

On decomposability and subdifferential of the tensor nuclear norm

Jiewen GUAN ^{*} Bo JIANG [†] Zhening LI [‡]

October 7, 2025

Abstract

We study the decomposability and the subdifferential of the tensor nuclear norm. Both concepts are well understood and widely applied in matrices but remain unclear for higher-order tensors. We show that the tensor nuclear norm admits a full decomposability over specific subspaces and determine the largest possible subspaces that allow the full decomposability. We derive novel inclusions of the subdifferential of the tensor nuclear norm and study its subgradients in a variety of subspaces of interest. All the results hold for tensors of an arbitrary order. As an immediate application, we establish the statistical performance of the tensor robust principal component analysis, the first such result for tensors of an arbitrary order.

Keywords: tensor nuclear norm, tensor spectral norm, decomposability, subdifferential, subspace, robust principal component analysis, exact recovery, random tensor

Mathematics Subject Classification (2020): 15A60, 15A69, 62B10, 68Q25, 90C25

1 Introduction

The matrix nuclear norm, as the convex envelope of the matrix rank, has found a wide range of applications, particularly in mathematical optimization to search for low-rank matrices. Its essential properties underlying the applications are the decomposability and the explicit characterization of the subdifferential. Specifically, the decomposability states that

$$\|\mathbf{T} + \mathbf{S}\|_* = \|\mathbf{T}\|_* + \|\mathbf{S}\|_* \text{ if } \mathbf{T}^T \mathbf{S} = \mathbf{O} \text{ and } \mathbf{T} \mathbf{S}^T = \mathbf{O} \quad (1)$$

and the explicit characterization of the subdifferential states that

$$\partial \|\mathbf{T}\|_* = \{\mathbf{U}\mathbf{V}^T + \mathbf{X} : \mathbf{X}^T \mathbf{U} = \mathbf{O}, \mathbf{X}\mathbf{V} = \mathbf{O}, \|\mathbf{X}\|_\sigma \leq 1\}, \quad (2)$$

where $\mathbf{T} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ is a compact singular value decomposition (SVD) and $\|\bullet\|_*$ and $\|\bullet\|_\sigma$ denote the nuclear norm and the spectral norm, respectively. They play a critical role in many matrix optimization problems and statistical learning, such as compressed sensing [45, Corollary 5], matrix recovery and completion [12, 13], multivariate regression [45, Corollary 3], phase retrieval [57, Chapter 10.4], recommender systems [57, Example 10.2], and robust principal component analysis (PCA) [11]; see a recent book [57] and the references therein for more examples.

^{*}Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong. Email: seemjwguan@gmail.com

[†]School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China. Email: isyebojiang@gmail.com

[‡]School of Mathematics and Physics, University of Portsmouth, Portsmouth PO1 3HF, United Kingdom. Email: zheningli@gmail.com

With the rapid developments of data sciences in various fields, the recent decade has witnessed a significant increase in research activities in tensors, the higher-order generalization of matrices. In the same spirit to matrices, the tensor nuclear norm is widely recognized as a convex surrogate for the intrinsic complexity of a tensor [17] and has triggered many applications in statistical learning such as tensor completion [61], tensor regression [50], and tensor robust PCA [19]. However, the nice properties of the matrix nuclear norm, (1) and (2), do not carry over directly to the tensor nuclear norm. This has made many low-rank tensor optimization problems difficult and statistical implications on low-rank tensors unclear. In fact, the only known result on the decomposability of the tensor nuclear norm is the so-called weak decomposability [50, Lemma 1] that only works for third-order tensors. Although it has found applications in tensor regression [50], it does not completely address the entire issue. For the subdifferential of the tensor nuclear norm, only two limited inclusions, [61, Lemma 1] and [62, Theorem 1], have been proposed. While these inclusions have been applied to analyze the statistical performance of tensor completion [61, 62], they are still unsatisfactory from the perspective of tensor analysis since many obvious subgradients have been excluded.

The main reason behind the full decomposability and the subdifferential characterization of the matrix nuclear norm is that every matrix admits an SVD. As a contrast, only a very special class of tensors admits SVDs. While this unfortunate fact is primarily responsible for the gaps of the properties, it indeed makes the problems subtler and more interesting. This paper aims to offer a more in-depth understanding of the decomposability and the subdifferential of the tensor nuclear norm.

As the first main result, we find that the tensor nuclear norm can in fact be fully decomposable if we carefully choose the tensor subspaces in which $\mathbf{T}$ and $\mathbf{S}$ reside as those in (1). It also directly points out why only a weak decomposability is possible for the subspace considered in [50, Lemma 1]. Our full decomposability applies to tensors of an arbitrary order, leading to a very natural generalization of the matrix case in (1). Moreover, we optimize the subspaces that allow the full decomposability and determine the largest possible such subspace pairs. The study has also resulted a dual byproduct, the decomposability of the tensor spectral norm, running in parallel to the tensor nuclear norm.

The full decomposability of the tensor nuclear norm offers a tool to study its subdifferential. As the next main result, we propose novel subdifferential inclusions of the tensor nuclear norm. They strictly enlarge the inclusion proposed in [62, Theorem 1] that is the only known subdifferential inclusion for tensors of an arbitrary order. In particular, we show that a full spectral ball, $\{\mathbf{X} : \|\mathbf{X}\|_\sigma \leq 1\}$, can be imposed in (2) rather than $\|\mathbf{X}\|_\sigma \leq \frac{2}{d(d-1)}$ imposed in [62, Theorem 1], where d is the order of the tensor. Our study indicates that there is no universal way to explicitly characterize the subdifferential of the tensor nuclear norm as that of the matrix nuclear norm in (2), supported by various approximations of the subdifferential and several interesting examples. Moreover, we investigate subgradients of the tensor nuclear norm and derive various bounds for the inclusion and exclusion of the subdifferential in all relevant subspaces.

We believe that these developments can be important to applications, no matter in theory or in practice. As a precursor, we propose an immediate application to analyze the statistical performance of the nuclear-norm-based tensor robust PCA that aims to recover a planted low-rank ground-truth tensor superposed by a sparse corruption. In the matrix case, the robust PCA is already a remarkable instance of modern compressed sensing and has found many interesting applications such as video surveillance, face recognition, and community detection; see, e.g., [11, 30, 59] and the references therein. A key component underlying its outstanding performance is the nuclear norm minimization that tends to reveal low-rank solutions. To the best of our knowledge, the statistical performance of

the tensor robust PCA has only been established for third-order tensors [19]. Our result applies to tensors of an arbitrary order and exactly recovers the matrix case [11, Theorem 1.1] when $d = 2$. Moreover, it enjoys some looser conditions for the exact recovery than those required in [19] when $d = 3$. One interesting insight from our analysis suggests that the conditions for the exact recovery of the tensor robust PCA of every order are highly likely to be identical.

The rest of this paper is organized as follows. We start with uniform notations, main concepts, and important properties of tensor operations and norms in Sections 2. As an essential part to the main results, we introduce tensor subspaces, projections, and tensor norms in subspaces in Section 3, presented in a self-contained way for readers with minimal background knowledge. The decomposability of the tensor nuclear norm is discussed in Section 4 along with the decomposability of the tensor spectral norm. The subdifferential of the tensor nuclear norm is discussed in Section 5. As an application, the statistical performance of the tensor robust PCA is analyzed in Section 6. Finally, we conclude this paper with some remarks and future research directions in Section 7.

2 Preparations

To support a better understanding of the theoretical results in later sections, we compile this section on uniform notations, necessary concepts and important properties of tensor operations, and tensor spectral and nuclear norms.

2.1 Basic notations

Throughout this paper, we uniformly adopt lowercase letters (e.g., x), boldface lowercase letters (e.g., $\mathbf{x} = (x_i)$), boldface capital letters (e.g., $\mathbf{X} = (x_{i_1 i_2 \dots i_d})$) to denote scalars, vectors, and d th order tensors (including matrices) with $d \geq 2$, respectively. Denote $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ to be the space of d th order real tensors of dimension $n_1 \times n_2 \times \dots \times n_d$. The same notation applies to a vector space and a matrix space when $d = 1$ and $d = 2$, respectively. Without loss of generality, we assume that $2 \leq n_1 \leq n_2 \leq \dots \leq n_d$ and that the order of the tensor space, d , is a fixed parameter. All Greek letters are used to denote constants. In particular, $\theta > 0$ and $\kappa > 0$ are some sufficiently small and sufficiently large constants that are not made explicit in Section 6, respectively.

The Frobenius inner product of two tensors $\mathbf{T}, \mathbf{S} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ is defined as

$$\langle \mathbf{T}, \mathbf{S} \rangle := \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \dots \sum_{i_d=1}^{n_d} t_{i_1 i_2 \dots i_d} s_{i_1 i_2 \dots i_d}.$$

Its induced Frobenius norm is naturally defined as $\|\mathbf{T}\|_2 := \sqrt{\langle \mathbf{T}, \mathbf{T} \rangle}$. When $d = 1$, the Frobenius norm reduces to the Euclidean norm of a vector. In a similar vein, we may define the ℓ_p -norm of a tensor (also known as the Hölder p -norm) for $1 \leq p \leq \infty$ by viewing the tensor as a vector, i.e.,

$$\|\mathbf{T}\|_p := \left(\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \dots \sum_{i_d=1}^{n_d} |t_{i_1 i_2 \dots i_d}|^p \right)^{\frac{1}{p}}.$$

Three specific Hölder p -norms are used in this paper, namely $\|\mathbf{T}\|_1 = \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \dots \sum_{i_d=1}^{n_d} |t_{i_1 i_2 \dots i_d}|$, $\|\mathbf{T}\|_2$ as the Frobenius norm, and $\|\mathbf{T}\|_\infty$ as the largest entry of $\mathbf{T}$ in absolute value.

All blackboard bold capital letters denote sets, such as $\mathbb{R}^n$, the Euclidean unit sphere $\mathbb{S}^n := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1\}$ embedded in $\mathbb{R}^n$, and the set of positive integers $\mathbb{N}$. We denote $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ to be the standard basis of $\mathbb{R}^n$. For any $n \in \mathbb{N}$, we let $[n] := \{1, 2, \dots, n\}$. We denote $\mathbb{I}^d := \{\mathbf{i} \in \mathbb{N}^d :$

$i_k \in [n_k] \forall k \in [d]$ to be the set of entry indices for the tensor space $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$. As a result, the set $\{\mathbf{e}_{i_1} \otimes \mathbf{e}_{i_2} \otimes \dots \otimes \mathbf{e}_{i_d} : \mathbf{i} \in \mathbb{I}^d\}$ becomes the standard basis of $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, where $\otimes$ denotes the vector outer product.

Finally, two most frequently used operations of sets are $\mathbf{p}$ for the orthogonal projection and $\mathbf{sp}$ for the span. Some other notations are self-explanatory, including $\dim$ for the dimension of a space, conv for the convex hull of a set, Pr for the probability measure, and Ex for the expectation.

2.2 Tensor operations

A tensor $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ has d modes. Fixing entry indices of $d - 1$ modes except mode k results a vector in $\mathbb{R}^{n_k}$, called a mode- k fiber. For matrices, mode-1 fibers are columns and mode-2 fibers are rows. The mode- k matricization of $\mathbf{T}$, denoted by $\mathbf{T}_{(k)} \in \mathbb{R}^{n_k \times \prod_{i \neq k} n_i}$, is to arrange mode- k fibers to be the columns of the resulting matrix. The mode- k product between $\mathbf{T}$ and a matrix $\mathbf{X} \in \mathbb{R}^{n \times n_k}$, denoted by $\mathbf{T} \times_k \mathbf{X} \in \mathbb{R}^{n_1 \times \dots \times n_{k-1} \times n \times n_{k+1} \times \dots \times n_d}$, changes every mode- k fiber of $\mathbf{T}$, say $\mathbf{v} \in \mathbb{R}^{n_k}$, to $\mathbf{X}\mathbf{v} \in \mathbb{R}^n$, a mode- k fiber of $\mathbf{T} \times_k \mathbf{X}$; in another word,

$$\mathbf{S} = \mathbf{T} \times_k \mathbf{X} \iff \mathbf{S}_{(k)} = \mathbf{X}\mathbf{T}_{(k)}.$$

The mode- k contraction of $\mathbf{T}$ by a vector $\mathbf{v} \in \mathbb{R}^{n_k}$ is a tensor $\mathbf{T} \times_k \mathbf{v} \in \mathbb{R}^{n_1 \times \dots \times n_{k-1} \times n_{k+1} \times \dots \times n_d}$ of order $d - 1$ under the same mode- k product rule by treating $\mathbf{v}$ as a $1 \times n_k$ matrix.

A rank-one tensor (also called a simple tensor) is a tensor that can be written as outer products of vectors, e.g., $\mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \dots \otimes \mathbf{x}_d$. It is easy to verify that

$$\|\mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \dots \otimes \mathbf{x}_d\|_2 = \prod_{k=1}^d \|\mathbf{x}_d\|_2.$$

Any tensor $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ uniquely defines a multilinear form

$$\mathbf{T}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_d) := \langle \mathbf{T}, \mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \dots \otimes \mathbf{x}_d \rangle = \mathbf{T} \times_1 \mathbf{x}_1 \times_2 \mathbf{x}_2 \dots \times_d \mathbf{x}_d$$

of vector entries $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_d)$ where $\mathbf{x}_k \in \mathbb{R}^{n_k}$ for $k \in [d]$. If any vector entry, say $\mathbf{x}_1$, is missing and replaced by a $\bullet$, then

$$\mathbf{T}(\bullet, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_d) = \mathbf{T} \times_2 \mathbf{x}_2 \times_3 \mathbf{x}_3 \dots \times_d \mathbf{x}_d \in \mathbb{R}^{n_1}$$

becomes a vector. Similarly, $\mathbf{T}(\bullet, \bullet, \mathbf{x}_3, \mathbf{x}_4, \dots, \mathbf{x}_d) \in \mathbb{R}^{n_1 \times n_2}$ is a matrix, and so on. For a thorough introduction to tensor operations, we refer interested readers to Kolda and Bader [32], Ballard and Kolda [1], and Nie [46, Chapter 11].

2.3 Tensor spectral norm and nuclear norm

Given a tensor $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, its spectral norm is defined as

$$\|\mathbf{T}\|_\sigma := \max\{\langle \mathbf{T}, \mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \dots \otimes \mathbf{x}_d \rangle : \|\mathbf{x}_k\|_2 = 1 \forall k \in [d]\}, \quad (3)$$

and its nuclear norm is defined as

$$\|\mathbf{T}\|_* := \min\left\{\sum_{i=1}^r |\lambda_i| : \mathbf{T} = \sum_{i=1}^r \lambda_i \mathbf{x}_1^i \otimes \mathbf{x}_2^i \otimes \dots \otimes \mathbf{x}_d^i, \|\mathbf{x}_k^i\|_2 = 1 \forall i \in [r] \text{ and } k \in [d], r \in \mathbb{N}\right\}; \quad (4)$$

see, e.g., [37, 49, 20] for more details. A decomposition $\sum_{i=1}^r \lambda_i \mathbf{x}_1^i \otimes \mathbf{x}_2^i \otimes \dots \otimes \mathbf{x}_d^i$ of $\mathbf{T}$ in (4) that attains $\|\mathbf{T}\|_*$ is called a nuclear decomposition. These two definitions reduce to the matrix spectral

norm and nuclear norm when $d = 2$. However, unlike their matrix counterparts, computing the tensor spectral norm [26] and nuclear norm [20] are both NP-hard when $d \geq 3$.

Written in terms of the multilinear form in (3), the tensor spectral norm enjoys a nice property, i.e.,

$$\begin{aligned}
\|\mathbf{T}\|_\sigma &= \max_{\|\mathbf{x}_k\|_2=1 \forall k \in [d]} \mathbf{T}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_d) \\
&= \max_{\|\mathbf{x}_k\|_2=1 \forall k \in [d-1]} \|\mathbf{T}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{d-1}, \bullet)\|_2 \\
&= \max_{\|\mathbf{x}_k\|_2=1 \forall k \in [d-2]} \|\mathbf{T}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{d-2}, \bullet, \bullet)\|_\sigma \\
&= \max_{\|\mathbf{x}_k\|_2=1 \forall k \in [d-3]} \|\mathbf{T}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{d-3}, \bullet, \bullet, \bullet)\|_\sigma,
\end{aligned} \tag{5}$$

and so on.

The tensor spectral norm is the dual norm of the tensor nuclear norm, and vice versa.

Lemma 2.1 *If two tensors $\mathbf{T}, \mathbf{S} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, then*

$$\langle \mathbf{T}, \mathbf{S} \rangle \leq \|\mathbf{T}\|_\sigma \|\mathbf{S}\|_*.$$

Moreover,

$$\|\mathbf{T}\|_\sigma = \max_{\|\mathbf{Z}\|_* \leq 1} \langle \mathbf{T}, \mathbf{Z} \rangle \text{ and } \|\mathbf{T}\|_* = \max_{\|\mathbf{Z}\|_\sigma \leq 1} \langle \mathbf{T}, \mathbf{Z} \rangle.$$

A proof can be found in, e.g., [38, Lemma 21]. An optimal tensor $\mathbf{Z}$ of either of the above two optimization problems is called a dual certificate of $\mathbf{T}$. In addition, for any nuclear decomposition $\mathbf{T} = \sum_{i=1}^r \lambda_i \mathbf{x}_1^i \otimes \mathbf{x}_2^i \otimes \dots \otimes \mathbf{x}_d^i$ with $\lambda_i > 0$ for $i \in [r]$ and any dual certificate $\mathbf{Z}$ with $\|\mathbf{Z}\|_\sigma = 1$ and $\|\mathbf{T}\|_* = \langle \mathbf{T}, \mathbf{Z} \rangle$, one always has $\langle \mathbf{Z}, \mathbf{x}_1^i \otimes \mathbf{x}_2^i \otimes \dots \otimes \mathbf{x}_d^i \rangle = 1$ for any $i \in [r]$; see [20, Lemma 4.1].

There are some trivial bounds relating the two norms to the Hölder p -norms mentioned earlier, i.e.,

$$\|\mathbf{T}\|_\infty \leq \|\mathbf{T}\|_\sigma \leq \|\mathbf{T}\|_2 \leq \|\mathbf{T}\|_* \leq \|\mathbf{T}\|_1 \tag{6}$$

for any tensor $\mathbf{T}$; see, e.g., [15, Proposition 4.2].

3 Tensors subspaces

The essential reason resulting the full decomposability of the tensor nuclear norm lies in properly chosen tensor subspaces. In this section, we introduce tensor subspaces and elaborate their notations as a key step to appreciate the main results. To provide a better picture of the tensor subspaces, we also discuss orthogonal transformations and projections for tensors. Finally, we prove key properties of the tensor spectral norm and nuclear norm over subspaces. These properties set a foundation for the main theoretical developments. The section is presented in a self-contained way for readers with minimal background knowledge.

3.1 Subspaces

We consider the general tensor space $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ of order d . Although a tensor space is also a vector space, we exclusively use $\mathbb{T}$ and $\mathbb{U}$ for tensor subspaces and $\mathbb{V}$ for vector subspaces. In particular, we use $\mathbb{V}_k$ to denote a subspace of $\mathbb{R}^{n_k}$ that corresponds to mode k of the tensor space.

Given vector subspaces $\mathbb{V}_k \subseteq \mathbb{R}^{n_k}$ for $k \in [d]$, we denote

$$\mathbb{T}((\mathbb{V}_k)_{k=1}^d) = \mathbb{T}(\mathbb{V}_1, \mathbb{V}_2, \dots, \mathbb{V}_d) := \text{sp} \left(\bigotimes_{k=1}^d \mathbb{V}_k \right) = \text{sp}(\mathbb{V}_1 \otimes \mathbb{V}_2 \otimes \dots \otimes \mathbb{V}_d)$$

to be the tensor subspace spanned by $(\mathbb{V}_1, \mathbb{V}_2, \dots, \mathbb{V}_d)$. Let $\mathbb{V}_k^\perp$ be the orthogonal complement of $\mathbb{V}_k$, i.e., $\mathbb{V}_k \oplus \mathbb{V}_k^\perp = \mathbb{R}^{n_k}$, where $\oplus$ stands for the direct sum. Given an index set $\mathbb{I} \subseteq [d]$, we denote

$$\mathbb{V}_k^\mathbb{I} := \begin{cases} \mathbb{V}_k^\perp & k \in \mathbb{I} \\ \mathbb{V}_k & k \notin \mathbb{I}. \end{cases}$$

For example, we always have $\mathbb{V}_k^\emptyset = \mathbb{V}_k$ and $\mathbb{V}_k^{[d]} = \mathbb{V}_k^\perp$ for $k \in [d]$. For any $\mathbb{I} \subseteq [d]$, we call

$$\mathbb{T}^\mathbb{I}((\mathbb{V}_k)_{k=1}^d) := \mathbb{T}((\mathbb{V}_k^\mathbb{I})_{k=1}^d) = \text{sp} \left(\bigotimes_{k=1}^d \mathbb{V}_k^\mathbb{I} \right)$$

a basic subspace of $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ defined by $(\mathbb{V}_1, \mathbb{V}_2, \dots, \mathbb{V}_d)$. In particular, we have $\mathbb{T}^\emptyset((\mathbb{V}_k)_{k=1}^d) = \mathbb{T}((\mathbb{V}_k)_{k=1}^d) = \text{sp}(\mathbb{V}_1 \otimes \mathbb{V}_2 \otimes \dots \otimes \mathbb{V}_d)$ and $\mathbb{T}^{[d]}((\mathbb{V}_k)_{k=1}^d) = \mathbb{T}((\mathbb{V}_k^\perp)_{k=1}^d) = \text{sp}(\mathbb{V}_1^\perp \otimes \mathbb{V}_2^\perp \otimes \dots \otimes \mathbb{V}_d^\perp)$. It is also easy to see that

$$\mathbb{T}^{[d] \setminus \mathbb{I}}((\mathbb{V}_k)_{k=1}^d) = \mathbb{T}^\mathbb{I}((\mathbb{V}_k^\perp)_{k=1}^d).$$

Given $\mathbb{V}_k \subseteq \mathbb{R}^{n_k}$ for $k \in [d]$, there are exactly 2^d basic subspaces of $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ and

$$\bigoplus_{\mathbb{I} \subseteq [d]} \mathbb{T}^\mathbb{I}((\mathbb{V}_k)_{k=1}^d) = \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}.$$

The following $\mathbb{U}$ subspaces are particularly important in our discussions, i.e.,

$$\begin{aligned} \mathbb{U}^\mathbb{I}((\mathbb{V}_k)_{k=1}^d) &:= \bigoplus_{\mathbb{I} \subseteq \mathbb{J} \subseteq [d]} \mathbb{T}^\mathbb{J}((\mathbb{V}_k)_{k=1}^d) = \text{sp} \left(\left\{ \bigotimes_{k=1}^d \mathbf{v}_k : \mathbf{v}_k \in \mathbb{V}_k^\perp \forall k \in \mathbb{I}, \mathbf{v}_k \in \mathbb{R}^{n_k} \forall k \notin \mathbb{I} \right\} \right), \\ \mathbb{U}_\mathbb{I}((\mathbb{V}_k)_{k=1}^d) &:= \mathbb{U}^\mathbb{I}((\mathbb{V}_k^\perp)_{k=1}^d) = \text{sp} \left(\left\{ \bigotimes_{k=1}^d \mathbf{v}_k : \mathbf{v}_k \in \mathbb{V}_k \forall k \in \mathbb{I}, \mathbf{v}_k \in \mathbb{R}^{n_k} \forall k \notin \mathbb{I} \right\} \right). \end{aligned}$$

For example, we have $\mathbb{U}^\emptyset((\mathbb{V}_k)_{k=1}^d) = \mathbb{U}_\emptyset((\mathbb{V}_k)_{k=1}^d) = \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, $\mathbb{U}^{[d]}((\mathbb{V}_k)_{k=1}^d) = \mathbb{T}((\mathbb{V}_k^\perp)_{k=1}^d)$, and $\mathbb{U}_{[d]}((\mathbb{V}_k)_{k=1}^d) = \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$.

As an illustrating example, if $\mathbb{V}_1$ is a subspace of $\mathbb{R}^{n_1}$ and $\mathbb{V}_2$ is a subspace of $\mathbb{R}^{n_2}$, we have in the matrix space $\mathbb{R}^{n_1 \times n_2}$ that

$$\begin{aligned} \mathbb{T}^\emptyset(\mathbb{V}_1, \mathbb{V}_2) &= \text{sp}(\mathbb{V}_1 \otimes \mathbb{V}_2), & \mathbb{U}^\emptyset(\mathbb{V}_1, \mathbb{V}_2) &= \text{sp}(\mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2}), & \mathbb{U}_\emptyset(\mathbb{V}_1, \mathbb{V}_2) &= \text{sp}(\mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2}), \\ \mathbb{T}^{\{1\}}(\mathbb{V}_1, \mathbb{V}_2) &= \text{sp}(\mathbb{V}_1^\perp \otimes \mathbb{V}_2), & \mathbb{U}^{\{1\}}(\mathbb{V}_1, \mathbb{V}_2) &= \text{sp}(\mathbb{V}_1^\perp \otimes \mathbb{R}^{n_2}), & \mathbb{U}_{\{1\}}(\mathbb{V}_1, \mathbb{V}_2) &= \text{sp}(\mathbb{V}_1 \otimes \mathbb{R}^{n_2}), \\ \mathbb{T}^{\{2\}}(\mathbb{V}_1, \mathbb{V}_2) &= \text{sp}(\mathbb{V}_1 \otimes \mathbb{V}_2^\perp), & \mathbb{U}^{\{2\}}(\mathbb{V}_1, \mathbb{V}_2) &= \text{sp}(\mathbb{R}^{n_1} \otimes \mathbb{V}_2^\perp), & \mathbb{U}_{\{2\}}(\mathbb{V}_1, \mathbb{V}_2) &= \text{sp}(\mathbb{R}^{n_1} \otimes \mathbb{V}_2), \\ \mathbb{T}^{\{1,2\}}(\mathbb{V}_1, \mathbb{V}_2) &= \text{sp}(\mathbb{V}_1^\perp \otimes \mathbb{V}_2^\perp), & \mathbb{U}^{\{1,2\}}(\mathbb{V}_1, \mathbb{V}_2) &= \text{sp}(\mathbb{V}_1^\perp \otimes \mathbb{V}_2^\perp), & \mathbb{U}_{\{1,2\}}(\mathbb{V}_1, \mathbb{V}_2) &= \text{sp}(\mathbb{V}_1 \otimes \mathbb{V}_2). \end{aligned}$$

For a given tensor $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, we denote $\text{sp}_k(\mathbf{T}) \subseteq \mathbb{R}^{n_k}$ to be the subspace spanned by the columns of $\mathbf{T}_{(k)}$, the mode- k matricization of $\mathbf{T}$. In particular, $\text{sp}_1(\mathbf{T})$ is the column space of $\mathbf{T}$ and $\text{sp}_2(\mathbf{T})$ is the row space of $\mathbf{T}$ if $\mathbf{T}$ is a matrix. We adopt the following shorthand notations for the tensor subspaces defined by $\mathbf{T}$,

$$\mathbb{T}^\mathbb{I}(\mathbf{T}) := \mathbb{T}^\mathbb{I}((\text{sp}_k(\mathbf{T}))_{k=1}^d), \quad \mathbb{U}^\mathbb{I}(\mathbf{T}) := \mathbb{U}^\mathbb{I}((\text{sp}_k(\mathbf{T}))_{k=1}^d), \quad \text{and} \quad \mathbb{U}_\mathbb{I}(\mathbf{T}) := \mathbb{U}_\mathbb{I}((\text{sp}_k(\mathbf{T}))_{k=1}^d).$$

We also simply denote $\mathbb{T}(\mathbf{T}) := \mathbb{T}^\emptyset(\mathbf{T})$.

3.2 Orthogonal transformations and projections

Orthogonal transformations preserve many matrix properties, in particular to this paper, keeping the spectral norm and nuclear norm unchanged. Given any matrix $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2}$, there is an orthogonal matrix $\mathbf{P} \in \mathbb{R}^{n_1 \times n_1}$ such that the nonzero entries of $\mathbf{P}\mathbf{T}$ only appear in the first r_1 rows where $r_1 = \dim(\text{sp}_1(\mathbf{T}))$. There is also an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{n_2 \times n_2}$ such that the nonzero entries of $\mathbf{T}\mathbf{Q}$ only appear in the first r_2 columns where $r_2 = \dim(\text{sp}_2(\mathbf{T}))$. Its obvious that $r_1 = r_2$, the rank of $\mathbf{T}$.

In fact, the same can be applied to tensors as well. Given a tensor $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ and a mode $k \in [d]$, there is an orthogonal matrix $\mathbf{P}_k \in \mathbb{R}^{n_k \times n_k}$, such that the entries of $\mathbf{T} \times_k \mathbf{P}_k$ are nonzero only when its mode- k index is at most r_k , where $r_k = \dim(\text{sp}_k(\mathbf{T}))$. Essentially, any mode- k fiber of $\mathbf{T}$, say $\mathbf{v} \in \mathbb{R}^{n_k}$, changes to $\mathbf{P}_k \mathbf{v} \in \mathbb{R}^{n_k}$ whose last $n_k - r_k$ entries become zeros. Therefore, by repeatedly applying mode- k products with orthogonal matrices $\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_d$, we obtain a tensor $\mathbf{S} = \mathbf{T} \times_1 \mathbf{P}_1 \times_2 \mathbf{P}_2 \cdots \times_d \mathbf{P}_d$ such that $s_{i_1 i_2 \dots i_d} = 0$ if there exists a mode index k with $i_k > \dim(\text{sp}_k(\mathbf{T}))$. Besides, this transformation is completely reversible, in particular, $\mathbf{T} = \mathbf{S} \times_1 \mathbf{P}_1^T \times_2 \mathbf{P}_2^T \cdots \times_d \mathbf{P}_d^T$. This orthogonal transformation for tensors is more like an expanded Tucker decomposition [32] whose core tensor is a shrunk version of $\mathbf{S}$ by deleting peripheral zero entries. While we may not explicitly apply this fact in our study, bearing this in mind makes many properties of tensor subspaces easier to understand. For example, when we consider a subspace $\mathbb{V}_k$ or $\text{sp}_k(\mathbf{T})$ of $\mathbb{R}^{n_k}$ with $\dim(\mathbb{V}_k) = r_k$, we can simply treat $\mathbb{V}_k$ as $\mathbb{R}^{r_k} \times \{0\}^{n_k - r_k}$ and $\mathbb{V}_k^\perp$ as $\{0\}^{r_k} \times \mathbb{R}^{n_k - r_k}$.

In fact, the matrix SVD makes the above even better. Since $r_1 = r_2$, SVD results the top-left $r_1 \times r_1$ submatrix of $\mathbf{S}$ to be diagonal with all of the singular values. With that structure, the decomposability of the matrix nuclear norm is straightforward and the subdifferential of the matrix nuclear norm admits an explicit and complete representation; see (1) and (2). However, the diagonal structure does not hold for tensors in general, and this makes the results of decomposability and subdifferential of the tensor nuclear norm unsatisfactory.

Given a subspace $\mathbb{V}$ of $\mathbb{R}^n$, the orthogonal projection operator $p_{\mathbb{V}}$ of a vector or a set of vectors in $\mathbb{R}^n$ is frequently used in this paper. The outer products of projection operators, in particular, $\bigotimes_{k=1}^d p_{\mathbb{V}_k}$ with $\mathbb{V}_k$ being a subspace of $\mathbb{R}^{n_k}$ for $k \in [d]$, of a tensor $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ is defined as follows: Given any rank-one decomposition $\mathbf{T} = \sum_{i=1}^r \mathbf{x}_1^i \otimes \mathbf{x}_2^i \otimes \cdots \otimes \mathbf{x}_d^i$,

$$\left(\bigotimes_{k=1}^d p_{\mathbb{V}_k} \right) (\mathbf{T}) := p_{\mathbb{T}((\mathbb{V}_k)_{k=1}^d)} (\mathbf{T}) = \sum_{i=1}^r p_{\mathbb{V}_1}(\mathbf{x}_1^i) \otimes p_{\mathbb{V}_2}(\mathbf{x}_2^i) \otimes \cdots \otimes p_{\mathbb{V}_d}(\mathbf{x}_d^i).$$

It is easy to check that the projection is invariant with respect to rank-one decompositions. In fact, if we let $\mathbf{P}_k \in \mathbb{R}^{n_k \times n_k}$ be the projection matrix of $p_{\mathbb{V}_k}$ for $k \in [d]$, then it is not hard to show that

$$\left(\bigotimes_{k=1}^d p_{\mathbb{V}_k} \right) (\mathbf{T}) = \mathbf{T} \times_1 \mathbf{P}_1 \times_2 \mathbf{P}_2 \cdots \times_d \mathbf{P}_d.$$

In another word, $\mathbf{T} \times_k \mathbf{P}_k$ projects all mode- k fibers of $\mathbf{T}$ onto the subspace $\mathbb{V}_k$. Therefore, we can take $\bigotimes_{k=1}^d p_{\mathbb{V}_k}$ as performing vector projections d times, in any order of the modes $1, 2, \dots, d$, as mode products allow swapping.

We state a property of the projection $\bigotimes_{k=1}^d p_{\mathbb{V}_k}$ below without proof.

Lemma 3.1 *If $\mathbb{V}_k$ is a subspace of $\mathbb{R}^{n_k}$ for $k \in [d]$, then the projection $\bigotimes_{k=1}^d p_{\mathbb{V}_k}$ is self-adjoint. Moreover,*

$$\mathbb{T}((\mathbb{V}_k)_{k=1}^d) = \{ \mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} : \text{sp}_k(\mathbf{T}) \subseteq \mathbb{V}_k \forall k \in [d] \}.$$

Finally, the norm of an operator p , in particular a projection or arithmetic operations of some projections in this paper, is defined as

$$\|p\| := \max\{\|p(\mathbf{T})\|_2 : \|\mathbf{T}\|_2 = 1, \mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}\}.$$

This includes the case of vector spaces when $d = 1$.

3.3 Tensor spectral and nuclear norms over subspaces

The definitions of the tensor spectral norm and nuclear norm in Section 2.3 involve working with all unit vectors. Instead, the following results state that, to evaluate the two norms of a tensor in $\mathbb{T}((\mathbb{V}_k)_{k=1}^d)$, it suffices to work with vectors in the subspaces $\mathbb{V}_k$'s.

Lemma 3.2 *If $\mathbb{V}_k$ is a subspace of $\mathbb{R}^{n_k}$ for $k \in [d]$ and $\mathbf{T} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$, then*

$$\|\mathbf{T}\|_\sigma = \max \left\{ \left\langle \mathbf{T}, \bigotimes_{k=1}^d \mathbf{x}_k \right\rangle : \mathbf{x}_k \in \mathbb{V}_k \cap \mathbb{S}^{n_k} \forall k \in [d] \right\}.$$

Hence, $\|p_{\mathbb{T}((\mathbb{V}_k)_{k=1}^d)}(\mathbf{T})\|_\sigma \leq \|\mathbf{T}\|_\sigma$ for any $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$.

Proof. Let $\|\mathbf{T}\|_\sigma = \langle \mathbf{T}, \bigotimes_{k=1}^d \mathbf{y}_k \rangle$, where $\mathbf{y}_k \in \mathbb{S}^{n_k}$ for $k \in [d]$. Since $\mathbf{T} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$, we have $(\bigotimes_{k=1}^d p_{\mathbb{V}_k})(\mathbf{T}) = \mathbf{T}$. As $\bigotimes_{k=1}^d p_{\mathbb{V}_k}$ is self-adjoint, we further have

$$\|\mathbf{T}\|_\sigma = \left\langle \left(\bigotimes_{k=1}^d p_{\mathbb{V}_k} \right) (\mathbf{T}), \bigotimes_{k=1}^d \mathbf{y}_k \right\rangle = \left\langle \mathbf{T}, \left(\bigotimes_{k=1}^d p_{\mathbb{V}_k} \right) \left(\bigotimes_{k=1}^d \mathbf{y}_k \right) \right\rangle = \left\langle \mathbf{T}, \bigotimes_{k=1}^d p_{\mathbb{V}_k}(\mathbf{y}_k) \right\rangle.$$

If there is some $p_{\mathbb{V}_k}(\mathbf{y}_k) = \mathbf{0}$, then $\|\mathbf{T}\|_\sigma = 0$. This makes $\mathbf{T} = \mathbf{O}$ and the equality trivially holds. Otherwise, we have $\|p_{\mathbb{V}_k}(\mathbf{y}_k)\|_2 \neq 0$ for every $k \in [d]$. By noticing that $\|p_{\mathbb{V}_k}(\mathbf{y}_k)\|_2 \leq \|\mathbf{y}_k\|_2 = 1$, we have

$$\begin{aligned} \|\mathbf{T}\|_\sigma &= \left\langle \mathbf{T}, \bigotimes_{k=1}^d p_{\mathbb{V}_k}(\mathbf{y}_k) \right\rangle \\ &= \left\langle \mathbf{T}, \bigotimes_{k=1}^d \frac{p_{\mathbb{V}_k}(\mathbf{y}_k)}{\|p_{\mathbb{V}_k}(\mathbf{y}_k)\|_2} \right\rangle \prod_{k=1}^d \|p_{\mathbb{V}_k}(\mathbf{y}_k)\|_2 \\ &\leq \max \left\{ \left\langle \mathbf{T}, \bigotimes_{k=1}^d \mathbf{x}_k \right\rangle : \mathbf{x}_k \in \mathbb{V}_k \cap \mathbb{S}^{n_k} \forall k \in [d] \right\} \\ &\leq \max \left\{ \left\langle \mathbf{T}, \bigotimes_{k=1}^d \mathbf{x}_k \right\rangle : \mathbf{x}_k \in \mathbb{S}^{n_k} \forall k \in [d] \right\} \\ &= \|\mathbf{T}\|_\sigma, \end{aligned}$$

implying the validity of the equality.

Finally, for any $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, as $p_{\mathbb{T}((\mathbb{V}_k)_{k=1}^d)}(\mathbf{T}) \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$, we have

$$\|p_{\mathbb{T}((\mathbb{V}_k)_{k=1}^d)}(\mathbf{T})\|_\sigma = \max \left\{ \left\langle \left(\bigotimes_{k=1}^d p_{\mathbb{V}_k} \right) (\mathbf{T}), \bigotimes_{k=1}^d \mathbf{x}_k \right\rangle : \mathbf{x}_k \in \mathbb{V}_k \cap \mathbb{S}^{n_k} \forall k \in [d] \right\}$$

$$\begin{aligned}
&= \max \left\{ \left\langle \mathbf{T}, \bigotimes_{k=1}^d \mathbf{p}_{\mathbb{V}_k}(\mathbf{x}_k) \right\rangle : \mathbf{x}_k \in \mathbb{V}_k \cap \mathbb{S}^{n_k} \forall k \in [d] \right\} \\
&= \max \left\{ \left\langle \mathbf{T}, \bigotimes_{k=1}^d \mathbf{x}_k \right\rangle : \mathbf{x}_k \in \mathbb{V}_k \cap \mathbb{S}^{n_k} \forall k \in [d] \right\} \\
&\leq \max \left\{ \left\langle \mathbf{T}, \bigotimes_{k=1}^d \mathbf{x}_k \right\rangle : \mathbf{x}_k \in \mathbb{S}^{n_k} \forall k \in [d] \right\} \\
&= \|\mathbf{T}\|_\sigma,
\end{aligned}$$

completing the final piece. $\square$

The orthogonal transformation of tensors discussed in Section 3.2 makes the above statement more intuitive. If $\mathbf{p}_{\mathbb{V}_k}$ simply keeps the first several entries unchanged and zeroing the remainder for every $k \in [d]$, to maximize $\langle \mathbf{T}, \bigotimes_{k=1}^d \mathbf{x}_k \rangle$ for a tensor $\mathbf{T} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$, an optimal $\mathbf{x}_k$ must not have a nonzero entry outside $\mathbb{V}_k$ for every $k \in [d]$.

The following is the dual version of Lemma 3.2.

Lemma 3.3 *If $\mathbb{V}_k$ is a subspace of $\mathbb{R}^{n_k}$ for $k \in [d]$ and $\mathbf{T} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$, then*

$$\|\mathbf{T}\|_* = \min \left\{ \sum_{i=1}^r |\lambda_i| : \mathbf{T} = \sum_{i=1}^r \lambda_i \bigotimes_{k=1}^d \mathbf{x}_k^i, \mathbf{x}_k^i \in \mathbb{V}_k \cap \mathbb{S}^{n_k} \forall i \in [r] \text{ and } k \in [d], r \in \mathbb{N} \right\}.$$

Furthermore, there exists a dual certificate $\mathbf{Z} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$ with $\|\mathbf{Z}\|_\sigma = 1$ such that $\langle \mathbf{T}, \mathbf{Z} \rangle = \|\mathbf{T}\|_*$.

Proof. We assume that $\mathbf{T} \neq \mathbf{O}$ as otherwise the results hold trivially. This means that $\dim(\mathbb{V}_k) \geq 1$ for every $k \in [d]$ and further $\mathbb{H} = \bigotimes_{k=1}^d (\mathbb{V}_k \cap \mathbb{S}^{n_k})$ is nonempty. It is easy to see that $\text{conv}(\mathbb{H})$ is convex, compact, and centrally symmetric. Let $\mathbb{H}_k \subseteq \mathbb{V}_k \cap \mathbb{S}^{n_k}$ be any normalized basis of $\mathbb{V}_k$ for $k \in [d]$. We have

$$\mathbb{T}((\mathbb{V}_k)_{k=1}^d) = \text{sp} \left(\bigotimes_{k=1}^d \mathbb{V}_k \right) = \text{sp} \left(\bigotimes_{k=1}^d \mathbb{H}_k \right) \subseteq \text{sp}(\mathbb{H}) \subseteq \text{sp}(\text{conv}(\mathbb{H})) \subseteq \text{sp} \left(\bigotimes_{k=1}^d \mathbb{V}_k \right),$$

implying that $\text{sp}(\text{conv}(\mathbb{H})) = \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$. As a result, $\text{conv}(\mathbb{H})$ must be the unit ball of some norm $\|\bullet\|_{\mathbb{H}} : \mathbb{T}((\mathbb{V}_k)_{k=1}^d) \rightarrow \mathbb{R}_+$.

We first claim that $\|\bigotimes_{k=1}^d \mathbf{x}_k\|_{\mathbb{H}} = 1$ if $\mathbf{x}_k \in \mathbb{V}_k \cap \mathbb{S}^{n_k}$ for any $k \in [d]$. As $\bigotimes_{k=1}^d \mathbf{x}_k \in \mathbb{H} \subseteq \text{conv}(\mathbb{H})$, it is obvious that $\|\bigotimes_{k=1}^d \mathbf{x}_k\|_{\mathbb{H}} \leq 1$. It then suffices to show that $\bigotimes_{k=1}^d \mathbf{x}_k$ is an extreme point of $\text{conv}(\mathbb{H})$. If this is not the case, then we can rewrite $\bigotimes_{k=1}^d \mathbf{x}_k = \lambda \mathbf{Y} + (1 - \lambda) \mathbf{Z}$ for some $\lambda \in (0, 1)$ and $\mathbf{Y}, \mathbf{Z} \in \text{conv}(\mathbb{H})$ with $\mathbf{Y} \neq \mathbf{Z}$. By Carathéodory's theorem, we can further rewrite

$$\bigotimes_{k=1}^d \mathbf{x}_k = \lambda \sum_{i=1}^{r_1} \alpha_i \mathbf{Y}_i + (1 - \lambda) \sum_{j=1}^{r_2} \beta_j \mathbf{Z}_j,$$

where $\mathbf{Y}_i, \mathbf{Z}_j \in \mathbb{H}$ and $\alpha_i, \beta_j > 0$ with $\sum_{i=1}^{r_1} \alpha_i = \sum_{j=1}^{r_2} \beta_j = 1$ for every $i \in [r_1]$ and $j \in [r_2]$. There must be some $\mathbf{Y}_i$ or $\mathbf{Z}_j$ that is not equal to $\bigotimes_{k=1}^d \mathbf{x}_k$, as otherwise it leads to $\mathbf{Y} = \mathbf{Z} = \bigotimes_{k=1}^d \mathbf{x}_k$. Assume without loss of generality that $\mathbf{Y}_1 \neq \bigotimes_{k=1}^d \mathbf{x}_k$. We observe that $\bigotimes_{k=1}^d \mathbf{x}_k - \lambda \alpha_1 \mathbf{Y}_1$ is not a

nonnegative multiple of $\lambda\alpha_1\mathbf{Y}_1$, as otherwise it leads to $\mathbf{Y}_1 = \bigotimes_{k=1}^d \mathbf{x}_k$. By the equality condition of the triangle inequality, we have

$$1 = \left\| \bigotimes_{k=1}^d \mathbf{x}_k \right\|_2 < \|\lambda\alpha_1\mathbf{Y}_1\|_2 + \left\| \bigotimes_{k=1}^d \mathbf{x}_k - \lambda\alpha_1\mathbf{Y}_1 \right\|_2 \leq \lambda \sum_{i=1}^{r_1} \alpha_i \|\mathbf{Y}_i\|_2 + (1-\lambda) \sum_{j=1}^{r_2} \beta_j \|\mathbf{Z}_j\|_2 = 1,$$

leading to a contradiction.

By vector scaling, any $\mathbf{T} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$ can be written as $\sum_{i=1}^r \lambda_i \bigotimes_{k=1}^d \mathbf{x}_k^i$ where $\mathbf{x}_k^i \in \mathbb{V}_k \cap \mathbb{S}^{n_k}$ for any $i \in [r]$ and $k \in [d]$. By the triangle inequality, we have $\|\mathbf{T}\|_{\mathbb{H}} \leq \sum_{i=1}^r |\lambda_i| \cdot \left\| \bigotimes_{k=1}^d \mathbf{x}_k^i \right\|_{\mathbb{H}} = \sum_{i=1}^r |\lambda_i|$, and so

$$\|\mathbf{T}\|_{\mathbb{H}} \leq \min \left\{ \sum_{i=1}^r |\lambda_i| : \mathbf{T} = \sum_{i=1}^r \lambda_i \bigotimes_{k=1}^d \mathbf{x}_k^i, \mathbf{x}_k^i \in \mathbb{V}_k \cap \mathbb{S}^{n_k} \forall i \in [r] \text{ and } k \in [d], r \in \mathbb{N} \right\}. \quad (7)$$

We next claim that (7) holds as an equality. Upon scaling, we only need to show that the equality holds for any $\mathbf{T} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$ with $\|\mathbf{T}\|_{\mathbb{H}} = 1$. Because $\|\mathbf{T}\|_{\mathbb{H}} = 1$ implies $\mathbf{T} \in \text{conv}(\mathbb{H})$, it follows that $\mathbf{T} = \sum_{i=1}^s \mu_i \bigotimes_{k=1}^d \mathbf{y}_k^i$, where $\mathbf{y}_k^i \in \mathbb{V}_k \cap \mathbb{S}^{n_k}$ and $\mu_i \geq 0$ with $\sum_{i=1}^s \mu_i = 1$ for $i \in [s]$ and $k \in [d]$. This decomposition is feasible in (7). As a result,

$$\begin{aligned} \|\mathbf{T}\|_{\mathbb{H}} &\leq \min \left\{ \sum_{i=1}^r |\lambda_i| : \mathbf{T} = \sum_{i=1}^r \lambda_i \bigotimes_{k=1}^d \mathbf{x}_k^i, \mathbf{x}_k^i \in \mathbb{V}_k \cap \mathbb{S}^{n_k} \forall i \in [r] \text{ and } k \in [d], r \in \mathbb{N} \right\} \\ &\leq \sum_{i=1}^s |\mu_i| = 1 = \|\mathbf{T}\|_{\mathbb{H}}. \end{aligned}$$

The final step is to claim the equivalence between $\|\bullet\|_{\mathbb{H}}$ and $\|\bullet\|_*$ over $\mathbb{T}((\mathbb{V}_k)_{k=1}^d)$. To start with, let us denote the dual norm of $\|\bullet\|_{\mathbb{H}}$ to be $\|\bullet\|_{\mathbb{H}'} : \mathbb{T}((\mathbb{V}_k)_{k=1}^d) \rightarrow \mathbb{R}_+$, i.e., $\|\mathbf{T}\|_{\mathbb{H}'} = \max_{\|\mathbf{Z}\|_{\mathbb{H}} \leq 1} \langle \mathbf{T}, \mathbf{Z} \rangle$ for $\mathbf{T} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$. Given any $\mathbf{T} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$, let $\mathbf{X} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$ be such that $\|\mathbf{T}\|_{\mathbb{H}'} = \langle \mathbf{T}, \mathbf{X} \rangle$ with $\|\mathbf{X}\|_{\mathbb{H}} \leq 1$, implying that $\mathbf{X} \in \text{conv}(\mathbb{H})$. Thus, $\mathbf{X}$ can be written as $\sum_{i=1}^{r_0} \gamma_i \mathbf{X}_i$, where $\mathbf{X}_i \in \mathbb{H}$ and $\gamma_i \geq 0$ with $\sum_{i=1}^{r_0} \gamma_i = 1$ for any $i \in [r_0]$. This further implies that

$$\|\mathbf{T}\|_{\mathbb{H}'} = \langle \mathbf{T}, \mathbf{X} \rangle = \left\langle \mathbf{T}, \sum_{i=1}^{r_0} \gamma_i \mathbf{X}_i \right\rangle \leq \left(\sum_{i=1}^{r_0} \gamma_i \right) \max_{\mathbf{Z} \in \mathbb{H}} \langle \mathbf{T}, \mathbf{Z} \rangle = \max_{\mathbf{Z} \in \mathbb{H}} \langle \mathbf{T}, \mathbf{Z} \rangle \leq \max_{\|\mathbf{Z}\|_{\mathbb{H}} \leq 1} \langle \mathbf{T}, \mathbf{Z} \rangle = \|\mathbf{T}\|_{\mathbb{H}'}.$$

As a result, for any $\mathbf{T} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$, we have

$$\|\mathbf{T}\|_{\mathbb{H}'} = \max_{\mathbf{Z} \in \mathbb{H}} \langle \mathbf{T}, \mathbf{Z} \rangle = \max_{\mathbf{x}_k \in \mathbb{V}_k \cap \mathbb{S}^{n_k} \forall k \in [d]} \left\langle \mathbf{T}, \bigotimes_{k=1}^d \mathbf{x}_k \right\rangle = \|\mathbf{T}\|_{\sigma},$$

where the last equality is due to Lemma 3.2. Now, by comparing (4) with that (7) holds as an equality, we have

$$\|\mathbf{T}\|_* \leq \|\mathbf{T}\|_{\mathbb{H}} = \max_{\|\mathbf{Z}\|_{\mathbb{H}'} \leq 1} \langle \mathbf{T}, \mathbf{Z} \rangle = \max_{\|\mathbf{Z}\|_{\sigma} \leq 1, \mathbf{Z} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)} \langle \mathbf{T}, \mathbf{Z} \rangle \leq \max_{\|\mathbf{Z}\|_{\sigma} \leq 1} \langle \mathbf{T}, \mathbf{Z} \rangle = \|\mathbf{T}\|_*.$$

Finally, given any $\mathbf{T} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$, let $\mathbf{S}$ be a dual certificate of $\|\mathbf{T}\|_*$, i.e., $\langle \mathbf{T}, \mathbf{S} \rangle = \|\mathbf{T}\|_*$ and $\|\mathbf{S}\|_{\sigma} = 1$. It follows from Lemma 3.2 that $\|\mathbf{p}_{\mathbb{T}((\mathbb{V}_k)_{k=1}^d)}(\mathbf{S})\|_{\sigma} \leq \|\mathbf{S}\|_{\sigma} = 1$. On the other hand, by Lemma 2.1,

$$\|\mathbf{p}_{\mathbb{T}((\mathbb{V}_k)_{k=1}^d)}(\mathbf{S})\|_{\sigma} \geq \frac{\langle \mathbf{T}, \mathbf{p}_{\mathbb{T}((\mathbb{V}_k)_{k=1}^d)}(\mathbf{S}) \rangle}{\|\mathbf{T}\|_*} = \frac{\langle \mathbf{p}_{\mathbb{T}((\mathbb{V}_k)_{k=1}^d)}(\mathbf{T}), \mathbf{S} \rangle}{\|\mathbf{T}\|_*} = \frac{\langle \mathbf{T}, \mathbf{S} \rangle}{\|\mathbf{T}\|_*} = 1.$$

Therefore, $\mathbf{p}_{\mathbb{T}((\mathbb{V}_k)_{k=1}^d)}(\mathbf{S}) \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$ is also a dual certificate of $\|\mathbf{T}\|_*$. $\square$

4 Decomposability of the tensor nuclear norm

The decomposability of the matrix nuclear norm has been well understood. It has given rise to many important results in machine learning and statistical estimation [57, Section 10]. Specifically, for two matrices $\mathbf{T}, \mathbf{S} \in \mathbb{R}^{n_1 \times n_2}$,

$$\|\mathbf{T} + \mathbf{S}\|_* = \|\mathbf{T}\|_* + \|\mathbf{S}\|_* \text{ if } \mathbf{T}^T \mathbf{S} = \mathbf{O} \text{ and } \mathbf{T} \mathbf{S}^T = \mathbf{O}.$$

In the notation of subspaces, it reads as

$$\|\mathbf{T} + \mathbf{S}\|_* = \|\mathbf{T}\|_* + \|\mathbf{S}\|_* \text{ for any } \mathbf{T} \in \mathbb{T}(\mathbb{V}_1, \mathbb{V}_2) \text{ and } \mathbf{S} \in \mathbb{T}(\mathbb{V}_1^\perp, \mathbb{V}_2^\perp),$$

where $\mathbb{V}_k$ is a subspace of $\mathbb{R}^{n_k}$ for $k \in [2]$. The main reason behind the above decomposability is that any matrix admits an SVD. After proper orthogonal transformations of the column space and the row space, a matrix becomes a diagonal one whose entries are its singular values. The nuclear norm, as the sum of singular values, automatically admits this type of decomposition.

The diagonalization fails to work for higher-order tensors in general. As a result, generalizing the decomposability to tensors remains unclear and unsatisfactory. To the best of our knowledge, the only known result is a weak decomposability [50, Lemma 2.1] that only works for third-order tensors. Specifically, the weak decomposability states that

$$\|\mathbf{T} + \mathbf{S}\|_* \geq \|\mathbf{T}\|_* + \frac{1}{2}\|\mathbf{S}\|_* \text{ for any } \mathbf{T} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^3) \text{ and } \mathbf{S} \in \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^\mathbb{I}((\mathbb{V}_k)_{k=1}^3), \quad (8)$$

where $\mathbb{V}_k$ is a subspace of $\mathbb{R}^{n_k}$ for $k \in [3]$.

We remark that this weak decomposability has found successful applications in high-dimensional tensor regression [50]. However, the full decomposability, $\|\mathbf{T} + \mathbf{S}\|_* = \|\mathbf{T}\|_* + \|\mathbf{S}\|_*$, turns out to be impossible over the subspace pair $\mathbb{T}((\mathbb{V}_k)_{k=1}^3)$ and $\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^\mathbb{I}((\mathbb{V}_k)_{k=1}^3)$ considered in (8). This is evidenced by the following example.

Example 4.1 Let $\mathbf{T} = \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 \in \mathbb{R}^{2 \times 2 \times 2}$ and $\mathbf{S} = \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_2 \otimes \mathbf{e}_2 \in \mathbb{R}^{2 \times 2 \times 2}$. It is obvious that $\mathbf{T} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^3)$ and $\mathbf{S} \in \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^\mathbb{I}((\mathbb{V}_k)_{k=1}^3)$, where $\mathbb{V}_k = \text{sp}(\mathbf{e}_1) \subseteq \mathbb{R}^2$ for $k \in [3]$. However,

$$3.078 \approx \|\mathbf{T} + \mathbf{S}\|_* < \|\mathbf{T}\|_* + \|\mathbf{S}\|_* \approx 1 + 3.162.$$

The above nuclear norms are computed by a fully polynomial-time approximation scheme designed in [28]. In fact, we even have $\|\mathbf{T} + \mathbf{S}\|_* < \|\mathbf{S}\|_*$ in Example 4.1, albeit $\mathbf{T}$ and $\mathbf{S}$ sit in two mutually orthogonal subspaces. This is a phenomenon that can never happen in the matrix case.

4.1 A natural decomposability

The weak but not full decomposability obviously raises a question on the subspace candidate that is orthogonal to $\mathbb{T}((\mathbb{V}_k)_{k=1}^3)$, i.e., why $\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^\mathbb{I}((\mathbb{V}_k)_{k=1}^3)$ has been chosen in (8). In the matrix space $\mathbb{R}^{n_1 \times n_2}$, the diagonalization provides a very clear picture. Among the four basic subspaces induced by $(\mathbb{V}_1, \mathbb{V}_2)$, i.e., $\mathbb{T}(\mathbb{V}_1, \mathbb{V}_2)$, $\mathbb{T}^{\{1\}}(\mathbb{V}_1, \mathbb{V}_2)$, $\mathbb{T}^{\{2\}}(\mathbb{V}_1, \mathbb{V}_2)$, and $\mathbb{T}^{\{1,2\}}(\mathbb{V}_1, \mathbb{V}_2)$ in Figure 1, the only candidate is $\mathbb{T}^{\{1,2\}}(\mathbb{V}_1, \mathbb{V}_2)$.

In the tensor space $\mathbb{R}^{n_1 \times n_2 \times n_3}$, however,

$$\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^\mathbb{I}((\mathbb{V}_k)_{k=1}^3) = \mathbb{T}(\mathbb{V}_1^\perp, \mathbb{V}_2^\perp, \mathbb{V}_3) \oplus \mathbb{T}(\mathbb{V}_1^\perp, \mathbb{V}_2, \mathbb{V}_3^\perp) \oplus \mathbb{T}(\mathbb{V}_1, \mathbb{V}_2^\perp, \mathbb{V}_3^\perp) \oplus \mathbb{T}(\mathbb{V}_1^\perp, \mathbb{V}_2^\perp, \mathbb{V}_3^\perp)$$

$\mathbb{V}_1^\perp$	$\mathbb{T}^{\{1\}}(\mathbb{V}_1, \mathbb{V}_2)$	$\mathbb{T}^{\{1,2\}}(\mathbb{V}_1, \mathbb{V}_2)$
$\mathbb{V}_1$	$\mathbb{T}(\mathbb{V}_1, \mathbb{V}_2)$	$\mathbb{T}^{\{2\}}(\mathbb{V}_1, \mathbb{V}_2)$
	$\mathbb{V}_2$	$\mathbb{V}_2^\perp$

Figure 1: The four basic subspaces of $\mathbb{R}^{n_1 \times n_2}$.

includes four out of the eight basic subspaces defined by $(\mathbb{V}_1, \mathbb{V}_2, \mathbb{V}_3)$, each of which spanned by at least two $\mathbb{V}_k^\perp$'s. While this candidate has only resulted a weak decomposability, it has its own theoretical merits that will be discussed later in Section 5. By making a compromise on the subspace size, it turns out that the full decomposability is indeed possible for some other subspaces. We start with the most restrictive candidate, $\mathbb{T}(\mathbb{V}_1^\perp, \mathbb{V}_2^\perp, \mathbb{V}_3^\perp)$, a natural generalization from the matrix case $\mathbb{T}(\mathbb{V}_1^\perp, \mathbb{V}_2^\perp)$. For comparison, an example of the subspaces of $\mathbb{R}^{n_1 \times n_2 \times n_3}$ mentioned earlier is offered in Figure 2.

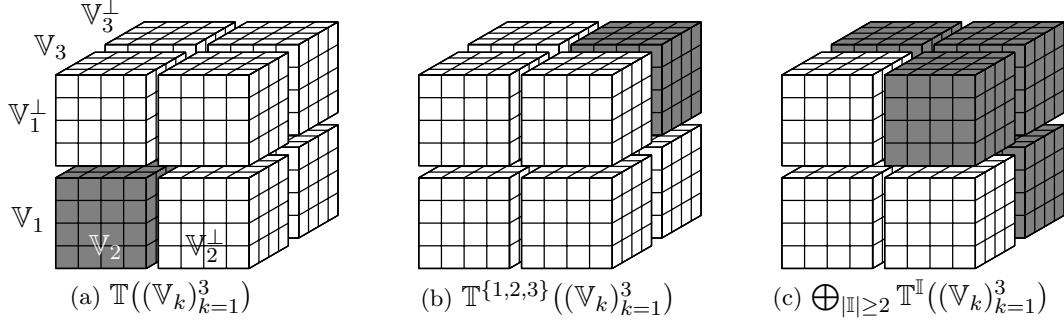


Figure 2: Subspaces of $\mathbb{R}^{n_1 \times n_2 \times n_3}$ presented by shaded blocks where each block represents a basic subspace of $\mathbb{R}^{n_1 \times n_2 \times n_3}$ and the union of all eight blocks represents $\mathbb{R}^{n_1 \times n_2 \times n_3}$ in each subfigure.

The main result in this subsection is as follows.

Theorem 4.2 *If $d \geq 2$ and $\mathbb{V}_k$ is a subspace of $\mathbb{R}^{n_k}$ for $k \in [d]$, then the nuclear norm is decomposable over $\mathbb{T}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbb{T}^{[d]}((\mathbb{V}_k)_{k=1}^d)$ of the tensor space $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, i.e.,*

$$\|\mathbf{T} + \mathbf{S}\|_* = \|\mathbf{T}\|_* + \|\mathbf{S}\|_* \text{ for any } \mathbf{T} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d) \text{ and } \mathbf{S} \in \mathbb{T}^{[d]}((\mathbb{V}_k)_{k=1}^d).$$

Proof. Let $\mathbf{T} = \sum_{i=1}^{r_1} \lambda_i \bigotimes_{k=1}^d \mathbf{x}_k^i$ be a nuclear decomposition, where $\mathbf{x}_k^i \in \mathbb{V}_k \cap \mathbb{S}^{n_k}$ for $i \in [r_1]$ and $k \in [d]$. The existence of such decomposition is guaranteed by Lemma 3.3. We may further assume that $\lambda_i > 0$ for any $i \in [r_1]$ simply by flipping signs and removing zeros. Similarly, let $\mathbf{S} = \sum_{i=1}^{r_2} \mu_i \bigotimes_{k=1}^d \mathbf{y}_k^i$ be a nuclear decomposition, where $\mu_i > 0$ for all $i \in [r_2]$ and $\mathbf{y}_k^i \in \mathbb{V}_k^\perp \cap \mathbb{S}^{n_k}$ for $i \in [r_2]$ and $k \in [d]$. It then follows by Lemma 3.3 again that there exist dual certificates $\mathbf{X} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbf{Y} \in \mathbb{T}((\mathbb{V}_k^\perp)_{k=1}^d)$ of $\|\mathbf{T}\|_*$ and $\|\mathbf{S}\|_*$, respectively.

With these preparations, we now show that

$$\mathbf{T} + \mathbf{S} = \sum_{i=1}^{r_1} \lambda_i \bigotimes_{k=1}^d \mathbf{x}_k^i + \sum_{i=1}^{r_2} \mu_i \bigotimes_{k=1}^d \mathbf{y}_k^i \quad (9)$$

is actually a nuclear decomposition and $\mathbf{X} + \mathbf{Y}$ is a dual certificate of $\|\mathbf{T} + \mathbf{S}\|_*$. It is clear that

$$\left\langle \mathbf{X} + \mathbf{Y}, \bigotimes_{k=1}^d \mathbf{x}_k^i \right\rangle = 1 + \left\langle \left(\bigotimes_{k=1}^d \mathbf{p}_{\mathbb{V}_k^\perp} \right) (\mathbf{Y}), \bigotimes_{k=1}^d \mathbf{x}_k^i \right\rangle = 1 + \left\langle \mathbf{Y}, \bigotimes_{k=1}^d \mathbf{p}_{\mathbb{V}_k^\perp} (\mathbf{x}_k^i) \right\rangle = 1 \quad (10)$$

for all $i \in [r_1]$. For the same reason, we also have $\langle \mathbf{X} + \mathbf{Y}, \bigotimes_{k=1}^d \mathbf{y}_k^i \rangle = 1$ for any $i \in [r_2]$. Hence,

$$\langle \mathbf{T} + \mathbf{S}, \mathbf{X} + \mathbf{Y} \rangle = \left\langle \sum_{i=1}^{r_1} \lambda_i \bigotimes_{k=1}^d \mathbf{x}_k^i + \sum_{i=1}^{r_2} \mu_i \bigotimes_{k=1}^d \mathbf{y}_k^i, \mathbf{X} + \mathbf{Y} \right\rangle = \sum_{i=1}^{r_1} \lambda_i + \sum_{i=1}^{r_2} \mu_i.$$

Since $\mathbf{X} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbf{Y} \in \mathbb{T}((\mathbb{V}_k^\perp)_{k=1}^d)$, we actually have $\|\mathbf{X} + \mathbf{Y}\|_\sigma = \max\{\|\mathbf{X}\|_\sigma, \|\mathbf{Y}\|_\sigma\} = 1$, due to a result dual to this decomposability; see Theorem 4.3 to be presented soon. This means that $\|\mathbf{T} + \mathbf{S}\|_* \geq \langle \mathbf{T} + \mathbf{S}, \mathbf{X} + \mathbf{Y} \rangle = \sum_{i=1}^{r_1} \lambda_i + \sum_{i=1}^{r_2} \mu_i$ by Lemma 2.1. On the other hand, by combining (9) with (4), the definition of the nuclear norm, we have $\|\mathbf{T} + \mathbf{S}\|_* \leq \sum_{i=1}^{r_1} \lambda_i + \sum_{i=1}^{r_2} \mu_i$. Therefore, $\|\mathbf{T} + \mathbf{S}\|_* = \sum_{i=1}^{r_1} \lambda_i + \sum_{i=1}^{r_2} \mu_i = \|\mathbf{T}\|_* + \|\mathbf{S}\|_*$. $\square$

The dual version of Theorem 4.2, i.e., the decomposability of the tensor spectral norm, is of independent interest. In particular, it has supported the above proof.

Theorem 4.3 *If $d \geq 2$ and $\mathbb{V}_k$ is a subspace of $\mathbb{R}^{n_k}$ for $k \in [d]$, then the spectral norm is decomposable over $\mathbb{T}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbb{T}^{[d]}((\mathbb{V}_k)_{k=1}^d)$ of the tensor space $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, i.e.,*

$$\|\mathbf{T} + \mathbf{S}\|_\sigma = \max\{\|\mathbf{T}\|_\sigma, \|\mathbf{S}\|_\sigma\} \text{ for any } \mathbf{T} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d) \text{ and } \mathbf{S} \in \mathbb{T}^{[d]}((\mathbb{V}_k)_{k=1}^d).$$

Proof. Since $(\bigotimes_{k=1}^d \mathbf{p}_{\mathbb{V}_k})(\mathbf{T}) = \mathbf{T}$ and $(\bigotimes_{k=1}^d \mathbf{p}_{\mathbb{V}_k^\perp})(\mathbf{S}) = \mathbf{S}$, we have

$$\begin{aligned} \|\mathbf{T} + \mathbf{S}\|_\sigma &= \max_{\mathbf{v}_k \in \mathbb{S}^{n_k} \forall k \in [d]} \left\langle \mathbf{T} + \mathbf{S}, \bigotimes_{k=1}^d \mathbf{v}_k \right\rangle \\ &= \max_{\mathbf{v}_k \in \mathbb{S}^{n_k} \forall k \in [d]} \left(\left\langle \left(\bigotimes_{k=1}^d \mathbf{p}_{\mathbb{V}_k} \right) (\mathbf{T}), \bigotimes_{k=1}^d \mathbf{v}_k \right\rangle + \left\langle \left(\bigotimes_{k=1}^d \mathbf{p}_{\mathbb{V}_k^\perp} \right) (\mathbf{S}), \bigotimes_{k=1}^d \mathbf{v}_k \right\rangle \right) \\ &= \max_{\mathbf{v}_k \in \mathbb{S}^{n_k} \forall k \in [d]} \left(\left\langle \mathbf{T}, \bigotimes_{k=1}^d \mathbf{p}_{\mathbb{V}_k} (\mathbf{v}_k) \right\rangle + \left\langle \mathbf{S}, \bigotimes_{k=1}^d \mathbf{p}_{\mathbb{V}_k^\perp} (\mathbf{v}_k) \right\rangle \right) \\ &\leq \max_{\mathbf{v}_k \in \mathbb{S}^{n_k} \forall k \in [d]} \left(\|\mathbf{T}\|_\sigma \prod_{k=1}^d \|\mathbf{p}_{\mathbb{V}_k} (\mathbf{v}_k)\|_2 + \|\mathbf{S}\|_\sigma \prod_{k=1}^d \|\mathbf{p}_{\mathbb{V}_k^\perp} (\mathbf{v}_k)\|_2 \right) \\ &\leq \max\{\|\mathbf{T}\|_\sigma, \|\mathbf{S}\|_\sigma\} \max_{\mathbf{v}_k \in \mathbb{S}^{n_k} \forall k \in [d]} \left(\prod_{k=1}^d \|\mathbf{p}_{\mathbb{V}_k} (\mathbf{v}_k)\|_2 + \prod_{k=1}^d \|\mathbf{p}_{\mathbb{V}_k^\perp} (\mathbf{v}_k)\|_2 \right) \\ &\leq \max\{\|\mathbf{T}\|_\sigma, \|\mathbf{S}\|_\sigma\} \max_{\mathbf{v}_k \in \mathbb{S}^{n_k} \forall k \in [2]} \left(\prod_{k=1}^2 \|\mathbf{p}_{\mathbb{V}_k} (\mathbf{v}_k)\|_2 + \prod_{k=1}^2 \|\mathbf{p}_{\mathbb{V}_k^\perp} (\mathbf{v}_k)\|_2 \right) \\ &\leq \max\{\|\mathbf{T}\|_\sigma, \|\mathbf{S}\|_\sigma\} \max_{\mathbf{v}_k \in \mathbb{S}^{n_k} \forall k \in [2]} \prod_{k=1}^2 \left\| \begin{pmatrix} \|\mathbf{p}_{\mathbb{V}_k} (\mathbf{v}_k)\|_2 \\ \|\mathbf{p}_{\mathbb{V}_k^\perp} (\mathbf{v}_k)\|_2 \end{pmatrix} \right\|_2 \\ &= \max\{\|\mathbf{T}\|_\sigma, \|\mathbf{S}\|_\sigma\}. \end{aligned}$$

On the other hand, let $\|\mathbf{T}\|_\sigma = \langle \mathbf{T}, \bigotimes_{k=1}^d \mathbf{x}_k \rangle$ with $\mathbf{x}_k \in \mathbb{V}_k \cap \mathbb{S}^{n_k}$ for $k \in [d]$ by Lemma 3.2. It is obvious that

$$\left\langle \mathbf{S}, \bigotimes_{k=1}^d \mathbf{x}_k \right\rangle = \left\langle \left(\bigotimes_{k=1}^d \mathbf{P}_{\mathbb{V}_k^\perp} \right) (\mathbf{S}), \bigotimes_{k=1}^d \mathbf{x}_k \right\rangle = \left\langle \mathbf{S}, \bigotimes_{k=1}^d \mathbf{P}_{\mathbb{V}_k^\perp} (\mathbf{x}_k) \right\rangle = 0,$$

where the last equality is due to $\mathbf{x}_k \in \mathbb{V}_k$ for every k . This means that $\|\mathbf{T} + \mathbf{S}\|_\sigma \geq \langle \mathbf{T} + \mathbf{S}, \bigotimes_{k=1}^d \mathbf{x}_k \rangle = \|\mathbf{T}\|_\sigma$. For the same reason, we also have $\|\mathbf{T} + \mathbf{S}\|_\sigma \geq \|\mathbf{S}\|_\sigma$. Therefore, $\|\mathbf{T} + \mathbf{S}\|_\sigma \geq \max\{\|\mathbf{T}\|_\sigma, \|\mathbf{S}\|_\sigma\}$, completing the whole proof. $\square$

We remark that the matrix case of Theorem 4.3 has not been seen before, to the best of our knowledge. Both Theorem 4.2 and Theorem 4.3 include and generalize the matrix case on the decomposability of the nuclear norm and spectral norm. It is also worth noting that the decomposability in both Theorem 4.2 and Theorem 4.3 is full, albeit for third-order tensors the subspace $\mathbb{T}^{[3]}((\mathbb{V}_k)_{k=1}^3)$ is smaller than $\bigoplus_{|\mathbb{I}| \geq 2, |\mathbb{I}| \leq [3]} \mathbb{T}^{|\mathbb{I}|}((\mathbb{V}_k)_{k=1}^3)$ in (8) where only a weak decomposability is possible. That being said, the subspace $\mathbb{T}^{[d]}((\mathbb{V}_k)_{k=1}^d)$ is still practically useful in, e.g., analyzing the tensor robust PCA in Section 6.

4.2 An improved decomposability

It is natural to ask whether the decomposability in Theorem 4.2 can be further improved, in the sense that the two subspaces $\mathbb{T}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbb{T}^{[d]}((\mathbb{V}_k)_{k=1}^d)$ can be enlarged without destroying the full decomposability of the nuclear norm. The answer is indeed affirmative. Let us first recall the notation of $\mathbb{U}$ -subspaces,

$$\begin{aligned} \mathbb{U}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d) &= \bigoplus_{\mathbb{I} \subseteq \mathbb{J} \subseteq [d]} \mathbb{T}^{\mathbb{J}}((\mathbb{V}_k)_{k=1}^d) = \text{sp} \left(\left\{ \bigotimes_{k=1}^d \mathbf{v}_k : \mathbf{v}_k \in \mathbb{V}_k^\perp \forall k \in \mathbb{I}, \mathbf{v}_k \in \mathbb{R}^{n_k} \forall k \notin \mathbb{I} \right\} \right), \\ \mathbb{U}_{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d) &= \mathbb{U}^{\mathbb{I}}((\mathbb{V}_k^\perp)_{k=1}^d) = \text{sp} \left(\left\{ \bigotimes_{k=1}^d \mathbf{v}_k : \mathbf{v}_k \in \mathbb{V}_k \forall k \in \mathbb{I}, \mathbf{v}_k \in \mathbb{R}^{n_k} \forall k \notin \mathbb{I} \right\} \right). \end{aligned}$$

An example of relevant subspaces of $\mathbb{R}^{n_1 \times n_2 \times n_3}$ for $\mathbb{I} = \{1, 2\}$ is presented in Figure 3.

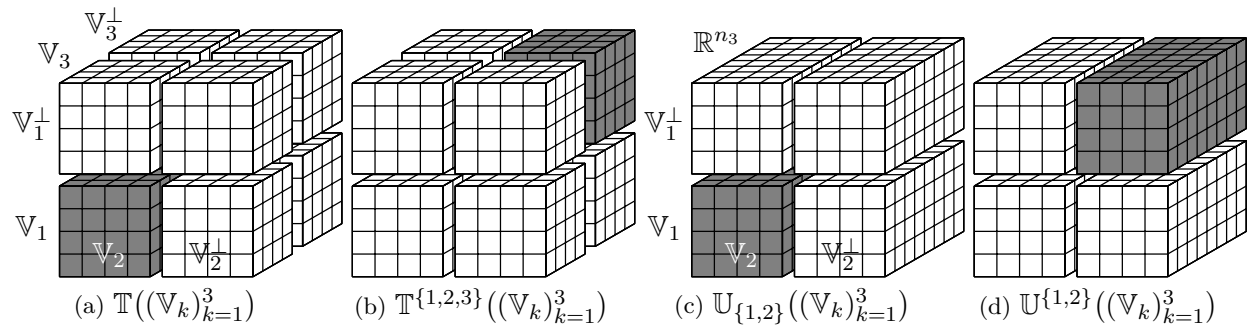


Figure 3: Subspaces of $\mathbb{R}^{n_1 \times n_2 \times n_3}$ presented by shaded blocks where the union of all blocks represents $\mathbb{R}^{n_1 \times n_2 \times n_3}$ in each subfigure.

For third-order tensors, the pair of subspaces shown in (a) and (b) of Figure 3 and required in the natural decomposability in Theorem 4.2 can now be enlarged to a pair of subspaces shown in

(c) and (d) of Figure 3, respectively. More generally, for higher-order tensors, the following result shows the full decomposability of the tensor nuclear norm over an improved subspace pair from Theorem 4.2.

Theorem 4.4 *If $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| \geq 2$ and $\mathbb{V}_k$ is a subspace of $\mathbb{R}^{n_k}$ for $k \in [d]$, then the nuclear norm is decomposable over $\mathbb{U}_{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbb{U}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$ of the tensor space $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, i.e.,*

$$\|\mathbf{T} + \mathbf{S}\|_* = \|\mathbf{T}\|_* + \|\mathbf{S}\|_* \text{ for any } \mathbf{T} \in \mathbb{U}_{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d) \text{ and } \mathbf{S} \in \mathbb{U}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d).$$

The proof of this result will be discussed slightly later.

It is important to remark and in fact easy to observe the monotonicity. The decomposability over $\mathbb{U}_{\mathbb{I}_1}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbb{U}^{\mathbb{I}_1}((\mathbb{V}_k)_{k=1}^d)$ implies the decomposability over $\mathbb{U}_{\mathbb{I}_2}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbb{U}^{\mathbb{I}_2}((\mathbb{V}_k)_{k=1}^d)$ if $\mathbb{I}_1 \subseteq \mathbb{I}_2$, under which $\mathbb{U}_{\mathbb{I}_2}((\mathbb{V}_k)_{k=1}^d) \subseteq \mathbb{U}_{\mathbb{I}_1}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbb{U}^{\mathbb{I}_2}((\mathbb{V}_k)_{k=1}^d) \subseteq \mathbb{U}^{\mathbb{I}_1}((\mathbb{V}_k)_{k=1}^d)$. Therefore, Theorem 4.4 is essentially for the strongest case $|\mathbb{I}| = 2$ whereas the weakest case is for $\mathbb{I} = [d]$, the result in Theorem 4.2. In a short word, the tensor nuclear norm is decomposable over a pair of subspaces if they have at least two disjoint modes. We say that a mode is disjoint for a pair of subspaces if the projections of the subspaces onto the vector space of that mode are orthogonal to each other.

The above point is particularly clear for the matrix case that is included in Theorem 4.4. Among the four basic subspaces $\mathbb{T}(\mathbb{V}_1, \mathbb{V}_2)$, $\mathbb{T}^{\{1\}}(\mathbb{V}_1, \mathbb{V}_2)$, $\mathbb{T}^{\{2\}}(\mathbb{V}_1, \mathbb{V}_2)$, and $\mathbb{T}^{\{1,2\}}(\mathbb{V}_1, \mathbb{V}_2)$, as well as any direct sum of some basic subspaces, only the pair $\mathbb{T}(\mathbb{V}_1, \mathbb{V}_2)$ and $\mathbb{T}^{\{1,2\}}(\mathbb{V}_1, \mathbb{V}_2)$ and the pair $\mathbb{T}^{\{1\}}(\mathbb{V}_1, \mathbb{V}_2)$ and $\mathbb{T}^{\{2\}}(\mathbb{V}_1, \mathbb{V}_2)$ have at least two disjoint modes, resulting their full decomposability. If a pair of subspaces share only one disjoint mode, such as $\mathbb{T}(\mathbb{V}_1, \mathbb{V}_2)$ and $\mathbb{T}^{\{2\}}(\mathbb{V}_1, \mathbb{V}_2)$, nothing can be said about the decomposability; see a simple example below. Therefore, the full decomposability of the tensor nuclear norm over $\mathbb{U}_{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbb{U}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$ is in fact in the maximal sense for any $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| = 2$.

Example 4.5 *Let $\mathbf{T} = \mathbf{e}_1 \otimes \mathbf{e}_1 \in \mathbb{R}^{2 \times 2}$ and $\mathbf{S} = \epsilon \mathbf{e}_1 \otimes \mathbf{e}_2 \in \mathbb{R}^{2 \times 2}$, where $\epsilon > 0$ is sufficiently small. It is obvious that $\mathbf{T} \in \mathbb{T}(\mathbb{V}_1, \mathbb{V}_2)$ and $\mathbf{S} \in \mathbb{T}(\mathbb{V}_1, \mathbb{V}_2^\perp)$, where $\mathbb{V}_1 = \mathbb{V}_2 = \text{sp}(\mathbf{e}_1)$. However, the inequality*

$$\sqrt{1 + \epsilon^2} = \|\mathbf{T} + \mathbf{S}\|_* \geq \|\mathbf{T}\|_* + \alpha \|\mathbf{S}\|_* = 1 + \alpha \epsilon = \sqrt{1 + 2\alpha\epsilon + \alpha^2\epsilon^2}$$

cannot hold for any $\alpha > 0$ that is independent of ϵ .

We would like to remark that Theorem 4.4 is also important in tensor analysis. As an immediate application, it helps to single out a large class of tensors whose upper bounds of the nuclear norm based on tensor partitions (see [35, Theorem 3.1] and [15, Theorem 3.1]) are tight. Specifically, it is stated in [35, Theorem 3.1] that if a tensor $\mathbf{T}$ is partitioned into any set of subtensors $\{\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_m\}$, then

$$\|(\|\mathbf{T}_1\|_\sigma, \|\mathbf{T}_2\|_\sigma, \dots, \|\mathbf{T}_m\|_\sigma)^T\|_\infty \leq \|\mathbf{T}\|_\sigma \leq \|(\|\mathbf{T}_1\|_\sigma, \|\mathbf{T}_2\|_\sigma, \dots, \|\mathbf{T}_m\|_\sigma)^T\|_2, \quad (11)$$

$$\|(\|\mathbf{T}_1\|_*, \|\mathbf{T}_2\|_*, \dots, \|\mathbf{T}_m\|_*)^T\|_2 \leq \|\mathbf{T}\|_* \leq \|(\|\mathbf{T}_1\|_*, \|\mathbf{T}_2\|_*, \dots, \|\mathbf{T}_m\|_*)^T\|_1. \quad (12)$$

Therefore, any tensor $\mathbf{T} = \mathbf{T}_1 + \mathbf{T}_2$, where $\mathbf{T}_1 \in \mathbb{U}_{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbf{T}_2 \in \mathbb{U}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$ with $|\mathbb{I}| = 2$, immediately becomes a tight example of the upper bound of (12) for $m = 2$. The number of blocks, m , can be further increased if we further decompose $\mathbf{T}_1$ and/or $\mathbf{T}_2$ properly. Besides, this class of tensors already includes all the examples discussed in [35, Section 3.2] as a proper subset. For the same reason, the dual version of Theorem 4.4, i.e., Theorem 4.7 to be presented soon, does a similar job to the tightness of the lower bound of (11).

The proof of Theorem 4.4 follows a similar structure to that of Theorem 4.2. It is an immediate consequence of the following two results, Lemma 4.6 and Theorem 4.7. The former can be shown in the same way as that in the proof of Theorem 4.2, and is thus left to interested readers.

Lemma 4.6 *If $d \geq 2$, $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| \geq 1$, and $\mathbb{V}_k$ is a subspace of $\mathbb{R}^{n_k}$ for $k \in [d]$, then (i) $\implies$ (ii), where*

- (i) $\|\mathbf{T} + \mathbf{S}\|_\sigma = \max\{\|\mathbf{T}\|_\sigma, \|\mathbf{S}\|_\sigma\}$ for any $\mathbf{T} \in \mathbb{U}_{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbf{S} \in \mathbb{U}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$;
- (ii) $\|\mathbf{T} + \mathbf{S}\|_* = \|\mathbf{T}\|_* + \|\mathbf{S}\|_*$ for any $\mathbf{T} \in \mathbb{U}_{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbf{S} \in \mathbb{U}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$.

We remark that the actual requirement of Lemma 4.6 is $|\mathbb{I}| \geq 1$ which is weaker than $|\mathbb{I}| \geq 2$ imposed in both Theorem 4.4 and Theorem 4.7. The condition $|\mathbb{I}| \geq 1$ is only required to show (10). As long as we have one mode $k \in \mathbb{I}$, it suffices to have $\bigotimes_{k=1}^d \mathbf{p}_{\mathbb{V}_k^\perp}(\mathbf{x}_k^i) = \mathbf{O}$ in (10) since $\mathbf{x}_k^i \in \mathbb{V}_k$. We believe that the reverse implication of Lemma 4.6 is also true, i.e., the two statements are in fact equivalent. Unfortunately, currently we are unable to verify this claim.

Theorem 4.7 is the dual version of Theorem 4.4 and is a generalization of Theorem 4.3.

Theorem 4.7 *If $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| \geq 2$ and $\mathbb{V}_k$ is a subspace of $\mathbb{R}^{n_k}$ for $k \in [d]$, then the spectral norm is decomposable over $\mathbb{U}_{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbb{U}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$ of the tensor space $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, i.e.,*

$$\|\mathbf{T} + \mathbf{S}\|_\sigma = \max\{\|\mathbf{T}\|_\sigma, \|\mathbf{S}\|_\sigma\} \text{ for any } \mathbf{T} \in \mathbb{U}_{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d) \text{ and } \mathbf{S} \in \mathbb{U}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d).$$

Proof. By monotonicity, it suffices to show the case of $|\mathbb{I}| = 2$. Without loss of generality, we assume that $\mathbb{I} = \{1, 2\}$. By Lemma 3.2, we have

$$\begin{aligned} & \|\mathbf{T} + \mathbf{S}\|_\sigma \\ &= \max_{\mathbf{v}_k \in \mathbb{S}^{n_k} \forall k \in [d]} \left(\left\langle \mathbf{T}, \bigotimes_{k=1}^d \mathbf{v}_k \right\rangle + \left\langle \mathbf{S}, \bigotimes_{k=1}^d \mathbf{v}_k \right\rangle \right) \\ &= \max_{\mathbf{v}_k \in \mathbb{S}^{n_k} \forall k \in [d]} \left(\left\langle \mathbf{T}, \mathbf{p}_{\mathbb{V}_1}(\mathbf{v}_1) \otimes \mathbf{p}_{\mathbb{V}_2}(\mathbf{v}_2) \otimes \bigotimes_{k=3}^d \mathbf{v}_k \right\rangle + \left\langle \mathbf{S}, \mathbf{p}_{\mathbb{V}_1^\perp}(\mathbf{v}_1) \otimes \mathbf{p}_{\mathbb{V}_2^\perp}(\mathbf{v}_2) \otimes \bigotimes_{k=3}^d \mathbf{v}_k \right\rangle \right) \\ &= \max_{\mathbf{v}_k \in \mathbb{S}^{n_k} \forall k \in [d] \setminus \{1\}} \sqrt{\left\| \mathbf{T} \left(\bullet, \mathbf{p}_{\mathbb{V}_2}(\mathbf{v}_2), \bigotimes_{k=3}^d \mathbf{v}_k \right) \right\|_2^2 + \left\| \mathbf{S} \left(\bullet, \mathbf{p}_{\mathbb{V}_2^\perp}(\mathbf{v}_2), \bigotimes_{k=3}^d \mathbf{v}_k \right) \right\|_2^2} \\ &= \max_{\mathbf{v}_k \in \mathbb{S}^{n_k} \forall k \in [d] \setminus \{1\}} \sqrt{\left\| \mathbf{T} \left(\bullet, \bullet, \bigotimes_{k=3}^d \mathbf{v}_k \right) \right\|_\sigma^2 \cdot \|\mathbf{p}_{\mathbb{V}_2}(\mathbf{v}_2)\|_2^2 + \left\| \mathbf{S} \left(\bullet, \bullet, \bigotimes_{k=3}^d \mathbf{v}_k \right) \right\|_\sigma^2 \cdot \|\mathbf{p}_{\mathbb{V}_2^\perp}(\mathbf{v}_2)\|_2^2} \\ &= \max_{\mathbf{v}_k \in \mathbb{S}^{n_k} \forall k \in [d] \setminus [2]} \sqrt{\max \left\{ \left\| \mathbf{T} \left(\bullet, \bullet, \bigotimes_{k=3}^d \mathbf{v}_k \right) \right\|_\sigma^2, \left\| \mathbf{S} \left(\bullet, \bullet, \bigotimes_{k=3}^d \mathbf{v}_k \right) \right\|_\sigma^2 \right\}} \\ &= \max\{\|\mathbf{T}\|_\sigma, \|\mathbf{S}\|_\sigma\}, \end{aligned}$$

where the third equality is due to Cauchy-Schwarz inequality and the fact that

$$\mathbf{T} \left(\bullet, \mathbf{p}_{\mathbb{V}_2}(\mathbf{v}_2), \bigotimes_{k=3}^d \mathbf{v}_k \right) \in \mathbb{V}_1 \text{ and } \mathbf{S} \left(\bullet, \mathbf{p}_{\mathbb{V}_2^\perp}(\mathbf{v}_2), \bigotimes_{k=3}^d \mathbf{v}_k \right) \in \mathbb{V}_1^\perp,$$

the fourth is due to the consistency of the matrix spectral norm, the second to last is due to $\|\mathbf{p}_{\mathbb{V}_2}(\mathbf{v}_2)\|_2^2 + \|\mathbf{p}_{\mathbb{V}_2^\perp}(\mathbf{v}_2)\|_2^2 = 1$, and the last is due to that $\max_{\mathbf{v}_k \in \mathbb{S}^{n_k} \forall k \in [d] \setminus [2]} \|\mathbf{T}(\bullet, \bullet, \bigotimes_{k=3}^d \mathbf{v}_k)\|_\sigma = \|\mathbf{T}\|_\sigma$ holds for any tensor $\mathbf{T}$; see (5). $\square$

Theorem 4.7 further implies a generalized decomposability of the nuclear norm.

Theorem 4.8 *If $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| \geq 2$ and $\mathbb{V}_k$ is a subspace of $\mathbb{R}^{n_k}$ for $k \in [d]$, then*

$$\|\mathbf{T}\|_* \geq \|\mathbf{p}_{\mathbb{U}_{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)}(\mathbf{T})\|_* + \|\mathbf{p}_{\mathbb{U}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)}(\mathbf{T})\|_* \text{ for any } \mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}.$$

Proof. Let $\mathbf{T} = \mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3$, where $\mathbf{T}_1 = \mathbf{p}_{\mathbb{U}_{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)}(\mathbf{T})$ and $\mathbf{T}_2 = \mathbf{p}_{\mathbb{U}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)}(\mathbf{T})$. As a result, $\mathbf{T}_1$, $\mathbf{T}_2$, and $\mathbf{T}_3$ reside in mutually orthogonal subspaces. By Lemma 3.3, there exist $\mathbf{Z}_1 \in \mathbb{U}_{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbf{Z}_2 \in \mathbb{U}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$ with $\|\mathbf{Z}_1\|_{\sigma} = \|\mathbf{Z}_2\|_{\sigma} = 1$ such that $\langle \mathbf{T}_1, \mathbf{Z}_1 \rangle = \|\mathbf{T}_1\|_*$ and $\langle \mathbf{T}_2, \mathbf{Z}_2 \rangle = \|\mathbf{T}_2\|_*$. By Theorem 4.7, we also know $\|\mathbf{Z}_1 + \mathbf{Z}_2\|_{\sigma} = \max\{\|\mathbf{Z}_1\|_{\sigma}, \|\mathbf{Z}_2\|_{\sigma}\} = 1$. These, together with the duality in Lemma 2.1, further imply that

$$\|\mathbf{T}\|_* \geq \langle \mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3, \mathbf{Z}_1 + \mathbf{Z}_2 \rangle = \langle \mathbf{T}_1, \mathbf{Z}_1 \rangle + \langle \mathbf{T}_2, \mathbf{Z}_2 \rangle = \|\mathbf{T}_1\|_* + \|\mathbf{T}_2\|_*,$$

where the first equality is due to that $\langle \mathbf{T}_i, \mathbf{Z}_j \rangle = 0$ for any $i \neq j$. $\square$

We remark that Theorem 4.8 includes Theorem 4.4 as a special case. Since $\mathbf{T}_3 = \mathbf{O}$ under the circumstances of Theorem 4.4, we have $\|\mathbf{T}\|_* \leq \|\mathbf{T}_1\|_* + \|\mathbf{T}_2\|_*$ by the triangle inequality and so Theorem 4.8 holds as an equality. It is also worth mentioning that Theorem 4.8 applies to any tensor instead of a direct sum of two tensors in mutually orthogonal subspaces. This broadens its applicability. For example, restricting to the matrix case, Theorem 4.8 reduces to

$$\|\mathbf{T}\|_* \geq \|\mathbf{p}_{\mathbb{T}(\mathbb{V}_1, \mathbb{V}_2)}(\mathbf{T})\|_* + \|\mathbf{p}_{\mathbb{T}(\mathbb{V}_1^{\perp}, \mathbb{V}_2^{\perp})}(\mathbf{T})\|_* \text{ if } \mathbb{V}_k \text{ is a subspace of } \mathbb{R}^{n_k} \text{ for } k \in [2].$$

We are not aware of such a generalization of the decomposability of the matrix nuclear norm, to the best of our knowledge. For tensor spaces, Theorem 4.8 also provides a new way to bound the tensor nuclear norm from below via the flexibility of $\mathbb{U}_{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbb{U}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$.

We conclude this section with a discussion of pairs of subspaces used in the decomposability of the tensor nuclear norm. Although not explicitly presented, the subspace $\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$ of $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ proposed in [62] actually implies (by a property similar to Lemma 5.1 to be discussed soon) another weak decomposability as that of $\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^3)$ in $\mathbb{R}^{n_1 \times n_2 \times n_3}$, i.e.,

$$\|\mathbf{T} + \mathbf{S}\|_* \geq \|\mathbf{T}\|_* + \frac{2}{d(d-1)} \|\mathbf{S}\|_* \text{ for any } \mathbf{T} \in \mathbb{T}((\mathbb{V}_k)_{k=1}^d) \text{ and } \mathbf{S} \in \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d).$$

The subspace in which $\mathbf{S}$ resides includes all the basic subspaces of $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ spanned by at least two $\mathbb{V}_k^{\perp}$'s. Therefore, this weak decomposability applies to the pair of subspaces, $\mathbb{T}((\mathbb{V}_k)_{k=1}^d)$ and $\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$, including 1 and $2^d - d - 1$ basic subspaces, respectively. The decomposability becomes (quadratically) weaker as d increases. By contrast, the natural and restrictive full decomposability in Theorem 4.2 applies to $\mathbb{T}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbb{T}^{[d]}((\mathbb{V}_k)_{k=1}^d)$, both being one basic subspace, whereas the improved full decomposability in Theorem 4.4 applies to $\mathbb{U}_{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$ and $\mathbb{U}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d)$, both including $2^{d-|\mathbb{I}|}$ basic subspaces and attaining the maximum 2^{d-2} when $|\mathbb{I}| = 2$.

5 Subdifferential of the tensor nuclear norm

The subdifferential of a convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at $\mathbf{x} \in \mathbb{R}^n$ is defined as

$$\partial f(\mathbf{x}) := \{\mathbf{z} \in \mathbb{R}^n : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{z}, \mathbf{y} - \mathbf{x} \rangle \forall \mathbf{y} \in \mathbb{R}^n\},$$

and the elements of $\partial f(\mathbf{x})$ are called subgradients; see, e.g., [52, Section 23] and [44, Section 1]. As an immediate consequence of [53, Corollary 8.25], the subdifferential of a norm $\|\bullet\|_{\diamond} : \mathbb{R}^n \rightarrow \mathbb{R}$ has a representation

$$\partial \|\mathbf{x}\|_{\diamond} = \{\mathbf{z} \in \mathbb{R}^n : \langle \mathbf{z}, \mathbf{x} \rangle = \|\mathbf{x}\|_{\diamond}, \|\mathbf{z}\|_{\diamond} \leq 1\},$$

where $\|\bullet\|_\circ$ is the dual norm of $\|\bullet\|_\diamond$.

It is well known that the subdifferential of the matrix nuclear norm has an explicit characterization. If $\mathbf{T} = \mathbf{U}\mathbf{D}\mathbf{V}^T \in \mathbb{R}^{n_1 \times n_2}$ is a compact SVD, then

$$\partial\|\mathbf{T}\|_* = \{\mathbf{U}\mathbf{V}^T + \mathbf{X} : \mathbf{X} \in \mathbb{T}^{\{1,2\}}(\mathbf{T}), \|\mathbf{X}\|_\sigma \leq 1\}; \quad (13)$$

see, e.g., [58, Example 2] and [33]. We remark that $\mathbf{U}\mathbf{V}^T$ is invariant over all compact SVDs of $\mathbf{T}$ due to Autonne's uniqueness [27, Theorem 2.6.5]. The importance of this representation has been well recognized in mathematical optimization and statistics.

In the tensor space, however, only two limited subdifferential inclusions of the nuclear norm are known in the literature, to the best of our knowledge. It was first stated in [61, Lemma 1] that for any third-order tensor $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$,

$$\mathbb{D}_1(\mathbf{T}) := \left\{ \mathbf{Z} + \mathbf{X} : \mathbf{Z} \in \mathbb{Z}(\mathbf{T}), \mathbf{X} \in \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T}), \|\mathbf{X}\|_\sigma \leq \frac{1}{2} \right\} \subseteq \partial\|\mathbf{T}\|_*, \quad (14)$$

where for any $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$,

$$\mathbb{Z}(\mathbf{T}) := \begin{cases} \{\mathbf{O}\} & \mathbf{T} = \mathbf{O} \\ \{\mathbf{Z} \in \mathbb{T}(\mathbf{T}) : \langle \mathbf{Z}, \mathbf{T} \rangle = \|\mathbf{T}\|_*, \|\mathbf{Z}\|_\sigma = 1\} & \mathbf{T} \neq \mathbf{O}. \end{cases} \quad (15)$$

Later, it was stated in [62, Theorem 1] that for any $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$,

$$\mathbb{D}_2(\mathbf{T}) := \left\{ \mathbf{Z} + \mathbf{X} : \mathbf{Z} \in \mathbb{Z}(\mathbf{T}), \mathbf{X} \in \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}(\mathbf{T}), \|\mathbf{X}\|_\sigma \leq \frac{2}{d(d-1)} \right\} \subseteq \partial\|\mathbf{T}\|_*. \quad (16)$$

We remark that although $\mathbb{D}_2(\mathbf{T})$ applies to tensors of an arbitrary order, it does not include $\mathbb{D}_1(\mathbf{T})$ as a special case. In fact, $\mathbb{D}_2(\mathbf{T})$ for $d = 3$ is strictly smaller than $\mathbb{D}_1(\mathbf{T})$. The two inclusions have been applied in tensor completions [61, 62]. It is also worth mentioning that $\mathbb{Z}(\mathbf{T})$ consists of exactly one element if $\mathbf{T}$ is a matrix, as a consequence of Von Neumann's trace inequality [51, Theorem 0.1].

5.1 Decomposability and subdifferential

To gain a better understanding of the subdifferential of the tensor nuclear norm, let us first establish its connections with the decomposability. This enables us to apply the results developed in Section 4.

Lemma 5.1 *If $d \geq 2$, $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| \geq 1$, and $\tau > 0$, then (i) $\implies$ (ii) $\implies$ (iii), where*

- (i) $\|\mathbf{T} + \mathbf{S}\|_\sigma \leq \max\{\|\mathbf{T}\|_\sigma, \|\mathbf{S}\|_\sigma/\tau\}$ for any $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ and $\mathbf{S} \in \mathbb{U}^{\mathbb{I}}(\mathbf{T})$;
- (ii) $\{\mathbf{Z} + \mathbf{X} : \mathbf{Z} \in \mathbb{Z}(\mathbf{T}), \mathbf{X} \in \mathbb{U}^{\mathbb{I}}(\mathbf{T}), \|\mathbf{X}\|_\sigma \leq \tau\} \subseteq \partial\|\mathbf{T}\|_*$ for any $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$;
- (iii) $\|\mathbf{T} + \mathbf{S}\|_* \geq \|\mathbf{T}\|_* + \tau\|\mathbf{S}\|_*$ for any $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ and $\mathbf{S} \in \mathbb{U}^{\mathbb{I}}(\mathbf{T})$.

Proof. Suppose that (i) holds. For any $\mathbf{Z} + \mathbf{X}$ in the left-hand-side set in (ii), $\mathbf{Z} \in \mathbb{Z}(\mathbf{T})$ implies that $\mathbb{U}^{\mathbb{I}}(\mathbf{T}) \subseteq \mathbb{U}^{\mathbb{I}}(\mathbf{Z})$ and so $\mathbf{X} \in \mathbb{U}^{\mathbb{I}}(\mathbf{T}) \subseteq \mathbb{U}^{\mathbb{I}}(\mathbf{Z})$. As a result, $\|\mathbf{Z} + \mathbf{X}\|_\sigma \leq \max\{\|\mathbf{Z}\|_\sigma, \|\mathbf{X}\|_\sigma/\tau\} \leq 1$. Therefore, for any $\mathbf{Y} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$,

$$\langle \mathbf{Z} + \mathbf{X}, \mathbf{Y} - \mathbf{T} \rangle = \langle \mathbf{Z} + \mathbf{X}, \mathbf{Y} \rangle - \langle \mathbf{Z}, \mathbf{T} \rangle - \langle \mathbf{X}, \mathbf{T} \rangle \leq \|\mathbf{Z} + \mathbf{X}\|_\sigma \|\mathbf{Y}\|_* - \|\mathbf{T}\|_* \leq \|\mathbf{Y}\|_* - \|\mathbf{T}\|_*, \quad (17)$$

where the first inequality is due to Lemma 2.1, (15), and the orthogonality between $\mathbf{T} \in \mathbb{T}(\mathbf{T})$ and $\mathbf{X} \in \mathbb{U}^{\mathbb{I}}(\mathbf{T})$. This shows that $\mathbf{Z} + \mathbf{X} \in \partial\|\mathbf{T}\|_*$.

Suppose that (ii) holds. For any $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ and $\mathbf{S} \in \mathbb{U}^{\mathbb{I}}(\mathbf{T})$, there exists an $\mathbf{X} \in \mathbb{U}^{\mathbb{I}}(\mathbf{T})$, such that $\langle \mathbf{X}, \mathbf{S} \rangle = \|\mathbf{S}\|_*$ and $\|\mathbf{X}\|_\sigma = 1$ by Lemma 3.3. As a result, $\mathbf{Z} + \tau\mathbf{X} \in \partial\|\mathbf{T}\|_*$ for any $\mathbf{Z} \in \mathbb{Z}(\mathbf{T})$. Therefore,

$$\|\mathbf{T} + \mathbf{S}\|_* \geq \|\mathbf{T}\|_* + \langle \mathbf{Z} + \tau\mathbf{X}, \mathbf{S} \rangle = \|\mathbf{T}\|_* + \langle \mathbf{Z}, \mathbf{S} \rangle + \tau\langle \mathbf{X}, \mathbf{S} \rangle = \|\mathbf{T}\|_* + \tau\|\mathbf{S}\|_*,$$

where the last equality is due to the orthogonality between $\mathbf{Z} \in \mathbb{Z}(\mathbf{T}) \subseteq \mathbb{T}(\mathbf{T})$ and $\mathbf{S} \in \mathbb{U}^{\mathbb{I}}(\mathbf{T})$. $\square$

In a nutshell, a decomposability of the spectral norm implies an inclusion of the subdifferential of the nuclear norm, which in turn implies a decomposability of the nuclear norm. We remark that Lemma 5.1 actually specializes to a universal version of Lemma 4.6, i.e., (i) $\implies$ (iii) when $\tau = 1$. It is worth noting that when $\tau = 1$, the inequalities in (i) and (iii) of Lemma 5.1 actually become equalities, i.e., $\|\mathbf{T} + \mathbf{S}\|_\sigma = \max\{\|\mathbf{T}\|_\sigma, \|\mathbf{S}\|_\sigma\}$ due to Lemma 3.2 and $\|\mathbf{T} + \mathbf{S}\|_* = \|\mathbf{T}\|_* + \|\mathbf{S}\|_*$ due to the triangle inequality. As a result of Lemma 5.1 with $\tau = 1$, the decomposability of the spectral norm discussed in Section 4, in particular Theorem 4.7, immediately provides new inclusions of the subdifferential of the nuclear norm.

Corollary 5.2 *If $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ and $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| \geq 2$, then*

$$\mathbb{D}^{\mathbb{I}}(\mathbf{T}) := \{\mathbf{Z} + \mathbf{X} : \mathbf{Z} \in \mathbb{Z}(\mathbf{T}), \mathbf{X} \in \mathbb{U}^{\mathbb{I}}(\mathbf{T}), \|\mathbf{X}\|_\sigma \leq 1\} \subseteq \partial\|\mathbf{T}\|_*. \quad (18)$$

When $d = 2$, Corollary 5.2 reduces to the maximal inclusion of the subdifferential of the matrix nuclear norm, i.e., (13). It is important to remark that $\mathbb{U}^{\mathbb{I}}(\mathbf{T})$, a properly chosen subspace for $\mathbf{X}$, has made the full stretch possible for $\mathbf{X}$ whose spectral norm can be as large as 1. Moreover, Corollary 5.2 offers the flexibility to choose any $\mathbb{U}^{\mathbb{I}}(\mathbf{T})$ as long as $|\mathbb{I}| \geq 2$. As a result, by combining all possible $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| = 2$ for higher-order tensors, we can have a larger inclusion.

Theorem 5.3 *If $d \geq 2$ and $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, then*

$$\begin{aligned} \mathbb{D}(\mathbf{T}) &:= \text{conv} \left(\bigcup_{|\mathbb{I}|=2, \mathbb{I} \subseteq [d]} \mathbb{D}^{\mathbb{I}}(\mathbf{T}) \right) \\ &= \left\{ \mathbf{Z} + \mathbf{X} : \mathbf{Z} \in \mathbb{Z}(\mathbf{T}), \mathbf{X} \in \text{conv} \left(\{\mathbf{Y} : \|\mathbf{Y}\|_\sigma \leq 1\} \cap \bigcup_{|\mathbb{I}|=2, \mathbb{I} \subseteq [d]} \mathbb{U}^{\mathbb{I}}(\mathbf{T}) \right) \right\} \\ &\subseteq \partial\|\mathbf{T}\|_*. \end{aligned} \quad (19)$$

Proof. Because any subdifferential is convex, the inclusion $\mathbb{D}(\mathbf{T}) \subseteq \partial\|\mathbf{T}\|_*$ is immediate by Corollary 5.2. The second equality can also be easily verified by noticing that $\mathbb{Z}(\mathbf{T})$ is convex (see the discussions before Example 5.7) and orthogonal to $\bigcup_{|\mathbb{I}|=2, \mathbb{I} \subseteq [d]} \mathbb{U}^{\mathbb{I}}(\mathbf{T})$. $\square$

We remark that $\bigcup_{|\mathbb{I}|=2, \mathbb{I} \subseteq [d]} \mathbb{U}^{\mathbb{I}}(\mathbf{T})$ does not include every direction of $\sum_{|\mathbb{I}|=2, \mathbb{I} \subseteq [d]} \mathbb{U}^{\mathbb{I}}(\mathbf{T})$. However, even when the former is intersected by a spectral ball, its convex hull does contain all directions of the latter, simply because

$$\text{conv} \left(\bigcup_{|\mathbb{I}|=2, \mathbb{I} \subseteq [d]} \mathbb{U}^{\mathbb{I}}(\mathbf{T}) \right) = \sum_{|\mathbb{I}|=2, \mathbb{I} \subseteq [d]} \mathbb{U}^{\mathbb{I}}(\mathbf{T}) = \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}(\mathbf{T}).$$

Before we analyze in detail the subdifferential of the tensor nuclear norm in Section 5.2, let us discuss the relations among $\mathbb{D}_1(\mathbf{T})$, $\mathbb{D}_2(\mathbf{T})$, and $\mathbb{D}(\mathbf{T})$. All these sets consist of subgradients in the form of $\mathbf{Z} + \mathbf{X}$, where $\mathbf{Z} \in \mathbb{Z}(\mathbf{T})$ and $\mathbf{X} \in \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$. There is no difference in terms of $\mathbf{Z}$ as it must be chosen exactly from $\mathbb{Z}(\mathbf{T})$ and there is also no difference in terms of the direction of $\mathbf{X}$ available from $\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$. However, the key difference among these three sets is the spectral size of $\mathbf{X}$, i.e., how large $\|\mathbf{X}\|_\sigma$ can be. The set $\mathbb{D}(\mathbf{T})$ clearly beats $\mathbb{D}_1(\mathbf{T})$ and $\mathbb{D}_2(\mathbf{T})$ for any $\mathbf{X} \in \bigcup_{|\mathbb{I}|=2, \mathbb{I} \subseteq [d]} \mathbb{U}^{\mathbb{I}}(\mathbf{T})$ where a full stretch is attainable, i.e., $\|\mathbf{X}\|_\sigma \leq 1$ instead of $\|\mathbf{X}\|_\sigma \leq \frac{1}{2}$ for $\mathbb{D}_1(\mathbf{T})$ and $\|\mathbf{X}\|_\sigma \leq \frac{2}{d(d-1)}$ for $\mathbb{D}_2(\mathbf{T})$. As a result, a lot of subgradients have been missed in $\mathbb{D}_1(\mathbf{T})$ and $\mathbb{D}_2(\mathbf{T})$. Moreover, $\frac{2}{d(d-1)}$ tends to zero as d tends to infinity.

Proposition 5.4 *If $d \geq 2$ and $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, then*

$$\begin{aligned} \mathbb{D}(\mathbf{T}) \setminus \mathbb{D}_1(\mathbf{T}) &\supseteq \left\{ \mathbf{Z} + \mathbf{X} : \mathbf{Z} \in \mathbb{Z}(\mathbf{T}), \mathbf{X} \in \bigcup_{|\mathbb{I}|=2, \mathbb{I} \subseteq [3]} \mathbb{U}^{\mathbb{I}}(\mathbf{T}), \frac{1}{2} < \|\mathbf{X}\|_\sigma \leq 1 \right\} \text{ if } d = 3, \\ \mathbb{D}(\mathbf{T}) \setminus \mathbb{D}_2(\mathbf{T}) &\supseteq \left\{ \mathbf{Z} + \mathbf{X} : \mathbf{Z} \in \mathbb{Z}(\mathbf{T}), \mathbf{X} \in \bigcup_{|\mathbb{I}|=2, \mathbb{I} \subseteq [d]} \mathbb{U}^{\mathbb{I}}(\mathbf{T}), \frac{2}{d(d-1)} < \|\mathbf{X}\|_\sigma \leq 1 \right\}. \end{aligned}$$

When $d = 3$, there does exist an $\mathbf{X} \in \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T}) \setminus \bigcup_{|\mathbb{I}|=2, \mathbb{I} \subseteq [3]} \mathbb{U}^{\mathbb{I}}(\mathbf{T})$ such that $\mathbf{Z} + \mathbf{X} \in \mathbb{D}(\mathbf{T})$ requires $\|\mathbf{X}\|_\sigma \leq \alpha$ for some $\alpha < \frac{1}{2}$, implying that $\mathbb{D}_1(\mathbf{T}) \setminus \mathbb{D}(\mathbf{T}) \neq \emptyset$ when $d = 3$. To be specific, the tensor $\mathbf{X}(\frac{1}{3})$ in Example 5.9, adding any $\mathbf{Z} \in \mathbb{Z}(\mathbf{T})$, resides on the boundary of $\mathbb{D}(\mathbf{T})$ but $\|\mathbf{X}(\frac{1}{3})\|_\sigma = \frac{2}{3\sqrt{3}} < \frac{1}{2}$. However, $\mathbb{D}(\mathbf{T})$ always includes $\mathbb{D}_2(\mathbf{T})$ as a proper subset for any $d \geq 3$. We believe that $\mathbb{D}(\mathbf{T})$ is more useful in applications.

Proposition 5.5 *If $d \geq 2$ and $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, then $\mathbb{D}_2(\mathbf{T}) \subseteq \mathbb{D}(\mathbf{T})$.*

Proof. Let $\mathbf{Z} + \mathbf{X} \in \mathbb{D}_2(\mathbf{T})$ where $\mathbf{Z} \in \mathbb{Z}(\mathbf{T})$ and $\mathbf{X} \in \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$ with $\|\mathbf{X}\|_\sigma \leq \frac{2}{d(d-1)}$. It suffices to show that $\mathbf{X} \in \text{conv}(\{\mathbf{Y} : \|\mathbf{Y}\|_\sigma \leq 1\} \cap \bigcup_{|\mathbb{I}|=2, \mathbb{I} \subseteq [d]} \mathbb{U}^{\mathbb{I}}(\mathbf{T}))$.

For any $\mathbb{I} = \{i, j\}$ with $1 \leq i < j \leq d$, denote $\mathbb{W}^{\mathbb{I}}(\mathbf{T}) = \mathbb{W}^{\mathbb{I}}((\text{sp}_k(\mathbf{T}))_{k=1}^d)$, where

$$\mathbb{W}^{\mathbb{I}}((\mathbb{V}_k)_{k=1}^d) := \text{sp} \left(\left\{ \bigotimes_{k=1}^d \mathbf{v}_k : \mathbf{v}_k \in \begin{cases} \mathbb{V}_k & k \in [j-1] \setminus \{i\} \\ \mathbb{R}^{n_k} & k \in [d] \setminus [j] \\ \mathbb{V}_k^\perp & k = i, j \end{cases} \right\} \right)$$

for any subspace $\mathbb{V}_k \subseteq \mathbb{R}^{n_k}$ for $k \in [d]$. It is easy to verify that $\bigoplus_{|\mathbb{I}|=2, \mathbb{I} \subseteq [d]} \mathbb{W}^{\mathbb{I}}(\mathbf{T}) = \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$ and $\mathbb{W}^{\mathbb{I}}(\mathbf{T}) \subseteq \mathbb{U}^{\mathbb{I}}(\mathbf{T})$. Since there are exactly $s = \frac{d(d-1)}{2}$ different index sets $\mathbb{I}$ satisfying $|\mathbb{I}| = 2$ and $\mathbb{I} \subseteq [d]$, we may denote them to be $\mathbb{I}_1, \mathbb{I}_2, \dots, \mathbb{I}_s$. As a result, we have $\mathbf{X} \in \bigoplus_{i=1}^s \mathbb{W}^{\mathbb{I}_i}(\mathbf{T})$ and so $\mathbf{X}$ can be uniquely decomposed as $\sum_{i=1}^s \mathbf{X}_i$, where $\mathbf{X}_i \in \mathbb{W}^{\mathbb{I}_i}(\mathbf{T})$ for $i \in [s]$.

It is important to observe that $\|\mathbf{X}_i\|_\sigma \leq \|\mathbf{X}\|_\sigma \leq \frac{1}{s}$ for any $i \in [s]$. To see why, let $\mathbb{I}_i = \{j_1, j_2\}$ with $j_1 < j_2$. By Lemma 3.2, there exist $\mathbf{v}_k \in \mathbb{S}^{n_k}$ with

$$\mathbf{v}_k \in \begin{cases} \text{sp}_k(\mathbf{T}) & k \in [j_2 - 1] \setminus \{j_1\} \\ \mathbb{R}^{n_k} & k \in [d] \setminus [j_2] \\ \text{sp}_k^\perp(\mathbf{T}) & k = j_1, j_2 \end{cases}$$

for $k \in [d]$, such that

$$\|\mathbf{X}_i\|_\sigma = \left\langle \mathbf{X}_i, \bigotimes_{k=1}^d \mathbf{v}_k \right\rangle = \left\langle \text{p}_{\mathbb{W}^{\mathbb{I}_i}(\mathbf{T})}(\mathbf{X}), \bigotimes_{k=1}^d \mathbf{v}_k \right\rangle = \left\langle \mathbf{X}, \text{p}_{\mathbb{W}^{\mathbb{I}_i}(\mathbf{T})} \left(\bigotimes_{k=1}^d \mathbf{v}_k \right) \right\rangle \leq \|\mathbf{X}\|_\sigma.$$

As a result, we have

$$\mathbf{X} = \frac{1}{s} \sum_{i=1}^s s\mathbf{X}_i \text{ with } \|s\mathbf{X}_i\|_\sigma \leq 1 \text{ and } s\mathbf{X}_i \in \mathbb{W}^{\mathbb{I}_i}(\mathbf{T}) \subseteq \mathbb{U}^{\mathbb{I}_i}(\mathbf{T}) \text{ for } i \in [s],$$

implying that $\mathbf{X} \in \text{conv}(\{\mathbf{Y} : \|\mathbf{Y}\|_\sigma \leq 1\} \cap \bigcup_{|\mathbb{I}|=2, \mathbb{I} \subseteq [d]} \mathbb{U}^{\mathbb{I}}(\mathbf{T}))$. $\square$

5.2 Subgradients in subspaces

A properly chosen subspace has not only provided better inclusions of the subdifferential of the tensor nuclear norm such as Corollary 5.2 and Theorem 5.3, but also resulted the full decomposability of the tensor nuclear norm such as Theorem 4.4. In this part we look into the details of subgradients of the tensor nuclear norm in various subspaces of interest. In particular, we look into the structure of the subdifferential and estimate the bounds of the spectral norm in relevant subspaces, in order to provide a class of tensors that must be subgradients and a class that cannot be.

Let us first examine the matrix case. Recall that $\partial\|\mathbf{T}\|_* = \{\mathbf{UV}^T + \mathbf{X} : \mathbf{X} \in \mathbb{T}^{\{1,2\}}(\mathbf{T}), \|\mathbf{X}\|_\sigma \leq 1\}$ if $\mathbf{T} = \mathbf{UDV}^T$ is a compact SVD. The projections of any subgradient $\mathbf{G} \in \partial\|\mathbf{T}\|_*$ onto the four basic subspaces, $\mathbb{T}(\mathbf{T})$, $\mathbb{T}^{\{1\}}(\mathbf{T})$, $\mathbb{T}^{\{2\}}(\mathbf{T})$, and $\mathbb{T}^{\{1,2\}}(\mathbf{T})$, behave exactly as follows:

- (i) $p_{\mathbb{T}(\mathbf{T})}(\mathbf{G}) = \mathbf{UV}^T$ which is unique although the SVD may not be unique;
- (ii) $p_{\mathbb{T}^{\{1\}}(\mathbf{T})}(\mathbf{G}) = \mathbf{O}$ and $p_{\mathbb{T}^{\{2\}}(\mathbf{T})}(\mathbf{G}) = \mathbf{O}$;
- (iii) $p_{\mathbb{T}^{\{1,2\}}(\mathbf{T})}(\mathbf{G})$ can be any matrix as long as $\|p_{\mathbb{T}^{\{1,2\}}(\mathbf{T})}(\mathbf{G})\|_\sigma \leq 1$.

Perhaps it is the last property that motivated the construction of the subspace $\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$ when the subdifferential of the tensor nuclear norm was first studied in [61, 62]. This is indeed a good choice as we will explain soon, but there are more interesting ones, such as $\mathbb{U}^{\mathbb{I}}(\mathbf{T})$ for any $|\mathbb{I}| \geq 2$ in Corollary 5.2. As the first fact we observed, none of the above three properties holds for higher-order tensors. Let us now examine the tensor case.

Given a nonzero $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, we have by definition that

$$\begin{aligned} \partial\|\mathbf{T}\|_* &= \{\mathbf{Y} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d} : \langle \mathbf{T}, \mathbf{Y} \rangle = \|\mathbf{T}\|_*, \|\mathbf{Y}\|_\sigma \leq 1\} \\ &= \left\{ \mathbf{Z} + \mathbf{X} : \mathbf{Z} \in \mathbb{Z}(\mathbf{T}), \mathbf{X} \in \bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}(\mathbf{T}), \|\mathbf{Z} + \mathbf{X}\|_\sigma \leq 1 \right\}, \end{aligned} \quad (20)$$

where $\mathbb{Z}(\mathbf{T}) = \{\mathbf{Z} \in \mathbb{T}(\mathbf{T}) : \langle \mathbf{Z}, \mathbf{T} \rangle = \|\mathbf{T}\|_*, \|\mathbf{Z}\|_\sigma = 1\}$ from (15). This natural decomposition of subgradients into $\mathbf{Z} + \mathbf{X}$ allows to examine $\mathbf{Z}$ and $\mathbf{X}$ separately in two mutually orthogonal subspaces but one needs to make sure $\|\mathbf{Z} + \mathbf{X}\|_\sigma \leq 1$. Let us first consider the set $\mathbb{Z}(\mathbf{T})$.

Lemma 5.6 *If $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, then $\mathbb{Z}(\mathbf{T})$ is nonempty, compact, and convex.*

Proof. It suffices to consider the case $\mathbf{T} \neq \mathbf{O}$. The nonemptiness and compactness follow easily from Lemma 3.3 and the definition in (15). For the convexity, given any $\mathbf{Z}_1, \mathbf{Z}_2 \in \mathbb{Z}(\mathbf{T})$ and $\alpha_1, \alpha_2 \geq 0$ with $\alpha_1 + \alpha_2 = 1$, we have

$$\|\mathbf{T}\|_* = \alpha_1 \|\mathbf{T}\|_* + \alpha_2 \|\mathbf{T}\|_* = \alpha_1 \langle \mathbf{Z}_1, \mathbf{T} \rangle + \alpha_2 \langle \mathbf{Z}_2, \mathbf{T} \rangle = \langle \alpha_1 \mathbf{Z}_1 + \alpha_2 \mathbf{Z}_2, \mathbf{T} \rangle \leq \|\alpha_1 \mathbf{Z}_1 + \alpha_2 \mathbf{Z}_2\|_\sigma \|\mathbf{T}\|_*,$$

implying that $\|\alpha_1 \mathbf{Z}_1 + \alpha_2 \mathbf{Z}_2\|_\sigma \geq 1$. By noticing that $\|\alpha_1 \mathbf{Z}_1 + \alpha_2 \mathbf{Z}_2\|_\sigma \leq \alpha_1 \|\mathbf{Z}_1\|_\sigma + \alpha_2 \|\mathbf{Z}_2\|_\sigma = 1$, we in fact have $\|\alpha_1 \mathbf{Z}_1 + \alpha_2 \mathbf{Z}_2\|_\sigma = 1$, implying that $\alpha_1 \mathbf{Z}_1 + \alpha_2 \mathbf{Z}_2 \in \mathbb{Z}(\mathbf{T})$. $\square$

Unlike the matrix case, the set $\mathbb{Z}(\mathbf{T})$, however, may not be a singleton when $d \geq 3$.

Example 5.7 Let $\mathbf{T} = \sum_{i=1}^3 \mathbf{e}_i \otimes \mathbf{e}_i \otimes \mathbf{e}_i \in \mathbb{R}^{3 \times 3 \times 3}$ and $\mathbf{Z}(t) = \mathbf{T} + t \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_3 \in \mathbb{R}^{3 \times 3 \times 3}$. It is obvious that $\mathbf{Z}(t) \in \mathbb{T}(\mathbf{T}) \in \mathbb{R}^{3 \times 3 \times 3}$. For any $-1 \leq t \leq 1$, it can be verified that $\|\mathbf{Z}(t)\|_\sigma = 1$ (see Lemma A.1) and $\langle \mathbf{Z}(t), \mathbf{T} \rangle = 3 = \|\mathbf{T}\|_*$. Therefore, we have $\mathbf{Z}(t) \in \mathbb{Z}(\mathbf{T})$ for any $-1 \leq t \leq 1$.

We next turn to the subspace $\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$ in (20) where $\mathbf{X}$ resides. Unlike the matrix case where the subspaces $\mathbb{T}^{\{1\}}(\mathbf{T})$ and $\mathbb{T}^{\{2\}}(\mathbf{T})$ are both intangible for any matrix $\mathbf{T}$, a nonzero $\mathbf{X}$ is in fact possible in, e.g., $\mathbb{T}^{\{3\}}(\mathbf{T})$, of a third-order tensor $\mathbf{T}$.

Example 5.8 Let $\mathbf{T} = \sum_{i=1}^2 \mathbf{e}_i \otimes \mathbf{e}_i \otimes \mathbf{e}_i \in \mathbb{R}^{2 \times 2 \times 3}$. We have $\text{sp}_1(\mathbf{T}) = \text{sp}_2(\mathbf{T}) = \mathbb{R}^2$ and $\text{sp}_3(\mathbf{T}) = \text{sp}(\mathbf{e}_1, \mathbf{e}_2)$. Let $\mathbf{Z} = \mathbf{T} \in \mathbb{Z}(\mathbf{T})$ since $\langle \mathbf{Z}, \mathbf{T} \rangle = 2 = \|\mathbf{T}\|_*$ and $\|\mathbf{Z}\|_\sigma = 1$. We have the following two observations in contrast.

- (i) Let $\mathbf{X}(t) = t \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_3 \in \mathbb{T}^{\{3\}}(\mathbf{T})$. For any $-1 \leq t \leq 1$, it can be verified that $\|\mathbf{Z} + \mathbf{X}(t)\|_\sigma = 1$ (see Lemma A.2). Therefore, $\mathbf{Z} + \mathbf{X}(t) \in \partial\|\mathbf{T}\|_*$ for any $-1 \leq t \leq 1$.
- (ii) Let $\mathbf{Y}(t) = t \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_3 \in \mathbb{T}^{\{3\}}(\mathbf{T})$. For any $t \neq 0$, it can be verified that $\|\mathbf{Z} + \mathbf{Y}(t)\|_\sigma > 1$ (see Lemma A.2). Therefore, $\mathbf{Z} + \mathbf{Y}(t) \notin \partial\|\mathbf{T}\|_*$ for any $t \neq 0$.

Although the inclusions discussed in Section 5.1 focus on basic subspaces that are spanned by at least two $\text{sp}_k^\perp(\mathbf{T})$'s, no matter for $\mathbb{D}_1(\mathbf{T})$, $\mathbb{D}_2(\mathbf{T})$, or $\mathbb{D}(\mathbf{T})$, it is important not to disregard basic subspaces that are spanned by only one $\text{sp}_k^\perp(\mathbf{T})$. Another interesting observation is that, for some subspace where $\mathbf{X}$ resides, e.g., $\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$ in $\mathbb{D}_1(\mathbf{T})$, the tensor $\mathbf{X}$ can surprisingly go beyond the full stretch, i.e., $\|\mathbf{X}\|_\sigma > 1$.

Example 5.9 Let $\mathbf{T} = \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 \in \mathbb{R}^{2 \times 2 \times 2}$. We have $\text{sp}_k(\mathbf{T}) = \text{sp}(\mathbf{e}_1)$ for $k \in [3]$. Let $\mathbf{Z} = \mathbf{T} \in \mathbb{Z}(\mathbf{T})$ since $\langle \mathbf{Z}, \mathbf{T} \rangle = 1 = \|\mathbf{T}\|_*$ and $\|\mathbf{Z}\|_\sigma = 1$.

Let $\mathbf{X}(t) = t(\mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1) \in \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$. It can be verified that $\|\mathbf{Z} + \mathbf{X}(t)\|_\sigma = 1$ if and only if $-1 \leq t \leq \frac{1}{2}$ (see Lemma A.3). Therefore, $\mathbf{Z} + \mathbf{X}(t) \in \partial\|\mathbf{T}\|_*$ for any $-1 \leq t \leq \frac{1}{2}$. However, it can also be verified that $\|\mathbf{X}(t)\|_\sigma = \frac{2|t|}{\sqrt{3}}$ (see Lemma A.3), in particular, $\|\mathbf{X}(-1)\|_\sigma = \frac{2}{\sqrt{3}} > 1$. Moreover, the allowed stretches of $\mathbf{X}(t)$ are different along two opposite directions, positive t making at most $\|\mathbf{X}(\frac{1}{2})\|_\sigma = \frac{1}{\sqrt{3}}$ while negative t leading to even $\|\mathbf{X}(-1)\|_\sigma = \frac{2}{\sqrt{3}}$.

The above examples make the subdifferential of the tensor nuclear norm much more complicated and interesting. From what we have observed, the basic subspace $\mathbb{T}^{[d]}(\mathbf{T})$ resembles the most to $\mathbb{T}^{\{1,2\}}(\mathbf{T})$ of the matrix space, evidenced by not only the full decomposability in Theorem 4.2 but also the inclusion of the subdifferential in Corollary 5.2 for $\mathbb{I} = [d]$. However, the above examples call attention that every $\mathbf{X} \in \bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$ should be taken care of. The rest of this subsection is focused on the size of $\|\mathbf{X}\|_\sigma$ in these subspaces.

Given a tensor $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ and a nonempty $\mathbb{J} \subseteq 2^{[d]} \setminus \emptyset$, it defines a subspace $\mathbb{X}^{\mathbb{J}}(\mathbf{T}) := \bigoplus_{\mathbb{I} \in \mathbb{J}} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$, a direct sum of $|\mathbb{J}|$ basic subspaces of $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ defined by $(\text{sp}_k(\mathbf{T}))_{k=1}^d$. It is also obvious that $\mathbb{X}^{\mathbb{J}}(\mathbf{T})$ is orthogonal to $\mathbb{T}(\mathbf{T}) = \mathbb{X}^{\{\emptyset\}}(\mathbf{T})$ since $\emptyset \notin \mathbb{J}$. Let us define

$$\begin{aligned} \underline{\tau}(\mathbb{X}^{\mathbb{J}}(\mathbf{T})) &:= \max\{t : \|\mathbf{Z} + \mathbf{X}\|_\sigma \leq 1 \forall \mathbf{Z} \in \mathbb{Z}(\mathbf{T}), \mathbf{X} \in \mathbb{X}^{\mathbb{J}}(\mathbf{T}), \|\mathbf{X}\|_\sigma \leq t\}, \\ \bar{\tau}(\mathbb{X}^{\mathbb{J}}(\mathbf{T})) &:= \max\{\|\mathbf{X}\|_\sigma : \mathbf{Z} \in \mathbb{Z}(\mathbf{T}), \mathbf{X} \in \mathbb{X}^{\mathbb{J}}(\mathbf{T}), \|\mathbf{Z} + \mathbf{X}\|_\sigma \leq 1\}, \end{aligned} \quad (21)$$

and further

$$\begin{aligned} \underline{\tau}(\mathbb{X}^{\mathbb{J}}) &:= \min\{\underline{\tau}(\mathbb{X}^{\mathbb{J}}(\mathbf{T})) : \mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}\}, \\ \bar{\tau}(\mathbb{X}^{\mathbb{J}}) &:= \max\{\bar{\tau}(\mathbb{X}^{\mathbb{J}}(\mathbf{T})) : \mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}\}. \end{aligned}$$

In other words, $\underline{\tau}(\mathbb{X}^{\mathbb{J}})$ is the tight lower bound of $\|\mathbf{X}\|_{\sigma}$ and $\bar{\tau}(\mathbb{X}^{\mathbb{J}})$ is the tight upper bound, in the sense that for any $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$,

$$\begin{aligned} \partial\|\mathbf{T}\|_* &\supseteq \{\mathbf{Z} + \mathbf{X} : \mathbf{Z} \in \mathbb{Z}(\mathbf{T}), \mathbf{X} \in \mathbb{X}^{\mathbb{J}}(\mathbf{T}), \|\mathbf{X}\|_{\sigma} \leq \underline{\tau}(\mathbb{X}^{\mathbb{J}})\}, \\ (\mathbb{T}(\mathbf{T}) \oplus \mathbb{X}^{\mathbb{J}}(\mathbf{T})) \cap \partial\|\mathbf{T}\|_* &\subseteq \{\mathbf{Z} + \mathbf{X} : \mathbf{Z} \in \mathbb{Z}(\mathbf{T}), \mathbf{X} \in \mathbb{X}^{\mathbb{J}}(\mathbf{T}), \|\mathbf{X}\|_{\sigma} \leq \bar{\tau}(\mathbb{X}^{\mathbb{J}})\}. \end{aligned}$$

As an example, the inclusion $\mathbb{D}_1(\mathbf{T}) \subseteq \partial\|\mathbf{T}\|_*$ in (14) implies that $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})\right) \geq \frac{1}{2}$ for any $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, or simply $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right) \geq \frac{1}{2}$. The inclusion $\mathbb{D}_2(\mathbf{T}) \subseteq \partial\|\mathbf{T}\|_*$ in (16) implies that $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \geq \frac{2}{d(d-1)}$. Although $\mathbb{D}(\mathbf{T})$ includes $\mathbb{D}_2(\mathbf{T})$ as a proper subset, it does not help to increase the lower bound of $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right)$. However, Corollary 5.2 and Theorem 4.7 actually imply that $\underline{\tau}(\mathbb{U}^{\mathbb{I}}) = \bar{\tau}(\mathbb{U}^{\mathbb{I}}) = 1$ for any $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| \geq 2$.

Some obvious monotonicity holds for the two bounds. If $\mathbb{J}_1 \subseteq \mathbb{J}_2$, then $\mathbb{X}^{\mathbb{J}_1} \subseteq \mathbb{X}^{\mathbb{J}_2}$ and so

$$\underline{\tau}(\mathbb{X}^{\mathbb{J}_1}) \geq \underline{\tau}(\mathbb{X}^{\mathbb{J}_2}) \text{ and } \bar{\tau}(\mathbb{X}^{\mathbb{J}_1}) \leq \bar{\tau}(\mathbb{X}^{\mathbb{J}_2}).$$

Moreover, for fixed $\mathbb{J}$, increasing the order d will decrease $\underline{\tau}(\mathbb{X}^{\mathbb{J}})$ but increase $\bar{\tau}(\mathbb{X}^{\mathbb{J}})$ in the weak sense, the same to increasing any dimension n_k while fixing the others.

We now present the main results on the two bounds for various tensor subspaces of interest.

Theorem 5.10 *In the tensor space $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ of order $d \geq 3$,*

- (i) $\underline{\tau}(\mathbb{T}^{\mathbb{I}}) = 0$ and $\bar{\tau}(\mathbb{T}^{\mathbb{I}}) = 1$ for any $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| = 1$;
- (ii) $\underline{\tau}(\mathbb{T}^{\mathbb{I}}) = \bar{\tau}(\mathbb{T}^{\mathbb{I}}) = 1$ for any $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| \geq 2$;
- (iii) $\underline{\tau}(\mathbb{U}^{\mathbb{I}}) = 0$ and $\bar{\tau}(\mathbb{U}^{\mathbb{I}}) = 1$ for any $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| = 1$;
- (iv) $\underline{\tau}(\mathbb{U}^{\mathbb{I}}) = \bar{\tau}(\mathbb{U}^{\mathbb{I}}) = 1$ for any $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| \geq 2$;
- (v) $\frac{1}{2} \leq \underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right) \leq \frac{1}{\sqrt{3}} \approx 0.577$ and $1.155 \approx \frac{2}{\sqrt{3}} \leq \bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right) \leq \frac{1+\sqrt{2}}{2} \approx 1.207$;
- (vi) $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right) = 0$ and $1.155 \approx \frac{2}{\sqrt{3}} \leq \bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right) \leq \frac{3}{2}$;
- (vii) $\frac{1}{3} \leq \underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}\right) \leq \frac{1}{2}$ and $1.207 \approx \frac{1+\sqrt{2}}{2} \leq \bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}\right) \leq \frac{1+\sqrt{3}}{2} \approx 1.366$;
- (viii) $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}\right) = 0$ and $1.207 \approx \frac{1+\sqrt{2}}{2} \leq \bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}\right) \leq \frac{8}{5}$;
- (ix) $\frac{2}{d(d-1)} \leq \underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \leq \frac{1}{2}$ and $1.207 \approx \frac{1+\sqrt{2}}{2} \leq \bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \leq 2$ for $d \geq 5$;
- (x) $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) = 0$ and $1.207 \approx \frac{1+\sqrt{2}}{2} \leq \bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \leq 2$ for $d \geq 5$.

Proof. The proof is given case by case. As a general notation, $\mathbf{T}$ is an arbitrary tensor or a specific example where $\underline{\tau}(\mathbb{X}^{\mathbb{J}}(\mathbf{T}))$ or $\bar{\tau}(\mathbb{X}^{\mathbb{J}}(\mathbf{T}))$ is concerned with and $\mathbf{Z} \in \mathbb{Z}(\mathbf{T})$.

- (i) When $d = 3$, the tensor $\mathbf{Z} + \mathbf{Y}(t) \in \mathbb{R}^{2 \times 2 \times 3}$ in Example 5.8 shows that $\underline{\tau}(\mathbb{T}^{\{3\}}) \leq 0$, implying that $\underline{\tau}(\mathbb{T}^{\{3\}}) = 0$ as it cannot be negative. By the monotonicity with respect to d and n_k , we have $\underline{\tau}(\mathbb{T}^{\mathbb{I}}) = 0$ for any $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| = 1$. Similarly, the tensor $\mathbf{Z} + \mathbf{X}(t) \in \mathbb{R}^{2 \times 2 \times 3}$ in Example 5.8 shows that $\bar{\tau}(\mathbb{T}^{\{3\}}) \geq 1$. By the monotonicity with respect to d and n_k , we have $\bar{\tau}(\mathbb{T}^{\mathbb{I}}) \geq 1$ for any $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| = 1$. By Lemma 3.2 and the requirement of a subgradient in (20), we have $\|\mathbf{X}\|_{\sigma} \leq \|\mathbf{Z} + \mathbf{X}\|_{\sigma} \leq 1$ for any $\mathbf{Z} \in \mathbb{Z}(\mathbf{T})$ and $\mathbf{X} \in \mathbb{T}^{\mathbb{I}}(\mathbf{T})$ with $\|\mathbf{Z} + \mathbf{X}\|_{\sigma} \leq 1$. This means that $\bar{\tau}(\mathbb{T}^{\mathbb{I}}) \leq 1$, implying that $\bar{\tau}(\mathbb{T}^{\mathbb{I}}) = 1$.

- (ii) Since $\mathbb{T}^{\mathbb{I}}(\mathbf{T}) \subseteq \mathbb{U}^{\mathbb{I}}(\mathbf{T})$ for any $\mathbf{T}$, the monotonicity with respect to $\mathbb{J}$ of $\mathbb{X}^{\mathbb{J}}$ implies that $\underline{\tau}(\mathbb{U}^{\mathbb{I}}) \leq \underline{\tau}(\mathbb{T}^{\mathbb{I}}) \leq \bar{\tau}(\mathbb{T}^{\mathbb{I}}) \leq \bar{\tau}(\mathbb{U}^{\mathbb{I}})$. The results then follow immediately from (iv).
- (iii) Since $\mathbb{T}^{\mathbb{I}}(\mathbf{T}) \subseteq \mathbb{U}^{\mathbb{I}}(\mathbf{T})$ for any $\mathbf{T}$, the monotonicity with respect to $\mathbb{J}$ of $\mathbb{X}^{\mathbb{J}}$ implies that $\underline{\tau}(\mathbb{U}^{\mathbb{I}}) \leq \underline{\tau}(\mathbb{T}^{\mathbb{I}}) = 0$ and $\bar{\tau}(\mathbb{U}^{\mathbb{I}}) \geq \bar{\tau}(\mathbb{T}^{\mathbb{I}}) = 1$ by (i). Therefore, $\underline{\tau}(\mathbb{U}^{\mathbb{I}}) = 0$ as it cannot be negative. On the other hand, by Lemma 3.2 and the requirement of a subgradient in (20), we have $\|\mathbf{X}\|_{\sigma} \leq \|\mathbf{Z} + \mathbf{X}\|_{\sigma} \leq 1$ for any $\mathbf{Z} \in \mathbb{Z}(\mathbf{T})$ and $\mathbf{X} \in \mathbb{U}^{\mathbb{I}}(\mathbf{T})$ with $\|\mathbf{Z} + \mathbf{X}\|_{\sigma} \leq 1$. This means that $\bar{\tau}(\mathbb{U}^{\mathbb{I}}) \leq 1$, implying that $\bar{\tau}(\mathbb{U}^{\mathbb{I}}) = 1$.
- (iv) This is an immediate consequence of Corollary 5.2 and Theorem 4.7.
- (v) $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right) \geq \frac{1}{2}$ is due to the inclusion $\mathbb{D}_1(\mathbf{T}) \subseteq \partial\|\mathbf{T}\|_*$; see also [61, Lemma 1]. $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right) \leq \frac{1}{\sqrt{3}}$ and $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right) \geq \frac{2}{\sqrt{3}}$ are due to $\mathbf{Z} + \mathbf{X}(\frac{1}{2})$ and $\mathbf{Z} + \mathbf{X}(-1)$ in Example 5.9, respectively. To upper bound $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right)$, we notice that for any $\mathbf{T}$,

$$\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T}) \subseteq \mathbb{T}^{\{2,3\}}(\mathbf{T}) \oplus \mathbb{U}^{\{1\}}(\mathbf{T}).$$

By the monotonicity and Lemma 5.11,

$$\begin{aligned} & \bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right) \\ & \leq \max\left\{\|\mathbf{X}_1 + \mathbf{X}_2\|_{\sigma} : \|\mathbf{Z} + \mathbf{X}_1 + \mathbf{X}_2\|_{\sigma} \leq 1, \mathbf{Z} \in \mathbb{T}(\mathbf{T}), \mathbf{X}_1 \in \mathbb{T}^{\{2,3\}}(\mathbf{T}), \mathbf{X}_2 \in \mathbb{U}^{\{1\}}(\mathbf{T})\right\} \\ & \leq \max\left\{x_1 y_2 z_2 + x_2 : x_1 y_2 z_2 + x_2 \leq 1 + x_1 y_1 z_1, \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{S}^2 \cap \mathbb{R}_+^2\right\} \\ & = \frac{1 + \sqrt{2}}{2}, \end{aligned}$$

where the last equality is computed in Lemma B.1.

- (vi) Since $\underline{\tau}(\mathbb{T}^{\mathbb{I}}) = 0$ for any $\mathbb{I} \subseteq [3]$ with $|\mathbb{I}| = 1$ from (i), $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right) = 0$ by the monotonicity with respect to $\mathbb{J}$ of $\mathbb{X}^{\mathbb{J}}$. $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right) \geq \frac{2}{\sqrt{3}}$ is due to the tensor $\mathbf{Z} + \mathbf{X}(-1)$ in Example 5.9. To upper bound $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right)$, we notice that for any $\mathbf{T}$,

$$\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T}) = \mathbb{T}^{\{3\}}(\mathbf{T}) \oplus (\mathbb{T}^{\{2\}}(\mathbf{T}) \oplus \mathbb{T}^{\{2,3\}}(\mathbf{T})) \oplus \mathbb{U}^{\{1\}}(\mathbf{T}).$$

By the monotonicity and a similar idea to Lemma 5.11,

$$\begin{aligned} & \bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right) \\ & \leq \max\left\{\|\mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3\|_{\sigma} : \|\mathbf{Z} + \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3\|_{\sigma} \leq 1, \mathbf{Z} \in \mathbb{T}(\mathbf{T}), \right. \\ & \quad \left. \mathbf{X}_1 \in \mathbb{T}^{\{3\}}(\mathbf{T}), \mathbf{X}_2 \in \mathbb{T}^{\{2\}}(\mathbf{T}) \oplus \mathbb{T}^{\{2,3\}}(\mathbf{T}), \mathbf{X}_3 \in \mathbb{U}^{\{1\}}(\mathbf{T})\right\} \\ & \leq \max\left\{x_1 y_1 z_2 + x_1 y_2 + x_2 : x_1 y_1 z_2 + x_1 y_2 + x_2 \leq 1 + x_1 y_1 z_1, \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{S}^2 \cap \mathbb{R}_+^2\right\} \\ & = \frac{3}{2}, \end{aligned}$$

where the last equality is computed in Lemma B.2.

- (vii) $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}\right) \geq \frac{1}{3}$ generalizes [61, Lemma 1] from $d = 3$ to $d = 4$ and its proof is given by Lemma C.2. $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}\right) \leq \frac{1}{2}$ and $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}\right) \geq \frac{1+\sqrt{2}}{2}$ are due to the tensors $\mathbf{Z} + \mathbf{X}(\frac{1}{3})$ and $\mathbf{Z} + \mathbf{X}(-\frac{1+\sqrt{2}}{3})$ in Example C.1, respectively. To upper bound $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}\right)$, we notice that for any $\mathbf{T}$,

$$\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}(\mathbf{T}) \subseteq \mathbb{T}^{\{3,4\}}(\mathbf{T}) \oplus (\mathbb{T}^{\{2\}} \oplus \mathbb{T}^{\{2,3\}} \oplus \mathbb{T}^{\{2,4\}} \oplus \mathbb{T}^{\{2,3,4\}})(\mathbf{T}) \oplus \mathbb{U}^{\{1\}}(\mathbf{T}).$$

By the monotonicity and a similar idea to Lemma 5.11,

$$\begin{aligned} & \bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}\right) \\ & \leq \max\left\{\|\mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3\|_{\sigma} : \begin{array}{l} \|\mathbf{Z} + \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3\|_{\sigma} \leq 1, \mathbf{Z} \in \mathbb{T}(\mathbf{T}), \mathbf{X}_1 \in \mathbb{T}^{\{3,4\}}(\mathbf{T}), \\ \mathbf{X}_2 \in (\mathbb{T}^{\{2\}} \oplus \mathbb{T}^{\{2,3\}} \oplus \mathbb{T}^{\{2,4\}} \oplus \mathbb{T}^{\{2,3,4\}})(\mathbf{T}), \mathbf{X}_3 \in \mathbb{U}^{\{1\}}(\mathbf{T}) \end{array}\right\} \\ & \leq \max\left\{x_1 y_1 z_2 w_2 + x_1 y_2 + x_2 : x_1 y_1 z_2 w_2 + x_1 y_2 + x_2 \leq 1 + x_1 y_1 z_1 w_1, \mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w} \in \mathbb{S}^2 \cap \mathbb{R}_+^2\right\} \\ & = \frac{1 + \sqrt{3}}{2}, \end{aligned}$$

where the last equality is obtained using a similar calculation to Lemma B.2.

- (viii) Since $\underline{\tau}(\mathbb{T}^{\mathbb{I}}) = 0$ for any $\mathbb{I} \subseteq [4]$ with $|\mathbb{I}| = 1$ from (i), $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}\right) = 0$ by the monotonicity with respect to $\mathbb{J}$ of $\mathbb{X}^{\mathbb{J}}$. $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}\right) \geq \frac{1+\sqrt{2}}{2}$ is due to the tensor $\mathbf{Z} + \mathbf{X}(-\frac{1+\sqrt{2}}{3})$ in Example C.1. To upper bound $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}\right)$, we notice that for any $\mathbf{T}$,

$$\begin{aligned} & \bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}(\mathbf{T}) \\ & = \mathbb{T}^{\{4\}}(\mathbf{T}) \oplus (\mathbb{T}^{\{3\}}(\mathbf{T}) \oplus \mathbb{T}^{\{3,4\}}(\mathbf{T})) \oplus (\mathbb{T}^{\{2\}} \oplus \mathbb{T}^{\{2,3\}} \oplus \mathbb{T}^{\{2,4\}} \oplus \mathbb{T}^{\{2,3,4\}})(\mathbf{T}) \oplus \mathbb{U}^{\{1\}}(\mathbf{T}). \end{aligned}$$

By the monotonicity and a similar idea to Lemma 5.11,

$$\begin{aligned} & \bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}\right) \\ & \leq \max\left\{\left\|\sum_{i=1}^4 \mathbf{X}_i\right\|_{\sigma} : \begin{array}{l} \|\mathbf{Z} + \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 + \mathbf{X}_4\|_{\sigma} \leq 1, \mathbf{Z} \in \mathbb{T}(\mathbf{T}), \mathbf{X}_2 \in \mathbb{T}^{\{3\}}(\mathbf{T}) \oplus \mathbb{T}^{\{3,4\}}(\mathbf{T}), \\ \mathbf{X}_1 \in \mathbb{T}^{\{4\}}(\mathbf{T}), \mathbf{X}_3 \in (\mathbb{T}^{\{2\}} \oplus \mathbb{T}^{\{2,3\}} \oplus \mathbb{T}^{\{2,4\}} \oplus \mathbb{T}^{\{2,3,4\}})(\mathbf{T}), \mathbf{X}_4 \in \mathbb{U}^{\{1\}}(\mathbf{T}) \end{array}\right\} \\ & \leq \max\left\{x_1 y_1 z_1 w_2 + x_1 y_1 z_2 + x_1 y_2 + x_2 : \begin{array}{l} x_1 y_1 z_1 w_2 + x_1 y_1 z_2 + x_1 y_2 + x_2 \leq 1 + x_1 y_1 z_1 w_1, \\ \mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w} \in \mathbb{S}^2 \cap \mathbb{R}_+^2 \end{array}\right\} \\ & = \frac{8}{5}, \end{aligned}$$

where the last equality is obtained using a similar calculation to Lemma B.2.

- (ix) $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \geq \frac{2}{d(d-1)}$ is due to the inclusion $\mathbb{D}_2(\mathbf{T}) \subseteq \partial\|\mathbf{T}\|_{*}$; see also [62, Theorem 1]. Both $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \leq \frac{1}{2}$ and $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \geq \frac{1+\sqrt{2}}{2}$ are due to (vii) and the monotonicity with respect to $\mathbb{J}$ of $\mathbb{X}^{\mathbb{J}}$ and d . Finally, $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \leq 2$ is a trivial bound since $\|\mathbf{X}\|_{\sigma} \leq \|\mathbf{X} + \mathbf{Z}\|_{\sigma} + \|\mathbf{Z}\|_{\sigma} \leq 2$ for any $\mathbf{Z} \in \mathbb{Z}(\mathbf{T})$ and $\mathbf{X} \in \mathbb{X}^{\mathbb{J}}(\mathbf{T})$ with $\|\mathbf{Z} + \mathbf{X}\|_{\sigma} \leq 1$ in (21).

- (x) Since $\tau(\mathbb{T}^{\mathbb{I}}) = 0$ for any $\mathbb{I} \subseteq [d]$ with $|\mathbb{I}| = 1$ from (i), $\tau\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) = 0$ by the monotonicity with respect to $\mathbb{J}$ of $\mathbb{X}^{\mathbb{J}}$. $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \geq \frac{1+\sqrt{2}}{2}$ is due to (viii) and the monotonicity with respect to $\mathbb{J}$ of $\mathbb{X}^{\mathbb{J}}$ and d . Finally, $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \leq 2$ is a trivial bound as in the proof of (ix).

The proof is complete. $\square$

The key idea to bound $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right)$ from above in (v) is to establish a link between a tensor optimization problem and a simple low-dimensional spherical optimization problem, i.e., Lemma 5.11. This link offers a tool of independent interest. While it is presented in a rather special structure, it can be easily extended to other similar structures, such as bounding $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right)$ in (vi), $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}\right)$ in (vii), and $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}\right)$ in (viii).

We remark that instead of using $\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T}) \subseteq \mathbb{T}^{\{2,3\}}(\mathbf{T}) \oplus \mathbb{U}^{\{1\}}(\mathbf{T})$ in the proof of (v) to bound $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right)$ from above, we may also use $\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T}) = \mathbb{T}^{\{1,3\}}(\mathbf{T}) \oplus \mathbb{T}^{\{2,3\}}(\mathbf{T}) \oplus \mathbb{U}^{\{1,2\}}(\mathbf{T})$ to obtain another upper bound via a similar argument. However, the bound is worse than that in (v), the same to other ways of partitioning $\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$. For the same reason, the ways of partitioning $\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$ in the proof of (vi), $\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$ in (vii), and $\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [4]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$ in (viii) are all the best.

Lemma 5.11 *If $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, then*

$$\begin{aligned} & \max \left\{ \|\mathbf{X}_1 + \mathbf{X}_2\|_{\sigma} : \|\mathbf{Z} + \mathbf{X}_1 + \mathbf{X}_2\|_{\sigma} \leq 1, \mathbf{Z} \in \mathbb{T}(\mathbf{T}), \mathbf{X}_1 \in \mathbb{T}^{\{2,3\}}(\mathbf{T}), \mathbf{X}_2 \in \mathbb{U}^{\{1\}}(\mathbf{T}) \right\} \\ & \leq \max \left\{ x_1 y_2 z_2 + x_2 : x_1 y_2 z_2 + x_2 \leq 1 + x_1 y_1 z_1, \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{S}^2 \cap \mathbb{R}_+^2 \right\}. \end{aligned}$$

Proof. Let us denote $(\mathbf{Z}, \mathbf{X}_1, \mathbf{X}_2)$ to be an optimal solution to the first optimization problem and let $\|\mathbf{X}_1 + \mathbf{X}_2\|_{\sigma} = \langle \mathbf{X}_1 + \mathbf{X}_2, \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \mathbf{v}_3 \rangle$ where $\mathbf{v}_k \in \mathbb{S}^{n_k}$ for $k \in [3]$. Let $\mathbb{V}_k = \text{sp}_k(\mathbf{T})$ for $k \in [3]$ and

$$\begin{aligned} a_1 &= \|\text{p}_{\mathbb{V}_1}(\mathbf{v}_1)\|_2, & b_1 &= \|\text{p}_{\mathbb{V}_2}(\mathbf{v}_2)\|_2, & c_1 &= \|\text{p}_{\mathbb{V}_3}(\mathbf{v}_3)\|_2, \\ a_2 &= \|\text{p}_{\mathbb{V}_1^{\perp}}(\mathbf{v}_1)\|_2, & b_2 &= \|\text{p}_{\mathbb{V}_2^{\perp}}(\mathbf{v}_2)\|_2, & c_2 &= \|\text{p}_{\mathbb{V}_3^{\perp}}(\mathbf{v}_3)\|_2. \end{aligned}$$

It is obvious that $\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2 = \|\mathbf{c}\|_2 = 1$ and $\mathbf{a}, \mathbf{b}, \mathbf{c} \geq \mathbf{0}$.

By applying Theorem 4.7 with $\mathbf{Z} + \mathbf{X}_1 \in \mathbb{U}_{\{1\}}(\mathbf{T})$ and $\mathbf{X}_2 \in \mathbb{U}^{\{1\}}(\mathbf{T})$, we have

$$\max \{ \|\mathbf{Z} + \mathbf{X}_1\|_{\sigma}, \|\mathbf{X}_2\|_{\sigma} \} = \|\mathbf{Z} + \mathbf{X}_1 + \mathbf{X}_2\|_{\sigma} \leq 1.$$

Again by applying Theorem 4.7 with $\mathbf{Z} \in \mathbb{U}_{\{2\}}(\mathbf{T})$ and $\mathbf{X}_1 \in \mathbb{U}^{\{2\}}(\mathbf{T})$, we also have

$$\max \{ \|\mathbf{Z}\|_{\sigma}, \|\mathbf{X}_1\|_{\sigma} \} = \|\mathbf{Z} + \mathbf{X}_1\|_{\sigma} \leq \max \{ \|\mathbf{Z} + \mathbf{X}_1\|_{\sigma}, \|\mathbf{X}_2\|_{\sigma} \} \leq 1.$$

As a result, $\|\mathbf{Z}\|_{\sigma}, \|\mathbf{X}_1\|_{\sigma}, \|\mathbf{X}_2\|_{\sigma} \leq 1$.

It follows from Lemma 3.2 that

$$\begin{aligned} \|\mathbf{X}_1 + \mathbf{X}_2\|_{\sigma} &= \langle \mathbf{X}_1, \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \mathbf{v}_3 \rangle + \langle \mathbf{X}_2, \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \mathbf{v}_3 \rangle \\ &= \langle \mathbf{X}_1, \text{p}_{\mathbb{V}_1}(\mathbf{v}_1) \otimes \text{p}_{\mathbb{V}_2^{\perp}}(\mathbf{v}_2) \otimes \text{p}_{\mathbb{V}_3^{\perp}}(\mathbf{v}_3) \rangle + \langle \mathbf{X}_2, \text{p}_{\mathbb{V}_1^{\perp}}(\mathbf{v}_1) \otimes \mathbf{v}_2 \otimes \mathbf{v}_3 \rangle \\ &\leq \|\mathbf{X}_1\|_{\sigma} \|\text{p}_{\mathbb{V}_1}(\mathbf{v}_1)\|_2 \|\text{p}_{\mathbb{V}_2^{\perp}}(\mathbf{v}_2)\|_2 \|\text{p}_{\mathbb{V}_3^{\perp}}(\mathbf{v}_3)\|_2 + \|\mathbf{X}_2\|_{\sigma} \|\text{p}_{\mathbb{V}_1^{\perp}}(\mathbf{v}_1)\|_2 \|\mathbf{v}_2\|_2 \|\mathbf{v}_3\|_2 \end{aligned}$$

$$\leq a_1 b_2 c_2 + a_2.$$

Moreover, by that

$$\langle \mathbf{Z}, \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \mathbf{v}_3 \rangle + \|\mathbf{X}_1 + \mathbf{X}_2\|_\sigma = \langle \mathbf{Z} + \mathbf{X}_1 + \mathbf{X}_2, \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \mathbf{v}_3 \rangle \leq \|\mathbf{Z} + \mathbf{X}_1 + \mathbf{X}_2\|_\sigma \leq 1$$

and applying Lemma 3.2 again, we also have

$$\|\mathbf{X}_1 + \mathbf{X}_2\|_\sigma \leq 1 - \langle \mathbf{Z}, \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \mathbf{v}_3 \rangle = 1 + \langle -\mathbf{Z}, \mathbf{p}_{\mathbb{V}_1}(\mathbf{v}_1) \otimes \mathbf{p}_{\mathbb{V}_2}(\mathbf{v}_2) \otimes \mathbf{p}_{\mathbb{V}_3}(\mathbf{v}_3) \rangle \leq 1 + a_1 b_1 c_1.$$

Let us now compare the two upper bounds of $\|\mathbf{X}_1 + \mathbf{X}_2\|_\sigma$, $a_1 b_2 c_2 + a_2$ and $1 + a_1 b_1 c_1$.

- (i) If $a_1 b_2 c_2 + a_2 > 1 + a_1 b_1 c_1$, consider $f(x) = a_1 b_2 \sqrt{1 - x^2} + a_2$ and $g(x) = 1 + a_1 b_1 x$ defined on $[0, 1]$. Since $f(x)$ is decreasing, $g(x)$ is increasing, $f(c_1) > g(c_1)$, and $f(1) = a_2 \leq 1 \leq g(1)$, there must exist some $d_1 \in [c_1, 1]$ such that $f(c_1) \geq f(d_1) = g(d_1) \geq g(c_1)$.
- (ii) If $a_1 b_2 c_2 + a_2 \leq 1 + a_1 b_1 c_1$, then by letting $d_1 = c_1$, we have $f(c_1) = f(d_1) \leq g(d_1) = g(c_1)$.

To summarize both cases, there always exists some $d_1 \in [c_1, 1]$, such that

$$\|\mathbf{X}_1 + \mathbf{X}_2\|_\sigma \leq \min\{f(c_1), g(c_1)\} \leq f(d_1) = a_1 b_2 d_2 + a_2 \leq g(d_1) = 1 + a_1 b_1 d_1, \quad (22)$$

where $d_2 = \sqrt{1 - d_1^2}$. Therefore, $(a_1, a_2, b_1, b_2, d_1, d_2)$ is feasible to the second optimization problem, which implies that $a_1 b_2 d_2 + a_2$ is no more than the optimal value of the second problem. This, together with the fact that the optimal value of the first optimization problem is no more than $a_1 b_2 d_2 + a_2$ by (22), directly shows the claimed result. $\square$

Let us take some time to digest the results in Theorem 5.10, focusing on the subspaces where $\mathbf{X}$ resides. For any basic subspace $\mathbb{T}^{\mathbb{I}}(\mathbf{T})$ with $|\mathbb{I}| \geq 2$, $\mathbf{Z} + \mathbf{X}$ is a subgradient if and only if $\|\mathbf{X}\|_\sigma \leq 1$, i.e., (ii). However, for a basic subspace $\mathbb{T}^{\mathbb{I}}(\mathbf{T})$ with $|\mathbb{I}| = 1$, nothing can be guaranteed and one has to check $\|\mathbf{Z} + \mathbf{X}\|_\sigma$ case by case (see Example 5.8), but any $\|\mathbf{X}\|_\sigma > 1$ definitely rules $\mathbf{Z} + \mathbf{X}$ out, i.e., (i). A direct sum of several basic subspaces can be a subspace that keeps the bounds perfect as long as the sum is a subset of $\mathbb{U}^{\mathbb{I}}(\mathbf{T})$ for some $|\mathbb{I}| = 2$, the largest subspace to be perfect, i.e., (iv). Although $\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}$ cannot keep the bounds perfect, it is the largest structure in the sense that every direction in $\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$ can make $\mathbf{Z} + \mathbf{X}$ a subgradient. This is perhaps the most interesting aspect of the subspace, in which the lower bound $\frac{2}{d(d-1)}$ provides an assurance of the spectral size of $\mathbf{X}$ albeit it can be conservative, i.e., (ix). For the two special cases $d = 3, 4$, we are able to nail both $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right)$ and $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right)$ down to a smaller range.

We remark in particular that $\frac{2}{d(d-1)}$, as the conservative lower bound of $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right)$ from $\mathbb{D}_2(\mathbf{T})$, was improved to $\frac{1}{2}$ in $\mathbb{D}_1(\mathbf{T})$, i.e., (v) for $d = 3$. It was also improved to $\frac{1}{3}$ in (vii) for $d = 4$, one of the most important results in Theorem 5.10. Both make us believe strongly that the $\frac{2}{d(d-1)}$ bound can be improved to $\frac{1}{d-1}$, i.e., the validity of (iii) in Conjecture 5.12 below.

We have not explored the cases for $d \geq 5$ in theory. However, we do have applied similar techniques to those used in Theorem 5.10 and then resorted to computer programs to find the optimal values of relevant problems. We list our findings as a conjecture below.

Conjecture 5.12 *In the tensor space $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ of order $d \geq 3$,*

- (i) $\underline{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right) = \frac{1}{\sqrt{3}}$ and $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right) = \frac{2}{\sqrt{3}}$;

- (ii) $\tau\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \leq \left(\frac{d-2}{d}\right)^{\frac{d-2}{2}}$ and in particular $\lim_{d \rightarrow \infty} \tau\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \leq \frac{1}{e}$;
- (iii) $\tau\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \geq \frac{1}{d-1}$;
- (iv) $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \leq \frac{2(d-2)}{d-1}$ for $d \geq 5$;
- (v) $\bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 1, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \leq \frac{2d}{d+1}$;
- (vi) $\lim_{d \rightarrow \infty} \bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}\right) \geq \bar{\tau}\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [200]} \mathbb{T}^{\mathbb{I}}\right) \geq 1.319$.

5.3 Final remark of the subdifferential

The original version of the subdifferential inclusion in [61, Lemma 1] states that

$$\bar{\mathbb{D}}_1(\mathbf{T}) := \left\{ \mathbf{Z} + \mathbf{p}_{\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})}(\mathbf{X}) : \mathbf{Z} \in \mathbb{Z}(\mathbf{T}), \mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}, \|\mathbf{X}\|_{\sigma} \leq \frac{1}{2} \right\} \subseteq \partial \|\mathbf{T}\|_*,$$

which is slightly different to the one that we structured

$$\mathbb{D}_1(\mathbf{T}) = \left\{ \mathbf{Z} + \mathbf{X} : \mathbf{Z} \in \mathbb{Z}(\mathbf{T}), \mathbf{X} \in \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T}), \|\mathbf{X}\|_{\sigma} \leq \frac{1}{2} \right\} \subseteq \partial \|\mathbf{T}\|_*.$$

They both restrict subgradients in the direct sum of $\mathbb{Z}(\mathbf{T})$ and $\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$. In order to restrict $\mathbf{X} \in \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$, $\mathbb{D}_1(\mathbf{T})$ simply asks $\|\mathbf{X}\|_{\sigma} \leq \frac{1}{2}$, whereas $\bar{\mathbb{D}}_1(\mathbf{T})$ asks $\|\mathbf{X} + \mathbf{Y}\|_{\sigma} \leq \frac{1}{2}$ for some $\mathbf{Y} \in \bigoplus_{|\mathbb{I}| \leq 1, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$, but $\mathbf{Y}$ has nothing to do with the subgradient $\mathbf{Z} + \mathbf{X}$ itself.

The study on the subdifferential in this section is presented in the structure of $\mathbb{D}_1(\mathbf{T})$, i.e., simply checking $\|\mathbf{X}\|_{\sigma}$, rather than checking $\min\{\|\mathbf{X} + \mathbf{Y}\|_{\sigma} : \mathbf{Y} \in \bigoplus_{|\mathbb{I}| \leq 1, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})\}$. For the perfect inclusion in Corollary 5.2, i.e.,

$$\mathbb{D}^{\mathbb{I}}(\mathbf{T}) = \{\mathbf{Z} + \mathbf{X} : \mathbf{Z} \in \mathbb{Z}(\mathbf{T}), \mathbf{X} \in \mathbb{U}^{\mathbb{I}}(\mathbf{T}), \|\mathbf{X}\|_{\sigma} \leq 1\} \subseteq \partial \|\mathbf{T}\|_* \text{ for any } \mathbb{I} \subseteq [d] \text{ with } |\mathbb{I}| \geq 2,$$

both structures lead to the same set. This is because $\mathbf{X} = \mathbf{p}_{\mathbb{U}^{\mathbb{I}}(\mathbf{T})}(\mathbf{X} + \mathbf{Y})$ and so $\|\mathbf{X}\|_{\sigma} \leq \|\mathbf{X} + \mathbf{Y}\|_{\sigma} \leq 1$ by Lemma 3.2. On the other hand, $\bar{\mathbb{D}}_1(\mathbf{T})$ includes $\mathbb{D}_1(\mathbf{T})$ as a proper subset since $\|\mathbf{X}\|_{\sigma} \leq \frac{1}{2}$ trivially implies that $\|\mathbf{X} + \mathbf{Y}\|_{\sigma} \leq \frac{1}{2}$ for $\mathbf{Y} = \mathbf{O}$ and the tensor $\mathbf{Z} + \mathbf{X}(\frac{1}{2})$ in Example 5.14 belongs to $\bar{\mathbb{D}}_1(\mathbf{T}) \setminus \mathbb{D}_1(\mathbf{T})$. Even though, both $\bar{\mathbb{D}}_1(\mathbf{T})$ and $\mathbb{D}_1(\mathbf{T})$ provide the lower bound $\frac{1}{2}$ of $\tau\left(\bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}\right)$ in Theorem 5.10. In fact, the constant $\frac{1}{2}$ in $\bar{\mathbb{D}}_1(\mathbf{T})$ turns out to be tight and this answers in the negative a comment raised in [61, Page 1039] to sharpen the constant.

Proposition 5.13 *The constant $\frac{1}{2}$ in $\bar{\mathbb{D}}_1(\mathbf{T})$ is tight.*

The proposition is an immediate consequence of the following example.

Example 5.14 *Let $\mathbf{T} = \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 \in \mathbb{R}^{2 \times 2 \times 2}$. We have $\text{sp}_k(\mathbf{T}) = \text{sp}(\mathbf{e}_1)$ for $k \in [3]$. The only tensor $\mathbf{Z} \in \mathbb{T}(\mathbf{T})$ that satisfies $\langle \mathbf{Z}, \mathbf{T} \rangle = \|\mathbf{T}\|_* = 1$ and $\|\mathbf{Z}\|_{\sigma} = 1$ is $\mathbf{Z} = \mathbf{T}$.*

Let $\mathbf{X}(t) = t(\mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1) \in \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$ and $\mathbf{Y}(t) = -t\mathbf{T} \in \bigoplus_{|\mathbb{I}| \leq 1, \mathbb{I} \subseteq [3]} \mathbb{T}^{\mathbb{I}}(\mathbf{T})$. It can be verified that $\|\mathbf{X}(t)\|_{\sigma} = \frac{2|t|}{\sqrt{3}}$ and $\|\mathbf{X}(t) + \mathbf{Y}(t)\|_{\sigma} = |t|$ for any $t \in \mathbb{R}$ (see Lemma A.3). However, it can also be verified that $\|\mathbf{Z} + \mathbf{X}(t)\|_{\sigma} \leq 1$ if and only if $-1 \leq t \leq \frac{1}{2}$ (see Lemma A.3), implying that $\mathbf{Z} + \mathbf{X}(t) \notin \partial \|\mathbf{T}\|_$ for any $t > \frac{1}{2}$.*

Perhaps the tightness of $\bar{\mathbb{D}}_1(\mathbf{T})$ provides another reason to consider the structure of $\mathbb{D}_1(\mathbf{T})$ as we strongly believe that the constant $\frac{1}{2}$ in $\mathbb{D}_1(\mathbf{T})$ can be improved to $\frac{1}{\sqrt{3}}$, i.e., (i) in Conjecture 5.12. We leave it to future works.

6 Tensor robust principal component analysis

In this section, we establish the statistical performance of the nuclear-norm-based tensor robust PCA as an immediate application of our theoretical developments. The main result that we apply is the new inclusion (19) in Theorem 5.3, more specifically, Corollary 5.2. It makes the study in the sequel as straightforward as that in the matrix case [11].

6.1 Model and main result

The tensor robust PCA aims to recover a low-rank ground-truth tensor $\mathbf{L} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \setminus \{\mathbf{O}\}$ that is superposed by a sparse corruption $\mathbf{S} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$. Specifically, it is to find the ground-truth $\mathbf{L}$ by solving the following convex optimization model

$$\min\{\|\mathbf{T}_1\|_* + \lambda\|\mathbf{T}_2\|_1 : \mathbf{T}_1 + \mathbf{T}_2 = \mathbf{L} + \mathbf{S}, \mathbf{T}_1, \mathbf{T}_2 \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}\}, \quad (23)$$

where $\lambda > 0$ is a balancing parameter. For a comprehensive introduction to the model, the readers are referred to [11] and a recent book [59]. Unfortunately, (23) is computationally intractable due to the NP-hardness to compute the tensor nuclear norm [20] albeit it is convex. However, understanding the statistical performance of (23) can still be of great importance, just as it is for relevant problems such as tensor completion [61] and tensor regression [50].

It should be noted that not every $\mathbf{L}$ is identifiable. For example, it is definitely impossible to recover a simultaneously low-rank and sparse $\mathbf{L}$ due to ambiguity. As a result, some standard assumptions on $\mathbf{L}$ and $\mathbf{S}$ are required.

Assumption 6.1 *The entries of $\mathbf{S}$ are independent random variables, each being zero with probability $1 - \rho$, positive with probability $\frac{\rho}{2}$, and negative with probability $\frac{\rho}{2}$, for a sufficiently small constant $\rho > 0$. There exist a constant $u_0 > 0$ and a sufficiently small constant $\theta_0 > 0$ such that*

$$\max_{k \in [d]} u_k \leq u_0, \quad r_0 := \max_{k \in [d]} r_k \leq \theta_0 \frac{(1 - \rho)n_1}{u_0 \ln^2 n_d}, \quad \text{and} \quad \min_{\mathbf{Z} \in \mathbb{Z}(\mathbf{L})} \|\mathbf{Z}\|_\infty \leq \sqrt{\frac{u_0 r_0}{n_1 n_d \ln^{\max\{2d-5, 0\}} n_d}},$$

where

$$r_k := \dim(\text{sp}_k(\mathbf{L})) \quad \text{and} \quad u_k := \frac{n_k}{r_k} \max_{i \in [n_k]} \|\mathbf{p}_{\text{sp}_k(\mathbf{L})}(\mathbf{e}_i)\|_2^2 \quad \text{for } k \in [d].$$

Recall $\mathbb{Z}(\mathbf{L}) = \{\mathbf{Z} \in \mathbb{T}(\mathbf{L}) : \langle \mathbf{Z}, \mathbf{L} \rangle = \|\mathbf{L}\|_*, \|\mathbf{Z}\|_\sigma = 1\}$ defined in (15) for $\mathbf{L} \neq \mathbf{O}$ and the assumption that $2 \leq n_1 \leq n_2 \leq \cdots \leq n_d$ without loss of generality. In Assumption 6.1, the first and last requirements of $\mathbf{L}$ are known as the incoherence conditions in the literature [11, 61, 19, 41]. The assumption on $\mathbf{S}$ means that its sparsity patterns follow Bernoulli distributions with parameter ρ but there is no assumption on the magnitudes of the entries of $\mathbf{S}$. We now state the main result in this section.

Theorem 6.2 *Under Assumption 6.1, the convex optimization model (23) with $\lambda = \frac{1}{\sqrt{n_d}}$ exactly recovers $\mathbf{L}$ and $\mathbf{S}$ with high probability for any fixed $d \geq 2$.*

An event with high probability is one whose probability depends on a certain number, which is n_d in our case, and tends to 1 as n_d tends to infinity, i.e., the probability of the event occurring can be made as close to 1 as desired. As mentioned in Theorem 6.2, the order of the tensor space, $d \geq 2$, is deemed as a fixed parameter.

To the best of our knowledge, Theorem 6.2 is the first result concerning the statistical behavior of (23) for tensors of an arbitrary order. Before proceeding to the proof, let us first compare our

result with several existing ones on tensor robust PCA based on nuclear norms. First of all, when $d = 2$, the special case of Theorem 6.2 exactly recovers the result on the matrix robust PCA; see [59, Theorem 5.3] and [11, Theorem 1.1].

There are many t-SVD-based methods in the literature that use the t-SVD [31] to define different tensor nuclear norms; see, e.g., [21, 41, 40]. These methods have a good computability and exhibit remarkable performance in numerical experiments but they only work for third-order tensors due to the inherent design mechanism of the t-SVD. This limitation has restricted the generality and versatility of the methods. By contrast, the model (23) is built on top of the vanilla tensor nuclear norm that applies to tensors of an arbitrary order. For third-order tensors, some of the conditions in these t-SVD-based methods seem a bit different from ours. Taking [41, Theorem 4.1] as an example, the balancing parameter λ therein is set as $\frac{1}{\sqrt{n_2 n_3}}$ while that in Theorem 6.2 is $\frac{1}{\sqrt{n_3}}$, which is consistent with the matrix case $\frac{1}{\sqrt{n_2}}$; see [11, Theorem 1.1]. As another example, in [41, Theorem 4.1], the counterpart of $\min_{\mathbf{Z} \in \mathbb{Z}(\mathbf{L})} \|\mathbf{Z}\|_\infty$ in its context needs to be bounded by $\sqrt{\frac{u_t r_t}{n_1 n_2 n_3^2}}$ where r_t is the tubal rank of $\mathbf{L}$ [31, Definition 4.4] while ours is $\sqrt{\frac{u_0 r_0}{n_1 n_3 \ln n_3}}$ in Assumption 6.1. The former seems more restrictive if r_t is regarded analogously to r_0 .

We are also aware of an unpublished work of tensor robust PCA based on the vanilla tensor nuclear norm [19]. We highlight that our result has some advantages. Similar to the t-SVD-based methods, the analysis in [19] also applies to third-order tensors only. Even for third-order tensors, some assumptions in [19] can be stricter than ours. For example, the requirements of $\min_{\mathbf{Z} \in \mathbb{Z}(\mathbf{L})} \|\mathbf{Z}\|_\infty$ in [19, Theorem 1] and in Theorem 6.2 are

$$\min_{\mathbf{Z} \in \mathbb{Z}(\mathbf{L})} \|\mathbf{Z}\|_\infty \leq \sqrt{\frac{u_d r_d}{n_1 n_2 n_3}} \text{ with } r_d = \sqrt{\frac{r_1 r_2 n_3 + r_1 r_3 n_2 + r_2 r_3 n_1}{n_1 + n_2 + n_3}} \text{ and } \min_{\mathbf{Z} \in \mathbb{Z}(\mathbf{L})} \|\mathbf{Z}\|_\infty \leq \sqrt{\frac{u_0 r_0}{n_1 n_3 \ln n_3}},$$

respectively. While it is difficult to compare the two upper bounds directly, in the case when $r_k = r$ and $n_k = n$ for $k \in [3]$, they become $O(\sqrt{\frac{r}{n^3}})$ and $O(\sqrt{\frac{r}{n^2 \ln n}})$, respectively. It shows that our bound is clearly better in this case. That being said, the rank requirements in [19, Theorem 1] and in Theorem 6.2 are

$$r_d \leq \theta_d \sqrt{\frac{n_1 n_2 n_3}{(n_1 + n_2 + n_3) \ln(n_1 n_2 n_3)}} \text{ and } r_0 \leq \theta_0 \frac{(1 - \rho) n_1}{u_0 \ln^2 n_3},$$

respectively. In the case when $n_k = n$ for $k \in [3]$, the two upper bounds become $O(\frac{n}{\sqrt{\ln n}})$ and $O(\frac{n}{\ln^2 n})$, respectively. It shows that our bound is slightly worse if r_d is regarded analogously to r_0 .

We remark that there are also a lot of works based on the so-called sum-of-nuclear-norms [39, 29], i.e., using $\sum_{k=1}^d \lambda_k \|\mathbf{T}_{(k)}\|_*$ as a tractable surrogate of $\|\mathbf{T}\|_*$. We do not compare Theorem 6.2 with these results since their assumptions involve quantities that are absent from our framework.

The rest of this section is devoted to the proof of Theorem 6.2. The overall framework of the proof is similar to that of the matrix robust PCA in [59, Chapter 5] but there are quite a lot of details to be dealt with for tensors.

6.2 Unique optimality

To prove Theorem 6.2, we start to characterize the conditions under which $(\mathbf{L}, \mathbf{S})$ is the unique optimal solution of (23). They are essentially the existence of a tensor in the relative interior of $\partial \|\mathbf{L}\|_* \cap \lambda \partial \|\mathbf{S}\|_1$.

To manipulate $\partial \|\mathbf{L}\|_*$, we resort to the inclusion in Corollary 5.2 with $\mathbb{I} = [d]$, i.e.,

$$\{\mathbf{Z} + \mathbf{X} : \mathbf{Z} \in \mathbb{Z}(\mathbf{L}), \mathbf{X} \in \mathbb{T}^{[d]}(\mathbf{L}), \|\mathbf{X}\|_\sigma \leq 1\} \subseteq \partial \|\mathbf{L}\|_*. \quad (24)$$

In order to alleviate the lengthy notation in derivations, let us denote $\mathbb{L} := \bigoplus_{|\mathbb{I}| \leq d-1, \mathbb{I} \subseteq [d]} \mathbb{T}^{\mathbb{I}}(\mathbf{L})$ as the direct sum of all basic subspaces defined by $\mathbf{L}$ except $\mathbb{T}^{[d]}(\mathbf{L})$, or equivalently $\mathbb{L}^\perp = \mathbb{T}^{[d]}(\mathbf{L})$ as the orthogonal complement.

To study $\partial\|\mathbf{S}\|_1$, we need to define subspaces based on the entries of a tensor. For any tensor $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, we denote $\mathbb{I}(\mathbf{T}) := \{\mathbf{i} \in \mathbb{I}^d : t_{\mathbf{i}} \neq 0\}$ to be the support set of $\mathbf{T}$, recalling that $\mathbb{I}^d = \{\mathbf{i} \in \mathbb{N}^d : i_k \in [n_k] \forall k \in [d]\}$. We use the notation $\mathbb{I}(\mathbf{S}) \sim \text{Bernoulli}(\rho)$ to denote the sampling process of $\mathbf{S}$ in Assumption 6.1, i.e., $\Pr\{\mathbf{i} \in \mathbb{I}(\mathbf{S})\} = \rho$ for every $\mathbf{i} \in \mathbb{I}^d$ independently. For any index set $\mathbb{I} \subseteq \mathbb{I}^d$, we also abuse $\mathbb{I}$ to denote a subspace of $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ under no ambiguity, i.e.,

$$\mathbb{I} = \text{sp}(\{\mathbf{e}_{i_1} \otimes \mathbf{e}_{i_2} \otimes \dots \otimes \mathbf{e}_{i_d} : \mathbf{i} \in \mathbb{I}\}).$$

Therefore, $\dim(\mathbb{I}) = |\mathbb{I}|$ where the former $\mathbb{I}$ is a subspace and the latter is an index set. This also makes $p_{\mathbb{I}}$ self-explanatory. For example, $p_{\emptyset}$ is the zero operator, $p_{\mathbb{I}^d}$ is the identity operator as $\mathbb{I}^d = \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, and $p_{\{\mathbf{i}\}} = p_{\text{sp}(\{\mathbf{e}_{i_1} \otimes \mathbf{e}_{i_2} \otimes \dots \otimes \mathbf{e}_{i_d}\})}$, i.e., zeroing out all but the $\mathbf{i}$ th entry. It is also obvious that $\mathbb{I}^\perp = \text{sp}(\{\mathbf{e}_{i_1} \otimes \mathbf{e}_{i_2} \otimes \dots \otimes \mathbf{e}_{i_d} : \mathbf{i} \in \mathbb{I}^d \setminus \mathbb{I}\})$. Let $\mathbf{E} := \text{sign}(\mathbf{S})$, which is a random tensor with entries being 0, 1, or -1 since $\mathbf{S}$ is random. It is easy to check (see also [5, Example 3.41]) that

$$\partial\|\mathbf{S}\|_1 = \{\mathbf{E} + \mathbf{F} : p_{\mathbb{I}(\mathbf{S})}(\mathbf{F}) = \mathbf{O}, \|\mathbf{F}\|_\infty \leq 1\}. \quad (25)$$

Lemma 6.3 *If $\mathbb{L} \cap \mathbb{I}(\mathbf{S}) = \{\mathbf{O}\}$, or equivalently $\|p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})}\| < 1$, then $(\mathbf{L}, \mathbf{S})$ is the unique optimal solution of (23) if there exists a $\mathbf{D} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ such that*

$$p_{\mathbb{L}}(\mathbf{D}) \in \mathbb{Z}(\mathbf{L}), \|p_{\mathbb{L}^\perp}(\mathbf{D})\|_\sigma < 1, p_{\mathbb{I}(\mathbf{S})}(\mathbf{D}) = \lambda \mathbf{E}, \text{ and } \|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{D})\|_\infty < \lambda.$$

Proof. Since any feasible solution of (23) can be written as $(\mathbf{L} + \mathbf{H}, \mathbf{S} - \mathbf{H})$ for some perturbation $\mathbf{H}$, it suffices to show that $\|\mathbf{L} + \mathbf{H}\|_* + \lambda\|\mathbf{S} - \mathbf{H}\|_1 > \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1$ for any $\mathbf{H} \neq \mathbf{O}$.

By Lemma 3.3, there exists an $\mathbf{X} \in \mathbb{L}^\perp$ with $\|\mathbf{X}\|_\sigma = 1$ such that $\langle p_{\mathbb{L}^\perp}(\mathbf{H}), \mathbf{X} \rangle = \|p_{\mathbb{L}^\perp}(\mathbf{H})\|_*$. As a result, $p_{\mathbb{L}}(\mathbf{D}) + \mathbf{X} \in \partial\|\mathbf{L}\|_*$ by (24). Let $\mathbf{F} \in \mathbb{I}^\perp(\mathbf{S})$ with $\|\mathbf{F}\|_\infty = 1$ such that $\langle p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{H}), \mathbf{F} \rangle = -\|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{H})\|_1$. It is obvious that $\mathbf{E} + \mathbf{F} \in \partial\|\mathbf{S}\|_1$ by (25). Thus, by the definition of subgradient,

$$\begin{aligned} & \|\mathbf{L} + \mathbf{H}\|_* + \lambda\|\mathbf{S} - \mathbf{H}\|_1 \\ & \geq \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 + \langle p_{\mathbb{L}}(\mathbf{D}) + \mathbf{X}, \mathbf{H} \rangle + \lambda\langle \mathbf{E} + \mathbf{F}, -\mathbf{H} \rangle \\ & = \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 + \langle \mathbf{X}, \mathbf{H} \rangle - \lambda\langle \mathbf{F}, \mathbf{H} \rangle + \langle p_{\mathbb{L}}(\mathbf{D}) - p_{\mathbb{I}(\mathbf{S})}(\mathbf{D}), \mathbf{H} \rangle \\ & = \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 + \|p_{\mathbb{L}^\perp}(\mathbf{H})\|_* + \lambda\|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{H})\|_1 + \langle p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{D}) - p_{\mathbb{L}^\perp}(\mathbf{D}), \mathbf{H} \rangle \\ & \geq \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 + (1 - \|p_{\mathbb{L}^\perp}(\mathbf{D})\|_\sigma) \|p_{\mathbb{L}^\perp}(\mathbf{H})\|_* + (\lambda - \|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{D})\|_\infty) \|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{H})\|_1 \\ & > \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 \end{aligned}$$

as long as $\max\{\|p_{\mathbb{L}^\perp}(\mathbf{H})\|_*, \|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{H})\|_1\} > 0$ since $\|p_{\mathbb{L}^\perp}(\mathbf{D})\|_\sigma < 1$ and $\|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{D})\|_\infty < \lambda$.

Suppose on the contrary that $\|p_{\mathbb{L}^\perp}(\mathbf{H})\|_* = \|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{H})\|_1 = 0$. This means that $\mathbf{H} \in \mathbb{L}$ and $\mathbf{H} \in \mathbb{I}(\mathbf{S})$, and so $\mathbf{H} \in \mathbb{L} \cap \mathbb{I}(\mathbf{S}) = \{\mathbf{O}\}$, a contradiction to $\mathbf{H} \neq \mathbf{O}$. $\square$

To guarantee the exact recovery by solving (23), it suffices to construct a dual certificate $\mathbf{D}$ that satisfies the conditions stated in Lemma 6.3. However, restricting $p_{\mathbb{L}}(\mathbf{D}) \in \mathbb{Z}(\mathbf{L})$ and $p_{\mathbb{I}(\mathbf{S})}(\mathbf{D}) = \lambda \mathbf{E}$ simultaneously can be demanding and in fact may not be possible. Thus, we need to relax these two conditions, from zero distance to $\frac{\lambda}{8}$ in terms of the Frobenius norm. As a compensation, the other two conditions need in turn to be restricted by shrinking their radii by half.

Lemma 6.4 *If $\|p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})}\| < \frac{1}{2}$ and $\lambda < 1$, then $(\mathbf{L}, \mathbf{S})$ is the unique optimal solution of (23) if there exists a $\mathbf{D} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ such that*

$$\min_{\mathbf{Z} \in \mathbb{Z}(\mathbf{L})} \|p_{\mathbb{L}}(\mathbf{D}) - \mathbf{Z}\|_2 \leq \frac{\lambda}{8}, \quad \|p_{\mathbb{L}^\perp}(\mathbf{D})\|_\sigma < \frac{1}{2}, \quad \|p_{\mathbb{I}(\mathbf{S})}(\mathbf{D}) - \lambda \mathbf{E}\|_2 \leq \frac{\lambda}{8}, \quad \text{and} \quad \|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{D})\|_\infty < \frac{\lambda}{2}.$$

Proof. We first derive a couple of technical bounds. For any $\mathbf{H} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, we have

$$\begin{aligned} \|p_{\mathbb{I}(\mathbf{S})}(\mathbf{H})\|_2 &\leq \|p_{\mathbb{I}(\mathbf{S})} p_{\mathbb{L}}(\mathbf{H})\|_2 + \|p_{\mathbb{I}(\mathbf{S})} p_{\mathbb{L}^\perp}(\mathbf{H})\|_2 \\ &\leq \frac{1}{2} \|\mathbf{H}\|_2 + \|p_{\mathbb{L}^\perp}(\mathbf{H})\|_2 \\ &\leq \frac{1}{2} (\|p_{\mathbb{I}(\mathbf{S})}(\mathbf{H})\|_2 + \|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{H})\|_2) + \|p_{\mathbb{L}^\perp}(\mathbf{H})\|_2, \end{aligned}$$

which implies that $\|p_{\mathbb{I}(\mathbf{S})}(\mathbf{H})\|_2 \leq \|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{H})\|_2 + 2\|p_{\mathbb{L}^\perp}(\mathbf{H})\|_2$. For the same reason, we also have $\|p_{\mathbb{L}}(\mathbf{H})\|_2 \leq \|p_{\mathbb{L}^\perp}(\mathbf{H})\|_2 + 2\|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{H})\|_2$.

Similar to the proof of Lemma 6.3, we have $\mathbf{Z} + \mathbf{X} \in \partial\|\mathbf{L}\|_*$, where $\mathbf{Z} \in \mathbb{Z}(\mathbf{L})$ is an optimal solution of $\min_{\mathbf{Z} \in \mathbb{Z}(\mathbf{L})} \|p_{\mathbb{L}}(\mathbf{D}) - \mathbf{Z}\|_2$ (the existence is guaranteed by Lemma 5.6) and $\mathbf{X} \in \mathbb{L}^\perp$ with $\|\mathbf{X}\|_\sigma = 1$ and $\langle p_{\mathbb{L}^\perp}(\mathbf{H}), \mathbf{X} \rangle = \|p_{\mathbb{L}^\perp}(\mathbf{H})\|_*$. We also have $\mathbf{E} + \mathbf{F} \in \partial\|\mathbf{S}\|_1$, where $\mathbf{F} \in \mathbb{I}^\perp(\mathbf{S})$ with $\|\mathbf{F}\|_\infty = 1$ and $\langle p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{H}), \mathbf{F} \rangle = -\|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{H})\|_1$. Therefore, for any $\mathbf{H} \neq \mathbf{O}$,

$$\begin{aligned} &\|\mathbf{L} + \mathbf{H}\|_* + \lambda \|\mathbf{S} - \mathbf{H}\|_1 \\ &\geq \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 + \langle \mathbf{Z} + \mathbf{X}, \mathbf{H} \rangle + \lambda \langle \mathbf{E} + \mathbf{F}, -\mathbf{H} \rangle \\ &= \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 + \langle \mathbf{X} - p_{\mathbb{L}^\perp}(\mathbf{D}) - \lambda \mathbf{F} + p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{D}) + (\mathbf{Z} - p_{\mathbb{L}}(\mathbf{D})) + (p_{\mathbb{I}(\mathbf{S})}(\mathbf{D}) - \lambda \mathbf{E}), \mathbf{H} \rangle \\ &\geq \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 + \left(1 - \frac{1}{2}\right) \|p_{\mathbb{L}^\perp}(\mathbf{H})\|_* + \left(\lambda - \frac{\lambda}{2}\right) \|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{H})\|_1 - \frac{\lambda}{8} \|p_{\mathbb{L}}(\mathbf{H})\|_2 - \frac{\lambda}{8} \|p_{\mathbb{I}(\mathbf{S})}(\mathbf{H})\|_2 \\ &\geq \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 + \frac{1}{2} \|p_{\mathbb{L}^\perp}(\mathbf{H})\|_* + \frac{\lambda}{2} \|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{H})\|_1 - \frac{3\lambda}{8} \|p_{\mathbb{L}^\perp}(\mathbf{H})\|_2 - \frac{3\lambda}{8} \|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{H})\|_2 \\ &\geq \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 + \left(\frac{1}{2} - \frac{3\lambda}{8}\right) \|p_{\mathbb{L}^\perp}(\mathbf{H})\|_* + \frac{\lambda}{8} \|p_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{H})\|_1, \end{aligned}$$

where the penultimate inequality is due to the technical bounds derived earlier and the last one is due to the trivial bounds among tensor norms (6). Since $\lambda < 1$, the desired result follows immediately by the condition $\|p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})}\| < \frac{1}{2} < 1$, i.e., $\mathbb{L} \cap \mathbb{I}(\mathbf{S}) = \{\mathbf{O}\}$. $\square$

The task of the remaining subsections is to study the concentration behavior of $\|p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})}\|$ and to construct a dual certificate $\mathbf{D}$ that satisfies the conditions in Lemma 6.4 with high probability. Using the idea in [11], we construct a low-rank part $\mathbf{D}_1$ and a sparse part $\mathbf{D}_2$ separately and then form the dual certificate $\mathbf{D} = \mathbf{D}_1 + \mathbf{D}_2$.

6.3 Concentration behavior of $\|p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})}\|$

In this subsection, we show that $\|p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})}\| \leq \delta$ holds with high probability for any $\delta \in (0, 1]$, a stronger result than that required in Lemma 6.4. As a necessary preparation, we first show that any standard basis of $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ is far away from the subspace $\mathbb{L}$ as a direct consequence of the incoherence conditions in Assumption 6.1.

Lemma 6.5 *If the condition $u_k = \frac{n_k}{r_k} \max_{i \in [n_k]} \|p_{\text{sp}_k(\mathbf{L})}(\mathbf{e}_i)\|_2^2 \leq u_0$ for $k \in [d]$ in Assumption 6.1 holds, then*

$$\max_{\mathbf{i} \in \mathbb{I}^d} \left\| p_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right) \right\|_2^2 \leq u_0 \sum_{k=1}^d \frac{r_k}{n_k}.$$

Proof. For any $\mathbf{i} \in \mathbb{I}^d$, we have

$$\begin{aligned}
& \left\| \mathbb{P}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right) \right\|_2^2 \\
&= \left\| \bigotimes_{k=1}^d \mathbf{e}_{i_k} - \bigotimes_{k=1}^d \mathbb{P}_{\text{sp}_k^\perp(\mathbf{L})}(\mathbf{e}_{i_k}) \right\|_2^2 \\
&= \left\| \sum_{\ell=1}^d \left(\left(\bigotimes_{k=1}^{\ell-1} \mathbb{P}_{\text{sp}_k^\perp(\mathbf{L})}(\mathbf{e}_{i_k}) \right) \otimes \left(\bigotimes_{k=\ell}^d \mathbf{e}_{i_k} \right) - \left(\bigotimes_{k=1}^{\ell} \mathbb{P}_{\text{sp}_k^\perp(\mathbf{L})}(\mathbf{e}_{i_k}) \right) \otimes \left(\bigotimes_{k=\ell+1}^d \mathbf{e}_{i_k} \right) \right) \right\|_2^2 \\
&= \sum_{\ell=1}^d \left\| \left(\bigotimes_{k=1}^{\ell-1} \mathbb{P}_{\text{sp}_k^\perp(\mathbf{L})}(\mathbf{e}_{i_k}) \right) \otimes (\mathbf{e}_{i_\ell} - \mathbb{P}_{\text{sp}_\ell^\perp(\mathbf{L})}(\mathbf{e}_{i_\ell})) \otimes \left(\bigotimes_{k=\ell+1}^d \mathbf{e}_{i_k} \right) \right\|_2^2 \\
&\leq \sum_{\ell=1}^d \left\| \mathbb{P}_{\text{sp}_\ell(\mathbf{L})}(\mathbf{e}_{i_\ell}) \right\|_2^2 \\
&\leq u_0 \sum_{\ell=1}^d \frac{r_\ell}{n_\ell},
\end{aligned}$$

where the last equality holds because all the rank-one tensors there are mutually orthogonal and the last inequality follows from the condition on u_k . $\square$

The most important result in this part is to bound the tail probability of $\|\mathbb{P}_{\mathbb{L}}(\mathbb{P}_{\mathbb{I}^d} - q^{-1} \mathbb{P}_{\mathbb{I}}) \mathbb{P}_{\mathbb{L}}\|$, i.e., Proposition 6.6. To achieve this, we treat a projection onto a subspace in $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, which is itself a linear operator, as a $\prod_{k=1}^d n_k$ by $\prod_{k=1}^d n_k$ matrix, and then apply the matrix Bernstein inequality [55, Theorem 1.4]: If $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m \in \mathbb{R}^{n \times n}$ are independent and self-adjoint random matrices such that $\text{Ex } \mathbf{X}_i = \mathbf{O}$ and $\text{emax}(\mathbf{X}_i) \leq s$ almost surely for any $i \in [m]$, then for any $t \geq 0$,

$$\Pr \left\{ \text{emax} \left(\sum_{i=1}^m \mathbf{X}_i \right) \geq t \right\} \leq n \exp \left(\frac{-3t^2}{6 \left\| \sum_{i=1}^m \text{Ex } \mathbf{X}_i^2 \right\|_\sigma + 2st} \right), \quad (26)$$

where $\text{emax}(\mathbf{X})$ denotes the largest eigenvalue of $\mathbf{X}$. In particular, $\text{emax}(\mathbf{X}) = \|\mathbf{X}\|_\sigma$ if $\mathbf{X}$ is positive semidefinite.

Proposition 6.6 *Under Assumption 6.1, if $\mathbb{I} \sim \text{Bernoulli}(q)$ and $t > 0$, then*

$$\Pr \{ \|\mathbb{P}_{\mathbb{L}}(\mathbb{P}_{\mathbb{I}^d} - q^{-1} \mathbb{P}_{\mathbb{I}}) \mathbb{P}_{\mathbb{L}}\| \geq t \} \leq \exp \left(\frac{-3t^2 q}{u_0(6 + 2t) \sum_{k=1}^d \frac{r_k}{n_k}} \right) \prod_{k=1}^d n_k.$$

Proof. We first notice that

$$\mathbb{P}_{\mathbb{L}}(\mathbb{P}_{\mathbb{I}^d} - q^{-1} \mathbb{P}_{\mathbb{I}}) \mathbb{P}_{\mathbb{L}} = \sum_{\mathbf{i} \in \mathbb{I}^d} \mathbb{P}_{\mathbb{L}}(\mathbb{P}_{\mathbf{i}} - q^{-1} \mathbb{P}_{\mathbb{I}}) \mathbb{P}_{\mathbb{L}},$$

which is a sum of independent, zero-mean, and positive semidefinite random linear operators as

$$\mathbb{P}_{\mathbb{L}}(\mathbb{P}_{\mathbf{i}} - q^{-1} \mathbb{P}_{\mathbb{I}}) \mathbb{P}_{\mathbb{L}} = \begin{cases} (1 - q^{-1}) \mathbb{P}_{\mathbb{L}} \mathbb{P}_{\mathbf{i}} \mathbb{P}_{\mathbb{L}} & \mathbf{i} \in \mathbb{I} \\ \mathbb{P}_{\mathbb{L}} \mathbb{P}_{\mathbf{i}} \mathbb{P}_{\mathbb{L}} & \mathbf{i} \notin \mathbb{I}. \end{cases} \quad (27)$$

In order to apply the matrix Bernstein inequality to bound its tail probability, we need to control both $\|\mathbf{p}_{\mathbb{L}}(\mathbf{p}_{\mathbf{i}} - q^{-1} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}) \mathbf{p}_{\mathbb{L}}\|$ and $\|\sum_{\mathbf{i} \in \mathbb{I}^d} \text{Ex}(\mathbf{p}_{\mathbb{L}}(\mathbf{p}_{\mathbf{i}} - q^{-1} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}) \mathbf{p}_{\mathbb{L}})^2\|$. In fact, for any $\mathbf{i} \in \mathbb{I}^d$,

$$\begin{aligned}
\|\mathbf{p}_{\mathbb{L}}(\mathbf{p}_{\mathbf{i}} - q^{-1} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}) \mathbf{p}_{\mathbb{L}}\| &\leq \max\{|1 - q^{-1}|, 1\} \|\mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}\| \\
&\leq q^{-1} \max_{\|\mathbf{X}\|_2 \leq 1} \langle \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}(\mathbf{X}), \mathbf{X} \rangle \\
&= q^{-1} \max_{\|\mathbf{X}\|_2 \leq 1} \langle \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}(\mathbf{X}), \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}(\mathbf{X}) \rangle \\
&= q^{-1} \max_{\|\mathbf{X}\|_2 \leq 1} \left\langle \bigotimes_{k=1}^d \mathbf{e}_{i_k}, \mathbf{p}_{\mathbb{L}}(\mathbf{X}) \right\rangle^2 \\
&= q^{-1} \left\| \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right) \right\|_2^2 \\
&\leq q^{-1} u_0 \sum_{k=1}^d \frac{r_k}{n_k},
\end{aligned}$$

where the last inequality follows from Lemma 6.5. On the other hand, since $\text{Ex}(\mathbf{p}_{\mathbb{L}}(q^{-1} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}) \mathbf{p}_{\mathbb{L}}) = \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}$ and the variance of $\mathbf{p}_{\mathbb{L}}(q^{-1} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}) \mathbf{p}_{\mathbb{L}}$ is no more than its second moment, we have

$$\begin{aligned}
\left\| \sum_{\mathbf{i} \in \mathbb{I}^d} \text{Ex}(\mathbf{p}_{\mathbb{L}}(\mathbf{p}_{\mathbf{i}} - q^{-1} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}) \mathbf{p}_{\mathbb{L}})^2 \right\| &\leq \left\| \sum_{\mathbf{i} \in \mathbb{I}^d} \text{Ex}(\mathbf{p}_{\mathbb{L}}(q^{-1} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}) \mathbf{p}_{\mathbb{L}})^2 \right\| \\
&= q^{-2} \left\| \sum_{\mathbf{i} \in \mathbb{I}^d} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}} \Pr\{\mathbf{i} \in \mathbb{I}\} \right\| \\
&= q^{-1} \left\| \sum_{\mathbf{i} \in \mathbb{I}^d} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}} \right\| \\
&= q^{-1} \max_{\|\mathbf{X}\|_2 \leq 1} \sum_{\mathbf{i} \in \mathbb{I}^d} \langle \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}(\mathbf{X}), \mathbf{X} \rangle \\
&= q^{-1} \max_{\|\mathbf{X}\|_2 \leq 1} \sum_{\mathbf{i} \in \mathbb{I}^d} \langle \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}(\mathbf{X}), \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}(\mathbf{X}) \rangle \\
&= q^{-1} \max_{\|\mathbf{X}\|_2 \leq 1} \sum_{\mathbf{i} \in \mathbb{I}^d} \left\langle \bigotimes_{k=1}^d \mathbf{e}_{i_k}, \mathbf{p}_{\mathbb{L}}(\mathbf{X}) \right\rangle^2 \left\| \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right) \right\|_2^2 \\
&\leq q^{-1} \max_{\mathbf{i} \in \mathbb{I}^d} \left\| \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right) \right\|_2^2 \max_{\|\mathbf{X}\|_2 \leq 1} \sum_{\mathbf{i} \in \mathbb{I}^d} \left\langle \bigotimes_{k=1}^d \mathbf{e}_{i_k}, \mathbf{p}_{\mathbb{L}}(\mathbf{X}) \right\rangle^2 \\
&\leq q^{-1} u_0 \sum_{k=1}^d \frac{r_k}{n_k},
\end{aligned}$$

where the last inequality is due to Lemma 6.5 and that

$$\sum_{\mathbf{i} \in \mathbb{I}^d} \left\langle \bigotimes_{k=1}^d \mathbf{e}_{i_k}, \mathbf{p}_{\mathbb{L}}(\mathbf{X}) \right\rangle^2 = \|\mathbf{p}_{\mathbb{L}}(\mathbf{X})\|_2^2.$$

The desired inequality then follows immediately from the matrix Bernstein inequality (26). $\square$

Proposition 6.6 immediately implies the following two results and shows that $\|P_{\mathbb{L}} P_{\mathbb{I}(\mathbf{S})}\| \leq \delta$ holds with high probability for any $\delta \in (0, 1]$.

Corollary 6.7 *Under Assumption 6.1, for any $\epsilon \in (0, 1]$, there exists a $\kappa_0 > 0$ depending on ϵ only, such that $\|P_{\mathbb{L}} P_{\mathbb{I}^\perp}\| \leq \sqrt{1 - q + q\epsilon}$ holds with high probability as long as $q \geq \kappa_0 \frac{d^2 u_0 r_0 \ln n_d}{n_1}$ and $\mathbb{I} \sim \text{Bernoulli}(q)$.*

Proof. By that

$$q \geq \kappa_0 \frac{d^2 u_0 r_0 \ln n_d}{n_1} \geq \kappa_0 u_0 \left(\sum_{k=1}^d \frac{r_k}{n_k} \right) \sum_{k=1}^d \ln n_k$$

and letting $t = \epsilon$ in Proposition 6.6, we have

$$\|P_{\mathbb{L}} P_{\mathbb{I}^\perp} P_{\mathbb{L}}\| = \|P_{\mathbb{L}} - P_{\mathbb{L}} P_{\mathbb{I}} P_{\mathbb{L}}\| \leq \|P_{\mathbb{L}} - q P_{\mathbb{L}}\| + q \|P_{\mathbb{L}}(P_{\mathbb{I}^d} - q^{-1} P_{\mathbb{I}}) P_{\mathbb{L}}\| \leq 1 - q + q\epsilon$$

holds with probability at least

$$\begin{aligned} 1 - \exp\left(\frac{-3\epsilon^2 q}{u_0(6+2\epsilon) \sum_{k=1}^d \frac{r_k}{n_k}}\right) \prod_{k=1}^d n_k &\geq 1 - \exp\left(\frac{-3\kappa_0 \epsilon^2 \sum_{k=1}^d \ln n_k}{6+2\epsilon}\right) \prod_{k=1}^d n_k \\ &= 1 - \left(\prod_{k=1}^d n_k\right)^{1 - \frac{3\kappa_0 \epsilon^2}{6+2\epsilon}}, \end{aligned}$$

a high probability for a large enough κ_0 . Notice that $q \geq \kappa_0 \frac{d^2 u_0 r_0 \ln n_d}{n_1}$ is a probability and so we must ensure that $\kappa_0 \frac{d^2 u_0 r_0 \ln n_d}{n_1} < 1$. This can always be guaranteed in high dimensions since we have $r_0 \leq \theta_0 \frac{(1-\rho)n_1}{u_0 \ln^2 n_d}$ in Assumption 6.1 and so $\frac{\kappa_0 d^2 u_0 r_0 \ln n_d}{n_1} \leq \frac{\kappa_0 d^2 \theta_0 (1-\rho)}{\ln n_d} \rightarrow 0$ as $n_d \rightarrow \infty$. Therefore,

$$\|P_{\mathbb{L}} P_{\mathbb{I}^\perp}\| = \sqrt{\|P_{\mathbb{L}} P_{\mathbb{I}^\perp} P_{\mathbb{L}}\|} = \sqrt{\|P_{\mathbb{L}} P_{\mathbb{I}^\perp} P_{\mathbb{L}}\|} \leq \sqrt{1 - q + q\epsilon}$$

holds with high probability. $\square$

Corollary 6.8 *For any $\delta \in (0, 1]$, there exists a $\theta > 0$ depending on δ only, such that $\|P_{\mathbb{L}} P_{\mathbb{I}(\mathbf{S})}\| \leq \delta$ holds with high probability as long as Assumption 6.1 holds for $\theta_0 \leq \theta$ and $\rho \leq \frac{\delta^2}{2-\delta^2}$.*

Proof. Since $\mathbb{I}(\mathbf{S}) \sim \text{Bernoulli}(\rho)$ by Assumption 6.1, we have $\mathbb{I}^\perp(\mathbf{S}) \sim \text{Bernoulli}(1 - \rho)$. Let κ_0 be the constant associated with $\epsilon = \frac{\delta^2}{2}$ in Corollary 6.7 and further let $\theta = \frac{\ln 2}{\kappa_0 d^2}$. Since $r_0 \leq \theta_0 \frac{(1-\rho)n_1}{u_0 \ln^2 n_d}$ in Assumption 6.1 and $\theta_0 \leq \theta \leq \frac{\ln n_d}{\kappa_0 d^2}$, we have $1 - \rho \geq \frac{u_0 r_0 \ln^2 n_d}{\theta_0 n_1} \geq \kappa_0 \frac{d^2 u_0 r_0 \ln n_d}{n_1}$, as required by Corollary 6.7. Therefore, if we let $q = 1 - \rho$ and $\epsilon = \frac{\delta^2}{2}$ in Corollary 6.7, then

$$\|P_{\mathbb{L}} P_{\mathbb{I}(\mathbf{S})}\| \leq \sqrt{1 - q + q\epsilon} = \sqrt{\rho + \frac{\delta^2(1-\rho)}{2}} \leq \sqrt{\frac{\delta^2}{2-\delta^2} \left(1 - \frac{\delta^2}{2}\right) + \frac{\delta^2}{2}} = \delta$$

holds with high probability. $\square$

6.4 Dual certificate: Low-rank part via the golfing scheme

In this subsection, we construct the low-rank part of the dual certificate via the golfing scheme [22, 11]; let us start with a couple of probability bounds before introducing the scheme. The first one is a direct consequence of [64, Theorem 2.1]: If d th order tensors $\mathbf{X} \in \mathbb{R}^{n \times n \times \dots \times n}$ and $\mathbf{Y} \in \{0, 1\}^{n \times n \times \dots \times n}$ with $\mathbf{X}$ given and $\mathbb{I}(\mathbf{Y}) \sim \text{Bernoulli}(q)$ for $q \geq \theta_1 \frac{\ln n}{n}$, then

$$\Pr\left\{\|\text{Ex}(\mathbf{X} \odot \mathbf{Y}) - \mathbf{X} \odot \mathbf{Y}\|_\sigma \leq \kappa_1 \|\mathbf{X}\|_\infty \sqrt{qn} \ln^{d-2} n\right\} \geq 1 - n^{-\kappa_2}, \quad (28)$$

where $\odot$ is the Hadamard product.

Corollary 6.9 *If $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ and $\mathbb{I} \sim \text{Bernoulli}(q)$ with $q \geq \theta_1 \frac{\ln n_d}{n_d}$, then*

$$\Pr\left\{\|(\mathbf{p}_{\mathbb{I}^d} - q^{-1} \mathbf{p}_{\mathbb{I}})(\mathbf{X})\|_\sigma \leq \kappa_1 \|\mathbf{X}\|_\infty \sqrt{q^{-1} n_d} \ln^{d-2} n_d\right\} \geq 1 - n_d^{-\kappa_2}.$$

Although (28) requires that the tensor space to be $n \times n \times \dots \times n$, we can embed any tensor in $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ into the space $\mathbb{R}^{n_d \times n_d \times \dots \times n_d}$ by appending zero entries. Obviously the embedding makes no changes to the spectral norm and the ℓ_∞ -norm. By inspecting the tensor $\mathbf{X} \odot \mathbf{Y}$ entrywisely, it is easy to see that

$$(\mathbf{p}_{\mathbb{I}^d} - q^{-1} \mathbf{p}_{\mathbb{I}})(\mathbf{X}) = q^{-1}(\text{Ex}(\mathbf{X} \odot \mathbf{Y}) - \mathbf{X} \odot \mathbf{Y}) \text{ if } \mathbf{Y} \in \{0, 1\}^{n \times n \times \dots \times n} \text{ and } \mathbb{I}(\mathbf{Y}) \sim \text{Bernoulli}(q).$$

Corollary 6.9 then follows immediately from (28).

We are ready to present the main probability bound in this subsection.

Lemma 6.10 *Under Assumption 6.1, if $\mathbf{X} \in \mathbb{L} \setminus \{\mathbf{O}\}$ and $\mathbb{I} \sim \text{Bernoulli}(q)$, then*

$$\Pr\left\{\|\mathbf{p}_{\mathbb{L}}(\mathbf{p}_{\mathbb{I}^d} - q^{-1} \mathbf{p}_{\mathbb{I}}) \mathbf{p}_{\mathbb{L}}(\mathbf{X})\|_\infty \geq t\right\} \leq 2 \exp\left(\frac{-3t^2 q}{u_0 \|\mathbf{X}\|_\infty (6\|\mathbf{X}\|_\infty + 2t) \sum_{k=1}^d \frac{r_k}{n_k}}\right) \prod_{k=1}^d n_k.$$

In particular, by letting $t = \frac{\|\mathbf{X}\|_\infty}{2}$ and $q \geq \kappa_3 \frac{d^2 u_0 r_0 \ln n_d}{n_1} \geq \kappa_3 u_0 \left(\sum_{k=1}^d \frac{r_k}{n_k}\right) \sum_{k=1}^d \ln n_k$,

$$\|\mathbf{p}_{\mathbb{L}}(\mathbf{p}_{\mathbb{I}^d} - q^{-1} \mathbf{p}_{\mathbb{I}}) \mathbf{p}_{\mathbb{L}}(\mathbf{X})\|_\infty \leq \frac{\|\mathbf{X}\|_\infty}{2}$$

holds with high probability.

Proof. In order to show the bound, we apply the scalar Bernstein inequality to $\mathbf{p}_{\mathbb{L}}(\mathbf{p}_{\mathbb{I}^d} - q^{-1} \mathbf{p}_{\mathbb{I}}) \mathbf{p}_{\mathbb{L}}(\mathbf{X})$ entrywisely. Since $\mathbf{X} \in \mathbb{L}$ and $\mathbf{X} = \sum_{\mathbf{i} \in \mathbb{I}^d} x_{\mathbf{i}} \bigotimes_{k=1}^d \mathbf{e}_{i_k}$, we have

$$\mathbf{p}_{\mathbb{L}}(\mathbf{p}_{\mathbb{I}^d} - q^{-1} \mathbf{p}_{\mathbb{I}}) \mathbf{p}_{\mathbb{L}}(\mathbf{X}) = \sum_{\mathbf{i} \in \mathbb{I}^d} \mathbf{p}_{\mathbb{L}}(\mathbf{p}_{\mathbf{i}} - q^{-1} \mathbf{p}_{\mathbb{I}} \mathbf{p}_{\mathbb{I}})(\mathbf{X}) = \sum_{\mathbf{i} \in \mathbb{I}^d} x_{\mathbf{i}} \mathbf{p}_{\mathbb{L}}(\mathbf{p}_{\mathbf{i}} - q^{-1} \mathbf{p}_{\mathbb{I}} \mathbf{p}_{\mathbb{I}}) \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k}\right).$$

As a result, for any $\mathbf{j} \in \mathbb{I}^d$,

$$(\mathbf{p}_{\mathbb{L}}(\mathbf{p}_{\mathbb{I}^d} - q^{-1} \mathbf{p}_{\mathbb{I}}) \mathbf{p}_{\mathbb{L}}(\mathbf{X}))_{\mathbf{j}} = \sum_{\mathbf{i} \in \mathbb{I}^d} x_{\mathbf{i}} \left\langle \mathbf{p}_{\mathbb{L}}(\mathbf{p}_{\mathbf{i}} - q^{-1} \mathbf{p}_{\mathbb{I}} \mathbf{p}_{\mathbb{I}}) \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k}\right), \bigotimes_{k=1}^d \mathbf{e}_{j_k} \right\rangle =: \sum_{\mathbf{i} \in \mathbb{I}^d} w_{\mathbf{i}}.$$

We first notice that $\text{Ex } w_{\mathbf{i}} = 0$ as $\text{Ex } \mathbf{p}_{\mathbb{I}} = q \mathbf{p}_{\mathbb{I}^d}$ and

$$\begin{aligned}
|w_{\mathbf{i}}| &\leq |x_{\mathbf{i}}| \max\{1, |1 - q^{-1}|\} \left| \left\langle \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right), \bigotimes_{k=1}^d \mathbf{e}_{j_k} \right\rangle \right| \\
&\leq q^{-1} \|\mathbf{X}\|_{\infty} \left| \left\langle \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right), \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{j_k} \right) \right\rangle \right| \\
&\leq q^{-1} \|\mathbf{X}\|_{\infty} \left\| \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right) \right\|_2 \left\| \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{j_k} \right) \right\|_2 \\
&\leq q^{-1} u_0 \|\mathbf{X}\|_{\infty} \sum_{k=1}^d \frac{r_k}{n_k},
\end{aligned}$$

where the last inequality follows from Lemma 6.5. On the other hand, as the variance is no more than the second moment, we have

$$\begin{aligned}
\sum_{\mathbf{i} \in \mathbb{I}^d} \text{Ex } w_{\mathbf{i}}^2 &\leq \sum_{\mathbf{i} \in \mathbb{I}^d} \text{Ex} \left(q^{-1} x_{\mathbf{i}} \left\langle \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{I}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right), \bigotimes_{k=1}^d \mathbf{e}_{j_k} \right\rangle \right)^2 \\
&= \sum_{\mathbf{i} \in \mathbb{I}^d} \left(q^{-1} x_{\mathbf{i}} \left\langle \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right), \bigotimes_{k=1}^d \mathbf{e}_{j_k} \right\rangle \right)^2 \Pr\{\mathbf{i} \in \mathbb{I}\} \\
&\leq q^{-1} \|\mathbf{X}\|_{\infty}^2 \sum_{\mathbf{i} \in \mathbb{I}^d} \left\langle \bigotimes_{k=1}^d \mathbf{e}_{i_k}, \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{j_k} \right) \right\rangle^2 \\
&= q^{-1} \|\mathbf{X}\|_{\infty}^2 \left\| \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{j_k} \right) \right\|_2^2 \\
&\leq q^{-1} u_0 \|\mathbf{X}\|_{\infty}^2 \sum_{k=1}^d \frac{r_k}{n_k}.
\end{aligned}$$

By applying the scalar Bernstein inequality [56, Theorem 2.8.4], we have

$$\Pr\left\{ \left| (\mathbf{p}_{\mathbb{L}}(\mathbf{p}_{\mathbb{I}^d} - q^{-1} \mathbf{p}_{\mathbb{I}}) \mathbf{p}_{\mathbb{L}}(\mathbf{X}))_{\mathbf{j}} \right| \geq t \right\} \leq 2 \exp \left(\frac{-3t^2 q}{u_0 \|\mathbf{X}\|_{\infty} (6\|\mathbf{X}\|_{\infty} + 2t) \sum_{k=1}^d \frac{r_k}{n_k}} \right).$$

Because the above inequality applies to any $\mathbf{j} \in \mathbb{I}^d$, we have

$$\Pr\left\{ \|\mathbf{p}_{\mathbb{L}}(\mathbf{p}_{\mathbb{I}^d} - q^{-1} \mathbf{p}_{\mathbb{I}}) \mathbf{p}_{\mathbb{L}}(\mathbf{X})\|_{\infty} \geq t \right\} \leq 2 \exp \left(\frac{-3t^2 q}{u_0 \|\mathbf{X}\|_{\infty} (6\|\mathbf{X}\|_{\infty} + 2t) \sum_{k=1}^d \frac{r_k}{n_k}} \right) \prod_{k=1}^d n_k$$

by the union bound. $\square$

Let us now introduce the golfing scheme. We decompose $\mathbb{I}^{\perp}(\mathbf{S}) = \bigcup_{j=1}^m \mathbb{I}^{\perp}(\mathbf{S}_j)$ with $m \in \mathbb{N}$ to be specified later, where $\mathbb{I}(\mathbf{S}_j) \sim \text{Bernoulli}(\varphi)$ for $j \in [m]$ are identical and independent of each other. Since $\Pr\{\mathbf{i} \notin \mathbb{I}^{\perp}(\mathbf{S}_j)\} = \varphi$, we have $\Pr\{\mathbf{i} \notin \bigcup_{j=1}^m \mathbb{I}^{\perp}(\mathbf{S}_j)\} = \varphi^m$ and so $\Pr\{\mathbf{i} \in \bigcup_{j=1}^m \mathbb{I}^{\perp}(\mathbf{S}_j)\} = 1 - \varphi^m$. As a result, we must have $1 - \varphi^m = 1 - \rho$, i.e., $\varphi = \sqrt[m]{\rho}$. Given a tensor $\mathbf{Z} \in \mathbb{Z}(\mathbf{L})$, the golfing scheme [22, 11] recursively defines

$$\mathbf{Z}_0 = \mathbf{O} \text{ and } \mathbf{Z}_j = \mathbf{Z}_{j-1} - (1 - \varphi)^{-1} \mathbf{p}_{\mathbb{I}^{\perp}(\mathbf{S}_j)}(\mathbf{p}_{\mathbb{L}}(\mathbf{Z}_{j-1}) - \mathbf{Z}) \text{ for } j \in [m]. \quad (29)$$

We expect $\mathbf{Z}_m$ to be a candidate for the low-rank part of the dual certificate.

Let us take a close look at the $\mathbf{Z}_j$'s. As $\mathbf{Z} \in \mathbb{Z}(\mathbf{L}) \in \mathbb{L}$, by treating $\mathbf{p}_{\mathbb{L}}(\mathbf{Z}_j) - \mathbf{Z} \in \mathbb{L}$ as a residual that is corrected iteratively in the process, we have

$$\mathbf{Z}_k = - \sum_{j=1}^k (1 - \varphi)^{-1} \mathbf{p}_{\mathbb{L}^\perp}(\mathbf{s}_j) (\mathbf{p}_{\mathbb{L}}(\mathbf{Z}_{j-1}) - \mathbf{Z}) \text{ for } k \in [m]. \quad (30)$$

By observing that for any $j \in [m]$

$$\begin{aligned} \mathbf{p}_{\mathbb{L}}(\mathbf{Z}_j) - \mathbf{Z} &= \mathbf{p}_{\mathbb{L}}(\mathbf{Z}_{j-1}) - \mathbf{Z} - (1 - \varphi)^{-1} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{L}^\perp}(\mathbf{s}_j) (\mathbf{p}_{\mathbb{L}}(\mathbf{Z}_{j-1}) - \mathbf{Z}) \\ &= (\mathbf{p}_{\mathbb{L}} - (1 - \varphi)^{-1} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{L}^\perp}(\mathbf{s}_j) \mathbf{p}_{\mathbb{L}}) (\mathbf{p}_{\mathbb{L}}(\mathbf{Z}_{j-1}) - \mathbf{Z}), \end{aligned}$$

we also have for any $k \in [m]$ that

$$\mathbf{p}_{\mathbb{L}}(\mathbf{Z}_k) - \mathbf{Z} = (\mathbf{p}_{\mathbb{L}} - (1 - \varphi)^{-1} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{L}^\perp}(\mathbf{s}_k) \mathbf{p}_{\mathbb{L}}) \cdots (\mathbf{p}_{\mathbb{L}} - (1 - \varphi)^{-1} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{L}^\perp}(\mathbf{s}_1) \mathbf{p}_{\mathbb{L}}) (-\mathbf{Z}). \quad (31)$$

We next provide some estimates of $\mathbf{Z}_m$.

Lemma 6.11 *Under Assumption 6.1, if $m = \kappa_4 \ln n_d$ and $\mathbf{Z} \in \mathbb{Z}(\mathbf{L})$, then*

$$\|\mathbf{p}_{\mathbb{L}}(\mathbf{Z}_m) - \mathbf{Z}\|_2 \leq \frac{\lambda}{8}, \quad \|\mathbf{Z}_m\|_\infty \leq 2(1 - \varphi)^{-1} \|\mathbf{Z}\|_\infty, \quad \text{and} \quad \|\mathbf{p}_{\mathbb{L}^\perp}(\mathbf{Z}_m)\|_\sigma \leq 2\kappa_1 \|\mathbf{Z}\|_\infty \sqrt{\frac{n_d}{1 - \varphi}} \ln^{d-2} n_d$$

hold with high probability, where $\mathbf{Z}_m$ is defined by (29) for this $\mathbf{Z}$.

Proof. Assumption 6.1 implies that $1 - \rho \geq \frac{u_0 r_0 \ln^2 n_d}{\theta_0 n_1}$ and the well-known Bernoulli inequality implies that

$$1 - \varphi = 1 - \rho^{\frac{1}{m}} = 1 - (1 - (1 - \rho))^{\frac{1}{m}} \geq 1 - \left(1 - \frac{1}{m}(1 - \rho)\right) = \frac{1 - \rho}{m}.$$

As result, for a sufficiently large $\kappa_3 > 0$,

$$1 - \varphi \geq \frac{u_0 r_0 \ln^2 n_d}{\theta_0 n_1 m} = \frac{u_0 r_0 \ln n_d}{\theta_0 \kappa_4 n_1} \geq \max \left\{ \theta_1 \frac{\ln n_d}{n_d}, \frac{\kappa_3 d^2 u_0 r_0 \ln n_d}{n_1} \right\}$$

as long as $\theta_0 > 0$ is sufficiently small. Therefore, $\mathbb{L}^\perp(\mathbf{s}_j)$ satisfies the conditions in Corollary 6.9 and Lemma 6.10 for any $j \in [m]$. Besides, by Proposition 6.6, $\|\mathbf{p}_{\mathbb{L}}(\mathbf{p}_{\mathbb{L}^\perp} - (1 - \varphi)^{-1} \mathbf{p}_{\mathbb{L}^\perp}(\mathbf{s}_j)) \mathbf{p}_{\mathbb{L}}\| \leq \frac{1}{2}$ holds with probability at least

$$1 - \exp \left(\frac{-3(1 - \varphi)}{28 u_0 \sum_{k=1}^d \frac{r_k}{n_k}} \right) \prod_{k=1}^d n_k \geq 1 - \exp \left(\frac{-3\kappa_3}{28} \sum_{k=1}^d \ln n_k \right) \prod_{k=1}^d n_k = 1 - \left(\prod_{k=1}^d n_k \right)^{1 - \frac{3\kappa_3}{28}} \geq 1 - \frac{1}{n_d}$$

for a sufficiently large κ_3 .

By the well-known bound between the spectral and Frobenius norms of a tensor (see e.g., [36])

$$\|\mathbf{T}\|_\sigma \geq \|\mathbf{T}\|_2 \left(\prod_{k=1}^{d-1} n_k \right)^{-\frac{1}{2}} \text{ for any } \mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d},$$

we have $\|\mathbf{Z}\|_2 \leq \sqrt{\prod_{k=1}^{d-1} n_k}$ since $\|\mathbf{Z}\|_\sigma = 1$. Therefore, by (31),

$$\|\mathbf{p}_\mathbb{L}(\mathbf{Z}_m) - \mathbf{Z}\|_2 \leq \left(\prod_{j=1}^m \|\mathbf{p}_\mathbb{L} - (1 - \varphi)^{-1} \mathbf{p}_\mathbb{L} \mathbf{p}_{\mathbb{I}^\perp}(\mathbf{s}_j) \mathbf{p}_\mathbb{L}\| \right) \|\mathbf{Z}\|_2 \leq \frac{1}{2^m} \sqrt{\prod_{k=1}^{d-1} n_k} \leq \frac{1}{8\sqrt{n_d}} = \frac{\lambda}{8}$$

holds with probability at least $(1 - \frac{1}{n_d})^m$, where the last inequality requires that $2^{m-3} \geq \sqrt{\prod_{k=1}^d n_k}$, guaranteed by $m = \kappa_4 \ln n_d$. This is a high probability since $(1 - \frac{1}{n_d})^m = (1 - \frac{1}{n_d})^{\kappa_4 \ln n_d}$ tends to 1 as n_d tends to infinity.

Using a similar argument, we have by Lemma 6.10 and (31) that $\|\mathbf{p}_\mathbb{L}(\mathbf{Z}_k) - \mathbf{Z}\|_\infty \leq \frac{\|\mathbf{Z}\|_\infty}{2^k}$ holds with high probability for any $k \in [m]$. This, together with (30), further implies that

$$\|\mathbf{Z}_m\|_\infty \leq (1 - \varphi)^{-1} \sum_{j=1}^m \|\mathbf{p}_{\mathbb{I}^\perp}(\mathbf{s}_j) (\mathbf{p}_\mathbb{L}(\mathbf{Z}_{j-1}) - \mathbf{Z})\|_\infty \leq (1 - \varphi)^{-1} \sum_{j=1}^m \frac{\|\mathbf{Z}\|_\infty}{2^{j-1}} \leq 2(1 - \varphi)^{-1} \|\mathbf{Z}\|_\infty$$

holds with high probability. Finally, by (30) again, we have

$$\begin{aligned} \|\mathbf{p}_{\mathbb{L}^\perp}(\mathbf{Z}_m)\|_\sigma &= \left\| \mathbf{p}_{\mathbb{L}^\perp} \left(- \sum_{j=1}^m (1 - \varphi)^{-1} \mathbf{p}_{\mathbb{I}^\perp}(\mathbf{s}_j) (\mathbf{p}_\mathbb{L}(\mathbf{Z}_{j-1}) - \mathbf{Z}) \right) \right\|_\sigma \\ &= \left\| \sum_{j=1}^m \mathbf{p}_{\mathbb{L}^\perp} \left(\mathbf{p}_\mathbb{L} - (1 - \varphi)^{-1} \mathbf{p}_{\mathbb{I}^\perp}(\mathbf{s}_j) \right) (\mathbf{p}_\mathbb{L}(\mathbf{Z}_{j-1}) - \mathbf{Z}) \right\|_\sigma \\ &= \left\| \sum_{j=1}^m \mathbf{p}_{\mathbb{L}^\perp} \left(\mathbf{p}_{\mathbb{I}^d} - (1 - \varphi)^{-1} \mathbf{p}_{\mathbb{I}^\perp}(\mathbf{s}_j) \right) (\mathbf{p}_\mathbb{L}(\mathbf{Z}_{j-1}) - \mathbf{Z}) \right\|_\sigma \\ &\leq \sum_{j=1}^m \left\| (\mathbf{p}_{\mathbb{I}^d} - (1 - \varphi)^{-1} \mathbf{p}_{\mathbb{I}^\perp}(\mathbf{s}_j)) (\mathbf{p}_\mathbb{L}(\mathbf{Z}_{j-1}) - \mathbf{Z}) \right\|_\sigma \\ &\leq \kappa_1 \sqrt{\frac{n_d}{1 - \varphi}} \ln^{d-2} n_d \sum_{j=1}^m \|\mathbf{p}_\mathbb{L}(\mathbf{Z}_{j-1}) - \mathbf{Z}\|_\infty \\ &\leq \kappa_1 \sqrt{\frac{n_d}{1 - \varphi}} \ln^{d-2} n_d \sum_{j=1}^m \frac{\|\mathbf{Z}\|_\infty}{2^{j-1}} \\ &\leq 2\kappa_1 \|\mathbf{Z}\|_\infty \sqrt{\frac{n_d}{1 - \varphi}} \ln^{d-2} n_d \end{aligned}$$

holds with high probability, where the last equality holds because $\mathbf{p}_\mathbb{L}(\mathbf{Z}_{j-1}) - \mathbf{Z} \in \mathbb{L}$ and the second inequality is due Corollary 6.9. The proof is then completed by an overall union bound. $\square$

We are now in a position to conclude this subsection.

Proposition 6.12 *Under Assumption 6.1, there exists a $\mathbf{D}_1 \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ such that*

$$\min_{\mathbf{Z} \in \mathbb{Z}(\mathbf{L})} \|\mathbf{p}_\mathbb{L}(\mathbf{D}_1) - \mathbf{Z}\|_2 \leq \frac{\lambda}{8}, \quad \|\mathbf{p}_{\mathbb{L}^\perp}(\mathbf{D}_1)\|_\sigma \leq \frac{1}{4}, \quad \mathbf{p}_\mathbb{I}(\mathbf{s})(\mathbf{D}_1) = \mathbf{O}, \quad \text{and} \quad \|\mathbf{D}_1\|_\infty \leq \frac{\lambda}{4}$$

hold with high probability.

Proof. By Assumption 6.1, there exists a $\mathbf{Z} \in \mathbb{Z}(\mathbf{L})$ such that

$$\|\mathbf{Z}\|_\infty \leq \sqrt{\frac{u_0}{n_1 n_d \ln^{\max\{2d-5,0\}} n_d}} \cdot \sqrt{\frac{\theta_0(1-\rho)n_1}{u_0 \ln^2 n_d}} = \sqrt{\frac{\theta_0(1-\rho)}{n_d \ln^{\max\{2d-3,2\}} n_d}}.$$

Let us consider $\mathbf{D}_1 = \mathbf{Z}_m$ defined by (29) for this $\mathbf{Z}$ with $m = \kappa_4 \ln n_d$. It is obvious from (30) that $\mathbf{p}_{\mathbb{I}(\mathbf{S})}(\mathbf{D}_1) = \mathbf{O}$ as $\mathbb{I}^\perp(\mathbf{S}_j) \subseteq \mathbb{I}^\perp(\mathbf{S})$ for any $j \in [m]$. Besides, $\min_{\mathbf{Z} \in \mathbb{Z}(\mathbf{L})} \|\mathbf{p}_{\mathbb{L}}(\mathbf{D}_1) - \mathbf{Z}\|_2 \leq \frac{\lambda}{8}$ has been shown in Lemma 6.11.

For the remaining two statements, by Lemma 6.11, we have

$$\|\mathbf{D}_1\|_\infty \leq \frac{2}{1-\varphi} \|\mathbf{Z}\|_\infty \leq \frac{2m}{1-\rho} \sqrt{\frac{\theta_0(1-\rho)}{n_d \ln^{\max\{2d-3,2\}} n_d}} \leq \left(\frac{2\kappa_4 \sqrt{\theta_0}}{\sqrt{(1-\rho)n_d \ln^{\max\{2d-5,0\}} n_d}} \right) \lambda \leq \frac{\lambda}{4}$$

for a sufficiently small θ_0 . Finally by Lemma 6.11 again, we have

$$\|\mathbf{p}_{\mathbb{L}^\perp}(\mathbf{D}_1)\|_\sigma \leq 2\kappa_1 \sqrt{\frac{\theta_0(1-\rho)}{n_d \ln^{\max\{2d-3,2\}} n_d}} \cdot \sqrt{\frac{n_d m}{1-\rho}} \ln^{d-2} n_d \leq \frac{2\kappa_1 \sqrt{\kappa_4 \theta_0}}{\sqrt{\ln^{\max\{5-2d,0\}} n_d}} \leq \frac{1}{4}$$

for a sufficiently small θ_0 . □

6.5 Dual certificate: Sparse part via the least squares method

In this subsection, we construct the sparse part of the dual certificate via the least squares method, aiming to achieve the following.

Proposition 6.13 *Under Assumption 6.1, there exists a $\mathbf{D}_2 \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ such that*

$$\mathbf{p}_{\mathbb{L}}(\mathbf{D}_2) = \mathbf{O}, \quad \|\mathbf{p}_{\mathbb{L}^\perp}(\mathbf{D}_2)\|_\sigma < \frac{1}{4}, \quad \mathbf{p}_{\mathbb{I}(\mathbf{S})}(\mathbf{D}_2) = \lambda \mathbf{E}, \quad \text{and} \quad \|\mathbf{p}_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{D}_2)\|_\infty < \frac{\lambda}{4}$$

hold with high probability.

To start with, let us consider the following convex optimization problem

$$\min\{\|\mathbf{D}\|_2^2 : \mathbf{p}_{\mathbb{L}}(\mathbf{D}) = \mathbf{O}, \mathbf{p}_{\mathbb{I}(\mathbf{S})}(\mathbf{D}) = \lambda \mathbf{E}\}.$$

The purpose is to make sure that any feasible solution satisfies the two equality conditions in Proposition 6.13 and that the norm minimization meets the other two conditions as well. Using the standard method of Lagrange multipliers, it is not difficult to show that the optimal solution of the problem is

$$\mathbf{D}_2 = \lambda \mathbf{p}_{\mathbb{L}^\perp} \sum_{k=0}^{\infty} (\mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{I}(\mathbf{S})})^k (\mathbf{E}).$$

The rest of this subsection is devoted to prove the two inequality conditions in Proposition 6.13 for $\mathbf{D}_2$, shown in Lemma 6.16 and Lemma 6.17, respectively.

To begin, we present a technical lemma that approximates the tensor spectral norm by using the so-called ϵ -net of the unit sphere, i.e., a set of unit vectors such that the Euclidean distance from any unit vector to the set is no more than ϵ .

Lemma 6.14 *If $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ and $\mathbb{E}_k$ is an ϵ -net of $\mathbb{S}^{n_k}$ for $k \in [d]$ and $\epsilon \in (0, \frac{1}{d})$, then*

$$\|\mathbf{T}\|_\sigma \leq \frac{1}{1 - d\epsilon} \max_{\mathbf{v}_k \in \mathbb{E}_k \forall k \in [d]} \langle \mathbf{T}, \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \dots \otimes \mathbf{v}_d \rangle.$$

In particular, by letting $\epsilon = \frac{1}{2d}$, there exist $\mathbb{E}_k \subseteq \mathbb{S}^{n_k}$ with $|\mathbb{E}_k| \leq (1 + 4d)^{n_k}$ for $k \in [d]$ such that

$$\|\mathbf{T}\|_\sigma \leq 2 \max_{\mathbf{v}_k \in \mathbb{E}_k \forall k \in [d]} \langle \mathbf{T}, \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \dots \otimes \mathbf{v}_d \rangle. \quad (32)$$

Proof. Let $\|\mathbf{T}\|_\sigma = \langle \mathbf{T}, \mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \dots \otimes \mathbf{x}_d \rangle$ where $\mathbf{x}_k \in \mathbb{S}^{n_k}$ for $k \in [d]$. By choosing $\mathbf{v}_k \in \mathbb{E}_k$ with $\|\mathbf{x}_k - \mathbf{v}_k\|_2 \leq \epsilon$ for $k \in [d]$, we have

$$\begin{aligned} & \|\mathbf{T}\|_\sigma - \langle \mathbf{T}, \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \dots \otimes \mathbf{v}_d \rangle \\ & \leq \left| \left\langle \mathbf{T}, \bigotimes_{k=1}^d \mathbf{x}_k \right\rangle - \left\langle \mathbf{T}, \bigotimes_{k=1}^d \mathbf{v}_k \right\rangle \right| \\ & = \left| \sum_{k=1}^d \left(\left\langle \mathbf{T}, \left(\bigotimes_{j=1}^{k-1} \mathbf{v}_j \right) \otimes \left(\bigotimes_{j=k}^d \mathbf{x}_j \right) \right\rangle - \left\langle \mathbf{T}, \left(\bigotimes_{j=1}^k \mathbf{v}_j \right) \otimes \left(\bigotimes_{j=k+1}^d \mathbf{x}_j \right) \right\rangle \right) \right| \\ & \leq \sum_{k=1}^d \|\mathbf{T}\|_\sigma \left(\prod_{j=1}^{k-1} \|\mathbf{v}_j\|_2 \right) \|\mathbf{x}_k - \mathbf{v}_k\|_2 \left(\prod_{j=k+1}^d \|\mathbf{x}_j\|_2 \right) \\ & \leq \epsilon d \|\mathbf{T}\|_\sigma, \end{aligned}$$

which further implies that

$$\|\mathbf{T}\|_\sigma \leq \frac{1}{1 - d\epsilon} \langle \mathbf{T}, \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \dots \otimes \mathbf{v}_d \rangle.$$

The result then follows by taking maximum over all $\mathbb{E}_k$'s.

In order to show (32), the existence of relevant ϵ -nets is required. In particular, it is stated in [56, Corollary 4.2.13] that for any $\epsilon > 0$, there exists an ϵ -net of $\mathbb{S}^n$ whose cardinality is no more than $(1 + \frac{2}{\epsilon})^n$. $\square$

For relevant studies on the tensor norms by sphere covering, we refer interested readers to [28, 25, 23].

We next bound the spectral norm of $\mathbf{E}$ by applying the result of ϵ -net.

Lemma 6.15 *Under Assumption 6.1, $\|\mathbf{E}\|_\sigma \leq \frac{\kappa_5}{\sqrt{-\ln \rho}} \sqrt{\sum_{k=1}^d n_k}$ holds with high probability.*

Proof. By the condition of $\mathbf{S}$ in Assumption 6.1 and $\mathbf{E} = \text{sign}(\mathbf{S})$, we have $(\text{Ex } |e_{\mathbf{i}}|^x)^{\frac{1}{x}} = \rho^{\frac{1}{x}}$ for any $\mathbf{i} \in \mathbb{I}^d$ and $x \geq 1$. The function $\rho^{\frac{1}{x}} / \sqrt{x}$ achieves the maximum over $[1, \infty)$ at $x = -2 \ln \rho$ as long as $\rho \leq \frac{1}{\sqrt{e}}$, which is certainly the case as ρ is assumed to be sufficiently small. As a result,

$$(\text{Ex } |e_{\mathbf{i}}|^x)^{\frac{1}{x}} \leq \sqrt{\frac{x}{-2e \ln \rho}} \text{ for any } x \geq 1.$$

This, together with [56, Proposition 2.5.2], further implies that

$$\text{Ex } \exp(t e_{\mathbf{i}}) \leq \exp\left(\frac{\kappa_6}{-2e \ln \rho} t^2\right) \text{ for any } \mathbf{i} \in \mathbb{I}^d \text{ and } t \in \mathbb{R}.$$

The result then follows directly by combining the above with [54, Lemma 1] and [54, Theorem 1]. $\square$

We are now ready to make two key claims to conclude this subsection.

Lemma 6.16 *Under Assumption 6.1, $\|p_{\mathbb{L}^\perp}(\mathbf{D}_2)\|_\sigma < \frac{1}{4}$ holds with high probability.*

Proof. We rewrite

$$p_{\mathbb{L}^\perp}(\mathbf{D}_2) = \lambda p_{\mathbb{L}^\perp}(\mathbf{E}) + \lambda p_{\mathbb{L}^\perp} \sum_{k=1}^{\infty} (p_{\mathbb{I}(\mathbf{S})} p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})})^k(\mathbf{E})$$

and bound their spectral norms separately. By Lemma 3.2 and Lemma 6.15,

$$\|\lambda p_{\mathbb{L}^\perp}(\mathbf{E})\|_\sigma \leq \lambda \|\mathbf{E}\|_\sigma \leq \frac{1}{\sqrt{n_d}} \cdot \frac{\kappa_5}{\sqrt{-\ln \rho}} \sqrt{\sum_{k=1}^d n_k} \leq \frac{\kappa_5 \sqrt{d}}{\sqrt{-\ln \rho}} \leq \frac{1}{8}$$

holds with high probability for a sufficiently small ρ . Besides, by Corollary 6.8, $\|p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})}\| \leq \delta$ holds with high probability for any $\delta \in (0, 1]$. Therefore, it suffices to show that

$$\Pr \left\{ \left\| \lambda p_{\mathbb{L}^\perp} \sum_{k=1}^{\infty} (p_{\mathbb{I}(\mathbf{S})} p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})})^k(\mathbf{E}) \right\|_\sigma < \frac{1}{8} \mid \|p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})}\| \leq \delta \right\}$$

is a high probability.

In fact, by Lemma 3.2 and (32) in Lemma 6.14,

$$\begin{aligned} \left\| \lambda p_{\mathbb{L}^\perp} \sum_{k=1}^{\infty} (p_{\mathbb{I}(\mathbf{S})} p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})})^k(\mathbf{E}) \right\|_\sigma &\leq \lambda \left\| \sum_{k=1}^{\infty} (p_{\mathbb{I}(\mathbf{S})} p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})})^k(\mathbf{E}) \right\|_\sigma \\ &\leq 2\lambda \max_{\mathbf{v}_k \in \mathbb{E}_k \forall k \in [d]} \left\langle \mathbf{E}, \sum_{k=1}^{\infty} (p_{\mathbb{I}(\mathbf{S})} p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})})^k \left(\bigotimes_{k=1}^d \mathbf{v}_k \right) \right\rangle, \end{aligned}$$

where $\mathbb{E}_k$'s are the ϵ -nets used in (32). For any given $\mathbb{I}(\mathbf{S})$, the nonzero entries of $\mathbf{E}$ are i.i.d. symmetric Bernoulli random variables (taking ± 1 with equal probability), according to Assumption 6.1. Hence, by Hoeffding's inequality [56, Theorem 2.2.2], we have

$$\begin{aligned} &\Pr \left\{ \left\| \lambda p_{\mathbb{L}^\perp} \sum_{k=1}^{\infty} (p_{\mathbb{I}(\mathbf{S})} p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})})^k(\mathbf{E}) \right\|_\sigma \geq t \mid \mathbb{I}(\mathbf{S}) \right\} \\ &\leq \Pr \left\{ \max_{\mathbf{v}_k \in \mathbb{E}_k \forall k \in [d]} \left\langle \mathbf{E}, \sum_{k=1}^{\infty} (p_{\mathbb{I}(\mathbf{S})} p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})})^k \left(\bigotimes_{k=1}^d \mathbf{v}_k \right) \right\rangle \geq \frac{t}{2\lambda} \mid \mathbb{I}(\mathbf{S}) \right\} \\ &\leq \sum_{\mathbf{v}_k \in \mathbb{E}_k \forall k \in [d]} \Pr \left\{ \left\langle \mathbf{E}, \sum_{k=1}^{\infty} (p_{\mathbb{I}(\mathbf{S})} p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})})^k \left(\bigotimes_{k=1}^d \mathbf{v}_k \right) \right\rangle \geq \frac{t}{2\lambda} \mid \mathbb{I}(\mathbf{S}) \right\} \\ &\leq \sum_{\mathbf{v}_k \in \mathbb{E}_k \forall k \in [d]} \exp \left(\frac{-t^2}{8\lambda^2 \left\| \sum_{k=1}^{\infty} (p_{\mathbb{I}(\mathbf{S})} p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})})^k \left(\bigotimes_{k=1}^d \mathbf{v}_k \right) \right\|_2^2} \right) \\ &\leq \exp \left(\frac{-t^2}{8\lambda^2 \left\| \sum_{k=1}^{\infty} (p_{\mathbb{I}(\mathbf{S})} p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})})^k \right\|_2^2} \right) \prod_{k=1}^d (1 + 4d)^{n_k}, \end{aligned}$$

where the last inequality is due to $\left\| \bigotimes_{k=1}^d \mathbf{v}_k \right\|_2 = 1$ and $|\mathbb{E}_k| \leq (1 + 4d)^{n_k}$ for $k \in [d]$.

As $\|p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})}\| \leq \delta$ implies that $\left\| \sum_{k=1}^{\infty} (p_{\mathbb{I}(\mathbf{S})} p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})})^k \right\| \leq \sum_{k=1}^{\infty} \delta^k = \frac{\delta}{1-\delta}$, by letting $t = \frac{1}{8}$ in the above bound and recalling $\lambda = \frac{1}{\sqrt{n_d}}$, we have

$$\Pr \left\{ \left\| \lambda p_{\mathbb{L}^\perp} \sum_{k=1}^{\infty} (p_{\mathbb{I}(\mathbf{S})} p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})})^k(\mathbf{E}) \right\|_\sigma \geq \frac{1}{8} \mid \|p_{\mathbb{L}} p_{\mathbb{I}(\mathbf{S})}\| \leq \delta \right\} \leq \exp \left(\frac{-(1-\delta)^2 n_d}{8^3 \delta^2} \right) \prod_{k=1}^d (1 + 4d)^{n_k}$$

$$\leq \exp\left(n_d \left(-\frac{(1-\delta)^2}{8^3 \delta^2} + d \ln(1+4d)\right)\right),$$

which is a low probability by choosing a sufficiently small δ such that $-\frac{(1-\delta)^2}{8^3 \delta^2} + d \ln(1+4d) < 0$. $\square$

Lemma 6.17 *Under Assumption 6.1, $\|\mathbf{p}_{\mathbb{I}^\perp}(\mathbf{S})(\mathbf{D}_2)\|_\infty < \frac{\lambda}{4}$ holds with high probability.*

Proof. To begin with, we observe that

$$\begin{aligned} \|\mathbf{p}_{\mathbb{I}^\perp}(\mathbf{S})(\mathbf{D}_2)\|_\infty &= \left\| \lambda \mathbf{p}_{\mathbb{I}^\perp}(\mathbf{S})(\mathbf{p}_{\mathbb{I}^d} - \mathbf{p}_{\mathbb{L}}) \sum_{k=0}^{\infty} (\mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{I}(\mathbf{S})})^k(\mathbf{E}) \right\|_\infty \\ &= \left\| -\lambda \mathbf{p}_{\mathbb{I}^\perp}(\mathbf{S}) \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{I}(\mathbf{S})} \sum_{k=0}^{\infty} (\mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{I}(\mathbf{S})})^k(\mathbf{E}) \right\|_\infty \\ &= \lambda \max_{\mathbf{i} \in \mathbb{I}^d} \left| \left\langle \mathbf{E}, \left(\sum_{k=0}^{\infty} (\mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{I}(\mathbf{S})})^k \right) \mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{I}^\perp}(\mathbf{S}) \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right) \right\rangle \right| \\ &= \lambda \max_{\mathbf{i} \in \mathbb{I}^\perp(\mathbf{S})} \left| \left\langle \mathbf{E}, \left(\sum_{k=0}^{\infty} (\mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{I}(\mathbf{S})})^k \right) \mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right) \right\rangle \right|. \end{aligned}$$

By Corollary 6.8, $\|\mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{I}(\mathbf{S})}\| \leq \delta$ holds with high probability for any $\delta \in (0, 1]$, which further implies that $\|\sum_{k=0}^{\infty} (\mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{I}(\mathbf{S})})^k\| \leq \sum_{k=0}^{\infty} \delta^k = \frac{1}{1-\delta}$. As a result, we have for any $\mathbf{i} \in \mathbb{I}^\perp(\mathbf{S})$ that

$$\left\| \left(\sum_{k=0}^{\infty} (\mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{I}(\mathbf{S})})^k \right) \mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right) \right\|_2^2 \leq \frac{\|\mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}}\|^2}{(1-\delta)^2} \left\| \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right) \right\|_2^2 \leq \frac{\delta^2 u_0 \sum_{k=1}^d \frac{r_k}{n_k}}{(1-\delta)^2},$$

where the last inequality is due to Lemma 6.5.

For any given $\mathbb{I}(\mathbf{S})$, the nonzero entries of $\mathbf{E}$ are i.i.d. symmetric Bernoulli random variables. By the two-sided version of Hoeffding's inequality [56, Theorem 2.2.5],

$$\begin{aligned} &\Pr\left\{ \|\mathbf{p}_{\mathbb{I}^\perp}(\mathbf{S})(\mathbf{D}_2)\|_\infty \geq \frac{\lambda}{4} \mid \mathbb{I}(\mathbf{S}) \right\} \\ &= \Pr\left\{ \max_{\mathbf{i} \in \mathbb{I}^\perp(\mathbf{S})} \left| \left\langle \mathbf{E}, \left(\sum_{k=0}^{\infty} (\mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{I}(\mathbf{S})})^k \right) \mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right) \right\rangle \right| \geq \frac{1}{4} \mid \mathbb{I}(\mathbf{S}) \right\} \\ &\leq \sum_{\mathbf{i} \in \mathbb{I}^\perp(\mathbf{S})} \Pr\left\{ \left| \left\langle \mathbf{E}, \left(\sum_{k=0}^{\infty} (\mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{I}(\mathbf{S})})^k \right) \mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right) \right\rangle \right| \geq \frac{1}{4} \mid \mathbb{I}(\mathbf{S}) \right\} \\ &\leq 2 \sum_{\mathbf{i} \in \mathbb{I}^\perp(\mathbf{S})} \exp\left(- \left(32 \left\| \left(\sum_{k=0}^{\infty} (\mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \mathbf{p}_{\mathbb{I}(\mathbf{S})})^k \right) \mathbf{p}_{\mathbb{I}(\mathbf{S})} \mathbf{p}_{\mathbb{L}} \left(\bigotimes_{k=1}^d \mathbf{e}_{i_k} \right) \right\|_2^2 \right)^{-1} \right) \\ &\leq 2 \exp\left(- \frac{(1-\delta)^2}{32 \delta^2 u_0 \sum_{k=1}^d \frac{r_k}{n_k}} \right) \prod_{k=1}^d n_k \\ &\leq 2 \exp\left(- \frac{(1-\delta)^2 \ln^2 n_d}{32 \delta^2 \theta_0 d (1-\rho)} + d \ln n_d \right), \end{aligned}$$

where the last inequality follows from Assumption 6.1. This is clearly a low probability by choosing a sufficiently small δ . $\square$

6.6 Putting everything together

We arrive at the following result by combining the results of previous two subsections with an overall union bound. This, together with Corollary 6.8 and Lemma 6.4, directly shows Theorem 6.2.

Proposition 6.18 *Under Assumption 6.1, there exists a $\mathbf{D} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ such that*

$$\min_{\mathbf{Z} \in \mathbb{Z}(\mathbf{L})} \|\mathbf{p}_{\mathbb{L}}(\mathbf{D}) - \mathbf{Z}\|_2 \leq \frac{\lambda}{8}, \quad \|\mathbf{p}_{\mathbb{L}^\perp}(\mathbf{D})\|_\sigma < \frac{1}{2}, \quad \|\mathbf{p}_{\mathbb{I}(\mathbf{S})}(\mathbf{D}) - \lambda \mathbf{E}\|_2 \leq \frac{\lambda}{8}, \quad \text{and} \quad \|\mathbf{p}_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{D})\|_\infty < \frac{\lambda}{2}$$

hold with high probability.

Proof. By Proposition 6.12 and Proposition 6.13,

$$\begin{aligned} \min_{\mathbf{Z} \in \mathbb{Z}(\mathbf{L})} \|\mathbf{p}_{\mathbb{L}}(\mathbf{D}_1 + \mathbf{D}_2) - \mathbf{Z}\|_2 &= \min_{\mathbf{Z} \in \mathbb{Z}(\mathbf{L})} \|\mathbf{p}_{\mathbb{L}}(\mathbf{D}_1) - \mathbf{Z}\|_2 \leq \frac{\lambda}{8}, \\ \|\mathbf{p}_{\mathbb{L}^\perp}(\mathbf{D}_1 + \mathbf{D}_2)\|_\sigma &\leq \|\mathbf{p}_{\mathbb{L}^\perp}(\mathbf{D}_1)\|_\sigma + \|\mathbf{p}_{\mathbb{L}^\perp}(\mathbf{D}_2)\|_\sigma < \frac{1}{2}, \\ \|\mathbf{p}_{\mathbb{I}(\mathbf{S})}(\mathbf{D}_1 + \mathbf{D}_2) - \lambda \mathbf{E}\|_2 &= \|\mathbf{p}_{\mathbb{I}(\mathbf{S})}(\mathbf{D}_2) - \lambda \mathbf{E}\|_2 = 0 \leq \frac{\lambda}{8}, \\ \|\mathbf{p}_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{D}_1 + \mathbf{D}_2)\|_\infty &\leq \|\mathbf{D}_1\|_\infty + \|\mathbf{p}_{\mathbb{I}^\perp(\mathbf{S})}(\mathbf{D}_2)\|_\infty < \frac{\lambda}{2} \end{aligned}$$

hold with high probability. The proof is completed by letting $\mathbf{D} = \mathbf{D}_1 + \mathbf{D}_2$. $\square$

7 Concluding remarks

In this paper, we systematically study the decomposability and the subdifferential of the tensor nuclear norm. We show that the tensor nuclear norm is decomposable over a pair of subspaces that have at least two disjoint modes, naturally generalizing the result for the matrix case. The same property holds for the tensor spectral norm as well. Based on the decomposability, we propose novel inclusions of the subdifferential of the tensor nuclear norm. They strictly enlarge the inclusion proposed in [62, Theorem 1], the only known subdifferential inclusion for tensors of an arbitrary order. While various bounds for the subdifferential of the tensor nuclear norm and several interesting examples suggest that there is no general way to explicitly characterize the subdifferential of the tensor nuclear norm in general as in the matrix case, we examine subgradients of the tensor nuclear norm in all relevant subspaces and estimate their inner and outer approximations.

We believe that these new results on the tensor nuclear norm have great potential in applications. For instance, the new inclusions of the subdifferential of the tensor nuclear norm can potentially be applied to analyze the statistical performance of a variety of nuclear-norm-based tensor learning problems. As a precursor, we study one immediate application, the tensor robust PCA. As shown in Theorem 6.2, the exact recovery applies to tensors of an arbitrary order for the first time in the literature, generalizing the result in the matrix case when $d = 2$. In light of our analysis, we propose a natural conjecture concerning the conditions for exact recovery of the tensor robust PCA.

Conjecture 7.1 *The factor $\ln^{\max\{2d-5, 0\}} n_d$ in Assumption 6.1 can be removed in Theorem 6.2.*

In fact, the conjecture holds true if the factor $\ln^{d-2} n_d$ in Corollary 6.9 can be removed. This has already been conjectured similarly in [64, Section 11]. If Conjecture 7.1 is true, then the conditions for exact recovery of the tensor robust PCA of every order, including the matrix case, would become identical.

While our study indicates that there is no general way to explicitly characterize the subdifferential of the tensor nuclear norm, it is definitely interesting to look into specific classes of tensors, in particular those from practical applications, such that a complete characterization is possible. Moreover, despite the remarkable statistical performance of the tensor robust PCA, the underlying optimization problem (23) is computationally intractable due to the NP-hardness to compute the tensor nuclear norm. Developing tractable tight approximations of the tensor robust PCA is important. For example, one may resort to the sum-of-squares relaxation; see, e.g., [3, 4] for its application to tensor completion and a related work [60].

Apart from practical applications, our developments may facilitate the study of a variety of other more in-depth properties and implications of the tensor nuclear norm. For example, they may help to deduce the C^2 -cone reducibility [6, Definition 3.135], subdifferential metric subregularity [18, Section 3.H], and twice epi-differentiability [53, Definition 13.6(b)] of the tensor nuclear norm; see, e.g., [16, Proposition 3.2], [63, Proposition 11] and [16, Proposition 3.8], and [24, Corollary 3.6] and [43], respectively for their matrix counterparts. Besides, the developments can also be useful to understand the neural collapse [48] in training tensor-based neural networks and the relations between various nuclear-norm-regularized tensor optimization problems and their Burer-Monteiro reparameterizations [9, 10]; see, e.g., [65, Appendix C] and [34, 8, 42, 47], respectively for their matrix counterparts.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China [Grants 72394364, 72394360, and 72171141]. The first author would like to express his sincere gratitude to Professor Simai He for his generous subsidy support when the first author was studying at School of Information Management and Engineering, Shanghai University of Finance and Economics, where part of the work by the first author was conducted.

A Computing tensor spectral norms

Lemma A.1 $\|\mathbf{Z}(t)\|_\sigma = 1$ if and only if $-1 \leq t \leq 1$, where

$$\mathbf{Z}(t) = \sum_{i=1}^3 \mathbf{e}_i \otimes \mathbf{e}_i \otimes \mathbf{e}_i + t \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_3 \in \mathbb{R}^{3 \times 3 \times 3}.$$

Proof. By the definition of the spectral norm, $\|\mathbf{Z}(t)\|_\sigma \geq \langle \mathbf{Z}(t), \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 \rangle = 1$ for any $t \in \mathbb{R}$ and $\|\mathbf{Z}(t)\|_\sigma \geq \langle \mathbf{Z}(t), \text{sign}(t) \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_3 \rangle = |t| > 1$ for any $|t| > 1$. It suffices to show that $\|\mathbf{Z}(t)\|_\sigma \leq 1$ for any $-1 \leq t \leq 1$.

To this end, consider the first-order optimality condition of the following problem

$$\|\mathbf{Z}(t)\|_\sigma = \max \left\{ \sum_{i=1}^3 x_i y_i z_i + t x_1 y_2 z_3 : \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{S}^3 \right\}. \quad (33)$$

As an optimization problem of a smooth function over the oblique manifold, it follows from [7, Proposition 4.5], [7, Proposition 4.6], [7, Exercise 3.67], and [7, (7.10)] that any local maximizer $(\mathbf{u}, \mathbf{v}, \mathbf{w})$ satisfies

$$\begin{cases} v_1 w_1 + t v_2 w_3 = \lambda u_1 \\ v_2 w_2 = \lambda u_2 \\ v_3 w_3 = \lambda u_3, \end{cases} \quad \begin{cases} u_1 w_1 = \lambda v_1 \\ u_2 w_2 + t u_1 w_3 = \lambda v_2 \\ u_3 w_3 = \lambda v_3, \end{cases} \quad \text{and} \quad \begin{cases} u_1 v_1 = \lambda w_1 \\ u_2 v_2 = \lambda w_2 \\ u_3 v_3 + t u_1 v_2 = \lambda w_3, \end{cases}$$

where $\lambda = \sum_{i=1}^3 u_i v_i w_i + t u_1 v_2 w_3$, i.e., the objective function value of $(\mathbf{u}, \mathbf{v}, \mathbf{w})$ in (33).

If $v_1 w_1 \neq 0$, then by $u_1 w_1 = \lambda v_1$ and $u_1 v_1 = \lambda w_1$, we have $\lambda^2 = u_1^2 \leq 1$, i.e., $|\lambda| \leq 1$. For the same reason, either $u_2 w_2 \neq 0$ or $u_3 v_3 \neq 0$ implies that $|\lambda| \leq 1$. It remains to consider the case that $v_1 w_1 = u_2 w_2 = u_3 v_3 = 0$, under which $|\lambda| = |\sum_{i=1}^3 u_i v_i w_i + t u_1 v_2 w_3| = |t u_1 v_2 w_3| \leq |t| \leq 1$ if $-1 \leq t \leq 1$. Therefore, the objective value of any local maximizer in (33) is no more than 1 if $-1 \leq t \leq 1$. $\square$

Lemma A.2 $\|\mathbf{Z} + \mathbf{X}(t)\|_\sigma = 1$ if and only if $-1 \leq t \leq 1$ and $\|\mathbf{Z} + \mathbf{Y}(t)\|_\sigma = 1$ if and only if $t = 0$, where

$$\mathbf{Z} = \sum_{i=1}^2 \mathbf{e}_i \otimes \mathbf{e}_i \otimes \mathbf{e}_i, \quad \mathbf{X}(t) = t \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_3, \quad \text{and} \quad \mathbf{Y}(t) = t \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_3, \quad \text{all in } \mathbb{R}^{2 \times 2 \times 3}.$$

Proof. By Lemma A.1, we have $\|\mathbf{U}(t)\|_\sigma = 1$ for any $-1 \leq t \leq 1$, where

$$\mathbf{U}(t) = \sum_{i=1}^3 \mathbf{e}_i \otimes \mathbf{e}_i \otimes \mathbf{e}_i + t \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_3 \in \mathbb{R}^{3 \times 3 \times 3}.$$

Since $\mathbf{Z} + \mathbf{X}(t)$ is a subtensor of $\mathbf{U}(t)$, it is easy to see (or by [35, Theorem 3.1]) that

$$\|\mathbf{Z} + \mathbf{X}(t)\|_\sigma \leq \|\mathbf{U}(t)\|_\sigma = 1 \quad \text{for any } -1 \leq t \leq 1.$$

This, together with the fact that

$$\begin{aligned} \|\mathbf{Z} + \mathbf{X}(t)\|_\sigma &\geq \langle \mathbf{Z} + \mathbf{X}(t), \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 \rangle = 1 \quad \text{for any } t \in \mathbb{R} \text{ and} \\ \|\mathbf{Z} + \mathbf{X}(t)\|_\sigma &\geq \langle \mathbf{Z} + \mathbf{X}(t), \text{sign}(t) \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_3 \rangle = |t| > 1 \quad \text{for any } |t| > 1 \end{aligned}$$

immediately implies the first result.

To show the second result, we observe that the vector $\mathbf{v}(t) = (1, 0, t)^\top$ is a subtensor of $\mathbf{Z} + \mathbf{Y}(t)$. Therefore, $\|\mathbf{Z} + \mathbf{Y}(t)\|_\sigma \geq \|\mathbf{v}(t)\|_\sigma = \sqrt{1 + t^2} > 1$ if $t \neq 0$. The conclusion then follows from the fact that $\|\mathbf{Z} + \mathbf{Y}(0)\|_\sigma = \|\mathbf{Z}\|_\sigma = 1$. $\square$

Lemma A.3 $\|\mathbf{X}(t)\|_\sigma = \frac{2|t|}{\sqrt{3}}$ for any $t \in \mathbb{R}$ and $\|\mathbf{Z} + \mathbf{X}(t)\|_\sigma = 1$ if and only if $-1 \leq t \leq \frac{1}{2}$, where

$$\mathbf{Z} = \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 \quad \text{and} \quad \mathbf{X}(t) = t(\mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1), \quad \text{both in } \mathbb{R}^{2 \times 2 \times 2}.$$

Proof. Since $\mathbf{X}(t)$ is a symmetric tensor, it follows from Banach's theorem (see e.g., [2] and [14, Corollary 4.2]) that

$$\begin{aligned} \|\mathbf{X}(t)\|_\sigma &= \max\{\langle \mathbf{X}(t), \mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x} \rangle : \mathbf{x} \in \mathbb{S}^2\} \\ &= \max\{3tx_1x_2^2 : x_1^2 + x_2^2 = 1\} \\ &= \max\{3tx(1 - x^2) : -1 \leq x \leq 1\} \\ &= \frac{2|t|}{\sqrt{3}}. \end{aligned}$$

$\mathbf{Z} + \mathbf{X}(t)$ is also a symmetric tensor. By Banach's theorem again,

$$\|\mathbf{Z} + \mathbf{X}(t)\|_\sigma = \max\{\langle \mathbf{Z} + \mathbf{X}(t), \mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x} \rangle : \mathbf{x} \in \mathbb{S}^2\}$$

$$\begin{aligned}
&= \max\{x_1^3 + 3tx_1x_2^2 : x_1^2 + x_2^2 = 1\} \\
&= \max\{x^3 + 3tx(1-x^2) : -1 \leq x \leq 1\}.
\end{aligned}$$

It can be calculated that

$$\max\{x^3 + 3tx(1-x^2) : -1 \leq x \leq 1\} = \begin{cases} 1 & -1 \leq t \leq \frac{1}{2} \\ 2\sqrt{\frac{t^3}{-1+3t}} & \text{otherwise.} \end{cases}$$

This, together with the fact that $2\sqrt{\frac{t^3}{-1+3t}} > 1$ for any $t \in (-\infty, -1) \cup (\frac{1}{2}, \infty)$, immediately implies that $\|\mathbf{Z} + \mathbf{X}(t)\|_\sigma = 1$ if and only if $-1 \leq t \leq \frac{1}{2}$. $\square$

We remark that $\|\mathbf{X}(\frac{1}{\sqrt{3}})\|_\sigma$ has been calculated in [20, Lemma 6.2] and can be easily extended to $\|\mathbf{X}(t)\|_\sigma$ for any t . $\|\mathbf{Z} + \mathbf{X}(-1)\|_\sigma$ has also been calculated in [20, Lemma 6.1]. In fact, the tensor $\mathbf{Z} + \mathbf{X}(-1)$ is known as an orthogonal tensor in the literature; see [36, Theorem 3.5] for more details. The result of $\|\mathbf{Z} + \mathbf{X}(t)\|_\sigma$ in Lemma A.3 provides a better understanding of the generalization of such a tensor.

B Computing optimization problems

Lemma B.1 $\max\{x_1y_2z_2 + x_2 : x_1y_2z_2 + x_2 \leq 1 + x_1y_1z_1, \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{S}^2 \cap \mathbb{R}_+^2\} = \frac{1+\sqrt{2}}{2}$.

Proof. Let v be the optimal value of the problem. By letting $x_1 = x_2 = \frac{\sqrt{2}}{2}$, $y_1 = z_1 = \sqrt{1 - \frac{\sqrt{2}}{2}}$, and $y_2 = z_2 = \sqrt{\frac{\sqrt{2}}{2}}$, we have $v \geq x_1y_2z_2 + x_2 = \frac{1+\sqrt{2}}{2}$. It remains to show that $v \leq \frac{1+\sqrt{2}}{2}$.

If $x_1y_2z_2 + x_2 \leq 1$, then $v \leq 1 \leq \frac{1+\sqrt{2}}{2}$. It remains to consider the case that $x_1y_2z_2 + x_2 > 1$. This also implies that $x_1 + x_2 \geq x_1y_2z_2 + x_2 > 1$ by the obvious fact that $x_i, y_i, z_i \in [0, 1]$ for $i \in [2]$. As a result,

$$\begin{aligned}
x_1y_2z_2 + x_2 \leq 1 + x_1y_1z_1 &\implies (x_1y_2z_2 + x_2 - 1)^2 \leq x_1^2y_1^2z_1^2 = x_1^2(1 - y_2^2)(1 - z_2^2) \\
&\implies 2x_1y_2z_2(x_2 - 1) + x_1^2(y_2^2 + z_2^2) \leq x_1^2 - (x_2 - 1)^2 \\
&\implies 2x_1y_2z_2(x_2 - 1) + 2x_1^2y_2z_2 \leq x_1^2 - (x_2 - 1)^2 \\
&\implies x_1y_2z_2 + x_2 \leq \frac{x_1^2 - (x_2 - 1)^2}{2(x_1 + x_2 - 1)} + x_2 = \frac{x_1x_2}{x_1 + x_2 - 1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
v &\leq \max\left\{\frac{x_1x_2}{x_1 + x_2 - 1} : x_1y_2z_2 + x_2 \leq 1 + x_1y_1z_1, x_1 + x_2 > 1, \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{S}^2 \cap \mathbb{R}_+^2\right\} \\
&\leq \max\left\{\frac{x_1x_2}{x_1 + x_2 - 1} : x_1 + x_2 > 1, \mathbf{x} \in \mathbb{S}^2 \cap \mathbb{R}_+^2\right\} = \frac{1 + \sqrt{2}}{2},
\end{aligned}$$

under the condition that $x_1y_2z_2 + x_2 > 1$. $\square$

Lemma B.2 $\max\{x_1y_1z_2 + x_1y_2 + x_2 : x_1y_1z_2 + x_1y_2 + x_2 \leq 1 + x_1y_1z_1, \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{S}^2 \cap \mathbb{R}_+^2\} = \frac{3}{2}$.

Proof. Let v be the optimal value of the problem. By letting $x_1 = \frac{\sqrt{3}}{2}$, $x_2 = \frac{1}{2}$, $y_1 = \frac{\sqrt{6}}{3}$, $y_2 = \frac{\sqrt{3}}{3}$, and $z_1 = z_2 = \frac{\sqrt{2}}{2}$, we have $v \geq x_1y_1z_2 + x_1y_2 + x_2 = \frac{3}{2}$. It remains to show that $v \leq \frac{3}{2}$.

If $x_1y_1z_2+x_1y_2+x_2 \leq 1$, then $v \leq 1 \leq \frac{3}{2}$. It remains to consider the case that $x_1y_1z_2+x_1y_2+x_2 > 1$. This, together with the constraint $x_1y_1z_2 + x_1y_2 + x_2 \leq 1 + x_1y_1z_1$, implies that

$$(x_1y_1z_2 + x_1y_2 + x_2 - 1)^2 \leq x_1^2y_1^2z_1^2 = 1 - x_2^2 - x_1^2y_2^2 - x_1^2y_1^2z_2^2 \leq 1 - \frac{(x_1y_1z_2 + x_1y_2 + x_2)^2}{3}.$$

By solving the above inequality with respect to $x_1y_1z_2 + x_1y_2 + x_2$, we get $x_1y_1z_2 + x_1y_2 + x_2 \leq \frac{3}{2}$, implying that $v \leq \frac{3}{2}$. $\square$

C Subdifferential of the nuclear norm for fourth-order tensors

The following example generalizes Example 5.9 from $d = 3$ to $d = 4$.

Example C.1 Let $\mathbf{T} = \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 \in \mathbb{R}^{2 \times 2 \times 2 \times 2}$. We have $\text{sp}_k(\mathbf{T}) = \text{sp}(\mathbf{e}_1)$ for $k \in [4]$. Let $\mathbf{Z} = \mathbf{T} \in \mathbb{Z}(\mathbf{T})$ since $\langle \mathbf{Z}, \mathbf{T} \rangle = 1 = \|\mathbf{T}\|_*$ and $\|\mathbf{Z}\|_\sigma = 1$.

Let $\mathbf{X}(t) = t\mathbf{e}_1 \otimes (\mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1) + t\mathbf{e}_2 \otimes (\mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1) \in \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [4]} \mathbb{T}^\mathbb{I}(\mathbf{T})$. It can be verified that $\|\mathbf{Z} + \mathbf{X}(t)\|_\sigma = 1$ if and only if $-\frac{1+\sqrt{2}}{3} \leq t \leq \frac{1}{3}$, following a similar proof to that of Lemma A.3. Therefore, $\mathbf{Z} + \mathbf{X}(t) \in \partial\|\mathbf{T}\|_*$ for any $-\frac{1+\sqrt{2}}{3} \leq t \leq \frac{1}{3}$. However, it can also be verified that $\|\mathbf{X}(t)\|_\sigma = \frac{3|t|}{2}$, again following a similar proof to that of Lemma A.3.

The following result generalizes [61, Lemma 1] from $d = 3$ to $d = 4$.

Lemma C.2 If $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4}$ is nonzero, then $\|\mathbf{Z} + \mathbf{X}\|_\sigma \leq 1$ for any $\mathbf{Z} \in \mathbb{Z}(\mathbf{T})$ and $\mathbf{X} \in \bigoplus_{|\mathbb{I}| \geq 2, \mathbb{I} \subseteq [4]} \mathbb{T}^\mathbb{I}(\mathbf{T})$ with $\|\mathbf{X}\|_\sigma \leq \frac{1}{3}$.

Proof. Let us denote

$$\begin{aligned} \mathbf{X}_i &= \text{p}_{\mathbb{T}^{\{i,4\}}(\mathbf{T})}(\mathbf{X}) \text{ for } i \in [3], \\ \mathbf{X}_4 &= \text{p}_{\mathbb{T}^{\{1,3\}}(\mathbf{T}) \oplus \mathbb{T}^{\{1,3,4\}}(\mathbf{T})}(\mathbf{X}), \\ \mathbf{X}_5 &= \text{p}_{\mathbb{T}^{\{2,3\}}(\mathbf{T}) \oplus \mathbb{T}^{\{2,3,4\}}(\mathbf{T})}(\mathbf{X}), \\ \mathbf{X}_6 &= \text{p}_{\mathbb{U}^{\{1,2\}}(\mathbf{T})}(\mathbf{X}). \end{aligned}$$

It is not difficult to see that $\mathbf{X} = \sum_{i=1}^6 \mathbf{X}_i$ and $\|\mathbf{X}_i\|_\sigma \leq \|\mathbf{X}\|_\sigma \leq \frac{1}{3}$ for $i \in [6]$ by Lemma 3.2.

Let $\mathbf{v}_k \in \mathbb{S}^{n_k}$ for $k \in [4]$ such that

$$\begin{aligned} \|\mathbf{Z} + \mathbf{X}\|_\sigma &= \langle \mathbf{Z} + \mathbf{X}, \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \mathbf{v}_3 \otimes \mathbf{v}_4 \rangle \\ &= \langle \mathbf{Z}, \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \mathbf{v}_3 \otimes \mathbf{v}_4 \rangle + \sum_{i=1}^6 \langle \mathbf{X}_i, \mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \mathbf{v}_3 \otimes \mathbf{v}_4 \rangle \\ &\leq a_1b_1c_1d_1 + \frac{1}{3}(a_2b_1c_1d_2 + a_1b_2c_1d_2 + a_1b_1c_2d_2 + a_2b_1c_2 + a_1b_2c_2 + a_2b_2), \end{aligned}$$

where the inequality is due to that $\|\mathbf{Z}\|_\sigma = 1$ and $\mathbf{Z} \in \mathbb{T}(\mathbf{T})$ as $\mathbf{Z} \in \mathbb{Z}(\mathbf{T})$, $\|\mathbf{X}_i\|_\sigma \leq \frac{1}{3}$ for $i \in [6]$, and

$$\begin{aligned} a_1 &= \|\text{p}_{\text{sp}_1(\mathbf{T})}(\mathbf{v}_1)\|_2, & b_1 &= \|\text{p}_{\text{sp}_2(\mathbf{T})}(\mathbf{v}_2)\|_2, & c_1 &= \|\text{p}_{\text{sp}_3(\mathbf{T})}(\mathbf{v}_3)\|_2, & d_1 &= \|\text{p}_{\text{sp}_4(\mathbf{T})}(\mathbf{v}_4)\|_2, \\ a_2 &= \|\text{p}_{\text{sp}_1^\perp(\mathbf{T})}(\mathbf{v}_1)\|_2, & b_2 &= \|\text{p}_{\text{sp}_2^\perp(\mathbf{T})}(\mathbf{v}_2)\|_2, & c_2 &= \|\text{p}_{\text{sp}_3^\perp(\mathbf{T})}(\mathbf{v}_3)\|_2, & d_2 &= \|\text{p}_{\text{sp}_4^\perp(\mathbf{T})}(\mathbf{v}_4)\|_2. \end{aligned}$$

Finally, by applying a similar proof to that of Lemma B.2, we obtain

$$\max_{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d} \in \mathbb{S}^2 \cap \mathbb{R}_+^2} \left(a_1b_1c_1d_1 + \frac{1}{3}(a_2b_1c_1d_2 + a_1b_2c_1d_2 + a_1b_1c_2d_2 + a_2b_1c_2 + a_1b_2c_2 + a_2b_2) \right) = 1,$$

implying that $\|\mathbf{Z} + \mathbf{X}\|_\sigma \leq 1$. $\square$

References

- [1] G. Ballard and T. G. Kolda. *Tensor Decompositions for Data Science*. Cambridge University Press, Cambridge, forthcoming.
- [2] S. Banach. Über homogene polynome in (L^2) . *Studia Mathematica*, 7(1):36–44, 1938.
- [3] B. Barak and A. Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *Proceedings of the 29th Annual Conference on Learning Theory*, pages 417–445, 2016.
- [4] B. Barak and A. Moitra. Noisy tensor completion via the sum-of-squares hierarchy. *Mathematical Programming*, 193(2):513–548, 2022.
- [5] A. Beck. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, 2017.
- [6] J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research and Financial Engineering. Springer Science & Business Media, New York, 2013.
- [7] N. Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, Cambridge, 2023.
- [8] N. Boumal and A. D. McRae. The usual smooth lift of the nuclear norm regularizer enjoys $2 \Rightarrow 1$. <https://www.racetothetbottom.xyz/posts/lift-regularizer-nuclear/>, 2024.
- [9] S. Burer and R. D. C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.
- [10] S. Burer and R. D. C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.
- [11] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011.
- [12] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.
- [13] 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.
- [14] B. Chen, S. He, Z. Li, and S. Zhang. Maximum block improvement and polynomial optimization. *SIAM Journal on Optimization*, 22(1):87–107, 2012.
- [15] B. Chen and Z. Li. On the tensor spectral p -norm and its dual norm via partitions. *Computational Optimization and Applications*, 75(3):609–628, 2020.
- [16] Y. Cui, C. Ding, and X. Zhao. Quadratic growth conditions for convex matrix optimization problems associated with spectral functions. *SIAM Journal on Optimization*, 27(4):2332–2355, 2017.
- [17] H. Derksen. On the nuclear norm and the singular value decomposition of tensors. *Foundations of Computational Mathematics*, 16(3):779–811, 2016.

- [18] A. L. Dontchev and R. T. Rockafellar. *Implicit Functions and Solution Mappings: A View from Variational Analysis*, volume 543 of *Springer Monographs in Mathematics*. Springer Science & Business Media, New York, first edition, 2009.
- [19] D. Driggs, S. Becker, and J. Boyd-Graber. Tensor robust principal component analysis: Better recovery with atomic norm regularization. *arXiv preprint arXiv:1901.10991*, 2019.
- [20] S. Friedland and L.-H. Lim. Nuclear norm of higher-order tensors. *Mathematics of Computation*, 87(311):1255–1281, 2018.
- [21] Q. Gao, P. Zhang, W. Xia, D. Xie, X. Gao, and D. Tao. Enhanced tensor RPCA and its application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2133–2140, 2020.
- [22] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.
- [23] J. Guan, S. He, B. Jiang, and Z. Li. ℓ_p -sphere covering and approximating nuclear p -norm. *Mathematics of Operations Research*, 2024. To appear.
- [24] J. He, C. Kan, and W. Song. Twice epi-differentiability of orthogonally invariant matrix functions and application. *arXiv preprint arXiv:2412.09898*, 2024.
- [25] S. He, H. Hu, B. Jiang, and Z. Li. Approximating tensor norms via sphere covering: Bridging the gap between primal and dual. *SIAM Journal on Optimization*, 33(3):2062–2088, 2023.
- [26] S. He, Z. Li, and S. Zhang. Approximation algorithms for homogeneous polynomial optimization with quadratic constraints. *Mathematical Programming*, 125(2):353–383, 2010.
- [27] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, second edition, 2012.
- [28] H. Hu, B. Jiang, and Z. Li. On spectral and nuclear norms of order three tensors with one fixed dimension. *SIAM Journal on Matrix Analysis and Applications*, 46(1):210–231, 2025.
- [29] 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.
- [30] B. Jiang, S. Ma, and S. Zhang. Low-M-rank tensor completion and robust tensor PCA. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1390–1404, 2018.
- [31] M. E. Kilmer, K. Braman, N. Hao, and R. C. Hoover. Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging. *SIAM Journal on Matrix Analysis and Applications*, 34(1):148–172, 2013.
- [32] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.
- [33] A. S. Lewis and H. S. Sendov. Nonsmooth analysis of singular values. Part I: Theory. *Set-Valued Analysis*, 13:213–241, 2005.
- [34] Q. Li, Z. Zhu, and G. Tang. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96, 2019.

- [35] Z. Li. Bounds on the spectral norm and the nuclear norm of a tensor based on tensor partitions. *SIAM Journal on Matrix Analysis and Applications*, 37(4):1440–1452, 2016.
- [36] Z. Li, Y. Nakatsukasa, T. Soma, and A. Uschmajew. On orthogonal tensors and best rank-one approximation ratio. *SIAM Journal on Matrix Analysis and Applications*, 39(1):400–425, 2018.
- [37] L.-H. Lim. Singular values and eigenvalues of tensors: A variational approach. In *Proceedings of the 1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, pages 129–132, 2005.
- [38] L.-H. Lim and P. Comon. Blind multilinear identification. *IEEE Transactions on Information Theory*, 60(2):1260–1280, 2013.
- [39] 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, 2012.
- [40] C. Lu. Transforms based tensor robust PCA: Corrupted low-rank tensors recovery via convex optimization. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*, pages 1145–1152, 2021.
- [41] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(04):925–938, 2020.
- [42] A. D. McRae. Low solution rank of the matrix LASSO under RIP with consequences for rank-constrained algorithms. *arXiv preprint arXiv:2404.12828*, 2024.
- [43] A. Mohammadi and E. Sarabi. Parabolic regularity of spectral functions. *Mathematics of Operations Research*, 50(3):2017–2046, 2025.
- [44] B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*, volume 330 of *Grundlehren der Mathematischen Wissenschaften, A Series of Comprehensive Studies in Mathematics*. Springer-Verlag Berlin, Heidelberg, first edition, 2006.
- [45] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.
- [46] J. Nie. *Moment and Polynomial Optimization*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, 2023.
- [47] W. Ouyang, T. K. Pong, and M.-C. Yue. Burer-Monteiro factorizability of nuclear norm regularized optimization. *arXiv preprint arXiv:2505.00349*, 2025.
- [48] V. Pappayan, X. Y. Han, and D. L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.
- [49] L. Qi. Eigenvalues of a real supersymmetric tensor. *Journal of Symbolic Computation*, 40(6):1302–1324, 2005.
- [50] G. Raskutti, M. Yuan, and H. Chen. Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics*, 47(3):1554–1584, 2019.

- [51] D. Rhea. The case of equality in the von Neumann trace inequality. <https://web.archive.org/web/20110723103501/http://www.drhea.net/wp-content/uploads/2011/01/vonNeumann.pdf>, 2011.
- [52] R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, Princeton, New Jersey, 1997.
- [53] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften, A Series of Comprehensive Studies in Mathematics*. Springer Berlin, Heidelberg, 2009.
- [54] R. Tomioka and T. Suzuki. Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*, 2014.
- [55] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.
- [56] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018.
- [57] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019.
- [58] G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992.
- [59] J. Wright and Y. Ma. *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. Cambridge University Press, Cambridge, 2022.
- [60] D. Xia and M. Yuan. On polynomial time methods for exact low-rank tensor completion. *Foundations of Computational Mathematics*, 19(6):1265–1313, 2019.
- [61] M. Yuan and C.-H. Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068, 2016.
- [62] M. Yuan and C.-H. Zhang. Incoherent tensor norms and their applications in higher order tensor completion. *IEEE Transactions on Information Theory*, 63(10):6753–6766, 2017.
- [63] Z. Zhou and A. M.-C. So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165:689–728, 2017.
- [64] Z. Zhou and Y. Zhu. Sparse random tensors: Concentration, regularization and applications. *Electronic Journal of Statistics*, 15(1):2483–2516, 2021.
- [65] Z. Zhu, T. Ding, J. Zhou, X. Li, C. You, J. Sulam, and Q. Qu. A geometric analysis of neural collapse with unconstrained features. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 29820–29834, 2021.