On Tensor Completion via Nuclear Norm Minimization

Ming Yuan* and Cun-Hui Zhang†
University of Wisconsin-Madison and Rutgers University

(April 21, 2015)

Abstract

Many problems can be formulated as recovering a low-rank tensor. Although an increasingly common task, tensor recovery remains a challenging problem because of the delicacy associated with the decomposition of higher order tensors. To overcome these difficulties, existing approaches often proceed by unfolding tensors into matrices and then apply techniques for matrix completion. We show here that such matricization fails to exploit the tensor structure and may lead to suboptimal procedure. More specifically, we investigate a convex optimization approach to tensor completion by directly minimizing a tensor nuclear norm and prove that this leads to an improved sample size requirement. To establish our results, we develop a series of algebraic and probabilistic techniques such as characterization of subdifferetial for tensor nuclear norm and concentration inequalities for tensor martingales, which may be of independent interests and could be useful in other tensor related problems.

*Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706. The research of Ming Yuan was supported in part by NSF Career Award DMS-1321692 and FRG Grant DMS-1265202, and NIH Grant 1U54AI117924-01.

†Department of Statistics and Biostatistics, Rutgers University, Piscataway, New Jersey 08854. The research of Cun-Hui Zhang was supported in part by NSF Grants DMS-1129626 and DMS-1209014

1 Introduction

Let $\boldsymbol{T} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ be an Nth order tensor, and Ω be a randomly sampled subset of $[d_1] \times \cdots \times [d_N]$ where $[d] = \{1, 2, \ldots, d\}$. The goal of tensor completion is to recover $\boldsymbol{T}$ when observing only entries $\boldsymbol{T}(\omega)$ for $\omega \in \Omega$. In particular, we are interested in the case when the dimensions $d_1, \ldots, d_N$ are large. Such a problem arises naturally in many applications. Examples include hyper-spectral image analysis (Li and Li, 2010), multi-energy computed tomography (Semerci et al., 2013), radar signal processing (Sidiropoulos and Nion, 2010), audio classification (Mesgarani, Slaney and Shamma, 2006) and text mining (Cohen and Collins, 2012) among numerous others. Common to these and many other problems, the tensor $\boldsymbol{T}$ can oftentimes be identified with a certain low-rank structure. The low-rankness entails reduction in degrees of freedom, and as a result, it is possible to recover $\boldsymbol{T}$ exactly even when the sample size $|\Omega|$ is much smaller than the total number, $d_1 d_2 \cdots d_N$, of entries in $\boldsymbol{T}$.

In particular, when $N = 2$, this becomes the so-called matrix completion problem which has received considerable amount of attention in recent years. See, e.g., Candès and Recht (2008), Candès and Tao (2009), Recht (2010), and Gross (2011) among many others. An especially attractive approach is through nuclear norm minimization:

$$\min_{\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}} \|\boldsymbol{X}\|_* \qquad \text{subject to } \boldsymbol{X}(\omega) = \boldsymbol{T}(\omega) \quad \forall \omega \in \Omega,$$

where the nuclear norm $\|\cdot\|_*$ of a matrix is given by

$$\|\boldsymbol{X}\|_* = \sum_{k=1}^{\min\{d_1, d_2\}} \sigma_k(\boldsymbol{X}),$$

and $\sigma_k(\cdot)$ stands for the kth largest singular value of a matrix. Denote by $\widehat{\boldsymbol{T}}$ the solution to the aforementioned nuclear norm minimization problem. As shown, for example, by Gross (2011), if an unknown $d_1 \times d_2$ matrix $\boldsymbol{T}$ of rank r is of low coherence with respect to the canonical basis, then it can be perfectly reconstructed by $\widehat{\boldsymbol{T}}$ with high probability whenever $|\Omega| \geq C(d_1 + d_2) r \log^2(d_1 + d_2)$, where C is a numerical constant. In other words, perfect recovery of a matrix is possible with observations from a very small fraction of entries in $\boldsymbol{T}$.

In many practical situations, we need to consider higher order tensors. The seemingly innocent task of generalizing these ideas from matrices to higher order tensor completion problems, however, turns out to be rather subtle, as basic notion such as rank, or singular value decomposition, becomes ambiguous for higher order tensors (e.g., Kolda and Bader, 2009; Hillar and Lim, 2013). A common strategy to overcome the challenges in dealing with high order tensors is to unfold them to matrices, and then resort to the usual nuclear norm minimization heuristics for matrices. To fix ideas, we shall focus on third order tensors ($N = 3$) in the rest of the paper although our techniques can be readily used to treat higher order tensor. Following the matricization approach, $\boldsymbol{T}$ can be reconstructed by the solution of the following convex program:

$$\min_{\boldsymbol{X}\in\mathbb{R}^{d_1\times d_2\times d_3}} \{\|X^{(1)}\|_* + \|X^{(2)}\|_* + \|X^{(3)}\|_*\} \qquad \text{subject to } \boldsymbol{X}(\omega) = \boldsymbol{T}(\omega) \quad \forall \omega \in \Omega,$$

where $X^{(j)}$ is a $d_j \times (\prod_{k\neq j} d_k)$ matrix whose columns are the mode-j fibers of $\boldsymbol{X}$. See, e.g., Liu et al. (2009), Signoretto, Lathauwer and Suykens (2010), Gandy et al. (2011), Tomioka, Hayashi and Kashima (2010), and Tomioka et al. (2011). In the light of existing results on matrix completion, with this approach, $\boldsymbol{T}$ can be reconstructed perfectly with high probability provided that

$$|\Omega| \geq C(d_1d_2r_3 + d_1r_2d_3 + r_1d_2d_3)\log^2(d_1 + d_2 + d_3)$$

uniformly sampled entries are observed, where r_j is the rank of $X^{(j)}$ and C is a numerical constant. See, e.g., Mu et al. (2013). It is of great interests to investigate if this sample size requirement can be improved by avoiding matricization of tensors. We show here that the answer indeed is affirmative and a more direct nuclear norm minimization formulation requires a smaller sample size to recover $\boldsymbol{T}$.

More specifically, write, for two tensors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d_1\times d_2\times d_3}$,

$$\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \sum_{\omega\in[d_1]\times[d_2]\times[d_3]} \boldsymbol{X}(\omega)\boldsymbol{Y}(\omega)$$

as their inner product. Define

$$\|\boldsymbol{X}\| = \max_{\boldsymbol{u}_j\in\mathbb{R}^{d_j}:\|\boldsymbol{u}_1\|=\|\boldsymbol{u}_2\|=\|\boldsymbol{u}_3\|=1} \langle \boldsymbol{X}, \boldsymbol{u}_1\otimes\boldsymbol{u}_2\otimes\boldsymbol{u}_3\rangle,$$

where, with slight abuse of notation, $\|\cdot\|$ also stands for the usual Euclidean norm for a vector, and for vectors $\boldsymbol{u}_j = (u_1^j, \ldots, u_{d_j}^j)^\top$,

$$\boldsymbol{u}_1\otimes\boldsymbol{u}_2\otimes\boldsymbol{u}_3 = (u_a^1u_b^2u_c^3)_{1\leq a\leq d_1, 1\leq b\leq d_2, 1\leq c\leq d_3}.$$

It is clear that the $\|\cdot\|$ defined above for tensors is a norm and can be viewed as an extension of the usual matrix spectral norm. Appealing to the duality between the spectral norm and nuclear norm in the matrix case, we now consider the following nuclear norm for tensors:

$$\|\boldsymbol{X}\|_* = \max_{\boldsymbol{Y}\in\mathbb{R}^{d_1\times d_2\times d_3}:\|\boldsymbol{Y}\|\leq 1} \langle \boldsymbol{Y}, \boldsymbol{X}\rangle.$$

It is clear that $\|\cdot\|_*$ is also a norm. We then consider reconstructing $\boldsymbol{T}$ via the solution to the following convex program:

$$\min_{\boldsymbol{X}\in\mathbb{R}^{d_1\times d_2\times d_3}} \|\boldsymbol{X}\|_* \qquad \text{subject to } \boldsymbol{X}(\omega)=\boldsymbol{T}(\omega) \quad \forall\omega\in\Omega.$$

We show that the sample size requirement for perfect recovery of a tensor with low coherence using this approach is

$$|\Omega| \geq C\left(r^2(d_1+d_2+d_3)+\sqrt{rd_1d_2d_3}\right)\text{polylog}\left(d_1+d_2+d_3\right),$$

where

$$r=\sqrt{(r_1r_2d_3+r_1r_3d_2+r_2r_3d_1)/(d_1+d_2+d_3)},$$

polylog(x) is a certain polynomial function of $\log(x)$, and C is a numerical constant. In particular, when considering (nearly) cubic tensors with d_1, d_2 and d_3 approximately equal to a common d, then this sample size requirement is essentially of the order $r^{1/2}(d\log d)^{3/2}$. In the case when the tensor dimension d is large while the rank r is relatively small, this can be a drastic improvement over the existing results based on matricizing tensors where the sample size requirement is $r(d\log d)^2$.

The high-level strategy to the investigation of the proposed nuclear norm minimization approach for tensors is similar, in a sense, to the treatment of matrix completion. Yet the analysis for tensors is much more delicate and poses significant new challenges because many of the well-established tools for matrices, either algebraic such as characterization of the subdifferential of the nuclear norm, or probabilistic such as concentration inequalities for martingales, do not exist for tensors. Some of these disparities can be bridged and we develop various tools to do so. Others are due to fundamental differences between matrices and higher order tensors, and we devise new strategies to overcome them. The tools and techniques we developed may be of independent interests and can be useful in dealing with other problems for tensors.

The rest of the paper is organized as follows. We first describe some basic properties of tensors and their nuclear norm necessary for our analysis in Section 2. Section 3 discusses the main architect of our analysis. The main probabilistic tools we use are concentration bounds for the sum of random tensors. Because the tensor spectral norm does not have the interpretation as an operator norm of a linear mapping between Hilbert spaces, the usual matrix Bernstein inequality cannot be directly applied. It turns out that different strategies are required for tensors of low rank and tensors with sparse support, and these results are presented in Sections 4 and 5 respectively. We conclude the paper with a few remarks in Section 6.

2 Tensor

We first collect some useful algebraic facts for tensors essential to our later analysis. Recall that the inner product between two third order tensors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d_1\times d_2\times d_3}$ is given by

$$\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \sum_{a=1}^{d_1}\sum_{b=1}^{d_2}\sum_{c=1}^{d_3} \boldsymbol{X}(a,b,c)\boldsymbol{Y}(a,b,c),$$

and $\|\boldsymbol{X}\|_{\rm HS} = \langle \boldsymbol{X}, \boldsymbol{X}\rangle^{1/2}$ is the usual Hilbert-Schmidt norm of $\boldsymbol{X}$. Another tensor norm of interest is the entrywise ℓ_∞ norm, or tensor max norm:

$$\|\boldsymbol{X}\|_{\max} = \max_{\omega\in[d_1]\times[d_2]\times[d_3]} |\boldsymbol{X}(\omega)|.$$

It is clear that for any the third order tensor $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2\times d_3}$,

$$\|\boldsymbol{X}\|_{\max} \le \|\boldsymbol{X}\| \le \|\boldsymbol{X}\|_{\rm HS} \le \|\boldsymbol{X}\|_*, \quad \text{and} \quad \|\boldsymbol{X}\|_{\rm HS}^2 \le \|\boldsymbol{X}\|_*\|\boldsymbol{X}\|.$$

We shall also encounter linear maps defined on tensors. Let $\mathcal{R} : \mathbb{R}^{d_1\times d_2\times d_3} \to \mathbb{R}^{d_1\times d_2\times d_3}$ be a linear map. We define the induced operator norm of $\mathcal{R}$ under tensor Hilbert-Schmidt norm as

$$\|\mathcal{R}\| = \max\left\{\|\mathcal{R}\boldsymbol{X}\|_{\rm HS} : \boldsymbol{X} \in \mathbb{R}^{d_1\times d_2\times d_3}, \|\boldsymbol{X}\|_{\rm HS} \le 1\right\}.$$

2.1 Decomposition and Projection

Consider the following tensor decomposition of $\boldsymbol{X}$ into rank-one tensors:

$$\boldsymbol{X} = [A, B, C] := \sum_{k=1}^{r} \boldsymbol{a}_k \otimes \boldsymbol{b}_k \otimes \boldsymbol{c}_k, \tag{1}$$

where $\boldsymbol{a}_k$s, $\boldsymbol{b}_k$s and $\boldsymbol{c}_k$s are the column vectors of matrices A, B and C respectively. Such a decomposition in general is not unique (see, e.g., Kruskal, 1989). However, the linear spaces spanned by columns of A, B and C respectively are uniquely defined.

More specifically, write $\boldsymbol{X}(\cdot, b, c) = (\boldsymbol{X}(1, b, c), \dots, \boldsymbol{X}(d_1, b, c))^\top$, that is the mode-1 fiber of $\boldsymbol{X}$. Define $\boldsymbol{X}(a, \cdot, c)$ and $\boldsymbol{X}(a, b, \cdot)$ in a similar fashion. Let

$$\begin{aligned}
\mathcal{L}_1(\boldsymbol{X}) &= \text{l.s.}\{\boldsymbol{X}(\cdot, b, c) : 1 \le b \le d_2, 1 \le c \le d_3\};\\
\mathcal{L}_2(\boldsymbol{X}) &= \text{l.s.}\{\boldsymbol{X}(a, \cdot, c) : 1 \le a \le d_1, 1 \le c \le d_3\};\\
\mathcal{L}_3(\boldsymbol{X}) &= \text{l.s.}\{\boldsymbol{X}(a, b, \cdot) : 1 \le a \le d_1, 1 \le b \le d_2\},
\end{aligned}$$

where l.s. represents the linear space spanned by a collection of vectors of conformable dimension. Then it is clear that the linear space spanned by the column vectors of A is

$\mathcal{L}_1(\boldsymbol{X})$, and similar statements hold true for the column vectors of B and C. In the case of matrices, both marginal linear spaces, $\mathcal{L}_1$ and $\mathcal{L}_2$ are necessarily of the same dimension as they are spanned by the respective singular vectors. For higher order tensors, however, this is typically not true. We shall denote by $r_j(\boldsymbol{X})$ the dimension of $\mathcal{L}_j(\boldsymbol{X})$ for $j = 1, 2$ and 3, which are often referred to the Tucker ranks of $\boldsymbol{X}$. Another useful notion of "tensor rank" for our purposes is

$$\overline{r}(\boldsymbol{X}) = \sqrt{\left(r_1(\boldsymbol{X})r_2(\boldsymbol{X})d_3 + r_1(\boldsymbol{X})r_3(\boldsymbol{X})d_2 + r_2(\boldsymbol{X})r_3(\boldsymbol{X})d_1\right)/d}.$$

where $d = d_1 + d_2 + d_3$, which can also be viewed as a generalization of the matrix rank to tensors. It is well known that the smallest value for r in the rank-one decomposition (1) is in $[\overline{r}(\boldsymbol{X}), \overline{r}^2(\boldsymbol{X})]$. See, e.g., Kolda and Bader (2009).

Let M be a matrix of size $d_0 \times d_1$. Marginal multiplication of M and a tensor $\boldsymbol{X}$ in the first coordinate yields a tensor of size $d_0 \times d_2 \times d_3$:

$$(M \times_1 \boldsymbol{X})(a, b, c) = \sum_{a'=1}^{d_1} M_{aa'} X(a', b, c).$$

It is easy to see that if $\boldsymbol{X} = [A, B, C]$, then $M \times_1 \boldsymbol{X} = [MA, B, C]$. Marginal multiplications $\times_2$ and $\times_3$ between a matrix of conformable size and $\boldsymbol{X}$ can be similarly defined.

Let $\boldsymbol{P}$ be arbitrary projection from $\mathbb{R}^{d_1}$ to a linear subspace of $\mathbb{R}^{d_1}$. It is clear from the definition of marginal multiplications, $[\boldsymbol{P}A, B, C]$ is also uniquely defined for tensor $\boldsymbol{X} = [A, B, C]$, that is, $[\boldsymbol{P}A, B, C]$ does not depend on the particular decomposition of A, B, C. Now let $\boldsymbol{P}_j$ be arbitrary projection from $\mathbb{R}^{d_j}$ to a linear subspace of $\mathbb{R}^{d_j}$. Define a tensor projection $\boldsymbol{P}_1 \otimes \boldsymbol{P}_2 \otimes \boldsymbol{P}_3$ on $\boldsymbol{X} = [A, B, C]$ as

$$(\boldsymbol{P}_1 \otimes \boldsymbol{P}_2 \otimes \boldsymbol{P}_3)\boldsymbol{X} = [\boldsymbol{P}_1 A, \boldsymbol{P}_2 B, \boldsymbol{P}_3 C].$$

We note that there is no ambiguity in defining $(\boldsymbol{P}_1 \otimes \boldsymbol{P}_2 \otimes \boldsymbol{P}_3)\boldsymbol{X}$ because of the uniqueness of marginal projections.

Recall that $\mathcal{L}_j(\boldsymbol{X})$ is the linear space spanned by the mode-j fibers of $\boldsymbol{X}$. Let $\boldsymbol{P}^j_{\boldsymbol{X}}$ be the projection from $\mathbb{R}^{d_j}$ to $\mathcal{L}_j(\boldsymbol{X})$, and $\boldsymbol{P}^j_{\boldsymbol{X}^\perp}$ be the projection to its orthogonal complement in $\mathbb{R}^{d_j}$. The following tensor projections will be used extensively in our analysis:

$$\begin{aligned}
\mathcal{Q}^0_{\boldsymbol{X}} &= \boldsymbol{P}^1_{\boldsymbol{X}} \otimes \boldsymbol{P}^2_{\boldsymbol{X}} \otimes \boldsymbol{P}^3_{\boldsymbol{X}}, & \mathcal{Q}^0_{\boldsymbol{X}^\perp} &= \boldsymbol{P}^1_{\boldsymbol{X}^\perp} \otimes \boldsymbol{P}^2_{\boldsymbol{X}^\perp} \otimes \boldsymbol{P}^3_{\boldsymbol{X}^\perp}, \\
\mathcal{Q}^1_{\boldsymbol{X}} &= \boldsymbol{P}^1_{\boldsymbol{X}^\perp} \otimes \boldsymbol{P}^2_{\boldsymbol{X}} \otimes \boldsymbol{P}^3_{\boldsymbol{X}}, & \mathcal{Q}^1_{\boldsymbol{X}^\perp} &= \boldsymbol{P}^1_{\boldsymbol{X}} \otimes \boldsymbol{P}^2_{\boldsymbol{X}^\perp} \otimes \boldsymbol{P}^3_{\boldsymbol{X}^\perp}, \\
\mathcal{Q}^2_{\boldsymbol{X}} &= \boldsymbol{P}^1_{\boldsymbol{X}} \otimes \boldsymbol{P}^2_{\boldsymbol{X}^\perp} \otimes \boldsymbol{P}^3_{\boldsymbol{X}}, & \mathcal{Q}^2_{\boldsymbol{X}^\perp} &= \boldsymbol{P}^1_{\boldsymbol{X}^\perp} \otimes \boldsymbol{P}^2_{\boldsymbol{X}} \otimes \boldsymbol{P}^3_{\boldsymbol{X}^\perp}, \\
\mathcal{Q}^3_{\boldsymbol{X}} &= \boldsymbol{P}^1_{\boldsymbol{X}} \otimes \boldsymbol{P}^2_{\boldsymbol{X}} \otimes \boldsymbol{P}^3_{\boldsymbol{X}^\perp}, & \mathcal{Q}^3_{\boldsymbol{X}^\perp} &= \boldsymbol{P}^1_{\boldsymbol{X}^\perp} \otimes \boldsymbol{P}^2_{\boldsymbol{X}^\perp} \otimes \boldsymbol{P}^3_{\boldsymbol{X}}, \\
\mathcal{Q}_{\boldsymbol{X}} &= \mathcal{Q}^0_{\boldsymbol{X}} + \mathcal{Q}^1_{\boldsymbol{X}} + \mathcal{Q}^2_{\boldsymbol{X}} + \mathcal{Q}^3_{\boldsymbol{X}}, & \mathcal{Q}_{\boldsymbol{X}^\perp} &= \mathcal{Q}^0_{\boldsymbol{X}^\perp} + \mathcal{Q}^1_{\boldsymbol{X}^\perp} + \mathcal{Q}^2_{\boldsymbol{X}^\perp} + \mathcal{Q}^3_{\boldsymbol{X}^\perp}.
\end{aligned}$$

2.2 Subdifferential of Tensor Nuclear Norm

One of the main technical tools in analyzing the nuclear norm minimization is the characterization of the subdifferntial of the nuclear norm. Such results are well known in the case of matrices. In particular, let $M = UDV^\top$ be the singular value decomposition of a matrix M, then the subdifferential of the nuclear norm at M is given by

$$\partial\|\cdot\|_*(M) = \{UV^\top + W : U^\top W = W^\top V = 0, \quad \text{and} \quad \|W\| \le 1\},$$

where with slight abuse of notion, $\|\cdot\|_*$ and $\|\cdot\|$ are the nuclear and spectral norms of matrices. See, e.g., Watson (1992). In other words, for any other matrix Y of conformable dimensions,

$$\|Y\|_* \ge \|M\|_* + \langle UV^\top + W, Y - M\rangle$$

if and only if $U^\top W = W^\top V = 0$ and $\|W\| \le 1$. Characterizing the subdifferential of the nuclear norm for higher order tensors is more subtle due to the lack of corresponding spectral decomposition.

A straightforward generalization of the above characterization may suggest that $\partial\|\cdot\|_*(\boldsymbol{X})$ be identified with

$$\{\boldsymbol{W} + \boldsymbol{W}^\perp : \boldsymbol{W}^\perp = \mathcal{Q}_{\boldsymbol{X}^\perp}\boldsymbol{W}^\perp, \|\boldsymbol{W}^\perp\| \le 1\},$$

for some $\boldsymbol{W}$ in the range of $\mathcal{Q}^0_{\boldsymbol{X}}$. It turns out that this in general is not true. As a simple counterexample, let

$$\boldsymbol{X} = \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 \otimes \boldsymbol{e}_1,$$

and

$$\boldsymbol{Y} = \sum_{1\le i,j,k\le 2} \boldsymbol{e}_i \otimes \boldsymbol{e}_j \otimes \boldsymbol{e}_k = (\boldsymbol{e}_1 + \boldsymbol{e}_2) \otimes (\boldsymbol{e}_1 + \boldsymbol{e}_2) \otimes (\boldsymbol{e}_1 + \boldsymbol{e}_2),$$

where $d_1 = d_2 = d_3 = 2$ and $\boldsymbol{e}_i$'s are the canonical basis of an Euclidean space. It is clear that $\|\boldsymbol{X}\|_* = 1$ and $\|\boldsymbol{Y}\|_* = 2\sqrt{2}$. Take $\boldsymbol{W}^\perp = \boldsymbol{U}/\|\boldsymbol{U}\|$ where

$$\boldsymbol{U} = \boldsymbol{e}_1 \otimes \boldsymbol{e}_2 \otimes \boldsymbol{e}_2 + \boldsymbol{e}_2 \otimes \boldsymbol{e}_1 \otimes \boldsymbol{e}_2 + \boldsymbol{e}_2 \otimes \boldsymbol{e}_2 \otimes \boldsymbol{e}_1. \tag{2}$$

As we shall show in the proof of Lemma 1 below, $\|\boldsymbol{U}\| = 2/\sqrt{3}$. It is clear that $\boldsymbol{W}^\perp = \mathcal{Q}_{\boldsymbol{X}^\perp}\boldsymbol{W}^\perp$ and $\|\boldsymbol{W}^\perp\| \le 1$. Yet,

$$\|\boldsymbol{X}\|_* + \langle \boldsymbol{Y} - \boldsymbol{X}, \boldsymbol{W} + \boldsymbol{W}^\perp\rangle = 1 + \langle \boldsymbol{Y} - \boldsymbol{X}, \boldsymbol{W}^\perp\rangle = 1 + 3\sqrt{3}/2 > 2\sqrt{2} = \|\boldsymbol{Y}\|_*,$$

for any $\boldsymbol{W}$ such that $\boldsymbol{W} = \mathcal{Q}^0_{\boldsymbol{X}}\boldsymbol{W}$.

Fortunately, for our purposes, the following relaxed characterization is sufficient.

Lemma 1 *For any third order tensor* $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$*, there exists a* $\boldsymbol{W} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ *such that* $\boldsymbol{W} = \mathcal{Q}^0_{\boldsymbol{X}} \boldsymbol{W}$*,* $\|\boldsymbol{W}\| = 1$ *and*

$$\|\boldsymbol{X}\|_* = \langle \boldsymbol{W}, \boldsymbol{X} \rangle.$$

Furthermore, for any $\boldsymbol{Y} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ *and* $\boldsymbol{W}^{\perp} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ *obeying* $\|\boldsymbol{W}^{\perp}\| \le 1/2$*,*

$$\|\boldsymbol{Y}\|_* \ge \|\boldsymbol{X}\|_* + \langle \boldsymbol{W} + \mathcal{Q}_{\boldsymbol{X}^\perp} \boldsymbol{W}^{\perp}, \boldsymbol{Y} - \boldsymbol{X} \rangle.$$

Proof of Lemma 1. If $\|\boldsymbol{W}\| \le 1$, then

$$\max_{\|\boldsymbol{u}_j\|=1} \langle \mathcal{Q}^0_{\boldsymbol{X}} \boldsymbol{W}, \boldsymbol{u}_1 \otimes \boldsymbol{u}_2 \otimes \boldsymbol{u}_3 \rangle \le \max_{\|\boldsymbol{u}_j\|=1} \langle \boldsymbol{W}, (\boldsymbol{P}^1_{\boldsymbol{X}} \boldsymbol{u}_1) \otimes (\boldsymbol{P}^2_{\boldsymbol{X}} \boldsymbol{u}_2) \otimes (\boldsymbol{P}^3_{\boldsymbol{X}} \boldsymbol{u}_3) \rangle \le 1.$$

It follows that

$$\|\boldsymbol{X}\|_* = \max_{\|\boldsymbol{W}\|=1} \langle \boldsymbol{W}, \boldsymbol{X} \rangle = \max_{\|\boldsymbol{W}\|=1} \langle \mathcal{Q}^0_{\boldsymbol{X}} \boldsymbol{W}, \boldsymbol{X} \rangle$$

is attained with a certain $\boldsymbol{W}$ satisfying $\|\boldsymbol{W}\| = \|\mathcal{Q}^0_{\boldsymbol{X}} \boldsymbol{W}\| = 1$.

Now consider a tensor $\boldsymbol{W}^{\perp}$ satisfying

$$\|\boldsymbol{W} + \mathcal{Q}_{\boldsymbol{X}^\perp} \boldsymbol{W}^{\perp}\| \le 1.$$

Because $\mathcal{Q}_{\boldsymbol{X}^\perp} \boldsymbol{X} = 0$, it follows from the definition of the tensor nuclear norm that

$$\langle \boldsymbol{W} + \mathcal{Q}_{\boldsymbol{X}^\perp} \boldsymbol{W}^{\perp}, \boldsymbol{Y} - \boldsymbol{X} \rangle \le \|\boldsymbol{Y}\|_* - \langle \boldsymbol{W}, \boldsymbol{X} \rangle = \|\boldsymbol{Y}\|_* - \|\boldsymbol{X}\|_*.$$

It remains to prove that $\|\boldsymbol{W}^{\perp}\| \le 1/2$ implies

$$\|\boldsymbol{W} + \mathcal{Q}_{\boldsymbol{X}^\perp} \boldsymbol{W}^{\perp}\| \le 1.$$

Recall that $\boldsymbol{W} = \mathcal{Q}^0_{\boldsymbol{X}} \boldsymbol{W}$, $\|\boldsymbol{W}^{\perp}\| \le 1/2$, and $\|\boldsymbol{u}_j\| = 1$. Then

$$\begin{aligned} & \langle \boldsymbol{W} + \mathcal{Q}_{\boldsymbol{X}^\perp} \boldsymbol{W}^{\perp}, \boldsymbol{u}_1 \otimes \boldsymbol{u}_2 \otimes \boldsymbol{u}_3 \rangle \\ \le\ & \|\mathcal{Q}^0_{\boldsymbol{X}} (\boldsymbol{u}_1 \otimes \boldsymbol{u}_2 \otimes \boldsymbol{u}_3)\|_* + \frac{1}{2} \|\mathcal{Q}_{\boldsymbol{X}^\perp} (\boldsymbol{u}_1 \otimes \boldsymbol{u}_2 \otimes \boldsymbol{u}_3)\|_* \\ \le\ & \prod_{j=1}^{3} \sqrt{1 - a_j^2} + \frac{1}{2} \Big(a_1 a_2 + a_1 a_3 \sqrt{1 - a_2^2} + a_2 a_3 \sqrt{1 - a_1^2} \Big). \end{aligned}$$

where $a_j = \|\boldsymbol{P}^j_{\boldsymbol{X}^\perp} \boldsymbol{u}_j\|_2$, for $j = 1, 2, 3$. Let $x = a_1 a_2$ and

$$y = \sqrt{(1 - a_1^2)(1 - a_2^2)}.$$

We have

$$\begin{aligned}\left(a_1\sqrt{1-a_2^2}+a_2\sqrt{1-a_1^2}\right)^2 &= a_1^2(1-a_2^2)+a_2^2(1-a_1^2)+2xy\\ &= a_1^2+a_2^2-2a_1^2a_2^2+2xy\\ &= 1-(y-x)^2.\end{aligned}$$

It follows that for any value of $a_3 \in (0,1)$,

$$\langle \boldsymbol{W}+\mathcal{Q}_{\boldsymbol{X}^\perp}\boldsymbol{W}^\perp, \boldsymbol{u}_1\otimes\boldsymbol{u}_2\otimes\boldsymbol{u}_3\rangle \le y\sqrt{1-a_3^2}+\frac{1}{2}\Big(x+a_3\sqrt{1-(y-x)^2}\Big).$$

This function of (x,y) is increasing in the smaller of x and y. For $x<y$, the maximum of x^2 given y^2 is attained when $a_1=a_2$ by simple calculation with the Lagrange multiplier. Similarly, for $y<x$, the maximum of y^2 given x^2 is attained when $a_1=a_2$. Thus, setting $a_1=a_2=a$, we find

$$\langle \boldsymbol{W}+\mathcal{Q}_{\boldsymbol{X}^\perp}\boldsymbol{W}^\perp, \boldsymbol{u}_1\otimes\boldsymbol{u}_2\otimes\boldsymbol{u}_3\rangle \le \max_{a_3,a}\Big\{(1-a^2)\sqrt{1-a_3^2}+\frac{1}{2}\Big(a^2+2a_3a\sqrt{1-a^2}\Big)\Big\}.$$

The above maximum is attained when $a_3=a$. Because $\sqrt{1-a^2}+a^2/2\le 1$, we have

$$\langle \boldsymbol{W}+\mathcal{Q}_{\boldsymbol{X}^\perp}\boldsymbol{W}^\perp, \boldsymbol{u}_1\otimes\boldsymbol{u}_2\otimes\boldsymbol{u}_3\rangle \le 1,$$

which completes the proof of the lemma.

The norm of $\boldsymbol{U}$ defined in (2) can be computed using a similar argument:

$$\begin{aligned}\|\boldsymbol{U}\| &= \max_{0\le a_1\le a_2\le a_3\le 1}\left(a_1a_2\sqrt{1-a_3^2}+a_1a_3\sqrt{1-a_2^2}+a_2a_3\sqrt{1-a_1^2}\right)\\ &= \max_{x,y,a_3}\left(x\sqrt{1-a_3^2}+a_3\sqrt{1-(y-x)^2}\right)\\ &= \max_{a,a_3} a\Big(a\sqrt{1-a_3^2}+2a_3\sqrt{1-a^2}\Big),\\ &= \max_a a\sqrt{a^2+4(1-a^2)},\end{aligned}$$

which yields $\|\boldsymbol{U}\|=2/\sqrt{3}$ with $a^2=2/3$. □

Note that Lemma 1 gives only sufficient conditions of the subgradient of tensor nuclear norm. Equivalently it states that

$$\partial\|\cdot\|_*(\boldsymbol{X}) \supseteq \big\{\boldsymbol{W}+\mathcal{Q}_{\boldsymbol{X}^\perp}\boldsymbol{W}^\perp : \|\boldsymbol{W}^\perp\|\le 1/2\big\}.$$

We note also that the constant 1/2 may be further improved. No attempt has been made here to sharpen the constant as it already suffices for our analysis.

2.3 Coherence

A central concept to matrix completion is coherence. Recall that the coherence of an r dimensional linear subspace U of $\mathbb{R}^k$ is defined as

$$\mu(U) = \frac{k}{r} \max_{1\le i\le k} \|\boldsymbol{P}_U \boldsymbol{e}_i\|^2 = \frac{\max_{1\le i\le k} \|\boldsymbol{P}_U \boldsymbol{e}_i\|^2}{k^{-1}\sum_{i=1}^k \|\boldsymbol{P}_U \boldsymbol{e}_i\|^2},$$

where $\boldsymbol{P}_U$ is the orthogonal projection onto U and $\boldsymbol{e}_i$'s are the canonical basis for $\mathbb{R}^k$. See, e.g., Candès and Recht (2008). We shall define the coherence of a tensor $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2\times d_3}$ as

$$\mu(\boldsymbol{X}) = \max\{\mu(\mathcal{L}_1(\boldsymbol{X})), \mu(\mathcal{L}_2(\boldsymbol{X})), \mu(\mathcal{L}_3(\boldsymbol{X}))\}.$$

It is clear that $\mu(\boldsymbol{X}) \ge 1$, since $\mu(U)$ is the ratio of the ℓ_∞ and length-normalized ℓ_2 norms of a vector.

Lemma 2 *Let $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2\times d_3}$ be a third order tensor. Then*

$$\max_{a,b,c} \|\mathcal{Q}_{\boldsymbol{X}}(\boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c)\|_{\mathrm{HS}}^2 \le \frac{\overline{r}^2(\boldsymbol{X})d}{d_1d_2d_3}\mu^2(\boldsymbol{X}).$$

Proof of Lemma 2. Recall that $\mathcal{Q}_{\boldsymbol{X}} = \mathcal{Q}^0_{\boldsymbol{X}} + \mathcal{Q}^1_{\boldsymbol{X}} + \mathcal{Q}^2_{\boldsymbol{X}} + \mathcal{Q}^3_{\boldsymbol{X}}$. Therefore,

$$\begin{aligned}
\|\mathcal{Q}_{\boldsymbol{X}}(\boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c)\|^2 &= \sum_{j,k=0}^{3} \langle \mathcal{Q}^j_{\boldsymbol{X}}(\boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c), \mathcal{Q}^k_{\boldsymbol{X}}(\boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c)\rangle \\
&= \sum_{j=0}^{3} \|\mathcal{Q}^j_{\boldsymbol{X}}(\boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c)\|^2 \\
&= \|\boldsymbol{P}^1_{\boldsymbol{X}}\boldsymbol{e}_a\|^2\|\boldsymbol{P}^2_{\boldsymbol{X}}\boldsymbol{e}_b\|^2\|\boldsymbol{P}^3_{\boldsymbol{X}}\boldsymbol{e}_c\|^2 + \|\boldsymbol{P}^1_{\boldsymbol{X}^\perp}\boldsymbol{e}_a\|^2\|\boldsymbol{P}^2_{\boldsymbol{X}}\boldsymbol{e}_b\|^2\|\boldsymbol{P}^3_{\boldsymbol{X}}\boldsymbol{e}_c\|^2 \\
&\quad + \|\boldsymbol{P}^1_{\boldsymbol{X}}\boldsymbol{e}_a\|^2\|\boldsymbol{P}^2_{\boldsymbol{X}^\perp}\boldsymbol{e}_b\|^2\|\boldsymbol{P}^3_{\boldsymbol{X}}\boldsymbol{e}_c\|^2 + \|\boldsymbol{P}^1_{\boldsymbol{X}}\boldsymbol{e}_a\|^2\|\boldsymbol{P}^2_{\boldsymbol{X}}\boldsymbol{e}_b\|^2\|\boldsymbol{P}^3_{\boldsymbol{X}^\perp}\boldsymbol{e}_c\|^2.
\end{aligned}$$

For brevity, write $r_j = r_j(\boldsymbol{X})$, and $\mu = \mu(\boldsymbol{X})$. Then

$$\begin{aligned}
\|\boldsymbol{P}^1_{\boldsymbol{X}}\boldsymbol{e}_a\|^2\|\boldsymbol{P}^2_{\boldsymbol{X}}\boldsymbol{e}_b\|^2\|\boldsymbol{P}^3_{\boldsymbol{X}}\boldsymbol{e}_c\|^2 + \|\boldsymbol{P}^1_{\boldsymbol{X}^\perp}\boldsymbol{e}_a\|^2\|\boldsymbol{P}^2_{\boldsymbol{X}}\boldsymbol{e}_b\|^2\|\boldsymbol{P}^3_{\boldsymbol{X}}\boldsymbol{e}_c\|^2 &\le \frac{r_2r_3\mu^2}{d_2d_3}; \\
\|\boldsymbol{P}^1_{\boldsymbol{X}}\boldsymbol{e}_a\|^2\|\boldsymbol{P}^2_{\boldsymbol{X}}\boldsymbol{e}_b\|^2\|\boldsymbol{P}^3_{\boldsymbol{X}}\boldsymbol{e}_c\|^2 + \|\boldsymbol{P}^1_{\boldsymbol{X}}\boldsymbol{e}_a\|^2\|\boldsymbol{P}^2_{\boldsymbol{X}^\perp}\boldsymbol{e}_b\|^2\|\boldsymbol{P}^3_{\boldsymbol{X}}\boldsymbol{e}_c\|^2 &\le \frac{r_1r_3\mu^2}{d_1d_3}; \\
\|\boldsymbol{P}^1_{\boldsymbol{X}}\boldsymbol{e}_a\|^2\|\boldsymbol{P}^2_{\boldsymbol{X}}\boldsymbol{e}_b\|^2\|\boldsymbol{P}^3_{\boldsymbol{X}}\boldsymbol{e}_c\|^2 + \|\boldsymbol{P}^1_{\boldsymbol{X}}\boldsymbol{e}_a\|^2\|\boldsymbol{P}^2_{\boldsymbol{X}}\boldsymbol{e}_b\|^2\|\boldsymbol{P}^3_{\boldsymbol{X}^\perp}\boldsymbol{e}_c\|^2 &\le \frac{r_1r_2\mu^2}{d_1d_2}.
\end{aligned}$$

As a result, for any $(a,b,c) \in [d_1]\times[d_2]\times[d_3]$,

$$\|\mathcal{Q}_{\boldsymbol{X}}(\boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c)\|^2 \le \frac{\mu^2(r_1r_2d_3 + r_1d_2r_3 + d_1r_2r_3)}{d_1d_2d_3},$$

which implies the desired statement. $\square$

Another measure of coherence for a tensor $\boldsymbol{X}$ is

$$\alpha(\boldsymbol{X}) := \sqrt{d_1 d_2 d_3/\overline{r}(\boldsymbol{X})}\|\boldsymbol{W}\|_{\max}$$

where $\boldsymbol{W}$ is such that $\boldsymbol{W} = \mathcal{Q}^0_{\boldsymbol{X}}\boldsymbol{W}$, $\|\boldsymbol{W}\| = 1$ and $\langle \boldsymbol{X}, \boldsymbol{W}\rangle = \|\boldsymbol{X}\|_*$ as described in Lemma 1. The quantity $\alpha(\boldsymbol{X})$ is related to $\mu(\boldsymbol{X})$ defined earlier and the spikiness

$$\widetilde{\alpha}(\boldsymbol{X}) = \sqrt{d_1 d_2 d_3}\|\boldsymbol{W}\|_{\max}/\|\boldsymbol{W}\|_{\mathrm{HS}}.$$

Lemma 3 *Let $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2\times d_3}$ be a third order tensor. Assume without loss of generality that $r_1(\boldsymbol{X}) \le r_2(\boldsymbol{X}) \le r_3(\boldsymbol{X})$. Then,*

$$\alpha^2(\boldsymbol{X}) \le \min\left\{r_1(\boldsymbol{X})r_2(\boldsymbol{X})r_3(\boldsymbol{X})\mu^3(\boldsymbol{X})/\overline{r}(\boldsymbol{X}), r_1(\boldsymbol{X})r_2(\boldsymbol{X})\widetilde{\alpha}^2(\boldsymbol{X})/\overline{r}(\boldsymbol{X})\right\}.$$

Moreover, if $\boldsymbol{X}$ admits a bi-orthogonal eigentensor decomposition $\sum_{i=1}^r \lambda_i(\boldsymbol{u}_i\otimes\boldsymbol{v}_i\otimes\boldsymbol{w}_i)$ with $\lambda_i \ne 0$ and $\boldsymbol{u}_i^\top\boldsymbol{u}_j = \boldsymbol{v}_i^\top\boldsymbol{v}_j\boldsymbol{w}_i^\top\boldsymbol{w}_j = \mathbb{I}\{i=j\}$ for $1 \le i,j \le r$, then $r_1(\boldsymbol{X}) = r_3(\boldsymbol{X}) = \overline{r}(\boldsymbol{X}) = r$, $\|X^{(1)}\|_* = \|\boldsymbol{X}\|_*$, and

$$\alpha(\boldsymbol{X}) = \widetilde{\alpha}(\boldsymbol{X}) \ge 1.$$

Proof of Lemma 3. Due to the conditions $\boldsymbol{W} = \mathcal{Q}^0_{\boldsymbol{X}}\boldsymbol{W}$ and $\|\boldsymbol{W}\| = 1$,

$$\begin{aligned}\|\boldsymbol{W}\|^2_{\max} &= \max_{a,b,c}|\langle \boldsymbol{W}, (\boldsymbol{P}^1_{\boldsymbol{X}}\boldsymbol{e}_a)\otimes(\boldsymbol{P}^2_{\boldsymbol{X}}\boldsymbol{e}_b)\otimes(\boldsymbol{P}^3_{\boldsymbol{X}}\boldsymbol{e}_c)\rangle|^2 \\ &\le \max_{a,b,c}\|\boldsymbol{P}^1_{\boldsymbol{X}}\boldsymbol{e}_a\|^2\|\boldsymbol{P}^2_{\boldsymbol{X}}\boldsymbol{e}_b\|^2\|\boldsymbol{P}^3_{\boldsymbol{X}}\boldsymbol{e}_c\|^2 \\ &\le r_1(\boldsymbol{X})r_2(\boldsymbol{X})r_3(\boldsymbol{X})\mu^3(\boldsymbol{X})/(d_1d_2d_3),\end{aligned}$$

which yields the upper bound for $\alpha(\boldsymbol{X})$ in terms of $\mu(\boldsymbol{X})$.

Because $\boldsymbol{W}$ is in the range of $\mathcal{Q}^0_{\boldsymbol{X}}$, $\mathcal{L}_j(\boldsymbol{W}) \subseteq \mathcal{L}_j(\boldsymbol{X})$. Therefore, $r_1(\boldsymbol{W}) \le r_1(\boldsymbol{X})$. Recall that $W^{(1)}$ is a $d_1 \times (d_2d_3)$ matrix whose columns are the mode-1 fibers of $\boldsymbol{W}$. Applying singular value decomposition to $W^{(1)}$ suggests that there are orthornomal vectors $\{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{r_1}\}$ in $\mathbb{R}^{d_1}$ and matrices $M_1,\ldots,M_{r_1} \in \mathbb{R}^{d_2\times d_3}$ such that $\langle M_j, M_k\rangle = 0$ if $j \ne k$, and

$$\boldsymbol{W} = \sum_{k=1}^{r_1(\boldsymbol{X})}\boldsymbol{u}_k\otimes M_k.$$

It is clear that $\|M_k\| \le \|\boldsymbol{W}\| = 1$, and $\mathrm{rank}(M_k) \le r_2(\boldsymbol{X})$. Therefore,

$$\|\boldsymbol{W}\|^2_{\mathrm{HS}} \le \sum_{k=1}^{r_1(\boldsymbol{X})}\|M_k\|^2_{\mathrm{HS}} \le r_1(\boldsymbol{X})r_2(\boldsymbol{X}).$$

This gives the upper bound for $\alpha(\boldsymbol{X})$ in terms of $\widetilde{\alpha}(\boldsymbol{X})$.

It remains to consider the case of $\boldsymbol{X} = \sum_{i=1}^r \lambda_i(\boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{w}_i)$. Obviously, by triangular inequality,

$$\|\boldsymbol{X}\|_* \le \sum_{i=1}^r \|\boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{w}_i\|_* = \sum_{i=1}^r \|\boldsymbol{w}_i\|.$$

On the other hand, let

$$\boldsymbol{W} = \sum_{i=1}^r \boldsymbol{u}_i \otimes \boldsymbol{v}_i \otimes (\boldsymbol{w}_i/\|\boldsymbol{w}_i\|).$$

Because

$$\|\boldsymbol{W}\| \le \max_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}:\|\boldsymbol{a}\|,\|\boldsymbol{b}\|,\|\boldsymbol{c}\|\le 1} \sum_{i=1}^n \left|(\boldsymbol{a}^\top \boldsymbol{u}_i)\text{trace}(\boldsymbol{c}\boldsymbol{b}^\top \boldsymbol{v}_i \boldsymbol{w}_i^\top/\|\boldsymbol{w}_i\|)\right| \le \|\boldsymbol{a}\|\|\boldsymbol{b}\boldsymbol{c}^\top\|_{\rm HS} \le 1,$$

we find

$$\|\boldsymbol{X}\|_* \ge \langle \boldsymbol{W}, \boldsymbol{X}\rangle = \sum_{i=1}^r \|\boldsymbol{w}_i\|,$$

which implies that $\boldsymbol{W}$ is dual to $\boldsymbol{X}$ and

$$\|\boldsymbol{X}\|_* = \sum_{i=1}^r \|\boldsymbol{w}_i\|,$$

where the rightmost hand side also equals to $\|X^{(1)}\|_*$ and $\|X^{(2)}\|_*$. The last statement now follows from the fact that $\|\boldsymbol{W}\|_{\rm HS}^2 = r$. □

As in the matrix case, exact recovery with observations on a small fraction of the entries is only possible for tensors with low coherence. In particular, we consider in this article the recovery of a tensor $\boldsymbol{T}$ obeying $\mu(\boldsymbol{T}) \le \mu_0$ and $\alpha(\boldsymbol{T}) \le \alpha_0$ for some $\mu_0, \alpha_0 \ge 1$.

3 Exact Tensor Recovery

We are now in position to study the nuclear norm minimization for tensor completion. Let $\widehat{\boldsymbol{T}}$ be the solution to

$$\min_{\boldsymbol{X}\in\mathbb{R}^{d_1\times d_2\times d_3}} \|\boldsymbol{X}\|_* \qquad \text{subject to } \mathcal{P}_\Omega \boldsymbol{X} = \mathcal{P}_\Omega \boldsymbol{T}, \tag{3}$$

where $\mathcal{P}_\Omega : \mathbb{R}^{d_1\times d_2\times d_3} \mapsto \mathbb{R}^{d_1\times d_2\times d_3}$ such that

$$(\mathcal{P}_\Omega \boldsymbol{X})(i,j,k) = \begin{cases} \boldsymbol{X}(i,j,k) & \text{if } (i,j,k)\in\Omega \\ 0 & \text{otherwise} \end{cases}.$$

Assume that Ω is a uniformly sampled subset of $[d_1]\times[d_2]\times[d_3]$. The goal is to determine what the necessary sample size is for successful reconstruction of $\boldsymbol{T}$ using $\widehat{\boldsymbol{T}}$ with high probability. In particular, we show that that with high probability, exact recovery can be achieved with nuclear norm minimization (3) if

$$|\Omega| \geq \left(\alpha_0\sqrt{rd_1d_2d_3} + \mu_0^2 r^2 d\right) \text{polylog}(d),$$

where $d = d_1 + d_2 + d_3$. More specifically, we have

Theorem 1 *Assume that $\mu(\boldsymbol{T}) \leq \mu_0$, $\alpha(\boldsymbol{T}) \leq \alpha_0$, and $\overline{r}(\boldsymbol{T}) = r$. Let Ω be a uniformly sampled subset of $[d_1] \times [d_2] \times [d_3]$ and $\widehat{\boldsymbol{T}}$ be the solution to (3). For $\beta > 0$, define*

$$q_1^* = \big(\beta + \log d\big)^2 \alpha_0^2 r \log d, \quad q_2^* = (1+\beta)(\log d)\mu_0^2 r^2.$$

Let $n = |\Omega|$. Suppose that for a sufficiently large numerical constant c_0,

$$n \geq c_0 \delta_2^{-1} \left[\sqrt{q_1^*(1+\beta)\delta_1^{-1} d_1 d_2 d_3} + q_1^* d^{1+\delta_1} + q_2^* d^{1+\delta_2}\right] \tag{4}$$

with certain $\{\delta_1, \delta_2\} \in [1/\log d, 1/2]$ and $\beta > 0$. Then,

$$\mathbb{P}\left\{\widehat{\boldsymbol{T}} \neq \boldsymbol{T}\right\} \leq d^{-\beta}. \tag{5}$$

In particular, for $\delta_1 = \delta_2 = (\log d)^{-1}$, (4) can be written as

$$n \geq C_{\mu_0,\alpha_0,\beta}\Big[(\log d)^3 \sqrt{rd_1d_2d_3} + \big\{r(\log d)^3 + r^2(\log d)\big\}d\Big]$$

with a constant $C_{\mu_0,\alpha_0,\beta}$ depending on $\{\mu_0, \alpha_0, \beta\}$ only.

For $d_1 \asymp d_2 \asymp d_3$ and fixed $\{\alpha_0, \mu_0, \delta_1, \delta_2, \beta\}$, the sample size requirement (4) becomes

$$n \asymp \sqrt{r}(d \log d)^{3/2},$$

provided $\max\{r(\log d)^3 d^{2\delta_1}, r^3 d^{2\delta_2}/(\log d)\} = O(d)$.

The high level idea of our strategy is similar to the matrix case – exact recovery of $\boldsymbol{T}$ is implied by the existence of a dual certificate $\boldsymbol{G}$ supported on Ω, that is $\mathcal{P}_\Omega \boldsymbol{G} = \boldsymbol{G}$, such that $\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{G} = \boldsymbol{W}$ and $\|\mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{G}\| < 1/2$. See, e.g., Gross (2011) and Recht (2011).

3.1 Recovery with a Dual Certificate

Write $\widehat{\boldsymbol{T}} = \boldsymbol{T} + \Delta$. Then, $\mathcal{P}_\Omega \Delta = 0$ and

$$\|\boldsymbol{T} + \Delta\|_* \leq \|\boldsymbol{T}\|_*.$$

Recall that, by Lemma 1, there exists a $\boldsymbol{W}$ obeying $\boldsymbol{W} = \mathcal{Q}^0_{\boldsymbol{T}}\boldsymbol{W}$ and $\|\boldsymbol{W}\| = 1$ such that

$$\|\boldsymbol{T} + \Delta\|_* \geq \|\boldsymbol{T}\|_* + \langle \boldsymbol{W} + \mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{W}^\perp, \Delta\rangle$$

for any $\boldsymbol{W}^\perp$ obeying $\|\boldsymbol{W}^\perp\| \leq 1/2$. Assume that a tensor $\boldsymbol{G}$ supported on Ω, that is $\mathcal{P}_\Omega\boldsymbol{G} = \boldsymbol{G}$, such that $\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{G} = \boldsymbol{W}$, and $\|\mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{G}\| < 1/2$. When $\|\mathcal{Q}_{\boldsymbol{T}^\perp}\Delta\|_* > 0$,

$$\begin{aligned}
\langle \boldsymbol{W} + \mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{W}^\perp, \Delta\rangle &= \langle \boldsymbol{W} + \mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{W}^\perp - \boldsymbol{G}, \Delta\rangle \\
&= \langle \boldsymbol{W} - \mathcal{Q}_{\boldsymbol{T}}\boldsymbol{G}, \Delta\rangle + \langle \boldsymbol{W}^\perp, \mathcal{Q}_{\boldsymbol{T}^\perp}\Delta\rangle - \langle \mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{G}, \mathcal{Q}_{\boldsymbol{T}^\perp}\Delta\rangle \\
&> \langle \boldsymbol{W}^\perp, \mathcal{Q}_{\boldsymbol{T}^\perp}\Delta\rangle - \frac{1}{2}\|\mathcal{Q}_{\boldsymbol{T}^\perp}\Delta\|_*
\end{aligned}$$

Take $\boldsymbol{W}^\perp = \boldsymbol{U}/2$ where

$$\boldsymbol{U} = \underset{\boldsymbol{X}:\|\boldsymbol{X}\|\leq 1}{\arg\max}\langle \boldsymbol{X}, \mathcal{Q}_{\boldsymbol{T}^\perp}\Delta\rangle.$$

We find that $\|\mathcal{Q}_{\boldsymbol{T}^\perp}\Delta\|_* > 0$ implies

$$\|\boldsymbol{T} + \Delta\|_* - \|\boldsymbol{T}\|_* \geq \langle \boldsymbol{W} + \mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{W}^\perp, \Delta\rangle > 0,$$

which contradicts with fact that $\widehat{\boldsymbol{T}}$ minimizes the nuclear norm. Thus, $\mathcal{Q}_{\boldsymbol{T}^\perp}\Delta = 0$, which then implies $\mathcal{Q}_{\boldsymbol{T}}\mathcal{Q}_\Omega\mathcal{Q}_{\boldsymbol{T}}\Delta = \mathcal{Q}_{\boldsymbol{T}}\mathcal{Q}_\Omega\Delta = 0$. When $\mathcal{Q}_{\boldsymbol{T}}\mathcal{Q}_\Omega\mathcal{Q}_{\boldsymbol{T}}$ is invertible in the range of $\mathcal{Q}_{\boldsymbol{T}}$, we also have $\mathcal{Q}_{\boldsymbol{T}}\Delta = 0$ and $\widehat{\boldsymbol{T}} = \boldsymbol{T}$.

With this in mind, it then suffices to seek such a dual certificate. In fact, it turns out that finding an "approximate" dual certificate is actually enough for our purposes.

Lemma 4 *Assume that*

$$\inf\left\{\|\mathcal{P}_\Omega\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}\|_{\rm HS} : \|\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}\|_{\rm HS} = 1\right\} \geq \sqrt{\frac{n}{2d_1d_2d_3}}. \tag{6}$$

If there exists a tensor $\boldsymbol{G}$ supported on Ω such that

$$\|\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{G} - \boldsymbol{W}\|_{\rm HS} < \frac{1}{4}\sqrt{\frac{n}{2d_1d_2d_3}} \qquad \text{and} \qquad \max_{\|\mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{X}\|_*=1}\langle \boldsymbol{G}, \mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{X}\rangle \leq 1/4, \tag{7}$$

then $\widehat{\boldsymbol{T}} = \boldsymbol{T}$.

Proof of Lemma 4. Write $\widehat{\boldsymbol{T}} = \boldsymbol{T} + \boldsymbol{\Delta}$, then $\mathcal{P}_\Omega\boldsymbol{\Delta} = 0$ and

$$\|\boldsymbol{T} + \boldsymbol{\Delta}\|_* \leq \|\boldsymbol{T}\|_*.$$

Recall that, by Lemma 1, there exists a $\boldsymbol{W}$ obeying $\boldsymbol{W} = \mathcal{Q}^0_{\boldsymbol{T}}\boldsymbol{W}$ and $\|\boldsymbol{W}\| = 1$ such that for any $\boldsymbol{W}^\perp$ obeying $\|\boldsymbol{W}^\perp\| \leq 1/2$,

$$\|\boldsymbol{T} + \boldsymbol{\Delta}\|_* \geq \|\boldsymbol{T}\|_* + \langle \boldsymbol{W} + \mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{W}^\perp, \boldsymbol{\Delta}\rangle.$$

Since $\langle \boldsymbol{G}, \boldsymbol{\Delta}\rangle = \langle \mathcal{P}_\Omega \boldsymbol{G}, \boldsymbol{\Delta}\rangle = \langle \boldsymbol{G}, \mathcal{P}_\Omega \boldsymbol{\Delta}\rangle = 0$ and $\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{W} = \boldsymbol{W}$,

$$\begin{aligned} 0 &\geq \langle \boldsymbol{W} + \mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{W}^\perp, \boldsymbol{\Delta}\rangle \\ &= \langle \boldsymbol{W} + \mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{W}^\perp - \boldsymbol{G}, \boldsymbol{\Delta}\rangle \\ &= \langle \mathcal{Q}_{\boldsymbol{T}}\boldsymbol{W} - \mathcal{Q}_{\boldsymbol{T}}\boldsymbol{G}, \boldsymbol{\Delta}\rangle + \langle \boldsymbol{W}^\perp, \mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{\Delta}\rangle - \langle \boldsymbol{G}, \mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{\Delta}\rangle \\ &\geq -\|\boldsymbol{W} - \mathcal{Q}_{\boldsymbol{T}}\boldsymbol{G}\|_{\rm HS}\|\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{\Delta}\|_{\rm HS} + \langle \boldsymbol{W}^\perp, \mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{\Delta}\rangle - \frac{1}{4}\|\mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{\Delta}\|_*. \end{aligned}$$

In particular, taking $\boldsymbol{W}^\perp$ satisfying $\|\boldsymbol{W}\| = 1/2$ and $\langle \boldsymbol{W}^\perp, \mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{\Delta}\rangle = \|\mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{\Delta}\|_*/2$, we find

$$\frac{1}{4}\|\mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{\Delta}\|_* \leq \|\boldsymbol{W} - \mathcal{Q}_{\boldsymbol{T}}\boldsymbol{G}\|_{\rm HS}\|\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{\Delta}\|_{\rm HS}.$$

Recall that $\mathcal{P}_\Omega\boldsymbol{\Delta} = \mathcal{P}_\Omega\mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{\Delta} + \mathcal{P}_\Omega\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{\Delta} = 0$. Thus, in view of the condition on $\mathcal{P}_\Omega$,

$$\frac{\|\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{\Delta}\|_{\rm HS}}{\sqrt{2d_1d_2d_3/n}} \leq \|\mathcal{P}_\Omega\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{\Delta}\|_{\rm HS} = \|\mathcal{P}_\Omega\mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{\Delta}\|_{\rm HS} \leq \|\mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{\Delta}\|_{\rm HS} \leq \|\mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{\Delta}\|_*. \tag{8}$$

Consequently,

$$\frac{1}{4}\|\mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{\Delta}\|_* \leq \sqrt{2d_1d_2d_3/n}\|\boldsymbol{W} - \mathcal{Q}_{\boldsymbol{T}}\boldsymbol{G}\|_{\rm HS}\|\mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{\Delta}\|_*.$$

Since

$$\sqrt{2d_1d_2d_3/n}\|\boldsymbol{W} - \mathcal{Q}_{\boldsymbol{T}}\boldsymbol{G}\|_{\rm HS} < 1/4,$$

we have $\|\mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{\Delta}\|_* = 0$. Together with (8), we conclude that $\boldsymbol{\Delta} = 0$, or equivalently $\widehat{\boldsymbol{T}} = \boldsymbol{T}$. $\square$

Equation (6) indicates the invertibility of $\mathcal{P}_\Omega$ when restricted to the range of $\mathcal{Q}_{\boldsymbol{T}}$. We argue first that this is true for "incoherent" tensors. To this end, we prove that

$$\Big\|\mathcal{Q}_{\boldsymbol{T}}\Big((d_1d_2d_3/n)\mathcal{P}_\Omega - \mathcal{I}\Big)\mathcal{Q}_{\boldsymbol{T}}\Big\| \leq 1/2$$

with high probability. This implies that as an operator in the range of $\mathcal{Q}_{\boldsymbol{T}}$, the spectral norm of $(d_1d_2d_3/n)\mathcal{Q}_{\boldsymbol{T}}\mathcal{P}_\Omega\mathcal{Q}_{\boldsymbol{T}}$ is contained in $[1/2, 3/2]$. Consequently, (6) holds because for any $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2\times d_3}$,

$$(d_1d_2d_3/n)\|\mathcal{P}_\Omega\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}\|_{\rm HS}^2 = \Big\langle \mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}, (d_1d_2d_3/n)\mathcal{Q}_{\boldsymbol{T}}\mathcal{P}_\Omega\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}\Big\rangle \geq \frac{1}{2}\|\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}\|_{\rm HS}^2.$$

Recall that $d = d_1 + d_2 + d_3$. We have

Lemma 5 *Assume $\mu(\boldsymbol{T}) \leq \mu_0$, $\overline{r}(\boldsymbol{T}) = r$, and Ω is uniformly sampled from $[d_1]\times[d_2]\times[d_3]$ without replacement. Then, for any $\tau > 0$,*

$$\mathbb{P}\left\{\big\|\mathcal{Q}_{\boldsymbol{T}}\big((d_1d_2d_3/n)\mathcal{P}_\Omega - \mathcal{I}\big)\mathcal{Q}_{\boldsymbol{T}}\big\| \geq \tau\right\} \leq 2r^2d\,\exp\Big(-\frac{\tau^2/2}{1+2\tau/3}\Big(\frac{n}{\mu_0^2r^2d}\Big)\Big).$$

In particular, taking $\tau = 1/2$ in Lemma 5 yields

$$\mathbb{P}\Big\{\text{ (6) holds }\Big\} \geq 1 - 2r^2d\,\exp\Big(-\frac{3}{32}\Big(\frac{n}{\mu_0^2r^2d}\Big)\Big).$$

3.2 Constructing a Dual Certificate

We now show that the "approximate" dual certificate as required by Lemma 4 can indeed be constructed. We use a strategy similar to the "golfing scheme" for the matrix case (see, e.g., Gross, 2011). Recall that Ω is a uniformly sampled subset of size $n = |\Omega|$ from $[d_1]\times[d_2]\times[d_3]$. The main idea is to construct an "approximate" dual certificate supported on a subset of Ω, and we do so by first constructing a random sequence with replacement from Ω. More specifically, we start by sampling (a_1, b_1, c_1) uniformly from Ω. We then sequentially sample (a_i, b_i, c_i) $(i = 2, \ldots, n)$ uniformly from the set of unique past observations, denoted by S_{i-1}, with probability $|S_{i-1}|/d_1d_2d_3$; and uniformly from $\Omega\setminus S_{i-1}$ with probability $1-|S_{i-1}|/d_1d_2d_3$. It is worth noting that, in general, there are replicates in the sequence $\{(a_j, b_j, c_j) : 1 \le j \le i\}$ and the set S_i consists of unique observations from the sequence so in general $|S_i| < i$. It is not hard to see that the sequence $(a_1, b_1, c_1), \ldots, (a_n, b_n, c_n)$ forms an independent and uniformly distributed (with replacement) sequence on $[d_1] \times [d_2] \times [d_3]$.

We now divide the sequence $\{(a_i, b_i, c_i) : 1 \le i \le n\}$ into n_2 subsequences of length n_1:

$$\Omega_k = \{(a_i, b_i, c_i) : (k-1)n_1 < i \le kn_1\},$$

for $k = 1, 2, \ldots, n_2$, where $n_1n_2 \le n$. Recall that $\boldsymbol{W}$ is such that $\boldsymbol{W} = \mathcal{Q}^0_{\boldsymbol{T}}\boldsymbol{W}$, $\|\boldsymbol{W}\| = 1$, and $\|\boldsymbol{T}\|_* = \langle \boldsymbol{T}, \boldsymbol{W}\rangle$. For brevity, write

$$\mathcal{P}_{(a,b,c)} : \mathbb{R}^{d_1\times d_2\times d_3} \to \mathbb{R}^{d_1\times d_2\times d_3}$$

as a linear operator that zeroes out all but the (a, b, c) entry of a tensor. Let

$$\mathcal{R}_k = \mathcal{I} - \frac{1}{n_1}\sum_{i=(k-1)n_1+1}^{kn_1}(d_1d_2d_3)\mathcal{P}_{(a_i,b_i,c_i)}$$

with $\mathcal{I}$ being the identity operator on tensors and define

$$\boldsymbol{G}_k = \sum_{\ell=1}^{k}\big(\mathcal{I} - \mathcal{R}_\ell\big)\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_{\ell-1}\mathcal{Q}_{\boldsymbol{T}}\cdots\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_1\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{W}, \quad \boldsymbol{G} = \boldsymbol{G}_{n_2}.$$

Since $(a_i, b_i, c_i) \in \Omega$, $\mathcal{P}_\Omega(\mathcal{I} - \mathcal{R}_k) = \mathcal{I} - \mathcal{R}_k$, so that $\mathcal{P}_\Omega\boldsymbol{G} = \boldsymbol{G}$. It follows from the definition of $\boldsymbol{G}_k$ that

$$\begin{aligned}\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{G}_k &= \sum_{\ell=1}^{k}(\mathcal{Q}_{\boldsymbol{T}} - \mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_\ell\mathcal{Q}_{\boldsymbol{T}})(\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_{\ell-1}\mathcal{Q}_{\boldsymbol{T}})\cdots(\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_1\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{W})\\ &= \boldsymbol{W} - (\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_k\mathcal{Q}_{\boldsymbol{T}})\cdots(\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_1\mathcal{Q}_{\boldsymbol{T}})\boldsymbol{W}\end{aligned}$$

and

$$\langle \boldsymbol{G}_k, \mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{X}\rangle = \Big\langle \sum_{\ell=1}^{k}\mathcal{R}_\ell(\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_{\ell-1}\mathcal{Q}_{\boldsymbol{T}})\cdots(\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_1\mathcal{Q}_{\boldsymbol{T}})\boldsymbol{W}, \mathcal{Q}_{\boldsymbol{T}^\perp}\boldsymbol{X}\Big\rangle.$$

Thus, condition (7) holds if

$$\|(\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_{n_2}\mathcal{Q}_{\boldsymbol{T}})\cdots(\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_1\mathcal{Q}_{\boldsymbol{T}})\boldsymbol{W}\|_{\rm HS} < \frac{1}{4}\sqrt{\frac{n}{2d_1d_2d_3}} \tag{9}$$

and

$$\Big\|\sum_{\ell=1}^{n_2}\mathcal{R}_\ell(\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_{\ell-1}\mathcal{Q}_{\boldsymbol{T}})\cdots(\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_1\mathcal{Q}_{\boldsymbol{T}})\boldsymbol{W}\Big\| < 1/4. \tag{10}$$

3.3 Verifying Conditions for Dual Certificate

We now prove that (9) and (10) hold with high probability for the approximate dual certificate constructed above. For this purpose, we need large deviation bounds for the average of certain iid tensors under the spectral and maximum norms.

Lemma 6 *Let $\{(a_i, b_i, c_i)\}$ be an independently and uniformly sampled sequence from $[d_1]\times[d_2]\times[d_3]$. Assume that $\mu(\boldsymbol{T})\le\mu_0$ and $\overline{r}(\boldsymbol{T})=r$. Then, for any fixed $k=1,2,\ldots,n_2$, and for all $\tau>0$,*

$$\mathbb{P}\Big\{\Big\|\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_k\mathcal{Q}_{\boldsymbol{T}}\Big\|\ge\tau\Big\}\le 2r^2d\exp\Big(-\frac{\tau^2/2}{1+2\tau/3}\Big(\frac{n_1}{\mu_0^2r^2d}\Big)\Big), \tag{11}$$

and

$$\max_{\|\boldsymbol{X}\|_{\max}=1}\mathbb{P}\Big\{\Big\|\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_k\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}\Big\|_{\max}\ge\tau\Big\}\le 2d_1d_2d_3\,\exp\Big(-\frac{\tau^2/2}{1+2\tau/3}\Big(\frac{n_1}{\mu_0^2r^2d}\Big)\Big). \tag{12}$$

Because

$$(d_1d_2d_3)^{-1/2}\|\boldsymbol{W}\|_{\rm HS}\le\|\boldsymbol{W}\|_{\max}\le\|\boldsymbol{W}\|\le 1,$$

Equation (9) holds if $\max_{1\le\ell\le n_2}\|\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_\ell\mathcal{Q}_{\boldsymbol{T}}\|\le\tau$ and

$$n_2\ge-\frac{1}{\log\tau}\log\Big(\sqrt{32}d_1d_2d_3n^{-1/2}\Big). \tag{13}$$

Thus, an application of (11) now gives the following bound:

$$\begin{aligned}\mathbb{P}\Big\{\text{ (9) holds }\Big\} &\ge 1-\mathbb{P}\{\|(\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_{n_2}\mathcal{Q}_{\boldsymbol{T}})\cdots(\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_1\mathcal{Q}_{\boldsymbol{T}})\|\ge\tau^{n_2}\}\\ &\le 1-\mathbb{P}\{\max_{1\le\ell\le n_2}\|\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_\ell\mathcal{Q}_{\boldsymbol{T}}\|\ge\tau\}\\ &\le 1-2n_2r^2d\exp\Big(-\frac{\tau^2/2}{1+2\tau/3}\Big(\frac{n_1}{\mu_0^2r^2d}\Big)\Big).\end{aligned}$$

Now consider Equation (10). Let $\boldsymbol{W}_\ell = \mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_\ell\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{W}_{\ell-1}$ for $\ell \geq 1$ with $\boldsymbol{W}_0 = \boldsymbol{W}$. Observe that (10) does not hold with at most probability

$$
\begin{aligned}
& \mathbb{P}\Big\{\Big\|\sum_{\ell=1}^{n_2}\mathcal{R}_\ell\boldsymbol{W}_{\ell-1}\Big\| \geq 1/4\Big\} \\
\leq\ & \mathbb{P}\Big\{\big\|\mathcal{R}_1\boldsymbol{W}_0\big\| \geq 1/8\Big\} + \mathbb{P}\Big\{\big\|\boldsymbol{W}_1\big\|_{\max} \geq \|\boldsymbol{W}\|_{\max}/2\Big\} \\
& +\mathbb{P}\Big\{\Big\|\sum_{\ell=2}^{n_2}\mathcal{R}_\ell\boldsymbol{W}_{\ell-1}\Big\| \geq 1/8,\ \big\|\boldsymbol{W}_1\big\|_{\max} < \|\boldsymbol{W}\|_{\max}/2\Big\} \\
\leq\ & \mathbb{P}\Big\{\big\|\mathcal{R}_1\boldsymbol{W}_0\big\| \geq 1/8\Big\} + \mathbb{P}\Big\{\big\|\boldsymbol{W}_1\big\|_{\max} \geq \|\boldsymbol{W}\|_{\max}/2\Big\} \\
& +\mathbb{P}\Big\{\big\|\mathcal{R}_2\boldsymbol{W}_1\big\| \geq 1/16,\ \big\|\boldsymbol{W}_1\big\|_{\max} < \|\boldsymbol{W}\|_{\max}/2\Big\} \\
& +\mathbb{P}\Big\{\big\|\boldsymbol{W}_2\big\|_{\max} \geq \|\boldsymbol{W}\|_{\max}/4,\ \big\|\boldsymbol{W}_1\big\|_{\max} < \|\boldsymbol{W}\|_{\max}/2\Big\} \\
& +\mathbb{P}\Big\{\Big\|\sum_{\ell=3}^{n_2}\mathcal{R}_\ell\boldsymbol{W}_{\ell-1}\Big\| \geq \frac{1}{16},\ \big\|\boldsymbol{W}_2\big\|_{\max} < \|\boldsymbol{W}\|_{\max}/4\Big\} \\
\leq\ & \sum_{\ell=1}^{n_2-1}\mathbb{P}\Big\{\big\|\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_\ell\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{W}_{\ell-1}\big\|_{\max} \geq \|\boldsymbol{W}\|_{\max}/2^\ell,\ \big\|\boldsymbol{W}_{\ell-1}\big\|_{\max} \leq \|\boldsymbol{W}\|_{\max}/2^{\ell-1}\Big\} \\
& +\sum_{\ell=1}^{n_2}\mathbb{P}\Big\{\big\|\mathcal{R}_\ell\boldsymbol{W}_{\ell-1}\big\| \geq 2^{-2-\ell},\ \big\|\boldsymbol{W}_{\ell-1}\big\|_{\max} \leq \|\boldsymbol{W}\|_{\max}/2^{\ell-1}\Big\}.
\end{aligned}
$$

Since $\{\mathcal{R}_\ell, \boldsymbol{W}_\ell\}$ are i.i.d., (12) with $\boldsymbol{X} = \boldsymbol{W}_{\ell-1}/\|\boldsymbol{W}_{\ell-1}\|_{\max}$ implies

$$
\begin{aligned}
& \mathbb{P}\Big\{\ (10)\text{ holds }\Big\} \\
\geq\ & 1 - n_2\max_{\substack{\boldsymbol{X}:\boldsymbol{X}=\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X} \\ \|\boldsymbol{X}\|_{\max}\leq 1}}\left(\mathbb{P}\Big\{\Big\|\mathcal{Q}_{\boldsymbol{T}}\mathcal{R}_1\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}\Big\|_{\max} > \frac{1}{2}\Big\} + \mathbb{P}\Big\{\Big\|\mathcal{R}_1\boldsymbol{X}\Big\| > \frac{1}{8\|\boldsymbol{W}\|_{\max}}\Big\}\right) \\
\geq\ & 1 - 2n_2d_1d_2d_3\exp\left(\frac{-(3/32)n_1}{\mu_0^2r^2d}\right) - n_2\max_{\substack{\boldsymbol{X}:\boldsymbol{X}=\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X} \\ \|\boldsymbol{X}\|_{\max}\leq\|\boldsymbol{W}\|_{\max}}}\mathbb{P}\Big\{\Big\|\mathcal{R}_1\boldsymbol{X}\Big\| > \frac{1}{8}\Big\}.
\end{aligned}
$$

The last term on the right hand side can be bounded using the following result.

Lemma 7 *Assume that $\alpha(\boldsymbol{T}) \leq \alpha_0$, $\overline{r}(\boldsymbol{T}) = r$ and $q_1^* = \big(\beta + \log d\big)^2\alpha_0^2 r\log d$. There exists a numerical constant $c_1 > 0$ such that for any constants $\beta > 0$ and $1/(\log d) \leq \delta_1 < 1$,*

$$n_1 \geq c_1\left[q_1^*d^{1+\delta_1} + \sqrt{q_1^*(1+\beta)\delta_1^{-1}d_1d_2d_3}\right] \tag{14}$$

implies

$$\max_{\substack{\boldsymbol{X}:\boldsymbol{X}=\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X} \\ \|\boldsymbol{X}\|_{\max}\leq\|\boldsymbol{W}\|_{\max}}}\mathbb{P}\Big\{\Big\|\mathcal{R}_1\boldsymbol{X}\Big\| \geq \frac{1}{8}\Big\} \leq d^{-\beta-1}, \tag{15}$$

where $\boldsymbol{W}$ is in the range of $\mathcal{Q}_{\boldsymbol{T}}^0$ such that $\|\boldsymbol{W}\| = 1$ and $\langle \boldsymbol{T}, \boldsymbol{W}\rangle = \|\boldsymbol{T}\|_*$.

3.4 Proof of Theorem 1

Since (7) is a consequence of (9) and (10), it follows from Lemmas 4, 5, 6 and 7 that for $\tau \in (0, 1/2]$ and $n \geq n_1 n_2$ satisfying conditions (13) and (14),

$$\begin{aligned}\mathbb{P}\Big\{\widehat{\boldsymbol{T}} \neq \boldsymbol{T}\Big\} \;\leq\;& 2r^2 d \exp\Big(-\frac{3}{32}\Big(\frac{n}{\mu_0^2 r^2 d}\Big)\Big) + 2n_2 r^2 d \exp\Big(-\frac{\tau^2/2}{1+2\tau/3}\Big(\frac{n_1}{\mu_0^2 r^2 d}\Big)\Big)\\ &+2n_2 d_1 d_2 d_3 \exp\Big(\frac{-(3/32)n_1}{\mu_0^2 r^2 d}\Big) + n_2 d^{-\beta-1}.\end{aligned}$$

We now prove Theorem 1 by setting $\tau = d^{-\delta_2/2}/2$, so that condition (13) can be written as $n_2 \geq c_2/\delta_2$. Assume without loss of generality $n_2 \leq d/2$ because large c_0 forces large d. For sufficiently large c_2', the right-hand side of the above inequality is no greater than $d^{-\beta}$ when

$$n_1 \geq c_2'(1+\beta)(\log d)\mu_0^2 r^2 d/(4\tau^2) = c_2' q_2^* d^{1+\delta_2}$$

holds as well as (14). Thus, (4) implies (5) for sufficiently large c_0. □

4 Concentration Inequalities for Low Rank Tensors

We now prove Lemmas 5 and 6, both involving tensors of low rank. We note that Lemma 5 concerns the concentration inequality for the sum of a sequence of dependent tensors whereas in Lemma 6, we are interested in a sequence of iid tensors.

4.1 Proof of Lemma 5

We first consider Lemma 5. Let (a_k, b_k, c_k) be sequentially uniformly sampled from Ω^* without replacement, $S_k = \{(a_j, b_j, c_j) : j \leq k\}$, and $m_k = d_1 d_2 d_3 - k$. Given S_k, the conditional expectation of $\mathcal{P}_{(a_{k+1}, b_{k+1}, c_{k+1})}$ is

$$\mathbb{E}\Big[\mathcal{P}_{(a_{k+1},b_{k+1},c_{k+1})}\Big| S_k\Big] = \frac{\mathcal{P}_{S_k^c}}{m_k}.$$

For $k = 1, \ldots, n$, define martingale differences

$$\mathcal{D}_k = d_1 d_2 d_3 (m_n/m_k)\mathcal{Q}_{\boldsymbol{T}}\Big(\mathcal{P}_{(a_k,b_k,c_k)} - \mathcal{P}_{S_{k-1}^c}/m_{k-1}\Big)\mathcal{Q}_{\boldsymbol{T}}.$$

Because $\mathcal{P}_{S_0^c} = \mathcal{I}$ and $S_n = \Omega$, we have

$$\mathcal{Q}_{\boldsymbol{T}}\mathcal{P}_\Omega\mathcal{Q}_{\boldsymbol{T}}/m_n \;=\; \frac{\mathcal{D}_n}{d_1 d_2 d_3 m_n} + \mathcal{Q}_{\boldsymbol{T}}(\mathcal{P}_{S_{n-1}^c}/m_{n-1})\mathcal{Q}_{\boldsymbol{T}}/m_n + \mathcal{Q}_{\boldsymbol{T}}\mathcal{P}_{S_{n-1}}\mathcal{Q}_{\boldsymbol{T}}/m_n$$

$$
\begin{aligned}
&= \frac{\mathcal{D}_n}{d_1 d_2 d_3 m_n} + \mathcal{Q}_{\boldsymbol{T}}(1/m_n - 1/m_{n-1}) + \mathcal{Q}_{\boldsymbol{T}}\mathcal{P}_{S_{n-1}}\mathcal{Q}_{\boldsymbol{T}}/m_{n-1} \\
&= \sum_{k=1}^n \frac{\mathcal{D}_k}{d_1 d_2 d_3 m_n} + \mathcal{Q}_{\boldsymbol{T}}(1/m_n - 1/m_0).
\end{aligned}
$$

Since $1/m_n - 1/m_0 = n/(d_1 d_2 d_3 m_n)$, it follows that

$$
\mathcal{Q}_{\boldsymbol{T}}(d_1 d_2 d_3/n)\mathcal{P}_\Omega \mathcal{Q}_{\boldsymbol{T}} - \mathcal{Q}_{\boldsymbol{T}} = \frac{1}{n}\sum_{k=1}^n \mathcal{D}_k.
$$

Now an application of the matrix martingale Bernstein inequality (see, e.g., Tropp, 2011) gives

$$
\mathbb{P}\Big\{\frac{1}{n}\Big\|\sum_{k=1}^n \mathcal{D}_k\Big\| > \tau\Big\} \le 2\,\mathrm{rank}(\mathcal{Q}_{\boldsymbol{T}})\exp\Big(\frac{-n^2\tau^2/2}{\sigma^2 + n\tau M/3}\Big),
$$

where M is a constant upper bound of $\|\mathcal{D}_k\|$ and σ^2 is a constant upper bound of

$$
\Big\|\sum_{k=1}^n \mathbb{E}\big[\mathcal{D}_k\mathcal{D}_k|S_{k-1}\big]\Big\|.
$$

Note that D_k are random self-adjoint operators.

Recall that $\mathcal{Q}_{\boldsymbol{T}}$ can be decomposed as a sum of orthogonal projections

$$
\begin{aligned}
\mathcal{Q}_{\boldsymbol{T}} &= (\mathcal{Q}^0_{\boldsymbol{T}} + \mathcal{Q}^1_{\boldsymbol{T}}) + \mathcal{Q}^2_{\boldsymbol{T}} + \mathcal{Q}^3_{\boldsymbol{T}} \\
&= \boldsymbol{I}\otimes \boldsymbol{P}^2_{\boldsymbol{T}} \otimes \boldsymbol{P}^3_{\boldsymbol{T}} + \boldsymbol{P}^1_{\boldsymbol{T}}\otimes \boldsymbol{P}^2_{\boldsymbol{T}^\perp}\otimes \boldsymbol{P}^3_{\boldsymbol{T}} + \boldsymbol{P}^1_{\boldsymbol{T}}\otimes \boldsymbol{P}^2_{\boldsymbol{T}}\otimes \boldsymbol{P}^3_{\boldsymbol{T}^\perp}.
\end{aligned}
$$

The rank of $\mathcal{Q}_{\boldsymbol{T}}$, or equivalently the dimension of its range, is given by

$$
d_1 r_2 r_3 + (d_2 - r_2) r_1 r_3 + (d_3 - r_3) r_1 r_2 \le \overline{r}^2 d.
$$

Hereafter, we shall write r_j for $r_j(\boldsymbol{T})$, μ for $\mu(\boldsymbol{T})$, and $\overline{r}$ for $\overline{r}(\boldsymbol{T})$ for brevity when no confusion occurs. Since $\mathbb{E}\big[\mathcal{D}_k\big|S_{k-1}\big] = 0$, the total variation is bounded by

$$
\begin{aligned}
&\max_{\|\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}\|_{\mathrm{HS}}=1}\sum_{k=1}^n \mathbb{E}\Big[\Big\langle \mathcal{D}_k\boldsymbol{X}, \mathcal{D}_k\boldsymbol{X}\Big\rangle\Big|S_{k-1}\Big] \\
\le\ &\max_{\|\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}\|_{\mathrm{HS}}=1}\sum_{k=1}^n \Big(d_1 d_2 d_3 (m_n/m_k)\Big)^2 \mathbb{E}\Big[\Big\langle \big(\mathcal{Q}_{\boldsymbol{T}}\mathcal{P}_{(a_k,b_k,c_k)}\mathcal{Q}_{\boldsymbol{T}}\big)^2\boldsymbol{X}, \boldsymbol{X}\Big\rangle\Big|S_{k-1}\Big] \\
\le\ &\sum_{k=1}^n \Big(d_1 d_2 d_3 (m_n/m_k)\Big)^2 m_{k-1}^{-1}\max_{\|\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}\|_{\mathrm{HS}}=1}\sum_{a,b,c}\Big\langle\big(\mathcal{Q}_{\boldsymbol{T}}\mathcal{P}_{(a,b,c)}\mathcal{Q}_{\boldsymbol{T}}\big)^2\boldsymbol{X},\boldsymbol{X}\Big\rangle.
\end{aligned}
$$

Since $m_n \le m_k$ and $\sum_{k=1}^n (m_n/m_k)/m_{k-1} = n/d_1 d_2 d_3$,

$$
\max_{\|\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}\|_{\mathrm{HS}}=1}\sum_{k=1}^n \mathbb{E}\Big[\Big\langle \mathcal{D}_k\boldsymbol{X}, \mathcal{D}_k\boldsymbol{X}\Big\rangle\Big|S_{k-1}\Big] \le n d_1 d_2 d_3 \max_{a,b,c}\Big\|\mathcal{Q}_{\boldsymbol{T}}\mathcal{P}_{(a,b,c)}\mathcal{Q}_{\boldsymbol{T}}\Big\|.
$$

It then follows that

$$
\begin{aligned}
\max_{a,b,c} \left\| \mathcal{Q}_{\boldsymbol{T}} \mathcal{P}_{(a,b,c)} \mathcal{Q}_{\boldsymbol{T}} \right\| &= \max_{\|\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}\|_{\rm HS}=1} \left\langle \mathcal{Q}_{\boldsymbol{T}} \mathcal{P}_{(a,b,c)} \mathcal{Q}_{\boldsymbol{T}} \boldsymbol{X}, \mathcal{Q}_{\boldsymbol{T}} \boldsymbol{X} \right\rangle \\
&= \max_{\|\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}\|_{\rm HS}=1} \left\langle \mathcal{Q}_{\boldsymbol{T}} \boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c, \mathcal{Q}_{\boldsymbol{T}} \boldsymbol{X} \right\rangle^2 \\
&\le \frac{\mu^2 \overline{r}^2 d}{d_1 d_2 d_3}.
\end{aligned}
$$

Consequently, we may take $\sigma^2 = n\mu_0^2\overline{r}^2 d$. Similarly,

$$
M \le \max_k d_1 d_2 d_3 (m_n/m_k) 2 \max_{a,b,c} \left\| \mathcal{Q}_{\boldsymbol{T}} \mathcal{P}_{(a,b,c)} \mathcal{Q}_{\boldsymbol{T}} \right\| \le 2\mu^2 \overline{r}^2 d.
$$

Inserting the expression and bounds for $\text{rank}(\mathcal{Q}_{\boldsymbol{T}})$, σ^2 and M into the Bernstein inequality, we find

$$
\mathbb{P}\left\{ \frac{1}{n} \left\| \sum_{k=1}^n \mathcal{D}_k \right\| > \tau \right\} \le 2(\overline{r}^2 d) \exp\left(\frac{-\tau^2/2}{1+2\tau/3} \left(\frac{n}{\mu^2 \overline{r}^2 d} \right) \right),
$$

which completes the proof because $\mu(\boldsymbol{T}) \le \mu_0$ and $\overline{r}(\boldsymbol{T}) = r$. □

4.2 Proof of Lemma 6.

In proving Lemma 6, we consider first (12). Let $\boldsymbol{X}$ be a tensor with $\|\boldsymbol{X}\|_{\max} \le 1$. Similar to before, write

$$
\mathcal{D}_i = d_1 d_2 d_3 \mathcal{Q}_{\boldsymbol{T}} \mathcal{P}_{(a_i,b_i,c_i)} - \mathcal{Q}_{\boldsymbol{T}}
$$

for $i = 1, \ldots, n_1$. Again, we shall also write μ for $\mu(\boldsymbol{T})$, and $\overline{r}$ for $\overline{r}(\boldsymbol{T})$ for brevity. Observe that for each point $(a, b, c) \in [d_1] \times [d_2] \times [d_3]$,

$$
\begin{aligned}
& \frac{1}{d_1 d_2 d_3} \left| \langle \boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c, \mathcal{D}_i \boldsymbol{X} \rangle \right| \\
= & \left| \langle \mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c), \mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{e}_{a_k} \otimes \boldsymbol{e}_{b_k} \otimes \boldsymbol{e}_{c_k}) \rangle X(a_k, b_k, c_k) - \langle \mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c), \mathcal{Q}_{\boldsymbol{T}} \boldsymbol{X} \rangle / (d_1 d_2 d_3) \right| \\
\le & \ 2 \max_{a,b,c} \|\mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c)\|_{\rm HS}^2 \|\boldsymbol{X}\|_{\max} \\
\le & \ 2\mu^2 \overline{r}^2 d / (d_1 d_2 d_3).
\end{aligned}
$$

Since the variance of a variable is no greater than the second moment,

$$
\begin{aligned}
& \mathbb{E} \left(\frac{1}{d_1 d_2 d_3} \left| \langle \boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c, \mathcal{Q}_{\boldsymbol{T}} \mathcal{D}_i \boldsymbol{X} \rangle \right| \right)^2 \\
\le & \ \mathbb{E} \left| \langle \mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c), \mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{e}_{a'} \otimes \boldsymbol{e}_{b'} \otimes \boldsymbol{e}_{c'}) \rangle X(a_k, b_k, c_k) \right|^2 \\
\le & \ \frac{1}{d_1 d_2 d_3} \sum_{a',b',c'} \left| \langle \mathcal{Q}_{\boldsymbol{T}} \boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c, \mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{e}_{a'} \otimes \boldsymbol{e}_{b'} \otimes \boldsymbol{e}_{c'}) \rangle \right|^2 \\
= & \ \frac{1}{d_1 d_2 d_3} \|\mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c)\|_{\rm HS}^2
\end{aligned}
$$

$$\leq \quad \mu^2\overline{r}^2 d/(d_1d_2d_3)^2.$$

Since $\langle \boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c, \mathcal{Q}_{\boldsymbol{T}}\mathcal{D}_i\boldsymbol{X}\rangle$ are iid random variables, the Bernstein inequality yields

$$\mathbb{P}\Big\{\Big|\frac{1}{n_1}\sum_{i=1}^{n_1}\langle \boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c, \mathcal{Q}_{\boldsymbol{T}}\mathcal{D}_i\boldsymbol{X}\rangle\Big| > \tau\Big\} \leq 2\exp\Big(-\frac{(n_1\tau)^2/2}{n_1\mu^2\overline{r}^2 d + n_1\tau 2\mu^2\overline{r}^2 d/3}\Big).$$

This yields (12) by the union bound.

The proof of (11) is similar, but the matrix Bernstein inequality is used. We equip $\mathbb{R}^{d_1\times d_2\times d_3}$ with the Hilbert-Schmidt norm so that it can be viewed as the Euclidean space. As linear maps in this Euclidean space, the operators D_i are just random matrices. Since the projection $\mathcal{P}_{(a,b,c)} : \boldsymbol{X} \to \langle \boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c, \boldsymbol{X}\rangle \boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c$ is of rank 1,

$$\|\mathcal{Q}_{\boldsymbol{T}}\mathcal{P}_{(a,b,c)}\mathcal{Q}_{\boldsymbol{T}}\| = \|\mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c)\|_{\rm HS}^2 \leq \mu^2\overline{r}^2 d/(d_1d_2d_3).$$

It follows that $\|\mathcal{D}_i\| \leq 2\mu^2\overline{r}^2 d$. Moreover, $\mathcal{D}_i$ is a self-adjoint operator and its covariance operator is bounded by

$$\begin{aligned}
\max_{\|\boldsymbol{X}\|_{\rm HS}=1}\mathbb{E}\big\|\mathcal{D}_i\boldsymbol{X}\big\|_{\rm HS}^2 &\leq (d_1d_2d_3)^2\max_{\|\boldsymbol{X}\|_{\rm HS}=1}\mathbb{E}\Big\langle \big(\mathcal{Q}_{\boldsymbol{T}}\mathcal{P}_{(a_k,b_k,c_k)}\mathcal{Q}_{\boldsymbol{T}}\big)^2\boldsymbol{X}, \boldsymbol{X}\Big\rangle \\
&\leq d_1d_2d_3\sum_{a,b,c}\|\mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c)\|_{\rm HS}^2\big\langle \boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c, \boldsymbol{X}\big\rangle^2 \\
&\leq d_1d_2d_3\max_{a,b,c}\|\mathcal{Q}_{\boldsymbol{T}}(\boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c)\|_{\rm HS}^2 \\
&= \mu^2\overline{r}^2 d
\end{aligned}$$

Consequently, by the matrix Bernstein inequality (Tropp, 2011),

$$\mathbb{P}\Big\{\Big\|\frac{1}{n_1}\sum_{i=1}^{n_1}\mathcal{D}_i\Big\| > \tau\Big\} \leq 2\,\mathrm{rank}(\mathcal{Q}_{\boldsymbol{T}})\exp\Big(\frac{\tau^2/2}{1+2\tau/3}\Big(\frac{n_1}{\mu^2\overline{r}^2 d}\Big)\Big).$$

This completes the proof due to the fact that $\mathrm{rank}(\mathcal{Q}_{\boldsymbol{T}}) \leq \overline{r}^2 d$. □

5 Concentration Inequalities for Sparse Tensors

We now derive probabilistic bounds for $\|\mathcal{R}_\ell\boldsymbol{X}\|$ when (a_i, b_i, c_i)s are iid vectors uniformly sampled from $[d_1]\times[d_2]\times[d_3]$ and $\boldsymbol{X} = \mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}$ with small $\|\boldsymbol{X}\|_{\max}$.

5.1 Symmetrization

We are interested in bounding

$$\max_{\boldsymbol{X}\in\mathscr{U}(\eta)}\mathbb{P}\left\{\left\|\frac{d_1d_2d_3}{n}\sum_{i=1}^{n}\mathcal{P}_{(a_i,b_i,c_i)}\boldsymbol{X} - \boldsymbol{X}\right\| \geq t\right\},$$

e.g. with (n,η,t) replaced by $(n_1,(\alpha_0\sqrt{r})\wedge\sqrt{d_1d_2d_3},1/8)$ in the proof of Lemma 7, where

$$\mathscr{U}(\eta)=\{\boldsymbol{X}:\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X}=\boldsymbol{X},\|\boldsymbol{X}\|_{\max}\le\eta/\sqrt{d_1d_2d_3}\}.$$

Our first step is symmetrization.

Lemma 8 *Let ϵ_is be a Rademacher sequence, that is a sequence of i.i.d. ϵ_i with $\mathbb{P}\{\epsilon_i=1\}=\mathbb{P}\{\epsilon_i=-1\}=1/2$. Then*

$$\begin{aligned}&\max_{\boldsymbol{X}\in\mathscr{U}(\eta)}\mathbb{P}\left\{\left\|\frac{d_1d_2d_3}{n}\sum_{i=1}^n\mathcal{P}_{(a_i,b_i,c_i)}\boldsymbol{X}-\boldsymbol{X}\right\|\ge t\right\}\\ \le\;&4\max_{\boldsymbol{X}\in\mathscr{U}(\eta)}\mathbb{P}\left\{\left\|\frac{d_1d_2d_3}{n}\sum_{i=1}^n\epsilon_i\mathcal{P}_{(a_i,b_i,c_i)}\boldsymbol{X}\right\|\ge t/2\right\}+4\exp\left(-\frac{nt^2/2}{\eta^2+2\eta t\sqrt{d_1d_2d_3}/3}\right).\end{aligned}$$

Proof of Lemma 8. Following a now standard Giné-Zinn type of symmetrization argument, we get

$$\begin{aligned}&\max_{\boldsymbol{X}\in\mathscr{U}(\eta)}\mathbb{P}\left\{\left\|\frac{d_1d_2d_3}{n}\sum_{i=1}^n\mathcal{P}_{(a_i,b_i,c_i)}\boldsymbol{X}-\boldsymbol{X}\right\|\ge t\right\}\\ \le\;&4\max_{\boldsymbol{X}\in\mathscr{U}(\eta)}\mathbb{P}\left\{\left\|\frac{d_1d_2d_3}{n}\sum_{i=1}^n\epsilon_i\mathcal{P}_{(a_i,b_i,c_i)}\boldsymbol{X}\right\|\ge t/2\right\}\\ &+2\max_{\boldsymbol{X}\in\mathscr{U}(\eta)}\max_{\|\boldsymbol{u}\|=\|\boldsymbol{v}\|=\|\boldsymbol{w}\|=1}\mathbb{P}\left\{\left\langle\boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w},\frac{d_1d_2d_3}{n}\sum_{i=1}^n\mathcal{P}_{(a_i,b_i,c_i)}\boldsymbol{X}-\boldsymbol{X}\right\rangle>t\right\}.\end{aligned}$$

See, e.g., Giné and Zinn (1984). It remains to bound the second quantity on the right-hand side. To this end, denote by

$$\xi_i=\left\langle\boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w},d_1d_2d_3\mathcal{P}_{(a_i,b_i,c_i)}\boldsymbol{X}-\boldsymbol{X}\right\rangle.$$

For $\|\boldsymbol{u}\|=\|\boldsymbol{v}\|=\|\boldsymbol{w}\|=1$ and $\boldsymbol{X}\in\mathscr{U}(\eta)$, ξ_i are iid variables with $\mathbb{E}\xi_i=0$, $|\xi_i|\le 2d_1d_2d_3\|\boldsymbol{X}\|_{\max}\le 2\eta\sqrt{d_1d_2d_3}$ and $\mathbb{E}\xi_i^2\le(d_1d_2d_3)\|\boldsymbol{X}\|_{\max}^2\le\eta^2$. Thus, the statement follows from the Bernstein inequality. □

In the light of Lemma 8, it suffices to consider bounding

$$\max_{\boldsymbol{X}\in\mathscr{U}(\eta)}\mathbb{P}\left\{\left\|\frac{d_1d_2d_3}{n}\sum_{i=1}^n\epsilon_i\mathcal{P}_{(a_i,b_i,c_i)}\boldsymbol{X}\right\|\ge t/2\right\}.$$

To this end, we use a thinning method to control the spectral norm of tensors.

5.2 Thinning of the spectral norm of tensors

Recall that the spectral norm of a tensor $\boldsymbol{Z} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is defined as

$$\|\boldsymbol{Z}\| = \max_{\substack{\boldsymbol{u}\in\mathbb{R}^{d_1},\boldsymbol{v}\in\mathbb{R}^{d_2},\boldsymbol{w}\in\mathbb{R}^{d_3}\\ \|\boldsymbol{u}\|\vee\|\boldsymbol{v}\|\vee\|\boldsymbol{w}\|\le 1}} \langle \boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w}, \boldsymbol{Z}\rangle.$$

We first use a thinning method to discretize maximization in the unit ball in $\mathbb{R}^{d_j}$ to the problem involving only vectors taking values 0 or $\pm 2^{-\ell/2}, \ell \le m_j := \lceil \log_2 d_j \rceil$, that is, binary "digitalized" vectors that belong to

$$\mathscr{B}_{m_j,d_j} \quad = \quad \{0, \pm 1, \pm 2^{-1/2}, \ldots, \pm 2^{-m_j/2}\}^d \cap \{\boldsymbol{u} \in \mathbb{R}^{d_j} : \|\boldsymbol{u}\| \le 1\}. \tag{16}$$

Lemma 9 *For any tensor* $\boldsymbol{Z} \in \mathbb{R}^{d_1\times d_2\times d_3}$,

$$\|\boldsymbol{Z}\| \le 8 \max\Big\{\langle \boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w}, \boldsymbol{Z}\rangle : \boldsymbol{u}\in\mathscr{B}_{m_1,d_1}, \boldsymbol{v}\in\mathscr{B}_{m_2,d_2}, \boldsymbol{w}\in\mathscr{B}_{m_3,d_3}\Big\},$$

where $m_j := \lceil \log_2 d_j \rceil$, $j = 1, 2, 3$.

Proof of Lemma 9. Denote by

$$C_{m,d} = \min_{\|\boldsymbol{a}\|=1} \max_{u\in\mathscr{B}_{m,d}} \boldsymbol{u}^\top \boldsymbol{a},$$

which bounds the effect of discretization. Let $\boldsymbol{X}$ be a linear mapping from $\mathbb{R}^d$ to a linear space equipped with a seminorm $\|\cdot\|$. Then, $\|\boldsymbol{X}\boldsymbol{u}\|$ can be written as the maximum of $\phi(\boldsymbol{X}\boldsymbol{u})$ over linear functionals $\phi(\cdot)$ of unit dual norm. Since $\max_{\|\boldsymbol{u}\|\le 1} \boldsymbol{u}^\top\boldsymbol{a} = 1$ for $\|\boldsymbol{a}\| = 1$, it follows from the definition of $C_{m,d}$ that

$$\max_{\|\boldsymbol{u}\|\le 1} \boldsymbol{u}^\top\boldsymbol{a} \le \|\boldsymbol{a}\| C_{m,d}^{-1} \max_{\boldsymbol{u}\in\mathscr{B}_{m,d}} \boldsymbol{u}^\top(\boldsymbol{a}/\|\boldsymbol{a}\|) = C_{m,d}^{-1} \max_{\boldsymbol{u}\in\mathscr{B}_{m,d}} \boldsymbol{u}^\top\boldsymbol{a}$$

for every $\boldsymbol{a} \in \mathbb{R}^d$ with $\|\boldsymbol{a}\| > 0$. Consequently, for any positive integer m,

$$\max_{\|\boldsymbol{u}\|\le 1} \|\boldsymbol{X}\boldsymbol{u}\| = \max_{\boldsymbol{a}:\boldsymbol{a}^\top\boldsymbol{v}=\phi(\boldsymbol{X}\boldsymbol{v})\forall\boldsymbol{v}} \max_{\|\boldsymbol{u}\|\le 1} \boldsymbol{a}^\top\boldsymbol{u} \le C_{m,d}^{-1} \max_{\boldsymbol{u}\in\mathscr{B}_{m,d}} \|\boldsymbol{X}\boldsymbol{u}\|.$$

An application of the above inequality to each coordinate yields

$$\|\boldsymbol{Z}\| \le C_{m_1,d_1}^{-1} C_{m_2,d_2}^{-1} C_{m_3,d_3}^{-1} \max_{\boldsymbol{u}\in\mathscr{B}_{m_1,d_1},\boldsymbol{v}\in\mathscr{B}_{m_2,d_2},\boldsymbol{w}\in\mathscr{B}_{m_3,d_3}} \langle \boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w}, \boldsymbol{Z}\rangle.$$

It remains to show that $C_{m_j,d_j} \ge 1/2$. To this end, we prove a stronger result that for any m and d,

$$C_{m,d}^{-1} \le \sqrt{2 + 2(d-1)/(2^m - 1)}.$$

Consider first a continuous version of $C_{m,d}$:

$$C'_{m,d} = \min_{\|\boldsymbol{a}\|=1} \max_{\boldsymbol{u}\in\mathscr{B}'_{m,d}} \boldsymbol{a}^\top \boldsymbol{u}.$$

where $\mathscr{B}'_{m,d} = \{t : t^2 \in [0,1] \setminus (0, 2^{-m})\}^d \cap \{\boldsymbol{u} : \|\boldsymbol{u}\| \le 1\}$. Without loss of generality, we confine the calculation to nonnegative ordered $\boldsymbol{a} = (a_1, \ldots, a_d)^\top$ satisfying $0 \le a_1 \le \ldots \le a_d$ and $\|\boldsymbol{a}\| = 1$. Let

$$k = \max\Big\{j : 2^m a_j^2 + \sum_{i=1}^{j-1} a_i^2 \le 1\Big\} \quad \text{and} \quad \boldsymbol{v} = \frac{(a_i I\{i > k\})_{d\times 1}}{\{1 - \sum_{i=1}^k a_i^2\}^{1/2}}.$$

Because $2^m v_{k+1}^2 = 2^m a_{k+1}^2/(1 - \sum_{i=1}^k a_i^2) \ge 1$, we have $\boldsymbol{v} \in \mathscr{B}'_{m,d}$. By the definition of k, there exists $x^2 \ge a_k^2$ satisfying

$$(2^m - 1)x^2 + \sum_{i=1}^k a_i^2 = 1.$$

It follows that

$$\sum_{i=1}^k a_i^2 = \frac{\sum_{i=1}^k a_i^2}{(2^m-1)x^2 + \sum_{i=1}^k a_i^2} \le \frac{kx^2}{(2^m-1)x^2 + kx^2} \le \frac{d-1}{2^m + d - 2}.$$

Because $\boldsymbol{a}^\top \boldsymbol{v} = (1 - \sum_{i=1}^k a_i^2)^{1/2}$ for this specific $\boldsymbol{v} \in \mathscr{B}'_{m,d}$, we get

$$C'_{m,d} \ge \min_{\|a\|_2=1} \Big(1 - \sum_{i=1}^k a_i^2\Big)^{1/2} \ge \Big(1 - \frac{d-1}{2^m+d-2}\Big)^{1/2} = \Big(\frac{2^m - 1}{2^m + d - 2}\Big)^{1/2}.$$

Now because every $\boldsymbol{v} \in \mathscr{B}'_{m,d}$ with nonnegative components matches a $\boldsymbol{u} \in \mathscr{B}_{m,d}$ with

$$\mathrm{sgn}(v_i)\sqrt{2}u_i \ge |v_i| \ge \mathrm{sgn}(v_i)u_i,$$

we find $C_{m,d} \ge C'_{m,d}/\sqrt{2}$. Consequently,

$$1/C_{m,d} \le \sqrt{2}/C'_{m,d} \le \sqrt{2}\{1 + (d-1)/(2^m-1)\}^{1/2}. \qquad \square$$

It follows from Lemma 9 that the spectrum norm $\|\boldsymbol{Z}\|$ is of the same order as the maximum of $\langle \boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w}, \boldsymbol{Z}\rangle$ over $\boldsymbol{u} \in \mathscr{B}_{m_1,d_1}$, $\boldsymbol{v} \in \mathscr{B}_{m_2,d_2}$ and $\boldsymbol{w} \in \mathscr{B}_{m_3,d_3}$. We will further decompose such tensors $\boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w}$ according to the absolute value of their entries and bound the entropy of the components in this decomposition.

5.3 Spectral norm of tensors with sparse support

Denote by D_j a "digitalization" operator such that $D_j(\boldsymbol{X})$ will zero out all entries of $\boldsymbol{X}$ whose absolute value is not $2^{-j/2}$, that is

$$D_j(\boldsymbol{X}) = \sum_{a,b,c} \mathbb{I}\Big\{|\langle \boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c, \boldsymbol{X}\rangle| = 2^{-j/2}\Big\} \langle \boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c, \boldsymbol{X}\rangle \boldsymbol{e}_a \otimes \boldsymbol{e}_b \otimes \boldsymbol{e}_c. \tag{17}$$

With this notation, it is clear that for $\boldsymbol{u} \in \mathscr{B}_{m_1,d_1}$, $\boldsymbol{v} \in \mathscr{B}_{m_2,d_2}$ and $\boldsymbol{w} \in \mathscr{B}_{m_3,d_3}$,

$$\langle \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}, \boldsymbol{X}\rangle = \sum_{j=0}^{m_1+m_2+m_3} \langle D_j(\boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}), \boldsymbol{X}\rangle.$$

The possible choice of $D_j(\boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w})$ in the above expression may be further reduced if $\boldsymbol{X}$ is sparse. More specifically, denote by

$$\text{supp}(\boldsymbol{X}) = \{\omega \in [d_1] \times [d_2] \times [d_3] : X(\omega) \neq 0\}.$$

Define the maximum aspect ratio of $\text{supp}(\boldsymbol{X})$ as

$$\nu_{\text{supp}(\boldsymbol{X})} = \max_{\ell=1,2,3} \max_{i_k : k \neq \ell} |\{i_\ell : (i_1, i_2, i_3) \in \text{supp}(\boldsymbol{X})\}|. \tag{18}$$

In other words, the quantity $\nu_{\text{supp}(\boldsymbol{X})}$ is the maximum ℓ_0 norm of the fibers of the third-order tensor. We observe first that, if $\text{supp}(\boldsymbol{X})$ is a uniformly sampled subset of $[d_1] \times [d_2] \times [d_3]$, then it necessarily has a small aspect ratio.

Lemma 10 *Let Ω be a uniformly sampled subset of $[d_1] \times [d_2] \times [d_3]$ without replacement. Let $d = d_1 + d_2 + d_3$, $p^* = \max(d_1, d_2, d_3)/(d_1 d_2 d_3)$, and $\nu_1 = (d^{\delta_1} enp^*) \vee \{(3+\beta)/\delta_1\}$ with a certain $\delta_1 \in [1/\log d, 1]$. Then,*

$$\mathbb{P}\Big\{\nu_\Omega \geq \nu_1\Big\} \leq d^{-\beta-1}/3.$$

Proof of Lemma 10. Let $p_1 = d_1/(d_1 d_2 d_3)$, $t = \log(\nu_1/(np^*)) \geq 1$, and

$$N_{i_2 i_3} = |\{i_\ell : (i_1, i_2, i_3) \in \Omega\}|.$$

Because $N_{i_2 i_3}$ follows the Hypergeometric$(d_1 d_2 d_3, d_1, n)$ distribution, its moment generating function is no greater than that of Binomial(n, p_1). Due to $p_1 \leq p^*$,

$$\mathbb{P}\{N_{i_2 i_3} \geq \nu_1\} \leq \exp\big(-t\nu_1 + np^*(e^t - 1)\big) \leq \exp\left(-\nu_1 \log(\nu_1/(enp^*))\right).$$

The condition on ν_1 implies $\nu_1 \log(\nu_1/(enp^*)) \ge (3+\beta)\log d$. By the union bound,

$$\mathbb{P}\{\max_{i_2,i_3} N_{i_2 i_3} \ge \nu_1\} \le d_2 d_3 d^{-3-\beta}.$$

By symmetry, the same tail probability bound also holds for $\max_{i_1,i_3} |\{i_2 : (i_1,i_2,i_3) \in \Omega\}|$ and $\max_{i_1,i_2} |\{i_3 : (i_1,i_2,i_3) \in \Omega\}|$, so that $\mathbb{P}\{\nu_\Omega \ge \nu_1\} \le (d_1 d_2 + d_1 d_3 + d_2 d_3) d^{-3-\beta}$. The conclusion follows from $d_1 d_2 + d_1 d_3 + d_2 d_3 \le d^2/3$. $\square$

We are now in position to further reduce the set of maximization in defining the spectrum norm of sparse tensors. To this end, denote for a block $A \times B \times C \subseteq [d_1] \times [d_2] \times [d_3]$,

$$h(A \times B \times C) = \min\Big\{\nu : |A| \le \nu|B||C|,\ |B| \le \nu|A||C|,\ |C| \le \nu|A||B|\Big\}. \tag{19}$$

It is clear that for any block $A \times B \times C$, there exists $\widetilde{A} \subseteq A$, $\widetilde{B} \subseteq B$ and $\widetilde{C} \subseteq C$ such that $h(\widetilde{A} \times \widetilde{B} \times \widetilde{C}) \le \nu_\Omega$ and

$$(A \times B \times C) \cap \Omega = (\widetilde{A} \times \widetilde{B} \times \widetilde{C}) \cap \Omega.$$

For $\boldsymbol{u} \in \mathscr{B}_{m_1,d_1}$, $\boldsymbol{v} \in \mathscr{B}_{m_2,d_2}$ and $\boldsymbol{w} \in \mathscr{B}_{m_3,d_3}$, let $A_{i_1} = \{a : u_a^2 = 2^{-i_1}\}$, $B_{i_2} = \{b : v_b^2 = 2^{-i_2}\}$ and $C_{i_3} = \{c : w_c^2 = 2^{-i_3}\}$, and define

$$\widetilde{D}_j(\boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}) = \sum_{(i_1,i_2,i_3): i_1+i_2+i_3=j} \mathcal{P}_{\widetilde{A}_{j,i_1} \times \widetilde{B}_{j,i_2} \times \widetilde{C}_{j,i_3}} D_j(\boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}) \tag{20}$$

where $\widetilde{A}_{j,i_1} \subseteq A_{j,i_1}$, $\widetilde{B}_{j,i_2} \subseteq B_{j,i_2}$ and $\widetilde{C}_{j,i_3} \subseteq C_{j,i_3}$ satisfying $h(\widetilde{A}_{j,i_1} \times \widetilde{B}_{j,i_2} \times \widetilde{C}_{j,i_3}) \le \nu_\Omega$ and

$$(A_{i_1} \times B_{i_2} \times C_{i_3}) \cap \Omega = (\widetilde{A}_{j,i_1} \times \widetilde{B}_{j,i_2} \times \widetilde{C}_{j,i_3}) \cap \Omega.$$

Because $\mathcal{P}_\Omega D_j(\boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w})$ is supported in $\cup_{(i_1,i_2,i_3): i_1+i_2+i_3=j} (A_{i_1} \times B_{i_2} \times C_{i_3}) \cap \Omega$, we have

$$\mathcal{P}_\Omega \widetilde{D}_j(\boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}) = \mathcal{P}_\Omega D_j(\boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}).$$

This observation, together with Lemma 9, leads to the following characterization of the spectral norm of a tensor support on a set with bounded aspect ratio.

Lemma 11 *Let $m_\ell = \lceil \log_2 d_\ell \rceil$ for $\ell = 1, 2, 3$, and $D_j(\cdot)$ and $\widetilde{D}_j(\cdot)$ be as in (17) and (20) respectively. Define*

$$\mathscr{B}^*_{\Omega,m_*} = \left\{ \sum_{0 \le j \le m_*} \widetilde{D}_j(\boldsymbol{u}_1 \otimes \boldsymbol{u}_2 \otimes \boldsymbol{u}_3) + \sum_{m_* < j \le m^*} D_j(\boldsymbol{u}_1 \otimes \boldsymbol{u}_2 \otimes \boldsymbol{u}_3) : \boldsymbol{u}_\ell \in \mathscr{B}_{m_\ell,d_\ell} \right\},$$

and $\mathscr{B}^*_{\nu,m_*} = \cup_{\nu_\Omega \le \nu} \mathscr{B}^*_{\Omega,m_*}$. Let $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ be a tensor with $\mathrm{supp}(\boldsymbol{X}) \subseteq \Omega$. For any $0 \le m_* \le m_1 + m_2 + m_3$ and $\nu \ge \nu_\Omega$,

$$\|\boldsymbol{X}\| \le 8 \max_{\boldsymbol{Y} \in \mathscr{B}^*_{\Omega,m_*}} \langle \boldsymbol{Y}, \boldsymbol{X} \rangle \le 8 \max_{\boldsymbol{Y} \in \mathscr{B}^*_{\nu,m_*}} \langle \boldsymbol{Y}, \boldsymbol{X} \rangle.$$

5.4 Entropy bounds

Essential to our argument are entropic bounds related to $\mathscr{B}^*_{\nu,m_*}$. It is clear that

$$\#\{D_j(\boldsymbol{u}) : \|\boldsymbol{u}\| \le 1\} \le \binom{d_j}{2^k \wedge d_j} 2^{2^k \wedge d_j} \le \exp((2^k \wedge d_j)(\log 2 + 1 + (\log(d_j/2^k))_+)),$$

so that by (16)

$$|\mathscr{B}_{m_j,d_j}| \le \prod_{k=0}^{m_j} \binom{d_j}{2^k \wedge d_j} 2^{2^k \wedge d_j} \le \exp\Big(d_j \sum_{\ell=1}^{\infty} 2^{-\ell}(\log 2 + 1 + \log(2^\ell))\Big) \le \exp(4.78\, d_j)$$

Consequently, due to $d = d_1 + d_2 + d_3$ and $4.78 \le 21/4$,

$$\big|\mathscr{B}^*_{\nu,m_*}\big| \le \prod_{j=1}^{3} |\mathscr{B}_{m_j,d_j}| \le e^{(21/4)d}. \tag{21}$$

We derive tighter entropy bounds for slices of $\mathscr{B}^*_{\nu,m_*}$ by considering

$$\mathscr{D}_{\nu,j,k} = \Big\{D_j(\boldsymbol{Y}) : \boldsymbol{Y} \in \mathscr{B}^*_{\nu,m_*}, \|D_j(\boldsymbol{Y})\|^2_{\mathrm{HS}} \le 2^{k-j}\Big\}.$$

Here and in the sequel, we suppress the dependence of $\mathscr{D}$ on quantities such as m_*, m_1, m_2, m_3 for brevity, when no confusion occurs.

Lemma 12 *Let $L(x,y) = \max\{1, \log(ey/x)\}$ and $\nu \ge 1$. For all $0 \le k \le j \le m^*$,*

$$\log\Big|\mathscr{D}_{\nu,j,k}\Big| \le (21/4) J(\nu, j, k), \tag{22}$$

where $J(\nu,j,k) = (j+2)\sqrt{\nu 2^{k-1}}\, L\big(\sqrt{\nu 2^{k-1}}, (j+2)d\big)$.

Proof of Lemma 12. We first bound the entropy of a single block. Let

$$\mathscr{D}^{(\text{block})}_{\nu,\ell} = \Big\{\mathrm{sgn}(u_a)\mathrm{sgn}(v_b)\mathrm{sgn}(w_c)\mathbb{I}\{(a,b,c) \in A \times B \times C\} : h(A \times B \times C) \le \nu,\ |A||B||C| = \ell\Big\}.$$

By the constraints on the size and aspect ratio of the block,

$$\max(|A|^2, |B|^2, |C|^2) \le \nu |A||B||C| \le \nu \ell.$$

By dividing $\mathscr{D}^{(\text{block})}_{\nu,\ell}$ into subsets according to $(\ell_1, \ell_2, \ell_3) = (|A|, |B|, |C|)$, we find

$$\Big|\mathscr{D}^{(\text{block})}_{\nu,\ell}\Big| \le \sum_{\ell_1\ell_2\ell_3 = \ell, \max(\ell_1,\ell_2,\ell_3) \le \sqrt{\nu\ell}} 2^{\ell_1+\ell_2+\ell_3} \binom{d_1}{\ell_1}\binom{d_2}{\ell_2}\binom{d_3}{\ell_3}$$

By the Stirling formula, for $i = 1, 2, 3$,

$$\log\left[\sqrt{2\pi\ell_i}2^{\ell_i}\binom{d_i}{\ell_i}\right] \le \ell_i L\big(\ell_i, 2d\big) \le \sqrt{\nu\ell}\ L\big(\sqrt{\nu\ell}, 2d\big).$$

We note that $k(k+1)/(2\sqrt{q^k})$ is no greater than 2.66, 1.16 and 1 respectively for $q = 2$, $q = 3$ and $q \ge 5$. Let $\ell = \prod_{j=1}^m q_j^{k_j}$ with distinct prime factors q_j. We get

$$|\{(\ell_1, \ell_2, \ell_3) : \ell_1\ell_2\ell_3 = \ell\}| = \prod_{j=1}^m \binom{k_j+1}{2} \le 2.66 \times 1.16 \prod_{j=1}^m \sqrt{q_j^{k_j}} \le \pi\ell^{1/2} \le \prod_{i=1}^3 \sqrt{2\pi\ell_i}.$$

It follows that

$$\left|\mathscr{D}_{\nu,\ell}^{(\text{block})}\right| \le \exp\Big(3\sqrt{\nu\ell}L\big(\sqrt{\nu\ell}, 2d\big)\Big). \tag{23}$$

Due to the constraint $i_1 + i_2 + i_3 = j$ in defining $\mathscr{B}^*_{\nu,m_*}$, for any $\boldsymbol{Y} \in \mathscr{B}^*_{\nu,m_*}$, $D_j(\boldsymbol{Y})$ is composed of at most $i^* = \binom{j+2}{2}$ blocks. Since the sum of the sizes of the blocks is bounded by 2^k, (23) yields

$$\begin{aligned}
\left|\mathscr{D}_{\nu,j,k}\right| &\le \sum_{\ell_1+\ldots+\ell_{i^*}\le 2^k} \prod_{i=1}^{i^*} \left|\mathscr{D}_{\nu\ell_i}^{(\text{block})}\right| \\
&\le \sum_{\ell_1+\ldots+\ell_{i^*}\le 2^k} \exp\Big(\sum_{i=1}^{i^*} 3\sqrt{\nu\ell_i}\ L\big(\sqrt{\nu\ell_i}, 2d\big)\Big) \\
&\le (2^k)^{i^*} \max_{\ell_1+\ldots+\ell_{i^*}\le 2^k} \exp\Big(\sum_{i=1}^{i^*} 3\sqrt{\nu\ell_i}\ L\big(\sqrt{\nu\ell_i}, 2d\big)\Big).
\end{aligned}$$

It follows from the definition of $L(x, y)$ and the Cauchy-Schwarz inequality that

$$\begin{aligned}
\sum_{i=1}^{i^*} \sqrt{\ell_i}\ L\big(\sqrt{\nu\ell_i}, 2d\big) &= \sum_{i=1}^{i^*} \sqrt{\ell_i}\Big(L\big(\sqrt{\nu 2^k}, 2d\big) + \log\Big(\sqrt{2^k/\ell_i}\Big)\Big) \\
&\le \sqrt{2^k}\Big(\sqrt{i^*}L\big(\sqrt{\nu 2^k}, 2d\big) + \sum_{i=1}^{i^*} \sqrt{\ell_i/2^k}\log\Big(\sqrt{2^k/\ell_i}\Big)\Big) \\
&\le \sqrt{i^* 2^k}\Big(L\big(\sqrt{\nu 2^k}, 2d\big) + \log\big(\sqrt{i^*}\big)\Big),
\end{aligned}$$

where the last inequality above follows from the fact that subject to $u_1, u_2 \ge 0$ and $u_1^2 + u_2^2 \le 2c^2 \le 2$, the maximum of $-u_1\log(u_1) - u_2\log(u_2)$ is attained at $u_1 = u_2 = c$. Consequently, since $i^* \le \binom{j+2}{2}$,

$$\log\left|\mathscr{D}_{\nu,j,k}\right| \le i^*\log\big(2^k\big) + 3\sqrt{i^*\nu 2^k}L\left(\sqrt{\nu 2^k}, 2d\sqrt{\tbinom{j+2}{2}}\right).$$

We note that $j \geq k$, $\sqrt{2^k} \geq k\sqrt{8}/3$, $\nu \geq 1$ and $xL(x,y)$ is increasing in x, so that

$$\frac{\sqrt{i^*\nu 2^k}L\Big(\sqrt{\nu 2^k}, 2d\sqrt{\binom{j+2}{2}}\Big)}{i^*\log\left(2^k\right)} \geq \frac{\sqrt{8}/3L\Big(k\sqrt{8}/2, 2d\sqrt{\binom{j+2}{2}}\Big)}{\sqrt{i^*}\log 2} \geq \frac{\sqrt{8}\log(ed)}{\sqrt{i^*}\log 8}.$$

Moreover, because $2^{m^*+3/2} \leq \sqrt{8}\prod_{j=1}^3(2d_j) \leq \sqrt{8}(2d/3)^3 \leq d^3$, we get $\sqrt{2i^*} \leq \sqrt{(j+1)(j+2)} \leq j+3/2 \leq m^*+3/2 \leq (3/\log 2)\log d$, so that the right-hand side of the above inequality is no smaller than $(4/\log 8)(\log 2)/3 = 4/9$. It follows that

$$\log\Big|\mathscr{D}_{\nu,j,k}\Big| \leq (3+9/4)\sqrt{i^*\nu 2^k}L\Big(\sqrt{\nu 2^k}, 2d\sqrt{\binom{j+2}{2}}\Big).$$

This yields (22) due to $i^* \leq \binom{j+2}{2}$. □

By (21), the entropy bound in Lemma 12 is useful only when $0 \leq k \leq j \leq m_*$, where

$$m_* = \min\Big\{x : x \geq m^* \quad \text{or} \quad J(\nu_1, x, x) \geq d\Big\}. \tag{24}$$

5.5 Probability bounds

We are now ready to derive a useful upper bound for

$$\max_{\boldsymbol{X}\in\mathscr{U}(\eta)} \mathbb{P}\left\{\left\|\frac{d_1d_2d_3}{n}\sum_{i=1}^n \epsilon_i\mathcal{P}_{(a_i,b_i,c_i)}\boldsymbol{X}\right\| \geq t\right\}.$$

Let $\boldsymbol{X} \in \mathscr{U}(\eta)$. For brevity, write $\boldsymbol{Z}_i = d_1d_2d_3\epsilon_i\mathcal{P}_{(a_i,b_i,c_i)}\boldsymbol{X}$ and

$$\boldsymbol{Z} = \frac{d_1d_2d_3}{n}\sum_{i=1}^n \epsilon_i\mathcal{P}_{(a_i,b_i,c_i)}\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i.$$

Let $\Omega = \{(a_i,b_i,c_i) : i \leq n\}$. In the light of Lemma 10, we shall proceed conditional on the event that $\nu_\Omega \leq \nu_1$ in this subsection. In this event, Lemma 11 yields

$$\begin{aligned} \|\boldsymbol{Z}\| &\leq 8\max_{\boldsymbol{Y}\in\mathscr{B}^*_{\nu_1,m_*}} \langle \boldsymbol{Y}, \boldsymbol{Z}\rangle \qquad (25)\\ &= 8\max_{\boldsymbol{Y}\in\mathscr{B}^*_{\nu_1,m_*}}\left(\sum_{0\leq j\leq m_*}\langle D_j(\boldsymbol{Y}), \boldsymbol{Z}\rangle + \langle S_*(\boldsymbol{Y}), \boldsymbol{Z}\rangle\right), \end{aligned}$$

where m_* is as in (24) with the given ν_1 and $S_*(\boldsymbol{Y}) = \sum_{j>m_*} D_j(\boldsymbol{Y})$.

Let $\boldsymbol{Y} \in \mathscr{B}^*_{\nu_1,m_*}$ and $\boldsymbol{Y}_j = D_j(\boldsymbol{Y})$. Recall that for $\boldsymbol{Y}_j \neq 0$, $2^{-j} \leq \|\boldsymbol{Y}_j\|_{\rm HS}^2 \leq 1$, so that $\boldsymbol{Y}_j \in \cup_{k=0}^j \mathscr{D}_{\nu_1,j,k}$. To bound the first term on the right-hand side of (25), consider

$$\mathbb{P}\left\{\max_{\boldsymbol{Y}_j\in\mathscr{D}_{\nu_1,j,k}\setminus\mathscr{D}_{\nu_1,j,k-1}} \langle \boldsymbol{Y}_j, \boldsymbol{Z}\rangle \geq t(m_*+2)^{-1/2}\|\boldsymbol{Y}_j\|_{\rm HS}\right\}$$

with $0 \le k \le j \le m_*$. Because $\|\boldsymbol{Z}_i\|_{\max} \le d_1d_2d_3\|\boldsymbol{X}\|_{\max} \le \eta\sqrt{d_1d_2d_3}$, for any $\boldsymbol{Y}_j \in \mathscr{D}_{\nu_1,j,k}$,

$$|\langle \boldsymbol{Y}_j, \boldsymbol{Z}_i\rangle| \le 2^{-j/2}\eta\sqrt{d_1d_2d_3}, \quad \mathbb{E}\langle \boldsymbol{Y}_j, \boldsymbol{Z}_i\rangle^2 \le \eta^2\|\boldsymbol{Y}_j\|_{\mathrm{HS}}^2 \le \eta^2 2^{k-j}.$$

Let $h_0(u) = (1+u)\log(1+u) - u$. By Bennet's inequality,

$$\begin{aligned} & \mathbb{P}\left\{\langle \boldsymbol{Y}_j, \boldsymbol{Z}\rangle \ge t2^{(k-j-1)/2}(m_*+2)^{-1/2}\right\} \\ \le\ & \exp\left(-\frac{n(\eta^2 2^{k-j})}{2^{-j}\eta^2 d_1d_2d_3}h_0\left(\frac{(t2^{(k-j-1)/2}(m_*+2)^{-1/2})(2^{-j/2}\eta\sqrt{d_1d_2d_3})}{\eta^2 2^{k-j}}\right)\right) \\ =\ & \exp\left(-\frac{n2^k}{d_1d_2d_3}h_0\left(\frac{t\sqrt{d_1d_2d_3}}{\eta\sqrt{(m_*+2)2^{k+1}}}\right)\right). \end{aligned}$$

Recall that $\mathscr{D}_{\nu,j,k} = \left\{\boldsymbol{Y}_j = D_j(\boldsymbol{Y}) : \boldsymbol{Y} \in \mathscr{B}^*_{\nu,m_*}, \|\boldsymbol{Y}_j\|_{\mathrm{HS}}^2 \le 2^{k-j}\right\}$. By Lemma 12,

$$\begin{aligned} & \mathbb{P}\left\{\max_{\boldsymbol{Y}_j \in \mathscr{D}_{\nu_1,j,k}\setminus\mathscr{D}_{\nu_1,j,k-1}} \langle \boldsymbol{Y}_j, \boldsymbol{Z}\rangle \ge t(m_*+2)^{-1/2}\|\boldsymbol{Y}_j\|_{\mathrm{HS}}\right\} \\ \le\ & |\mathscr{D}_{\nu_1,j,k}| \max_{\boldsymbol{Y}_j \in \mathscr{D}_{\nu_1,j,k}} \mathbb{P}\left\{\langle \boldsymbol{Y}_j, \boldsymbol{Z}\rangle \ge t2^{(k-j-1)/2}(m_*+2)^{-1/2}\right\} \\ \le\ & \exp\left((21/4)J(\nu_1,j,k) - \frac{n2^k}{d_1d_2d_3}h_0\left(\frac{t\sqrt{d_1d_2d_3}}{\eta\sqrt{(m_*+2)2^{k+1}}}\right)\right). \end{aligned} \tag{26}$$

Let $L_k = 1 \vee \log(ed(m_*+2)/\sqrt{\nu_1 2^{k-1}})$. By the definition of $J(\nu,j,k)$ in Lemma 12,

$$J(\nu_1,j,k) \le J(\nu_1,m_*,k) = (m_*+2)\sqrt{2^{k-1}\nu_1}L_k \le J(\nu_1,m_*,m_*) = d.$$

Let $x \ge 1$ and t_1 be a constant satisfying

$$nt_1 \ge 24\eta(m_*+2)^{3/2}\sqrt{\nu_1 d_1d_2d_3}, \quad nxt_1^2 h_0(1) \ge 12\eta^2(m_*+2)^2\sqrt{e}dL_0. \tag{27}$$

We prove that for all $x \ge 1$ and $0 \le k \le j \le m_*$

$$\frac{n2^k}{d_1d_2d_3}h_0\left(\frac{xt_1\sqrt{d_1d_2d_3}}{\eta\sqrt{(m_*+2)2^{k+1}}}\right) \ge 6xJ(\nu_1,m_*,k) \ge 6xJ(\nu_1,j,k). \tag{28}$$

Consider the following three cases:

$$\begin{aligned} &\text{Case 1:} && \frac{xt_1\sqrt{d_1d_2d_3}}{\eta\sqrt{(m_*+2)2^{k+1}}} \ge \left[\max\left\{1, \frac{ed(m_*+2)}{\sqrt{\nu_1 2^{k-1}}}\right\}\right]^{1/2}, \\ &\text{Case 2:} && 1 < \frac{xt_1\sqrt{d_1d_2d_3}}{\eta\sqrt{(m_*+2)2^{k+1}}} \le \left[\frac{ed(m_*+2)}{\sqrt{\nu_1 2^{k-1}}}\right]^{1/2}, \\ &\text{Case 3:} && \frac{xt_1\sqrt{d_1d_2d_3}}{\eta\sqrt{(m_*+2)2^{k+1}}} \le 1. \end{aligned}$$

Case 1: Due to $h_0(u) \geq (u/2)\log(1+u)$ for $u \geq 1$ and the lower bound for nt_1,

$$\begin{aligned} \frac{n2^k}{d_1d_2d_3} h_0\left(\frac{xt_1\sqrt{d_1d_2d_3}}{\eta\sqrt{(m_*+2)2^{k+1}}}\right) &\geq \frac{n2^k}{d_1d_2d_3}\left(\frac{xt_1\sqrt{d_1d_2d_3}}{\eta\sqrt{(m_*+2)2^{k+1}}}\right)\frac{L_k}{4} \\ &\geq 6x(m_*+2)\sqrt{\nu_1 2^{k-1}}L_k. \end{aligned}$$

Case 2: Due to $(m_*+2)\sqrt{2^{k-1}\nu_1} \leq d$, we have

$$\frac{1}{\sqrt{d_1d_2d_3}} \geq \frac{xt_1\sqrt{e}(m_*+2)}{\eta\sqrt{(m_*+2)2^{k+1}}}\left[\frac{ed(m_*+2)}{\sqrt{\nu_1 2^{k-1}}}\right]^{-1} = \frac{xt_1\sqrt{\nu_1 2^{k-1}}}{\eta d\sqrt{e(m_*+2)2^{k+1}}}.$$

Thus, due to $h_0(u) \geq uh_0(1)$ for $u \geq 1$, we have

$$\begin{aligned} \frac{n2^k}{d_1d_2d_3} h_0\left(\frac{xt_1\sqrt{d_1d_2d_3}}{\eta\sqrt{(m_*+2)2^{k+1}}}\right) &\geq \left(\frac{xt_1\sqrt{\nu_1 2^{k-1}}}{\eta d\sqrt{e(m_*+2)2^{k+1}}}\right)\frac{n\sqrt{2^{k-1}}xt_1h_0(1)}{\eta\sqrt{m_*+2}} \\ &= \frac{nx^2t_1^2h_0(1)\sqrt{\nu_1 2^{k-1}}}{2\eta^2(m_*+2)\sqrt{e}d}. \end{aligned}$$

Because $nxt_1^2h_0(1) \geq 12\eta^2(m_*+2)^2\sqrt{e}dL_0$ and $L_0 \geq L_k$, it follows that

$$\frac{n2^k}{d_1d_2d_3} h_0\left(\frac{xt_1\sqrt{d_1d_2d_3}}{\eta\sqrt{(m_*+2)2^{k+1}}}\right) \geq 6x(m_*+2)\sqrt{\nu_1 2^{k-1}}L_k.$$

Case 3: Due to $h_0(u) \geq u^2h_0(1)$ for $0 \leq u \leq 1$ and $nxt_1^2h_0(1) \geq 12\eta^2(m_*+2)d$, we have

$$\frac{n2^k}{d_1d_2d_3} h_0\left(\frac{xt_1\sqrt{d_1d_2d_3}}{\eta\sqrt{(m_*+2)2^{k+1}}}\right) \geq \frac{nx^2t_1^2h_0(1)}{2\eta^2(m_*+2)} \geq 6xd \geq 6xJ(\nu_1, m_*, k).$$

Thus, (28) holds in all three cases.

It follows from (26) and (28) that for t_1 satisfying (27) and all $x \geq 1$

$$\mathbb{P}\left\{\max_{\boldsymbol{Y}_j \in \mathscr{D}_{\nu_1,j,k}\setminus\mathscr{D}_{\nu_1,j,k-1}} \langle \boldsymbol{Y}_j, \boldsymbol{Z}\rangle \geq \frac{xt_1\|\boldsymbol{Y}_j\|_{\mathrm{HS}}}{\sqrt{m_*+2}}\right\} \leq \exp\left(-(6x-21/4)J(\nu_1, m_*, k)\right).$$

We note that $J(\nu_1, m_*, k) \geq J(1, m_*, 0) \geq (m_*+2)\log(ed(m_*+2))$ by the monotonicity of $x\log(y/x)$ for $x \in [1, y]$ and $\sqrt{\nu_1 2^{k-1}} \geq 1$. Summing over $0 \leq k \leq j \leq m_*$, we find by the union bound that

$$\begin{aligned} &\mathbb{P}\left\{\max_{0\leq j\leq m_*}\max_{\boldsymbol{Y}\in\mathscr{B}^*_{\nu_1,m_*}:D_j(\boldsymbol{Y})\neq 0}\frac{\langle D_j(\boldsymbol{Y}), \boldsymbol{Z}\rangle}{\|D_j(\boldsymbol{Y})\|_{\mathrm{HS}}} \geq \frac{xt_1}{(m_*+2)^{1/2}}\right\} \\ \leq\ &\binom{m_*+2}{2}\{ed(m_*+2)\}^{-(6x-21/4)(m_*+2)} \end{aligned}$$

For the second term on the right-hand side of (25), we have

$$|\langle S_*(\boldsymbol{Y}), \boldsymbol{Z}_i\rangle| \leq 2^{-m_*/2}\eta\sqrt{d_1d_2d_3}, \quad \mathbb{E}\left(\langle S_*(\boldsymbol{Y}), \boldsymbol{Z}_i\rangle\right)^2 \leq \eta^2\|S_*(\boldsymbol{Y})\|_{\mathrm{HS}}^2.$$

As in the proof of (26) and (28), $\log|\mathscr{B}^*_{\nu_1,m_*}| \le 5d = 5J(\nu_1, m_*, m_*)$ implies

$$\begin{aligned} & \mathbb{P}\left\{\max_{\boldsymbol{Y}\in\mathscr{B}^*_{\nu_1,m_*}:S_*(\boldsymbol{Y})\neq 0} \frac{\langle S_*(\boldsymbol{Y}), \boldsymbol{Z}\rangle}{\|S_*(\boldsymbol{Y})\|_{\rm HS}} \ge \frac{xt_1}{(m_*+2)^{1/2}}\right\} \\ \le\ & (m^* - m_*)\{ed(m_*+2)\}^{-(6x-21/4)(m_*+2)} \end{aligned}$$

By Cauchy Schwartz inequality, for any $\boldsymbol{Y} \in \mathscr{B}^*_{\nu_1,m_*}$,

$$\frac{\|S_*(\boldsymbol{Y})\|_{\rm HS} + \sum_{0\le j\le m_*}\|D_j(\boldsymbol{Y})\|_{\rm HS}}{(m_*+2)^{1/2}} \le \left(\|S_*(\boldsymbol{Y})\|^2_{\rm HS} + \sum_{0\le j\le m_*}\|D_j(\boldsymbol{Y})\|^2_{\rm HS}\right)^{1/2} \le 1.$$

Thus, by (25), for all t_1 satisfying (27) and $x \ge 1$,

$$\begin{aligned} & \mathbb{P}\left\{\max_{\boldsymbol{Y}\in\mathscr{B}^*_{\nu_1,m_*}} \langle \boldsymbol{Y}, \boldsymbol{Z}\rangle \ge xt_1\right\} \\ \le\ & \left\{\binom{m_*+2}{2} + m^* - m_*\right\}\{ed(m_*+2)\}^{-(6x-21/4)(m_*+2)}. \end{aligned} \tag{29}$$

We now have the following probabilistic bound via (29) and Lemma 10.

Lemma 13 *Let ν_1 be as in Lemma 10, $x \ge 1$, t_1 as in (27) and m_* in (24). Then,*

$$\begin{aligned} & \max_{\boldsymbol{X}\in\mathscr{U}(\eta)} \mathbb{P}\left\{\left\|\frac{d_1d_2d_3}{n}\sum_{i=1}^n \epsilon_i \mathcal{P}_{(a_i,b_i,c_i)}\boldsymbol{X}\right\| \ge xt_1\right\} \\ & \le \left\{\binom{m_*+2}{2} + m^* - m_*\right\}\{ed(m_*+2)\}^{-(6x-21/4)(m_*+2)} + d^{-\beta-1}/3. \end{aligned}$$

5.6 Proof of Lemma 7

We are now in position to prove Lemma 7. Let $\Omega_1 = \{(a_i, b_i, c_i), i \le n_1\}$. By the definition of the tensor coherence $\alpha(\boldsymbol{X})$ and the conditions on $\alpha(\boldsymbol{T})$ and $\overline{r}(\boldsymbol{T})$, we have $\|\boldsymbol{W}\|_{\max} \le \eta/\sqrt{d_1d_2d_3}$ with $\eta = (\alpha_0\sqrt{r}) \wedge \sqrt{d_1d_2d_3}$, so that in the light of Lemmas 8, 10 and 13,

$$\begin{aligned} & \max_{\substack{\boldsymbol{X}:\boldsymbol{X}=\mathcal{Q}_{\boldsymbol{T}}\boldsymbol{X} \\ \|\boldsymbol{X}\|_{\max}\le\|\boldsymbol{W}\|_{\max}}} \mathbb{P}\left\{\left\|\frac{d_1d_2d_3}{n_1}\sum_{i=1}^{n_1}\mathcal{P}_{(a_i,b_i,c_i)}\boldsymbol{X} - \boldsymbol{X}\right\| \ge \frac{1}{8}\right\} \\ \le\ & 4\exp\left(-\frac{n_1(1/16)^2/2}{\eta^2 + (2/3)\eta(1/16)\sqrt{d_1d_2d_3}}\right) \\ & + \left\{\binom{m_*+2}{2} + m^* - m_*\right\}\{ed(m_*+2)\}^{-(6x-21/4)(m_*+2)} + d^{-\beta-1}/3 \end{aligned} \tag{30}$$

with $t = 1/8$ in Lemma 8 and the ν_1 in Lemma 10, provided that $xt_1 \le 1/16$ and

$$n_1t_1 \ge 24\eta(m_*+2)^{3/2}\sqrt{\nu_1d_1d_2d_3}, \quad n_1xt_1^2h_0(1) \ge 12\eta^2(m_*+2)^2\sqrt{e}dL_0.$$

Thus, the right-hand side of (30) is no greater than $d^{-\beta-1}$ for certain $x > 1$ and t_1 satisfying these conditions when

$$
\begin{aligned}
n_1 \;\geq\; & c_1'\Big(1+\frac{1+\beta}{m_*+2}\Big)\Big(\eta\sqrt{(m_*+2)^3\nu_1 d_1 d_2 d_3} + \eta^2(m_*+2)^2 L_0 d\Big) \\
& + c_1'(1+\beta)(\log d)\Big(\eta^2 + \eta\sqrt{d_1 d_2 d_3}\Big)
\end{aligned}
$$

for a sufficiently large constant c_1'. Because $\sqrt{2^{m_*-1}} \leq J(\nu_1, m_*, m_*) = d$, it suffices to have

$$
n_1 \geq c_1''\Big[\big(\beta+\log d\big)\Big\{\eta\sqrt{(\log d)\nu_1 d_1 d_2 d_3} + \eta^2(\log d)^2 d\Big\} + (1+\beta)(\log d)\eta\sqrt{d_1 d_2 d_3}\Big].
$$

When the sample size is n_1, $\nu_1 = (d^{\delta_1} e n_1 \max(d_1,d_2,d_3)/(d_1 d_2 d_3)) \vee \{(3+\beta)/\delta_1\}$ in Lemma 10 with $\delta_1 \in [1/\log d, 1]$. When $\nu_1 = d^{\delta_1} e n_1 \max(d_1,d_2,d_3)/(d_1 d_2 d_3)$, $n_1 \geq x\sqrt{\nu_1 d_1 d_2 d_3}$ iff $n_1 \geq x^2 d^{\delta_1} e \max(d_1,d_2,d_3)$. Thus, it suffices to have

$$
\begin{aligned}
n_1 \;\geq\; & c_1''\Big[\big(\beta+\log d\big)^2\eta^2(\log d)d^{1+\delta_1} + \big(\beta+\log d\big)\eta\sqrt{(\log d)(3+\beta)\delta_1^{-1} d_1 d_2 d_3}\Big] \\
& + c_1''\Big[\big(\beta+\log d\big)\eta^2(\log d)^2 d + (1+\beta)(\log d)\eta\sqrt{d_1 d_2 d_3}\Big].
\end{aligned}
$$

Due to $\sqrt{(1+\beta)(\log d)} \leq 1+\beta+\log d$, the quantities in the second line in the above inequality is absorbed into those in the first line. Consequently, with $\eta = \alpha_0\sqrt{r}$, the stated sample size is sufficient. □

6 Concluding Remarks

In this paper, we study the performance of nuclear norm minimization in recovering a large tensor with low Tucker ranks. Our results demonstrate the benefits of not treating tensors as matrices despite its popularity.

Throughout the paper, we have focused primarily on third order tensors. In principle, our technique can also be used to treat higher order tensors although the analysis is much more tedious and the results quickly become hard to describe. Here we outline a considerably simpler strategy which yields similar sample size requirement as the vanilla nuclear norm minimization. The goal is to illustrate some unique and interesting phenomena associated with higher order tensors. The idea is similar to matricization – instead of unfolding a Nth order tensor into a matrix, we unfold it into a cubic or nearly cubic third order tensor. To fix ideas, we shall restrict our attention to hyper cubic Nth order tensors with $d_1 = \ldots = d_N =: d$ and $r_1(\boldsymbol{T}), r_2(\boldsymbol{T}), \ldots, r_N(\boldsymbol{T})$ are bounded from above by a constant. The discussion can be straightforwardly extended to more general situations. In this case, the resulting third order tensor will have dimensions either $d^{\lfloor N/3 \rfloor}$ or $d^{\lfloor N/3 \rfloor+1}$, and Tucker ranks again bounded.

Here $\lfloor x \rfloor$ stands for the integer part of x. Our results on third order tensor then suggests a sample size requirement of

$$n \asymp d^{N/2}\text{polylog}(d).$$

This is to be compared with a matricization approach that unfolds an Nth order tensor to a (nearly) square matrix (see, e.g., Mu et al., 2013) – where the sample size requirement is $d^{\lceil N/2 \rceil}\text{polylog}(d)$. It is interesting to notice that, in this special case, unfolding a higher order tensor to a third order tensor is preferable to matricization when N is odd.

The main focus of this article is to provide a convex solution to tensor completion that yields a tighter sample size requirement than those based upon matricization. Although the nuclear norm minimization we studied is convex, an efficient implementation has so far eluded us. Indeed, as recently shown by Hillar and Lim (2011), most common tensor problems are NP hard, including computing the tensor spectral or nuclear norm as we defined. However, one should interpret such pessimistic results with caution. First of all, NP-hardness does not rule out the possibility of efficient ways to approximate it. One relevant example is tensor decomposition, a problem very much related to our approach here. Although also a NP hard problem in general, there are many algorithms that can do tensor decomposition efficiently in practice. Moreover, the NP hardness result is for evaluating spectral norm of an arbitrary tensor. Yet for our purposes, we may only need to do so for a smaller class of tensors, particularly those in a neighborhood around the truth. It is possible that evaluating tensor norm in a small neighborhood around the truth can be done in polynomial time. One such example can actually be easily deducted from Lemma 3 of our manuscript, where we show that for bi-orthogonal tensors, nuclear norm can be evaluated in polynomial time.

The research on tensor completion is still in its infancy. In addition to the computational challenges, there are a lot of important questions yet to be addressed. For example, it is recently shown by Mu et al. (2013) that exact recovery is possible with sample size requirement of the order $(r^3 + rd)\text{polylog}(d)$ through direct rank minimization. It remains unclear whether or not such a sample size requirement can be achieved by a convex approach. Our hope is that the results from the current work may lead to further understanding of the nature of tensor related issues, and at some point satisfactory answers to such open problems.

References

Candès, E.J. and Recht, B. (2008), Exact matrix completion via convex optimization, *Foundations of Computational Mathematics*, **9**, 717-772.

Candès, E.J. and Tao, T. (2009), The power of convex relaxation: Near-optimal matrix completion, *IEEE Transactions on Information Theory*, **56(5)**, 2053-2080.

Gandy, S., Recht, B. and Yamada, I. (2011), Tensor completion and low-n-rank tensor recovery via convex optimization, *Inverse Problems*, **27(2)**, 025010.

Giné, E. and Zinn, J. (1984), Some limit theorems for empirical processes, *The Annals of Probability*, **12(4)**, 929-989.

Gross, D. (2011), Recovering low-rank matrices from few coefficients in any basis, *IEEE Transaction on Information Theory*, **57**, 1548-1566.

Hillar, C. and Lim, L.H. (2013), Most tensor problems are NP-hard, *Journal of the ACM*, **60(6)**, Art. 45.

Kolda, T.G. and Bader, B.W. (2009), Tensor decompositions and applications, *SIAM Review*, **51(3)**, 455-500.

Kruskal, J. B. (1989), Rank, decomposition, and uniqueness for 3-way and N-way arrays, in *"Multiway data analysis"*, North-Holland, Amsterdam, pp. 7-18.

Li, N. and Li, B. (2010), Tensor completion for on-board compression of hyperspectral images, In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, 517-520.

Liu, J., Musialski, P., Wonka, P. and Ye, J. (2009), Tensor completion for estimating missing values in visual data, In *ICCV*, 2114-2121.

Mu, C., Huang, B., Wright, J. and Goldfarb, D. (2013), Square deal: lower bounds and improved relaxations for tensor recovery, arXiv: 1307.5870.

Recht, B. (2011), A simpler approach to matrix completion, *Journal of Machine Learning Research*, **12**, 3413-3430.

Semerci, O., Hao, N., Kilmer, M. and Miller, E. (2013), Tensor based formulation and nuclear norm regularizatin for multienergy computed tomography, to appear in *IEEE Transactions on Image Processing.*

Sidiropoulos N.D. and Nion, N. (2010), Tensor algebra and multi-dimensional harmonic retrieval in signal processing for mimo radar, *IEEE Trans. on Signal Processing*, **58(11)**, 5693-5705.

Signoretto, M., De Lathauwer, L. and Suykens, J. (2010), Nuclear norms for tensors and their use for convex multilinear estimation.

Signoretto, M., Van de Plas, R., De Moor, B. and Suykens, J. (2011), Tensor versus matrix completion: A comparison with application to spectral data, *IEEE SPL*, **18(7)**, 403-406.

Tomioka, R., Hayashi, K. and Kashima, H. (2010), Estimation of low-rank tensors via convex optimization, arXiv preprint arXiv:1010.0789.

Tomioka, R., Suzuki, T., Hayashi, K. and Kashima, H. (2011), Statistical performance of convex tensor decomposition, *Advances in Neural Information Processing Systems (NIPS)*, 137.

Tropp, J. (2012), User-friendly tail bounds for sums of random matrices, *Foundations of Computational Mathematics*, **12**, 389-434.

Watson, G. A. (1992), Characterization of the subdifferential of some matrix norms. *Linear Algebra Appl.*, **170**, 33-45.