# Low-rank Tensor Recovery

**Dissertation**

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Željka Stojanac

aus

Zagreb, Kroatien

Bonn, 2016

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der
Rheinischen Friedrich-Wilhelms Universität Bonn

1. Gutachter: Prof. Dr. Holger Rauhut
2. Gutachter: Prof. Dr. André Uschmajew

Tag der Promotion: 01.09.2016

Erscheinungsjahr: 2016

# Abstract

It has been noticed that many real world signals are sparse or compressible (approximately sparse) in a certain basis or frame. This observation has led to the theory of compressive sensing which enables the recovery of sparse vectors from a number of measurements smaller than the vector length via efficient algorithms. Since solving the affine sparse minimization problem

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 = \#\{i : \mathbf{x}(i) \neq 0\} \leq s \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x},$$

is in general NP-hard, tractable alternatives have been suggested and analyzed. Recovery algorithms include optimization methods such as *$\ell_1$-minimization* (also known as basis pursuit) and quadratically constrained $\ell_1$-minimization (also called basis pursuit denoising), greedy methods such as orthogonal matching pursuit (OMP) and compressive sampling matching pursuit (CoSaMP), and thresholding-based methods which include iterative hard thresholding algorithm (IHT) and fast iterative shrinkage-thresholding algorithm (FISTA). The concept of compressive sensing extends to recovery of two-dimensional signals where the aim is to reconstruct a low-rank matrix from an underdetermined linear system. This extension is not just theoretically interesting, but has also been applied to various engineering problems where the signal is of low-rank. For example, this problem arises in quantum state tomography, computer vision, collaborative filtering, and matrix completion (where a task is to recover a low-rank matrix from few known entries). Since the affine rank minimization problem is NP-hard to solve, the idea of $\ell_1$-minimization for sparse vector recovery has been extended to *nuclear norm minimization* for low-rank matrix recovery (where the nuclear norm of a matrix is the sum of its singular values). Several other algorithms have been adapted to the matrix scenario including iterative hard thresholding algorithm (IHT), iteratively reweighted least squares minimization (IRLS), and Atomic Decomposition for Minimium Rank Approximation (ADMiRA) which extends CoSaMP for sparse vector recovery. Additionally, other approaches and algorithms have been suggested including the alternating projections algorithm and approaches developed particularly for matrix completion involving the left and right singular vectors along the Grassman manifold.

Both sparse vector and low-rank matrix recovery have been extensively investigated. Several efficient algorithms – we emphasize that the above lists are highly incomplete – have been provided together with theoretical guarantees on the convergence of the algorithms and the (often optimal, or optimal up to log factors) bounds on the number of measurements required for successful signal recovery.

However, in many applications such as machine learning, video compression, and seismology, the signals of interest are tensors. In particular, in seismology a signal is a five dimensional object with two spatial coordinates, two receiver coordinates, and a time coordinate, see for example [38]. In this thesis, we consider a further extension of compressive sensing to low-rank tensor recovery.

The aim is to reconstruct a $d$th order low-rank tensor from a number of linear measurements much smaller than the ambient dimension of the tensor. Several approaches to low-rank tensor recovery have been suggested to this end. However, presently there is no completely satisfactory theory available for these methods. That is, the method is either not tractable, or the recovery results quantifying the minimal number of measurements are non-optimal or even non-existent. This is due to the several difficulties that arise when passing from matrices to higher order tensors. We describe these difficulties in detail below.

We first discuss several tensor decompositions and corresponding notions of tensor rank that have been introduced in the literature. The decomposition which arises naturally when passing from two-dimensional to $d$-dimensional objects is the *canonical decomposition* (or *CP-decomposition*). One defines the CP-rank of an order-$d$ tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ – similarly to the matrix scenario – as the minimal number of rank one tensors (which are outer products of $d$ vectors of appropriate dimensions) that sum up to the original tensor $\mathbf{X}$. Additionally, one can define a tensor nuclear norm as the analog of the matrix nuclear norm, i.e., for $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$

$$\|\mathbf{X}\|_* = \min \left\{ \sum_{k=1}^{r} |c_k| : \mathbf{X} = \sum_{k=1}^{r} c_k \, \mathbf{u}_k^1 \otimes \mathbf{u}_k^2 \otimes \cdots \otimes \mathbf{u}_k^d, \; r \in \mathbb{N}, \right.$$
$$\left. \left\|\mathbf{u}_k^i\right\|_2 = 1, \text{ for all } i \in \{1, 2, \ldots, d\}, k \in \{1, 2, \ldots, r\} \right\}.$$

Unfortunately, the set or rank-$r$ tensors is not closed (for $r > 1$) and thus computing the tensor rank, the canonical decomposition, and the nuclear norm of a tensor is in general NP-hard. In particular, in contrast to the matrix case, the affine tensor nuclear norm minimization is NP-hard to solve and therefore one needs to develop different approaches to low-rank tensor recovery.

The Tucker decomposition, the tensor train (TT) decomposition, and the hierarchical Tucker (HT) decomposition are tensor decompositions which, in contrast to the CP decomposition, can be computed efficiently via sequential singular value decompositions.

The *Tucker decomposition* and its normalized version, called the *higher-order singular value decomposition (HOSVD)*, have been applied for example in chemical analysis, psychometrics, signal processing, and computer vision. Here a $d$-th order tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ is represented as

$$\mathbf{X}(i_1, i_2, \ldots, i_d) = \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \cdots \sum_{k_d=1}^{r_d} \mathbf{C}(k_1, k_2, \ldots, k_d) \mathbf{U}_1(i_1, k_1) \mathbf{U}_2(i_2, k_2) \cdots \mathbf{U}_d(i_d, k_d),$$

where $\mathbf{C} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}$ is a $d$-th order tensor and $\mathbf{U}_i \in \mathbb{R}^{n_i \times r_i}$, with $r_i < n_i$ for all $i \in \{1, 2, \ldots, d\}$ are matrices. Recall that the rank of a matrix equals the number of its independent columns (or rows). For a $d$th order tensor, we call a vector obtained by fixing all the entries of the tensor except the $k$-th one, *the mode-k fiber.* (For $d = 2$, the mode-1 fibers are columns and the mode-2 fibers are rows of the matrix.) However, in the tensor scenario, the number of independent mode-$k$ fibers, denoted by $r_k$, is in general different for different $k \in \{1, 2, \ldots, d\}$. Thus, the Tucker-rank (or HOSVD rank) of an order-$d$ tensor is a $d$-dimensional vector $\mathbf{r} = (r_1, r_2, \ldots, r_d)$. A downside of this decomposition is its storage complexity $\mathcal{O}\left(ndr + r^d\right)$ where $n = \max\{n_i : i \in \{1, 2, \ldots, d\}\}$ and $r = \max\{r_i : i \in \{1, 2, \ldots, d\}\}$, i.e., it suffers from the *curse of dimensionality* (the exponential dependence in $d$). Thus, without further sparsity of the core tensor, the Tucker format is only useful for low order tensors (i.e., for tensors of order three).

A recently introduced decomposition that can be considered as a compromise between the canonical and the Tucker format is the *Hierarchical Tucker decomposition (HT-decomposition).* An HT-decomposition is induced by a binary dimension tree $T_I$, which can be described as follows. The root of a $T_I$ is a set $t_D = \{1, 2, \ldots, d\}$ and every non-leaf node $t$ has exactly two sons – the left son $t_1$ and the right son $t_2$ – satisfying $t = t_1 \dot\cup t_2$. Additionally, for all $s_1 \in t_1$, $s_2 \in t_2$ it holds that $s_1 < s_2$. A special case of the HT-decomposition, where every non-leaf node of the binary dimension tree has a left son which is a leaf, is the *Tensor train decomposition (TT-decomposition).* Both TT and HT-decompositions are computable and do not suffer from the curse of dimensionality. In particular, the TT-decomposition of an order-$d$ tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ consists of $d-1$ order-3 tensors $\mathbf{G}_i \in \mathbb{R}^{r_{i-1} \times n_i \times r_i}$ with $i \in \{1, 2, \ldots, d-1\}$ and $r_0 = r_d := 1$, or element-wise

$$\mathbf{X}(i_1, i_2, \ldots, i_d) = \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \cdots \sum_{k_{d-1}=1}^{r_{d-1}} \mathbf{G}_1(i_1, k_1) \mathbf{G}_2(k_1, i_2, k_2) \cdots \mathbf{G}_d(k_{d-1}, i_d).$$

In quantum physics the TT-decomposition is known under the name *matrix product states (MPS).* The TT-rank of an order-$d$ tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ is a $(d-1)$-dimensional vector $\mathbf{r} = (r_1, r_2, \ldots, r_{d-1})$, where $r_i$ is a rank of the specific tensor matricization. To be precise, $r_i = \operatorname{rank}\left(\mathbf{X}^{\{1,2,\ldots,i\}}\right)$, where the first $i$ indexes enumerate the rows and the last $d-i$ indexes enumerate the columns of the matrix $\mathbf{X}^{\{1,2,\ldots,i\}}$. The general HT-decomposition is slightly more complicated and considers different tensor matricizations. However, often it is assumed that the corresponding dimension binary tree is a *balanced tree.* That is, each non-leaf mode $t$ has two sons $t_1$ and $t_2$ of almost the same size. In particular, without loss of generality, one assumes that $|t_1| = \lceil \frac{|t|}{2} \rceil$ and $|t_2| = \lfloor \frac{|t|}{2} \rfloor$. The corresponding HT-rank is the set of $2d-2$ ranks $r_i$ of the corresponding matricizations. Let $r$ be the largest rank in this set (i.e., $r = \max\{r_i : i \in \{1, 2, \ldots, 2d-2\}\}$) and let $n := \max\{n_i : i \in \{1, 2, \ldots, d\}\}$. For an order-$d$ tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ , the overall complexity for storing the required data is $\mathcal{O}\left(ndr + dr^3\right)$ for an HT-decomposition with a balanced tree, as opposed to $\mathcal{O}\left(dnr^2\right)$ for the TT-decomposition.

Recall that the canonical tensor decomposition would be a natural generalization of the matrix singular value decomposition. However, it is in general NP-hard to compute, and therefore other tensor decompositions have been developed. Although Tucker, TT, and HT-decompositions can be computed efficiently (via sequential singular value decompositions) and the corresponding tensor ranks are well-defined quantities, they suffer from several disadvantages that cause problems in analyses of the algorithms for low-rank tensor recovery. For example, it is unknown how to obtain the best rank-$\mathbf{r}$ approximation of a given tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$. However, it is possible to compute a quasi-best rank-$\mathbf{r}$ approximation $\mathcal{H}_{\mathbf{r}}(\mathbf{X})$ satisfying

$$\|\mathbf{X} - \mathcal{H}_{\mathbf{r}}(\mathbf{X})\|_F \leq \mathcal{O}(\sqrt{d}) \inf\left\{\|\mathbf{X} - \mathbf{Y}\|_F : \operatorname{rank}(\mathbf{Y}) \leq \mathbf{r}\right\}. \tag{0.1}$$

This approximation is obtained via truncation procedures.

In the analysis of the algorithms for sparse vector recovery and low-rank matrix recovery, it is often used (directly or implicitly) that we know how to obtain the best $s$-sparse approximation and the best rank-$r$ approximation of a given vector or a matrix, respectively. Additionally, it is often used that a $2s$-sparse vector and a $2r$-rank matrix can be represented as a sum of two mutually orthogonal $s$-sparse vectors and two mutually orthogonal rank-$r$ matrices, respectively.

So far, the best known result for tensors states that an HOSVD-rank $2\mathbf{r}$ tensor can be represented as a sum of $2^d$ pairwise orthogonal rank-$\mathbf{r}$ tensors. For TT and HT decompositions – to the best of our knowledge – an analogous result is unavailable. Recall that a computable complete analog of the matrix singular value decomposition for tensors does not exist and we can only obtain a quasi-best rank-$\mathbf{r}$ approximation of a given tensor. These tensor peculiarities cause problems in the analysis of algorithms for low-rank tensor recovery. They also give an insight into why there are no completely satisfactory results available for low-rank tensor recovery to this end and why there is a need for new methods when passing from matrix to tensor scenario.

Two new approaches to low-rank tensor recovery are presented in this thesis. The first approach is a convex optimization approach that could be considered as a tractable higher-order generalization of $\ell_1$-minimization for sparse vector recovery and nuclear norm minimization for low-rank matrix recovery. It is based on *theta bodies*, a recent tool developed in real algebraic geometry. It requires computing the reduced Gröbner basis with respect to the graded reverse lexicographic (grevlex) ordering of the polynomial ideal $\mathcal{J}_d$ in $\mathbb{R}\left[x_{11\ldots1}, x_{11\ldots2}, \ldots, x_{n_1n_2\ldots n_d}\right]$ whose real algebraic variety (the set of points where the ideal vanishes) $\nu_{\mathbb{R}}\left(\mathcal{J}_d\right)$ consists of all rank-one Frobenius-norm-one tensors in $\mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$. Here, each variable represents a tensor entry. We consider the canonical tensor format and the corresponding tensor nuclear norm which are in general NP-hard to compute, as already mentioned. Notice that the convex hull of $\nu_{\mathbb{R}}\left(\mathcal{J}_d\right)$ corresponds to the unit-tensor-nuclear-norm ball. Theta bodies provide sum-of-squares hierarchical relaxations of this convex set. The $\theta_k$-norms are defined via their unit balls and they are nested, i.e., they satisfy $\|\mathbf{X}\|_* \geq \cdots \geq \|\mathbf{X}\|_{\theta_k} \geq \|\mathbf{X}\|_{\theta_{k-1}} \geq \cdots \geq \|\mathbf{X}\|_{\theta_1}$, for all $\mathbf{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$. These norms can be computed via semidefinite programming. First we compute the reduced Gröbner basis with respect to the grevlex ordering of the ideal $\mathcal{J}_d$ – and in particular of $\mathcal{J}_3$. Then the semidefinite program for computing the tensor $\theta_1$-norm of an order-3 tensor as well as the semidefinite program for a low-rank third order tensor recovery via $\theta_1$-norm minimization are provided explicitly. We perform numerical experiments for third-order tensor recovery with Gaussian measurement ensembles via $\theta_1$-norm minimization. In our experiments, rank-one and rank-two tensors could always be recovered from a number of measurements significantly smaller than the ambient dimension of the corresponding tensor. Thus, our new theta-body approach seems to be very promising. In future, we would like to provide the theoretical guarantees for low-rank tensor recovery via $\theta_k$-norm minimization.

The theta body method can also be applied to the vector and the matrix scenario. That is, the corresponding unit $\theta_k$-norm balls form a hierarchical set of relaxations of the unit $\ell_1$-norm ball (in the vector scenario) and of the unit matrix nuclear norm ball (in the matrix scenario). However, in these cases, the method does not lead to new vector and matrix norms, respectively. In particular, we show that for a vector $\mathbf{x} \in \mathbb{R}^n$ and the corresponding $\theta_k$-norms denoted by $\|\cdot\|_{\theta_k,v}$ it holds that $\|\mathbf{x}\|_1 = \|\mathbf{x}\|_{\theta_k,v}$, for all $k$. Similarly, for a matrix $\mathbf{X} \in \mathbb{R}^{n_1\times n_2}$ and the corresponding $\theta_k$-norms denoted by $\|\cdot\|_{\theta_k,m}$ it holds that $\|\mathbf{X}\|_* = \|\mathbf{X}\|_{\theta_k,m}$, for all $k$.

In the above approach we focused on the polynomial ideals $\mathcal{J}_d$ whose real algebraic variety $\nu_{\mathbb{R}}(\mathcal{J}_d)$ consists of all order-$d$ rank-one Frobenius-norm-one tensors. Omitting the last condition (i.e., the tensor normalization), leads to ideals $\mathcal{I}_{2,d}$ that – to the best of our knowledge – have not been considered before. These ideals can be seen as the natural generalization of the *determinantal ideal* $\mathcal{I}_2$ to tensors. A determinantal ideal $\mathcal{I}_t$ is a polynomial ideal generated by all order-$t$ minors of the matrix of indeterminates (matrix of unknowns). Equivalently, the real algebraic variety of determinantal ideal $\mathcal{I}_t$ contains all rank-$(t-1)$ matrices. Determinantal ideals and objects related to them have been widely studied in real algebraic geometry and commutative algebra for the last three decades. Recall that we have computed the reduced Gröbner basis $\boldsymbol{\mathcal{G}}_d$ of the polynomial ideal $\mathcal{J}_d$ with respect to the grevlex ordering. Directly from this result, we obtain the reduced Gröbner basis of the polynomial ideal $\mathcal{I}_{2,d}$. To be precise, the leading term of the polynomial $g_d = x_{11\ldots1}^2 + x_{11\ldots2}^2 + \cdots + x_{n_1n_2\cdots n_d}^2 - 1 \in \boldsymbol{\mathcal{G}}_d$ which promotes the Frobenius-norm-one constraint is relatively prime with the leading term of every other polynomial in $\boldsymbol{\mathcal{G}}_d \backslash \{g_d\}$. Consequently, the set $\boldsymbol{\mathcal{G}}_d \backslash \{g_d\}$ is the reduced Gröbner basis with respect to the grevlex ordering of the higher-order determinantal ideal $\mathcal{I}_{2,d}$. In other words, our computation of the reduced Gröbner basis of polynomial ideal $\mathcal{J}_d$ could be considered as the first result on *higher-order determinantal ideals* $\mathcal{I}_{t,d}$ (whose real algebraic variety contains all rank-$(t-1)$ order-$d$ tensors), where $d \geq 3$.

Our second approach to low-rank tensor recovery is a generalization of the iterative hard thresholding algorithm (IHT) for sparse vectors and low-rank matrices to the tensor setting. We consider recovery of tensors which are of low-rank in the HOSVD, the TT, and more generally the HT format. The analyses of these algorithms are based on an appropriate notion of the restricted isometry property for tensors (tensor RIP or TRIP). We show that subgaussian measurement ensembles satisfy TRIP with high probability under an (almost) optimal condition on the number of measurements. This includes Gaussian and Bernoulli measurement ensembles. Additionally, we show that partial Fourier maps combined with random sign flips of the tensor entries also satisfy TRIP with high probability. The crucial step in IHT algorithms consists in computing the projection of the current iterate onto the manifold of low-rank matrices/tensors or space of sparse vectors. In contrast to the vector/matrix case, it is NP-hard to compute the projection (best approximation) exactly – regardless of the choice of tensor format. Thus, we compute its quasi-best approximation $\mathcal{H}_{\mathbf{r}}(\mathbf{X})$ introduced in (0.1) by a truncation procedure. Due to the tensor peculiarities discussed above, we obtain a partial convergence result with an additional assumption on the tensor truncation operator. To illustrate our theoretical results, we perform numerical experiments for recovery of randomly generated low HOSVD-rank third order tensors via the classical tensor IHT algorithm and the normalized tensor IHT algorithm. In our experiments, we consider Gaussian maps, tensor completion, and partial Fourier maps combined with random sign flips of the tensor entries. Our numerical results indicate that the tensor IHT algorithm in practice performs better than our theory can currently guarantee.

# Acknowledgements

I would like to express my deepest gratitude to my advisor, Prof. Dr. Holger Rauhut, for giving me the opportunity to do research under his supervision. His support, understanding, advices, and fruitful discussions – both career-wise and life-wise – throughout my PhD has been truly invaluable.

I would also like to thank the whole group at RWTH Aachen University – namely Maryia Kabanava, Sjoerd Dirksen, Ulrich Terstiege, Jean-Luc Bouchot, Alexander Stollenwerk, Hans Christian Jung, and Jonathan Fell – as well as my former colleagues Tino Ullrich, Felix Krahmer, Max Hügel, and Ulaş Ayaz for the inspiration, support, and friendly atmosphere. I always appreciated and cherished the time we have spent together in working environment and especially in a purely social setting.

My sincere appreciation is extended to Prof. Dr. Bernd Sturmfels, Dr. Daniel Plaumann, Prof. Dr. Shmuel Friedland, Prof. Dr. Aldo Conca, James Saunderson, and Hamza Fawzi for helpful discussions and useful remarks on the algebraic geometry side of my research. I would also like to thank Prof. Dr. Reinhold Schneider for the work, support, and patience he showed during our collaboration. Additionally, I would like to thank Sjoerd Dirksen, Claudio Cacciapuoti, and Matthew Fickus for their helpful comments and remarks on parts of my thesis.

I would also like to extend my deepest gratitude to my family. Without their encouragement, unconditional love, and continued support, I would not have the courage to follow this path. They have always been there for me and I thank them for that. I dedicate my thesis to my closest family.

Finally, I acknowledge funding by the European Research Council through the grant StG 258926.

# Contents

# List of Figures

CHAPTER 1

# Introduction to compressive sensing and low-rank matrix recovery

In this chapter we introduce compressive sensing (sparse vector recovery) and its generalization to matrices, i.e., low-rank matrix recovery. We introduce three criteria that ensure successful recovery in compressive sensing including its versions adapted to matrix scenario – namely, the restricted isometry property, the null space property, and the coherence. In the compressive sensing scenario we concentrate on two algorithms: *$\ell_1$-minimization* which is a convex optimization approach and *iterative hard thresholding algorithm* which is an iterative thresholding based method. In Section 1.2 we introduce the generalizations of these algorithms to matrices – namely, *nuclear norm minimization* and *matrix iterative hard thresholding algorithm*. Later on, in Chapter 4 and Chapter 5 we extend these algorithms to the tensor scenario. However, either the generalization of the algorithm is not straightforward (in the convex optimization case) or the proof of complete convergence of the algorithm is not straightforward (in the iterative hard thresholding case). In fact, for the iterative hard thresholding algorithm we have a convergence guarantee with an additional assumption on thresholding operator. Even more, so far there is no complete theory available for low-rank tensor recovery. Either the methods are not tractable, or the recovery results quantifying the minimal number of measurements are non-optimal or even non-existent. That is why, throughout this chapter, we are going to put an emphasis on certain vector and matrix properties which play a crucial role in the proofs of convergence of above mentioned algorithms. Additionally, we stress why the proofs can not be easily extended to tensor scenario due to the properties of tensors and why there is consequently a need for new approaches in low-rank tensor recovery.

## 1.1. Recovery of $s$-sparse vectors

Compressive sensing (also known as compressed sensing or compressive sampling) is a recently introduced technique for efficiently acquiring and reconstructing signals from linear measurements. In the most basic setting, the observed data $\mathbf{y} \in \mathbb{C}^m$ is connected to the signal $\mathbf{x} \in \mathbb{C}^N$ via

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

For example, in signal and image processing application, $\mathbf{x}$ is a signal that one would like to reconstruct from measured data. Traditional reconstruction techniques suggest that the number of measurements $m$ should be at least as large as the signal length $N$. If $m < N$, then it is known from classical linear algebra that the above linear system is underdetermined and that there are infinitely many solutions. This is also connected to the Shannon sampling theorem, which states that the sampling rate of a continuous-time signal must be twice its highest frequency in order to ensure reconstruction, see [140].

However, many real-world signals are sparse (most of its components are zero) or compressible (approximately sparse) in a certain basis or frame. Compression techniques such as JPEG, MPEG, and MP3 are based on this observation. In essence, these techniques store a compressed version of the measured signal by only retaining the locations and the magnitude of the components that are not zero (or above some threshold). The theory of compressive sensing uses the knowledge of sparsity (or compressibility) of the signal and reconstructs the signal (or its compressed version) via significantly fewer measurements than the ambient dimension of the signal (i.e., $m \ll N$). Thus, the main difficulty lies in locating the nonzero entries of the signal (or signal entries above some threshold), since they are not known in advance.

Two main questions arise in compressive sensing theory:

- Which measurement matrices $\mathbf{A} \in \mathbb{C}^{m \times N}$ are suitable for compressive sensing?
- How can we efficiently reconstruct sparse or approximately sparse signals $\mathbf{x} \in \mathbb{C}^N$ from $\mathbf{y} = \mathbf{A}\mathbf{x}$?

Throughout, let $\mathbb{K}$ denote either $\mathbb{R}$ or $\mathbb{C}$. A vector $\mathbf{x} \in \mathbb{K}^N$ is an $s$-sparse vector if at most $s$ of its entries are different from zero, i.e., if $\|\mathbf{x}\|_0 := |\{i : \mathbf{x}(i) \neq 0\}| \leq s$. The goal of compressive sensing is to reconstruct a sparse vector $\mathbf{x} \in \mathbb{K}^N$ from the measurement vector $\mathbf{y} \in \mathbb{K}^m$ given by $\mathbf{y} = \mathbf{A}\mathbf{x}$ and a measurement matrix $\mathbf{A} \in \mathbb{K}^{m \times N}$ with $m \ll N$. Unfortunately, the natural approach for reconstructing the signal $\mathbf{x}$ by solving the optimization problem

$$\min_{\mathbf{z} \in \mathbb{K}^N} \|\mathbf{z}\|_0 \quad \text{s.t.} \quad \mathbf{A}\mathbf{z} = \mathbf{y} \tag{1.1}$$

is a combinatorial problem known to be NP-hard in general.

In the last decades, several algorithms have been introduced to solve the above problem in special cases. In particular, as we will see later, the recovery guarantees depend on both the choice of the reconstruction algorithm, as well as the choice of the measurement ensemble. Presently, the best known recovery guarantees are always achieved for *random* measurement ensembles.

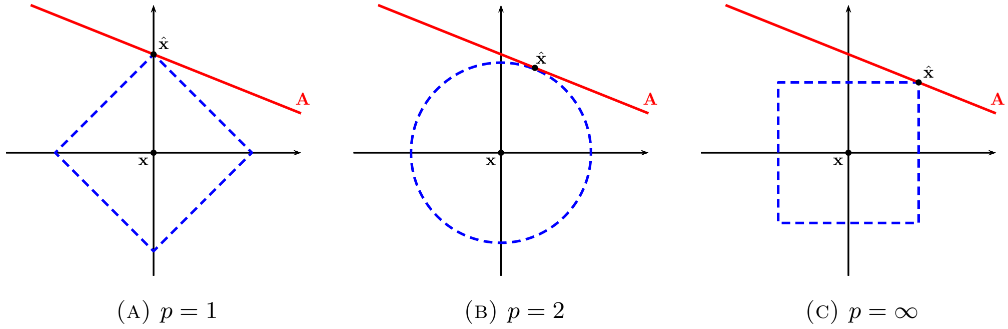
Let us first focus on a convex optimization approach serving as proxy for (1.1). It has been shown that the convex optimization problem (also called *basis pursuit*)

$$\min_{\mathbf{z} \in \mathbb{K}^N} \|\mathbf{z}\|_1 \quad \text{s.t.} \quad \mathbf{A}\mathbf{z} = \mathbf{y}, \tag{1.2}$$

where $\|\mathbf{z}\|_1 = \sum_{i=1}^N |z_i|$ denotes the $\ell_1$-norm of the vector $\mathbf{z}$, reconstructs $\mathbf{x}$ exactly under suitable conditions on $\mathbf{A}$. In fact, this approach has been popularized in signal processing by Donoho et al. [33], even though similar ideas existed earlier in other areas of research, see for example [138]. For an intuitive understanding why $\ell_1$-norm promotes sparsity (instead of the $\ell_2$-norm and the $\ell_\infty$-norm) see Figure 1.1. In mathematical terms, the figure presents a solution of the problem

$$\arg\min_{\mathbf{z} \in \mathbb{K}^N} \|\mathbf{z}\|_p \quad \text{s.t.} \quad \mathbf{A}\mathbf{z} = \mathbf{y},$$

where $p = 1$ in Figure 1.1a, $p = 2$ in Figure 1.1b, and $p = \infty$ in Figure 1.1c. That is, the reconstructed vector $\hat{\mathbf{x}}$ is among all the vectors in the set $\boldsymbol{\mathcal{H}}_{\mathbf{A},\mathbf{y}} := \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{y}\}$ the one with the smallest $\ell_p$-norm. As it can be seen from the figures, the vector $\hat{\mathbf{x}}$ is a sparse vector only in the $\ell_1$-norm case. In particular, in $\mathbb{R}^2$ a sparse vector is a solution of $\ell_2$-minimization only in the case when $\boldsymbol{\mathcal{H}}_{\mathbf{A},\mathbf{y}}$ is parallel to the coordinate axis. In the $\ell_\infty$ scenario, either a sparse vector is not a solution or it is not a unique solution of $\ell_\infty$-minimization (when $\boldsymbol{\mathcal{H}}_{\mathbf{A},\mathbf{y}}$ is parallel to the coordinate



(A) $p = 1$ (B) $p = 2$ (C) $p = \infty$

FIGURE 1.1. Best approximation of a point $\mathbf{x} \in \mathbb{R}^{2\times 2}$ by one dimensional subspace using the $\ell_p$-norm.

axis). Thus, the figures suggest that the $\ell_1$-norm promotes sparsity, whereas the $\ell_2$-norm as well as the $\ell_\infty$-norm do not.

Notice that the constraint in (1.2) will be satisfied for every vector $\mathbf{z} = \mathbf{x}+\mathbf{v}$, where $\mathbf{v} \in \ker(\mathbf{A})$ since

$$\mathbf{Az} = \mathbf{Ax} + \mathbf{Av} = \mathbf{y} + \mathbf{0} = \mathbf{y}.$$

Thus, to ensure that $\mathbf{x}$ is the unique solution of (1.2), we need to pose certain restrictions on the null space (kernel) of $\mathbf{A}$. Let $\boldsymbol{\mathcal{S}} \subseteq [N] = \{1,\ldots,N\}$ be a set of cardinality $|\boldsymbol{\mathcal{S}}| = s \leq N$. With $\mathbf{v}_{\boldsymbol{\mathcal{S}}} \in \mathbb{K}^N$ we denote a vector which coincides with $\mathbf{v} \in \mathbb{K}^N$ on the entries in the set $\boldsymbol{\mathcal{S}}$ and is extended to zero outside $\boldsymbol{\mathcal{S}}$. In addition, $\mathbf{A}_{\boldsymbol{\mathcal{S}}} \in \mathbb{K}^{m\times s}$ denotes the submatrix containing the columns of $\mathbf{A} \in \mathbb{K}^{m\times N}$ indexed by $\boldsymbol{\mathcal{S}}$.

**Definition 1.1** ([25, 60])**.** A matrix $\mathbf{A} \in \mathbb{K}^{m\times N}$ is said to satisfy the *null space property* (NSP) relative to a set $\boldsymbol{\mathcal{S}} \subset [N]$ if

$$\|\mathbf{v}_{\boldsymbol{\mathcal{S}}}\|_1 < \|\mathbf{v}_{\boldsymbol{\mathcal{S}}^c}\|_1, \qquad \text{for all } \mathbf{v} \in \ker(\mathbf{A}) \setminus \{\mathbf{0}\},$$

where $\boldsymbol{\mathcal{S}}^c$ denotes the complement of the set $\boldsymbol{\mathcal{S}}$. A matrix $\mathbf{A}$ satisfies the null space property of order $s$ if it satisfies the null space property relative to any set $\boldsymbol{\mathcal{S}} \subset [N]$ with $|\boldsymbol{\mathcal{S}}| \leq s$.

The following theorem states that the null space property is a necessary and sufficient condition for exact recovery of all sparse vectors via basis pursuit.

**Theorem 1.2** ([60])**.** Given a matrix $\mathbf{A} \in \mathbb{K}^{m\times N}$, every $s$-sparse vector $\mathbf{x} \in \mathbb{K}^N$ is the unique solution of (1.2) with $\mathbf{y} = \mathbf{Ax}$ if and only if $\mathbf{A}$ satisfies the null space property of order $s$.

This condition or a version of this condition is often used to establish uniform recovery results, see for example [35, 71, 89, 146]. However, in practice, it is difficult to check whether a given matrix $\mathbf{A}$ satisfies the null space property. In the following, we present another criterion – introduced for the first time in [26] – which guarantees signal recovery.

**Definition 1.3** (RIP, [26])**.** Let $\mathbf{A} \in \mathbb{K}^{m\times N}$ be a measurement matrix. The restricted isometry constant (RIC) $\delta_s$ of a matrix $\mathbf{A}$ is the smallest $0 < \delta \leq 1$ such that

$$(1-\delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{Ax}\|_2^2 \leq (1+\delta)\|\mathbf{x}\|_2^2, \quad \text{for all } s\text{-sparse vectors } \mathbf{x}. \tag{1.3}$$

We say that $\mathbf{A}$ satisfies the RIP (restricted isometry property) at sparsity level $s$ if $\delta_s$ is bounded by a sufficiently small constant.

By definition, the sequence of restricted isometry constants is nondecreasing, that is

$$\delta_1 \leq \delta_2 \leq \ldots \leq \delta_s \leq \delta_{s+1} \leq \ldots \leq \delta_N.$$

Equivalently, the RIC is given by

$$\delta_s = \max_{\boldsymbol{\mathcal{S}} \subset [N], |\boldsymbol{\mathcal{S}}| \leq s} \left\| \mathbf{A}_{\boldsymbol{\mathcal{S}}}^* \mathbf{A}_{\boldsymbol{\mathcal{S}}} - \mathbf{I} \right\|_{2 \to 2}, \tag{1.4}$$

where $\|\cdot\|_{2\to 2}$ denotes the operator norm. This notion of the RIP is often used in deriving recovery guarantees for compressive sensing algorithms.

To show the equivalence of (1.3) and (1.4), we start with the following observation

$$\begin{aligned} (1-\delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta)\|\mathbf{x}\|_2^2 \quad & \text{for all } s\text{-sparse vectors} \\ \Leftrightarrow \left| \|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right| \leq \delta \|\mathbf{x}\|_2^2 \quad & \text{for all } s\text{-sparse vectors} \\ \Leftrightarrow \left| \|\mathbf{A}_{\boldsymbol{\mathcal{S}}}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right| \leq \delta \|\mathbf{x}\|_2^2 \quad & \text{for all } \boldsymbol{\mathcal{S}} \subset [N], |\boldsymbol{\mathcal{S}}| \leq s, \text{ and for all } \mathbf{x} \in \mathbb{K}^s. \end{aligned}$$

Fix any $\boldsymbol{\mathcal{S}} \subset [N]$ with $|\boldsymbol{\mathcal{S}}| = s$. Expanding the left hand side in the above inequality gives for all $\mathbf{x} \in \mathbb{K}^s$

$$\|\mathbf{A}_{\boldsymbol{\mathcal{S}}}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 = \langle \mathbf{A}_{\boldsymbol{\mathcal{S}}}\mathbf{x}, \mathbf{A}_{\boldsymbol{\mathcal{S}}}\mathbf{x} \rangle - \langle \mathbf{x}, \mathbf{x} \rangle = \langle (\mathbf{A}_{\boldsymbol{\mathcal{S}}}^* \mathbf{A}_{\boldsymbol{\mathcal{S}}} - \mathbf{I})\mathbf{x}, \mathbf{x} \rangle.$$

Since the operator $\mathbf{A}_{\boldsymbol{\mathcal{S}}}^* \mathbf{A}_{\boldsymbol{\mathcal{S}}} - \mathbf{I}$ is Hermitian,

$$\|\mathbf{A}_{\boldsymbol{\mathcal{S}}}^* \mathbf{A}_{\boldsymbol{\mathcal{S}}} - \mathbf{I}\|_{2\to 2} = \max_{\mathbf{x} \in \mathbb{K}^s \setminus \{\mathbf{0}\}} \frac{\langle (\mathbf{A}_{\boldsymbol{\mathcal{S}}}^* \mathbf{A}_{\boldsymbol{\mathcal{S}}} - \mathbf{I})\mathbf{x}, \mathbf{x} \rangle}{\|\mathbf{x}\|_2^2} \leq \delta.$$

The equality holds, as $\delta_s$ is the smallest such $\delta$.

Next, we present a result showing the importance of the RIP in sparse vector recovery.

**Theorem 1.4** ([26])**.** Suppose that the $2s$-th restricted isometry constant of a matrix $\mathbf{A} \in \mathbb{K}^{m \times N}$ satisfies $\delta_{2s} < 1$. Let $\mathbf{x}_0$ be any $s$-sparse vector and let $\mathbf{y} := \mathbf{A}\mathbf{x}_0$. Then $\mathbf{x}_0$ is the only $s$-sparse vector satisfying $\mathbf{A}\mathbf{x} = \mathbf{y}$.

Proof. We prove the theorem by contradiction. Assume that there exists an $s$-sparse vector $\mathbf{x}$ different from $\mathbf{x}_0$ and satisfying $\mathbf{A}\mathbf{x} = \mathbf{y}$. Then $\mathbf{z} := \mathbf{x} - \mathbf{x}_0 \in \ker(\mathbf{A}) \setminus \{\mathbf{0}\}$ and $\mathbf{z}$ is $2s$-sparse. But then

$$0 = \|\mathbf{A}\mathbf{z}\|_2^2 \geq (1 - \delta_{2s})\|\mathbf{z}\|_2^2 > 0$$

which is a contradiction. □

The following theorem on sparse vector recovery via basis pursuit under the RIP assumption is stated without proof.

**Theorem 1.5** ([18, 170])**.** Suppose that the $2s$-th restricted isometry constant of the matrix $\mathbf{A} \in \mathbb{K}^{m \times N}$ satisfies

$$\delta_{2s} < \frac{1}{\sqrt{2}}.$$

Then every $s$-sparse vector $\mathbf{x} \in \mathbb{K}^N$ is the unique solution of

$$\min_{\mathbf{z} \in \mathbb{K}^N} \|\mathbf{z}\|_1 \quad \text{such that} \quad \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}.$$

So far, we have seen that matrices $\mathbf{A} \in \mathbb{K}^{m \times N}$ satisfying the RIP condition are suitable measurement matrices for compressive sensing. In the following we argue that such matrices exist.

In particular, random subgaussian matrices satisfy the RIP condition on sparsity level $s$ with high probability when $m \geq s\ln(eN/s)$. On the other hand, all known deterministic constructions of matrices satisfying the RIP require that $m \gtrsim s^2$ (see for example [47, 84]) or at least $m \gtrsim s^{2-\varepsilon}$ for some small constant $\varepsilon > 0$, see [13, 14].

**Definition 1.6** ([60])**.** Let $\mathbf{A} \in \mathbb{R}^{m\times N}$ be a random matrix, i.e., a matrix having random variables as its entries.

(1) If the entries of $\mathbf{A}$ are independent Rademacher variables (taking values ±1 with equal probability), then $\mathbf{A}$ is called a *Bernoulli random matrix.*

(2) If the entries of $\mathbf{A}$ are independent standard Gaussian random variables, then $\mathbf{A}$ is called a *Gaussian random matrix.*

(3) If the entries of $\mathbf{A}$ are independent mean-zero subgaussian random variables with variance 1 and the same subgaussian parameters $\beta, \kappa$, i.e.,

$$\mathbb{P}\left(\left|\mathbf{A}\left(j,k\right)\right| \geq t\right) \leq \beta e^{-\kappa t^2}, \quad \text{for all } t > 0, j \in [m], k \in [N],$$

then $\mathbf{A}$ is called a subgaussian random matrix. Equivalently, $\mathbf{A}$ is a subgaussian random matrix if for some constant $c$ independent of $j, k$ and $N$ it holds that

$$\mathbb{E}\left[e^{\theta\mathbf{A}(j,k)}\right] \leq e^{c\theta^2}, \quad \text{for all } \theta \in \mathbb{K}, j \in [m], k \in [N].$$

Clearly, Gaussian and Bernoulli random matrices are subgaussian random matrices. The following theorem states that Gaussian and Bernoulli random matrices $\mathbf{A} \in \mathbb{C}^{m\times N}$ satisfy the RIP condition with high probability if the number of rows $m$ is large enough.

**Theorem 1.7** ([60])**.** *Let $\mathbf{A}$ be an $m \times N$ Gaussian or Bernoulli random matrix. Then there exists a universal constant $C > 0$ such that the restricted isometry constant of $\frac{1}{\sqrt{m}}\mathbf{A}$ satisfies $\delta_s \leq \delta$ with probability at least $1 - \varepsilon$ provided*

$$m \geq C\delta^{-2}\left(s\ln\left(eN/s\right) + \ln\left(2\varepsilon^{-1}\right)\right). \tag{1.5}$$

If the locations of non-zero entries in an $s$-sparse vector $\mathbf{x}$ are known, one needs only $s$ independent measurements to recover $\mathbf{x}$. Thus, the factor $s$ is necessary in the above bound. However, in general, these locations are not known. Even more, the bound (1.5) on the number of measurements for sparse vector recovery is in fact optimal. That is, the logarithmic factor $\ln(N/s)$ cannot be improved, see [59, 63].

For practical purposes, one would like to obtain matrices $\mathbf{A}$ with structure satisfying RIP with high probability (with optimal or almost optimal bounds) which are also efficient in the sense that only $O(N\log N)$ operations are required to compute $\mathbf{Ax}$. In particular, it is known that random partial Fourier matrices $\mathbb{C}^{m\times N}$ (obtained by randomly choosing $m$ rows of the $N \times N$ discrete Fourier matrix) satisfy the RIP with high probability if $m \gtrsim s\log^2(s)\log(N)$, see [12, 77]. In [95] it has been shown that a random partial circulant matrix generated by a Rademacher vector $\varepsilon$ satisfies the RIP with high probability provided that $m \gtrsim s\log^2(s)\log^2(N)$. For the definition of partial Fourier and partial circulant matrices see Subsection A.1. In [2] the authors construct RIP-optimal efficient matrices for the regime $m \leq N^{1/2-\mu}$ (for an arbitrarily small $\mu$). In particular, their construction of an RIP matrix is of the form

$$\mathbf{A} = \mathbf{BHD}_1\mathbf{HD}_2\cdots\mathbf{HD}_r,$$

where $\mathbf{H}$ is a Hadamard or a Fourier transform, $\mathbf{D}_i$ is a diagonal matrix with random $\{+1,-1\}$ on the diagonal for all $i$, and $\mathbf{B}$ is any $m \times N$ matrix with orthonormal rows.

So far we have introduced the null space property (NSP), which is a necessary and sufficient condition for sparse vector recovery, and the restricted isometry property (RIP), which is a sufficient condition for sparse vector recovery. Verifying whether a given matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ satisfies the RIP or NSP property is in general NP-hard, see [155]. However, we have shown how to generate a random matrix which will satisfy the RIP condition with high probability. We now introduce the *coherence* which is another criterion that leads to recovery guarantees. It is a very simple measure of suitability of the measurement matrix. The coherence of a given matrix is easily computable which is an advantage in comparison to other two conditions. In general, the smaller the coherence, the better the recovery algorithm performs.

**Definition 1.8** ([60])**.** Let $\mathbf{A} \in \mathbb{K}^{m\times N}$ be a matrix with $\ell_2$-normalized columns $\mathbf{a}_{\cdot 1}, \mathbf{a}_{\cdot 2}, \ldots, \mathbf{a}_{\cdot N}$, i.e., $\|\mathbf{a}_{\cdot i}\|_2 = 1$ for all $i \in [N]$. The *coherence* $\mu = \mu(\mathbf{A})$ of the matrix $\mathbf{A}$ is defined as

$$\mu := \max_{1\leq i\neq j\leq N} |\langle \mathbf{a}_{\cdot i}, \mathbf{a}_{\cdot j}\rangle| .$$

The coherence $\mu$ of the matrix $\mathbf{A} \in \mathbb{K}^{m\times N}$ with $\ell_2$-normalized columns satisfies

$$\sqrt{\frac{N-m}{m(N-1)}} \leq \mu \leq 1.$$

The upper bound follows since $|\langle \mathbf{a}_{\cdot i}, \mathbf{a}_{\cdot j}\rangle| \leq \|\mathbf{a}_{\cdot i}\|_2 \|\mathbf{a}_{\cdot j}\|_2$ and the columns are unit normed. For the bound below, we refer the interested reader to [60] and we note that the bound is tight. In particular, it is achieved if and only if the columns of the matrix form an equiangular tight frame. However, for most pairs $(m, N)$, constructing such a frame is an open problem [148, 165].

Next, we introduce the more general concept of $\ell_1$-coherence function which includes the usual coherence for the parameter $s = 1$.

**Definition 1.9** ([60])**.** Let $\mathbf{A} \in \mathbb{C}^{m\times N}$ be a matrix with $\ell_2$-normalized columns $\mathbf{a}_{\cdot 1}, \mathbf{a}_{\cdot 2}, \ldots, \mathbf{a}_{\cdot N}$, i.e., $\|\mathbf{a}_{\cdot i}\|_2 = 1$ for all $i \in [N]$. The *$\ell_1$-coherence function* $\mu_1$ of the matrix $\mathbf{A}$ is defined for $s \in [N-1]$ by

$$\mu_1(s) := \max_{i\in[N]} \max \left\{ \sum_{j\in\boldsymbol{\mathcal{S}}} |\langle \mathbf{a}_{\cdot i}, \mathbf{a}_{\cdot j}\rangle| : \boldsymbol{\mathcal{S}} \subset [N], |\boldsymbol{\mathcal{S}}| = s, i \notin \boldsymbol{\mathcal{S}} \right\}.$$

Notice that for $s \in [N-1]$

$$\mu \leq \mu_1(s) \leq s\mu,$$

and more generally that for $s, t \in [N-1]$ with $s + t \leq N - 1$,

$$\max\{\mu_1(s), \mu_1(t)\} \leq \mu_1(s+t) \leq \mu_1(s) + \mu_1(t).$$

The following theorem gives a recovery guarantee under the assumption that the $\ell_1$-coherence function of the measurement matrix is small enough.

**Theorem 1.10** ([60])**.** Let $\mathbf{A} \in \mathbb{K}^{m\times N}$ be a matrix with $\ell_2$-normalized columns. If

$$\mu_1(s) + \mu_1(s-1) < 1$$

then every $s$-sparse vector $\mathbf{x} \in \mathbb{K}^N$ is exactly recovered from the measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x}$ via basis pursuit.

Since $\mu_1(s) \leq s\mu$, the above theorem is true also if coherence $\mu$ of the measurement matrix $\mathbf{A}$ with $\ell_2$-normalized columns satisfies $\mu(2s-1) < 1$. The next lemma connects the RIP and the coherence property.

**Lemma 1.11** ([42]). If $\mathbf{A} \in \mathbb{C}^{m \times N}$ has unit-norm columns and coherence $\mu$, then $\mathbf{A}$ satisfies the RIP of order $s$ with $\delta_s = (s-1)\mu$, for all $s < 1/\mu + 1$.

In particular, if $\mathbf{A} \in \mathbb{C}^{m \times N}$ has unit norm columns and coherence $\mu < 1$, then $\mathbf{A}$ satisfies the RIP of order two with $\delta_2 = \mu$. A consequence of the above lemma and Theorem 1.5 is that the coherence is also a sufficient condition for sparse vector recovery.

In more realistic scenarios, the signals we aim to recover are not sparse, but approximately sparse. In such cases, we would like to recover a signal with an error controlled by its distance to $s$-sparse vectors. In the literature, this property is usually called *stability of reconstruction scheme with respect to sparsity defect.* In particular, it is known that the basis pursuit is stable under a slightly strengthened version of the null space property.

**Definition 1.12.** A matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ is said to satisfy the *stable null space property* with constant $0 < \rho < 1$ relative to a set $\boldsymbol{\mathcal{S}} \subset [N]$ if

$$\|\mathbf{v}_{\boldsymbol{\mathcal{S}}}\|_1 \leq \rho \|\mathbf{v}_{\boldsymbol{\mathcal{S}}^c}\|_1 \qquad \text{for all } \mathbf{v} \in \ker(\mathbf{A}).$$

A matrix $\mathbf{A}$ satisfies the stable null space property of order $s$ with constant $0 < \rho < 1$ if it satisfies the stable null space property with constant $0 < \rho < 1$ relative to any set $\boldsymbol{\mathcal{S}} \subset [N]$ with $|\boldsymbol{\mathcal{S}}| \leq s$.

The following theorem is a well-known stability result for basis pursuit.

**Theorem 1.13** ([60]). Suppose that $\mathbf{A} \in \mathbb{C}^{m \times N}$ satisfies the *stable null space property of order $s$ with constant* $0 < \rho < 1$. Then, for any $\mathbf{x} \in \mathbb{C}^N$, a solution $\mathbf{x}^{\#}$ of

$$\min_{\mathbf{z}} \|\mathbf{z}\|_1 \quad \text{s.t.} \quad \mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x}$$

approximates the vector $\mathbf{x}$ with $\ell_1$-error

$$\left\|\mathbf{x} - \mathbf{x}^{\#}\right\|_1 \leq \frac{2(1+\rho)}{(1-\rho)} \sigma_s(\mathbf{x})_1,$$

where $\sigma_s(\mathbf{x})_1 = \inf\{\|\mathbf{w} - \mathbf{x}\|_1 : \mathbf{w} \in \mathbb{C}^N, \|\mathbf{w}\|_0 \leq s\}$.

Additionally, in real applications, measured data will be corrupted by noise since sensing devices do not have infinite precision. As a consequence, the measurement vector $\mathbf{y} \in \mathbb{C}^m$ is only an approximation to $\mathbf{A}\mathbf{x} \in \mathbb{C}^m$, with

$$\|\mathbf{A}\mathbf{x} - \mathbf{y}\| \leq \eta,$$

for some $\eta \geq 0$ and some norm $\|\cdot\|$ on $\mathbb{C}^m$ – typically the $\ell_1$-norm or the $\ell_2$-norm. In such situations, we would like the reconstruction algorithm to recover a signal whose distance to output vector $\mathbf{x}^{\#}$ is controlled by measurement error $\eta \geq 0$. It is well-known that robustness of the convex optimization problem

$$\min_{\mathbf{z} \in \mathbb{C}^N} \|\mathbf{z}\|_1 \quad \text{s.t.} \quad \|\mathbf{A}\mathbf{z} - \mathbf{y}\| \leq \eta,$$

is guaranteed under the following additional strengthening of the null space property.

**Definition 1.14.** A matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ is said to satisfy the *robust null space property* (with respect to $\|\cdot\|$) with constants $0 < \rho < 1$ and $\tau > 0$ relative to a set $\boldsymbol{\mathcal{S}} \subset [N]$ if

$$\|\mathbf{v}_{\boldsymbol{\mathcal{S}}}\|_1 \leq \rho \|\mathbf{v}_{\boldsymbol{\mathcal{S}}^c}\|_1 + \tau \|\mathbf{A}\mathbf{v}\| \quad \text{ for all } \mathbf{v} \in \mathbb{C}^N.$$

A matrix $\mathbf{A}$ satisfies the robust null space property of order $s$ with constants $0 < \rho < 1$ and $\tau > 0$ if it satisfies the robust null space property with constants $\rho$, $\tau$ relative to any set $\boldsymbol{\mathcal{S}} \subset [N]$ with $|\boldsymbol{\mathcal{S}}| \leq s$.

The following theorem includes Theorem 1.13 as the special noiseless case, i.e., when $\eta = 0$.

**Theorem 1.15** ([60]). Suppose that matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ satisfies the robust null space property of order $s$ with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $\mathbf{x} \in \mathbb{C}^N$, a solution $\mathbf{x}^{\#}$ of

$$\min_{\mathbf{z}\in\mathbb{C}^N} \|\mathbf{z}\|_1 \quad \text{s.t.} \quad \|\mathbf{A}\mathbf{z} - \mathbf{y}\| \leq \eta,$$

where $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, and $\|\mathbf{e}\| \leq \eta$ approximates the vector $\mathbf{x}$ with $\ell_1$-error

$$\left\|\mathbf{x} - \mathbf{x}^{\#}\right\|_1 \leq \frac{2(1+\rho)}{(1-\rho)}\sigma_s(\mathbf{x})_1 + \frac{4\tau}{1-\rho}\eta.$$

Finally, we present a compressible vector recovery result with noisy measurements under the assumption that the measurement matrix $\mathbf{A}$ satisfies the RIP. Thus, the $\ell_1$-minimization is robust to noise also under the RIP assumption.

**Theorem 1.16** ([60]). Suppose that the $2s$th restricted isometry constant of the matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ satisfies

$$\delta_{2s} < \frac{4}{\sqrt{41}} \approx 0.6246.$$

Then, for any $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{y} \in \mathbb{C}^m$ with $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \eta$, a solution $\mathbf{x}^*$ of

$$\min_{\mathbf{z}\in\mathbb{C}^N} \|\mathbf{z}\|_1 \quad \text{s.t.} \quad \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_2 \leq \eta$$

approximates the vector $\mathbf{x}$ with error

$$\|\mathbf{x} - \mathbf{x}^*\|_1 \leq C\sigma_s(\mathbf{x})_1 + D\sqrt{s}\eta,$$

where constants $C, D > 0$ depend only on $\delta_{2s}$.

So far we have focused on the $\ell_1$-minimization approach to sparse vector recovery. However, many algorithms have been developed to solve the sparse vector recovery problem. In fact, recovery algorithms for compressed sensing problem can be roughly divided in three categories: convex optimization methods, greedy methods, and thresholding based methods.

Optimization algorithms include *$\ell_1$-minimization* (also called basis pursuit) [33] and *quadratically constrained $\ell_1$-minimization* (also called *basis pursuit denoising*) [154]. These minimization problems can be solved by standard methods from convex optimization, such as interior-point methods, see [15, 121]. Additionally, specialized numerical methods have been developed for solving $\ell_1$-minimization problems such as *Homotopy method* or *modified LARS* [50, 55, 122, 123], *Chambolle and Pock's Primal-Dual Algorithm* [31], and *iteratively reweighted least squares algorithm* [41].

*Orthogonal matching pursuit* (OMP) [157] is a greedy method which builds the support of the reconstructed $s$-sparse vector iteratively by adding one index to the current support set at each iteration. Needell and Tropp in [118] introduced another greedy method called *Compressive sampling matching pursuit algorithm* (CoSaMP). Greedy methods include also *regularized orthogonal matching pursuit* [119, 120], *subspace pursuit algorithm* [39], and others.

Finally, thresholding based methods include *iterative hard thresholding algorithm* [10], *iterative soft thresholding algorithm* – also called *iterative shrinkage-thresholding algorithm* (ISTA) [40, 56, 57], *fast iterative shrinkage thresholding algorithm* (FISTA) [6], and others. In the next subsection we analyze in detail *iterative hard thresholding algorithm* and its normalized version introduced by the same authors in [11]. In Chapter 5 we focus on versions of this algorithm adapted to the tensor scenario.

Before passing to the iterative hard thresholding algorithm we remark that there are two types of results available in compressive sensing for random measurement matrices. *Uniform recovery* results state that with high probability on the choice of the random matrix, *all* sparse signals can be recovered using the same matrix. On the other hand, *nonuniform recovery* results state that a *fixed* sparse signal can be recovered with high probability using a random draw of the measurement matrix. In this chapter and throughout this thesis, the theory of compressed sensing is presented in more detail with the focus on uniform recovery. For nonuniform recovery results, we refer the interested reader to [3, 24, 32].

**1.1.1. Iterative hard thresholding algorithm.** The iterative hard thresholding (IHT) algorithm is an iterative method for solving the system $\mathbf{Az} = \mathbf{y}$ knowing that the solution is $s$-sparse, see Algorithm 1.1. Instead of solving $\mathbf{Az} = \mathbf{y}$, the algorithm is solving the quadratic system $\mathbf{A}^*\mathbf{Az} = \mathbf{A}^*\mathbf{y}$ which can be interpreted as the fixed point equation

$$\mathbf{z} = \mathbf{z} - \mathbf{A}^*\mathbf{Az} + \mathbf{A}^*\mathbf{y} = \mathbf{z} + \mathbf{A}^*\left(\mathbf{y} - \mathbf{Az}\right).$$

Since we search for an $s$-sparse solution, in the $j$th iteration of the algorithm we compute the best $s$-sparse approximation of the vector $\mathbf{u}^j := \mathbf{x}^j + \mathbf{A}^*\left(\mathbf{y} - \mathbf{Ax}^j\right)$ – denoted by $\mathcal{H}_s\left(\mathbf{u}^j\right)$. The best $s$-sparse approximation $\mathcal{H}_s\left(\mathbf{x}\right)$ of a vector $\mathbf{x}$ is obtained by keeping the $s$ largest in magnitude (absolute values) entries of $\mathbf{x}$. Therefore, in the $j$th iteration of the algorithm we obtain the vector $\mathbf{x}^{j+1} = \mathcal{H}_s\left(\mathbf{u}^j\right)$.

**Algorithm 1.1.** Iterative hard thresholding (IHT) algorithm

```
Input: Measurement matrix $\mathbf{A} \in \mathbb{K}^{m \times N}$, measurement vector $\mathbf{y} \in \mathbb{K}^m$,
        sparsity level $s$.
Initialization: sparse vector $\mathbf{x}^0$, typically $\mathbf{x}^0 = \mathbf{0}$, $j = 0$.
Iteration: repeat until the stopping criterion is met at $j = \bar{j}$
        $\mathbf{x}^{j+1} = \mathcal{H}_s\left(\mathbf{x}^j + \mathbf{A}^*\left(\mathbf{y} - \mathbf{Ax}^j\right)\right)$
        $j = j + 1$
Output: $s$-sparse vector $\mathbf{x}^{\#} = \mathbf{x}^{\bar{j}}$
```

A typical stopping criterion for both the IHT (see Algorithm 1.1) and the normalized IHT (NIHT) algorithm (see Algorithm 1.2) is $\|\mathbf{y} - \mathbf{Ax}^{\bar{j}}\|_2 \leq \varepsilon$, for a chosen tolerance $\varepsilon > 0$.

The following result gives a criterion for the convergence of the IHT algorithm under the assumption that the measurement matrix satisfies the RIP condition.

**Theorem 1.17** ([60])**.** For $a \in (0,1)$, let $\mathbf{A} \in \mathbb{K}^{m \times N}$ satisfy the restricted isometry property with

$$\delta_{3s} < \frac{a}{2} \tag{1.6}$$

and let $\mathbf{x} \in \mathbb{K}^N$ be $s$-sparse. Given noisy measurements $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, the vector $\mathbf{x}^{j+1}$ obtained in the $j$-th iteration of the IHT algorithm satisfies

$$\left\|\mathbf{x}^{j+1} - \mathbf{x}\right\|_2 \leq a^{j+1} \left\|\mathbf{x}^0 - \mathbf{x}\right\|_2 + \frac{b(a)}{1-a} \left\|\mathbf{e}\right\|_2,$$

where $b(a) = 2\sqrt{1 + \frac{a}{2}}$. Consequently, if $\mathbf{e} \neq 0$ then after at most $j^* = \lceil \log_{1/a}(\left\|\mathbf{x}^0 - \mathbf{x}\right\|_2 / \left\|\mathbf{e}\right\|_2) \rceil$ iterations, $\mathbf{x}^{j+1}$ estimates $\mathbf{x}$ with accuracy

$$\left\|\mathbf{x}^{j+1} - \mathbf{x}\right\|_F \leq \frac{1 + a + b(a)}{1-a} \left\|\mathbf{e}\right\|_2. \tag{1.7}$$

Proof. It is enough to show that $\mathbf{x}^{j+1}$ obtained in iteration $j$ satisfies

$$\left\|\mathbf{x}^{j+1} - \mathbf{x}\right\|_2 \leq a \left\|\mathbf{x}^{j} - \mathbf{x}\right\|_2 + b \left\|\mathbf{e}\right\|_2, \quad j \geq 0.$$

Since $\mathbf{x}^{j+1}$ is the best $s$-sparse approximation to the vector $\mathbf{u}^j = \mathbf{x}^j + \mathbf{A}^*\left(\mathbf{y} - \mathbf{A}\mathbf{x}^j\right) = \mathbf{x}^j + \mathbf{A}^*\mathbf{A}\left(\mathbf{x} - \mathbf{x}^j\right) + \mathbf{A}^*\mathbf{e}$,

$$\left\|\mathbf{u}^j - \mathbf{x}^{j+1}\right\|_2^2 \leq \left\|\mathbf{u}^j - \mathbf{x}\right\|_2^2. \tag{1.8}$$

Expanding the left hand side we obtain

$$\begin{aligned}\left\|\mathbf{u}^j - \mathbf{x}^{j+1}\right\|_2^2 &= \left\|\left(\mathbf{u}^j - \mathbf{x}\right) - \left(\mathbf{x}^{j+1} - \mathbf{x}\right)\right\|_2^2 \\ &= \left\|\mathbf{u}^j - \mathbf{x}\right\|_2^2 + \left\|\mathbf{x}^{j+1} - \mathbf{x}\right\|_2^2 - 2\operatorname{Re}\left\langle \mathbf{u}^j - \mathbf{x}, \mathbf{x}^{j+1} - \mathbf{x}\right\rangle. \end{aligned} \tag{1.9}$$

Let $\boldsymbol{\mathcal{T}}$ denote the union of the support sets of the vectors $\mathbf{x}^j - \mathbf{x}$ and $\mathbf{x}^{j+1} - \mathbf{x}$, i.e.,

$$\boldsymbol{\mathcal{T}} := \operatorname{supp}\left(\mathbf{x}^j - \mathbf{x}\right) \cup \operatorname{supp}\left(\mathbf{x}^{j+1} - \mathbf{x}\right) \quad \text{and} \quad |\boldsymbol{\mathcal{T}}| \leq 3s.$$

In the following $\mathbf{v}_{\boldsymbol{\mathcal{T}}} \in \mathbb{C}^{|\boldsymbol{\mathcal{T}}|}$ denotes the restriction of a vector $\mathbf{v} \in \mathbb{C}^N$ to indices in $\boldsymbol{\mathcal{T}}$. From (1.8) and (1.9) it follows that

$$\begin{aligned}\left\|\mathbf{x}^{j+1} - \mathbf{x}\right\|_2^2 &\leq 2\operatorname{Re}\left\langle \mathbf{u}^j - \mathbf{x}, \mathbf{x}^{j+1} - \mathbf{x}\right\rangle = 2\operatorname{Re}\left\langle (\mathbf{I} - \mathbf{A}^*\mathbf{A})\left(\mathbf{x}^j - \mathbf{x}\right) + \mathbf{A}^*\mathbf{e}, \mathbf{x}^{j+1} - \mathbf{x}\right\rangle \\ &\leq 2\left|\left\langle (\mathbf{I} - \mathbf{A}^*\mathbf{A})\left(\mathbf{x}^j - \mathbf{x}\right), \mathbf{x}^{j+1} - \mathbf{x}\right\rangle\right| + 2\left|\left\langle \mathbf{A}^*\mathbf{e}, \mathbf{x}^{j+1} - \mathbf{x}\right\rangle\right| \\ &= 2\left|\left\langle \mathbf{x}^j - \mathbf{x}, \mathbf{x}^{j+1} - \mathbf{x}\right\rangle - \left\langle \mathbf{A}\left(\mathbf{x}^j - \mathbf{x}\right), \mathbf{A}\left(\mathbf{x}^{j+1} - \mathbf{x}\right)\right\rangle\right| + 2\left|\left\langle \mathbf{e}, \mathbf{A}(\mathbf{x}^{j+1} - \mathbf{x})\right\rangle\right| \\ &= 2\left|\left\langle \left(\mathbf{x}^j - \mathbf{x}\right)_{\boldsymbol{\mathcal{T}}}, \left(\mathbf{x}^{j+1} - \mathbf{x}\right)_{\boldsymbol{\mathcal{T}}}\right\rangle - \left\langle \mathbf{A}_{\boldsymbol{\mathcal{T}}}\left(\mathbf{x}^j - \mathbf{x}\right)_{\boldsymbol{\mathcal{T}}}, \mathbf{A}_{\boldsymbol{\mathcal{T}}}\left(\mathbf{x}^{j+1} - \mathbf{x}\right)_{\boldsymbol{\mathcal{T}}}\right\rangle\right| \\ &+ 2\left|\left\langle \mathbf{e}, \mathbf{A}_{\boldsymbol{\mathcal{T}}}(\mathbf{x}^{j+1} - \mathbf{x})\right\rangle\right| \\ &= 2\left|\left\langle \left(\mathbf{x}^j - \mathbf{x}\right)_{\boldsymbol{\mathcal{T}}}, (\mathbf{I} - \mathbf{A}_{\boldsymbol{\mathcal{T}}}^*\mathbf{A}_{\boldsymbol{\mathcal{T}}})\left(\mathbf{x}^{j+1} - \mathbf{x}\right)_{\boldsymbol{\mathcal{T}}}\right\rangle\right| + 2\left|\left\langle \mathbf{e}, \mathbf{A}_{\boldsymbol{\mathcal{T}}}(\mathbf{x}^{j+1} - \mathbf{x})\right\rangle\right| \\ &\leq 2\left\|\left(\mathbf{x}^j - \mathbf{x}\right)_{\boldsymbol{\mathcal{T}}}\right\|_2 \left\|\mathbf{I} - \mathbf{A}_{\boldsymbol{\mathcal{T}}}^*\mathbf{A}_{\boldsymbol{\mathcal{T}}}\right\|_{2\to 2} \left\|\left(\mathbf{x}^{j+1} - \mathbf{x}\right)_{\boldsymbol{\mathcal{T}}}\right\|_2 + 2\left\|\mathbf{e}\right\|_2 \left\|\mathbf{A}_{\boldsymbol{\mathcal{T}}}(\mathbf{x}^{j+1} - \mathbf{x})\right\|_2 \\ &\leq 2\delta_{3s}\left\|\mathbf{x}^j - \mathbf{x}\right\|_2 \left\|\mathbf{x}^{j+1} - \mathbf{x}\right\|_2 + 2\sqrt{1 + \delta_{3s}}\left\|\mathbf{e}\right\|_2 \left\|\mathbf{x}^{j+1} - \mathbf{x}\right\|_2, \end{aligned}$$

where the last inequality follows from $|\boldsymbol{\mathcal{T}}| \leq 3s$ and (1.4). If $\left\|\mathbf{x}^{j+1} - \mathbf{x}\right\|_2 > 0$, then dividing the above inequality by $\left\|\mathbf{x}^{j+1} - \mathbf{x}\right\|_2$, we obtain

$$\left\|\mathbf{x}^{j+1} - \mathbf{x}\right\|_2 \leq 2\delta_{3s}\left\|\mathbf{x}^j - \mathbf{x}\right\|_2 + 2\sqrt{1 + \delta_{3s}}\left\|\mathbf{e}\right\|_2.$$

This concludes the proof (since $2\delta_{3s} < a$ by assumption (1.6)). □

**Remark 1.18.** This theorem also applies in a case where the vector $\mathbf{x}$ is compressible (i.e., approximately $s$-sparse). By splitting $\mathbf{x} = \mathbf{x}_s + \mathbf{x}_c$ into the best $s$-sparse approximation $\mathbf{x}_s$ and a remainder term $\mathbf{x}_c$, we obtain

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e} = \mathbf{A}\mathbf{x}_s + \mathbf{A}\mathbf{x}_c + \mathbf{e} = \mathbf{A}\mathbf{x}_s + \widetilde{\mathbf{e}},$$

where $\widetilde{\mathbf{e}} = \mathbf{A}\mathbf{x}_c + \mathbf{e}$. Then the theorem is applied to $\widetilde{\mathbf{e}}$ instead of $\mathbf{e}$ and (1.7) results in the error estimate

$$\left\|\mathbf{x}^{j+1} - \mathbf{x}_s\right\|_F \leq \frac{1 + a + b(a)}{1 - a} \left\|\mathbf{A}\mathbf{x}_c + \mathbf{e}\right\|_2. \tag{1.10}$$

We can further estimate the right hand side of (1.10). To obtain a better estimate, we have to consider $2s$-sparse vectors. This leads us to the following theorem.

**Theorem 1.19** ([60])**.** For $a \in (0, 1)$, let $\mathbf{A} \in \mathbb{K}^{m\times N}$ satisfy the restricted isometry property with

$$\delta_{6s} < \frac{a}{2}.$$

Then for all $\mathbf{x} \in \mathbb{C}^N$, $\mathbf{e} \in \mathbb{C}^N$, the sequence $(\mathbf{x}^j)_j$ defined by the IHT algorithm with $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, $\mathbf{x}^0 = \mathbf{0}$, and $\mathcal{H}_s$ replaced by $\mathcal{H}_{2s}$ satisfies

$$\left\|\mathbf{x} - \mathbf{x}^{j+1}\right\|_2 \leq \frac{C}{\sqrt{s}}\sigma_s(\mathbf{x})_1 + D \left\|\mathbf{e}\right\|_2 + 2\rho^{j+1} \left\|\mathbf{x}\right\|_2, \tag{1.11}$$

where constants $C, D > 0$, and $\rho \in (0, 1)$ depend only on $a$.

The following lemma is used in the above proof.

**Lemma 1.20** ([60])**.** Suppose $\mathbf{A} \in \mathbb{C}^{m\times N}$ has restricted isometry constant $\delta_s < 1$. Given $\tau > 0$, $\xi \geq 0$, and $\mathbf{e} \in \mathbb{C}^m$, assume that two vectors $\mathbf{x}, \mathbf{x}' \in \mathbb{C}^N$ satisfy

$$\left\|\mathbf{x}'\right\|_0 \leq 2s \quad \text{and} \quad \left\|\mathbf{x}_{\boldsymbol{\mathcal{T}}} - \mathbf{x}'\right\|_2 \leq \tau \left\|\mathbf{A}\mathbf{x}_{\boldsymbol{\mathcal{T}}^c} + \mathbf{e}\right\|_2 + \xi,$$

where $\boldsymbol{\mathcal{T}}$ denotes an index set of $2s$ largest in magnitude entries of $\mathbf{x}$. Then

$$\left\|\mathbf{x} - \mathbf{x}'\right\|_2 \leq \frac{C}{\sqrt{s}}\sigma_s(\mathbf{x})_1 + \tau \left\|\mathbf{e}\right\|_2 + \xi,$$

where $C = 1 + \sqrt{2}\tau$.

Proof.

$$\begin{aligned}\left\|\mathbf{x} - \mathbf{x}'\right\|_2 &= \left\|\mathbf{x}_{\boldsymbol{\mathcal{T}}^c} + \mathbf{x}_{\boldsymbol{\mathcal{T}}} - \mathbf{x}'\right\|_2 \leq \left\|\mathbf{x}_{\boldsymbol{\mathcal{T}}^c}\right\|_2 + \left\|\mathbf{x}_{\boldsymbol{\mathcal{T}}} - \mathbf{x}'\right\|_2 \leq \left\|\mathbf{x}_{\boldsymbol{\mathcal{T}}^c}\right\|_2 + \tau \left\|\mathbf{A}\mathbf{x}_{\boldsymbol{\mathcal{T}}^c} + \mathbf{e}\right\|_2 + \xi \\ &\leq \left\|\mathbf{x}_{\boldsymbol{\mathcal{T}}^c}\right\|_2 + \tau \left\|\mathbf{A}\mathbf{x}_{\boldsymbol{\mathcal{T}}^c}\right\|_2 + \tau \left\|\mathbf{e}\right\|_2 + \xi. \end{aligned} \tag{1.12}$$

Let $\boldsymbol{\mathcal{S}} \subset \boldsymbol{\mathcal{T}}$ denote the index set of $s$ largest absolute entries of $\mathbf{x}$. Then

$$\left\|\mathbf{x}_{\boldsymbol{\mathcal{T}}^c}\right\|_2 = \sigma_s(\mathbf{x}_{\boldsymbol{\mathcal{S}}^c})_2 \leq \frac{1}{\sqrt{s}} \left\|\mathbf{x}_{\boldsymbol{\mathcal{S}}^c}\right\|_1 = \frac{1}{\sqrt{s}}\sigma_s(\mathbf{x})_1, \tag{1.13}$$

where the above inequality follows from the following observation. Let $\mathbf{x}^*$ denote the nonincreasing rearrangement of the vector $\mathbf{x}$. That is, $x_1^* \geq x_2^* \geq \cdots \geq x_N^* \geq 0$ and there exists a permutation $\pi : [N] \to [N]$ with $x_j^* = \left|x_{\pi(j)}\right|$, for all $j \in [N]$. Then it holds

$$\sigma_s(\mathbf{x}_{\boldsymbol{\mathcal{S}}^c})_2^2 = \sum_{j=2s+1}^{N} (x_j^*)^2 \leq x_{2s}^* \sum_{j=2s+1}^{N} x_j^* \leq x_{2s}^* \left\|\mathbf{x}_{\boldsymbol{\mathcal{S}}^c}\right\|_1 \leq \frac{1}{s} \sum_{j=s+1}^{2s} x_j^* \left\|\mathbf{x}_{\boldsymbol{\mathcal{S}}^c}\right\|_1 \leq \frac{1}{s} \left\|\mathbf{x}_{\boldsymbol{\mathcal{S}}^c}\right\|_1^2.$$

Set $\boldsymbol{\mathcal{S}}_1 := \boldsymbol{\mathcal{T}} \backslash \boldsymbol{\mathcal{S}}$. Let us partition the complement of $\boldsymbol{\mathcal{T}}$ as $\boldsymbol{\mathcal{T}}^c = \boldsymbol{\mathcal{S}}_2 \cup \boldsymbol{\mathcal{S}}_3 \cup \ldots$ as

$$\boldsymbol{\mathcal{S}}_2 := \text{index set of } s \text{ largest absolute entries of } \mathbf{x} \text{ in } \boldsymbol{\mathcal{T}}^c$$

$$\boldsymbol{\mathcal{S}}_3 := \text{index set of } s \text{ largest absolute entries of } \mathbf{x} \text{ in } (\boldsymbol{\mathcal{T}} \cup \boldsymbol{\mathcal{S}}_2)^c, \text{ etc.}$$

Applying the RIP assumption, we obtain the estimate

$$\left\| \mathbf{A}\mathbf{x}_{\boldsymbol{\mathcal{T}}^c} \right\|_2 \leq \sum_{k \geq 2} \left\| \mathbf{A}\mathbf{x}_{\boldsymbol{\mathcal{S}}_k} \right\|_2 \leq \sqrt{1+\delta_s} \sum_{k \geq 2} \left\| \mathbf{x}_{\boldsymbol{\mathcal{S}}_k} \right\|_2 . \tag{1.14}$$

Since

$$\frac{1}{\sqrt{s}} \left\| \mathbf{x}_{\boldsymbol{\mathcal{S}}_k} \right\|_2 = \left[ \frac{1}{s} \sum_{\ell \in \boldsymbol{\mathcal{S}}_k} x_\ell^2 \right]^{1/2} \leq \max_{\ell \in \boldsymbol{\mathcal{S}}_k} |x_\ell| \leq \min_{p \in \boldsymbol{\mathcal{S}}_{k-1}} |x_p| \leq \frac{1}{s} \sum_{p \in \boldsymbol{\mathcal{S}}_{k-1}} |x_p| = \frac{1}{s} \left\| \mathbf{x}_{\boldsymbol{\mathcal{S}}_{k-1}} \right\|_1 \tag{1.15}$$

we have

$$\sum_{k \geq 2} \left\| \mathbf{x}_{\boldsymbol{\mathcal{S}}_k} \right\|_2 \leq \sum_{\ell \geq 1} \frac{1}{\sqrt{s}} \left\| \mathbf{x}_{\boldsymbol{\mathcal{S}}_\ell} \right\|_1 = \frac{1}{\sqrt{s}} \left\| \mathbf{x}_{\boldsymbol{\mathcal{S}}^c} \right\|_1 = \frac{1}{\sqrt{s}} \sigma_s(\mathbf{x})_1. \tag{1.16}$$

Plugging (1.16) in (1.14) leads to the estimate

$$\left\| \mathbf{A}\mathbf{x}_{\boldsymbol{\mathcal{T}}^c} \right\|_2 \leq \frac{\sqrt{1+\delta_s}}{\sqrt{s}} \sigma_s(\mathbf{x})_1 \leq \frac{\sqrt{2}}{\sqrt{s}} \sigma_s(\mathbf{x})_1. \tag{1.17}$$

Substituting (1.13) and (1.17) in (1.12) we obtain the estimate

$$\left\| \mathbf{x} - \mathbf{x}' \right\|_2 \leq \frac{1}{\sqrt{s}} \sigma_s(\mathbf{x})_1 + \tau \frac{\sqrt{2}}{\sqrt{s}} \sigma_s(\mathbf{x})_1 + \tau \left\| \mathbf{e} \right\|_2 + \xi = \frac{1}{\sqrt{s}} \left( 1 + \sqrt{2}\tau \right) \sigma_s(\mathbf{x})_1 + \tau \left\| \mathbf{e} \right\|_2 + \xi.$$

□

Proof of Theorem 1.19. Theorem 1.17 implies that there exists $0 < \rho < 1$ and $\tau > 0$ depending only on $a$ such that, for any $j \geq 0$,

$$\left\| \mathbf{x}_{\boldsymbol{\mathcal{T}}} - \mathbf{x}^{j+1} \right\|_2 \leq \tau \left\| \mathbf{A}\mathbf{x}_{\boldsymbol{\mathcal{T}}^c} + \mathbf{e} \right\|_2 + \rho^{j+1} \left\| \mathbf{x}_{\boldsymbol{\mathcal{T}}} \right\|_2 ,$$

where $\boldsymbol{\mathcal{T}}$ denotes an index set of $2s$ largest in magnitude entries of $\mathbf{x}$. Then Lemma 1.20 with $\mathbf{x}' = \mathbf{x}^{j+1}$ and $\xi = \rho^{j+1} \left\| \mathbf{x}_{\boldsymbol{\mathcal{T}}} \right\|_2 \leq \rho^{j+1} \left\| \mathbf{x} \right\|_2$, implies that

$$\left\| \mathbf{x} - \mathbf{x}^{j+1} \right\|_2 \leq \frac{C}{\sqrt{s}} \sigma_s(\mathbf{x})_1 + D \left\| \mathbf{e} \right\|_2 + 2\rho^{j+1} \left\| \mathbf{x} \right\|_2 , \tag{1.18}$$

where $C, D > 0$ depend only on $\tau$, hence only on $\delta_{6s}$. □

The crucial point in the proof of Theorem 1.17 is step (1.8). That is, the fact that we know how to obtain the best $s$-sparse approximation of a given vector $\mathbf{x} \in \mathbb{K}^N$. As it will be seen later, in the tensor case it is not known how to efficiently obtain the best rank-$\mathbf{r}$ approximation of a given tensor – regardless of the choice of tensor rank. We present another recovery result for the IHT algorithm without a proof. Essentially, small coherence of the measurement matrix guarantees the success of the IHT algorithm.

**Theorem 1.21** ([60])**.** Let $\mathbf{A} \in \mathbb{K}^{m \times N}$ be a matrix with $\ell_2$-normalized columns and $\mathbf{y} = \mathbf{A}\mathbf{x}$. If

$$\mu_1(2s) < 1/2 \text{ – in particular if } \mu < \frac{1}{4s}$$

then for every $s$-sparse vector $\mathbf{x} \in \mathbb{K}^N$, the sequence $(\mathbf{x}^j)_j$ generated by the IHT algorithm convergences linearly to $\mathbf{x}$.

The normalized iterative hard thresholding (NIHT) algorithm (see Algorithm 1.2) is a version of the IHT algorithm with a different stepsize. In particular, the vector update in the $j$th iteration is of the form

$$\mathbf{x}^{j+1} = \mathcal{H}_s(\mathbf{x}^j + \mu_j \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^j)),$$

for a precisely defined stepsize $\mu_j$. To motivate the stepsize in the NIHT algorithm, we recall the original sparse approximation problem we want to solve. Assume we are given a measurement matrix $\mathbf{A} \in \mathbb{C}^{m\times N}$ and a measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x}$ knowing that the signal $\mathbf{x}$ we want to recover is at most $s$-sparse (or is well approximated by an $s$-sparse vector). This leads to the following $s$-sparse constrained optimization problem which is NP-hard in general

$$\min_{\mathbf{x}} \frac{1}{2}\left\|\mathbf{y} - \mathbf{A}\mathbf{x}\right\|_2^2 \quad \text{s.t.} \quad \left\|\mathbf{x}\right\|_0 \leq s.$$

Gradient descent methods can be used to solve $f(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{y} - \mathbf{A}\mathbf{x}\right\|_2^2$. Starting with an initial guess $\bar{\mathbf{x}}_0$, the gradient $\nabla f(x) = \mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{y})$, and adding a stepsize $\mu_j$ leads to the sequence $(\bar{\mathbf{x}}^j)_j$

$$\bar{\mathbf{x}}^{j+1} = \bar{\mathbf{x}}^j + \mu_j \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^j).$$

Thus, iterative hard thresholding algorithm could be considered as a gradient descent method. Since the signal $\mathbf{x}$ we seek to recover is $s$-sparse and a current iterate is not necessary $s$-sparse, we additionally apply the thresholding operator $\mathcal{H}_s$ on the iterates. This leads to the sequence $(\mathbf{x}^j)_j$ of the more generalized IHT algorithm with

$$\mathbf{x}^{j+1} = \mathcal{H}_s(\mathbf{x}^j + \mu_j \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^j)).$$

As already mentioned, if $\mu_j = 1$ for all iterations $j$ we get the IHT algorithm presented as Algorithm 1.1. However, one can try to improve the performance of the algorithm by choosing carefully the parameter $\mu_j$ in each step of the algorithm.

Let us assume that in iteration $j$ we have identified the correct support $\boldsymbol{\mathcal{U}}_j$. That is, $\boldsymbol{\mathcal{U}}_j$ is the support of the best $s$-term approximation to $\mathbf{x}$. Since the support in this case is fixed, we can calculate the optimal stepsize (i.e., the stepsize that maximally reduces the error $\frac{1}{2}\left\|\mathbf{y} - \mathbf{A}\mathbf{x}^j\right\|_2^2$ in each iteration) [66] and obtain

$$\mu_j = \frac{\left\|\mathbf{A}^*_{\boldsymbol{\mathcal{U}}_j}(\mathbf{y} - \mathbf{A}\mathbf{x}^j)\right\|_2^2}{\left\|\mathbf{A}_{\boldsymbol{\mathcal{U}}_j}\mathbf{A}^*_{\boldsymbol{\mathcal{U}}_j}(\mathbf{y} - \mathbf{A}\mathbf{x}^j)\right\|_2^2}.$$

This stepsize is then used for the NIHT algorithm, see Algorithm 1.2. The NIHT algorithm was introduced in [11] by the same authors who introduced the IHT algorithm.

Similar convergence guarantees are available also for the NIHT algorithm. We present here a recovery result under the assumption that the measurement matrix satisfies the RIP.

**Theorem 1.22.** For $a \in (0,1)$, let $\mathbf{A} \in \mathbb{K}^{m\times N}$ satisfy the restricted isometry property with

$$\delta_{3s} < \frac{a}{a+4}$$

and let $\mathbf{x} \in \mathbb{K}^N$ be $s$-sparse vector. Given noisy measurements $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, the vector $\mathbf{x}^{j+1}$ obtained in the $j$-th iteration of the NIHT algorithm satisfies

$$\left\|\mathbf{x}^{j+1} - \mathbf{x}\right\|_2 \leq a^{j+1}\left\|\mathbf{x}^0 - \mathbf{x}\right\|_2 + \frac{b(a)}{1-a}\left\|\mathbf{e}\right\|_2,$$

where $b(a) = 2\frac{\sqrt{1+\delta_{3s}}}{1-\delta_s}$. Consequently, if $\mathbf{e} \neq 0$ then after at most $j^* = \lceil \log_{1/a}(\left\|\mathbf{x}^0 - \mathbf{x}\right\|_2 / \|\mathbf{e}\|_2) \rceil$ iterations, $\mathbf{x}^{j+1}$ estimates $\mathbf{x}$ with accuracy

$$\left\|\mathbf{x}^{j+1} - \mathbf{x}\right\|_F \leq \frac{1+a+b(a)}{1-a} \|\mathbf{e}\|_2 .$$

Proof. The proof of the theorem is analogous to the proof of Theorem 1.17. In particular, we get the estimate

$$\left\|\mathbf{x}^{j+1} - \mathbf{x}\right\|_2 \leq 2 \left\|\mathbf{I} - \mu_j \mathbf{A}_{\mathcal{T}}^* \mathbf{A}_{\mathcal{T}}\right\|_{2\to 2} \left\|\mathbf{x}^j - \mathbf{x}\right\|_2 + 2\mu_j \sqrt{1+\delta_{3s}} \|\mathbf{e}\|_2 .$$

It remains to bound terms $\left\|\mathbf{I} - \mu_j \mathbf{A}_{\mathcal{T}}^* \mathbf{A}_{\mathcal{T}}\right\|_{2\to 2}$ and $\mu_j$. This is done analogously as in the proof of Theorem 1.41. □

**Algorithm 1.2.** Normalized iterative hard thresholding (NIHT) algorithm

```
Input: Measurement matrix $\mathbf{A} \in \mathbb{K}^{m \times N}$, measurement vector $\mathbf{y} \in \mathbb{K}^m$,
        sparsity level $s$.
Initialization: sparse vector $\mathbf{x}^0$, typically $\mathbf{x}^0 = \mathbf{0}$, $\boldsymbol{\mathcal{U}}^0 = \mathrm{supp}\left(\mathcal{H}_s\left(\mathbf{A}^*\left(\mathbf{y}\right)\right)\right)$, $j = 0$.
Iteration: repeat until the stopping criterion is met at $j = \bar{j}$
        $\mu_j = \frac{\left\|\mathbf{A}_{\boldsymbol{\mathcal{U}}_j}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^j)\right\|_2^2}{\left\|\mathbf{A}_{\boldsymbol{\mathcal{U}}_j}\mathbf{A}_{\boldsymbol{\mathcal{U}}_j}^*(\mathbf{y}-\mathbf{A}\mathbf{x}^j)\right\|_2^2}$.
        $\mathbf{x}^{j+1} = \mathcal{H}_s\left(\mathbf{x}^j + \mu_j \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^j)\right)$
        $\boldsymbol{\mathcal{U}}^{j+1} = \mathrm{supp}\left(\mathbf{x}^{j+1}\right)$
        $j = j + 1$
Output: $s$-sparse vector $\mathbf{x}^{\#} = \mathbf{x}^{\bar{j}}$
```

**Remark 1.23.** In [112] extensive numerical experiments have been conducted regarding the iterative hard thresholding algorithm. In particular, the authors considered versions of the IHT algorithm with iterates defined via

$$\mathbf{x}^{j+1} = \mathcal{H}_s(\mathbf{x}^j + \mu \mathbf{A}^*(\mathbf{y} - \mathbf{A}\mathbf{x}^j))$$

with a fixed stepsize $\mu$. The authors suggest setting the parameter $\mu = 0.65$ since they obtained the best performance of IHT for this stepsize.

An undesirable property of the IHT algorithm is that it is sensitive to scaling of the matrix $\mathbf{A}$. This has also been supported by extensive numerical experiments in [11]. However, the NIHT algorithm is invariant under arbitrary scaling of the matrix $\mathbf{A}$. Even more, the numerical experiments conducted in [11] suggest that the NIHT algorithm also has better average performance.

## 1.2. Low-rank matrix recovery

Low-rank matrix recovery builds on ideas from the theory of compressive sensing. The goal of low-rank matrix recovery is to reconstruct a matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ of rank at most $r \leq \min\{n_1, n_2\}$ from the measurement vector $\mathbf{y} = \mathcal{A}(\mathbf{X})$, where $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is a known linear operator with $m \ll n_1 n_2$. This problem appears in many applications, namely recommender systems [135, 144] (which includes also the famous *Netflix Prize* [7]), quantum state tomography [74], phase retrieval [21, 73, 100] etc. We remark that by the Riesz representation theorem, for any linear operator $\mathcal{A} : \mathbb{K}^{n_1 \times n_2} \to \mathbb{K}^m$, there exist a unique set of matrices $\{\mathbf{A}_i \in \mathbb{K}^{n_1 \times n_2}\}_{i=1}^m$ such that $(\mathcal{A}(\mathbf{X}))_i = \operatorname{tr}(\mathbf{X}\mathbf{A}_i^*)$, for all $i \in [m]$ and for all $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$. We refer to the matrices $\{\mathbf{A}_i\}_{i=1}^m$ as sensing matrices of the linear operator $\mathcal{A}$.

Similarly to the sparse vector recovery, the natural approach of finding the solution of the optimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}} \operatorname{rank}(\mathbf{Z}) \quad \text{s.t.} \quad \mathcal{A}(\mathbf{Z}) = \mathbf{y},$$

is NP-hard in general. Let $\boldsymbol{\sigma}$ denote the vector of singular values of a matrix $\mathbf{Z}$. Notice that $\|\boldsymbol{\sigma}\|_0 = \operatorname{rank}(\mathbf{Z})$. The theory of compressive sensing suggests that the $\ell_1$-norm minimization is a good proxy for the $\ell_0$-minimization problem. This results in $\|\boldsymbol{\sigma}\|_1 = \|\mathbf{Z}\|_*$, where $\|\mathbf{Z}\|_* = \operatorname{tr}\left(\sqrt{\mathbf{Z}^*\mathbf{Z}}\right)$ denotes the nuclear norm (also known as the trace norm or the Schatten 1-norm) of a matrix $\mathbf{Z}$. Even more, it has been shown that solving the convex optimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}} \|\mathbf{Z}\|_* \quad \text{s.t.} \quad \mathcal{A}(\mathbf{Z}) = \mathbf{y}, \tag{1.19}$$

reconstructs $\mathbf{X}$ exactly under suitable conditions on $\mathcal{A}$. The required number of measurements scales as $m \geq Cr \max\{n_1, n_2\}$ for Gaussian measurement ensembles [22, 133]. Much more generally, this bound holds also for ensembles with four finite moments, i.e., it is enough to require that the sensing matrices $\{\mathbf{A}_k\}_{k=1}^m$ are independent copies of a random matrix $\mathbf{A}$ satisfying $\mathbb{E}[\mathbf{A}(i,j)] = 0$, $\mathbb{E}\left[\mathbf{A}^2(i,j)\right] = 1$, and $\mathbb{E}\left[\mathbf{A}^4(i,j)\right] \leq C_4$ for all $i, j$ and some constant $C_4$, see [88]. Additionally, this bound is optimal since $n_1 r + n_2 r - r^2 \sim Cr \max\{n_1, n_2\}$ is the number of degrees of freedom to describe a rank-$r$ matrix of dimensions $n_1 \times n_2$.

The existence of the singular value decomposition (SVD) of matrices plays a crucial role in the proofs of convergence of the algorithms designed for low-rank matrix recovery. In particular, unlike in the general tensor case, due to the existence of the SVD the best rank-$r$ approximation of a given matrix can be computed efficiently and a $2r$-rank matrix can be decomposed into a sum of two mutually orthogonal rank-$r$ matrices.

Recall that for compressive sensing we introduced three different concepts which guarantee recovery via efficient algorithms: the restricted isometry property (RIP), the null space property (NSP), and the coherence. In the following we introduce analogous conditions for low-rank matrix recovery, followed by the corresponding recovery guarantees for nuclear norm minimization.

**Definition 1.24** (Matrix-RIP, [133])**.** Let $\mathcal{A} \colon \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ be a linear measurement map. For every integer $r$ with $1 \leq r \leq \min\{n_1, n_2\}$ the $r$-th matrix restricted isometry constant $\delta_r$ of $\mathcal{A}$ is the smallest $0 < \delta_r$ such that

$$(1 - \delta_r) \|\mathbf{X}\|_F^2 \leq \|\mathcal{A}(\mathbf{X})\|_2^2 \leq (1 + \delta_r) \|\mathbf{X}\|_F^2$$

holds for all matrices $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ of rank at most $r$. We say that $\mathcal{A}$ satisfies the matrix-RIP at rank $r$ and level $\delta_r$ if $\delta_r$ is sufficiently small.

The following two recovery theorems present the power of the matrix-RIP. The first theorem is an analogue of Theorem 1.4 for sparse vector recovery.

**Theorem 1.25** ([133])**.** Let $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ be a linear map satisfying $\delta_{2r} < 1$ and let $\mathbf{y} := \mathcal{A}(\mathbf{X}_0)$. Then $\mathbf{X}_0$ is the only rank-$r$ matrix satisfying $\mathcal{A}(\mathbf{X}) = \mathbf{y}$.

Proof. We prove the theorem by contradiction. Thus, assume that there exists a rank-$r$ matrix $\mathbf{X}$ different from $\mathbf{X}_0$ satisfying $\mathcal{A}(\mathbf{X}) = \mathbf{y}$. Then $\mathbf{Z} := \mathbf{X}_0 - \mathbf{X} \in \ker(\mathcal{A}) \setminus \{\mathbf{0}\}$ and $\operatorname{rank}(\mathbf{Z}) \leq 2r$. But then

$$0 = \|\mathcal{A}(\mathbf{Z})\|_2^2 \geq (1 - \delta_{2r}) \|\mathbf{Z}\|_F^2 > 0$$

which is a contradiction. □

The proof of the following theorem is presented in details. We will emphasize the importance of the existence of the singular value decomposition in obtaining the desired result. As it will be seen later in Chapter 2, such a decomposition does not exist for tensors, at least not one that can be computed efficiently. This causes significant difficulties in extending the theory to low-rank tensor recovery.

**Theorem 1.26** ([133])**.** Let $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ be a linear map with $\delta_{5r} < \frac{1}{5}$ and let $\mathbf{y} := \mathcal{A}(\mathbf{X}_0)$. Then $\mathbf{X}_0$ is the unique solution of

$$\underset{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}}{\arg\min} \ \|\mathbf{Z}\|_* \quad \text{s.t.} \quad \mathcal{A}(\mathbf{Z}) = \mathbf{y}. \tag{1.20}$$

**Remark 1.27.** We remark that the bound $\delta_{5r} < \frac{1}{5}$ in Theorem 1.26 is not optimal. In fact, it is known that the optimal bound which guarantees that the nuclear norm minimization stably recovers a low-rank matrix is $\delta_{2r} < 1/\sqrt{2}$, see [18].

To prove the theorem we use the following technical lemma.

**Lemma 1.28** ([133])**.** Let $\mathbf{A}$ and $\mathbf{B}$ be matrices of dimensions $m \times n$ with $\operatorname{rank}(\mathbf{A}) < \min\{m, n\}$. Then there exist matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ such that

**P.1**: $\mathbf{B} = \mathbf{B}_1 + \mathbf{B}_2$

**P.2**: $\operatorname{rank}(\mathbf{B}_1) \leq 2 \operatorname{rank}(\mathbf{A})$

**P.3**: $\mathbf{A}\mathbf{B}_2^T = \mathbf{0}$ and $\mathbf{A}^T\mathbf{B}_2 = \mathbf{0}$

**P.4**: $\langle \mathbf{B}_1, \mathbf{B}_2 \rangle = 0$

**P.5**: $\|\mathbf{A} + \mathbf{B}_2\|_* = \|\mathbf{A}\|_* + \|\mathbf{B}_2\|_*$.

Proof. Let

$$\mathbf{A} = \mathbf{U} \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{V}^T$$

be the singular value decomposition of the matrix $\mathbf{A}$ and define

$$\hat{\mathbf{B}} := \mathbf{U}^T \mathbf{B} \mathbf{V} = \begin{pmatrix} \hat{\mathbf{B}}_{11} & \hat{\mathbf{B}}_{12} \\ \hat{\mathbf{B}}_{21} & \hat{\mathbf{B}}_{22} \end{pmatrix},$$

where $\hat{\mathbf{B}}_{11}$ is of the same size as $\boldsymbol{\Sigma}$. It can be easily verified that matrices

$$\mathbf{B}_1 := \mathbf{U} \begin{pmatrix} \hat{\mathbf{B}}_{11} & \hat{\mathbf{B}}_{12} \\ \hat{\mathbf{B}}_{21} & \mathbf{0} \end{pmatrix} \mathbf{V}^T \quad \text{and} \quad \mathbf{B}_2 := \mathbf{U} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{B}}_{22} \end{pmatrix} \mathbf{V}^T$$

satisfy the conditions **P.1**:-**P.4**:.

Let $\mathbf{A} = \mathbf{U}_1 \boldsymbol{\Sigma} \mathbf{V}_1^T$ with $\operatorname{rank}(\mathbf{A}) = r_1$ and $\mathbf{B}_2 = \mathbf{W} \boldsymbol{\Gamma} \mathbf{Z}^T$ with $\operatorname{rank}(\mathbf{B}_2) = r_2$ be the reduced singular value decomposition of the matrices $\mathbf{A}$ and $\mathbf{B}_2$, respectively.

Let $\mathbf{u}_{\cdot i}, \mathbf{v}_{\cdot i}, \mathbf{w}_{\cdot i}, \mathbf{z}_{\cdot i}$ denote the columns of matrices $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{Z}$, respectively. By **P.3**:,

$$\mathbf{0} = \mathbf{A}\mathbf{B}_2^T = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sigma_i \gamma_j \left(\mathbf{v}_{\cdot i}^T \mathbf{z}_{\cdot j}\right) \mathbf{u}_{\cdot i} \mathbf{w}_{\cdot j}^T.$$

Since the matrices $\left\{\mathbf{u}_{\cdot i}\mathbf{w}_{\cdot j}^T : i \in [r_1], j \in [r_2]\right\}$ are pairwise orthonormal, it follows that $\mathbf{v}_{\cdot i}^T \mathbf{z}_{\cdot j} = \mathbf{0}$, for all $i \in [r_1]$, $j \in [r_2]$. Similarly, from $\mathbf{A}^T\mathbf{B}_2 = \mathbf{0}$ we obtain that $\mathbf{u}_{\cdot i}^T \mathbf{w}_{\cdot j} = 0$, for all $i \in [r_1]$, $j \in [r_2]$. Thus

$$\mathbf{A} + \mathbf{B}_2 = \begin{pmatrix} \mathbf{U}_1 & \mathbf{W} \end{pmatrix} \begin{pmatrix} \mathbf{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma} \end{pmatrix} \begin{pmatrix} \mathbf{V}_1 & \mathbf{Z} \end{pmatrix}^T$$

forms the singular value decomposition of the matrix $\mathbf{A}+\mathbf{B}_2$ and $\left\|\mathbf{A}+\mathbf{B}_2\right\|_* = \left\|\mathbf{A}\right\|_* + \left\|\mathbf{B}_2\right\|_*$. □

Notice that the proof of the above lemma relies heavily on the existence of the singular value decomposition (SVD). Already in the first step of the proof we assume that such a decomposition exists.

Proof of Theorem 1.26. Let $\mathbf{X}_*$ be any solution of (1.20). Then by optimality of $\mathbf{X}_*$, we have $\left\|\mathbf{X}_0\right\|_* \geq \left\|\mathbf{X}_*\right\|_*$. Let $\mathbf{R} := \mathbf{X}_* - \mathbf{X}_0 \in \ker(\mathcal{A})$. Applying Lemma 1.28 to the matrices $\mathbf{X}_0$ and $\mathbf{R}$, it follows that there exist two matrices $\mathbf{R}_0$ and $\mathbf{R}_c$ such that

**S.1**: $\mathbf{R} = \mathbf{R}_0 + \mathbf{R}_c$
**S.2**: $\operatorname{rank}(\mathbf{R}_0) \leq 2 \operatorname{rank}(\mathbf{X}_0) = 2r$
**S.3**: $\mathbf{X}_0\mathbf{R}_c^T = \mathbf{0}$, $\mathbf{X}_0^T\mathbf{R}_c = \mathbf{0}$
**S.4**: $\langle \mathbf{R}_0, \mathbf{R}_c \rangle = 0$
**S.5**: $\left\|\mathbf{X}_0 + \mathbf{R}_c\right\|_* = \left\|\mathbf{X}_0\right\|_* + \left\|\mathbf{R}_c\right\|_*$.

Then

$$\left\|\mathbf{X}_0\right\|_* \geq \left\|\mathbf{X}_*\right\|_* = \left\|\mathbf{X}_0 + \mathbf{R}\right\|_* \geq \left\|\mathbf{X}_0 + \mathbf{R}_c\right\|_* - \left\|\mathbf{R}_0\right\|_* = \left\|\mathbf{X}_0\right\|_* + \left\|\mathbf{R}_c\right\|_* - \left\|\mathbf{R}_0\right\|_*,$$

where in the last equality we applied **S.5**:. Subtracting $\left\|\mathbf{X}_0\right\|_*$ in the above inequality leads to

$$\left\|\mathbf{R}_0\right\|_* \geq \left\|\mathbf{R}_c\right\|_*. \tag{1.21}$$

Next we partition $\mathbf{R}_c$ into a sum of matrices $\mathbf{R}_1, \mathbf{R}_2, \ldots$ each of rank at most $3r$. Let $\mathbf{R}_c = \mathbf{U}\operatorname{diag}(\boldsymbol{\sigma})\mathbf{V}^T$ be the singular value decomposition of the matrix $\mathbf{R}_c$, where $\boldsymbol{\sigma}$ is a vector of the corresponding singular values in descending order. For each $i \geq 1$ define the index set $\boldsymbol{\mathcal{I}}_i = \{3r(i-1)+1, \ldots, 3ri\}$ and let $\mathbf{R}_i := \mathbf{U}_{\boldsymbol{\mathcal{I}}_i} \operatorname{diag}(\boldsymbol{\sigma}_{\boldsymbol{\mathcal{I}}_i}) \mathbf{V}_{\boldsymbol{\mathcal{I}}_i}^T$ (notice that additionally, $\langle \mathbf{R}_i, \mathbf{R}_j \rangle = 0$, whenever $i \neq j$). By construction, we have

$$\sigma_k \leq \frac{1}{3r} \sum_{j \in \boldsymbol{\mathcal{I}}_i} \boldsymbol{\sigma}_{\boldsymbol{\mathcal{I}}_i}(j), \quad \text{for all } k \in \boldsymbol{\mathcal{I}}_{i+1}$$

which implies $\left\|\mathbf{R}_{i+1}\right\|_F^2 \leq \frac{1}{3r}\left\|\mathbf{R}_i\right\|_*^2$. We can then compute bound

$$\sum_{j \geq 2} \left\|\mathbf{R}_j\right\|_F \leq \frac{1}{\sqrt{3r}} \sum_{j \geq 1} \left\|\mathbf{R}_j\right\|_* = \frac{1}{\sqrt{3r}} \left\|\mathbf{R}_c\right\|_* \leq \frac{1}{\sqrt{3r}} \left\|\mathbf{R}_0\right\|_* \leq \frac{\sqrt{2r}}{\sqrt{3r}} \left\|\mathbf{R}_0\right\|_F,$$

where the last inequality follows from **S.2**:.

Noticing that $\operatorname{rank}(\mathbf{R}_0 + \mathbf{R}_1) \leq 5r$ and putting our estimates together, we obtain the inequality

$$\begin{aligned} 0 = \|\mathcal{A}(\mathbf{R})\|_2 &\geq \|\mathcal{A}(\mathbf{R}_0 + \mathbf{R}_1)\|_2 - \sum_{j\geq 2} \|\mathcal{A}(\mathbf{R}_j)\|_2 \\ &\geq \sqrt{1-\delta_{5r}}\, \|\mathbf{R}_0 + \mathbf{R}_1\|_F - \sqrt{1+\delta_{3r}} \sum_{j\geq 2} \|\mathbf{R}_j\|_F \\ &\geq \left( \sqrt{1-\delta_{5r}} - \sqrt{\frac{2}{3}}\sqrt{1+\delta_{3r}} \right) \|\mathbf{R}_0\|_F . \end{aligned}$$

Therefore, if the factor on the right hand side is positive, then $\|\mathbf{R}_0\|_F = 0$, that is $\mathbf{R}_0 = \mathbf{0}$, which by (1.21) further implies that $\mathbf{R}_c = \mathbf{0}$. Thus $\mathbf{X}_* = \mathbf{X}_0$. The right hand side is positive when

$$3\delta_{3r} + 2\delta_{5r} < 1.$$

Since $\delta_{3r} \leq \delta_{5r}$ and by assumption $\delta_{5r} < \frac{1}{5}$, the statement follows. □

The construction of the $\mathbf{R}_i$'s above is again based on the existence of the singular value decomposition. In the following we show that linear operators satisfying the matrix-RIP exist. In particular, we can construct linear operators $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ that, with high probability, satisfy the RIP with optimal bounds on $m$. Together with Theorem 1.26, this implies that the low-rank matrix recovery is possible via nuclear norm minimization, i.e., via (1.19).

**Definition 1.29.** A linear map $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ with a corresponding matrix representation $\mathbf{A} \in \mathbb{R}^{m \times n_1 n_2}$ is a *Gaussian measurement ensemble* if each row $\mathbf{a}_{i\cdot}$ of $\mathbf{A}$ contains independent identically distributed $\mathcal{N}(0, 1/m)$ entries (and the $\mathbf{a}_{i\cdot}$'s are independent from each other).

**Theorem 1.30** ([22])**.** Fix $0 \leq \delta < 1$ and let $\mathcal{A}$ be a random measurement ensemble obeying the following condition: for any given $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ and any fixed $0 < t < 1$,

$$\mathbb{P}\left( \left| \|\mathcal{A}(\mathbf{X})\|_2^2 - \|\mathbf{X}\|_F^2 \right| > t \|\mathbf{X}\|_F^2 \right) \leq C e^{-ct^2 m} \tag{1.22}$$

for fixed constants $C, c > 0$. Then there exist constants $D, d > 0$ (which may depend on $t$) so that, if $m \geq D \max\{n_1, n_2\}\, r$, the measurement ensemble $\mathcal{A}$ satisfies $\delta_r \leq \delta$ with probability exceeding $1 - Ce^{-d\delta^2 m}$.

The concentration bound (1.22) is valid for various random measurement ensembles. For example, if $\mathcal{A}$ is Gaussian measurement ensemble, we have (see [22] for details)

$$\mathbb{P}\left( \left| \|\mathcal{A}(\mathbf{X})\|_2^2 - \|\mathbf{X}\|_F^2 \right| > t \|\mathbf{X}\|_F^2 \right) \leq 2e^{-\frac{m}{2}\left(t^2/2 - t^3/2\right)}. \tag{1.23}$$

As a consequence, a Gaussian measurement ensemble satisfies the matrix-RIP with constant $\delta_r \leq \delta \in (0,1)$ with high probability provided

$$m \geq C_\delta \max\{n_1, n_2\} r, \quad \text{with } C_\delta \sim C\delta^{-2},$$

where $C_\delta$ denotes a constant depending on $\delta$. Together, with Theorem 1.26, we obtain an exact low-rank matrix recovery result via nuclear norm minimization. Additionally, a linear map $\mathcal{A}$ satisfies the inequality (1.23) in the case where each row $\mathbf{a}_{i\cdot}$ has i.i.d. entries that take values $\pm\frac{1}{\sqrt{m}}$ with equal probability, or if $\mathcal{A}$ is a random projection [1, 133]. Further, $\mathcal{A}$ satisfies (1.22) if the rows $\mathbf{a}_{i\cdot}$ contain subgaussian entries (properly normalized) [162], although in this case the constants involved depend on the parameters of the subgaussian entries.

The proof of Theorem 1.30 uses a covering argument and, in particular, $\varepsilon$-nets.

**Definition 1.31** ([163])**.** A set $\mathcal{N}_\varepsilon^{\boldsymbol{\mathcal{X}}} \subset \boldsymbol{\mathcal{X}}$, where $\boldsymbol{\mathcal{X}}$ is a subset of a normed space, is called an $\varepsilon$-net of $\boldsymbol{\mathcal{X}}$ with respect to the norm $\|\cdot\|$ if for each $v \in \boldsymbol{\mathcal{X}}$, there exists $v_0 \in \mathcal{N}_\varepsilon^{\boldsymbol{\mathcal{X}}}$ with $\|v_0 - v\| \leq \varepsilon$. The minimal cardinality of an $\varepsilon$-net of $\boldsymbol{\mathcal{X}}$ with respect to the norm $\|\cdot\|$ is denoted by $\mathcal{N}(\boldsymbol{\mathcal{X}}, \|\cdot\|, \varepsilon)$ and is called the covering number of $\boldsymbol{\mathcal{X}}$ (at scale $\varepsilon$).

In the proof of Theorem 1.30 the following result on $\varepsilon$-nets is used.

**Lemma 1.32** ([163])**.** Let $\varepsilon \in (0,1)$. For any set $\boldsymbol{\mathcal{X}}$ there always exists an $\varepsilon$-net $\mathcal{N}_\varepsilon^{\boldsymbol{\mathcal{X}}}$ with respect to a norm $\|\cdot\|$ satisfying $\mathcal{N}_\varepsilon^{\boldsymbol{\mathcal{X}}} \subset \boldsymbol{\mathcal{X}}$ and

$$\left|\mathcal{N}_\varepsilon^{\boldsymbol{\mathcal{X}}}\right| \leq \frac{\operatorname{Vol}\left(\boldsymbol{\mathcal{X}} + \frac{\varepsilon}{2}\boldsymbol{\mathcal{B}}\right)}{\operatorname{Vol}\left(\frac{\varepsilon}{2}\boldsymbol{\mathcal{B}}\right)},$$

where $\frac{\varepsilon}{2}\boldsymbol{\mathcal{B}}$ is an $\varepsilon/2$ ball with respect to the norm $\|\cdot\|$ and $\boldsymbol{\mathcal{X}} + \frac{\varepsilon}{2}\boldsymbol{\mathcal{B}} = \left\{x + y : x \in \mathcal{N}_\varepsilon^{\boldsymbol{\mathcal{X}}}, y \in \frac{\varepsilon}{2}\boldsymbol{\mathcal{B}}\right\}$. Specifically, if $\boldsymbol{\mathcal{X}}$ is a subset of the unit ball in $d$ dimensions then $\boldsymbol{\mathcal{X}} + \frac{\varepsilon}{2}\boldsymbol{\mathcal{B}}$ is contained in the $\left(1 + \frac{\varepsilon}{2}\right)$-ball and thus

$$\left|\mathcal{N}_\varepsilon^{\boldsymbol{\mathcal{X}}}\right| \leq \frac{(1+\varepsilon/2)^d}{(\varepsilon/2)^d} = \left(1 + \frac{2}{\varepsilon}\right)^d < (3/\varepsilon)^d,$$

where the last inequality follows since $\varepsilon < 1$. We always require that $\mathcal{N}_\varepsilon^{\boldsymbol{\mathcal{X}}} \subset \boldsymbol{\mathcal{X}}$.

The first step in the proof of Theorem 1.30 is to compute a covering number of rank-$r$ matrices, for a fixed rank $r$.

**Lemma 1.33** (Covering number for low-rank matrices, [22])**.** Let

$$\boldsymbol{\mathcal{S}}_r = \left\{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2} : \operatorname{rank}(\mathbf{X}) \leq r, \|\mathbf{X}\|_F = 1\right\}.$$

Then there exists an $\varepsilon$-net $\mathcal{N}_\varepsilon^{\boldsymbol{\mathcal{S}}_r}$ for the Frobenius norm obeying

$$\mathcal{N}(\boldsymbol{\mathcal{S}}_r, \|\cdot\|_F, \varepsilon) \leq (9/\varepsilon)^{(n_1+n_2+1)r}.$$

Proof. For a matrix $\mathbf{X} \in \boldsymbol{\mathcal{S}}_r$, let $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*$ denote the reduced singular value decomposition. Since $\|\mathbf{X}\|_F = 1$, it follows that also $\|\boldsymbol{\Sigma}\|_F = 1$. Our argument constructs an $\varepsilon$-net for $\boldsymbol{\mathcal{S}}_r$ by covering the set of permissible $\mathbf{U}$, $\mathbf{V}$, and $\boldsymbol{\Sigma}$. We work in the simpler case where $n_1 = n_2 = n$ since the general case is a straightforward modification.

Let $\boldsymbol{\mathcal{D}} = \left\{\mathbf{D} \in \mathbb{R}^{r \times r} : \mathbf{D} \text{ diagonal}, \|\mathbf{D}\|_F = 1, \mathbf{D}(i,i) \geq 0 \text{ for all } i \in [n]\right\}$. We take $\mathcal{N}_{\varepsilon/3}^{\boldsymbol{\mathcal{D}}}$ to be an $\varepsilon/3$-net for $\boldsymbol{\mathcal{D}}$ with $\mathcal{N}(\boldsymbol{\mathcal{D}}, \|\cdot\|_F, \varepsilon/3) \leq (9/\varepsilon)^r$. Next, let $\boldsymbol{\mathcal{O}}_{n,r} = \left\{\mathbf{U} \in \mathbb{R}^{n \times r} : \mathbf{U}^*\mathbf{U} = \mathbf{I}\right\}$. To cover $\boldsymbol{\mathcal{O}}_{n,r}$, it is beneficial to use the norm $\|\cdot\|_{1,2}$ defined as

$$\|\mathbf{X}\|_{1,2} = \max_i \|\mathbf{x}_{\cdot i}\|_2,$$

where $\mathbf{x}_{\cdot i}$ denotes the $i$th column of $\mathbf{X}$. Let $\boldsymbol{\mathcal{Q}}_{n,r} = \left\{\mathbf{X} \in \mathbb{R}^{n \times r} : \|\mathbf{X}\|_{1,2} \leq 1\right\}$. It is easy to see that $\boldsymbol{\mathcal{O}}_{n,r} \subset \boldsymbol{\mathcal{Q}}_{n,r}$ since the columns of an orthogonal matrix are unit normed. We have seen that there is an $\varepsilon/3$-net $\mathcal{N}_{\varepsilon/3}^{\boldsymbol{\mathcal{O}}_{n,r}}$ for $\boldsymbol{\mathcal{O}}_{n,r}$ obeying $\mathcal{N}\left(\boldsymbol{\mathcal{O}}_{n,r}, \|\cdot\|_{1,2}, \varepsilon/3\right) \leq (9/\varepsilon)^{nr}$.

We now let $\overline{\boldsymbol{\mathcal{S}}}_r = \left\{\overline{\mathbf{U}}\,\overline{\boldsymbol{\Sigma}}\,\overline{\mathbf{V}}^* : \overline{\mathbf{U}}, \overline{\mathbf{V}} \in \mathcal{N}_{\varepsilon/3}^{\boldsymbol{\mathcal{O}}_{n,r}}, \overline{\boldsymbol{\Sigma}} \in \mathcal{N}_{\varepsilon/3}^{\boldsymbol{\mathcal{D}}}\right\}$ and remark that

$$\left|\overline{\boldsymbol{\mathcal{S}}}_r\right| \leq \left[\mathcal{N}\left(\boldsymbol{\mathcal{O}}_{n,r}, \|\cdot\|_{1,2}, \varepsilon/3\right)\right]^2 \mathcal{N}(\boldsymbol{\mathcal{D}}, \|\cdot\|_F, \varepsilon/3) \leq (9/\varepsilon)^{(2n+1)r}.$$

If we show that for all $\mathbf{X} \in \boldsymbol{\mathcal{S}}_r$ there exists $\overline{\mathbf{X}} \in \overline{\boldsymbol{\mathcal{S}}}_r$ with $\left\|\mathbf{X} - \overline{\mathbf{X}}\right\|_F \leq \varepsilon$ then $\overline{\boldsymbol{\mathcal{S}}}_r$ is an $\varepsilon$-net set for $\boldsymbol{\mathcal{S}}_r$ and thus, $\mathcal{N}\left(\boldsymbol{\mathcal{S}}_r, \|\cdot\|_F, \varepsilon\right) \leq \left|\overline{\boldsymbol{\mathcal{S}}}_r\right| \leq (9/\varepsilon)^{(2n+1)r}$.

Fix $\mathbf{X} \in \boldsymbol{\mathcal{S}}_r$ and let $\mathbf{X} = \mathbf{U\Sigma V}^*$ be its singular value decomposition. Then there exists $\overline{\mathbf{X}} = \overline{\mathbf{U}}\,\overline{\mathbf{\Sigma}}\,\overline{\mathbf{V}}^* \in \overline{\boldsymbol{\mathcal{S}}}_r$ with $\overline{\mathbf{U}}, \overline{\mathbf{V}} \in \mathcal{N}_{\varepsilon/3}^{\boldsymbol{\mathcal{O}}_{n,r}}$, $\overline{\mathbf{\Sigma}} \in \mathcal{N}_{\varepsilon/3}^{\boldsymbol{\mathcal{D}}}$ obeying $\left\|\mathbf{U} - \overline{\mathbf{U}}\right\|_{1,2} \leq \varepsilon/3$, $\left\|\mathbf{V} - \overline{\mathbf{V}}\right\|_{1,2} \leq \varepsilon/3$, and $\left\|\mathbf{\Sigma} - \overline{\mathbf{\Sigma}}\right\|_F \leq \varepsilon/3$. This gives

$$\begin{aligned}
\left\|\mathbf{X} - \overline{\mathbf{X}}\right\|_F &= \left\|\mathbf{U\Sigma V}^* - \overline{\mathbf{U}}\,\overline{\mathbf{\Sigma}}\,\overline{\mathbf{V}}^*\right\|_F \\
&= \left\|\mathbf{U\Sigma V}^* - \overline{\mathbf{U}}\mathbf{\Sigma V}^* + \overline{\mathbf{U}}\mathbf{\Sigma V}^* - \overline{\mathbf{U}}\,\overline{\mathbf{\Sigma}}\mathbf{V}^* + \overline{\mathbf{U}}\,\overline{\mathbf{\Sigma}}\mathbf{V}^* - \overline{\mathbf{U}}\,\overline{\mathbf{\Sigma}}\,\overline{\mathbf{V}}^*\right\|_F \\
&\leq \left\|\left(\mathbf{U} - \overline{\mathbf{U}}\right)\mathbf{\Sigma V}^*\right\|_F + \left\|\overline{\mathbf{U}}\left(\mathbf{\Sigma} - \overline{\mathbf{\Sigma}}\right)\mathbf{V}^*\right\|_F + \left\|\overline{\mathbf{U}}\,\overline{\mathbf{\Sigma}}\left(\mathbf{V} - \overline{\mathbf{V}}\right)^*\right\|_F.
\end{aligned}$$

For the first term, note that since $\mathbf{V}$ is an orthogonal matrix, $\left\|\left(\mathbf{U} - \overline{\mathbf{U}}\right)\mathbf{\Sigma V}^*\right\|_F = \left\|\left(\mathbf{U} - \overline{\mathbf{U}}\right)\mathbf{\Sigma}\right\|_F$ and

$$\left\|\left(\mathbf{U} - \overline{\mathbf{U}}\right)\mathbf{\Sigma}\right\|_F^2 = \sum_{1\leq i\leq r} \mathbf{\Sigma}\left(i,i\right)^2 \left\|\overline{\mathbf{u}}_{\cdot i} - \mathbf{u}_{\cdot i}\right\|_2^2 \leq \left\|\mathbf{\Sigma}\right\|_F^2 \left\|\mathbf{U} - \overline{\mathbf{U}}\right\|_{1,2}^2 \leq \left(\varepsilon/3\right)^2.$$

Hence, $\left\|\left(\mathbf{U} - \overline{\mathbf{U}}\right)\mathbf{\Sigma V}^*\right\|_F \leq \varepsilon/3$. The same argument gives $\left\|\overline{\mathbf{U}}\,\overline{\mathbf{\Sigma}}\left(\mathbf{V} - \overline{\mathbf{V}}\right)^*\right\|_F \leq \varepsilon/3$. To bound the middle term, observe that $\left\|\overline{\mathbf{U}}\left(\mathbf{\Sigma} - \overline{\mathbf{\Sigma}}\right)\mathbf{V}^*\right\|_F = \left\|\mathbf{\Sigma} - \overline{\mathbf{\Sigma}}\right\|_F \leq \varepsilon/3$. This completes the proof. □

We are now ready to present the proof of Theorem 1.30.

Proof of Theorem 1.30. The proof is essentially the same as the proof of Lemma 4.3 in [133]. We begin by showing that with high probability $\mathcal{A}$ is an approximate isometry on a net for $\boldsymbol{\mathcal{S}}_r$. Lemma 1.33 with $\varepsilon = \delta/\left(4\sqrt{2}\right)$ gives

$$\mathcal{N}\left(\boldsymbol{\mathcal{S}}_r, \|\cdot\|_F, (\delta/4\sqrt{2})\right) \leq \left(36\sqrt{2}/\delta\right)^{(n_1+n_2+1)r}.$$

Then it follows from (1.22) together with the union bound that

$$\begin{aligned}
\mathbb{P}\left(\max_{\overline{\mathbf{X}} \in \mathcal{N}_{\delta/4\sqrt{2}}^{\boldsymbol{\mathcal{S}}_r}} \left|\left\|\mathcal{A}\left(\overline{\mathbf{X}}\right)\right\|_2^2 - \left\|\overline{\mathbf{X}}\right\|_F^2\right| > \delta/2\right) &\leq \left|\overline{\boldsymbol{\mathcal{S}}}_r\right| Ce^{-c\delta^2 m} \\
&\leq \left(36\sqrt{2}/\delta\right)^{(n_1+n_2+1)r} Ce^{-c\delta^2 m} \\
&= Ce^{(n_1+n_2+1)r\log\left(36\sqrt{2}/\delta\right) - c\delta^2 m} \\
&\leq Ce^{-dm},
\end{aligned}$$

where $d = c - \frac{\log\left(36\sqrt{2}/\delta\right)}{\bar{D}} > 0$ if we choose $\bar{D} > \log\left(36\sqrt{2}/\delta\right)/c$ and $m \geq D\delta^{-2}\max\{n_1, n_2\}r \geq \bar{D}\delta^{-2}\left(n_1 + n_2 + 1\right)r$ (if we choose for example $D = 3\bar{D}$ for constant).

Now suppose that

$$\max_{\overline{\mathbf{X}} \in \mathcal{N}_{\delta/4\sqrt{2}}^{\boldsymbol{\mathcal{S}}_r}} \left|\left\|\mathcal{A}\left(\overline{\mathbf{X}}\right)\right\|_2^2 - \left\|\overline{\mathbf{X}}\right\|_F^2\right| \leq \delta/2 \tag{1.24}$$

(which occurs with probability at least $1 - Ce^{-dm}$). We begin by showing that the upper bound in the RIP condition holds. Set

$$\kappa_r = \sup_{\mathbf{X} \in \boldsymbol{\mathcal{S}}_r} \left\|\mathcal{A}\left(\mathbf{X}\right)\right\|_2.$$

For any $\mathbf{X} \in \boldsymbol{\mathcal{S}}_r$, there exists $\overline{\mathbf{X}} \in \mathcal{N}^{\boldsymbol{\mathcal{S}}_r}_{\delta/4\sqrt{2}}$ with $\left\|\mathbf{X} - \overline{\mathbf{X}}\right\|_F \leq \delta/\left(4\sqrt{2}\right)$ and, therefore, by (1.24)

$$\left\|\mathcal{A}\left(\mathbf{X}\right)\right\|_2 \leq \left\|\mathcal{A}\left(\mathbf{X} - \overline{\mathbf{X}}\right)\right\|_2 + \left\|\mathcal{A}\left(\overline{\mathbf{X}}\right)\right\|_2 \leq \left\|\mathcal{A}\left(\mathbf{X} - \overline{\mathbf{X}}\right)\right\|_2 + 1 + \delta/2. \tag{1.25}$$

Put $\Delta\mathbf{X} = \mathbf{X} - \overline{\mathbf{X}}$ and note that $\operatorname{rank}\left(\Delta\mathbf{X}\right) \leq 2r$. Write $\Delta\mathbf{X} = \Delta\mathbf{X}_1 + \Delta\mathbf{X}_2$, where $\langle \Delta\mathbf{X}_1, \Delta\mathbf{X}_2 \rangle = 0$ and $\operatorname{rank}\left(\Delta\mathbf{X}_i\right) \leq r$ for $i = 1, 2$ (for example by splitting the SVD). Note that $\Delta\mathbf{X}_1/\left\|\Delta\mathbf{X}_1\right\|_F$, $\Delta\mathbf{X}_2/\left\|\Delta\mathbf{X}_2\right\|_F \in \boldsymbol{\mathcal{S}}_r$. Thus

$$\left\|\mathcal{A}\left(\Delta\mathbf{X}\right)\right\|_2 \leq \left\|\mathcal{A}\left(\Delta\mathbf{X}_1\right)\right\|_2 + \left\|\mathcal{A}\left(\Delta\mathbf{X}_2\right)\right\|_2 \leq \kappa_r\left(\left\|\Delta\mathbf{X}_1\right\|_F + \left\|\Delta\mathbf{X}_2\right\|_F\right). \tag{1.26}$$

Now $\left\|\Delta\mathbf{X}_1\right\|_F + \left\|\Delta\mathbf{X}_2\right\|_F \leq \sqrt{2}\left\|\Delta\mathbf{X}\right\|_F$ which follows from $\left\|\Delta\mathbf{X}_1\right\|_F^2 + \left\|\Delta\mathbf{X}_2\right\|_F^2 = \left\|\Delta\mathbf{X}\right\|_F^2$. Also, $\left\|\Delta\mathbf{X}\right\|_F \leq \delta/\left(4\sqrt{2}\right)$ leading to $\left\|\mathcal{A}\left(\Delta\mathbf{X}\right)\right\|_2 \leq \kappa_r\delta/4$. Plugging this into (1.25) leads to

$$\left\|\mathcal{A}\left(\mathbf{X}\right)\right\|_2 \leq \kappa_r\delta/4 + 1 + \delta/2.$$

Since this holds for all $\mathbf{X} \in \boldsymbol{\mathcal{S}}_r$, we have $\kappa_r \leq \kappa_r\delta/4 + 1 + \delta/2$. Thus, $\kappa_r \leq \left(1 + \delta/2\right)/\left(1 - \delta/4\right) \leq 1 + \delta$, where in the last inequality we used that $\delta \in (0, 1)$, which essentially completes the upper bound. Now that this is established, the lower bound follows from

$$\left\|\mathcal{A}\left(\mathbf{X}\right)\right\|_2 \geq \left\|\mathcal{A}\left(\overline{\mathbf{X}}\right)\right\|_2 - \left\|\mathcal{A}\left(\Delta\mathbf{X}\right)\right\|_2 \geq 1 - \delta/2 - \left(1 + \delta\right)\delta/4 \geq 1 - \delta.$$

Thus, we have shown that for all $\mathbf{X} \in \boldsymbol{\mathcal{S}}_{\mathbf{r}}$ holds

$$\left(1 - \delta\right)\left\|\mathbf{X}\right\|_F \leq \left\|\mathcal{A}\left(\mathbf{X}\right)\right\|_2 \leq \left(1 + \delta\right)\left\|\mathbf{X}\right\|_F$$

which can easily be translated into the desired version of the RIP bound by taking squares and renaming $\delta$. □

Notice that, right after the inequality (1.25) the following property of the matrices is used. For any $2r$-rank matrix $\mathbf{X}$ there exist two matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ such that

(1) $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2$
(2) $\operatorname{rank}\left(\mathbf{X}_1\right) = \operatorname{rank}\left(\mathbf{X}_2\right) = r$
(3) $\langle \mathbf{X}_1, \mathbf{X}_2 \rangle = 0$.

Additionally, if $\mathbf{X} = \sum_{i=1}^{2r} \sigma_i \mathbf{u}_{\cdot i}\mathbf{v}_{\cdot i}^*$ is the singular value decomposition of a matrix $\mathbf{X}$, then the matrices

$$\mathbf{X}_1 = \sum_{i=1}^{r} \sigma_i \mathbf{u}_{\cdot i}\mathbf{v}_{\cdot i}^* \quad \text{and} \quad \mathbf{X}_2 = \sum_{i=r+1}^{2r} \sigma_i \mathbf{u}_{\cdot i}\mathbf{v}_{\cdot i}^*$$

satisfy the conditions (1)-(3). This again shows the power of the singular value decomposition.

**Remark 1.34.** We remark that the proofs of the RIP for random linear maps $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ and random matrices $\mathbf{A} \in \mathbb{R}^{m \times N}$ via generic chaining do not use the above properties of matrices and the analogue properties of vectors, respectively. For more details, see for example [48, 95, 127]. Even more, in Chapter 5 to show that certain measurement maps $\mathcal{A} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^m$ satisfy the notion of RIP for tensors we use the results obtained in [48, 95, 127].

We continue with a matrix analogue of the null space property introduced in Definition 1.1 which guarantees efficient recovery of low-rank matrices via nuclear norm minimization.

**Definition 1.35** ([60])**.** A linear map $\mathcal{A} : \mathbb{K}^{n_1 \times n_2} \to \mathbb{K}^m$ is said to satisfy the *stable rank null space property of order* $r$ *with constant* $0 < \rho < 1$ if

$$\sum_{j=1}^{r} \sigma_j \left(\mathbf{M}\right) < \rho \sum_{j=r+1}^{\min\{n_1,n_2\}} \sigma_j \left(\mathbf{M}\right), \quad \text{for all } \mathbf{M} \in \ker\left(\mathcal{A}\right) \backslash \{\mathbf{0}\},$$

where $\sigma_1 \left(\mathbf{M}\right) \geq \ldots \geq \sigma_{\min\{n_1,n_2\}} \left(\mathbf{M}\right) \geq 0$ denote the singular values of a matrix $\mathbf{M}$.

If $\rho = 1$, we say that $\mathcal{A}$ satisfies the *rank null space property of order* $r$.

We remark that the rank null space property of order $r$ is defined through the singular value decomposition of matrices in the kernel of $\mathcal{A}$. The following recovery theorems display the power of the rank null space property for low-rank matrix recovery. We present the results without proofs.

**Theorem 1.36** ([60])**.** Given a linear map $\mathcal{A} : \mathbb{K}^{n_1 \times n_2} \to \mathbb{K}^m$ every matrix $\mathbf{X} \in \mathbb{K}^{n_1 \times n_2}$ of rank at most $r$ is the unique solution of

$$\min_{\mathbf{Z} \in \mathbb{K}^{n_1 \times n_2}} \left\|\mathbf{Z}\right\|_* \quad \text{such that} \quad \mathcal{A}\left(\mathbf{Z}\right) = \mathcal{A}\left(\mathbf{X}\right) \tag{1.27}$$

if and only if $\mathcal{A}$ satisfies the rank null space property of order $r$.

The following theorem gives a recovery guarantee for the nuclear norm minimization problem under the assumption that the linear operator satisfies the stable rank null space property of order $r$.

**Theorem 1.37** ([60])**.** Let $\mathcal{A} : \mathbb{K}^{n_1 \times n_2} \to \mathbb{K}^m$ be a linear measurement map satisfying the stable rank null space property of order $r$ with constant $0 < \rho < 1$. Let $\mathbf{X}^{\#}$ be a solution of the nuclear norm minimization problem (1.27). Then

$$\left\|\mathbf{X} - \mathbf{X}^{\#}\right\|_* \leq \frac{2\left(1+\rho\right)}{1-\rho} \sum_{\ell=r+1}^{\min\{n_1,n_2\}} \sigma_\ell \left(\mathbf{X}\right).$$

Recovery algorithms with the assumption that the corresponding linear operator satisfies the (stable) rank null space property or a version of it have been studied in several papers. For example, in the papers [58] and [114] different versions of the *iterative reweighted least squares algorithm* have been analyzed. In the paper [27] the *matrix completion* problem has been treated. This problem is a special case of the nuclear norm minimization problem, namely

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}} \left\|\mathbf{Z}\right\|_* \quad \text{such that} \quad \mathcal{P}_{\boldsymbol{\mathcal{M}}}\left(\mathbf{Z}\right) = \mathcal{P}_{\boldsymbol{\mathcal{M}}}\left(\mathbf{X}\right),$$

where $\mathcal{P}_{\boldsymbol{\mathcal{M}}} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ denotes the orthogonal projection onto the subspace of matrices which vanish outside of $\boldsymbol{\mathcal{M}}$. That is, $\mathbf{Y} = \mathcal{P}_{\boldsymbol{\mathcal{M}}}\left(\mathbf{X}\right)$ is defined as

$$\mathbf{Y}\left(i,j\right) = \begin{cases} \mathbf{X}\left(i,j\right), & \text{if } \left(i,j\right) \in \boldsymbol{\mathcal{M}} \\ 0, & \text{otherwise.} \end{cases}$$

Similarly to the vector case, a version of the coherence for low-rank matrix recovery has been introduced.

**Definition 1.38** ([23])**.** Let $\mathbf{U}$ be a subspace of $\mathbb{K}^n$ of dimension $r$ and $\mathbf{P_U}$ be the orthogonal projection onto $\mathbf{U}$. Then the coherence of $\mathbf{U}$ (relative to the standard basis $(\mathbf{e}_i)$) is defined to be

$$\mu\left(\mathbf{U}\right) := \frac{n}{r} \max_{1 \leq i \leq n} \left\|\mathbf{P_U}\mathbf{e}_i\right\|_2^2.$$

To state the main result of the paper [23], we introduce two assumptions about an $n_1 \times n_2$ matrix $\mathbf{X}$ whose SVD is given by $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^* = \sum_{i=1}^{r} \sigma_i \mathbf{u}_{\cdot i}\mathbf{v}_{\cdot i}^*$.

**N1** The coherences obey $\max\left(\mu\left(\mathbf{U}\right), \mu\left(\mathbf{V}\right)\right) \leq \mu_0$, for some positive $\mu_0$.

**N2** The $n_1 \times n_2$ matrix $\sum_{i=1}^{r} \mathbf{u}_{\cdot i}\mathbf{v}_{\cdot i}^*$ has a maximum entry bounded by $\mu_1\sqrt{r/\left(n_1 n_2\right)}$ in absolute value for some positive $\mu_1$.

The parameters $\mu_0$ and $\mu_1$ may depend on $r$, $n_1$, and $n_2$. Moreover, note that **N2** always holds with $\mu_1 = \mu_0\sqrt{r}$ (to see this, apply the Cauchy-Schwarz inequality to the entries of $\sum_{i=1}^{r} \mathbf{u}_{\cdot i}\mathbf{v}_{\cdot i}^*$).

The following recovery result is stated under the coherence assumption without a proof.

**Theorem 1.39** ([23])**.** *Let $\mathbf{X}$ be an $n_1 \times n_2$ matrix of rank $r$ obeying* **N1** *and* **N2** *and let $n = \max\{n_1, n_2\}$. Suppose that we observe $m$ entries of $\mathbf{X}$ whose locations are sampled uniformly at random. Then there exist constants $C, c$ such that if*

$$m \geq C \max\{\mu_1^2, \mu_0^{1/2}\mu_1, \mu_0 n^{1/4}\} nr \left(\beta \log n\right)$$

*for some $\beta > 2$, then the minimizer of the problem*

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_* \qquad \text{subject to} \qquad \mathbf{Z}\left(i,j\right) = \mathbf{X}\left(i,j\right),\ \left(i,j\right) \in \boldsymbol{\mathcal{M}}$$

*is unique and equal to $\mathbf{X}$ with probability at least $1 - cn^{-\beta}$. For $r \leq \mu_0^{-1} n^{1/5}$ this estimate can be improved to*

$$m \geq C\mu_0 n^{6/5} r \left(\beta \log n\right)$$

*with the same probability of success.*

Several subsequent papers analyzing matrix completion appeared under the assumption that the coherence (or a version of the coherence) of the low-rank matrix we want to recover is small enough. For example, the approach in paper [27] is based on a slightly different assumption on the matrix $\mathbf{X}$ called the *strong incoherence property.* In addition, papers [34, 72] present provable matrix completion results for incoherent matrices under a uniform sampling model via nuclear norm minimization.

Other algorithms applied to low-rank matrix recovery have followed, including the *local descent method* [91], the *alternating projections algorithm* [86], and the *Atomic Decomposition for Minimum Rank Approximation (ADMiRA)* [103].

**Algorithm 1.3.** Iterative hard thresholding algorithm for matrices

```
Input: Measurement map $\mathcal{A}: \mathbb{K}^{n_1 \times n_2} \to \mathbb{K}^m$, measurement vector $\mathbf{y} \in \mathbb{K}^m$,
        rank $r$.
Initialization: low-rank matrix $\mathbf{X}^0$, typically $\mathbf{X}^0 = \mathcal{H}_r^M\left(\mathcal{A}^*\left(\mathbf{y}\right)\right)$, $j = 0$.
Repeat until the stopping criterion is met at $j = \bar{j}$
        $\mathbf{X}^{j+1} = \mathcal{H}_r^M\left(\mathbf{X}^j + \mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right)\right)$
        $j = j + 1$
Output: rank-$r$ matrix $\mathbf{X}^\# = \mathbf{X}^{\bar{j}}$
```

**1.2.1. Matrix iterative hard thresholding algorithm.** In this subsection we introduce several versions of the IHT algorithm adapted matrix scenario. The thresholding operator $\mathcal{H}_r^M$

**Algorithm 1.4.** Normalized iterative hard thresholding algorithm for matrices

```
Input: Measurement map $\mathcal{A} : \mathbb{K}^{n_1 \times n_2} \to \mathbb{K}^m$, measurement vector $\mathbf{y} \in \mathbb{K}^m$,
       rank $r$.
Initialization: low-rank matrix $\mathbf{X}^0$, typically $\mathbf{X}^0 = \mathcal{H}_r^M\left(\mathcal{A}^*\left(\mathbf{y}\right)\right)$, $j = 0$.
Repeat until the stopping criterion is met at $j = \overline{j}$
        Set the projection operator $\mathbf{P}_{\mathbf{U}}^j := \mathbf{U}_j \mathbf{U}_j^*$
        Compute the stepsize $\mu_j^{\mathbf{U}} := \frac{\left\|\mathbf{P}_{\mathbf{U}}^j \mathcal{A}^*\left(\mathbf{y}-\mathcal{A}\left(\mathbf{X}^j\right)\right)\right\|_F^2}{\left\|\mathcal{A}\left(\mathbf{P}_{\mathbf{U}}^j \mathcal{A}^*\left(\mathbf{y}-\mathcal{A}\left(\mathbf{X}^j\right)\right)\right)\right\|_2^2}$
        $\mathbf{X}^{j+1} = \mathcal{H}_r^M\left(\mathbf{X}^j + \mu_j^{\mathbf{U}} \mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right)\right)$
        Let $\mathbf{U}_{j+1}$ be the top $r$ left singular vectors of $\mathbf{X}^{j+1}$
        $j = j + 1$
Output: rank-$r$ matrix $\mathbf{X}^{\#} = \mathbf{X}^{\overline{j}}$
```

returns the best rank-$r$ approximation of a given matrix. If $\mathbf{X} = \sum_{i=1}^{\min\{n_1,n_2\}} \sigma_i \mathbf{u}_{\cdot i} \mathbf{v}_{\cdot i}^T$ is the singular value decomposition of a matrix $\mathbf{X}$, with singular values $\{\sigma_i\}_{i=1}^{\min\{n_1,n_2\}}$ arranged in decreasing order, then by the Eckart-Young theorem [54, 87] the best rank-$r$ approximation of the matrix $\mathbf{X}$ is $\mathcal{H}_r^M\left(\mathbf{X}\right) = \sum_{i=1}^{r} \sigma_i \mathbf{u}_{\cdot i} \mathbf{v}_{\cdot i}^T$. The matrix IHT algorithm is presented in Algorithm 1.3.

Matrix NIHT algorithm is motivated similarly to the NIHT algorithm for compressive sensing. When the matrix IHT algorithm is converging to a minimum rank solution $\mathbf{X}^{\#}$, each of the singular values and singular vectors of the current estimate must also be converging to the singular values and singular vectors of $\mathbf{X}^{\#}$. If the singular vectors of the iterate $\mathbf{X}^j$ have been correctly identified, then the update is being used to improve the singular values. Let $\mathbf{X}^j = \mathbf{U}_j \mathbf{\Sigma}_j \mathbf{V}_j^*$ be the singular value decomposition of the rank-$r$ iterate $\mathbf{X}^j$ and let $\mathbf{P}_{\mathbf{U}}^j := \mathbf{U}_j \mathbf{U}_j^*$ and $\mathbf{P}_{\mathbf{V}}^j := \mathbf{V}_j \mathbf{V}_j^*$ denote the projection onto the left and right singular vector spaces, respectively. A search direction can be projected onto the span of singular vectors by applying $\mathbf{P}_{\mathbf{U}}^j$ from the left and by applying $\mathbf{P}_{\mathbf{V}}^j$ from the right. For instance, the projective negative gradient descent direction is given by $\mathbf{W}_j^{\mathbf{UV}} := \mathbf{P}_{\mathbf{U}}^j \mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right) \mathbf{P}_{\mathbf{V}}^j$. Alternatively, a search direction can be projected to the span of just left or right singular vectors leading to the search directions $\mathbf{W}_j^{\mathbf{U}} := \mathbf{P}_{\mathbf{U}}^j \mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right)$ and $\mathbf{W}_j^{\mathbf{V}} := \mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right) \mathbf{P}_{\mathbf{V}}^j$, respectively. This projected directions should not be used as the update direction since they would not allow the iterates to converge to the lowest rank solution unless the projected directions are already correctly identified. However, as in the case of NIHT for compressive sensing, we use them for selecting the stepsize. Thus, this leads to three choices for the stepsize – namely, $\mu_j^{\mathbf{U}} := \frac{\left\|\mathbf{P}_{\mathbf{U}}^j \mathcal{A}^*\left(\mathbf{y}-\mathcal{A}\left(\mathbf{X}^j\right)\right)\right\|_F^2}{\left\|\mathcal{A}\left(\mathbf{P}_{\mathbf{U}}^j \mathcal{A}^*\left(\mathbf{y}-\mathcal{A}\left(\mathbf{X}^j\right)\right)\right)\right\|_2^2}$, $\mu_j^{\mathbf{V}} := \frac{\left\|\mathcal{A}^*\left(\mathbf{y}-\mathcal{A}\left(\mathbf{X}^j\right)\right)\mathbf{P}_{\mathbf{V}}^j\right\|_F^2}{\left\|\mathcal{A}\left(\mathcal{A}^*\left(\mathbf{y}-\mathcal{A}\left(\mathbf{X}^j\right)\right)\mathbf{P}_{\mathbf{V}}^j\right)\right\|_2^2}$, and $\mu_j^{\mathbf{UV}} := \frac{\left\|\mathbf{P}_{\mathbf{U}}^j \mathcal{A}^*\left(\mathbf{y}-\mathcal{A}\left(\mathbf{X}^j\right)\right)\mathbf{P}_{\mathbf{V}}^j\right\|_F^2}{\left\|\mathcal{A}\left(\mathbf{P}_{\mathbf{U}}^j \mathcal{A}^*\left(\mathbf{y}-\mathcal{A}\left(\mathbf{X}^j\right)\right)\mathbf{P}_{\mathbf{V}}^j\right)\right\|_2^2}$.

**Remark 1.40.** In paper [152] numerical experiments on matrix completion have been performed comparing the three versions of the NIHT algorithm. Their numerical results suggest that in general, the version of NIHT with stepsize $\mu_j^{\mathbf{U}}$ outperforms the other two variants. In particular, the version of NIHT with this stepsize was able to recover matrices of the same or larger rank than the other two variants. Thus, in the following we focus on NIHT with stepsize $\mu_j^{\mathbf{U}}$ presented in Algorithm 1.4 which in the following we just refer to as NIHT algorithm. However, we remark that the convergence guarantees of the other variants can be obtained analogously.

The following theorem gives a convergence criterion for the normalized IHT algorithm, presented in Algorithm 1.4, under the matrix-RIP assumption.

**Theorem 1.41** ([152])**.** For $a \in (0,1)$, let $\mathcal{A} : \mathbb{K}^{n_1 \times n_2} \to \mathbb{K}^m$ satisfy the matrix RIP with

$$\delta_{3r} < \frac{a}{a+4}$$

and let $\mathbf{X} \in \mathbb{K}^{n_1 \times n_2}$ be a matrix of rank at most $r$. Given measurements $\mathbf{y} = \mathcal{A}(\mathbf{X})$, matrix $\mathbf{X}^{j+1}$ obtained in the $j$-th iteration of matrix NIHT algorithm satisfies

$$\left\|\mathbf{X}^{j+1} - \mathbf{X}\right\|_F \leq a^j \left\|\mathbf{X}^0 - \mathbf{X}\right\|_F .$$

In other words, the sequence $(\mathbf{X}^j)_j$ produced by matrix NIHT algortihm converges linearly to $\mathbf{X}$.

Proof. Let $\mathbf{X}_0$ be the original rank $r$ matrix, i.e. $\mathbf{y} = \mathcal{A}(\mathbf{X}_0)$. With $\mathbf{W}^j := \mathbf{X}^j + \mu_j^{\mathbf{U}} \mathcal{A}^* \left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right)$ we denote the intermediate update to $\mathbf{X}^j$ in NIHT.

By Eckart-Young theorem,

$$\left\|\mathbf{W}^j - \mathbf{X}^{j+1}\right\|_F^2 \leq \left\|\mathbf{W}^j - \mathbf{X}_0\right\|_F^2$$

Expanding the left hand side we obtain

$$\begin{aligned}\left\|\mathbf{W}^j - \mathbf{X}_0\right\|_F^2 \geq \left\|\mathbf{W}^j - \mathbf{X}^{j+1}\right\|_F^2 &= \left\|\mathbf{W}^j - \mathbf{X}_0 + \mathbf{X}_0 - \mathbf{X}^{j+1}\right\|_F^2 \\ &= \left\|\mathbf{W}^j - \mathbf{X}_0\right\|_F^2 + \left\|\mathbf{X}_0 - \mathbf{X}^{j+1}\right\|_F^2 + 2\left\langle \mathbf{W}^j - \mathbf{X}_0, \mathbf{X}_0 - \mathbf{X}^{j+1}\right\rangle .\end{aligned}$$

Canceling the term $\left\|\mathbf{W}^j - \mathbf{X}_0\right\|_F$ in the above inequality leads to

$$\begin{aligned}\left\|\mathbf{X}_0 - \mathbf{X}^{j+1}\right\|_F^2 &\leq 2\left\langle \mathbf{W}^j - \mathbf{X}_0, \mathbf{X}^{j+1} - \mathbf{X}_0\right\rangle \\ &= 2\left\langle \mathbf{X}^j + \mu_j^{\mathbf{U}} \mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right) - \mathbf{X}_0, \mathbf{X}^{j+1} - \mathbf{X}_0\right\rangle \\ &= 2\left\langle \mathbf{X}^j - \mathbf{X}_0, \mathbf{X}^{j+1} - \mathbf{X}_0\right\rangle - 2\mu_j^{\mathbf{U}}\left\langle \mathcal{A}^*\left(\mathcal{A}\left(\mathbf{X}^j\right) - \mathbf{y}\right), \mathbf{X}^{j+1} - \mathbf{X}_0\right\rangle \\ &= 2\left\langle \mathbf{X}^j - \mathbf{X}_0, \mathbf{X}^{j+1} - \mathbf{X}_0\right\rangle - 2\mu_j^{\mathbf{U}}\left\langle \mathcal{A}^*\mathcal{A}\left(\mathbf{X}^j - \mathbf{X}_0\right), \mathbf{X}^{j+1} - \mathbf{X}_0\right\rangle \\ &= 2\left\langle \mathbf{X}^j - \mathbf{X}_0, \mathbf{X}^{j+1} - \mathbf{X}_0\right\rangle - 2\mu_j^{\mathbf{U}}\left\langle \mathcal{A}\left(\mathbf{X}^j - \mathbf{X}_0\right), \mathcal{A}\left(\mathbf{X}^{j+1} - \mathbf{X}_0\right)\right\rangle . \qquad (1.28)\end{aligned}$$

Let $\mathbf{Q}_j \in \mathbb{K}^{m \times 3r}$ be a matrix whose columns form orthonormal basis of the space spanned by the columns spaces of $\mathbf{X}_0, \mathbf{X}^j, \mathbf{X}^{j+1}$ and let $\mathbf{P}_{\mathbf{Q}}^j := \mathbf{Q}_j \mathbf{Q}_j^*$ be a projection operator onto a column space of $\mathbf{Q}_j$. Thus, $\mathbf{P}_{\mathbf{Q}}^j \mathbf{X}_0 = \mathbf{X}_0$, $\mathbf{P}_{\mathbf{Q}}^j \mathbf{X}^j = \mathbf{X}^j$, and $\mathbf{P}_{\mathbf{Q}}^j \mathbf{X}^{j+1} = \mathbf{X}^{j+1}$.

Next, we define a linear operator $\mathcal{A}_{\mathbf{Q}}^j(\mathbf{Z}) = \mathcal{A}\left(\mathbf{P}_{\mathbf{Q}}^j \mathbf{Z}\right)$ which is obtained by replacing the sensing matrices $\{\mathbf{A}_\ell\}_{\ell=1}^m$ of $\mathcal{A}$ with the sensing matrices $\left\{\mathbf{P}_{\mathbf{Q}}^j \mathbf{A}_\ell\right\}_{\ell=1}^m$ and the corresponding adjoint

$$\mathcal{A}_{\mathbf{Q}}^{j*}(\mathbf{y}) = \sum_{\ell=1}^m \mathbf{y}(\ell)\left(\mathbf{P}_{\mathbf{Q}}^j \mathbf{A}_\ell\right).$$

We continue the estimation (1.28)

$$\begin{aligned}\left\|\mathbf{X}^{j+1} - \mathbf{X}_0\right\|_F^2 &\leq 2\left\langle \mathbf{X}^j - \mathbf{X}_0, \mathbf{X}^{j+1} - \mathbf{X}_0\right\rangle - 2\mu_j^{\mathbf{U}}\left\langle \mathcal{A}\left(\mathbf{X}^j - \mathbf{X}_0\right), \mathcal{A}\left(\mathbf{X}^{j+1} - \mathbf{X}_0\right)\right\rangle \\ &= 2\left\langle \mathbf{X}^j - \mathbf{X}_0, \mathbf{X}^{j+1} - \mathbf{X}_0\right\rangle - 2\mu_j^{\mathbf{U}}\left\langle \mathcal{A}_{\mathbf{Q}}^j\left(\mathbf{X}^j - \mathbf{X}_0\right), \mathcal{A}_{\mathbf{Q}}^j\left(\mathbf{X}^{j+1} - \mathbf{X}_0\right)\right\rangle \\ &= 2\left\langle \mathbf{X}^j - \mathbf{X}_0, \left(\mathbf{X}^{j+1} - \mathbf{X}_0\right) - \mu_j^{\mathbf{U}} \mathcal{A}_{\mathbf{Q}}^{j*}\mathcal{A}_{\mathbf{Q}}^j\left(\mathbf{X}^{j+1} - \mathbf{X}_0\right)\right\rangle \\ &= 2\left\langle \mathbf{X}^j - \mathbf{X}_0, \left(\mathbf{I} - \mu_j^{\mathbf{U}} \mathcal{A}_{\mathbf{Q}}^{j*}\mathcal{A}_{\mathbf{Q}}^j\right)\left(\mathbf{X}^{j+1} - \mathbf{X}_0\right)\right\rangle\end{aligned}$$

$$\leq 2 \left\| \mathbf{I} - \mu_j^{\mathbf{U}} \mathcal{A}_{\mathbf{Q}}^{j*} \mathcal{A}_{\mathbf{Q}}^{j} \right\|_{2\to 2} \left\| \mathbf{X}^j - \mathbf{X}_0 \right\|_F \left\| \mathbf{X}^{j+1} - \mathbf{X}_0 \right\|_F .$$

Canceling the term $\left\| \mathbf{X}^{j+1} - \mathbf{X}_0 \right\|_F$ in the above inequality leads to

$$\left\| \mathbf{X}^{j+1} - \mathbf{X}_0 \right\|_F \leq 2 \left\| \mathbf{I} - \mu_j^{\mathbf{U}} \mathcal{A}_{\mathbf{Q}}^{j*} \mathcal{A}_{\mathbf{Q}}^{j} \right\|_{2\to 2} \left\| \mathbf{X}^j - \mathbf{X}_0 \right\|_F . \tag{1.29}$$

It remains to bound the term $\left\| \mathbf{I} - \mu_j^{\mathbf{U}} \mathcal{A}_{\mathbf{Q}}^{j*} \mathcal{A}_{\mathbf{Q}}^{j} \right\|_{2\to 2}$ by a constant smaller than $a/2$. Since $\mathbf{P}_{\mathbf{Q}}^j$ maps onto rank-$3r$ matrices, the matrix RIP implies that every eigenvalue of the self-adjoint operator $\mathcal{A}_{\mathbf{Q}}^{j*} \mathcal{A}_{\mathbf{Q}}^{j}$ is contained in the interval $[1-\delta_{3r}, 1+\delta_{3r}]$.

Additionally, we can bound the stepsize $\mu_j^{\mathbf{U}}$

$$\frac{1}{1+\delta_r} \leq \mu_j^{\mathbf{U}} = \frac{\left\| \mathbf{P}_{\mathbf{U}}^j \mathcal{A}^* \left( \mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right) \right\|_F^2}{\left\| \mathcal{A}\left( \mathbf{P}_{\mathbf{U}}^j \mathcal{A}^* \left( \mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right)\right) \right\|_2^2} \leq \frac{1}{1-\delta_r}.$$

Therefore, every eigenvalue of operator $\mathbf{I} - \mu_j^{\mathbf{U}} \mathcal{A}_{\mathbf{Q}}^{j*} \mathcal{A}_{\mathbf{Q}}^{j}$ is contained in $\left[1 - \frac{1+\delta_{3r}}{1-\delta_r}, 1 - \frac{1-\delta_{3r}}{1+\delta_r}\right]$. Since the magnitude of the lower bound is greater than the upper bound and since $\delta_{3r} < \frac{a}{a+4}$ it follows that

$$\left\| \mathbf{I} - \mu_j^{\mathbf{U}} \mathcal{A}_{\mathbf{Q}}^{j*} \mathcal{A}_{\mathbf{Q}}^{j} \right\|_{2\to 2} \leq \frac{1+\delta_{3r}}{1-\delta_r} - 1 < \frac{a}{2}$$

which concludes the proof. □

**Remark 1.42.** Let $a \in (0,1)$. The proof that matrix IHT algorithm (Algorithm 1.3) recovers any rank-$r$ matrix under the assumption that the isometry constant of $\mathcal{A}$ satisfies $\delta_{3r} < a/2$ can be done in an analogous way. To be more precise, the proof is the same as the proof of Theorem 1.41 up to (1.29). As before, one has to bound the term $\left\| \mathbf{I} - \mu_j^{\mathbf{U}} \mathcal{A}_{\mathbf{Q}}^{j*} \mathcal{A}_{\mathbf{Q}}^{j} \right\|_{2\to 2}$ by a constant smaller than $\frac{a}{2}$, where $\mu_j^{\mathbf{U}} = 1$. However, notice that

$$\begin{aligned} \left\| \mathbf{I} - \mathcal{A}_{\mathbf{Q}}^{j*} \mathcal{A}_{\mathbf{Q}}^{j} \right\|_{2\to 2} &= \sup_{\|\mathbf{X}\|_F = 1} \left| \left\langle (\mathbf{I} - \mathcal{A}_{\mathbf{Q}}^{j*} \mathcal{A}_{\mathbf{Q}}^{j})\mathbf{X}, \mathbf{X} \right\rangle \right| = \sup_{\|\mathbf{X}\|_F = 1} \left| \|\mathbf{X}\|_F^2 - \|\mathcal{A}_{\mathbf{Q}}^{j}(\mathbf{X})\|_2^2 \right| \\ &\leq \sup_{\|\mathbf{X}\|_F = 1, \operatorname{rank}(\mathbf{X}) \leq 3r} \left| \|\mathbf{X}\|_F^2 - \|\mathcal{A}(\mathbf{X})\|_2^2 \right| = \delta_{3r}. \end{aligned}$$

Applying the assumption $\delta_{3r} < \frac{a}{2}$ concludes the proof.

Theorem 1.41 can also be extended to the more general setting, with noisy measurements.

**Theorem 1.43.** For $a \in (0,1)$, let $\mathcal{A} : \mathbb{K}^{n_1 \times n_2} \to \mathbb{K}^m$ satisfy the matrix RIP with

$$\delta_{3r} < \frac{a}{a+4}$$

and let $\mathbf{X} \in \mathbb{K}^{n_1 \times n_2}$ be a matrix of rank at most $r$. Given the noisy measurements $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \mathbf{e}$ for some $\mathbf{e} \in \mathbb{K}^m$, the matrix $\mathbf{X}^{j+1}$ produced by matrix NIHT algortihm in $j$th iteration satisfies

$$\left\| \mathbf{X}^{j+1} - \mathbf{X} \right\|_F \leq a^j \left\| \mathbf{X}^0 - \mathbf{X} \right\|_F + \frac{b(a)}{1-a} \|\mathbf{e}\|_2 ,$$

where $b(a) := 2\frac{\sqrt{1+a/2}}{1-a/2}$. Consequently, if $\mathbf{e} \neq \mathbf{0}$ then after at most $j^* = \lceil \log_{1/a}(\left\| \mathbf{X}^0 - \mathbf{X} \right\|_F / \|\mathbf{e}\|_2)$ iterations, $\mathbf{X}^{j+1}$ estimates $\mathbf{X}$ with accuracy

$$\left\| \mathbf{X}^{j^*+1} - \mathbf{X} \right\|_F \leq \frac{1+a+b(a)}{1-a} \|\mathbf{e}\|_2 .$$

The proof of this theorem is analogous to the proof of Theorem 1.41. In estimate (1.28) we get an additional term $2\mu_j^{\mathbf{U}} \left\langle \mathcal{A}^*\mathbf{e}, \mathbf{X}^{j+1} - \mathbf{X}_0 \right\rangle$. Estimating this term leads to

$$\begin{aligned}\left\langle \mathcal{A}^*\mathbf{e}, \mathbf{X}^{j+1} - \mathbf{X}_0 \right\rangle = \left\langle \mathbf{e}, \mathcal{A}(\mathbf{X}^{j+1} - \mathbf{X}_0) \right\rangle &\leq \left\| \mathcal{A}(\mathbf{X}^{j+1} - \mathbf{X}_0) \right\|_F \|\mathbf{e}\|_2 \\ &\leq \sqrt{1+\delta_{3r}} \left\| \mathbf{X}^{j+1} - \mathbf{X} \right\|_F \|\mathbf{e}\|_2 .\end{aligned}$$

Finally, the estimate (1.29) translates into

$$\left\| \mathbf{X}^{j+1} - \mathbf{X}_0 \right\|_F \leq 2 \left\| \mathbf{I} - \mu_j^{\mathbf{U}} \mathcal{A}_{\mathbf{Q}}^{j*} \mathcal{A}_{\mathbf{Q}}^{j} \right\|_{2\to 2} \left\| \mathbf{X}^j - \mathbf{X}_0 \right\|_F + 2\mu_j^{\mathbf{U}} \sqrt{1+\delta_{3r}} \|\mathbf{e}\|_2 .$$

Plugging into all the assumptions leads to the stated result.

Similarly to the vector scenario, we can state a result also if the tensor $\mathbf{X}$ is not necessarily rank-$r$.

**Theorem 1.44.** Fix $a \in (0,1)$. Let $\mathcal{A} : \mathbb{K}^{n_1 \times n_2} \to \mathbb{K}^m$ satisfy the matrix RIP with

$$\delta_{6s} < \frac{a}{a+4}$$

and let $\mathbf{X} \in \mathbb{K}^{n_1 \times n_2}$. Then for all $\mathbf{X} \in \mathbb{K}^{n_1 \times n_2}$, $\mathbf{e} \in \mathbb{C}^m$, the sequence $(\mathbf{X}^j)_j$ defined by NIHT algorithm, with $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \mathbf{e}$, $\mathbf{X}^0 = \mathbf{0}$, and $\mathcal{H}_{\mathbf{r}}^M$ replaced by $\mathcal{H}_{\mathbf{2r}}^M$ satisfies

$$\left\| \mathbf{X} - \mathbf{X}^{j+1} \right\|_F \leq \frac{C}{\sqrt{s}} \|\mathbf{X} - \mathbf{X}_{\mathbf{r}}\|_* + D \|\mathbf{e}\|_2 + 2\rho^{j+1} \|\mathbf{X}\|_F ,$$

where constants $C, D > 0$ and $\rho \in (0,1)$ depend only on $a$.

The theorem is a consequence of Theorem 1.43 and the following lemma.

**Lemma 1.45.** Suppose $\mathcal{A} : \mathbb{K}^{n_1 \times n_2} \to \mathbb{K}^N$ has restricted isometry constants $\delta_s < 1$. Given $\tau > 0$, $\xi \geq 0$, and $\mathbf{e} \in \mathbb{K}^m$ assume that two matrices $\mathbf{X}, \mathbf{X}' \in \mathbb{K}^{n_1 \times n_2}$ satisfy

$$\operatorname{rank}(\mathbf{X}') \leq 2r \quad \text{and} \quad \|\mathbf{X}_{2r} - \mathbf{X}'\|_F \leq \tau \|\mathcal{A}(\mathbf{X}_{c,2r}) + \mathbf{e}\|_2 + \xi,$$

where $\mathbf{X}_{2r}$ denotes the best rank-$2r$ approximation to $\mathbf{X}$ and $\mathbf{X}_{c,2r} = \mathbf{X} - \mathbf{X}_{2r}$. Then

$$\|\mathbf{X} - \mathbf{X}'\|_F \leq \frac{C}{\sqrt{r}} \|\mathbf{X} - \mathbf{X}_r\|_* + \tau \|\mathbf{e}\|_2 + \xi, \tag{1.30}$$

where $\mathbf{X}_r$ denotes the best rank-$r$ approximation to $\mathbf{X}$ and constant $C = 1 + \sqrt{2}\tau$.

Proof. The proof is analogous to the proof of Lemma 1.20. □

In paper [85] a slightly more generalized version of the matrix IHT algorithm is presented under the name *Singular Value Projection* (SVP). In the SVP algorithm a stepsize $\mu_j$ is given in advance and often is a fixed constant $\mu$. Thus, the matrix IHT algorithm is a special case of the SVP algorithm with $\mu_j = 1$, for all iterations $j$. The stopping criterion is met in iteration $j$ if

$$\left\| \mathcal{A}(\mathbf{X}^{j+1}) - \mathbf{y} \right\|_2^2 \leq \varepsilon,$$

where a fixed tolerance $\varepsilon > 0$ is chosen beforehand. The main result of the paper is presented next.

**Theorem 1.46** ([85])**.** Fix tolerance $\varepsilon \geq 0$. Suppose that isometry constant of $\mathcal{A}$ satisfies $\delta_{2r} < \frac{1}{3}$ and let $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \mathbf{e}$ for a rank-$r$ matrix $\mathbf{X}$ and error vector $\mathbf{e} \in \mathbb{R}^m$. Then SVP with stepsize

$\mu_j = \frac{1}{1+\delta_{2r}}$ outputs a matrix $\mathbf{X}^*$ of rank at most $r$ satisfying

$$\|\mathcal{A}(\mathbf{X}^*) - \mathbf{y}\|_2^2 \leq C \|\mathbf{e}\|^2 + \varepsilon \quad \text{and} \quad \|\mathbf{X} - \mathbf{X}^*\|_F^2 \leq \frac{C \|\mathbf{e}\|^2 + \varepsilon}{1 - \delta_{2r}}$$

in at most $\lceil \frac{1}{\log(1/D)} \log \frac{\|\mathbf{y}\|^2}{2(C\|\mathbf{e}\|^2+\varepsilon)} \rceil$ iterations for universal constants $C, D$.

**Remark 1.47.** Notice that theoretical guarantees are better for the SVP algorithm (including the IHT algorithm) than for the matrix NIHT algorithm. Additionally, every iteration of the NIHT algorithm requires computing an additional singular value decomposition to obtain the parameter $\mu_j$, which consequently makes each iteration of the algorithm slower compared to the alternative versions of the matrix IHT. However, in paper [152] several numerical experiments concerning matrix completion have been performed comparing these algorithms. Recall that extensive numerical experiments for the IHT algorithm for compressive sensing suggest to fix the stepsize $\mu_j = 0.65$, see [112]. This choice of the stepsize has been shown to be effective also in matrix completion, see [152]. Thus, the authors compare the SVP algorithm with $\mu_j = 0.65$ and the matrix NIHT algorithm. They observed that the matrix NIHT algorithm is typically able to recover the same or larger rank than SVP with $\mu_j = 0.65$ for the same stopping criteria, and that NIHT converges faster than SVP (except for ranks when SVP is not able to recover the matrix and NIHT successfully recovers it). In other words, numerical experiments suggest that matrix NIHT algorithm typically performs better than the SVP. However, to recover matrices of large rank (larger than it is possible for SVP), comes with the cost of very slow convergence and thus, there is a need for accelerated variants of NIHT.

## 1.3. Low-rank tensor recovery

So far we have seen several results for sparse vector recovery and low-rank matrix recovery. In both scenarios we have presented some known results. In particularly, we have put an emphasis on convex optimization approach – $\ell_1$-minimization and nuclear norm minimization, respectively – and on iterative approach – (normalized) iterative hard thresholding algorithm and its matrix variant. However, there is a significant interest in going one step further and extending the theory to low-rank tensor recovery. Applications include image and video inpainting [108], reflectance data recovery [108], and machine learning [136]. The goal of low-rank tensor recovery is to reconstruct a low-rank $d$th order tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ from the linear measurement map $\mathcal{A} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^m$ and the measurement vector $\mathbf{y} = \mathcal{A}(\mathbf{X})$, with $m \ll n_1 n_2 \cdots n_d$. In particular, we want to solve the optimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}} \operatorname{rank}(\mathbf{Z}) \quad \text{s.t.} \quad \mathcal{A}(\mathbf{Z}) = \mathbf{y}. \tag{1.31}$$

Unlike in the matrix case, there are different notions of tensor rank which are induced by different tensor decompositions, see Chapter 2. One possibility is to define a rank of a $d$th order tensor analogously to the matrix case – as the minimal number of rank-one order-$d$ tensors that sum up to the original tensor. A tensor $\mathbf{Y} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ is a rank-one tensor if there exist $d$ vectors $\mathbf{u}_i \in \mathbb{R}^{n_i}$ such that $\mathbf{Y}(i_1, i_2, \ldots, i_d) = \mathbf{u}_1(i_1)\, \mathbf{u}_2(i_2) \cdots \mathbf{u}_d(i_d)$, for all $i_k \in [n_k]$ and $k \in [d]$. This notion of tensor rank is called the *CP-rank* or *canonical rank*. In addition, one can also define the corresponding notion of the tensor nuclear norm, denoted by $\|\cdot\|_*$. Inspired by the matrix case, one would expect that solving the minimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}} \|\mathbf{Z}\|_* \quad \text{s.t.} \quad \mathcal{A}(\mathbf{Z}) = \mathbf{y} \tag{1.32}$$

would be a good proxy for (1.31) under certain assumptions on linear operator $\mathcal{A}$. However, computing the CP-rank of a tensor is already NP-hard in general, see [61, 82]. In particular, the set of rank-$r$ tensors is not closed for $r \geq 2$, [81, 101]. Thus, not only solving the optimization problem (1.31), but also solving (1.32) is NP-hard in general. Therefore, although one could analyze the tensor nuclear norm minimization problem (which has been done in [169]), this is not particularly interesting from the computational point of view. Additionally, one could define the null space property in an analogous way as in the matrix scenario, however this leads to the same problem since the CP-decomposition is in general NP-hard to compute. Several generalizations of the coherence/incoherence condition have been introduced in [4, 83, 169] to provide recovery results for tensor completion. A variety of approaches to low-rank tensor recovery already well known in the community are discussed in Chapter 3.

In Chapter 4 we develop a new convex optimization approach to low-rank tensor recovery. We present – to the best of our knowledge – new tensor norms called *theta norms.* These norms are based on a recent tool in the real algebraic geometry – *theta bodies* – which in general provide a sum-of-square nested closed convex relaxations of a given convex set. In our scenario, this convex set is a unit-tensor-nuclear-norm ball. Thus, this method will lead to a set $(\boldsymbol{\mathcal{B}}_{\theta_k})_k$ satisfying

$$\boldsymbol{\mathcal{B}}_{\theta_1} \supseteq \boldsymbol{\mathcal{B}}_{\theta_2} \supseteq \cdots \supseteq \boldsymbol{\mathcal{B}}_{\theta_k} \supseteq \boldsymbol{\mathcal{B}}_{\theta_{k+1}} \supseteq \cdots \supseteq \left\{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} : \|\mathbf{X}\|_* \leq 1\right\}.$$

We define the $\theta_k$-norms via its unit-norm balls. More precisely, the set $\boldsymbol{\mathcal{B}}_{\theta_k}$ is a unit-tensor-$\theta_k$-norm ball, i.e.,

$$\left\{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} : \|\mathbf{X}\|_{\theta_k} \leq 1\right\} = \boldsymbol{\mathcal{B}}_{\theta_k} \quad \text{for all } k.$$

By the theory already developed for theta bodies, it is known that for a given order-$d$ tensor its $\theta_k$-norm can be computed via semidefinite programming. However for simplicity, we only derive a semidefinite programming for computing a $\theta_1$-norm for a $d$-th order tensor. We also give a semidefinite program for low-rank tensor recovery via $\theta_1$-norm minimization. Our numerical experiments show that $\theta_1$-norm minimization successfully recovers a low-rank tensor from few linear measurements and therefore, seems to be a promising approach. However, presently we do not have any theoretical guarantees. That is, the minimal number of measurements that ensures the recovery of a low-rank tensor via $\theta_k$-norm minimization, still remains an open question.

In Chapter 5 we analyze several versions of the iterative hard thresholding algorithm for tensors (TIHT algorithms). As will be seen later in Chapter 2, to compute the best rank-$\mathbf{r}$ approximation of a given tensor is in general NP-hard – independently of the notion of tensor rank. (Here, we do not treat the CP-decomposition since it is already in general NP-hard to compute it). This causes significant difficulties in the analyses of the algorithms. Recall that in compressive sensing and low-rank matrix recovery the operators $\mathcal{H}_s$ and $\mathcal{H}_r^M$ give the best $s$-sparse approximation and the best rank-$r$ approximation of a given vector and matrix, respectively. This is exploited in the convergence proofs of the corresponding IHT algorithms. In the tensor scenario, however, we can only compute a rank-$\mathbf{r}$ approximation $\mathcal{H}_{\mathbf{r}}(\mathbf{X})$ of a tensor $\mathbf{X}$ satisfying

$$\|\mathbf{X} - \mathcal{H}_{\mathbf{r}}(\mathbf{X})\|_F \leq C(d) \|\mathbf{X} - \mathbf{X}_{\text{BEST}}\|_F \quad \text{with } C(d) = \mathcal{O}(\sqrt{d}),$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\mathbf{X}_{\mathrm{BEST}}$ denotes the best rank-$\mathbf{r}$ approximation of a tensor $\mathbf{X}$. In addition, in the proof of Theorem 1.30, in step (1.26) Lemma 1.28 was applied. For the HOSVD-format, we provide a similar result, see Lemma A.5. However, the number of tensors needed to decompose an HOSVD-2$\mathbf{r}$ rank tensor into a sum of rank-$\mathbf{r}$ tensors scales exponentially in the dimension. This leads to significantly worse convergence bounds than in the matrix scenario. For other tensor decompositions (TT-format and HT-format) – to the best of our knowledge – it is unknown how to obtain an analogous result. Thus, we use another tool developed in [95] and later improved in [49] to provide a partial convergence result for TIHT and normalized TIHT algorithm (and for HOSVD, TT, and HT-decomposition). That is, assuming that in every iteration $j$ of the TIHT algorithm we have

$$\left\|\mathbf{Y}^j - \mathcal{H}_{\mathbf{r}}\left(\mathbf{Y}^j\right)\right\|_F \leq (1+\varepsilon)\left\|\mathbf{X} - \mathbf{Y}^j\right\|_F,$$

where $\mathbf{X}$ denotes the original tensor satisfying $\mathcal{A}(\mathbf{X}) = \mathbf{y}$ and $\varepsilon \in [0,1)$ is small enough, we prove a linear convergence of the algorithm. The analysis is based on an appropriate notion of the tensor restricted isometry property (TRIP) – similarly to the matrix scenario. Additionally, we show that subgaussian measurement ensembles and partial Fourier ensembles combined with random sign flips of the tensor entries satisfy TRIP with high probability. Lastly, we present numerical results for low HOSVD-rank order-3 tensor recovery via Partial Fourier map combined with random sign flips of the tensor entries, Gaussian map, and tensor completion.

CHAPTER 2

# Tensors

In this chapter several tensor decompositions are introduced – namely, the *canonical* (CP or CANDECOMP/PARAFAC) decomposition, the *Tucker decomposition*, the *tensor train decomposition* (TT-decomposition), and more generally the *hierarchical Tucker* (HT) *decomposition*. As already mentioned in the previous chapter, several difficulties arise when one considers tensors of order $d \geq 3$. For example, a tractable decomposition – which would be analogous to the singular value decomposition for matrices – does not exist for tensors. This causes significant difficulties in analyzing the algorithms for the low-rank tensor recovery and providing theoretical guarantees for the convergence of the algorithms.

In this section, we will often use the MATLAB notion for better readability. That is, for a matrix $\mathbf{U} \in \mathbb{R}^{n_1 \times n_2}$, its $i$th row and $j$th column are denoted by $\mathbf{U}(i,:)$ and $\mathbf{U}(:,j)$, respectively. In addition, the set $\{1, 2 \ldots, n\}$ will be denoted by $[n]$.

Here, a slightly more extensive introduction to tensors is provided than in our book chapter [130].

## 2.1. Tensor product spaces

We start with some preliminaries. In the sequel, we consider mainly the real field $\mathbb{K} = \mathbb{R}$, although most parts are easy to extend to the complex case. We confine ourselves to finite dimensional linear spaces $V_i = \mathbb{R}^{n_i}$ from which the tensor product space

$$\mathcal{H}_d = \bigotimes_{i=1}^{d} V_i := \bigotimes_{i=1}^{d} \mathbb{R}^{n_i},$$

is built [75]. If it is not stated explicitly, the $V_i = \mathbb{R}^{n_i}$ are supplied with the canonical basis $\{\mathbf{e}_1^i, \ldots, \mathbf{e}_{n_i}^i\}$ of the vector space $\mathbb{R}^{n_i}$. Then any $\mathbf{X} \in \mathcal{H}_d$ can be represented as

$$\mathbf{X} \quad = \quad \sum_{\mu_1=1}^{n_1} \sum_{\mu_2=1}^{n_2} \ldots \sum_{\mu_d=1}^{n_d} \mathbf{X}\left(\mu_1, \mu_2, \ldots, \mu_d\right) \; \mathbf{e}_{\mu_1}^1 \otimes \mathbf{e}_{\mu_2}^2 \otimes \cdots \otimes \mathbf{e}_{\mu_d}^d \; .$$

The tensor $\mathbf{X}$ is called *dth order tensor* or *order-d tensor*. Using this basis, with a slight abuse of notation, we can identify $\mathbf{X} \in \mathcal{H}_d$ with its representation by a $d$-variate function, often called *hyper matrix*,

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_d) \mapsto \mathbf{X}\left(\mu_1, \mu_2, \ldots, \mu_d\right) \in \mathbb{R}, \quad \mu_i \in [n_i]\,, i \in [d]\,,$$

depending on discrete variables, usually called indices $\mu_i$, and $\boldsymbol{\mu}$ is called a multi-index. The actual representation of $\mathbf{X} \in \mathcal{H}_d$ clearly depends on the chosen bases of $V_1, \ldots, V_d$. The number of possibly nonzero entries in the representation of $\mathbf{X}$ is $n_1 n_2 \cdots n_d = \mathcal{O}(n^d)$, with $n = \max\{n_i : i \in [d]\}$. That is, it grows exponentially in the dimension $d$. This is often referred to as the *curse of dimensions.* We equip the linear space $\mathcal{H}_d$ with the $\ell_2$-norm (also called Frobenius norm) $\|\mathbf{X}\|_F = \sqrt{\langle \mathbf{X}, \mathbf{X} \rangle}$

induced by the inner product

$$\langle \mathbf{X}, \mathbf{Y} \rangle := \sum_{\mu_1=1}^{n_1} \sum_{\mu_2=1}^{n_2} \cdots \sum_{\mu_d=1}^{n_d} \mathbf{X}(\mu_1, \mu_2, \ldots, \mu_d)\, \mathbf{Y}(\mu_1, \mu_2, \ldots, \mu_d).$$

Throughout this chapter, all tensor contractions or various tensor–tensor products are either defined explicitly, by summation over the corresponding indices, or by introducing the corresponding tensor matricizations and performing matrix–matrix products.

The *vectorization* of a tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ is a linear transformation that converts a tensor into a column vector. Its result is denoted by $\operatorname{vec}(\mathbf{X}) \in \mathbb{R}^{n_1 n_2 \cdots n_d \times 1}$. The ordering of an elements in $\operatorname{vec}(\mathbf{X})$ is not important as long as it is consistent.

The *matricization* (also called *flattening*) is an operation that transforms a tensor into a matrix. For a $d$th order tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ and an ordered subset $\boldsymbol{\mathcal{S}} \subseteq [d]$, an $\boldsymbol{\mathcal{S}}$-matricization $\mathbf{X}^{\boldsymbol{\mathcal{S}}} \in \mathbb{R}^{\prod_{k \in \boldsymbol{\mathcal{S}}} n_k \times \prod_{\ell \in \boldsymbol{\mathcal{S}}^c} n_\ell}$ is defined element-wise as

$$\mathbf{X}^{\boldsymbol{\mathcal{S}}}\left((i_k)_{k \in \boldsymbol{\mathcal{S}}}; (i_\ell)_{\ell \in \boldsymbol{\mathcal{S}}^c}\right) = \mathbf{X}(i_1, i_2, \ldots, i_d).$$

That is, the indexes in the set $\boldsymbol{\mathcal{S}}$ define the rows of a matrix and the indexes in the set $\boldsymbol{\mathcal{S}}^c = [d] \setminus \boldsymbol{\mathcal{S}}$ define the columns.

For a singleton $\boldsymbol{\mathcal{S}} = \{i\}$ (for $i \in [d]$), the $\{i\}$-matricization is called the *$i$-th unfolding.*

*Fibers* are a higher order analogue of matrix rows and columns. For $k \in [d]$, the *mode-$k$ fiber* $\mathbf{x}_{i_1 \ldots i_{k-1} i_{k+1} \ldots i_d} \in \mathbb{R}^{n_k}$ of a $d$th order tensor is defined element-wise as

$$\mathbf{x}_{i_1 \ldots i_{k-1} i_{k+1} \ldots i_d}(i_k) = \mathbf{X}(i_1, i_2, \ldots, i_d), \quad \text{for all } k \in [n_k].$$

Next, we introduce the *k-th mode product* which is a product between a tensor and a matrix of appropriate dimensions.

**Definition 2.1** ($k$-mode product)**.** For a tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, a matrix $\mathbf{A} \in \mathbb{R}^{J \times n_k}$, and $k \in [d]$, the *$k$-mode product* of $\mathbf{X}$ and $\mathbf{A}$

$$\mathbf{X} \times_k \mathbf{A} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_{k-1} \times J \times n_{k+1} \times \cdots \times n_d}$$

is defined element-wise as

$$(\mathbf{X} \times_k \mathbf{A})(i_1, \ldots, i_{k-1}, j, i_{k+1}, \ldots, i_d) = \sum_{i_k=1}^{n_k} \mathbf{X}(i_1, i_2, \ldots, i_d)\, \mathbf{A}(j, i_k).$$

For a tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ and matrices $\mathbf{A} \in \mathbb{R}^{J \times n_j}$, $\mathbf{B} \in \mathbb{R}^{K \times n_k}$, $\mathbf{C} \in \mathbb{R}^{L \times K}$ it holds

$$\begin{aligned} \mathbf{X} \times_j \mathbf{A} \times_k \mathbf{B} &= \mathbf{X} \times_k \mathbf{B} \times_j \mathbf{A}, \quad \text{whenever } j \neq k \\ \mathbf{X} \times_k \mathbf{B} \times_k \mathbf{C} &= \mathbf{X} \times_k \mathbf{CB}. \end{aligned}$$

Notice that the singular value decomposition (SVD decomposition) of a matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ can be written using the above notation as $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \boldsymbol{\Sigma} \times_1 \mathbf{U} \times_2 \mathbf{V}$.

**2.1.1. Subspace approximation.** The essence of the classical Tucker format is that, given a tensor $\mathbf{X}$ and a rank-tuple $\mathbf{r} = (r_1, r_2, \ldots, r_d)$, one is searching for optimal subspaces $U_i \subset \mathbb{R}^{n_i}$ such that

$$\|\mathbf{X} - \mathbf{Y}\|_F, \text{ where } \mathbf{Y} \in U_1 \otimes \cdots \otimes U_d,$$

is minimized over $U_1, \dots, U_d$ with $\dim U_i = r_i$, for all $i \in [d]$. Equivalently, we are looking for the corresponding basis $\left\{\mathbf{u}^i_{k_i}\right\}_{k_i}$ of $U_i$, which can be written in the form

$$\mathbf{u}^i_{k_i} := \sum_{\mu_i=1}^{n_i} \mathbf{U}_i(\mu_i, k_i)\mathbf{e}^i_{\mu_i}, \quad k_i \in [r_i], r_i < n_i, \tag{2.1}$$

where $\mathbf{U}_i(\mu_i, k_i) \in \mathbb{R}$, for each coordinate direction $i \in [d]$. The matrix $\mathbf{U}_i$ is the matrix representation of the subspace $U_i$. With a slight abuse of notation we often identify the basis vector with its representation

$$\mathbf{u}^i_{k_i} \simeq \big(\mu_i \mapsto \mathbf{U}_i(\mu_i, k_i)\big), \quad \mu_i \in [n_i], \quad k_i \in [r_i],$$

i.e., a discrete function or an $n_i$-tuple. This concept of subspace approximation can be used for an approximation of a single tensor $\mathbf{X}$ in tensor product spaces. Given the bases $\left\{\mathbf{u}^i_{k_i}\right\}_{k_i}$, the tensor $\mathbf{X}$ can be represented by

$$\mathbf{X} = \sum_{k_1=1}^{r_1} \dots \sum_{k_d=1}^{r_d} \mathbf{C}(k_1, \dots, k_d)\, \mathbf{u}^1_{k_1} \otimes \dots \otimes \mathbf{u}^d_{k_d} \in \bigotimes_{i=1}^{d} U_i \subset \mathcal{H}_d = \bigotimes_{i=1}^{d} \mathbb{R}^{n_i}. \tag{2.2}$$

This representation can be written using the *k-mode product* (see Definition 2.1) as

$$\mathbf{X} = \mathbf{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times \dots \times_d \mathbf{U}_d,$$

where $\mathbf{U}_k$ is the matrix representation of the subspace $U_k$.

In case where $\left\{\mathbf{u}^i_{k_i}\right\}_{k_i}$'s form orthonormal bases, the core tensor $\mathbf{C} \in \bigotimes_{i=1}^d \mathbb{R}^{r_i}$ is given entry-wise by

$$\mathbf{C}(k_1, \dots, k_d) = \langle \mathbf{X}, \mathbf{u}^1_{k_1} \otimes \dots \otimes \mathbf{u}^d_{k_d} \rangle.$$

We call a representation of the form (2.2) with some $\mathbf{u}^i_{k_i}$, and tensor $\mathbf{C}$ the Tucker representation, and the Tucker representations the Tucker format. In this formal parametrization, the upper limit of the sums may be larger than the ranks and $\left\{\mathbf{u}^i_{k_i}\right\}_{k_i}$ may not be linearly independent. Noticing that the Tucker representation of a tensor is not uniquely defined, we are interested in some normal form, see Subsection 2.1.5.

**2.1.2. Hierarchical tensor representation.** The *hierarchical Tucker format* (HT) in the form introduced by Hackbusch and Kühn in [76], extends the idea of subspace approximation to a hierarchical or multi-level framework. Let us proceed in a hierarchical way. We first consider $V_1 \otimes V_2 = \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2}$ or preferably the subspaces $U_1 \otimes U_2$ introduced in the previous section. For the approximation of $\mathbf{X} \in \mathcal{H}_d$ we only need a subspace $U_{\{1,2\}} \subset U_1 \otimes U_2$ with dimension $r_{\{1,2\}} < r_1 r_2$. Indeed, $V_{\{1,2\}}$ is defined through a new basis

$$V_{\{1,2\}} = \text{span}\, \{\mathbf{u}^{\{1,2\}}_{k_{\{1,2\}}} : k_{\{1,2\}} \in \left[r_{\{1,2\}}\right]\},$$

with basis vectors given by

$$\mathbf{u}^{\{1,2\}}_{k_{\{1,2\}}} = \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \mathbf{u}^{\{1,2\}}(k_{\{1,2\}}, k_1, k_2)\, \mathbf{u}^1_{k_1} \otimes \mathbf{u}^2_{k_2}, \quad k_{\{1,2\}} \in \left[r_{\{1,2\}}\right].$$

One may continue in several ways, e.g., by building a subspace $U_{\{1,2,3\}} \subset U_{\{1,2\}} \otimes U_3 \subset U_1 \otimes U_2 \otimes U_3 \subset V_1 \otimes V_2 \otimes V_3$, or $U_{\{1,2,3,4\}} \subset U_{\{1,2\}} \otimes U_{\{3,4\}}$, where $U_{\{3,4\}}$ is defined analogously to $U_{\{1,2\}}$ and so on.

For a systematic treatment, this approach can be cast into the framework of a partition tree, with leaves $\{1\}, \ldots, \{d\}$ (simply abbreviated here by $1, \ldots, d$), root $D := \{1, 2, \ldots, d\}$ (often denoted also as $t_{\text{root}}$), and vertices $t \subset D := \{1, \ldots, d\}$, corresponding to the partition $t = t_1 \cup t_2$, $t_1 \cap t_2 = \emptyset$. Without loss of generality, we can assume that $i < j$, for all $i \in t_1$, $j \in t_2$. We call $t_1, t_2$ the *sons* of the *father* $t$ and $D$ is called the *root* of the tree. In the example above we have $t = \{1, 2, 3\} = t_1 \cup t_2 = \{1, 2\} \cup \{3\}$, where $t_1 = \{1, 2\}$ and $t_2 = \{3\}$.

Restricting the partition tree to a binary tree so that every interior node $t$ (i.e. $t \neq \{i\}$) contains two sons is often the common choice, which leads to TT, and more generally HT format. Let $t_1, t_2 \subset D$ be the two sons of $t \subset D$. Then $U_t \subset U_{t_1} \otimes U_{t_2}$ is defined via the basis

$$\mathbf{u}_\ell^t = \sum_{i=1}^{r_{t_1}} \sum_{j=1}^{r_{t_2}} \mathbf{B}_\alpha(\ell, i, j)\, \mathbf{u}_i^{t_1} \otimes \mathbf{u}_j^{t_2}. \tag{2.3}$$

For non-leaf nodes $t$, the subspaces $U_t$ can be also considered as matrices $\mathbf{U}_t \in \mathbb{R}^{n_t \times r_t}$ with $\mathbf{U}_t\,(:, \ell) = \mathbf{u}_\ell^t$ and $n_t = \prod_{\ell \in t} n_\ell$. Without loss of generality, all the basis vectors, e.g., $\{\mathbf{u}_\ell^t : \ell = 1, \ldots, r_t\}$, can be constructed to be orthonormal, as long as $t$ is not the root ($t \neq D$). The tensors $(\ell, i, j) \mapsto \mathbf{B}_t(i, j, \ell)$ are called *transfer* or *component tensors*. For a leaf $\{i\} \simeq i$, the matrix $(\mu_i, k_i) \mapsto \mathbf{U}_i(\mu_i, k_i)$ in (2.1) is called an *i-frame*. The component tensor $\mathbf{B}_D = \mathbf{B}_{\{1,\ldots,d\}}$ at the root is called the *root tensor*.

Since the matrices $\mathbf{U}_t$ are too large, we avoid computing them. We store only the $\alpha$-frames $\mathbf{U}_\alpha$ for all leaves $\alpha \in [d]$, and the *transfer* or *component tensors* which, for fixed $\ell = 1, \ldots, r_t$, can also be casted into *transfer matrices* $(i, j) \mapsto [\mathbf{B}_t(\ell)]\,(i, j) \in \mathbb{R}^{r_{t_1} \times r_{t_2}}$.

With $\boldsymbol{\mathcal{I}}\,(T_I)$ we denote the set of all interior (non-leaf) nodes and with $\boldsymbol{\mathcal{L}}\,(T_I)$ we denote the set of all leaves of the corresponding partition tree $T_I$.

**Proposition 2.2** ([75]). A tensor $\mathbf{X} \in \mathcal{H}_d$ is completely parametrized by the transfer tensors $\mathbf{B}_t$, $t \in \boldsymbol{\mathcal{I}}\,(T_I)$ and $\alpha$-frames $\mathbf{U}_\alpha$, $\alpha \in \boldsymbol{\mathcal{L}}\,(T_I)$, i.e., by a multi-linear function $\tau$

$$(\{\mathbf{B}_t : t \in \boldsymbol{\mathcal{I}}\,(T_I)\}, \{\mathbf{U}_\alpha : \alpha \in \boldsymbol{\mathcal{L}}\,(T_I)\}) \mapsto \mathbf{X} = \tau\big(\{\mathbf{B}_t : t \in \boldsymbol{\mathcal{I}}\,(T_I)\}, \{\mathbf{U}_\alpha : \alpha \in \boldsymbol{\mathcal{L}}\,(T_I)\}\big).$$

A tensor whose partition tree is presented in Figure 2.1 is completely parametrized by

$$\left(\left\{\mathbf{B}_{\{1,2\}}, \mathbf{B}_{\{1,2,3\}}, \mathbf{B}_{\{4,5\}}, \mathbf{B}_{\{1,2,3,4,5\}}\right\}, \{\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{U}_4, \mathbf{U}_5\}\right).$$

An HT-decomposition $(\{\mathbf{B}_t : t \in \boldsymbol{\mathcal{I}}\,(T_I)\}, \{\mathbf{U}_\alpha : \alpha \in \boldsymbol{\mathcal{L}}\,(T_I)\})$ satisfying

$$\mathbf{U}_\alpha^T \mathbf{U}_\alpha = \mathbf{I}, \quad \text{for all leaves } \alpha \in [d]$$
$$\mathbf{B}_t^{\{1,2\}^T} \mathbf{B}_t^{\{1,2\}} = \mathbf{I}, \quad \text{for all } t \in \boldsymbol{\mathcal{I}}(T_I) \backslash t_{\text{root}}$$

is called an orthogonal HT-decomposition.

Indeed $\tau$ is defined by applying (2.3) recursively. Since $\mathbf{B}_t$ depends bi-linearly on $\mathbf{B}_{t_1}$ and $\mathbf{B}_{t_2}$, the composite function $\tau$ is multi-linear in its arguments $\mathbf{B}_t$ and $\mathbf{U}_\alpha$.

**Remark 2.3.** In the literature, tensors are often considered as vectors over product index sets. For this purpose, the $d$-fold product index set is introduced

$$I = \boldsymbol{\mathcal{I}}_1 \times \boldsymbol{\mathcal{I}}_2 \times \cdots \times \boldsymbol{\mathcal{I}}_d, \quad \text{where } \boldsymbol{\mathcal{I}}_\mu := \{1, 2, \ldots, n_\mu\}, \text{ for } \mu \in [d]$$

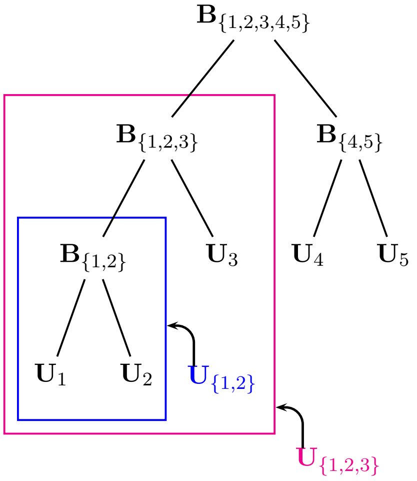


FIGURE 2.1. Hierarchical tensor representation of an order 5 tensor

and we write $\mathbf{X} \in \mathbb{R}^{I}$. The subscript of the dimensional tree $T_I$ refers to the above defined index set. By further defining the index sets

$$I^{(\mu)} := \boldsymbol{\mathcal{I}}_1 \times \cdots \times \boldsymbol{\mathcal{I}}_{\mu-1} \times \boldsymbol{\mathcal{I}}_{\mu+1} \times \cdots \times \boldsymbol{\mathcal{I}}_d,$$

the corresponding tensor unfolding $\mathbf{X}^{\mu}$ is in $\mathbb{R}^{\boldsymbol{\mathcal{I}}_\mu \times I^{(\mu)}}$.

The HT-rank $\mathbf{r} = (r_t)_{t \in T_I}$ of a tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ can be obtained via corresponding matricizations of a tensor $\mathbf{X}$. That is,

$$r_t = \operatorname{rank}\left(\mathbf{X}^t\right), \quad \text{for all } t \in T_I.$$

The set of tensors $\mathbf{X} \in \mathcal{H}_d$ of given HT-rank $\mathbf{r}$ will be denoted by $\mathcal{M}_{\mathbf{r}}$. The set of all tensors of rank $\mathbf{s}$ at most $\mathbf{r}$ (i.e., $s_\alpha \leq r_\alpha$ for all $\alpha \in T_I$) will be denoted by $\mathcal{M}_{\leq \mathbf{r}}$.

Unlike the matrix case, it is possible that $\mathcal{M}_{\mathbf{r}} = \emptyset$ for some tuples $\mathbf{r}$, see [30]. However, since the TIHT algorithm presented in Chapter 5 works on a closed nonempty set $\mathcal{M}_{\leq \mathbf{r}}$, this issue does not concern us.

In contrast to the canonical format (presented later in (2.5)), also known as CP and CANDECOMP/PARAFAC, see [46, 94] and the border rank problem [101], in the present setting the rank is a well defined quantity. This fact makes the present concept highly attractive for tensor recovery. Additionally, if $\mathbf{X}$ is a rank $\mathbf{r}$ tensor then there exists a component tensor $\mathbf{B}_\alpha$ of the form (2.3) where $\ell = 1, \ldots, r_\alpha$.

It is well known that the set of all matrices $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ of rank at most $r$ is a set of common zeros of multi-variate polynomials, i.e., an algebraic variety (see e.g. Chapter 4.) The set $\mathcal{M}_{\leq \mathbf{r}}$ is the set of all tensors $\mathbf{X} \in \mathcal{H}_d$, where the matrices $\mathbf{U}_t$ have a rank at most $r_t$. Therefore, it is again a set of common zeros of multivariate polynomials.

**Data complexity** Let $n := \max\{n_i : i = 1, \ldots, d\}$, $r := \max\{r_\alpha : \alpha \in T_I\}$. Then the number of data required for the representation is $\mathcal{O}(ndr + dr^3)$, in particular it does not scale exponentially with respect to the order $d$.

**2.1.3. Tensor trains and matrix product representation.** We now highlight a special case of hierarchical tensor representations – the *tensor trains* (TT tensors) – defined by taking $U_{\{1,\ldots,p+1\}} \subset U_{\{1,\ldots,p\}} \otimes V_{\{p+1\}}$. The TT tensors, developed by Oseledets and Tyrtyshnikov in [125, 126], are also known in quantum physics as *matrix product states* (MPS) [164]. Therein, we abbreviate $i \simeq \{i, \ldots, d\}$ and consider the unbalanced tree (see Figure 2.2)

$$T_I = \{\{1, 2, 3, \ldots, d\}, \{1\}, \{2, 3, \ldots, d\}, \{2\}, \{3, \ldots, d\}, \{3\}, \ldots, \{d-1, d\}, \{d\}\}.$$

The $\alpha$-frame $\mathbf{U}_\alpha$ for a leaf $\alpha \in \{\{1\}, \{2\}, \ldots, \{d-1\}\}$ is usually defined as the identity matrix of appropriate size and therefore the tensor $\mathbf{X} \in \mathcal{H}_d$ is completely parametrized by the transfer tensors $(\mathbf{B}_t)_{t \in \mathcal{I}(T_I)}$ and the $d$-frame $\mathbf{U}_{\{d\}}$. Applying a recursive construction, the tensor $\mathbf{X}$ can be written as

$$\begin{aligned}(\mu_1, \ldots, \mu_d) \quad &\mapsto \quad \mathbf{X}(\mu_1, \ldots, \mu_d) \\ &= \quad \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \cdots \sum_{k_{d-1}=1}^{r_{d-1}} \mathbf{B}_1(\mu_1, k_1)\mathbf{B}_2(k_1, \mu_2, k_2) \cdots \mathbf{B}_d(k_{d-1}, \mu_d),\end{aligned}$$

where $\mathbf{B}_d := \mathbf{U}_{\{d\}}$ and we used the abbreviation $i \simeq \{i, i+1, \ldots, d\}$ for all interior nodes $\{i, i+1, \ldots, d\}$ (i.e., the transfer tensor $\mathbf{B}_i = \mathbf{B}_{\{i,i+1,\ldots,d\}}$). Introducing the matrices $\mathbf{G}_i(\mu_i) \in \mathbb{R}^{r_{i-1} \times r_i}$,

$$[\mathbf{G}_i(\mu_i)]\,(k_{i-1}, k_i) = \mathbf{B}_i(k_{i-1}, \mu_i, k_i), \quad 2 \leq i \leq d-1,$$

and with the convention $r_0 = r_d = 1$

$$[\mathbf{G}_1(\mu_1)]\,(k_1) = \mathbf{B}_1(\mu_1, k_1) \quad \text{and} \quad [\mathbf{G}_d(\mu_d)]\,(k_{d-1}) = \mathbf{B}_d(k_{d-1}, \mu_d),$$

the formula (2.4) can be rewritten entry-wise by matrix–matrix products

$$\mathbf{X}(\mu_1, \ldots, \mu_d) = \mathbf{G}_1(\mu_1) \cdots \mathbf{G}_i(\mu_i) \cdots \mathbf{G}_d(\mu_d) = \tau(\mathbf{B}_1, \ldots, \mathbf{B}_d). \tag{2.4}$$

This representation is by no means unique. For example, if $\{\mathbf{M}_i\}_{i=1}^{d-1}$ is a set of invertible matrices of appropriate dimension, then

$$\begin{aligned}\mathbf{X}\,(\mu_1, \mu_2, \ldots, \mu_d) &= \mathbf{G}_1(\mu_1) \cdots \mathbf{G}_i(\mu_i) \cdots \mathbf{G}_d(\mu_d) \\ &= \mathbf{G}_1(\mu_1)\mathbf{M}_1\mathbf{M}_1^{-1} \cdots \mathbf{M}_{i-1}\mathbf{M}_{i-1}^{-1}\mathbf{G}_i(\mu_i)\mathbf{M}_i\mathbf{M}_i^{-1} \cdots \mathbf{M}_{d-1}\mathbf{M}_{d-1}^{-1}\mathbf{G}_d(\mu_d) \\ &= \overline{\mathbf{G}}_1(\mu_1) \cdots \overline{\mathbf{G}}_i(\mu_i) \cdots \overline{\mathbf{G}}_d(\mu_d),\end{aligned}$$

where

$$\begin{aligned}&\overline{\mathbf{G}}_1(\mu_1) := \mathbf{G}_1(\mu_1)\mathbf{M}_1, \quad \overline{\mathbf{G}}_d(\mu_d) := \mathbf{M}_{d-1}^{-1}\mathbf{G}_d(\mu_d) \\ &\overline{\mathbf{G}}_i(\mu_i) := \mathbf{M}_{i-1}^{-1}\mathbf{G}_i(\mu_i)\mathbf{M}_i, \quad \text{for all } i = 2, \ldots, d-1.\end{aligned}$$

Let $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ be a TT-rank $\mathbf{r} = (r_1, r_2, \ldots, r_{d-1})$ tensor. Then the entries of a rank vector $\mathbf{r} = (r_1, r_2, \ldots, r_{d-1})$ can be obtained via the ranks of the appropriate matricizations. That is,

$$r_k = \operatorname{rank}\left(\mathbf{X}^{\{1,\ldots,k\}}\right), \quad \text{for all } k \in [d-1]\,.$$

**Data complexity:** Let $n := \max\{n_i : i = 1, \ldots, d\}$, $r := \max\{r_j : j \in [d-1]\}$. Then the number of data required for the presentation is $\mathcal{O}(dnr^2)$. Notice, however, that representing a tensor whose tree is a TT-tree in an HT-format requires $\mathcal{O}(dnr + (d-1)r^3)$ number of data. Thus, in Chapter 5, to compute the covering number related to the TT-tensors, we use the latter

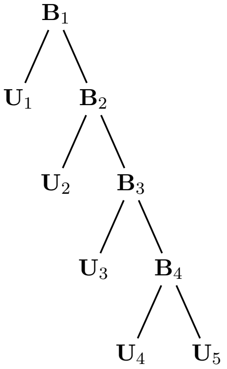


FIGURE 2.2. TT representation of an order 5 tensor with abbreviation $i \simeq \{i, \ldots, d\}$ for the interior nodes.

representation (since it will lead to better covering bounds). Additionally, computing a single entry of a tensor in the TT-format requires the matrix multiplication of $d$ matrices of size at most $r \times r$. This can be performed in $\mathcal{O}(ndr^3)$ operations.

Since the parametrization $\tau$ can be written in the simple matrix product form (2.4), we will consider the TT format often as a prototype model, and use it frequently for our explanations. We remark that most of the properties can easily be extended to the general hierarchical case with straightforward modifications, see [75].

**Canonical format (CP-format):** The CP decomposition factorizes a tensor into a sum of component rank-one tensors. A $d$th order tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ is a rank-one tensor if there exist vectors $\mathbf{u}^1 \in \mathbb{R}^{n_1}, \mathbf{u}^2 \in \mathbb{R}^{n_2}, \ldots, \mathbf{u}^d \in \mathbb{R}^{n_d}$ such that $\mathbf{X} = \mathbf{u}^1 \otimes \mathbf{u}^2 \otimes \cdots \otimes \mathbf{u}^d$ or element-wise

$$\mathbf{X}\left(i_1, i_2, \ldots, i_d\right) = \mathbf{u}^1\left(i_1\right) \mathbf{u}^2\left(i_2\right) \cdots \mathbf{u}^d\left(i_d\right).$$

For example, given a $d$th order tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, we wish to write it as

$$\mathbf{X} = \sum_{k=1}^{R} \mathbf{u}_k^1 \otimes \mathbf{u}_k^2 \otimes \cdots \otimes \mathbf{u}_k^d, \tag{2.5}$$

where $R$ is a positive integer and $\mathbf{u}_k^\ell \in \mathbb{R}^{n_\ell}$, for all $k \in [R]$ and $\ell \in [d]$. Element-wise, (2.5) is written as

$$\mathbf{X}\left(i_1, i_2, \ldots, i_d\right) = \sum_{k=1}^{R} \mathbf{u}_k^1(i_1)\, \mathbf{u}_k^2(i_2)\, \cdots\, \mathbf{u}_k^d(i_d).$$

A CP-rank (or canonical rank) of a tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, similarly to the matrix case, is the smallest number of rank-one tensors that sum up to $\mathbf{X}$. Then the analog of the matrix nuclear norm for tensors is

$$\|\mathbf{X}\|_* = \min \left\{ \sum_{k=1}^{r} |c_k| : \mathbf{X} = \sum_{k=1}^{r} c_k\, \mathbf{u}_k^1 \otimes \mathbf{u}_k^2 \otimes \cdots \otimes \mathbf{u}_k^d,\ r \in \mathbb{N}, \right.$$
$$\left. \left\|\mathbf{u}_k^i\right\|_2 = 1,\ \text{for all } i \in [d], k \in [r] \right\}.$$

Unfortunately, computing the canonical rank of a tensor, as well as computing the nuclear norm of a tensor is in general NP-hard, see [61, 82].

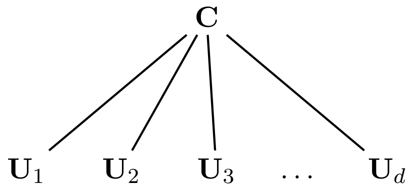


FIGURE 2.3. Tucker representation of an order-$d$ tensor

**2.1.4. Higher order singular value decomposition.** Let us provide more details about the rather classical higher order singular value decomposition of an order-$d$ tensor. Above we have considered only binary dimension trees $T_I$, but we can extend the considerations also to $N$-ary trees with $N \geq 3$. The $d$-ary tree $T_I = \{\{1,\dots,d\},\{1\},\{2\},\dots,\{d\}\}$ (the tree with a root with $d$ sons $i \simeq \{i\}$, see also Figure 2.3) induces the Tucker decomposition and the corresponding higher order singular value decomposition (HOSVD). The Tucker decomposition was first introduced by Tucker in 1963 [158] and has been refined later on in many works, see, e.g., [104, 158, 159]. Additionally, it has been applied in chemical analysis [79], psychometrics [92], signal processing [44, 116], computer vision [161], etc.

**Definition 2.4** (Tucker decomposition)**.** Given a tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, the decomposition

$$\mathbf{X} = \mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times \cdots \times_d \mathbf{U}_d,$$

or element-wise

$$\mathbf{X}\left(\mu_1,\mu_2,\dots,\mu_d\right) = \sum_{k_1=1}^{r_1}\sum_{k_2=1}^{r_2}\dots\sum_{k_d=1}^{r_d}\mathbf{C}\left(k_1,k_2,\dots,k_d\right)\mathbf{u}^1_{k_1}\left(\mu_1\right)\mathbf{u}^2_{k_2}\left(\mu_2\right)\cdots\mathbf{u}^d_{k_d}\left(\mu_d\right),$$

$r_i \leq n_i$, $i \in [d]$, is called a Tucker decomposition. The tensor $\mathbf{C} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}$ is called a core tensor and the $\mathbf{u}^i_{k_i} \in \mathbb{R}^{n_i}$ for $k_i \in [r_i]$, form a basis of the subspace $U_i \subset \mathbb{R}^{n_i}$. They can also be considered as $i$-frames $\mathbf{U}_i \in \mathbb{R}^{n_i \times r_i}$.

Notice that the Tucker decomposition is highly non-unique. For an $i \in [d]$ and an invertible matrix $\mathbf{Q}_i \in \mathbb{R}^{r_i \times r_i}$, one can define a matrix $\overline{\mathbf{U}}_i = \mathbf{U}_i\mathbf{Q}_i$ and a tensor $\overline{\mathbf{C}}_i$

$$\overline{\mathbf{C}}_i\left(k_1,k_2,\dots,k_d\right) = \sum_{\overline{k}_i=1}^{r_i}\mathbf{C}\left(k_1,k_2,\dots,\overline{k}_i,\dots k_d\right)\mathbf{Q}_i^{-1}\left(\overline{k}_i,k_i\right)$$

such that the tensor $\mathbf{X}$ can also be written as

$$\mathbf{X}\left(\mu_1,\mu_2,\dots,\mu_d\right) = \sum_{k_1=1}^{r_1}\sum_{k_2=1}^{r_2}\dots\sum_{k_d=1}^{r_d}\overline{\mathbf{C}}_i\left(k_1,k_2,\dots,k_d\right)\mathbf{u}^1_{k_1}\left(\mu_1\right)\mathbf{u}^2_{k_2}\left(\mu_2\right)\cdots\overline{\mathbf{u}}^i_{k_i}\left(\mu_i\right)\cdots\mathbf{u}^d_{k_d}\left(\mu_d\right).$$

Similarly to the matrix case and the singular value decomposition, one can impose orthogonality conditions on the matrices $\mathbf{U}_i$, for all $i \in [d]$, i.e., we assume that $\{\mathbf{u}^i_{k_i} : k_i \in [r_i]\}$ are orthonormal bases. However, unlike in the matrix scenario, in this case one does not obtain a super-diagonal core tensor $\mathbf{C}$.

**Definition 2.5** (HOSVD decomposition)**.** The HOSVD decomposition of a given tensor $\mathbf{X} \in \mathcal{H}_d$ is a special case of the Tucker decomposition where

- the bases $\{\mathbf{u}^i_{k_i} \in \mathbb{R}^{n_i} : k_i \in [r_i]\}$ are orthogonal and normalized, for all $i \in [d]$,

- the tensor $\mathbf{C} \in \mathcal{H}_d$ is all orthogonal, i.e., $\langle \mathbf{C}_{k_i=p}, \mathbf{C}_{k_i=q} \rangle = 0$, for all $i \in [d]$ and whenever $p \neq q$,
- the subtensors of the core tensor $\mathbf{C}$ are ordered according to their Frobenius norm, i.e., $\|\mathbf{C}_{k_i=1}\|_F \geq \|\mathbf{C}_{k_i=2}\|_F \geq \cdots \geq \|\mathbf{C}_{k_i=n_i}\|_F \geq 0$, for all $i \in [d]$.

Here, the subtensor $\mathbf{C}_{k_i=p} \in \mathbb{R}^{n_1 \times \cdots \times n_{i-1} \times n_{i+1} \times \cdots \times n_d}$ is a tensor of order $d-1$ obtained by fixing the $k_i$-th mode in the tensor $\mathbf{C}$ to $p$. That is, it is defined element-wise as

$$\mathbf{C}_{k_i=p}\left(\mu_1, \ldots, \mu_{i-1}, \mu_{i+1}, \ldots, \mu_d\right) = \mathbf{C}\left(\mu_1, \ldots, \mu_{i-1}, p, \mu_{i+1}, \ldots, \mu_d\right).$$

The Tucker rank $\mathbf{r} = (r_1, r_2, \ldots, r_d)$ of a tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ can be obtained via the unfoldings. That is,

$$r_k = \operatorname{rank}\left(\mathbf{X}^{\{k\}}\right), \quad \text{for all } k \in [d].$$

Since the core tensor contains $r_1 \cdots r_d \sim r^d$, $r := \max\{r_i : i \in [d]\}$, possibly nonzero entries, this concept does not prevent the number of free parameters from scaling exponentially with the dimensions $\mathcal{O}(r^d)$. Setting $n := \max\{n_i : i \in [d]\}$, the overall complexity for storing the required data (including the basis vectors) is bounded by $\mathcal{O}(ndr + r^d)$. Since $n_i$ is replaced by $r_i$, one obtains a compression $\frac{r_1}{n_1} \cdots \frac{r_d}{n_d} \sim \left(\frac{r}{n}\right)^d$. Without further sparsity of the core tensor the Tucker format is appropriate for low order tensors $d < 4$.

**Algorithm 2.1.** Tucker's method for computing a rank-$(r_1, r_2, \ldots, r_d)$ Tucker decomposition, also known as HOSVD.

```
Input: $\mathbf{X} \in \mathcal{H}_d$ or Tucker rank $\mathbf{r} = (r_1, r_2, \ldots, r_d)$.
for $i = 1, \ldots, d$ do
    Compute a singular value decomposition of $\mathbf{X}^{\{i\}} := \overline{\mathbf{U}}_i \overline{\mathbf{\Sigma}}_i \overline{\mathbf{V}}_i^T$.
    Set $\mathbf{U}_i := \overline{\mathbf{U}}_i\left(:, [r_i]\right)$.
end for
Set $\mathbf{S} := \mathbf{X} \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \times \cdots \times_d \mathbf{U}_d^T$.
Output: HOSVD decomposition $\mathbf{X} = \mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times \cdots \times_d \mathbf{U}_d$.
```

The HOSVD decomposition can be computed via SVDs of appropriate unfoldings $\mathbf{U}^{\{i\}}$, see e.g. [94] and Algorithm 2.1. The more general hierarchical Tucker (HT) decomposition can be computed vis successive SVDs, see the following subsection. For more information on these decompositions, we refer the interested reader to [43, 75].

**2.1.5. Tensor SVD algorithm and truncation.** The singular value decomposition of the matricization $\mathbf{X}^t$, $t \in T_I$, factorizes the tensor $\mathbf{X}$ into two parts. Thereby, we separate the tree into two subtrees. Each part can be treated independently in an analogous way as before by applying the singular value decomposition. This procedure can be continued in a way such that one ends up with an explicit description of the component tensors. There are several sequential orders one can proceed, including top-down and bottom-up strategies. We will call these procedures tensor SVD algorithms. As long as no approximation (i.e., no truncation) has been applied during the corresponding SVDs, at the end one obtains an exact recovery of the original tensor. The situation changes if we apply truncations (via thresholding). Then the result may depend on the way and on the order we proceed as well as on the variant of the thresholding procedure.

In order to become more explicit let us present a procedure for obtaining the HT-decomposition via the tensor SVD algorithm for the model example of a TT-tensor [124], already introduced in [164] for the matrix product representation. Without truncations the Algorithm 2.2 provides an exact reconstruction with a TT representation provided that the multi-linear rank $\mathbf{s} = (s_1, \ldots, s_{d-1})$ is chosen large enough. In general, the $s_i$'s can be chosen to be larger than the dimensions $n_i$. Via inspecting the ranks of the relevant matricizations, the multilinear rank $\mathbf{s}$ may be determined a priori.

**Algorithm 2.2.** TT-SVD algorithm

```
**Input:** $\mathbf{X} \in \mathcal{H}_d$ **of multi-linear rank** $\mathbf{s} = (s_1, \ldots, s_{d-1})$, $s_0 := 1$, $s_d := 1$.
**Initialization: Set** $\mathbf{M}_1 = \mathbf{X}^{\{1,2\}}$.
**for** $i = 1, \ldots, d-1$ **do**
    **Compute the SVD of** $\mathbf{M}_i$:
        $\mathbf{M}_i = \mathbf{G}_i^{\{1,2\}} \mathbf{\Sigma}_i \mathbf{D}_i^{\{1\}}$ or element-wise
        $\mathbf{M}_i\left((k_{i-1}, \mu_i); (\mu_{i+1}, \ldots, \mu_d)\right) = \sum_{k_i=1}^{s_i} \sigma_{k_i}^i \mathbf{G}_i(k_{i-1}, \mu_i, k_i) \mathbf{D}_i(k_i, \mu_{i+1}, \ldots, \mu_d)$,
        where $\left(\sigma_{k_i}^i\right)_{k_i}$ is the monotonically decreasing sequence of singular values of $\mathbf{M}^i$.
    **Set** $\mathbf{V}_{i+1}^{\{1\}} := \mathbf{\Sigma}_i \mathbf{D}_i^{\{1\}}$ **and** $\mathbf{M}_{i+1} := \mathbf{V}_{i+1}^{\{1,2\}}$.
**end for**
**Set** $[\mathbf{G}_d(\mu_d)](k_{d-1}) := \mathbf{M}_d(k_{d-1}, \mu_d)$ **and** $[\mathbf{G}_i(\mu_i)](k_{i-1}, k_i) = \mathbf{G}_i(k_{i-1}, \mu_i, k_i)$ for $i \in [d]$.
**Output**: Decomposition $\mathbf{X}(\mu_1, \mu_2, \ldots, \mu_d) = \mathbf{G}_1(\mu_1)\mathbf{G}_2(\mu_2) \cdots \mathbf{G}_d(\mu_d)$.
```

Let us notice that the present algorithm is not the only way to use multiple singular value decompositions in order to obtain a hierarchical representation of $\mathbf{X}$ for the given tree, here a TT representation. For example, one may start at the right end separating $\mathbf{B}_d$ first and so on. The procedure above provides some *normal form* of the tensor.

Let us now explain hard thresholding on the example of a TT tensor. This procedure remains essentially the same to the TT-SVD algorithm – presented in Algorithm 2.2 – (and more generally HT-SVD algorithm) with the only difference that we apply a thresholding to a target rank $\mathbf{r} = (r_i)_{i=1}^{d-1}$ with $r_i \leq s_i$ at each step of the for loop by setting $\sigma_{k_i}^i = 0$ for all $k_i > r_i$, $i \in [d-1]$ (where $\left(\sigma_{k_i}^i\right)_{k_i}$ is the monotonically decreasing sequence of singular values of $\mathbf{M}_i$). This results in a matrix $\mathbf{M}_{i,\varepsilon}$ satisfying $\|\mathbf{M}_i - \mathbf{M}_{i,\varepsilon_i}\|_F = \epsilon_i = \sqrt{\sum_{k_i > r_i} (\sigma_{k_i}^i)^2}$. By the *hard thresholding* procedure presented above, one obtains a unique approximate tensor

$$\mathbf{X}_\epsilon := \mathcal{H}_\mathbf{r}(\mathbf{X})$$

of multi-linear rank $\mathbf{r}$ within a guaranteed error bound

$$\|\mathbf{X}_\epsilon - \mathbf{X}\|_F \leq \sum_{i=1}^{d-1} \epsilon_i \ .$$

In contrast to the matrix case, this approximation $\mathbf{X}_\epsilon$, however, may not be the best rank $\mathbf{r}$ approximation of $\mathbf{X}$, which is in fact NP-hard to compute [61, 82]. A more evolved analysis shows the following quasi-optimal error bound.

The procedure introduced above can be modified to apply for general hierarchical tensor representations. In Appendix A we introduce three procedures due to Grasedyck [70], namely the *root-to-leaves truncation* (Algorithm A.1), the *leaves-to-root truncation* (Algorithm A.2), and *truncation via projections* (Theorem A.6).

Let $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ and let $\mathbf{X}^{\alpha} = \mathbf{U}_{\alpha} \mathbf{\Sigma}_{\alpha} \mathbf{V}_{\alpha}^{T}$ be the singular value decomposition of the tensor unfolding $\mathbf{X}^{\alpha}$ with $\alpha \in [d]$ and $\mathbf{U}_{\alpha} \in \mathbb{R}^{n_{\alpha} \times n_{\alpha}}$. Then the truncation of $\mathbf{X}$ to Tucker rank $\mathbf{r} = (r_1, r_2, \ldots, r_d)$ is defined by

$$\mathcal{H}_{\mathbf{r}}\left(\mathbf{X}\right) := \mathbf{X} \times_1 \tilde{\mathbf{U}}_1 \tilde{\mathbf{U}}_1^T \times_2 \tilde{\mathbf{U}}_2 \tilde{\mathbf{U}}_2^T \times \cdots \times_d \tilde{\mathbf{U}}_d \tilde{\mathbf{U}}_d^T,$$

where $\tilde{\mathbf{U}}_k$ is the matrix of first $r_k$ columns of $\mathbf{U}_k$. This definition was introduced in [43].

Next, we present a result on truncation error.

**Theorem 2.6.** Let $\mathbf{X}_{\epsilon} = \mathcal{H}_{\mathbf{r}}(\mathbf{X})$. Then there exists $C(d) = \mathcal{O}(\sqrt{d})$, such that $\mathbf{X}_{\epsilon}$ satisfies the quasi-optimal error bound

$$\inf\{\|\mathbf{X} - \mathbf{Y}\|_F : \mathbf{Y} \in \mathcal{M}_{\leq \mathbf{r}}\} \leq \|\mathbf{X} - \mathcal{H}_{\mathbf{r}}(\mathbf{X})\|_F \leq C(d) \inf\{\|\mathbf{X} - \mathbf{Y}\|_F : \mathbf{Y} \in \mathcal{M}_{\leq \mathbf{r}}\} \, .$$

The constant satisfies $C(d) = \sqrt{d}$ for the Tucker format [70], $C(d) = \sqrt{d-1}$ for the TT format [125], and $C(d) = \sqrt{2d-3}$ for a balanced tree (and truncation via projections) in the HT-format [70].

When we consider the HT format in the sequel, we have in mind that we have fixed our tensor SVD method choosing one of the several variants.

CHAPTER 3

# Other approaches to low-rank tensor recovery

The problem of low-rank tensor recovery is an interesting subject from both the theoretical and the application point of view. On one side, it is a natural generalization of the sparse vector and low-rank matrix recovery problem. On the other side, estimating a low-rank tensor has applications in many different areas such as machine learning [137], video compression [108], and seismic data interpolation [38].

The aim of low-rank tensor recovery is to reconstruct a low-rank tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ from linear measurements $\mathbf{y} = \mathcal{A}(\mathbf{X})$, where $\mathcal{A} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^m$ with $m \ll n_1 n_2 \cdots n_d$. In Chapter 2, different notions of tensor rank corresponding to different tensor decompositions have been introduced. Thus, to analyze low-rank tensor recovery it is necessary to first fix the tensor decomposition and a corresponding notion of rank. Similarly to the matrix case, the natural approach of finding the solution of the optimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}} \operatorname{rank}(\mathbf{Z}) \quad \text{s.t.} \quad \mathcal{A}(\mathbf{Z}) = \mathbf{y} \tag{3.1}$$

is NP-hard to compute in general – regardless of the choice of the tensor decomposition, see [81]. Based on the experience on the low-rank matrix recovery, one would expect that the convex optimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}} \|\mathbf{Z}\| \quad \text{s.t.} \quad \mathcal{A}(\mathbf{Z}) = \mathbf{y}, \tag{3.2}$$

where $\|\cdot\|$ denotes a suitable norm, would be a promising approach. However, for the CP-decomposition and the corresponding notion of the tensor nuclear norm, also (3.2) is in general NP-hard to compute, see [61, 82]. A recent paper [169] on tensor completion for third order tensors provides the bounds for low-rank tensor recovery via tensor nuclear norm minimization. The analysis is based on a tensor version of coherence. That is, the first step is to compute coherences of linear subspaces spanned by columns of the unfoldings. The tensor coherence then corresponds to the maximum of these three subspace coherences. Theoretical results are significantly better than the ones provided by other approaches – the sample size requirement for robust third order tensor completion in $\mathbb{R}^{n \times n \times n}$ is $\mathcal{O}(r^{1/2}(n \log n)^{3/2})$. Still, the computation remains NP-hard.

In Chapter 4 we introduce another approach related to the same decomposition. We provide convex relaxations of the unit-tensor-nuclear-norm ball which lead to the "new" tensor norms called *theta norms*. In this case, the optimization problem in (3.2) (with $\|\cdot\|$ corresponding to any of the theta norms) can be solved via semidefinite programming. A similar approach based on sum-of-squares relaxations and resulting also in the "new" tensor norms was suggested in [4]. This approach is based on the Lassere's relaxations (see Remark B.2 for the difference between theta bodies and Lassere's relaxations). In particular, at the sixth level of Lassere's hierarchy $m = \tilde{O}(rn^{3/2})$ number of measurements is required to recover a tensor in $\mathbb{R}^{n \times n \times n}$ and rank at most $r$ via tensor completion. However, the method is highly non-scalable, i.e., it runs in exponential time (since it requires solving optimization problems at the sixth level of Lassere's hierarchy).

A recent paper [90] develops parallel algorithms for tensor completion in the CP-format and provides local convergence results. The authors generalize the well known parallel approaches for matrix completion based on ALS (alternating least-squares) [153, 171], CCD (cyclic coordinate descent) [128, 167, 168] and SGD (stochastic gradient descent) [64, 134, 153] to tensor completion. For other approaches to low-rank tensor recovery and tensor completion considering the CP-decomposition, see e.g. [107, 115].

Recall that the Tucker rank of a $d$th order tensor $\mathbf{X}$ is a vector $\mathbf{r} = (r_1, r_2, \ldots, r_d)$, with the $k$-th entry corresponding to the rank of the $k$th unfolding. That is, $r_k = \operatorname{rank}\left(\mathbf{X}^{\{k\}}\right)$ for all $k \in [d]$. Then the minimization problem (3.1) can be transferred into

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}} \sum_{i=1}^{d} \operatorname{rank}\left(\mathbf{Z}^{\{i\}}\right) \quad \text{s.t.} \quad \mathcal{A}(\mathbf{Z}) = \mathbf{y}. \tag{3.3}$$

This suggests a natural, tractable convex approach to the recovery of low-rank tensors

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}} \sum_{i=1}^{d} \lambda_i \left\|\mathbf{Z}^{\{i\}}\right\|_* \quad \text{s.t.} \quad \mathcal{A}(\mathbf{Z}) = \mathbf{y}, \tag{3.4}$$

with some positive weights $\lambda_i$ (usually $\lambda_i = 1$ or $1/d$, for all $i \in [d]$) and where $\left\|\mathbf{Z}^{\{i\}}\right\|_* = \sum_j \boldsymbol{\sigma}_i(j)$ denotes the sum of the singular values of the $i$-th unfolding of the tensor $\mathbf{Z}$. This approach has been originally proposed in [108] as the first approach suggested for the low-rank recovery, it has been widely studied in [62, 83, 141, 142, 156], and applied to various datasets in imaging in [98, 105, 106, 139]. For notational purposes we introduce the set

$$\mathcal{T}_{d,n,r} = \left\{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} : n = n_i \text{ and } \operatorname{rank}\left(\mathbf{X}^{\{i\}}\right) = r, \text{ for all } i \in [d]\right\}.$$

In paper [156], the authors show that a tensor $\mathbf{X} \in \mathcal{T}_{d,n,r}$ can be recovered from Gaussian measurements provided that the number of observations scales like $m \geq Crn^{d-1}$. Additionally, it has been shown that this number of measurements is necessary, see [115]. However, to describe a generic tensor from $\mathcal{T}_{d,n,r}$, we need at most $r^d + drn$ parameters. Thus, there is a substantial gap between the intrinsic degrees of freedom of a generic tensor in $\mathcal{T}_{d,n,r}$ and the necessary number of measurements to ensure recovery. We present the main result from paper [115].

**Theorem 3.1** ([115]). Let $\mathbf{X}_0 \in \mathbb{R}^{n \times n \times \cdots \times n}$ be an order-$d$ tensor. Consider an optimization problem for fixed $j = \left\lceil \frac{d}{2} \right\rceil$

$$\min_{\mathbf{X}} \left\|\mathbf{X}^{\{1 \ldots j\}}\right\|_* \quad \text{subject to} \quad \mathcal{A}(\mathbf{X}) = \mathcal{A}(\mathbf{X}_0), \tag{3.5}$$

where $\mathbf{X}^{\{1 \ldots j\}}$ denotes the $\{1 \ldots j\}$-th matricization of a tensor $\mathbf{X}$. Then if

- $\mathbf{X}_0$ has a CP-rank $r_{cp}$, a sufficient number of measurements to recover $\mathbf{X}_0$ with high probability via (3.5) is $m \geq Cr_{cp}n^j$.
- $\mathbf{X}_0$ has a Tucker rank $\mathbf{r} = (r, r, \ldots, r)$ (i.e., $\mathbf{X}_0 \in \mathcal{T}_{d,n,r}$), a sufficient number of measurements to recover $\mathbf{X}_0$ with high probability via (3.5) is $m \geq Cr^j n^j$.

The above result follows from the low-rank matrix recovery results in [32] and noting that

$$\operatorname{rank}\left(\mathbf{X}^{\{1\ldots j\}}\right) \leq r_{cp} \quad \text{and} \quad \operatorname{rank}\left(\mathbf{X}^{\{1\ldots j\}}\right) \leq \min\left\{\prod_{i=1}^{j} r_i, \prod_{i=j+1}^{d} r_i\right\},$$

where $\mathbf{r} = (r_1, r_2, \ldots, r_d)$ denotes the Tucker rank. Interestingly and possibly surprisingly, the above theorem states that, at least theoretically, the necessary number of measurements scales equally for (3.4) and (3.5). That is, to recover a low-rank $d$th order tensor it is enough to consider only $\{1, \ldots, j\}$-th matricization with $j = \left\lceil \frac{d}{2} \right\rceil$. Clearly, in this approach the structure of a tensor is completely lost (since we are basically recovering a low-rank matrix instead of a low-rank tensor).

Notice that for the TT-decomposition and, more generally, the HT-decomposition, it is possible to define the minimization problems analogously to (3.3) and (3.4) – by considering the corresponding matricizations related to the TT and HT rank, respectively. However, the paper [115] shows that the corresponding convex minimization programs will not, in general, improve the recovery result given in Theorem 3.1.

In papers [51, 52] Baraniuk and Duarte have introduced a new convex optimization approach for tensor completion based on the Kronecker product (see Definition A.2) and Tucker decomposition. In several applications (see [20] for details), for example compressive sensing MRI with multiscale scanning and Hyper-spectral compressive imaging [52], the measurements can be taken in a multilinear way (which explains the use of the Tucker decomposition). That is, by using linear operators in each mode separately as follows

$$\mathbf{Y} = \mathbf{X} \times_1 \mathbf{D}_1 \times_2 \mathbf{D}_2 \times_3 \mathbf{D}_3,$$

which can be written in terms of vectorized tensors (and Kronecker product $\otimes_K$) as

$$\mathbf{y} = (\mathbf{D}_3 \otimes_K \mathbf{D}_2 \otimes_K \mathbf{D}_1)\, \mathbf{x},$$

where $\mathbf{x} = \operatorname{vec}(\mathbf{X})$ and $\mathbf{y} = \operatorname{vec}(\mathbf{Y})$. Thus, in this scenario the following convex optimization program is considered

$$\min_{\mathbf{z}} \|\mathbf{z}\|_0 \quad \text{s.t.} \quad \mathbf{y} = (\mathbf{D}_3 \otimes_K \mathbf{D}_2 \otimes_K \mathbf{D}_1)\, \mathbf{z},$$

where the original tensor $\mathbf{X}$ satisfies $\mathbf{y} = (\mathbf{D}_3 \otimes_K \mathbf{D}_2 \otimes_K \mathbf{D}_1)\, \mathbf{x}$ and, as before, $\mathbf{y}$ is known, as well as the matrices $\mathbf{D}_i$. The authors also provide the recovery guarantees. However, the analysis is still based on the matrix completion results. Recently, Caiafa and Cichocki have presented three versions of the OMP algorithm adapted to the Kronecker setting, see [19, 20]. The best result is obtained for the N-BOMP (N-way Block OMP) algorithm – see Algorithm 3.1 – under the assumption of the "block-tensor sparsity". In Algorithm 3.1 the $i_n$-th column of the matrix $\mathbf{D}_n$ is denoted by $\mathbf{D}_n(:, i_n)$ and $\mathbf{X}_{\boldsymbol{\mathcal{I}}_1 \times \boldsymbol{\mathcal{I}}_2 \times \cdots \times \boldsymbol{\mathcal{I}}_d}$ denotes a subtensor of $\mathbf{X}$ obtained by keeping in $k$-mode only the indexes in $\boldsymbol{\mathcal{I}}_k$ for all $k \in [d]$.

**Definition 3.2** ([19])**.** A multidimensional signal $\mathbf{Y} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_d}$ is $(S_1, S_2, \ldots, S_d)$-block sparse with respect to the factors $\mathbf{D}_n \in \mathbb{R}^{I_n \times M_n}$ $(n = 1, 2, \ldots, d)$ if it admits a Tucker representation based only on few $S_n$ selected columns of each factor $(S_n \leq M_n)$. That is, if $\boldsymbol{\mathcal{I}}_n = \left[i_n^1, i_n^2, \ldots, i_n^{S_n}\right]$ denotes a subset of indices for the mode n $(n = 1, 2, \ldots, d)$, then

$$\mathbf{Y} = \mathbf{X} \times_1 \mathbf{D}_1 \times_2 \mathbf{D}_2 \times \cdots \times_d \mathbf{D}_d,$$

with $\mathbf{X}(i_1, i_2, \ldots, i_d) = 0$, for all $(i_1, i_2, \ldots, i_d) \notin \boldsymbol{\mathcal{I}}_1 \times \boldsymbol{\mathcal{I}}_2 \times \cdots \times \boldsymbol{\mathcal{I}}_d$.

**Algorithm 3.1.** N-BOMP Algorithm

```
Input: mode-$n$ dictionaries $\{\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_d\}$ with $\mathbf{D}_n \in \mathbb{R}^{I_n \times M_n}$,
       signal $\mathbf{Y} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_d}$,
       maximum number of nonzero entries $K_{\max}$, tolerance $\varepsilon$
Initialization: $\boldsymbol{\mathcal{I}}_n = \emptyset$, for $n \in [d]$, $\mathbf{R} = \mathbf{Y}$, $\mathbf{X} = \mathbf{0}$, $j = 1$.
while $|\boldsymbol{\mathcal{I}}_1| \, |\boldsymbol{\mathcal{I}}_2| \cdots |\boldsymbol{\mathcal{I}}_d| \leq K_{\max}$ and $\|\mathbf{R}\|_F > \varepsilon$ do
    $(i_1^j, i_2^j, \ldots, i_d^j) = \arg\max_{(i_1,i_2,\ldots,i_d)} \left|\mathbf{R} \times_1 \mathbf{D}_1^T(:,i_1) \times_2 \mathbf{D}_2^T(:,i_2) \times \cdots \times_d \mathbf{D}_d^T(:,i_d)\right|$.
    $\boldsymbol{\mathcal{I}}_n = \boldsymbol{\mathcal{I}}_n \cup \{i_n^j\}$ $(n \in [d])$, $\mathbf{B}_n = \mathbf{D}_n(:, \boldsymbol{\mathcal{I}}_n)$.
    $\mathbf{a} = \arg\min_{\mathbf{u}} \|(\mathbf{B}_d \otimes \mathbf{B}_{d-1} \otimes \cdots \otimes \mathbf{B}_1)\,\mathbf{u} - \mathbf{y}\|_2^2$.
    $\mathbf{R} = \mathbf{Y} - \mathbf{A} \times_1 \mathbf{B}_1 \times_2 \mathbf{B}_2 \times \cdots \times_d \mathbf{B}_d$, with $\mathbf{a} = \mathrm{vec}(\mathbf{A})$.
    $j = j + 1$.
end while
Output: Sparse representation $\mathbf{Y} \approx \mathbf{X} \times_1 \mathbf{D}_1 \times_2 \mathbf{D}_2 \times \cdots \times_d \mathbf{D}_d$ with
        $\mathbf{X}(i_1, i_2, \ldots, i_d) = 0$, for all $(i_1, i_2, \ldots, i_d) \notin \boldsymbol{\mathcal{I}}_1 \times \boldsymbol{\mathcal{I}}_2 \times \cdots \times \boldsymbol{\mathcal{I}}_d$
        (with nonzero entries given by $\mathbf{X}_{\boldsymbol{\mathcal{I}}_1 \times \boldsymbol{\mathcal{I}}_2 \times \cdots \times \boldsymbol{\mathcal{I}}_d} = \mathbf{A}$.)
```

The analysis of the N-BOMP algorithm is based on the coherences of the dictionaries $\mathbf{D}_1$, $\mathbf{D}_2$, $\ldots, \mathbf{D}_d$.

**Theorem 3.3** (N-BOMP Performance Guarantee, [19])**.** Given the decomposition $\mathbf{Y} = \mathbf{X} \times_1 \mathbf{D}_1 \times_2 \mathbf{D}_2 \times \cdots \times_d \mathbf{D}_d$, with a fixed tensor $\mathbf{Y} \in \mathbb{R}^{I \times I \times \cdots \times I}$ and known dictionaries $\mathbf{D}_n \in \mathbb{R}^{I \times M}$ having coherences $\mu_n$ $(n = 1, 2, \ldots, d)$, if an $(S, S, \ldots, S)$-block sparse solution exists satisfying

$$(S\mu)^d < 2 - (1 + (S-1)\,\mu)^d\,, \tag{3.6}$$

with $\mu = \max\{\mu_1, \mu_2, \ldots, \mu_d\}$, then the N-BOMP algorithm (Algorithm 3.1) is guaranteed to find this sparse representation in $K$ iterations with $S \leq K \leq dS$.

The condition (3.6) is quite strong and it demands that the coherence for each dictionary $\mathbf{D}_n$ is quite small, i.e., close to zero. In addition, the total cost of the algorithm applied to the case as in the above theorem when $S \ll I < M$ is $IdS(M)^d$, see [20].

We remark that this approach has been applied also to the sparse vector recovery, where the signal $\mathbf{x} \in \mathbb{C}^N$ is rewritten as a matrix $\mathbf{X} \in \mathbb{C}^{n_1 \times n_2}$ with $N = n_1 n_2$, see [160].

Several approaches from compressive sensing and low-rank matrix recovery have been generalized and analyzed for low-rank tensor recovery and tensor completion. For example, in [99, 145] Riemannian optimization for tensor completion is suggested and in [109] generalized higher-order iteration algorithm (gHOI) for tensor completion which is a generalization of ADMM (alternating direction method of multipliers) is developed.

In Chapter 5, we present the *iterative hard thresholding algorithm* and the *normalized iterative hard thresholding algorithm* adapted to the tensor scenario. Unfortunately, due to the properties of $d$th order tensors presented in detail in previous chapter, only partial convergence results are provided. That is, showing either local convergence (see [130]) or global convergence with an additional assumption on the truncation operator $\mathcal{H}_{\mathbf{r}}$.

CHAPTER 4

# Low-rank tensor recovery via tensor theta norms

The tensor nuclear norm minimization, which could be considered as the natural convex optimization approach to the low-rank tensor recovery, is NP-hard in general – see the beginning of Chapter 3. Recent convex optimization approaches suggest minimizing sums of nuclear norms of specific matricizations. In this way, the theory developed for low-rank matrix recovery can be applied to the tensor scenario. However, as already mentioned earlier, in this scenario the minimal number of measurements needed for low-rank tensor recovery scales exponentially in the dimension and thus, does not provide completely satisfying results. In this chapter we present an alternative convex optimization approach. We introduce hierarchical relaxations (supersets) $\boldsymbol{\mathcal{B}}_k$ of the unit-tensor-nuclear-norm ball which provide us with – to the best of our knowledge – new tensor norms called $\theta_k$*-norms*. More precisely, the sets $\boldsymbol{\mathcal{B}}_k$ satisfy

$$\{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} : \|\mathbf{X}\|_* \leq 1\} \subseteq \cdots \subseteq \boldsymbol{\mathcal{B}}_k \subseteq \boldsymbol{\mathcal{B}}_{k-1} \subseteq \cdots \subseteq \boldsymbol{\mathcal{B}}_1$$

and $\{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} : \|\mathbf{X}\|_{\theta_k} \leq 1\} = \boldsymbol{\mathcal{B}}_k$, for all $k$. Thus, $\theta_k$-norms satisfy $\|\mathbf{X}\|_* \geq \cdots \geq \|\mathbf{X}\|_{\theta_k} \geq \|\mathbf{X}\|_{\theta_{k-1}} \geq \cdots \geq \|\mathbf{X}\|_{\theta_1}$, for all $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$.

This approach was first suggested in [32] and it is based on *theta bodies* of an appropriately defined polynomial ideal $\mathcal{J}_d$ in $\mathbb{R}[\mathbf{x}] = \mathbb{R}[x_{11\ldots1}, x_{11\ldots2}, \ldots, x_{n_1 n_2 \ldots n_d}]$. Theta bodies are recently introduced tool from real algebraic geometry with a special case introduced first by Lovász in [110] and in full generality in [9, 68]. We treat each entry of a tensor as a polynomial variable. The idea is to define a polynomial ideal $\mathcal{J}_d$ which vanishes on the set $\nu_{\mathbb{R}}(\mathcal{J}_d)$ of all rank-one Frobenius-norm-one $d$th-order tensors. This is achieved by taking all order-two minors of all matricizations (to satisfy the rank-one condition) and a polynomial $\sum_{i_1, i_2, \ldots, i_d} x_{i_1 i_2 \ldots i_d}^2 - 1$ (to satisfy the unit norm condition) as its basis. In this scenario the convex hull of the set $\nu_{\mathbb{R}}(\mathcal{J}_d)$ forms the unit-tensor-nuclear-norm ball. For $k \in \mathbb{N}$, we define the tensor $\theta_k$-norm via its unit-norm ball as already explained above. That is, the set $\boldsymbol{\mathcal{B}}_k$ is the $k$th theta body of the polynomial ideal $\mathcal{J}_d$. By the theory already developed for theta bodies, all theta norms can be computed via semidefinite programing. However, to build this semidefinite program it is required to compute the reduced Gröbner basis (with respect to a particular monomial ordering – in this case the graded reverse lexicographic ordering) of the polynomial ideal $\mathcal{J}_d$. In fact, in this chapter we provide the reduced Gröbner basis with respect to the graded reverse lexicographic ordering of the polynomial ideal $J_d$, with $d \geq 3$. Interestingly, in the matrix scenario, the theta body approach does not lead to the new matrix norms. That is, we prove that all matrix $\theta_k$-norms are equal and coincide with the matrix nuclear norm.

The theta body method has lead us to the polynomial ideals $\mathcal{J}_{t,d}$ in $\mathbb{R}[\mathbf{x}]$ generated by all order-$t$ minors of all matricizations of tensor $\mathbf{X}$ of indeterminates. These ideals could be considered as a natural higher-order generalization of the determinantal ideals $\mathcal{I}_t$ in $\mathbb{R}[x_{11}, x_{12}, \ldots, x_{n_1 n_2}]$

generated by all order-$t$ minors of matrix $\mathbf{X}$ of indeterminates. Determinantal ideals have already been studied in real algebraic geometry and commutative algebra. However, the ideals $\mathcal{J}_{t,d}$ – up to the best of our knowledge – have not been considered before. In this chapter, we additionally compute the reduced Gröbner basis of the polynomial ideals $\mathcal{J}_{2,d}$.

Finally, we provide a semidefinite program for computing tensor $\theta_1$-norm and for low-rank third-order tensor recovery via $\theta_1$-norm minimization. Our numerical experiments presented in Section 4.5 indicate that the low-rank tensor recovery via $\theta_1$-norm minimization is a promising approach.

Since we treat each entry of a tensor as a polynomial variable, for better readability, the $(\alpha_1, \alpha_2, \ldots, \alpha_d)$-th entry of a tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ will be denoted in this chapter as $X_{\alpha_1\alpha_2\ldots\alpha_d}$ instead of $\mathbf{X}(\alpha_1, \alpha_2, \ldots, \alpha_d)$. For simplicity, the subscripts $\alpha_1\alpha_2 \cdots \alpha_d$ and $\beta_1\beta_2 \cdots \beta_d$ will often be denoted by $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. In particular, instead of writing $x_{\alpha_1\alpha_2\ldots\alpha_d}x_{\beta_1\beta_2\ldots\beta_d}$, we often just write $x_{\boldsymbol{\alpha}}x_{\boldsymbol{\beta}}$. Below, we will use the grevlex ordering of monomials indexed by subscripts $\boldsymbol{\alpha}$, which in particular requires to define an ordering for such subscripts. We make the agreement that $x_{11\ldots11} > x_{11\ldots12} > \ldots > x_{11\ldots1n_d} > x_{11\ldots21} > \ldots > x_{1n_2\ldots n_d} > \ldots > x_{n_1n_2\ldots n_d}$.

The following results are contained in our paper [132]. However, in this chapter we use a more standard notation from commutative algebra and real algebraic geometry – especially, when referring to determinantal ideals.

## 4.1. Introduction to theta bodies

The computation of the theta bodies and the corresponding $\theta_k$-norms requires several definitions and tools from real algebraic geometry which are collected in Appendix B. In particular, in Subsection B.2 we introduce Gröbner bases of polynomial ideals and several monomial orderings including the *graded reverse lexicographic* (or *grevlex*) *ordering* which is used throughout this chapter. For an intuition behind the theta bodies and in general sum-of-squares certificates, see Subsection B.1.

For a nonzero polynomial $f = \sum_{\boldsymbol{\alpha}} a_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} = \sum_{\alpha_1,\alpha_2,\ldots,\alpha_n} a_{\alpha_1,\alpha_2,\ldots,\alpha_n} x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$ in $\mathbb{R}[\mathbf{x}] = \mathbb{R}[x_1, x_2, \ldots, x_n]$ and a monomial order $>$, we denote

a) the multidegree of $f$ with $\operatorname{multideg}(f) = \arg\max\left(\mathbf{x}^{\boldsymbol{\alpha}} : a_{\boldsymbol{\alpha}} \neq 0, \boldsymbol{\alpha} \in \mathbb{Z}^n_{\geq 0}\right)$,
(the maximum is taken with respect to the fixed monomial ordering)

b) the leading coefficient of $f$ with $\operatorname{LC}(f) = a_{\operatorname{multideg}(f)} \in \mathbb{R}$,

c) the leading monomial of $f$ with $\operatorname{LM}(f) = \mathbf{x}^{\operatorname{multideg}(f)}$,

d) the leading term of $f$ with $\operatorname{LT}(f) = \operatorname{LC}(f) \cdot \operatorname{LM}(f)$.

Let $f(x_1, x_2) = -x_1^3 x_2 + 3x_2^2 - 2x_1x_2 + 4x_1 + 1$ be a polynomial in $\mathbb{R}[x_1, x_2]$. Then the multidegree, the leading coefficient, the leading monomial, and the leading term of $f$ with respect to the *grevlex ordering* induced by variable ordering $x_1 > x_2$, see Definition B.3, are $\operatorname{multideg}(f) = (3, 1)$, $\operatorname{LC}(f) = -1$, $\operatorname{LM}(f) = x_1^3 x_2$, and $\operatorname{LT}(f) = -x_1^3 x_2$, respectively.

Let $\mathcal{J} \subset \mathbb{R}[\mathbf{x}] = \mathbb{R}[x_1, x_2, \ldots, x_n]$ be a polynomial ideal. Its real algebraic variety is the set of points $\mathbf{x} \in \mathbb{R}^n$ at which all polynomials of the ideal $\mathcal{J}$ vanish, i.e.,

$$\nu_{\mathbb{R}}(\mathcal{J}) = \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) = 0, \text{ for all } f \in \mathcal{J}\}.$$

By Hilbert's basis theorem [37] every polynomial ideal in $\mathbb{R}[\mathbf{x}]$ has a finite generating set. Thus, we may assume that $\mathcal{J}$ is generated by a set $\boldsymbol{\mathcal{F}} = \{f_1, f_2, \ldots, f_k\}$ of polynomials in $\mathbb{R}[\mathbf{x}]$ and write

$$\mathcal{J} = \langle f_1, f_2, \ldots, f_k \rangle = \left\langle \{f_i\}_{i\in[k]} \right\rangle \quad \text{or simply} \quad \mathcal{J} = \langle \boldsymbol{\mathcal{F}} \rangle .$$

Its real algebraic variety is the set

$$\nu_{\mathbb{R}}(\mathcal{J}) = \{\mathbf{x} \in \mathbb{R}^n : f_i(\mathbf{x}) = 0 \text{ for all } i \in [k]\}.$$

Throughout the chapter, $\mathbb{R}[\mathbf{x}]_k$ denotes the set of polynomials of degree at most $k$ and a degree-one polynomial will be called a linear polynomial. A very useful certificate for the positivity of polynomials is contained in the following definition.

**Definition 4.1** ([68])**.** Let $\mathcal{J}$ be an ideal in $\mathbb{R}[\mathbf{x}]$. A polynomial $f \in \mathbb{R}[\mathbf{x}]$ is *$k$-sos mod $\mathcal{J}$* if there exists a finite set of polynomials $h_1, h_2, \ldots, h_t \in \mathbb{R}[\mathbf{x}]_k$ such that $f \equiv \sum_{j=1}^{t} h_j^2 \mod \mathcal{J}$, i.e, if $f - \sum_{j=1}^{t} h_j^2 \in \mathcal{J}$.

A special case of theta bodies was first introduced by Lovász in [110] and in full generality they appeared in [68]. Later, they have been analyzed in [67, 69]. The definitions and theorems in the remainder of the section are taken from [68].

**Definition 4.2** (Theta body)**.** Let $\mathcal{J} \subseteq \mathbb{R}[\mathbf{x}]$ be an ideal. For a positive integer $k$, the *$k$-th theta body of $\mathcal{J}$* is defined as

$$\mathrm{TH}_k(\mathcal{J}) := \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \geq 0 \text{ for every linear } f \text{ that is } k\text{-sos mod } \mathcal{J}\}.$$

We say that an ideal $\mathcal{J} \subseteq \mathbb{R}[\mathbf{x}]$ is $\mathrm{TH}_k$-*exact* if $\mathrm{TH}_k(\mathcal{J})$ equals $\overline{\mathrm{conv}(\nu_{\mathbb{R}}(\mathcal{J}))}$, i.e., the closure of the convex hull of $\nu_{\mathbb{R}}(\mathcal{J})$.

Theta bodies are closed convex sets, while $\mathrm{conv}(\nu_{\mathbb{R}}(\mathcal{J}))$ may not necessarily be closed and by definition,

$$\mathrm{TH}_1(\mathcal{J}) \supseteq \mathrm{TH}_2(\mathcal{J}) \supseteq \cdots \supseteq \mathrm{TH}_{k-1}(\mathcal{J}) \supseteq \mathrm{TH}_k(\mathcal{J}) \supseteq \cdots \supseteq \mathrm{conv}(\nu_{\mathbb{R}}(\mathcal{J})). \tag{4.1}$$

The theta-body sequence of an ideal $\mathcal{J}$ can converge (finitely or asymptotically), if at all, only to $\overline{\mathrm{conv}(\nu_{\mathbb{R}}(\mathcal{J}))}$. More results regarding such guarantees can be found in [68, 69]. However, to the best of our knowledge, none of the existing guarantees apply to the cases discussed below.

Given any polynomial, it is possible to check whether it is $k$-sos mod $\mathcal{J}$ using a Gröbner basis of $\mathcal{J}$ and semidefinite programming. However, using this definition in practice requires the knowledge of all linear polynomials (possibly infinitely many) that are $k$-sos mod $\mathcal{J}$. To overcome this difficulty, we need an alternative description of $\mathrm{TH}_k(\mathcal{J})$ discussed next.

As in [9], we assume that there are no linear polynomials in the ideal $\mathcal{J}$. Otherwise, some variable $x_i$ would be congruent to a linear combination of other variables modulo $\mathcal{J}$ and we could work in a smaller polynomial ring $\mathbb{R}[x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]$. Therefore, $\mathbb{R}[\mathbf{x}]_1 / \mathcal{J} \cong \mathbb{R}[\mathbf{x}]_1$ and $\{1 + \mathcal{J}, x_1 + \mathcal{J}, \ldots, x_n + \mathcal{J}\}$ can be completed to a basis $\boldsymbol{\mathcal{B}}$ of $\mathbb{R}[\mathbf{x}]/\mathcal{J}$. Recall that the degree of an equivalence class $f + \mathcal{J}$, denoted by $\deg(f + \mathcal{J})$, is the smallest degree of an element in its class. We assume that each element in the basis $\boldsymbol{\mathcal{B}} = \{f_i + \mathcal{J}\}$ of $\mathbb{R}[\mathbf{x}]/\mathcal{J}$ is represented by the polynomial whose degree equals the degree of its equivalence class, i.e., $\deg(f_i + \mathcal{J}) = \deg(f_i)$. In addition, we assume that $\boldsymbol{\mathcal{B}} = \{f_i + \mathcal{J}\}$ is ordered so that $\deg(f_{i+1}) > \deg(f_i)$, where $>$ is a fixed monomial ordering. Further, we define the set $\boldsymbol{\mathcal{B}}_k$

$$\boldsymbol{\mathcal{B}}_k := \{f + \mathcal{J} \in \boldsymbol{\mathcal{B}} : \deg(f + \mathcal{J}) \leq k\}.$$

**Definition 4.3** (Theta basis)**.** Let $\mathcal{J} \subseteq \mathbb{R}[\mathbf{x}]$ be an ideal. A basis $\boldsymbol{\mathcal{B}} = \{f_0 + \mathcal{J}, f_1 + \mathcal{J}, \ldots\}$ of $\mathbb{R}[\mathbf{x}]/\mathcal{J}$ is a *$\theta$-basis* if it has the following properties

1) $\boldsymbol{\mathcal{B}}_1 = \{1 + \mathcal{J}, x_1 + \mathcal{J}, \ldots, x_n + \mathcal{J}\}$,
2) if $\deg(f_i + \mathcal{J}), \deg(f_j + \mathcal{J}) \leq k$ then $f_i f_j + \mathcal{J}$ is in the $\mathbb{R}$-span of $\boldsymbol{\mathcal{B}}_{2k}$.

As in [9, 68] we consider only monomial bases $\boldsymbol{\mathcal{B}}$ of $\mathbb{R}[\mathbf{x}]/\mathcal{J}$, i.e., bases $\boldsymbol{\mathcal{B}}$ such that $f_i$ is a monomial, for all $f_i + \mathcal{J} \in \boldsymbol{\mathcal{B}}$.

To determine a $\theta$-basis, we first need to compute the reduced Gröbner basis $\boldsymbol{\mathcal{G}}$ of the ideal $\mathcal{J}$, see Definition B.5 and Definition B.6. The set $\boldsymbol{\mathcal{B}}$ will satisfy the second property in the definition of the theta basis if the reduced Gröbner basis $\boldsymbol{\mathcal{G}}$ is with respect to an ordering which first compares the total degree. Therefore, throughout the paper we use the graded reverse monomial (grevlex) ordering (see Definition B.3), although also the graded lexicographic ordering (see Definition B.3) would be appropriate.

A technique to compute a $\theta$-basis $\boldsymbol{\mathcal{B}}$ of $\mathbb{R}[\mathbf{x}]/\mathcal{J}$ consists in taking $\boldsymbol{\mathcal{B}}$ to be the set of equivalence classes of the standard monomials of the corresponding initial ideal

$$\mathcal{J}_{\text{initial}} = \left\langle \{\text{LT}(f)\}_{f \in \mathcal{J}} \right\rangle = \left\langle \{\text{LT}(g_i)\}_{i \in [s]} \right\rangle,$$

where $\boldsymbol{\mathcal{G}} = \{g_1, g_2, \ldots, g_s\}$ is the reduced Gröbner basis of the ideal $\mathcal{J}$. In other words, a set $\boldsymbol{\mathcal{B}} = \{f_0 + \mathcal{J}, f_1 + \mathcal{J}, \ldots\}$ will be a $\theta$-basis of $\mathbb{R}[\mathbf{x}]/\mathcal{J}$ if it contains all $f_i + \mathcal{J}$ such that

1) $f_i$ is a monomial
2) $f_i$ is not divisible by any of the monomials in the set $\{\text{LT}(g_i) : i \in [s]\}$.

The next important tool we need is the *combinatorial moment matrix* of $\mathcal{J}$. To this end, we fix a $\theta$-basis $\boldsymbol{\mathcal{B}} = \{f_i + \mathcal{J}\}$ of $\mathbb{R}[\mathbf{x}]/\mathcal{J}$ and define $[\mathbf{x}]_{\boldsymbol{\mathcal{B}}_k}$ to be the column vector formed by all elements of $\boldsymbol{\mathcal{B}}_k$ in order. Then $[\mathbf{x}]_{\boldsymbol{\mathcal{B}}_k} [\mathbf{x}]_{\boldsymbol{\mathcal{B}}_k}^T$ is a square matrix indexed by $\boldsymbol{\mathcal{B}}_k$ and its $(i, j)$-entry is equal to $f_i f_j + \mathcal{J}$. By definition of a theta basis, the entries of $[\mathbf{x}]_{\boldsymbol{\mathcal{B}}_k} [\mathbf{x}]_{\boldsymbol{\mathcal{B}}_k}^T$ lie in the $\mathbb{R}$-span of $\boldsymbol{\mathcal{B}}_{2k}$. Let $\{\lambda_{i,j}^l\}$ be the unique set of real numbers such that $f_i f_j + \mathcal{J} = \sum_{f_l + \mathcal{J} \in \boldsymbol{\mathcal{B}}_{2k}} \lambda_{i,j}^l (f_l + \mathcal{J})$.

The theta bodies $\text{TH}_k(\mathcal{J})$ can be characterized via the combinatorial moment matrix as stated in the next result from [68]. Thus, these matrices will be the basis for computing the new tensor norms and recovering low-rank tensors via the new tensor norm minimization introduced below via the semidefinite programming.

**Definition 4.4.** Let $\mathcal{J}, \boldsymbol{\mathcal{B}}$ and $\{\lambda_{i,j}^l\}$ be as above. Let $\mathbf{y}$ be a real vector indexed by $\boldsymbol{\mathcal{B}}_{2k}$ with its first entry $y_0 = 1$ indexed by the basis element $1 + \mathcal{J}$. The *$k$-th combinatorial moment matrix* $\mathbf{M}_{\boldsymbol{\mathcal{B}}_k}(\mathbf{y})$ of $\mathcal{J}$ is the real matrix indexed by $\boldsymbol{\mathcal{B}}_k$ whose $(i, j)$-entry is $[\mathbf{M}_{\boldsymbol{\mathcal{B}}_k}(\mathbf{y})]_{i,j} = \sum_{f_l + \mathcal{J} \in \boldsymbol{\mathcal{B}}_{2k}} \lambda_{i,j}^l y_l$.

**Theorem 4.5.** The $k$-th theta body of $\mathcal{J}$, $\text{TH}_k(\mathcal{J})$, is the closure of

$$\boldsymbol{\mathcal{Q}}_{\boldsymbol{\mathcal{B}}_k}(\mathcal{J}) = \pi_{\mathbb{R}^n} \left\{ \mathbf{y} \in \mathbb{R}^{\boldsymbol{\mathcal{B}}_{2k}} : \mathbf{M}_{\boldsymbol{\mathcal{B}}_k}(\mathbf{y}) \succeq 0, y_0 = 1 \right\},$$

where $\pi_{\mathbb{R}^n}$ denotes the projection onto the variables $y_1 = y_{x_1 + \mathcal{J}}, \ldots, y_n = y_{x_n + \mathcal{J}}$.

Algorithm 4.1 shows a step-by-step procedure for computing $\text{TH}_k(\mathcal{J})$.

**Algorithm 4.1.** Algorithm for computing $\mathrm{TH}_k(\mathcal{J})$

```
**Input: An ideal** $\mathcal{J} \in \mathbb{R}[\mathbf{x}] = \mathbb{R}[x_1, x_2, \ldots, x_n]$.
**Compute the reduced Gröbner basis of the ideal** $\mathcal{J}$.
**Compute a** $\theta$**-basis** $\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{B}}_1 \cup \boldsymbol{\mathcal{B}}_2 \cup \ldots = \{f_0 + \mathcal{J}, f_1 + \mathcal{J}, \ldots\}$ **of** $\mathbb{R}[\mathbf{x}]/\mathcal{J}$
                                                        **(see Definition 4.3)**
**Compute the combinatorial moment matrix** $\mathbf{M}_{\boldsymbol{\mathcal{B}}_k}(\mathbf{y})$:
    $[\mathbf{x}]_{\boldsymbol{\mathcal{B}}_k} = \{\text{all elements of } \boldsymbol{\mathcal{B}}_k \text{ in order}\}$
    $(\mathbf{X}_{\boldsymbol{\mathcal{B}}_k})_{i,j} = \left([\mathbf{x}]_{\boldsymbol{\mathcal{B}}_k}[\mathbf{x}]_{\boldsymbol{\mathcal{B}}_k}^T\right)_{i,j} = f_i f_j + \mathcal{J} = \sum_{f_l+\mathcal{J}\in\boldsymbol{\mathcal{B}}_{2k}} \lambda_{i,j}^l (f_l + \mathcal{J})$
    $[\mathbf{M}_{\boldsymbol{\mathcal{B}}_k}(\mathbf{y})]_{i,j} = \sum_{f_l+\mathcal{J}\in\boldsymbol{\mathcal{B}}_{2k}} \lambda_{i,j}^l y_l$    **(linearize using** $\mathbf{y} = (1, y_1, y_2, \ldots, y_{|\boldsymbol{\mathcal{B}}_{2k}|-1})$**)**
**Output:** $\mathrm{TH}_k(\mathcal{J})$ **is the closure of**
    $\boldsymbol{\mathcal{Q}}_{\boldsymbol{\mathcal{B}}_k}(\mathcal{J}) = \pi_{\mathbb{R}^n}\left\{\mathbf{y} \in \mathbb{R}^{\boldsymbol{\mathcal{B}}_{2k}} : \mathbf{M}_{\boldsymbol{\mathcal{B}}_k}(\mathbf{y}) \succeq 0, y_0 = 1\right\}$.
```

## 4.2. Matrix case

First, we consider the unit-matrix-nuclear-norm ball and provide its hierarchical relaxations via theta bodies. The $k$-th relaxation defines a matrix unit $\theta_k$-norm ball with the property

$$\|\mathbf{X}\|_{\theta_k} \leq \|\mathbf{X}\|_{\theta_{k+1}}, \qquad \text{for all } \mathbf{X} \in \mathbb{R}^{m\times n} \text{ and all } k \in \mathbb{N}.$$

However, we will show that all these $\theta_k$-norms coincide with the matrix nuclear norm.

The first step in computing hierarchical relaxations of the unit-matrix-nuclear-norm ball consists in finding a polynomial ideal $\mathcal{J}$ such that its real algebraic variety (the set of points at which all polynomials of the ideal vanish) coincides with the set of all rank-one, Frobenius-norm-one matrices

$$\nu_{\mathbb{R}}(\mathcal{J}) = \left\{\mathbf{X} \in \mathbb{R}^{m\times n} : \|\mathbf{X}\|_F = 1, \operatorname{rank}(\mathbf{X}) = 1\right\}. \tag{4.2}$$

Recall that the convex hull of this set is the nuclear norm ball. The following lemma states the elementary fact that a nonzero matrix is a rank-one matrix if and only if all its order-two minors are zero. (The determinant of an $r$-by-$r$ submatrix of $\mathbf{A}$ is called an *order-$r$ minor of* $\mathbf{A}$ or a *minor of an order $r$ of* $\mathbf{A}$.)

For notational purposes, we define the following polynomials in $\mathbb{R}[\mathbf{x}] = \mathbb{R}[x_{11}, x_{12}, \ldots, x_{mn}]$

$$g(\mathbf{x}) = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}^2 - 1 \text{ and } f_{ijkl}(\mathbf{x}) = x_{il}x_{kj} - x_{ij}x_{kl} \quad \text{for } 1 \leq i < k \leq m,\ 1 \leq j < l \leq n. \tag{4.3}$$

**Lemma 4.6.** Let $\mathbf{X} \in \mathbb{R}^{m\times n}\setminus\{\mathbf{0}\}$. Then $\mathbf{X}$ is a rank-one, Frobenius-norm-one matrix if and only if

$$\mathbf{X} \in \boldsymbol{\mathcal{R}} := \{\mathbf{X} : g(\mathbf{X}) = 0 \text{ and } f_{ijkl}(\mathbf{X}) = 0 \text{ for all } i < k, j < l\}. \tag{4.4}$$

Proof. If $\mathbf{X} \in \mathbb{R}^{m\times n}$ is a rank-one matrix with $\|\mathbf{X}\|_F = 1$, then by definition there exist two vectors $\mathbf{u} \in \mathbb{R}^m$ and $\mathbf{v} \in \mathbb{R}^n$ such that $X_{ij} = u_i v_j$ for all $i \in [m]$, $j \in [n]$ and $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$. Thus

$$X_{ij}X_{kl} - X_{il}X_{kj} = u_i v_j u_k v_l - u_i v_l u_k v_j = 0 \quad \text{and} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2 = \sum_{i=1}^{m} u_i^2 \sum_{j=1}^{n} v_j^2 = 1.$$

For the converse, let $\mathbf{x}_{\cdot i}$ represent the $i$-th column of a matrix $\mathbf{X} \in \boldsymbol{\mathcal{R}}$. Then, for all $j, l \in [n]$ with $j < l$, it holds

$$X_{ml} \cdot \mathbf{x}_{\cdot j} - X_{mj} \cdot \mathbf{x}_{\cdot l} = X_{ml} \cdot \begin{bmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{mj} \end{bmatrix} - X_{mj} \cdot \begin{bmatrix} X_{1l} \\ X_{2l} \\ \vdots \\ X_{ml} \end{bmatrix} = \begin{bmatrix} X_{1j}X_{ml} - X_{1l}X_{mj} \\ X_{2j}X_{ml} - X_{2l}X_{mj} \\ \vdots \\ X_{mj}X_{ml} - X_{mj}X_{ml} \end{bmatrix} = \mathbf{0},$$

since $X_{ij}X_{ml} = X_{il}X_{mj}$ for all $i \in [m-1]$ by the definition of $\boldsymbol{\mathcal{R}}$. Thus, the columns of the matrix $\mathbf{X}$ span a one-dimensional space, i.e., the matrix $\mathbf{X}$ is a rank-one matrix. From $\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2 - 1 = 0$ it follows that the matrix $\mathbf{X}$ is normalized, i.e., $\|\mathbf{X}\|_F = 1$. □

**Remark 4.7.** The set of polynomials $\{f_{ijkl} : 1 \leq i < k \leq m, 1 \leq j < l \leq n\}$ defined in (4.3) is equivalent to the set $\{f^{(\boldsymbol{\alpha},\boldsymbol{\beta})} : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{S}}^2\}$, where

$$f^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{x}) = x_{\boldsymbol{\alpha}}x_{\boldsymbol{\beta}} - x_{\boldsymbol{\alpha}\vee\boldsymbol{\beta}}x_{\boldsymbol{\alpha}\wedge\boldsymbol{\beta}} \quad \text{and} \quad \boldsymbol{\mathcal{S}}^2 = \{(\boldsymbol{\alpha},\boldsymbol{\beta}) : 1 \leq \alpha_1 < \beta_1 \leq m, 1 \leq \beta_2 < \alpha_2 \leq n\},$$

with $[\boldsymbol{\alpha} \vee \boldsymbol{\beta}]_i = \max\{\alpha_i, \beta_i\}$ and $[\boldsymbol{\alpha} \wedge \boldsymbol{\beta}]_i = \min\{\alpha_i, \beta_i\}$ for $i \in \{1, 2\}$. These polynomials were introduced for the first time in [80] and are known in real algebraic geometry and commutative algebra as Hibi relations.

As a consequence of Lemma 4.6, the set of rank-one, Frobenius-norm-one matrices coincides with the real algebraic variety $\nu_{\mathbb{R}}(\mathcal{J}_{M_{mn}})$ for the ideal $\mathcal{J}_{M_{mn}}$ generated by the polynomials $g$ and $f_{ijkl}$. That is,

$$\mathcal{J}_{M_{mn}} = \langle \boldsymbol{\mathcal{G}}_{M_{mn}} \rangle \quad \text{with } \boldsymbol{\mathcal{G}}_{M_{mn}} = \{g(\mathbf{x})\} \cup \{f_{ijkl}(\mathbf{x}) : 1 \leq i < k \leq m,\ 1 \leq j < l \leq n\}. \tag{4.5}$$

Recall that the convex hull of the set $\boldsymbol{\mathcal{R}}$ in (4.4) forms the unit-nuclear-norm ball and by the definition of the theta bodies,

$$\overline{\operatorname{conv}(\nu_{\mathbb{R}}(\mathcal{J}_{M_{mn}}))} \subseteq \cdots \subseteq \mathrm{TH}_{k+1}(\mathcal{J}_{M_{mn}}) \subseteq \mathrm{TH}_k(\mathcal{J}_{M_{mn}}) \subseteq \cdots \subseteq \mathrm{TH}_1(\mathcal{J}_{M_{mn}}).$$

Therefore, the theta bodies form closed, convex hierarchical relaxations of the unit-matrix-nuclear-norm ball. In addition, the theta body $\mathrm{TH}_k(\mathcal{J}_{M_{mn}})$ is symmetric, $\mathrm{TH}_k(\mathcal{J}_{M_{mn}}) = -\mathrm{TH}_k(\mathcal{J}_{M_{mn}})$. Thus, it defines a unit ball of a norm that we call the *$\theta_k$-norm.*

The next result shows that the generating set of the ideal $\mathcal{J}_{M_{mn}}$ introduced above is a Gröbner basis for $\mathcal{J}_{M_{mn}}$.

**Theorem 4.8.** The set $\boldsymbol{\mathcal{G}}_{M_{mn}}$ forms the reduced Gröbner basis of the ideal $\mathcal{J}_{M_{mn}}$ with respect to the grevlex order.

Proof. The set $\boldsymbol{\mathcal{G}}_{M_{mn}}$ is clearly a basis for the ideal $\mathcal{J}_{M_{mn}}$. By Proposition B.12, we only need to check whether the $S$-polynomial (see Definition B.7) satisfies $S(p,q) \rightarrow_{\boldsymbol{\mathcal{G}}_{M_{mn}}} 0$ for all $p, q \in \boldsymbol{\mathcal{G}}_{M_{mn}}$ whenever the leading monomials $\mathrm{LM}(p)$ and $\mathrm{LM}(q)$ are not relatively prime. Here, $S(p,q) \rightarrow_{\boldsymbol{\mathcal{G}}_{M_{mn}}} 0$ means that $S(p,q)$ reduces to 0 modulo $\boldsymbol{\mathcal{G}}_{M_{mn}}$, see Definition B.9.

Notice that $\mathrm{LM}(g) = x_{11}^2$ and $\mathrm{LM}(f_{ijkl}) = x_{il}x_{kj}$ are relatively prime, for all $1 \leq i < k \leq m$ and $1 \leq j < l \leq n$. Therefore, we only need to show that $S(f_{ijkl}, f_{\hat{i}\hat{j}\hat{k}\hat{l}}) \rightarrow_{\boldsymbol{\mathcal{G}}_{M_{mn}}} 0$ whenever the leading monomials $\mathrm{LM}(f_{ijkl})$ and $\mathrm{LM}(f_{\hat{i}\hat{j}\hat{k}\hat{l}})$ are not relatively prime. First we consider

$$f_{ijkl}(\mathbf{x}) = x_{il}x_{kj} - x_{ij}x_{kl} \quad \text{and} \quad f_{i\hat{j}\hat{k}l}(\mathbf{x}) = x_{il}x_{\hat{k}\hat{j}} - x_{i\hat{j}}x_{\hat{k}l}$$

for $1 \leq i < k < \hat{k} \leq m$, $1 \leq j < \hat{j} < l \leq n$. The $S$-polynomial is then of the form

$$\begin{aligned} S(f_{ijkl}, f_{i\hat{j}\hat{k}l}) = x_{\hat{k}\hat{j}} f_{ijkl}(\mathbf{x}) - x_{kj} f_{i\hat{j}\hat{k}l}(\mathbf{x}) &= -x_{ij}x_{kl}x_{\hat{k}\hat{j}} + x_{i\hat{j}}x_{\hat{k}l}x_{kj} \\ &= x_{\hat{k}l} f_{ijk\hat{j}}(\mathbf{x}) - x_{ij} f_{k\hat{j}\hat{k}l}(\mathbf{x}) \in \mathcal{J}_{M_{mn}} \end{aligned}$$

so that $S(f_{ijkl}, f_{i\hat{j}\hat{k}l}) \to_{\boldsymbol{\mathcal{G}}_{M_{mn}}} 0$. The remaining cases are treated with similar arguments.

In order to show that $\boldsymbol{\mathcal{G}}_{M_{mn}}$ is the reduced Gröbner basis (see Definition B.6) for the ideal $\mathcal{J}_{M_{mn}}$, we first notice that $\mathrm{LC}(f) = 1$ for all $f \in \boldsymbol{\mathcal{G}}_{M_{mn}}$. In addition, the leading monomial of $f \in \boldsymbol{\mathcal{G}}_{M_{mn}}$ is always of degree two and there are no two different polynomials $f_i, f_j \in \boldsymbol{\mathcal{G}}_{M_{mn}}$ such that $\mathrm{LM}(f_i) = \mathrm{LM}(f_j)$. Therefore, $\boldsymbol{\mathcal{G}}_{M_{mn}}$ is the reduced Gröbner basis of the ideal $\mathcal{J}_{M_{mn}}$ with respect to the grevlex order. □

The Gröbner basis $\boldsymbol{\mathcal{G}}_{M_{mn}}$ of $\mathcal{J}_{M_{mn}} = \langle \boldsymbol{\mathcal{G}}_{M_{mn}} \rangle$ yields the $\theta$-basis of $\mathbb{R}[\mathbf{x}]/\mathcal{J}_{M_{mn}}$. For the sake of simplicity, we only provide its elements up to degree two,

$$\begin{aligned} \boldsymbol{\mathcal{B}}_1 &= \{1 + \mathcal{J}_{M_{mn}}, x_{11} + \mathcal{J}_{M_{mn}}, x_{12} + \mathcal{J}_{M_{mn}}, \ldots, x_{mn} + \mathcal{J}_{M_{mn}}\} \\ \boldsymbol{\mathcal{B}}_2 &= \boldsymbol{\mathcal{B}}_1 \cup \{x_{ij}x_{kl} + \mathcal{J}_{M_{mn}} : (i,j,k,l) \in \boldsymbol{\mathcal{S}}_{\boldsymbol{\mathcal{B}}_2}\}, \end{aligned}$$

where $\boldsymbol{\mathcal{S}}_{\boldsymbol{\mathcal{B}}_2} = \{(i,j,k,l) : 1 \leq i \leq k \leq m, 1 \leq j \leq l \leq n\} \setminus (1,1,1,1)$. Given the $\theta$-basis, the theta body $\mathrm{TH}_k(\mathcal{J}_{M_{mn}})$ is well-defined. We formally introduce the associated norm next.

**Definition 4.9.** The matrix $\theta_k$*-norm*, denoted by $\|\cdot\|_{\theta_k}$, is the norm induced by the $k$-theta body $\mathrm{TH}_k(\mathcal{J}_{M_{mn}})$, i.e.,

$$\|\mathbf{X}\|_{\theta_k} = \inf\{r : \mathbf{X} \in r\,\mathrm{TH}_k(\mathcal{J}_{M_{mn}})\}.$$

The $\theta_k$-norm can be computed with the help of Theorem 4.5, i.e., as

$$\|\mathbf{X}\|_{\theta_k} = \min_t t \quad \text{subject to} \quad \mathbf{X} \in t\boldsymbol{\mathcal{Q}}_{\boldsymbol{\mathcal{B}}_k}(\mathcal{J}_{M_{mn}}).$$

Given the moment matrix $\mathbf{M}_{\boldsymbol{\mathcal{B}}_k}(\mathbf{y})$ associated with $\mathcal{J}_{M_{mn}}$, this minimization program is equivalent to the semidefinite program

$$\min_{t \in \mathbb{R}, \mathbf{y} \in \mathbb{R}^{\boldsymbol{\mathcal{B}}_{2k}}} t \quad \text{subject to} \quad \mathbf{M}_{\boldsymbol{\mathcal{B}}_k}(\mathbf{y}) \succcurlyeq 0, y_0 = t, \mathbf{y}_{\boldsymbol{\mathcal{B}}_1} = \mathbf{X}. \tag{4.6}$$

The last constraint might require some explanation. The vector $\mathbf{y}_{\boldsymbol{\mathcal{B}}_1}$ denotes the restriction of $\mathbf{y}$ to the indices in $\boldsymbol{\mathcal{B}}_1$, where the latter can be identified with the set $[m] \times [n]$ indexing the matrix entries. Therefore, $\mathbf{y}_{\boldsymbol{\mathcal{B}}_1} = \mathbf{X}$ means componentwise $y_{x_{11}+\mathcal{J}_{M_{mn}}} = X_{11}$, $y_{x_{12}+\mathcal{J}_{M_{mn}}} = X_{12}$, $\ldots, y_{x_{mn}+\mathcal{J}_{M_{mn}}} = X_{mn}$. For the purpose of illustration, we focus on the $\theta_1$-norm in $\mathbb{R}^{2\times 2}$ in Subsection 4.2.1 below, and provide a step-by-step procedure for building the corresponding semidefinite program in (4.6).

Notice that the number of elements in $\boldsymbol{\mathcal{B}}_1$ is $mn + 1$, and in $\boldsymbol{\mathcal{B}}_2 \backslash \boldsymbol{\mathcal{B}}_1$ is $\frac{m\cdot(m+1)}{2} \cdot \frac{n\cdot(n+1)}{2} - 1 \sim \frac{(mn)^2}{2}$. That is, the number of elements of the $\theta$-basis restricted to the degree two scales polynomially in the total number of matrix entries $mn$. Therefore, the computational complexity of the SDP in (4.6) is polynomial in $mn$.

We show next that the theta body $\mathrm{TH}_1(\mathcal{J}_{M_{mn}})$ and hence, all $\mathrm{TH}_k(\mathcal{J}_{M_{mn}})$ for $k \in \mathbb{N}$, coincide with the matrix-unit-nuclear-norm ball. To this end, the following lemma provides expressions for the boundary of the matrix-unit-nuclear-norm ball.

**Lemma 4.10.** Let $\boldsymbol{\mathcal{O}}_c$ ($\boldsymbol{\mathcal{O}}_r$) denote the set of all matrices $\mathbf{M} \in \mathbb{R}^{n \times m}$ with orthonormal columns (rows), i.e., $\boldsymbol{\mathcal{O}}_c = \left\{\mathbf{M} \in \mathbb{R}^{n \times m} : \mathbf{M}^T\mathbf{M} = \mathbf{I}_m\right\}$ and $\boldsymbol{\mathcal{O}}_r = \left\{\mathbf{M} \in \mathbb{R}^{n \times m} : \mathbf{M}\mathbf{M}^T = \mathbf{I}_n\right\}$. Then

$$\left\{\mathbf{X} \in \mathbb{R}^{m \times n} : \|\mathbf{X}\|_* \leq 1\right\} = \left\{\mathbf{X} \in \mathbb{R}^{m \times n} : \operatorname{tr}(\mathbf{M}\mathbf{X}) \leq 1, \text{ for all } \mathbf{M} \in \boldsymbol{\mathcal{O}}_c \cup \boldsymbol{\mathcal{O}}_r\right\}.$$

**Remark 4.11.** Notice that $\boldsymbol{\mathcal{O}}_c = \emptyset$ for $m > n$ and $\boldsymbol{\mathcal{O}}_r = \emptyset$ for $m < n$.

Proof. It suffices to treat the case $m \leq n$ because $\|\mathbf{X}\|_* = \left\|\mathbf{X}^T\right\|_*$ for all matrices $\mathbf{X}$, and $\mathbf{M} \in \boldsymbol{\mathcal{O}}_r$ if and only if $\mathbf{M}^T \in \boldsymbol{\mathcal{O}}_c$. Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ such that $\|\mathbf{X}\|_* \leq 1$ and let $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ be its singular value decomposition. For $\mathbf{M} \in \boldsymbol{\mathcal{O}}_c$, the spectral norm satisfies $\|\mathbf{M}\|_{2 \to 2} \leq 1$ and therefore, using the fact that the nuclear norm is the dual of the spectral norm, see e.g. [8, p. 96] and Example C.4,

$$\operatorname{tr}(\mathbf{M}\mathbf{X}) \leq \|\mathbf{M}\|_{2 \to 2} \cdot \|\mathbf{X}\|_* \leq \|\mathbf{X}\|_* \leq 1.$$

For the converse, let $\mathbf{X} \in \mathbb{R}^{m \times n}$ be such that $\operatorname{tr}(\mathbf{M}\mathbf{X}) \leq 1$, for all $\mathbf{M} \in \boldsymbol{\mathcal{O}}_c$. Let $\mathbf{X} = \mathbf{U}\overline{\boldsymbol{\Sigma}}\,\overline{\mathbf{V}}^T$ denote its reduced singular value decomposition, i.e., $\mathbf{U}, \overline{\boldsymbol{\Sigma}} \in \mathbb{R}^{m \times m}$ and $\overline{\mathbf{V}} \in \mathbb{R}^{n \times m}$ with $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \overline{\mathbf{V}}^T\overline{\mathbf{V}} = \mathbf{I}_m$. Since $\mathbf{M} := \overline{\mathbf{V}}\mathbf{U}^T \in \boldsymbol{\mathcal{O}}_c$, it follows that

$$1 \geq \operatorname{tr}(\mathbf{M}\mathbf{X}) = \operatorname{tr}(\overline{\mathbf{V}}\mathbf{U}^T\mathbf{U}\boldsymbol{\Sigma}\overline{\mathbf{V}}^T) = \operatorname{tr}(\boldsymbol{\Sigma}) = \|\mathbf{X}\|_*.$$

This completes the proof. □

Next, using Lemma 4.10, we show that the theta body $\mathrm{TH}_1(\mathcal{J}_{M_{mn}})$ equals the matrix-unit-nuclear-norm ball.

**Theorem 4.12.** The polynomial ideal $\mathcal{J}_{M_{mn}}$ defined in (4.5) satisfies

$$\mathrm{TH}_1\left(\mathcal{J}_{M_{mn}}\right) = \overline{\operatorname{conv}\left(\{\mathbf{x} : g(\mathbf{x}) = 0, f_{ijkl}(\mathbf{x}) = 0 \text{ for all } i < k, j < l\}\right)}.$$

In other words,

$$\left\{\mathbf{X} \in \mathbb{R}^{m \times n} : \mathbf{X} \in \mathrm{TH}_1\left(\mathcal{J}_{M_{mn}}\right)\right\} = \left\{\mathbf{X} \in \mathbb{R}^{m \times n} : \|\mathbf{X}\|_* \leq 1\right\}.$$

Proof. By the definition of $\mathrm{TH}_1(\mathcal{J}_{M_{mn}})$, it is enough to show that the boundary of the unit-nuclear-norm ball can be written as 1-sos mod $\mathcal{J}_{M_{mn}}$. By Lemma 4.10 this means that the polynomial $1 - \sum_{i=1}^m \sum_{j=1}^n x_{ij}M_{ji}$ is 1-sos mod $\mathcal{J}_{M_{mn}}$, for all $\mathbf{M} \in \boldsymbol{\mathcal{O}}_c \cup \boldsymbol{\mathcal{O}}_r$. We start by fixing $\mathbf{M} = \begin{pmatrix} \mathbf{I}_m \\ \mathbf{0} \end{pmatrix}$ in case $m \leq n$ and $\mathbf{M} = \begin{pmatrix} \mathbf{I}_n & \mathbf{0} \end{pmatrix}$ in case $m > n$, where $\mathbf{I}_k \in \mathbb{R}^{k \times k}$ is the identity matrix. For this choice of $\mathbf{M}$, we need to show that $1 - \sum_{i=1}^{\ell} x_{ii}$ is 1-sos mod $\mathcal{J}_{M_{mn}}$, where $\ell = \min\{m, n\}$. Note that

$$\begin{aligned} 1 - \sum_{i=1}^{\ell} x_{ii} =& \frac{1}{2}\left[\left(1 - \sum_{i=1}^{\ell} x_{ii}\right)^2 + \left(1 - \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}^2\right) + \sum_{i<j\leq\ell}\left(x_{ij} - x_{ji}\right)^2\right. \\ &\left. -2\sum_{i<j\leq\ell}\left(x_{ii}x_{jj} - x_{ij}x_{ji}\right) + \sum_{i=1}^{m}\sum_{j=m+1}^{n} x_{ij}^2 + \sum_{i=n+1}^{m}\sum_{j=1}^{n} x_{ij}^2\right], \end{aligned}$$

since

$$\left(1 - \sum_{i=1}^{\ell} x_{ii}\right)^2 =1 - 2\sum_{i=1}^{\ell} x_{ii} + \sum_{i=1}^{\ell}\sum_{j=1}^{\ell} x_{ii}x_{jj} = 1 - 2\sum_{i=1}^{\ell} x_{ii} + 2\sum_{i<j\leq\ell} x_{ii}x_{jj} + \sum_{i=1}^{\ell} x_{ii}^2,$$

$$1-\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}^2+\sum_{i=1}^{m}\sum_{j=m+1}^{n}x_{ij}^2+\sum_{i=n+1}^{m}\sum_{j=1}^{n}x_{ij}^2=1-\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}x_{ij}^2=1-\sum_{i<j\leq\ell}\left(x_{ij}^2+x_{ji}^2\right)-\sum_{i=1}^{\ell}x_{ii}^2,$$

and

$$\begin{aligned}\sum_{i<j\leq\ell}\left(x_{ij}-x_{ji}\right)^2-2\sum_{i<j\leq\ell}\left(x_{ii}x_{jj}-x_{ij}x_{ji}\right)&=\sum_{i<j\leq\ell}\left(x_{ij}^2+x_{ji}^2-2x_{ij}x_{ji}-2x_{ii}x_{jj}+2x_{ij}x_{ji}\right)\\&=\sum_{i<j\leq\ell}\left(x_{ij}^2+x_{ji}^2\right)-2\sum_{i<j\leq\ell}x_{ii}x_{jj}.\end{aligned}$$

Therefore, $1-\sum_{i=1}^{\ell}x_{ii}$ is 1-sos mod $\mathcal{J}_{M_{mn}}$, since the polynomials $1-\sum_{i=1}^{\ell}x_{ii}$, $x_{ij}-x_{ji}$, $x_{ij}$, and $x_{ji}$ are linear and the polynomials $1-\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}^2$ and $2\left(x_{ii}x_{jj}-x_{ij}x_{ji}\right)$ are contained in the ideal, for all $i<j\leq\ell$.

Next, we define the transformed variables

$$x'_{ij}=\begin{cases}\sum_{k=1}^{m}M_{ik}x_{kj} & \text{if } m\leq n,\\ \sum_{k=1}^{n}x_{ik}M_{kj} & \text{if } m>n.\end{cases}$$

Since $x'_{ij}$ is a linear combination of $\{x_{kj}\}_{k=1}^{m}\cup\{x_{ik}\}_{k=1}^{n}$ for every $i\in[m]$ and $j\in[n]$, the linearity of the polynomials $1-\sum_{i=1}^{\ell}x'_{ii}$, $x'_{ij}-x'_{ji}$, $x'_{ij}$, and $x'_{ji}$ is preserved for all $i<j$. It remains to show that the ideal is invariant under this transformation. For the polynomial $1-\sum_{i=1}^{m}\sum_{j=1}^{n}{x'_{ij}}^2$ this is clear since $\mathbf{M}\in\mathbb{R}^{n\times m}$ has unitary columns if $m\leq n$ and unitary rows if $m\geq n$. Moreover, if $m\leq n$ the polynomial $x'_{ii}x'_{jj}-x'_{ij}x'_{ji}$ is contained in the ideal $\mathcal{J}$ since

$$x'_{ii}x'_{jj}-x'_{ij}x'_{ji}=\sum_{k=1}^{m}\sum_{l=1}^{m}M_{ik}M_{jl}\left(x_{ki}x_{lj}-x_{kj}x_{li}\right)$$

and the polynomials $x_{ki}x_{lj}-x_{kj}x_{li}$ are contained in $\mathcal{J}_{M_{mn}}$ for all $i<j\leq m$. Similarly, if $m\geq n$ the polynomial $x'_{ii}x'_{jj}-x'_{ij}x'_{ji}$ is in the ideal since

$$x'_{ii}x'_{jj}-x'_{ij}x'_{ji}=\sum_{k=1}^{n}\sum_{l=1}^{n}M_{ki}M_{lj}\left(x_{ik}x_{jl}-x_{il}x_{jk}\right)$$

and polynomials $x_{ik}x_{jl}-x_{il}x_{jk}$ are in the ideal, for all $i<j\leq n$. □

The following corollary is a direct consequence of Theorem 4.12 and the nestedness property (4.1) of the theta bodies.

**Corollary 4.13.** The matrix $\theta_1$-norm coincides with the matrix nuclear norm, i.e.,

$$\left\|\mathbf{X}\right\|_*=\left\|\mathbf{X}\right\|_{\theta_1},\qquad\text{for all }\mathbf{X}\in\mathbb{R}^{m\times n}.$$

In other words, ideal $\mathcal{J}_{M_{mn}}$ is TH$_1$-exact, i.e.,

$$\mathrm{TH}_1\left(\mathcal{J}_{M_{mn}}\right)=\mathrm{TH}_2\left(\mathcal{J}_{M_{mn}}\right)=\cdots=\overline{\mathrm{conv}\left(\nu_{\mathbb{R}}\left(\mathcal{J}_{M_{mn}}\right)\right)}.$$

**Remark 4.14.** The ideal (4.5) is not the only choice that satisfies (4.2). For example, in [32] the following polynomial ideal was suggested

$$\mathcal{J}=\left\langle\left\{x_{ij}-u_iv_j\right\}_{i\in[m],j\in[n]},\sum_{i=1}^{m}u_i^2-1,\sum_{j=1}^{n}v_j^2-1\right\rangle$$

in $\mathbb{R}[\mathbf{x},\mathbf{u},\mathbf{v}] = \mathbb{R}[x_{11},\ldots,x_{mn},u_1,\ldots,u_m,v_1,\ldots,v_n]$. Some tedious computations reveal the reduced Gröbner basis $\boldsymbol{\mathcal{G}}$ of the ideal $\mathcal{J}$ with respect to the grevlex (and grlex) ordering,

$$\begin{aligned}
\boldsymbol{\mathcal{G}} = &\left\{ g_1^{i,j} = x_{ij} - u_i v_j : i \in [m]\,,\, j \in [n] \right\} \bigcup \left\{ g_2 = \sum_{i=1}^m u_i^2 - 1, g_3 = \sum_{j=1}^n v_j^2 - 1 \right\} \\
&\bigcup \left\{ g_4^{i,j,k} = x_{ij}u_k - x_{kj}u_i : 1 \le i < k \le m,\, j \in [n] \right\} \bigcup \left\{ g_5^j = \sum_{i=1}^m x_{ij}u_i - v_j : j \in [n] \right\} \\
&\bigcup \left\{ g_6^{i,j,k} = x_{ij}v_k - x_{ik}v_j : i \in [m]\,,\, 1 \le j < k \le n \right\} \bigcup \left\{ g_7^i = \sum_{j=1}^n x_{ij}v_j - u_i : i \in [m] \right\} \\
&\bigcup \left\{ g_8^{i,j} = \sum_{k=1}^n x_{ik}x_{jk} - u_i u_j : 1 \le i < j \le m \right\} \bigcup \left\{ g_9^{i,j} = \sum_{k=1}^m x_{ki}x_{kj} - v_i v_j : 1 \le i < j \le n \right\} \\
&\bigcup \left\{ g_{10}^i = \sum_{j=1}^n x_{ij}^2 - u_i^2 : 2 \le i \le m \right\} \bigcup \left\{ g_{11}^j = \sum_{i=1}^m x_{ij}^2 - v_j^2 : 2 \le j \le n \right\} \\
&\bigcup \left\{ g_{12}^{i,j,k,l} = x_{ij}x_{kl} - x_{il}x_{kj} : 1 \le i < k \le m,\, 1 \le j < l \le n \right\} \\
&\bigcup \left\{ g_{13} = x_{11}^2 - \sum_{i=2}^m \sum_{j=2}^n x_{ij}^2 + \sum_{i=2}^m u_i^2 + \sum_{j=2}^n v_j^2 - 1 \right\}.
\end{aligned}$$

Obviously, this Gröbner basis is much more complicated than the one of the ideal $\mathcal{J}_{M_{mn}}$ introduced above. Therefore, computations (both theoretical and numerical) with this alternative ideal seem to be more demanding. In any case, the variables $\{u_i\}_{i=1}^m$ and $\{v_j\}_{j=1}^n$ are only auxiliary, so one would like to eliminate these from the above Gröbner basis (see for example [37] for the elimination procedure). By doing so, one obtains the Gröbner basis $\boldsymbol{\mathcal{G}}_{M_{mn}}$ defined in (4.5). Notice that $\sum_{i=1}^m \sum_{j=1}^n x_{ij}^2 - 1 = g_{13} + \sum_{i=2}^m g_{10}^i + \sum_{j=2}^n g_{11}^j$ together with $\{g_{12}^{i,j,k,l}\}$ form the basis $\boldsymbol{\mathcal{G}}_{M_{mn}}$.

**Remark 4.15.** The polynomial ideal $\overline{\mathcal{J}}_{M_{mn}} = \langle \{f_{ijkl} : i < k, j < l\} \rangle$ is a special case of *determinantal ideals* already studied in real algebraic geometry and commutative algebra, see [16, 147]. Let $\mathbb{K}$ be a field and $\mathbf{X}$ be a matrix of indeterminates over $\mathbb{K}$. For $t \in \mathbb{N}$, $t \le \min\{m,n\}$, the determinantal ideal $\mathcal{I}_t$ is generated by all $t$-minors (minors of order $t$ or $t$-th order minors) of the matrix $\mathbf{X}$. Consequently, the real algebraic variety of $\mathcal{I}_t$, called *determinantal variety*, is the set of all rank-$(t-1)$ matrices. In particular, the ideal $\overline{\mathcal{J}}_{M_{mn}}$ coincides with the ideal $\mathcal{I}_2$. Sturmfels in [147] proved that the set of all $t$-minors of $\mathbf{X}$ is the reduced Gröbner basis of $\mathcal{I}_t$ with respect to the lexicographic order induced from the variable order $x_{1n_2} > x_{1,n_2-1} > \ldots > x_{11} > x_{2n_2} > x_{2,n_2-1} > \ldots > x_{21} > \ldots > x_{n_1n_2} > \ldots x_{n_11}$. The author used the Knuth-Robinson-Schensted correspondence KRS – the technique introduced in [93] – to avoid applying Buchberger's algorithm. Afterwards, several authors independently proved that $t$-minors of $\mathbf{X}$ form a Gröbner basis of $\mathcal{I}_t$ regardless of the choice of the monomial order, see [29, 111, 117]. Still, to the best of our knowledge – the presented results regarding the ideal $\mathcal{J}_{M_{mn}}$ are new.

**4.2.1. The $\theta_1$-norm in $\mathbb{R}^{2\times 2}$.** For the sake of illustration, we consider the specific example of $2\times 2$ matrices and provide the corresponding semidefinite program for the computation of the $\theta_1$-norm explicitly. Let us denote the corresponding polynomial ideal in $\mathbb{R}[\mathbf{x}] = \mathbb{R}[x_{11},x_{12},x_{21},x_{22}]$

| $1$ | $x_{11}$ | $x_{12}$ | $x_{21}$ | $x_{22}$ | $x_{11}x_{12}$ | $x_{11}x_{21}$ | $x_{11}x_{22}$ | $x_{12}^2$ | $x_{12}x_{22}$ | $x_{21}^2$ | $x_{21}x_{22}$ | $x_{22}^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y_0$ | $x_{11}$ | $x_{12}$ | $x_{21}$ | $x_{22}$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | $y_7$ | $y_8$ |

TABLE 4.1. Linearization of the elements of $\boldsymbol{\mathcal{B}}_2$ for matrix $2 \times 2$ case. The polynomial $f$ in the first row refers to the element $f + \mathcal{J} \in \boldsymbol{\mathcal{B}}_2$.

simply by

$$\mathcal{J} = \mathcal{J}_{M_{22}} = \left\langle x_{12}x_{21} - x_{11}x_{22},\, x_{11}^2 + x_{12}^2 + x_{21}^2 + x_{22}^2 - 1 \right\rangle .$$

The associated real algebraic variety is of the form

$$\nu_{\mathbb{R}}\left(\mathcal{J}\right) = \left\{ \mathbf{x} : x_{12}x_{21} = x_{11}x_{22},\, x_{11}^2 + x_{12}^2 + x_{21}^2 + x_{22}^2 = 1 \right\}$$

and corresponds to the set of rank-one Frobenius-norm-one matrices in $\mathbb{R}^{2\times 2}$. Its convex hull consists of matrices $\mathbf{X} \in \mathbb{R}^{2\times 2}$ with $\|\mathbf{X}\|_* \leq 1$. According to Theorem 4.8, the reduced Gröbner basis $\boldsymbol{\mathcal{G}}$ of $\mathcal{J}$ with respect to the grevlex order is

$$\boldsymbol{\mathcal{G}} = \left\{ g_1 = x_{12}x_{21} - x_{11}x_{22},\, g_2 = x_{11}^2 + x_{12}^2 + x_{21}^2 + x_{22}^2 - 1 \right\}$$

with the corresponding $\theta$-basis $\boldsymbol{\mathcal{B}}$ of $\mathbb{R}\left[\mathbf{x}\right]/\mathcal{J}$ restricted to the degree two given as

$$\begin{aligned}
\boldsymbol{\mathcal{B}}_1 &= \left\{1 + \mathcal{J}, x_{11} + \mathcal{J}, x_{12} + \mathcal{J}, x_{21} + \mathcal{J}, x_{22} + \mathcal{J}\right\} \\
\boldsymbol{\mathcal{B}}_2 &= \boldsymbol{\mathcal{B}}_1 \cup \left\{x_{11}x_{12} + \mathcal{J}, x_{11}x_{21} + \mathcal{J}, x_{11}x_{22} + \mathcal{J}, x_{12}^2 + \mathcal{J}, x_{12}x_{22} + \mathcal{J},\right. \\
&\qquad\qquad\qquad\qquad \left. x_{21}^2 + \mathcal{J}, x_{21}x_{22} + \mathcal{J}, x_{22}^2 + \mathcal{J}\right\} .
\end{aligned}$$

The set $\boldsymbol{\mathcal{B}}_2$ consists of all monomials of degree at most two which are not divisible by a leading term of any of the polynomials inside the Gröbner basis $\boldsymbol{\mathcal{G}}$. For example, $x_{11}x_{12} + \mathcal{J}$ is an element of the theta basis $\boldsymbol{\mathcal{B}}$, but $x_{11}^2 + \mathcal{J}$ is not since $x_{11}^2$ is divisible by $\mathrm{LT}(g_2)$.

Linearizing the elements of $\boldsymbol{\mathcal{B}}_2$ results in Table 4.1, where the monomials $f$ in the first row stand for an element $f + \mathcal{J} \in \boldsymbol{\mathcal{B}}_2$. Therefore, $[\mathbf{x}]_{\boldsymbol{\mathcal{B}}_1} = \left(1 + \mathcal{J}, x_{11} + \mathcal{J}, x_{12} + \mathcal{J}, x_{21} + \mathcal{J}, x_{22} + \mathcal{J}\right)^T$ and the following combinatorial moment matrix $\mathbf{M}_{\boldsymbol{\mathcal{B}}_1}\left(\mathbf{x}, \mathbf{y}\right)$, see Definition 4.4, is given as

$$\mathbf{M}_{\boldsymbol{\mathcal{B}}_1}\left(\mathbf{x}, \mathbf{y}\right) = \begin{bmatrix} y_0 & x_{11} & x_{12} & x_{21} & x_{22} \\ x_{11} & -y_4 - y_6 - y_8 + y_0 & y_1 & y_2 & y_3 \\ x_{12} & y_1 & y_4 & y_3 & y_5 \\ x_{21} & y_2 & y_3 & y_6 & y_7 \\ x_{22} & y_3 & y_5 & y_7 & y_8 \end{bmatrix} .$$

For instance, the entry $(2, 2)$ of $[\mathbf{x}]_{\boldsymbol{\mathcal{B}}_1} [\mathbf{x}]_{\boldsymbol{\mathcal{B}}_1}^T$ is of the form $x_{11}^2 + \mathcal{J} = -x_{12}^2 - x_{21}^2 - x_{22}^2 + 1 + \mathcal{J}$, where we exploit the second property in Definition 4.3 and the fact that $g_2 \in \mathcal{J}$. Replacing $x_{12}^2 + \mathcal{J}$ by $y_4$, etc. as in Table 4.1, yields the stated expression for $\left[\mathbf{M}_{\boldsymbol{\mathcal{B}}_1}\left(\mathbf{x}, \mathbf{y}\right)\right]_{2,2}$.

By Theorem 4.5, the first theta body $\mathrm{TH}_1\left(\mathcal{J}\right)$ is the closure of

$$\boldsymbol{\mathcal{Q}}_{\boldsymbol{\mathcal{B}}_1}\left(\mathcal{J}\right) = \pi_{\mathbf{x}} \left\{ \left(\mathbf{x}, \mathbf{y}\right) \in \mathbb{R}^{\boldsymbol{\mathcal{B}}_2} : \mathbf{M}_{\boldsymbol{\mathcal{B}}_1}\left(\mathbf{x}, \mathbf{y}\right) \succeq 0,\, y_0 = 1 \right\} ,$$

where $\pi_{\mathbf{x}}$ represents the projection onto the variables $\mathbf{x}$, i.e., the projection onto $x_{11}$, $x_{12}$, $x_{21}$, $x_{22}$. Furthermore, the $\theta_1$-norm of a matrix $\mathbf{X} \in \mathbb{R}^{2\times 2}$ induced by the $\mathrm{TH}_1\left(\mathcal{J}\right)$ and denoted by $\|\cdot\|_{\theta_1}$ can be computed as

$$\|\mathbf{X}\|_{\theta_1} = \inf_t t \quad \text{s.t.} \quad \mathbf{X} \in t\boldsymbol{\mathcal{Q}}_{\boldsymbol{\mathcal{B}}_1}\left(\mathcal{J}\right)$$

which is equivalent to

$$\inf_{t\in\mathbb{R},\mathbf{y}\in\mathbb{R}^8} t \quad \text{s.t.} \quad \mathbf{M} = \begin{bmatrix} t & X_{11} & X_{12} & X_{21} & X_{22} \\ X_{11} & -y_4 - y_6 - y_8 + t & y_1 & y_2 & y_3 \\ X_{12} & y_1 & y_4 & y_3 & y_5 \\ X_{21} & y_2 & y_3 & y_6 & y_7 \\ X_{22} & y_3 & y_5 & y_7 & y_8 \end{bmatrix} \succeq \mathbf{0}. \tag{4.7}$$

Notice that $\mathrm{tr}(\mathbf{M}) = 2t$. By Theorem 4.12, the above program is equivalent to the standard semidefinite program for computing the nuclear norm of a given matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$

$$\min_{\mathbf{W},\mathbf{Z}} \frac{1}{2}\left(\mathrm{tr}\left(\mathbf{W}\right) + \mathrm{tr}\left(\mathbf{Z}\right)\right) \quad \text{s.t.} \quad \begin{bmatrix} W_{11} & W_{12} & X_{11} & X_{12} \\ W_{12} & W_{22} & X_{21} & X_{22} \\ X_{11} & X_{21} & Z_{11} & Z_{12} \\ X_{22} & X_{22} & Z_{12} & Z_{22} \end{bmatrix} \succeq \mathbf{0}.$$

Notice that the matrix $\mathbf{M}$ in (4.7) can be written as the following sum

$$\mathbf{M} = t\cdot\mathbf{M}_0 + \sum_{i=1}^{2}\sum_{j=1}^{2} X_{ij}\cdot\mathbf{M}_{ij} + \sum_{k=1}^{8} y_k\cdot\mathbf{M}_k,$$

where

$$\mathbf{M}_0 = \begin{bmatrix} 1&0&0&0&0\\0&1&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0 \end{bmatrix}, \quad \mathbf{M}_1 = \begin{bmatrix} 0&0&0&0&0\\0&0&1&0&0\\0&1&0&0&0\\0&0&0&0&0\\0&0&0&0&0 \end{bmatrix}, \mathbf{M}_2 = \begin{bmatrix} 0&0&0&0&0\\0&0&0&1&0\\0&0&0&0&0\\0&1&0&0&0\\0&0&0&0&0 \end{bmatrix}, \quad \mathbf{M}_3 = \begin{bmatrix} 0&0&0&0&0\\0&0&0&0&1\\0&0&0&1&0\\0&0&1&0&0\\0&1&0&0&0 \end{bmatrix},$$

$$\mathbf{M}_4 = \begin{bmatrix} 0&0&0&0&0\\0&-1&0&0&0\\0&0&1&0&0\\0&0&0&0&0\\0&0&0&0&0 \end{bmatrix}, \quad \mathbf{M}_5 = \begin{bmatrix} 0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&1\\0&0&0&0&0\\0&0&1&0&0 \end{bmatrix}, \mathbf{M}_6 = \begin{bmatrix} 0&0&0&0&0\\0&-1&0&0&0\\0&0&0&0&0\\0&0&0&1&0\\0&0&0&0&0 \end{bmatrix}, \quad \mathbf{M}_7 = \begin{bmatrix} 0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&1\\0&0&0&1&0 \end{bmatrix},$$

$$\mathbf{M}_8 = \begin{bmatrix} 0&0&0&0&0\\0&-1&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&1 \end{bmatrix}, \mathbf{M}_{11} = \begin{bmatrix} 0&1&0&0&0\\1&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0 \end{bmatrix}, \mathbf{M}_{12} = \begin{bmatrix} 0&0&1&0&0\\0&0&0&0&0\\1&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0 \end{bmatrix}, \quad \mathbf{M}_{21} = \begin{bmatrix} 0&0&0&1&0\\0&0&0&0&0\\0&0&0&0&0\\1&0&0&0&0\\0&0&0&0&0 \end{bmatrix},$$

$$\mathbf{M}_{22} = \begin{bmatrix} 0&0&0&0&1\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\1&0&0&0&0 \end{bmatrix}.$$

## 4.3. The tensor $\theta_k$-norm

We now turn to the tensor case and study the hierarchical closed convex relaxations of the unit-tensor-nuclear-norm ball defined via theta bodies. Since in the matrix case all $\theta_k$-norms are equal to the matrix nuclear norm, their generalizations to the tensor case may all be viewed as natural generalizations of the tensor-nuclear norm. The focus is mostly on the $\theta_1$-norm whose unit norm ball is the largest in this hierarchical sequence of relaxations. Unlike in the matrix case, the $\theta_1$-norm defines a new tensor norm, that – to the best of our knowledge – has not been studied before.

In the matrix scenario, all the theta norms are equal to the unit matrix nuclear norm ball. In other words, the set of theta bodies converges to the unit matrix nuclear norm ball. By definition of the theta bodies, we have the following relation for unit $\theta_k$ norms in $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$

$$\begin{aligned}\{\mathbf{X} : \|\mathbf{X}\|_{\theta_1} \leq 1\} \supseteq \cdots \supseteq \{\mathbf{X} : \|\mathbf{X}\|_{\theta_k} \leq 1\}& \\ \supseteq \{\mathbf{X} : \|\mathbf{X}\|_{\theta_{k+1}} \leq 1\} \supseteq \cdots \supseteq \{\mathbf{X} : \|\mathbf{X}\|_{*} \leq 1\}&.\end{aligned}$$

The analysis of convergence of the theta bodies in the tensor scenario is left for future research. However, we do not expect the finite convergence to the unit tensor nuclear norm ball (otherwise computing the tensor nuclear norm would not be NP-hard).

The polynomial ideal will be generated by the order-two minors of tensor matricizations, where each variable corresponds to one tensor entry. As we will see, a tensor is rank-one if and only if all order-two minors of the unfoldings (matricizations) vanish. In the order three case one considers all three unfoldings. However, there are several possibilities for the order $d$ case when $d \geq 4$. In fact, a $d$th-order tensor is rank-one if all order-two minors of all unfoldings vanish so that it may be enough to consider only the unfoldings. However, one may as well consider the ideal generated by all order-two minors of *all* matricizations or one may consider a subset of matricizations including all unfoldings. Indeed, any tensor format – and thereby any notion of tensor rank – corresponds to a set of matricizations and in this way, one may associate a $\theta_k$-norm to a certain tensor format. In particular, for Tucker format one considers all order-two minors of all the unfoldings and for the TT-format all order-two minors of all the matricizations $\mathbf{X}^{\{1,2,\ldots,k\}}$ with $k \in [d-1]$. (See Chapter 2 for some background on various tensor formats.) However, as we will show later, the corresponding reduced Gröbner basis with respect to the grevlex order does not depend on the choice of the tensor format.

Below, we consider first the special case of third-order tensors. In Subsection 4.3.2 we treat the general $d$th order case.

**4.3.1. Third-order tensors.** As described above, we will consider the order-two minors of all the unfoldings of a third-order tensor. Our notation requires the following sets of subscripts

$$\begin{aligned}\boldsymbol{\mathcal{S}}_1 &= \left\{(\boldsymbol{\alpha},\boldsymbol{\beta}) : 1 \leq \alpha_1 < \beta_1 \leq n_1,\, 1 \leq \beta_2 < \alpha_2 \leq n_2,\, 1 \leq \beta_3 \leq \alpha_3 \leq n_3\right\},\\ \boldsymbol{\mathcal{S}}_2 &= \left\{(\boldsymbol{\alpha},\boldsymbol{\beta}) : 1 \leq \alpha_1 \leq \beta_1 \leq n_1,\, 1 \leq \beta_2 < \alpha_2 \leq n_2,\, 1 \leq \alpha_3 < \beta_3 \leq n_3\right\},\\ \boldsymbol{\mathcal{S}}_3 &= \left\{(\boldsymbol{\alpha},\boldsymbol{\beta}) : 1 \leq \alpha_1 < \beta_1 \leq n_1,\, 1 \leq \alpha_2 \leq \beta_2 \leq n_2,\, 1 \leq \beta_3 < \alpha_3 \leq n_3\right\},\\ \overline{\boldsymbol{\mathcal{S}}}_i &= \left\{(\boldsymbol{\alpha},\boldsymbol{\beta}) : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{S}}_i \text{ and } \alpha_j \neq \beta_j, \text{ for all } j \in [3]\right\}, \quad \text{for all } i \in [3]\,.\end{aligned}$$

The following polynomials $f^{(\boldsymbol{\alpha},\boldsymbol{\beta})}$ in $\mathbb{R}[\mathbf{x}] = \mathbb{R}[x_{111}, x_{112}, \ldots, x_{n_1 n_2 n_3}]$ correspond to a subset of all order-two minors of all tensor unfoldings,

$$\begin{aligned}f^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) &= x_{\boldsymbol{\alpha}} x_{\boldsymbol{\beta}} - x_{\boldsymbol{\alpha}\vee\boldsymbol{\beta}} x_{\boldsymbol{\alpha}\wedge\boldsymbol{\beta}}, \quad (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{S}} := \boldsymbol{\mathcal{S}}_1 \cup \boldsymbol{\mathcal{S}}_2 \cup \boldsymbol{\mathcal{S}}_3\\ g_3(\mathbf{x}) &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} x_{ijk}^2 - 1,\end{aligned}$$

where $[\boldsymbol{\alpha} \vee \boldsymbol{\beta}]_i = \max\{\alpha_i, \beta_i\}$ and $[\boldsymbol{\alpha} \wedge \boldsymbol{\beta}]_i = \min\{\alpha_i, \beta_i\}$. In particular, the following order-two minor of $\mathbf{X}^{\{1\}}$ is not contained in $\left\{f^{(\boldsymbol{\alpha},\boldsymbol{\beta})} : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{S}}\right\}$

$$f = x_{\boldsymbol{\alpha}} x_{\boldsymbol{\beta}} - x_{\hat{\boldsymbol{\alpha}}} x_{\hat{\boldsymbol{\beta}}}, \quad \text{where } \hat{\boldsymbol{\alpha}} = (\alpha_1, \beta_2, \beta_3)\,, \hat{\boldsymbol{\beta}} = (\beta_1, \alpha_2, \alpha_3) \text{ and } (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \overline{\boldsymbol{\mathcal{S}}}_3.$$

We remark that in real algebraic geometry and commutative algebra, polynomials $f^{(\boldsymbol{\alpha},\boldsymbol{\beta})}$ are known as Hibi relations, see [80].

**Lemma 4.16.** A tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is a rank-one, Frobenius-norm-one tensor if and only if

$$g_3(\mathbf{X}) = 0 \text{ and } f^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{X}) = 0 \quad \text{for all} \quad (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{S}}. \tag{4.8}$$

Proof. Sufficiency of (4.8) follows directly from the definition of the rank-one Frobenius-norm-one tensors. For necessity, the first step is to show that mode-1 fibers (columns) span one-dimensional space in $\mathbb{R}^{n_1}$. To this end, we note that for $\beta_2 \leq \alpha_2$ and $\beta_3 \leq \alpha_3$, the fibers $\mathbf{x}_{\cdot \alpha_2 \alpha_3}$ and $\mathbf{x}_{\cdot \beta_2 \beta_3}$ satisfy

$$-X_{n_1 \alpha_2 \alpha_3} \begin{bmatrix} X_{1\beta_2\beta_3} \\ X_{2\beta_2\beta_3} \\ \vdots \\ X_{n_1\beta_2\beta_3} \end{bmatrix} + X_{n_1\beta_2\beta_3} \begin{bmatrix} X_{1\alpha_2\alpha_3} \\ X_{2\alpha_2\alpha_3} \\ \vdots \\ X_{n_1\alpha_2\alpha_3} \end{bmatrix} = \begin{bmatrix} -X_{1\beta_2\beta_3}X_{n_1\alpha_2\alpha_3} + X_{1\beta_2\beta_3}X_{n_1\alpha_2\alpha_3} \\ -X_{2\beta_2\beta_3}X_{n_1\alpha_2\alpha_3} + X_{2\beta_2\beta_3}X_{n_1\alpha_2\alpha_3} \\ \vdots \\ -X_{n_1\beta_2\beta_3}X_{n_1\alpha_2\alpha_3} + X_{n_1\beta_2\beta_3}X_{n_1\alpha_2\alpha_3} \end{bmatrix} = \mathbf{0},$$

where we used that $f^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{X}) = 0$ for all $(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{S}}$. From $g_3(\mathbf{X}) = 0$ it follows that the tensor $\mathbf{X}$ is normalized.

Using similar arguments, one argues that mode-2 fibers (rows) and mode-3 fibers span one dimensional spaces in $\mathbb{R}^{n_2}$ and $\mathbb{R}^{n_3}$, respectively. This completes the proof. □

A third-order tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is rank-one if and only if all three unfoldings $\mathbf{X}^{\{1\}} \in \mathbb{R}^{n_1 \times n_2 n_3}$, $\mathbf{X}^{\{2\}} \in \mathbb{R}^{n_2 \times n_1 n_3}$, and $\mathbf{X}^{\{3\}} \in \mathbb{R}^{n_3 \times n_1 n_2}$ are rank-one matrices. Notice that $f^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{X}) = 0$ for all $(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{S}}_\ell$ is equivalent to the statement that the $\ell$-th unfolding $\mathbf{X}^{\{\ell\}}$ is a rank-one matrix, i.e., that all its order-two minors vanish, for all $\ell \in [3]$. In order to define relaxations of the unit-tensor-nuclear-norm ball we introduce the polynomial ideal $\mathcal{J}_3 \subset \mathbb{R}[\mathbf{x}] = \mathbb{R}[x_{111}, x_{112,}, \ldots, x_{n_1 n_2 n_3}]$ as the one generated by

$$\boldsymbol{\mathcal{G}}_3 = \left\{ f^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{S}} \right\} \cup \{g_3(\mathbf{x})\}, \tag{4.9}$$

i.e., $\mathcal{J}_3 = \langle \boldsymbol{\mathcal{G}}_3 \rangle$. Its real algebraic variety equals the set of rank-one third-order tensors with unit Frobenius norm and its convex hull coincides with the unit-tensor-nuclear-norm ball. The next result provides the Gröbner basis of $\mathcal{J}_3$.

**Theorem 4.17.** The basis $\boldsymbol{\mathcal{G}}_3$ defined in (4.9) forms the reduced Gröbner basis of the ideal $\mathcal{J}_3 = \langle \boldsymbol{\mathcal{G}}_3 \rangle$ with respect to the grevlex order.

Proof. Similarly to the proof of Theorem 4.8 we need to show that $S(p,q) \rightarrow_{\boldsymbol{\mathcal{G}}_3} 0$ for all polynomials $p, q \in \boldsymbol{\mathcal{G}}_3$ whose leading terms are not relatively prime. The leading monomials with respect to the grevlex ordering are given by

$$\begin{aligned} \mathrm{LM}(g_3) &= x_{111}^2 \\ \text{and } \mathrm{LM}(f^{(\boldsymbol{\alpha},\boldsymbol{\beta})}) &= x_{\boldsymbol{\alpha}} x_{\boldsymbol{\beta}}, \quad (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{S}}. \end{aligned}$$

The leading terms of $g_3$ and $f^{(\boldsymbol{\alpha},\boldsymbol{\beta})}$ are always relatively prime. First we consider two distinct polynomials $f, g \in \{f^{(\boldsymbol{\alpha},\boldsymbol{\beta})} : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{S}}_3\}$. Let $f = f^{(\boldsymbol{\alpha},\boldsymbol{\beta})}$ and $g = f^{(\boldsymbol{\alpha},\overline{\boldsymbol{\beta}})}$ for $(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \overline{\boldsymbol{\mathcal{S}}}_3$, where $\overline{\boldsymbol{\beta}} = (\beta_1, \alpha_2, \beta_3)$. That is,

$$f(\mathbf{x}) = x_{\boldsymbol{\alpha}} x_{\boldsymbol{\beta}} - x_{\boldsymbol{\alpha} \vee \boldsymbol{\beta}} x_{\boldsymbol{\alpha} \wedge \boldsymbol{\beta}}, \qquad g(\mathbf{x}) = x_{\boldsymbol{\alpha}} x_{\overline{\boldsymbol{\beta}}} - x_{\boldsymbol{\alpha} \vee \overline{\boldsymbol{\beta}}} x_{\boldsymbol{\alpha} \wedge \overline{\boldsymbol{\beta}}}.$$

| | $\mathbf{X} \in \mathbb{R}^{2\times 2\times 2}$ | $\Vert\mathbf{X}^{\{1\}}\Vert_*$ | $\Vert\mathbf{X}^{\{2\}}\Vert_*$ | $\Vert\mathbf{X}^{\{3\}}\Vert_*$ | $\Vert\mathbf{X}\Vert_{\theta_1}$ |
|---|---|---|---|---|---|
| 1 | $\left[\begin{array}{cc\|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right]$ | 2 | 2 | 2 | 2 |
| 2 | $\left[\begin{array}{cc\|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}\right]$ | 2 | 2 | $\sqrt{2}$ | 2 |
| 3 | $\left[\begin{array}{cc\|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right]$ | 2 | $\sqrt{2}$ | 2 | 2 |
| 4 | $\left[\begin{array}{cc\|cc} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right]$ | $\sqrt{2}$ | 2 | 2 | 2 |
| 5 | $\left[\begin{array}{cc\|cc} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{array}\right]$ | $\sqrt{2}+1$ | $\sqrt{2}+1$ | $\sqrt{2}+1$ | 3 |

Table 4.2. Matrix nuclear norms of unfoldings and $\theta_1$-norm of tensors $\mathbf{X} \in \mathbb{R}^{2\times 2\times 2}$, which are represented in the second column as $\mathbf{X} = [\mathbf{X}(:,:,1) \,|\, \mathbf{X}(:,:,2)]$. The third, the fourth, and the fifth column represent the nuclear norms of the first, the second, and the third unfolding of a tensor $\mathbf{X}$, respectively. The last column contains the numerically computed $\theta_1$-norm.

Since $\boldsymbol{\alpha} \wedge \boldsymbol{\beta} = \boldsymbol{\alpha} \wedge \overline{\boldsymbol{\beta}}$ and $f^{(\boldsymbol{\beta},\boldsymbol{\alpha}\vee\overline{\boldsymbol{\beta}})} \in \{f^{(\boldsymbol{\alpha},\boldsymbol{\beta})} : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{S}}_2\}$, then

$$S(f,g) = x_{\boldsymbol{\alpha}\wedge\boldsymbol{\beta}}\left(-x_{\overline{\boldsymbol{\beta}}}x_{\boldsymbol{\alpha}\vee\boldsymbol{\beta}} + x_{\boldsymbol{\beta}}x_{\boldsymbol{\alpha}\vee\overline{\boldsymbol{\beta}}}\right) = x_{\boldsymbol{\alpha}\wedge\boldsymbol{\beta}} f^{(\boldsymbol{\beta},\boldsymbol{\alpha}\vee\overline{\boldsymbol{\beta}})} \to_{\boldsymbol{\mathcal{G}}_3} 0.$$

Next we show that $S(f,g) \in \mathcal{J}_3$, for $f \in \left\{f^{(\boldsymbol{\alpha},\boldsymbol{\beta})} : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{S}}_2\right\}$ and $g \in \left\{f^{(\boldsymbol{\alpha},\boldsymbol{\beta})} : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{S}}_1\right\}$. Let $f = f^{(\boldsymbol{\alpha},\hat{\boldsymbol{\beta}})}$ with $\hat{\boldsymbol{\beta}} = (\alpha_1,\beta_2,\beta_3)$ and $g = f^{(\boldsymbol{\alpha},\tilde{\boldsymbol{\beta}})}$ with $\tilde{\boldsymbol{\beta}} = (\beta_1,\beta_2,\alpha_3)$, where $(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \overline{\boldsymbol{\mathcal{S}}}_2$. Since $x_{\boldsymbol{\alpha}\wedge\hat{\boldsymbol{\beta}}} = x_{\boldsymbol{\alpha}\wedge\tilde{\boldsymbol{\beta}}}$, $f^{(\hat{\boldsymbol{\beta}},\boldsymbol{\alpha}\vee\tilde{\boldsymbol{\beta}})} \in \left\{f^{(\boldsymbol{\alpha},\boldsymbol{\beta})} : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{S}}_3\right\}$, and $f^{(\boldsymbol{\alpha}\vee\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}})} \in \left\{f^{(\boldsymbol{\alpha},\boldsymbol{\beta})} : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{S}}_1\right\}$

$$S(f,g) = x_{\boldsymbol{\alpha}\wedge\hat{\boldsymbol{\beta}}}\left(-x_{\tilde{\boldsymbol{\beta}}}x_{\boldsymbol{\alpha}\vee\hat{\boldsymbol{\beta}}} + x_{\hat{\boldsymbol{\beta}}}x_{\boldsymbol{\alpha}\vee\tilde{\boldsymbol{\beta}}}\right) = x_{\boldsymbol{\alpha}\wedge\hat{\boldsymbol{\beta}}}\left(f^{(\hat{\boldsymbol{\beta}},\boldsymbol{\alpha}\vee\tilde{\boldsymbol{\beta}})} - f^{(\boldsymbol{\alpha}\vee\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}})}\right) \to_{\boldsymbol{\mathcal{G}}_3} 0.$$

For the remaining cases one proceeds similarly. In order to show that $\boldsymbol{\mathcal{G}}_3$ is the reduced Gröbner basis, one uses the same arguments as in the proof of Theorem 4.8. □

**Remark 4.18.** The above Gröbner basis $\boldsymbol{\mathcal{G}}_3$ is obtained by taking a particular subset of all order-two minors of all three unfoldings of the tensor $\mathbf{X} \in \mathbb{R}^{n_1\times n_2\times n_3}$ (not considering the same minor twice). One might think that the $\theta_1$-norm obtained in this way corresponds to a (weighted) sum of the nuclear norms of the unfoldings, which has been used in [62, 83] for tensor recovery. The examples of cubic tensors $\mathbf{X} \in \mathbb{R}^{2\times 2\times 2}$ presented in Table 4.2 show that this is not the case. Assuming that $\theta_1$-norm is a linear combination of the nuclear norm of the unfoldings, there exist $\alpha$, $\beta$, $\gamma \in \mathbb{R}$ such that $\alpha\Vert\mathbf{X}^{\{1\}}\Vert_* + \beta\Vert\mathbf{X}^{\{2\}}\Vert_* + \gamma\Vert\mathbf{X}^{\{3\}}\Vert_* = \Vert\mathbf{X}\Vert_{\theta_1}$. From the first and the second tensor in Table 4.2 we obtain $\gamma = 0$. Similarly, the first and the third tensor, and the first and the fourth tensor give $\beta = 0$ and $\alpha = 0$, respectively. Thus, the $\theta_1$-norm does not coincide with a weighted sum of the nuclear norms of the unfoldings. In addition, the last tensor shows that the $\theta_1$-norm does not equal the maximum of the nuclear norms of the unfoldings.

Theorem 4.17 states that $\mathcal{G}_3$ is the reduced Gröbner basis of the ideal $J_3$ generated by all order-two minors of all matricizations of an order-three tensor. That is, $J_3$ is generated by the following polynomials

$$f^{\{1\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) = -x_{\alpha_1\alpha_2\alpha_3}x_{\beta_1\beta_2\beta_3} + x_{\alpha_1\beta_2\beta_3}x_{\beta_1\alpha_2\alpha_3}, \quad \text{for } (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{1\}}$$

$$\begin{aligned}
f^{\{2\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) &= -x_{\alpha_1\alpha_2\alpha_3}x_{\beta_1\beta_2\beta_3} + x_{\beta_1\alpha_2\beta_3}x_{\alpha_1\beta_2\alpha_3}, \quad \text{for } (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{2\}} \\
f^{\{3\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) &= -x_{\alpha_1\alpha_2\alpha_3}x_{\beta_1\beta_2\beta_3} + x_{\beta_1\beta_2\alpha_3}x_{\alpha_1\alpha_2\beta_3}, \quad \text{for } (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{3\}},
\end{aligned}$$

where $\left\{f^{\{k\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{k\}}\right\}$ is the set of all order-two minors of the $k$th unfolding and

$$\boldsymbol{\mathcal{T}}^{\{k\}} = \left\{(\boldsymbol{\alpha},\boldsymbol{\beta}) : \alpha_k \neq \beta_k,\, \overline{\boldsymbol{\alpha}} \neq \overline{\boldsymbol{\beta}}, \text{where } \overline{\alpha}_k = \overline{\beta}_k = 0, \overline{\alpha}_\ell = \alpha_\ell, \overline{\beta}_\ell = \beta_\ell\right\}.$$

For $(\boldsymbol{\alpha},\boldsymbol{\beta})$, with $x_{\boldsymbol{\alpha}^{\{k\}}}x_{\boldsymbol{\beta}^{\{k\}}}$ we denote the monomial where $\alpha_k^{\{k\}} = \alpha_k$, $\beta_k^{\{k\}} = \beta_k$, and $\alpha_\ell^{\{k\}} = \beta_\ell$, $\beta_\ell^{\{k\}} = \alpha_\ell$, for all $\ell \in [d]\backslash\{k\}$. Notice that $f^{\{k\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) = f^{\{k\}}_{(\boldsymbol{\beta},\boldsymbol{\alpha})}(\mathbf{x}) = -f^{\{k\}}_{\left(\boldsymbol{\alpha}^{\{k\}},\boldsymbol{\beta}^{\{k\}}\right)}(\mathbf{x}) = -f^{\{k\}}_{\left(\boldsymbol{\beta}^{\{k\}},\boldsymbol{\alpha}^{\{k\}}\right)}(\mathbf{x})$, for all $(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{k\}}$, and all $k \in [3]$. This redundancy simplifies the proof of the next theorem. Let us now consider a TT-format and a corresponding notion of tensor rank. Recall that a TT-rank of an order three tensor is a vector $\mathbf{r} = (r_1, r_2)$ where $r_1 = \operatorname{rank}(\mathbf{X}^{\{1\}})$ and $r_2 = \operatorname{rank}(\mathbf{X}^{\{1,2\}})$. Consequently, we consider an ideal $J_{3,\mathrm{TT}}$ generated by all order-two minors of matricizations $\mathbf{X}^{\{1\}}$ and $\mathbf{X}^{\{1,2\}}$ of the order-3 tensor. That is, ideal $J_{3,\mathrm{TT}}$ is generated by the following polynomials

$$\begin{aligned}
f^{\{1\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) &= -x_{\alpha_1\alpha_2\alpha_3}x_{\beta_1\beta_2\beta_3} + x_{\alpha_1\beta_2\beta_3}x_{\beta_1\alpha_2\alpha_3}, \quad \text{for } (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{1\}} \\
f^{\{1,2\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) &= -x_{\alpha_1\alpha_2\alpha_3}x_{\beta_1\beta_2\beta_3} + x_{\alpha_1\alpha_2\beta_3}x_{\beta_1\beta_2\alpha_3}, \quad \text{for } (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{1,2\}},
\end{aligned}$$

where $\boldsymbol{\mathcal{T}}^{\{1,2\}} = \{(\boldsymbol{\alpha},\boldsymbol{\beta}) : (\alpha_1,\alpha_2,0) > (\beta_1,\beta_2,0)\,,\, \alpha_3 < \beta_3\}$.

**Theorem 4.19.** The polynomial ideals $J_3$ and $J_{3,\mathrm{TT}}$ are equal.

**Remark 4.20.** As a consequence, $\mathcal{G}_3$ is also the reduced Gröbner basis for the ideal $J_{3,\mathrm{TT}}$ with respect to the grevlex ordering.

Proof. Notice that $\left(\mathbf{X}^{\{3\}}\right)^T = \mathbf{X}^{\{1,2\}}$ and therefore

$$\left\{f^{\{3\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{3\}}\right\} = \left\{f^{\{1,2\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{1,2\}}\right\}.$$

Therefore, it is enough to show that $f^{\{2\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})} \in J_{3,\mathrm{TT}}$, for all $(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{2\}}$. By definition of $\boldsymbol{\mathcal{T}}^{\{2\}}$, we have that $\alpha_2 \neq \beta_2$ and $(\alpha_1, 0, \alpha_3) \neq (\beta_1, 0, \beta_3)$. We can assume that $\alpha_3 \neq \beta_3$, since otherwise $f^{\{2\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})} = f^{\{1\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}$. Analogously, $\alpha_1 \neq \beta_1$ since otherwise $f^{\{2\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})} = f^{\{1,2\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}$. Consider the following polynomials

$$\begin{aligned}
f(\mathbf{x}) &= -x_{\alpha_1\alpha_2\alpha_3}x_{\beta_1\beta_2\beta_3} + x_{\beta_1\alpha_2\beta_3}x_{\alpha_1\beta_2\alpha_3}, \quad (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{2\}} \\
g(\mathbf{x}) &= -x_{\beta_1\beta_2\alpha_3}x_{\alpha_1\alpha_2\beta_3} + x_{\beta_1\alpha_2\beta_3}x_{\alpha_1\beta_2\alpha_3}, \quad (\beta_1,\beta_2,\alpha_3,\alpha_1,\alpha_2,\beta_3) \in \boldsymbol{\mathcal{T}}^{\{1\}} \\
h(\mathbf{x}) &= -x_{\alpha_1\alpha_2\alpha_3}x_{\beta_1\beta_2\beta_3} + x_{\alpha_1\alpha_2\beta_3}x_{\beta_1\beta_2\alpha_3}, \quad (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{1,2\}}
\end{aligned}$$

Thus, we have that $f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}) \in J_{3,\mathrm{TT}}$. □

**4.3.2. The theta norm for general $d$th-order tensors.** Let us now consider $d$th-order tensors in $\mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ for general $d \geq 4$. Our approach relies again on the fact that a tensor $\mathbf{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ is rank-one if and only if all its matricizations are rank-one matrices, or equivalently, if all order-two minors of each matricization vanish.

The description of the polynomial ideal generated by the second order minors of all the matricizations of a tensor $\mathbf{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ unfortunately requires some technical notation. Again, we do not need all such minors in the generating set that we introduce next. In fact, this generating set will turn out to be the reduced Gröbner basis of the ideal.

Similarly to before, the entry $(\alpha_1, \alpha_2, \ldots, \alpha_d)$ of a tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ corresponds to the variable $x_{\alpha_1 \alpha_2 \cdots \alpha_d}$ or simply $x_{\boldsymbol{\alpha}}$. We aim at introducing a set of polynomials of the form

$$f_d^{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) := -x_{\boldsymbol{\alpha}\wedge\boldsymbol{\beta}} x_{\boldsymbol{\alpha}\vee\boldsymbol{\beta}} + x_{\boldsymbol{\alpha}} x_{\boldsymbol{\beta}} \tag{4.10}$$

which will generate the desired polynomial ideal. These polynomials correspond to a subset of all order-two minors of all the possible $d$th-order tensor matricizations. The set $\boldsymbol{\mathcal{S}}$ denotes the indices where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ differ. Since for an order-two minor of a matricization $\mathbf{X}^{\boldsymbol{\mathcal{M}}}$ the sets $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ need to differ in at least two indices, $\boldsymbol{\mathcal{S}}$ is contained in

$$\boldsymbol{\mathcal{S}}_{[d]} := \{\boldsymbol{\mathcal{S}} \subset [d] : 2 \leq |\boldsymbol{\mathcal{S}}| \leq d\}.$$

Given the set $\boldsymbol{\mathcal{S}}$ of different indices, we require all non-empty subsets $\boldsymbol{\mathcal{M}} \subseteq \boldsymbol{\mathcal{S}}$ of possible indices which are "switched" between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for forming the minors in (4.10). This implies that, without loss of generality,

$$\begin{aligned} \alpha_j &> \beta_j, && \text{for all } j \in \boldsymbol{\mathcal{M}} \\ \alpha_k &< \beta_k, && \text{for all } k \in \boldsymbol{\mathcal{S}} \backslash \boldsymbol{\mathcal{M}}. \end{aligned}$$

That is, the same minor is obtained if we require that $\alpha_j < \beta_j$ for all $j \in \boldsymbol{\mathcal{M}}$ and $\alpha_k > \beta_k$ for all $k \in \boldsymbol{\mathcal{S}} \backslash \boldsymbol{\mathcal{M}}$ since the set of all two-minors of $\mathbf{X}^{\boldsymbol{\mathcal{M}}}$ coincides with the set of all two-minors of $\mathbf{X}^{\boldsymbol{\mathcal{S}} \backslash \boldsymbol{\mathcal{M}}}$.

For $\boldsymbol{\mathcal{S}} \in \boldsymbol{\mathcal{S}}_{[d]}$, we define $e_{\boldsymbol{\mathcal{S}}} := \min\{p : p \in \boldsymbol{\mathcal{S}}\}$. The set $\boldsymbol{\mathcal{M}}$ corresponds to an associated matricization $\mathbf{X}^{\boldsymbol{\mathcal{M}}}$. The set of possible subsets $\boldsymbol{\mathcal{M}}$ is given as

$$\boldsymbol{\mathcal{P}}_{\boldsymbol{\mathcal{S}}} = \begin{cases} \left\{\boldsymbol{\mathcal{M}} \subset \boldsymbol{\mathcal{S}} : |\boldsymbol{\mathcal{M}}| \leq \lfloor \frac{|\boldsymbol{\mathcal{S}}|}{2} \rfloor \right\} \backslash \{\emptyset\}, & \text{if } |\boldsymbol{\mathcal{S}}| \text{ is odd,} \\ \left\{\boldsymbol{\mathcal{M}} \subset \boldsymbol{\mathcal{S}} : |\boldsymbol{\mathcal{M}}| \leq \lfloor \frac{|\boldsymbol{\mathcal{S}}|-1}{2} \rfloor \right\} \cup \left\{\boldsymbol{\mathcal{M}} \subset \boldsymbol{\mathcal{S}} : |\boldsymbol{\mathcal{M}}| = \frac{|\boldsymbol{\mathcal{S}}|}{2}, e_{\boldsymbol{\mathcal{S}}} \in \boldsymbol{\mathcal{M}} \right\} \backslash \{\emptyset\}, & \text{if } |\boldsymbol{\mathcal{S}}| \text{ is even.} \end{cases}$$

Notice that $\boldsymbol{\mathcal{P}}_{\boldsymbol{\mathcal{S}}} \cup \boldsymbol{\mathcal{P}}_{\boldsymbol{\mathcal{S}}^c} \cup \{\emptyset\} \cup \boldsymbol{\mathcal{S}}$ with $\boldsymbol{\mathcal{P}}_{\boldsymbol{\mathcal{S}}^c} := \{\boldsymbol{\mathcal{M}} : \boldsymbol{\mathcal{S}} \backslash \boldsymbol{\mathcal{M}} \in \boldsymbol{\mathcal{P}}_{\boldsymbol{\mathcal{S}}}\}$ forms the power set of $\boldsymbol{\mathcal{S}}$. The constraint on the size of $\boldsymbol{\mathcal{M}}$ in the definition of $\boldsymbol{\mathcal{P}}_{\boldsymbol{\mathcal{S}}}$ is motivated by the fact that the role of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be switched and lead to the same polynomial $f_d^{(\boldsymbol{\alpha},\boldsymbol{\beta})}$.

Thus, for $\boldsymbol{\mathcal{S}} \in \boldsymbol{\mathcal{S}}_{[d]}$ and $\boldsymbol{\mathcal{M}} \in \boldsymbol{\mathcal{P}}_{\boldsymbol{\mathcal{S}}}$, we define a set

$$\begin{aligned} \boldsymbol{\mathcal{T}}_d^{\boldsymbol{\mathcal{S}},\boldsymbol{\mathcal{M}}} := \{(\boldsymbol{\alpha},\boldsymbol{\beta}) : &\alpha_i = \beta_i, \text{ for all } i \notin \boldsymbol{\mathcal{S}} \\ &\alpha_j > \beta_j, \text{ for all } j \in \boldsymbol{\mathcal{M}} \\ &\alpha_k < \beta_k, \text{ for all } k \in \boldsymbol{\mathcal{S}} \backslash \boldsymbol{\mathcal{M}}\}. \end{aligned}$$

For notational purposes, we define

$$\{f_d^{\boldsymbol{\mathcal{S}}}\} = \cup_{\boldsymbol{\mathcal{M}} \in \boldsymbol{\mathcal{P}}_{\boldsymbol{\mathcal{S}}}} \{f_d^{(\boldsymbol{\alpha},\boldsymbol{\beta})} : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}_d^{\boldsymbol{\mathcal{S}},\boldsymbol{\mathcal{M}}}\} \quad \text{for } \boldsymbol{\mathcal{S}} \in \boldsymbol{\mathcal{S}}_{[d]}.$$

Since we are interested in Frobenius-norm-one tensors, we also introduce the polynomial

$$g_d(\mathbf{x}) = \sum_{\alpha_1=1}^{n_1} \sum_{\alpha_2=1}^{n_2} \ldots \sum_{\alpha_d=1}^{n_d} x_{\alpha_1 \alpha_2 \ldots \alpha_d}^2 - 1.$$

Our polynomial ideal is then generated by the polynomials in

$$\mathcal{G}_d = \bigcup_{\mathcal{S} \in \mathcal{S}_{[d]}} \{f_d^{\mathcal{S}}\} \cup \{g_d\} \subset \mathbb{R}[\mathbf{x}] = \mathbb{R}[x_{11\dots1}, x_{11\dots2}, \dots, x_{n_1 n_2 \dots n_d}], \quad \text{i.e.,} \quad \mathcal{J}_d = \langle \mathcal{G}_d \rangle.$$

As in the special case of the third-order tensors, not all order-two minors corresponding to all the matricizations are contained in the generating set $\mathcal{G}_d$ due to the non-leading term of $f_d^{(\boldsymbol{\alpha},\boldsymbol{\beta})}$ (since $[\boldsymbol{\alpha} \wedge \boldsymbol{\beta}](i) \leq [\boldsymbol{\alpha} \vee \boldsymbol{\beta}](i)$, for all $i \in [d]$). Nevertheless all the order-two minors are contained in the ideal $\mathcal{J}_d$ as it will also be revealed by the proof of Theorem 4.21 below. For instance, $h(\mathbf{x}) = -x_{1234}x_{2343} + x_{1243}x_{2334}$ – corresponding to a minor of the matricization $\mathbf{X}^{\mathcal{M}}$ for $\mathcal{M} = \{1, 2\}$ – does not belong to $\mathcal{G}_4$, but it does belong to the ideal $\mathcal{J}_4$. Moreover, it is straightforward to verify that all polynomials in $\mathcal{G}_d$ differ from each other.

The algebraic variety of $\mathcal{J}_d$ consists of all rank-one Frobenius-norm-one order-$d$ tensors as desired, and its convex hull yields the tensor nuclear norm ball.

**Theorem 4.21.** The set $\mathcal{G}_d$ forms the reduced Gröbner basis of the ideal $\mathcal{J}_d$ with respect to the grevlex order.

Proof of Theorem 4.21. Again, we use Buchberger's criterion stated in Theorem B.11. First notice that the polynomials $g_d$ and $f_d^{(\boldsymbol{\alpha},\boldsymbol{\beta})}$ are always relatively prime, since $\text{LM}(g_d) = x_{11\dots1}^2$ and $\text{LM}(f_d^{(\boldsymbol{\alpha},\boldsymbol{\beta})}) = x_{\boldsymbol{\alpha}} x_{\boldsymbol{\beta}}$ for $(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \mathcal{T}_d^{\mathcal{M},\mathcal{S}}$, where $\mathcal{S} \in \mathcal{S}_{[d]}$ and $\mathcal{M} \in \mathcal{P}_{\mathcal{S}}$. Therefore, we need to show that $S(f_1, f_2) \to_{\mathcal{G}_d} 0$, for all $f_1, f_2 \in \mathcal{G}_d \backslash \{g_d\}$ with $f_1 \neq f_2$. To this end, we analyze the division algorithm on $\langle \mathcal{G}_d \rangle$.

Let $f_1, f_2 \in \mathcal{G}_d$ with $f_1 \neq f_2$. Then it holds $\text{LM}(f_1) \neq \text{LM}(f_2)$. If these leading monomials are not relatively prime, the $S$-polynomial is of the form

$$S(f_1, f_2) = x_{\boldsymbol{\alpha}^1} x_{\boldsymbol{\alpha}^2} x_{\boldsymbol{\alpha}^3} - x_{\bar{\boldsymbol{\alpha}}^1} x_{\bar{\boldsymbol{\alpha}}^2} x_{\bar{\boldsymbol{\alpha}}^3}$$

with $\{\alpha_k^1, \alpha_k^2, \alpha_k^3\} = \{\bar{\alpha}_k^1, \bar{\alpha}_k^2, \bar{\alpha}_k^3\}$ for all $k \in [d]$.

The step-by-step procedure of the division algorithm for our scenario is presented in Algorithm 4.2. We will show that the algorithm eventually stops and that step 2) is feasible, i.e., that there always exist $k$ and $\ell$ such that line 7 of Algorithm 4.2 holds – provided that $S^i \neq 0$. (In fact, the purpose of the algorithm is to achieve the condition that in the $i$th iteration of the algorithm $\hat{\alpha}_k^{1,i} \leq \hat{\alpha}_k^{2,i} \leq \hat{\alpha}_k^{3,i}$, for all $k \in [d]$.) This will show then that $S(f_1, f_2) \to_{\mathcal{G}_d} 0$.

Before passing to the general proof, we illustrate the division algorithm on an example for $d = 4$. The experienced reader may skip this example.

Let $f_1(\mathbf{x}) := f_4^{(1212,2123)}(\mathbf{x}) = -x_{1112}x_{2223} + x_{1212}x_{2123} \in \mathcal{G}_4$ (with the corresponding sets $\mathcal{S} = \{1, 2, 3, 4\}$, $\mathcal{M} = \{2\}$) and $f_2(\mathbf{x}) := f_4^{(3311,2123)}(\mathbf{x}) = -x_{2111}x_{3323} + x_{3311}x_{2123} \in \mathcal{G}_4$ (with the corresponding sets $\mathcal{S} = \{1, 2, 3, 4\}$, $\mathcal{M} = \{1, 2\}$). We will show that $S(f_1, f_2) = -x_{1112}x_{2223}x_{3311} + x_{1212}x_{2111}x_{3323} \to_{\mathcal{G}_4} 0$ by going through the division algorithm.

In iteration $i = 0$ we set $S^0 = S(f_1, f_2) = -x_{1112}x_{2223}x_{3311} + x_{1212}x_{2111}x_{3323}$. The leading monomial is $\text{LM}(S^0) = x_{1112}x_{2223}x_{3311}$, the leading coefficient is $\text{LC}(S^0) = -1$, and the non-leading monomial is $\text{NLM}(S^0) = x_{1212}x_{2111}x_{3323}$. Among the two options for choosing a pair of indexes $(\boldsymbol{\alpha}^{1,0}, \boldsymbol{\alpha}^{2,0})$ in step 2), we decide to take $\boldsymbol{\alpha}^{1,0} = 1112$ and $\boldsymbol{\alpha}^{2,0} = 3311$ which leads to the set $\mathcal{M}_0 = \{4\}$. The polynomial $x_{\boldsymbol{\alpha}^{1,0}} x_{\boldsymbol{\alpha}^{2,0}} - x_{\boldsymbol{\alpha}^{1,0} \wedge \boldsymbol{\alpha}^{2,0}} x_{\boldsymbol{\alpha}^{1,0} \vee \boldsymbol{\alpha}^{2,0}}$ then equals the polynomial

**Algorithm 4.2.** The division algorithm on the ideal $\langle \boldsymbol{\mathcal{G}}_d \rangle$.

```
1:  Input: polynomials $f_1, f_2 \in \boldsymbol{\mathcal{G}}_d$
2:  Set $S^0 = S(f_1, f_2) = x_{\boldsymbol{\alpha}^1} x_{\boldsymbol{\alpha}^2} x_{\boldsymbol{\alpha}^3} - x_{\bar{\boldsymbol{\alpha}}^1} x_{\bar{\boldsymbol{\alpha}}^2} x_{\bar{\boldsymbol{\alpha}}^3}$, $i = 0$
3:      while $S^i \neq 0$ do
4:          1) Let $\mathrm{LM}(S^i) = x_{\hat{\boldsymbol{\alpha}}^{1,i}} x_{\hat{\boldsymbol{\alpha}}^{2,i}} x_{\hat{\boldsymbol{\alpha}}^{3,i}}$ and $\mathrm{NLM}(S^i) = \left|S^i - \mathrm{LT}(S^i)\right|$
5:          2) Find indices $\boldsymbol{\alpha}^{1,i}, \boldsymbol{\alpha}^{2,i} \in \{\hat{\boldsymbol{\alpha}}^{1,i}, \hat{\boldsymbol{\alpha}}^{2,i}, \hat{\boldsymbol{\alpha}}^{3,i}\}$ such that there exist
6:              at least one $k$ and at least one $\ell$ for which
7:                  $\alpha_k^{1,i} < \alpha_k^{2,i}$ and $\alpha_\ell^{1,i} > \alpha_\ell^{2,i}$ s.t. $\boldsymbol{\mathcal{M}}_i := \left\{\ell \in [d] : \alpha_\ell^{1,i} > \alpha_\ell^{2,i}\right\} \in \boldsymbol{\mathcal{P}}_{\boldsymbol{\mathcal{S}}}$,
8:              where $\boldsymbol{\mathcal{S}} := \left\{k \in [d] : \alpha_k^{1,i} \neq \alpha_k^{2,i}\right\}$
9:              and let $\boldsymbol{\alpha}^{3,i}$ be the remaining index in $\{\hat{\boldsymbol{\alpha}}^{1,i}, \hat{\boldsymbol{\alpha}}^{2,i}, \hat{\boldsymbol{\alpha}}^{3,i}\} \backslash \{\boldsymbol{\alpha}^{1,i}, \boldsymbol{\alpha}^{2,i}\}$.
10:         3) Divide $S^i$ by $f_d^{(\boldsymbol{\alpha}^{1,i}, \boldsymbol{\alpha}^{2,i})} = x_{\boldsymbol{\alpha}^{1,i}} x_{\boldsymbol{\alpha}^{2,i}} - x_{\boldsymbol{\alpha}^{1,i} \wedge \boldsymbol{\alpha}^{2,i}} x_{\boldsymbol{\alpha}^{1,i} \vee \boldsymbol{\alpha}^{2,i}}$ to obtain
11:                 $S^i = \mathrm{LC}(S^i)\big[x_{\boldsymbol{\alpha}^{3,i}}(-x_{\boldsymbol{\alpha}^{1,i} \wedge \boldsymbol{\alpha}^{2,i}} x_{\boldsymbol{\alpha}^{1,i} \vee \boldsymbol{\alpha}^{2,i}} + x_{\boldsymbol{\alpha}^{1,i}} x_{\boldsymbol{\alpha}^{2,i}})$
12:                     $+ x_{\boldsymbol{\alpha}^{1,i} \wedge \boldsymbol{\alpha}^{2,i}} x_{\boldsymbol{\alpha}^{1,i} \vee \boldsymbol{\alpha}^{2,i}} x_{\boldsymbol{\alpha}^{3,i}} - \mathrm{NLM}(S^i)\big]$.
13:         4) Define
14:                     $S^{i+1} := x_{\boldsymbol{\alpha}^{1,i} \wedge \boldsymbol{\alpha}^{2,i}} x_{\boldsymbol{\alpha}^{1,i} \vee \boldsymbol{\alpha}^{2,i}} x_{\boldsymbol{\alpha}^{3,i}} - \mathrm{NLM}(S^i)$.
15:         5) $i = i + 1$
16:     end while
```

$f_4^{(1112,3311)}(\mathbf{x}) = -x_{1111}x_{3312} + x_{1112}x_{3311} \in \boldsymbol{\mathcal{G}}_4$ and we can write

$$S^0 = -1 \cdot \Big( x_{2223}\left(-x_{1111}x_{3312} + x_{1112}x_{3311}\right) + \underbrace{x_{1111}x_{2223}x_{3312} - x_{1212}x_{2111}x_{3323}}_{= S^1} \Big).$$

The leading and non-leading monomials of $S^1$ are $\mathrm{LM}(S^1) = x_{1111}x_{2223}x_{3312}$ and $\mathrm{NLM}(S^1) = x_{1212}x_{2111}x_{3323}$, respectively, while $\mathrm{LC}(S^1) = 1$. The only option for a pair of indices as in line 7 of Algorithm 4.2 is $\boldsymbol{\alpha}^{1,1} = 3312, \boldsymbol{\alpha}^{2,1} = 2223$, so that the set $\boldsymbol{\mathcal{M}}_1 = \{1, 2\}$. The divisor $x_{\boldsymbol{\alpha}^{1,1}} x_{\boldsymbol{\alpha}^{2,1}} - x_{\boldsymbol{\alpha}^{1,1} \wedge \boldsymbol{\alpha}^{2,1}} x_{\boldsymbol{\alpha}^{1,1} \vee \boldsymbol{\alpha}^{2,1}}$ in the step 4) equals $f_4^{(3312,2223)}(\mathbf{x}) = -x_{2212}x_{3323} + x_{3312}x_{2223} \in \boldsymbol{\mathcal{G}}_4$ and we obtain

$$S^1 = 1 \cdot \Big( x_{1111}\left(-x_{2212}x_{3323} + x_{2223}x_{3312}\right) + \underbrace{x_{1111}x_{2212}x_{3323} - x_{1212}x_{2111}x_{3323}}_{= S^2} \Big).$$

The index sets of the monomial $x_{\boldsymbol{\alpha}^1} x_{\boldsymbol{\alpha}^2} x_{\boldsymbol{\alpha}^3} = x_{1111}x_{2212}x_{3323}$ in $S^2$ satisfy

$$\alpha_k^1 \leq \alpha_k^2 \leq \alpha_k^3 \quad \text{for all } k \in [4]$$

and therefore it is the non-leading monomial of $S^2$, i.e., $\mathrm{NLM}(S^2) = x_{1111}x_{2212}x_{3323}$. Thus, $\mathrm{LM}(S^2) = x_{1212}x_{2111}x_{3323}$ and $\mathrm{LC}(S^2(f_1, f_2)) = -1$. Now the only option for a pair of indices as in step 2) is $\boldsymbol{\alpha}^{1,2} = 2111$, $\boldsymbol{\alpha}^{2,2} = 1212$ with $\boldsymbol{\mathcal{M}}_2 = \{1\}$. This yields

$$S^2 = -1 \cdot \Big( x_{3323}\left(-x_{1111}x_{2212} + x_{2111}x_{1212}\right) + \underbrace{x_{1111}x_{2212}x_{3323} - x_{1111}x_{2212}x_{3323}}_{= S^3 = 0} \Big).$$

Thus, the division algorithm stops and we obtained after three steps

$$S(f_1, f_2) = S^0 = \mathrm{LC}(S^0) x_{2223} f_4^{(1112,3311)}(\mathbf{x}) + \mathrm{LC}(S^0)\,\mathrm{LC}(S^1) x_{1111} f_4^{(3312,2223)}(\mathbf{x})$$
$$+ \mathrm{LC}(S^0)\,\mathrm{LC}(S^1)\,\mathrm{LC}(S^2) x_{3323} f_4^{(2111,1212)}(\mathbf{x}).$$

Thus, $S(f_1, f_2) \to_{\boldsymbol{\mathcal{G}}_4} 0$.

Let us now return to the general proof. We first show that there always exist indices $\boldsymbol{\alpha}^{1,i}, \boldsymbol{\alpha}^{2,i}$ satisfying line 7 of Algorithm 4.2 unless $S^i = 0$. We start by setting $\mathbf{x}^{\boldsymbol{\alpha}_i} = x_{\hat{\boldsymbol{\alpha}}^{1,i}} x_{\hat{\boldsymbol{\alpha}}^{2,i}} x_{\hat{\boldsymbol{\alpha}}^{3,i}}$ with $x_{\hat{\boldsymbol{\alpha}}^{1,i}} \geq x_{\hat{\boldsymbol{\alpha}}^{2,i}} \geq x_{\hat{\boldsymbol{\alpha}}^{3,i}}$ to be the leading monomial and $\mathbf{x}^{\boldsymbol{\beta}_i}$ to be the non-leading monomial of $S^i$. The existence of a polynomial $h \in \boldsymbol{\mathcal{G}}_d$ such that $\mathrm{LM}(h)$ divides $\mathrm{LM}(S^i) = x_{\hat{\boldsymbol{\alpha}}^{1,i}} x_{\hat{\boldsymbol{\alpha}}^{2,i}} x_{\hat{\boldsymbol{\alpha}}^{3,i}} = \mathbf{x}^{\boldsymbol{\alpha}_i}$ is equivalent to the existence of $\boldsymbol{\alpha}^{1,i}, \boldsymbol{\alpha}^{2,i} \in \left\{ \hat{\boldsymbol{\alpha}}^{1,i}, \hat{\boldsymbol{\alpha}}^{2,i}, \hat{\boldsymbol{\alpha}}^{3,i} \right\}$ such that there exists at least one $k$ and at least one $\ell$ for which $\alpha_k^{1,i} < \alpha_k^{2,i}$ and $\alpha_\ell^{1,i} > \alpha_\ell^{2,i}$. If such pair does not exist in iteration $i$, we have

$$\hat{\alpha}_k^{1,i} \leq \hat{\alpha}_k^{2,i} \leq \hat{\alpha}_k^{3,i} \quad \text{for all } k \in [d]\,. \tag{4.11}$$

We claim that this cannot happen if $S^i \neq 0$. In fact, (4.11) would imply that the monomial $\mathbf{x}^{\boldsymbol{\alpha}_i} = x_{\hat{\boldsymbol{\alpha}}^{1,i}} x_{\hat{\boldsymbol{\alpha}}^{2,i}} x_{\hat{\boldsymbol{\alpha}}^{3,i}}$ is the smallest monomial $x_{\boldsymbol{\beta}} x_{\boldsymbol{\gamma}} x_{\boldsymbol{\eta}}$ (with respect to the grevlex order) which satisfies

$$\{\beta_k, \gamma_k, \eta_k\} = \{\hat{\alpha}_k^{1,i}, \hat{\alpha}_k^{2,i}, \hat{\alpha}_k^{3,i}\} \quad \text{for all } k \in [d]\,.$$

However, then $\mathbf{x}^{\boldsymbol{\alpha}_i}$ would not be the leading monomial by definition of the grevlex order, which leads to a contradiction. Hence, we can always find indices $\boldsymbol{\alpha}^{1,i}, \boldsymbol{\alpha}^{2,i}$ satisfying line 7 in step 2) of Algorithm 4.2 unless $S^i = 0$.

Next we show that the division algorithm always stops in a finite number of steps. We start with iteration $i = 0$ and assume that $S^0 \neq 0$. We choose $\boldsymbol{\alpha}^{1,0}, \boldsymbol{\alpha}^{2,0}, \boldsymbol{\alpha}^{3,0}$ as in step 2) of Algorithm 4.2. Then we divide the polynomial $S^0$ by a polynomial $h \in \boldsymbol{\mathcal{G}}_d$ such that $\mathrm{LM}(h) = x_{\boldsymbol{\alpha}^{1,0}} x_{\boldsymbol{\alpha}^{2,0}}$. The polynomial $h \in \boldsymbol{\mathcal{G}}_d$ is defined as in step 3) of the algorithm, i.e.,

$$h(\mathbf{x}) = f_d^{\left(\boldsymbol{\alpha}^{1,0}, \boldsymbol{\alpha}^{2,0}\right)} = x_{\boldsymbol{\alpha}^{1,0}} x_{\boldsymbol{\alpha}^{2,0}} - x_{\boldsymbol{\alpha}^{1,0} \wedge \boldsymbol{\alpha}^{2,0}} x_{\boldsymbol{\alpha}^{1,0} \vee \boldsymbol{\alpha}^{2,0}} \in \boldsymbol{\mathcal{G}}_d.$$

The division of $S^0$ by $h$ results in

$$S^0 = \mathrm{LC}(S^0) \Big( x_{\boldsymbol{\alpha}^{3,0}} \cdot f_d^{\left(\boldsymbol{\alpha}^{1,0}, \boldsymbol{\alpha}^{2,0}\right)} + \underbrace{x_{\boldsymbol{\alpha}^{1,0} \wedge \boldsymbol{\alpha}^{2,0}} x_{\boldsymbol{\alpha}^{1,0} \vee \boldsymbol{\alpha}^{2,0}} x_{\boldsymbol{\alpha}^{3,0}} - \mathrm{NLM}(S^0)}_{= S^1} \Big).$$

Note that by construction

$$\left[\boldsymbol{\alpha}^{1,0} \wedge \boldsymbol{\alpha}^{2,0}\right]_k \leq \left[\boldsymbol{\alpha}^{1,0} \vee \boldsymbol{\alpha}^{2,0}\right]_k \quad \text{for all } k \in [d]\,. \tag{4.12}$$

If $S^1 \neq 0$, then in the following iteration $i = 1$ we can assume $\mathrm{LM}(S^1) = x_{\boldsymbol{\alpha}^{1,0} \wedge \boldsymbol{\alpha}^{2,0}} x_{\boldsymbol{\alpha}^{1,0} \wedge \boldsymbol{\alpha}^{2,0}} x_{\boldsymbol{\alpha}^{3,0}}$. Due to (4.12), a pair $\boldsymbol{\alpha}^{1,1}, \boldsymbol{\alpha}^{2,1}$ as in line 7 of Algorithm 4.2 can be either $\boldsymbol{\alpha}^{1,0} \wedge \boldsymbol{\alpha}^{2,0}, \boldsymbol{\alpha}^{3,0}$ or $\boldsymbol{\alpha}^{1,0} \vee \boldsymbol{\alpha}^{2,0}, \boldsymbol{\alpha}^{3,0}$. Let us assume the former. Then this iteration results in

$$S^1 = \mathrm{LC}(S^1) \Big( x_{\boldsymbol{\alpha}^{3,1}} \cdot f_d^{\left(\boldsymbol{\alpha}^{1,1}, \boldsymbol{\alpha}^{2,1}\right)} + \underbrace{x_{\boldsymbol{\alpha}^{1,1} \wedge \boldsymbol{\alpha}^{2,1}} x_{\boldsymbol{\alpha}^{1,1} \vee \boldsymbol{\alpha}^{2,1}} x_{\boldsymbol{\alpha}^{3,1}} - \mathrm{NLM}(S^0)}_{= S^2} \Big)$$

with

$$\left[\boldsymbol{\alpha}^{1,1} \wedge \boldsymbol{\alpha}^{2,1}\right]_k \leq \left[\boldsymbol{\alpha}^{3,1}\right]_k, \left[\boldsymbol{\alpha}^{1,1} \vee \boldsymbol{\alpha}^{2,1}\right]_k \quad \text{for all } k \in [d]\,, \text{ and } x_{\boldsymbol{\alpha}^{3,1}} = x_{\boldsymbol{\alpha}^{1,0} \vee \boldsymbol{\alpha}^{2,0}}.$$

Next, if $S^2 \neq 0$ and $\mathrm{LM}(S^2) = x_{\boldsymbol{\alpha}^{1,1} \wedge \boldsymbol{\alpha}^{2,1}} x_{\boldsymbol{\alpha}^{1,1} \vee \boldsymbol{\alpha}^{2,1}} x_{\boldsymbol{\alpha}^{3,1}}$ then a pair of indices satisfying line 7 of Algorithm 4.2 must be $\boldsymbol{\alpha}^{1,1} \vee \boldsymbol{\alpha}^{2,1}, \boldsymbol{\alpha}^{3,1}$ so that the iteration ends up with

$$S^2 = \mathrm{LC}(S^2) \Big( x_{\boldsymbol{\alpha}^{3,2}} \cdot f_d^{\left(\boldsymbol{\alpha}^{1,2}, \boldsymbol{\alpha}^{2,2}\right)} + \underbrace{x_{\boldsymbol{\alpha}^{1,2} \wedge \boldsymbol{\alpha}^{2,2}} x_{\boldsymbol{\alpha}^{1,2} \vee \boldsymbol{\alpha}^{2,2}} x_{\boldsymbol{\alpha}^{3,2}} - \mathrm{NLM}(S^0)}_{= S^3} \Big)$$

such that

$$\left[\boldsymbol{\alpha}^{3,2}\right]_k \leq \left[\boldsymbol{\alpha}^{1,2} \wedge \boldsymbol{\alpha}^{2,2}\right]_k \leq \left[\boldsymbol{\alpha}^{1,2} \vee \boldsymbol{\alpha}^{2,2}\right]_k \quad \text{for all } k \in [d]\,, \text{ and } x_{\boldsymbol{\alpha}^{3,2}} = x_{\boldsymbol{\alpha}^{1,1} \wedge \boldsymbol{\alpha}^{2,1}}.$$

Thus, in iteration $i = 3$ the leading monomial $\mathrm{LM}(S^3)$ must be $\mathrm{NLM}(S^0)$ (unless $S^3 = 0$).

A similar analysis can be performed on the monomial $\mathrm{NLM}(S^0)$ and therefore the algorithm stops after at most 6 iterations. The division algorithm results in

$$S(f_1, f_2) = \sum_{i=0}^{p} \left( \prod_{j=0}^{i} \mathrm{LC}(S^j) \right) x_{\boldsymbol{\alpha}^{3,i}} \cdot f_d^{\left(\boldsymbol{\alpha}^{1,i}, \boldsymbol{\alpha}^{2,i}\right)},$$

where $f_d^{\left(\boldsymbol{\alpha}^{1,i}, \boldsymbol{\alpha}^{2,i}\right)} = -x_{\boldsymbol{\alpha}^{1,i} \wedge \boldsymbol{\alpha}^{2,i}} x_{\boldsymbol{\alpha}^{1,i} \vee \boldsymbol{\alpha}^{2,i}} + x_{\boldsymbol{\alpha}^{1,i}} x_{\boldsymbol{\alpha}^{2,i}} \in \boldsymbol{\mathcal{G}}_d$ and $p \leq 5$. All the cases that we left out above are treated in a similar way. This shows that $\boldsymbol{\mathcal{G}}_d$ is a Gröbner basis of $\mathcal{J}_d$.

In order to show that $\boldsymbol{\mathcal{G}}_d$ is the *reduced* Gröbner basis of $\mathcal{J}_d$, first notice that $\mathrm{LC}(g) = 1$ for all $g \in \boldsymbol{\mathcal{G}}_d$. Furthermore, the leading term of any polynomial in $\boldsymbol{\mathcal{G}}_d$ is of degree two. Thus, it is enough to show that for every pair of different polynomials $f_d^{(\boldsymbol{\alpha}^1, \boldsymbol{\beta}^1)}, f_d^{(\boldsymbol{\alpha}^2, \boldsymbol{\beta}^2)} \in \boldsymbol{\mathcal{G}}_d$ (related to $\boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{M}}_1$ and $\boldsymbol{\mathcal{S}}_2, \boldsymbol{\mathcal{M}}_2$, respectively) it holds that $\mathrm{LM}(f_d^{(\boldsymbol{\alpha}^1, \boldsymbol{\beta}^1)}) \neq \mathrm{LM}(f_d^{(\boldsymbol{\alpha}^2, \boldsymbol{\beta}^2)})$ with $(\boldsymbol{\alpha}^k, \boldsymbol{\beta}^k) \in \boldsymbol{\mathcal{T}}_d^{\boldsymbol{\mathcal{S}}_k, \boldsymbol{\mathcal{M}}_k}$ for $k = 1, 2$. But this follows from the fact that all elements of $\boldsymbol{\mathcal{G}}_d$ are different as remarked before the statement of the theorem. □

We define the tensor $\theta_k$-norm analogously to the matrix scenario.

**Definition 4.22.** The tensor $\theta_k$*-norm*, denoted by $\|\cdot\|_{\theta_k}$, is the norm induced by the $k$-theta body $\mathrm{TH}_k\left(J_d\right)$, i.e.,

$$\|\mathbf{X}\|_{\theta_k} = \inf\left\{r : \mathbf{X} \in r\, \mathrm{TH}_k\left(J_d\right)\right\}.$$

The $\theta_k$-norm can be computed with the help of Theorem 4.5, i.e., as

$$\|\mathbf{X}\|_{\theta_k} = \min t \quad \text{subject to } \mathbf{X} \in t\mathbf{Q}_{\mathcal{B}_k}(J_{M_{mn}}).$$

Given the moment matrix $\mathbf{M}_{\mathcal{B}_k}[\mathbf{y}]$ associated with $J$, this minimization program is equivalent to the semidefinite program

$$\min_{t \in \mathbb{R}, \mathbf{y} \in \mathbb{R}^{\mathcal{B}_k}} t \quad \text{subject to} \quad \mathbf{M}_{\mathcal{B}_k}[\mathbf{y}] \succcurlyeq 0, y_0 = t, \mathbf{y}_{\mathcal{B}_1} = \mathbf{X}. \tag{4.13}$$

We have focused on the polynomial ideal generated by all second order minors of all matricizations of the tensor. One may also consider a subset of all possible matricizations corresponding to various tensor decompositions and notions of tensor rank. For example, the Tucker(HOSVD)-rank (corresponding to the Tucker or HOSVD decomposition) of a $d$th-order tensor $\mathbf{X}$ is a $d$-dimensional vector $\mathbf{r}_{HOSVD} = (r_1, r_2, \ldots, r_d)$ such that $r_i = \operatorname{rank}\left(\mathbf{X}^{\{i\}}\right)$ for all $i \in [d]$, see Subsection 2.1.4. Thus, we can define an ideal $J_{d,\mathrm{HOSVD}}$ generated by all second order minors of unfoldings $\mathbf{X}^{\{k\}}$, for $k \in [d]$.

The tensor train (TT) decomposition is another popular approach for tensor computations [125]. The corresponding TT-rank of a $d$th-order tensor $\mathbf{X}$ is a $(d-1)$-dimensional vector $\mathbf{r}_{TT} = (r_1, r_2, \ldots, r_{d-1})$ such that $r_i = \operatorname{rank}\left(\mathbf{X}^{\{1,\ldots,i\}}\right)$, $i \in [d-1]$. By taking into account only minors of order two of the matricizations $\boldsymbol{\tau} \in \{\{1\}, \{1,2\}, \ldots, \{1,2,\ldots,d-1\}\}$, one may introduce a corresponding polynomial ideal $J_{d,\mathrm{TT}}$.

**Theorem 4.23.** The polynomial ideals $J_d$, $J_{d,\mathrm{HOSVD}}$, and $J_{d,\mathrm{TT}}$ are equal, for all $d \geq 3$.

PROOF. Let $\boldsymbol{\tau} \subset [d]$ represents a matricization and let $\boldsymbol{\Delta} := \{(\boldsymbol{\alpha},\boldsymbol{\beta}) : \alpha_i, \beta_i \in [n_i]\,, \text{ for all } i \in [d]\}$. Similarly to the case of order-three tensors, for $(\boldsymbol{\alpha},\boldsymbol{\beta})$, with $x_{\boldsymbol{\alpha}^{\boldsymbol{\tau}}} x_{\boldsymbol{\beta}^{\boldsymbol{\tau}}}$ we denote the monomial where $\alpha^{\boldsymbol{\tau}}_k = \alpha_k$, $\beta^{\boldsymbol{\tau}}_k = \beta_k$ for all $k \in \boldsymbol{\tau}$ and $\alpha^{\boldsymbol{\tau}}_\ell = \beta_\ell$, $\beta^{\boldsymbol{\tau}}_\ell = \alpha_\ell$ for all $\ell \in \boldsymbol{\tau}^c = [d] \backslash \boldsymbol{\tau}$. Additionally, for $(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\Delta}$, with $x_{\boldsymbol{\alpha}^{\boldsymbol{\tau},\mathbf{0}}} x_{\boldsymbol{\beta}^{\boldsymbol{\tau},\mathbf{0}}}$ we denote the monomial where $\alpha^{\boldsymbol{\tau},\mathbf{0}}_k = \alpha_k$, $\beta^{\boldsymbol{\tau},\mathbf{0}}_k = \beta_k$ for all $k \in \boldsymbol{\tau}$ and $\alpha^{\boldsymbol{\tau},\mathbf{0}}_\ell = \beta^{\boldsymbol{\tau},\mathbf{0}}_\ell = 0$ for all $\ell \in \boldsymbol{\tau}^c = [d] \backslash \boldsymbol{\tau}$. The corresponding order-two minors are defined as

$$f^{\boldsymbol{\tau}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) = -x_{\boldsymbol{\alpha}} x_{\boldsymbol{\beta}} + x_{\boldsymbol{\alpha}^{\boldsymbol{\tau}}} x_{\boldsymbol{\beta}^{\boldsymbol{\tau}}}, \quad (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\boldsymbol{\tau}}.$$

We define the set $\boldsymbol{\mathcal{T}}^{\boldsymbol{\tau}}$ as

$$\boldsymbol{\mathcal{T}}^{\boldsymbol{\tau}} = \left\{ (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\Delta} : \boldsymbol{\alpha}^{\boldsymbol{\tau},\mathbf{0}} \neq \boldsymbol{\beta}^{\boldsymbol{\tau},\mathbf{0}},\, \boldsymbol{\alpha}^{\boldsymbol{\tau}^c,\mathbf{0}} \neq \boldsymbol{\beta}^{\boldsymbol{\tau}^c,\mathbf{0}} \right\}.$$

Similarly to the case of order-three tensors, notice that $f^{\boldsymbol{\tau}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) = f^{\boldsymbol{\tau}}_{(\boldsymbol{\beta},\boldsymbol{\alpha})}(\mathbf{x}) = -f^{\boldsymbol{\tau}}_{(\boldsymbol{\alpha}^{\boldsymbol{\tau}},\boldsymbol{\beta}^{\boldsymbol{\tau}})}(\mathbf{x}) = -f^{\boldsymbol{\tau}}_{(\boldsymbol{\beta}^{\boldsymbol{\tau}},\boldsymbol{\alpha}^{\boldsymbol{\tau}})}(\mathbf{x})$, for all $(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\boldsymbol{\tau}}$. First, we show that $J_d = J_{d,\text{HOSVD}}$ by showing that $f^{\boldsymbol{\tau}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) \in J_{d,\text{HOSVD}}$, for all $(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\boldsymbol{\tau}}$ and all $|\boldsymbol{\tau}| \geq 2$. Without loss of generality, we can assume that $\alpha_i \neq \beta_i$, for all $i \in \boldsymbol{\tau}$ since otherwise we can consider the matricization $\boldsymbol{\tau} \backslash \{i : \alpha_i = \beta_i\}$. Additionally, by definition of $\boldsymbol{\mathcal{T}}^{\boldsymbol{\tau}}$, there exists at least one $\ell \in \boldsymbol{\tau}^c$ such that $\alpha_\ell \neq \beta_\ell$. Let $\boldsymbol{\tau} = \{t_1, t_2, \ldots, t_k\}$ with $t_i < t_{i+1}$, for all $i \in [k-1]$ and $k \geq 2$. Next, fix $(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \mathcal{T}^{\boldsymbol{\tau}}$ and define $\boldsymbol{\alpha}^0 = \boldsymbol{\alpha}$ and $\boldsymbol{\beta}^{\mathbf{0}} = \boldsymbol{\beta}$. Algorithm 4.3 results in polynomials $g_k \in J_{3,\text{TT}}$ such that $f^{\boldsymbol{\tau}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) = \sum_{i=1}^k g_i(\mathbf{x})$. This is clear, since

$$\sum_{i=1}^k g_i = \sum_{i=1}^k \left( -x_{\boldsymbol{\alpha}^{i-1}} x_{\boldsymbol{\beta}^{i-1}} + x_{\boldsymbol{\alpha}^i} x_{\boldsymbol{\beta}^i} \right) = -x_{\boldsymbol{\alpha}^0} x_{\boldsymbol{\beta}^0} + x_{\boldsymbol{\alpha}^k} x_{\boldsymbol{\beta}^k} = f^{\boldsymbol{\tau}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}).$$

By the definition of polynomials $g_k$ it is obvious that $g_i \in \left\{ f^{\{i\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{i\}} \right\}$, for all $i \in [k]$. Next, we show that $J_d = J_{d,\text{TT}}$. Since $J_d = J_{d,\text{HOSVD}}$, it is enough to show that $f^{\{k\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})} \in J_{d,\text{TT}}$, for all $(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{k\}}$ and all $k \in [d]$. By definition of $J_{d,\text{TT}}$ this is true for $k = 1$. Fix $k \in \{2, 3, \ldots, d\}$, $(\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{k\}}$ and consider a polynomial $f(\mathbf{x}) = f^{\{k\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x})$ corresponding to the second order minor of the matricization $\mathbf{X}^{\{k\}}$. By definition of $\boldsymbol{\mathcal{T}}^{\{k\}}$, $\alpha_k \neq \beta_k$ and there exists an index $i \in [d] \backslash \{k\}$ such that $\alpha_i \neq \beta_i$. Assume that $i > k$. Define the polynomials $g(\mathbf{x}) \in \boldsymbol{\mathcal{R}}^{\{1,2,\ldots,k\}} := \left\{ f^{\{1,2,\ldots,k\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{1,2,\ldots,k\}} \right\}$ and $h(\mathbf{x}) \in \boldsymbol{\mathcal{R}}^{\{1,2,\ldots,k-1\}} := \left\{ f^{\{1,2,\ldots,k-1\}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x}) : (\boldsymbol{\alpha},\boldsymbol{\beta}) \in \boldsymbol{\mathcal{T}}^{\{1,2,\ldots,k-1\}} \right\}$ as

**Algorithm 4.3.** Algorithm for proving that $J_d = J_{d,\text{TT}}$

```
Input: An ideal $J_{d,\mathbf{TT}} \in \mathbb{R}[\mathbf{x}]$, polynomial $f^{\boldsymbol{\tau}}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathbf{x})$
        with $\boldsymbol{\alpha}^0 = \boldsymbol{\alpha}$, $\boldsymbol{\beta}^0 = \boldsymbol{\beta}$, $\boldsymbol{\tau} = \{t_1, t_2, \ldots, t_k\}$, where $k \geq 2$
  for $i = 1 : k$
    Define $\boldsymbol{\alpha}^i$ and $\boldsymbol{\beta}^i$ as
      $\alpha^i_j := \begin{cases} \beta^{i-1}_j & \text{if } j = t_i, \\ \alpha^{i-1}_j & \text{otherwise} \end{cases}$ and $\beta^i_j := \begin{cases} \alpha^{i-1}_j & \text{if } j = t_i, \\ \beta^{i-1}_j & \text{otherwise.} \end{cases}$
    Define polynomial $g_i(\mathbf{x}) := -x_{\boldsymbol{\alpha}^{i-1}} x_{\boldsymbol{\beta}^{i-1}} + x_{\boldsymbol{\alpha}^i} x_{\boldsymbol{\beta}^i}$ .
  end for
Output: Polynomials $g_1, g_2, \ldots, g_k$.
```

$$g(\mathbf{x}) = -x_{\boldsymbol{\alpha}} x_{\boldsymbol{\beta}} + x_{\boldsymbol{\alpha}^{\{1,2,\ldots,k\}}} x_{\boldsymbol{\beta}^{\{1,2,\ldots,k\}}}$$

$$h(\mathbf{x}) = -x_{\boldsymbol{\alpha}^{\{1,2,\ldots,k\}}} x_{\boldsymbol{\beta}^{\{1,2,\ldots,k\}}} + x_{\boldsymbol{\alpha}^{\{1,2,\ldots,k\}\{1,2,\ldots,k-1\}}} x_{\boldsymbol{\beta}^{\{1,2,\ldots,k\}\{1,2,\ldots,k-1\}}}$$

Since $x_{\boldsymbol{\alpha}^{\{1,2,\ldots,k\}\{1,2,\ldots,k-1\}}} x_{\boldsymbol{\beta}^{\{1,2,\ldots,k\}\{1,2,\ldots,k-1\}}} = x_{\boldsymbol{\alpha}^{\{k\}}} x_{\boldsymbol{\beta}^{\{k\}}}$, we obtained that $f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x})$ and thus $f \in J_{d,\mathrm{TT}}$. For the other case, i.e., when $i < k$ notice that $f(\mathbf{x}) = g_1(\mathbf{x}) + h_1(\mathbf{x})$, where

$$\begin{aligned} g_1(\mathbf{x}) &= -x_{\boldsymbol{\alpha}} x_{\boldsymbol{\beta}} + x_{\boldsymbol{\alpha}^{\{1,2,\ldots,k-1\}}} x_{\boldsymbol{\beta}^{\{1,2,\ldots,k-1\}}} \in \boldsymbol{\mathcal{R}}^{\{1,2,\ldots,k-1\}} \\ h_1(\mathbf{x}) &= -x_{\boldsymbol{\alpha}^{\{1,2,\ldots,k-1\}}} x_{\boldsymbol{\beta}^{\{1,2,\ldots,k-1\}}} + x_{\boldsymbol{\alpha}^{\{1,2,\ldots,k-1\}\{1,2,\ldots,k\}}} x_{\boldsymbol{\beta}^{\{1,2,\ldots,k-1\}\{1,2,\ldots,k\}}} \\ &= -x_{\boldsymbol{\alpha}^{\{1,2,\ldots,k\}}} x_{\boldsymbol{\beta}^{\{1,2,\ldots,k\}}} + x_{\boldsymbol{\alpha}^{\{k\}}} x_{\boldsymbol{\beta}^{\{k\}}} \in \boldsymbol{\mathcal{R}}^{\{1,2,\ldots,k\}}. \end{aligned}$$

□

**Remark 4.24.** Fix a decomposition tree $T_I$ which generates a particular HT-decomposition and consider the ideal $J_{d,\mathrm{HT},T_I}$ generated by all second order minors corresponding to the matricizations induced by the tree $T_I$. In a similar way as above, one can obtain that $J_{d,\mathrm{HT},T_I}$ equals to $J_d$.

## 4.4. Computational complexity

The computational complexity of the semidefinite programs for computing the tensor $\theta_1$-norm or for minimizing the $\theta_1$-norm subject to a linear constraint depends polynomially on the number of variables, i.e., on the size of $\boldsymbol{\mathcal{B}}_2$, and on the dimension of the corresponding moment matrix $\mathbf{M}$. We claim that the overall complexity scales polynomially in $n^d$, where for simplicity we consider $d$th-order tensors in $\mathbb{R}^{n \times n \times \cdots \times n}$. Therefore, in contrast to the tensor nuclear norm minimization which is NP-hard in general for $d \geq 3$, the tensor recovery via $\theta_1$-norm minimization is tractable.

Indeed, the dimension of the moment matrix $\mathbf{M}$ is $(1 + n^d) \times (1 + n^d)$ (see also (4.7) for matrices in $\mathbb{R}^{2\times 2}$) and if $a = n^d$ denotes the total number of entries of a tensor $\mathbf{X} \in \mathbb{R}^{n \times n \times \cdots \times n}$, then the number of the variables is at most $\frac{a\cdot(a+1)}{2} \sim \boldsymbol{\mathcal{O}}(a^2)$ which is polynomial in $a$. (A more precise counting does not give a substantially better estimate.)

**Symmetric tensors.** We may reduce the complexity of our semidefinite program by reducing to tensors possessing symmetries. Of course, in practice this requires additional information about the tensors to be recovered. For example, let us consider the case of $d$th-order symmetric-tensors, i.e., tensors $\mathbf{X} \in \mathbb{R}^{n \times n \times \cdots \times n}$ such that $X_{\alpha_1 \alpha_2 \ldots \alpha_d} = X_{\sigma(\alpha_1)\sigma(\alpha_2)\cdots\sigma(\alpha_d)}$ for all the possible permutations $\sigma : \{\alpha_1, \alpha_2, \ldots, \alpha_d\} \to \{\alpha_1, \alpha_2, \ldots, \alpha_d\}$. In this scenario, the size of the semidefinite program for computing the $\theta_1$-norm is $(a+1) \times (a+1)$, where

$$a = \binom{n+d-1}{d} \leq \left(e\frac{n+d-1}{d}\right)^d = e^d \left(1 + \frac{n-1}{d}\right)^d.$$

The above inequality uses the general estimate $\binom{p}{q} \leq (ep/q)^q$, see Lemma A.1. The number of variables in the corresponding semidefinite program for computing the $\theta_1$-norm equals the number of monomials $x_{\boldsymbol{\alpha}} x_{\hat{\boldsymbol{\alpha}}}$ such that $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_d \leq \hat{\alpha}_1 \leq \ldots \leq \hat{\alpha}_d$, excluding the monomial $x_{11\ldots 1}^2 = \mathrm{LM}(g_d)$, which is

$$\binom{n+2d-1}{2d} - 1 \leq e^{2d}\left(1 + \frac{n-1}{2d}\right)^{2d}.$$

We leave the study of low-rank symmetric tensor recovery via $\theta_k$-minimization to the future investigation.

### 4.5. Numerical experiments

Let us now empirically study the performance of low-rank tensor recovery via $\theta_1$-norm minimization via numerical experiments, where we concentrate on third-order tensors. Given the measurements $\mathbf{b} = \mathcal{A}(\mathbf{X})$ of a low-rank tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, where $\mathcal{A} : \mathbb{R}^{n_1 \times n_2 \times n_3} \to \mathbb{R}^m$ is a linear measurement map, we aim at reconstructing $\mathbf{X}$ as the solution of the minimization program

$$\min \|\mathbf{Z}\|_{\theta_1} \quad \text{subject to} \quad \mathcal{A}(\mathbf{Z}) = \mathbf{b}. \tag{4.14}$$

As outlined in Section 4.1, the $\theta_1$-norm of a tensor $\mathbf{Z}$ can be computed as the minimizer of the semidefinite program

$$\min_{t,\mathbf{y}} t \quad \text{subject to} \quad \mathbf{M}(t,\mathbf{y},\mathbf{Z}) \succcurlyeq 0,$$

where $\mathbf{M}(t,\mathbf{y},\mathbf{X}) = \mathbf{M}_{\boldsymbol{\mathcal{B}}_1}(t,\mathbf{X},\mathbf{y})$ is the moment matrix of order 1 associated to the ideal $\mathcal{J}_3$, see Theorem 4.17. This moment matrix for $\mathcal{J}_3$ is explicitly given by

$$\mathbf{M}\left(t,\mathbf{y},\mathbf{X}\right) = t\mathbf{M}_0 + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} X_{ijk}\mathbf{M}_{ijk} + \sum_{p=2}^{9}\sum_{q=1}^{|\mathbf{M}^p|} y_\ell \mathbf{M}^p_{h_p(q)},$$

where $\ell = \sum_{r=2}^{p-1} |\mathbf{M}^r| + q$, $\mathbf{M}^p = \{\mathbf{M}^p_{\widetilde{I}}\}$, and the matrices $\mathbf{M}_0, \mathbf{M}_{ijk}$ and $\mathbf{M}^p_{\widetilde{I}}$ are provided in Table 4.3. For $p \in \{2,3,\ldots,9\}$, the function $h_p$ denotes an arbitrary but fixed bijection $\{1,2,\ldots,|\mathbf{M}^p|\} \mapsto \{(i,\hat{\imath},j,\hat{\jmath},k,\hat{k})\}$, where $\widetilde{I} = (i,\hat{\imath},j,\hat{\jmath},k,\hat{k})$ is in the range of the last column of Table 4.3. As discussed in Section 4.1 for the general case, the $\theta_1$-norm minimization problem (4.14) is then equivalent to the semidefinite program

$$\min_{t,\mathbf{y},\mathbf{Z}} t \quad \text{subject to} \quad \mathbf{M}\left(t,\mathbf{y},\mathbf{Z}\right) \succeq 0 \quad \text{and} \quad \mathcal{A}(\mathbf{Z}) = \mathbf{b}. \tag{4.15}$$

For our experiments, the linear mapping is defined as $(\mathcal{A}\left(\mathbf{X}\right))_k = \langle \mathbf{X}, \mathbf{A}_k \rangle$, $k \in [m]$, with independent Gaussian random tensors $\mathbf{A}_k \in \mathbb{R}^{n_1 \times n_2 \times n_3}$. That is, all entries of $\mathbf{A}_k$ are independent $\mathcal{N}\left(0, \frac{1}{m}\right)$ random variables. We choose the rank-one tensors $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ as $\mathbf{X} = \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w}$, where each entry of the vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ is taken independently from the normal distribution $\mathcal{N}\left(0,1\right)$. The rank-two tensors $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ are generated as the sum of two random rank-one tensors. With $\mathcal{A}$ and $\mathbf{X}$ given, we compute $\mathbf{b} = \mathcal{A}(\mathbf{X})$, run the semidefinite program (4.15) and compare its minimizer with the original low-rank tensor $\mathbf{X}$. For a given set of parameters (namely, dimensions $n_1, n_2, n_3$, number of measurements $m$, and rank $r$), we repeat this experiment 200 times and record the empirical success rate of recovering the original tensor, where we say that recovery is successful if the element-wise reconstruction error is at most $10^{-6}$. We use MATLAB (R2008b) for these numerical experiments, including SeDuMi_1.3 for solving the semidefinite programs.

Table 4.4 summarizes the results of our numerical tests for cubic and non-cubic rank-one and rank-two tensors and several choices of the dimensions. Here, the number $m_{\max}$ denotes the maximal number of measurements for which not even one out of 200 generated tensors is recovered and $m_{\min}$ denotes the minimal number of measurements for which all 200 tensors are recovered. The fifth column in Table 4.4 represents the number of independent measurements which are always sufficient for the recovery of a tensor of an arbitrary rank. For illustration, we present the average cpu time (in seconds) for solving the semidefinite programs via SeDuMi_1.3 in the last

| | $\theta$-basis | position $(p,q)$ in the matrix | $M_{pq}$ | Range of $\widetilde{I} = (i,\hat{i},j,\hat{j},k,\hat{k})$ |
|---|---|---|---|---|
| $\mathbf{M}_0$ | $1$ | $(1,1)\,,(2,2)$ | $1$ | |
| $\mathbf{M}_{ijk}$ | $x_{ijk}$ | $(1,f(i,j,k))$ | $1$ | $i \in [n_1]\,, j \in [n_2]\,, k \in [n_3]$ |
| $\mathbf{M}^2_{\widetilde{I}}$ | $x^2_{ijk}$ | $(2,2)$ | $-1$ | |
| | | $(f(i,j,k),f(i,j,k))$ | $1$ | $\{i \in [n_1]\,, j \in [n_2]\,, k \in [n_3]\}$<br>$\setminus \{i=j=k=1\}$ |
| $\mathbf{M}^3_{\widetilde{I}}$ | $x_{i\hat{j}k}x_{ij\hat{k}}$ | $(f(i,j,k),f(i,\hat{j},\hat{k})),(f(i,j,\hat{k}),f(i,\hat{j},k))$ | $1$ | $i \in [n_1]\,, j < \hat{j}, k < \hat{k}$ |
| $\mathbf{M}^4_{\widetilde{I}}$ | $x_{ijk}x_{\hat{i}\hat{j}\hat{k}}$ | $(f(i,j,k),f(\hat{i},\hat{j},\hat{k})),(f(i,\hat{j},k),f(\hat{i},j,\hat{k}))$ | $1$ | |
| | | $(f(i,\hat{j},\hat{k}),f(\hat{i},j,k)),(f(i,j,\hat{k}),f(\hat{i},\hat{j},k))$ | $1$ | $i < \hat{i}, j < \hat{j}, k < \hat{k}$ |
| $\mathbf{M}^5_{\widetilde{I}}$ | $x_{ijk}x_{\hat{i}j\hat{k}}$ | $(f(i,j,k),f(\hat{i},j,\hat{k})),(f(i,j,\hat{k}),f(\hat{i},j,k))$ | $1$ | $i < \hat{i}, j \in [n_2]\,, k < \hat{k}$ |
| $\mathbf{M}^6_{\widetilde{I}}$ | $x_{ijk}x_{\hat{i}\hat{j}k}$ | $(f(i,j,k),f(\hat{i},\hat{j},k)),(f(i,\hat{j},k),f(\hat{i},j,k))$ | $1$ | $i < \hat{i}, j < \hat{j}, k \in [n_3]$ |
| $\mathbf{M}^7_{\widetilde{I}}$ | $x_{\hat{i}jk}x_{ijk}$ | $(f(i,j,k),f(\hat{i},j,k))$ | $1$ | $i < \hat{i}, j \in [n_2]\,, k \in [n_3]$ |
| $\mathbf{M}^8_{\widetilde{I}}$ | $x_{i\hat{j}k}x_{ijk}$ | $(f(i,j,k),f(i,\hat{j},k))$ | $1$ | $i \in [n_1]\,, j < \hat{j}, k \in [n_3]$ |
| $\mathbf{M}^9_{\widetilde{I}}$ | $x_{ij\hat{k}}x_{ijk}$ | $(f(i,j,k),f(i,j,\hat{k}))$ | $1$ | $i \in [n_1]\,, j \in [n_2]\,, k < \hat{k}$ |

Table 4.3. The matrices involved in the definition of the moment matrix $\mathbf{M}\,(t,\mathbf{y},\mathbf{X})$. Due to the symmetry only the upper triangle part of the matrices is specified. The other non-specified entries of the matrices $\mathbf{M} \in \mathbb{R}^{(n_1n_2n_3+1)\times(n_1n_2n_3+1)}$ from the first column are equal to zero. The matrix $\mathbf{M}$ corresponds to the element $g + \mathcal{J}_3$ of the $\theta$-basis specified in the second column. The index $\widetilde{I} = (i,\hat{i},j,\hat{j},k,\hat{k})$ is in the range of the last column. The function $f : \mathbb{N}^3 \to \mathbb{N}$ is defined as $f\,(i,j,k) = (i-1)n_2n_3 + (j-1)n_3 + k + 1$.

column. We remark that no attempt of accelerating the optimization algorithm has been made. This task is left for future research.

| $n_1 \times n_2 \times n_3$ | rank | $m_{\max}$ | $m_{\min}$ | $n_1n_2n_3$ | cpu(sec) |
|---|---|---|---|---|---|
| $2 \times 2 \times 3$ | 1 | 4 | 12 | 12 | 0.1976 |
| $3 \times 3 \times 3$ | 1 | 6 | 19 | 27 | 0.3705 |
| $3 \times 4 \times 5$ | 1 | 11 | 30 | 60 | 6.6600 |
| $4 \times 4 \times 4$ | 1 | 11 | 32 | 64 | 7.2818 |
| $4 \times 5 \times 6$ | 1 | 18 | 42 | 120 | 129.4804 |
| $5 \times 5 \times 5$ | 1 | 18 | 43 | 125 | 138.9040 |
| $3 \times 4 \times 5$ | 2 | 27 | 56 | 60 | 7.5494 |
| $4 \times 4 \times 4$ | 2 | 26 | 56 | 64 | 8.6525 |
| $4 \times 5 \times 6$ | 2 | 41 | 85 | 120 | 192.5787 |

Table 4.4. Numerical results for low-rank tensor recovery in $\mathbb{R}^{n_1\times n_2\times n_3}$.

Except for very small tensor dimensions, we can always recover rank-one (or rank-two) tensors from a number of measurements which is significantly smaller than the dimension of the corresponding tensor space. Therefore, low-rank tensor recovery via $\theta_1$-minimization seems to be a promising approach. Of course, it remains to investigate the recovery performance theoretically.

We remark that we have used standard MATLAB packages for convex optimization to obtain the numerical experiments. To obtain better performance, new optimization methods should be developed specifically to solve our optimization problem, or more generally, to solve the sum-of-squares polynomial problems. We expect this to be possible and the resulting algorithms to give much better performance results since we have shown that in the matrix scenario all theta norms correspond to the matrix nuclear norm. The state-of-the-art algorithms developed for the matrix scenario can compute the matrix nuclear norm and can solve the matrix nuclear norm minimization problem for matrices of large dimensions. The theory developed in this chapter together with the first numerical results should encourage the development into this direction.

CHAPTER 5

# Tensor Iterative Hard Thresholding Algorithm

In this chapter we introduce an iterative approach to low-rank tensor recovery from a small number of linear measurements. In particular, we introduce and analyze several versions of the iterative hard thresholding algorithm (IHT) adapted to tensor formats – namely, to higher order singular value decomposition (HOSVD), tensor train (TT) decomposition, and hierarchical Tucker (HT) decomposition. We provide a partial convergence result for these algorithms based on the assumption that the measurement map satisfies a variant of the restricted isometry property (tensor RIP or TRIP) – similarly to compressive sensing and low-rank matrix recovery. That is, we show that if the measurement map satisfies the TRIP and if the thresholding operator satisfies an additional condition which can not be guaranteed in advance, the TIHT algorithm converges linearly. Since different tensor decompositions induce different notions of tensor rank, the notion of TRIP has to be adapted to the tensor decomposition at hand. Next, we show that partial Fourier maps combined with random sign flips of tensor entries and subgaussian measurement ensembles satisfy TRIP with high probability. Furthermore, for subgaussian measurement ensembles this is true under almost optimal bounds on the number of measurements (optimal up to the log factors) depending on the tensor format. Lastly, we present numerical experiments for low-HOSVD-rank third order tensor recovery with partial Fourier measurements combined with random sign flips of tensor entries, Gaussian measurement ensembles, and tensor completion via tensor IHT algorithms.

The tensor IHT (TIHT) algorithm – similarly to the IHT for compressive sensing and IHT for low-rank matrix recovery – consists of the following two steps. Given measurements $\mathbf{y} = \mathcal{A}(\mathbf{X})$, one starts with an initial tensor $\mathbf{X}_0$ (usually $\mathbf{X}_0 = \mathbf{0}$), and iteratively computes (for $j = 1, 2, \ldots$)

$$\begin{aligned} \mathbf{Y}^j &= \mathbf{X}^j + \mu_j \mathcal{A}^* \left( \mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right), \\ \mathbf{X}^{j+1} &= \mathcal{H}_{\mathbf{r}}\left(\mathbf{Y}^j\right). \end{aligned}$$

The parameter $\mu_j$ is a suitable stepsize parameter and $\mathcal{H}_{\mathbf{r}}(\mathbf{Z})$ returns a rank-$\mathbf{r}$ approximation of a tensor $\mathbf{Z}$ within the given tensor format obtained via successive SVDs (see Subsection 2.1.5). Recall that in every step of the iterative hard thresholding algorithm for compressive sensing and low-rank matrix recovery, one computes the best $s$-sparse and the best rank-$r$ approximation of a given vector and a given matrix, respectively. This fact is exploited to prove the convergence of the IHT algorithms. Unfortunately, obtaining the best rank-$\mathbf{r}$ approximation of a given tensor is in general NP-hard – regardless of the choice of tensor decomposition, see [81, 82]. Nevertheless, $\mathcal{H}_{\mathbf{r}}$ computes a quasi-best approximation in the sense that

$$\|\mathbf{Z} - \mathcal{H}_{\mathbf{r}}(\mathbf{Z})\|_F \leq C(d) \|\mathbf{Z} - \mathbf{Z}_{\text{BEST}}\|_F \quad \text{with } C(d) = \mathcal{O}(\sqrt{d}), \tag{5.1}$$

and $\mathbf{Z}_{\text{BEST}}$ denotes the best rank-$\mathbf{r}$ approximation of $\mathbf{Z}$, see Theorem 2.6 for the exact bound for each of the tensor decompositions.

As already mentioned, our analyses of the TIHT algorithms build on the assumption that the linear operator $\mathcal{A}$ satisfies a variant of the restricted isometry property adapted to the tensor decomposition at hand (HOSVD, TT, or HT decomposition). Our analyses require additionally that at each iteration it holds

$$\left\|\mathbf{Y}^j - \mathbf{X}^{j+1}\right\|_F \leq (1+\varepsilon)\left\|\mathbf{Y}^j - \mathbf{X}_{\mathbf{r}}\right\|_F, \tag{5.2}$$

where $\varepsilon$ is a small non-negative number close to 0 and $\mathbf{X}_{\mathbf{r}}$ is the best rank-$\mathbf{r}$ approximation to $\mathbf{X}$ – the tensor to be recovered (satisfying $\mathbf{y} = \mathcal{A}(\mathbf{X})$), see Theorem 5.4 for details. In particular, if $\mathbf{X}$ is exactly of rank $\mathbf{r}$, then $\mathbf{X}_{\mathbf{r}} = \mathbf{X}$. Unfortunately, (5.1) only guarantees that

$$\left\|\mathbf{Y}^j - \mathbf{X}^{j+1}\right\|_F \leq C(d)\left\|\mathbf{Y}^j - \mathbf{Y}^j_{\text{BEST}}\right\|_F \leq C(d)\left\|\mathbf{Y}^j - \mathbf{X}_{\mathbf{r}}\right\|_F.$$

Since $C(d) = \mathcal{O}(\sqrt{d})$ regardless of the tensor format we consider here, the condition (5.2) cannot be guaranteed a priori. However, our numerical experiments indicate that usually a much better low-rank approximation to $\mathbf{Y}^j$ is computed than the one guaranteed in (5.1). Removing the assumption (5.2) as well as computing efficiently a better rank-$\mathbf{r}$ approximation than $\mathcal{H}_{\mathbf{r}}\left(\mathbf{Y}^j\right)$ seems to be a very difficult, if not impossible, task. As discussed in Chapter 3, there are no completely rigorous results for tensor recovery via efficient algorithms that work for a near optimal number of measurements. However, TIHT algorithm is easy to implement and fast, and additionally, our TRIP bounds give some hints on what the optimal number of measurements – depending on the tensor format – should be.

Another contribution consists in an analysis of TRIP related to the tensor formats HOSVD, TT, and HT for random measurement maps – namely, partial Fourier maps combined with the random sign flips of tensor entries and subgaussian maps. In particular, we show that subgaussian linear maps $\mathcal{A} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^m$ satisfy TRIP at rank $\mathbf{r}$ and level $\delta_{\mathbf{r}}$ with probability exceeding $1-\varepsilon$ provided that

$$\begin{aligned} m &\geq C_1 \delta_{\mathbf{r}}^{-2} \max\left\{\left(r^d + dnr\right)\log(d), \log\left(\varepsilon^{-1}\right)\right\}, \quad \text{for HOSVD,} \\ m &\geq C_2 \delta_{\mathbf{r}}^{-2} \max\left\{\left((d-1)r^3 + dnr\right)\log(dr), \log\left(\varepsilon^{-1}\right)\right\}, \quad \text{for TT and HT,} \end{aligned}$$

where $C_1, C_2 > 0$ are universal constants, $n = \max\{n_i : i \in [d]\}$, $r = \max\{r_t : t \in T_I\}$, and $T_I$ is the corresponding dimensional tree. Up to the logarithmic factors, these bounds match the number of degrees of freedom of a rank-$\mathbf{r}$ tensor in the particular format, and therefore are almost optimal. For linear maps $\mathcal{A} : \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{C}^m$ that are constructed by composing random sign flips of the tensor entries with a $d$-dimensional Fourier transform followed by random subsampling we provide a similar result, see Theorem 5.11 for details.

The following results can be found in our paper [131].

## 5.1. Analysis of iterative hard thresholding

We now pass to our iterative hard thresholding algorithms. For each tensor format (HOSVD, TT, and HT format), we let $\mathcal{H}_{\mathbf{r}}$ be a corresponding low-rank projection operator as described previously. Given measurements $\mathbf{y} = \mathcal{A}(\mathbf{X})$ of a low-rank tensor $\mathbf{X}$, or $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \mathbf{e}$ if the measurements are noisy, the iterative thresholding algorithm starts with an initial guess $\mathbf{X}^0$ (often

$\mathbf{X}^0 = \mathbf{0}$) and performs the iterations

$$\begin{aligned}\mathbf{Y}^j &= \mathbf{X}^j + \mu_j \mathcal{A}^* \left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right), \\ \mathbf{X}^{j+1} &= \mathcal{H}_{\mathbf{r}}(\mathbf{Y}^j).\end{aligned}$$

We analyze two variants of the algorithm which only differ by the choice of the step lengths $\mu_j$.

- **Classical TIHT** (CTIHT) uses simply $\mu_j = 1$, see [10] and Subsection 1.1.1 for the sparse recovery variant and [85] and Subsection 1.2.1 for low-rank matrix recovery variant.
- **Normalized TIHT** (NTIHT) uses

$$\mu_j = \frac{\left\|\mathcal{M}^j\left(\mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right)\right)\right\|_F^2}{\left\|\mathcal{A}\left(\mathcal{M}^j\left(\mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right)\right)\right)\right\|_2^2}, \tag{5.3}$$

see [11] and Subsection 1.1.1 for the sparse vector and [152] and Subsection 1.2.1 for the matrix variant.

Here, the operator $\mathcal{M}^j : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ depends on the choice of the tensor format and is computed via projections onto spaced spanned by left singular vectors of several matricizations of the current iterate $\mathbf{X}^j$. This choice of $\mu_j$ is motivated by the fact that in the sparse vector recovery scenario, the corresponding choice of the step length maximally decreases the residual if the support set does not change in this iteration, see discussion after Theorem 1.21 and [11].

Let us describe the operator $\mathcal{M}^j : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ appearing in (5.3). For the sake of illustration we first specify it for the matrix case, i.e. when $d = 2$. Let $\mathbf{P}^j_{\mathbf{U}_1}$ and $\mathbf{P}^j_{\mathbf{U}_2}$ be the projectors onto the top $r$ left and right singular vector spaces of $\mathbf{X}^j$, respectively. Then $\mathcal{M}^j(\mathbf{Z}) = \mathbf{P}^j_{\mathbf{U}_1}\mathbf{Z}\mathbf{P}^j_{\mathbf{U}_2}$ for a matrix $\mathbf{Z}$ so that (5.3) yields

$$\mu_j = \frac{\left\|\mathbf{P}^j_{\mathbf{U}_1}\mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right)\mathbf{P}^j_{\mathbf{U}_2}\right\|_F^2}{\left\|\mathcal{A}\left(\mathbf{P}^j_{\mathbf{U}_1}\mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right)\mathbf{P}^j_{\mathbf{U}_2}\right)\right\|_2^2}.$$

For the general tensor case, let $\mathbf{U}_{i,j}$ be the left singular vectors of the matricizations $\mathbf{X}^{j\{i\}}$, $\mathbf{X}^{j\{1,\ldots,i\}}$, $\mathbf{X}^{j\,T_I(i)}$ in case of HOSVD, TT, HT decomposition with the corresponding ordered tree $T_I$, respectively. The corresponding projection operators are given as $\mathbf{P}^j_{\mathbf{U}_i} := \hat{\mathbf{U}}_{i,j}\hat{\mathbf{U}}^*_{i,j}$, where $\hat{\mathbf{U}}_{i,j} = \mathbf{U}_{i,j}\left(:, [r_i]\right)$, with $r_i = r_{T_I(i)}$ in the HT case. Then in the case of the HOSVD decomposition we define

$$\mathcal{M}^j\left(\mathbf{Z}\right) = \mathbf{Z} \times_1 \mathbf{P}^j_{\mathbf{U}_1} \times_2 \mathbf{P}^j_{\mathbf{U}_2} \times \cdots \times_d \mathbf{P}^j_{\mathbf{U}_d}. \tag{5.4}$$

In order to define the operator $\mathcal{M}^j$ for the TT decomposition we use the $k$-mode product introduced in Definition 2.1. The TT decomposition of a $d$-th order tensor $\mathbf{Z}$ can be written as

$$\begin{aligned}\mathbf{Z}\left(i_1, i_2, \ldots, i_d\right) &= \mathbf{Z}_1(i_1)\mathbf{Z}_2(i_2)\cdots\mathbf{Z}_d(i_d) \\ &= \mathbf{Z}_d \times_1 \left(\mathbf{Z}_{d-1} \times_1 \left(\cdots \left(\mathbf{Z}_2 \times_1 \mathbf{Z}_1\right)^{\{1,2\}} \cdots\right)^{\{1,2\}}\right)^{\{1,2\}}\left((i_1, i_2, \ldots, i_{d-1}), i_d\right).\end{aligned}$$

Then the operator $\mathcal{M}^j : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ is defined as $\mathcal{M}^j\left(\mathbf{Z}\right) = \mathcal{T}_{\text{vec}}(\hat{\mathbf{Z}})$, where

$$\hat{\mathbf{Z}} := \left(\mathbf{Z}_d \times_1 \mathbf{P}^j_{\mathbf{U}_{d-1}}\left(\mathbf{Z}_{d-1} \times_1 \mathbf{P}^j_{\mathbf{U}_{d-2}}\left(\cdots \mathbf{P}^j_{\mathbf{U}_2}\left(\mathbf{Z}_2 \times_1 \mathbf{P}^j_{\mathbf{U}_1}\mathbf{Z}_1\right)^{\{1,2\}} \cdots\right)^{\{1,2\}}\right)^{\{1,2\}}\right)^{\{1,2\}}$$

and $\mathcal{T}_{\text{vec}}(\mathbf{x}) \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ represents the tensorized version of a vector $\mathbf{x}$. More precisely, the operator $\mathcal{T}_{\text{vec}}$ transforms a vector $\mathbf{x} \in \mathbb{R}^{n_1 n_2 \cdots n_d}$ into a $d$th order tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, i.e., $\mathcal{T}_{\text{vec}}(\text{vec}(\mathbf{X})) = \mathbf{X}$.

Using the general $k$-mode product, one can define the operator $\mathcal{M}^j$ for the general HT-decomposition by applying the above procedure in an analogous way. In the normalized version of the tensor iterative hard thresholding algorithm (NTIHT algorithm), one computes the projection operators $\mathbf{P}^j_{\mathbf{U}_i}$ in each iteration $j$. To accomplish this, the tensor decomposition has to be computed one extra time in each iteration which makes one iteration of the NTIHT algorithm substantially slower in comparison to the CTIHT algorithm. However, we are able to provide better convergence results for the NTIHT than for the CTIHT algorithm.

**Remark 5.1.** For the normalized matrix iterative hard thresholding algorithm three different stepsizes have been introduced – namely, $\mu_j^{\mathbf{U}} := \frac{\left\|\mathbf{P}^j_{\mathbf{U}}\mathcal{A}^*\left(\mathbf{y}-\mathcal{A}\left(\mathbf{X}^j\right)\right)\right\|_F^2}{\left\|\mathcal{A}\left(\mathbf{P}^j_{\mathbf{U}}\mathcal{A}^*(\mathbf{y}-\mathcal{A}(\mathbf{X}^j))\right)\right\|_2^2}$, $\mu_j^{\mathbf{V}} := \frac{\left\|\mathcal{A}^*\left(\mathbf{y}-\mathcal{A}\left(\mathbf{X}^j\right)\right)\mathbf{P}^j_{\mathbf{V}}\right\|_F^2}{\left\|\mathcal{A}\left(\mathcal{A}^*(\mathbf{y}-\mathcal{A}(\mathbf{X}^j))\mathbf{P}^j_{\mathbf{V}}\right)\right\|_2^2}$, and $\mu_j^{\mathbf{UV}} := \frac{\left\|\mathbf{P}^j_{\mathbf{U}}\mathcal{A}^*\left(\mathbf{y}-\mathcal{A}\left(\mathbf{X}^j\right)\right)\mathbf{P}^j_{\mathbf{V}}\right\|_F^2}{\left\|\mathcal{A}\left(\mathbf{P}^j_{\mathbf{U}}\mathcal{A}^*(\mathbf{y}-\mathcal{A}(\mathbf{X}^j))\mathbf{P}^j_{\mathbf{V}}\right)\right\|_2^2}$, where $\mathbf{P}^j_{\mathbf{U}}$ and $\mathbf{P}^j_{\mathbf{V}}$ denote the projection onto the left and right singular spaces of $\mathbf{X}^j$, respectively (see Subsection 1.2.1 for more details). They were motivated by the following three search directions $\mathbf{W}_j^{\mathbf{U}} := \mathbf{P}^j_{\mathbf{U}}\mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right)$, $\mathbf{W}_j^{\mathbf{V}} := \mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right)\mathbf{P}^j_{\mathbf{V}}$, and $\mathbf{W}_j^{\mathbf{UV}} := \mathbf{P}^j_{\mathbf{U}}\mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right)\mathbf{P}^j_{\mathbf{V}}$. Notice that each of these three directions result in a rank $r$ matrix. However, to obtain the rank $\mathbf{r}$ tensor, we need to consider all projections $\mathbf{P}^j_{\mathbf{U}_k}$ in (5.4). As it will be seen later, the fact that $\mathcal{M}^j$ returns a rank $\mathbf{r}$ tensor is also used in the proof of partial convergence for the NTIHT algorithm to provide bounds for $\mu_j$. Consequently, we consider only the stepsize $\mu_j$ defined in (5.3).

The available analysis of the IHT algorithm for recovery of sparse vectors [10] and low-rank matrices [85] is based on the restricted isometry property (RIP). Therefore, we start by introducing an analog for tensors, which we call the tensor restricted isometry property (TRIP). Since different tensor decomposition induce different notions of tensor rank, they also induce different notions of the TRIP.

**Definition 5.2** (TRIP)**.** Let $\mathcal{A}\colon \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^m$ be a measurement map. Then for a fixed tensor decomposition and a corresponding rank tuple $\mathbf{r}$, the tensor restricted isometry constant $\delta_{\mathbf{r}}$ of $\mathcal{A}$ is the smallest quantity such that

$$(1-\delta_{\mathbf{r}})\left\|\mathbf{X}\right\|_F^2 \leq \left\|\mathcal{A}\left(\mathbf{X}\right)\right\|_2^2 \leq (1+\delta_{\mathbf{r}})\left\|\mathbf{X}\right\|_F^2$$

holds for all tensors $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ of rank at most $\mathbf{r}$.

We say that $\mathcal{A}$ satisfies the TRIP at rank $\mathbf{r}$ if $\delta_{\mathbf{r}}$ is bounded by a sufficiently small constant between 0 and 1. When referring to a particular tensor decomposition we use the notions HOSVD-TRIP, TT-TRIP, and HT-TRIP.

The following theorem shows that TRIP is a sufficient condition for low-rank tensor recovery. It is an analogue of Theorem 1.4 for compressive sensing and Theorem 1.25 for low-rank matrix recovery.

**Theorem 5.3.** Let $\mathcal{A} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^m$ be a linear map that satisfies TRIP at rank $2\mathbf{r}$ and level $\delta_{2\mathbf{r}} < 1$. Let $\mathbf{X}_0 \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ be a rank-$\mathbf{r}$ tensor and let $\mathbf{y} := \mathcal{A}\left(\mathbf{X}_0\right)$. Then $\mathbf{X}_0$ is the only rank-$\mathbf{r}$ tensor satisfying $\mathcal{A}\left(\mathbf{X}\right) = \mathbf{y}$.

Proof. We prove the theorem by contradiction. Thus, assume that there exists a rank $\mathbf{r}$ tensor $\mathbf{X}$, $\mathbf{X} \neq \mathbf{X}_0$ satisfying $\mathcal{A}(\mathbf{X}) = \mathbf{y}$. Then $\mathbf{Z} := \mathbf{X}_0 - \mathbf{X} \in \ker(\mathcal{A}) \setminus \{\mathbf{0}\}$ and $\operatorname{rank}(\mathbf{Z}) \leq 2\mathbf{r}$. But then

$$0 = \|\mathcal{A}(\mathbf{Z})\|_2^2 \geq (1 - \delta_{2r}) \|\mathbf{Z}\|_F^2 > 0$$

which is a contradiction. □

Under the TRIP of the measurement operator $\mathcal{A}$, we prove partial convergence results for the two versions of the TIHT algorithm. Depending on some number $a \in (0,1)$, the operator norm and the restricted isometry constants of $\mathcal{A}$, and on the version of TIHT, we define

$$\delta(a) = \begin{cases} \frac{a}{4} & \text{for CTIHT,} \\ \frac{a}{a+8} & \text{for NTIHT,} \end{cases}$$

$$\varepsilon(a) = \begin{cases} \frac{a^2}{17\left(1+\sqrt{1+\delta_{3\mathbf{r}}}\|\mathcal{A}\|_{2\to2}\right)^2} & \text{for CTIHT,} \\ \frac{a^2(1-\delta_{3\mathbf{r}})^2}{17\left(1-\delta_{3\mathbf{r}}+\sqrt{1+\delta_{3\mathbf{r}}}\|\mathcal{A}\|_{2\to2}\right)^2} & \text{for NTIHT,} \end{cases} \tag{5.5}$$

$$b(a) = \begin{cases} 2\sqrt{1+\delta_{3\mathbf{r}}} + \sqrt{4\varepsilon(a) + 2\varepsilon(a)^2}\, \|\mathcal{A}\|_{2\to2} & \text{for CTIHT,} \\ 2\frac{\sqrt{1+\delta_{3\mathbf{r}}}}{1-\delta_{3\mathbf{r}}} + \sqrt{4\varepsilon(a) + 2\varepsilon(a)^2}\frac{1}{1-\delta_{3\mathbf{r}}}\, \|\mathcal{A}\|_{2\to2} & \text{for NTIHT.} \end{cases} \tag{5.6}$$

**Theorem 5.4.** For $a \in (0,1)$, let $\mathcal{A} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^m$ satisfy TRIP (for a fixed tensor format) with

$$\delta_{3\mathbf{r}} < \delta(a)$$

and let $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ be a tensor of rank at most $\mathbf{r}$. Given measurements $\mathbf{y} = \mathcal{A}(\mathbf{X})$, the sequence $(\mathbf{X}^j)_j$ produced by CTIHT or NTIHT converges to $\mathbf{X}$ if

$$\left\|\mathbf{Y}^j - \mathbf{X}^{j+1}\right\|_F \leq (1 + \varepsilon(a)) \left\|\mathbf{Y}^j - \mathbf{X}\right\|_F \quad \text{for all } j = 1, 2, \ldots . \tag{5.7}$$

If the measurements are noisy, $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \mathbf{e}$ for some $\mathbf{e} \in \mathbb{R}^m$, and if (5.7) holds, then

$$\left\|\mathbf{X}^{j+1} - \mathbf{X}\right\|_F \leq a^j \left\|\mathbf{X}^0 - \mathbf{X}\right\|_F + \frac{b(a)}{1-a} \|\mathbf{e}\|_2 \quad \text{for all } j = 1, 2, \ldots . \tag{5.8}$$

Consequently, if $\mathbf{e} \neq \mathbf{0}$ then after at most $j^* := \lceil \log_{1/a} \left( \left\|\mathbf{X}^0 - \mathbf{X}\right\|_F / \|\mathbf{e}\|_2 \right) \rceil$ iterations, $\mathbf{X}^{j^*+1}$ estimates $\mathbf{X}$ with accuracy

$$\left\|\mathbf{X}^{j^*+1} - \mathbf{X}\right\|_F \leq \frac{1 - a + b(a)}{1-a} \|\mathbf{e}\|_2 . \tag{5.9}$$

**Remark 5.5.** (a) The unpleasant part of the theorem is that condition (5.7) cannot be checked. It is implied by the stronger condition

$$\left\|\mathbf{Y}^j - \mathbf{X}^{j+1}\right\|_F \leq (1 + \varepsilon(a)) \left\|\mathbf{Y}^j - \mathbf{Y}^j_{\text{BEST}}\right\|_F ,$$

where $\mathbf{Y}^j_{\text{BEST}}$ is the best rank-$\mathbf{r}$ approximation of $\mathbf{Y}^j$, since the best approximation $\mathbf{Y}^j_{\text{BEST}}$ is by definition a better rank-$\mathbf{r}$ approximation to $\mathbf{Y}^j$ than $\mathbf{X}$. Due to Theorem 2.6 we can only guarantee that this condition holds with $(1+\varepsilon(a))$ replaced by $C(d) \asymp \sqrt{d}$, but the proof of the theorem only works for $(1+\varepsilon(a))$. In fact, $\varepsilon(a)$ is close to 0 as $\|\mathcal{A}\|_{2\to2}$ scales like $\sqrt{n_1 \cdot n_2 \cdots n_d / m}$ for reasonable measurement maps with $\delta_{3\mathbf{r}} < 1$, see below. However, the approximation guarantees for $\mathcal{H}_{\mathbf{r}}$ are only worst case estimates and one may expect that usually much better approximations are computed that satisfy (5.7), which only requires a comparison of the computed approximation error of the Frobenius distance of $\mathbf{Y}^j$ to $\mathbf{X}$ rather than to $\mathbf{Y}^j_{\text{BEST}}$. In fact, during the initial iterations one is usually still far from the original tensor $\mathbf{X}$ so that (5.7) will hold. In any case, the algorithms work in practice so that the theorem may explain why this is the case.

(b) The corresponding theorem [152] (see also Theorem 1.43) for the matrix recovery case applies also to approximately low-rank matrices – not only to exactly low rank matrices – and provides approximation guarantees also for this case. This is in principle also contained in our theorem by splitting $\mathbf{X} = \mathbf{X}_{\text{BEST}} + \mathbf{X}_c$ into the best rank-$\mathbf{r}$ approximation and a remainder term $\mathbf{X}_c$, and writing

$$\mathbf{y} = \mathcal{A}(\mathbf{X}) + \mathbf{e} = \mathcal{A}(\mathbf{X}_{\text{BEST}}) + \mathcal{A}(\mathbf{X}_c) + \mathbf{e} = \mathcal{A}(\mathbf{X}_{\text{BEST}}) + \widetilde{\mathbf{e}},$$

where $\widetilde{\mathbf{e}} = \mathcal{A}(\mathbf{X}_c) + \mathbf{e}$. Then the theorem may be applied to $\widetilde{\mathbf{e}}$ instead of $\mathbf{e}$ and (5.9) gives the error estimate

$$\left\|\mathbf{X}^{j^*+1} - \mathbf{X}_{\text{BEST}}\right\|_F \leq \frac{1-a+b(a)}{1-a}\|\mathcal{A}(\mathbf{X}_c) + \mathbf{e}\|_2.$$

In the matrix case, the right hand side can be further estimated by a sum of three terms (exploiting the restricted isometry property), one of them being the nuclear norm of $\mathbf{X}_c$, i.e., the error of best rank-$r$ approximation in the nuclear norm, see Theorem 1.44. In the tensor case, a similar estimate is problematic, in particular, the analogue of the nuclear norm approximation error is unclear.

(c) In [130] local convergence of a class of algorithms including iterative hard thresholding has been shown. That is, once an iterate $\mathbf{X}^j$ is close enough to the original $\mathbf{X}$ then convergence is guaranteed. (The theorem in [130] requires $\mathcal{H}_{\mathbf{r}}$ to be a retraction on the manifold of rank-$\mathbf{r}$ tensors which is in fact true [99, 145].) Unfortunately, the distance to $\mathbf{X}$ which ensures local convergence depends on the curvature at $\mathbf{X}$ of the manifold of rank-$\mathbf{r}$ tensors and is therefore unknown a-priori. Nevertheless, together with Theorem 5.4, we conclude that the initial iterations decrease the distance to the original $\mathbf{X}$ (if the initial distance is large enough), and if the iterates become sufficiently close to $\mathbf{X}$, then we are guaranteed convergence. The theoretical question remains about the "intermediate phase", i.e., whether the iterates always do come close enough to $\mathbf{X}$ at some point.

(d) In [78], Hedge, Indyk, and Schmidt find a way to deal with approximate projections onto model sets satisfying a relation like $\|\mathbf{Z} - \mathcal{H}_{\mathbf{r}}(\mathbf{Z})\|_F \leq C_d \|\mathbf{Z} - \mathbf{Z}_{\text{BEST}}\|_F$ within iterative hard thresholding algorithms by working with a second approximate projection $\widetilde{\mathcal{H}}_{\mathbf{r}}$ satisfying a *head approximation guarantee* of the form $\|\widetilde{\mathcal{H}}_{\mathbf{r}}(\mathbf{X})\|_F \geq c\|\mathbf{X}\|_F$ for some constant $c > 0$. Unfortunately, we were only able to find such head approximations for the tensor formats at hand with constants $c$ that scale unfavorably with $r$ and the dimensions $n_1, \ldots, n_d$, so that in the end one arrives only at trivial estimates for the minimal number of required measurements.

Proof of Theorem 5.4. We proceed similar to the corresponding proofs for the sparse vector and matrix recovery case, see Theorem 1.17 and Theorem 1.41, respectively. The fact that (5.7) only holds with an additional $\varepsilon = \varepsilon(a)$ requires extra care.

It follows from assumption (5.7) that

$$(1+\varepsilon)^2 \left\|\mathbf{Y}^j - \mathbf{X}\right\|_F^2 \geq \left\|\mathbf{Y}^j - \mathbf{X}^{j+1}\right\|_F^2 = \left\|\mathbf{Y}^j - \mathbf{X} + \mathbf{X} - \mathbf{X}^{j+1}\right\|_F^2$$

$$= \left\|\mathbf{Y}^j - \mathbf{X}\right\|_F^2 + \left\|\mathbf{X} - \mathbf{X}^{j+1}\right\|_F^2 + 2\left\langle \mathbf{Y}^j - \mathbf{X}, \mathbf{X} - \mathbf{X}^{j+1}\right\rangle.$$

Subtracting $\left\|\mathbf{Y}^j - \mathbf{X}\right\|_F^2$ and using $\mathbf{Y}^j = \mathbf{X}^j - \mu_j \mathcal{A}^*\left(\mathcal{A}\left(\mathbf{X}^j\right) - \mathbf{y}\right) = \mathbf{X}^j - \mu_j \mathcal{A}^*\mathcal{A}\left(\mathbf{X}^j - \mathbf{X}\right) + \mu_j \mathcal{A}^*\mathbf{e}$ gives

$$\begin{aligned}
\left\|\mathbf{X}^{j+1} - \mathbf{X}\right\|_F^2 \leq & \, 2\left\langle \mathbf{Y}^j - \mathbf{X}, \mathbf{X}^{j+1} - \mathbf{X}\right\rangle + \left(2\varepsilon + \varepsilon^2\right)\left\|\mathbf{Y}^j - \mathbf{X}\right\|_F^2 \\
= & \, 2\left\langle \mathbf{X}^j - \mathbf{X}, \mathbf{X}^{j+1} - \mathbf{X}\right\rangle - 2\mu_j\left\langle \mathcal{A}^*\mathcal{A}\left(\mathbf{X}^j - \mathbf{X}\right), \mathbf{X}^{j+1} - \mathbf{X}\right\rangle \\
& + 2\mu_j\left\langle \mathcal{A}^*\mathbf{e}, \mathbf{X}^{j+1} - \mathbf{X}\right\rangle + \left(2\varepsilon + \varepsilon^2\right)\left\|\mathbf{Y}^j - \mathbf{X}\right\|_F^2 \\
= & \, 2\left\langle \mathbf{X}^j - \mathbf{X}, \mathbf{X}^{j+1} - \mathbf{X}\right\rangle - 2\mu_j\left\langle \mathcal{A}\left(\mathbf{X}^j - \mathbf{X}\right), \mathcal{A}\left(\mathbf{X}^{j+1} - \mathbf{X}\right)\right\rangle \\
& + 2\mu_j\left\langle \mathbf{e}, \mathcal{A}\left(\mathbf{X}^{j+1} - \mathbf{X}\right)\right\rangle + \left(2\varepsilon + \varepsilon^2\right)\left\|\mathbf{Y}^j - \mathbf{X}\right\|_F^2 \\
\leq & \, 2\left\langle \mathbf{X}^j - \mathbf{X}, \mathbf{X}^{j+1} - \mathbf{X}\right\rangle - 2\mu_j\left\langle \mathcal{A}\left(\mathbf{X}^j - \mathbf{X}\right), \mathcal{A}\left(\mathbf{X}^{j+1} - \mathbf{X}\right)\right\rangle \\
& + 2\mu_j\sqrt{1 + \delta_{3\mathbf{r}}}\left\|\mathbf{X}^{j+1} - \mathbf{X}\right\|_F \left\|\mathbf{e}\right\|_2 + \left(2\varepsilon + \varepsilon^2\right)\left\|\mathbf{Y}^j - \mathbf{X}\right\|_F^2, \qquad (5.10)
\end{aligned}$$

where the last inequality is valid since $\operatorname{rank}\left(\mathbf{X}^{j+1} - \mathbf{X}\right) \leq 2\mathbf{r} \leq 3\mathbf{r}$ so that

$$\left\langle \mathbf{e}, \mathcal{A}\left(\mathbf{X}^{j+1} - \mathbf{X}\right)\right\rangle \leq \left\|\mathcal{A}\left(\mathbf{X}^{j+1} - \mathbf{X}\right)\right\|_2 \left\|\mathbf{e}\right\|_2 \leq \sqrt{1 + \delta_{3\mathbf{r}}}\left\|\mathbf{X}^{j+1} - \mathbf{X}\right\|_F \left\|\mathbf{e}\right\|_2.$$

Now let $U^j$ be the subspace of $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ spanned by the tensors $\mathbf{X}$, $\mathbf{X}^j$, and $\mathbf{X}^{j+1}$ and denote by $\mathcal{Q}^j : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to U^j$ the orthogonal projection onto $U^j$. Then $\mathcal{Q}^j\left(\mathbf{X}\right) = \mathbf{X}$, $\mathcal{Q}^j\left(\mathbf{X}^j\right) = \mathbf{X}^j$, and $\mathcal{Q}^j\left(\mathbf{X}^{j+1}\right) = \mathbf{X}^{j+1}$. Clearly, the rank of $\mathcal{Q}^j\left(\mathbf{Y}\right)$ is at most $3\mathbf{r}$ for all $\mathbf{Y} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$. Further, we define the operator $\mathcal{A}^j_{\mathbf{Q}} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ by $\mathcal{A}^j_{\mathbf{Q}}\left(\mathbf{Z}\right) = \mathcal{A}\left(\mathcal{Q}^j\left(\mathbf{Z}\right)\right)$ for $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$.

With these notions the estimate (5.10) is continued as

$$\begin{aligned}
\left\|\mathbf{X}^{j+1} - \mathbf{X}\right\|_F^2 \leq & \, 2\left\langle \mathbf{X}^j - \mathbf{X}, \mathbf{X}^{j+1} - \mathbf{X}\right\rangle - 2\mu_j\left\langle \mathcal{A}^j_{\mathbf{Q}}\left(\mathbf{X}^j - \mathbf{X}\right), \mathcal{A}^j_{\mathbf{Q}}\left(\mathbf{X}^{j+1} - \mathbf{X}\right)\right\rangle \\
& + 2\mu_j\sqrt{1 + \delta_{3\mathbf{r}}}\left\|\mathbf{X}^{j+1} - \mathbf{X}\right\|_F \left\|\mathbf{e}\right\|_2 + \left(2\varepsilon + \varepsilon^2\right)\left\|\mathbf{Y}^j - \mathbf{X}\right\|_F^2 \\
= & \, 2\left\langle \mathbf{X}^j - \mathbf{X}, \left(\mathbf{X}^{j+1} - \mathbf{X}\right) - \mu_j \mathcal{A}^{j*}_{\mathbf{Q}}\mathcal{A}^j_{\mathbf{Q}}\left(\mathbf{X}^{j+1} - \mathbf{X}\right)\right\rangle \\
& + 2\mu_j\sqrt{1 + \delta_{3\mathbf{r}}}\left\|\mathbf{X}^{j+1} - \mathbf{X}\right\|_F \left\|\mathbf{e}\right\|_2 + \left(2\varepsilon + \varepsilon^2\right)\left\|\mathbf{Y}^j - \mathbf{X}\right\|_F^2 \\
= & \, 2\left\langle \mathbf{X}^j - \mathbf{X}, \left(\mathbf{I} - \mu_j \mathcal{A}^{j*}_{\mathbf{Q}}\mathcal{A}^j_{\mathbf{Q}}\right)\left(\mathbf{X}^{j+1} - \mathbf{X}\right)\right\rangle \\
& + 2\mu_j\sqrt{1 + \delta_{3\mathbf{r}}}\left\|\mathbf{X}^{j+1} - \mathbf{X}\right\|_F \left\|\mathbf{e}\right\|_2 + \left(2\varepsilon + \varepsilon^2\right)\left\|\mathbf{Y}^j - \mathbf{X}\right\|_F^2 \\
\leq & \, 2\left\|\mathbf{I} - \mu_j \mathcal{A}^{j*}_{\mathbf{Q}}\mathcal{A}^j_{\mathbf{Q}}\right\|_{2\to 2}\left\|\mathbf{X}^j - \mathbf{X}\right\|_F \left\|\mathbf{X}^{j+1} - \mathbf{X}\right\|_F \\
& + 2\mu_j\sqrt{1 + \delta_{3\mathbf{r}}}\left\|\mathbf{X}^{j+1} - \mathbf{X}\right\|_F \left\|\mathbf{e}\right\|_2 + \left(2\varepsilon + \varepsilon^2\right)\left\|\mathbf{Y}^j - \mathbf{X}\right\|_F^2. \qquad (5.11)
\end{aligned}$$

The last term can be bounded by

$$\begin{aligned}
\left\|\mathbf{Y}^j - \mathbf{X}\right\|_F = & \left\|\mathbf{X}^j - \mu_j \mathcal{A}^*\mathcal{A}\left(\mathbf{X}^j - \mathbf{X}\right) + \mu_j \mathcal{A}^*\mathbf{e} - \mathbf{X}\right\|_F \\
= & \left\|\left(\mathbf{X}^j - \mathbf{X}\right) - \mu_j \mathcal{A}^*\mathcal{A}\left(\mathbf{X}^j - \mathbf{X}\right) + \mu_j \mathcal{A}^*\mathbf{e}\right\|_F \\
= & \left\|\left(\mathbf{I} - \mu_j \mathcal{A}^*\mathcal{A}\right)\left(\mathbf{X}^j - \mathbf{X}\right) + \mu_j \mathcal{A}^*\mathbf{e}\right\|_F = \left\|\left(\mathbf{I} - \mu_j \mathcal{A}^*\mathcal{A}^j_{\mathbf{Q}}\right)\left(\mathbf{X}^j - \mathbf{X}\right) + \mu_j \mathcal{A}^*\mathbf{e}\right\|_F \\
\leq & \left\|\mathbf{I} - \mu_j \mathcal{A}^*\mathcal{A}^j_{\mathbf{Q}}\right\|_{2\to 2}\left\|\mathbf{X}^j - \mathbf{X}\right\|_F + \mu_j\left\|\mathcal{A}^*\mathbf{e}\right\|_F \\
\leq & \left(1 + \mu_j\left\|\mathcal{A}\right\|_{2\to 2}\left\|\mathcal{A}^j_{\mathbf{Q}}\right\|_{2\to 2}\right)\left\|\mathbf{X}^j - \mathbf{X}\right\|_F + \mu_j\left\|\mathcal{A}\right\|_{2\to 2}\left\|\mathbf{e}\right\|_2
\end{aligned}$$

$$\leq \left(1+\mu_j\sqrt{1+\delta_{3\mathbf{r}}}\,\|\mathcal{A}\|_{2\to 2}\right)\left\|\mathbf{X}^j-\mathbf{X}\right\|_F+\mu_j\,\|\mathcal{A}\|_{2\to 2}\,\|\mathbf{e}\|_2\,. \tag{5.12}$$

Using that $(u+v)^2 \leq 2\left(u^2+v^2\right)$ for all $u, v \in \mathbb{R}$, we obtain the estimate

$$\left\|\mathbf{Y}^j-\mathbf{X}\right\|_F^2 \leq 2\left(1+\mu_j\sqrt{1+\delta_{3\mathbf{r}}}\,\|\mathcal{A}\|_{2\to 2}\right)^2\left\|\mathbf{X}^j-\mathbf{X}\right\|_F^2+2\mu_j^2\,\|\mathcal{A}\|_{2\to 2}^2\,\|\mathbf{e}\|_2^2\,. \tag{5.13}$$

Combining inequalities (5.11) and (5.13) yields

$$\begin{aligned}
&\left\|\mathbf{X}^{j+1}-\mathbf{X}\right\|_F^2 \leq 2\left\|\mathbf{I}-\mu_j\mathcal{A}_{\mathbf{Q}}^{j*}\mathcal{A}_{\mathbf{Q}}^{j}\right\|_{2\to 2}\left\|\mathbf{X}^j-\mathbf{X}\right\|_F\left\|\mathbf{X}^{j+1}-\mathbf{X}\right\|_F\\
&+2\mu_j\sqrt{1+\delta_{3\mathbf{r}}}\left\|\mathbf{X}^{j+1}-\mathbf{X}\right\|_F\|\mathbf{e}\|_2+2\left(2\varepsilon+\varepsilon^2\right)\left(1+\mu_j\sqrt{1+\delta_{3\mathbf{r}}}\,\|\mathcal{A}\|_{2\to 2}\right)^2\left\|\mathbf{X}^j-\mathbf{X}\right\|_F^2\\
&+2\left(2\varepsilon+\varepsilon^2\right)\mu_j^2\,\|\mathcal{A}\|_{2\to 2}^2\,\|\mathbf{e}\|_2^2\,.
\end{aligned}$$

This implies that there exist $\alpha, \beta, \gamma \in [0,1]$ such that $\alpha+\beta+\gamma \leq 1$ and

$$(1-\alpha-\beta-\gamma)\left\|\mathbf{X}^{j+1}-\mathbf{X}\right\|_F^2 \leq 2\left\|\mathbf{I}-\mu_j\mathcal{A}_{\mathbf{Q}}^{j*}\mathcal{A}_{\mathbf{Q}}^{j}\right\|_{2\to 2}\left\|\mathbf{X}^j-\mathbf{X}\right\|_F\left\|\mathbf{X}^{j+1}-\mathbf{X}\right\|_F \tag{5.14}$$

$$\alpha\left\|\mathbf{X}^{j+1}-\mathbf{X}\right\|_F^2 \leq 2\mu_j\sqrt{1+\delta_{3\mathbf{r}}}\left\|\mathbf{X}^{j+1}-\mathbf{X}\right\|_F\|\mathbf{e}\|_2 \tag{5.15}$$

$$\beta\left\|\mathbf{X}^{j+1}-\mathbf{X}\right\|_F^2 \leq 2\left(2\varepsilon+\varepsilon^2\right)\left(1+\mu_j\sqrt{1+\delta_{3\mathbf{r}}}\,\|\mathcal{A}\|_{2\to 2}\right)^2\left\|\mathbf{X}^j-\mathbf{X}\right\|_F^2, \tag{5.16}$$

$$\gamma\left\|\mathbf{X}^{j+1}-\mathbf{X}\right\|_F^2 \leq 2\left(2\varepsilon+\varepsilon^2\right)\mu_j^2\,\|\mathcal{A}\|_{2\to 2}^2\,\|\mathbf{e}\|_2^2\,. \tag{5.17}$$

Canceling one power of $\left\|\mathbf{X}^{j+1}-\mathbf{X}\right\|_F$ in inequalities (5.14) and (5.15), taking the square root of the inequalities (5.16) and (5.17), and summation of all resulting inequalities yields

$$\begin{aligned}
&\left\|\mathbf{X}^{j+1}-\mathbf{X}\right\|_F\\
&\leq f(\beta,\gamma)\left(2\left\|\mathbf{I}-\mu_j\mathcal{A}_{\mathbf{Q}}^{j*}\mathcal{A}_{\mathbf{Q}}^{j}\right\|_{2\to 2}+\sqrt{4\varepsilon+2\varepsilon^2}\left(1+\mu_j\sqrt{1+\delta_{3\mathbf{r}}}\,\|\mathcal{A}\|_{2\to 2}\right)\right)\left\|\mathbf{X}^j-\mathbf{X}\right\|_F\\
&+f(\beta,\gamma)\left(2\mu_j\sqrt{1+\delta_{3\mathbf{r}}}+\sqrt{4\varepsilon+2\varepsilon^2}\mu_j\,\|\mathcal{A}\|_{2\to 2}\right)\|\mathbf{e}\|_2
\end{aligned} \tag{5.18}$$

with $f(\beta,\gamma)=(1-\beta+\sqrt{\beta}-\gamma+\sqrt{\gamma})^{-1}$. Notice that $f$ is positive and strictly less than 1 on $[0,1]\times[0,1]$ and will therefore be omitted in the following.

Let us now specialize to CTIHT where $\mu_j = 1$. Since $\mathcal{A}_{\mathbf{Q}}^{j}$ is the restriction of $\mathcal{A}$ to the space $U^j$ which contains only tensors of rank at most $3\mathbf{r}$, we have (with $\mathbf{I}$ denoting the identity operator on $U^j$)

$$\begin{aligned}
\|\mathbf{I}-\mu_j\mathcal{A}_{\mathbf{Q}}^{j*}\mathcal{A}_{\mathbf{Q}}^{j}\|_{2\to 2} = \|\mathbf{I}-\ \mathcal{A}_{\mathbf{Q}}^{j*}\mathcal{A}_{\mathbf{Q}}^{j}\|_{2\to 2} &= \sup_{\mathbf{X}\in U_j:\|\mathbf{X}\|_F=1}\left|\|\mathbf{X}\|_F^2-\|\mathcal{A}(\mathbf{X})\|_2^2\right|\\
&\leq \sup_{\mathbf{X}:\operatorname{rank}(\mathbf{X})\leq 3\mathbf{r},\|\mathbf{X}\|_F=1}\left|\|\mathbf{X}\|_F^2-\|\mathcal{A}(\mathbf{X})\|_2^2\right| = \delta_{3\mathbf{r}}.
\end{aligned}$$

Plugging $\mu_j = 1$ and above estimate into (5.18) yields

$$\begin{aligned}
\left\|\mathbf{X}^{j+1}-\mathbf{X}\right\|_F &\leq \left(2\left\|\mathbf{I}-\mathcal{A}_{\mathbf{Q}}^{j*}\mathcal{A}_{\mathbf{Q}}^{j}\right\|_{2\to 2}+\sqrt{4\varepsilon+2\varepsilon^2}\left(1+\sqrt{1+\delta_{3\mathbf{r}}}\,\|\mathcal{A}\|_{2\to 2}\right)\right)\left\|\mathbf{X}^j-\mathbf{X}\right\|_F\\
&\quad+\left(2\sqrt{1+\delta_{3\mathbf{r}}}+\sqrt{4\varepsilon+2\varepsilon^2}\,\|\mathcal{A}\|_{2\to 2}\right)\|\mathbf{e}\|_2\\
&\leq \left(2\delta_{3\mathbf{r}}+\sqrt{4\varepsilon+2\varepsilon^2}\left(1+\sqrt{1+\delta_{3\mathbf{r}}}\,\|\mathcal{A}\|_{2\to 2}\right)\right)\left\|\mathbf{X}^j-\mathbf{X}\right\|_F\\
&\quad+\left(2\sqrt{1+\delta_{3\mathbf{r}}}+\sqrt{4\varepsilon+2\varepsilon^2}\,\|\mathcal{A}\|_{2\to 2}\right)\|\mathbf{e}\|_2\,.
\end{aligned}$$

Setting $\kappa := 1 + \sqrt{1+\delta_{3\mathbf{r}}}\|\mathcal{A}\|_{2\to 2} > 1$, the bound $\delta_{3\mathbf{r}} \leq a/4$ with $a < 1$ and the definition of $\varepsilon = \varepsilon(a)$ in (5.5) yield

$$2\delta_{3\mathbf{r}} + \sqrt{4\varepsilon + 2\varepsilon^2}\left(1 + \sqrt{1+\delta_{3\mathbf{r}}}\left\|\mathcal{A}\right\|_{2\to 2}\right) \leq \frac{a}{2} + \sqrt{\frac{4a^2}{17\kappa^2} + \frac{2a^4}{17^2\kappa^4}}\kappa \leq a\left(\frac{1}{2} + \sqrt{\frac{4}{17} + \frac{2}{17^2}}\right) < a.$$

Thus, with the definition (5.6) of $b = b(a)$ for CTIHT we obtain

$$\left\|\mathbf{X}^{j+1} - \mathbf{X}\right\|_F \leq a\left\|\mathbf{X}^j - \mathbf{X}\right\|_F + b\left\|\mathbf{e}\right\|_2.$$

Iterating this inequality leads to (5.8), which implies a recovery accuracy of $\left\|\mathbf{X}^{j+1} - \mathbf{X}\right\|_F \leq \frac{1-a+b}{1-a}\left\|\mathbf{e}\right\|_2$ if $a^j\left\|\mathbf{X}^0 - \mathbf{X}\right\|_F \leq \left\|\mathbf{e}\right\|_2$. Hence, if $\mathbf{e} \neq \mathbf{0}$ then after $j^* := \lceil \log_{1/a}\left(\left\|\mathbf{X}^0 - \mathbf{X}\right\|_F / \left\|\mathbf{e}\right\|_2\right)\rceil$ iterations, (5.9) holds.

Let us now consider the variant NTIHT. Since the image of the operator $\mathcal{M}^j$ is contained in the set of rank-$\mathbf{r}$ tensors, the tensor restricted isometry property yields

$$\frac{1}{1+\delta_{\mathbf{r}}} \leq \mu_j = \frac{\left\|\mathcal{M}^j\left(\mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right)\right)\right\|_F^2}{\left\|\mathcal{A}\left(\mathcal{M}^j\left(\mathcal{A}^*\left(\mathbf{y} - \mathcal{A}\left(\mathbf{X}^j\right)\right)\right)\right)\right\|_2^2} \leq \frac{1}{1-\delta_{\mathbf{r}}}. \tag{5.19}$$

Since $\mathcal{Q}^j$ maps onto rank-$3\mathbf{r}$ tensors, the TRIP implies that every eigenvalue of $\mathcal{A}_{\mathbf{Q}}^{j*}\mathcal{A}_{\mathbf{Q}}^{j}$ is contained in the interval $[1-\delta_{3\mathbf{r}}, 1+\delta_{3\mathbf{r}}]$. Therefore, every eigenvalue of $\mathbf{I} - \mu_j\mathcal{A}_{\mathbf{Q}}^{j*}\mathcal{A}_{\mathbf{Q}}^{j}$ is contained in $[1 - \frac{1+\delta_{3\mathbf{r}}}{1-\delta_{\mathbf{r}}}, 1 - \frac{1-\delta_{3\mathbf{r}}}{1+\delta_{\mathbf{r}}}]$. The magnitude of the lower end point is greater than that of the upper end point, giving the operator norm bound

$$\left\|\mathbf{I} - \mu_j\mathcal{A}_{\mathbf{Q}}^{j*}\mathcal{A}_{\mathbf{Q}}^{j}\right\|_{2\to 2} \leq \frac{1+\delta_{3\mathbf{r}}}{1-\delta_{\mathbf{r}}} - 1 \leq \frac{1+\delta_{3\mathbf{r}}}{1-\delta_{3\mathbf{r}}} - 1.$$

Hence, plugging the upper bound on $\mu_j$ in (5.19) and the above inequality into (5.18) leads to

$$\begin{aligned}\left\|\mathbf{X}^{j+1} - \mathbf{X}\right\|_F &\leq \left(2\left\|\mathbf{I} - \mu_j\mathcal{A}_{\mathbf{Q}}^{j*}\mathcal{A}_{\mathbf{Q}}^{j}\right\|_{2\to 2} + \sqrt{4\varepsilon + 2\varepsilon^2}\left(1 + \mu_j\sqrt{1+\delta_{3\mathbf{r}}}\left\|\mathcal{A}\right\|_{2\to 2}\right)\right)\left\|\mathbf{X}^j - \mathbf{X}\right\|_F \\ &\quad + \left(2\mu_j\sqrt{1+\delta_{3\mathbf{r}}} + \sqrt{4\varepsilon + 2\varepsilon^2}\mu_j\left\|\mathcal{A}\right\|_{2\to 2}\right)\left\|\mathbf{e}\right\|_2 \\ &\leq \left(2\left(\frac{1+\delta_{3\mathbf{r}}}{1-\delta_{3\mathbf{r}}} - 1\right) + \sqrt{4\varepsilon + 2\varepsilon^2}\left(1 + \frac{\sqrt{1+\delta_{3\mathbf{r}}}}{1-\delta_{3\mathbf{r}}}\left\|\mathcal{A}\right\|_{2\to 2}\right)\right)\left\|\mathbf{X}^j - \mathbf{X}\right\|_F \\ &\quad + \left(2\frac{\sqrt{1+\delta_{3\mathbf{r}}}}{1-\delta_{3\mathbf{r}}} + \frac{\sqrt{4\varepsilon + 2\varepsilon^2}}{1-\delta_{3\mathbf{r}}}\left\|\mathcal{A}\right\|_{2\to 2}\right)\left\|\mathbf{e}\right\|_2.\end{aligned}$$

Setting $\nu := 1 + \frac{\sqrt{1+\delta_{3\mathbf{r}}}}{1-\delta_{3\mathbf{r}}}\|\mathcal{A}\|_{2\to 2} \geq 1$, using $\delta_{3\mathbf{r}} \leq a/(a+8)$ and the definition (5.5) of $\varepsilon = \varepsilon(a) = a^2/(17\nu^2)$, gives

$$2\left(\frac{1+\delta_{3\mathbf{r}}}{1-\delta_{3\mathbf{r}}} - 1\right) + \sqrt{4\varepsilon + 2\varepsilon^2}\left(1 + \frac{\sqrt{1+\delta_{3\mathbf{r}}}}{1-\delta_{3\mathbf{r}}}\left\|\mathcal{A}\right\|_{2\to 2}\right) \leq \frac{a}{2} + \nu\sqrt{\frac{4a^2}{17\nu^2} + \frac{2a^2}{17^2\nu^4}} < a$$

so that with the definition of $b(a)$ in (5.6) we arrive at

$$\left\|\mathbf{X}^{j+1} - \mathbf{X}\right\|_F \leq a\left\|\mathbf{X}^j - \mathbf{X}\right\|_F + b(a)\left\|\mathbf{e}\right\|_2.$$

The proof is concluded in the same way as for CTIHT. □

**Remark 5.6.** For the noiseless scenario where $\left\|\mathbf{e}\right\|_2 = 0$, one may work with a slightly improved definition of $\varepsilon(a)$. In fact, (5.12) implies then

$$\left\|\mathbf{Y}^j - \mathbf{X}\right\|_F \leq \left(1 + \mu_j\sqrt{1+\delta_{3\mathbf{r}}}\left\|\mathcal{A}\right\|_{2\to 2}\right)\left\|\mathbf{X}^j - \mathbf{X}\right\|_F.$$

Following the proof in the same way as above, one finds that the constant 17 in the definition (5.5) of $\varepsilon(a)$ can be improved to 9.

## 5.2. Tensor RIP

Now that we have shown a (partial) convergence result for the TIHT algorithm based on TRIP, the question arises which measurement maps satisfy TRIP under suitable conditions on the number of measurements in terms of the rank $\mathbf{r}$, the order $d$ and the dimensions $n_1, \ldots, n_d$. As common in compressive sensing and low-rank matrix recovery, we study this question for random measurement maps. We concentrate first on subgaussian measurement maps and consider maps based on partial random Fourier transform afterwards.

A random variable $X$ is called $L$-subgaussian if there exists a constant $L > 0$ such that

$$\mathbb{E}\left[\exp\left(tX\right)\right] \leq \exp\left(L^2 t^2/2\right)$$

holds for all $t \in \mathbb{R}$. We call $\mathcal{A} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^m$ an $L$-subgaussian measurement ensemble if all elements of $\mathcal{A}$, interpreted as a tensor in $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d \times m}$, are independent mean-zero, variance one, $L$-subgaussian variables. Gaussian and Bernoulli random measurement ensembles where the entries are standard normal distributed random variables and Rademacher ±1 variables (i.e., taking the values +1 and −1 with equal probability), respectively, are special cases of 1-subgaussian measurement ensembles.

**Theorem 5.7.** Fix one of the tensor formats HOSVD, TT, HT (with decomposition tree $T_I$). For $\delta, \varepsilon \in (0,1)$, a random draw of an $L$-subgaussian measurement ensemble $\mathcal{A} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^m$ satisfies $\delta_{\mathbf{r}} \leq \delta$ with probability at least $1 - \varepsilon$ provided that

$$\begin{aligned} \text{HOSVD:} \quad & m \geq C_1 \delta^{-2} \max\left\{\left(r^d + dnr\right)\log\left(d\right), \log\left(\varepsilon^{-1}\right)\right\}, \\ \text{TT \& HT:} \quad & m \geq C_2 \delta^{-2} \max\left\{\left((d-1)r^3 + dnr\right)\log\left(dr\right), \log\left(\varepsilon^{-1}\right)\right\}, \end{aligned}$$

where $n = \max\{n_i : i \in [d]\}$, $r = \max\{r_t : t \in T_I\}$. The constants $C_1, C_2, C_3 > 0$ only depend on the subgaussian parameter $L$.

One may generalize the above theorem to situations where it is no longer required that all entries of the tensor $\mathcal{A}$ are independent, but only that the sensing tensors $\mathbf{A}_i \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, $i = 1, \ldots, m$, are independent. We refer to [48] for details, in particular to Corollary 5.4 and Example 5.8. Furthermore, we note that the term $dnr$ in all bounds for $m$ may be refined to $\sum_{i=1}^d n_i r_i$.

The proof of Theorem 5.7 uses $\varepsilon$-nets and covering numbers, see Subsection A.3 and [163] for more background on this topic. To recall the notation, an $\varepsilon$-net of the set $\boldsymbol{\mathcal{X}}$ with respect to the norm $\|\cdot\|$ is denoted by $\mathcal{N}_\varepsilon^{\boldsymbol{\mathcal{X}}}$. The covering number of $\boldsymbol{\mathcal{X}}$ (at scale $\varepsilon$) is denoted by $\mathcal{N}\left(\boldsymbol{\mathcal{X}}, \|\cdot\|, \varepsilon\right)$.

It is crucial for the proof of Theorem 5.7 to estimate the covering numbers of the set of unit Frobenius norm rank-$\mathbf{r}$ tensors with respect to the different tensor formats. We start with the HOSVD.

**Lemma 5.8** (Covering numbers related to HOSVD)**.** The covering numbers of

$$\boldsymbol{\mathcal{S}}_{\mathbf{r}} = \left\{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} : \operatorname{rank}_{\text{HOSVD}}\left(\mathbf{X}\right) \leq \mathbf{r}, \|\mathbf{X}\|_F = 1\right\}$$

with respect to the Frobenius norm satisfy

$$\mathcal{N}\left(\boldsymbol{\mathcal{S}}_{\mathbf{r}}, \|\cdot\|_F, \varepsilon\right) \leq \left(3\left(d+1\right)/\varepsilon\right)^{r_1 r_2 \cdots r_d + \sum_{i=1}^{d} n_i r_i}. \tag{5.20}$$

Proof. The proof follows a similar strategy as the one of [22, Lemma 3.1]. The HOSVD decomposition $\mathbf{X} = \mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times \cdots \times_d \mathbf{U}_d$ of any $\mathbf{X} \in \boldsymbol{\mathcal{S}}_{\mathbf{r}}$ obeys $\|\mathbf{S}\|_F = 1$. Our argument constructs an $\varepsilon$-net for $\boldsymbol{\mathcal{S}}_{\mathbf{r}}$ by covering the sets of matrices $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_d$ with orthonormal columns and the set of unit Frobenius norm tensors $\mathbf{S}$. For simplicity we assume that $n_1 = n_2 = \ldots = n_d = n$ and $r_1 = r_2 = \ldots = r_d = r$ since the general case requires only a straightforward modification.

The set $\boldsymbol{\mathcal{D}}$ of all-orthogonal $d$-th order tensors $\mathbf{X} \in \mathbb{R}^{r \times r \times \cdots \times r}$ with unit Frobenius norm is contained in $\boldsymbol{\mathcal{F}} = \{\mathbf{X} \in \mathbb{R}^{r \times r \times \cdots \times r} : \|\mathbf{X}\|_F = 1\}$. Lemma A.8 therefore provides an $\varepsilon/\left(d+1\right)$-net $\mathcal{N}^{\boldsymbol{\mathcal{F}}}_{\varepsilon/(d+1)}$ with respect to the Frobenius norm of cardinality $\left|\mathcal{N}^{\boldsymbol{\mathcal{F}}}_{\varepsilon/(d+1)}\right| \leq \left(3\left(d+1\right)/\varepsilon\right)^{r^d}$. For covering $\boldsymbol{\mathcal{O}}_{n,r} = \{\mathbf{U} \in \mathbb{R}^{n \times r} : \mathbf{U}^*\mathbf{U} = \mathbf{I}\}$, it is beneficial to use the norm $\|\cdot\|_{1,2}$ defined as

$$\|\mathbf{V}\|_{1,2} = \max_i \|\mathbf{V}\left(:, i\right)\|_2,$$

where $\mathbf{V}\left(:, i\right)$ denotes the $i$-th column of $\mathbf{V}$. Since the elements of $\boldsymbol{\mathcal{O}}_{n,r}$ have normed columns, it holds $\boldsymbol{\mathcal{O}}_{n,r} \subset \boldsymbol{\mathcal{Q}}_{n,r} = \left\{\mathbf{V} \in \mathbb{R}^{n \times r} : \|\mathbf{V}\|_{1,2} \leq 1\right\}$. Lemma A.8 gives $\mathcal{N}\left(\boldsymbol{\mathcal{O}}_{n,r}, \|\cdot\|_{1,2}, \varepsilon/\left(d+1\right)\right) \leq \left(3\left(d+1\right)/\varepsilon\right)^{nr}$, i.e., there exists an $\varepsilon/\left(d+1\right)$-net $\mathcal{N}^{\boldsymbol{\mathcal{O}}_{n,r}}_{\varepsilon/(d+1)}$ of this cardinality.

Then the set

$$\mathcal{N}^{\boldsymbol{\mathcal{S}}_{\mathbf{r}}}_{\varepsilon} := \left\{\overline{\mathbf{S}} \times_1 \overline{\mathbf{U}}_1 \times_2 \overline{\mathbf{U}}_2 \times \cdots \times_d \overline{\mathbf{U}}_d : \overline{\mathbf{S}} \in \mathcal{N}^{\boldsymbol{\mathcal{D}}}_{\varepsilon/(d+1)} \text{ and } \overline{\mathbf{U}}_i \in \mathcal{N}^{\boldsymbol{\mathcal{O}}_{n,r}}_{\varepsilon/(d+1)} \text{ for all } i \in [d]\right\},$$

obeys

$$\left|\mathcal{N}^{\boldsymbol{\mathcal{S}}_{\mathbf{r}}}_{\varepsilon}\right| \leq \mathcal{N}\left(\boldsymbol{\mathcal{D}}, \|\cdot\|_F, \varepsilon/\left(d+1\right)\right) \left[\mathcal{N}\left(\boldsymbol{\mathcal{O}}_{n,r}, \|\cdot\|_{1,2}, \varepsilon/\left(d+1\right)\right)\right]^d \leq \left(3\left(d+1\right)/\varepsilon\right)^{r^d + dnr}.$$

It remains to show that $\mathcal{N}^{\boldsymbol{\mathcal{S}}_{\mathbf{r}}}_{\varepsilon}$ is an $\varepsilon$-net for $\boldsymbol{\mathcal{S}}_{\mathbf{r}}$, i.e., that for all $\mathbf{X} \in \boldsymbol{\mathcal{S}}_{\mathbf{r}}$ there exists $\overline{\mathbf{X}} \in \mathcal{N}^{\boldsymbol{\mathcal{S}}_{\mathbf{r}}}_{\varepsilon}$ with $\left\|\mathbf{X} - \overline{\mathbf{X}}\right\|_F \leq \varepsilon$. To this end, we fix $\mathbf{X} \in \boldsymbol{\mathcal{S}}_{\mathbf{r}}$ and decompose $\mathbf{X}$ as $\mathbf{X} = \mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times \cdots \times_d \mathbf{U}_d$. Then there exists $\overline{\mathbf{X}} = \overline{\mathbf{S}} \times_1 \overline{\mathbf{U}}_1 \times_2 \overline{\mathbf{U}}_2 \times \cdots \times_d \overline{\mathbf{U}}_d \in \mathcal{N}^{\boldsymbol{\mathcal{S}}_{\mathbf{r}}}_{\varepsilon}$ with $\overline{\mathbf{U}}_i \in \mathcal{N}^{\boldsymbol{\mathcal{O}}_{n,r}}_{\varepsilon/(d+1)}$, for all $i \in [d]$ and $\overline{\mathbf{S}} \in \mathcal{N}^{\boldsymbol{\mathcal{D}}}_{\varepsilon/(d+1)}$ obeying $\left\|\mathbf{U}_i - \overline{\mathbf{U}}_i\right\|_{1,2} \leq \varepsilon/\left(d+1\right)$, for all $i \in [d]$ and $\left\|\mathbf{S} - \overline{\mathbf{S}}\right\|_F \leq \varepsilon/\left(d+1\right)$. This gives

$$\begin{aligned}
\left\|\mathbf{X} - \overline{\mathbf{X}}\right\|_F &= \left\|\mathbf{S} \times_1 \mathbf{U}_1 \times \cdots \times_d \mathbf{U}_d - \overline{\mathbf{S}} \times_1 \overline{\mathbf{U}}_1 \times \cdots \times_d \overline{\mathbf{U}}_d\right\|_F \\
&= \big\|\mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times \cdots \times_d \mathbf{U}_d \pm \mathbf{S} \times_1 \mathbf{U}_1 \times \mathbf{U}_2 \times \cdots \times_{d-1} \mathbf{U}_{d-1} \times_d \overline{\mathbf{U}}_d \\
&\quad \pm \mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times \cdots \times_{d-2} \mathbf{U}_{d-2} \times_{d-1} \overline{\mathbf{U}}_{d-1} \times_d \overline{\mathbf{U}}_d \\
&\quad \pm \cdots \pm \mathbf{S} \times_1 \overline{\mathbf{U}}_1 \times \cdots \times_d \overline{\mathbf{U}}_d - \overline{\mathbf{S}} \times_1 \overline{\mathbf{U}}_1 \times \cdots \times_d \overline{\mathbf{U}}_d\big\|_F \\
&\leq \left\|\mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times \cdots \times_d \left(\mathbf{U}_d - \overline{\mathbf{U}}_d\right)\right\|_F \\
&\quad + \left\|\mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times \cdots \times_{d-1} \left(\mathbf{U}_{d-1} - \overline{\mathbf{U}}_{d-1}\right) \times_d \overline{\mathbf{U}}_d\right\|_F \\
&\quad + \cdots + \left\|\mathbf{S} \times_1 \left(\mathbf{U}_1 - \overline{\mathbf{U}}_1\right) \times_2 \overline{\mathbf{U}}_2 \times \cdots \times_d \overline{\mathbf{U}}_d\right\|_F \\
&\quad + \left\|\left(\mathbf{S} - \overline{\mathbf{S}}\right) \times_1 \overline{\mathbf{U}}_1 \times_2 \overline{\mathbf{U}}_2 \times \cdots \times_d \overline{\mathbf{U}}_d\right\|_F.
\end{aligned} \tag{5.21}$$

For the first $d$ terms in the above estimation note that by unitarity $\sum_{i_j} \mathbf{U}_j\left(i_j, k_j\right) \mathbf{U}_j\left(i_j, l_j\right) = \delta_{k_j l_j}$ and $\sum_{i_j} \overline{\mathbf{U}}_j\left(i_j, k_j\right) \overline{\mathbf{U}}_j\left(i_j, l_j\right) = \delta_{k_j l_j}$, for all $j \in [d]$, and $\left\langle \mathbf{S}_{i_j = k_j}, \mathbf{S}_{i_j = l_j}\right\rangle = 0$ for all $j \in [d]$ whenever $k_j \neq l_j$. Therefore, we obtain

$$\left\|\mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times \cdots \times_{j-1} \mathbf{U}_{j-1} \times_j \left(\mathbf{U}_j - \overline{\mathbf{U}}_j\right) \times_{j+1} \overline{\mathbf{U}}_{j+1} \times \cdots \times_d \overline{\mathbf{U}}_d\right\|_F^2$$

$$
\begin{aligned}
&= \sum_{i_1,\dots,i_d} \left[\left(\mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times \cdots \times_{j-1} \mathbf{U}_{j-1} \times_j \left(\mathbf{U}_j - \overline{\mathbf{U}}_j\right) \times_{j+1} \overline{\mathbf{U}}_{j+1} \times \cdots \times_d \overline{\mathbf{U}}_d\right)(i_1,\dots,i_d)\right]^2 \\
&= \sum_{i_1,\dots,i_d} \sum_{k_1,\dots,k_d} \sum_{l_1,\dots,l_d} \mathbf{S}\left(k_1,\dots,k_d\right) \mathbf{S}\left(l_1,\dots,l_d\right) \mathbf{U}_1\left(i_1,k_1\right) \mathbf{U}_1\left(i_1,l_1\right) \mathbf{U}_2\left(i_2,k_2\right) \mathbf{U}_2\left(i_2,l_2\right) \\
&\cdot \ldots \cdot \left(\mathbf{U}_j - \overline{\mathbf{U}}_j\right)\left(i_j,k_j\right) \left(\mathbf{U}_j - \overline{\mathbf{U}}_j\right)\left(i_j,l_j\right) \cdot \ldots \cdot \overline{\mathbf{U}}_d\left(i_d,k_d\right) \overline{\mathbf{U}}_d\left(i_d,l_d\right) \\
&= \sum_{i_j} \sum_{k_1,\dots,k_d} \sum_{l_j} \mathbf{S}\left(k_1,\dots,k_j,\dots,k_d\right) \mathbf{S}\left(k_1,\dots,l_j,\dots,k_d\right) \left(\mathbf{U}_j - \overline{\mathbf{U}}_j\right)\left(i_j,k_j\right) \left(\mathbf{U}_j - \overline{\mathbf{U}}_j\right)\left(i_j,l_j\right) \\
&= \sum_{i_j} \sum_{k_1,k_2,\dots,k_d} \mathbf{S}\left(k_1,k_2,\dots,k_d\right)^2 \left(\left(\mathbf{U}_j - \overline{\mathbf{U}}_j\right)\left(i_j,k_j\right)\right)^2 \leq \left\|\mathbf{U}_j - \overline{\mathbf{U}}_j\right\|_{1,2}^2 \left\|\mathbf{S}\right\|_F^2 = \left\|\mathbf{U}_j - \overline{\mathbf{U}}_j\right\|_{1,2}^2 \\
&\leq \left(\varepsilon / \left(d+1\right)\right)^2 .
\end{aligned}
$$

In order to bound the last term in (5.21), observe that the unitarity of the matrices $\overline{\mathbf{U}}_i$ gives

$$
\left\|\left(\mathbf{S} - \overline{\mathbf{S}}\right) \times_1 \overline{\mathbf{U}}_1 \times \cdots \times_d \overline{\mathbf{U}}_d\right\|_F = \left\|\mathbf{S} - \overline{\mathbf{S}}\right\|_F \leq \varepsilon / \left(d+1\right).
$$

This completes the proof. □

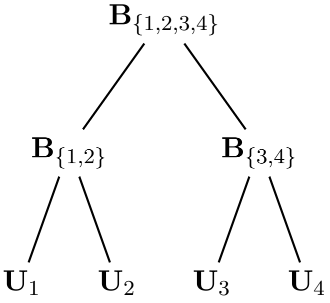


FIGURE 5.1. Tree for the HT-decomposition with $d = 4$

Next, we bound the covering numbers related to the HT decomposition, which includes the TT decomposition as a special case.

**Lemma 5.9** (Covering numbers related to HT-decomposition)**.** For a given HT-tree $T_I$, the covering numbers of the set of unit norm, rank-$\mathbf{r}$ tensors

$$
\boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\mathrm{HT}} = \left\{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} : \mathrm{rank}_{\mathrm{HT}}\left(\mathbf{X}\right) \leq \mathbf{r}_{\mathrm{HT}}, \left\|\mathbf{X}\right\|_F = 1\right\}
$$

satisfy

$$
\mathcal{N}\left(\boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\mathrm{HT}}, \left\|\cdot\right\|_F, \varepsilon\right) \leq \left(3(2d-1)\sqrt{r}/\varepsilon\right)^{\sum_{t \in \boldsymbol{\mathcal{I}}(T_I)} r_t r_{t_1} r_{t_2} + \sum_{i=1}^d r_i n_i} \quad \text{for } 0 \leq \varepsilon \leq 1, \tag{5.22}
$$

where $r = \max\left\{r_t : t \in T_I\right\}$, and $t_1$, $t_2$ are the left and the right son of a node $t$, respectively.

The proof requires a non-standard orthogonalization of the HT-decomposition. (The standard orthogonalization leads to worse bounds, in both TT and HT case.) We say that a tensor $\mathbf{B}_t \in \mathbb{C}^{r_t \times r_{t_1} \times r_{t_2}}$ is *right-orthogonal* if $\left(\mathbf{B}_t^{\{2,3\}}\right)^T \mathbf{B}_t^{\{2,3\}} = \mathbf{I}_{r_t}$. We call an HT-decomposition *right-orthogonal* if all transfer tensors $\mathbf{B}_t$, for $t \in \boldsymbol{\mathcal{I}}(T_I) \backslash \{t_{\mathrm{root}}\}$, i.e. except for the root, are *right orthogonal* and all frames $\mathbf{U}_i$ have orthogonal columns. For the sake of simple notation, we write the right-orthogonal HT-decomposition of a tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4}$ with the corresponding

HT-tree as in Figure 5.1 as

$$\mathbf{X} = \mathbf{B}_{\{1,2,3,4\}} \bigtriangledown \left(\mathbf{B}_{\{1,2\}} \bigtriangledown \mathbf{U}_1 \bigtriangledown \mathbf{U}_2\right) \bigtriangledown \left(\mathbf{B}_{\{3,4\}} \bigtriangledown \mathbf{U}_3 \bigtriangledown \mathbf{U}_4\right). \tag{5.23}$$

In fact, the above decomposition can be written as

$$\mathbf{X} = \mathbf{B}_{\{1,2,3,4\}} \bigtriangledown \left(\mathbf{B}_{\{1,2\}} \times_2 \mathbf{U}_1 \times_3 \mathbf{U}_2\right) \bigtriangledown \left(\mathbf{B}_{\{3,4\}} \times_2 \mathbf{U}_3 \times_3 \mathbf{U}_4\right).$$

since $\mathbf{U}_i$ is a matrix for all $i \in [4]$. However, for simplicity, we are going to use the notation as in (5.23). A right-orthogonal HT-decomposition can be obtained as follows from the standard orthogonal HT-decomposition (see [70]), where in particular, all frames $\mathbf{U}_i$ have orthonormal columns.

We first compute the QR-decomposition of the flattened transfer tensors $\mathbf{B}_t^{\{2,3\}} = \mathbf{Q}_t^{\{2,3\}}\mathbf{R}_t$ for all nodes $t$ at the highest possible level $\ell = p-1$. The level $\ell$ of the tree is defined as the set of all nodes having the distance of exactly $\ell$ to the root. We denote the level $\ell$ of the tree $T_I$ as $T_I^\ell = \{t \in T_I : \text{level}(t) = \ell\}$. (For example, for tree $T_I$ as in Figure 5.1, $T_I^0 = \{\{1,2,3,4\}\}$, $T_I^1 = \{\{1,2\},\{3,4\}\}$, $T_I^2 = \{\{1\},\{2\},\{3\},\{4\}\}$.) The $\mathbf{Q}_t$'s are then right-orthogonal by construction. In order to obtain a representation of the same tensor, we have to replace the tensors $\mathbf{B}_{t'}$ with nodes at lower level $p-2$ by $\bar{\mathbf{B}}_{t'} = \mathbf{B}_{t'} \times_2 \mathbf{R}_{t_{\text{left}}} \times_3 \mathbf{R}_{t_{\text{right}}}$, where $t_{\text{left}}$ corresponds to the left son of $t'$ and $t_{\text{right}}$ to the right son. We continue by computing the QR-decompositions of $\bar{\mathbf{B}}_{t'}^{\{2,3\}}$ with $t'$ at level $p-2$ and so on until we finally updated the root $\mathbf{B}_{\{1,2,\ldots,d\}}$ (which may remain the only non right-orthogonal transfer tensor). We illustrate this right-orthogonalization process for an HT-decomposition of the form (5.23) related to the HT-tree of Figure 5.1:

$$\begin{aligned}
\mathbf{X} &= \mathbf{B}_{\{1,2,3,4\}} \bigtriangledown \left(\mathbf{B}_{\{1,2\}} \bigtriangledown \mathbf{U}_1 \bigtriangledown \mathbf{U}_2\right) \bigtriangledown \left(\mathbf{B}_{\{3,4\}} \bigtriangledown \mathbf{U}_3 \bigtriangledown \mathbf{U}_4\right) \\
&= \mathbf{B}_{\{1,2,3,4\}} \bigtriangledown \left(\left[\mathbf{Q}_{\{1,2\}} \times_1 \mathbf{R}_{\{1,2\}}\right] \bigtriangledown \mathbf{U}_1 \bigtriangledown \mathbf{U}_2\right) \bigtriangledown \left(\left[\mathbf{Q}_{\{3,4\}} \times_1 \mathbf{R}_{\{3,4\}}\right] \bigtriangledown \mathbf{U}_3 \bigtriangledown \mathbf{U}_4\right) \\
&= \left[\mathbf{B}_{\{1,2,3,4\}} \times_2 \mathbf{R}_{\{1,2\}} \times_3 \mathbf{R}_{\{3,4\}}\right] \bigtriangledown \left(\mathbf{Q}_{\{1,2\}} \bigtriangledown \mathbf{U}_1 \bigtriangledown \mathbf{U}_2\right) \bigtriangledown \left(\mathbf{Q}_{\{3,4\}} \bigtriangledown \mathbf{U}_3 \bigtriangledown \mathbf{U}_4\right).
\end{aligned}$$

The second identity is easily verified by writing out the expressions with index notation. The last expression is a right-orthogonal HT decomposition with root tensor $\overline{\mathbf{B}}_{\{1,2,3,4\}} = \mathbf{B}_{\{1,2,3,4\}} \times_2 \mathbf{R}_{\{1,2\}} \times_3 \mathbf{R}_{\{3,4\}}$.

Proof of Lemma 5.9. For the sake of better readability, we will show the result for the special case of the order-4 HT-decomposition as in Figure 5.1 as well as for the special case of the TT decomposition for arbitrary $d$. The general case is then done analogously.

For the HT-tree $T_I$ as in Figure 5.1 we have $T_I = \{\{1,2,3,4\},\{1,2\},\{1\},\{2\},\{3,4\},\{3\},\{4\}\}$ and the number of nodes is $|T_I| = 2d-1 = 7$. We have to show that for $T_I$ as in Figure 5.1, the covering numbers of

$$\boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\mathrm{HT}} = \left\{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} : \operatorname{rank}_{\mathrm{HT}}(\mathbf{X}) \leq \mathbf{r}_{\mathrm{HT}}, \|\mathbf{X}\|_F = 1\right\},$$

satisfy

$$\mathcal{N}\left(\boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\mathrm{HT}}, \|\cdot\|_F, \varepsilon\right) \leq \left(21\sqrt{r}/\varepsilon\right)^{r_{\{1,2,3,4\}}r_{\{1,2\}}r_{\{3,4\}} + r_{\{1,2\}}r_1r_2 + r_{\{3,4\}}r_3r_4 + \sum_{i=1}^4 r_in_i} \quad \text{for } 0 \leq \varepsilon \leq 1.$$

For simplicity, we treat the case that $r_t = r$ for all $t \in T_I$ and $n_i = n$ for $i \in [4]$. We will use the right-orthogonal HT-decomposition introduced above and we cover the admissible components $\mathbf{U}_i$ and $\mathbf{B}_t$ in (5.23) separately, for all $t \in T_I$ and $i \in [4]$.

We introduce the set of right-orthogonal tensors $\boldsymbol{\mathcal{O}}_{r,r,r}^{\text{right}} = \left\{ \mathbf{U} \in \mathbb{R}^{r \times r \times r} : {\mathbf{U}^{\{2,3\}}}^T \mathbf{U}^{\{2,3\}} = \mathbf{I}_r \right\}$ which we will cover with respect to the norm

$$\|\mathbf{U}\|_{F,1} := \max_i \|\mathbf{U}(i,:,:)\|_F .$$

The set $\boldsymbol{\mathcal{Q}}_{r,r,r}^{\text{right}} := \left\{ \mathbf{X} \in \mathbb{R}^{r \times r \times r} : \|\mathbf{X}\|_{F,1} \le 1 \right\}$ contains $\boldsymbol{\mathcal{O}}_{r,r,r}^{\text{right}}$. Thus, by Lemma A.8 there is an $\varepsilon / (7\sqrt{r})$-set $\mathcal{N}_{\varepsilon/(7\sqrt{r})}^{\boldsymbol{\mathcal{O}}_{r,r,r}^{\text{right}}}$ for $\boldsymbol{\mathcal{O}}_{r,r,r}^{\text{right}}$ obeying

$$\left| \mathcal{N}_{\varepsilon/(7\sqrt{r})}^{\boldsymbol{\mathcal{O}}_{r,r,r}^{\text{right}}} \right| \le \left(3 \cdot 7\sqrt{r}/\varepsilon\right)^{r^3} = \left(21\sqrt{r}/\varepsilon\right)^{r^3} .$$

For the frames $\mathbf{U}_i \in \mathbb{R}^{n \times r}$ with $i \in [4]$, we define the set $\boldsymbol{\mathcal{O}}_{n,r} = \left\{ \mathbf{U} \in \mathbb{R}^{n \times r} : \mathbf{U}^T \mathbf{U} = \mathbf{I}_r \right\}$ which we cover with respect to

$$\|\mathbf{U}\|_{1,2} := \max_i \|\mathbf{U}(:,i)\|_2 .$$

Clearly, $\boldsymbol{\mathcal{O}}_{n,r} \subseteq \boldsymbol{\mathcal{Q}}_{n,r} := \left\{ \mathbf{X} \in \mathbb{R}^{n \times r} : \|\mathbf{X}\|_{1,2} \le 1 \right\}$ since the elements of an orthonormal set are unit normed. Again by Lemma A.8, there is an $\varepsilon/(7\sqrt{r})$-set $\mathcal{N}_{\varepsilon/(7\sqrt{r})}^{\boldsymbol{\mathcal{O}}_{n,r}}$ for $\boldsymbol{\mathcal{O}}_{n,r}$ obeying

$$\left| \mathcal{N}_{\varepsilon/(7\sqrt{r})}^{\boldsymbol{\mathcal{O}}_{n,r}} \right| \le \left(21\sqrt{r}/\varepsilon\right)^{nr} .$$

Finally, to cover $\mathbf{B}_{\{1,2,3,4\}}$, we define the set $\boldsymbol{\mathcal{F}}_{r,r} = \left\{ \mathbf{X} \in \mathbb{R}^{1 \times r \times r} : \|\mathbf{X}\|_F = 1 \right\}$ which has an $\varepsilon/(7\sqrt{r})$-net $\mathcal{N}_{\varepsilon/(7\sqrt{r})}^{\boldsymbol{\mathcal{F}}_{r,r}}$ of cardinality at most $(21\sqrt{r}/\varepsilon)^{r^2}$. We now define

$$\begin{aligned} \mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\text{HT}}} := \Big\{ & \overline{\mathbf{B}}_{\{1,2,3,4\}} \triangledown \left( \overline{\mathbf{B}}_{\{1,2\}} \triangledown \overline{\mathbf{U}}_1 \triangledown \overline{\mathbf{U}}_2 \right) \triangledown \left( \overline{\mathbf{B}}_{\{3,4\}} \triangledown \overline{\mathbf{U}}_3 \triangledown \overline{\mathbf{U}}_4 \right) : \\ & \overline{\mathbf{B}}_{\{1,2\}}, \overline{\mathbf{B}}_{\{3,4\}} \in \mathcal{N}_{\varepsilon/(7\sqrt{r})}^{\boldsymbol{\mathcal{O}}_{r,r,r}^{\text{right}}}, \overline{\mathbf{B}}_{\{1,2,3,4\}} \in \mathcal{N}_{\varepsilon/(7\sqrt{r})}^{\boldsymbol{\mathcal{F}}_{r,r}}, \overline{\mathbf{U}}_i \in \mathcal{N}_{\varepsilon/(7\sqrt{r})}^{\boldsymbol{\mathcal{O}}_{n,r}} \text{ for all } i \in [4] \Big\} \end{aligned}$$

and remark that

$$\mathcal{N}\left( \boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\text{HT}}, \|\cdot\|_F, \varepsilon \right) \le \left| \mathcal{N}_{\varepsilon/(7\sqrt{r})}^{\boldsymbol{\mathcal{O}}_{r,r,r}^{\text{right}}} \right|^2 \left| \mathcal{N}_{\varepsilon/(7\sqrt{r})}^{\boldsymbol{\mathcal{O}}_{n,r}} \right|^4 \left| \mathcal{N}_{\varepsilon/(7\sqrt{r})}^{\boldsymbol{\mathcal{F}}_{r,r}} \right| \le \left(21\sqrt{r}/\varepsilon\right)^{3r^3 + 4nr} .$$

It remains to show that for any $\mathbf{X} \in \boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\text{HT}}$ there exists $\overline{\mathbf{X}} \in \mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\text{HT}}}$ such that $\left\|\mathbf{X} - \overline{\mathbf{X}}\right\|_F \le 1$. For $\mathbf{X} = \mathbf{B}_{\{1,2,3,4\}} \triangledown \left( \mathbf{B}_{\{1,2\}} \triangledown \mathbf{U}_1 \triangledown \mathbf{U}_2 \right) \triangledown \left( \mathbf{B}_{\{3,4\}} \triangledown \mathbf{U}_3 \triangledown \mathbf{U}_4 \right)$, we choose $\overline{\mathbf{X}} = \overline{\mathbf{B}}_{\{1,2,3,4\}} \triangledown \left( \overline{\mathbf{B}}_{\{1,2\}} \triangledown \overline{\mathbf{U}}_1 \triangledown \overline{\mathbf{U}}_2 \right) \triangledown \left( \overline{\mathbf{B}}_{\{3,4\}} \triangledown \overline{\mathbf{U}}_3 \triangledown \overline{\mathbf{U}}_4 \right) \in \mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\text{HT}}}$ such that $\overline{\mathbf{B}}_{\{1,2,3,4\}} \in \boldsymbol{\mathcal{F}}_{r,r}$, $\overline{\mathbf{B}}_{\{1,2\}}, \overline{\mathbf{B}}_{\{3,4\}} \in \boldsymbol{\mathcal{O}}_{r,r,r}^{\text{right}}$, $\overline{\mathbf{U}}_i \in \boldsymbol{\mathcal{O}}_{n,r}$ for all $i \in [4]$ and

$$\begin{aligned} & \left\|\mathbf{U}_i - \overline{\mathbf{U}}_i\right\|_{1,2} \le \frac{\varepsilon}{7\sqrt{r}} \quad \text{for all } i \in [4], \\ & \left\|\mathbf{B}_{\{1,2,3,4\}} - \overline{\mathbf{B}}_{\{1,2,3,4\}}\right\|_F \le \frac{\varepsilon}{7\sqrt{r}}, \\ & \left\|\mathbf{B}_{\{1,2\}} - \overline{\mathbf{B}}_{\{1,2\}}\right\|_{F,1} \le \frac{\varepsilon}{7\sqrt{r}}, \quad \text{and } \left\|\mathbf{B}_{\{3,4\}} - \overline{\mathbf{B}}_{\{3,4\}}\right\|_{F,1} \le \frac{\varepsilon}{7\sqrt{r}}. \end{aligned}$$

Applying the triangle inequality results in

$$\left\|\mathbf{X} - \overline{\mathbf{X}}\right\|_F \le \left\| \mathbf{B}_{\{1,2,3,4\}} \triangledown \left( \mathbf{B}_{\{1,2\}} \triangledown \mathbf{U}_1 \triangledown \mathbf{U}_2 \right) \triangledown \left( \mathbf{B}_{\{3,4\}} \triangledown \mathbf{U}_3 \triangledown \left( \mathbf{U}_4 - \overline{\mathbf{U}}_4 \right) \right) \right\|_F \tag{5.24}$$

$$
\begin{aligned}
&+ \left\|\mathbf{B}_{\{1,2,3,4\}} \triangledown \left(\mathbf{B}_{\{1,2\}} \triangledown \mathbf{U}_1 \triangledown \mathbf{U}_2\right) \triangledown \left(\mathbf{B}_{\{3,4\}} \triangledown \left(\mathbf{U}_3 - \overline{\mathbf{U}}_3\right) \triangledown \overline{\mathbf{U}}_4\right)\right\|_F \\
&+ \left\|\mathbf{B}_{\{1,2,3,4\}} \triangledown \left(\mathbf{B}_{\{1,2\}} \triangledown \mathbf{U}_1 \triangledown \mathbf{U}_2\right) \triangledown \left(\left(\mathbf{B}_{\{3,4\}} - \overline{\mathbf{B}}_{\{3,4\}}\right) \triangledown \overline{\mathbf{U}}_3 \triangledown \overline{\mathbf{U}}_4\right)\right\|_F \qquad (5.25) \\
&+ \cdots + \left\|\left(\mathbf{B}_{\{1,2,3,4\}} - \overline{\mathbf{B}}_{\{1,2,3,4\}}\right) \triangledown \left(\overline{\mathbf{B}}_{\{1,2\}} \triangledown \overline{\mathbf{U}}_1 \triangledown \overline{\mathbf{U}}_2\right) \triangledown \left(\overline{\mathbf{B}}_{\{3,4\}} \triangledown \overline{\mathbf{U}}_3 \triangledown \overline{\mathbf{U}}_4\right)\right\|_F . \qquad (5.26)
\end{aligned}
$$

To estimate (5.24), we use orthogonality of $\mathbf{U}_i$, $i \in [4]$, and the right-orthogonality of $\mathbf{B}_{\{1,2\}}$, $\mathbf{B}_{\{3,4\}}$ to obtain

$$
\begin{aligned}
&\left\|\mathbf{B}_{\{1,2,3,4\}} \triangledown \left(\mathbf{B}_{\{1,2\}} \triangledown \mathbf{U}_1 \triangledown \mathbf{U}_2\right) \triangledown \left(\mathbf{B}_{\{3,4\}} \triangledown \mathbf{U}_3 \triangledown \left(\mathbf{U}_4 - \overline{\mathbf{U}}_4\right)\right)\right\|_F^2 \\
&= \sum_{i_1,\ldots,i_4} \sum_{\substack{j_1,\ldots,j_4 \\ k_1,\ldots,k_4}} \sum_{\substack{j_{12}, \\ k_{12}}} \sum_{\substack{j_{34}, \\ k_{34}}} \mathbf{B}_{\{1,2,3,4\}}\left(1, j_{12}, j_{34}\right) \mathbf{B}_{\{1,2,3,4\}}\left(1, k_{12}, k_{34}\right) \mathbf{B}_{\{1,2\}}\left(j_{12}, j_1, j_2\right) \\
&\cdot \mathbf{B}_{\{1,2\}}\left(k_{12}, k_1, k_2\right) \mathbf{U}_1\left(i_1, j_1\right) \mathbf{U}_1\left(i_1, k_1\right) \mathbf{U}_2\left(i_2, j_2\right) \mathbf{U}_2\left(i_2, k_2\right) \mathbf{B}_{\{3,4\}}\left(j_{34}, j_3, j_4\right) \\
&\cdot \mathbf{B}_{\{3,4\}}\left(k_{34}, k_3, k_4\right) \mathbf{U}_3\left(i_3, j_3\right) \mathbf{U}_3\left(i_3, k_3\right) \left(\mathbf{U}_4 - \overline{\mathbf{U}}_4\right)\left(i_4, j_4\right) \left(\mathbf{U}_4 - \overline{\mathbf{U}}_4\right)\left(i_4, k_4\right) \\
&= \sum_{i_4} \sum_{\substack{j_3, j_4 \\ k_4}} \sum_{j_{12}} \sum_{\substack{j_{34}, \\ k_{34}}} \mathbf{B}_{\{1,2,3,4\}}\left(1, j_{12}, j_{34}\right) \mathbf{B}_{\{1,2,3,4\}}\left(1, j_{12}, k_{34}\right) \mathbf{B}_{\{3,4\}}\left(j_{34}, j_3, j_4\right) \mathbf{B}_{\{3,4\}}\left(k_{34}, j_3, k_4\right) \\
&\cdot \left(\mathbf{U}_4 - \overline{\mathbf{U}}_4\right)\left(i_4, j_4\right) \left(\mathbf{U}_4 - \overline{\mathbf{U}}_4\right)\left(i_4, k_4\right) = \left\langle \Delta\mathbf{U}_4, \square\mathbf{B}_{\{3,4\}} \right\rangle \leq \left\|\Delta\mathbf{U}_4\right\|_{2\to 2} \left\|\square\mathbf{B}_{\{3,4\}}\right\|_*
\end{aligned}
$$

where

$$
\begin{aligned}
\Delta\mathbf{U}_4\left(j_4, k_4\right) &= \sum_{i_4} \left(\mathbf{U}_4 - \overline{\mathbf{U}}_4\right)\left(i_4, j_4\right) \left(\mathbf{U}_4 - \overline{\mathbf{U}}_4\right)\left(i_4, k_4\right) = \left[(\mathbf{U}_4 - \overline{\mathbf{U}}_4)^T (\mathbf{U}_4 - \overline{\mathbf{U}}_4)\right]\left(j_4, k_4\right), \\
\square\mathbf{B}_{\{3,4\}}\left(j_4, k_4\right) &= \sum_{j_3} \sum_{j_{12}} \sum_{j_{34}, k_{34}} \mathbf{B}_{\{1,2,3,4\}}\left(1, j_{12}, j_{34}\right) \mathbf{B}_{\{1,2,3,4\}}\left(1, j_{12}, k_{34}\right) \\
&\quad \cdot \mathbf{B}_{\{3,4\}}\left(j_{34}, j_3, j_4\right) \mathbf{B}_{\{3,4\}}\left(k_{34}, j_3, k_4\right).
\end{aligned}
$$

Since the Frobenius norm dominates the spectral norm, we have

$$
\left\|\Delta\mathbf{U}_4\right\|_{2\to 2} = \left\|\mathbf{U}_4 - \overline{\mathbf{U}}_4\right\|_{2\to 2}^2 \leq \left\|\mathbf{U}_4 - \overline{\mathbf{U}}_4\right\|_F^2 \leq r \left\|\mathbf{U}_4 - \overline{\mathbf{U}}_4\right\|_{1,2}^2 .
$$

Since $\square\overline{\mathbf{B}}_{\{3,4\}}$ is symmetric and positive semidefinite, it holds

$$
1 = \left\|\overline{\mathbf{X}}\right\|_F^2 = \left\langle \mathbf{I}, \square\overline{\mathbf{B}}_{\{3,4\}} \right\rangle = \operatorname{tr}\left(\square\overline{\mathbf{B}}_{\{3,4\}}\right) = \left\|\square\overline{\mathbf{B}}_{\{3,4\}}\right\|_* .
$$

Hence,

$$
\left\|\mathbf{B}_{\{1,2,3,4\}} \triangledown \left(\mathbf{B}_{\{1,2\}} \triangledown \mathbf{U}_1 \triangledown \mathbf{U}_2\right) \triangledown \left(\mathbf{B}_{\{3,4\}} \triangledown \mathbf{U}_3 \triangledown \left(\mathbf{U}_4 - \overline{\mathbf{U}}_4\right)\right)\right\|_F \leq \sqrt{r} \left\|\mathbf{U}_4 - \overline{\mathbf{U}}_4\right\|_{1,2} \leq \frac{\varepsilon}{7}.
$$

A similar procedure leads to the estimates

$$
\begin{aligned}
&\left\|\mathbf{B}_{\{1,2,3,4\}} \triangledown \left(\mathbf{B}_{\{1,2\}} \triangledown \left(\mathbf{U}_1 - \overline{\mathbf{U}}_1\right) \triangledown \overline{\mathbf{U}}_2\right) \triangledown \left(\overline{\mathbf{B}}_{\{3,4\}} \triangledown \overline{\mathbf{U}}_3 \triangledown \overline{\mathbf{U}}_4\right)\right\|_F \leq \sqrt{r} \left\|\mathbf{U}_1 - \overline{\mathbf{U}}_1\right\|_{1,2} \leq \frac{\varepsilon}{7}, \\
&\left\|\mathbf{B}_{\{1,2,3,4\}} \triangledown \left(\mathbf{B}_{\{1,2\}} \triangledown \mathbf{U}_1 \triangledown \left(\mathbf{U}_2 - \overline{\mathbf{U}}_2\right)\right) \triangledown \left(\overline{\mathbf{B}}_{\{3,4\}} \triangledown \overline{\mathbf{U}}_3 \triangledown \overline{\mathbf{U}}_4\right)\right\|_F \leq \sqrt{r} \left\|\mathbf{U}_2 - \overline{\mathbf{U}}_2\right\|_{1,2} \leq \frac{\varepsilon}{7}, \\
&\left\|\mathbf{B}_{\{1,2,3,4\}} \triangledown \left(\mathbf{B}_{\{1,2\}} \triangledown \mathbf{U}_1 \triangledown \mathbf{U}_2\right) \triangledown \left(\mathbf{B}_{\{3,4\}} \triangledown \left(\mathbf{U}_3 - \overline{\mathbf{U}}_3\right) \triangledown \overline{\mathbf{U}}_4\right)\right\|_F \leq \sqrt{r} \left\|\mathbf{U}_3 - \overline{\mathbf{U}}_3\right\|_{1,2} \leq \frac{\varepsilon}{7}.
\end{aligned}
$$

Since $\overline{\mathbf{U}}_i$ is orthogonal for all $i \in [4]$ and $\overline{\mathbf{B}}_{\{1,2\}}, \overline{\mathbf{B}}_{\{3,4\}}$ are right-orthogonal, we similarly estimate (5.25),

$$
\left\|\mathbf{B}_{\{1,2,3,4\}} \triangledown \left(\mathbf{B}_{\{1,2\}} \triangledown \mathbf{U}_1 \triangledown \mathbf{U}_2\right) \triangledown \left(\left(\mathbf{B}_{\{3,4\}} - \overline{\mathbf{B}}_{\{3,4\}}\right) \triangledown \overline{\mathbf{U}}_3 \triangledown \overline{\mathbf{U}}_4\right)\right\|_F^2
$$

$$
\begin{aligned}
&= \sum_{i_1,\ldots,i_4} \sum_{\substack{j_1,\ldots,j_4\\ k_1,\ldots,k_4}} \sum_{j_{12},k_{12}} \sum_{j_{34},k_{34}} \mathbf{B}_{\{1,2,3,4\}}\left(1, j_{12}, j_{34}\right) \mathbf{B}_{\{1,2,3,4\}}\left(1, k_{12}, k_{34}\right) \mathbf{B}_{\{1,2\}}\left(j_{12}, j_1, j_2\right) \\
&\cdot \mathbf{B}_{\{1,2\}}\left(k_{12}, k_1, k_2\right) \mathbf{U}_1\left(i_1, j_1\right) \mathbf{U}_1\left(i_1, k_1\right) \mathbf{U}_2\left(i_2, j_2\right) \mathbf{U}_2\left(i_2, k_2\right) \left(\mathbf{B}_{\{3,4\}} - \overline{\mathbf{B}}_{\{3,4\}}\right)\left(j_{34}, j_3, j_4\right) \\
&\cdot \left(\mathbf{B}_{\{3,4\}} - \overline{\mathbf{B}}_{\{3,4\}}\right)\left(k_{34}, k_3, k_4\right) \overline{\mathbf{U}}_3\left(i_3, j_3\right) \overline{\mathbf{U}}_3\left(i_3, k_3\right) \overline{\mathbf{U}}_4\left(i_4, j_4\right) \overline{\mathbf{U}}_4\left(i_4, k_4\right) \\
&= \sum_{j_3,j_4} \sum_{j_{12}} \sum_{j_{34},k_{34}} \mathbf{B}_{\{1,2,3,4\}}\left(1, j_{12}, j_{34}\right) \mathbf{B}_{\{1,2,3,4\}}\left(1, j_{12}, k_{34}\right) \left(\mathbf{B}_{\{3,4\}} - \overline{\mathbf{B}}_{\{3,4\}}\right)\left(j_{34}, j_3, j_4\right) \\
&\cdot \left(\mathbf{B}_{\{3,4\}} - \overline{\mathbf{B}}_{\{3,4\}}\right)\left(k_{34}, j_3, j_4\right) = \left\langle \Delta\mathbf{B}_{\{3,4\}}, \Box\mathbf{B}_{\{1,2,3,4\}} \right\rangle \leq \left\|\Delta\mathbf{B}_{\{3,4\}}\right\|_{2\to 2} \left\|\Box\mathbf{B}_{\{1,2,3,4\}}\right\|_{*}
\end{aligned}
$$

where

$$
\begin{aligned}
\Delta\mathbf{B}_{\{3,4\}}\left(j_{34}, k_{34}\right) &= \sum_{j_3,j_4} \left(\mathbf{B}_{\{3,4\}} - \overline{\mathbf{B}}_{\{3,4\}}\right)\left(j_{34}, j_3, j_4\right) \left(\mathbf{B}_{\{3,4\}} - \overline{\mathbf{B}}_{\{3,4\}}\right)\left(k_{34}, j_3, j_4\right) \\
&= \left[\left(\mathbf{B}^{\{2,3\}}_{\{3,4\}} - \overline{\mathbf{B}}^{\{2,3\}}_{\{3,4\}}\right)^T \left(\mathbf{B}^{\{2,3\}}_{\{3,4\}} - \overline{\mathbf{B}}^{\{2,3\}}_{\{3,4\}}\right)\right]\left(j_{34}, k_{34}\right) \\
\Box\mathbf{B}_{\{1,2,3,4\}}\left(j_{34}, k_{34}\right) &= \sum_{j_{12}} \mathbf{B}_{\{1,2,3,4\}}\left(1, j_{12}, j_{34}\right) \mathbf{B}_{\{1,2,3,4\}}\left(1, j_{12}, k_{34}\right).
\end{aligned}
$$

The spectral norm of $\Delta\mathbf{B}_{\{3,4\}}$ can be estimated as

$$
\left\|\Delta\mathbf{B}_{\{3,4\}}\right\|_{2\to 2} = \left\|\mathbf{B}^{\{2,3\}}_{\{3,4\}} - \overline{\mathbf{B}}^{\{2,3\}}_{\{3,4\}}\right\|^2_{2\to 2} \leq \left\|\mathbf{B}_{\{3,4\}} - \overline{\mathbf{B}}_{\{3,4\}}\right\|^2_F \leq r \left\|\mathbf{B}_{\{3,4\}} - \overline{\mathbf{B}}_{\{3,4\}}\right\|^2_{F,1}.
$$

Since $\Box\overline{\mathbf{B}}_{\{1,2,3,4\}}$ is symmetric and positive semidefinite

$$
1 = \left\|\overline{\mathbf{X}}\right\|^2_F = \left\langle \mathbf{I}, \Box\overline{\mathbf{B}}_{\{1,2,3,4\}} \right\rangle = \operatorname{tr}\left(\Box\overline{\mathbf{B}}_{\{1,2,3,4\}}\right) = \left\|\Box\overline{\mathbf{B}}_{\{1,2,3,4\}}\right\|_{*}.
$$

Hence,

$$
\begin{aligned}
&\left\|\mathbf{B}_{\{1,2,3,4\}} \triangledown \left(\mathbf{B}_{\{1,2\}} \triangledown \mathbf{U}_1 \triangledown \mathbf{U}_2\right) \triangledown \left(\left(\mathbf{B}_{\{3,4\}} - \overline{\mathbf{B}}_{\{3,4\}}\right) \triangledown \overline{\mathbf{U}}_3 \triangledown \overline{\mathbf{U}}_4\right)\right\|_F \\
&\qquad\qquad \leq \sqrt{r} \left\|\mathbf{B}_{\{3,4\}} - \overline{\mathbf{B}}_{\{3,4\}}\right\|_{F,1} \leq \frac{\varepsilon}{7}.
\end{aligned}
$$

A similar procedure leads to the following estimates

$$
\begin{aligned}
&\left\|\left(\mathbf{B}_{\{1,2,3,4\}} - \overline{\mathbf{B}}_{\{1,2,3,4\}}\right) \triangledown \left(\overline{\mathbf{B}}_{\{1,2\}} \triangledown \overline{\mathbf{U}}_1 \triangledown \overline{\mathbf{U}}_2\right) \triangledown \left(\overline{\mathbf{B}}_{\{3,4\}} \triangledown \overline{\mathbf{U}}_3 \triangledown \overline{\mathbf{U}}_4\right)\right\|_F \\
&\qquad\qquad \leq \left\|\mathbf{B}_{\{1,2,3,4\}} - \overline{\mathbf{B}}_{\{1,2,3,4\}}\right\|_F \leq \frac{\varepsilon}{7}, \\
&\left\|\mathbf{B}_{\{1,2,3,4\}} \triangledown \left(\left(\mathbf{B}_{\{1,2\}} - \overline{\mathbf{B}}_{\{1,2\}}\right) \triangledown \overline{\mathbf{U}}_1 \triangledown \overline{\mathbf{U}}_2\right) \triangledown \left(\overline{\mathbf{B}}_{\{3,4\}} \triangledown \overline{\mathbf{U}}_3 \triangledown \overline{\mathbf{U}}_4\right)\right\|_F \\
&\qquad\qquad \leq \sqrt{r} \left\|\mathbf{B}_{\{1,2\}} - \overline{\mathbf{B}}_{\{1,2\}}\right\|_{F,1} \leq \frac{\varepsilon}{7}.
\end{aligned}
$$

Plugging the bounds into (5.26) completes the proof for the HT-tree of Figure 5.1.

Let us now consider the TT-decomposition for tensors of order $d \geq 3$ as illustrated in Figure 5.2. We start with a right-orthogonal decomposition (see also the discussion after Lemma 5.9) of the form

$$
\begin{aligned}
\mathbf{X}\left(i_1, i_2, \ldots, i_d\right) = &\sum_{j_1, j_{23\ldots d}} \sum_{j_2, j_{3\ldots d}} \cdots \sum_{j_{d-1,d}, j_d} \mathbf{B}_{\{1,2,\ldots,d\}}\left(1, j_1, j_{23\ldots d}\right) \mathbf{U}_1\left(i_1, j_1\right) \\
&\cdot \mathbf{B}_{\{2,3,\ldots,d\}}\left(j_{23\ldots d}, j_2, j_{3\ldots d}\right) \mathbf{U}_2\left(i_2, j_2\right) \cdots \mathbf{B}_{\{d-1,d\}}\left(j_{d-1,d}, j_{d-1}, j_d\right) \\
&\cdot \mathbf{U}_{d-1}\left(i_{d-1}, j_{d-1}\right) \mathbf{U}_d\left(i_d, j_d\right).
\end{aligned}
$$

As for the general HT-decomposition, we write this as

$$\mathbf{X} = \mathbf{B}_{\{1,2,3,\ldots,d\}} \triangledown \mathbf{U}_1 \triangledown \left(\mathbf{B}_{\{2,3,\ldots,d\}} \triangledown \mathbf{U}_2 \triangledown \left(\cdots \triangledown \left(\mathbf{B}_{\{d-1,d\}} \triangledown \mathbf{U}_{d-1} \triangledown \mathbf{U}_d\right) \cdots\right)\right). \tag{5.27}$$

As above, we cover each set of admissible components $\mathbf{U}_i$, $\mathbf{B}_t$ separately, and then combine these components in order to obtain a covering of

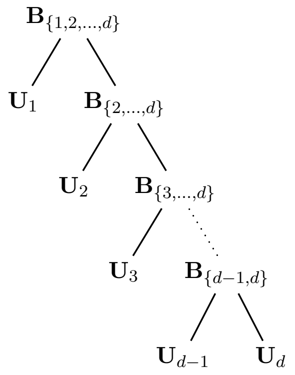


FIGURE 5.2. TT decomposition

$$\boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\mathrm{TT}} = \left\{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} : \operatorname{rank}_{\mathrm{TT}}(\mathbf{X}) \leq \mathbf{r}_{\mathrm{TT}}, \|\mathbf{X}\|_F = 1\right\}$$

with respect to the Frobenius norm, that is, we form

$$\begin{aligned}\mathcal{N}_\varepsilon^{\boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\mathrm{TT}}} := \Big\{&\overline{\mathbf{B}}_{\{1,2,3,\ldots,d\}} \triangledown \overline{\mathbf{U}}_1 \triangledown \left(\overline{\mathbf{B}}_{\{2,3,\ldots,d\}} \triangledown \overline{\mathbf{U}}_2 \triangledown \left(\cdots \triangledown \left(\overline{\mathbf{B}}_{\{d-1,d\}} \triangledown \overline{\mathbf{U}}_{d-1} \triangledown \overline{\mathbf{U}}_d\right) \cdots\right)\right) : \\ &\overline{\mathbf{U}}_i \in \mathcal{N}^{\boldsymbol{\mathcal{O}}_{n,r}}_{\varepsilon/\left((2d-1)\sqrt{r}\right)}, \overline{\mathbf{B}}_{\{1,\ldots,d\}} \in \mathcal{N}^{\boldsymbol{\mathcal{F}}_{r,r}}_{\varepsilon/\left((2d-1)\sqrt{r}\right)}, \overline{\mathbf{B}}_{\{j,j+1,\ldots,d\}} \in \mathcal{N}^{\boldsymbol{\mathcal{O}}^{\mathrm{right}}_{r,r,r}}_{\varepsilon/\left((2d-1)\sqrt{r}\right)}, \\ &\qquad\qquad i \in [d-1], j = 2, \ldots, d-1\Big\}.\end{aligned}$$

In order to show that $\mathcal{N}_\varepsilon^{\boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\mathrm{TT}}}$ forms an $\varepsilon$-net of $\boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\mathrm{TT}}$ we choose an arbitrary $\mathbf{X} \in \boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\mathrm{TT}}$ with right-orthogonal decomposition of the form (5.27) and for each $\mathbf{U}_i$ and $\mathbf{B}_{\{j,\ldots,d\}}$ the closest corresponding points $\overline{\mathbf{U}}_i \in \mathcal{N}^{\boldsymbol{\mathcal{O}}_{n,r}}_{\varepsilon/\left((2d-1)\sqrt{r}\right)}$, $\overline{\mathbf{B}}_{\{1,\ldots,d\}} \in \mathcal{N}^{\boldsymbol{\mathcal{F}}_{r,r}}_{\varepsilon/\left((2d-1)\sqrt{r}\right)}$, $\overline{\mathbf{B}}_{\{j,j+1,\ldots,d\}} \in \mathcal{N}^{\boldsymbol{\mathcal{O}}^{\mathrm{right}}_{r,r,r}}_{\varepsilon/\left((2d-1)\sqrt{r}\right)}$, $j = 2, \ldots, d-1$ resulting in $\mathbf{X} \in \mathcal{N}_\varepsilon^{\boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\mathrm{TT}}}$. The triangle inequality yields

$$\begin{aligned}\left\|\mathbf{X} - \overline{\mathbf{X}}\right\|_F \leq &\left\|\mathbf{B}_{\{1,2,\ldots,d\}} \triangledown \mathbf{U}_1 \triangledown \left(\mathbf{B}_{\{2,\ldots,d\}} \triangledown \cdots \left(\mathbf{B}_{\{d-1,d\}} \triangledown \mathbf{U}_{d-1} \triangledown \left(\mathbf{U}_d - \overline{\mathbf{U}}_d\right)\right) \cdots\right)\right\|_F \\ &+ \left\|\mathbf{B}_{\{1,2,\ldots,d\}} \triangledown \mathbf{U}_1 \triangledown \left(\mathbf{B}_{\{2,\ldots,d\}} \triangledown \cdots \triangledown \left(\mathbf{B}_{\{d-1,d\}} \triangledown \left(\mathbf{U}_{d-1} - \overline{\mathbf{U}}_{d-1}\right) \triangledown \overline{\mathbf{U}}_d\right) \cdots\right)\right\|_F + \cdots \\ &+ \left\|\left(\mathbf{B}_{\{1,2,\ldots,d\}} - \overline{\mathbf{B}}_{\{1,2,\ldots,d\}}\right) \triangledown \overline{\mathbf{U}}_1 \triangledown \left(\overline{\mathbf{B}}_{\{2,\ldots,d\}} \triangledown \cdots \triangledown \left(\overline{\mathbf{B}}_{\{d-1,d\}} \triangledown \overline{\mathbf{U}}_{d-1} \triangledown \overline{\mathbf{U}}_d\right) \cdots\right)\right\|_F.\end{aligned} \tag{5.28}$$

We need to bound terms of the form (for $q \in [d]$ and $p \in [d-1]$)

$$\left\|\mathbf{B}_{\{1,2,\ldots,d\}} \triangledown \mathbf{U}_1 \triangledown \cdots \triangledown \left(\mathbf{B}_{\{q,q+1,\ldots,d\}} \triangledown \left(\mathbf{U}_q - \overline{\mathbf{U}}_q\right) \triangledown \left(\overline{\mathbf{B}}_{\{q+1,\ldots,d\}} \triangledown \cdots \triangledown \overline{\mathbf{U}}_d\right) \cdots\right)\right\|_F, \tag{5.29}$$

$$\text{and } \left\|\mathbf{B}_{\{1,2,\ldots,d\}} \triangledown \mathbf{U}_1 \triangledown \cdots \triangledown \mathbf{U}_{p-1} \triangledown \left(\left(\mathbf{B}_{\{p,p+1,\ldots,d\}} - \overline{\mathbf{B}}_{\{p,p+1,\ldots,d\}}\right) \triangledown \overline{\mathbf{U}}_p \triangledown \left(\cdots \triangledown \overline{\mathbf{U}}_d\right) \cdots\right)\right\|_F. \tag{5.30}$$

To estimate (5.29), we use orthogonality of $\mathbf{U}_q$, $\overline{\mathbf{U}}_q$, $q \in [d]$, and right-orthogonality of $\mathbf{B}_{\{p,p+1\ldots,d\}}$, $\overline{\mathbf{B}}_{\{p,p+1,\ldots,d\}}$, $p = 2, 3, \ldots, d-1$, to obtain

$$\left\|\mathbf{B}_{\{1,2,\ldots,d\}} \triangledown \mathbf{U}_1 \triangledown \cdots \triangledown \left(\mathbf{B}_{\{q,q+1,\ldots,d\}} \triangledown \left(\mathbf{U}_q - \overline{\mathbf{U}}_q\right) \triangledown \left(\overline{\mathbf{B}}_{\{q+1,\ldots,d\}} \triangledown \cdots \triangledown \overline{\mathbf{U}}_d\right) \cdots\right)\right\|_F^2$$

$$
\begin{aligned}
&= \sum_{i_1,\dots,i_d} \sum_{\substack{j_1,\dots,j_d\\ k_1,\dots,k_d}} \sum_{\substack{j_{23\dots d},\\ j_{3\dots d},\\ \dots,\\ j_{d-1,d}}} \sum_{\substack{k_{23\dots d},\\ k_{3\dots d},\\ \dots,\\ k_{d-1,d}}} \mathbf{B}_{\{1,2,\dots,d\}}\left(1, j_1, j_{23\dots d}\right) \mathbf{B}_{\{1,2,\dots,d\}}\left(1, k_1, k_{23\dots d}\right) \mathbf{U}_1\left(i_1, j_1\right) \mathbf{U}_1\left(i_1, k_1\right) \\
&\cdots \mathbf{B}_{\{q,q+1,\dots,d\}}\left(j_{q,q+1\dots d}, j_q, j_{q+1\dots d}\right) \mathbf{B}_{\{q,q+1,\dots,d\}}\left(k_{q,q+1\dots d}, k_q, k_{q+1\dots d}\right) \\
&\cdot \left(\mathbf{U}_q - \overline{\mathbf{U}}_q\right)\left(i_q, j_q\right) \left(\mathbf{U}_q - \overline{\mathbf{U}}_q\right)\left(i_q, k_q\right) \overline{\mathbf{B}}_{\{q+1,\dots,d\}}\left(j_{q+1\dots d}, j_{q+1}, j_{q+2\dots d}\right) \\
&\cdot \overline{\mathbf{B}}_{\{q+1,\dots,d\}}\left(k_{q+1\dots d}, k_{q+1}, k_{q+2\dots d}\right) \cdots \overline{\mathbf{U}}_d\left(i_d, j_d\right) \overline{\mathbf{U}}_d\left(i_d, k_d\right) \\
&= \sum_{i_q} \sum_{\substack{j_1,\dots,j_q\\ k_q}} \sum_{\substack{j_{23\dots d},\\ j_{3\dots d},\dots,\\ j_{q+1\dots d}}} \sum_{\substack{k_{23\dots d},\\ k_{3\dots d},\dots,\\ k_{q\dots d}}} \mathbf{B}_{\{1,2,\dots,d\}}\left(1, j_1, j_{23\dots d}\right) \mathbf{B}_{\{1,2,\dots,d\}}\left(1, j_1, k_{23\dots d}\right) \\
&\cdots \overline{\mathbf{B}}_{\{q,q+1,\dots,d\}}\left(j_{q,q+1\dots d}, j_q, j_{q+1\dots d}\right) \overline{\mathbf{B}}_{\{q,q+1,\dots,d\}}\left(k_{q,q+1\dots d}, k_q, j_{q+1\dots d}\right) \\
&\cdot \left(\mathbf{U}_q - \overline{\mathbf{U}}_q\right)\left(i_q, j_q\right) \left(\mathbf{U}_q - \overline{\mathbf{U}}_q\right)\left(i_q, k_q\right) \\
&= \left\langle \Delta\mathbf{U}_q, \square\overline{\mathbf{B}}_{\{q,q+1,\dots,d\}} \right\rangle \leq \left\| \Delta\mathbf{U}_q \right\|_{2\to 2} \left\| \square\overline{\mathbf{B}}_{\{q,q+1,\dots,d\}} \right\|_* ,
\end{aligned}
$$

where

$$
\begin{aligned}
\Delta\mathbf{U}_q\left(j_q, k_q\right) &= \sum_{i_q} \left(\mathbf{U}_q - \overline{\mathbf{U}}_q\right)\left(i_q, j_q\right) \left(\mathbf{U}_q - \overline{\mathbf{U}}_q\right)\left(i_q, k_q\right), \\
\square\overline{\mathbf{B}}_{\{q,q+1,\dots,d\}}\left(j_q, k_q\right) &= \sum_{j_1,\dots,j_{q-1}} \sum_{\substack{j_{23\dots d},\\ j_{3\dots d},\\ \dots,j_{q+1,\dots,d}}} \sum_{\substack{k_{23\dots d},\\ k_{3\dots d},\\ \dots,k_{q,\dots,d}}} \mathbf{B}_{\{1,2,\dots,d\}}\left(1, j_1, j_{23\dots d}\right) \mathbf{B}_{\{1,2,\dots,d\}}\left(1, j_1, k_{23\dots d}\right) \\
&\cdots \mathbf{B}_{\{q-1,q,\dots,d\}}\left(j_{q-1,q\dots d}, j_{q-1}, j_{q\dots d}\right) \mathbf{B}_{\{q-1,q,\dots,d\}}\left(k_{q-1,q\dots d}, j_{q-1}, k_{q\dots d}\right) \\
&\cdot \overline{\mathbf{B}}_{\{q,q+1,\dots,d\}}\left(j_{q,q+1\dots d}, j_q, j_{q+1\dots d}\right) \overline{\mathbf{B}}_{\{q,q+1,\dots,d\}}\left(k_{q,q+1\dots d}, k_q, j_{q+1\dots d}\right).
\end{aligned}
$$

We have

$$
\left\| \Delta\mathbf{U}_q \right\|_{2\to 2} = \left\| \mathbf{U}_q - \overline{\mathbf{U}}_q \right\|_{2\to 2}^2 \leq \left\| \mathbf{U}_q - \overline{\mathbf{U}}_q \right\|_F^2 \leq r \left\| \mathbf{U}_q - \overline{\mathbf{U}}_q \right\|_{1,2}^2 .
$$

Since $\square\overline{\mathbf{B}}_{\{q,q+1,\dots,d\}}$ is symmetric and positive semidefinite

$$
1 = \left\| \overline{\mathbf{X}} \right\|_F^2 = \left\langle \mathbf{I}, \square\overline{\mathbf{B}}_{\{q,q+1,\dots,d\}} \right\rangle = \operatorname{tr}\left( \square\overline{\mathbf{B}}_{\{q,q+1,\dots,d\}} \right) = \left\| \square\overline{\mathbf{B}}_{\{q,q+1,\dots,d\}} \right\|_* .
$$

Hence,

$$
\begin{aligned}
&\left\| \mathbf{B}_{\{1,2,\dots,d\}} \triangledown \mathbf{U}_1 \triangledown \cdots \triangledown \left( \mathbf{B}_{\{q,q+1,\dots,d\}} \triangledown \left( \mathbf{U}_q - \overline{\mathbf{U}}_q \right) \triangledown \left( \overline{\mathbf{B}}_{\{q+1,\dots,d\}} \triangledown \cdots \triangledown \overline{\mathbf{U}}_d \right) \cdots \right) \right\|_F \\
&\qquad \leq \sqrt{r} \left\| \mathbf{U}_q - \overline{\mathbf{U}}_q \right\|_{1,2} \leq \frac{\varepsilon}{2d-1}.
\end{aligned}
$$

In a similar way, distinguishing the cases $p = 1$ and $p = 2, \dots, d-1$, we estimate terms of the form (5.30) as

$$
\left\| \mathbf{B}_{\{1,2,\dots,d\}} \triangledown \mathbf{U}_1 \triangledown \cdots \triangledown \mathbf{U}_{p-1} \triangledown \left( \left( \mathbf{B}_{\{p,p+1,\dots,d\}} - \overline{\mathbf{B}}_{\{p,p+1,\dots,d\}} \right) \triangledown \overline{\mathbf{U}}_p \triangledown \left( \cdots \triangledown \overline{\mathbf{U}}_d \right) \cdots \right) \right\|_F \leq \frac{\varepsilon}{2d-1}.
$$

Plugging the bounds into (5.28) completes the proof for the TT decomposition. □

The proof of Theorem 5.7 also requires a recent deviation bound [49, 95] for random variables of the form $X = \sup_{\mathbf{B}\in\boldsymbol{\mathcal{B}}} \left| \left\| \mathbf{B}\boldsymbol{\xi} \right\|_2^2 - \mathbb{E} \left\| \mathbf{B}\boldsymbol{\xi} \right\|_2^2 \right|$ in terms of a complexity parameter of the set of matrices $\boldsymbol{\mathcal{B}}$ involving covering numbers. In order to state it, we introduce the radii of a set of matrices $\boldsymbol{\mathcal{B}}$ in the Frobenius norm, the operator norm, and the Schatten-4 norm as

$$d_F(\boldsymbol{\mathcal{B}}) := \sup_{\mathbf{B}\in\boldsymbol{\mathcal{B}}} \|\mathbf{B}\|_F\,,\; d_{2\to 2}(\boldsymbol{\mathcal{B}}) := \sup_{\mathbf{B}\in\boldsymbol{\mathcal{B}}} \|\mathbf{B}\|_{2\to 2}\,,\; d_4(\boldsymbol{\mathcal{B}}) := \sup_{\mathbf{B}\in\boldsymbol{\mathcal{B}}} \|\mathbf{B}\|_{S_4} = \sup_{\mathbf{B}\in\boldsymbol{\mathcal{B}}} \left(\operatorname{tr}\left(\mathbf{B}^T\mathbf{B}\right)^2\right)^{1/4}.$$

The complexity parameter is Talagrand's $\gamma_2$-functional $\gamma_2(\boldsymbol{\mathcal{B}}, \|\cdot\|_{2\to2})$. We do not give the precise definition here, but refer to [151] for details. For us, it is only important that it can be bounded in terms of covering numbers via a Dudley type integral [53, 151] as

$$\gamma_2(\boldsymbol{\mathcal{B}}, \|\cdot\|_{2\to2}) \leq C \int_0^{d_{2\to2}(\boldsymbol{\mathcal{B}})} \sqrt{\log \mathcal{N}(\boldsymbol{\mathcal{B}}, \|\cdot\|_{2\to2}\,, u)}du. \tag{5.31}$$

We will use the following result from [49, Theorem 6.5] which is a slightly refined version of the main result of [95].

**Theorem 5.10.** Let $\boldsymbol{\mathcal{B}}$ be a set of matrices, and let $\boldsymbol{\xi}$ be a random vector whose entries $\xi_j$ are independent, mean-zero, variance 1 and $L$-subgaussian random variables. Set

$$\begin{aligned} E &= \gamma_2(\boldsymbol{\mathcal{B}}, \|\cdot\|_{2\to2})\left(\gamma_2(\boldsymbol{\mathcal{B}}, \|\cdot\|_{2\to2}) + d_F(\boldsymbol{\mathcal{B}})\right) + d_F(\boldsymbol{\mathcal{B}})\, d_{2\to2}(\boldsymbol{\mathcal{B}}) \\ V &= d_4^2(\boldsymbol{\mathcal{B}})\,, \text{ and } U = d_{2\to2}^2(\boldsymbol{\mathcal{B}})\,. \end{aligned}$$

Then, for $t > 0$,

$$\mathbb{P}\left(\sup_{\mathbf{B}\in\boldsymbol{\mathcal{B}}} \left|\|\mathbf{B}\boldsymbol{\xi}\|_2^2 - \mathbb{E}\|\mathbf{B}\boldsymbol{\xi}\|_2^2\right| \geq c_1 E + t\right) \leq 2\exp\left(-c_2 \min\left\{\frac{t^2}{V^2}, \frac{t}{U}\right\}\right).$$

The constants $c_1, c_2$ only depend on $L$.

Proof of Theorem 5.7. We write

$$\mathcal{A}(\mathbf{X}) = \mathbf{V}_{\mathbf{X}}\boldsymbol{\xi},$$

where $\boldsymbol{\xi}$ is an $L$-subgaussian random vector of length $n_1 n_2 \cdots n_d m$ and $\mathbf{V}_{\mathbf{X}}$ is the $m \times n_1 n_2 \cdots n_d m$ block-diagonal matrix

$$\mathbf{V}_{\mathbf{X}} = \frac{1}{\sqrt{m}} \begin{bmatrix} \mathbf{x}^T & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}^T & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}^T \end{bmatrix},$$

with $\mathbf{x}$ being the vectorized version of the tensor $\mathbf{X}$. With this notation the restricted isometry constant is given by

$$\delta_{\mathbf{r}} = \sup_{\mathbf{X}\in\boldsymbol{\mathcal{T}}} \left|\|\mathbf{V}_{\mathbf{X}}\boldsymbol{\xi}\|_2^2 - \mathbb{E}\|\mathbf{V}_{\mathbf{X}}\boldsymbol{\xi}\|_2^2\right|,$$

where in the HOSVD case $\boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{S}}_{\mathbf{r}} = \{\mathbf{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d} : \operatorname{rank}_{\text{HOSVD}}(\mathbf{X}) \leq \mathbf{r}, \|\mathbf{X}\|_F = 1\}$, and $\boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\text{HT}} = \{\mathbf{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d} : \operatorname{rank}_{\text{HT}}(\mathbf{X}) \leq \mathbf{r}, \|\mathbf{X}\|_F = 1\}$ in the HT-case (including the TT case). Theorem 5.10 provides a general probabilistic bound for expressions in the form of the right hand side above in terms of the diameters $d_F(\boldsymbol{\mathcal{B}})$, $d_{2\to2}(\boldsymbol{\mathcal{B}})$, and $d_4(\boldsymbol{\mathcal{B}})$ of the set $\boldsymbol{\mathcal{B}} := \{\mathbf{V}_{\mathbf{X}} : \mathbf{X} \in \boldsymbol{\mathcal{T}}\}$, as well as in terms of Talagrand's functional $\gamma_2(\boldsymbol{\mathcal{B}}, \|\cdot\|_{2\to2})$. It is straightforward to see that $d_F(\boldsymbol{\mathcal{B}}) = 1$, since $\|\mathbf{X}\|_F = 1$, for all $\mathbf{X} \in \boldsymbol{\mathcal{T}}$. Furthermore, for all $\mathbf{X} \in \boldsymbol{\mathcal{T}}$,

$$m\mathbf{V}_{\mathbf{X}}\mathbf{V}_{\mathbf{X}}^T = \begin{bmatrix} \mathbf{x}^T\mathbf{x} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}^T\mathbf{x} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}^T\mathbf{x} \end{bmatrix} = \begin{bmatrix} \|\mathbf{x}\|_2^2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \|\mathbf{x}\|_2^2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \|\mathbf{x}\|_2^2 \end{bmatrix} = \mathbf{I}_m, \tag{5.32}$$

so that $\|\mathbf{V}_{\mathbf{X}}\|_{2\to 2} = \frac{1}{\sqrt{m}}$ and $d_{2\to 2}(\boldsymbol{\mathcal{B}}) = \frac{1}{\sqrt{m}}$. (Since the operator norm of a block-diagonal matrix is the maximum of the operator norm of its diagonal blocks we obtain

$$\|\mathbf{V}_{\mathbf{X}}\|_{2\to 2} = \frac{1}{\sqrt{m}} \|\mathbf{x}\|_2 = \frac{1}{\sqrt{m}} \|\mathbf{X}\|_F .) \tag{5.33}$$

From the cyclicity of the trace and (5.32) it follows that

$$\|\mathbf{V}_{\mathbf{X}}\|_{S_4}^4 = \operatorname{tr}\left[\left(\mathbf{V}_{\mathbf{X}}^T\mathbf{V}_{\mathbf{X}}\right)^2\right] = \operatorname{tr}\left[\left(\mathbf{V}_{\mathbf{X}}\mathbf{V}_{\mathbf{X}}^T\right)^2\right] = \operatorname{tr}\left[\left(\frac{1}{m}\mathbf{I}_m\right)^2\right] = \operatorname{tr}\left(\frac{1}{m^2}\mathbf{I}_m\right) = \frac{1}{m},$$

for all $\mathbf{V}_{\mathbf{X}} \in \boldsymbol{\mathcal{B}}$. Thus, $d_4^2(\boldsymbol{\mathcal{B}}) = \sup_{\mathbf{V}_{\mathbf{X}} \in \boldsymbol{\mathcal{B}}} \|\mathbf{V}_{\mathbf{X}}\|_{S_4}^2 = \frac{1}{\sqrt{m}}$. Using observation (5.33), the bound of the $\gamma_2$-functional via the Dudley type integral (5.31) yields

$$\gamma_2\left(\boldsymbol{\mathcal{B}}, \|\cdot\|_{2\to 2}\right) \leq C\frac{1}{\sqrt{m}} \int_0^1 \sqrt{\log\left(\mathcal{N}\left(\boldsymbol{\mathcal{S}}_{\mathbf{r}}, \|\cdot\|_F, u\right)\right)}\, du, \tag{5.34}$$

where $\boldsymbol{\mathcal{S}}_{\mathbf{r}}$ is replaced by $\boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\mathrm{HT}}$ in the HT case.

Let us first continue with the HOSVD case. Using the bound (5.20) for $\mathcal{N}\left(\boldsymbol{\mathcal{S}}_{\mathbf{r}}, \|\cdot\|_F, u\right)$ and the triangle inequality we reach

$$\begin{aligned} \gamma_2\left(\boldsymbol{\mathcal{B}}, \|\cdot\|_{2\to 2}\right) &\leq C\frac{1}{\sqrt{m}} \int_0^1 \sqrt{\left(r_1 r_2 \cdots r_d + \sum_{i=1}^d n_i r_i\right) \log\left(3\left(d+1\right)/u\right)}\, du \\ &= C\sqrt{\frac{r_1 r_2 \cdots r_d + \sum_{i=1}^d n_i r_i}{m}} \int_0^1 \sqrt{\log\left(d+1\right) + \log\left(3/u\right)}\, du \\ &\leq C\sqrt{\frac{r_1 r_2 \cdots r_d + \sum_{i=1}^d n_i r_i}{m}} \left(\sqrt{\log\left(d+1\right)} + \int_0^1 \sqrt{\log\left(3/u\right)}\, du\right) \\ &\leq \tilde{C}\sqrt{\frac{\left(r_1 r_2 \cdots r_d + \sum_{i=1}^d n_i r_i\right)\log\left(d\right)}{m}} \leq \tilde{C}\sqrt{\frac{(r^d + dnr)\log(d)}{m}}, \end{aligned} \tag{5.35}$$

where $r := \max\left\{r_i : i \in [d]\right\}$ and $n := \max\left\{n_i : i \in [d]\right\}$.

Let us now consider the HT case (including the TT case). Using the bound (5.34) of the $\gamma_2$-functional via Dudley type integral and the covering number bound (5.22) for $\mathcal{N}\left(\boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\mathrm{HT}}, \|\cdot\|_F, u\right)$, we obtain

$$\begin{aligned} \gamma_2\left(\boldsymbol{\mathcal{B}}, \|\cdot\|_{2\to 2}\right) &\leq C\frac{1}{\sqrt{m}} \int_0^1 \sqrt{\log\left(\mathcal{N}\left(\boldsymbol{\mathcal{S}}_{\mathbf{r}}^{\mathrm{HT}}, \|\cdot\|_F, u\right)\right)}\, du \\ &\leq C\frac{1}{\sqrt{m}} \sqrt{\sum_{t \in \boldsymbol{\mathcal{I}}(T_I)} r_t r_{t_1} r_{t_2} + \sum_{i=1}^d r_i n_i} \cdot \int_0^1 \sqrt{\log\left(3(2d-1)\sqrt{r}/u\right)}\, du. \\ &\leq \tilde{C}_1 \sqrt{\frac{\left(\sum_{t \in \boldsymbol{\mathcal{I}}(T_I)} r_t r_{t_1} r_{t_2} + \sum_{i=1}^d r_i n_i\right)\log\left((2d-1)\sqrt{r}\right)}{m}} \end{aligned}$$

$$\leq \tilde{C}_1 \sqrt{\frac{((d-1)r^3 + dnr)\log((2d-1)\sqrt{r})}{m}}. \tag{5.36}$$

In order to apply Theorem 5.10 we note that

$$\begin{aligned} E &= \gamma_2(\boldsymbol{\mathcal{B}}, \|\cdot\|_{2\to 2})(\gamma_2(\boldsymbol{\mathcal{B}}, \|\cdot\|_{2\to 2}) + d_F(\boldsymbol{\mathcal{B}})) + d_F(\boldsymbol{\mathcal{B}})\, d_{2\to 2}(\boldsymbol{\mathcal{B}}) \\ &= \gamma_2^2(\boldsymbol{\mathcal{B}}, \|\cdot\|_{2\to 2}) + \gamma_2(\boldsymbol{\mathcal{B}}, \|\cdot\|_{2\to 2}) + \frac{1}{\sqrt{m}}, \\ V &= d_4^2(\boldsymbol{\mathcal{B}}) = \frac{1}{\sqrt{m}}, \qquad U = d_{2\to 2}^2(\boldsymbol{\mathcal{B}}) = \frac{1}{m}. \end{aligned}$$

The bound on $m$ of Theorem 5.7 ensures that $c_1 E \leq \delta/2$ and that $2\exp\left(-c_2 \min\left\{\frac{t^2}{V^2}, \frac{t}{U}\right\}\right) \leq \varepsilon$ with $t = \delta/2$ (provided constants are chosen appropriately). Therefore, the claim follows from Theorem 5.10. □

## 5.3. Random Fourier measurements

While subgaussian measurements often provide benchmark guarantees in compressive sensing and low-rank recovery in terms of the minimal number of required measurements, they lack of any structure and therefore are of limited use in practice. In particular, no fast multiplication routines are available for them. In order to overcome such limitations, structured random measurement matrices have been studied in compressive sensing [28, 60, 96, 129] and low-rank matrix recovery [23, 27, 58, 96] and almost optimal recovery guarantees have been shown.

In this section, we extend one particular construction of a randomized Fourier transform from the matrix case [58, Section 1] to the tensor case. The measurement map

$$\mathcal{A}: \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{C}^m, \quad \mathcal{A} = \frac{1}{\sqrt{m}} \mathcal{R}_{\boldsymbol{\Omega}} \mathcal{F}_d \mathcal{D}$$

is the composition of a random sign flip map $\mathcal{D}: \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ defined componentwise as $\mathcal{D}(\mathbf{X})(j_1, \ldots, j_d) = \epsilon_{j_1,\ldots,j_d} \mathbf{X}(j_1, \ldots, j_d)$ with the $\epsilon_{j_1,\ldots,j_d}$ being independent ±1 Rademacher variables, a $d$-dimensional Fourier transform

$$\begin{aligned} &\mathcal{F}_d: \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}, \\ &\qquad \mathcal{F}_d(\mathbf{X})(j_1, \ldots, j_d) = \sum_{k_1=1}^{n_1} \cdots \sum_{k_d=1}^{n_d} \mathbf{X}(k_1, \ldots, k_d)\, e^{-2\pi i \sum_{\ell=1}^{d} \frac{k_\ell j_\ell}{n_\ell}}, \end{aligned}$$

and a random subsampling operator $\mathcal{R}_{\boldsymbol{\Omega}}: \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{C}^{\boldsymbol{\Omega}} = \mathbb{C}^m$, $\mathcal{R}_{\boldsymbol{\Omega}}(\mathbf{X})_{\mathbf{j}} = \mathbf{X}(\mathbf{j})$ for $\mathbf{j} \in \boldsymbol{\Omega} \subset [n_1] \times \cdots \times [n_d]$, where $\boldsymbol{\Omega}$ is selected uniformly at random among all subsets of $[n_1] \times \cdots \times [n_d]$ of cardinality $m$. Instead of the $d$-dimensional Fourier transform, we can also use the 1-dimensional Fourier transform applied to the vectorized version of a tensor $\mathbf{X}$ without changes in the results below. Since the Fourier transform can be applied quickly in $\mathcal{O}(n^d \log n^d)$, $n = \max\{n_\ell : \ell \in [d]\}$, operations using the FFT, the map $\mathcal{A}$ runs with this computational complexity – as opposed to the trivial running time of $\mathcal{O}(n^{2d})$ for unstructured measurement maps. By vectorizing tensors in $\mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$, the map $\mathcal{A}$ can be written as a partial random Fourier matrices with randomized column signs.

The randomized Fourier map $\mathcal{A}$ satisfies the TRIP for an almost optimal number of measurements as shown by the next result.

**Theorem 5.11.** Let $\mathcal{A} : \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{C}^m$ be the randomized Fourier map described above. Then $\mathcal{A}$ satisfies TRIP with tensor restricted isometry constant $\delta_{\mathbf{r}}$ with probability exceeding $1 - 2e^{-\eta}$ as long as

$$m \geq C\delta_{\mathbf{r}}^{-1} (1+\eta) \log^2(n^d) \max\left\{\delta_{\mathbf{r}}^{-1} (1+\eta) \log^2(n^d), f(n,d,r)\right\}, \tag{5.37}$$

where

$$f(n,d,r) = \left(r^d + dnr\right) \log(d) \quad \text{for the HOSVD case },$$
$$f(n,d,r) = \left(dr^3 + dnr\right) \log(dr) \quad \text{for the TT and HT case,}$$

$n = \max\{n_i : i \in [d]\}$ and $r = \max\{r_t : t \in T_I\}$.

To prove Theorem 5.11 we use a special case of Theorem 3.3 in [127] (included in Appendix A as Theorem A.4) for the partial Fourier matrix with randomized column signs, which generalizes the main result of [97]. Using that the Gaussian width of a set $\mathcal{T}$ is equivalent to $\gamma_2(\mathcal{T}, \|\cdot\|_2)$ by Talagrand's majorizing theorem [149, 150], this result reads in our notation as follows.

**Theorem 5.12.** Let $\mathcal{T} \subset \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ and let $\mathcal{A} : \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{C}^m$ be the randomized Fourier map as described above. Then for $0 < \delta < 1$

$$\sup_{\mathbf{X} \in \mathcal{T}} \left| \|\mathcal{A}(\mathbf{X})\|_2^2 - \|\mathbf{X}\|_2^2 \right| \leq \delta \cdot \left(d_F\left(\mathcal{T}\right)\right)^2,$$

holds with probability at least $1 - 2e^{-\eta}$ as long as

$$m \geq C\delta^{-2} (1+\eta)^2 \left(\log(n_1 \cdots n_d)\right)^4 \max\left\{1, \frac{\gamma_2^2\left(\mathcal{T}, \|\cdot\|_F\right)}{\left(d_F\left(\mathcal{T}\right)\right)^2}\right\}. \tag{5.38}$$

Proof of Theorem 5.11. We use $\mathcal{T} = \mathcal{S}_{\mathbf{r}}$ and $\mathcal{T} = \mathcal{S}_{\mathbf{r}}^{\mathrm{HT}}$ and recall that $d_F(\mathcal{T}) = 1$. Moreover, $\gamma_2(\mathcal{T}, \|\cdot\|_F)$ has been estimated in (5.35) and (5.36). By distinguishing cases, one then verifies that (5.38) implies (5.37) so that Theorem 5.12 implies Theorem 5.11. □

Using recent improved estimates for the standard RIP for random partial Fourier matrices [12, 77] in connection with techniques from [127] it may be possible to improve Theorem 5.12 and thereby (5.37) in terms of logarithmic factors.

## 5.4. Numerical results

We present numerical results for recovery of third order tensors $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and the HOSVD format which illustrate that tensor iterative hard thresholding works very well despite the fact that we only have a partial recovery result. We ran experiments for both versions (CTIHT and NTIHT) of the algorithm and for Gaussian random measurement maps, randomized Fourier measurement maps (where $\mathbf{X} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$), and tensor completion, i.e., recovery from randomly chosen entries of the tensor. (No theoretical investigations are yet available for the latter scenario).

For other related numerical results, we refer to papers [45, 65], where they have considered a slightly different versions of the tensor iterative hard thresholding algorithm and compared it with NTIHT.

We consider recovery of a cubic tensor, i.e., $n_1 = n_2 = n_3 = 10$, with equal and unequal ranks of its unfoldings, respectively, (first and second experiment) and of a non-cubic tensor $\mathbf{X} \in \mathbb{R}^{6 \times 10 \times 15}$ with equal ranks of the unfoldings, i.e., $r_1 = r_2 = r_3 = r$ (third experiment). For

| type | tensor dimensions | rank | CTIHT-$\overline{n}$ | CPU time in sec |
|---|---|---|---|---|
| Fourier | $100 \times 100 \times 100$ | $(1,1,1)$ | 10 | 16.2709 |
| | $100 \times 100 \times 100$ | $(1,1,1)$ | 20 | 14.9761 |
| | $100 \times 100 \times 100$ | $(5,5,5)$ | 10 | 31.8866 |
| | $100 \times 100 \times 100$ | $(5,5,5)$ | 20 | 26.3486 |
| | $100 \times 100 \times 100$ | $(7,7,7)$ | 20 | 27.2222 |
| | $100 \times 100 \times 100$ | $(10,10,10)$ | 20 | 36.3950 |
| Fourier | $200 \times 200 \times 200$ | $(1,1,1)$ | 10 | 142.2105 |

Table 5.1. The numerical experiments are run on a personal computer with processor Intel(R) Core(TM) i7-2600 CPU @ 3.40 GHz on Windows 7 Professional Platform (with 64-bit operating system) and 8 GB RAM; $\overline{n}$ denotes the percentage of measurements – the number of measurements equals to $m = \lceil n_1 n_2 n_3 \frac{\overline{n}}{100} \rceil$;

fixed tensor dimensions $n_1 \times n_2 \times n_3$, fixed HOSVD-rank $\mathbf{r} = (r_1, r_2, r_3)$ and a fixed number of measurements $m$ we performed 200 simulations. We say that an algorithm successfully recovers the original tensor $\mathbf{X}_0$ if the reconstruction $\mathbf{X}^{\#}$ satisfies $\left\|\mathbf{X}_0 - \mathbf{X}^{\#}\right\|_F < 10^{-3}$ for Gaussian measurement maps and Fourier measurement ensembles, and $\mathbf{X}^{\#}$ such that $\left\|\mathbf{X}_0 - \mathbf{X}^{\#}\right\|_F < 2.5 \cdot 10^{-3}$ for tensor completion. The algorithm stops in iteration $j$ if $\left\|\mathbf{X}^{j+1} - \mathbf{X}^{j}\right\|_F < 10^{-4}$ in which case we say that the algorithm converged, or it stops if it reached 5000 iterations.

A Gaussian linear mapping $\mathcal{A} : \mathbb{R}^{n_1 \times n_2 \times n_3} \to \mathbb{R}^m$ is defined by tensors $\mathbf{A}_k \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ via $[\mathcal{A}(\mathbf{X})](k) = \langle \mathbf{X}, \mathbf{A}_k \rangle$, for all $k \in [m]$, where the entries of the tensors $\mathbf{A}_k$ are i.i.d. Gaussian $\mathcal{N}\left(0, \frac{1}{m}\right)$. The tensor $\mathbf{X}^0 \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ of rank $\mathbf{r} = (r_1, r_2, r_3)$ is generated via its Tucker decomposition $\mathbf{X}^0 = \mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$: Each of the elements of the tensor $\mathbf{S}$ is taken independently from the normal distribution $\mathcal{N}(0,1)$, and the components $\mathbf{U}_k \in \mathbb{R}^{n_k \times r_k}$ are the first $r_k$ left singular vectors of a matrix $\mathbf{M}_k \in \mathbb{R}^{n_k \times n_k}$ whose elements are also drawn independently from the normal distribution $\mathcal{N}(0,1)$.

We have used the toolbox TensorLab [143] for computing the HOSVD decomposition of a given tensor and the truncation operator $\mathcal{H}_r$. By exploiting the Fast Fourier Transform (FFT), the measurement operator $\mathcal{A}$ from Section 5.3 related to the Fourier transform and its adjoint $\mathcal{A}^*$ can be applied efficiently which leads to reasonable run-times for comparably large tensor dimensions, see Table 5.1.

The numerical results for low-rank tensor recovery obtained via the NTIHT algorithm and Gaussian measurement maps are presented in Figures 5.3, 5.4, and 5.5. In Figure 5.3 and 5.4 we present the recovery results for low-rank tensors of size $10 \times 10 \times 10$. The horizontal axis represents the number of measurements taken with respect to the number of degrees of freedom of an arbitrary tensor of this size. To be more precise, for a tensor of size $n_1 \times n_2 \times n_3$, the number $\bar{n}$ on the horizontal axis represents $m = \left\lceil n_1 n_2 n_2 \frac{\bar{n}}{100} \right\rceil$ measurements. The vertical axis represents the percentage of successful recovery.

Finally, in Table 5.2 we present numerical results for third order tensor recovery via the CTIHT and the NTIHT algorithm. We consider Gaussian measurement maps, Fourier measurement ensembles, and tensor completion. With $m_0$ we denote the minimal number of measurements that are necessary to get full recovery and with $m_1$ we denote the maximal number of measurements for which we do not manage to recover any out of 200 tensors.

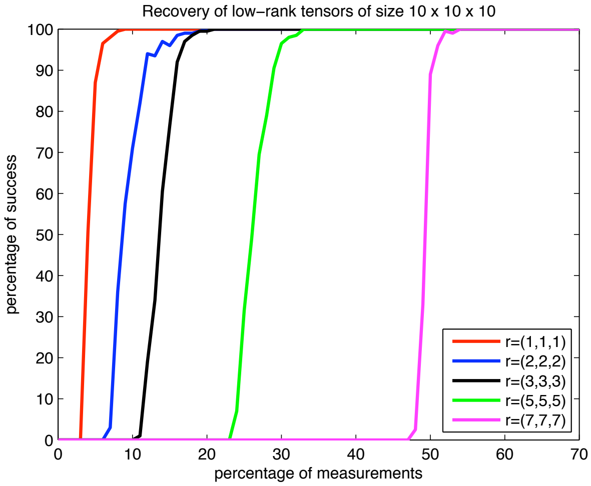


FIGURE 5.3. Recovery of low HOSVD-rank 10 x 10 x 10 tensors of rank $\mathbf{r} = (r, r, r)$ via NTIHT

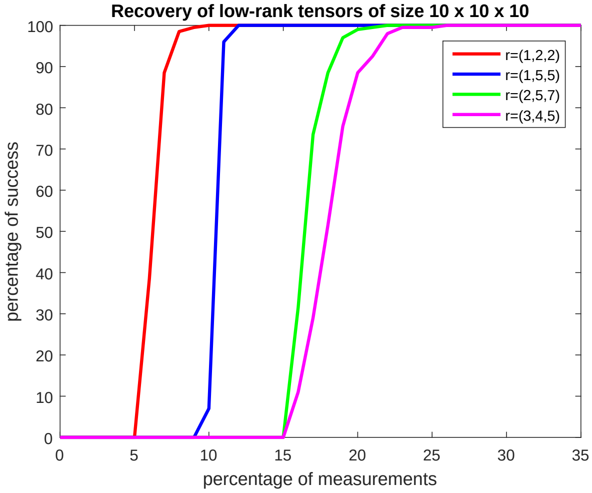


FIGURE 5.4. Recovery of low HOSVD-rank $10 \times 10 \times 10$ tensors of different unfolding ranks via NTIHT

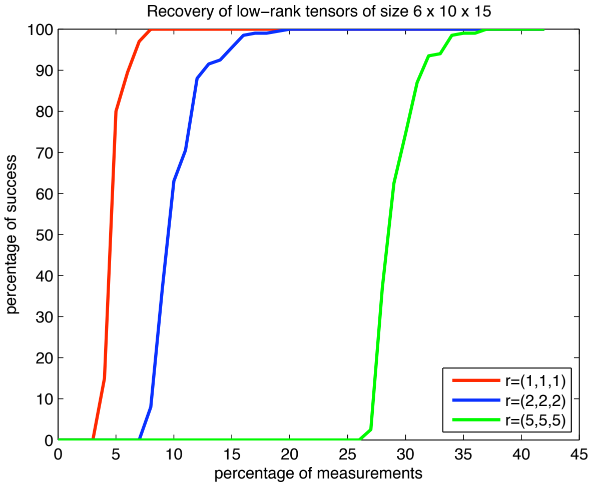


FIGURE 5.5. Recovery of low HOSVD-rank $6 \times 10 \times 15$ tensors of rank $\mathbf{r} = (r, r, r)$ via NTIHT

| type | tensor dimensions | rank | NTIHT-$\overline{n}_0$ | NTIHT-$\overline{n}_1$ | CTIHT-$\overline{n}_0$ | CTIHT-$\overline{n}_1$ |
|---|---|---|---|---|---|---|
| Gaussian | $10 \times 10 \times 10$ | $(1,1,1)$ | 8 | 3 | 24 | 6 |
| | $10 \times 10 \times 10$ | $(2,2,2)$ | 20 | 6 | 39 | 21 |
| | $10 \times 10 \times 10$ | $(3,3,3)$ | 21 | 11 | 60 | 40 |
| | $10 \times 10 \times 10$ | $(5,5,5)$ | 33 | 23 | $-$ | $-$ |
| | $10 \times 10 \times 10$ | $(7,7,7)$ | 53 | 47 | $-$ | $-$ |
| Gaussian | $10 \times 10 \times 10$ | $(1,2,2)$ | 10 | 5 | 34 | 16 |
| | $10 \times 10 \times 10$ | $(1,5,5)$ | 12 | 9 | 57 | 37 |
| | $10 \times 10 \times 10$ | $(2,5,7)$ | 20 | 15 | 83 | 64 |
| | $10 \times 10 \times 10$ | $(3,4,5)$ | 23 | 15 | 83 | 62 |
| Gaussian | $6 \times 10 \times 15$ | $(1,1,1)$ | 9 | 3 | 25 | 8 |
| | $6 \times 10 \times 15$ | $(2,2,2)$ | 20 | 7 | 44 | 27 |
| | $6 \times 10 \times 15$ | $(5,5,5)$ | 34 | 26 | $-$ | $-$ |
| Fourier | $10 \times 10 \times 10$ | $(1,1,1)$ | 16 | 3 | 15 | 8 |
| | $10 \times 10 \times 10$ | $(2,2,2)$ | 11 | 6 | 25 | 16 |
| | $10 \times 10 \times 10$ | $(3,3,3)$ | 16 | 14 | 31 | 26 |
| | $10 \times 10 \times 10$ | $(5,5,5)$ | 29 | 26 | 43 | 40 |
| | $10 \times 10 \times 10$ | $(7,7,7)$ | 51 | 48 | 50 | 49 |
| Fourier | $10 \times 10 \times 10$ | $(1,2,2)$ | 10 | 5 | 21 | 14 |
| | $10 \times 10 \times 10$ | $(1,5,5)$ | 16 | 12 | 31 | 25 |
| | $10 \times 10 \times 10$ | $(2,5,7)$ | 21 | 18 | 37 | 33 |
| | $10 \times 10 \times 10$ | $(3,4,5)$ | 21 | 18 | 37 | 33 |
| Fourier | $6 \times 10 \times 15$ | $(1,1,1)$ | 12 | 3 | 16 | 9 |
| | $6 \times 10 \times 15$ | $(2,2,2)$ | 13 | 8 | 25 | 20 |
| | $6 \times 10 \times 15$ | $(5,5,5)$ | 32 | 29 | 45 | 42 |
| completion | $10 \times 10 \times 10$ | $(1,1,1)$ | 17 | 2 | 27 | 2 |
| | $10 \times 10 \times 10$ | $(2,2,2)$ | 43 | 8 | 45 | 13 |
| | $10 \times 10 \times 10$ | $(3,3,3)$ | 37 | 12 | 32 | 16 |
| | $10 \times 10 \times 10$ | $(5,5,5)$ | 44 | 24 | 50 | 30 |
| | $10 \times 10 \times 10$ | $(7,7,7)$ | 71 | 46 | 84 | 54 |
| completion | $10 \times 10 \times 10$ | $(1,2,2)$ | 33 | 6 | 38 | 10 |
| | $10 \times 10 \times 10$ | $(1,5,5)$ | 57 | 15 | 58 | 21 |
| | $10 \times 10 \times 10$ | $(2,5,7)$ | 35 | 17 | 47 | 24 |
| | $10 \times 10 \times 10$ | $(3,4,5)$ | 36 | 17 | 41 | 22 |
| completion | $6 \times 10 \times 15$ | $(1,1,1)$ | 20 | 3 | 33 | 8 |
| | $6 \times 10 \times 15$ | $(2,2,2)$ | 47 | 10 | 51 | 14 |
| | $6 \times 10 \times 15$ | $(5,5,5)$ | 46 | 27 | 51 | 33 |

Table 5.2. Recovery results for low-rank matrix recovery via Gaussian measurement maps, Fourier measurement ensembles and tensor completion for NTIHT and CTIHT algorithm. An algorithm successfully recovers the sensed tensor $\mathbf{X}_0$ if it returns a tensor $\mathbf{X}^{\#}$ such that $\left\|\mathbf{X}_0 - \mathbf{X}^{\#}\right\|_F < 10^{-3}$ for Gaussian measurement maps and Fourier ensembles, and $\mathbf{X}^{\#}$ such that $\left\|\mathbf{X}_0 - \mathbf{X}^{\#}\right\|_F < 2.5 \cdot 10^{-3}$ for tensor completion. $\overline{n}_0$: minimal percentage of measurements needed to get hundred percent recovery; $\overline{n}_1$: maximal percentage of measurements for which recover is not successful for all out of 200 tensors. That is, the number of measurements is $m_i = \lceil n_1 n_2 n_3 \frac{\overline{n}_i}{100} \rceil$, for $i = 0, 1$. $-$ means that we did not manage to recover all 200 tensors with percentage of measurements less than $\overline{n} = 100$;

CHAPTER 6

# Conclusion and future work

In Chapter 1 we have introduced compressive sensing and low-rank matrix recovery. We have focused mostly on convex optimization approaches and on iterative hard thresholding algorithm since in Chapter 4 and Chapter 5 we have extended these approaches to low-rank tensor recovery.

In Chapter 2 we have introduced different tensor formats. Several tensor properties have been discussed which cause significant difficulties in the analyses of algorithms for low-rank tensor recovery. In particular, the CP-decomposition (canonical decomposition), which could be considered as a natural generalization of the matrix singular value decomposition, is in general NP-hard to compute. Consequently, the CP-rank and the corresponding tensor nuclear norm are also NP-hard to compute. This has led to the development of other tensor formats – Tucker (HOSVD) format, TT-format, and HT-format. For these decompositions, the tensor rank is a vector and a well defined quantity. That is, the entries of the ranks equal to the ranks of the corresponding tensor matricizations. Unfortunately, computing the best rank-$\mathbf{r}$ approximation of a given tensor remains NP-hard – regardless of the choice of tensor format.

In Chapter 3 several previous approaches to low-rank tensor recovery have been introduced with the corresponding theoretical results. Unfortunately, none of the approaches are complete from both applicational and theoretical point of view. More precisely, either the methods are not tractable, or the recovery results quantifying the minimal number of measurements are non-optimal or even non-existent.

A new convex optimization approach to low-rank tensor recovery has been introduced in Chapter 4. This approach is based on theta bodies of the appropriately defined polynomial ideal which induce new tensor norms called theta norms. The $\theta_k$-norm of a given tensor can be computed via semidefinite programming. Additionally, a unit-$\theta_k$-norm ball is a superset of the unit-tensor-nuclear-norm ball for all $k$ and the theta norms satisfy $\|\mathbf{X}\|_{\theta_1} \leq \|\mathbf{X}\|_{\theta_2} \leq \cdots \leq \|\mathbf{X}\|_{\theta_{k-1}} \leq \|\mathbf{X}\|_{\theta_k} \leq \cdots \leq \|\mathbf{X}\|_*$, for all $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$. We have shown that in the matrix case, all $\theta_k$-norms coincide with the matrix nuclear norm. We have provided a semidefinite program for computing the $\theta_1$-norm of a given order-three tensor as well as the semidefinite program for low-rank order-three tensor recovery via $\theta_1$-norm minimization. The numerical results presented in Section 4.5 indicate that the low-rank tensor recovery via $\theta_1$-norm is a promising approach. In the future we would like to describe the boundary of the unit-$\theta_k$-norm balls, to develop faster algorithms for computing $\theta_k$-norm of a given tensor and for recovery of low-rank tensors via the same norm, and to quantify the number of measurements needed for low-rank tensor recovery via $\theta_k$-norm minimization.

The theta-body approach requires computing the reduced Gröbner basis with respect to the grevlex ordering of the polynomial ideal $\mathcal{J}_d$ whose real algebraic variety is the set of all order-$d$ rank-one Frobenius-norm-one tensors. In the matrix scenario, the polynomial ideal $\mathcal{J}_2$ is related to the determinantal ideal often denoted in literature by $\mathcal{I}_2$. Recall that the real algebraic variety of the determinantal ideal $\mathcal{I}_t$ with $t \geq 2$ is the set of all rank-(t-1) matrices. Thus, in the tensor scenario, the ideal $\mathcal{I}_{2,d}$ generated by all order-two minors of all tensor matricizations (i.e., the real algebraic variety of $\mathcal{I}_{2,d}$ is the set of all rank-one tensors) could be considered as a natural generalization of determinantal ideal $\mathcal{I}_2$. The polynomial ideals $\mathcal{I}_{2,d}$ to the best of our knowledge have not been considered before. (One can also generalize in an analogous way determinantal ideals $\mathcal{I}_t$ with $t \geq 3$.) We have also computed the reduced Gröbner basis with respect to the grevlex ordering of the polynomial ideal $\mathcal{I}_{2,d}$. Determinatal ideals have been the central topic throughout the last three decades in both algebraic geometry and commutative algebra. In future, we would like to extend the results of determinatal ideals $\mathcal{I}_t$ to the higher-order determinantal ideals $\mathcal{I}_{t,d}$ that naturally arose in our research.

In Chapter 5 we have introduced and analyzed several versions of the iterative hard thresholding (IHT) algorithm – namely, classical tensor iterative hard thresholding algorithm (CTIHT) and its normalized version (NTIHT) – adapted to the tensor decomposition at hand. Here, the Tucker, TT, and HT decompositions have been considered. The TIHT algorithms are iterative thresholding based methods to low-rank tensor recovery. The analysis of these algorithms is based on the corresponding notion of the tensor restricted isometry property (TRIP). We have proved that partial Fourier maps combined with random sign flips of the tensor entries and subgaussian measurement ensembles satisfy TRIP with high probability. Under the assumption that the measurement map satisfies TRIP, we have provided a partial convergence of the TIHT algorithms. More precisely, we proved that the algorithms converge linearly if the thresholding operator satisfies a specific condition which can not be guaranteed a priori. This condition is required since the best rank-$\mathbf{r}$ approximation of a given tensor is in general NP-hard to compute and the thresholding operator computes only its quasi-best rank-$\mathbf{r}$ approximation. In spite of this additional assumption on the thresholding operator needed for the theoretical guarantees, our numerical experiments for third-order low-HOSVD-rank tensor recovery suggest that TIHT algorithms perform well in practice. Providing theoretical guarantees for tensor completion as well as for the complete convergence of the TIHT algorithms is left for future research.

CHAPTER 7

# Summary

In this thesis we have considered a further extension of compressive sensing and low-rank matrix recovery to low-rank tensor recovery. The aim is to reconstruct an order-$d$ low-rank tensor from a number of linear measurements much smaller than its ambient dimension. As expected, the natural approach of finding the solution of the optimization problem

$$\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}} \operatorname{rank}(\mathbf{Z}) \quad \text{s.t.} \quad \mathcal{A}(\mathbf{Z}) = \mathbf{y},$$

where $\mathcal{A} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^m$ is a known linear operator and $\mathbf{y} \in \mathbb{R}^m$ is a measurement vector with $m \ll n_1 n_2 \cdots n_d$, is in general NP-hard. Although several approaches to low-rank tensor recovery have already been suggested, due to the certain tensor properties there is no completely satisfactory theory available for these methods. Either the method is not tractable, or the recovery results quantifying the minimal number of measurements are non-optimal or even non-existent.

We have presented two new approaches to low-rank tensor recovery. The first approach was a convex optimization approach and could be considered as a tractable extension of $\ell_1$-minimization for sparse vector recovery and nuclear norm minimization for matrix recovery to tensor scenario. It is based on theta bodies – a recently introduced tool from real algebraic geometry. This approach required computing the reduced Gröbner basis of the polynomial ideal $\mathcal{J}_d$ in $\mathbb{R}[x_{11\ldots1}, x_{11\ldots2}, \ldots, x_{n_1 n_2 \ldots n_d}]$ with respect to the graded reverse lexicographic (grevlex) ordering. The corresponding real algebraic variety $\nu_{\mathbb{R}}(\mathcal{J}_d) = \{\mathbf{x} : f(\mathbf{x}) = 0, \text{ for all } f \in \mathcal{J}_d\}$ is the set of all rank-one Frobenius-norm-one tensors in $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$. We have treated each variable as a tensor entry. We have considered the canonical format and the corresponding tensor nuclear norm which are in general NP-hard to compute. Theta bodies $\mathrm{TH}_k(\mathcal{J}_d)$ are closed convex sets containing the closure of the convex hull of $\nu_{\mathbb{R}}(\mathcal{J}_d)$ denoted by $\overline{\operatorname{conv}(\nu_{\mathbb{R}}(\mathcal{J}_d))}$ and they satisfy

$$\mathrm{TH}_1(\mathcal{J}_d) \supseteq \mathrm{TH}_2(\mathcal{J}_d) \supseteq \cdots \supseteq \mathrm{TH}_{k-1}(\mathcal{J}_d) \supseteq \mathrm{TH}_k(\mathcal{J}_d) \supseteq \cdots \supseteq \operatorname{conv}(\nu_{\mathbb{R}}(\mathcal{J}_d)).$$

Since $\operatorname{conv}(\nu_{\mathbb{R}}(\mathcal{J}_d))$ is the unit-tensor-nuclear-norm-ball, every theta body provides its convex closed relaxation. This has allowed us to define new tensor norms – $\theta_k$-norms – via their unit norm balls. That is,

$$\left\{\mathbf{X} : \|\mathbf{X}\|_{\theta_k} \leq 1\right\} = \mathrm{TH}_k(\mathcal{J}_d), \quad \text{for all } k \in \mathbb{N}.$$

All $\theta_k$-norms can be computed via semidefinite programming. However, for simplicity, we have provided only a semidefinite program for computing the $\theta_1$-norm of a given order-3 tensor. We have shown that in the matrix scenario (i.e., when $d = 2$) all $\theta_k$-norms are equal and coincide with the matrix nuclear norm. In the tensor scenario, however, we have obtained – to the best of our knowledge – new tensor norms. We have also provided a semidefinite program for low-rank tensor recovery via $\theta_1$-norm minimization, i.e., a semidefinite program for

$$\min_{\mathbf{Z}\in\mathbb{R}^{n_1\times n_2\times n_3}} \|\mathbf{Z}\|_{\theta_1} \quad \text{s.t.} \quad \mathcal{A}(\mathbf{Z})=\mathbf{y}.$$

In our numerical experiments we have always recovered rank-one and rank-two tensors of order-3 via Gaussian measurement ensembles from a number of measurements significantly smaller than the ambient dimension of the corresponding tensor. Therefore, low-rank tensor recovery via $\theta_1$-norm minimization seems to be a promising approach. Additionally, we have also briefly considered the ideal $\mathcal{I}_d$ whose real algebraic variety contains all rank-one order-$d$ tensors (not necessarily of Frobenius-norm-one). These ideals that – to the best of our knowledge – have not been considered before could be considered as a natural higher-order generalization of the determinantal ideal $\mathcal{I}_2$. We have gone one step further and defined generalized determinantal ideals $\mathcal{I}_{t,d}$ whose real algebraic variety contains all rank-$(t-1)$ order-$d$ tensors. We call these ideals higher-order determinantal ideals. We have also computed the reduced Gröbner basis $\boldsymbol{\mathcal{G}}_{2,d}$ of the polynomial ideal $\mathcal{I}_{2,d}$ with respect to the grevlex ordering. To be more precise, we have shown that $\boldsymbol{\mathcal{G}}_{2,d}=\boldsymbol{\mathcal{G}}_2\backslash\{g_d=\sum_{i_1,i_2,\ldots,i_d} x^2_{i_1 i_2\ldots i_d}-1\}$, where $\boldsymbol{\mathcal{G}}_d$ is the reduced Gröbner basis of the ideal $\mathcal{J}_d$ with respect to the grevlex ordering. This is – to the best of our knowledge – the first result available for the higher-order determinantal ideals $\mathcal{I}_{t,d}$ with $d\geq 3$.

The second approach was a generalization of the iterative hard thresholding algorithm (IHT algorithm) for sparse vector and low-rank matrix recovery to the tensor scenario (tensor IHT or TIHT algorithm). We have considered the Tucker format, the tensor train (TT) decomposition, and the hierarchical Tucker (HT) decomposition. The crucial step of the IHT algorithms consists in taking a projection onto the set of sparse vectors or the manifold of low-rank matrices/tensors. Unlike in the vector and the matrix scenario, it is NP-hard to compute the best rank-$\mathbf{r}$ approximation of a given tensor – regardless of the choice of tensor format. Even more, sometimes the best rank-$\mathbf{r}$ approximation does not even exist. Therefore, in this step of the algorithm we have computed a quasi-best rank-$\mathbf{r}$ approximation $\mathcal{H}_{\mathbf{r}}(\mathbf{X})$ of a given tensor $\mathbf{X}$ which can be done efficiently. The analysis of the algorithm was based, similarly to the vector and the matrix scenario, on the version of the restricted isometry property (tensor RIP or TRIP) adapted to the tensor decomposition at hand. We have showed that subgaussian measurement ensembles satisfy TRIP with high probability under an almost optimal condition on the number of measurements. We have also proved that partial Fourier maps combined with random sign flips of the tensor entries satisfy TRIP with high probability. Under the assumption that the linear operator satisfies TRIP and under an additional assumption on the thresholding operator we have provided a linear convergence result for the TIHT algorithm. In spite of the additional condition on the thresholding operator (which can not be guaranteed a priori) required for theoretical guarantees, our numerical results indicated that the algorithm performs well in practice. That is, we have always recovered a low-HOSVD-rank third-order tensor via Partial Fourier maps combined with random sign flips of tensor entries, tensor completion, and Gaussian measurement ensembles from a much smaller number of measurements than its ambient dimension.

# APPENDIX A

In this appendix we collect several basic definitions and results related to matrices (Subsection A.1), tensors (Subsection A.2), and covering numbers (Subsection A.3).

Before passing to (random) matrices, we present a general estimate used in Subsection 4.4.

**Lemma A.1** ([60])**.** Integers $n \geq k > 0$ satisfy

$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k.$$

Proof. First notice that for $k \in \mathbb{N}$ it holds that

$$e^k = \sum_{\ell=0}^{\infty} \frac{k^\ell}{\ell!} \geq \frac{k^k}{k!}.$$

The estimate then follows from

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} \leq \frac{n^k}{k!} = \frac{k^k}{k!}\frac{n^k}{k^k} \leq \left(\frac{en}{k}\right)^k.$$

□

## A.1. (Random) matrices

**Definition A.2** (Kronecker product)**.** Let $\mathbf{A} \in \mathbb{C}^{m \times n}$ and $\mathbf{B} \in \mathbb{C}^{p \times q}$. The Kronecker product of two matrices $\mathbf{A}$ and $\mathbf{B}$ is the matrix $\mathbf{A} \otimes_K \mathbf{B} \in \mathbb{C}^{mp \times nq}$ defined by

$$\mathbf{A} \otimes_K \mathbf{B} = \begin{pmatrix} \mathbf{A}(1,1)\mathbf{B} & \mathbf{A}(1,2)\mathbf{B} & \cdots & \mathbf{A}(1,n)\mathbf{B} \\ \mathbf{A}(2,1)\mathbf{B} & \mathbf{A}(2,2)\mathbf{B} & \cdots & \mathbf{A}(2,n)\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}(m,1)\mathbf{B} & \mathbf{A}(m,2)\mathbf{B} & \cdots & \mathbf{A}(m,n)\mathbf{B} \end{pmatrix}.$$

The Fourier matrix $\mathbf{F}_n \in \mathbb{C}^{n \times n}$ is the most important complex matrix in applied mathematics since it is used to define the Fourier transform. The fast Fourier transform (FFT) reduces the multiplication by $\mathbf{F}_n$ from $n^2$ to roughly $n(\log_2 n)$ multiplications.

The Fourier matrix $\mathbf{F}_n \in \mathbb{C}^{n \times n}$ is defined element-wise as $\mathbf{F}_n(j,k) = \frac{1}{\sqrt{n}}\omega^{(j-1)(k-1)}$, where $\omega = e^{-\frac{2\pi i}{n}}$, or equivalently

$$\mathbf{F}_n = \frac{1}{\sqrt{n}} \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(n-1)} \\ 1 & \omega^3 & \omega^6 & \cdots & \omega^{3(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{n-1} & \omega^{2(n-1)} & \cdots & \omega^{(n-1)(n-1)} \end{pmatrix}.$$

The vector $\hat{\mathbf{x}} = \mathbf{F}_n\mathbf{x}$ is called the discrete Fourier transform of $\mathbf{x}$. Matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ obtained by choosing $m$ rows independently and uniformly at random from $\mathbf{F}_n$ is called random partial Fourier matrix.

In compressive sensing, instead of the Fourier matrix, often the *nonequispaced* Fourier matrix is used. That is, a matrix $\hat{\mathbf{F}}_n := \sqrt{n}\mathbf{F}_n$.

Partial circulant matrices are another class of structured random matrices. In such set-up for a vector $\mathbf{c} = (c_0, c_1, \ldots, c_{n-1})^T \in \mathbb{C}^n$, the $n \times n$ circulant matrix $\mathbf{A} = \mathbf{A}(\mathbf{c})$ is of the form

$$\mathbf{A} = \begin{pmatrix} c_0 & c_{n-1} & \cdots & c_2 & c_1 \\ c_1 & c_0 & \cdots & c_3 & c_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{n-2} & c_{n-3} & \cdots & c_0 & c_{n-1} \\ c_{n-1} & c_{n-2} & \cdots & c_1 & c_0 \end{pmatrix}.$$

Partial circulant matrix $\mathbf{A}^{\boldsymbol{\Omega}} = \mathbf{A}^{\boldsymbol{\Omega}}(\mathbf{c}) \in \mathbb{C}^{m\times N}$ is a submatrix of $\mathbf{A}$ consisting of the rows indexed by $\boldsymbol{\Omega}$. Additionally, if $\mathbf{c}$ is a Rademacher vector $\varepsilon$ (a vector of independent random variables taking $+1$ and $-1$ with equal probability), then $\mathbf{A}^{\boldsymbol{\Omega}}(\varepsilon)$ is called random partial circulant matrix.

We remark that circulant matrices can be diagonalized using the discrete Fourier transform. Thus, there exists an algorithm – using FFT – such that the (circulant) matrix-vector multiplication is of complexity $\mathcal{O}(n \log n)$.

In the following we focus on a more general type of structured matrices called *SORS* (*Subsampled Orthogonal with Randomized Signs*) *matrices* that have been studied in [127]. We also present their main theorem that we apply in Chapter 5.

**Definition A.3** ([127])**.** Let $\mathbf{F} \in \mathbb{R}^{N\times N}$ denote a matrix obeying

$$\mathbf{F}^*\mathbf{F} = \mathbf{I} \quad \text{and} \quad \max_{i,j} |\mathbf{F}(i,j)| \leq \frac{\Delta}{\sqrt{N}}. \tag{A.1}$$

Define the random subsampled matrix $\mathbf{H} \in \mathbb{R}^{m\times N}$ with i.i.d. (independently identically distributed) rows chosen uniformly at random from the rows of $\mathbf{F}$. Now we define the Subsampled Orthogonal with Random Signs (SORS) measurement ensemble as $\mathbf{A} = \mathbf{HD}$, where $\mathbf{D} \in \mathbb{R}^{N\times N}$ is a random diagonal matrix with the diagonal entries i.i.d. $\pm 1$ with equal probability.

For example, an $N \times N$ nonequispaced Fourier matrix and (normalized) Fourier matrix satisfy the conditions (A.1) in the above definition with $\Delta = \sqrt{N}$ and $\Delta = 1$, respectively. To present the theorem, we need to introduce some notation. With $d_2(\boldsymbol{\mathcal{T}})$ we denote the diameter of a given set $\boldsymbol{\mathcal{T}}$, i.e., $d_2(\boldsymbol{\mathcal{T}}) = \sup_{\mathbf{v}\in\boldsymbol{\mathcal{T}}} \|\mathbf{v}\|_2$. The Gaussian width $\omega(\boldsymbol{\mathcal{T}})$ of a set $\boldsymbol{\mathcal{T}}$ is defined as $\omega(\boldsymbol{\mathcal{T}}) = \mathbb{E}\left[\sup_{\mathbf{v}\in\boldsymbol{\mathcal{T}}} \mathbf{g}^T\mathbf{v}\right]$, where $\mathbf{g} \in \mathbb{R}^N$ is a standard Gaussian random vector.

**Theorem A.4** ([127])**.** Let $\boldsymbol{\mathcal{T}} \subset \mathbb{R}^N$ and suppose $\widetilde{\mathbf{A}} = \sqrt{\frac{N}{m}}\mathbf{A}$, where $\mathbf{A} \in \mathbb{R}^{m\times N}$ is selected from the SORS distribution of Definition A.3. Then,

$$\sup_{\mathbf{x}\in\boldsymbol{\mathcal{T}}} \left| \left\|\widetilde{\mathbf{A}}\mathbf{x}\right\|_2^2 - \|\mathbf{x}\|_2^2 \right| \leq \max\left\{\delta, \delta^2\right\} \cdot \left(d_2(\boldsymbol{\mathcal{T}})\right)^2,$$

holds with probability at least $1 - 2e^{-\eta}$ as long as

$$m \geq C\delta^{-2}\Delta^2 (1+\eta)^2 (\log N)^4 \max\left\{1, \frac{\omega^2(\boldsymbol{\mathcal{T}})}{\left(d_2(\boldsymbol{\mathcal{T}})\right)^2}\right\}.$$

## A.2. Tensors

The following lemma shows a way of decomposing the HOSVD-rank $2\mathbf{r}$ $d$th-order tensor in a sum of pairwise orthogonal (entry-wise) at most HOSVD-rank $\mathbf{r}$ tensors.

**Lemma A.5.** Any tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ of HOSVD-rank $2\mathbf{r} = (2r_1, 2r_2, \ldots, 2r_d)$ can be decomposed into a set of $2^d$ tensors $\{\mathbf{X}_i\}_{i=1}^{2^d}$ s.t.

$$\begin{aligned}
&\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_{2^d}, \\
&\operatorname{rank}_{\mathrm{HOSVD}}(\mathbf{X}_p) \leq \mathbf{r}, \text{ for all } p \in \left[2^d\right] \\
&\langle \mathbf{X}_p, \mathbf{X}_q \rangle = 0, \text{ whenever } p \neq q.
\end{aligned}$$

Proof. The HOSVD decomposition of a tensor $\mathbf{X}$ is of the form

$$\mathbf{X} = \mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times \cdots \times_d \mathbf{U}_d,$$

where $\mathbf{S} \in \mathbb{R}^{2r_1 \times 2r_2 \times \cdots \times 2r_d}$ and $\mathbf{U}_i \in \mathbb{R}^{n_i \times 2r_i}$, for all $i \in [d]$.

Let

$$\begin{aligned}
&\mathbf{U}_i^0(:, [r_i]) = \mathbf{U}_i(:, [r_i]), && \mathbf{U}_i^0(:, [2r_i] \setminus [r_i]) = \mathbf{0} \\
&\mathbf{U}_i^1(:, [r_i]) = \mathbf{0}, && \mathbf{U}_i^1(:, [2r_i] \setminus [r_i]) = \mathbf{U}_i(:, [2r_i] \setminus [r_i]), \text{ for all } i \in [d].
\end{aligned}$$

In other words, let $\mathbf{U}_i^0$ be a matrix identical to $\mathbf{U}_i$ on the first $r_i$ columns and zero otherwise and let $\mathbf{U}_i^1$ be a matrix identical to $\mathbf{U}_i$ on the last $r_i$ columns and zero otherwise.

Notice that, since $\mathbf{U}_i = \mathbf{U}_i^0 + \mathbf{U}_i^1$ for all $i \in [d]$,

$$\mathbf{X} = \mathbf{S} \times_1 \left(\mathbf{U}_1^0 + \mathbf{U}_1^1\right) \times_2 \left(\mathbf{U}_2^0 + \mathbf{U}_2^1\right) \times \cdots \times_d \left(\mathbf{U}_d^0 + \mathbf{U}_d^1\right).$$

Next, define

$$\mathbf{X}_p := \mathbf{X}_{p_1 \cdot 2^{d-1} + \cdots + p_{d-1} \cdot 2 + p_d + 1} = \mathbf{S} \times_1 \mathbf{U}_1^{p_1} \times_2 \mathbf{U}_2^{p_2} \times \cdots \times_d \mathbf{U}_d^{p_d}, \text{ for all } p_1, p_2, \ldots, p_d \in \{0, 1\}.$$

Then $\operatorname{rank}(\mathbf{X}_p) \leq \mathbf{r}$, for all $p$, and

$$\langle \mathbf{X}_p, \mathbf{X}_q \rangle = \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_d=1}^{n_d} \mathbf{X}_p(i_1, i_2, \ldots, i_d)\, \mathbf{X}_q(i_1, i_2, \ldots, i_d) = 0, \text{ whenever } p \neq q.$$

We only show that $\mathbf{X}_1$ has rank at most $\mathbf{r}$ since the proofs for all other $\mathbf{X}_p$ with $p \in \left[2^d\right]$ are analogous. Recall that

$$\mathbf{X}_1 = \mathbf{S} \times_1 \mathbf{U}_1^0 \times_2 \mathbf{U}_2^0 \times \cdots \times_d \mathbf{U}_d^0,$$

or elementwise

$$\begin{aligned}
\mathbf{X}_1(i_1, i_2, \ldots, i_d) &= \sum_{j_1=1}^{2r_1} \sum_{j_2=1}^{2r_2} \cdots \sum_{j_d=1}^{2r_d} \mathbf{S}(j_1, j_2, \ldots, j_d)\, \mathbf{U}_1^0(i_1, j_1)\, \mathbf{U}_2^0(i_2, j_2) \cdots \mathbf{U}_d^0(i_d, j_d) \\
&= \sum_{j_1=1}^{r_1} \sum_{j_2=1}^{r_2} \cdots \sum_{j_d=1}^{r_d} \mathbf{S}(j_1, j_2, \ldots j_d)\, \mathbf{U}_1(i_1, j_1)\, \mathbf{U}_2(i_2, j_2) \cdots \mathbf{U}_d(i_d, j_d).
\end{aligned}$$

First, notice that the above decomposition is not necessarily its HOSVD (for example, the tensor $\mathbf{S}([r_1], [r_2], \ldots, [r_d])$ does not have to be all-orthogonal). The first unfolding of $\mathbf{X}_1$ is of the form

$$\mathbf{X}_1^{\{1\}}\left(i_1;(i_2,i_3,\ldots,i_d)\right)=\mathbf{X}_1\left(i_1,i_2,i_3,\ldots,i_d\right)$$
$$=\sum_{j_1=1}^{r_1}\mathbf{U}_1\left(i_1,j_1\right)\sum_{j_2=1}^{r_2}\sum_{j_3=1}^{r_3}\cdots\sum_{j_d=1}^{r_d}\mathbf{S}\left(j_1,j_2,j_3,\ldots,j_d\right)\mathbf{U}_2\left(i_2,j_2\right)\mathbf{U}_3\left(i_3,j_3\right)\cdots\mathbf{U}_d\left(i_d,j_d\right).$$

Define the matrix $\mathbf{M}_1\in\mathbb{R}^{r_1\times n_2n_3\cdots n_d}$ element-wise as

$$\mathbf{M}_1\left(j_1;(i_2,i_3,\ldots,i_d)\right):=\sum_{j_2=1}^{r_2}\sum_{j_3=1}^{r_3}\cdots\sum_{j_d=1}^{r_d}\mathbf{S}\left(j_1,j_2,j_3,\ldots,j_d\right)\mathbf{U}_2^0\left(i_2,j_2\right)\mathbf{U}_3^0\left(i_3,j_3\right)\cdots\mathbf{U}_d^0\left(i_d,j_d\right)$$

and matrix $\overline{\mathbf{U}}_1\in\mathbb{R}^{n_1\times r_1}$ as a submatrix of $\mathbf{U}_1$ containing the first $r_1$ columns. Then we can write

$$\mathbf{X}_1^{\{1\}}=\overline{\mathbf{U}}_1\mathbf{M}_1.$$

Since $\operatorname{rank}\left(\overline{\mathbf{U}}_1\right)\leq r_1$, we deduce that $\operatorname{rank}\left(\mathbf{X}_1^{\{1\}}\right)\leq r_1$. Similarly, we can show that $\operatorname{rank}\left(\mathbf{X}_1^{\{k\}}\right)\leq r_k$, for all the other $k\in[d]$, which proves the first statement.

For the second statement of the theorem notice that

$$\langle\mathbf{X}_p,\mathbf{X}_q\rangle=\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\cdots\sum_{i_d=1}^{n_d}\mathbf{X}_p\left(i_1,i_2,\ldots,i_d\right)\mathbf{X}_q\left(i_1,i_2,\ldots,i_d\right)$$
$$=\sum_{j_1,j_2,\ldots,j_d}\sum_{k_1,k_2,\ldots,k_d}\mathbf{S}\left(j_1,j_2,\ldots,j_d\right)\mathbf{S}\left(k_1,k_2,\ldots,k_d\right)\sum_{i_1}\mathbf{U}_1^{p_1}\left(i_1,j_1\right)\mathbf{U}_1^{q_1}\left(i_1,k_1\right)$$
$$\cdot\sum_{i_2}\mathbf{U}_2^{p_2}\left(i_2,j_2\right)\mathbf{U}_2^{q_2}\left(i_2,k_2\right)\cdots\sum_{i_d}\mathbf{U}_d^{p_d}\left(i_d,j_d\right)\mathbf{U}_d^{q_d}\left(i_d,k_d\right).$$

Since $p\neq q$, for at least one $i$, $p_i\neq q_i$. Without loss of generality, we can assume that $i=1$ and $p_1=0,q_1=1$. From the definition of $\mathbf{U}_1^0$ and $\mathbf{U}_1^1$ it is clear that $\sum_{i_1}\mathbf{U}_1^0\left(i_1,j_1\right)\mathbf{U}_1^1\left(i_1,k_1\right)=0$, for all $j_1,k_1\in[2r_1]$ which proves the claim. □

In the following, three algorithms for HT-truncation suggested in [70] are presented. In the following, as in paper [70], we consider tensors as vectors over product index sets. For this purpose we introduce

$$I:=\boldsymbol{\mathcal{I}}_1\times\boldsymbol{\mathcal{I}}_2\times\cdots\times\boldsymbol{\mathcal{I}}_d,\quad\boldsymbol{\mathcal{I}}_\mu:=\{1,2,\ldots,n_\mu\}\quad(\text{with }\mu\in[d])$$

and we write that tensor $\mathbf{X}\in\mathbb{R}^I$. The $I$ defined above corresponds to the subscript of the corresponding dimensional tree $T_I$.

**Theorem A.6** ([70]). Let $T_I$ be a dimension tree and $\mathbf{X}\in\mathbb{R}^I$. Let $\mathbf{X}_{\mathrm{BEST}}$ denote the best HT-rank $\mathbf{r}=\{r_t\}_{t\in T_I}$ approximation of $\mathbf{X}$ and let $\pi_t$ be the orthogonal frame projection for the $t$-frame $\mathbf{U}_t$ that consists of the left singular vectors of $\mathbf{X}^t$ corresponding to the $r_t$ largest singular values $\sigma_{t,i}$ of $\mathbf{X}^t$ (i.e., $\left(\pi_t\mathbf{X}\right)^t:=\mathbf{U}_t\mathbf{U}_t^T\mathbf{X}^t$ and $\pi_{\{1,\ldots,d\}}\mathbf{X}=\mathbf{X}$). Then for any order of the projections $\pi_t$, $t\in T_I$, the following holds

$$\left\|\mathbf{X}-\prod_{t\in T_I}\pi_t\mathbf{X}\right\|_F\leq\sqrt{\sum_{t\in T_I}\sum_{i>r_t}\sigma_{t,i}^2}\leq\sqrt{2d-2}\left\|\mathbf{X}-\mathbf{X}_{\mathrm{BEST}}\right\|_F.$$

Additionally, for the root node $t=\{1,\ldots,d\}$ with sons $t_1$ and $t_2$, combining the projections $\pi_{t_1}$ and $\pi_{t_2}$ into a single projection via the SVD, leads to the improved bound $\sqrt{2d-3}$ (instead of $\sqrt{2d-2}$).

We call the above algorithm *truncation via projections.*

We explain the notation in Algorithm A.1 and Algorithm A.2. The $i$-th column of a matrix $\mathbf{U}$ is denoted with $\mathbf{U}(:,i)$. Also, recall that $\boldsymbol{\mathcal{L}}(T_I)$ and $\boldsymbol{\mathcal{I}}(T_I)$ denote the set of all leaf nodes and all interior nodes of the tree $T_I$, respectively.

**Algorithm A.1.** root-to-leaves truncation of arbitrary tensors to HT-format

```
Input: tensor $\mathbf{X} \in \mathbb{R}^{I}$, dimension tree $T_I$ (depth $p > 0$),
        target representation rank $(r_t)_{t \in T_I}$;
for each singleton $\alpha \in \boldsymbol{\mathcal{L}}(T_I)$ do
    Compute an SVD of $\mathbf{X}^{\alpha}$ and store the dominant $r_\alpha$ left singular vectors in the
    columns of the $\alpha$-frame $\mathbf{U}_\alpha$.
end for
for $\ell = p-1, \ldots, 0$ do
    for each mode cluster $t \in \boldsymbol{\mathcal{I}}(T_I)$ on level $\ell$ do
        Compute an SVD of $\mathbf{X}^{t}$ and store the dominant $r_t$ left singular vectors in the
        columns of the $t$-frame $\mathbf{U}_t$.
        Let $\mathbf{U}_{t_1}$ and $\mathbf{U}_{t_2}$ denote the frames for the successors of $t$ on level $\ell + 1$.
        Compute the entries of the transfer tensor:
            $\mathbf{B}_t(i,j,\nu) := \langle \mathbf{U}_t(:,i), \mathbf{U}_{t_1}(:,j) \otimes \mathbf{U}_{t_2}(:,\nu) \rangle$
    end for
end for
Compute the entries of the root (with sons $t_1$, $t_2$) transfer tensors
        $\mathbf{B}_{\{1,\ldots,d\}}(1,j,\nu) := \langle \mathbf{X}^{\{1,2,\ldots,d\}}, \mathbf{U}_{t_1}(:,j) \otimes \mathbf{U}_{t_2}(:,\nu) \rangle$
return HT-rank-$\mathbf{r} = \{r_t\}_{t \in T_I}$ approximation $\mathbf{X}_{HT}$: $\left( \{\mathbf{U}_t\}_{t \in \boldsymbol{\mathcal{L}}(T_I)}, \{\boldsymbol{\mathcal{B}}_t\}_{t \in \boldsymbol{\mathcal{I}}(T_I)} \right)$.
```

## A.3. Covering numbers

The proofs of several theorems in Chapter 5 use $\varepsilon$-nets and covering numbers.

**Definition A.7** ([163])**.** A set $\mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{X}}} \subset \boldsymbol{\mathcal{X}}$, where $\boldsymbol{\mathcal{X}}$ is a subset of a normed space, is called an $\varepsilon$-net of $\boldsymbol{\mathcal{X}}$ with respect to the norm $\|\cdot\|$ if for each $v \in \boldsymbol{\mathcal{X}}$, there exists $v_0 \in \mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{X}}}$ with $\|v_0 - v\| \leq \varepsilon$. The minimal cardinality of an $\varepsilon$-net of $\boldsymbol{\mathcal{X}}$ with respect to the norm $\|\cdot\|$ is denoted by $\mathcal{N}(\boldsymbol{\mathcal{X}}, \|\cdot\|, \varepsilon)$ and is called the covering number of $\boldsymbol{\mathcal{X}}$ (at scale $\varepsilon$).

In Chapter 5 the following result is used frequently.

**Lemma A.8** ([163])**.** Let $\varepsilon \in (0,1)$. For any set $\boldsymbol{\mathcal{X}}$ there always exists an $\varepsilon$-net $\mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{X}}}$ with respect to a norm $\|\cdot\|$ satisfying $\mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{X}}} \subset \boldsymbol{\mathcal{X}}$ and

$$\left|\mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{X}}}\right| \leq \frac{\operatorname{Vol}\left(\boldsymbol{\mathcal{X}} + \frac{\varepsilon}{2}\boldsymbol{\mathcal{B}}\right)}{\operatorname{Vol}\left(\frac{\varepsilon}{2}\boldsymbol{\mathcal{B}}\right)},$$

where $\frac{\varepsilon}{2}\boldsymbol{\mathcal{B}}$ is an $\varepsilon/2$ ball with respect to the norm $\|\cdot\|$ and $\boldsymbol{\mathcal{X}} + \frac{\varepsilon}{2}\boldsymbol{\mathcal{B}} = \left\{x + y : x \in \mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{X}}}, y \in \frac{\varepsilon}{2}\boldsymbol{\mathcal{B}}\right\}$. Specifically, if $\boldsymbol{\mathcal{X}}$ is a subset of the unit ball in $d$ dimensions then $\boldsymbol{\mathcal{X}} + \frac{\varepsilon}{2}\boldsymbol{\mathcal{B}}$ is contained in the $\left(1 + \frac{\varepsilon}{2}\right)$-ball and thus

$$\left|\mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{X}}}\right| \leq \frac{(1 + \varepsilon/2)^d}{(\varepsilon/2)^d} = \left(1 + \frac{2}{\varepsilon}\right)^d < (3/\varepsilon)^d,$$

**Algorithm A.2.** leaves-to-root truncation of arbitrary tensors to HT-format

```
Input: tensor $\mathbf{X} \in \mathbb{R}^{I}$, dimension tree $T_I$ (depth $p > 0$),
        target representation rank $(r_t)_{t \in T_I}$;
for each singleton $\alpha \in \boldsymbol{\mathcal{L}}(T_I)$ do
    Compute an SVD of $\mathbf{X}^{\alpha}$ and store the dominant $r_\alpha$ left singular vectors in the
    columns of the $\alpha$-frame $\mathbf{U}_\alpha$.
end for
    Compute the core tensor $\mathbf{C}_p := \mathbf{X} \times_1 \mathbf{U}_{\{1\}}^T \times \cdots \times_d \mathbf{U}_{\{d\}}^T$ .
for $\ell = p-1, \dots, 0$ do
    Initialize $\mathbf{C}_\ell := \mathbf{C}_{\ell+1}$.
    for each mode cluster $t \in \boldsymbol{\mathcal{I}}(T_I)$ on level $\ell$ do
        Compute an SVD of $\left(\mathbf{C}_{\ell+1}\right)^{t}$ and store the dominant $r_t$ left singular vectors in the
        columns of the $t$-frame $\mathbf{U}_t \in \mathbb{R}^{r_{t_1} r_{t_2} \times r_t}$. Let $\mathbf{U}_{t_1}$ and $\mathbf{U}_{t_2}$ denote
        the corresponding frames for the successors $t_1$, $t_2$ of $t$ on level $\ell + 1$.
        Compute the entries of the transfer tensor
            $\mathbf{B}_t\,(i, j, \nu) := \langle \mathbf{U}_t(:, i), \mathbf{U}_{t_1}(:, j) \otimes \mathbf{U}_{t_2}(:, \nu) \rangle$
        Update the core tensor $\mathbf{C}_\ell := \mathbf{C}_\ell \times_t \mathbf{U}_t^T$
    end for
end for
return HT-rank $\mathbf{r} = \{r_t\}_{t \in T_I}$ approximation $\mathbf{X}_{HT}$: $\left(\{\mathbf{U}_\alpha\}_{\alpha \in \boldsymbol{\mathcal{L}}(T_I)}, \{\boldsymbol{\mathcal{B}}_t\}_{t \in \boldsymbol{\mathcal{I}}(T_I)}\right)$.
```

where the last inequality follows since $\varepsilon < 1$. We always require that $\mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{X}}} \subset \boldsymbol{\mathcal{X}}$.

Next we prove a special case of above lemma for $\boldsymbol{\mathcal{X}}$ being the unit Euclidean sphere in $d$-dimensions denoted by $\boldsymbol{\mathcal{S}}^{d-1}$. The proof for Lemma A.8 follows by similar arguments. With $\boldsymbol{\mathcal{B}}_d(\mathbf{x}, \varepsilon)$ we denote the Euclidean ball in $d$ dimensions, centered at $\mathbf{x}$ of radius $\varepsilon$.

**Lemma A.9** (Covering number of the sphere, [163])**.** The unit Euclidean sphere $\boldsymbol{\mathcal{S}}^{d-1}$ equipped with the Euclidean metric satisfies

$$\left|\mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}^{d-1}}\right| \leq \left(1 + \frac{2}{\varepsilon}\right)^{d} \quad \text{for every } \varepsilon > 0.$$

Proof. Let us fix $\varepsilon > 0$ and choose $\mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}^{d-1}}$ to be a maximal $\varepsilon$-separated subset of $\boldsymbol{\mathcal{S}}^{d-1}$. (In other words, $\|\mathbf{x} - \mathbf{y}\|_2 \geq \varepsilon$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}^{d-1}}$, $\mathbf{x} \neq \mathbf{y}$, and no other subset of $\boldsymbol{\mathcal{S}}^{d-1}$ containing $\mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}^{d-1}}$ has this property.)

The maximality property implies that $\mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}^{d-1}}$ is an $\varepsilon$-net of $\boldsymbol{\mathcal{S}}^{d-1}$. Otherwise, there would exist $\mathbf{x} \in \boldsymbol{\mathcal{S}}^{d-1} \backslash \mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}^{d-1}}$ such that $\|\mathbf{x} - \mathbf{y}\|_2 > \varepsilon$, for all $\mathbf{y} \in \mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}^{d-1}}$. But then $\mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}^{d-1}} \cup \{\mathbf{x}\}$ would be an $\varepsilon$-separated set (which is in contradiction with $\mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}^{d-1}}$ being the maximal $\varepsilon$-separated subset of $\boldsymbol{\mathcal{S}}^{d-1}$).

The separation property implies that the balls centered at the points in $\mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}^{d-1}}$ of radius $\varepsilon/2$ are disjoint. Additionally, they lie in the ball centered at origin of radius $(1 + \varepsilon/2)$. That is,

$$\cup_{\mathbf{x} \in \mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}^{d-1}}} \boldsymbol{\mathcal{B}}_d\left(\mathbf{x}, \frac{\varepsilon}{2}\right) \subseteq \boldsymbol{\mathcal{B}}_d(\mathbf{0}, 1 + \frac{\varepsilon}{2}).$$

Comparing the volume and applying that $\operatorname{Vol}(\boldsymbol{\mathcal{B}}_d(\mathbf{0},r)) = r^d \operatorname{Vol}(\boldsymbol{\mathcal{B}}_d(\mathbf{0},1)) = r^d$ gives

$$\left|\mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}^{d-1}}\right| \cdot \left(\frac{\varepsilon}{2}\right)^d \leq \left(1+\frac{\varepsilon}{2}\right)^d,$$

which implies that

$$\left|\mathcal{N}_{\varepsilon}^{\boldsymbol{\mathcal{S}}^{d-1}}\right| \leq \left(\frac{1+\frac{\varepsilon}{2}}{\frac{\varepsilon}{2}}\right)^d = \left(1+\frac{2}{\varepsilon}\right)^d.$$

□

# APPENDIX B

In the following, we provide an intuition behind the sum-of-squares certificates. In particular, we focus on the theta bodies in Subsection B.1 and in Subsection B.2 we introduce some basic definitions and results related to Gröbner bases.

## B.1. Intuition behind the sum-of-squares certificates

A central problem in optimization is to find the maximum value of a linear function over a set $\boldsymbol{\mathcal{S}} \in \mathbb{R}^n$. That is, solving

$$\max_{\mathbf{x}} \langle \mathbf{c}, \mathbf{x} \rangle \quad \text{s.t.} \quad \mathbf{x} \in \boldsymbol{\mathcal{S}}$$

which is equivalent to solving

$$\max_{\mathbf{x}} \langle \mathbf{c}, \mathbf{x} \rangle \quad \text{s.t.} \quad \mathbf{x} \in \overline{\text{conv}\,(\boldsymbol{\mathcal{S}})},$$

where $\overline{\text{conv}\,(\boldsymbol{\mathcal{S}})}$ denotes the closure of the convex hull of the set $\boldsymbol{\mathcal{S}}$. For example, in linear programming the set $\boldsymbol{\mathcal{S}}$ is a polyhedron $\boldsymbol{\mathcal{S}} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$. We are interested in the case where $\boldsymbol{\mathcal{S}}$ is a real algebraic set, i.e., a set of all real solutions to a finite set of polynomials. In particular, for a given polynomial ideal $\mathcal{J}$ generated by a finite basis $\{f_1, f_2, \dots, f_m\}$, the set $\boldsymbol{\mathcal{S}}$ we are interested in is the real algebraic variety of the ideal $\mathcal{J}$ denoted by $\nu_{\mathbb{R}}(\mathcal{J})$. That is, the set $\boldsymbol{\mathcal{S}}$ is of the form

$$\boldsymbol{\mathcal{S}} := \nu_{\mathbb{R}}\,(\mathcal{J}) = \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) = 0, \forall\, f \in \mathcal{J}\} = \{\mathbf{x} \in \mathbb{R}^n : f_i(\mathbf{x}) = 0, \forall\, i \in [m]\}\,.$$

Recall that every closed convex set $\boldsymbol{\mathcal{C}} \subset \mathbb{R}^n$ is the intersection of closed half-spaces that contain it. That is,

$$\begin{aligned} \boldsymbol{\mathcal{C}} &= \cap \{\boldsymbol{\mathcal{H}} : \boldsymbol{\mathcal{H}} \text{ closed half-space, } \boldsymbol{\mathcal{C}} \subseteq \boldsymbol{\mathcal{H}}\} && \text{(B.1)} \\ &= \{\mathbf{x} : \mathbf{x} \in \boldsymbol{\mathcal{H}} \text{ for every closed half-space } \boldsymbol{\mathcal{H}} \text{ satisfying } \boldsymbol{\mathcal{C}} \subseteq \boldsymbol{\mathcal{H}}\}. && \text{(B.2)} \end{aligned}$$

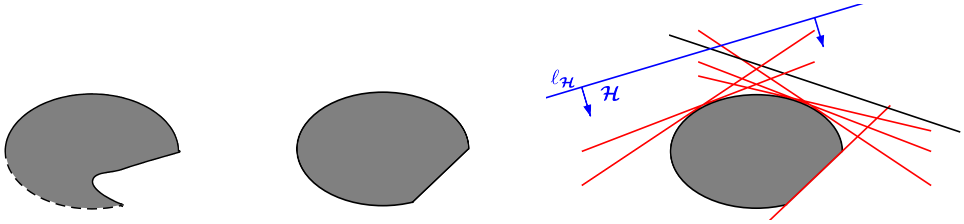
By definition, every closed half-space $\boldsymbol{\mathcal{H}} \subset \mathbb{R}^n$ is identified with the affine function $\ell_{\boldsymbol{\mathcal{H}}}$ which specifies the corresponding hyperplane $\{\mathbf{x} : \ell_{\boldsymbol{\mathcal{H}}}(\mathbf{x}) = 0\}$. (There exists a one-to-one correspondence between the affine functions and closed half-spaces.) That is, for every closed half-space $\boldsymbol{\mathcal{H}}$ there exist $a_0^{\boldsymbol{\mathcal{H}}}, a_1^{\boldsymbol{\mathcal{H}}}, a_2^{\boldsymbol{\mathcal{H}}}, \dots, a_n^{\boldsymbol{\mathcal{H}}} \in \mathbb{R}$ (not all zero) such that

$$\{\mathbf{x} \in \boldsymbol{\mathcal{H}}\} = \left\{\mathbf{x} : \ell_{\boldsymbol{\mathcal{H}}}\,(\mathbf{x}) \geq 0, \ell_{\boldsymbol{\mathcal{H}}}\,(\mathbf{x}) = a_0^{\boldsymbol{\mathcal{H}}} + a_1^{\boldsymbol{\mathcal{H}}} x_1 + a_2^{\boldsymbol{\mathcal{H}}} x_2 + \cdots + a_n^{\boldsymbol{\mathcal{H}}} x_n\right\}.$$

Thus, set $\boldsymbol{\mathcal{C}}$ defined in (B.2) can be expressed as

$$\boldsymbol{\mathcal{C}} = \{\mathbf{x} : \ell_{\boldsymbol{\mathcal{H}}}\,(\mathbf{x}) \geq 0 \text{ for all affine } \ell_{\boldsymbol{\mathcal{H}}} \text{ s.t. } \ell_{\boldsymbol{\mathcal{H}}}|_{\boldsymbol{\mathcal{C}}} \geq 0\}\,. \tag{B.3}$$

Additionally, it is enough to consider only the affine functions $\ell_{\boldsymbol{\mathcal{H}}}$ satisfying $\ell_{\boldsymbol{\mathcal{H}}}|_{\boldsymbol{\mathcal{C}}} \geq 0$ which define the supporting hyperplanes $\ell_{\boldsymbol{\mathcal{H}}}(\mathbf{x}) = 0$ of the set $\boldsymbol{\mathcal{C}}$ – see Figure B.1. Recall that a closed half-space $\boldsymbol{\mathcal{H}} = \{\mathbf{x} : \ell_{\boldsymbol{\mathcal{H}}}(\mathbf{x}) \geq 0\}$ containing $\boldsymbol{\mathcal{C}}$ where $\ell_{\boldsymbol{\mathcal{H}}}(\mathbf{x}) = 0$ is the supporting hyperplane of the set $\boldsymbol{\mathcal{C}}$



(A) set $\boldsymbol{\mathcal{C}}$ (B) $\overline{\operatorname{conv}(\boldsymbol{\mathcal{C}})}$ (C) hyperplanes $\{\mathbf{x} : \ell_{\boldsymbol{\mathcal{H}}}(\mathbf{x}) = 0\}$

FIGURE B.1. Representation of $\overline{\operatorname{conv}(\boldsymbol{\mathcal{C}})}$ as the intersection of possibly infinitely many half-spaces $\boldsymbol{\mathcal{H}}$ defined via affine functions $\ell_{\boldsymbol{\mathcal{H}}}$. The supporting hyperplanes $\{\mathbf{x} : \ell_{\boldsymbol{\mathcal{H}}}(\mathbf{x}) = 0\}$ are denoted in red in the third figure.

satisfies $\boldsymbol{\mathcal{C}} \subseteq \boldsymbol{\mathcal{H}}$ and $\boldsymbol{\mathcal{C}} \cap \boldsymbol{\mathcal{H}} \neq \emptyset$. However, for a general set $\boldsymbol{\mathcal{C}}$, it can be a difficult task to determine all its supporting hyperplanes.

Let us now return to our set of interest $\boldsymbol{\mathcal{S}} = \nu_{\mathbb{R}}(\mathcal{J})$, where $\mathcal{J}$ is a polynomial ideal. By setting $\boldsymbol{\mathcal{C}}$ to be the smallest closed convex superset of $\boldsymbol{\mathcal{S}} = \nu_{\mathbb{R}}(\mathcal{J})$, i.e. $\boldsymbol{\mathcal{C}} = \overline{\operatorname{conv}(\boldsymbol{\mathcal{S}})} = \overline{\operatorname{conv}(\nu_{\mathbb{R}}(\mathcal{J}))}$, (B.3) gives

$$\overline{\operatorname{conv}(\nu_{\mathbb{R}}(\mathcal{J}))} = \left\{\mathbf{x} \in \mathbb{R}^n : \ell(\mathbf{x}) \geq 0 \text{ for all } \ell \text{ affine s.t. } \ell|_{\overline{\operatorname{conv}(\nu_{\mathbb{R}}(\mathcal{J}))}} \geq 0\right\}. \tag{B.4}$$

Notice that if $\ell(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \nu_{\mathbb{R}}(\mathcal{J})$, then $\ell(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \overline{\operatorname{conv}(\nu_{\mathbb{R}}(\mathcal{J}))}$. Thus, in (B.4) it is enough to consider only affine polynomials $\ell$ satisfying $\ell|_{\nu_{\mathbb{R}}(\mathcal{J})} \geq 0$. However, already checking for a single polynomial $\ell \in \mathbb{R}[\mathbf{x}]$ whether it is nonnegative on a set $\nu_{\mathbb{R}}(\mathcal{J})$ can be a difficult task. The idea is to relax the condition $\ell|_{\nu_{\mathbb{R}}(\mathcal{J})} \geq 0$ into something which is easier to verify at the risk of losing some of the affine $\ell$'s in (B.4) and obtaining a superset of $\overline{\operatorname{conv}(\nu_{\mathbb{R}}(\mathcal{J}))}$. One possibility to obtain the hierarchy of the convex relaxations is restricting only to the affine polynomials which are *k-sos mod* $\mathcal{J}$, i.e., to the polynomials that can be written as

$$\ell(\mathbf{x}) = \sigma(\mathbf{x}) + p(\mathbf{x}), \qquad \text{where } \sigma \in \Sigma_{2k},\ p \in \mathcal{J}, \tag{B.5}$$

with $\Sigma_{2k}$ denoting the sum of squares (sos) polynomials of degree at most $2k$ in $\mathbb{R}[\mathbf{x}]$. That is,

$$\sigma \in \Sigma_{2k} \quad \text{if} \quad \exists\, h_1, h_2, \ldots, h_t \text{ with } \deg(h_1), \ldots, \deg(h_t) \leq k \text{ s.t. } \sigma(\mathbf{x}) = \sum_{i=1}^{t} h_i^2(\mathbf{x}).$$

Nonnegativity in (B.5) is guaranteed since $\sigma(\mathbf{x}) \geq 0$ and $p(\mathbf{x}) = 0$, for all $\mathbf{x} \in \nu_{\mathbb{R}}(\mathcal{J})$. These relaxations defined for all $k \in \mathbb{N}$ are called theta bodies [9, 68]. More precisely, for $k \in \mathbb{N}$, the *k-th theta body* is the set

$$\mathrm{TH}_k(\mathcal{J}) = \{\mathbf{x} \in \mathbb{R}^n : \ell(\mathbf{x}) \geq 0, \text{ for all } \ell \text{ affine that are } k\text{-sos mod } \mathcal{J}\}.$$

Notice that, by definition, theta bodies satisfy

$$\overline{\operatorname{conv}(\nu_{\mathbb{R}}(\mathcal{J}))} \subseteq \cdots \subseteq \mathrm{TH}_k(\mathcal{J}) \subseteq \mathrm{TH}_{k-1}(\mathcal{J}) \subseteq \cdots \subseteq \mathrm{TH}_1(\mathcal{J}). \tag{B.6}$$

In the following example we compute the theta body relaxations of the unit $\ell_1$-ball in $\mathbb{R}^2$. In particular, we show that the first theta body coincides with the unit $\ell_1$-ball which together with (B.6) further implies that all theta bodies are equal in this scenario and coincide with the unit $\ell_1$-ball.

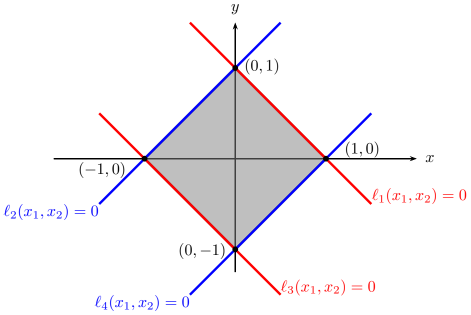


FIGURE B.2. Representation of the unit $\ell_1$-ball together with its four supporting hyperplanes $\{(x_1,x_2) : \ell_i(x_1,x_2)=0\}$ for all $i \in [4]$, with $\ell_i$ defined in (B.7). The corresponding half-spaces are $\boldsymbol{\mathcal{H}}_i = \{(x_1,x_2) : \ell_i(x_1,x_2) \geq 0\}$, for all $i \in [4]$.

**Example B.1** (Relaxations of the unit $\ell_1$-ball in $\mathbb{R}^2$)**.** First, notice that the unit $\ell_1$-ball can be written as a convex hull of its extreme points $\boldsymbol{\mathcal{S}} := \{(1,0),(-1,0),(0,-1),(0,1)\}$, see Figure B.2.. Thus, following the discussion above, we define a polynomial ideal $\mathcal{J}_1$ such that its real algebraic variety $\nu_{\mathbb{R}}(\mathcal{J}_1)$ coincides with the set $\boldsymbol{\mathcal{S}}$. Clearly, for every $\mathbf{x} \in \boldsymbol{\mathcal{S}}$ it holds that $x_1, x_2 \in \{0,1,-1\}$ which is ensured by the constraint $f_i(\mathbf{x}) = 0$, where $f_i(\mathbf{x}) = x_i(x_i-1)(x_i+1)$, for $i = 1,2$. Additionally, every $\mathbf{x} \in \boldsymbol{\mathcal{S}}$ satisfies $|x_1|+|x_2| = 1$ which is guaranteed by the constraint $f_3(\mathbf{x}) = 0$, where $f_3(\mathbf{x}) = x_1^2 + x_2^2 - 1$.

Thus, one option to define the polynomial ideal $\mathcal{J}_1 \in \mathbb{R}[\mathbf{x}] = \mathbb{R}[x_1,x_2]$ such that $\nu_{\mathbb{R}}(\mathcal{J}_1) = \boldsymbol{\mathcal{S}}$ is

$$\mathcal{J}_1 = \langle f_1, f_2, f_3 \rangle = \left\langle x_1(x_1-1)(x_1+1), x_2(x_2-1)(x_2+1), x_1^2 + x_2^2 - 1 \right\rangle .$$

From Figure B.2 it is clear that there are only four supporting hyperplanes $\ell_i(\mathbf{x}) = 0$ of the unit $\ell_1$-ball corresponding to the following four affine functions

$$\begin{aligned} \ell_1(x_1,x_2) &= -x_1 - x_2 + 1 \\ \ell_2(x_1,x_2) &= x_1 - x_2 + 1 \\ \ell_3(x_1,x_2) &= x_1 + x_2 + 1 \\ \ell_4(x_1,x_2) &= -x_1 + x_2 + 1. \end{aligned} \tag{B.7}$$

In other words,

$$\overline{\operatorname{conv}(\nu_{\mathbb{R}}(\mathcal{J}_1))} = \{\mathbf{x} \in \mathbb{R}^2 : \ell_i(\mathbf{x}) \geq 0, \quad \text{for all } i \in [4]\}.$$

In addition, notice that

$$\ell_i(x_1,x_2) = \sigma_i(x_1,x_2) + p_i(x_1,x_2) = \frac{1}{2}\ell_i^2(x_1,x_2) + (-1)^i x_1 x_2 - \frac{1}{2}(x_1^2 + x_2^2 - 1), \qquad \text{for all } i \in [4],$$

where $\sigma_i(x_1,x_2) = \frac{1}{2}\ell_i^2(x_1,x_2) \in \Sigma_2$ and $p_i(x_1,x_2) = (-1)^i x_1 x_2 - \frac{1}{2}(x_1^2 + x_2^2 - 1) \in \mathcal{J}_1$, for all $i \in [4]$ (since $x_1 x_2 = x_1 x_2 \cdot f_3 - x_2 \cdot f_1 - x_1 \cdot f_2 \in \mathcal{J}_1$). That is, the polynomial $\ell_i$ is 1-sos mod $\mathcal{J}_1$, for all $i \in [4]$. Thus, in this scenario, all the theta body relaxations coincide with the unit $\ell_1$-ball.

At the end of Subsection B.2, we generalize the above result to the unit-$\ell_1$-norm balls in $\mathbb{R}^n$, with $n \in \mathbb{N}$.

**Remark B.2.** *Lasserre's method* is another sum-of-squares method for obtaining the hierarchy of convex relaxations of the set $\overline{\operatorname{conv}(\nu_{\mathbb{R}}(\mathcal{J}))}$, see [102]. Let $\mathcal{J}$ be a polynomial ideal generated by the basis $\{f_1, f_2, \ldots, f_m\}$. In this scenario, instead of considering all affine polynomials nonnegative on $\nu_{\mathbb{R}}(\mathcal{J})$, one considers affine polynomials $\ell$ of the form

$$\ell\left(\mathbf{x}\right) = \sigma\left(\mathbf{x}\right) + \sum_{i=1}^{m} g_i\left(\mathbf{x}\right) f_i\left(\mathbf{x}\right), \tag{B.8}$$

where $\sigma \in \Sigma_{2k}$ and $g_i f_i \in \mathbb{R}\left[\mathbf{x}\right]_{2k}$, for a fixed positive integer $k$. In this case, polynomial $\ell$ is called *k-sos mod* $\{f_1, f_2, \ldots, f_m\}$. Notice that the the polynomial $\ell$ in (B.8) is nonnegative on $\nu_{\mathbb{R}}(\mathcal{J})$, since $\sigma(\mathbf{x}) \geq 0$ and $f_i(\mathbf{x}) = 0$, for all $\mathbf{x} \in \nu_{\mathbb{R}}(\mathbf{x})$ and all $i \in [m]$.

Although theta bodies and Lasserre's method are closely related, in general they result in different sets of hierarchical relaxations. Lasserre's relaxations depend on the choice of the basis of the ideal. Therefore, different bases will in general provide different sets of hierarchical relaxations. On the other side, the theta bodies do not depend on the choice of the basis of the ideal and thus are more natural if one is interested in the geometry of $\nu_{\mathbb{R}}\left(\mathcal{J}\right)$ and $\operatorname{conv}\left(\nu_{\mathbb{R}}\left(\mathcal{J}\right)\right)$. However, one has to compute a Gröbner basis of $\mathcal{J}$ which can be a challenging task – Buchberger's algorithm for computing a Gröbner basis of a given polynomial ideal has a double exponential worst case complexity, see [5].

In addition, notice that even if $\{f_1, f_2, \ldots, f_m\}$ is a Gröbner basis of the ideal, Lasserre's relaxations in general differ from theta bodies since for a fixed positive integer $k$, Lasserre's relaxation has additional degree restriction on the polynomial $p\left(\mathbf{x}\right) = \sum_{i=1}^{m} g_i\left(\mathbf{x}\right) f_i\left(\mathbf{x}\right) \in \mathcal{J}$.

We remark that tensor completion via Lassere's relaxations has been analyzed in [4].

## B.2. Gröbner bases

In this section we present the monomial orderings and the Gröbner bases. All the results are stated without proofs which can be found together with the definitions in [36, 37].

To compute a Gröbner basis of a polynomial ideal in $\mathbb{R}\left[\mathbf{x}\right] = \mathbb{R}\left[x_1, x_2, \ldots, x_n\right]$ we need to fix a monomial ordering. In the following we introduce the *lexicographic (lex)*, *graded lexicographic (grlex)*, and *graded reverse lexicographic (grevlex) ordering.* For further details on monomial orderings, we refer the interested reader to [36, 37].

**Definition B.3.** Let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n) \in \mathbb{Z}_{\geq 0}^n$ and the vector difference $\boldsymbol{\alpha} - \boldsymbol{\beta} \in \mathbb{Z}^n$. With $\mathbf{x}^{\boldsymbol{\alpha}}$ we denote the monomial $x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$. Then we write

1) $\mathbf{x}^{\boldsymbol{\alpha}} >_{lex} \mathbf{x}^{\boldsymbol{\beta}}$ if in $\boldsymbol{\alpha} - \boldsymbol{\beta}$ the leftmost nonzero entry is positive.
2) $\mathbf{x}^{\boldsymbol{\alpha}} >_{grlex} \mathbf{x}^{\boldsymbol{\beta}}$ if $|\boldsymbol{\alpha}| = \sum_{i=1}^{n} \alpha_i > |\boldsymbol{\beta}| = \sum_{i=1}^{n} \beta_i$ or $|\boldsymbol{\alpha}| = |\boldsymbol{\beta}|$ and $\boldsymbol{\alpha} >_{lex} \boldsymbol{\beta}$.
3) $\mathbf{x}^{\boldsymbol{\alpha}} >_{grevlex} \mathbf{x}^{\boldsymbol{\beta}}$ if $|\boldsymbol{\alpha}| > |\boldsymbol{\beta}|$ or $|\boldsymbol{\alpha}| = |\boldsymbol{\beta}|$ and the rightmost nonzero entry of $\boldsymbol{\alpha} - \boldsymbol{\beta}$ is negative.

Notice that all three monomial orderings induce the variable order $x_1 > x_2 > \cdots > x_n$.

**Example B.4.** We order the terms of polynomial $f(x, y, z) = 5x^3 - 7xz^2 + 2xy^2 - 3y^2z + yz - z^4 \in \mathbb{R}\left[x, y, z\right]$ in decreasing order with respect to different monomial orderings

(1) *lex:* $5x^3 + 2xy^2 - 7xz^2 - 3y^2z + yz - z^4$

(2) *grlex:* $-z^4 + 5x^3 + 2xy^2 - 7xz^2 - 3y^2z + yz$

(3) *grevlex:* $-z^4 + 5x^3 + 2xy^2 - 3y^2z - 7xz^2 + yz$.

A Gröbner basis is a particular kind of generating set of a polynomial ideal. It was first introduced in 1965 in the Phd thesis of Buchberger [17] together with the algorithm for transforming a given generator set of a polynomial ideal into a Gröbner basis, see Algorithm B.1.

**Definition B.5** (Gröbner basis)**.** Fix a monomial order. A basis $\boldsymbol{\mathcal{G}} = \{g_1, g_2, \ldots, g_s\}$ of a polynomial ideal $\mathcal{J} \subset \mathbb{R}[\mathbf{x}]$ is a *Gröbner basis* (or standard basis) if for all $f \in \mathbb{R}[\mathbf{x}]$ there exist **unique** $r \in \mathbb{R}[\mathbf{x}]$ and $g \in \mathcal{J}$ s.t.

$$f = g + r \tag{B.9}$$

and no monomial of $r$ is divisible by any of the leading monomials in $\boldsymbol{\mathcal{G}}$, i.e., by any of the $\mathrm{LM}(g_1), \mathrm{LM}(g_2), \ldots, \mathrm{LM}(g_s)$.

Notice that since the remainder $r$ in the above definition is unique, a Gröbner basis can be used to determine whether a certain polynomial belongs to an ideal. That is, a polynomial $f \in \mathbb{R}[\mathbf{x}]$ as in (B.9) is in the ideal $\mathcal{J}$ **if and only if** $r = 0$. A Gröbner basis is also one of the main computational tools in solving systems of polynomial equations [37] and in the elimination theory [37]. A Gröbner basis is not unique, but the reduced version (defined below) is.

**Definition B.6.** The *reduced Gröbner basis* for a polynomial ideal $\mathcal{J} \in \mathbb{R}[\mathbf{x}]$ is a Gröbner basis $\boldsymbol{\mathcal{G}} = \{g_1, g_2, \ldots, g_s\}$ for $\mathcal{J}$ such that

1) $\mathrm{LC}(g_i) = 1$, for all $i \in [s]$.
2) no monomial of $g_i$ lies in $\langle \mathrm{LT}(\boldsymbol{\mathcal{G}} \backslash \{g_i\}) \rangle$, for all $i \in [s]$.

In other words, a Gröbner basis $\boldsymbol{\mathcal{G}} = \{g_1, g_2, \ldots, g_s\}$ of $\mathcal{J}$ is the reduced Gröbner basis of $\mathcal{J}$ if for all $i \in [s]$ the polynomial $g_i \in \boldsymbol{\mathcal{G}}$ is monic (i.e., $\mathrm{LC}(g_i) = 1$) and its leading monomial $\mathrm{LM}(g_i)$ does not divide $\mathrm{LM}(g_j)$, for any $j \neq i$.

With $\overline{f}^{\boldsymbol{\mathcal{F}}}$ we denote the remainder on division of $f$ by the ordered $k$-tuple $\boldsymbol{\mathcal{F}} = (f_1, f_2, \ldots, f_k)$. If $\boldsymbol{\mathcal{F}}$ is a Gröbner basis for an ideal $\langle f_1, f_2, \ldots, f_k \rangle$, then we can regard $\boldsymbol{\mathcal{F}}$ as a set without any particular order by Definition B.5. Therefore, $\overline{f}^{\boldsymbol{\mathcal{G}}} = r$ in Definition B.5.

Next we define the $S$-polynomial of given polynomials $f$ and $g$. The $S$-polynomial plays an important role in the Buchberger's algorithm for computing a Gröbner basis of a given polynomial ideal.

**Definition B.7.** Let $f, g \in \mathbb{R}[\mathbf{x}]$ be a nonzero polynomials.

(1) If $\mathrm{multideg}(f) = \boldsymbol{\alpha}$ and $\mathrm{multideg}(g) = \boldsymbol{\beta}$, then let $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_n)$, where $\gamma_i = \max\{\alpha_i, \beta_i\}$, for every $i$. We call $\mathbf{x}^{\boldsymbol{\gamma}}$ the least common multiple of $\mathrm{LM}(f)$ and $\mathrm{LM}(g)$ written $\mathbf{x}^{\boldsymbol{\gamma}} = \mathrm{LCM}(\mathrm{LM}(f), \mathrm{LM}(g))$.

(2) The $S$-polynomial of $f$ and $g$ is the combination

$$S(f, g) = \frac{\mathbf{x}^{\boldsymbol{\gamma}}}{\mathrm{LT}(f)} f - \frac{\mathbf{x}^{\boldsymbol{\gamma}}}{\mathrm{LT}(g)} g.$$

The following result follows directly from Definition B.5 and Division algorithm [37].

**Corollary B.8.** Fix a monomial ordering and let $\boldsymbol{\mathcal{G}} = \{g_1, g_2, \ldots, g_s\} \subset \mathbb{R}[\mathbf{x}]$ be a Gröbner basis of a polynomial ideal $\mathcal{J}$. A polynomial $f \in \mathbb{R}[\mathbf{x}]$ is in the ideal $\mathcal{J}$ if it can be written in the form $f = a_1 g_1 + a_2 g_2 + \ldots + a_s g_s$, where $a_i \in \mathbb{R}[\mathbf{x}]$, for all $i \in [s]$, such that whenever

$$a_i g_i \neq 0,$$

we have

$$\operatorname{multideg}(f) \geq \operatorname{multideg}(a_i g_i).$$

The following definition is important in computing a Gröbner basis of a polynomial ideal.

**Definition B.9.** Fix a monomial order and let $\boldsymbol{\mathcal{G}} = \{g_1, g_2, \ldots, g_s\} \subset \mathbb{R}[\mathbf{x}]$. Given $f \in \mathbb{R}[\mathbf{x}]$, we say that $f$ reduces to zero modulo $\boldsymbol{\mathcal{G}}$ and write

$$f \to_{\boldsymbol{\mathcal{G}}} 0$$

if $f$ can be written in the form $f = a_1 g_1 + a_2 g_2 + \ldots + a_s g_s$, where $a_i \in \mathbb{R}[\mathbf{x}]$ for all $i \in [s]$, s.t. whenever

$$a_i g_i \neq 0,$$

we have

$$\operatorname{multideg}(f) \geq \operatorname{multideg}(a_i g_i).$$

Assume that $\boldsymbol{\mathcal{G}}$ in the above definition is a Gröbner basis of a given ideal $\mathcal{J}$. Then a polynomial $f$ is in the ideal $\mathcal{J}$ if and only if $f$ reduces to zero modulo $\boldsymbol{\mathcal{G}}$. In other words, for a Gröbner basis $\boldsymbol{\mathcal{G}}$,

$$f \to_{\boldsymbol{\mathcal{G}}} 0 \qquad \text{if and only if} \qquad \overline{f}^{\boldsymbol{\mathcal{G}}} = 0.$$

**Theorem B.10** (Buchberger's algorithm)**.** Let $\mathcal{J} = \langle f_1, f_2, \ldots, f_k \rangle \neq \{0\}$ be a polynomial ideal. Then a Gröbner basis of $\mathcal{J}$ can be constructed in a finite number of steps via Buchberger's algorithm (presented in Algorithm B.1).

By the Hilbert Basis Theorem every polynomial ideal $\mathcal{J} \subset \mathbb{R}[\mathbf{x}] = \mathbb{R}[x_1, x_2, \ldots, x_n]$ has a finite basis. By the above theorem, a Gröbner basis of a polynomial ideal always exists and can be computed from a basis of a corresponding polynomial ideal in a finite number of steps via Buchberger's algorithm, see Algorithm B.1 and [17, 36, 37]. Therefore, a Gröbner basis is also finite.

**Algorithm B.1.** Buchberger's algorithm

```
Input: Basis $\boldsymbol{\mathcal{F}} = (f_1, f_2, \ldots, f_k)$ for an ideal $\mathcal{J}$.
  $\boldsymbol{\mathcal{G}} := \boldsymbol{\mathcal{F}}$.
  repeat
    $\boldsymbol{\mathcal{G}}' := \boldsymbol{\mathcal{G}}$.
    for each pair $\{p, q\}$, $p \neq q$ in $\boldsymbol{\mathcal{G}}'$
      $S := \overline{S(p,q)}^{\boldsymbol{\mathcal{G}}'}$
      if $S \neq 0$
        $\boldsymbol{\mathcal{G}} := \boldsymbol{\mathcal{G}} \cup \{S\}$.
      end if
    end for
  until $\boldsymbol{\mathcal{G}} = \boldsymbol{\mathcal{G}}'$.
Output: Gröbner basis $\boldsymbol{\mathcal{G}} = \{g_1, g_2, \ldots, g_s\}$ for $\mathcal{J}$ with $\boldsymbol{\mathcal{F}} \subset \boldsymbol{\mathcal{G}}$.
```

The following theorem gives a criterion for checking whether a given basis of a polynomial ideal is a Gröbner basis.

**Theorem B.11** (Buchberger's criterion)**.** A basis $\mathcal{G} = \{g_1, g_2, \ldots, g_s\}$ for a polynomial ideal $\mathcal{J} \subset \mathbb{R}[\mathbf{x}]$ is a Gröbner basis if and only if $S(g_i, g_j) \rightarrow_{\mathcal{G}} 0$, for all $i \neq j$.

Computing whether $S(g_i, g_j) \rightarrow_{\mathcal{G}} 0$ for all the possible pairs of polynomials in the basis $\mathcal{G}$ can be a tedious task. The following proposition tells us for which pairs of polynomials this is not needed.

**Proposition B.12.** Given a finite set $\mathcal{G} \subset \mathbb{R}[\mathbf{x}]$, suppose that we have $f, g \in \mathcal{G}$ s.t. the leading monomials of $f$ and $g$ are relatively prime, i.e.,

$$\mathrm{LCM}(\mathrm{LM}(f), \mathrm{LM}(g)) = \mathrm{LM}(f)\,\mathrm{LM}(g).$$

Then $S(f, g) \rightarrow_{\mathcal{G}} 0$.

Therefore, to prove that the set $\mathcal{G} \subset \mathbb{R}[\mathbf{x}]$ is a Gröbner basis, it is enough to show that $S(g_i, g_j) \rightarrow_{\mathcal{G}} 0$ for those $i < j$ where $\mathrm{LM}(g_i)$ and $\mathrm{LM}(g_j)$ are not relatively prime.

**Example B.13** (Ideal $\mathcal{J}_{M_{22}}$ from Chapter 4)**.** Let $\mathcal{J} \subset \mathbb{R}[\mathbf{x}] = \mathbb{R}[x_{11}, x_{12}, x_{21}, x_{22}]$ be an ideal defined as $\mathcal{J} = \langle f, g \rangle$, where

$$\begin{aligned} f(\mathbf{x}) &= x_{12}x_{21} - x_{11}x_{22} \\ g(\mathbf{x}) &= x_{11}^2 + x_{12}^2 + x_{21}^2 + x_{22}^2 - 1. \end{aligned}$$

The variable order is $x_{11} > x_{12} > x_{21} > x_{22}$ regardless of the choice of monomial ordering. In the following we compute a Gröbner basis $\mathcal{G}$ of $\mathcal{J}$ (with variable order $x_{11} > x_{12} > x_{21} > x_{22}$) with respect to the

(1) *graded reverse lexicographic order*: Since the leading monomials of polynomials $f$ and $g$ ($\mathrm{LM}(f) = x_{12}x_{21}$ and $\mathrm{LM}(g) = x_{11}^2$) are relatively prime, then by Proposition B.12 the Gröbner basis $\mathcal{G}_{\text{grevlex}}$ with respect to the grevlex order is $\mathcal{G}_{\text{grevlex}} = \{f, g\}$.

(2) *lexicographic order*: We start by defining $\mathcal{G} = \{f, g\}$ and computing the $S$-polynomial $S(f, g)$

$$\begin{aligned} h_1(\mathbf{x}) = S(f, g) &= \frac{x_{11}^2 x_{22}}{-x_{11}x_{22}}(-x_{11}x_{22} + x_{12}x_{21}) - \frac{x_{11}^2 x_{22}}{x_{11}^2}(x_{11}^2 + x_{12}^2 + x_{21}^2 + x_{22}^2 - 1) \\ &= -x_{11}x_{12}x_{21} - x_{12}^2 x_{22} - x_{21}^2 x_{22} - x_{22}^3 + x_{22}. \end{aligned}$$

Since $\mathrm{LM}(h_1) = x_{11}x_{12}x_{21}$ is not divisible by any of the leading monomials of $f$ and $g$, we include the polynomial $h_1$ in the basis $\mathcal{G}$. Next, we compute the $S$-polynomial of $f$ and $h_1$

$$h_2(\mathbf{x}) = S(f, h_1) = -x_{12}^2 x_{21}^2 - x_{12}^2 x_{22}^2 - x_{21}^2 x_{22}^2 - x_{22}^4 + x_{22}^2.$$

Similarly to before, since $\mathrm{LM}(h_2) = x_{12}^2 x_{21}^2$ is not divisible by any of the leading monomials of $f$, $g$, and $h_1$, we add the polynomial $h_2$ in the basis $\mathcal{G}$ which is now of the form $\mathcal{G} = \{f, g, h_1, h_2\}$. By above, $S(f, g) \rightarrow_{\mathcal{G}} 0$ and $S(f, h_1) \rightarrow_{\mathcal{G}} 0$. By Proposition B.12, we have that $S(f, h_2) \rightarrow_{\mathcal{G}} 0$ and $S(g, h_2) \rightarrow_{\mathcal{G}} 0$. It remains to compute $\overline{S(g, h_1)}^{\mathcal{G}}$ and $\overline{S(h_1, h_2)}^{\mathcal{G}}$

$$\begin{aligned} S(g, h_1) &= f \cdot \left(x_{12}^2 + x_{21}^2 + x_{22}^2 - 1\right) \rightarrow_{\mathcal{G}} 0 \\ S(h_1, h_2) &= x_{22} \cdot f \cdot \left(x_{12}^2 + x_{21}^2 + x_{22}^2 - 1\right) \rightarrow_{\mathcal{G}} 0. \end{aligned}$$

Therefore, $\mathcal{G}_{\text{lex}} := \{f, g, h_1, h_2\}$ is the Gröbner basis of the ideal $\mathcal{J}$ with respect to the lexicographic order.

One can show that $\mathcal{G}_{\text{lex}} = \{f, g, h_1, h_2\}$ as in (2) is also the Gröbner basis of the ideal $\mathcal{J}$ with respect to the *grlex* order. However, $\mathcal{G}_{\text{lex}}$ is not the reduced Gröbner basis of $\mathcal{J}$, since $\text{LC}(f) = \text{LC}(h_1) = \text{LC}(h_2) = -1$. However, $\mathcal{G}'_{\text{lex}} = \{-f, g, -h_1, -h_2\}$ is the reduced Gröbner basis of the ideal $\mathcal{J}$ with respect to the *lexicographic* and *grlex* ordering. On the other hand, $\mathcal{G}_{\text{grevlex}} = \{f, g\}$ is already the reduced Gröbner basis of $\mathcal{J}$ with respect to the *grevlex* ordering.

**Example B.14** (Continuing the Example B.1)**.** Recall that we are considering an ideal $\mathcal{J}_1$ and the corresponding basis $\mathcal{B}_1$

$$\mathcal{J}_1 = \langle \mathcal{B}_1 \rangle = \left\langle x_1(x_1-1)(x_1+1), x_2(x_2-1)(x_2+1), x_1^2 + x_2^2 - 1 \right\rangle .$$

The basis $\mathcal{B}_1$ is not a Gröbner basis of the ideal $\mathcal{J}_2$ with respect to the grevlex ordering since $\overline{S(x_1^2 + x_2^2 - 1, x_1^3 - x_1)}^{\mathcal{B}_1} = x_1 x_2^2$. However, recall that we have showed that $x_1 x_2 \in \mathcal{J}_2$ and thus $\mathcal{B}_2 := \mathcal{B}_1 \cup \{x_1 x_2\}$ defines a new basis of $\mathcal{J}_2$. Additionally, $S(x_1^2 + x_2^2 - 1, x_1^3 - x_1) \to_{\mathcal{B}_2} 0$. One can verify that $\mathcal{B}_2$ is a Gröbner basis but it is not the reduced Gröbner basis of the ideal $\mathcal{J}_2$ since the leading monomial $\text{LM}(x_1^2 + x_2^2 + 1) = x_1^2$ divides the leading monomial $\text{LM}(x_1^3 - x_1) = x_1^3$. Moreover, $x_1^3 - x_1 = x_1 \cdot (x_1^2 + x_2^2 - 1) - x_2 \cdot (x_1 x_2)$.

However, eliminating the polynomial $x_1^3 - x_1$ from the basis $\mathcal{B}_2$ leads to the basis

$$\mathcal{G}_1 = \left\{ x_2(x_2-1)(x_2+1), x_1^2 + x_2^2 - 1, x_1 x_2 \right\}$$

which is the reduced Gröbner basis of the ideal $\mathcal{J}_1$.

In Example B.1 we have shown that the theta body relaxations of the unit-$\ell_1$-norm ball in $\mathbb{R}^2$ coincide with the unit-$\ell_1$-norm ball. Consequently, the theta body relaxations do not provide new vector norms. In the following we show that this is true also for the unit-$\ell_1$-norm ball in $\mathbb{R}^n$, with $n \in \mathbb{N}$.

**Example B.15** (the unit-$\ell_1$-norm ball in $\mathbb{R}^n$)**.** It is easy to see that $\mathbf{x} \in \mathbb{R}^n$ is an extreme point of the unit-$\ell_1$-norm ball if and only if

$$\begin{aligned}
&x_1^2 + x_2^2 + \cdots + x_n^2 = 1 \\
&x_i \in \{0, -1, 1\}, \quad \text{for all } i \in [n] \\
&x_j \cdot x_k = 0, \quad \text{for all } 1 \leq j < k \leq n.
\end{aligned}$$

Following the example above, we define an ideal $\mathcal{J}_n = \langle \mathcal{G}_n \rangle$ through its basis

$$\begin{aligned}
\mathcal{G}_n = \{ & g(\mathbf{x}) = x_1^2 + x_2^2 + \cdots + x_n^2 - 1, \\
& f_i(\mathbf{x}) = x_i(x_i - 1)(x_i + 1), \quad i \in \{2, 3, \ldots, n\}, \\
& h_{jk}(\mathbf{x}) = x_j x_k, \quad 1 \leq j < k \leq n\}.
\end{aligned}$$

Similarly to the previous example, we omitted $f_1(\mathbf{x}) = x_1^3 - x_1$ from $\mathcal{G}_n$ since $x_1^3 - x_1 = x_1 \cdot g(\mathbf{x}) - \sum_{k=2}^n h_{1k}(\mathbf{x}) \cdot x_k$. Clearly, the real algebraic variety $\nu_{\mathbb{R}}(\mathcal{J}_n)$ of the ideal $\mathcal{J}_n$ is the set of all extreme points of the unit-$\ell_1$-norm ball in $\mathbb{R}^n$. We claim that $\mathcal{G}_n$ is also the reduced Gröbner basis of the ideal $\mathcal{J}_n$ with respect to the grevlex ordering. Notice that the leading terms of $g$ and $f_i$, as well as the leading terms of $g$ and $h_{jk}$ with $j \geq 2$ are relatively prime. In the following we compute the remaining $S$-polynomials.

$$S(g,h_{1k}) = x_2^2x_k + \cdots + x_n^2x_k - x_k = f_k + \sum_{i=2}^{k-1} x_ih_{ik} + \sum_{i=k+1}^{n} x_ih_{ki} \to_{\boldsymbol{\mathcal{G}}_n} 0, \quad \text{for } k \in \{2,\ldots,n\}$$

$$S(f_j,h_{jk}) = S(f_k,h_{jk}) = -x_jx_k = -h_{jk} \to_{\boldsymbol{\mathcal{G}}_n} 0, \quad \text{for } 1 \le j < k \le n.$$

Additionally, all the polynomials in the set $\boldsymbol{\mathcal{G}}_n$ are monic. Also, no monomial of $p \in \boldsymbol{\mathcal{G}}_n$ lies in $\langle \mathrm{LT}(\boldsymbol{\mathcal{G}}\backslash\{p\})\rangle$, for all $p \in \boldsymbol{\mathcal{G}}_n$. Thus, set $\boldsymbol{\mathcal{G}}_n$ is the reduced Gröbner basis of the polynomial $\mathcal{J}_n$ with respect to the grevlex ordering.

Next, we compute the theta bodies $\mathrm{TH}_k(\mathcal{J}_n)$, for $k \in \mathbb{N}$. Recall that a vector $\mathbf{x}$ is in the unit-$\ell_1$-norm ball in $\mathbb{R}^n$ if and only if

$$x_1^2 + x_2^2 + \cdots + x_n^2 \le 1.$$

Thus, the supporting hyperplanes of the unit-$\ell_1$-norm ball in $\mathbb{R}^n$ are of the form $\ell_i(\mathbf{x}) = 0$, where

$$\ell_i(\mathbf{x}) = (-1)^{i_1}x_1 + (-1)^{i_2}x_2 + \cdots + (-1)^{i_n}x_n + 1 \quad \text{for all } i \in \{0,1,2,\ldots,2^n-1\},$$

and $i_1i_2\cdots i_n$ is the binary representation of the number $i$. The unit-$\ell_1$-norm ball $\boldsymbol{\mathcal{B}}^n$ in $\mathbb{R}^n$ is then of the form

$$\boldsymbol{\mathcal{B}}^n = \overline{\mathrm{conv}(\nu_{\mathbb{R}}(\mathcal{J}_n))} = \{\mathbf{x} : \ell_i(\mathbf{x}) \ge 0, \quad \text{for all } i \in \{0,1,2,\ldots,2^n-1\}\}$$

and the theta bodies $\mathrm{TH}_k(\mathcal{J}_n)$ are defined as

$$\mathrm{TH}_k(\mathcal{J}_n) = \{\mathbf{x} : \ell_i(\mathbf{x}) \ge 0, \quad \text{for all } i \in \{0,1,2,\ldots,2^n-1\} \text{ s.t. } \ell_i \text{ is } k\text{-sos mod } \mathcal{J}_n\}.$$

For every $i \in \{0,1,2,\ldots,2^n-1\}$ it holds that

$$\ell_i = \frac{1}{2}\ell_i^2 - \left(\frac{1}{2}g + \sum_{j<k}(-1)^{i_j+i_k}h_{jk}\right)$$

where $\frac{1}{2}\ell_i^2 \in \Sigma_2$ and $p = -\left(\frac{1}{2}g + \sum_{j<k}(-1)^{i_j+i_k}h_{jk}\right) \in \mathcal{J}_n$. Thus, the polynomial $\ell_i$ is 1-sos mod $\mathcal{J}_n$ for every $i$ and the unit-$\ell_1$-norm ball coincides with the first theta body $\mathrm{TH}_1(\mathcal{J}_n)$. Consequently, in this scenario, all theta bodies are equal to the unit-$\ell_1$-norm ball.

# APPENDIX C

## C.1. Semidefinite programming

Semidefinite programming is a part of convex programming where one minimizes (or maximizes) a linear objective function over a spectahedron (intersection of the cone of positive semidefinite matrices with an affine space).

Semidefinite programming is a relatively new field which recently gained a lot of interest since many practical problems in combinatorial optimization can be modeled or approximated by semidefinite programs (SDPs). In addition, since SDPs are a special case of cone programming, they can be solved efficiently via interior point methods, see [166].

A general SDP (also called a primal problem) is of the form

$$\min_{\mathbf{X}\in\boldsymbol{\mathcal{S}}^n} \langle \mathbf{C},\mathbf{X}\rangle_{\boldsymbol{\mathcal{S}}^n} \quad \text{subject to} \quad \langle \mathbf{A}_i,\mathbf{X}\rangle_{\boldsymbol{\mathcal{S}}^n} = \mathbf{b}\left(i\right),\ i\in[m]$$
$$\mathbf{X}\succeq \mathbf{0}, \tag{P1}$$

where $\boldsymbol{\mathcal{S}}^n = \left\{\mathbf{X}\in\mathbb{R}^{n\times n} : \mathbf{X}=\mathbf{X}^T\right\}$ and $\langle \mathbf{A},\mathbf{B}\rangle_{\boldsymbol{\mathcal{S}}^n} = \operatorname{tr}\left(\mathbf{B}^T\mathbf{A}\right) = \sum_{i=1}^n\sum_{j=1}^n \mathbf{A}\left(i,j\right)\mathbf{B}\left(i,j\right)$.

The corresponding dual problem is

$$\max_{\mathbf{y}\in\mathbb{R}^m} \langle \mathbf{b},\mathbf{y}\rangle_{\mathbb{R}^m} \quad \text{subject to} \quad \sum_{i=1}^m \mathbf{y}\left(i\right)\mathbf{A}_i \preceq \mathbf{C}, \tag{D1}$$

where $\langle\cdot,\cdot\rangle_{\mathbb{R}^m}$ denotes the standard $\ell_2$-inner product. It is possible to start with a different primal problem (maximizing instead of minimizing a linear functional), i.e.

$$\max_{\mathbf{X}\in\boldsymbol{\mathcal{S}}^n} \langle \mathbf{C},\mathbf{X}\rangle_{\boldsymbol{\mathcal{S}}^n} \quad \text{subject to} \quad \langle \mathbf{A}_i,\mathbf{X}\rangle_{\boldsymbol{\mathcal{S}}^n} = \mathbf{b}\left(i\right),\ i\in[m]$$
$$\mathbf{X}\succeq \mathbf{0}. \tag{P2}$$

Noticing that $\max_{\mathbf{X}\in\boldsymbol{\mathcal{S}}^n}\langle \mathbf{C},\mathbf{X}\rangle_{\boldsymbol{\mathcal{S}}^n} = -\min_{\mathbf{X}\in\boldsymbol{\mathcal{S}}^n}\langle -\mathbf{C},\mathbf{X}\rangle_{\boldsymbol{\mathcal{S}}^n}$ leads to an equivalent primal problem

$$-\min_{\mathbf{X}\in\boldsymbol{\mathcal{S}}^n} \langle -\mathbf{C},\mathbf{X}\rangle_{\boldsymbol{\mathcal{S}}^n} \quad \text{such that} \quad \langle \mathbf{A}_i,\mathbf{X}\rangle_{\boldsymbol{\mathcal{S}}^n} = \mathbf{b}\left(i\right),\ i\in[m]$$
$$\mathbf{X}\succeq \mathbf{0}.$$

Then the corresponding dual problem is

$$-\max_{\mathbf{y}\in\mathbb{R}^m} \langle \mathbf{b},\mathbf{y}\rangle_{\mathbb{R}^m} \quad \text{s.t.} \quad \sum_{i=1}^m \mathbf{y}\left(i\right)\mathbf{A}_i \preceq -\mathbf{C} \Leftrightarrow \min_{\mathbf{y}\in\mathbb{R}^m} \langle \mathbf{b},-\mathbf{y}\rangle_{\mathbb{R}^m} \quad \text{s.t.} \quad \sum_{i=1}^m \left(-\mathbf{y}(i)\right)\mathbf{A}_i \succeq \mathbf{C}.$$

Applying the substitution $\mathbf{w} := -\mathbf{y}$ leads to an equivalent SDP

$$\min_{\mathbf{w}} \langle \mathbf{b},\mathbf{w}\rangle_{\mathbb{R}^m} \quad \text{s.t.} \quad \sum_{i=1}^m \mathbf{w}(i)\mathbf{A}_i \succeq \mathbf{C}. \tag{D2}$$

**Definition C.1** ([113])**.** Let $\mathbf{X} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix}$ be a symmetric matrix in $\mathbb{R}^{n \times n}$, where $\mathbf{A}$ is invertible. Then the matrix

$$\mathbf{S} = \mathbf{C} - \mathbf{B}^T \mathbf{A}^{-1} \mathbf{B}$$

is the *Schur complement of the block* $\mathbf{A}$ *in* $\mathbf{X}$.

The Schur complement of $\mathbf{A}$ in $\mathbf{X}$ is closely connected to the positive definiteness of the block matrix $\mathbf{X}$.

**Proposition C.2** ([15])**.** Let $\mathbf{X} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix}$ be a symmetric matrix in $\mathbb{R}^{n \times n}$. If $\mathbf{A} \succ \mathbf{0}$, then

$$\mathbf{X} \succeq \mathbf{0} \text{ if and only if } \mathbf{S} \succeq \mathbf{0}, \tag{C.1}$$

where $\mathbf{S}$ is the Schur complement of $\mathbf{A}$ in $\mathbf{X}$.

**Example C.3** (Matrix operator norm minimization, see [15, 133])**.** We want to compute the operator norm of a matrix $\mathbf{Z} \in \mathbb{R}^{m \times n} \backslash \{\mathbf{0}\}$. In the following, $\lambda_{\max}(\mathbf{Z})$ denotes the largest eigenvalue of matrix $\mathbf{Z}$ and $\mathbf{I}_n \in \mathbb{R}^{n \times n}$ denotes the identity matrix. We obtain the following equivalent statements

$$\|\mathbf{Z}\|_{2\to2} \leq t \Leftrightarrow \lambda_{\max}\left(\mathbf{Z}^T\mathbf{Z}\right) \leq t^2 \Leftrightarrow \max_{\|\mathbf{v}\|_2=1} \mathbf{v}^T\mathbf{Z}^T\mathbf{Z}\mathbf{v} \leq t^2 \Leftrightarrow \max_{\|\mathbf{v}\|_2=1} \left(\mathbf{v}^T\mathbf{Z}^T\mathbf{Z}\mathbf{v} - t^2\mathbf{v}^T\mathbf{v}\right) \leq 0$$
$$\Leftrightarrow -\max_{\|\mathbf{v}\|_2=1} \left(\mathbf{v}^T\mathbf{Z}^T\mathbf{Z}\mathbf{v} - t^2\mathbf{v}^T\mathbf{v}\right) \geq 0 \Leftrightarrow \min_{\|\mathbf{v}\|_2=1} \left(-\mathbf{v}^T\mathbf{Z}^T\mathbf{Z}\mathbf{v} + t^2\mathbf{v}^T\mathbf{v}\right) \geq 0$$
$$\Leftrightarrow \min_{\|\mathbf{v}\|_2=1} \mathbf{v}^T\left(t^2\mathbf{I}_n - \mathbf{Z}^T\mathbf{Z}\right)\mathbf{v} \geq 0 \Leftrightarrow t^2\mathbf{I}_n - \mathbf{Z}^T\mathbf{Z} \succeq \mathbf{0}.$$

Applying (C.1) leads to the following SDP for computing the operator norm of a matrix $\mathbf{Z}$

$$\min_{t\geq0} t \quad \text{such that} \quad \begin{bmatrix} t\mathbf{I}_m & \mathbf{Z} \\ \mathbf{Z}^T & t\mathbf{I}_n \end{bmatrix} \succeq \mathbf{0}.$$

In the following example we show that the matrix nuclear norm is a dual norm of the matrix operator norm. Recall that in an inner product space, the dual norm $\|\cdot\|_d$ of a given norm $\|\cdot\|$ always exists and is defined as

$$\|\mathbf{X}\|_d := \max_{\mathbf{Y}} \left\{\langle \mathbf{X}, \mathbf{Y} \rangle : \|\mathbf{Y}\| \leq 1\right\}. \tag{C.2}$$

In addition, the dual norm of the norm $\|\cdot\|_d$ is the original norm $\|\cdot\|$. The matrix space $\mathbb{R}^{m \times n}$ is equipped with the inner product $\langle \mathbf{X}, \mathbf{Y} \rangle = \operatorname{tr}\left(\mathbf{Y}^T\mathbf{X}\right)$.

**Example C.4** (Dual norm of the matrix operator norm is the nuclear norm, see [133])**.** From Example C.3, a semidefinite characterization of the operator norm is

$$\|\mathbf{Z}\|_{2\to2} = \min_{t\geq0} t \quad \text{such that} \quad \begin{bmatrix} t\mathbf{I}_m & \mathbf{Z} \\ \mathbf{Z}^T & t\mathbf{I}_n \end{bmatrix} \succeq \mathbf{0}. \tag{C.3}$$

Let $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the reduced singular value decomposition of the matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, with $\mathbf{U} \in \mathbb{R}^{m \times r}$, diagonal $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$, and $\mathbf{V} \in \mathbb{R}^{n \times r}$, where $r$ denotes the rank of the matrix $\mathbf{X}$. For a matrix $\mathbf{Y} := \mathbf{U}\mathbf{V}^T$ it holds that $\|\mathbf{Y}\|_{2\to2} = 1$ and $\langle \mathbf{X}, \mathbf{Y} \rangle = \operatorname{tr}\left(\mathbf{Y}^T\mathbf{X}\right) = \operatorname{tr}\left(\mathbf{V}\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}\right) = \operatorname{tr}(\mathbf{\Sigma}) = \sum_{i=1}^r \sigma_i(\mathbf{X}) = \|\mathbf{X}\|_*$, where $\{\sigma_i(\mathbf{X})\}_{i=1}^r$ is the set of the singular values of the matrix $\mathbf{X}$. Therefore, by (C.2) it holds that $\|\mathbf{X}\|_d \geq \|\mathbf{X}\|_*$, for all $\mathbf{X} \in \mathbb{R}^{m \times n}$.

For the upper bound on the dual norm we use the SDP duality theory. Notice that by (C.3)

$$\max_{\mathbf{Y}} \left\{ \langle \mathbf{X}, \mathbf{Y} \rangle : \|\mathbf{Y}\|_{2\to 2} \leq 1 \right\} \Leftrightarrow \max_{\mathbf{Y}} \operatorname{tr}\left(\mathbf{Y}^T\mathbf{X}\right) \quad \text{s.t.} \quad \begin{bmatrix} \mathbf{I}_m & \mathbf{Y} \\ \mathbf{Y}^T & \mathbf{I}_n \end{bmatrix} \succeq \mathbf{0}. \tag{C.4}$$

Define matrices $\mathbf{Z} := \begin{bmatrix} \mathbf{I}_m & \mathbf{Y} \\ \mathbf{Y}^T & \mathbf{I}_n \end{bmatrix}$, $\mathbf{C} := \frac{1}{2}\begin{bmatrix} \mathbf{0} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{0} \end{bmatrix}$, $\mathbf{A}_1 := \begin{bmatrix} \frac{1}{m}\mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$, and $\mathbf{A}_2 := \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{n}\mathbf{I}_n \end{bmatrix}$. Noticing that $\langle \mathbf{Z}, \mathbf{A}_1 \rangle = \langle \mathbf{Z}, \mathbf{A}_2 \rangle = 1$ leads to a semidefinite program

$$\max_{\mathbf{Z}} \langle \mathbf{C}, \mathbf{Z} \rangle \quad \text{such that} \quad \langle \mathbf{A}_1, \mathbf{Z} \rangle = 1,\ \langle \mathbf{A}_2, \mathbf{Z} \rangle = 1,\ \mathbf{Z} \succeq \mathbf{0}. \tag{C.5}$$

Notice that every feasible matrix $\mathbf{Y}$ for (C.4) induces a feasible $\mathbf{Z}$ for (C.5). The dual problem corresponding to (C.5) is

$$\min_{\mathbf{y}} \left(\mathbf{y}(1) + \mathbf{y}(2)\right) \quad \text{such that} \quad \mathbf{y}(1)\,\mathbf{A}_1 + \mathbf{y}(2)\,\mathbf{A}_2 \succeq \mathbf{C} \tag{C.6}$$

which is equivalent to

$$\min_{\mathbf{y}} \left(\mathbf{y}(1) + \mathbf{y}(2)\right) \quad \text{such that} \quad \begin{bmatrix} \frac{2\mathbf{y}(1)}{m}\mathbf{I}_m & -\mathbf{X} \\ -\mathbf{X}^T & \frac{2\mathbf{y}(2)}{n}\mathbf{I}_n \end{bmatrix} \succeq \mathbf{0}.$$

For matrices $\mathbf{A} \in \mathbb{C}^{m\times m}$, $\mathbf{X} \in \mathbb{C}^{m\times n}$, and $\mathbf{C} \in \mathbb{C}^{n\times n}$ it holds that $\begin{bmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{C} \end{bmatrix} \succeq \mathbf{0}$ if and only if $\begin{bmatrix} \mathbf{A} & -\mathbf{X} \\ -\mathbf{X}^T & \mathbf{C} \end{bmatrix} \succeq \mathbf{0}$ since

$$\begin{bmatrix} \mathbf{y}_1^T & \mathbf{y}_2^T \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1^T & -\mathbf{y}_2^T \end{bmatrix} \begin{bmatrix} \mathbf{A} & -\mathbf{X} \\ -\mathbf{X}^T & \mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ -\mathbf{y}_2 \end{bmatrix}, \quad \text{for } \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}.$$

Thus, the dual problem (C.6) is equivalent to

$$\min_{\mathbf{y}} \left(\mathbf{y}(1) + \mathbf{y}(2)\right) \quad \text{such that} \quad \begin{bmatrix} \frac{2\mathbf{y}(1)}{m}\mathbf{I}_m & \mathbf{X} \\ \mathbf{X}^T & \frac{2\mathbf{y}(2)}{n}\mathbf{I}_n \end{bmatrix} \succeq \mathbf{0}.$$

By defining the matrices $\mathbf{W}_1 := \frac{2\mathbf{y}(1)}{m}\mathbf{I}_m$ and $\mathbf{W}_2 := \frac{2\mathbf{y}(2)}{n}\mathbf{I}_n$ we obtain the following semidefinite program

$$\min_{\mathbf{W}_1,\mathbf{W}_2} \frac{1}{2}\left(\operatorname{tr}(\mathbf{W}_1) + \operatorname{tr}(\mathbf{W}_2)\right) \quad \text{such that} \quad \begin{bmatrix} \mathbf{W}_1 & \mathbf{X} \\ \mathbf{X}^T & \mathbf{W}_2 \end{bmatrix} \succeq \mathbf{0}. \tag{C.7}$$

Recall that $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ and set $\mathbf{W}_1 := \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^T$ and $\mathbf{W}_2 := \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T$. Then the tuple $(\mathbf{W}_1, \mathbf{W}_2, \mathbf{Z})$ is feasible for (C.7) since

$$\begin{bmatrix} \mathbf{W}_1 & \mathbf{X} \\ \mathbf{X}^T & \mathbf{W}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} \boldsymbol{\Sigma} \begin{bmatrix} \mathbf{U}^T & \mathbf{V}^T \end{bmatrix} \succeq \mathbf{0},$$

and $\boldsymbol{\Sigma}$ is a diagonal matrix with non-negative entries. Furthermore, the objective function satisfies $\frac{1}{2}\operatorname{tr}(\mathbf{W}_1 + \mathbf{W}_2) = \|\mathbf{X}\|_*$. Since any feasible solution of (C.7) provides an upper bound for (C.4), we have that the dual norm is less or equal to the nuclear norm which concludes the proof.

Let $\boldsymbol{\sigma} \in \mathbb{R}^r$ denote the vector of singular vales of a given matrix $\mathbf{X}$. By duality theory in vector space $\mathbb{R}^r$ it holds that $\ell_1$-norm and $\ell_\infty$-norm are dual to each other (see [15]). Then, it follows that

$$\|\boldsymbol{\sigma}\|_1 = \sum_{i=1}^{r} \sigma_i = \|\mathbf{X}\|_* \text{ is a dual norm to } \|\boldsymbol{\sigma}\|_\infty = \max_{1\leq i\leq r} \boldsymbol{\sigma}\left(i\right) = \|\mathbf{X}\|_{2\to 2}\,.$$

Therefore, the above example is consistent with the duality theory in $\mathbb{R}^r$, see also [15].

# Bibliography

[1] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.

[2] N. Ailon and H. Rauhut. Fast and RIP-Optimal Transforms. *Discrete & Computational Geometry*, 52(4):780–798, 2014.

[3] U. Ayaz and H. Rauhut. Nonuniform sparse recovery with subgaussian matrices. *Electronic Transactions on Numerical Analysis*, 41:167–178, 2014.

[4] B. Barak and A. Moitra. Noisy Tensor Completion via the Sum-of-Squares Hierarchy. In *Proceedings of the 29th Annual Conference on Learning Theory (COLT), Columbia University, New-York City, USA, 23-26 June 2016*, pages 417–445, 2016.

[5] D. Bayer and D. Mumford. What Can Be Computed in Algebraic Geometry? In *Computational Algebraic Geometry and Commutative Algebra*, pages 1–48. University Press, 1992.

[6] A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[7] J. Bennett and S. Lanning. The Netflix Prize. In *KDD Cup and Workshop in conjunction with KDD*, 2007.

[8] R. Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, 1997.

[9] G. Blekherman, P. A. Parrilo, and R. R. Thomas. *Semidefinite optimization and convex algebraic geometry*. SIAM, 2013.

[10] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

[11] T. Blumensath and M. E. Davies. Normalized Iterative Hard Thresholding: Guaranteed Stability and Performance. *Journal of Selected Topics in Signal Processing*, 4(2):298–309, 2010.

[12] J. Bourgain. An improved estimate in the restricted isometry problem. In B. Klartag and E. Milman, editors, *Geometric Aspects of Functional Analysis*, volume 2116 of *Lecture Notes in Mathematics*, pages 65–70. Springer International Publishing, 2014.

[13] J. Bourgain, S. J. Dilworth, K. Ford, S. Konyagin, and D. Kutzarova. Breaking the $\mathrm{k}^2$ barrier for explicit RIP matrices. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 637–644, 2011.

[14] J. Bourgain, S. J. Dilworth, K. Ford, S. Konyagin, and D. Kutzarova. Explicit constructions of RIP matrices and related problems. *Duke Mathematical Journal*, 159(1):145–185, 2011.

[15] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 1st edition, 2004.

[16] W. Bruns and A. Conca. Gröbner Bases and Determinantal Ideals. In J. Herzog and V. Vuletescu, editors, *Commutative Algebra, Singularities and Computer Algebra*, volume 115 of *NATO Science Series*, pages 9–66. Springer Netherlands, 2003.

[17] B. Buchberger. Bruno Buchberger's PhD thesis 1965: An algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal. *Journal of Symbolic Computation*, 41(3-4):475–511, 2006.

[18] T. T. Cai and A. Zhang. Sparse Representation of a Polytope and Recovery of Sparse Signals and Low-Rank Matrices. *IEEE Transactions on Information Theory*, 60(1):122–132, 2014.

[19] C. F. Caiafa and A. Cichocki. Computing Sparse Representations of Multidimensional Signals Using Kronecker Bases. *Neural Computation*, 25(1):186–220, 2013.

[20] C. F. Caiafa and A. Cichocki. Multidimensional compressed sensing and their applications. *WIREs Data Mining and Knowledge Discovery*, 3(6):355–380, 2013.

[21] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase Retrieval via Matrix Completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013.

[22] E. J. Candès and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[23] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[24] E. J. Candès and B. Recht. Simple bounds for recovering low-complexity models. *Mathematical Programming*, 141(1-2):577–589, 2013.

[25] E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[26] E. J. Candès and T. Tao. Decoding by Linear Programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[27] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[28] E. J. Candès, T. Tao, and J. K. Romberg. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[29] L. Caniglia, J. A. Guccione, and J. J. Guccione. Ideals of generic minors. *Communications in Algebra*, 18(8):2633–2640, 1990.

[30] E. Carlini and J. Kleppe. Ranks derived from multilinear maps. *Journal of Pure and Applied Algebra*, 215(8):1999–2004, 2011.

[31] A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[32] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[33] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic Decomposition by Basis Pursuit. *SIAM Review*, 43(1):129–159, 2001.

[34] Y. Chen. Incoherence-Optimal Matrix Completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

[35] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, pages 211–231, 2009.

[36] D. A. Cox, J. Little, and D. O'Shea. *Using Algebraic Geometry*. Springer, 2nd edition, 2005.

[37] D. A. Cox, J. Little, and D. O'Shea. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra, 3/e (Undergraduate Texts in Mathematics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[38] C. Da Silva and F. J. Herrmann. Optimization on the Hierarchical Tucker manifold – applications to tensor completion. *Linear Algebra and its Applications*, 481:131–173, 2015.

[39] W. Dai and O. Milenkovic. Subspace Pursuit for Compressive Sensing Signal Reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, 2009.

[40] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

[41] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010.

[42] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok. Introduction to compressed sensing. In *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2011.

[43] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[44] L. De Lathauwer and J. Vandewalle. Dimensionality reduction in higher-order signal processing and rank-$(R_1, R_2, ..., R_N)$ reduction in multilinear algebra. *Linear Algebra and its Applications*, 391:31–55, 2004.

[45] J. H. de Morais Goulart and G. Favier. An iterative hard thresholding algorithm with improved convergence for low-rank tensor recovery. In *2015 European Signal Processing Conference (EUSIPCO 2015)*, Nice, France, 2015. Accepted for publication in the Proceedings of the European Signal Processing Conference (EUSIPCO) 2015.

[46] V. de Silva and L.-H. Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1084–1127, 2008.

[47] R. A. DeVore. Deterministic constructions of compressed sensing matrices. *Journal of Complexity*, 23(46):918–925, 2007. Festschrift for the 60th Birthday of Henryk Woniakowski.

[48] S. Dirksen. Dimensionality Reduction with Subgaussian Matrices: A Unified Theory. *Foundations of Computational Mathematics*, pages 1–30, 2015.

[49] S. Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20(53):1–29, 2015.

[50] D. L. Donoho and Y. Tsaig. Fast Solution of $\ell_1$-Norm Minimization Problems When the Solution May Be Sparse. *IEEE Transactions on Information Theory*, 54(11):4789–4812, 2008.

[51] M. F. Duarte and R. G. Baraniuk. Kronecker product matrices for compressive sensing. In *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 3650–3653. IEEE, 2010.

[52] M. F. Duarte and R. G. Baraniuk. Kronecker compressive sensing. *IEEE Transactions on Image Processing*, 21(2):494–504, 2012.

[53] R. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

[54] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[55] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[56] M. Fornasier and H. Rauhut. Iterative thresholding algorithms. *Applied and Computational Harmonic Analysis*, 25(2):187–208, 2008.

[57] M. Fornasier and H. Rauhut. Recovery algorithms for vector-valued data with joint sparsity constraints. *SIAM Journal on Numerical Analysis*, 46(2):577–613, 2008.

[58] M. Fornasier, H. Rauhut, and R. Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM Journal on Optimization*, 21(4):1614–1640, 2011.

[59] S. Foucart, A. Pajor, H. Rauhut, and T. Ullrich. The Gelfand widths of $\ell_p$-balls for $0 < p \leq 1$. *Journal of Complexity*, 26(6):629–640, 2010.

[60] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.

[61] S. Friedland and L.-H. Lim. Nuclear Norm of Higher-Order Tensors. *arXiv preprint arXiv:1410.6072*, 2014.

[62] S. Gandy, B. Recht, and I. Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):19pp, 2011.

[63] A. Y. Garnaev and E. D. Gluskin. On widths of the Euclidean ball. *Soviet Mathematics, Doklady*, 30:200–204, 1984.

[64] R. Gemulla, E. Nijkamp, P. J. Haas, and Y. Sismanis. Large-scale Matrix Factorization with Distributed Stochastic Gradient Descent. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 69–77, New York, NY, USA, 2011. ACM.

[65] J. Geng, X. Yang, X. Wang, and L. Wang. An Accelerated Iterative Hard Thresholding Method for Matrix Completion. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 8(7):141–150, 2015.

[66] G. H. Golub and C. F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[67] J. Gouveia, M. Laurent, P. A. Parrilo, and R. R. Thomas. A new semidefinite programming hierarchy for cycles in binary matroids and cuts in graphs. *Mathematical Programming*, 133(1):203–225, 2012.

[68] J. Gouveia, P. A. Parrilo, and R. R. Thomas. Theta Bodies for Polynomial Ideals. *SIAM Journal on Optimization*, 20(4):2097–2118, 2010.

[69] F. Grande and R. Sanyal. Theta rank, levelness, and matroid minors. *arXiv preprint arXiv:1408.1262*, 2014.

[70] L. Grasedyck. Hierarchical Singular Value Decomposition of Tensors. *SIAM Journal on Matrix Analysis and Applications*, 31(4):2029–2054, 2010.

[71] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(12):3320–3325, 2003.

[72] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[73] D. Gross, F. Krahmer, and R. Kueng. Improved Recovery Guarantees for Phase Retrieval from Coded Diffraction Patterns. *Applied and Computational Harmonic Analysis*, pages –, 2015.

[74] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010.

[75] W. Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 42 of *Springer series in computational mathematics*. Springer, Heidelberg, 2012.

[76] W. Hackbusch and S. Kühn. A new scheme for the tensor representation. *Journal of Fourier Analysis and Applications*, 15(5):706–722, 2009.

[77] I. Haviv and O. Regev. The restricted isometry property of subsampled fourier matrices. In R. Krauthgamer, editor, *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 288–297. SIAM, 2016.

[78] C. Hegde, P. Indyk, and L. Schmidt. Approximation algorithms for model-based compressive sensing. *IEEE Transactions on Information Theory*, 61(9):5129–5147, 2015.

[79] R. Henrion. N-way principal component analysis: theory, algorithms and applications. *Chemometrics and intelligent laboratory systems*, 25(1):1–23, 1994.

[80] T. Hibi. Distributive lattices, affine semigroup rings and algebras with straightening laws. Commutative algebra and combinatorics, US-Jap. joint Semin., Kyoto/Jap. 1985, Advanced Studies in Pure Mathematics 11, 93-109 (1987), 1987.

[81] C. J. Hillar and L.-H. Lim. Most Tensor Problems Are NP-Hard. *Journal of the ACM*, 60(6):45:1–45:39, 2013.

[82] J. Håstad. Tensor rank is NP-complete. *Journal of Algorithms*, 11(4):644–654, 1990.

[83] B. Huang, C. Mu, D. Goldfarb, and J. Wright. Provable models for robust low-rank tensor completion. *Pacific Journal of Optimization*, 11(2):339–364, 2015.

[84] M. A. Iwen. Simple deterministically constructible RIP matrices with sublinear fourier sampling requirements. In *43rd Annual Conference on Information Sciences and Systems*, pages 870–875, 2009.

[85] P. Jain, R. Meka, and I. S. Dhillon. Guaranteed Rank Minimization via Singular Value Projection. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 937–945. Curran Associates, Inc., 2010.

[86] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank Matrix Completion Using Alternating Minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 665–674, New York, NY, USA, 2013. ACM.

[87] R. Johnson. On a theorem stated by Eckart and Young. *Psychometrika*, 28(3):259–263, 1963.

[88] M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege. Stable low-rank matrix recovery via null space properties. *arXiv preprint arXiv:1507.07184*, 2015.

[89] M. Kabanava and H. Rauhut. Analysis $\ell_1$-recovery with Frames and Gaussian Measurements. *Acta Applicandae Mathematicae*, 140(1):173–195, 2015.

[90] L. Karlsson, D. Kressner, and A. Uschmajew. Parallel algorithms for tensor completion in the CP format. *Parallel Computing*, pages –, 2015.

[91] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[92] H. Kiers and I. Van Mechelen. Three-way component analysis: Principles and illustrative application. *Psychological methods*, 6(1):84–110, 2001.

[93] D. E. Knuth. Permutations, matrices, and generalized Young tableaux. *Pacific Journal of Mathematics*, 34(3):709–727, 1970.

[94] T. G. Kolda and B. W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500, 2009.

[95] F. Krahmer, S. Mendelson, and H. Rauhut. Suprema of Chaos Processes and the Restricted Isometry Property. *Communications on Pure and Applied Mathematics*, 67(11):1877–1904, 2014.

[96] F. Krahmer and H. Rauhut. Structured random measurements in signal processing. *GAMM-Mitteilungen*, 37(2):217–238, 2014.

[97] F. Krahmer and R. Ward. New and improved Johnson-Lindenstrauss embeddings via the Restricted Isometry Property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.

[98] N. Kreimer and M. D. Sacchi. Nuclear norm minimization and tensor completion in exploration seismology. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 4275–4279. IEEE, 2013.

[99] D. Kressner, M. Steinlechner, and B. Vandereycken. Low-rank tensor completion by Riemannian optimization. *BIT Numerical Mathematics*, 54(2):447–468, 2014.

[100] R. Kueng, H. Rauhut, and U. Terstiege. Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis*, pages –, 2015.

[101] J. Landsberg. *Tensors: geometry and applications*, volume 128. American Mathematical Society, 2012.

[102] J. B. Lasserre. Convexity in semialgebraic geometry and polynomial optimization. *SIAM Journal on Optimization*, 19(4):1995–2014, 2009.

[103] K. Lee and Y. Bresler. ADMiRA: Atomic Decomposition for Minimum Rank Approximation. *IEEE Transactions on Information Theory*, 56(9):4402–4416, 2010.

[104] J. Levin. Three-mode factor analysis. *Dissertation Abstracts International*, 24:5530–5531, 1963.

[105] N. Li and B. Li. Tensor completion for on-board compression of hyperspectral images. In *Proceedings of the International Conference on Image Processing, ICIP 2010, September 26-29, Hong Kong, China*, pages 517–520. IEEE, 2010.

[106] Y. Li, J. Yan, Y. Zhou, and J. Yang. Optimum Subspace Learning and Error Correction for Tensors. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III*, volume 6313 of *Lecture Notes in Computer Science*, pages 790–803. Springer, 2010.

[107] L.-H. Lim and P. Comon. Multiarray signal processing: Tensor decomposition meets compressed sensing. *Comptes Rendus Mecanique*, 338(6):311–320, 2010.

[108] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor Completion for Estimating Missing Values in Visual Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013.

[109] Y. Liu, F. Shang, W. Fan, J. Cheng, and H. Cheng. Generalized Higher-Order Orthogonal Iteration for Tensor Decomposition and Completion. In *Advances in Neural Information Processing Systems*, pages 1763–1771, 2014.

[110] L. Lovász. On the Shannon Capacity of a Graph. *IEEE Transaction on Information Theory*, 25(1):1–7, 2006.

[111] Y. Ma. On The Minors Defined By A Generic Matrix. *Journal of Symbolic Computation*, 18(6):503–518, 1994.

[112] A. Maleki and D. L. Donoho. Optimally tuned iterative reconstruction algorithms for compressed sensing. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):330–341, 2010.

[113] C. D. Meyer, editor. *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

[114] K. Mohan and M. Fazel. Iterative Reweighted Algorithms for Matrix Rank Minimization. *Journal of Machine Learning Research*, 13(1):3441–3473, 2012.

[115] C. Mu, B. Huang, J. Wright, and D. Goldfarb. Square Deal: Lower Bounds and Improved Relaxations for Tensor Recovery. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Proceedings*, pages 73–81. JMLR.org, 2014.

[116] D. Muti and S. Bourennane. Multidimensional filtering based on a tensor approach. *Signal Processing*, 85(12):2338–2353, 2005.

[117] H. Narasimhan. The irreducibility of ladder determinantal varieties. *Journal of Algebra*, 102(1):162–185, 1986.

[118] D. Needell and J. A. Tropp. CoSaMP: Iterative Signal Recovery from Incomplete and Inaccurate Samples. *Communications of the ACM*, 53(12):93–100, 2010.

[119] D. Needell and R. Vershynin. Uniform Uncertainty Principle and Signal Recovery via Regularized Orthogonal Matching Pursuit. *Foundations of Computational Mathematics*, 9(3):317–334, 2009.

[120] D. Needell and R. Vershynin. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):310–316, 2010.

[121] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer series in operations research and financial engineering. Springer, New York, NY, 2. edition, 2006.

[122] M. R. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403, 2000.

[123] M. R. Osborne, B. Presnell, and B. A. Turlach. On the LASSO and Its Dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

[124] I. V. Oseledets. A new tensor decomposition. *Doklady Mathematics*, 80(1):495–496, 2009.

[125] I. V. Oseledets. Tensor-Train Decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

[126] I. V. Oseledets and E. E. Tyrtyshnikov. Breaking the curse of dimensionality, or how to use SVD in many dimensions. *SIAM Journal on Scientific Computing*, 31(5):3744–3759, 2009.

[127] S. Oymak, B. Recht, and M. Soltanolkotabi. Isometric sketching of any set via Restricted Isometry Property. *arXiv preprint arXiv:1506.03521*, 2015.

[128] I. Pilászy, D. Zibriczky, and D. Tikk. Fast Als-based Matrix Factorization for Explicit and Implicit Feedback Datasets. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 71–78, New York, NY, USA, 2010. ACM.

[129] H. Rauhut. Compressive sensing and structured random matrices. In M. Fornasier, editor, *Theoretical foundations and numerical methods for sparse recovery*, volume 9 of *Radon Series on Computational and Applied Mathematics*, pages 1–92. deGruyter, 2010.

[130] H. Rauhut, R. Schneider, and Ž. Stojanac. Tensor completion in hierarchical tensor representations. In H. Boche, R. Calderbank, G. Kutyniok, and J. Vybiral, editors, *Compressed sensing and its applications*. Springer, 2015.

[131] H. Rauhut, R. Schneider, and Ž. Stojanac. Low rank tensor recovery via iterative hard thresholding. *arXiv preprint arXiv:1602.05217*, 2016.

[132] H. Rauhut and Ž. Stojanac. Tensor theta norms and low rank recovery. *arXiv preprint arXiv:1505.05175*, 2015.

[133] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review*, 52(3):471–501, 2010.

[134] B. Recht and C. Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

[135] J. D. M. Rennie and N. Srebro. Fast Maximum Margin Matrix Factorization for Collaborative Prediction. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 713–719, New York, NY, USA, 2005. ACM.

[136] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil. Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1444–1452, 2013.

[137] B. Romera-Paredes and M. Pontil. A New Convex Relaxation for Tensor Completion. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2967–2975, 2013.

[138] F. Santosa and W. W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986.

[139] O. Semerci, N. Hao, M. E. Kilmer, and E. L. Miller. Tensor-Based Formulation and Nuclear Norm Regularization for Multienergy Computed Tomography. *IEEE Transactions on Image Processing*, 23(4):1678–1693, 2014.

[140] C. E. Shannon. Communication in the presence of noise. *Proceedings of the Institute of Radio Engineers*, 37(1):10–21, 1949.

[141] M. Signoretto, L. De Lathauwer, and J. A. K. Suykens. Nuclear norms for tensors and their use for convex multilinear estimation. Technical Report 10-186 ESAT-SISTA, K.U.Leuven, Leuven, Belgium, 2010.

[142] M. Signoretto, D. Q. Tran, L. De Lathauwer, and J. A. K. Suykens. Learning with tensors: a framework based on convex optimization and spectral regularization. *Machine Learning*, 94(3):303–351, 2014.

[143] L. Sorber, M. Van Barel, and L. De Lathauwer. Tensorlab v2.0. *Available online*, http://www.tensorlab.net/, January 2014.

[144] N. Srebro. *Learning with matrix factorizations.* PhD thesis, Cambridge, MA, USA, 2004. AAI0807530.

[145] M. Steinlechner. Riemannian optimization for high-dimensional tensor completion. Technical Report MATHICSE 5.2015, EPFL Lausanne, Switzerland, 2015.

[146] M. Stojnic. Recovery thresholds for $\ell_1$ optimization in binary compressed sensing. In *Proceedings of 2010 IEEE International Symposium on Information Theory, Austin, TX, USA, June 13-18, 2010*, ISIT 2010, pages 1593–1597. IEEE, 2010.

[147] B. Sturmfels. Gröbner bases and Stanley decompositions of determinantal rings. *Mathematische Zeitschrift*, 205(1):137–144, 1990.

[148] M. A. Sustik, J. A. Tropp, I. S. Dhillon, and R. W. Heath Jr. On the existence of equiangular tight frames. *Linear Algebra and its Applications*, 426(23):619–635, 2007.

[149] M. Talagrand. Regularity of gaussian processes. *Acta Mathematica*, 159(1):99–149, 1987.

[150] M. Talagrand. Majorizing measures without measures. *The Annals of Probability*, 29(1):411–417, 2001.

[151] M. Talagrand. *Upper and lower bounds for stochastic processes*, volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Heidelberg, 2014.

[152] J. Tanner and K. Wei. Normalized iterative hard thresholding for matrix completion. *SIAM Journal on Scientific Computing*, 35(5):S104–S125, 2013.

[153] C. Teflioudi, F. Makari, and R. Gemulla. Distributed matrix completion. In *12th International Conference on Data Mining (ICDM), 2012 IEEE. Proceedings of the meeting held 10-13 December 2012, Brussels, Belgium*, pages 655–664, 2012.

[154] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society (Series B)*, 58(1):267–288, 1996.

[155] A. M. Tillmann and M. E. Pfetsch. The Computational Complexity of the Restricted Isometry Property, the Nullspace Property, and Related Concepts in Compressed Sensing. *IEEE Transactions on Information Theory*, 60(2):1248–1259, 2014.

[156] R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima. Statistical Performance of Convex Tensor Decomposition. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 972–980, 2011.

[157] J. A. Tropp. Greed is Good: Algorithmic Results for Sparse Approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2006.

[158] L. R. Tucker. Implications of factor analysis of three-way matrices for measurement of change. *Problems in Measuring Change*, pages 122–137, 1963.

[159] L. R. Tucker. The extension of factor analysis to three-dimensional matrices. In H. Gulliksen and N. Frederiksen, editors, *Contributions to mathematical psychology*, pages 110–127. Holt, Rinehart and Winston, New York, 1964.

[160] R. Vanderbei, H. Liu, L. Wang, and K. Lin. Optimization for Compressed Sensing: the Simplex Method and Kronecker Sparsification. *arXiv preprint arXiv:1312.4426*, 2013.

[161] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: TensorFaces. *Lecture Notes in Computer Science*, 2350:447–460, 2002.

[162] R. Vershynin. On large random almost euclidean bases. *Acta Mathematica Universitatis Comenianae. New Series*, 69(2):137–144, 2000.

[163] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing*, pages 210–268. Cambridge University Press, 2012. Cambridge Books Online.

[164] G. Vidal. Efficient classical simulation of slightly entangled quantum computations. *Physical Review Letters*, 91(14):147902, 2003.

[165] S. Waldron. On the construction of equiangular frames from graphs. *Linear Algebra and its Applications*, 431(11):2228–2242, 2009.

[166] S. J. Wright. *Primal-dual Interior-point Methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1997.

[167] H.-F. Yu, C.-J. Hsieh, S. Si, and I. S. Dhillon. Scalable Coordinate Descent Approaches to Parallel Matrix Factorization for Recommender Systems. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, pages 765–774, Washington, DC, USA, 2012. IEEE Computer Society.

[168] H.-F. Yu, C.-J. Hsieh, S. Si, and I. S. Dhillon. Parallel matrix factorization for recommender systems. *Knowledge and Information Systems*, 41(3):793–819, 2014.

[169] M. Yuan and C.-H. Zhang. On Tensor Completion via Nuclear Norm Minimization. *Foundations of Computational Mathematics*, pages 1–38, 2015.

[170] S. Zhou, L. Kong, Z. Luo, and N. Xiu. New RIC Bounds via l_q-minimization with 0<q<=1 in Compressed Sensing. *arXiv preprint arXiv:1308.0455*, 2013.

[171] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan. Large-Scale Parallel Collaborative Filtering for the Netflix Prize. In R. Fleischer and J. Xu, editors, *Algorithmic Aspects in Information and Management*, volume 5034 of *Lecture Notes in Computer Science*, pages 337–348. Springer Berlin Heidelberg, 2008.