices". In: *Journal of Functional Analysis* 260.4 (2011), pp. 1096–1105.

[Woj10] P Wojtaszczyk. "Stability and instance optimality for Gaussian measurements in compressed sensing". In: *Foundations of Computational Mathematics* 10.1 (2010), pp. 1–13.

[Woo14] David P. Woodruff. "Sketching as a Tool for Numerical Linear Algebra". In: *Foundations and Trends® in Theoretical Computer Science* 10.1–2 (2014), pp. 1–157. ISSN: 1551-305X.

[WS96] Jeffrey Wood and John Shawe-Taylor. "Representation theory and invariant neural networks". In: *Discrete Applied Mathematics* 69.1-2 (1996), pp. 33–60.

[XXC12] Junyuan Xie, Linli Xu, and Enhong Chen. "Image denoising and inpainting with deep neural networks". In: *Advances in Neural Information Processing Systems* 25 (2012).

[Zbo+18] Jure Zbontar et al. *fastMRI: An Open Dataset and Benchmarks for Accelerated MRI.* 2018. arXiv: `1811.08839 [physics, stat]`.

[Zha+17] K. Zhang et al. "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising". In: *IEEE Transactions on Image Processing* 26.7 (2017), pp. 3142–3155.

[Zie12] Günter M Ziegler. *Lectures on Polytopes.* Vol. 152. Springer Science & Business Media, 2012.

[ZYY16] Hui Zhang, Ming Yan, and Wotao Yin. "One condition for solution uniqueness and robustness of both l1-synthesis and l1-analysis minimizations". In: *Advances in Computational Mathematics* 42.6 (2016), pp. 1381–1399.