Foundations of Computational Mathematics (2019) 19:1265–1313
https://doi.org/10.1007/s10208-018-09408-6

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

On Polynomial Time Methods for Exact Low-Rank Tensor Completion

Dong Xia[1] · **Ming Yuan**[2]

Published online: 7 January 2019

Abstract
In this paper, we investigate the sample size requirement for exact recovery of a high-order tensor of low rank from a subset of its entries. We show that a gradient descent algorithm with initial value obtained from a spectral method can, in particular, reconstruct a $d \times d \times d$ tensor of multilinear ranks (r, r, r) with high probability from as few as $O(r^{7/2}d^{3/2}\log^{7/2} d + r^7 d \log^6 d)$ entries. In the case when the ranks $r = O(1)$, our sample size requirement matches those for nuclear norm minimization (Yuan and Zhang in Found Comput Math 1031–1068, 2016), or alternating least squares assuming orthogonal decomposability (Jain and Oh in Advances in Neural Information Processing Systems, pp 1431–1439, 2014). Unlike these earlier approaches, however, our method is efficient to compute, is easy to implement, and does not impose extra structures on the tensor. Numerical results are presented to further demonstrate the merits of the proposed approach.

Keywords Concentration inequality · Matrix completion · Nonconvex optimization · Polynomial time complexity · Tensor completion · Tensor rank · U-statistics

Mathematics Subject Classification Primary 90C25; Secondary 90C59 · 15A52

Communicated by Thomas Strohmer.

Ming Yuan's research was supported in part by NSF Grant DMS-1721584.

✉ Ming Yuan
ming.yuan@columbia.edu

[1] Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

[2] Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027, USA

1 Introduction

Let $\mathbf{T} \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ be a kth-order tensor. The goal of tensor completion is to recover $\mathbf{T}$ based on a subset of its entries $\{T(\omega) : \omega \in \Omega\}$ for some $\Omega \subset [d_1] \times \cdots \times [d_k]$ where $[d] = \{1, 2, \ldots, d\}$. The problem of tensor completion has attracted a lot of attention in recent years due to its wide range of applications. See, e.g., Li and Li [19], Sidiropoulos and Nion [29], Tomioka et al. [30], Gandy et al. [13], Cohen and Collins [7], Liu et al. [20], Anandkumar et al. [2], Mu et al. [23], Semerci et al. [28], Yuan and Zhang [33] and references therein. In particular, the second-order (matrix) case has been extensively studied. See, e.g., Candèes and Recht [5], Keshavan et al. [17], Candès and Tao [6], Gross [14], Recht [26] among many others. One of the main revelations from these studies is that, although the matrix completion problem is in general NP-hard, it is possible to develop tractable algorithms to achieve exact recovery with high probability. Naturally, one asks if the same can be said for higher-order tensors. This seemingly innocent task of generalizing from second- to higher-order tensors turns out to be rather delicate.

The challenges in dealing with higher-order tensors come from both computational and theoretical fronts. On the one hand, many of the standard operations for matrices become prohibitively expensive to compute for higher-order tensors. A notable example is the computation of tensor spectral norm. For second-order tensors, or matrices, the spectral norm is merely its largest singular value and can be computed with little effort. Yet, this is no longer the case for higher-order tensors where computing the spectral norm is NP-hard in general (see, e.g., [15]). On the other hand, many of the mathematical tools, either algebraic such as characterization of the subdifferential of the nuclear norm or probabilistic such as concentration inequalities, essential to the analysis of matrix completion, are still under development for higher-order tensors. There is a fast-growing literature to address both issues, and much progress has been made in both fronts in the past several years.

When it comes to higher-order tensor completion, an especially appealing idea is to first unfold a tensor into a matrix and then treat it using techniques for matrix completion. Notable examples include Tomioka et al. [30], Gandy et al. [13], Liu et al. [20], Mu et al. [23] among others. As shown recently by Yuan and Zhang [33], these approaches, although easy to implement, may require an unnecessarily large amount of entries to be observed to ensure exact recovery. As an alternative, Yuan and Zhang [33] established a sample size requirement for recovering a third-order tensor via nuclear norm minimization and showed that a $d \times d \times d$ tensor with multilinear ranks (r, r, r) can be recovered exactly with high probability with as few as $O((r^{1/2}d^{3/2} + r^2 d)(\log d)^2)$ entries observed. Perhaps more surprisingly, Yuan and Zhang [34] later showed that the dependence on d (e.g., the factor $d^{3/2}$) remains the same for higher-order tensors, and we can reconstruct a kth-order cubic tensor with as few as $O((r^{(k-1)/2}d^{3/2} + r^{k-1}d)(\log d)^2)$ entries for any $k \geq 3$ when minimizing a more specialized nuclear norm devised to take into account the incoherence. These sample size requirements drastically improve those based on unfolding which typically require a sample size of the order $r^{\lfloor k/2 \rfloor} d^{\lceil k/2 \rceil} \mathrm{polylog}(d)$ (see, e.g., [23]). Although both nuclear norm minimization approaches are based on convex optimization, they are also NP-hard to compute in general. Many approximate algorithms have also been

proposed in recent years with little theoretical justification. See, e.g., Kressner et al. [18], Rauhut and Stojanac [24], and Rauhut et al. [25]. It remains unknown whether there exist polynomial time algorithms that can recover a low-rank tensor exactly with similar sample size requirements. The goal of the present article is to fill in the gap between these two strands of research by developing a computationally efficient approach with a tight sample size requirement for completing a third-order tensor.

In particular, we show that there are polynomial time algorithms that can reconstruct a $d_1 \times d_2 \times d_3$ tensor with multilinear ranks (r_1, r_2, r_3) from as few as

$$O\left(r_1 r_2 r_3 (r d_1 d_2 d_3)^{1/2} \log^{7/2} d + (r_1 r_2 r_3)^2 r d \log^6 d\right)$$

entries where $r = \max\{r_1, r_2, r_3\}$ and $d = \max\{d_1, d_2, d_3\}$. This sample size requirement matches those for tensor nuclear norm minimization in terms of its dependence on the dimensions d_1, d_2 and d_3 although it is inferior in terms of its dependence on the ranks r_1, r_2 and r_3. This makes our approach especially attractive in practice because we are primarily interested in high-dimension (large d) and low-rank (small r) instances. In particular, when $r = O(1)$, our algorithms can recover a tensor exactly based on $O(d^{3/2} \log^{7/2} d)$ observed entries, which is nearly identical to that based on nuclear norm minimization. It is also worth noting that the sample size requirement we obtained is comparable to those for orthogonally decomposable tensors [16]. Unlike matrices, orthogonal decomposability is a rather restrictive assumption for higher-order tensors and our results suggest it may not be necessary after all.

It is known that the problem of tensor completion can be cast as optimization over a direct product of Grassmannians (see, e.g., [18]). The high-level idea behind our development is similar to those used earlier by Keshavan et al. [17] for matrix completion: If we can start with an initial value sufficiently close to the truth, then a small number of observed entries can ensure the convergence of typical optimization algorithms on Grassmannians such as gradient descent to the truth. Yet, the implementation of this strategy is much more delicate and poses significant new challenges when moving from matrices to tensors.

At the core of our method is the initialization of the linear subspaces in which the fibers of a tensor reside. In the matrix case, a natural way to do so is by singular value decomposition, a tool that is no longer available for higher-order tensors. An obvious solution is the so-called high-order singular value decomposition that unfolds tensors into matrices and then applies the usual singular value decomposition. This, however, requires an unnecessarily large sample size. To overcome this problem, we propose an alternative approach to estimating the singular spaces of the matrix unfoldings of a tensor. Our method is based on a carefully constructed estimate of the second moment of appropriate unfolding of a tensor, which can be viewed as a matrix version U-statistics. We show that the eigenspace of the proposed estimate concentrates around the true singular spaces of the matrix unfolding more sharply than the usual singular value decomposition-based approaches and therefore leads to consistent estimate with tighter sample size requirement.

The fact that there exist polynomial time algorithms to estimate a tensor consistently, not exactly, with $O(d^{3/2}\text{polylog}(r, \log d))$ observed entries was first recognized by

Barak and Moitra [3]. Their approach is based on sum-of-square relaxations of tensor nuclear norm. Although polynomial time solvable in principle, their method requires solving a semidefinite program of size $d^3 \times d^3$ and is not amenable to practical implementation. In contrast, our approach is essentially based on the spectral decomposition of a $d \times d$ matrix and can be computed fairly efficiently. Very recently, in independent work and under further restrictions on the tensor ranks, Montanari and Sun [22] showed that a spectral method different from ours can also achieve consistency with $O(d^{3/2}\mathrm{polylog}(r, \log d))$ observed entries. The rate of concentration for their estimate, however, is slower than ours, and as a result, it is unclear if it provides a sufficiently accurate initial value for the exact recovery with the said sample size.

Once a good initial value is obtained, we consider reconstructing a tensor by optimizing on a direct product of Grassmannians locally. To this end, we consider a simple gradient descent algorithm adapted for our purposes. The main architecture of our argument is similar to those taken by Keshavan et al. [17] for matrix completion. We argue that the objective function, in a suitable neighbor around the truth and including the initial value, behaves like a parabola. As a result, the gradient descent algorithm necessarily converges locally to a stationary point. We then show that the true tensor is indeed the only stationary point in the neighborhood and therefore the algorithm recovers the truth. To prove these statements for higher-order tensors, however, requires a number of new probabilistic tools for tensors, and we do so by establishing several new concentration bounds, building upon those from Yuan and Zhang [33,34].

The rest of the paper is organized as follows: We first review necessary concepts and properties of tensors for our purpose in the next section. Section 3 describes our main result with the initialization and local optimization steps being treated in detail in Sects. 4 and 5, respectively. Numerical experiments presented in Sect. 6 complement our theoretical development. We conclude with some discussions and remarks in Sect. 7. Proofs of the main results are relegated to Sect. 8.

2 Preliminaries

To describe our treatment of low-rank tensor completion, we first review a few basic and necessary facts and properties of tensors. In what follows, we shall denote a tensor or matrix by a boldfaced uppercase letter, and its entries the same uppercase letter in normal font with appropriate indices. Similarly, a vector will be denoted by a boldfaced lowercase letter, and its entries by the same letter in normal font. For notational simplicity, we shall focus primarily on third-order ($k = 3$) tensors, although our discussion can mostly be extended to higher-order tensor in a straightforward fashion. Subtle differences in treatment between third- and higher-order tensors will be discussed in Sect. 7.

The goal of tensor completion is to recover a tensor from partial observations of its entries. The problem is obviously underdetermined in general. To this end, we focus here on tensors that are of low multilinear ranks.

For a tensor $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, define the matrix $\mathcal{M}_1(\mathbf{A}) \in \mathbb{R}^{d_1 \times (d_2 d_3)}$ by the entries

$$\mathcal{M}_1(\mathbf{A})(i_1, (i_2 - 1)d_3 + i_3) = A(i_1, i_2, i_3), \quad \forall i_1 \in [d_1], i_2 \in [d_2], i_3 \in [d_3].$$

In other words, the columns of $\mathcal{M}_1(\mathbf{A})$ are the mode-1 fibers, $\{(A(i_1, i_2, i_3))_{i_1\in[d_1]} : i_2 \in [d_2], i_3 \in [d_3]\}$, of $\mathbf{A}$. We can define $\mathcal{M}_2$ and $\mathcal{M}_3$ in the same fashion. It is clear that $\mathcal{M}_j : \mathbb{R}^{d_1\times d_2\times d_3} \to \mathbb{R}^{d_j\times(d_1d_2d_3/d_j)}$ is a vector space isomorphism and often referred to as matricization or unfolding. The multilinear ranks of $\mathbf{A}$ are given by

$$\begin{aligned}
r_1(\mathbf{A}) &= \operatorname{rank}(\mathcal{M}_1(\mathbf{A})) = \dim(\operatorname{span}\{(A(i_1, i_2, i_3))_{i_1\in[d_1]} : i_2 \in [d_2], i_3 \in [d_3]\}),\\
r_2(\mathbf{A}) &= \operatorname{rank}(\mathcal{M}_2(\mathbf{A})) = \dim(\operatorname{span}\{(A(i_1, i_2, i_3))_{i_2\in[d_2]} : i_1 \in [d_1], i_3 \in [d_3]\}),\\
r_3(\mathbf{A}) &= \operatorname{rank}(\mathcal{M}_3(\mathbf{A})) = \dim(\operatorname{span}\{(A(i_1, i_2, i_3))_{i_3\in[d_3]} : i_1 \in [d_1], i_2 \in [d_2]\}).
\end{aligned}$$

Note that, in general, $r_1(\mathbf{A}) \neq r_2(\mathbf{A}) \neq r_3(\mathbf{A})$.

Let $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ be the left singular vectors of $\mathcal{M}_1(\mathbf{A})$, $\mathcal{M}_2(\mathbf{A})$ and $\mathcal{M}_3(\mathbf{A})$, respectively. It is not hard to see that there exists a so-called core tensor $\mathbf{C} \in \mathbb{R}^{r_1(\mathbf{A})\times r_2(\mathbf{A})\times r_3(\mathbf{A})}$ such that

$$\mathbf{A} = \sum_{j_1=1}^{r_1(\mathbf{A})}\sum_{j_2=1}^{r_2(\mathbf{A})}\sum_{j_3=1}^{r_3(\mathbf{A})} C(j_1, j_2, j_3)(\mathbf{u}_{j_1} \otimes \mathbf{v}_{j_2} \otimes \mathbf{w}_{j_3}), \tag{1}$$

where $\mathbf{u}_j$, $\mathbf{v}_j$ and $\mathbf{w}_j$ are the jth column of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$, respectively, and

$$\mathbf{x} \otimes \mathbf{y} \otimes \mathbf{z} := (x_{i_1} y_{i_2} z_{i_3})_{i_1\in[d_1], i_2\in[d_2], i_3\in[d_3]},$$

is a so-called rank-one tensor. Following the notation from de Silva and Lim [10], (1) can also be more compactly represented as a trilinear multiplication:

$$\mathbf{A} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C} := \mathbf{C} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W},$$

where the marginal product $\times_1 : \mathbb{R}^{r_1\times r_2\times r_3} \times \mathbb{R}^{d_1\times r_1} \to \mathbb{R}^{d_1\times r_2\times r_3}$ is given by

$$\mathbf{A} \times_1 \mathbf{B} = \left(\sum_{j_1=1}^{r_1} A(j_1, j_2, j_3)B(i_1, j_1)\right)_{i_1\in[d_1], j_2\in[r_2], j_3\in[r_3]},$$

and $\times_2$ and $\times_3$ are similarly defined.

The collection of all tensors of dimension $d_1 \times d_2 \times d_3$ whose multilinear ranks are at most $\mathbf{r} = (r_1, r_2, r_3)$ can be written as

$$\mathcal{A}(\mathbf{r}) = \left\{(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \cdot \mathbf{C} : \mathbf{X} \in \mathcal{V}(d_1, r_1), \mathbf{Y} \in \mathcal{V}(d_2, r_2), \mathbf{Z} \in \mathcal{V}(d_3, r_3), \mathbf{C} \in \mathbb{R}^{r_1\times r_2\times r_3}\right\},$$

where $\mathcal{V}(d, r)$ is the Stiefel manifold of orthonormal r-frames in $\mathbb{R}^d$. In fact, any tensor $\mathbf{A} \in \mathcal{A}(\mathbf{r})$ can be identified with a r_1-dimensional linear subspace in $\mathbb{R}^{d_1}$, a r_2-dimensional linear subspace in $\mathbb{R}^{d_2}$, a r_3-dimensional linear subspace in $\mathbb{R}^{d_3}$ and a core tensor in $\mathbb{R}^{r_1\times r_2\times r_3}$ so that $\mathcal{A}(\mathbf{r})$ is isomorphic to $\mathcal{G}(d_1, r_1) \times \mathcal{G}(d_2, r_2) \times \mathcal{G}(d_3, r_3) \times \mathbb{R}^{r_1\times r_2\times r_3}$ where $\mathcal{G}(d, r)$ is the Grassmannian of r-dimensional linear subspaces in $\mathbb{R}^d$.

Another common way of defining tensor ranks is through the so-called CP decomposition which expresses a tensor as the sum of the smallest possible number of rank-one tensors. The number of rank-one tensors in the CP decomposition of a tensor is commonly referred to as its CP rank. It is not hard to see that for a tensor of multilinear ranks (r_1, r_2, r_3), its CP rank is necessarily between $\max\{r_1, r_2, r_3\}$ and $\min\{r_1r_2, r_1r_3, r_2r_3\}$. We shall focus here primarily on multilinear ranks because it allows for stable numerical computation, as well as refined theoretical analysis. In addition, we can view a tensor of CP rank r also as a tensor with multilinear ranks no greater than (r, r, r). This allows us to straightforwardly translate the current result to tensors of low CP rank. However, it is worth noting this may lead to suboptimal dependence on r.

In addition to being of low rank, another essential property that $\mathbf{T}$ needs to satisfy so that we can possibly recover it from a uniformly sampled subset of its entries is the incoherence of linear subspaces spanned by its fibers (see, e.g., [5]). More specifically, let $\mathcal{X}$ be a r-dimensional linear subspace in $\mathbb{R}^d$ and $\mathbf{P}_\mathcal{X} : \mathbb{R}^d \to \mathbb{R}^d$ be its projection matrix. We can define the coherence for $\mathcal{X}$ as

$$\mu(\mathcal{X}) = \frac{d}{r} \max_{1\le i\le d} \|\mathbf{P}_\mathcal{X}\mathbf{e}_i\|^2,$$

where $\mathbf{e}_i$ is the ith canonical basis of an Euclidean space, that is, it is a vector whose ith entry is one and all other entries are zero. Note that

$$\mu(\mathcal{X}) = \frac{\max_{1\le i\le d} \|\mathbf{P}_\mathcal{X}\mathbf{e}_i\|^2}{d^{-1}\sum_{i=1}^d \|\mathbf{P}_\mathcal{X}\mathbf{e}_i\|^2},$$

for

$$\sum_{i=1}^d \|\mathbf{P}_\mathcal{X}\mathbf{e}_i\|^2 = \mathrm{trace}(\mathbf{P}_\mathcal{X}) = r.$$

Now for a tensor $\mathbf{A} \in \mathbb{R}^{d_1\times d_2\times d_3}$, denote by $\mathcal{U}(\mathbf{A})$ the linear space spanned by its mode-1 fibers, $\mathcal{V}(\mathbf{A})$ mode-2 fibers, and $\mathcal{W}(\mathbf{A})$ mode-3 fibers. With slight abuse of notation, we define the coherence of $\mathbf{A}$ as

$$\mu(\mathbf{A}) = \max\{\mu(\mathcal{U}(\mathbf{A})), \mu(\mathcal{V}(\mathbf{A})), \mu(\mathcal{W}(\mathbf{A}))\}.$$

Incoherence as defined here is a natural requirement for tensor recovery. It ensures that each fiber contains similar amount of information about the whole tensor and therefore allows for its recovery even if no entry of a particular fiber is observed. In particular, for any rank-(r_1, r_2, r_3) tensor $\mathbf{A} \in \mathbb{R}^{d_1\times d_2\times d_3}$,

$$\max_{\omega\in[d_1]\times[d_2]\times[d_3]} |A(\omega)| \le \sqrt{\frac{r_1r_2r_3}{d_1d_2d_3}} \cdot [\mu(\mathbf{A})]^{3/2}\|\mathbf{A}\|_\mathrm{F},$$

so, in a certain sense, the simultaneous incoherence rules out situations where some entries might be dominating and missing them could prevent us from reconstructing the original tensor.

In what follows, we shall also encounter various tensor norms. Recall that the vector space inner product between two tensors $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is defined as

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

The corresponding norm, referred to as Frobenius norm, for a tensor $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is given by

$$\|\mathbf{A}\|_{\mathrm{F}} := \langle \mathbf{A}, \mathbf{A} \rangle^{1/2}.$$

We can also define the spectral norm of $\mathbf{A}$ as

$$\|\mathbf{A}\| := \sup_{\mathbf{u}_j \in \mathbb{R}^{d_j} : \|\mathbf{u}_1\| = \|\mathbf{u}_2\| = \|\mathbf{u}_3\| = 1} \langle \mathbf{A}, \mathbf{u}_1 \otimes \mathbf{u}_2 \otimes \mathbf{u}_3 \rangle,$$

where, with slight abuse of notation, we write $\|\cdot\|$ both as the spectral norm for a tensor and as the usual ℓ_2 norm for a vector for brevity. The nuclear norm is the dual of spectral norm:

$$\|\mathbf{A}\|_\star = \sup_{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2 \times d_3}, \|\mathbf{X}\| \le 1} \langle \mathbf{A}, \mathbf{X} \rangle.$$

Another norm of interest is the max norm or the entrywise sup norm of $\mathbf{A}$:

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

The following relationships among these norms are immediate and stated here for completeness. We shall make use of them without mentioning throughout the rest of our discussion.

Lemma 1 *For any* $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$,

$$\|\mathbf{A}\|_{\max} \le \|\mathbf{A}\| \le \|\mathbf{A}\|_{\mathrm{F}} \le \sqrt{r_1(\mathbf{A}) r_2(\mathbf{A}) r_3(\mathbf{A})} \|\mathbf{A}\|,$$

and

$$\|\mathbf{A}\|_\star \le \min \left\{ \sqrt{r_1(\mathbf{A}) r_2(\mathbf{A})}, \sqrt{r_1(\mathbf{A}) r_3(\mathbf{A})}, \sqrt{r_2(\mathbf{A}) r_3(\mathbf{A})} \right\} \|\mathbf{A}\|_{\mathrm{F}}.$$

The proof of Lemma 1 is included in Appendix A for completeness. We are now in a position to describe our approach to tensor completion.

3 Tensor Completion

Assume that $\mathbf{T}$ has multilinear ranks $\mathbf{r} := (r_1, r_2, r_3)$ and coherence at most μ_0; we want to recover $\mathbf{T}$ based on $\big(\omega_i, T(\omega_i)\big)$ for $i = 1, 2, \ldots, n$ where ω_i are independently and uniformly drawn from $[d_1] \times [d_2] \times [d_3]$. This sampling scheme is often referred to the Bernoulli model, or sampling with replacement (see, e.g., [14,26]). Another commonly considered scheme is the so-called uniform sampling without replacement where we observe $T(\omega)$ for $\omega \in \Omega$ and Ω is a uniformly sampled subset of $[d_1] \times [d_2] \times [d_3]$ with size $|\Omega| = n$. It is known that both sampling schemes are closely related in that, given a uniformly sampled subset Ω of size n, one can always create a sample $\omega_i \in \Omega$, $i = 1, \ldots, n$ so that ω_is follow the Bernoulli model. This connection ensures that any method that works for Bernoulli model necessarily works for uniform sampling without replacement as well. From a technical point of view, it has been demonstrated that working with the Bernoulli model leads to considerably simpler arguments for a number of matrix or tensor completion approaches. See, e.g., Gross [14], Recht [26], Yuan and Zhang [33], among others. For these reasons, we shall focus on the Bernoulli model in the current work.

A natural way to solve this problem is through the following optimization:

$$\min_{\mathbf{A} \in \mathcal{A}(\mathbf{r})} \frac{1}{2} \|\mathcal{P}_\Omega(\mathbf{A} - \mathbf{T})\|_{\mathrm{F}}^2,$$

where the linear operator $\mathcal{P}_\Omega : \mathbb{R}^{d_1 \times d_2 \times d_3} \to \mathbb{R}^{d_1 \times d_2 \times d_3}$ is given by

$$\mathcal{P}_\Omega \mathbf{X} = \sum_{i=1}^{n} \mathcal{P}_{\omega_i} \mathbf{X},$$

and $\mathcal{P}_\omega \mathbf{X}$ is a $d_1 \times d_2 \times d_3$ tensor whose ω entry is $X(\omega)$ and other entries are zero. Equivalently, we can reconstruct $\mathbf{T} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{G}$ by $\widehat{\mathbf{T}} := (\widehat{\mathbf{U}}, \widehat{\mathbf{V}}, \widehat{\mathbf{W}}) \cdot \widehat{\mathbf{G}}$ where the tuple $(\widehat{\mathbf{U}}, \widehat{\mathbf{V}}, \widehat{\mathbf{W}}, \widehat{\mathbf{G}})$ solves

$$\min_{\mathbf{X} \in \mathcal{V}(d_1, r_1), \mathbf{Y} \in \mathcal{V}(d_2, r_2), \mathbf{Z} \in \mathcal{V}(d_3, r_3), \mathbf{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}} \frac{1}{2} \|\mathcal{P}_\Omega((\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \cdot \mathbf{C} - \mathbf{T})\|_{\mathrm{F}}^2. \tag{2}$$

Recall that $\mathbf{X} \otimes \mathbf{Y} \otimes \mathbf{Z}$ is a sixth-order tensor of dimension $d_1 \times d_2 \times d_3 \times r_1 \times r_2 \times r_3$. With slight abuse of notation, for any $\omega \in [d_1] \times [d_2] \times [d_3]$, denote by $(\mathbf{X} \otimes \mathbf{Y} \otimes \mathbf{Z})(\omega)$ a third-order tensor with the first three indices of $\mathbf{X} \otimes \mathbf{Y} \otimes \mathbf{Z}$ fixed at ω. By the first-order optimality condition, we get

$$\sum_{i=1}^{n} \langle (\mathbf{X} \otimes \mathbf{Y} \otimes \mathbf{Z})(\omega_i), \mathbf{C} \rangle \, (\mathbf{X} \otimes \mathbf{Y} \otimes \mathbf{Z})(\omega_i) = \sum_{i=1}^{n} T(\omega_i)(\mathbf{X} \otimes \mathbf{Y} \otimes \mathbf{Z})(\omega_i),$$

so that

$$\operatorname{vec}(\mathbf{C}) = \left(\sum_{i=1}^{n} \operatorname{vec}((\mathbf{X}\otimes\mathbf{Y}\otimes\mathbf{Z})(\omega_i))\operatorname{vec}((\mathbf{X}\otimes\mathbf{Y}\otimes\mathbf{Z})(\omega_i))^{\top}\right)^{-1} \times \left(\sum_{i=1}^{n} T(\omega_i)\operatorname{vec}((\mathbf{X}\otimes\mathbf{Y}\otimes\mathbf{Z})(\omega_i))\right). \tag{3}$$

Here, we assumed implicitly that $n \geq r_1 r_2 r_3$. In general, there may be multiple minimizers to (2) and we can replace the inverse by the Moore–Penrose pseudoinverse to yield a solution. Plugging it back to (2) suggests that $(\widehat{\mathbf{U}}, \widehat{\mathbf{V}}, \widehat{\mathbf{W}})$ is the solution to

$$\max_{\mathbf{X}\in\mathcal{V}(d_1,r_1),\mathbf{Y}\in\mathcal{V}(d_2,r_2),\mathbf{Z}\in\mathcal{V}(d_3,r_3)} F(\mathbf{X},\mathbf{Y},\mathbf{Z}),$$

where

$$F(\mathbf{X},\mathbf{Y},\mathbf{Z}) = \left(\sum_{i=1}^{n} T(\omega_i)\operatorname{vec}((\mathbf{X}\otimes\mathbf{Y}\otimes\mathbf{Z})(\omega_i))\right)^{\top} \times \left(\sum_{i=1}^{n} \operatorname{vec}((\mathbf{X}\otimes\mathbf{Y}\otimes\mathbf{Z})(\omega_i))\operatorname{vec}((\mathbf{X}\otimes\mathbf{Y}\otimes\mathbf{Z})(\omega_i))^{\top}\right)^{-1} \times \left(\sum_{i=1}^{n} T(\omega_i)\operatorname{vec}((\mathbf{X}\otimes\mathbf{Y}\otimes\mathbf{Z})(\omega_i))\right).$$

Let $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{Q}_1$, $\tilde{\mathbf{Y}} = \mathbf{Y}\mathbf{Q}_2$ and $\tilde{\mathbf{Z}} = \mathbf{Z}\mathbf{Q}_3$, where $\mathbf{Q}_j \in \mathcal{O}(r_j)$ and $\mathcal{O}(r)$ is the set of $r \times r$ orthonormal matrices. It is easy to verify that

$$F(\mathbf{X},\mathbf{Y},\mathbf{Z}) = F(\tilde{\mathbf{X}},\tilde{\mathbf{Y}},\tilde{\mathbf{Z}})$$

so that it suffices to optimize $F(\mathbf{X},\mathbf{Y},\mathbf{Z})$ over

$$(\mathbf{X},\mathbf{Y},\mathbf{Z}) \in (\mathcal{V}(d_1,r_1)/\mathcal{O}(r_1)) \times (\mathcal{V}(d_2,r_2)/\mathcal{O}(r_2)) \times (\mathcal{V}(d_3,r_3)/\mathcal{O}(r_3)).$$

Recall that $\mathcal{V}(d,r)/\mathcal{O}(r) \cong \mathcal{G}(d,r)$, the Grassmannian of r-dimensional linear subspace in $\mathbb{R}^d$. Optimizing F can then be cast as an optimization problem over a direct product of Grassmannian manifolds, a problem that has been well studied in the literature. See, e.g., Absil et al. [1]. In particular, (quasi-)Newton (see, e.g., [12,27]), gradient descent (see, e.g., [17]) and conjugate gradient (see, e.g., [18]) methods have all been proposed previously to solve optimization problems similar to the one we consider here.

There are two prerequisites for any of these methods to be successful. The highly nonconvex nature of the optimization problem dictates that even if any of the aforementioned iterative algorithms converges, it could only converge to a local optimum.

Therefore, a good initial value is critical. This unfortunately is an especially challenging task for tensors. For example, if we consider random initial values, then a prohibitively large number, in fact exponential in d, of seeds would be required to ensure the existence of a good starting point. Alternatively, in the second-order or matrix case, Keshavan et al. [17] suggests a singular value decomposition-based approach for initialization. The method, however, cannot be directly applied for higher-order tensors as similar type of spectral decomposition becomes NP-hard to compute [15]. To address this challenge, we propose here a new spectral method that is efficient to compute and at the same time is guaranteed to produce an initial value sufficiently close to the optimal value.

With the initial value coming from a neighborhood near the truth, any of the aforementioned methods could then be applied in principle. In order for them to converge to the truth, we need to make sure that the objective function F behaves well in the neighborhood. In particular, we shall show that, when n is sufficiently large, F behaves like a parabola in a neighborhood around the truth and therefore ensures the local convergence of algorithms such as gradient descent.

We shall address both aspects, initialization and local convergence, separately in the next two sections. In summary, we can obtain a sample size requirement for exact recovery of $\mathbf{T}$ via polynomial time algorithms. As in the matrix case, the sample size requirement depends on notions of condition number of $\mathbf{T}$. Recall that the condition number for a matrix $\mathbf{A}$ is given by $\kappa(\mathbf{A}) = \sigma_{\max}(\mathbf{A})/\sigma_{\min}(\mathbf{A})$ where $\sigma_{\max}$ and $\sigma_{\min}$ are the largest and smallest nonzero singular values of $\mathbf{A}$, respectively. We can straightforwardly generalize the concept to a third-order tensor $\mathbf{A}$ as:

$$\kappa(\mathbf{A}) = \frac{\max\{\sigma_{\max}(\mathcal{M}_1(\mathbf{A})), \sigma_{\max}(\mathcal{M}_2(\mathbf{A})), \sigma_{\max}(\mathcal{M}_3(\mathbf{A}))\}}{\min\{\sigma_{\min}(\mathcal{M}_1(\mathbf{A})), \sigma_{\min}(\mathcal{M}_2(\mathbf{A})), \sigma_{\min}(\mathcal{M}_3(\mathbf{A}))\}}.$$

Our main result can then be summarized as follows:

Theorem 1 *Assume that $\mathbf{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is a rank-(r_1, r_2, r_3) tensor whose coherence is bounded by $\mu(\mathbf{T}) \le \mu_0$ and condition number is bounded by $\kappa(\mathbf{T}) \le \kappa_0$. Then there exists a polynomial time algorithm that recovers $\mathbf{T}$ exactly based on $\{(\omega_i, T(\omega_i)) : 1 \le i \le n\}$, with probability at least $1 - d^{-\alpha}$ if ω_is are independently and uniformly sampled from $[d_1] \times [d_2] \times [d_3]$ and*

$$n \ge C\left\{\alpha^3 \mu_0^3 \kappa_0^4 r_1 r_2 r_3 (r d_1 d_2 d_3)^{1/2} \log^{7/2} d + \alpha^6 \mu_0^6 \kappa_0^8 (r_1 r_2 r_3)^2 r d \log^6 d\right\}, \quad (4)$$

for a universal constant $C > 0$, and an arbitrary constant $\alpha \ge 1$, where $d = \max\{d_1, d_2, d_3\}$ and $r = \max\{r_1, r_2, r_3\}$.

In particular, we shall show that the following algorithm indeed achieves the sample size requirement given by Theorem 1.

The next two sections will be devoted to the analysis of the second-order spectral algorithm and gradient descent algorithm, respectively. These results, together with the polynomial time complexity of both algorithms, immediately imply the validity of Algorithm 1 and hence Theorem 1.

Algorithm 1 Tensor completion

Run the second-order spectral algorithm (Algorithm 2) to initialized $\mathbf{U}$, and similarly $\mathbf{V}$ and $\mathbf{W}$. Denote these initial values as $(\mathbf{U}^{(0)}, \mathbf{V}^{(0)}, \mathbf{W}^{(0)})$.
2: Run the gradient descent algorithm (Algorithm 3) with initial value $(\mathbf{U}^{(0)}, \mathbf{V}^{(0)}, \mathbf{W}^{(0)})$. Denote the output by $\widehat{\mathbf{T}}$.
Return $\mathbf{T}$.

4 Second-Order Method for Estimating Singular Spaces

We now describe a spectral algorithm that produces good initial values for $\mathbf{U}$ and $\mathbf{V}$ and $\mathbf{W}$ based on $\mathcal{P}_\Omega \mathbf{T}$. To fix ideas, we focus on estimating $\mathbf{U}$. $\mathbf{V}$ and $\mathbf{W}$ can be treated in an identical fashion. Denote

$$\widehat{\mathbf{T}} = \frac{d_1 d_2 d_3}{n} \mathcal{P}_\Omega \mathbf{T}.$$

It is clear that $\mathbb{E}(\widehat{\mathbf{T}}) = \mathbf{T}$ so that $\mathcal{M}_1(\widehat{\mathbf{T}})$ is an unbiased estimate of $\mathcal{M}_1(\mathbf{T})$. Recall that $\mathbf{U}$ is the left singular vectors of $\mathcal{M}_1(\mathbf{T})$; it is therefore natural to consider estimating $\mathbf{U}$ by the leading singular vectors of $\mathcal{M}_1(\widehat{\mathbf{T}})$. The main limitation of this naïve approach is its inability to take advantage of the fact that $\mathcal{M}_1(\widehat{\mathbf{T}})$ may be unbalanced in that $d_1 \ll d_2 d_3$, and the quality of an estimate of $\mathbf{U}$ is driven largely by the greater dimension $(d_2 d_3)$ although we are only interested in estimating the singular space in a lower-dimensional (d_1) space.

To specifically address this issue, we consider here a different technique for estimating singular spaces from a noisy matrix, which is more powerful when the underlying matrix is unbalanced in that it is either very fat or very tall. More specifically, let $\mathbf{M} \in \mathbb{R}^{m_1 \times m_2}$ be a rank r matrix. Our goal is to estimate the left singular space of $\mathbf{M}$ based on n pairs of observations $\{(\omega_i, \mathbf{M}(\omega_i)) : 1 \le i \le n\}$ where ω_is are independently and uniformly sampled from $[m_1] \times [m_2]$. Recall that $\mathbf{U}$ is also the eigenspace of $\mathbf{M}\mathbf{M}^\top$ which is of dimension $m_1 \times m_1$. Instead of estimating $\mathbf{M}$, we shall consider instead estimating $\mathbf{M}\mathbf{M}^\top$. To this end, write $\mathbf{X}_i = (m_1 m_2)\mathcal{P}_{\omega_i}\mathbf{M}$, that is a $m_1 \times m_2$ matrix whose ω_i entry is $(m_1 m_2)\mathbf{M}(\omega_i)$, and other entries are zero. It is clear that $\mathbb{E}(\mathbf{X}_i) = \mathbf{M}$. We shall then consider estimating $\mathbf{N} := \mathbf{M}\mathbf{M}^\top$ by

$$\widehat{\mathbf{N}} := \frac{1}{n(n-1)} \sum_{i<j} (\mathbf{X}_i \mathbf{X}_j^\top + \mathbf{X}_j \mathbf{X}_i^\top) \tag{5}$$

Note that $\mathbf{X}_i$ has only a single nonzero entry so that each summand on the right-hand side of (5) can be computed in constant time. In total, computing $\widehat{\mathbf{N}}$ has the time complexity of $O_p(n^2)$. Our first result shows that $\widehat{\mathbf{N}}$ could be a very good estimate of $\mathbf{N}$ even in situations when $n \ll m_2$.

Theorem 2 *Let $\mathbf{M} \in \mathbb{R}^{m_1 \times m_2}$ and $\mathbf{X}_i = (m_1 m_2)\mathcal{P}_{\omega_i}\mathbf{M}$ $(i = 1, 2, \ldots, n)$, where ω_is are independently and uniformly sampled from $[m_1] \times [m_2]$. There exists an absolute constant $C > 0$ such that for any $\alpha > 1$, if*

$$n \geq \frac{8}{3}\frac{(\alpha+1)\log m}{\min\{m_1, m_2\}}, \quad m := \max\{m_1, m_2\} \geq 2$$

then

$$\begin{aligned}
&\|\widehat{\mathbf{N}} - \mathbf{M}\mathbf{M}^\top\| \\
&\quad \leq C \cdot \alpha^2 \cdot \frac{m_1^{3/2} m_2^{3/2} \log m}{n} \\
&\qquad \times \left[\left(1 + \frac{m_1}{m_2}\right)^{1/2} + \frac{m_1^{1/2} m_2^{1/2}}{n} + \left(\frac{n}{m_2 \log m}\right)^{1/2}\right] \cdot \|\mathbf{M}\|_{\max}^2,
\end{aligned}$$

with probability at least $1 - m^{-\alpha}$*, where* $\widehat{\mathbf{N}}$ *is given by* (5).

In particular, if $\|\mathbf{M}\|_{\max} = O((m_1 m_2)^{-1/2})$, then $\|\widehat{\mathbf{N}} - \mathbf{M}\mathbf{M}^\top\| \to_p 0$ as soon as $n \gg \left((m_1 m_2)^{1/2} + m_1\right) \log m$. This is in contrast to estimating $\mathbf{M}$. As shown by Recht [26],

$$\widehat{\mathbf{M}} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$$

is a consistent estimate of $\mathbf{M}$ in spectral norm if $n \gg m \log m$. The two sample size requirements differ when $m_1 \ll m_2$ in which case $\widehat{\mathbf{N}}$ is still a consistent estimate of $\mathbf{M}\mathbf{M}^\top$, yet $\widehat{\mathbf{M}}$ is no longer a consistent estimate of $\mathbf{M}$ if $(m_1 m_2)^{1/2} \log m_2 \ll n \ll m_2 \log m_2$.

Equipped with Theorem 2, we can now address the initialization of $\mathbf{U}$ (and similarly $\mathbf{V}$ and $\mathbf{W}$). Instead of estimating it by the singular vectors of $\mathcal{M}_1(\widehat{\mathbf{T}})$, we shall do so based on an estimate of $\mathcal{M}_1(\mathbf{T})\mathcal{M}_1(\mathbf{T})$. With slight abuse of notation, write $\mathbf{X}_i = (d_1 d_2 d_3)\mathcal{M}_1(\mathcal{P}_{\omega_i}\mathbf{T})$ and

$$\widehat{\mathbf{N}} := \frac{1}{n(n-1)}\sum_{i<j}(\mathbf{X}_i\mathbf{X}_j^\top + \mathbf{X}_j\mathbf{X}_i^\top).$$

We shall then estimate $\mathbf{U}$ by the leading r left singular vectors of $\widehat{\mathbf{N}}$, hereafter denoted by $\widehat{\mathbf{U}}$.

As we are concerned with the linear spaces spanned by the column vector of $\mathbf{U}$ and $\widehat{\mathbf{U}}$, respectively, we can measure the estimation error by the projection distance defined over Grassmannian:

$$d_{\mathrm{p}}(\mathbf{U}, \widehat{\mathbf{U}}) := \frac{1}{\sqrt{2}}\|\mathbf{U}\mathbf{U}^\top - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\|_{\mathrm{F}}.$$

The following result is an immediate consequence of Theorem 2 and Davis–Kahan theorem, and its proof is deferred to Appendix.

Corollary 1 *Assume that* $\mathbf{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ *is a rank-*(r_1, r_2, r_3) *tensor whose coherence is bounded by* $\mu(\mathbf{T}) \le \mu_0$ *and condition number is bounded by* $\kappa(\mathbf{T}) \le \kappa_0$. *Let* $\mathbf{U}$ *be the left singular vectors of* $\mathcal{M}_1(\mathbf{T})$ *and* $\widehat{\mathbf{U}}$ *be defined as above, then there exist absolute constants* $C_1, C_2 > 0$ *such that for any* $\alpha > 1$, *if*

$$n \ge C_1 \left(\alpha (d_1 d_2 d_3)^{1/2} + d_1 \log d \right),$$

then

$$d_{\mathrm{p}}(\mathbf{U}, \widehat{\mathbf{U}}) \le C_2 \alpha^2 \mu_0^3 \kappa_0^2 r_1^{3/2} r_2 r_3 \left(\frac{(d_1 d_2 d_3)^{1/2} \log d}{n} + \sqrt{\frac{d_1 \log d}{n}} \right),$$

with probability at least $1 - d^{-\alpha}$.

In the light of Corollary 1, $\widehat{\mathbf{U}}$ (and similarly $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{W}}$) is a consistent estimate of $\mathbf{U}$ whenever

$$n \gg \left[r_1^{3/2} r_2 r_3 (d_1 d_2 d_3)^{1/2} + r_1^3 r_2^2 r_3^2 d \right] \log d.$$

In addition, it is worth noting that $\widehat{\mathbf{U}}$ can be computed effectively via truncated singular value decomposition. Since $\widehat{\mathbf{N}}$ is a $d_1 \times d_1$ matrix, the time complexity for doing so is $O(d_2 d_3 + r_1^2 d_1)$.

In order to be used as an initial value in our algorithm for optimizing F, we also need to make sure that $\widehat{\mathbf{U}}$ is incoherence. However, this may not always be the case. Fortunately, because $\widehat{\mathbf{U}}$ is close to an incoherent basis $\mathbf{U}$, we can readily derive an initial value that is both incoherent and remains close to $\mathbf{U}$, an observation made earlier by Keshavan et al. [17]. In particular, an initial value for optimizing F can be obtained via the following algorithm.

Algorithm 2 Second-order spectral estimate of $\mathbf{U}$

Compute

$$\widehat{\mathbf{N}} := \frac{1}{n(n-1)} \sum_{i<j} (\mathbf{X}_i \mathbf{X}_j^\top + \mathbf{X}_j \mathbf{X}_i^\top).$$

where $\mathbf{X}_i = (d_1 d_2 d_3) \mathcal{M}_1(\mathcal{P}_{\omega_i} \mathbf{T})$.

2: Compute the truncated SVD for $\widehat{\mathbf{N}}$ and denote by $\widehat{\mathbf{U}}$ be the top r leading left singular vectors. Let $\tilde{\mathbf{U}}$ be a $d_1 \times r$ matrix whose ith row is given by

$$\tilde{\mathbf{U}}^{(i)} = \frac{\widehat{\mathbf{U}}^{(i)}}{\|\widehat{\mathbf{U}}^{(i)}\|} \cdot \min\{\|\widehat{\mathbf{U}}^{(i)}\|, \sqrt{\mu_0 r}\}, \quad i = 1, \ldots, d_1,$$

where $\widehat{\mathbf{U}}^{(i)}$ is the ith row vector of $\widehat{\mathbf{U}}$.

4: Return $\tilde{\mathbf{U}}(\tilde{\mathbf{U}}^\top \tilde{\mathbf{U}})^{-1/2}$.

Following the discussion earlier, the running time of Algorithm 2 is $O(n^2)$ under the settings of Corollary 1.

5 Exact Recovery by Optimizing Locally

Now that a good initial value sufficiently close to $(\mathbf{U}, \mathbf{V}, \mathbf{W})$ is identified, we can then proceed to optimize

$$F(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \min_{\mathbf{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}} \frac{1}{2} \|\mathcal{P}_\Omega((\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \cdot \mathbf{C} - \mathbf{T})\|_{\mathrm{F}}^2$$

locally. To this end, we argue that F indeed is well behaved in a neighborhood around $(\mathbf{U}, \mathbf{V}, \mathbf{W})$ so that such a local optimization is amenable to computation. For brevity, write

$$\mathcal{J}(d_1, d_2, d_3, r_1, r_2, r_3) := \mathcal{G}(d_1, r_1) \times \mathcal{G}(d_2, r_2) \times \mathcal{G}(d_3, r_3).$$

We can also generalize the projection distance d_{p} on Grassmannian to $\mathcal{J}(d_1, d_2, d_3, r_1, r_2, r_3)$ as follows:

$$d_{\mathrm{p}}\left((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\right) = d_{\mathrm{p}}(\mathbf{U}, \mathbf{X}) + d_{\mathrm{p}}(\mathbf{V}, \mathbf{Y}) + d_{\mathrm{p}}(\mathbf{W}, \mathbf{Z}).$$

We shall focus, in particular, on a neighborhood around $(\mathbf{U}, \mathbf{V}, \mathbf{W})$ that are incoherent:

$$\mathcal{N}(\delta, \mu) = \Big\{ (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in \mathcal{J}(d_1, d_2, d_3, r_1, r_2, r_3) : d_{\mathrm{p}}\left((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\right) \le \delta, \\ \text{and } \max\{\mu(\mathbf{X}), \mu(\mathbf{Y}), \mu(\mathbf{Z})\} \le \mu \Big\}$$

For a third-order tensor $\mathbf{A}$, denote

$$\Lambda_{\max}(\mathbf{A}) = \max\left\{\sigma_{\max}(\mathcal{M}_1(\mathbf{A})), \sigma_{\max}(\mathcal{M}_2(\mathbf{A})), \sigma_{\max}(\mathcal{M}_3(\mathbf{A}))\right\},$$

and

$$\Lambda_{\min}(\mathbf{A}) = \min\left\{\sigma_{\min}(\mathcal{M}_1(\mathbf{A})), \sigma_{\min}(\mathcal{M}_2(\mathbf{A})), \sigma_{\min}(\mathcal{M}_3(\mathbf{A}))\right\}.$$

Theorem 3 *Let $\mathbf{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ be a rank-(r_1, r_2, r_3) tensor such that*

$$\mu(\mathbf{T}) \le \mu_0, \quad \Lambda_{\min}(\mathbf{T}) \ge \underline{\Lambda}, \quad \Lambda_{\max}(\mathbf{T}) \le \overline{\Lambda}, \quad \text{and} \quad \kappa(\mathbf{T}) \le \kappa_0.$$

There exist absolute constants $C_1, C_2, C_3, C_4, C_5 > 0$ such that for any $\alpha > 1$ and $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in \mathcal{N}\big(C_1(\alpha\kappa_0 \log d)^{-1}, 4\mu_0\big)$,

$$C_2\left(\|\mathbf{G} - \mathbf{C}\|_{\mathrm{F}}^2 + \underline{\Lambda}^2 d_{\mathrm{p}}^2\left((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\right)\right) \le \frac{d_1 d_2 d_3}{n} F(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$$

$$\leq C_3 \alpha \overline{\Lambda}^2 d_{\mathrm{p}}^2 \left((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\right) \log d, \tag{6}$$

and

$$\frac{d_1 d_2 d_3}{n} \|\mathrm{grad}\ F(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\|_{\mathrm{F}} \geq C_4 \left(\underline{\Lambda}^2 d_{\mathrm{p}} \left((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\right)\right), \tag{7}$$

with probability at least $1 - 3d^{-\alpha}$*, provided that*

$$n \geq C_5 \left\{\alpha^3 \mu_0^{3/2} \kappa_0^4 r (r_1 r_2 r_3 d_1 d_2 d_3)^{1/2} \log^{7/2} d + \alpha^6 \mu_0^3 \kappa_0^8 r_1 r_2 r_3 r^2 d \log^6 d\right\}$$

where $\mathbf{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ *is given by* (3).

Theorem 3 shows that the objective function F behaves like a parabola in $\mathcal{N}(\delta, 4\mu_0)$ for sufficiently small δ, and furthermore, $(\mathbf{U}, \mathbf{V}, \mathbf{W})$ is the unique stationary point in $\mathcal{N}(\delta, 4\mu_0)$. This implies that a gradient descent type of algorithm may be employed to optimize F within $\mathcal{N}(\delta, 4\mu_0)$. In particular, to fix ideas, we shall focus here on a simple gradient descent type of algorithms similar to the popular choice of matrix completion algorithm proposed by Keshavan et al. [17]. As suggested by Keshavan et al. [17], to guarantee that the coherence condition is satisfied, a penalty function is imposed so that the objective function becomes:

$$\tilde{F}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) := F(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) + G(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$$

where

$$G(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) := \rho \sum_{j_1=1}^{d_1} G_0\Big(\frac{d_1 \|\mathbf{x}_{j_1}\|^2}{3\mu_0 r_1}\Big) + \rho \sum_{j_2=1}^{d_2} G_0\Big(\frac{d_2 \|\mathbf{y}_{j_2}\|^2}{3\mu_0 r_2}\Big) + \rho \sum_{j_3=1}^{d_3} G_0\Big(\frac{d_3 \|\mathbf{z}_{j_3}\|^2}{3\mu_0 r_3}\Big)$$

and

$$G_0(z) = \begin{cases} 0, & \text{if } z \leq 1 \\ e^{(z-1)^2} - 1, & \text{if } z \geq 1. \end{cases}$$

It turns out that, with a sufficiently large $\rho > 0$, we can ensure low coherence at all iterations in a gradient descent algorithm. More specifically, let $\mathbf{B} \in \mathbb{R}^{d \times r}$ be an element of the tangent space at $\mathbf{A} \in \mathcal{G}(d, r)$ and $\mathbf{B} = \mathbf{L}\boldsymbol{\Theta}\mathbf{R}^\top$ be its singular value decomposition. The geodesic starting from $\mathbf{A}$ in the direction $\mathbf{B}$ is defined as $\mathcal{H}(\mathbf{A}, \mathbf{B}, t) = \mathbf{A}\mathbf{R}\cos(\boldsymbol{\Theta} t)\mathbf{R}^\top + \mathbf{L}\sin(\boldsymbol{\Theta} t)\mathbf{R}^\top$ for $t \geq 0$. Interested readers are referred to Edelman et al. [11] for further details on the differential geometry of Grassmannians. The gradient descent algorithm on the direct product of Grassmannians is given below:

Algorithm 3 Gradient descent algorithm on Grassmannians (GoG)

Set up values of max _Iteration, tolerance $\varepsilon_{\text{tol}} > 0$, parameter $\gamma = \frac{\delta}{4}$, step counter $k = 0$ and initial value $(\mathbf{X}^{(0)}, \mathbf{Y}^{(0)}, \mathbf{Z}^{(0)})$.
while $k <$ max _Iteration **do**
  Compute the negative gradient $(\mathbf{D}_{\mathbf{X}}^{(k)}, \mathbf{D}_{\mathbf{Y}}^{(k)}, \mathbf{D}_{\mathbf{Z}}^{(k)}) = -\text{grad}\, \tilde{F}(\mathbf{X}^{(k)}, \mathbf{Y}^{(k)}, \mathbf{Z}^{(k)})$
  For $t \geq 0$, denote the geodesics

$$\mathbf{X}^{(k)}(t) = \mathcal{H}(\mathbf{X}^{(k)}, \mathbf{D}_{\mathbf{X}}^{(k)}, t)$$
$$\mathbf{Y}^{(k)}(t) = \mathcal{H}(\mathbf{Y}^{(k)}, \mathbf{D}_{\mathbf{Y}}^{(k)}, t)$$
$$\mathbf{Z}^{(k)}(t) = \mathcal{H}(\mathbf{Z}^{(k)}, \mathbf{D}_{\mathbf{Z}}^{(k)}, t)$$

  Minimize $t \mapsto \tilde{F}(\mathbf{X}^{(k)}(t), \mathbf{Y}^{(k)}(t), \mathbf{Z}^{(k)}(t))$ for $t \geq 0$, subject to

$$d_{\text{p}}\big((\mathbf{X}^{(k)}(t), \mathbf{Y}^{(k)}(t), \mathbf{Z}^{(k)}(t)), (\mathbf{X}^{(0)}, \mathbf{Y}^{(0)}, \mathbf{Z}^{(0)})\big) \leq \gamma.$$

  Set $\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)}(t_k)$, $\mathbf{Y}^{(k+1)} = \mathbf{Y}^{(k)}(t_k)$ and $\mathbf{Z}^{(k+1)} = \mathbf{Z}^{(k)}(t_k)$ where t_k is the minimal solution. Set $k = k + 1$.
  if $d_{\text{p}}\big((\mathbf{X}^{(k)}(t), \mathbf{Y}^{(k)}(t), \mathbf{Z}^{(k)}(t)), (\mathbf{X}^{(k-1)}, \mathbf{Y}^{(k-1)}, \mathbf{Z}^{(k-1)})\big) \leq \varepsilon_{\text{tol}}$ **then**
    break;
  end if
end while
Return $F(\mathbf{X}^{(k)}, \mathbf{Y}^{(k)}, \mathbf{Z}^{(k)})$.

Our next result shows that this algorithm indeed converges to $(\mathbf{U}, \mathbf{V}, \mathbf{W})$ when an appropriate initial value is provided.

Theorem 4 *Let* $\mathbf{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ *be a rank-*(r_1, r_2, r_3) *tensor such that*

$$\mu(\mathbf{T}) \leq \mu_0, \quad \Lambda_{\max}(\mathbf{T}) \leq \overline{\Lambda}, \quad \text{and} \quad \kappa(\mathbf{T}) \leq \kappa_0.$$

Then there exist absolute constants $C_1, C_2, C_3 > 0$ *such that for any* $\alpha > 1$, *if*

$$\rho \geq C_1 \alpha n (d_1 d_2 d_3)^{-1} \overline{\Lambda}^2 \log d,$$
$$(\mathbf{X}^{(0)}, \mathbf{Y}^{(0)}, \mathbf{Z}^{(0)}) \in \mathcal{N}\big(C_2(\alpha \kappa_0^2 \log d)^{-1}, 3\mu_0\big),$$

and

$$n \geq C_3 \left\{ \alpha^3 \mu_0^{3/2} \kappa_0^4 r (r_1 r_2 r_3 d_1 d_2 d_3)^{1/2} \log^{7/2} d + \alpha^6 \mu_0^3 \kappa_0^8 r_1 r_2 r_3 r^2 d \log^6 d \right\},$$

then Algorithm 3 initiated with $(\mathbf{X}^{(0)}, \mathbf{Y}^{(0)}, \mathbf{Z}^{(0)})$ *converges to* $\mathbf{T}$ *with probability at least* $1 - d^{-\alpha}$.

Theorem 4 shows that the gradient descent algorithm presented here indeed converges to the true tensor. In the light of the explicit formulas for the gradient and geodesic, each iteration of the algorithm has a time complexity $O(r_1^2 d_1 + r_2^2 d_2 + r^3 d_3 + r_1 r_2 r_3)$. The total computational cost of our method depends on the convergence rate of the gradient descent algorithm. Our experience with the numerical

experiments as reported in the next section seems to suggest a linear convergence rate as often expected of similar algorithms. A more rigorous investigation of the rate of convergence for the gradient descent algorithm is beyond the scope of the current work, and we shall leave it for future investigation.

6 Numerical Experiments

To complement our theoretical developments, we also conducted several sets of numerical experiments to investigate the performance of the proposed approach. In particular, we focus on recovering a cubic tensor $\mathbf{T} \in \mathbb{R}^{d\times d\times d}$ with multilinear ranks $r_1 = r_2 = r_3 = r$ from n randomly sampled entries. To fix ideas, we focus on completing orthogonal decomposable tensors in this section, i.e., the core tensor $\mathbf{G} \in \mathbb{R}^{r\times r\times r}$ is diagonal. Note that even though our theoretical analysis requires $n \gtrsim r^{7/2}d^{3/2}$, our simulation results seem to suggest that our approach can be successful for as few as $O(\sqrt{r}d^{3/2})$ observed entries. To this end, we shall consider sample size $n = \alpha\sqrt{r}d^{3/2}$ for some $\alpha > 0$.

More specifically, we consider $\mathbf{T} = d\sum_{k=1}^{r} \mathbf{u}_k \otimes \mathbf{v}_k \otimes \mathbf{w}_k \in \mathbb{R}^{d\times d\times d}$ with $d = 50, 100$ and $r = 2, 3, 4, 5$. The orthonormal vectors $\{\mathbf{u}_k, k = 1, \ldots, r\}, \{\mathbf{v}_k, k = 1, \ldots, r\}, \{\mathbf{w}_k, k = 1, \ldots, r\}$ are obtained from the eigenspace of randomly generated standard Gaussian matrices which guarantee the incoherence conditions based on the delocalization property of eigenvectors of Gaussian random matrices. For each choice of r and $\alpha = \frac{n}{\sqrt{r}d^{3/2}}$, the gradient descent algorithm from Sect. 5 with initialization described in Sect. 4 is applied in 50 simulation runs. We claim that the underlying tensor is successfully recovered if the returned tensor $\widehat{\mathbf{T}}$ satisfies that $\|\widehat{\mathbf{T}}-\mathbf{T}\|_{\mathrm{F}}/\|\mathbf{T}\|_{\mathrm{F}} \leq 10^{-7}$. The reconstruction rates are given in Figs. 1 and 2. It suggests that approximately when $n \geq 7\sqrt{r}d^{3/2}$, the algorithm reconstructed the true tensor with near certainty.

As mentioned before, in addition to the gradient descent algorithm described in Sect. 5, several other algorithms can also be applied to optimize $F(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ locally. A notable example is the geometrical conjugate gradient descent algorithm on Riemannian manifolds proposed by Kressner et al. [18]. Although we have focused on the analysis of the gradient descent algorithm, we believe similar results could also be established for these other algorithms as well. In essence, the success of these methods is determined by the quality of the initialization, which the method from Sect. 4 could be readily applied. We leave the more rigorous theoretical analysis for future work; we conducted a set of numerical experiments to illustrate the similarity between these optimization algorithms while highlighting the crucial role of initialization.

We considered a similar setup as before with $d = 50, r = 5$ and $d = 100, r = 3$. We shall refer to our method as GoG and the geometrical conjugate gradient descent algorithm as GeoCG, for brevity. Note that the GeoCG algorithm was proposed without considering the theoretical requirement on the sample size and the algorithm is initiated with a random guess. We first tested both algorithms with a reliable initialization as proposed in Sect. 4. That is, we started with $\widehat{\mathbf{U}}, \widehat{\mathbf{V}}, \widehat{\mathbf{W}}$ obtained from the spectral algorithm and let $\widehat{\mathbf{C}} \in \mathbb{R}^{r\times r\times r}$ be the minimizer of (2). Then, the GeoCG(Spectral) algorithm is initialized from the starting point $\widehat{\mathbf{A}}^{(0)} = (\widehat{\mathbf{U}}, \widehat{\mathbf{V}}, \widehat{\mathbf{W}}) \cdot \widehat{\mathbf{C}}$. For each $\alpha =$

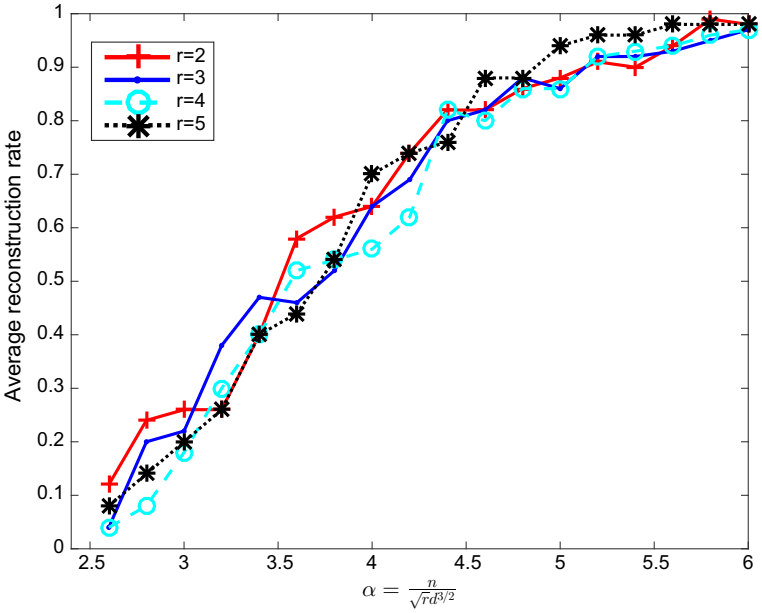

Fig. 1 Average reconstruction rate of the proposed approach when $d = 50$ and $r = 2, 3, 4, 5$. For each r and α, the algorithm is repeated for 50 times

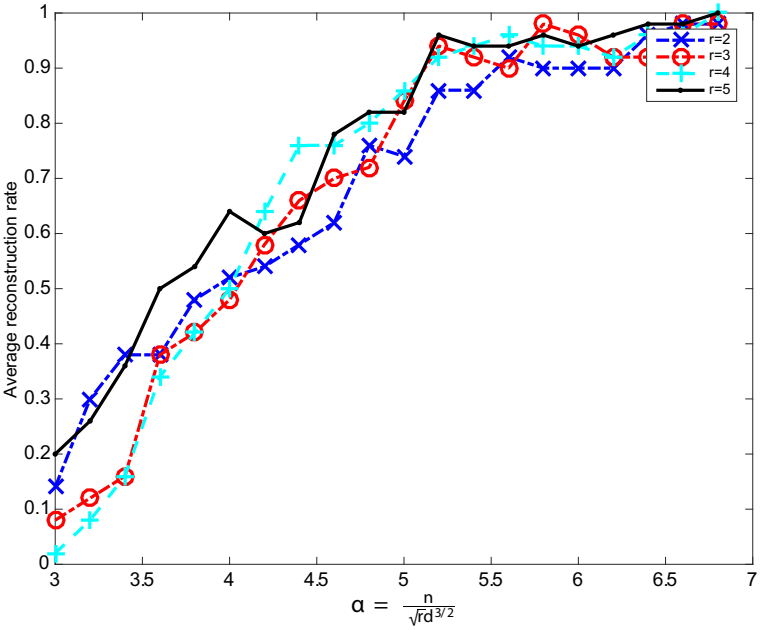

Fig. 2 Average reconstruction rate of the proposed approach when $d = 100$ and $r = 2, 3, 4, 5$. For each r and α, the algorithm is repeated for 50 times

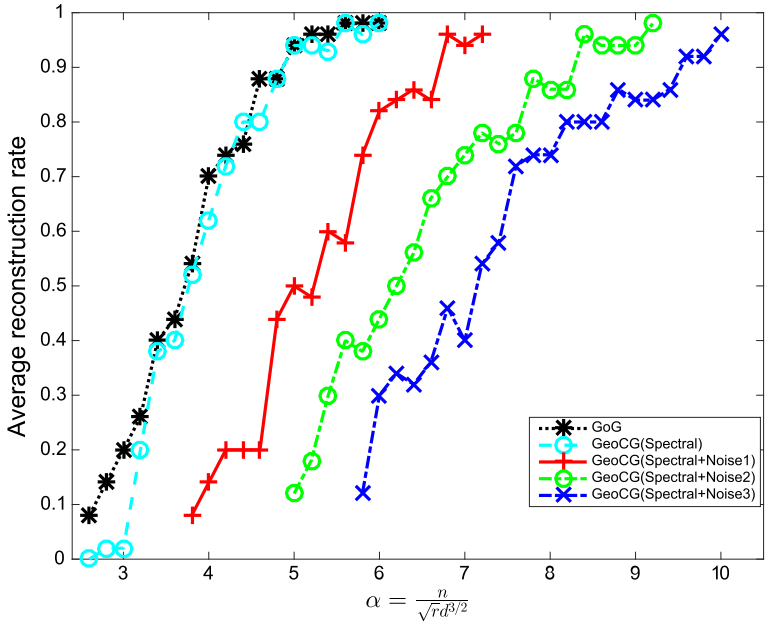

Fig. 3 Comparison between GoG and GeoCG algorithms when $d = 50$ and $r = 5$. The successful rates of GeoCG algorithm depend on the initialization. Here GeoCG(Spectral) means that the GeoCG algorithm is initialized with the spectral methods as GoG algorithm. The black and cyan curves show that GoG and GeoCG algorithms perform similarly when both are initialized with spectral methods. Here GeoCG(Spectral + Noise$X$) means that GeoCG algorithm is initialized with spectral methods plus random perturbation. If X is larger, the perturbation is larger and the initialization is further away from the truth, in which case the reconstruction rate decreases

$\frac{n}{\sqrt{r}d^{3/2}}$, the GeoCG algorithm is repeated for 50 times. The reconstruction rates are as shown in the Cyan curves in Figs. 3 and 4. It is clear that both algorithms perform well and are comparable.

To illustrate that successful recovery hinges upon the appropriate initialization, we now consider applying GeoCG algorithm with a randomly perturbed spectral initialization. More specifically, the GeoCG algorithm is initialized with $\widehat{\mathbf{A}}^{(0)} + \sigma\mathbf{Z}$ where $\mathbf{Z} \in \mathbb{R}^{d\times d\times d}$ is a random tensor with i.i.d. standard normal entries and $\sigma > 0$ represents the noise level. Figures 3 and 4 show that the reconstruction rate decreases when σ gets larger.

These observations confirm the insights from our theoretical development: That the objective function F is well behaved locally and therefore with appropriate initialization can lead to successful recovery of low-rank tensors.

7 Discussion

In this paper, we proved that with $n \geq C\mu_0^3 r_1 r_2 r_3 (rd_1d_2d_3)^{1/2}\log^{7/2}(d)$ uniformly sampled entries, a tensor $\mathbf{T}$ of multilinear ranks (r_1, r_2, r_3) can be recovered with high

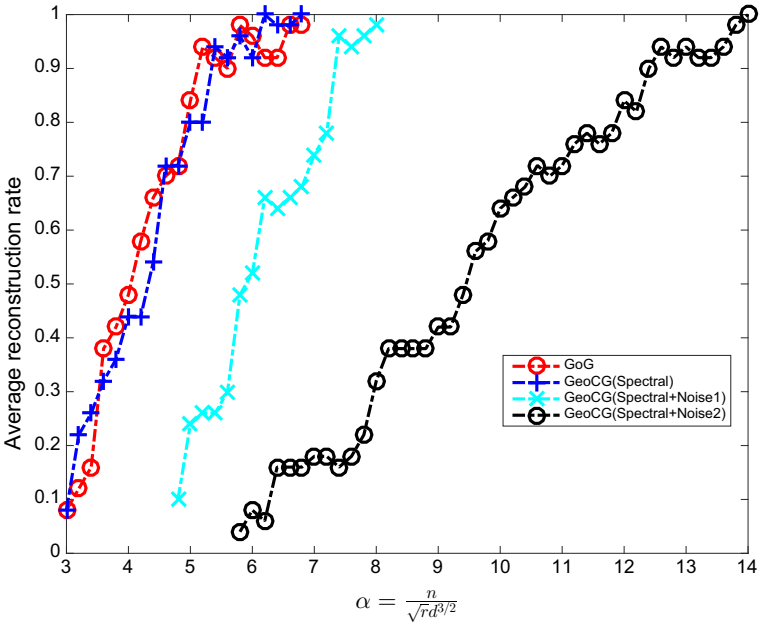

Fig. 4 Comparison between GoG and GeoCG algorithm when $d = 100$ and $r = 3$. The successful rates of GeoCG algorithm depend on the initialization

probability with a polynomial time algorithm. In doing so, we argue that the underlying optimization problem is well behaved in a neighborhood around the truth and therefore, the sample size requirement is largely driven by our ability to initialize the algorithm appropriately. To this end, a new spectral method based on estimating the second moment of tensor unfoldings is proposed. In the low-rank case, e.g., $r = O(1)$, this sample size requirement is essentially of the same order as $d^{3/2}$, up to a polynomial of $\log d$ term. This matches the sample size requirement for nuclear norm minimization which is NP-hard to compute in general. An argument put forth by Barak and Moitra [3] suggests that such a dependence on the dimension may be optimal for polynomial time algorithms unless a more efficient algorithm exists for the 3-SAT problem.

Even though our framework is established for third-order tensors, it can be naturally extended to higher-order tensors. Indeed, to complete a kth-order tensor $\mathbf{T} \in \mathbb{R}^{d \times d \times \ldots \times d}$ with multilinear ranks $(r, r, \ldots, r)$, we can apply similar type of algorithms for optimizing over product of Grassmannians. In order to ensure exact recovery, we can start with similar initialization where we unfold the tensor to $d \times d^{k-1}$ matrices. Following an identical argument, it can be derived in the same fashion that the sample size requirement for exact recovery now becomes

$$n \geq C d^{k/2} \text{polylog}(r, \log d)$$

for some constant $C > 0$. Unlike the third-order case, the dependence on the dimensionality ($d^{k/2}$) is worse than the nuclear norm minimization ($d^{3/2}$) for $k > 3$. See

Yuan and Zhang [34]. In general, it remains unclear whether the requirement of $d^{k/2}$ is the best attainable for polynomial time algorithms for $k > 3$.

In the current work, we are concerned with the reconstruction of a tensor when its entries are observed exactly. In many applications, however, these observations are often made with error. The presence of measurement errors changes the nature of the problem as it in general rules out the possibility of exact recovery. Instead, the focus is on how well we can estimate or approximate the tensor based on the noisy observations. The two problems, albeit closely connected, pose fundamentally different challenges. In particular, it is essential in exact recovery that we match all observed entries, but doing so in the presence of measurement error typically leads to suboptimal estimates. In general, exact recovery is more stringent than seeking an approximation. While it is essential to exact recovery that the tensor is of low rank, oftentimes a good approximation can still be obtained even if the underlying tensor is only approximately low rank.

8 Proofs

Throughout the proofs, we shall use C and similarly C_1, C_2, etc., to denote generic numerical positive constants that may take different values at each appearance.

8.1 Proof of Theorem 1

In view of Theorem 4, the proof of Theorem 1 is immediate if $(\mathbf{X}^{(0)}, \mathbf{Y}^{(0)}, \mathbf{Z}^{(0)}) \in \mathcal{N}\big(C(\alpha\kappa_0^2 \log d)^{-1}, 3\mu_0\big)$. Clearly, under the conditions on n given in Theorem 1, the top singular vectors $(\widehat{\mathbf{U}}, \widehat{\mathbf{V}}, \widehat{\mathbf{W}})$ satisfy that

$$d_{\mathrm{p}}\Big((\widehat{\mathbf{U}}, \widehat{\mathbf{V}}, \widehat{\mathbf{W}}), (\mathbf{U}, \mathbf{V}, \mathbf{W})\Big) \le C(\alpha\kappa_0^2 \log d)^{-1}$$

with probability at least $1 - 3d^{-\alpha}$. The fact that $(\mathbf{X}^{(0)}, \mathbf{Y}^{(0)}, \mathbf{Z}^{(0)}) \in \mathcal{N}\big(C(\alpha\kappa_0^2 \log d)^{-1}, 3\mu_0\big)$ then follows immediately from Remark 6.2 of Keshavan et al. [17].

8.2 Proof of Theorem 2

Using a standard decoupling technique for U-statistics, we get

$$\mathbb{P}(\|\widehat{\mathbf{N}} - \mathbf{N}\| > t) \le 15\mathbb{P}(15\|\tilde{\mathbf{N}} - \mathbf{N}\| > t)$$

for any $t > 0$, where

$$\tilde{\mathbf{N}} := \frac{1}{2n(n-1)} \sum_{i \neq j} (\mathbf{X}_i \mathbf{Y}_j^\top + \mathbf{Y}_j \mathbf{X}_i^\top),$$

and $\{\mathbf{Y}_i : 1 \le i \le n\}$ is an independent copy of $\{\mathbf{X}_i : 1 \le i \le n\}$. We shall then focus, in what follows, on bounding $\mathbb{P}(\|\tilde{\mathbf{N}} - \mathbf{N}\| > t)$. See, e.g., Theorem 1 of de la Peña and Montgomery-Smith [9] or Theorem 3.4.1 of de la Pena and Giné [8].

Observe that

$$\tilde{\mathbf{N}} = \frac{1}{2n(n-1)}(\mathbf{S}_{1n}\mathbf{S}_{2n}^\top + \mathbf{S}_{2n}\mathbf{S}_{1n}^\top) - \frac{1}{2n(n-1)}\sum_{i=1}^{n}(\mathbf{X}_i\mathbf{Y}_i^\top + \mathbf{Y}_i\mathbf{X}_i^\top),$$

where

$$\mathbf{S}_{1k} = \sum_{i=1}^{k}\mathbf{X}_i, \quad \text{and} \quad \mathbf{S}_{2k} = \sum_{i=1}^{k}\mathbf{Y}_i.$$

An application of Chernoff bound yields that, with probability at least $1 - m^{-\alpha}$,

$$\|\mathbf{S}_{1n}\|_{\ell_\infty} \le (3\alpha + 7)m_1 m_2 \|\mathbf{M}\|_{\max}\left(\frac{n}{m_2} + \log m\right)$$

for any $\alpha > 0$, where

$$\|\mathbf{S}_{1n}\|_{\ell_\infty} := \max_{1\le j\le m_2}\sum_{1\le i\le m_1}\left|(\mathbf{S}_{1n})_{ij}\right|.$$

See, e.g., proof of Theorem 2 in Yuan and Zhang [34]. Denote this event by $\mathcal{E}_1$. On the other hand, as shown by Recht [26] (Theorem 4), with probability at least $1 - m^{-\alpha}$,

$$\left\|\frac{1}{n}\mathbf{S}_{1n} - \mathbf{M}\right\| \le \sqrt{\frac{8(\alpha+1)m_1 m_2 m\log m}{3n}}\|\mathbf{M}\|_{\max}.$$

Denote this event by $\mathcal{E}_2$. Write $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$. It is not hard to see that for any $t \ge 0$,

$$\mathbb{P}\left\{\left\|\tilde{\mathbf{N}} - \mathbf{N}\right\| > t\right\} \le \mathbb{P}\left\{\left\|\tilde{\mathbf{N}} - \mathbf{N}\right\| > t \bigcap \mathcal{E}\right\} + 3m^{-\alpha}$$

We shall now proceed to bound the first probability on the right-hand side.

Write

$$\begin{aligned}\tilde{\mathbf{N}} - \mathbf{N} = {} & \frac{1}{2n(n-1)}\left[(\mathbf{S}_{1n} - n\mathbf{M})(\mathbf{S}_{2n} - n\mathbf{M})^\top + (\mathbf{S}_{2n} - n\mathbf{M})(\mathbf{S}_{1n} - n\mathbf{M})^\top\right] \\ & + \frac{1}{2(n-1)}\left[\mathbf{M}(\mathbf{S}_{2n} - n\mathbf{M})^\top + (\mathbf{S}_{2n} - n\mathbf{M})\mathbf{M}^\top\right] \\ & + \frac{1}{2(n-1)}\left[\mathbf{M}(\mathbf{S}_{1n} - n\mathbf{M})^\top + (\mathbf{S}_{1n} - n\mathbf{M})\mathbf{M}^\top\right] \\ & - \frac{1}{2n(n-1)}\sum_{i=1}^{n}(\mathbf{X}_i\mathbf{Y}_i^\top + \mathbf{Y}_i\mathbf{X}_i^\top - 2\mathbf{M}\mathbf{M}^\top)\end{aligned}$$

$$=: \mathbf{A}_1 + \mathbf{A}_2 + \mathbf{A}_3 + \mathbf{A}_4.$$

We bound each of the four terms on the rightmost side separately. For brevity, write

$$\Delta_{1k} = \mathbf{S}_{1k} - k\mathbf{M}, \quad \text{and} \quad \Delta_{2k} = \mathbf{S}_{2k} - k\mathbf{M}.$$

We begin with

$$\mathbf{A}_1 = \frac{1}{2n(n-1)} \left(\Delta_{1n} \Delta_{2n}^\top + \Delta_{2n} \Delta_{1n}^\top \right).$$

By Markov inequality, for any $\lambda > 0$,

$$\mathbb{P}\left\{ \|\mathbf{A}_1\| > t \bigcap \mathcal{E} \right\} \le \mathbb{P}\left\{ \operatorname{tr} \exp(\lambda \mathbf{A}_1) > \exp(\lambda t) \bigcap \mathcal{E} \right\} \le e^{-\lambda t} \mathbb{E}\left(\operatorname{tr} \exp[\lambda \mathbf{A}_1] \mathbf{1}_{\mathcal{E}} \right).$$

Repeated use of Golden–Thompson inequality yields

$$\begin{aligned}
\mathbb{E}\left(\operatorname{tr} \exp[\lambda \mathbf{A}_1] \mathbf{1}_{\mathcal{E}} \right) &= \mathbb{E}\left(\mathbb{E}\left\{ \operatorname{tr} \exp[\lambda \mathbf{A}_1] \mathbf{1}_{\mathcal{E}} \middle| \mathbf{S}_{1n} \right\} \right) \\
&\le \mathbb{E}\Bigg(\mathbb{E}\left\{ \operatorname{tr} \exp\left[\frac{\lambda}{2n(n-1)} (\Delta_{1n} \Delta_{2,n-1}^\top + \Delta_{2,n-1} \Delta_{1n}^\top) \right] \mathbf{1}_{\mathcal{E}} \middle| \mathbf{S}_{1n} \right\} \\
&\quad \times \left\| \mathbb{E}\left\{ \exp\left[\frac{\lambda}{2n(n-1)} (\Delta_{1n} (\mathbf{Y}_n - \mathbf{M})^\top + (\mathbf{Y}_n - \mathbf{M}) \Delta_{1n}^\top) \right] \mathbf{1}_{\mathcal{E}} \middle| \mathbf{S}_{1n} \right\} \right\| \Bigg) \\
&\le \cdots\cdots \\
&\le m \mathbb{E}\left(\left\| \mathbb{E}\left\{ \exp\left[\frac{\lambda}{2n(n-1)} (\Delta_{1n} (\mathbf{Y}_n - \mathbf{M})^\top + (\mathbf{Y}_n - \mathbf{M}) \Delta_{1n}^\top) \right] \mathbf{1}_{\mathcal{E}} \middle| \mathbf{S}_{1n} \right\} \right\|^n \right)
\end{aligned}$$

By triangular inequality,

$$\begin{aligned}
&\left\| \frac{\lambda}{2n(n-1)} \left[\Delta_{1n} (\mathbf{Y}_n - \mathbf{M})^\top + (\mathbf{Y}_n - \mathbf{M}) \Delta_{1n}^\top \right] \right\| \\
&\quad \le \frac{\lambda}{n(n-1)} \left(\|\Delta_{1n} \mathbf{Y}_n^\top\| + \|\Delta_{1n} \mathbf{M}^\top\| \right).
\end{aligned}$$

Under the event $\mathcal{E}_1$ with the upper bound of $\|\mathbf{S}_{1n}\|_{\ell_\infty}$,

$$\begin{aligned}
\|\Delta_{1n} \mathbf{Y}_n^\top\| &\le \|\mathbf{S}_{1n} \mathbf{Y}_n^\top\| + n \|\mathbf{M} \mathbf{Y}_n^\top\| \\
&\le (3\alpha + 7) m_1^2 m_2^2 \|\mathbf{M}\|_{\max}^2 \left(\frac{n}{m_2} + \log m \right) + n m_1 m_2 \|\mathbf{M}\|_{\max} \|\mathbf{M}\|.
\end{aligned}$$

On the other hand, under the event $\mathcal{E}_2$,

$$\|\Delta_{1n} \mathbf{M}^\top\| \le \|\Delta_{1n}\| \|\mathbf{M}\| \le \sqrt{\frac{8}{3} n(\alpha+1) m_1 m_2 m \log m} \|\mathbf{M}\|_{\max} \|\mathbf{M}\|$$

Recall that

$$n \cdot \min\{m_1, m_2\} \geq \frac{8}{3}(\alpha+1)\log m.$$

Then

$$\begin{aligned}
&\left\| \frac{\lambda}{2n(n-1)} \left[\Delta_{1n}(\mathbf{Y}_n - \mathbf{M})^\top + (\mathbf{Y}_n - \mathbf{M})\Delta_{1n}^\top \right] \right\| \\
&\quad \leq \frac{\lambda}{n(n-1)} \left((3\alpha+7) m_1^2 m_2^2 \|\mathbf{M}\|_{\max}^2 \left(\frac{n}{m_2} + \log m \right) + 2nm_1 m_2 \|\mathbf{M}\|_{\max} \|\mathbf{M}\| \right).
\end{aligned}$$

Therefore, for any

$$\lambda \leq n(n-1) \left((3\alpha+7) m_1^2 m_2^2 \|\mathbf{M}\|_{\max}^2 \left(\frac{n}{m_2} + \log m \right) + 2nm_1 m_2 \|\mathbf{M}\|_{\max} \|\mathbf{M}\| \right)^{-1},$$

we get

$$\begin{aligned}
&\mathbb{E}\left\{ \exp\left[\frac{\lambda}{2n(n-1)} \left[\Delta_{1n}(\mathbf{Y}_n - \mathbf{M})^\top + (\mathbf{Y}_n - \mathbf{M})\Delta_{1n}^\top \right] \right] \mathbf{1}_{\mathcal{E}} \middle| \mathbf{S}_{1n} \right\} \\
&\quad \preceq \mathbf{I}_{m_1} + \mathbb{E}\left\{ \left[\frac{\lambda}{2n(n-1)} \left[\Delta_{1n}(\mathbf{Y}_n - \mathbf{M})^\top + (\mathbf{Y}_n - \mathbf{M})\Delta_{1n}^\top \right] \right]^2 \mathbf{1}_{\mathcal{E}} \middle| \mathbf{S}_{1n} \right\} \\
&\quad = \mathbf{I}_{m_1} + \mathbb{E}\left\{ \left[\frac{\lambda}{2n(n-1)} \left(\Delta_{1n}\mathbf{Y}_n^\top + \mathbf{Y}_n\Delta_{1n}^\top \right) \right]^2 \mathbf{1}_{\mathcal{E}} \middle| \mathbf{S}_{1n} \right\} \\
&\qquad - \left[\frac{\lambda^2}{4n^2(n-1)^2} \left(\Delta_{1n}\mathbf{M}^\top + \mathbf{M}\Delta_{1n}^\top \right)^2 \mathbf{1}_{\mathcal{E}} \middle| \mathbf{S}_{1n} \right] \\
&\quad \preceq \mathbf{I}_{m_1} + \mathbb{E}\left\{ \left[\frac{\lambda}{2n(n-1)} \left(\Delta_{1n}\mathbf{Y}_n^\top + \mathbf{Y}_n\Delta_{1n}^\top \right) \right]^2 \mathbf{1}_{\mathcal{E}} \middle| \mathbf{S}_{1n} \right\} \\
&\quad \preceq \mathbf{I}_{m_1} + \frac{\lambda^2 m_1 m_2 \|\mathbf{M}\|_{\max}^2}{4n^2(n-1)^2} \left[(m_1+2)\Delta_{1n}\Delta_{1n}^\top + \mathrm{tr}(\Delta_{1n}\Delta_{1n}^\top)\mathbf{I}_{m_1} \right] \mathbf{1}_{\mathcal{E}}
\end{aligned}$$

where in the first inequality, we used the facts that

$$\exp(\mathbf{A}) \preceq \mathbf{I}_d + \mathbf{A} + \mathbf{A}^2$$

for any $\mathbf{A} \in \mathbb{R}^{d\times d}$ such that $\|\mathbf{A}\| \leq 1$, and

$$\mathbb{E}\left\{ \left[\frac{\lambda}{2n(n-1)} \left[\Delta_{1n}(\mathbf{Y}_n - \mathbf{M})^\top + (\mathbf{Y}_n - \mathbf{M})\Delta_{1n}^\top \right] \right] \mathbf{1}_{\mathcal{E}} \middle| \mathbf{S}_{1n} \right\} = 0.$$

Recall that

$$\mathrm{tr}(\Delta_{1n}\Delta_{1n}^\top) \leq m_1 \|\Delta_{1n}\Delta_{1n}^\top\|.$$

This implies that

$$\begin{aligned}&\left\|\mathbb{E}\left\{\exp\left[\frac{\lambda}{2n(n-1)}\left[\Delta_{1n}(\mathbf{Y}_n-\mathbf{M})^\top+(\mathbf{Y}_n-\mathbf{M})\Delta_{1n}^\top\right]\right]\mathbf{1}_{\mathcal{E}}\middle|\mathbf{S}_{1n}\right\}\right\|\\&\quad\le 1+\frac{\lambda^2\|\mathbf{M}\|_{\max}^2 m_1^2 m_2}{2n^2(n-1)^2}\|\Delta_{1n}\Delta_{1n}^\top\|\mathbf{1}_{\mathcal{E}}\\&\quad\le 1+\frac{8(\alpha+1)\lambda^2\|\mathbf{M}\|_{\max}^4 m_1^3 m_2^2 m\log m}{3n(n-1)^2},\end{aligned}$$

where the last inequality follows from the definition of $\mathcal{E}_2$. Thus,

$$\mathbb{E}\,\mathrm{tr}\exp\left[\lambda\mathbf{A}_1\mathbf{1}_{\mathcal{E}}\right]\le m\cdot\exp\left[\lambda^2\frac{16(\alpha+1)\|\mathbf{M}\|_{\max}^4 m_1^3 m_2^2 m\log m}{3(n-1)^2}\right].$$

Taking

$$\begin{aligned}\lambda=\min\Bigg\{&\frac{3(n-1)^2 t}{64(\alpha+1)\|\mathbf{M}\|_{\max}^4 m_1^3 m_2^2 m\log m},\\&\frac{n(n-1)}{(6\alpha+14)m_1^2 m_2^2\|\mathbf{M}\|_{\max}^2(n/m_2+\log m)},\frac{n(n-1)}{4nm_1m_2\|\mathbf{M}\|_{\max}\|\mathbf{M}\|}\Bigg\}\end{aligned}$$

yields

$$\begin{aligned}\mathbb{P}\left\{\|\mathbf{A}_1\|>t\bigcap\mathcal{E}\right\}\le\exp\Bigg(-\min\Bigg\{&\frac{3(n-1)^2t^2}{128(\alpha+1)\|\mathbf{M}\|_{\max}^4 m_1^3 m_2^2 m\log m},\\&\frac{n(n-1)t}{(12\alpha+28)m_1^2m_2^2\|\mathbf{M}\|_{\max}^2(n/m_2+\log m)},\frac{n(n-1)t}{8nm_1m_2\|\mathbf{M}\|_{\max}\|\mathbf{M}\|}\Bigg\}\Bigg)\end{aligned}$$

We now proceed to bound $\mathbf{A}_2$ and $\mathbf{A}_3$. Both terms can be treated in an identical fashion, and we shall consider only $\mathbf{A}_2$ here to fix ideas. As before, it can be derived that

$$\begin{aligned}&\mathbb{P}\left\{\|\mathbf{A}_2\|>t\bigcap\mathcal{E}\right\}\le m\cdot\exp(-\lambda t)\\&\quad\times\left\|\mathbb{E}\left\{\exp\left[\frac{\lambda}{2(n-1)}(\mathbf{M}(\mathbf{Y}_n-\mathbf{M})^\top+(\mathbf{Y}_n-\mathbf{M})\mathbf{M}^\top)\right]\mathbf{1}_{\mathcal{E}}\right\}\right\|^n.\end{aligned}$$

By taking

$$\lambda\le\frac{n-1}{\|\mathbf{M}\|^2+m_1m_2\|\mathbf{M}\|\|\mathbf{M}\|_{\max}},$$

we can ensure

$$\left\| \frac{\lambda}{2(n-1)} (\mathbf{M}(\mathbf{Y}_n - \mathbf{M})^\top + (\mathbf{Y}_n - \mathbf{M})\mathbf{M}^\top) \right\| \leq \frac{\lambda}{n-1} \left(\|\mathbf{M}\|^2 + m_1 m_2 \|\mathbf{M}\| \|\mathbf{M}\|_{\max} \right) \leq 1.$$

If this is the case, we can derive as before that

$$\begin{aligned} &\left\| \mathbb{E} \left\{ \exp \left[\frac{\lambda}{2(n-1)} (\mathbf{M}(\mathbf{Y}_n - \mathbf{M})^\top + (\mathbf{Y}_n - \mathbf{M})\mathbf{M}^\top) \right] \mathbf{1}_{\mathcal{E}} \right\} \right\| \\ &\quad \leq 1 + \left\| \mathbb{E} \left\{ \left[\frac{\lambda}{2(n-1)} (\mathbf{M}(\mathbf{Y}_n - \mathbf{M})^\top + (\mathbf{Y}_n - \mathbf{M})\mathbf{M}^\top) \right]^2 \mathbf{1}_{\mathcal{E}} \right\} \right\| \\ &\quad \leq 1 + \left\| \mathbb{E} \left\{ \left[\frac{\lambda}{2(n-1)} (\mathbf{M}\mathbf{Y}_n^\top + \mathbf{Y}_n \mathbf{M}^\top) \right]^2 \mathbf{1}_{\mathcal{E}} \right\} \right\| \\ &\quad \leq 1 + \frac{\lambda^2 m_1^2 m_2 \|\mathbf{M}\|_{\max}^2 \|\mathbf{M}\|^2}{2(n-1)^2}. \end{aligned}$$

In particular, taking

$$\lambda = \min \left\{ \frac{n-1}{2\|\mathbf{M}\|^2}, \frac{n-1}{2m_1 m_2 \|\mathbf{M}\| \|\mathbf{M}\|_{\max}}, \frac{(n-1)^2 t}{n m_1^2 m_2 \|\mathbf{M}\|_{\max}^2 \|\mathbf{M}\|^2} \right\}$$

yields

$$\begin{aligned} &\mathbb{P} \left\{ \|\mathbf{A}_2\| > t \bigcap \mathcal{E} \right\} \\ &\quad \leq \exp \left(- \min \left\{ \frac{(n-1)t}{4\|\mathbf{M}\|^2}, \frac{(n-1)t}{2m_1 m_2 \|\mathbf{M}\| \|\mathbf{M}\|_{\max}}, \frac{(n-1)^2 t^2}{2n m_1^2 m_2 \|\mathbf{M}\|_{\max}^2 \|\mathbf{M}\|^2} \right\} \right). \end{aligned}$$

Finally, we treat $\mathbf{A}_4$. Observe that

$$\begin{aligned} \|\mathbf{X}_i \mathbf{Y}_i^\top + \mathbf{Y}_i \mathbf{X}_i^\top - 2\mathbf{M}\mathbf{M}^\top\| &\leq 2\|\mathbf{X}_i \mathbf{Y}_i^\top\| + 2\|\mathbf{M}\|^2 \\ &\leq 2m_1^2 m_2^2 \|\mathbf{M}\|_{\max}^2 + 2\|\mathbf{M}\|^2 \\ &\leq 4m_1^2 m_2^2 \|\mathbf{M}\|_{\max}^2, \end{aligned}$$

where the last inequality follows from the fact that $\|\mathbf{M}\| \leq \|\mathbf{M}\|_{\mathrm{F}} \leq \sqrt{m_1 m_2} \|\mathbf{M}\|_{\max}$. On the other hand,

$$\mathbb{E} \left(\mathbf{X}_i \mathbf{Y}_i^\top + \mathbf{Y}_i \mathbf{X}_i^\top - 2\mathbf{M}\mathbf{M}^\top \right)^2 \preceq \mathbb{E} \left(\mathbf{X}_i \mathbf{Y}_i^\top + \mathbf{Y}_i \mathbf{X}_i^\top \right)^2 \preceq 2(m_1 + 1) m_1^2 m_2^3 \|\mathbf{M}\|_{\max}^4 \mathbf{I}.$$

An application of matrix Bernstein inequality (e.g., Theorem 6.1 of [31]) yields

$$\begin{aligned}\mathbb{P}\{\|\mathbf{A}_4\| > t \cap \mathcal{E}\} &\le \mathbb{P}\{\|\mathbf{A}_4\| > t\} \\ &\le m_1 \exp\left(-\frac{n^2(n-1)^2t^2/2}{2n(m_1+1)m_1^2m_2^3\|\mathbf{M}\|_{\max}^4 + 4m_1^2m_2^2\|\mathbf{M}\|_{\max}^2 t/3}\right).\end{aligned}$$

Putting the probability bounds for $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4$ together, we have

$$\mathbb{P}\{\|\tilde{\mathbf{N}} - \mathbf{N}\| > t/15\} \le \sum_{k=1}^{4} \mathbb{P}\{\|\mathbf{A}_k\| > t/60 \cap \mathcal{E}\} + \mathbb{P}\{\mathcal{E}^c\} \le 7m^{-\alpha}$$

by taking

$$\begin{aligned}t = C \cdot \alpha^2 \cdot \frac{m_1^{3/2}m_2^{3/2}\log m}{n} \\ \times \left[\left(1 + \frac{m_1}{m_2}\right)^{1/2} + \frac{m_1^{1/2}m_2^{1/2}}{n} + \left(\frac{n}{m_2 \log m}\right)^{1/2}\right] \cdot \|\mathbf{M}\|_{\max}^2,\end{aligned}$$

for some $C \ge 1680$. This immediately implies that

$$\mathbb{P}\big\{\|\widehat{\mathbf{N}} - \mathbf{N}\| \ge t\big\} \le 105m^{-\alpha}.$$

The proof is then concluded by replacing α with $\alpha + \log_m 105$ and adjusting the constant C accordingly.

8.3 Proof of Theorem 3

Let $\mathbf{P}_\mathbf{U}$, $\mathbf{P}_\mathbf{V}$ and $\mathbf{P}_\mathbf{W}$ be the projection matrices onto the column spaces of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$, respectively. Denote by $\mathbf{Q}_\mathbf{T} : \mathbb{R}^{d_1 \times d_2 \times d_3} \to \mathbb{R}^{d_1 \times d_2 \times d_3}$ a linear operator such that for any $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$,

$$\begin{aligned}\mathbf{Q}_\mathbf{T}\mathbf{A} := (\mathbf{P}_\mathbf{U}, \mathbf{P}_\mathbf{V}, \mathbf{P}_\mathbf{W}) \cdot \mathbf{A} + (\mathbf{P}_\mathbf{U}^\perp, \mathbf{P}_\mathbf{V}, \mathbf{P}_\mathbf{W}) \cdot \mathbf{A} + (\mathbf{P}_\mathbf{U}, \mathbf{P}_\mathbf{V}^\perp, \mathbf{P}_\mathbf{W}) \cdot \mathbf{A} \\ + (\mathbf{P}_\mathbf{U}, \mathbf{P}_\mathbf{V}, \mathbf{P}_\mathbf{W}^\perp) \cdot \mathbf{A},\end{aligned}$$

where $\mathbf{P}_\mathbf{U}^\perp = I - \mathbf{P}_\mathbf{U}$, and $\mathbf{P}_\mathbf{V}^\perp$ and $\mathbf{P}_\mathbf{W}^\perp$ are defined similarly. We shall also write $\mathbf{Q}_\mathbf{T}^\perp = \mathcal{I} - \mathbf{Q}_\mathbf{T}$ where $\mathcal{I}$ is the identity map.

Basic facts about Grassmannians Before proceeding, we shall first review some basic facts about the Grassmannians necessary for our proof. For further details, interested readers are referred to Edelman et al. [11] (Section 2). To fix ideas, we shall focus on $\mathbf{U} \in \mathcal{G}(d_1, r_1)$. The tangent space of $\mathcal{G}(d_1, r_1)$ at $\mathbf{U}$, denoted by $\mathcal{T}_\mathbf{U} \subset \mathbb{R}^{d_1 \times r_1}$, can be identified with the property $\mathbf{U}^\top \mathbf{D}_\mathbf{U} = \mathbf{0}$. The geodesic path from $\mathbf{U}$ to another

point $\mathbf{X} \in \mathcal{G}(d_1, r_1)$ with respect to the canonical Riemann metric can be explicitly expressed as:

$$\mathbf{X}(t) = \mathbf{U}\mathbf{R}_{\mathbf{U}} \cos(\mathbf{\Theta}_{\mathbf{U}} t)\mathbf{R}_{\mathbf{U}}^{\top} + \mathbf{L}_{\mathbf{U}} \sin(\mathbf{\Theta}_{\mathbf{U}} t)\mathbf{R}_{\mathbf{U}}^{\top}, \quad 0 \le t \le 1$$

for some $\mathbf{D}_{\mathbf{U}} \in \mathcal{T}_{\mathbf{U}}$ and $\mathbf{D}_{\mathbf{U}} = \mathbf{L}_{\mathbf{U}}\mathbf{\Theta}_{\mathbf{U}}\mathbf{R}_{\mathbf{U}}^{\top}$ is its thin singular value decomposition. We can identify $\mathbf{X}(0) = \mathbf{U}$ and $\mathbf{X}(1) = \mathbf{X}$. The diagonal element of $\mathbf{\Theta}_{\mathbf{U}}$ lies in $[-\pi/2, \pi/2]$ and can be viewed as the principle angle between $\mathbf{U}$ and $\mathbf{X}$.

It is easy to check

$$d_{\mathrm{p}}(\mathbf{U}, \mathbf{X}) = \|\sin \mathbf{\Theta}_{\mathbf{U}}\|_{\mathrm{F}} \quad \text{and} \quad \|\mathbf{\Delta}_{\mathbf{X}}\|_{\mathrm{F}} = \|\mathbf{U} - \mathbf{X}\|_{\mathrm{F}} = 2\|\sin(\mathbf{\Theta}_{\mathbf{U}}/2)\|_{\mathrm{F}}.$$

Note that for any $\theta \in [0, \pi/2]$,

$$\frac{\theta}{2} \le \sqrt{2}\sin(\theta/2) \le \sin\theta \le 2\sin(\theta/2) \le \theta.$$

This implies that

$$d_{\mathrm{p}}(\mathbf{U}, \mathbf{X}) \le \|\Delta_{\mathbf{X}}\|_{\mathrm{F}} \le \sqrt{2} d_{\mathrm{p}}(\mathbf{U}, \mathbf{X}).$$

Moreover,

$$\|\mathbf{U}^{\top}\mathbf{\Delta}_{\mathbf{X}}\|_{\mathrm{F}} = \|\cos(\mathbf{\Theta}_{\mathbf{U}}) - \mathbf{I}\|_{\mathrm{F}} = 4\|\sin^2(\mathbf{\Theta}_{\mathbf{U}}/2)\|_{\mathrm{F}} \le 2\|\sin \mathbf{\Theta}_{\mathbf{U}}\|_{\mathrm{F}}^2 = 2d_{\mathrm{p}}^2(\mathbf{U}, \mathbf{X}).$$

With slight abuse of notation, write $\mathbf{D}_{\mathbf{X}} = \frac{d\mathbf{X}(t)}{dt}\big|_{t=1} \in \mathcal{T}_{\mathbf{X}}$. $\mathbf{D}_{\mathbf{X}}$ can be more explicitly expressed as

$$\mathbf{D}_{\mathbf{X}} = -\mathbf{U}\mathbf{R}_{\mathbf{U}}\mathbf{\Theta}_{\mathbf{U}} \sin \mathbf{\Theta}_{\mathbf{U}}\mathbf{R}_{\mathbf{U}}^{\top} + \mathbf{L}_{\mathbf{U}}\mathbf{\Theta}_{\mathbf{U}} \cos \mathbf{\Theta}_{\mathbf{U}}\mathbf{R}_{\mathbf{U}}^{\top}.$$

It is clear that

$$\|\mathbf{D}_{\mathbf{X}}\|_{\mathrm{F}}^2 = \|\mathbf{\Theta}_{\mathbf{U}} \sin \mathbf{\Theta}_{\mathbf{U}}\|_{\mathrm{F}}^2 + \|\mathbf{\Theta}_{\mathbf{U}} \cos \mathbf{\Theta}_{\mathbf{U}}\|_{\mathrm{F}}^2 = \|\mathbf{\Theta}_{\mathbf{U}}\|_{\mathrm{F}}^2,$$

so that

$$d_{\mathrm{p}}(\mathbf{U}, \mathbf{X}) \le \|\mathbf{D}_{\mathbf{X}}\|_{\mathrm{F}} \le 2d_{\mathrm{p}}(\mathbf{U}, \mathbf{X}).$$

A couple of other useful relations can also be derived:

$$\begin{aligned}\|\mathbf{D}_{\mathbf{X}} - \mathbf{\Delta}_{\mathbf{X}}\|_{\mathrm{F}}^2 &= \|\mathbf{\Theta}_{\mathbf{U}}\|_{\mathrm{F}}^2 + 4\|\sin(\mathbf{\Theta}_{\mathbf{U}}/2)\|_{\mathrm{F}}^2 - 2\langle \mathbf{\Theta}_{\mathbf{U}}, \sin \mathbf{\Theta}_{\mathbf{U}}\rangle \\ &\le \|\mathbf{\Theta}_{\mathbf{U}} - 2\sin(\mathbf{\Theta}_{\mathbf{U}}/2)\|_{\mathrm{F}}^2 \le d_{\mathrm{p}}^4(\mathbf{U}, \mathbf{X}),\end{aligned}$$

and

$$\|\mathbf{U}^{\top}\mathbf{D}_{\mathbf{X}}\|_{\mathrm{F}} = \|\mathbf{\Theta}_{\mathbf{U}} \sin \mathbf{\Theta}_{\mathbf{U}}\|_{\mathrm{F}} \le 2\|\sin \mathbf{\Theta}_{\mathbf{U}}\|_{\mathrm{F}}^2 = 2d_{\mathrm{p}}^2(\mathbf{U}, \mathbf{X}).$$

Lower bound of $F(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ in Eq. (6). Note that

$$F(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \frac{1}{2}\left\|\mathcal{P}_\Omega\left(\widehat{\mathbf{T}} - \mathbf{T}\right)\right\|_{\mathrm{F}}^2 \geq \frac{1}{4}\left\|\mathcal{P}_\Omega \mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}} - \mathbf{T})\right\|_{\mathrm{F}}^2 - \frac{1}{2}\left\|\mathcal{P}_\Omega \mathbf{Q}_{\mathbf{T}}^{\perp}(\widehat{\mathbf{T}})\right\|_{\mathrm{F}}^2, \quad (8)$$

where

$$\widehat{\mathbf{T}} = (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \cdot \mathbf{C}$$

and $\mathbf{C}$ is given by (3). To derive the lower bound in the first statement, we shall lower-bound $\|\mathcal{P}_\Omega \mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}} - \mathbf{T})\|_{\mathrm{F}}^2$ and upper-bound $\|\mathcal{P}_\Omega \mathbf{Q}_{\mathbf{T}}^{\perp}(\widehat{\mathbf{T}})\|_{\mathrm{F}}^2$.

By Lemma 5 of Yuan and Zhang [33], if $n \geq C_1 \alpha \mu_0^2 r^2 d \log d$, then

$$\mathbb{P}\left\{\left\|\mathbf{Q}_{\mathbf{T}}\left(\mathcal{I} - \frac{d_1 d_2 d_3}{n}\mathcal{P}_\Omega\right)\mathbf{Q}_{\mathbf{T}}\right\| \geq \frac{1}{2}\right\} \leq d^{-\alpha},$$

where the operator norm is induced by the Frobenius norm, or the vectorized ℓ_2 norm. Denote this event by $\mathcal{E}_1$. We shall now proceed under $\mathcal{E}_1$. On event $\mathcal{E}_1$,

$$\|\mathcal{P}_\Omega \mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}} - \mathbf{T})\|_{\mathrm{F}}^2 \geq \left\langle \mathcal{P}_\Omega \mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}} - \mathbf{T}), \mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}} - \mathbf{T})\right\rangle \geq \frac{n}{2d_1 d_2 d_3}\|\mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}} - \mathbf{T})\|_{\mathrm{F}}^2.$$

Recall that

$$\begin{aligned}\mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}} - \mathbf{T}) = {} & (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot (\mathbf{G} - \mathbf{C}) + (\boldsymbol{\Delta}_{\mathbf{X}}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C} + (\mathbf{U}, \boldsymbol{\Delta}_{\mathbf{Y}}, \mathbf{W}) \cdot \mathbf{C} \\ & + (\mathbf{U}, \mathbf{V}, \boldsymbol{\Delta}_{\mathbf{Z}}) \cdot \mathbf{C}, \end{aligned} \quad (9)$$

where

$$\boldsymbol{\Delta}_{\mathbf{X}} := \mathbf{X} - \mathbf{U}, \quad \boldsymbol{\Delta}_{\mathbf{Y}} := \mathbf{Y} - \mathbf{V}, \quad \text{and} \quad \boldsymbol{\Delta}_{\mathbf{Z}} := \mathbf{Z} - \mathbf{W}.$$

Therefore,

$$\begin{aligned}\|\mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}} - \mathbf{T})\|_{\mathrm{F}}^2 = {} & \|(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot (\mathbf{G} - \mathbf{C})\|_{\mathrm{F}}^2 + \|(\boldsymbol{\Delta}_{\mathbf{X}}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C}\|_{\mathrm{F}}^2 + \|(\mathbf{U}, \boldsymbol{\Delta}_{\mathbf{Y}}, \mathbf{W}) \cdot \mathbf{C}\|_{\mathrm{F}}^2 \\ & + \|(\mathbf{U}, \mathbf{V}, \boldsymbol{\Delta}_{\mathbf{Z}}) \cdot \mathbf{C}\|_{\mathrm{F}}^2 + 2\langle (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot (\mathbf{G} - \mathbf{C}), (\boldsymbol{\Delta}_{\mathbf{X}}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C}\rangle \\ & + 2\langle (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot (\mathbf{G} - \mathbf{C}), (\mathbf{U}, \boldsymbol{\Delta}_{\mathbf{Y}}, \mathbf{W}) \cdot \mathbf{C}\rangle \\ & + 2\langle (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot (\mathbf{G} - \mathbf{C}), (\mathbf{U}, \mathbf{V}, \boldsymbol{\Delta}_{\mathbf{Z}}) \cdot \mathbf{C}\rangle \\ & + 2\langle (\boldsymbol{\Delta}_{\mathbf{X}}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C}, (\mathbf{U}, \boldsymbol{\Delta}_{\mathbf{Y}}, \mathbf{W}) \cdot \mathbf{C}\rangle \\ & + 2\langle (\boldsymbol{\Delta}_{\mathbf{X}}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C}, (\mathbf{U}, \mathbf{V}, \boldsymbol{\Delta}_{\mathbf{Z}}) \cdot \mathbf{C}\rangle \\ & + 2\langle (\mathbf{U}, \boldsymbol{\Delta}_{\mathbf{Y}}, \mathbf{W}) \cdot \mathbf{C}, (\mathbf{U}, \mathbf{V}, \boldsymbol{\Delta}_{\mathbf{Z}}) \cdot \mathbf{C}\rangle. \end{aligned}$$

It is clear that

$$\|(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot (\mathbf{G} - \mathbf{C})\|_{\mathrm{F}}^2 = \|\mathbf{G} - \mathbf{C}\|_{\mathrm{F}}^2.$$

We now bound each of the remaining terms on the right-hand side separately.

Note that

$$\begin{aligned}\|(\boldsymbol{\Delta}_{\mathbf{X}}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C}\|_{\mathrm{F}}^2 &\geq \frac{1}{2}\|(\boldsymbol{\Delta}_{\mathbf{X}}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{G}\|_{\mathrm{F}}^2 - \|(\boldsymbol{\Delta}_{\mathbf{X}}, \mathbf{V}, \mathbf{W}) \cdot (\mathbf{C} - \mathbf{G})\|_{\mathrm{F}}^2 \\ &\geq \frac{1}{2}\sigma_{\min}^2(\mathcal{M}_1(\mathbf{G}))\|\boldsymbol{\Delta}_{\mathbf{X}}\|_{\mathrm{F}}^2 - \sigma_{\max}^2(\mathcal{M}_1(\mathbf{C} - \mathbf{G}))\|\boldsymbol{\Delta}_{\mathbf{X}}\|_{\mathrm{F}}^2 \\ &\geq \frac{1}{2}\sigma_{\min}^2(\mathcal{M}_1(\mathbf{G}))\|\boldsymbol{\Delta}_{\mathbf{X}}\|_{\mathrm{F}}^2 - \|\mathbf{C} - \mathbf{G}\|_{\mathrm{F}}^2\|\boldsymbol{\Delta}_{\mathbf{X}}\|_{\mathrm{F}}^2 \\ &= \frac{1}{2}\sigma_{\min}^2(\mathcal{M}_1(\mathbf{T}))\|\boldsymbol{\Delta}_{\mathbf{X}}\|_{\mathrm{F}}^2 - \|\mathbf{C} - \mathbf{G}\|_{\mathrm{F}}^2\|\boldsymbol{\Delta}_{\mathbf{X}}\|_{\mathrm{F}}^2\end{aligned}$$

Similarly,

$$\|(\mathbf{U}, \boldsymbol{\Delta}_{\mathbf{Y}}, \mathbf{W}) \cdot \mathbf{C}\|_{\mathrm{F}}^2 \geq \frac{1}{2}\sigma_{\min}^2(\mathcal{M}_2(\mathbf{T}))\|\boldsymbol{\Delta}_{\mathbf{Y}}\|_{\mathrm{F}}^2 - \|\mathbf{C} - \mathbf{G}\|_{\mathrm{F}}^2\|\boldsymbol{\Delta}_{\mathbf{Y}}\|_{\mathrm{F}}^2,$$

and

$$\|(\mathbf{U}, \mathbf{V}, \boldsymbol{\Delta}_{\mathbf{Z}}) \cdot \mathbf{C}\|_{\mathrm{F}}^2 \geq \frac{1}{2}\sigma_{\min}^2(\mathcal{M}_3(\mathbf{T}))\|\boldsymbol{\Delta}_{\mathbf{Z}}\|_{\mathrm{F}}^2 - \|\mathbf{C} - \mathbf{G}\|_{\mathrm{F}}^2\|\boldsymbol{\Delta}_{\mathbf{Z}}\|_{\mathrm{F}}^2.$$

On the other hand,

$$\begin{aligned}&|\langle(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot (\mathbf{G} - \mathbf{C}), (\boldsymbol{\Delta}_{\mathbf{X}}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C}\rangle| \\ &\quad = |\langle(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot (\mathbf{G} - \mathbf{C}), (\mathbf{P}_{\mathbf{U}}\boldsymbol{\Delta}_{\mathbf{X}}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C}\rangle| \\ &\quad \leq \|(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot (\mathbf{G} - \mathbf{C})\|_{\mathrm{F}}\,\|(\mathbf{P}_{\mathbf{U}}\boldsymbol{\Delta}_{\mathbf{X}}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C}\|_{\mathrm{F}} \\ &\quad \leq \|\mathbf{G} - \mathbf{C}\|_{\mathrm{F}}\|\mathbf{P}_{\mathbf{U}}\boldsymbol{\Delta}_{\mathbf{X}}\|_{\mathrm{F}}\|\mathbf{C}\| \\ &\quad \leq 2\|\mathbf{C}\|\|\mathbf{G} - \mathbf{C}\|_{\mathrm{F}}d_{\mathrm{p}}^2(\mathbf{U}, \mathbf{X}).\end{aligned}$$

Observe that

$$\|\mathbf{C}\| \leq \|\mathbf{G}\| + \|\mathbf{G} - \mathbf{C}\| \leq \|\mathbf{G}\| + \|\mathbf{G} - \mathbf{C}\|_{\mathrm{F}} = \|\mathbf{T}\| + \|\mathbf{G} - \mathbf{C}\|_{\mathrm{F}}.$$

We get

$$\begin{aligned}|\langle(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot (\mathbf{G} - \mathbf{C}), (\boldsymbol{\Delta}_{\mathbf{X}}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C}\rangle| &\leq 2\|\mathbf{T}\|\|\mathbf{G} - \mathbf{C}\|_{\mathrm{F}}d_{\mathrm{p}}^2(\mathbf{U}, \mathbf{X}) \\ &\quad + 2\|\mathbf{G} - \mathbf{C}\|_{\mathrm{F}}^2d_{\mathrm{p}}^2(\mathbf{U}, \mathbf{X}).\end{aligned}$$

Similarly,

$$\begin{aligned}|\langle(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot (\mathbf{G} - \mathbf{C}), (\mathbf{U}, \boldsymbol{\Delta}_{\mathbf{Y}}, \mathbf{W}) \cdot \mathbf{C}\rangle| &\leq 2\|\mathbf{T}\|\|\mathbf{G} - \mathbf{C}\|_{\mathrm{F}}d_{\mathrm{p}}^2(\mathbf{V}, \mathbf{Y}) \\ &\quad + 2\|\mathbf{G} - \mathbf{C}\|_{\mathrm{F}}^2d_{\mathrm{p}}^2(\mathbf{V}, \mathbf{Y})\end{aligned}$$

and

$$|\langle(\mathbf{U},\mathbf{V},\mathbf{W})\cdot(\mathbf{G}-\mathbf{C}),(\mathbf{U},\mathbf{V},\boldsymbol{\Delta}_{\mathbf{Z}})\cdot\mathbf{C}\rangle| \le 2\|\mathbf{T}\|\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}} d_{\mathrm{p}}^2(\mathbf{W},\mathbf{Z}) + 2\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}^2 d_{\mathrm{p}}^2(\mathbf{W},\mathbf{Z}).$$

Finally, we note that

$$\begin{aligned}
&|\langle(\boldsymbol{\Delta}_{\mathbf{X}},\mathbf{V},\mathbf{W})\cdot\mathbf{C},(\mathbf{U},\boldsymbol{\Delta}_{\mathbf{Y}},\mathbf{W})\cdot\mathbf{C}\rangle| \\
&\quad = |\langle(\mathbf{P}_{\mathbf{U}}\boldsymbol{\Delta}_{\mathbf{X}},\mathbf{V},\mathbf{W})\cdot\mathbf{C},(\mathbf{U},\mathbf{P}_{\mathbf{V}}\boldsymbol{\Delta}_{\mathbf{Y}},\mathbf{W})\cdot\mathbf{C}\rangle| \\
&\quad \le \|\mathbf{C}\|^2\|\mathbf{P}_{\mathbf{U}}\boldsymbol{\Delta}_{\mathbf{X}}\|_{\mathrm{F}}\|\mathbf{P}_{\mathbf{V}}\boldsymbol{\Delta}_{\mathbf{Y}}\|_{\mathrm{F}} \\
&\quad \le 4\left(\|\mathbf{T}\|+\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}\right)^2 d_{\mathrm{p}}^2(\mathbf{U},\mathbf{X}) d_{\mathrm{p}}^2(\mathbf{V},\mathbf{Y}).
\end{aligned}$$

And similarly,

$$|\langle(\boldsymbol{\Delta}_{\mathbf{X}},\mathbf{V},\mathbf{W})\cdot\mathbf{C},(\mathbf{U},\mathbf{V},\boldsymbol{\Delta}_{\mathbf{Z}})\cdot\mathbf{C}\rangle| \le 4\left(\|\mathbf{T}\|+\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}\right)^2 d_{\mathrm{p}}^2(\mathbf{U},\mathbf{X}) d_{\mathrm{p}}^2(\mathbf{W},\mathbf{Z}),$$

and

$$|\langle(\mathbf{U},\boldsymbol{\Delta}_{\mathbf{Y}},\mathbf{W})\cdot\mathbf{C},(\mathbf{U},\mathbf{V},\boldsymbol{\Delta}_{\mathbf{Z}})\cdot\mathbf{C}\rangle| \le 4\left(\|\mathbf{T}\|+\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}\right)^2 d_{\mathrm{p}}^2(\mathbf{V},\mathbf{Y}) d_{\mathrm{p}}^2(\mathbf{W},\mathbf{Z}).$$

Putting all these bounds together, we get

$$\begin{aligned}
\|\mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}}-\mathbf{T})\|_{\mathrm{F}}^2 \ge\ & \|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}^2 + \left(\frac{\Lambda_{\min}^2}{2} - \|\mathbf{C}-\mathbf{G}\|_{\mathrm{F}}^2\right) \\
&\times\left(\|\boldsymbol{\Delta}_{\mathbf{X}}\|_{\mathrm{F}}^2+\|\boldsymbol{\Delta}_{\mathbf{Y}}\|_{\mathrm{F}}^2+\|\boldsymbol{\Delta}_{\mathbf{Z}}\|_{\mathrm{F}}^2\right) \\
&- 4\|\mathbf{T}\|\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}} d_{\mathrm{p}}^2((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})) \\
&- 4\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}^2 d_{\mathrm{p}}^2((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})) \\
&- 8\left(\|\mathbf{T}\|+\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}\right)^2 d_{\mathrm{p}}^4((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})),
\end{aligned}$$

where, with slight abuse of notation, we write

$$\Lambda_{\min} := \min\left\{\sigma_{\min}(\mathcal{M}_1(\mathbf{T})),\sigma_{\min}(\mathcal{M}_2(\mathbf{T})),\sigma_{\min}(\mathcal{M}_3(\mathbf{T}))\right\}.$$

Recall that

$$\|\boldsymbol{\Delta}_{\mathbf{X}}\|_{\mathrm{F}} \ge d_{\mathrm{p}}(\mathbf{X},\mathbf{U}), \quad \|\boldsymbol{\Delta}_{\mathbf{Y}}\|_{\mathrm{F}} \ge d_{\mathrm{p}}(\mathbf{Y},\mathbf{V}), \quad \text{and} \quad \|\boldsymbol{\Delta}_{\mathbf{Z}}\|_{\mathrm{F}} \ge d_{\mathrm{p}}(\mathbf{Z},\mathbf{W}),$$

so that

$$\|\boldsymbol{\Delta}_{\mathbf{X}}\|_{\mathrm{F}}^2+\|\boldsymbol{\Delta}_{\mathbf{Y}}\|_{\mathrm{F}}^2+\|\boldsymbol{\Delta}_{\mathbf{Z}}\|_{\mathrm{F}}^2 \ge \frac{1}{3} d_{\mathrm{p}}^2((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})).$$

We can further bound $\|\mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}}-\mathbf{T})\|_{\mathrm{F}}^2$ by

$$\begin{aligned}\|\mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}}-\mathbf{T})\|_{\mathrm{F}}^2 \geq\ & \|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}^2 \\ & + \left(\frac{\Lambda_{\min}^2}{6} - 5\|\mathbf{C}-\mathbf{G}\|_{\mathrm{F}}^2 - 4\|\mathbf{T}\|\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}\right) \\ & \times d_{\mathrm{p}}^2((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})) \\ & -16\left(\|\mathbf{T}\|^2 + \|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}^2\right) d_{\mathrm{p}}^4((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z}))\end{aligned}$$

Note that

$$\Lambda_{\min} \geq \kappa_0^{-1}\Lambda_{\max}(\mathbf{T}) \geq \kappa_0^{-1}\|\mathbf{T}\|.$$

If $d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big) \leq C(\alpha\kappa_0\log d)^{-1}$ for a sufficiently small C, we can ensure that

$$\|\mathbf{T}\| d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big) \leq \frac{\Lambda_{\min}}{16}.$$

This implies that

$$\begin{aligned}\|\mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}}-\mathbf{T})\|_{\mathrm{F}}^2 \geq\ & \frac{5}{8}\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}^2 \\ & + \Big(\frac{\Lambda_{\min}^2}{12} - 4\|\mathbf{T}\|\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}\Big) d_{\mathrm{p}}^2\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big).\end{aligned}$$

We have thus proved that under the event $\mathcal{E}_1$,

$$\begin{aligned}\|\mathcal{P}_\Omega\mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}}-\mathbf{T})\|_{\mathrm{F}}^2 \geq\ & \frac{5n}{16d_1d_2d_3}\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}^2 \\ & + \frac{n}{2d_1d_2d_3}\Big(\frac{\Lambda_{\min}^2}{12} - 4\|\mathbf{T}\|\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}\Big) \\ & d_{\mathrm{p}}^2\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big). \qquad (10)\end{aligned}$$

Now consider upper-bounding $\|\mathcal{P}_\Omega\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}\|_{\mathrm{F}}^2$. By Chernoff bound, it is easy to see that with probability $1-d^{-\alpha}$,

$$\max_{\omega\in[d_1]\times[d_2]\times[d_3]}\sum_{i=1}^{n}\mathbb{I}(\omega_i=\omega) \leq C\alpha\log d$$

for some constant $C>0$. Denote this event by $\mathcal{E}_2$. Under this event

$$\|\mathcal{P}_\Omega\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}\|_{\mathrm{F}}^2 \leq C(\alpha\log d)\left\langle\mathcal{P}_\Omega\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}},\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}\right\rangle.$$

To this end, it suffices to obtain upper bounds of

$$\left|\left\langle \mathcal{P}_\Omega \mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}, \mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}\right\rangle\right| \le \frac{n}{d_1d_2d_3}\|\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_{\mathrm{F}}^2 + \left|\left\langle \mathcal{P}_\Omega \mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}, \mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}\right\rangle - \frac{n}{d_1d_2d_3}\|\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_{\mathrm{F}}^2\right|.$$

For $\gamma_1, \gamma_2 > 0$, define

$$\mathcal{K}(\gamma_1, \gamma_2) := \left\{\mathbf{A} \in \mathbb{R}^{d_1\times d_2\times d_3} : \|\mathbf{A}\|_{\mathrm{F}} \le 1, \|\mathbf{A}\|_{\max} \le \gamma_1, \|\mathbf{A}\|_\star \le \gamma_2\right\}.$$

Consider the following empirical process:

$$\beta_n(\gamma_1, \gamma_2) := \sup_{\mathbf{A}\in\mathcal{K}(\gamma_1,\gamma_2)} \left|\frac{1}{n}\langle \mathcal{P}_\Omega \mathbf{A}, \mathbf{A}\rangle - \frac{1}{d_1d_2d_3}\|\mathbf{A}\|_{\mathrm{F}}^2\right|.$$

Obviously,

$$\left|\left\langle \mathcal{P}_\Omega \mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}, \mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}\right\rangle\right| \le \frac{n}{d_1d_2d_3}\|\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_{\mathrm{F}}^2 + n\|\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_{\mathrm{F}}^2\beta_n\Big(\frac{\|\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_{\max}}{\|\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_{\mathrm{F}}}, \frac{\|\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_\star}{\|\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_F}\Big).$$

We now appeal to the following lemma whose proof is given in Appendix C.

Lemma 2 *Given $0 < \delta_1^- < \delta_1^+$, $0 < \delta_2^- < \delta_2^+$ and $t \ge 1$, let*

$$\bar{t} = t + \log\big(\log_2(\delta_1^+/\delta_1^-) + \log_2(\delta_2^+/\delta_2^-) + 3\big).$$

Then there exists a universal constant $C > 0$ such that with probability at least $1-e^{-t}$, the following bound holds for all $\gamma_1 \in [\delta_1^-, \delta_1^+]$ and all $\gamma_2 \in [\delta_2^-, \delta_2^+]$

$$\beta_n(\gamma_1, \gamma_2) \le C\gamma_1\gamma_2\Big(\sqrt{\frac{d}{nd_1d_2d_3}}\log d + \frac{\log^{3/2} d}{n}\Big) + 2\gamma_1\sqrt{\frac{\bar{t}}{nd_1d_2d_3}} + 2\gamma_1^2\frac{\bar{t}}{n}$$

For any $\mathbf{A} \in \mathbb{R}^{d_1\times d_2\times d_3}$, we have $\frac{\|\mathbf{A}\|_{\max}}{\|\mathbf{A}\|_{\mathrm{F}}} \in [1/\sqrt{d_1d_2d_3}, 1]$ and $\frac{\|\mathbf{A}\|_\star}{\|\mathbf{A}\|_{\mathrm{F}}} \in [1, d]$; we apply Lemma 2 with $\delta_1^- = \frac{1}{d_1d_2d_3}$, $\delta_1^+ = 1$, $\delta_2^- = 1$ and $\delta_2^+ = d$. By setting $t = \alpha\log d$ with $\bar{t} = t + \log\big(\log_2(d_1) + \log_2(d_2) + \log_2(d_3) + \log_2(d) + 3\big) \le 6\alpha\log d$, we obtain that with probability at least $1 - d^{-\alpha}$, for all $\gamma_1 \in [(d_1d_2d_3)^{-1}, 1]$ and $\gamma_2 \in [1, d]$,

$$\begin{aligned}\beta_n(\gamma_1, \gamma_2) \le{}& C_1\alpha\gamma_1\gamma_2\Big(\sqrt{\frac{d}{nd_1d_2d_3}}\log d + \frac{\log^{3/2} d}{n}\Big)\\ &+ C_1\sqrt{\alpha}\gamma_1\sqrt{\frac{\log d}{nd_1d_2d_3}} + C_1\alpha\gamma_1^2\frac{\log d}{n}.\end{aligned}$$

Denote this event by $\mathcal{E}_3$. Under $\mathcal{E}_3$, for any $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$,

$$\|\mathbf{A}\|_{\mathrm{F}}^2 \beta_n\Big(\frac{\|\mathbf{A}\|_{\max}}{\|\mathbf{A}\|_{\mathrm{F}}}, \frac{\|\mathbf{A}\|_\star}{\|\mathbf{A}\|_F}\Big) \le C_1\alpha\|\mathbf{A}\|_{\max}\|\mathbf{A}\|_\star\Big(\sqrt{\frac{d}{nd_1d_2d_3}}\log d + \frac{\log^{3/2} d}{n}\Big) + C_1\alpha\|\mathbf{A}\|_{\max}\|\mathbf{A}\|_{\mathrm{F}}\sqrt{\frac{\log d}{nd_1d_2d_3}} + C_1\alpha\|\mathbf{A}\|_{\max}^2\frac{\log d}{n}.$$

This implies that

$$\langle \mathcal{P}_\Omega\mathbf{A}, \mathbf{A}\rangle \le \frac{n}{d_1d_2d_3}\|\mathbf{A}\|_{\mathrm{F}}^2 + C\alpha\|\mathbf{A}\|_{\max}\|\mathbf{A}\|_\star\Big(\sqrt{\frac{nd}{d_1d_2d_3}}\log d + \log^{3/2} d\Big). \quad (11)$$

We shall now focus on $\mathcal{E}_3$ and obtain

$$\Big\langle \mathcal{P}_\Omega\mathbf{Q}_{\mathbf{T}}^\perp\widehat{\mathbf{T}}, \mathbf{Q}_{\mathbf{T}}^\perp\widehat{\mathbf{T}}\Big\rangle \le \frac{n}{d_1d_2d_3}\|\mathbf{Q}_{\mathbf{T}}^\perp\widehat{\mathbf{T}}|_{\mathrm{F}}^2 + C\alpha\|\mathbf{Q}_{\mathbf{T}}^\perp\widehat{\mathbf{T}}|_{\max}\|\mathbf{Q}_{\mathbf{T}}^\perp\widehat{\mathbf{T}}|_\star\Big(\sqrt{\frac{nd}{d_1d_2d_3}}\log d + \log^{3/2} d\Big). \quad (12)$$

It remains to bound $\|\mathbf{Q}_{\mathbf{T}}^\perp\widehat{\mathbf{T}}\|_{\max}$, $\|\mathbf{Q}_{\mathbf{T}}^\perp\widehat{\mathbf{T}}|_\star$ and $\|\mathbf{Q}_{\mathbf{T}}^\perp\widehat{\mathbf{T}}\|_{\mathrm{F}}$. Recall that

$$\mathbf{Q}_{\mathbf{T}}^\perp\widehat{\mathbf{T}} = (\mathbf{P}_{\mathbf{U}}^\perp\mathbf{X}, \mathbf{P}_{\mathbf{V}}^\perp\mathbf{Y}, \mathbf{Z})\cdot\mathbf{C} + (\mathbf{P}_{\mathbf{U}}^\perp\mathbf{X}, \mathbf{Y}, \mathbf{P}_{\mathbf{W}}^\perp\mathbf{Z})\cdot\mathbf{C} + (\mathbf{X}, \mathbf{P}_{\mathbf{V}}^\perp\mathbf{Y}, \mathbf{P}_{\mathbf{W}}^\perp\mathbf{Z})\cdot\mathbf{C} + (\mathbf{P}_{\mathbf{U}}^\perp\mathbf{X}, \mathbf{P}_{\mathbf{V}}^\perp\mathbf{Y}, \mathbf{P}_{\mathbf{W}}^\perp\mathbf{Z})\cdot\mathbf{C}.$$

Recall that $\Lambda_{\max}(\mathbf{C}) := \max\{\|\mathcal{M}_k(\mathbf{C})\|, k = 1, 2, 3\}$. Clearly, $\Lambda_{\max}(\mathbf{C}) \le \Lambda_{\max} + \|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}$ where, with slight abuse of notation, we write $\Lambda_{\max} := \Lambda_{\max}(\mathbf{T})$ for brevity. Then,

$$\begin{aligned}\|\mathbf{Q}_{\mathbf{T}}^\perp\widehat{\mathbf{T}}|_{\mathrm{F}} &\le \big(\Lambda_{\max} + \|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}\big)\\ &\times\Big(\|\mathbf{P}_{\mathbf{U}}^\perp\mathbf{X}\|_{\mathrm{F}}\|\mathbf{P}_{\mathbf{V}}^\perp\mathbf{Y}\|_{\mathrm{F}} + \|\mathbf{P}_{\mathbf{U}}^\perp\mathbf{X}\|_{\mathrm{F}}\|\mathbf{P}_{\mathbf{W}}^\perp\mathbf{Z}\|_{\mathrm{F}} + \|\mathbf{P}_{\mathbf{W}}^\perp\mathbf{Z}\|_{\mathrm{F}}\|\mathbf{P}_{\mathbf{V}}^\perp\mathbf{Y}\|_{\mathrm{F}}\Big)\\ &+\big(\Lambda_{\max} + \|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}\big)\|\mathbf{P}_{\mathbf{U}}^\perp\mathbf{X}\|_{\mathrm{F}}\|\mathbf{P}_{\mathbf{V}}^\perp\mathbf{Y}\|_{\mathrm{F}}\|\mathbf{P}_{\mathbf{W}}^\perp\mathbf{Z}\|_{\mathrm{F}}.\end{aligned}$$

Observe that

$$\|\mathbf{P}_{\mathbf{U}}^\perp\mathbf{X}\|_{\mathrm{F}} = \|\mathbf{P}_{\mathbf{U}}^\perp\boldsymbol{\Delta}_{\mathbf{X}}\|_{\mathrm{F}} \le \|\boldsymbol{\Delta}_{\mathbf{X}}\|_{\mathrm{F}} \le \sqrt{2}d_{\mathrm{p}}(\mathbf{U}, \mathbf{X})$$

and

$$d_{\mathrm{p}}\big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\big) \le (C\alpha\kappa_0\log d)^{-1}.$$

Therefore,

$$\begin{aligned}\|\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_{\mathrm{F}} &\le \big(\Lambda_{\max}+\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}\big)\Big(2d_{\mathrm{p}}^2\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big)\\ &\quad+2\sqrt{2}d_{\mathrm{p}}^3\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big)\Big)\\ &\le 3\big(\Lambda_{\max}+\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}\big)d_{\mathrm{p}}^2\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big).\end{aligned}$$

It is clear that

$$\max_{k=1,2,3}\big\{\operatorname{rank}(\mathcal{M}_k(\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}))\big\}\le 4r.$$

By Lemma 1,

$$\|\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_{\star}\le 4r\|\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_{\mathrm{F}}\le 12r\big(\Lambda_{\max}+\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}\big)d_{\mathrm{p}}^2\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big).$$

Because of the incoherence condition

$$\max\{\mu(\mathbf{\Delta}_{\mathbf{X}}),\mu(\mathbf{\Delta}_{\mathbf{Y}}),\mu(\mathbf{\Delta}_{\mathbf{Z}})\}\le 9\mu_0,$$

we get

$$\|\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_{\max}\le 54\big(\Lambda_{\max}+\|\mathbf{C}-\mathbf{G}\|_{\mathrm{F}}\big)\mu_0^{3/2}\sqrt{\frac{r_1r_2r_3}{d_1d_2d_3}}.$$

By putting the bounds of $\|\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_{\mathrm{F}}$, $\|\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_{\max}$ and $\|\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_{\star}$ into (12), we conclude that on event $\mathcal{E}_3$,

$$\begin{aligned}\Big\langle\mathcal{P}_{\Omega}\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}},\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}\Big\rangle &\le \frac{9n}{d_1d_2d_3}\big(\Lambda_{\max}+\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}\big)^2d_{\mathrm{p}}^4\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big)\\ &\quad+C_1\Big(\alpha r(\Lambda_{\max}+\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}})^2\mu_0^{3/2}\sqrt{\frac{r_1r_2r_3}{d_1d_2d_3}}\Big(\sqrt{\frac{nd}{d_1d_2d_3}}\log d+\log^{3/2}d\Big)\Big)\\ &\quad\times d_{\mathrm{p}}^2\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big)\end{aligned}\tag{13}$$

for a universal constant $C_1>0$. If $d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big)\le(C_2\alpha\kappa_0\log d)^{-1}$ and

$$n\ge C_2\Big(\alpha^4\mu_0^3\kappa_0^4r^2r_1r_2r_3d\log^4d+\alpha^2\mu_0^{3/2}\kappa_0^2r(r_1r_2r_3d_1d_2d_3)^{1/2}\log^{5/2}d\Big).$$

The above upper bound can be simplified as

$$\begin{aligned}\Big\langle\mathcal{P}_{\Omega}\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}},\mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}\Big\rangle &\le \frac{n}{8C\alpha d_1d_2d_3\log d}\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}^2\\ &\quad+\frac{n}{96C\alpha d_1d_2d_3\log d}\Lambda_{\min}^2d_{\mathrm{p}}^2\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big).\end{aligned}\tag{14}$$

Therefore, under $\mathcal{E}_2 \cap \mathcal{E}_3$,

$$\|\mathcal{P}_\Omega \mathbf{Q}_{\mathbf{T}}^{\perp} \widehat{\mathbf{T}}\|_{\mathrm{F}}^2 \le \frac{n}{8d_1d_2d_3}\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}^2 + \frac{n}{96d_1d_2d_3}\Lambda_{\min}^2 d_{\mathrm{p}}^2((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})). \tag{15}$$

Combining (8), (10) and (15), we conclude that

$$\begin{aligned} F(\mathbf{X},\mathbf{Y},\mathbf{Z}) \ge{}& \frac{n}{64d_1d_2d_3}\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}^2 \\ &+ \frac{n}{d_1d_2d_3}\Big(\frac{\Lambda_{\min}^2}{192} - \|\mathbf{T}\|\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}\Big) d_{\mathrm{p}}^2((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})), \end{aligned} \tag{16}$$

with probability at least

$$\mathbb{P}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3\} \ge 1 - 3d^{-\alpha}.$$

Before concluding the proof of lower bound, we first develop a comparable upper bound.

Upper bound of $F(\mathbf{X},\mathbf{Y},\mathbf{Z})$ in Eq. (6). Let

$$\tilde{\mathbf{T}} = (\mathbf{X},\mathbf{Y},\mathbf{Z})\cdot\mathbf{G}.$$

By definition of $\widehat{\mathbf{T}}$,

$$\begin{aligned} F(\mathbf{X},\mathbf{Y},\mathbf{Z}) = \frac{1}{2}\|\mathcal{P}_\Omega(\widehat{\mathbf{T}}-\mathbf{T})\|_{\mathrm{F}}^2 \le \frac{1}{2}\|\mathcal{P}_\Omega(\tilde{\mathbf{T}}-\mathbf{T})\|_{\mathrm{F}}^2 \le \|\mathcal{P}_\Omega\mathbf{Q}_{\mathbf{T}}(\tilde{\mathbf{T}}-\mathbf{T})\|_{\mathrm{F}}^2 \\ + \|\mathcal{P}_\Omega\mathbf{Q}_{\mathbf{T}}^{\perp}\tilde{\mathbf{T}}\|_{\mathrm{F}}^2 \end{aligned}$$

Again, by Lemma 5 of Yuan and Zhang [33], on event $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\begin{aligned} \|\mathcal{P}_\Omega\mathbf{Q}_{\mathbf{T}}(\tilde{\mathbf{T}}-\mathbf{T})\|_{\mathrm{F}}^2 &\le C(\alpha\log d)\Big\langle \mathcal{P}_\Omega\mathbf{Q}_{\mathbf{T}}(\tilde{\mathbf{T}}-\mathbf{T}), \mathbf{Q}_{\mathbf{T}}(\tilde{\mathbf{T}}-\mathbf{T})\Big\rangle \\ &\le \frac{3C\alpha n\log d}{2d_1d_2d_3}\|\mathbf{Q}_{\mathbf{T}}(\tilde{\mathbf{T}}-\mathbf{T})\|_{\mathrm{F}}^2. \end{aligned}$$

Recall that

$$\mathbf{Q}_{\mathbf{T}}(\tilde{\mathbf{T}}-\mathbf{T}) = (\mathbf{\Delta}_{\mathbf{X}},\mathbf{V},\mathbf{W})\cdot\mathbf{G} + (\mathbf{U},\mathbf{\Delta}_{\mathbf{Y}},\mathbf{W})\cdot\mathbf{G} + (\mathbf{U},\mathbf{V},\mathbf{\Delta}_{\mathbf{Z}})\cdot\mathbf{G}.$$

We have

$$\begin{aligned} &\|\mathbf{Q}_{\mathbf{T}}(\tilde{\mathbf{T}}-\mathbf{T})\|_{\mathrm{F}}^2 \le 3 \\ &\times\Big(\|(\mathbf{\Delta}_{\mathbf{X}},\mathbf{V},\mathbf{W})\cdot\mathbf{G}\|_{\mathrm{F}}^2 + \|(\mathbf{U},\mathbf{\Delta}_{\mathbf{Y}},\mathbf{W})\cdot\mathbf{G}\|_{\mathrm{F}}^2 + \|(\mathbf{U},\mathbf{V},\mathbf{\Delta}_{\mathbf{Z}})\cdot\mathbf{G}\|_{\mathrm{F}}^2\Big). \end{aligned}$$

Note that

$$\|(\mathbf{\Delta_X}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{G}\|_{\mathrm{F}}^2 \le \sigma_{\max}^2(\mathcal{M}_1(\mathbf{G}))\|\mathbf{\Delta_X}\|_{\mathrm{F}}^2 \le \Lambda_{\max}^2\|\mathbf{\Delta_X}\|_{\mathrm{F}}^2.$$

Similar bounds hold for $\|(\mathbf{U}, \mathbf{\Delta_Y}, \mathbf{W}) \cdot \mathbf{G}\|_{\mathrm{F}}^2$ and $\|(\mathbf{U}, \mathbf{V}, \mathbf{\Delta_Z}) \cdot \mathbf{G}\|_{\mathrm{F}}^2$. We get on event $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\|\mathcal{P}_\Omega \mathbf{Q_T}(\tilde{\mathbf{T}} - \mathbf{T})\|_{\mathrm{F}}^2 \le \frac{9C\alpha n \log d}{d_1 d_2 d_3}\Lambda_{\max}^2 d_{\mathrm{p}}^2((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})). \tag{17}$$

On the other hand, following the same argument for bounding $\|\mathcal{P}_\Omega \mathbf{Q}_{\widehat{\mathbf{T}}}^{\perp}\widehat{\mathbf{T}}\|_{\mathrm{F}}^2$ as in (15), we can show that

$$\|\mathcal{P}_\Omega \mathbf{Q}_{\mathbf{T}}^{\perp}\tilde{\mathbf{T}}\|_{\mathrm{F}}^2 \le C\alpha \log d\left\langle \mathcal{P}_\Omega \mathbf{Q}_{\mathbf{T}}^{\perp}\tilde{\mathbf{T}}, \mathbf{Q}_{\mathbf{T}}^{\perp}\tilde{\mathbf{T}}\right\rangle \le \frac{n}{96 d_1 d_2 d_3}\Lambda_{\min}^2 d_{\mathrm{p}}^2((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})),$$

under the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$. In summary, we get on event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,

$$\frac{d_1 d_2 d_3}{n} F(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \le 10C\alpha\Lambda_{\max}^2 d_{\mathrm{p}}^2((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})) \log d. \tag{18}$$

The bounds (16) and (18) imply that

$$\begin{aligned}&\frac{n}{64 d_1 d_2 d_3}\|\mathbf{G} - \mathbf{C}\|_{\mathrm{F}}^2 + \frac{n}{d_1 d_2 d_3}\Big(\frac{\Lambda_{\min}^2}{192} - \|\mathbf{T}\|\|\mathbf{G} - \mathbf{C}\|_{\mathrm{F}}\Big) d_{\mathrm{p}}^2\big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\big) \\ &\le F(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \le \frac{10C\alpha n}{d_1 d_2 d_3}\Lambda_{\max}^2 d_{\mathrm{p}}^2\big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\big) \log d\end{aligned}$$

which guarantees that

$$\|\mathbf{G} - \mathbf{C}\|_{\mathrm{F}} \le C(\alpha \log d)^{1/2}\Lambda_{\max} d_{\mathrm{p}}\big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\big). \tag{19}$$

Recall that $\Lambda_{\max} \le \bar{\Lambda}$ and $\Lambda_{\min} \ge \underline{\Lambda}$. We conclude that on event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,

$$\begin{aligned}&\frac{1}{128}\|\mathbf{G} - \mathbf{C}\|_{\mathrm{F}}^2 + \frac{1}{384}\underline{\Lambda}^2 d_{\mathrm{p}}^2\big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\big) \\ &\le \frac{d_1 d_2 d_3}{n} F(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \le C(\alpha \log d)\bar{\Lambda}^2 d_{\mathrm{p}}^2\big((\mathbf{X}, \mathbf{Y}, \mathbf{Z}), (\mathbf{U}, \mathbf{V}, \mathbf{W})\big).\end{aligned}$$

Lower bound of $\|\mathrm{grad}\ F(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\|_{\mathrm{F}}$ **in Eq. (7)** Observe that

$$\|\mathrm{grad}\ F(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\|_{\mathrm{F}} \ge \frac{\langle \mathrm{grad}\ F(\mathbf{X}, \mathbf{Y}, \mathbf{Z}), (\mathbf{D_X}, \mathbf{D_Y}, \mathbf{D_Z})\rangle}{\big(\|\mathbf{D_X}\|_{\mathrm{F}}^2 + \|\mathbf{D_Y}\|_{\mathrm{F}}^2 + \|\mathbf{D_Z}\|_{\mathrm{F}}^2\big)^{1/2}}. \tag{20}$$

Write

$$\mathbf{H} = (\mathbf{D_X}, \mathbf{Y}, \mathbf{Z}) \cdot \mathbf{C} + (\mathbf{X}, \mathbf{D_Y}, \mathbf{Z}) \cdot \mathbf{C} + (\mathbf{X}, \mathbf{Y}, \mathbf{D_Z}) \cdot \mathbf{C}.$$

Then

$$\langle \text{grad } F(\mathbf{X},\mathbf{Y},\mathbf{Z}),(\mathbf{D_X},\mathbf{D_Y},\mathbf{D_Z})\rangle = \left\langle \mathcal{P}_\Omega(\widehat{\mathbf{T}}-\mathbf{T}),\mathbf{H}\right\rangle.$$

Denote

$$\mathbf{H}_1 = (\mathbf{D_X},\mathbf{V},\mathbf{W})\cdot\mathbf{C} + (\mathbf{U},\mathbf{D_Y},\mathbf{W})\cdot\mathbf{C} + (\mathbf{U},\mathbf{V},\mathbf{D_Z})\cdot\mathbf{C}$$

and

$$\begin{aligned}\mathbf{H}_2 := {} & (\mathbf{D_X},\boldsymbol{\Delta_Y},\mathbf{W})\cdot\mathbf{C} + (\mathbf{D_X},\mathbf{V},\boldsymbol{\Delta_Z})\cdot\mathbf{C} + (\mathbf{D_X},\boldsymbol{\Delta_Y},\boldsymbol{\Delta_Z})\cdot\mathbf{C} + (\boldsymbol{\Delta_X},\mathbf{D_Y},\mathbf{W})\cdot\mathbf{C}\\ & + (\mathbf{U},\mathbf{D_Y},\boldsymbol{\Delta_Z})\cdot\mathbf{C} + (\boldsymbol{\Delta_X},\mathbf{D_Y},\boldsymbol{\Delta_Z})\cdot\mathbf{C} + (\boldsymbol{\Delta_X},\mathbf{V},\mathbf{D_Z})\cdot\mathbf{C} + (\mathbf{U},\boldsymbol{\Delta_Y},\mathbf{D_Z})\cdot\mathbf{C}\\ & + (\boldsymbol{\Delta_X},\boldsymbol{\Delta_Y},\mathbf{D_Z})\cdot\mathbf{C}.\end{aligned}$$

Then, $\mathbf{H} = \mathbf{H}_1 + \mathbf{H}_2$ and $\mathbf{Q_T}\mathbf{H}_1 = \mathbf{H}_1$. We write

$$\left\langle \mathcal{P}_\Omega(\widehat{\mathbf{T}}-\mathbf{T}),\mathbf{H}\right\rangle = \left\langle \mathcal{P}_\Omega\mathbf{Q_T}(\widehat{\mathbf{T}}-\mathbf{T}),\mathbf{H}_1\right\rangle + \left\langle \mathcal{P}_\Omega\mathbf{Q}_\mathbf{T}^\perp\widehat{\mathbf{T}},\mathbf{H}_1\right\rangle + \left\langle \mathcal{P}_\Omega(\widehat{\mathbf{T}}-\mathbf{T}),\mathbf{H}_2\right\rangle.$$

Since $\mathbf{Q_T}\mathbf{H}_1 = \mathbf{H}_1$, we can show that under the event $\mathcal{E}_1$,

$$\left\langle \mathcal{P}_\Omega\mathbf{Q_T}(\widehat{\mathbf{T}}-\mathbf{T}),\mathbf{H}_1\right\rangle \geq \frac{d_1d_2d_3}{2n}\left\langle \mathbf{Q_T}(\widehat{\mathbf{T}}-\mathbf{T}),\mathbf{H}_1\right\rangle.$$

Based on the lower bound of $\left\langle \mathbf{Q_T}(\widehat{\mathbf{T}}-\mathbf{T}),\mathbf{H}_1\right\rangle$ proved in Appendix D, we conclude that on event $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3$,

$$\left\langle \mathcal{P}_\Omega\mathbf{Q_T}(\widehat{\mathbf{T}}-\mathbf{T}),\mathbf{H}_1\right\rangle \geq \frac{n}{8d_1d_2d_3}\zeta_1 \geq \frac{\Lambda_{\min}^2}{128}\frac{n}{d_1d_2d_3}d_{\mathrm{p}}^2\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big) \tag{21}$$

where $\zeta_1 := \|(\boldsymbol{\Delta_X},\mathbf{V},\mathbf{W})\cdot\mathbf{C} + (\mathbf{U},\boldsymbol{\Delta_Y},\mathbf{W})\cdot\mathbf{C} + (\mathbf{U},\mathbf{V},\boldsymbol{\Delta_Z})\cdot\mathbf{C}\|_{\mathrm{F}}^2$ with (see Appendix D)

$$\zeta_1 \geq \frac{1}{16}\Lambda_{\min}^2 d_{\mathrm{p}}^2\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big) \tag{22}$$

on event $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3$. Moreover, by Cauchy–Schwarz inequality

$$\left|\left\langle \mathcal{P}_\Omega\mathbf{Q}_\mathbf{T}^\perp\widehat{\mathbf{T}},\mathbf{H}_1\right\rangle\right| \leq \left\langle \mathcal{P}_\Omega\mathbf{Q}_\mathbf{T}^\perp\widehat{\mathbf{T}},\mathbf{Q}_\mathbf{T}^\perp\widehat{\mathbf{T}}\right\rangle^{1/2}\langle\mathcal{P}_\Omega\mathbf{H}_1,\mathbf{H}_1\rangle^{1/2}.$$

Observe that $\mathbf{Q_T}\mathbf{H}_1 = \mathbf{H}_1$. Therefore, under the event $\mathcal{E}_1\cap\mathcal{E}_2$,

$$\langle\mathcal{P}_\Omega\mathbf{Q_T}\mathbf{H}_1,\mathbf{Q_T}\mathbf{H}_1\rangle^{1/2} \leq \sqrt{\frac{3n}{2d_1d_2d_3}}\|\mathbf{H}_1\|_{\mathrm{F}}.$$

Recall the upper bound of $\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}$ as in (19) which implies that $\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}} \le \Lambda_{\min}/2$ if

$$d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big) \le (C\alpha\kappa_0 \log d)^{-1}$$

for a large enough $C > 0$. As a result, on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,

$$\frac{\Lambda_{\min}}{2} \le \Lambda_{\min}(\mathbf{C}) \le \Lambda_{\max}(\mathbf{C}) \le 2\Lambda_{\max} \tag{23}$$

Then, on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,

$$\begin{aligned}\|\mathbf{H}_1\|_{\mathrm{F}} &\le \|(\boldsymbol{\Delta}_{\mathbf{X}},\mathbf{V},\mathbf{W})\cdot\mathbf{C} + (\mathbf{U},\boldsymbol{\Delta}_{\mathbf{Y}},\mathbf{W})\cdot\mathbf{C} + (\mathbf{U},\mathbf{V},\boldsymbol{\Delta}_{\mathbf{Z}})\cdot\mathbf{C}\|_{\mathrm{F}}\\ &\quad + \|(\boldsymbol{\Delta}_{\mathbf{X}}-\mathbf{D}_{\mathbf{X}},\mathbf{V},\mathbf{W})\cdot\mathbf{C} + (\mathbf{U},\boldsymbol{\Delta}_{\mathbf{Y}}-\mathbf{D}_{\mathbf{Y}},\mathbf{W})\cdot\mathbf{C} + (\mathbf{U},\mathbf{V},\boldsymbol{\Delta}_{\mathbf{Z}}-\mathbf{D}_{\mathbf{Z}})\cdot\mathbf{C}\|_{\mathrm{F}}\\ &\le \sqrt{\zeta_1} + 2\Lambda_{\max}\big(\|\boldsymbol{\Delta}_{\mathbf{X}}-\mathbf{D}_{\mathbf{X}}\|_{\mathrm{F}} + \|\boldsymbol{\Delta}_{\mathbf{Y}}-\mathbf{D}_{\mathbf{Y}}\|_{\mathrm{F}} + \|\boldsymbol{\Delta}_{\mathbf{Z}}-\mathbf{D}_{\mathbf{Z}}\|_{\mathrm{F}}\big)\\ &\le \sqrt{\zeta_1} + \sqrt{\zeta_1}8\kappa_0 d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big) \le 2\sqrt{\zeta_1}\end{aligned}$$

where we used the lower bound of ζ_1 in (22). Moreover, it suffices to apply bound (13) and (19) to $\Big\langle \mathcal{P}_\Omega \mathbf{Q}_{\widehat{\mathbf{T}}}^{\perp}\widehat{\mathbf{T}}, \mathbf{Q}_{\widehat{\mathbf{T}}}^{\perp}\widehat{\mathbf{T}}\Big\rangle$. It is easy to check that as long as

$$d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big) \le (C_1\alpha\kappa_0 \log d)^{-1}$$

and

$$n \ge C_1\Big(\alpha^3\kappa_0^2\mu_0^{3/2} r(r_1r_2r_3d_1d_2d_3)^{1/2}\log^{7/2} d + \alpha^6\kappa_0^4\mu_0^3 r^2 r_1r_2r_3 d\log^6 d\Big)$$

for a sufficiently large C_1,

$$\Big\langle \mathcal{P}_\Omega \mathbf{Q}_{\widehat{\mathbf{T}}}^{\perp}\widehat{\mathbf{T}}, \mathbf{Q}_{\widehat{\mathbf{T}}}^{\perp}\widehat{\mathbf{T}}\Big\rangle^{1/2} \le \sqrt{\frac{n}{d_1d_2d_3}}\frac{\Lambda_{\min}}{128\sqrt{6}} d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big), \tag{24}$$

under the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$. Due to the lower bound on ζ_1 in (22),

$$\begin{aligned}\Big|\Big\langle \mathcal{P}_\Omega \mathbf{Q}_{\widehat{\mathbf{T}}}^{\perp}\widehat{\mathbf{T}}, \mathbf{H}_1\Big\rangle\Big| &\le \sqrt{6}\sqrt{\frac{n}{d_1d_2d_3}}\sqrt{\zeta_1}\sqrt{\frac{n}{d_1d_2d_3}}\frac{\Lambda_{\min}}{128\sqrt{6}} d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big)\\ &\le \frac{n}{32d_1d_2d_3}\zeta_1,\end{aligned} \tag{25}$$

under the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$. It remains to control $\big|\langle \mathcal{P}_\Omega(\widehat{\mathbf{T}}-\mathbf{T}), \mathbf{H}_2\rangle\big|$. The following fact (Cauchy–Schwarz inequality) on $\mathcal{E}_2$ is obvious

$$\big|\langle \mathcal{P}_\Omega(\widehat{\mathbf{T}}-\mathbf{T}), \mathbf{H}_2\rangle\big| \le \langle \mathcal{P}_\Omega(\widehat{\mathbf{T}}-\mathbf{T}), \widehat{\mathbf{T}}-\mathbf{T}\rangle^{1/2} \langle \mathcal{P}_\Omega\mathbf{H}_2, \mathbf{H}_2\rangle^{1/2}. \tag{26}$$

On event $\mathcal{E}_3$, by (11)

$$\langle \mathcal{P}_\Omega \mathbf{H}_2, \mathbf{H}_2 \rangle \le \frac{n}{d_1 d_2 d_3} \|\mathbf{H}_2\|_{\mathrm{F}}^2 + n \|\mathbf{H}_2\|_{\mathrm{F}}^2 \beta_n \Big(\frac{\|\mathbf{H}_2\|_{\max}}{\|\mathbf{H}_2\|_{\mathrm{F}}}, \frac{\|\mathbf{H}_2\|_\star}{\|\mathbf{H}_2\|_{\mathrm{F}}} \Big).$$

It is clear that

$$\begin{aligned} \|\mathbf{H}_2\|_{\mathrm{F}} &\le 4\Lambda_{\max} \big(\|\mathbf{\Delta}_{\mathbf{X}}\|_{\mathrm{F}} + \|\mathbf{\Delta}_{\mathbf{Y}}\|_{\mathrm{F}} + \|\mathbf{\Delta}_{\mathbf{Z}}\|_{\mathrm{F}} \big) \big(\|\mathbf{D}_{\mathbf{X}}\|_{\mathrm{F}} + \|\mathbf{D}_{\mathbf{Y}}\|_{\mathrm{F}} + \|\mathbf{D}_{\mathbf{Z}}\|_{\mathrm{F}} \big) \\ &\le 8\sqrt{2} \Lambda_{\max} d_{\mathrm{p}}^2 \big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \big). \end{aligned}$$

Meanwhile, by Appendix E,

$$\|\mathbf{H}_2\|_{\mathrm{F}} \le 4\sqrt{6\zeta_1} d_{\mathrm{p}} \big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \big) + 24 \Lambda_{\max} d_{\mathrm{p}}^3 \big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \big).$$

Moreover, by Lemma 1, $\|\mathbf{H}_2\|_\star \le 18r \|\mathbf{H}_2\|_{\mathrm{F}}$. By Remark 8.1 of Keshavan et al. [17],

$$\max\{\mu(\mathbf{D}_{\mathbf{X}}), \mu(\mathbf{D}_{\mathbf{Y}}), \mu(\mathbf{D}_{\mathbf{Z}})\} \le 55\mu_0.$$

Thus, $\|\mathbf{H}_2\|_{\max} \le C_1 \Lambda_{\max} \mu_0^{3/2} \sqrt{\frac{r_1 r_2 r_3}{d_1 d_2 d_3}}$ for an absolute constant $C_1 > 0$. Applying (11), on the event $\mathcal{E}_3$,

$$\begin{aligned} \langle \mathcal{P}_\Omega \mathbf{H}_2, \mathbf{H}_2 \rangle &\le \frac{n}{d_1 d_2 d_3} \|\mathbf{H}_2\|_{\mathrm{F}}^2 + C\alpha \|\mathbf{H}_2\|_{\max} \|\mathbf{H}_2\|_\star \Big(\sqrt{\frac{nd}{d_1 d_2 d_3}} \log d + \log^{3/2} d \Big) \\ &\le C \cdot \Big\{ \frac{n}{d_1 d_2 d_3} \Lambda_{\max}^2 d_{\mathrm{p}}^6 \big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \big) \\ &\quad + \frac{n}{d_1 d_2 d_3} \zeta_1 d_{\mathrm{p}}^2 \big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \big) \\ &\quad + \alpha r \mu_0^{3/2} \Lambda_{\max}^2 \sqrt{\frac{r_1 r_2 r_3}{d_1 d_2 d_3}} \Big(\sqrt{\frac{nd}{d_1 d_2 d_3}} \log d + \log^{3/2} d \Big) \\ &\quad \times d_{\mathrm{p}}^2 \big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \big) \Big\}. \end{aligned}$$

If

$$d_{\mathrm{p}} \big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \big) \le (C_1 \alpha \kappa_0 \log d)^{-1}$$

and

$$n \ge C_1 \Big(\alpha^3 \mu_0^{3/2} \kappa_0^4 r (r_1 r_2 r_3 d_1 d_2 d_3)^{1/2} \log^{7/2} d + \alpha^6 \mu_0^3 \kappa_0^8 r^2 r_1 r_2 r_3 d \log^6 d \Big),$$

then the above bound can be simplified as

$$\langle \mathcal{P}_\Omega \mathbf{H}_2, \mathbf{H}_2 \rangle \le \frac{n}{d_1 d_2 d_3} \Big(\frac{1}{5000^2 C^2 \alpha^2 \log^2 d} \frac{\Lambda_{\min}^4}{\Lambda_{\max}^2} + C\zeta_1 \Big) d_{\mathrm{p}}^2 \big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \big).$$

Moreover by (24), on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,

$$\begin{aligned}\left\langle \mathcal{P}_\Omega(\widehat{\mathbf{T}}-\mathbf{T}), \widehat{\mathbf{T}}-\mathbf{T}\right\rangle^{1/2} &\le \|\mathcal{P}_\Omega(\widehat{\mathbf{T}}-\mathbf{T})\|_{\mathrm{F}} \\ &\le \|\mathcal{P}_\Omega \mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}}-\mathbf{T})\|_{\mathrm{F}} + \|\mathcal{P}_\Omega \mathbf{Q}_{\mathbf{T}}^{\perp}\widehat{\mathbf{T}}|_{\mathrm{F}} \\ &\le \sqrt{\frac{3C\alpha n \log d}{2d_1d_2d_3}}\|\mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}}-\mathbf{T})\|_{\mathrm{F}} \\ &\quad + \sqrt{\frac{n}{d_1d_2d_3}}\frac{\Lambda_{\min}}{128\sqrt{6}} d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big) \\ &\le 5\sqrt{\frac{n}{d_1d_2d_3}}\Lambda_{\max}(C\alpha\log d) d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big),\end{aligned}$$

where we used the following fact that, in the light of (9), (19), (23),

$$\|\mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}}-\mathbf{T})\|_{\mathrm{F}} \le \|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}} + 2\Lambda_{\max} d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big).$$

Finally, on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, by (26),

$$\begin{aligned}\left\langle \mathcal{P}_\Omega(\widehat{\mathbf{T}}-\mathbf{T}), \mathbf{H}_2\right\rangle &\le \left\langle \mathcal{P}_\Omega(\widehat{\mathbf{T}}-\mathbf{T}), \widehat{\mathbf{T}}-\mathbf{T}\right\rangle^{1/2} \left\langle \mathcal{P}_\Omega \mathbf{H}_2, \mathbf{H}_2\right\rangle^{1/2} \\ &\le \frac{5}{5000}\frac{n}{d_1d_2d_3}\left(\Lambda_{\min}^2 + C\alpha\Lambda_{\max}\sqrt{\zeta_1}\log d\right) \\ &\quad \times d_{\mathrm{p}}^2\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big) \\ &\le \frac{n}{32d_1d_2d_3}\zeta_1, \end{aligned} \tag{27}$$

where we used bound (22) and the fact that

$$d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big) \le (C\alpha\kappa_0\log d)^{-1}.$$

Putting (21), (25), (27) together, we conclude that on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,

$$\begin{aligned}\langle \mathrm{grad}\, F(\mathbf{X},\mathbf{Y},\mathbf{Z}), (\mathbf{D}_{\mathbf{X}},\mathbf{D}_{\mathbf{Y}},\mathbf{D}_{\mathbf{Z}})\rangle &= \left\langle \mathcal{P}_\Omega(\widehat{\mathbf{T}}-\mathbf{T}), \mathbf{H}\right\rangle \\ &\ge \frac{n}{16d_1d_2d_3}\zeta_1 \\ &\ge \frac{n}{256d_1d_2d_3}\Lambda_{\min}^2 d_{\mathrm{p}}^2\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big).\end{aligned}$$

Moreover, note that

$$\|\mathbf{D}_{\mathbf{X}}\|_{\mathrm{F}} + \|\mathbf{D}_{\mathbf{Y}}\|_{\mathrm{F}} + \|\mathbf{D}_{\mathbf{Z}}\|_{\mathrm{F}} \le 2d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big).$$

By (20), we obtain

$$\frac{d_1 d_2 d_3}{n} \|\text{grad } F(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\|_{\text{F}} \geq \frac{\Lambda_{\min}^2}{512} d_{\text{p}}\big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\big),$$

which concludes the proof since $\Lambda_{\min} \geq \underline{\Lambda}$.

8.4 Proof of Theorem 4

We first note that the additional penalty function we imposed on F does not change its local behavior in that Theorem 3 still holds if we replace F with $\tilde{F}$. In the light of Theorem 3, the first statement remains true for $\tilde{F}$ simply due to our choice of ρ. We now argue that the second statement also holds for $\tilde{F}$, more specifically,

$$\begin{aligned} &\frac{d_1 d_2 d_3}{n} \left\|\text{grad } \tilde{F}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\right\|_{\text{F}} \\ &\quad \geq \frac{1}{512} \underline{\Lambda}^2 d_{\text{p}}\Big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\Big), \end{aligned}$$

Observe that

$$\begin{aligned} &\|\text{grad } \tilde{F}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\|_{\text{F}} \\ &\geq \frac{\Big\langle \text{grad } F(\mathbf{X}, \mathbf{Y}, \mathbf{Z}), (\mathbf{D}_{\mathbf{X}}, \mathbf{D}_{\mathbf{Y}}, \mathbf{D}_{\mathbf{Z}})\Big\rangle + \Big\langle \text{grad } G(\mathbf{X}, \mathbf{Y}, \mathbf{Z}), (\mathbf{D}_{\mathbf{X}}, \mathbf{D}_{\mathbf{Y}}, \mathbf{D}_{\mathbf{Z}})\Big\rangle}{\|\mathbf{D}_{\mathbf{X}}\|_{\text{F}} + \|\mathbf{D}_{\mathbf{Y}}\|_{\text{F}} + \|\mathbf{D}_{\mathbf{Z}}\|_{\text{F}}}. \end{aligned}$$

In proving Theorem 3, we showed that

$$\frac{d_1 d_2 d_3}{n} \frac{\Big\langle \text{grad } F(\mathbf{X}, \mathbf{Y}, \mathbf{Z}), (\mathbf{D}_{\mathbf{X}}, \mathbf{D}_{\mathbf{Y}}, \mathbf{D}_{\mathbf{Z}})\Big\rangle}{\|\mathbf{D}_{\mathbf{X}}\|_{\text{F}} + \|\mathbf{D}_{\mathbf{Y}}\|_{\text{F}} + \|\mathbf{D}_{\mathbf{Z}}\|_{\text{F}}} \geq \frac{1}{512} \underline{\Lambda}^2 d_{\text{p}}\Big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\Big).$$

It therefore suffices to show that

$$\langle \text{grad } G(\mathbf{X}, \mathbf{Y}, \mathbf{Z}), (\mathbf{D}_{\mathbf{X}}, \mathbf{D}_{\mathbf{Y}}, \mathbf{D}_{\mathbf{Z}})\rangle \geq 0.$$

This follows the argument from Section 8.2 of Keshavan et al. [17] and is omitted for brevity.

Now that Theorem 3 holds for $\tilde{F}$, we know that $\tilde{F}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ has a unique stationary point in $\mathcal{N}(\delta, 4\mu_0)$ at $(\mathbf{U}, \mathbf{V}, \mathbf{W})$ for $\delta \leq (C\alpha\kappa_0 \log d)^{-1}$. Again, by a similar argument as that from the proof of Theorem 1.2 from Keshavan et al. [17], it can be shown that all iterates $(\mathbf{X}^{(k)}, \mathbf{Y}^{(k)}, \mathbf{Z}^{(k)}) \in \mathcal{N}(\delta/10, 4\mu_0)$ and therefore Algorithm 3 is just gradient descent with exact line search in $\mathcal{N}(\delta/10, 4\mu_0)$. This suggests that Algorithm 3 must converge to the unique stationary point $(\mathbf{U}, \mathbf{V}, \mathbf{W})$. See, e.g., Chapter 8 of Luenberger and Ye [21].

A Proof of Lemma 1

The first claim is straightforward. It suffices to prove the second claim. Let $\mathbf{A} = (\mathbf{U}, \mathbf{V}, \mathbf{W})\cdot\mathbf{C}$ with $\mathbf{C} \in \mathbb{R}^{r_1(\mathbf{A})\times r_2(\mathbf{A})\times r_3(\mathbf{A})}$ being the core tensor. Clearly, $\|\mathbf{A}\|_\star = \|\mathbf{C}\|_\star$ and $\|\mathbf{A}\|_{\mathrm{F}} = \|\mathbf{C}\|_{\mathrm{F}}$. Denote by $\mathbf{C}_1, \ldots, \mathbf{C}_{r_1(\mathbf{A})} \in \mathbb{R}^{r_2(\mathbf{A})\times r_3(\mathbf{A})}$ the mode-1 slices of $\mathbf{C}$. By convexity of nuclear norm,

$$\|\mathbf{C}\|_\star \le \|\mathbf{C}_1\|_\star + \cdots + \|\mathbf{C}_{r_1(\mathbf{A})}\|_\star.$$

As a result,

$$\begin{aligned}\|\mathbf{C}\|_\star^2 &\le r_1(\mathbf{A})\big(\|\mathbf{C}_1\|_\star^2 + \cdots + \|\mathbf{C}_{r_1(\mathbf{A})}\|_\star^2\big)\\ &\le r_1(\mathbf{A})\big(r_2(\mathbf{A}) \wedge r_3(\mathbf{A})\big)\big(\|\mathbf{C}_1\|_{\mathrm{F}}^2 + \ldots + \|\mathbf{C}_{r_1(\mathbf{A})}\|_{\mathrm{F}}^2\big)\\ &= r_1(\mathbf{A})\big(r_2(\mathbf{A}) \wedge r_3(\mathbf{A})\big)\|\mathbf{C}\|_{\mathrm{F}}^2.\end{aligned}$$

Therefore,

$$\|\mathbf{C}\|_\star \le \sqrt{r_1(\mathbf{A})\min\{r_2(\mathbf{A}), r_3(\mathbf{A})\}}\|\mathbf{C}\|_{\mathrm{F}}.$$

By the same process on mode-2 and mode-3 slices of $\mathbf{C}$, we obtain

$$\|\mathbf{C}\|_\star \le \sqrt{r_2(\mathbf{A})\min\{r_1(\mathbf{A}), r_3(\mathbf{A})\}}\|\mathbf{C}\|_{\mathrm{F}},$$

and

$$\|\mathbf{C}\|_\star \le \sqrt{r_3(\mathbf{A})\min\{r_1(\mathbf{A}), r_2(\mathbf{A})\}}\|\mathbf{C}\|_{\mathrm{F}},$$

which concludes the proof.

B Proof of Corollary 1

By Davis–Kahan theorem (see, e.g., Theorem 2 of [32]),

$$d_{\mathrm{p}}\big(\widehat{\mathbf{U}}, \mathbf{U}\big) \le \frac{2\sqrt{r_1}\|\widehat{\mathbf{N}} - \mathbf{M}\mathbf{M}^\top\|}{\sigma_{\min}(\mathbf{M}\mathbf{M}^\top)}.$$

By choosing $m_1 = d_1, m_2 = d_2d_3$ in Theorem 2 and noticing that $n \ge C_1(\alpha + 1)(d_1d_2d_3)^{1/2}$, then

$$\begin{aligned}\|\widehat{\mathbf{N}} - \mathbf{M}\mathbf{M}^\top\| &\le C\alpha^2\frac{(d_1d_2d_3)^{3/2}\log d}{n}\\ &\times\left[\left(1 + \frac{d_1}{d_2d_3}\right)^{1/2} + \left(\frac{n}{d_2d_3\log d}\right)^{1/2}\right]\|\mathbf{M}\|_{\max}^2\end{aligned}$$

with probability at least $1-d^{-\alpha}$. It suffices to control $\|\mathbf{M}\|_{\max}$. Recall that $\mu(\mathbf{T}) \le \mu_0$; then,

$$\|\mathbf{M}\|_{\max} = \|\mathbf{T}\|_{\max} \le \|\mathbf{T}\|\mu_0^{3/2}\left(\frac{r_1 r_2 r_3}{d_1 d_2 d_3}\right)^{1/2}.$$

It is clear by definition that

$$\|\mathbf{T}\|^2/\sigma_{\min}(\mathbf{M}\mathbf{M}^\top) \le \kappa^2(\mathbf{T}) \le \kappa_0^2.$$

As a result, the following bound holds with probability at least $1-d^{-\alpha}$,

$$\begin{aligned} d_{\mathrm{p}}\big(\widehat{\mathbf{U}}, \mathbf{U}\big) &\le 2C\alpha^2\mu_0^3\kappa_0^2 r_1^{3/2} r_2 r_3 \frac{(d_1 d_2 d_3)^{1/2}\log d}{n} \\ &\quad \times \left[\left(1+\frac{d_1}{d_2 d_3}\right)^{1/2} + \left(\frac{n}{d_2 d_3 \log d}\right)^{1/2}\right] \\ &\le 2C\alpha^2\mu_0^3\kappa_0^2 r_1^{3/2} r_2 r_3 \left[\frac{(d_1 d_2 d_3)^{1/2}\log d}{n} + \frac{d_1 \log d}{n} + \left(\frac{d_1 \log d}{n}\right)^{1/2}\right]. \end{aligned}$$

The claim then follows.

C Proof of Lemma 2

For simplicity, define a random tensor $\mathbf{E} \in \{0,1\}^{d_1\times d_2\times d_3}$ based on $\omega \in [d_1]\times[d_2]\times[d_3]$ such that $\mathbf{E}(\omega) = 1$ and all the other entries are 0s. Let $\mathbf{E}_1, \ldots, \mathbf{E}_n$ be i.i.d. copies of $\mathbf{E}$. Equivalently, we write

$$\beta_n(\gamma_1,\gamma_2) = \sup_{\mathbf{A}\in\mathcal{K}(\gamma_1,\gamma_2)} \left|\frac{1}{n}\sum_{i=1}^n \langle \mathbf{A}, \mathbf{E}_i\rangle^2 - \mathbb{E}\langle \mathbf{A}, \mathbf{E}\rangle^2\right|$$

which is the upper bound of an empirical process indexed by $\mathcal{K}(\gamma_1,\gamma_2)$. Define $\delta_{1,j} = 2^j\delta_1^-$ for $j = 0,1,2,\ldots,\lfloor \log \frac{\delta_1^+}{\delta_1^-}$ and $\delta_{2,k} = 2^k\delta_2^-$ for $k = 0,1,2,\ldots,\lfloor \log \frac{\delta_2^+}{\delta_2^-}$. For each j,k, we derive the upper bound of $\beta_n(\gamma_1,\gamma_2)$ with $\gamma_1 \in [\delta_{1,j},\delta_{1,j+1}]$ and $\gamma_2 \in [\delta_{2,k},\delta_{2,k+1}]$. Following the union argument, we can make the bound uniformly true for $\gamma_1 \in [\delta_1^-,\delta_1^+]$ and $\gamma_2 \in [\delta_2^-,\delta_2^+]$.

Consider $\gamma_1 \in [\delta_{1,j},\delta_{1,j+1}]$, $\gamma_2 \in [\delta_{2,k},\delta_{2,k+1}]$, and observe that

$$\sup_{\mathbf{A}\in\mathcal{K}(\gamma_1,\gamma_2)} \left|\langle \mathbf{A}, \mathbf{E}\rangle^2 - \mathbb{E}\langle \mathbf{A}, \mathbf{E}\rangle^2\right| \le \gamma_1^2.$$

Moreover,

$$\sup_{\mathbf{A}\in\mathcal{K}(\gamma_1,\gamma_2)} \mathrm{Var}\big(\langle \mathbf{A}, \mathbf{E}\rangle^2\big) \le \sup_{\mathbf{A}\in\mathcal{K}(\gamma_1,\gamma_2)} \mathbb{E}\langle \mathbf{A}, \mathbf{E}\rangle^4 \le \frac{\gamma_1^2\|\mathbf{A}\|_{\mathrm{F}}^2}{d_1d_2d_3} \le \frac{\gamma_1^2}{d_1d_2d_3}.$$

Applying Bousquet's version of Talagrand's concentration inequality [4], with probability at least $1-e^{-t}$ for all $t \ge 0$,

$$\beta_n(\gamma_1,\gamma_2) \le 2\mathbb{E}\beta_n(\gamma_1,\gamma_2) + 2\gamma_1\sqrt{\frac{t}{nd_1d_2d_3}} + 2\gamma_1^2\frac{t}{n}.$$

By the symmetrization inequality,

$$\mathbb{E}\beta_n(\gamma_1,\gamma_2) \le 2\mathbb{E}\sup_{\mathbf{A}\in\mathcal{K}(\gamma_1,\gamma_2)}\Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle \mathbf{A}, \mathbf{E}_i\rangle^2\Big|,$$

where $\varepsilon_1,\dots,\varepsilon_n$ are i.i.d Rademacher random variables. Since $|\langle \mathbf{A}, \mathbf{E}\rangle| \le \gamma_1$, by the contraction inequality,

$$\mathbb{E}\beta_n(\gamma_1,\gamma_2) \le 4\gamma_1\mathbb{E}\sup_{\mathbf{A}\in\mathcal{K}(\gamma_1,\gamma_2)}\Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle \mathbf{A}, \mathbf{E}_i\rangle\Big|.$$

Denote $\mathbf{\Gamma} = n^{-1}\sum_{i=1}^n \varepsilon_i\mathbf{E}_i \in \mathbb{R}^{d_1\times d_2\times d_3}$. Then,

$$\mathbb{E}\sup_{\mathbf{A}\in\mathcal{K}(\gamma_1,\gamma_2)}\Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle \mathbf{A}, \mathbf{E}_i\rangle\Big| \le \mathbb{E}\sup_{\mathbf{A}\in\mathcal{K}(\gamma_1,\gamma_2)}\|\mathbf{\Gamma}\|\|\mathbf{A}\|_\star \le \gamma_2\mathbb{E}\|\mathbf{\Gamma}\|.$$

It is not difficult to show that

$$\mathbb{E}\|\mathbf{\Gamma}\| \le C\Big(\sqrt{\frac{d}{nd_1d_2d_3}}\log d + \frac{\log^{3/2} d}{n}\Big).$$

See, e.g., Lemma 8 of Yuan and Zhang [33]. The above bound holds as long as

$$n \ge C\Big\{\mu_0(r_1r_2r_3d_1d_2d_3)^{1/2}\log^{3/2} d + \mu_0^2 r_1r_2r_3 d\log^2 d\Big\}.$$

As a result, with probability at least $1-e^{-t}$,

$$\beta_n(\gamma_1,\gamma_2) \le C\gamma_1\gamma_2\Big(\sqrt{\frac{d}{nd_1d_2d_3}}\log d + \frac{\log^{3/2} d}{n}\Big) + 2\gamma_1\sqrt{\frac{t}{nd_1d_2d_3}} + 2\gamma_1^2\frac{t}{n}$$

for $\gamma_1 \in [\delta_{1,j}, \delta_{1,j+1}]$ and $\gamma_2 \in [\delta_{2,k}, \delta_{2,k+1}]$. Now, consider all the combinations of j and k, and we can make the upper bound uniform for all j and k with adjusting t to $\bar{t}$, and C to $2C$.

D Proof of lower bound of $\langle \mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}} - \mathbf{T}), \mathbf{H}_1 \rangle$

Recall that

$$\langle \mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}} - \mathbf{T}), \mathbf{H}_1 \rangle = \Big\langle (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot (\mathbf{C} - \mathbf{G}) + (\mathbf{\Delta_X}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C} + (\mathbf{U}, \mathbf{\Delta_Y}, \mathbf{W}) \cdot \mathbf{C} + (\mathbf{U}, \mathbf{V}, \mathbf{\Delta_Z}) \cdot \mathbf{C}, (\mathbf{D_X}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C} + (\mathbf{U}, \mathbf{D_Y}, \mathbf{W}) \cdot \mathbf{C} + (\mathbf{U}, \mathbf{V}, \mathbf{D_Z}) \cdot \mathbf{C} \Big\rangle.$$

Clearly, the right-hand side can be written as $\zeta_1 + \zeta_2 + \zeta_3$ where

$$\begin{aligned}
\zeta_1 &= \|(\mathbf{\Delta_X}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C} + (\mathbf{U}, \mathbf{\Delta_Y}, \mathbf{W}) \cdot \mathbf{C} + (\mathbf{U}, \mathbf{V}, \mathbf{\Delta_Z}) \cdot \mathbf{C}\|_{\mathrm{F}}^2 \\
\zeta_2 &= \Big\langle (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot (\mathbf{C} - \mathbf{G}), (\mathbf{D_X}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C} + (\mathbf{U}, \mathbf{D_Y}, \mathbf{W}) \cdot \mathbf{C} + (\mathbf{U}, \mathbf{V}, \mathbf{D_Z}) \cdot \mathbf{C} \Big\rangle \\
\zeta_3 &= \Big\langle \mathbf{\Delta_X}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C} + (\mathbf{U}, \mathbf{\Delta_Y}, \mathbf{W}) \cdot \mathbf{C} + (\mathbf{U}, \mathbf{V}, \mathbf{\Delta_Z}) \cdot \mathbf{C}, (\mathbf{D_X} - \mathbf{\Delta_X}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C} \\
&\quad + (\mathbf{U}, \mathbf{D_Y} - \mathbf{\Delta_Y}, \mathbf{W}) \cdot \mathbf{C} + (\mathbf{U}, \mathbf{V}, \mathbf{D_Z} - \mathbf{\Delta_Z}) \cdot \mathbf{C} \Big\rangle.
\end{aligned}$$

Clearly,

$$\begin{aligned}
\zeta_1 &\geq \|(\mathbf{\Delta_X}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C}\|_{\mathrm{F}}^2 + \|(\mathbf{U}, \mathbf{\Delta_Y}, \mathbf{W}) \cdot \mathbf{C}\|_{\mathrm{F}}^2 + \|(\mathbf{U}, \mathbf{V}, \mathbf{\Delta_Z}) \cdot \mathbf{C}\|_{\mathrm{F}}^2 \\
&\quad - 2\Lambda_{\max}^2(\mathbf{C}) \Big(\|\mathbf{U}^\top \mathbf{\Delta_X}\|_{\mathrm{F}} \|\mathbf{V}^\top \mathbf{\Delta_Y}\|_{\mathrm{F}} + \|\mathbf{U}^\top \mathbf{\Delta_X}\|_{\mathrm{F}} \|\mathbf{W}^\top \mathbf{\Delta_Z}\|_{\mathrm{F}} + \|\mathbf{V}^\top \mathbf{\Delta_Y}\|_{\mathrm{F}} \|\mathbf{W}^\top \mathbf{\Delta_Z}\|_{\mathrm{F}} \Big) \\
&\geq \Lambda_{\min}^2(\mathbf{C}) \Big(\|\mathbf{\Delta_X}\|_{\mathrm{F}}^2 + \|\mathbf{\Delta_Y}\|_{\mathrm{F}}^2 + \|\mathbf{\Delta_Z}\|_{\mathrm{F}}^2 \Big) - 8\Lambda_{\max}^2(\mathbf{C}) d_{\mathrm{p}}^4\big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\big)
\end{aligned}$$

where we used the fact that

$$\|\mathbf{U}^\top \mathbf{\Delta_X}\|_{\mathrm{F}} \leq 2 d_{\mathrm{p}}^2(\mathbf{U}, \mathbf{X}).$$

Recall from (23) that on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, we have

$$\frac{\Lambda_{\min}}{2} \leq \Lambda_{\min}(\mathbf{C}) \leq \Lambda_{\max}(\mathbf{C}) \leq 2\Lambda_{\max}.$$

Then

$$\zeta_1 \geq \frac{1}{12} \Lambda_{\min}^2 d_{\mathrm{p}}^2\big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\big) - 32\Lambda_{\max}^2 d_{\mathrm{p}}^4\big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\big).$$

It also implies that on the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,

$$\zeta_1 \geq \frac{1}{2} \Big(\|(\mathbf{\Delta_X}, \mathbf{V}, \mathbf{W}) \cdot \mathbf{C}\|_{\mathrm{F}}^2 + \|(\mathbf{U}, \mathbf{\Delta_Y}, \mathbf{W}) \cdot \mathbf{C}\|_{\mathrm{F}}^2 + \|(\mathbf{U}, \mathbf{V}, \mathbf{\Delta_Z}) \cdot \mathbf{C}\|_{\mathrm{F}}^2 \Big). \quad (28)$$

We can control $|\zeta_3|$ in the same fashion. Indeed,

$$\begin{aligned}
|\zeta_3|^2 &\leq |\zeta_1| \Lambda_{\max}^2(\mathbf{C}) (\|\mathbf{D_X} - \mathbf{\Delta_X}\|_{\mathrm{F}}^2 + \|\mathbf{D_Y} - \mathbf{\Delta_Y}\|_{\mathrm{F}}^2 + \|\mathbf{D_Z} - \mathbf{\Delta_Z}\|_{\mathrm{F}}^2) \\
&\leq 4|\zeta_1| \Lambda_{\max}^2 d_{\mathrm{p}}^4\big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\big).
\end{aligned}$$

If

$$d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big) \le (C\alpha\kappa_0 \log d)^{-1}$$

for large $C > 0$, then under the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,

$$\zeta_1 \ge \frac{1}{16}\Lambda_{\min}^2 d_{\mathrm{p}}^2\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big) \quad \text{and} \quad |\zeta_3| \le \frac{\zeta_1}{4}$$

To control ζ_2, recall that $\mathbf{X}^\top \mathbf{D}_{\mathbf{X}} = \mathbf{0}$, $\mathbf{Y}^\top \mathbf{D}_{\mathbf{Y}} = \mathbf{0}$ and $\mathbf{Z}^\top \mathbf{D}_{\mathbf{Z}} = \mathbf{0}$. Then,

$$\begin{aligned}
|\zeta_2| &\le |\langle(\boldsymbol{\Delta}_{\mathbf{X}},\mathbf{V},\mathbf{W})\cdot(\mathbf{C}-\mathbf{G}),(\mathbf{D}_{\mathbf{X}},\mathbf{V},\mathbf{W})\cdot\mathbf{C}\rangle| \\
&\quad + |\langle(\mathbf{U},\boldsymbol{\Delta}_{\mathbf{Y}},\mathbf{W})\cdot(\mathbf{C}-\mathbf{G}),(\mathbf{U},\mathbf{D}_{\mathbf{Y}},\mathbf{W})\cdot\mathbf{C}\rangle| \\
&\quad + |\langle(\mathbf{U},\mathbf{V},\boldsymbol{\Delta}_{\mathbf{Z}})\cdot(\mathbf{C}-\mathbf{G}),(\mathbf{U},\mathbf{V},\mathbf{D}_{\mathbf{Z}})\cdot\mathbf{C}\rangle| \\
&\le 2\|\mathbf{C}-\mathbf{G}\|_{\mathrm{F}}\Big\{\Big(\|(\boldsymbol{\Delta}_{\mathbf{X}},\mathbf{V},\mathbf{W})\cdot\mathbf{C}\|_{\mathrm{F}} + \|\mathbf{U},\boldsymbol{\Delta}_{\mathbf{Y}},\mathbf{W})\cdot\mathbf{C}\|_{\mathrm{F}} + \|(\mathbf{U},\mathbf{V},\boldsymbol{\Delta}_{\mathbf{Z}})\cdot\mathbf{C}\|_{\mathrm{F}}\Big) \\
&\quad + \Lambda_{\max}(\mathbf{C})\Big(\|\mathbf{D}_{\mathbf{X}}-\boldsymbol{\Delta}_{\mathbf{X}}\|_{\mathrm{F}} + \|\mathbf{D}_{\mathbf{Y}}-\boldsymbol{\Delta}_{\mathbf{Y}}\|_{\mathrm{F}} + \|\mathbf{D}_{\mathbf{Z}}-\boldsymbol{\Delta}_{\mathbf{Z}}\|_{\mathrm{F}}\Big)\Big\} d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big) \\
&\le 2\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}}\sqrt{\zeta_1}\, d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big) \\
&\quad + 4\|\mathbf{C}-\mathbf{G}\|_{\mathrm{F}}\Lambda_{\max} d_{\mathrm{p}}^3\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big).
\end{aligned}$$

Recall from (19) that under the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,

$$\|\mathbf{G}-\mathbf{C}\|_{\mathrm{F}} \le C\Lambda_{\max}(\alpha \log d)^{1/2} d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big).$$

Therefore, $|\zeta_2| \le \zeta_1/2$ in view of the lower bound of ζ_1. In summary, under the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,

$$\langle \mathbf{Q}_{\mathbf{T}}(\widehat{\mathbf{T}}-\mathbf{T}), \mathbf{H}_1\rangle \ge \frac{1}{4}\zeta_1 \ge \frac{1}{64}\Lambda_{\min}^2 d_{\mathrm{p}}^2\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big).$$

E Upper bound of $\|\mathbf{H}_2\|_{\mathrm{F}}$

It is shown in (28) that if $d_{\mathrm{p}}\big((\mathbf{U},\mathbf{V},\mathbf{W}),(\mathbf{X},\mathbf{Y},\mathbf{Z})\big) \le (C\alpha\kappa_0 \log d)^{-1}$, then

$$\zeta_1 \ge \frac{1}{2}\Big(\|(\boldsymbol{\Delta}_{\mathbf{X}},\mathbf{V},\mathbf{W})\cdot\mathbf{C}\|_{\mathrm{F}}^2 + \|(\mathbf{U},\boldsymbol{\Delta}_{\mathbf{Y}},\mathbf{W})\cdot\mathbf{C}\|_{\mathrm{F}}^2 + \|(\mathbf{U},\mathbf{V},\boldsymbol{\Delta}_{\mathbf{Z}})\cdot\mathbf{C}\|_{\mathrm{F}}^2\Big).$$

Observe that

$$\|(\boldsymbol{\Delta}_{\mathbf{X}},\mathbf{V},\mathbf{W})\cdot\mathbf{C}\|_{\mathrm{F}}^2 = \|\mathcal{M}_2(\mathbf{C})(\boldsymbol{\Delta}_{\mathbf{X}}\otimes\mathbf{W})\|_{\mathrm{F}} = \|\mathcal{M}_3(\mathbf{C})(\boldsymbol{\Delta}_{\mathbf{X}}\otimes\mathbf{V})\|_{\mathrm{F}}$$

which implies that

$$\zeta_1 \geq \frac{1}{6}\Big(\|\mathcal{M}_2(\mathbf{C})(\boldsymbol{\Delta}_\mathbf{X} \otimes \mathbf{W})\|_\mathrm{F} + \|\mathcal{M}_3(\mathbf{C})(\mathbf{U} \otimes \boldsymbol{\Delta}_\mathbf{Y})\|_\mathrm{F} + \|\mathcal{M}_1(\mathbf{C})(\mathbf{V} \otimes \boldsymbol{\Delta}_\mathbf{Z})\|_\mathrm{F}\Big)^2$$

By definition of $\mathbf{H}_2$, we obtain

$$\begin{aligned}\|\mathbf{H}_2\|_\mathrm{F} \leq\ & \|\mathcal{M}_1(\mathbf{C})(\boldsymbol{\Delta}_\mathbf{Y} \otimes \mathbf{W})\|_\mathrm{F}\|\mathbf{D}_\mathbf{X}\|_\mathrm{F} + \|\mathcal{M}_1(\mathbf{C})(\mathbf{V} \otimes \boldsymbol{\Delta}_\mathbf{Z})\|_\mathrm{F}\|\mathbf{D}_\mathbf{X}\|_\mathrm{F} \\ &+ \|\mathcal{M}_2(\mathbf{C})(\boldsymbol{\Delta}_\mathbf{X} \otimes \mathbf{W})\|_\mathrm{F}\|\mathbf{D}_\mathbf{Y}\|_\mathrm{F} + \|\mathcal{M}_2(\mathbf{C})(\mathbf{U} \otimes \boldsymbol{\Delta}_\mathbf{Z})\|_\mathrm{F}\|\mathbf{D}_\mathbf{Y}\|_\mathrm{F} \\ &+ \|\mathcal{M}_3(\mathbf{C})(\boldsymbol{\Delta}_\mathbf{X} \otimes \mathbf{V})\|_\mathrm{F}\|\mathbf{D}_\mathbf{Z}\|_\mathrm{F} + \|\mathcal{M}_3(\mathbf{C})(\mathbf{U} \otimes \boldsymbol{\Delta}_\mathbf{Y})\|_\mathrm{F}\|\mathbf{D}_\mathbf{Z}\|_\mathrm{F} \\ &+ 24\Lambda_{\max} d_\mathrm{p}^3\big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\big)\end{aligned}$$

where we used the fact $\Lambda_{\max}(\mathbf{C}) \leq 2\Lambda_{\max}$ from (23). Clearly,

$$\begin{aligned}\|\mathbf{H}_2\|_\mathrm{F} &\leq 2\sqrt{6\zeta_1}\big(\|\mathbf{D}_\mathbf{X}\|_\mathrm{F} + \|\mathbf{D}_\mathbf{Y}\|_\mathrm{F} + \|\mathbf{D}_\mathbf{Z}\|_\mathrm{F}\big) + 24\Lambda_{\max} d_\mathrm{p}^3\big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\big) \\ &\leq 4\sqrt{6\zeta_1} d_\mathrm{p}\big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\big) + 24\Lambda_{\max} d_\mathrm{p}^3\big((\mathbf{U}, \mathbf{V}, \mathbf{W}), (\mathbf{X}, \mathbf{Y}, \mathbf{Z})\big).\end{aligned}$$

References

1. P. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.
2. Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.
3. Boaz Barak and Ankur Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *29th Annual Conference on Learning Theory*, pages 417–445, 2016.
4. Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.
5. Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.
6. Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.
7. S. Cohen and M. Collins. Tensor decomposition for fast parsing with latent-variable PCFGS. In *Advances in Neural Information Processing Systems*, 2012.
8. Victor de la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 1999.
9. Victor H de la Peña and Stephen J Montgomery-Smith. Decoupling inequalities for the tail probabilities of multivariate U-statistics. *The Annals of Probability*, pages 806–816, 1995.
10. Vin de Silva and Lek-Heng Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1084–1127, 2008.
11. Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.
12. Lars Elden and Berkant Savas. A Newton-Grassmann method for computing the best multilinear rank-(r_1, r_2, r_3) approximation of a tensor. *SIAM Journal on Matrix Analysis and Applications*, 31(2):248–271, 2009.
13. Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.
14. David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.
15. C. Hillar and Lek-Heng Lim. Most tensor problems are NP-hard. *Journal of ACM*, 60(6):45, 2013.

16. Prateek Jain and Sewoong Oh. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, pages 1431–1439, 2014.
17. Raghunandan H Keshavan, Sewoong Oh, and Andrea Montanari. Matrix completion from a few entries. In *2009 IEEE International Symposium on Information Theory*, pages 324–328. IEEE, 2009.
18. Daniel Kressner, Michael Steinlechner, and Bart Vandereycken. Low-rank tensor completion by Riemannian optimization. *BIT Numerical Mathematics*, 54(2):447–468, 2014.
19. N. Li and B. Li. Tensor completion for on-board compression of hyperspectral images. In *17th IEEE International Conference on Image Processing (ICIP)*, pages 517–520, 2010.
20. Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013.
21. David G Luenberger and Yinyu Ye. *Linear and nonlinear programming*, volume 228. Springer, 2015.
22. Andrea Montanari and Nike Sun. Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics*, 2016.
23. Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved convex relaxations for tensor recovery. *Journal of Machine Learning Research*, 1:1–48, 2014.
24. Holger Rauhut and Željka Stojanac. Tensor theta norms and low rank recovery. *arXiv preprint* arXiv:1505.05175, 2015.
25. Holger Rauhut, Reinhold Schneider, and Zeljka Stojanac. Low rank tensor recovery via iterative hard thresholding. *arXiv preprint* arXiv:1602.05217, 2016.
26. Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.
27. Berkant Savas and Lek-Heng Lim. Quasi-newton methods on Grassmannians and multilinear approximations of tensors. *SIAM Journal on Matrix Analysis and Applications*, 32(6):3352–3393, 2010.
28. O. Semerci, N. Hao, M. Kilmer, and E. Miller. Tensor-based formulation and nuclear norm regularization for multienergy computed tomography. *IEEE Transactions on Image Processing*, 23:1678–1693, 2014.
29. N.D. Sidiropoulos and N. Nion. Tensor algebra and multi-dimensional harmonic retrieval in signal processing for mimo radar. *IEEE Transactions on Signal Processing*, 58:5693–5705, 2010.
30. Ryota Tomioka, Kohei Hayashi, and Hisashi Kashima. Estimation of low-rank tensors via convex optimization. *arXiv preprint* arXiv:1010.0789, 2010.
31. Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.
32. Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.
33. Ming Yuan and Cun-Hui Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, pages 1031–1068, 2016.
34. Ming Yuan and Cun-Hui Zhang. Incoherent tensor norms and their applications in higher order tensor completion. *IEEE Transactions on Information Theory*, 63(10):6753–6766, 2017.