Information and Inference: A Journal of the IMA (2022) **00**, 1–42
https://doi.org/10.1093/imaiai/iaac002

A super-resolution framework for tensor decomposition

Qiuwei Li[†]
Damo Academy, Alibaba Group (US), Bellevue, WA, USA

Ashley Prater
Air Force Research Laboratory, Information Directorate Rome, NY, USA

Lixin Shen
Department of Mathematics, Syracuse University, Syracuse, NY, USA

and

Gongguo Tang
Department of Electrical, Computer & Energy Engineering, University of Colorado, Boulder, CO, USA
[†]Corresponding author: liqiuweiss@gmail.com

[Received on 8 August 2020; revised on 30 September 2021; accepted on 7 February 2022]

This work considers a super-resolution framework forovercomplete tensor decomposition. Specifically, we view tensor decomposition as a super-resolution problem of recovering a sum of Dirac measures on the sphere and solve it by minimizing a continuous analog of the ℓ_1 norm on the space of measures. The optimal value of this optimization defines the tensor nuclear norm. Similar to the separation condition in the super-resolution problem, by explicitly constructing a dual certificate, we develop incoherence conditions of the tensor factors so that they form the unique optimal solution of the continuous analog of ℓ_1 norm minimization. Remarkably, the derived incoherence conditions are satisfied with high probability by random tensor factors uniformly distributed on the sphere, implying global identifiability of random tensor factors.

Keywords: atomic norm minimization; dual certificate; nonconvex; tensor decomposition; tensor nuclear norm; super resolution.

1. Introduction

Tensors provide natural representations for massive multi-mode datasets encountered in many applications including image and video processing [6], collaborative filtering [31], array signal processing [52], convolutional networks design [27] and psychometrics [53]. Tensor methods also form the backbone of many machine learning, signal processing and statistical algorithms, including independent component analysis [14], latent graphical model learning [2], dictionary learning [3] and Gaussian mixture estimation [51]. The utility of tensors in such diverse applications is mainly due to the ability to identify *overcomplete*, *non-orthogonal* factors from tensor data as already suggested by Kruskal's theorem [35]. This is known as tensor decomposition, which describes the problem of decomposing a tensor into a linear combination of a small number of rank-1 tensors. The identifiability of tensor factors is in sharp contrast to the inherent ambiguous nature of matrix decompositions without additional assumptions such as orthogonality and non-negativity.

In addition to its practical applicability, tensor decomposition is also of fundamental theoretical interest in solving linear inverse problems involving low-rank tensors. For one thing, theoretical results for tensor decomposition inform what types of rank-1 tensor combinations are identifiable given full observations. For another, a dual polynomial is constructed to certify a particular decomposition, which is useful in investigating the regularization power of the tensor nuclear norm for tensor inverse problems, including tensor completion, tensor denoising and robust tensor principal component analysis. We expect that the *dual certificate* constructed in this work will play an important role in these tensor inverse problems similar to that of the subdifferential characterization of matrix nuclear norm in matrix completion and low-rank matrix recovery [13, 50].

1.1 *The tensor decomposition problem*

In this work, we focus on third-order nonsymmetric tensors that can be decomposed into a linear combination of unit-norm, rank-1 tensors of the form $\boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} \in \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2} \otimes \mathbb{R}^{n_3}$. More precisely, consider the following nonsymmetric tensor decomposition

$$\mathcal{T} = \sum_{p=1}^{r} \lambda_p^\star \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star. \tag{1.1}$$

Through this work, we assume the rank-1 tensor factors $\{(\boldsymbol{u}_p^\star, \boldsymbol{v}_p^\star, \boldsymbol{w}_p^\star)\}$ are living on the unit spheres and might be *overcomplete*, that is, r is potentially greater than the individual tensor dimensions n_1, n_2 and n_3. Without loss of generality, we assume that the coefficients $\{\lambda_p^\star\}$ are positive as their signs can be absorbed into the factors.

Problem 1.1 The tensor decomposition problem is the inverse problem of retrieving those ground-truth rank-1 tensor factors $\{(\boldsymbol{u}_p^\star, \boldsymbol{v}_p^\star, \boldsymbol{w}_p^\star)\}_{p=1}^r$ from the tensor data $\mathcal{T}$ in (1.1) [36].

1.2 *The super-resolution framework*

Tensor decomposition is an extremely challenging problem [29]. This is because we lack proper theories for basic tensor concepts and operations such as singular values, vectors and singular value decompositions. To address these challenging issues, we will consider a *super-resolution* framework for tensor decomposition. More precisely, we can view tensor decomposition as a problem of *measure estimation* from moments. This is because we can rewrite the tensor decomposition (1.1) as a integral on the unit spheres $\mathbb{K} := \mathbb{S}^{n_1-1} \times \mathbb{S}^{n_2-1} \times \mathbb{S}^{n_3-1}$:

$$\mathcal{T} = \int_{\mathbb{K}} \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} \, \mathrm{d}\, \mu^\star. \tag{1.2}$$

and then the problem of retrieving the rank-1 tensor factors $\{(\boldsymbol{u}_p^\star, \boldsymbol{v}_p^\star, \boldsymbol{w}_p^\star)\}$ from the observed tensor entries in $\mathcal{T}$ is equivalent to recovering a linear combination of Dirac measures defined on the unit spheres $\mathbb{K}$:

$$\mu^\star = \sum_{p=1}^{r} \lambda_p^\star \delta(\boldsymbol{u} - \boldsymbol{u}_p^\star, \boldsymbol{v} - \boldsymbol{v}_p^\star, \boldsymbol{w} - \boldsymbol{w}_p^\star) \tag{1.3}$$

Several advantages are offered by this super-resolution framework. First, it provides a natural way to extend the ℓ_1 norm minimization in finding sparse representations for finite dictionaries [20] to tensor

decomposition. By viewing the set of rank-1 tensors $\mathcal{A} = \{\boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} : (\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \in \mathbb{K}\}$ as a dictionary with an infinite number of atoms, this formulation allows us to find a *sparse*[1] representation of $\boldsymbol{\mathcal{T}}$ by minimizing the ℓ_1 norm of the representation coefficients with respect to the dictionary $\mathcal{A}$. More precisely, we recover $\mu^\star$ from the tensor $\boldsymbol{\mathcal{T}}$ by solving a continuous analog of ℓ_1 norm minimization (a.k.a. the total mass minimization over the space of measures)

$$\text{minimize}_{\ \mu \in \mathcal{M}_+(\mathbb{K})} \mu(\mathbb{K}) \ \text{ subject to } \boldsymbol{\mathcal{T}} = \int_{\mathbb{K}} \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} \, \mathrm{d}\, \mu \tag{1.4}$$

where $\mathcal{M}_+(\mathbb{K})$ is the set of (non-negative) Borel measures on $\mathbb{K}$, and $\mu(\mathbb{K})$ is the total measure/mass of the set $\mathbb{K}$ measured by the Borel measure $\mu \in \mathcal{M}_+(\mathbb{K})$. Second, the optimal value of the total mass minimization defines precisely the *tensor nuclear norm* [25, Proposition 3.1), which is a special case of atomic norms [15, Eq. (2)) corresponding to the atomic set $\mathcal{A}$. The tensor nuclear norm is useful in many tensor inverse problems, such as tensor completion [10] and robust tensor principal component analysis [45].

2. Main results

The main focus of this work is on characterizing the conditions when the tensor factors $\{(\boldsymbol{u}_p^\star, \boldsymbol{v}_p^\star, \boldsymbol{w}_p^\star)\}_{p=1}^r$ correspond to the unique optimal solution of the continuous analog of ℓ_1 norm minimization (1.4), which is extension of the incoherence condition in matrix completion problem [13], the minimum separation condition in mathematical super resolution [12] and the wrap-around distance condition in line spectral estimation [54]. More precisely, we develop the following three assumptions, namely incoherence condition, bounded spectral norm condition and Gram isometry condition. For ease of exposition, in what follows, these assumptions and the main result of this work will be presented for square tensors with $n_1 = n_2 = n_3 = n$.

Assumption I: Incoherence condition.

$$\Delta := \max_{p \neq q} \max\{|\langle \boldsymbol{u}_p^\star, \boldsymbol{u}_q^\star \rangle|, |\langle \boldsymbol{v}_p^\star, \boldsymbol{v}_q^\star \rangle|, |\langle \boldsymbol{w}_p^\star, \boldsymbol{w}_q^\star \rangle|\} \leq \frac{\tau(\log n)}{\sqrt{n}}, \tag{2.1}$$

where $\tau(\cdot)$ is a polynomial function of its argument.[2]

Assumption II: Bounded spectral norm condition.

$$\max\{\|\mathbf{U}\|, \|\mathbf{V}\|, \|\mathbf{W}\|\} \leq 1 + c\sqrt{\frac{r}{n}} \tag{2.2}$$

for some constant $c > 0$, where $\mathbf{U} := \begin{bmatrix} \boldsymbol{u}_1^\star & \cdots & \boldsymbol{u}_r^\star \end{bmatrix}$, $\mathbf{V} := \begin{bmatrix} \boldsymbol{v}_1^\star & \cdots & \boldsymbol{v}_r^\star \end{bmatrix}$, and $\mathbf{W} := \begin{bmatrix} \boldsymbol{w}_1^\star & \cdots & \boldsymbol{w}_r^\star \end{bmatrix}$.
Assumption III: Gram isometry condition.

$$\|(\mathbf{U}^\top \mathbf{U}) \odot (\mathbf{V}^\top \mathbf{V}) - \mathbf{I}\| \leq \kappa(\log n) \frac{\sqrt{r}}{n}, \tag{2.3}$$

[1] The decomposition (1.1) is sparse, because in most practical scenarios, r is much smaller than the product $n_1 n_2 n_3$.

[2] That being said, $\tau(\cdot)$ is of the form $\tau(x) = a_m x^m + a_{m-1} x^{m-1} + \ldots + a_2 x^2 + a_1 x + a_0$ with some (positive) real numbers a's being the coefficients of the polynomial and some positive integer m being the degree of the polynomial.

where $\kappa(\cdot)$ is a polynomial function of its argument. Similar bounds hold for $\mathbf{U}, \mathbf{W}$ and $\mathbf{V}, \mathbf{W}$.

THEOREM 2.1. Suppose the target tensor $\mathcal{T} \in \mathbb{R}^{n\times n\times n}$ admits a decomposition (1.1) with the normalized tensor factors $\{(\boldsymbol{u}_p^\star, \boldsymbol{v}_p^\star, \boldsymbol{w}_p^\star)\}_{p=1}^r$ satisfying Assumptions I, II, III and

$$r \le \frac{n^{17/16}}{32c^2\sqrt{15\tau(\log n)}} \tag{2.4}$$

with the polynomial $\tau(\cdot)$ given in (2.1), the constant c given in (2.2), and n being large enough. Then the true factors $\{(\boldsymbol{u}_p^\star, \boldsymbol{v}_p^\star, \boldsymbol{w}_p^\star)\}_{p=1}^r$ correspond to the unique optimal solution of the continuous analog of ℓ_1 norm minimization (1.4) up to a sign ambiguity.

A few remarks follow. Firstly, since $r = O\big(n^{17/16}/\sqrt{\tau(\log n)}\big) \gg n$, total mass minimization is guaranteed to recover *overcomplete* tensor decompositions. Secondly, the incoherence condition is reasonable as we argue in the following. Tensor decomposition using total mass minimization is an atomic decomposition problem. The latter determines the conditions under which a decomposition in terms of atoms in an atomic set $\mathcal{A}$ achieves the corresponding atomic norm. For example, the singular value decomposition is an atomic decomposition for the set of unit-norm, rank-one matrices. Finally, if the incoherence bound in Assumption I is further strengthened to $O(\frac{1}{n\alpha(\log n)})$ for some polynomial $\alpha(\cdot)$, then Assumptions II and III are consequences of Assumption I. So if the rank-one factors of an overcomplete tensor are incoherent enough, without needing Assumptions II and III, its CP decomposition can always be uniquely identified.

We note that Assumptions I, II and III hold with high probability if the tensor factors are generated independently according to uniform distributions on the unit spheres ([1], Lemmas 25, 31).

COROLLARY 2.1. If the true tensor factors $\{(\boldsymbol{u}_p^\star, \boldsymbol{v}_p^\star, \boldsymbol{w}_p^\star)\}_{p=1}^r$ in (1.1) are uniformly distributed on the unit spheres, and if r satisfies (2.4), then with high probability, the true tensor factors correspond to the unique optimal solution of the continuous analog of ℓ_1 norm minimization (1.4) up to a sign ambiguity.

3. Prior art and inspirations

Despite the advantages provided by tensor methods in many applications, their widespread adoption has been slow due to inherent computational intractability. Although the decomposition (1.1) is a multi-mode generalization of the singular value decomposition for matrices, extracting the decomposition from a given tensor is a nontrivial problem that is still under active investigation (cf. [18, 34]). Indeed, even determining the rank of a third-order tensor is an NP-hard problem [29]. A common strategy used to compute a tensor decomposition is to apply an alternating minimization scheme. Although efficient, this approach has the drawback of not providing global convergence guarantees [18]. Recently, an approach combining alternating minimization with power iteration has gained popularity due to its ability to guarantee the tensor decomposition results under certain assumptions [1, 33].

Tensor decomposition is a special case of atomic decomposition, which is to determine when a decomposition with respect to some given atomic set $\mathcal{A}$ achieves the atomic norm [15]. For finite atomic sets, it is now well-known that if the atoms satisfy certain conditions such as the restricted isometry property, then a sparse decomposition achieves the atomic norm [11]. For the set of rank-1, unit-norm matrices, the atomic norm (the matrix nuclear norm), is achieved by orthogonal decompositions [50]. When the atoms are complex sinusoids parameterized by the frequency, Candès and Fernandez-Granda

showed that atomic decomposition is solved by atoms with well-separated frequencies [12]. Similar separation conditions also show up when the atoms are translations of a known waveform [16, 21, 56], spherical harmonics [5] and radar signals parameterized by translations and modulations [28]. Tang and Shah in [57] employed the same atomic norm idea but focused on symmetric tensors. In addition, the result of [57] does not apply to overcomplete decompositions. Under a set of conditions, including the incoherence condition ensuring the separation of tensor factors, this work characterizes a class of *nonsymmetric* and *overcomplete* tensor decompositions that achieve the tensor nuclear norm $\|\mathcal{T}\|_*$.

Another closely related line of work is matrix recovery [19] and tensor recovery. Low-rank matrix recovery based on the idea of nuclear norm minimization has received a great deal of attention in recent years [13, 49, 50]. A direct generalization of this approach to tensors would have been using tensor nuclear norm to perform low-rank tensor recovery. However, this approach was not pursued due to the NP-hardness of computing the tensor nuclear norm [29] and the lack of analysis tools for tensor problems. The mainstream tensor recovery approaches are based on various forms of matricization [6, 26, 47]. Alternating minimization can also be applied to tensor recovery with performance guarantees established in recent work [32]. More recently, gradient descent with a good initialization is applied to the noisy symmetric tensor completion and achieves near-optimal statistical guarantees [10]. Note that all the above mentioned works study the low-rank tensor recovery problems, i.e. the number of rank-1 tensor factors is less than the factor size n. While in general calculating tensor decomposition is NP-hard, the theoretical computer science community has developed some interesting algorithms for overcomplete tensor decomposition. For example, Anandkumar *et al*. [1, 1, 2] apply the iterative power method with good initialization to the overcomplete tensor decomposition problem and provide guarantees for the linear-overcomplete case (i.e. $r \leq \beta n$). In addition to these local search algorithms such as gradient descent, power method and alternating minimization, another line of algorithms for overcomplete tensor decomposition are based on the sum-of-squares (SoS) semidefinite programming (SDP) hierarchy [30, 46, 48]. Although the SoS relaxation approaches provide provable guarantees for overcomplete tensor decomposition, they are essentially SDPs, which is not scalable to high-dimensional tensors.

In contrast, we expect that the atomic norm, when specialized to tensors, will achieve the information theoretical limit for tensor completion as it does for compressive sensing, matrix completion [19, 49] and line spectral estimation with missing data [54]. Given a set of atoms, the atomic norm is an abstraction of ℓ_1-type regularization that favors simple models. Using the notion of descent cones, Chandrasekaran *et al*. in [15] argued that the atomic norm is the best possible convex proxy for recovering simple models. Particularly, atomic norms are shown in many problems beyond compressive sensing and matrix completion to be able to recover simple models from minimal number of linear measurements. For example, when specialized to the atomic set formed by complex exponentials, the atomic norm can recover signals having sparse representations in the continuous frequency domain with the number of measurements approaching the information theoretic limit without noise [54] as well as achieving near minimax denoising performance [55]. Continuous frequency estimation using the atomic norm is also an instance of measure estimation from (trigonometric) moments.

4. Tensor decomposition, atomic norms and duality

In this work, we view tensor decomposition in the frameworks of both atomic norms and measure estimation. The unit sphere of $\mathbb{R}^n$ is denoted by $\mathbb{S}^{n-1}$ and the direct product of three unit spheres $\mathbb{S}^{n-1} \times \mathbb{S}^{n-1} \times \mathbb{S}^{n-1}$ by $\mathbb{K}$. The tensor atomic set is denoted by $\mathcal{A} = \{\boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} : (\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \in \mathbb{K}\}$ parameterized

by the set $\mathbb{K}$, where $\boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}$ is a rank-1 tensor with the (i,j,k)th entry being $u_i v_j w_k$. For any tensor $\mathcal{T}$, its atomic norm with respect to $\mathcal{A}$ is defined by ([15], Eq. (2))

$$\begin{aligned}\|\mathcal{T}\|_{\mathcal{A}} &= \inf\{t : \mathcal{T} \in t\,\mathrm{conv}(\mathcal{A})\} \\ &= \inf\left\{\sum_p \lambda_p : \mathcal{T} = \sum_p \lambda_p \boldsymbol{u}_p \otimes \boldsymbol{v}_p \otimes \boldsymbol{w}_p, \lambda_p > 0, (\boldsymbol{u}_p, \boldsymbol{v}_p, \boldsymbol{w}_p) \in \mathbb{K}\right\}, \end{aligned} \tag{4.1}$$

where $\mathrm{conv}(\mathcal{A})$ is the convex hull of the atomic set $\mathcal{A}$, and a scalar multiplying a set scales every element in the set. Therefore, the tensor atomic norm is the minimal ℓ_1 norm of its expansion coefficients among all valid expansions in terms of unit-norm, rank-1 tensors. The atomic norm $\|\mathcal{T}\|_{\mathcal{A}}$ defined in (4.1) is also called the tensor nuclear norm and denoted by $\|\mathcal{T}\|_*$ in ([25], Eq. (2.7)). We will use these two names and notations interchangeably in the following. The way of defining the tensor nuclear norm is precisely the same as that of defining the matrix nuclear norm.

We argue that the two lines in the definition (4.1) are consistent and are also equivalent to (1.4) as follows. Since $\mathrm{conv}(\mathcal{A}) = \{\mathcal{T} : \mathcal{T} = \int_{\mathbb{K}} \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}\, \mathrm{d}\,\mu, \mu \in \mathcal{M}_+(\mathbb{K}), \mu(\mathbb{K}) \le 1\}$, the first line in the definition (4.1) implies that $\|\mathcal{T}\|_{\mathcal{A}}$ is equal to the optimal value of (1.4). Compared with the measure optimization (1.4), the feasible region of the minimization defining the atomic norm in the second line of (4.1) is restricted to discrete measures. However, these two optimizations share the same optimal value as a consequence of Carathéodory's convex hull theorem, which states that if a point $\boldsymbol{x} \in \mathbb{R}^d$ lies in the convex hull of a set, then $\boldsymbol{x}$ can be written as a convex combination of at most $d+1$ points of that set ([4], Theorem 2.3). Since $\mathcal{T} \in \|\mathcal{T}\|_{\mathcal{A}}\,\mathrm{conv}(\mathcal{A}) = \mathrm{conv}(\|\mathcal{T}\|_{\mathcal{A}}\mathcal{A})$, $\mathcal{T}$ can be expressed as a convex combination of at most n^3+1 points of the set $\|\mathcal{T}\|_{\mathcal{A}}\mathcal{A}$, implying that the optimal value is achieved by a discrete measure with support size at most n^3+1. This argument establishes that the two lines in (4.1) as well as the measure optimization (1.4) are equivalent. Therefore, the atomic norm framework and the measure optimization framework are two different formulations of the same problem, with the former setting the stage in the finite dimensional space and the latter in the infinite-dimensional space of measures.

Given an abstract atomic set, the problem of atomic decomposition seeks the conditions under which a decomposition in terms of the given atoms achieves the atomic norm. In this sense, the tensor decomposition considered in this work is an atomic decomposition problem.

4.1 *Duality*

Duality plays an important role in analyzing atomic tensor decomposition. We again approach duality from both perspectives of atomic norms and measure estimation.

First, we find the dual problem of the optimization problem (1.4). Given $\mathcal{Q}, \mathcal{T} \in \mathbb{R}^{n\times n\times n}$, we define the tensor inner product $\langle \mathcal{Q}, \mathcal{T}\rangle := \sum_{i,j,k} Q_{ijk} T_{ijk}$. Standard Lagrangian analysis shows that the dual problem of (1.4) is the following semi-infinite program, which has an infinite number of constraints:

$$\begin{aligned} &\underset{\mathcal{Q}\in\mathbb{R}^{n\times n\times n}}{\text{maximize}}\ \langle \mathcal{Q}, \mathcal{T}\rangle \\ &\text{subject to}\ \langle \mathcal{Q}, \boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w}\rangle \le 1, \forall (\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}) \in \mathbb{K} \end{aligned} \tag{4.2}$$

The polynomial $q(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}) := \langle \mathcal{Q}, \boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w}\rangle = \sum_{i,j,k} Q_{ijk} u_i v_j w_k$ corresponding to a dual feasible solution $\mathcal{Q}$ of (4.2) is called a dual polynomial. The dual polynomial associated with an optimal dual

solution can be used to certify the optimality of a particular decomposition, as demonstrated by the following proposition.

PROPOSITION 4.1. Suppose the set of rank-1 tensors $\{\boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star\}_{p=1}^r$ given in (1.1) are linearly independent. If there exists a dual solution $\mathcal{Q} \in \mathbb{R}^{n\times n\times n}$ to (4.2) such that the corresponding dual polynomial $q : \mathbb{K} \to \mathbb{R}$

$$q(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) := \langle \mathcal{Q}, \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}\rangle \tag{4.3}$$

satisfies the following *Boundedness and Interpolation Property (BIP):*

$$q(\boldsymbol{u}_p^\star, \boldsymbol{v}_p^\star, \boldsymbol{w}_p^\star) = 1 \text{ for } p \in [r] \text{ (Interpolation)} \tag{4.4a}$$

$$q(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) < 1 \text{ in } \mathbb{K} \setminus S^\star \text{ (Boundedness)} \tag{4.4b}$$

where $[r] := \{1, \ldots, r\}$ and

$$S^\star := \{(a_p\boldsymbol{u}_p^\star, b_p\boldsymbol{v}_p^\star, c_p\boldsymbol{w}_p^\star) : |a_p| = |b_p| = |c_p| = a_pb_pc_p = 1, p \in [r]\}, \tag{4.5}$$

then $\mu^\star$ given in (1.3) is the unique optimal solution to (1.4) up to sign ambiguity.

Proof. In view of (4.2), any $\mathcal{Q}$ that satisfies the BIP in (4.4) is a dual feasible solution. We also have

$$\langle \mathcal{Q}, \mathcal{T}\rangle = \left\langle \mathcal{Q}, \sum_{p=1}^r \lambda_p^\star \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star \right\rangle = \sum_{p=1}^r \lambda_p^\star \langle \mathcal{Q}, \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star\rangle = \sum_{p=1}^r \lambda_p^\star q(\boldsymbol{u}_p^\star, \boldsymbol{v}_p^\star, \boldsymbol{w}_p^\star) = \mu^\star(\mathbb{K})$$

establishing a zero-duality gap of the primal-dual feasible solution $(\mu^\star, \mathcal{Q})$. As a consequence, $\mu^\star$ is a primal optimal solution to (1.4) and $\mathcal{Q}$ is a dual optimal solution to (4.2).

For uniqueness, suppose $\hat{\mu}$ is another primal optimal solution to (1.4). If $\hat{\mu}(\mathbb{K} \setminus S^\star) > 0$, then

$$\mu^\star(\mathbb{K}) = \langle \mathcal{Q}, \mathcal{T}\rangle = \left\langle \mathcal{Q}, \int_{\mathbb{K}} \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} \, \mathrm{d}\, \hat{\mu} \right\rangle < \hat{\mu}(S^\star) + \int_{\mathbb{K}\setminus S^\star} 1 \, \mathrm{d}\, \hat{\mu} = \hat{\mu}(\mathbb{K})$$

contradicting the optimality of $\hat{\mu}$. So all optimal solutions are supported on $S^\star$. To remove the sign ambiguity, we can assume an optimal solution is supported on $\{\boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star\}_{p=1}^r$. Since $\{\boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star\}_{p=1}^r$ are linearly independent by assumption, the coefficients $\lambda_p^\star$ can be uniquely determined from solving the linear system of equations encoded in $T = \sum_{p=1}^r \lambda_p^\star \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star$. This proves the uniqueness (up to sign ambiguity). □

4.2 *Dual certificate and subdifferential*

The dual optimal solution $\mathcal{Q}$ satisfying the BIP is called a *dual certificate*, which is used frequently as the starting point to derive several atomic decomposition and super-resolution results [5, 12, 54, 57]. In Section 5, we will explicitly construct a *dual certificate* to prove Theorem 2.1. In this subsection, we will relate the *dual certificate* with the subdifferential of the tensor nuclear norm.

First, the dual norm of the tensor nuclear norm, i.e. the tensor spectral norm, of a tensor $\mathcal{Q}$ is given by

$$\|\mathcal{Q}\| := \sup_{\mathcal{T}:\|\mathcal{T}\|_*\leq 1} \langle \mathcal{Q}, \mathcal{T}\rangle = \sup_{(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})\in\mathbb{K}} \langle \mathcal{Q}, \boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w}\rangle. \tag{4.6}$$

The equality is due to the fact that the atomic set $\mathcal{A}$ are the extreme points of the unit nuclear norm ball $\{\mathcal{T} : \|\mathcal{T}\|_* \leq 1\}$. In light of the spectral norm definition, we rewrite the dual problem (4.2) as

$$\underset{\mathcal{Q}\in\mathbb{R}^{n\times n\times n}}{\text{maximize}}\ \langle \mathcal{Q}, \mathcal{T}\rangle \text{ subject to } \|\mathcal{Q}\| \leq 1, \tag{4.7}$$

which is precisely the definition of the dual norm of the tensor spectral norm, i.e. the tensor nuclear norm.

The subdifferential (the set of subgradients) of the tensor nuclear norm is defined by ([24], Definition B.20)

$$\partial\|\cdot\|_*(\mathcal{T}) = \{\mathcal{Q}\in\mathbb{R}^{n\times n\times n} : \|\mathcal{R}\|_* \geq \|\mathcal{T}\|_* + \langle \mathcal{R}-\mathcal{T}, \mathcal{Q}\rangle, \text{for all}\mathcal{R}\in\mathbb{R}^{n\times n\times n}\}, \tag{4.8}$$

which has an equivalent representation ([59], Section 1)

$$\partial\|\cdot\|_*(\mathcal{T}) = \left\{\mathcal{Q}\in\mathbb{R}^{n\times n\times n} : \|\mathcal{T}\|_* = \langle \mathcal{Q}, \mathcal{T}\rangle, \|\mathcal{Q}\| \leq 1\right\}. \tag{4.9}$$

For $\mathcal{T}$ having an atomic decomposition given in (1.1), it can be established that the defining properties of subdifferential (4.9) are equivalent to

$$\langle \mathcal{Q}, \boldsymbol{u}_p^\star\otimes\boldsymbol{v}_p^\star\otimes\boldsymbol{w}_p^\star\rangle = 1, \text{ for } p\in[r] \tag{4.10a}$$

$$\langle \mathcal{Q}, \boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w}\rangle \leq 1, \text{ for } (\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})\in\mathbb{K} \tag{4.10b}$$

We recognize that the BIP (4.4) is a strengthened version of the subdifferential conditions (4.10). Therefore, a *dual certificate*, i.e. any $\mathcal{Q}$ satisfying the BIP, is an element of the subdifferential $\partial\|\cdot\|_*(\mathcal{T})$. The BIP in fact means that $\mathcal{Q}$ is an interior point of $\partial\|\cdot\|_*(\mathcal{T})$. Our proof strategy for Theorem 2.1 is to construct such an interior point in Section 5. This is in contrast to the matrix case, for which we have an explicit characterization of the entire subdifferential of the nuclear norm using the singular value decomposition (more explicit than the one given in (4.9)). More specifically, suppose $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ is the (compact) singular value decomposition of $\mathbf{X}\in\mathbb{R}^{m\times n}$ with $\mathbf{U}\in\mathbb{R}^{m\times r}, \mathbf{V}\in\mathbb{R}^{n\times r}$ and $\boldsymbol{\Sigma}$ being an $r\times r$ diagonal matrix. Then the subdifferential of the matrix nuclear norm at $\mathbf{X}$ is given by ([50], Eq. (2.9))

$$\partial\|\cdot\|_*(\mathbf{X}) = \{\mathbf{U}\mathbf{V}^\top + \mathbf{W} : \mathbf{U}^\top\mathbf{W} = \mathbf{0}, \mathbf{W}\mathbf{V} = \mathbf{0}, \|\mathbf{W}\| \leq 1\}.$$

It is challenging to obtain such a characterization for tensors unless the tensor admits an orthogonal decomposition.

4.3 *Extension: regularization using tensor nuclear norm*

Independent from practical considerations, we investigate tensor decomposition for theoretical reasons. Similar to regularizing matrix inverse problems using the matrix nuclear norm, the tensor nuclear norm can be used to regularize tensor inverse problems. Suppose we observe an unknown low-rank tensor

$\mathcal{T}^\star$ through the linear measurement model $\boldsymbol{y} = \mathcal{B}(\mathcal{T}^\star)$, we would like to recover the tensor $\mathcal{T}^\star$ from the observation $\boldsymbol{y}$. For instance, when $\mathcal{B}$ samples the individual entries of $\mathcal{T}^\star$, we are looking at a tensor completion problem. Remarkably, Yuan and Zhang exploited the tensor nuclear norm approach to tensor completion and improved the state-of-the-art sample complexity in the seminal work [60]. We propose recovering $\mathcal{T}^\star$ by solving

$$\underset{\mathcal{T}\in\mathbb{R}^{n\times n\times n}}{\text{minimize}} \|\mathcal{T}\|_* \ \text{ subject to } \boldsymbol{y} = \mathcal{B}(\mathcal{T}) \tag{4.11}$$

which favors a low-rank solution. To establish recoverability, we can construct a *dual certificate* $\mathcal{Q}$ of the form $\mathcal{B}^*(\boldsymbol{\lambda})$, whose corresponding dual polynomial satisfies the BIP. Here $\mathcal{B}^*$ is the adjoint operator of $\mathcal{B}$. When the operator $\mathcal{B}$ is random, the concentration of measure guarantees that we can construct a *dual certificate* $\mathcal{B}^*(\boldsymbol{\lambda})$ that is close to the one constructed in the full data case. This fact can then be exploited to verify the BIP of $\mathcal{B}^*(\boldsymbol{\lambda})$ and to establish exact recovery. When the atoms are complex exponentials parameterized by continuous frequencies, this strategy is adopted to establish the compressed sensing off the grid result (the completion problem) [54] building upon the dual polynomial constructed for the super-resolution problem (the full data case) [12]. It shows that the number of random linear measurements required for exact recovery approaches the information theoretical limit. In addition to exact recovery from noise-free measurements, the *dual certificate* for the full data case can also be utilized to derive near-minimax denoising performance [7, 55], approximate support recovery [22, 38] and robust recovery from observations corrupted by outliers [23, 58]. We expect that the dual polynomial constructed for tensor decomposition will play a similar role for tensor inverse problems, enabling the development of tensor results parallel to their matrix counterparts such as matrix completion, denoising and robust principal component analysis. We leave these as our future work.

5. Proof of Theorem 2.1

5.1 *Proof outline*

The proof of Theorem 2.1 relies on the construction of a dual polynomial that satisfies the BIP (4.4). Towards that end, we first partition $\mathbb{K}$ into the far region (controlled by Lemma 5.4) and the near region. To control the dual polynomial in the near region, we use an angular parametrization to further divide it into near vertex region (controlled by Lemma 5.6) and near band region (controlled by Lemma 5.7). In the end, we can show the constructed dual polynomial satisfies the BIP in the whole region. We summarize the proof map on the right.

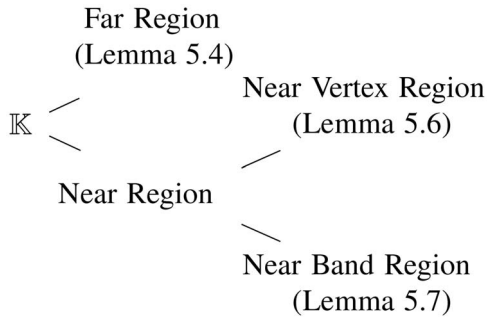
5.2 *Minimal energy construction*

Since the BIP (4.4) (especially the Boundedness property (4.4b)) is hard to enforce directly, we start from a candidate *dual certificate* or pre-certificate $\mathcal{Q}$ in the subdifferntial set $\partial\|\mathcal{T}\|_*$ defined by (4.10):

$$\langle \mathcal{Q}, \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star \rangle = 1, \text{ for } p \in [r]$$
$$\langle \mathcal{Q}, \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} \rangle \le 1, \text{ for } (\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \in \mathbb{K}$$

which essentially characterizes the optimal solution set of following optimization

$$\text{maximize}_{(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})\in\mathbb{K}} \ \langle \mathcal{Q}, \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} \rangle \tag{5.1}$$



Then applying the Karush–Kuhn–Tucker (KKT) conditions to the constrained optimization (5.1), we can further relax the subdifferential conditions (4.10) to a set of linear constraints.

LEMMA 5.1. The following conditions are necessary for (4.10):

$$\begin{aligned}
&\sum_{j,k} Q_{ijk}\boldsymbol{v}_p^\star(j)\boldsymbol{w}_p^\star(k) = \boldsymbol{u}_p^\star(i), \forall i \in [n], \forall p \in [r];\\
&\sum_{i,k} Q_{ijk}\boldsymbol{u}_p^\star(i)\boldsymbol{w}_p^\star(k) = \boldsymbol{v}_p^\star(j), \forall i \in [n], \forall p \in [r];\\
&\sum_{i,j} Q_{ijk}\boldsymbol{u}_p^\star(i)\boldsymbol{v}_p^\star(j) = \boldsymbol{w}_p^\star(k), \forall i \in [n], \forall p \in [r]
\end{aligned} \quad \overset{or}{\Longleftrightarrow} \quad \begin{aligned}
&\mathcal{Q}\times_2\boldsymbol{v}_p^\star\times_3\boldsymbol{w}_p^\star = \boldsymbol{u}_p^\star, \forall p \in [r];\\
&\mathcal{Q}\times_1\boldsymbol{u}_p^\star\times_3\boldsymbol{w}_p^\star = \boldsymbol{v}_p^\star, \forall p \in [r];\\
&\mathcal{Q}\times_1\boldsymbol{u}_p^\star\times_2\boldsymbol{v}_p^\star = \boldsymbol{w}_p^\star, \forall p \in [r]
\end{aligned} \tag{5.2}$$

where $\{\times_k\}$ are the k-mode tensor-vector product [34] whose definitions are apparent from context.

The proof of Lemma 5.1 is given in Appendix A.

Apparently, the subdifferential conditions (4.10) is necessary for the BIP (4.4), but generally not sufficient, by comparing the second line of (4.10) and the Boundedness Property (4.4b). Indeed, as we argued before, any $\mathcal{Q}$ satisfying the BIP is an interior point of the subdifferential $\partial\|\cdot\|_*(\mathcal{T})$. To satisfy the Boundedness Property (4.4b), we further minimize the energy $\|\mathcal{Q}\|_F^2 = \sum_{ijk} Q_{ijk}^2$ in the hope that this will push $q(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})$ towards zero such that $\mathcal{Q}$ is an interior point of $\partial\|\cdot\|_*(\mathcal{T})$. Thus, we propose solving the following *minimum-energy* problem to obtain a pre-certificate:

$$\underset{Q}{\text{minimize}} \ \frac{1}{2}\|\mathcal{Q}\|_F^2 \ \text{ subject to (5.2)} \tag{5.3}$$

LEMMA 5.2. (Explicit form of the pre-certificate) The solution of the least-norm problem (5.3) has the form (normal equation)

$$\mathcal{Q} = \sum_{p=1}^{r}(\boldsymbol{\alpha}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star + \boldsymbol{u}_p^\star \otimes \boldsymbol{\beta}_p^\star \otimes \boldsymbol{w}_p^\star + \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{\gamma}_p^\star) \tag{5.4}$$

with the unknown coefficients $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$ being chosen such that $\mathcal{Q}$ in (5.4) satisfies (5.2). So we get an explicit form of a pre-certificate

$$q(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \langle \mathcal{Q}, \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} \rangle$$
$$= \sum_{p=1}^{r} [\langle \boldsymbol{\alpha}_p^\star, \boldsymbol{u} \rangle \langle \boldsymbol{v}_p^\star, \boldsymbol{v} \rangle \langle \boldsymbol{w}_p^\star, \boldsymbol{w} \rangle + \langle \boldsymbol{u}_p^\star, \boldsymbol{u} \rangle \langle \boldsymbol{\beta}_p^\star, \boldsymbol{v} \rangle \langle \boldsymbol{w}_p^\star, \boldsymbol{w} \rangle + \langle \boldsymbol{u}_p^\star, \boldsymbol{u} \rangle \langle \boldsymbol{v}_p^\star, \boldsymbol{v} \rangle \langle \boldsymbol{\gamma}_p^\star, \boldsymbol{w} \rangle]. \tag{5.5}$$

The proof of Lemma 5.2 is given in Appendix B.

To obtain some intuition of what these dual-polynomial coefficients $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$ would look like, let us assume $\{\boldsymbol{u}_p^\star\}_{p=1}^r$, $\{\boldsymbol{v}_p^\star\}_{p=1}^r$, $\{\boldsymbol{w}_p^\star\}_{p=1}^r$ are almost orthogonal and plug the explicit form of $\mathcal{Q}$ (5.4) into the first equation in (5.2)

$$\boldsymbol{\alpha}_p^\star + \boldsymbol{u}_p^\star \langle \boldsymbol{\beta}_p^\star, \boldsymbol{v}_p^\star \rangle + \boldsymbol{u}_p^\star \langle \boldsymbol{\gamma}_p^\star, \boldsymbol{w}_p^\star \rangle \approx \boldsymbol{u}_p^\star. \tag{5.6}$$

Then multiplying $\boldsymbol{u}_p^{\star\top}$ on both sides gives

$$\langle \boldsymbol{\alpha}_p^\star, \boldsymbol{u}_p^\star \rangle + \langle \boldsymbol{\beta}_p^\star, \boldsymbol{v}_p^\star \rangle + \langle \boldsymbol{\gamma}_p^\star, \boldsymbol{w}_p^\star \rangle \approx 1. \tag{5.7}$$

Finally combining (5.6) and (5.7) together with the symmetry property of (5.4), we get these coefficients $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$ are located approximately at $\{\boldsymbol{u}_p^\star/3, \boldsymbol{v}_p^\star/3, \boldsymbol{w}_p^\star/3\}_{p=1}^r$. The accurate description of this phenomenon is given by the following lemma with the proof listed in Appendix C.

LEMMA 5.3 (Control the dual polynomial coefficients). Under Assumptions II and III together with $r = o(n^2/\kappa(\log n)^2)$, the following estimates are valid for sufficiently large n:

$$\left\| \mathbf{A} - \frac{1}{3}\mathbf{U} \right\| \le 2\kappa(\log n) \left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}} \right);$$
$$\left\| \mathbf{B} - \frac{1}{3}\mathbf{V} \right\| \le 2\kappa(\log n) \left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}} \right);$$
$$\left\| \mathbf{C} - \frac{1}{3}\mathbf{W} \right\| \le 2\kappa(\log n) \left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}} \right)$$

where the norm $\|\cdot\|$ is the matrix spectral norm and

$$\mathbf{A} = \left[\boldsymbol{\alpha}_1^\star, \cdots, \boldsymbol{\alpha}_r^\star\right], \mathbf{B} = \left[\boldsymbol{\beta}_1^\star, \cdots, \boldsymbol{\beta}_r^\star\right], \mathbf{C} = \left[\boldsymbol{\gamma}_1^\star, \cdots, \boldsymbol{\gamma}_r^\star\right], \mathbf{U} = \left[\boldsymbol{u}_1^\star, \cdots, \boldsymbol{u}_r^\star\right],$$
$$\mathbf{V} = \left[\boldsymbol{v}_1^\star, \cdots, \boldsymbol{v}_r^\star\right], \mathbf{W} = \left[\boldsymbol{w}_1^\star, \cdots, \boldsymbol{w}_r^\star\right].$$

5.3 *Far region*

For a parameter $\delta \in (0, 1)$, the far region is defined by

$$\mathcal{F}(\delta) := \bigcap_{p=1}^{r} \{(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \in \mathbb{K} : |\langle \boldsymbol{u}, \boldsymbol{u}_p^\star \rangle| \le \delta \text{ or} |\langle \boldsymbol{v}, \boldsymbol{v}_p^\star \rangle| \le \delta \text{ or} |\langle \boldsymbol{w}, \boldsymbol{w}_p^\star \rangle| \le \delta\}, \tag{5.8}$$

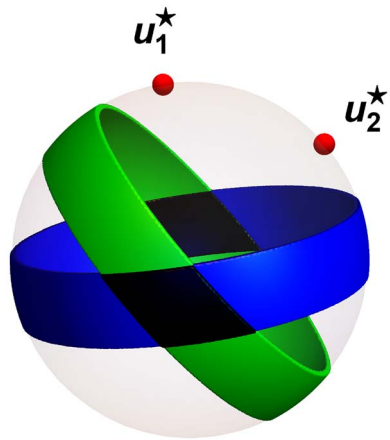


FIG. 1. Projection of the far region in the $\boldsymbol{u}$ coordinate. The blue band represents the region $\{\boldsymbol{u} : |\langle \boldsymbol{u}, \boldsymbol{u}_1^\star\rangle| \le \delta\}$ that is far away from $\boldsymbol{u}_1^\star$, while the green region $\{\boldsymbol{u} : |\langle \boldsymbol{u}, \boldsymbol{u}_2^\star\rangle| \le \delta\}$ is the far-region associated with $\boldsymbol{u}_2^\star$. The far region is their intersection $\bigcap_{p=1}^2 \{\boldsymbol{u} : |\langle \boldsymbol{u}, \boldsymbol{u}_p^\star\rangle| \le \delta\}$, consisting of the two black diamonds.

which consists of points $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ in $\mathbb{K}$ that are far away (in the angular sense) from

$$\mathbb{S}^\star := \{(\pm \boldsymbol{u}_p^\star, \pm \boldsymbol{v}_p^\star, \pm \boldsymbol{w}_p^\star) : p = 1, \dots, r\} \tag{5.9}$$

in at least one coordinate of $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$. For $n = 3$ and $r = 2$, the far region projected onto the unit sphere $\{\boldsymbol{u} : \|\boldsymbol{u}\|_2 = 1\}$ is shown in Figure 1.

5.3.1 *Controlling in far region* Instead of bounding the dual polynomial q directly, we will bound its absolute value $|q|$. To obtain some intuition of how to bound it, we rewrite the explicit form (5.5) as follows

$$\begin{aligned} &q(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \\ &= \sum_{p=1}^r \left[\langle \boldsymbol{\alpha}_p^\star - \frac{1}{3}\boldsymbol{u}_p^\star, \boldsymbol{u}\rangle \langle \boldsymbol{v}_p^\star, \boldsymbol{v}\rangle \langle \boldsymbol{w}_p^\star, \boldsymbol{w}\rangle + \langle \boldsymbol{u}_p^\star, \boldsymbol{u}\rangle \langle \boldsymbol{\beta}_p^\star - \frac{1}{3}\boldsymbol{v}_p^\star, \boldsymbol{v}\rangle \langle \boldsymbol{w}_p^\star, \boldsymbol{w}\rangle + \langle \boldsymbol{u}_p^\star, \boldsymbol{u}\rangle \langle \boldsymbol{v}_p^\star, \boldsymbol{v}\rangle \langle \boldsymbol{\gamma}_p^\star - \frac{1}{3}\boldsymbol{w}_p^\star, \boldsymbol{w}\rangle \right] \end{aligned} \tag{5.10}$$

$$+ \sum_{p=1}^r \langle \boldsymbol{u}_p^\star, \boldsymbol{u}\rangle \langle \boldsymbol{v}_p^\star, \boldsymbol{v}\rangle \langle \boldsymbol{w}_p^\star, \boldsymbol{w}\rangle. \tag{5.11}$$

The main idea is first using the closeness of $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$ and $\{\boldsymbol{u}_p^\star/3, \boldsymbol{v}_p^\star/3, \boldsymbol{w}_p^\star/3\}_{p=1}^r$ to bound (5.10) and then using angular-distance between $\mathcal{F}(\delta)$ and $(\boldsymbol{u}_p^\star, \boldsymbol{v}_p^\star, \boldsymbol{w}_p^\star), \forall p$ to bound (5.11).

The accurate argument is made by the following lemma with the proof given in Appendix D.

LEMMA 5.4 (Controlling in far region). Under Assumptions I, II, III, if $r \ll n^{1.25}$ and $r \le \frac{n}{24\delta c^2}$ for $\delta \in (0, \frac{1}{24}]$, then for sufficiently large n, we have $|q(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})| < 1$ in $\mathcal{F}(\delta)$.

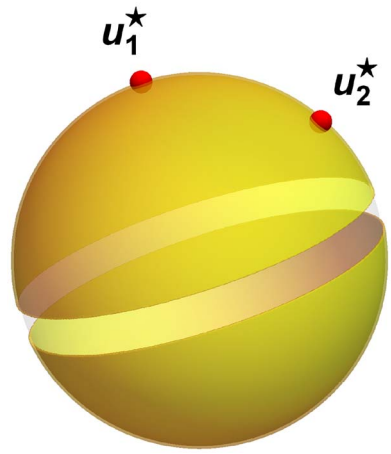


FIG. 2. The two yellow spherical caps form the near region $\mathcal{N}_1(\delta)$ around the point $(\boldsymbol{u}_1^\star, \boldsymbol{v}_1^\star, \boldsymbol{w}_1^\star)$ projected onto the $\boldsymbol{u}$ coordinates. $\mathcal{N}_2(\delta)$, which is not shown here, consists of another two spherical caps. The union of $\mathcal{N}_1(\delta), \mathcal{N}_2(\delta)$ and the far region $\mathcal{F}(d)$ shown in Figure 1 will cover the entire sphere $\{\boldsymbol{u} : \|\boldsymbol{u}\| = 1\}$.

5.4 *Near region*

For the union of the far and near regions to cover the entire region $\mathbb{K}$, we define the near region as

$$\mathcal{N}(\delta) := \mathbb{K} \setminus \mathcal{F}(\delta) = \bigcup_{p=1}^{r} \{(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \in \mathbb{K} : |\langle \boldsymbol{u}_p^\star, \boldsymbol{u}\rangle| \geq \delta, |\langle \boldsymbol{v}_p^\star, \boldsymbol{v}\rangle| \geq \delta, |\langle \boldsymbol{w}_p^\star, \boldsymbol{w}\rangle| \geq \delta\} := \bigcup_{p=1}^{r} \mathcal{N}_p(\delta) \tag{5.12}$$

with each individual near region $\mathcal{N}_p(\delta)$ close to $(\boldsymbol{u}_p^\star, \boldsymbol{v}_p^\star, \boldsymbol{w}_p^\star)$ in all coordinate of $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$.

For $n = 3$, $r = 2$, we plot the near region $\mathcal{N}_1(\delta)$ projected onto the sphere $\{\boldsymbol{u} : \|\boldsymbol{u}\|_2 = 1\}$ in Figure 2.

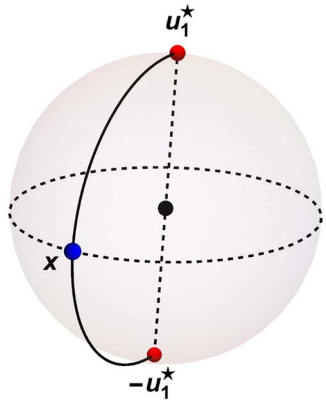
5.4.1 *Angular parametrization of near region* In order to show the dual polynomial satisfying the BIP in the entire near region $\mathcal{N}(\delta)$, we use the 'Divide-and-conquer' idea to bound the dual polynomial in each individual near region $\mathcal{N}_p(\delta)$ for $p \in [r]$. The main technique used to control each individual near region is applying angular parametrization to each individual near region.

As the domain $\mathbb{K}$ is essentially a direct product of spheres, we re-parameterize each individual near region $\mathcal{N}_p(\delta)$ in the angular sense. Without loss of generality, let us consider $p = 1$. Pick $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in \mathbb{K}$ such that $\boldsymbol{x} \perp \boldsymbol{u}_1^\star, \boldsymbol{y} \perp \boldsymbol{v}_1^\star, \boldsymbol{z} \perp \boldsymbol{w}_1^\star$ and consider the parameterized points

$$(\boldsymbol{u}(\theta_1), \boldsymbol{v}(\theta_2), \boldsymbol{w}(\theta_3)) \in \mathbb{K} \quad \text{with} \quad \begin{cases} \boldsymbol{u}(\theta_1) = \boldsymbol{u}_1^\star \cos(\theta_1) + \boldsymbol{x} \sin(\theta_1) \\ \boldsymbol{v}(\theta_2) = \boldsymbol{v}_1^\star \cos(\theta_2) + \boldsymbol{y} \sin(\theta_2) \\ \boldsymbol{w}(\theta_3) = \boldsymbol{w}_1^\star \cos(\theta_3) + \boldsymbol{z} \sin(\theta_3). \end{cases} \tag{5.13}$$

When θ_1 ranges from 0 to π, $\boldsymbol{u}(\theta_1)$ traces out a 2D semi-circle that starts at $\boldsymbol{u}_1^\star$, passes through $\boldsymbol{x}$, and finally reaches $-\boldsymbol{u}_1^\star$; while for a fixed $\theta_1 \in [0, \pi]$, the set $\bigcup_{\boldsymbol{x} \perp \boldsymbol{u}_1^\star} \{\boldsymbol{u}(\theta_1)\}$ parameterizes all the points on $\mathbb{S}^{n-1}$ having an angle of θ_1 with $\boldsymbol{u}_1^\star$. The same properties hold for $\boldsymbol{v}(\theta_2)$ and $\boldsymbol{w}(\theta_3)$. This parametrization

projected onto the $\boldsymbol{u}$ coordinate is shown on right.



In fact, using this angular parametrization, the individual near region $\mathcal{N}_1(\delta)$ in (5.12) can be expressed as

$$\mathcal{N}_1(\delta) = \bigcup_{(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}):\boldsymbol{x}\perp\boldsymbol{u}_1^\star,\boldsymbol{y}\perp\boldsymbol{v}_1^\star,\boldsymbol{z}\perp\boldsymbol{w}_1^\star} \{(\boldsymbol{u}(\theta_1),\boldsymbol{v}(\theta_2),\boldsymbol{w}(\theta_3)) : |\cos(\theta_i)| \geq \delta, \theta_i \in [0,\pi], i = 1,2,3\}. \tag{5.14}$$

Proposition 5.1 (Near angular region). For any $\delta \in (0,1)$, the near region $\mathcal{N}_1(\delta)$ is contained in the following set

$$\mathcal{N}_1(\delta) \subset \bigcup_{(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}):\boldsymbol{x}\perp\boldsymbol{u}_1^\star,\boldsymbol{y}\perp\boldsymbol{v}_1^\star,\boldsymbol{z}\perp\boldsymbol{w}_1^\star} \{(\boldsymbol{u}(\theta_1),\boldsymbol{v}(\theta_2),\boldsymbol{w}(\theta_3)) : (\theta_1,\theta_2,\theta_3) \in \mathbb{N}(\delta)\} \tag{5.15}$$

with the near angular region $\mathbb{N}(\delta)$ defined by

$$\mathbb{N}(\delta) := \left\{(\theta_1,\theta_2,\theta_3) : \theta_i \in \left[0, \frac{\pi}{2} - \delta\right] \cup \left[\frac{\pi}{2} + \delta, \pi\right], i = 1,2,3\right\}. \tag{5.16}$$

Proof. Since the function $|\cos(\theta)|$ is symmetric at $\frac{\pi}{2}$ on the interval $[0,\pi]$ and is decreasing on $[0,\pi/2]$, we know that $\{\theta : |\cos(\theta)| \geq \delta\} \cap [0,\pi] = [0, \arccos(\delta)] \cup [\pi - \arccos(\delta), \pi]$. Note that $\arccos(\delta) = \frac{\pi}{2} - \arcsin(\delta)$ and $\delta < \arcsin(\delta)$, so we get $\{\theta : |\cos(\theta)| \geq \delta\} \cap [0,\pi] \subset [0, \frac{\pi}{2} - \delta] \cup [\frac{\pi}{2} + \delta, \pi]$. The inclusion (5.15) follows from (5.14) immediately. □

The near angular region $\mathbb{N}(\delta)$ contains the eight cubes with side length $\frac{\pi}{2} - \delta$, located at the eight corners of the cube $[0,\pi] \times [0,\pi] \times [0,\pi]$. Moreover, one can see that the smaller the parameter δ is, the larger the near angular region $\mathbb{N}(\delta)$ will be. In particular, when δ approaches to zero, the near angular region $\mathbb{N}(\delta)$ becomes the whole cube $\mathbb{N}(0) = [0,\pi] \times [0,\pi] \times [0,\pi]$. The near angular region $\mathbb{N}(\delta)$ is plotted in Figure 3.

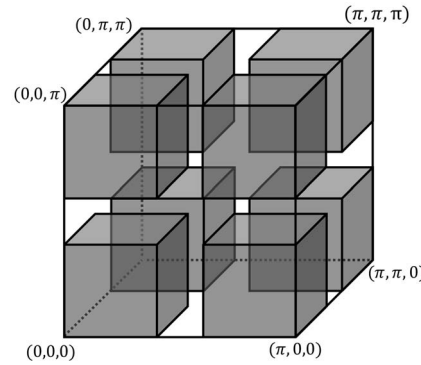


FIG. 3. The eight gray cubes of side-length $\pi/2-\delta$ at the corners form the near angular region $\mathbb{N}(\delta)$.

5.4.2 *Angular parametrization of dual polynomial* Evaluating the dual polynomial $q(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})$ at $(\boldsymbol{u}(\theta_1),\boldsymbol{v}(\theta_2),\boldsymbol{w}(\theta_3))$ in (5.13), we get the angular dual polynomial $F(\theta_1,\theta_2,\theta_3) := q(\boldsymbol{u}(\theta_1),\boldsymbol{v}(\theta_2),\boldsymbol{w}(\theta_3))$ as

$$\begin{aligned} F(\theta_1,\theta_2,\theta_3) =& q(\boldsymbol{u}_1^\star,\boldsymbol{v}_1^\star,\boldsymbol{w}_1^\star)\cos(\theta_1)\cos(\theta_2)\cos(\theta_3) + q(\boldsymbol{u}_1^\star,\boldsymbol{v}_1^\star,\boldsymbol{z})\cos(\theta_1)\cos(\theta_2)\sin(\theta_3)\\ &+ q(\boldsymbol{u}_1^\star,\boldsymbol{y},\boldsymbol{w}_1^\star)\cos(\theta_1)\sin(\theta_2)\cos(\theta_3) + q(\boldsymbol{x},\boldsymbol{v}_1^\star,\boldsymbol{w}_1^\star)\sin(\theta_1)\cos(\theta_2)\cos(\theta_3)\\ &+ q(\boldsymbol{u}_1^\star,\boldsymbol{y},\boldsymbol{z})\cos(\theta_1)\sin(\theta_2)\sin(\theta_3) + q(\boldsymbol{x},\boldsymbol{v}_1^\star,\boldsymbol{z})\sin(\theta_1)\cos(\theta_2)\sin(\theta_3)\\ &+ q(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}_1^\star)\sin(\theta_1)\sin(\theta_2)\cos(\theta_3) + q(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\sin(\theta_1)\sin(\theta_2)\sin(\theta_3). \end{aligned} \tag{5.17}$$

Among these eight terms, the first term is $\cos(\theta_1)\cos(\theta_2)\cos(\theta_3)$ since $q(\boldsymbol{u}_1^\star,\boldsymbol{v}_1^\star,\boldsymbol{w}_1^\star)=1$. The next three terms involving one sine function are zero as, for example,

$$q(\boldsymbol{u}_1^\star,\boldsymbol{v}_1^\star,\boldsymbol{z}) = \mathcal{Q}\times_1\boldsymbol{u}_1^\star\times_2\boldsymbol{v}_1^\star\times_3\boldsymbol{z} = \boldsymbol{w}_1^\star\times_3\boldsymbol{z} = {\boldsymbol{w}_1^\star}^\top\boldsymbol{z} = 0,$$

where we have used $Q\times_1\boldsymbol{u}_1^\star\times_2\boldsymbol{v}_1^\star = \boldsymbol{w}_1^\star$ and the third equality of (5.2). Hence, we get a more concise form of F:

$$\begin{aligned} F(\theta_1,\theta_2,\theta_3) =& \cos(\theta_1)\cos(\theta_2)\cos(\theta_3) + q(\boldsymbol{u}_1^\star,\boldsymbol{y},\boldsymbol{z})\cos(\theta_1)\sin(\theta_2)\sin(\theta_3) + q(\boldsymbol{x},\boldsymbol{v}_1^\star,\boldsymbol{z})\sin(\theta_1)\cos(\theta_2)\\ &\sin(\theta_3) + q(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}_1^\star)\sin(\theta_1)\sin(\theta_2)\cos(\theta_3) + q(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\sin(\theta_1)\sin(\theta_2)\sin(\theta_3). \end{aligned} \tag{5.18}$$

By further bounding the other quantities $q(\boldsymbol{u}_1^\star,\boldsymbol{y},\boldsymbol{z})$, $q(\boldsymbol{x},\boldsymbol{v}_1^\star,\boldsymbol{z})$, $q(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}_1^\star)$ and $q(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$, we get the following lemma to uniformly upper-bound $F(\theta_1,\theta_2,\theta_3)$ with the proof given in Appendix E.

LEMMA 5.5 (Upper bound of angular dual polynomial). Under Assumptions I, II, III, if $r \le n^{1.25-1.5r_c}$ with $r_c \in (0,\frac{1}{6})$, then for sufficiently large n, we have

$$|F(\theta_1,\theta_2,\theta_3)| \le |\cos(\theta_1)\cos(\theta_2)\cos(\theta_3)| + |\sin(\theta_1)\sin(\theta_2)\sin(\theta_3)| + \frac{4}{3}\tau(\log n)n^{-r_c}. \tag{5.19}$$

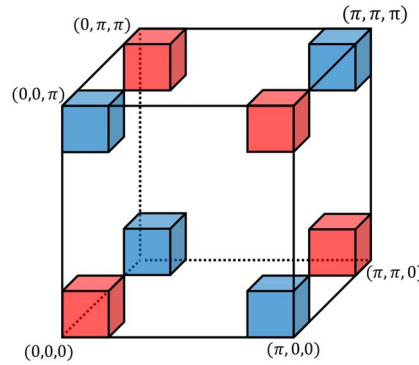


FIG. 4. The eight colored cubes of size $\delta_v \times \delta_v \times \delta_v$ form the near vertex region $\mathbb{N}_v(\delta_v)$: the red ones are corresponding to the vertexes in $\mathbb{S}^\star$, whereas the blue ones are corresponding to other vertexes in the cube.

5.4.3 *Angular parametrization of Boundedness and Interpolation Property* By Proposition 5.1, a sufficient condition for the BIP (4.4) to hold in the individual near region $\mathcal{N}_1(\delta)$, is the following Angular BIP (Angular-BIP):

$$F(\theta_1,\theta_2,\theta_3) = 1 \text{ in} \mathbb{S}^\star \quad \text{(Angular Interpolation)} \tag{5.20a}$$

$$F(\theta_1,\theta_2,\theta_3) < 1 \text{ in} \mathbb{N}(\delta) \setminus \mathbb{S}^\star \quad \text{(Angular Boundedness)} \tag{5.20b}$$

with $\mathbb{S}^\star := \{(0,0,0),(0,\pi,\pi),(\pi,0,\pi),(\pi,\pi,0)\}$ such that $\boldsymbol{u}(\theta_1)\otimes\boldsymbol{v}(\theta_2)\otimes\boldsymbol{w}(\theta_3) = \boldsymbol{u}_1^\star\otimes\boldsymbol{v}_1^\star\otimes\boldsymbol{w}_1^\star$ for any $(\theta_1,\theta_2,\theta_3)\in\mathbb{S}^\star$.

Similar as before, the Angular Interpolation property (5.20a) is a consequence of the construction process. In the rest of the paper, we will focus on showing the Angular Boundedness property (5.20b). Specifically, we will divide the near angular region into near vertex region and near band region, and then control the angular dual polynomial F in both near vertex region and near band region.

5.4.4 *Near vertex region* The near vertex region, denoted by $\mathbb{N}_v(\delta_v)$, is defined as the union of the eight small cubes all with side length δ_v in 8 corners of the cube $[0,\pi]^3$. We plot the near vertex region $\mathbb{N}_v(\delta_v)$ in Figure 4. Comparing with the definition of the near angular region $\mathbb{N}(\cdot)$, the near vertex region is also an near angular region but with a different parameter:

$$\mathbb{N}_v(\delta_v) = \mathbb{N}(\frac{\pi}{2} - \delta_v). \tag{5.21}$$

Without loss of generality, we can always assume the near vertex region $\mathbb{N}_v(\delta_v)$ is included in the near angular region $\mathbb{N}(\delta)$; otherwise, we only need to show the Angular-BIP holds in $\mathbb{N}_v(\delta_v)$. This assumption together with (5.21) implies

$$\delta_v \le \frac{\pi}{2} - \delta. \tag{5.22}$$

Note that $\pi/2-\delta$ is the side length of the corner-cubes in $\mathbb{N}(\delta)$.

CONTROLLING IN NEAR VERTEX REGION. To control the angular dual polynomial F in the near vertex region $\mathbb{N}_v(\delta_v)$, we further classify the eight small cubes in $\mathbb{N}_v(\delta_v)$ into two groups depending on if their vertices are in $\mathbb{S}^\star$ or not.

LEMMA 5.6 (Controlling in near vertex region). Under Assumptions I, II, III, if $r \ll n^{1.25}$, then for any $\xi_i \in \left(-\frac{\sqrt{2}-1}{3}, \frac{\sqrt{2}-1}{3}\right)$, we have

$$F(\theta_1+\xi_1, \theta_2+\xi_2, \theta_3+\xi_3) \le 1 \tag{5.23}$$

for $(\theta_1,\theta_2,\theta_3) \in \{(0,0,0),(0,\pi,\pi),(\pi,0,\pi),(\pi,\pi,0)\}$ and

$$F(\theta_1+\xi_1, \theta_2+\xi_2, \theta_3+\xi_3) < 0 \tag{5.24}$$

for $(\theta_1,\theta_2,\theta_3) \in \{(\pi,\pi,\pi),(\pi,0,0),(0,\pi,0),(0,0,\pi)\}$. Here, equality in (5.23) holds only if $\xi_1 = \xi_2 = \xi_3 = 0$.

The proof of Lemma 5.6 is in Appendix F.

REMARK 5.1. Lemma 5.6 proves the Angular-BIP holds in the near vertex region $\mathbb{N}_v(\delta_v)$ with $\delta_v = \frac{\sqrt{2}-1}{3}$:

$$F(\theta_1,\theta_2,\theta_3) = 1 \text{ in} \mathbb{S}^\star$$
$$F(\theta_1,\theta_2,\theta_3) < 1 \text{ in} \mathbb{N}_v(\delta_v) \setminus \mathbb{S}^\star$$

5.4.5 *Near band region.* The near band region is introduced to cover the remaining region $\mathbb{N}(\delta) \setminus \mathbb{N}_v(\delta_v)$. Invoking the definitions of the near angular region (5.16) and the near vertex region (5.21):

$$\mathbb{N}(\delta) = \left\{(\theta_1,\theta_2,\theta_3) : \theta_i \in \left[0, \frac{\pi}{2}-\delta\right] \cup \left[\frac{\pi}{2}+\delta, \pi\right]\right\}$$
$$\mathbb{N}_v(\delta_v) = \left\{(\theta_1,\theta_2,\theta_3) : \theta_i \in [0,\delta_v] \cup [\pi-\delta_v, \pi]\right\}$$

we have

$$\mathbb{N}(\delta) \setminus \mathbb{N}_v(\delta_v) = \left\{(\theta_1,\theta_2,\theta_3) : \theta_i \in \left(\delta_v, \frac{\pi}{2}-\delta\right) \cup \left(\frac{\pi}{2}+\delta, \pi-\delta_v\right)\right\} \cap \mathbb{N}(\delta), \tag{5.25}$$

which is nonempty since $\delta_v \le \pi/2 - \delta$ by the assumption (5.22). We plot the remaining region $\mathbb{N}(\delta) \setminus \mathbb{N}_v(\delta_v)$ projected onto the (θ_1,θ_2)-coordinates in Figure 5.

To let the near band region cover $\mathbb{N}(\delta) \setminus \mathbb{N}_v(\delta_v)$, we define it as

$$\mathbb{N}_b(\delta_b) := \left\{(\theta_1,\theta_2,\theta_3) : \theta_i \in \left(\delta_b, \frac{\pi}{2}-\delta_b\right) \cup \left(\frac{\pi}{2}+\delta_b, \pi-\delta_b\right), i = 1,2,3\right\}. \tag{5.26}$$

We plot the near band region $\mathbb{N}_b(\delta_b)$ projected onto the (θ_1,θ_2)-coordinates in Figure 6.

REMARK 5.2. From (5.25) and (5.26), we have $\mathbb{N}_b(\delta_b)$ covers $\mathbb{N}(\delta) \setminus \mathbb{N}_v(\delta_v)$ if $\delta_b \le \min\{\delta_v, \delta\}$, or equivalently,

$$\mathbb{N}(\delta) \subset \mathbb{N}_b(\delta_b) \cup \mathbb{N}_v(\delta_v), \quad \text{if} \delta_b \le \min\{\delta_v, \delta\}. \tag{5.27}$$

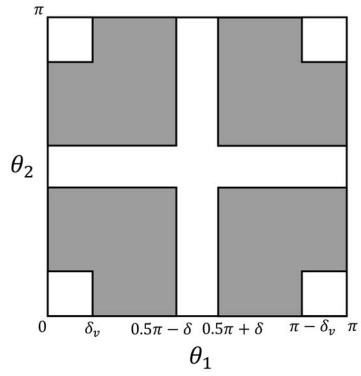


FIG. 5. The remaining region $\mathbb{N}(\delta) \setminus \mathbb{N}_v(\delta_v)$ projected onto the (θ_1, θ_2)-coordinates.

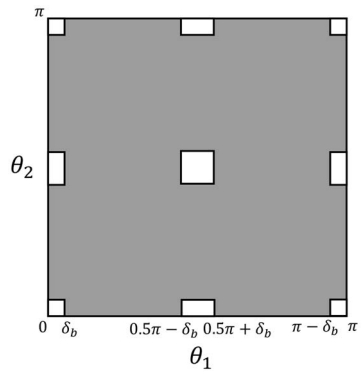


FIG. 6. The near band region $\mathbb{N}_b(\delta_b)$ projected onto the (θ_1, θ_2)-coordinates.

CONTROLLING IN NEAR BAND REGION. We start with the uniform upper-bound in Lemma 5.5:

$$
\begin{aligned}
&|F(\theta_1,\theta_2,\theta_3)| \le |\cos(\theta_1)\cos(\theta_2)\cos(\theta_3)| + |\sin(\theta_1)\sin(\theta_2)\sin(\theta_3)| + \frac{4}{3}\tau(\log n)n^{-r_c} \\
&\le \frac{1}{3}(|\cos(\theta_1)|^3 + |\cos(\theta_2)|^3 + |\cos(\theta_3)|^3) + \frac{1}{3}(|\sin(\theta_1)|^3 + |\sin(\theta_2)|^3 + |\sin(\theta_3)|^3) + \frac{4}{3}\tau(\log n)n^{-r_c} \\
&\le \frac{1}{3}(|\cos(\theta_i)|^3 + |\sin(\theta_i)|^3) + \frac{2}{3} + \frac{4}{3}\tau(\log n)n^{-r_c},\ \forall i \in \{1,2,3\}
\end{aligned} \tag{5.28}
$$

where the first inequality follows from (5.19) in Lemma 5.5 (under Assumptions I–III and $r \le n^{1.25-1.5r_c}$ with $r_c \in (0, \frac{1}{6})$), the second inequality follows from the inequality of arithmetic and geometric means, and the last one is a consequence of $|\sin(\theta)|^3 + |\cos(\theta)|^3 \le 1$. So, $|F(\theta_1,\theta_2,\theta_3)| < 1$ in $\mathbb{N}_b(\delta_b)$ if

$$
|\cos(\theta_i)|^3 + |\sin(\theta_i)|^3 < 1 - 4\tau(\log n)n^{-r_c} \tag{5.29}
$$

for *some* $i \in \{1,2,3\}$. The final result is summarized in the following lemma, with the proof listed in Appendix G.

LEMMA 5.7 (Controlling in near band region). Under Assumptions I, II, III, if $r \le n^{1.25-1.5r_c}$ with $r_c \in (0, \frac{1}{6})$, then for sufficiently large n, we have $|F(\theta_1,\theta_2,\theta_3)| < 1$ in $\mathbb{N}_b(\delta_b)$ for $\delta_b = \sqrt{\frac{80\tau(\log n)}{3}} n^{-0.5r_c}$.

5.4.6 *Combining the near vertex region and near band region.* Finally the Angular-BIP (5.20) follows from Lemma 5.6 and Lemma 5.7 if the union of the near vertex region $\mathbb{N}_v(\delta_v)$ and the near band region $\mathbb{N}_b(\delta_b)$ covers the near angular region $\mathbb{N}(\delta)$:

$$\mathbb{N}(\delta) \subset \mathbb{N}_v(\delta_v) \cup \mathbb{N}_b(\delta_b).$$

From (5.27), this happens when

$$\delta_b \le \min\{\delta, \delta_v\},$$

which is equivalent to

$$\delta_b \le \delta, \tag{5.30}$$

since $\delta_b = \sqrt{\frac{80\tau(\log n)}{3}} n^{-0.5r_c} \ll \frac{\sqrt{2}-1}{3} = \delta_v$.

Then by Proposition 5.1, q satisfies the BIP in $\mathcal{N}_1(\delta)$. Similar results apply to all individual near region $\mathcal{N}_p(\delta)$, for $p \in [r]$. Therefore, we claim the BIP holds in the whole near region $\mathcal{N}(\delta) = \bigcup_{p=1}^r \mathcal{N}_p(\delta)$.

LEMMA 5.8 (Near-region bound). Under Assumptions I, II, III, if $r \le n^{1.25-1.5r_c}$ with $r_c \in (0, \frac{1}{6})$, then for sufficiently large n, the dual polynomial q satisfies the BIP in $\mathcal{N}(\delta)$ for any $\delta \ge \delta_b$.

5.5 *Combining the far region and near region*

Combining Lemma 5.4 (for far region) and Lemma 5.8 (for near region), we conclude that the BIP holds in the whole domain $\mathbb{K}$ if Assumptions I, II, III are satisfied and

$$r \le \frac{n}{24\delta c^2} \text{ for} \delta \in [\delta_b, \frac{1}{24}] \quad \text{and} \quad r \le n^{1.25-1.5r_c} \text{ for} r_c \in (0, \frac{1}{6}). \tag{5.31}$$

Then letting $\delta = \delta_b$ (to maximize r) and $r_c = \frac{1}{8}$, the requirements (5.31) on r are reduced to the desired bound (2.4): $r \le \frac{n^{17/16}}{32c^2\sqrt{15\tau(\log n)}}$. The proof of Theorem 2.1 is completed.

6. Computational method

Theorem 2.1 shows that when the tensor factors $\{(\boldsymbol{u}_p^\star, \boldsymbol{v}_p^\star, \boldsymbol{w}_p^\star)\}_{p=1}^r$ satisfy Assumptions I, II, III, we can recover the tensor decomposition of r up to the order of $n^{17/16}$ by solving the convex, infinite-dimensional optimization (1.4). However, as a measure optimization problem, optimization problem

(1.4) is not directly solvable on a computer. In this section, we first propose a computational method based on the popular Burer–Monteiro factorization method [9] and then test it by numerical experiments.

THEOREM 6.1. Suppose the decomposition that achieves the tensor nuclear norm $\|\mathcal{T}\|_*$ involves r terms and $\tilde{r} \geq r$, then $\|\mathcal{T}\|_*$ is equal to the optimal value of the following optimization:

$$\underset{\{\boldsymbol{u}_p,\boldsymbol{v}_p,\boldsymbol{w}_p\}_{p=1}^{\tilde{r}}}{\text{minimize}} \sum_{p=1}^{\tilde{r}} \frac{1}{3}\left(\|\boldsymbol{u}_p\|_2^3 + \|\boldsymbol{v}_p\|_2^3 + \|\boldsymbol{w}_p\|_2^3\right) \text{ subject to } \mathcal{T} = \sum_{p=1}^{\tilde{r}} \boldsymbol{u}_p \otimes \boldsymbol{v}_p \otimes \boldsymbol{w}_p \tag{6.1}$$

Proof. Suppose the tensor nuclear norm is achieved by the decomposition

$$\mathcal{T} = \sum_{p=1}^{r} \lambda_p^\star \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star.$$

Then, we note that $\{\lambda_p^{\star 1/3}\boldsymbol{u}_p^\star, \lambda_p^{\star 1/3}\boldsymbol{v}_p^\star, \lambda_p^{\star 1/3}\boldsymbol{w}_p^\star\}_{p=1}^{\tilde{r}}$ forms a feasible solution to (6.1) when $\tilde{r} = r$. When $\tilde{r} > r$, we can zero-pad the remaining factors $\{\boldsymbol{u}_p, \boldsymbol{v}_p, \boldsymbol{w}_p\}_{p=r+1}^{\tilde{r}}$. The objective function value at this feasible solution is $\frac{1}{3}(\sum_{p=1}^{\tilde{r}} 3\lambda_p^\star) = \|\mathcal{T}\|_*$. This shows that $\|\mathcal{T}\|_*$ is greater than the optimal value of (6.1).

To show the other, suppose an optimal solution of (6.1) is $\{\boldsymbol{u}_p, \boldsymbol{v}_p, \boldsymbol{w}_p\}_{p=1}^{\tilde{r}}$. Define $\lambda_p := \|\boldsymbol{u}_p\|_2\|\boldsymbol{v}_p\|_2\|\boldsymbol{w}_p\|_2$, for $p \in [\tilde{r}]$. Then,

$$\mathcal{T} = \sum_{p:\lambda_p \neq 0} \lambda_p \frac{\boldsymbol{u}_p}{\|\boldsymbol{u}_p\|_2} \otimes \frac{\boldsymbol{v}_p}{\|\boldsymbol{v}_p\|_2} \otimes \frac{\boldsymbol{w}_p}{\|\boldsymbol{w}_p\|_2}.$$

By definition of the tensor nuclear norm (4.1), we have

$$\|\mathcal{T}\|_* \leq \sum_{p:\lambda_p \neq 0} \lambda_p = \sum_{p=1}^{\tilde{r}} \lambda_p = \sum_{p=1}^{\tilde{r}} \|\boldsymbol{u}_p\|_2\|\boldsymbol{v}_p\|_2\|\boldsymbol{w}_p\|_2 \leq \frac{1}{3}\sum_{p=1}^{\tilde{r}}\left[\|\boldsymbol{u}_p\|_2^3 + \|\boldsymbol{v}_p\|_2^3 + \|\boldsymbol{w}_p\|_2^3\right],$$

which is the optimal value of (6.1). Therefore, the optimal value of (6.1) is equal to $\|\mathcal{T}\|_*$. □

Theorem 6.1 implies that when an upper bound on r is known, we can solve the nonlinear (and non-convex) program (6.1) to compute the tensor nuclear norm (and obtain the corresponding decomposition). Despite the nonconvex nature of (6.1), numerical simulations suggest that the ADMM approach [8] has superior performance in solving (6.1).

7. Numerical experiments and beyond

Now we perform some numerical results to test the performance of the proposed Burer–Monteiro factorization method. In particular, we will examine the phase transition of the rate of success for the ADMM implementation of the proposed Burer–Monteiro factorization approach (6.1) with random

initialization. To illustrate the superiority of the proposed tensor nuclear norm approach, we compare it with the Least Squares formulation, that is, the L2 error minimization problem

$$\underset{\{\boldsymbol{u}_p,\boldsymbol{v}_p,\boldsymbol{w}_p\}_{p=1}^{\tilde{r}}}{\text{minimize}} \left\| \boldsymbol{\mathcal{T}} - \sum_{p=1}^{\tilde{r}} \boldsymbol{u}_p \otimes \boldsymbol{v}_p \otimes \boldsymbol{w}_p \right\|_F^2. \tag{7.1}$$

In the experiments, the r tensor factors $\{(\boldsymbol{u}_p^\star, \boldsymbol{v}_p^\star, \boldsymbol{w}_p^\star)\}_{p=1}^r$ were generated following i.i.d. Gaussian distribution, and then each $\boldsymbol{u}_p^\star, \boldsymbol{v}_p^\star, \boldsymbol{w}_p^\star$ was normalized to have a unit norm. We set the coefficients $\lambda_p^\star = (1 + \varepsilon_p^2)/2$, where ε_p is chosen from the standard normal distribution, to ensure a minimal coefficient of at least $1/2$. With the generated ground-truth factors $\{(\boldsymbol{u}_p^\star, \boldsymbol{v}_p^\star, \boldsymbol{w}_p^\star)\}_{p=1}^r$ and coefficients $\{\lambda_p\}_{p=1}^r$, we generated the tensor $\boldsymbol{\mathcal{T}} = \sum_{p=1}^r \lambda_p^\star \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star$. To generate the phase transition plot, we varied the dimension n and factor-number r, and for each fixed (r, n) pair, 20 instances of such tensor were generated. We then ran the ADMM algorithm to minimize (6.1), and ran LBFGS to minimize the L2 error function (7.1), from the same random initialization. We remark that the global minimum value of the minimization (6.1) keeps the same for any $\tilde{r} \geq r$ but the global minimum solution doesn't. Therefore, to find all the true tensor factors, we choose $\tilde{r} = r$ in both methods. For each instance, we declared success if the relative recovery error $\text{Err}\left(\left\{\left(\widehat{\boldsymbol{u}}_p, \widehat{\boldsymbol{v}}_p, \widehat{\boldsymbol{w}}_p\right)\right\}_{p=1}^r\right)$ of the output tensor factors $\left\{\left(\widehat{\boldsymbol{u}}_p, \widehat{\boldsymbol{v}}_p, \widehat{\boldsymbol{w}}_p\right)\right\}_{p=1}^r$ (after removing sign and permutation ambiguities) is within 10^{-3} where

$$\text{Err}\left(\{(\widehat{\boldsymbol{u}}_p, \widehat{\boldsymbol{v}}_p, \widehat{\boldsymbol{w}}_p)\}_{p=1}^r\right) := \sum_{p=1}^r \left(\frac{\|\widehat{\boldsymbol{u}}_p - \boldsymbol{u}_p^\star\|_2}{\|\boldsymbol{u}_p^\star\|_2} + \frac{\|\widehat{\boldsymbol{v}}_p - \boldsymbol{v}_p^\star\|_2}{\|\boldsymbol{v}_p^\star\|_2} + \frac{\|\widehat{\boldsymbol{w}}_p - \boldsymbol{w}_p^\star\|_2}{\|\boldsymbol{w}_p^\star\|_2} \right).$$

We plot the experiment results in Figure 7, which shows that the proposed method is clearly superior compared to the traditional Least Squares method.

Due to the nonconvexity nature of the two tensor decomposition formulations, we believe that the performance gain achieved by the proposed ADMM approach is because the optimization landscape of the Least Squares formulation of tensor decomposition is not as good as that of the tensor nuclear norm formulation (6.1). We therefore conjecture that the tensor nuclear norm is crucial in flatting out the spurious local minima and high-order saddle points so that it helps to provide a benign optimization landscape, e.g. 'strict saddle property', i.e. every critical point is either a strict saddle (where the Hessian has negative eigenvalues) or a global minimizer, see [17, 37, 39–44, 61–63] for more literature of landscape analysis. In contrast, the Least Squares formulation of tensor decomposition doesn't satisfy the 'strict saddle property'. To verify this conjecture, we perform some preliminary analysis. To simplify the notations and analysis, we consider the symmetric case as a first step. We believe the nonsymmetric case will have similar properties. More precisely, we consider the following L2 loss function

$$g(\mathbf{U}) = \frac{1}{6} \|\boldsymbol{u}_1 \otimes^3 + \boldsymbol{u}_2 \otimes^3 - \boldsymbol{a}_1 \otimes^3 - \boldsymbol{a}_2 \otimes^3 \|_F^2 \tag{7.2}$$

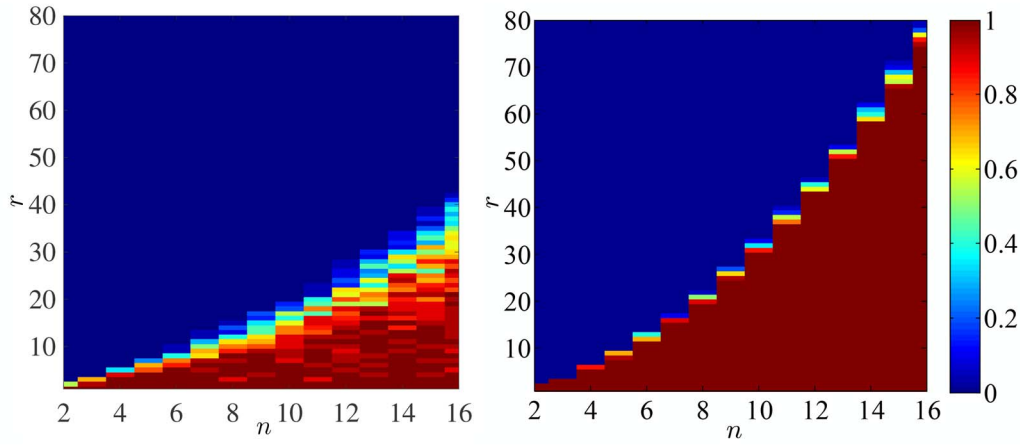


FIG. 7. Rate of success using Least Squares method (left) and ADMM implementation of (6.1) (right) for tensor decomposition, respectively.

where $\boldsymbol{u}^{\otimes 3} := \boldsymbol{u} \otimes \boldsymbol{u} \otimes \boldsymbol{u}$ and $\mathbf{U} := [\boldsymbol{u}_1, \boldsymbol{u}_2] \in \mathbb{R}^{2\times 2}$. For this function, the critical points are given by the following equation:

$$\nabla g(\mathbf{U}) = \mathbf{U}[(\mathbf{U}^\top \mathbf{U}) \odot (\mathbf{U}^\top \mathbf{U})] - \mathbf{A}[(\mathbf{A}^\top \mathbf{U}) \odot (\mathbf{A}^\top \mathbf{U})] = \mathbf{0}. \tag{7.3}$$

For simplicity, assume $\mathbf{A} = \mathbf{I}$ the identity matrix. Then the stationary equation reduces to

$$\mathbf{U}[(\mathbf{U}^\top \mathbf{U}) \odot (\mathbf{U}^\top \mathbf{U})] = \mathbf{U} \odot \mathbf{U} \tag{7.4}$$

Directly solving the above equation (7.4) (through MATHEMATICA) generates three sets of solutions.

Case I

$$\mathbf{U}_1(x) = \begin{bmatrix} 0 & 0 \\ \sqrt[3]{1-x^3} & x \end{bmatrix} \text{ or } \begin{bmatrix} \sqrt[3]{1-x^3} & x \\ 0 & 0 \end{bmatrix}$$

where $x \in \mathbb{R}$. In this case, we then compute the eigenvalues of the Hessian matrix:

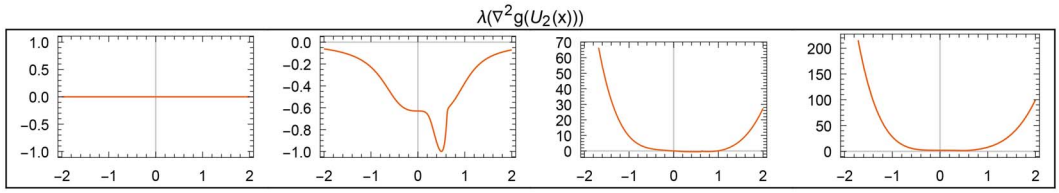
$$\boldsymbol{\lambda}(\nabla^2 g(U_1(x))) = \begin{bmatrix} 0 \\ 0 \\ \underbrace{\left(x^4 - x^3\sqrt[3]{1-x^3} + \sqrt[3]{1-x^3}\right)}_{\geq 3(2^{\frac{2}{3}})} \\ 3\underbrace{\left(x^4 - x^3\sqrt[3]{1-x^3} + \sqrt[3]{1-x^3}\right)}_{\geq 3(2^{\frac{2}{3}})} \end{bmatrix},$$

where $\boldsymbol{\lambda}(\cdot)$ denotes the eigenvalue list of its argument. We conclude that the first set of critical points $\mathbf{U}_1(x)$ are neither strict saddle points (since Hessian matrix has no negative eigenvalues) nor global minima (since it is not a permuted version of Identify matrix). Therefore, the Least Squares formulation of tensor decomposition doesn't satisfy the 'strict saddle property'. In addition, since the Hessian doesn't have negative curvature at these critical points, the iterative algorithm such as gradient descent easily gets trapped by these points. This explains the relatively poor performance of the Least Squares formulation of tensor decomposition in Figure 7.

Case II

$$\mathbf{U}_2(x) = \begin{bmatrix} \sqrt[3]{0.25 - x^3} & x \\ \sqrt[3]{0.25 - x^3} & x \end{bmatrix}$$

For this set of critical points, there are no closed-form eigenvalues of its Hessian matrix. For convenience, we plot the four eigenvalues as a function of x.



We therefore conclude that this set of critical points are strict saddle points at least when $x \in [-2, 2]$ because the Hessian matrix has a negative eigenvalue for $x \in [-2, 2]$.

Case III

$$\mathbf{U}_3(x) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

It is clear that the zero point is a special critical point and it is actually a high-order saddle point, because the Hessian matrix at zero point is also zero.

Although the 2-by-2 case is very simple, it provides some evidence that the Least Squares formulation of tensor decomposition doesn't satisfy the 'strict saddle property' and has high-order saddle points.

8. Conclusion

By explicitly constructing a dual certificate, we derive similar incoherence conditions (as the separation conditions in super-resolution problem) for a tensor decomposition to achieve the tensor nuclear norm. This implies that the infinite dimensional total mass minimization can globally identify those decompositions satisfying the developed incoherence conditions. Computational method based on Burer-Monteiro factorization approach is used to solve the measure optimization. Numerical experiments show that the Burer–Monteiro factorization approach achieves amazingly superior performance. Future work will analyze the nonconvex landscape of the Burer–Monteiro factorization approach.

Data Availability Statements

No new data were generated or analyzed in support of this research.

Funding

National Science Foundation [DMS-1913039 to LS, CCF-2203060, CCF-2106834 to GT].

References

1. Anandkumar, A., Ge, R. & Janzamin, M. (2015) Learning overcomplete latent variable models through tensor methods. (Grünwald, Peter and Hazan, Elad and Kale, Satyen eds), *Proceedings of The 28th Conference on Learning Theory*. Paris, France: PMLR, **40**, pp. 36–112.
2. Anandkumar, A., Ge, R., & Janzamin, M. Analyzing tensor power method dynamics in overcomplete regime. *Journal of Machine Learning Research*, **18**(22):1–40, 2017.
3. Barak, B., Kelner, J. A. & Steurer, D. (2015) Dictionary learning and tensor decomposition via the sum-of-squares method. *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. Portland, Oregon, USA: ACM, **9**, pp. 143–151.
4. Barvinok, A. I. (2002) A course in convexity. *American Mathematical Society*, vol. **54**. https://dblp.org/rec/books/daglib/0012694.bib
5. Bendory, T., Dekel, S., & Feuer, A. Super-resolution on the sphere using convex optimization. *Signal Processing, IEEE Transactions on*, **63**(9):2253–2262, 2015.
6. Bengua, J. A., Phien, H. N., Tuan, H. D., & Do, M. N. Efficient tensor completion for color image and video recovery: Low-rank tensor train. *IEEE Trans. Image Process.*, **26**(5):2466–2479, 2017.
7. Bhaskar, B. N., Tang, G., & Recht, B. Atomic norm denoising with applications to line spectral estimation. *Signal Processing, IEEE Transactions on*, **61**(23):5987–5999, 2013.
8. Boyd, S. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, **3**(1):1–122, 2011.
9. Burer, S. & Monteiro, R. D. C. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Math. Programming*, **95**(2):329–357, February 2003.
10. Cai, C., Gen Li, H., Poor, V. & Chen, Y. (2019) Nonconvex low-rank symmetric tensor completion from noisy data. H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alchü-Buc, E. Fox & R. Garnett eds) *Advances in neural information processing systems*. Curran Associates, Inc.,
11. Candès, E. J. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, **346**(9-10):589–592, May 2008.
12. Candès, E. J & Fernandez-Granda, C. Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, Wiley Online Library, **67**(6):906–956, June 2014.
13. Candès, E. J. & Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, **9**(6):717–772, 2009.
14. Cardoso, J.-F. (1989) Source separation using higher order moments. *International Conference on Acoustics, Speech, and Signal Processing*, vol. **4**, pp. 2109–2112.
15. Chandrasekaran, V., Recht, B., Parrilo, P. A., & Willsky, A. S. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, **12**(6):805–849, 2012.
16. Chi, Y. & Chen, Y. Compressive two-dimensional harmonic retrieval via atomic norm minimization. *IEEE Trans. Signal Process.*, **63**(4):1030–1042, 2014.
17. Chi, Y., Lu, Y. M., & Chen, Y. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Trans. Signal Process.*, **67**(20):5239–5269, 2019.
18. Comon, P. Tensor decompositions, state of the art and applications. *IMA Conf. Mathematics in Signal Processing*, May 2009.

19. Davenport, M. A. & Romberg, J. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, **10**(4):608–622, 2016.
20. Donoho, D. L. & Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. *Proceedings of the National Academy of Sciences*, **100**(5):2197–2202. doi: 10.1073/pnas.0437847100.
21. Eftekhari, A., Tanner, J., Thompson, A., Toader, B. & Tyagi, H. (2019) Sparse non-negative super-resolution-simplified and stabilised. *Applied and Computational Harmonic Analysis*, **50**:216–280. doi: https://doi.org/10.1016/j.acha.2019.08.004.
22. Fernandez-Granda, C. (2013) Support detection in super-resolution. *Proceedings of the 10th International Conference on Sampling Theory and Applications (SampTA 2013)*, pp. 145–148.
23. Fernandez-Granda, C., Tang, G., Wang, X. & Zheng, L. (2017) Demixing sines and spikes: Robust spectral super-resolution in the presence of outliers. *Information and Inference: A Journal of the IMA*, **7**(1):105–168.
24. Foucart, S. & Rauhut, H. (2013) *A mathematical introduction to compressive sensing*. *Applied and Numerical Harmonic Analysis. Springer*. New York:: Birkhäuser Basel, 0817649476.
25. Friedland, S. & Lim, L.-H. Nuclear norm of higher-order tensors. *Math. Comp.*, **87**:1255–1281, 2017.
26. Gandy, S., Recht, B., & Yamada, I. (2011) Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, IOP Publishing, **27**(2):1–19.
27. Ge, R., Lee, J. D., & Ma, T. Learning one-hidden-layer neural networks with landscape design. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.
28. Heckel, R., Morgenshtern, V. I., & Soltanolkotabi, M. Super-resolution radar. *Information and Inference: A Journal of the IMA*, **5**(1):22–75, 2016.
29. Hillar, C. J. & Lim, L.-H. Most tensor problems are NP-Hard. *Journal of the ACM (JACM)*, **60**(6):45–39, November 2013.
30. Hopkins, S. B., Schramm, T. & Shi, J. (2019) A robust spectral algorithm for overcomplete tensor decomposition. *Conference on Learning Theory*.PMLR, pp. 1683–1722.
31. Hou, J. & Qian, H. (2017) Collaboratively filtering malware infections: a tensor decomposition approach. *Proceedings of the ACM Turing 50th Celebration Conference-China*, vol. **28**.ACM.
32. Huang, B., Cun, M., Goldfarb, D. & Wright, J. (2014) Provable low-rank tensor recovery. *Optimization-Online*, **4252**, 1.
33. Jain, P. & Sewoong, O. Provable tensor factorization with missing data. (Z. Ghahramani and M. Welling and C. Cortes and N. Lawrence and K. Q. Weinberger eds), *Advances in Neural Information Processing Systems*, urran Associates, Inc. **27**, pages 1431–1439, 2014.
34. Kolda, T. G. & Bader, B. W. Tensor decompositions and applications. *SIAM Rev.*, **51**(3):455–500, 2009.
35. Kruskal, J. B. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.*, **18**(2):95–138, January 1977.
36. Li, Q., Prater, A., Shen, L. & Tang, G. (2015) Overcomplete tensor decomposition via convex optimization. *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*.Cancun, Mexico: IEEE, pp. 53–56.
37. Li, Q. & Tang, G. (2017) Convex and nonconvex geometries of symmetric tensor factorization. *In Asilomar Conference on Signals, Systems, and Computers*, 1.
38. Li, Q. & Tang, G. (2018) Approximate support recovery of atomic line spectral estimation: A tale of resolution and precision. *Appl. Comput. Harmon. Anal.*, 1.
39. Li, Q., Zhu, Z. & Tang, G. (2017) Geometry of factored nuclear norm regularization. arXiv preprint arXiv:1704.01265.
40. Li, Q., Zhu, Z., & Tang, G. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, **8**(1):51–96, 2018.

41. Li, Q., Zhu, Z. & Tang, G. (2019) Alternating minimizations converge to second-order optimal solutions. (Chaudhuri, Kamalika and Salakhutdinov, Ruslan eds), *International Conference on Machine Learning*. PMLR, pp. 3935–3943.
42. Li, Q., Zhu, Z., Tang, G. & Wakin, M. B. (2019) *Provable bregman-divergence based methods for nonconvex and non-lipschitz problems*arXiv preprint arXiv:1904.09712.
43. Li, S. & Li, Q. (2022) Local and global convergence of general burer-monteiro tensor optimizations. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. **36**.
44. Li, S., Li, Q., Zhu, Z., Tang, G. & Wakin, M. B. (2020) The global geometry of centralized and distributed low-rank matrix recovery without regularization. *IEEE Signal Processing Letters*, **27**, 1400–1404.
45. Canyi, L., Feng, J., Chen, Y., Liu, W., Lin, Z. & Yan, S. (2016) Tensor robust principal component analysis: exact recovery of corrupted low-rank tensors via convex optimization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Los Alamitos, CA, USA: IEEE Computer Society, pp. 5249–5257.
46. Ma, T., Shi, J., & Steurer, D. Polynomial-time tensor decompositions with sum-of-squares. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 438–446. IEEE, 2016.
47. Cun, M., Huang, B., Wright, J., & Goldfarb, D. Square deal: lower bounds and improved relaxations for tensor recovery. In *International Conference on Machine Learning*, pages 73–81, 2014.
48. Potechin, A. & Steurer, D. Exact tensor completion with sum-of-squares. In *Conference on Learning Theory*, pages 1619–*1673*. PMLR, 2017.
49. Recht, B. A simpler approach to matrix completion. *Journal of Machine Learning Research*, **12**(Dec):3413–3430, 2011.
50. Recht, B., Fazel, M., & Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, **52**(3):471–501, August 2010.
51. Sedghi, H., Janzamin, M. & Anandkumar, A. (2016) Provable tensor methods for learning mixtures of generalized linear models. *In Artificial Intelligence and Statistics*, 1223–1231.
52. Sidiropoulos, N. D., De Lathauwer, L., Xiao, F., Huang, K., Papalexakis, E. E., & Faloutsos, C. Tensor decomposition for signal processing and machine learning. *IEEE Trans. Signal Process.*, **65**(13):3551–3582, 2017.
53. Smilde, A., Bro, R. & Geladi, P. (2005) *Multi-Way Analysis: Applications in the Chemical Sciences*.John Wiley & Sons.
54. Gongguo Tang, B. N. Bhaskar, P. S., & Recht, B. Compressed sensing off the grid. *Information Theory, IEEE Transactions on*, **59**(11):7465–7490, 2013.
55. Tang, G., Bhaskar, B. N., & Recht, B. Near minimax line spectral estimation. *IEEE Transactions on Information Theory*, **61**(1):499–512, 2015.
56. Tang, G. & Recht, B. Atomic decomposition of mixtures of translation-invariant signals. In *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing CAMSAP*, Saint Martin: IEEE, December 2013.
57. Tang, G. & Shah, P. (2015) Guaranteed tensor decomposition: a moment approach. (Francis Bach & David Blei eds), *International Conference on Machine Learning*. Lille, France: PMLR.
58. Tang, G., Shah, P., Bhaskar, B. N., & Recht, B. Robust line spectral estimation. In *2014 48th Asilomar Conference on Signals, Systems and Computers*, pages 301–305. Pacific Grove, CA, USA: IEEE, 2014.
59. Watson, G. A. Characterization of the subdifferential of some matrix norms. *Linear Algebra Appl.*, **170**:33–45, June 1992.
60. Yuan, M. & Zhang, C.-H. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, **16**(4):1031–1068, 2016.
61. Zhu, Z., Li, Q., Tang, G., & Wakin, M. B. Global optimality in low-rank matrix optimization. *IEEE Trans. Signal Process.*, **66**(13):3614–3628, 2018.
62. Zhu, Z., Li, Q., Tang, G., & Wakin, M. B. The global optimization geometry of low-rank matrix optimization. *IEEE Transactions on Information Theory*, **67**(2):1308–1331, 2021.
63. Zhu, Z., Li, Q., Yang, X., Tang, G. & Wakin, M. B. (2019) Distributed low-rank matrix factorization with exact consensus. *Advances in Neural Information Processing Systems*, **32**, 8422–8432.

A. Proof of Lemma 5.1

Proof. From the KKT conditions of the constrained optimization (5.1), we have the partial derivatives of its Lagrangian

$$\mathcal{L}(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w},a,b,c) = q(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}) - a(\|\boldsymbol{u}\|_2^2 - 1) - b(\|\boldsymbol{v}\|_2^2 - 1) - c(\|\boldsymbol{w}\|_2^2 - 1)$$

at $\boldsymbol{u} = \boldsymbol{u}_p^\star$, $\boldsymbol{v} = \boldsymbol{v}_p^\star$, and $\boldsymbol{w} = \boldsymbol{w}_p^\star$, $p = 1,\dots,r$, must vanish. Therefore,

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{u}_p^\star,\boldsymbol{v}_p^\star,\boldsymbol{w}_p^\star,a,b,c)}{\partial \boldsymbol{u}} &= \frac{\partial q(\boldsymbol{u}_p^\star,\boldsymbol{v}_p^\star,\boldsymbol{w}_p^\star)}{\partial \boldsymbol{u}} - 2a\boldsymbol{u}_p^\star = 0,\\
\frac{\partial \mathcal{L}(\boldsymbol{u}_p^\star,\boldsymbol{v}_p^\star,\boldsymbol{w}_p^\star,a,b,c)}{\partial \boldsymbol{v}} &= \frac{\partial q(\boldsymbol{u}_p^\star,\boldsymbol{v}_p^\star,\boldsymbol{w}_p^\star)}{\partial \boldsymbol{v}} - 2b\boldsymbol{v}_p^\star = 0,\\
\frac{\partial \mathcal{L}(\boldsymbol{u}_p^\star,\boldsymbol{v}_p^\star,\boldsymbol{w}_p^\star,a,b,c)}{\partial \boldsymbol{w}} &= \frac{\partial q(\boldsymbol{u}_p^\star,\boldsymbol{v}_p^\star,\boldsymbol{w}_p^\star)}{\partial \boldsymbol{w}} - 2c\boldsymbol{w}_p^\star = 0.
\end{aligned} \tag{A.1}$$

Hence, $2a = \langle \frac{\partial q(\boldsymbol{u}_p^\star,\boldsymbol{v}_p^\star,\boldsymbol{w}_p^\star)}{\partial \boldsymbol{u}}, \boldsymbol{u}_p^\star\rangle$, $2b = \langle \frac{\partial q(\boldsymbol{u}_p^\star,\boldsymbol{v}_p^\star,\boldsymbol{w}_p^\star)}{\partial \boldsymbol{v}}, \boldsymbol{v}_p^\star\rangle$, and $2c = \langle \frac{\partial q(\boldsymbol{u}_p^\star,\boldsymbol{v}_p^\star,\boldsymbol{w}_p^\star)}{\partial \boldsymbol{w}}, \boldsymbol{w}_p^\star\rangle$. Note that q satisfies the Interpolation condition and $\frac{\partial q(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})}{\partial \boldsymbol{u}(i)} = \sum_{j,k} Q_{ijk}\boldsymbol{v}(j)\boldsymbol{w}(k)$, we have that

$$2a = \sum_{i,j,k} Q_{ijk}\boldsymbol{u}_p^\star(i)\boldsymbol{v}_p^\star(j)\boldsymbol{w}_p^\star(k) = q(\boldsymbol{u}_p^\star,\boldsymbol{v}_p^\star,\boldsymbol{w}_p^\star) = 1.$$

That is $a = 1/2$. With similar arguments, one can show that $b = c = 1/2$. The conclusion of this lemma follows from (A.1). □

B. Proof of Lemma 5.2

Proof. First, the Lagrangian form of (5.3) is

$$\begin{aligned}
\mathcal{L}(\mathcal{Q},\{\boldsymbol{\alpha}_p^\star,\boldsymbol{\beta}_p^\star,\boldsymbol{\gamma}_p^\star\}_{p=1}^r) &= \frac{1}{2}\|\mathcal{Q}\|_F^2 - \sum_{p=1}^r \left(\mathcal{Q}\times_1\boldsymbol{\alpha}_p^\star\times_2\boldsymbol{v}_p^\star\times_3\boldsymbol{w}_p^\star + \mathcal{Q}\times_1\boldsymbol{u}_p^\star\times_2\boldsymbol{\beta}_p^\star\times_3\boldsymbol{w}_p^\star + \mathcal{Q}\times_1\boldsymbol{u}_p^\star\times_2\boldsymbol{v}_p^\star\times_3\boldsymbol{\gamma}_p^\star\right)\\
&= \frac{1}{2}\|\mathcal{Q}\|_F^2 - \left\langle \mathcal{Q}, \sum_{p=1}^r \boldsymbol{\alpha}_p^\star\otimes\boldsymbol{v}_p^\star\otimes\boldsymbol{w}_p^\star + \boldsymbol{u}_p^\star\otimes\boldsymbol{\beta}_p^\star\otimes\boldsymbol{w}_p^\star + \boldsymbol{u}_p^\star\otimes\boldsymbol{v}_p^\star\otimes\boldsymbol{\gamma}_p^\star\right\rangle
\end{aligned}$$

with the Lagrangian multipliers $\{\boldsymbol{\alpha}_p^\star,\boldsymbol{\beta}_p^\star,\boldsymbol{\gamma}_p^\star\}_{p=1}^r$ to be chosen such that $\mathcal{Q}$ satisfies (5.2). Then, by the KKT necessary conditions, the solution of the least-norm problem (5.3) should satisfy

$$\mathbf{0} = \frac{\partial \mathcal{L}(\mathcal{Q},\{\boldsymbol{\alpha}_p^\star,\boldsymbol{\beta}_p^\star,\boldsymbol{\gamma}_p^\star\}_{p=1}^r)}{\partial \mathcal{Q}} = \mathcal{Q} - \sum_{p=1}^r \left(\boldsymbol{\alpha}_p^\star\otimes\boldsymbol{v}_p^\star\otimes\boldsymbol{w}_p^\star + \boldsymbol{u}_p^\star\otimes\boldsymbol{\beta}_p^\star\otimes\boldsymbol{w}_p^\star + \boldsymbol{u}_p^\star\otimes\boldsymbol{v}_p^\star\otimes\boldsymbol{\gamma}_p^\star\right).$$

□

C. Proof of Lemma 5.3

Proof. We need to find coefficients $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$ so that

$$\mathcal{Q} = \sum_{p=1}^{r} \left(\boldsymbol{\alpha}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star + \boldsymbol{u}_p^\star \otimes \boldsymbol{\beta}_p^\star \otimes \boldsymbol{w}_p^\star + \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{\gamma}_p^\star \right)$$

satisfies

$$\begin{aligned} \mathcal{Q} \times_2 \boldsymbol{v}_p^\star \times_3 \boldsymbol{w}_p^\star &= \boldsymbol{u}_p^\star, \quad \forall p \in [r], \\ \mathcal{Q} \times_1 \boldsymbol{u}_p^\star \times_3 \boldsymbol{w}_p^\star &= \boldsymbol{v}_p^\star, \quad \forall p \in [r], \\ \mathcal{Q} \times_1 \boldsymbol{u}_p^\star \times_2 \boldsymbol{v}_p^\star &= \boldsymbol{w}_p^\star, \quad \forall p \in [r]. \end{aligned} \tag{C.1}$$

□

C.1 *An iteration scheme*

We adopt the following *iterative scheme* to find such $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$:

$$\begin{aligned} \boldsymbol{\alpha}_q^{t+1} &= \boldsymbol{\alpha}_q^t - \rho \left(\mathcal{Q}_1^t \times_2 \boldsymbol{v}_p^\star \times_3 \boldsymbol{w}_q^\star - \boldsymbol{u}_q^\star \right), \quad q \in [r], \\ \boldsymbol{\beta}_q^{t+1} &= \boldsymbol{\beta}_q^t - \rho \left(\mathcal{Q}_2^t \times_1 \boldsymbol{u}_p^\star \times_3 \boldsymbol{w}_q^\star - \boldsymbol{v}_q^\star \right), \quad q \in [r], \\ \boldsymbol{\gamma}_q^{t+1} &= \boldsymbol{\gamma}_q^t - \rho \left(\mathcal{Q}_3^t \times_1 \boldsymbol{u}_p^\star \times_2 \boldsymbol{v}_q^\star - \boldsymbol{w}_q^\star \right), \quad q \in [r], \end{aligned} \tag{C.2}$$

initialized by $\boldsymbol{\alpha}_q^0 = \frac{1}{3}\boldsymbol{u}_q^\star$, $\boldsymbol{\beta}_q^0 = \frac{1}{3}\boldsymbol{v}_q^\star$, and $\boldsymbol{\gamma}_q^0 = \frac{1}{3}\boldsymbol{w}_q^\star$ with $q \in [r]$. Here the parameter ρ is a step size to be chosen later and the tensors

$$\begin{aligned} \mathcal{Q}_1^t &:= \sum_{p=1}^{r} \left(\boldsymbol{\alpha}_p^t \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star + \boldsymbol{u}_p^\star \otimes \boldsymbol{\beta}_p^\star \otimes \boldsymbol{w}_p^\star + \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{\gamma}_p^\star \right), \\ \mathcal{Q}_2^t &:= \sum_{p=1}^{r} \left(\boldsymbol{\alpha}_p^t \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star + \boldsymbol{u}_p^\star \otimes \boldsymbol{\beta}_p^t \otimes \boldsymbol{w}_p^\star + \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{\gamma}_p^\star \right), \\ \mathcal{Q}_3^t &:= \sum_{p=1}^{r} \left(\boldsymbol{\alpha}_p^t \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star + \boldsymbol{u}_p^\star \otimes \boldsymbol{\beta}_p^t \otimes \boldsymbol{w}_p^\star + \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{\gamma}_p^t \right). \end{aligned} \tag{C.3}$$

Note that the above iterative scheme is for theoretical analysis only as we used $\{\boldsymbol{\alpha}_p^\star, \boldsymbol{\beta}_p^\star, \boldsymbol{\gamma}_p^\star\}_{p=1}^r$ in the definitions of $\mathcal{Q}_1^t$, $\mathcal{Q}_2^t$ and $\mathcal{Q}_3^t$.

C.2 *Convergence analysis*

We next establish the convergence of the iterations (C.2). Plugging the tensor eigenvalue equations (C.1) into (C.2) followed by subtracting the true solutions from both sides yields for $q \in [r]$

$$\begin{aligned}
\boldsymbol{\alpha}_q^{t+1} - \boldsymbol{\alpha}_q^\star &= \boldsymbol{\alpha}_q^t - \boldsymbol{\alpha}_q^\star - \rho[\boldsymbol{\mathcal{Q}}_1^t - \boldsymbol{\mathcal{Q}}]\times_2 \boldsymbol{v}_q^\star \times_3 \boldsymbol{w}_q^\star,\\
\boldsymbol{\beta}_q^{t+1} - \boldsymbol{\beta}_q^\star &= \boldsymbol{\beta}_q^t - \boldsymbol{\beta}_q^\star - \rho[\boldsymbol{\mathcal{Q}}_2^t - \boldsymbol{\mathcal{Q}}]\times_1 \boldsymbol{u}_q^\star \times_3 \boldsymbol{w}_q^\star,\\
\boldsymbol{\gamma}_q^{t+1} - \boldsymbol{\gamma}_q^\star &= \boldsymbol{\gamma}_q^t - \boldsymbol{\gamma}_q^\star - \rho[\boldsymbol{\mathcal{Q}}_3^t - \boldsymbol{\mathcal{Q}}]\times_1 \boldsymbol{u}_q^\star \times_2 \boldsymbol{v}_q^\star.
\end{aligned} \tag{C.4}$$

Then plugging the definitions of $\boldsymbol{\mathcal{Q}}_1^t, \boldsymbol{\mathcal{Q}}_2^t, \boldsymbol{\mathcal{Q}}_3^t$ (C.3) into (C.4) and using the following matrix notations

$$\begin{aligned}
\mathbf{A}^t &:= \left[\boldsymbol{\alpha}_1^t, \cdots, \boldsymbol{\alpha}_r^t\right], \mathbf{A} := \left[\boldsymbol{\alpha}_1^\star, \cdots, \boldsymbol{\alpha}_r^\star\right],\\
\mathbf{B}^t &:= \left[\boldsymbol{\alpha}_1^t, \cdots, \boldsymbol{\alpha}_r^t\right], \mathbf{B} := \left[\boldsymbol{\alpha}_1^\star, \cdots, \boldsymbol{\alpha}_r^\star\right],\\
\mathbf{C}^t &:= \left[\boldsymbol{\gamma}_1^t, \cdots, \boldsymbol{\gamma}_r^t\right], \mathbf{C} := \left[\boldsymbol{\gamma}_1^\star, \cdots, \boldsymbol{\gamma}_r^\star\right],
\end{aligned}$$

we have

$$\begin{aligned}
\mathbf{A}^{t+1} - \mathbf{A} =& (\mathbf{A}^t - \mathbf{A})\Big(\mathbf{I} - \rho\Big[(\mathbf{V}^\top\mathbf{V}) \odot (\mathbf{W}^\top\mathbf{W})\Big]\Big),\\
\mathbf{B}^{t+1} - \mathbf{B} =& (\mathbf{B}^t - \mathbf{B})\big(\mathbf{I} - \rho[(\mathbf{U}^\top\mathbf{U}) \odot (\mathbf{W}^\top\mathbf{W})]\big) - \rho\mathbf{V}\Big[((\mathbf{A}^t - \mathbf{A})^\top\mathbf{U}) \odot (\mathbf{W}^\top\mathbf{W})\Big],\\
\mathbf{C}^{t+1} - \mathbf{C} =& (\mathbf{C}^t - \mathbf{C})(\mathbf{I} - \rho[(\mathbf{U}^\top\mathbf{U}) \odot (\mathbf{V}^\top\mathbf{V})])\\
&- \rho\mathbf{W}\Big\{[((\mathbf{A}^t - \mathbf{A})^\top\mathbf{U}) \odot (\mathbf{V}^\top\mathbf{V})] + [(\mathbf{U}^\top\mathbf{U}) \odot ((\mathbf{B}^t - \mathbf{B})^\top\mathbf{V})]\Big\}.
\end{aligned} \tag{C.5}$$

Denoting $e_a^t = \|\mathbf{A}^t - \mathbf{A}\|, e_b^t = \|\mathbf{B}^t - \mathbf{B}\|, e_c^t = \|\mathbf{C}^t - \mathbf{C}\|$ and

$$\tilde{\rho} := \rho \min \begin{Bmatrix} \lambda_{\min}((\mathbf{V}^\top\mathbf{V}) \odot (\mathbf{W}^\top\mathbf{W})) \\ \lambda_{\min}((\mathbf{U}^\top\mathbf{U}) \odot (\mathbf{W}^\top\mathbf{W})) \\ \lambda_{\min}((\mathbf{U}^\top\mathbf{U}) \odot (\mathbf{V}^\top\mathbf{V})) \end{Bmatrix},$$

it follows from (C.5) that

$$\begin{aligned}
e_a^{t+1} &\le (1 - \tilde{\rho})e_a^t,\\
e_b^{t+1} &\le \rho\|\mathbf{U}\|\|\mathbf{V}\|\|\mathbf{W}\|^2 e_a^t + (1 - \tilde{\rho})e_b^t,\\
e_c^{t+1} &\le \rho\|\mathbf{U}\|^2\|\mathbf{V}\|\|\mathbf{W}\|e_a^t + \rho\|\mathbf{U}\|^2\|\mathbf{V}\|\|\mathbf{W}\|e_b^t + (1 - \tilde{\rho})e_c^t,
\end{aligned} \tag{C.6}$$

where we have used that $\|\mathbf{P} \odot \mathbf{Q}\| \le \|\mathbf{P} \otimes \mathbf{Q}\| = \|\mathbf{P}\|\|\mathbf{Q}\|$. Converting (C.6) into matrix form gives

$$\begin{bmatrix} e_a^{t+1} \\ e_b^{t+1} \\ e_c^{t+1} \end{bmatrix} \le \begin{bmatrix} 1 - \tilde{\rho} & 0 & 0 \\ \rho\|\mathbf{U}\|\|\mathbf{V}\|\|\mathbf{W}\|^2 & 1 - \tilde{\rho} & 0 \\ \rho\|\mathbf{U}\|\|\mathbf{W}\|\|\mathbf{V}\|^2 & \rho\|\mathbf{U}\|^2\|\mathbf{V}\|\|\mathbf{W}\| & 1 - \tilde{\rho} \end{bmatrix} \begin{bmatrix} e_a^t \\ e_b^t \\ e_c^t \end{bmatrix},$$

where the lower triangular system matrix share the same value

$$\eta = 1 - \tilde{\rho} \in \left[1 - \rho\left(1 + \frac{\kappa(\log n)\sqrt{r}}{n}\right), 1 - \rho\left(1 - \frac{\kappa(\log n)\sqrt{r}}{n}\right)\right] \subset (0, 1) \tag{C.7}$$

where '$\in$' follows from applying Weyl's inequality to (2.3) in Assumption III and '$\subset$' holds for any $\rho \in \left(0, (1+\frac{\kappa(\log n)\sqrt{r}}{n})^{-1}\right)$.

Therefore, the error sequence (e_a^t, e_b^t, e_c^t) is convergent to $(0,0,0)$ geometrically with a rate $\eta \in (0,1)$. Thus,

$$\lim_{t\to\infty}(\mathbf{A}^t, \mathbf{B}^t, \mathbf{C}^t) = (\mathbf{A}, \mathbf{B}, \mathbf{C}).$$

C.3 *Convergence of* $\{\|\mathbf{A}^t - \mathbf{A}^{t-1}\|\}, \{\|\mathbf{B}^t - \mathbf{B}^{t-1}\|\}, \{\|\mathbf{C}^t - \mathbf{C}^{t-1}\|\}$

Subtracting the following two consecutive iterations for $\{\mathbf{A}^t\}$ in (C.5):

$$\mathbf{A}^{t+1} - \mathbf{A} = (\mathbf{A}^t - \mathbf{A})\Big(\mathbf{I} - \rho\Big[(\mathbf{V}^\top\mathbf{V}) \odot (\mathbf{W}^\top\mathbf{W})\Big]\Big)$$

$$\mathbf{A}^t - \mathbf{A} = (\mathbf{A}^{t-1} - \mathbf{A})\Big(\mathbf{I} - \rho\Big[(\mathbf{V}^\top\mathbf{V}) \odot (\mathbf{W}^\top\mathbf{W})\Big]\Big) \quad \Longrightarrow \mathbf{A}^{t+1} - \mathbf{A}^t = (\mathbf{A}^t - A^{t-1})$$

$$\Big(\mathbf{I} - \rho\Big[(\mathbf{V}^\top\mathbf{V}) \odot (\mathbf{W}^\top\mathbf{W})\Big]\Big).$$

Similar manipulations applied to $\{\mathbf{B}^t\}$ and $\{\mathbf{C}^t\}$ lead to

$$\mathbf{B}^{t+1} - \mathbf{B}^t = (\mathbf{B}^t - \mathbf{B}^{t-1})(\mathbf{I} - \rho\Big[(\mathbf{U}^\top\mathbf{U}) \odot (\mathbf{W}^\top\mathbf{W})\Big]) - \rho\mathbf{V}\Big[((\mathbf{A}^t - \mathbf{A}^{t-1})^\top\mathbf{U}) \odot (\mathbf{W}^\top\mathbf{W})\Big],$$

$$\mathbf{C}^{t+1} - \mathbf{C}^t = (\mathbf{C}^t - \mathbf{C}^{t-1})(\mathbf{I} - \rho\Big[(\mathbf{U}^\top\mathbf{U}) \odot (\mathbf{V}^\top\mathbf{V})\Big])$$

$$- \rho\mathbf{W}\Big\{\Big[((\mathbf{A}^t - \mathbf{A}^{t-1})^\top\mathbf{U}) \odot (\mathbf{V}^\top\mathbf{V})\Big] + \Big[(\mathbf{U}^\top\mathbf{U}) \odot ((\mathbf{B}^t - \mathbf{B}^{t-1})^\top\mathbf{V})\Big]\Big\}$$

Defining $\hat{e}_a^t = \|\mathbf{A}^t - \mathbf{A}^{t-1}\|, \hat{e}_b^t = \|\mathbf{B}^t - \mathbf{B}^{t-1}\|, \hat{e}_c^t = \|\mathbf{C}^t - \mathbf{C}^{t-1}\|$, we can get the same form as (C.6) and therefore claim that $(\hat{e}_a^t, \hat{e}_b^t, \hat{e}_c^t)$ converge to $(0,0,0)$ geometrically with the same rate $\eta \in (0,1)$ in (C.7).

C.4 *Controlling the accumulative errors*

The geometric convergence of $\{\|\mathbf{C}^t - \mathbf{C}^{t-1}\|\}$ implies

$$\|\mathbf{C}^t - \mathbf{C}^{t-1}\| \le \eta^{t-1}\|\mathbf{C}^1 - \mathbf{C}^0\|$$

which implies that

$$\|\mathbf{C}^t - \mathbf{C}^0\| \le \sum_{s=0}^{t-1}\|\mathbf{C}^{s+1} - \mathbf{C}^s\| \le \sum_{s=0}^{t-1}\eta^s\|\mathbf{C}^1 - \mathbf{C}^0\| \le \frac{1}{1-\eta}\|\mathbf{C}^1 - \mathbf{C}^0\|.$$

Let t go to infinity:

$$\|\mathbf{C} - \mathbf{C}^0\| \le \frac{1}{1-\eta}\|\mathbf{C}^1 - \mathbf{C}^0\|. \tag{C.8}$$

We next bound $\|\mathbf{C}^1 - \mathbf{C}^0\|$. From (C.2), we have

$$\boldsymbol{\gamma}_q^1 - \boldsymbol{\gamma}_q^0 = \rho(\mathcal{Q}_3^0 \times_1 \boldsymbol{u}_q^\star \times_2 \boldsymbol{v}_q^\star - \boldsymbol{w}_q^\star) = \rho\left(\sum_{p=1}^{r}\langle \boldsymbol{u}_p^\star, \boldsymbol{u}_q^\star\rangle\langle \boldsymbol{v}_p^\star, \boldsymbol{v}_q^\star\rangle \boldsymbol{w}_p^\star - \boldsymbol{w}_q^\star\right)$$

$$\Longrightarrow \mathbf{C}^1 - \mathbf{C}^0 = \rho\mathbf{W}((\mathbf{U}^\top\mathbf{U}) \odot (\mathbf{V}^\top\mathbf{V}) - \mathbf{I}).$$

Then from Assumptions II and III, we have

$$\|\mathbf{C}^1 - \mathbf{C}^0\| \le \rho\|\mathbf{W}\|\|(\mathbf{U}^\top\mathbf{U}) \odot (\mathbf{V}^\top\mathbf{V}) - \mathbf{I}\| \le \rho\left(1 + c\sqrt{\frac{r}{n}}\right)\frac{\kappa(\log n)\sqrt{r}}{n}. \tag{C.9}$$

Combining All Finally, combining (C.7), (C.8) and (C.9) and using $\mathbf{C}_0 = \frac{1}{3}\mathbf{W}$, we have

$$\left\|\mathbf{C} - \frac{1}{3}\mathbf{W}\right\| \le \frac{1 + c\sqrt{\frac{r}{n}}}{1 - \frac{\kappa(\log n)\sqrt{r}}{n}}\frac{\kappa(\log n)\sqrt{r}}{n} \le 2\left(1 + c\sqrt{\frac{r}{n}}\right)\frac{\kappa(\log n)\sqrt{r}}{n} = 2\kappa(\log n)\left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}}\right)$$

where the second inequality follows from the assumption $r = o(n^2/\kappa(\log n)^2)$, which implies $1 - \frac{\kappa(\log n)\sqrt{r}}{n} \ge \frac{1}{2}$ for a sufficiently large n. Similar arguments and bounds apply to $\|\mathbf{A} - \frac{1}{3}\mathbf{U}\|$ and $\|\mathbf{B} - \frac{1}{3}\mathbf{V}\|$.

D. Proof of Lemma 5.4

Proof. The following lemma is required in the proof of Lemma 5.4. Let us first admit Lemma D.1 to prove Lemma 5.4. Since q is the sum of two parts given in (5.10) and (5.11), to bound $|q|$, we will control these parts separately. □

LEMMA D.1. Under Assumptions I and II, if $r \le n^{1.25-1.5r_c}$ with$r_c \in (0, 1/6)$, then for any integer $p \ge 3$,

$$\|\mathbf{U}^\top\|_{2\to p} \le 1 + \frac{1}{p}\tau(\log n)n^{-r_c}$$

The same bounds hold for $\mathbf{V}$ and $\mathbf{W}$. Here, we define $\|\mathbf{H}\|_{2\to p} := \sup\{\|\mathbf{H}\boldsymbol{x}\|_p : \boldsymbol{x} \in \mathbb{S}^{n-1}\}$.

Proof. Proof of Lemma D.1 See Appendix D.1. □

Bounding absolute value of (5.10):

$$\begin{aligned}
\sum_{p=1}^{r} |\langle \boldsymbol{\alpha}_p^\star - \frac{1}{3}\boldsymbol{u}_p^\star, \boldsymbol{u}\rangle\langle \boldsymbol{v}_p^\star, \boldsymbol{v}\rangle\langle \boldsymbol{w}_p^\star, \boldsymbol{w}\rangle| &\le \sqrt{\sum_{p=1}^{r}\langle \boldsymbol{\alpha}_p^\star - \frac{1}{3}\boldsymbol{u}_p^\star, \boldsymbol{u}\rangle^2}\sqrt{\sum_{p=1}^{r}\langle \boldsymbol{v}_p^\star, \boldsymbol{v}\rangle^2\langle \boldsymbol{w}_p^\star, \boldsymbol{w}\rangle^2} \\
&\le \sqrt{\sum_{p=1}^{r}\langle \boldsymbol{\alpha}_p^\star - \frac{1}{3}\boldsymbol{u}_p^\star, \boldsymbol{u}\rangle^2}\sqrt[4]{\sum_{p=1}^{r}\langle \boldsymbol{v}_p^\star, \boldsymbol{v}\rangle^4}\sqrt[4]{\sum_{p=1}^{r}\langle \boldsymbol{w}_p^\star, \boldsymbol{w}\rangle^4} \\
&= \|(\mathbf{A} - \frac{1}{3}\mathbf{U})^\top\boldsymbol{u}\|_2\|\mathbf{V}^\top\boldsymbol{v}\|_4\|\mathbf{W}^\top\boldsymbol{w}\|_4 \\
&\le \|\mathbf{A} - \frac{1}{3}\mathbf{U}\|\|\mathbf{V}^\top\|_{2\to 4}\|\mathbf{W}^\top\|_{2\to 4} \\
&\le 2\kappa(\log n)\left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}}\right)(1 + o(1)) \\
&= o(1),
\end{aligned}$$

where the last second line follows from Lemma 5.3 and Lemma D.1 when $r \ll n^{1.25}$ (by letting r_c in '$r \ll n^{1.25-r_c}$' approach to zero). The last line holds for $r \ll \frac{n^{1.5}}{\kappa(\log n)}$.

Similar bounds hold for the other two terms in (5.10).

Bounding the absolute value of (5.11): First of all, for any $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \in \mathcal{F}(\delta)$, there exists a division of $[r] = \Omega_u \cup \Omega_v \cup \Omega_w$ such that

$$\begin{aligned} &|\langle \boldsymbol{u}_p^\star, \boldsymbol{u}\rangle| \le \delta, \quad \forall p \in \Omega_u, \\ &|\langle \boldsymbol{v}_p^\star, \boldsymbol{v}\rangle| \le \delta, \quad \forall p \in \Omega_v, \\ &|\langle \boldsymbol{w}_p^\star, \boldsymbol{u}\rangle| \le \delta, \quad \forall p \in \Omega_w. \end{aligned} \tag{D.1}$$

We will denote by $\mathbf{U}_{\Omega_u}$ the submatrix of $\mathbf{U}$ forming from those columns of $\mathbf{U}$ with indexes in Ω_u. Similarly, we can define $\mathbf{V}_{\Omega_v}$ and $\mathbf{W}_{\Omega_w}$. With these preparation, we have that

$$\begin{aligned} \sum_{p=1}^{r} |\langle \boldsymbol{u}_p^\star, \boldsymbol{u}\rangle \langle \boldsymbol{v}_p^\star, \boldsymbol{v}\rangle \langle \boldsymbol{w}_p^\star \boldsymbol{w}\rangle| &= \sum_{p \in \Omega_u \cup \Omega_v \cup \Omega_w} |\langle \boldsymbol{u}_p^\star, \boldsymbol{u}\rangle \langle \boldsymbol{v}_p^\star, \boldsymbol{v}\rangle \langle \boldsymbol{w}_p^\star \boldsymbol{w}\rangle| \\ &\le \delta(\|\mathbf{V}_{\Omega_u}\|\|\mathbf{W}_{\Omega_u}\| + \|\mathbf{U}_{\Omega_v}\|\|\mathbf{W}_{\Omega_v}\| + \|\mathbf{U}_{\Omega_w}\|\|\mathbf{V}_{\Omega_w}\|) \\ &\le 3\delta\left(1 + c\sqrt{\frac{r}{n}}\right)^2 \\ &\le 12\delta \max\{1, c^2 r/n\} \\ &\le \frac{1}{2}, \end{aligned}$$

where the first inequality follows from (D.1) and $\sum_{p\in\Omega_u} |\langle \boldsymbol{v}_p^\star, \boldsymbol{v}\rangle \langle \boldsymbol{w}_p^\star, \boldsymbol{w}\rangle| \le \|\mathbf{V}_{\Omega_u}\|\|\mathbf{W}_{\Omega_u}\|$, etc. The second inequality uses the fact that the spectral norm of any submatrix is smaller than the original one and Assumption II. The last inequality holds when $\delta \le \frac{1}{24}$ and and $r \le n/(24\delta c^2)$.

Combining All Under Assumptions I, II, III, if $r \ll n^{1.25}$ and $r \le \frac{n}{24\delta c^2}$ for $\delta \in (0, \frac{1}{24}]$, we have $|q| \le o(1) + \frac{1}{2} < 1$ in $\mathcal{F}(\delta)$ for sufficiently large n.

D.1 *Proof of Lemma D.1*

The proof refines the one for Lemma 4 of [1]. We only prove it for $\mathbf{U}$ since the same arguments apply to $\mathbf{W}$ and $\mathbf{V}$. We start with a general integer $p \ge 3$.

$$\|\mathbf{U}^\top\|_{2\to p} = \sup_{\boldsymbol{x}\in\mathbb{S}^{n-1}} \|\mathbf{U}^\top \boldsymbol{x}\|_p := \|\mathbf{U}^\top \boldsymbol{x}^\star\|_p \tag{D.2}$$

where we define $\boldsymbol{x}^\star \in \mathbb{S}^{n-1}$ to be the optimal solution of $\sup_{\boldsymbol{x}\in\mathbb{S}^{n-1}} \|\mathbf{U}^\top \boldsymbol{x}\|_p^p$. Further note that

$$\|\mathbf{U}^\top \boldsymbol{x}^\star\|_p^p = \|\mathbf{U}_S^\top \boldsymbol{x}^\star\|_p^p + \|\mathbf{U}_{S^c}^\top \boldsymbol{x}^\star\|_p^p \tag{D.3}$$

where S denotes the indices of the largest (in absolute value) L entries of $\mathbf{U}^\top \boldsymbol{x}^\star$ and $\mathbf{U}_S$ denotes the column submatrix of $\mathbf{U}$ indexed by S. Similar notations apply to its complement set $S^c = [r] \setminus S$.

Bound the first term:

$$\|\mathbf{U}_S^\top \boldsymbol{x}^\star\|_p^p \le \|\mathbf{U}_S^\top \boldsymbol{x}^\star\|_2^2 \le \|\mathbf{U}_S \mathbf{U}_S^\top\| \le 1 + \sum_{i\in S\setminus\{j\}} |\langle \boldsymbol{u}_i, \boldsymbol{u}_j\rangle| \le 1 + (L-1)\frac{\tau(\log n)}{\sqrt{n}}. \tag{D.4}$$

Note this upper-bound is independent of p. Here, the first inequality is because $|\boldsymbol{u}_i^{\star\top}\boldsymbol{x}^\star| \le \|\boldsymbol{u}_i^\star\|_2 \|\boldsymbol{x}^\star\|_2 = 1$ and the last second inequality follows from Gershgorin's circle theorem. Finally, the last inequality is from Assumption I and L being the cardinality of the set S.

Bound the second term: First note that

$$\min_{i\in S} |\boldsymbol{u}_i^\top \boldsymbol{x}^\star|^2 \le \frac{1}{L}\sum_{i\in S} |\boldsymbol{u}_i^\top \boldsymbol{x}^\star|^2 \le \frac{1}{L}\|\mathbf{U}_S\mathbf{U}_S^\top\| \|\boldsymbol{x}^\star\|_2^2 \le \frac{1}{L}(1+o(1)) \le \frac{2}{L}$$

for sufficiently large n. The last second inequality follows from (D.4) and an additional assumption on L

$$(L-1)\frac{\tau(\log n)}{\sqrt{n}} = o(1). \tag{D.5}$$

We conclude that

$$\max_{i\in S^c} |\boldsymbol{u}_i^\top \boldsymbol{x}^\star|^2 \le \min_{i\in S} |\boldsymbol{u}_i^\top \boldsymbol{x}^\star|^2 \le \frac{2}{L},$$

since S consists of the indices of the L largest (in absolute value) elements of $\mathbf{U}^\top \boldsymbol{x}^\star$. As a consequence, we have

$$\begin{aligned}\|\mathbf{U}_{S^c}^\top \boldsymbol{x}^\star\|_p^p &= \sum_{i\notin S} |\boldsymbol{u}_i^\top \boldsymbol{x}^\star|^p \\ &\le \big(\max_{i\notin S} |\boldsymbol{u}_i^\top \boldsymbol{x}^\star|^{p-2}\big) \sum_{i\notin S} |\boldsymbol{u}_i^\top \boldsymbol{x}|^2 \quad = \big(\max_{i\notin S} |\boldsymbol{u}_i^\top \boldsymbol{x}^\star|^{p-2}\big)\|\mathbf{U}_{S^c}^\top x^\star\|_2^2 \le \left(\frac{2}{L}\right)^{\frac{p}{2}-1}\left(1 + c\sqrt{\frac{r}{n}}\right)^2\end{aligned} \tag{D.6}$$

where the last inequality follows from the fact that $\|\mathbf{U}_{S^c}^\top \boldsymbol{x}^\star\|_2^2 \le \|\mathbf{U}_{S^c}\|^2 \le \|\mathbf{U}\|^2 \le (1 + c\sqrt{\frac{r}{n}})^2$ by Assumption II. Furthermore, since $(1 + c\sqrt{\frac{r}{n}})^2 \le 4\max\{1, c^2\frac{r}{n}\}$, $c^2\frac{r}{n} \le c^2 n^{0.25-1.5r_c}$ from the condition of $r \le n^{1.25-1.5r_c}$, and $1 \ll c^2 n^{0.25-1.5r_c}$ for $r_c \in (0, 1/6)$, we have $(1 + c\sqrt{\frac{r}{n}})^2 \le 4c^2 n^{0.25-1.5r_c}$ for $r_c \in (0, 1/6)$. So from (D.6), we get

$$\|\mathbf{U}_{S^c}^\top \boldsymbol{x}^\star\|_p^p \le 4\left(\frac{2}{L}\right)^{\frac{p}{2}-1} c^2 n^{0.25-1.5r_c}. \tag{D.7}$$

From (D.3), (D.4) and (D.7), we have

$$\|\mathbf{U}^\top \boldsymbol{x}^\star\|_p^p \le 1 + (L-1)\frac{\tau(\log n)}{\sqrt{n}} + 4\left(\frac{2}{L}\right)^{\frac{p}{2}-1} c^2 n^{0.25-1.5r_c}.$$

By choosing $L = \left\lceil \frac{1}{2} n^{0.5-r_c} \right\rceil \Rightarrow \begin{cases} L \le \frac{1}{2} n^{0.5-r_c} + 1 \\ L \ge \frac{1}{2} n^{0.5-r_c} \end{cases}$ (which satisfies the condition (D.5)), we have that

$$\|\mathbf{U}^\top \boldsymbol{x}^\star\|_p^p \le 1 + \frac{1}{2}\tau(\log n) n^{-r_c} + 4^{\frac{p}{2}} c^2 n^{(\frac{3}{4}-\frac{p}{4})+(\frac{p}{2}-\frac{5}{2})r_c}.$$

Then from the assumptions $p \ge 3$ and $r_c \in (0, \frac{1}{6})$, we get

$$\left(\frac{3}{4} - \frac{p}{4}\right) + \left(\frac{p}{2} - \frac{5}{2}\right) r_c \le \left(\frac{3}{4} - \frac{p}{4}\right) 6 r_c + \left(\frac{p}{2} - \frac{5}{2}\right) r_c = (2-p) r_c \le -r_c. \tag{D.8}$$

So, we have

$$\|\mathbf{U}^\top \boldsymbol{x}^\star\|_p^p \le 1 + \left(\frac{1}{2}\tau(\log n) + 4^{\frac{p}{2}} c^2\right) n^{-r_c}.$$

Since $4^{\frac{p}{2}} c^2 \ll \frac{1}{2}\tau(\log n)$ and $(1+t)^{1/p} \le 1 + \frac{1}{p} t$ for all $t \ge 0$, then

$$\|\mathbf{U}^\top \boldsymbol{x}^\star\|_p \le 1 + \frac{1}{p}\tau(\log n) n^{-r_c}$$

holds for any $p \ge 3$. This completes the proof since $\|\mathbf{U}^\top\|_{2\to p} = \|\mathbf{U}^\top \boldsymbol{x}^\star\|_p$ by (D.2).

E. Proof of Lemma 5.5

Proof. We start by the angular dual polynomial

$$\begin{aligned} q(\boldsymbol{u}(\theta_1), \boldsymbol{v}(\theta_2), \boldsymbol{w}(\theta_3)) = &\cos(\theta_1)\cos(\theta_2)\cos(\theta_3) \\ &+ q(\boldsymbol{u}_1^\star, \boldsymbol{y}, \boldsymbol{z})\cos(\theta_1)\sin(\theta_2)\sin(\theta_3) + q(\boldsymbol{x}, \boldsymbol{v}_1^\star, \boldsymbol{z})\sin(\theta_1)\cos(\theta_2)\sin(\theta_3) \\ &+ q(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}_1^\star)\sin(\theta_1)\sin(\theta_2)\cos(\theta_3) + q(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})\sin(\theta_1)\sin(\theta_2)\sin(\theta_3). \end{aligned}$$

To bound q, we only need to bound the coefficients $q(\boldsymbol{u}_1^\star, \boldsymbol{y}, \boldsymbol{z})$, $q(\boldsymbol{x}, \boldsymbol{v}_1^\star, \boldsymbol{z})$, $q(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}_1^\star)$, and $q(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$.

We first show that $q(\boldsymbol{u}_1^\star, \boldsymbol{y}, \boldsymbol{z})$, $q(\boldsymbol{x}, \boldsymbol{v}_1^\star, \boldsymbol{z})$, and $q(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}_1^\star)$ are close to zero. To see this, we examine

$$q(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}_1^\star) = \sum_{p=1}^{r} [\langle \boldsymbol{\alpha}_p^\star, \boldsymbol{x}\rangle \langle \boldsymbol{v}_p^\star, \boldsymbol{y}\rangle \langle \boldsymbol{w}_p^\star, \boldsymbol{w}_1^\star\rangle + \langle \boldsymbol{u}_p^\star, \boldsymbol{x}\rangle \langle \boldsymbol{\beta}_p^\star, \boldsymbol{y}\rangle \langle \boldsymbol{w}_p^\star, \boldsymbol{w}_1^\star\rangle + \langle \boldsymbol{u}_p^\star, \boldsymbol{x}\rangle \langle \boldsymbol{v}_p^\star, \boldsymbol{y}\rangle \langle \boldsymbol{\gamma}_p^\star, \boldsymbol{w}_1^\star\rangle]$$

$$= \boldsymbol{x}^\top [\mathbf{A}\ \text{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top + \mathbf{U}\ \text{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star)\mathbf{B}^\top + \mathbf{U}\ \text{diag}\ (\mathbf{C}^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top] \boldsymbol{y}$$

$$= \boldsymbol{x}^\top \left(\mathbf{A}\ \text{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\boldsymbol{u}_1^\star \boldsymbol{v}_1^\star + \mathbf{U}\ \text{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star)\mathbf{B}^\top - \frac{1}{3}\boldsymbol{u}_1^\star \boldsymbol{v}_1^\star + \mathbf{U}\ \text{diag}\ (\mathbf{C}^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\boldsymbol{u}_1^\star \boldsymbol{v}_1^\star\right) \boldsymbol{y},$$

since $\boldsymbol{x} \perp \boldsymbol{u}_1^\star, \boldsymbol{y} \perp \boldsymbol{v}_1^\star$. This implies

$$\begin{aligned} |q(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}_1^\star)| \le &\left\| \mathbf{A}\ \text{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\boldsymbol{u}_1^\star \boldsymbol{v}_1^\star \right\| \\ &+ \left\| \mathbf{U}\ \text{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star)\mathbf{B}^\top - \frac{1}{3}\boldsymbol{u}_1^\star \boldsymbol{v}_1^\star \right\| + \left| \boldsymbol{x}^\top \left(\mathbf{U}\ \text{diag}\ (\mathbf{C}^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\boldsymbol{u}_1^\star \boldsymbol{v}_1^\star\right) \boldsymbol{y} \right|. \end{aligned}$$

□

We first bound $\left\|\mathbf{A}\ \mathrm{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\boldsymbol{u}_1^\star \boldsymbol{v}_1^\star\right\|$.

$$\begin{aligned}\left\|\mathbf{A}\ \mathrm{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\boldsymbol{u}_1^\star \boldsymbol{v}_1^{\star\top}\right\| \le& \left\|\mathbf{A}\ \mathrm{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\mathbf{U}\ \mathrm{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top\right\| \\ &+ \left\|\frac{1}{3}\mathbf{U}\ \mathrm{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\boldsymbol{u}_1^\star \boldsymbol{v}_1^{\star\top}\right\| \\ \le& \left\|\mathbf{A} - \frac{1}{3}\mathbf{U}\right\| \|\ \mathrm{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star)\|\|\mathbf{V}\| + \frac{1}{3}\|\mathbf{U}\|\|\ \mathrm{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star - \mathbf{e}_1)\|\mathbf{V}^\top\| \\ \le& 2\kappa(\log n)\left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}}\right)\left(1 + c\sqrt{\frac{r}{n}}\right) + \frac{\tau(\log n)}{3\sqrt{n}}\left(1 + c\sqrt{\frac{r}{n}}\right)^2 \\ =& \left[2\kappa(\log n)\frac{\sqrt{r}}{n} + \frac{\tau(\log n)}{3\sqrt{n}}\right]\left(1 + c\sqrt{\frac{r}{n}}\right)^2,\end{aligned}$$

where the third inequality first uses the facts $\|\ \mathrm{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star)\| = 1$ and $\|\ \mathrm{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star - \mathbf{e}_1)\| = \max_{p\neq 1}|\langle \boldsymbol{w}_p^\star, \boldsymbol{w}_1^\star\rangle|$ and then follows from Assumptions I and II and Lemma 5.3.

Similarly,

$$\left\|\mathbf{U}\ \mathrm{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star)\mathbf{B}^\top - \frac{1}{3}\boldsymbol{u}_1^\star \boldsymbol{v}_1^\star\right\| \le \left[2\kappa(\log n)\frac{\sqrt{r}}{n} + \frac{\tau(\log n)}{3\sqrt{n}}\right]\left(1 + c\sqrt{\frac{r}{n}}\right)^2.$$

The similar arguments also apply to bounding $|\boldsymbol{x}^\top(\mathbf{U}\ \mathrm{diag}\ (\mathbf{C}^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\boldsymbol{u}_1^\star \boldsymbol{v}_1^\star)\boldsymbol{y}|$. Note that

$$\begin{aligned}\boldsymbol{x}^\top\left(\mathbf{U}^\star\ \mathrm{diag}\ (\mathbf{C}^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\boldsymbol{u}_1^\star \boldsymbol{v}_1^{\star\top}\right)\boldsymbol{y} =& \boldsymbol{x}^\top(\mathbf{U}\ \mathrm{diag}\ ((\mathbf{C} - \mathbf{W}/3)^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top)\boldsymbol{y} \\ &+ \frac{1}{3}\boldsymbol{x}^\top(\mathbf{U}\ \mathrm{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star - \mathbf{e}_1)\mathbf{V}^\top)\boldsymbol{y}\end{aligned}$$

and the first term can be rewritten as

$$\begin{aligned}\boldsymbol{x}^\top(\mathbf{U}\ \mathrm{diag}\ ((\mathbf{C} - \mathbf{W}/3)^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top)\boldsymbol{y} &= \sum_{i=1}^r \boldsymbol{x}^\top\left((\mathbf{c}_i - \boldsymbol{w}_i/3)^\top \boldsymbol{w}_1^\star \boldsymbol{u}_i \boldsymbol{v}_i^\top\right)\boldsymbol{y} \\ &= \sum_{i=1}^r (\boldsymbol{x}^\top \boldsymbol{u}_i)(\boldsymbol{v}_i^\top \boldsymbol{y})(\mathbf{c}_i - \boldsymbol{w}_i/3)^\top \boldsymbol{w}_1^\star) \\ &= \boldsymbol{x}^\top \sum_{i=1}^r \left(\boldsymbol{u}_i(\boldsymbol{v}_i^\top \boldsymbol{y})(\mathbf{c}_i - \boldsymbol{w}_i/3)^\top\right)\boldsymbol{w}_1^\star \\ &= \boldsymbol{x}^\top\left(\mathbf{U}\ \mathrm{diag}\ (\mathbf{V}^\top \boldsymbol{y})(\mathbf{C} - \mathbf{W}/3)^\top\right)\boldsymbol{w}_1^\star,\end{aligned}$$

and so

$$\left|\boldsymbol{x}^\top\left(\mathbf{U}^\star\ \mathrm{diag}\ (\mathbf{C}^\top \boldsymbol{w}_1^\star)\mathbf{V}^\top - \frac{1}{3}\boldsymbol{u}_1^\star \boldsymbol{v}_1^{\star\top}\right)\boldsymbol{y}\right| \le \|\mathbf{U}\|\|\ \mathrm{diag}\ (\mathbf{V}^\top \boldsymbol{y})\|\|\mathbf{C} - \mathbf{W}/3\| + \frac{1}{3}\|\mathbf{U}\|\|\ \mathrm{diag}\ (\mathbf{W}^\top \boldsymbol{w}_1^\star - \mathbf{e}_1)\|\mathbf{V}^\top\|.$$

Finally, we obtain

$$|q(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}_1^\star)| \le \left[6\kappa(\log n)\frac{\sqrt{r}}{n} + \frac{\tau(\log n)}{\sqrt{n}}\right]\left(1 + c\sqrt{\frac{r}{n}}\right)^2 = O\left(\frac{\kappa(\log n)\sqrt{r}}{n}, \frac{\tau(\log n)}{\sqrt{n}}, \frac{\kappa(\log n) r^{1.5}}{n^2}, \frac{\tau(\log n) r}{n^{1.5}}\right)$$
$$= O\left(\frac{\kappa(\log n)}{n^{3/8+\frac{3}{4}r_c}}, \frac{\tau(\log n)}{n^{5/8-\frac{3}{4}r_c}}, \frac{\kappa(\log n)}{n^{1/8+\frac{9}{4}r_c}}, \frac{\tau(\log n)}{n^{\frac{1}{4}+1.5r_c}}\right)$$
$$= O(\kappa(\log n) n^{-3r_c}, \tau(\log n) n^{-3r_c}) = o(n^{-2r_c})$$

with the notation $O(f(n), g(n)) := \max\{O(f(n)), O(g(n))\}$. Then the last second line holds if $r \le n^{1.25-1.5r_c}$ and the last line follows from the assumption $r_c \in (0, 1/6)$.

The same bound holds for $|q(\boldsymbol{x}, \boldsymbol{v}_1^\star, \boldsymbol{z})|$ and $|q(\boldsymbol{u}_1^\star, \boldsymbol{y}, \boldsymbol{z})|$.

The coefficient of the last term of (5.17) is $q(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ and its absolute value is bounded by the tensor spectral norm of $\boldsymbol{\mathcal{Q}}$ and should be close to constant as $\boldsymbol{\mathcal{Q}}$ is close to $\sum_{p=1}^r \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star$, the spectral norm of which is $1 + O(n^{-r_c})$ by the following lemma.

LEMMA E.1. Under Assumptions I and II, and if $r \le n^{1.25-1.5r_c}$ with $r_c \in (0, 1/6)$,

$$\left\| \sum_{p=1}^r \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star \right\| \le 1 + \frac{5}{4}\tau(\log n) n^{-r_c}.$$

Proof of Lemma E.1

$$\begin{aligned}
\left\| \sum_{p=1}^r \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star \right\| &= \sup_{(\boldsymbol{a},\mathbf{b},\mathbf{c})\in\mathbb{K}} \langle \mathbf{U}^\top \boldsymbol{a}, (\mathbf{V}^\top \mathbf{b}) \odot (\mathbf{W}^\top \mathbf{c}) \rangle \\
&\le \sup_{(\boldsymbol{a},\mathbf{b},\mathbf{c})\in\mathbb{K}} \|\mathbf{U}^\top \boldsymbol{a}\|_3 \|(\mathbf{V}^\top \mathbf{b}) \odot (\mathbf{W}^\top \mathbf{c})\|_{3/2} \\
&\le \sup_{(\boldsymbol{a},\mathbf{b},\mathbf{c})\in\mathbb{K}} \|\mathbf{U}^\top \boldsymbol{a}\|_3 \|\mathbf{V}^\top \mathbf{b}\|_3 \|\mathbf{W}^\top \boldsymbol{v}\|_3 \\
&\le \|\mathbf{U}^\top\|_{2\to 3} \|\mathbf{V}^\top\|_{2\to 3} \|\mathbf{W}^\top\|_{2\to 3} \\
&\le \left(1 + \frac{1}{3}\tau(\log n) n^{-r_c}\right)^3 \\
&= 1 + \tau(\log n) n^{-r_c} + \frac{1}{3}\tau(\log n)^2 n^{-r_c} + \frac{1}{9}\tau(\log n)^3 n^{-3r_c} \le 1 + \frac{5}{4}\tau(\log n) n^{-r_c},
\end{aligned}$$

where the first inequality follows from Hölder's inequality and the second follows from Cauchy's inequality. The fourth follows from Lemma D.1 when $r \le n^{1.25-1.5r_c}$ with $r_c \in (0, \frac{1}{6})$. The last holds since $\frac{1}{3}\tau(\log n)^2 n^{-r_c} + \frac{1}{9}\tau(\log n)^3 n^{-3r_c} \ll \frac{1}{4} n^{-r_c}$. □

It remains to bound the difference between $\mathcal{Q}$ and $\sum_{p=1}^{r} \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star$:

$$\begin{aligned}
&\left\| \mathbf{Q} - \sum_{p=1}^{r} \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star \right\| \\
&\leq \underbrace{\left\| \sum_{p=1}^{r} (\boldsymbol{\alpha}_p^\star - \frac{1}{3}\boldsymbol{u}_p^\star) \otimes \boldsymbol{v}_p^\star \otimes \boldsymbol{w}_p^\star \right\|}_{\Pi_1} + \underbrace{\left\| \sum_{p=1}^{r} \boldsymbol{u}_p^\star \otimes (\boldsymbol{\beta}_p^\star - \frac{1}{3}\boldsymbol{v}_p^\star) \otimes \boldsymbol{w}_p^\star \right\|}_{\Pi_2} + \underbrace{\left\| \sum_{p=1}^{r} \boldsymbol{u}_p^\star \otimes \boldsymbol{v}_p^\star \otimes (\boldsymbol{\gamma}_p^\star - \frac{1}{3}\boldsymbol{w}_p^\star) \right\|}_{\Pi_3}
\end{aligned}$$

First we bound Π_1:

$$\begin{aligned}
\Pi_1 &= \sup_{(\boldsymbol{a},\mathbf{b},\mathbf{c})\in\mathbb{K}} \langle (\mathbf{A} - \frac{1}{3}\mathbf{U})^\top \boldsymbol{a}, (\mathbf{V}^\top \mathbf{b}) \odot (\mathbf{W}^\top \mathbf{c}) \rangle \\
&\leq \sup_{(\boldsymbol{a},\mathbf{b},\mathbf{c})\in\mathbb{K}} \|(\mathbf{A} - \frac{1}{3}\mathbf{U})^\top \boldsymbol{x}\|_2 \|(\mathbf{V}^\top \mathbf{b}) \odot (\mathbf{W}^\top \mathbf{c})\|_2 \\
&\leq \sup_{(\boldsymbol{a},\mathbf{b},\mathbf{c})\in\mathbb{K}} \|(\mathbf{A} - \frac{1}{3}\mathbf{U})^\top \boldsymbol{x}\|_2 \|(\mathbf{V}^\top \mathbf{b})\|_4 \|(\mathbf{W}^\top \mathbf{c})\|_4 \\
&\leq \|\mathbf{A} - \frac{1}{3}\mathbf{U}\| \|\mathbf{V}^\top\|_{2\to 4} \|\mathbf{W}^\top\|_{2\to 4} \\
&\leq 2\kappa(\log n)\left(\frac{\sqrt{r}}{n} + c\frac{r}{n^{1.5}}\right)(1 + o(1)) \leq 8\kappa(\log n) \max\left\{\frac{\sqrt{r}}{n}, c\frac{r}{n^{1.5}}\right\} \leq 8\kappa(\log n) n^{-3r_c} = o(n^{-2r_c})
\end{aligned}$$

where the first and second inequalities follows from Cauchy's inequality and the fourth inequality follows from Lemma 5.3 and Lemma D.1 when $r \ll n^{1.25}$. The last inequality follows by plugging $r \leq n^{1.25-1.5r_c}$ with $r_c \in (0, \frac{1}{6})$.

The same bound also holds for Π_2 and Π_3.

Combining All If $r \leq n^{1.25-1.5r_c}$ with$r_c \in (0, 1/6)$, we have

$$\begin{aligned}
&|q(\boldsymbol{u}_1^\star, \boldsymbol{y}, \boldsymbol{z})| = o(n^{-2r_c}), \\
&|q(\boldsymbol{x}, \boldsymbol{v}_1^\star, \boldsymbol{z})| = o(n^{-2r_c}), \\
&|q(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}_1^\star)| = o(n^{-2r_c}), \\
&|q(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})| \leq 1 + \frac{5}{4}\tau(\log n) n^{-r_c} + o(n^{-2r_c})
\end{aligned} \tag{E.1}$$

which together with (5.17) gives

$$\begin{aligned}
|q(\boldsymbol{u}(\theta_1), \boldsymbol{v}(\theta_2), \boldsymbol{w}(\theta_3))| \leq& |\cos(\theta_1)\cos(\theta_2)\cos(\theta_3)| + |\sin(\theta_1)\sin(\theta_2)\sin(\theta_3)| + \frac{5}{4}\tau(\log n) n^{-r_c} \\
&+ o(n^{-2r_c}) \leq |\cos(\theta_1)\cos(\theta_2)\cos(\theta_3)| + |\sin(\theta_1)\sin(\theta_2)\sin(\theta_3)| + \frac{4}{3}\tau(\log n) n^{-r_c}
\end{aligned}$$

where the last inequality follows from $o(n^{-2r_c}) \ll \frac{1}{12}\tau(\log n) n^{-r_c}$.

F. Proof of Lemma 5.6

Proof. Recall that

$$\begin{aligned} F(\theta_1,\theta_2,\theta_3) &= \cos(\theta_1)\cos(\theta_2)\cos(\theta_3) + q(\boldsymbol{u}_1^\star,\boldsymbol{y},\boldsymbol{z})\cos(\theta_1)\sin(\theta_2)\sin(\theta_3) \\ &+ q(\boldsymbol{x},\boldsymbol{v}_1^\star,\boldsymbol{z})\sin(\theta_1)\cos(\theta_2)\sin(\theta_3) + q(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}_1^\star)\sin(\theta_1)\sin(\theta_2)\cos(\theta_3) \\ &+ q(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\sin(\theta_1)\sin(\theta_2)\sin(\theta_3). \end{aligned} \tag{F.1}$$

The points of special interest are the eight vertices of the cube $[0,\pi]\times[0,\pi]\times[0,\pi]$, i.e.

$$\{(\theta_1,\theta_2,\theta_3):\theta_i\in\{0,\pi\}, i=1,2,3\}$$

which we classify into two sets:

(1) The first set of vertices involve an even number of π: $(0,0,0)$, $(0,\pi,\pi)$, $(\pi,0,\pi)$, $(\pi,\pi,0)$;

(2) The second set of vertices involve an odd number of π: $(\pi,0,0)$, $(0,\pi,0)$, $(0,0,\pi)$, (π,π,π).

□

F.1 *Control the first vertex set*

For the first set of points, we only show that

$$F(\theta_1+\xi_1,\theta_2+\xi_2,\theta_3+\xi_3)\le 1,\quad \forall\xi_i\in\left(-\frac{\sqrt{2}-1}{3},\frac{\sqrt{2}-1}{3}\right)\bigcup\left(\frac{\pi}{2}-\frac{\sqrt{2}-1}{3},\frac{\pi}{2}+\frac{\sqrt{2}-1}{3}\right)$$

holds for $(\theta_1,\theta_2,\theta_3)=(0,0,0)$. The same arguments apply to the other cases $(\pi,0,\pi)$, $(0,\pi,\pi)$, $(\pi,\pi,0)$ since (F.1) implies

$$F(\xi_1,\xi_2,\xi_3) \quad = F(\xi_1,\pi+\xi_2,\pi+\xi_3) = F(\pi+\xi_1,\xi_2,\pi+\xi_3) = F(\pi+\xi_1,\pi+\xi_2,\xi_3)$$

for all $\xi_1,\xi_2\,\xi_3\in\mathbb{R}$.

Let us apply the first-order Taylor expansion to $F(\theta_1,\theta_2,\theta_3)$ over some smaller cube $[-\theta_0,\theta_0]\times[-\theta_0,\theta_0]\times[-\theta_0,\theta_0]$ with $\theta_0\in(0,\pi/2)$ to be determined later,

$$F(\theta_1,\theta_2,\theta_3) = F(0,0,0)+\boldsymbol{\theta}^\top\nabla F(\xi_1,\xi_2,\xi_3)\ge 1-\|\boldsymbol{\theta}\|_1\sup_{|\xi_1|,|\xi_2|,|\xi_3|\le\theta_0}\|\nabla F(\xi_1,\xi_2,\xi_3)\|_\infty,$$

where $\boldsymbol{\theta}=\begin{bmatrix}\theta_1 & \theta_2 & \theta_3\end{bmatrix}^\top$. Since

$$\begin{aligned} \frac{\partial}{\partial\theta_1}F(\xi_1,\xi_2,\xi_3) =& -\sin(\xi_1)\cos(\xi_2)\cos(\xi_3) - q(\boldsymbol{u}_1^\star,\boldsymbol{y},\boldsymbol{z})\sin(\xi_1)\sin(\xi_2)\sin(\xi_3) \\ &+ q(\boldsymbol{x},\boldsymbol{v}_1^\star,\boldsymbol{z})\cos(\xi_1)\cos(\xi_2)\sin(\xi_3) + q(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}_1^\star)\cos(\xi_1)\sin(\xi_2)\cos(\xi_3) \\ &+ q(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\cos(\xi_1)\sin(\xi_2)\sin(\xi_3), \end{aligned}$$

we have

$$\begin{aligned} \left|\frac{\partial}{\partial\theta_1}F(\xi_1,\xi_2,\xi_3)\right| \le& |\sin(\theta_0)| + o(1)(|\sin(\theta_0)|^3+2|\sin(\theta_0)|) + (1+o(1))|\sin(\theta_0)|^2 \\ \le& |\sin(\theta_0)| + |\sin(\theta_0)|^2 + o(1) \le 3|\sin(\theta_0)| \end{aligned}$$

where the first inequality follows from (E.1), and so

$$\begin{aligned}|q(\boldsymbol{u}_1^\star, \boldsymbol{y}, \boldsymbol{z})| &= o(1),\\ |q(\boldsymbol{x}, \boldsymbol{v}_1^\star, \boldsymbol{z})| &= o(1),\\ |q(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}_1^\star)| &= o(1),\\ |q(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})| &= 1 + o(1)\end{aligned} \tag{F.2}$$

under Assumptions I–III and $r \ll n^{1.25}$ (by letting r_c in "$r \ll n^{1.25-r_c}$" approach to zero). The last inequality uses the facts that $|\sin(\theta_0)|^2 \le |\sin(\theta_0)|$ and $o(1) \le |\sin(\theta_0)|$ for sufficiently large n. The same bound holds for $\left|\frac{\partial}{\partial\theta_2}F(\xi_1,\xi_2,\xi_3)\right|$ and $\left|\frac{\partial}{\partial\theta_3}F(\xi_1,\xi_2,\xi_3)\right|$. We therefore have

$$F(\theta_1,\theta_2,\theta_3) \ge 1 - 3\|\boldsymbol{\theta}\|_1 |\sin(\theta_0)| \ge 1 - 9\theta_0^2. \tag{F.3}$$

Let us recall the integral form of the second-order Taylor expansion of $F(\theta_1,\theta_2,\theta_3)$:

$$F(\theta_1,\theta_2,\theta_3) = F(0,0,0) + \boldsymbol{\theta}^\top \nabla F(0,0,0) + \int_0^1 \frac{t^2}{2}\boldsymbol{\theta}^\top \nabla^2 F(t\theta_1, t\theta_2, t\theta_3)\boldsymbol{\theta}\ \mathrm{d}\, t$$

As a consequence of the construction process of the dual polynomial, we have $F(0,0,0) = 1$ and $\nabla F(0,0,0) = 0$, implying

$$F(\theta_1,\theta_2,\theta_3) = 1 + \int_0^1 \frac{t^2}{2}\boldsymbol{\theta}^\top \nabla^2 F(t\theta_1, t\theta_2, t\theta_3)\boldsymbol{\theta}\ \mathrm{d}\, t$$

Therefore, as long as the Hessian matrix $\nabla^2 F$ is negative definite over the region $[-\theta_0,\theta_0]^3$ for some $\theta_0 > 0$, then $F(\theta_1,\theta_2,\theta_3) \le 1$ for any $(\theta_1,\theta_2,\theta_3) \in [-\theta_0,\theta_0]^3$ with equality holds only if $(\theta_1,\theta_2,\theta_3) = (0,0,0)$.

We next estimate the Hessian matrix $\nabla^2 F(\xi_1,\xi_2,\xi_3)$. Direct computation gives

$$\nabla^2 F(\xi_1,\xi_2,\xi_3) = \begin{bmatrix} -F(\xi_1,\xi_2,\xi_3) & * & * \\ * & -F(\xi_1,\xi_2,\xi_3) & * \\ * & * & -F(\xi_1,\xi_2,\xi_3)\end{bmatrix}$$

whose off-diagonal elements are nonsymmetric partial derivatives of F, for example,

$$\begin{aligned}\frac{\partial^2}{\partial\theta_1\partial\theta_2}F(\xi_1,\xi_2,\xi_3) = \sin(\xi_1)\sin(\xi_2)\cos(\xi_3) &- q(\boldsymbol{u}_1^\star,\boldsymbol{y},\boldsymbol{z})\sin(\xi_1)\cos(\xi_2)\sin(\xi_3)\\ &+ q(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w}_1^\star)\cos(\xi_1)\cos(\xi_2)\cos(\xi_3)\\ &- q(\boldsymbol{x},\boldsymbol{v}_1^\star,\boldsymbol{z})\cos(\xi_1)\sin(\xi_2)\sin(\xi_3)\\ &+ q(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\cos(\xi_1)\cos(\xi_2)\sin(\xi_3),\end{aligned}$$

which implies by (F.2) that for any $|\xi_i| \le \theta_0, i = 1,2,3$,

$$\begin{aligned}\left|\frac{\partial^2}{\partial\xi_1\partial\xi_2}F(\xi_1,\xi_2,\xi_3)\right| &\le |\sin(\theta_0)|^2 + o(1)(1 + 2|\sin(\theta_0)|^2) + (1+o(1))|\sin(\theta_0)|\\ &\le |\sin(\theta_0)| + |\sin(\theta_0)|^2 + o(1) \le 3|\sin(\theta_0)|.\end{aligned}$$

The same bound holds for other mixed partial derivatives $\left|\frac{\partial^2}{\partial\xi_i\partial\xi_j}F(\xi_1,\xi_2,\xi_3)\right|$ with $i,j=1,2,3$ and $i\neq j$.

To make $\nabla^2 F(\xi_1,\xi_2,\xi_3)$ negative definite, by Gershgorin's circle theorem and the bound (F.3), we only need

$$-F(\xi_1,\xi_2,\xi_3)+6|\sin(\theta_0)|\leq -1+9\theta_0^2+6\theta_0<0$$

which holds for any $\theta_0\in(\frac{-\sqrt{2}-1}{3},\frac{\sqrt{2}-1}{3})$, including $(\frac{-\sqrt{2}+1}{3},\frac{\sqrt{2}-1}{3})$. This completes the first part of the proof.

F.2 *Control the second vertex set*

Similarly as before, we first show

$$F(\pi+\xi_1,\pi+\xi_2,\pi+\xi_3)<0,\ \forall|\xi_i|<\frac{\sqrt{2}-1}{3}.$$

It follows from the intermediate result (F.3):

$$F(\xi_1,\xi_2,\xi_3)\geq 1-9\theta_0^2>0,\ \forall|\xi_i|\leq\theta_0$$

by recognizing that $F(\pi+\xi_1,\pi+\xi_2,\pi+\xi_3)=-F(\xi_1,\xi_2,\xi_3),\forall\xi_1,\xi_2,\xi_3$ and choosing $\theta_0=(\sqrt{2}-1)/3$.

Finally, we claim the same conclusion applies to the remaining three cases since

$$F(\pi+\xi_1,\pi+\xi_2,\pi+\xi_3)=F(\pi+\xi_1,\xi_2,\xi_3)=F(\xi_1,\pi+\xi_2,\xi_3)=F(\xi_1,\xi_2,\pi+\xi_3)$$

for all $\xi_1,\xi_2,\xi_3\in\mathbb{R}$.

G. Proof of Lemma 5.7

Proof. First, solve for θ such that

$$|\cos(\theta)^3|+|\sin(\theta)|^3<1-4\tau(\log n)n^{-r_c}. \tag{G.1}$$

To this end, we define $f(\theta):=|\cos(\theta)^3|+|\sin(\theta)|^3$ for $\theta\in[0,\pi]$. It can be verified directly that f is symmetric around $\frac{\pi}{2}$ on $[0,\pi]$, symmetric around $\frac{\pi}{4}$ on $[0,\frac{\pi}{2}]$ and strictly decreasing on $[0,\frac{\pi}{4}]$. Since $1-4\tau(\log n)n^{-r_c}\in(0,1)$, there exists a unique $\varpi\in(0,\frac{\pi}{4})$ such that $f(\varpi)=1-4\tau(\log n)n^{-r_c}\in(0,1)$. Thus, the inequality (G.1) holds on $(\varpi,\frac{\pi}{2}-\varpi)\cup(\frac{\pi}{2}+\varpi,\pi-\varpi)$. □

To have an approximation of ϖ, we need the following lemma.

LEMMA G.1. Let f and g be any two real functions with g being strictly decreasing in some interval (α,β) and satisfying $g(x)\geq f(x),\forall x\in(\alpha,\beta)$. Suppose both equations $f(x)=b$ and $g(x)=b$ admit one root in $[\alpha,\beta]$, denoted by x_f and x_g, respectively. Then $x_g\geq x_f$.

Proof of Lemma G.1. Since $g(x)>g(x_f)\geq f(x_f)=b$ for any $x\in[\alpha,x_f)$, $g(x_g)=b$ could only happen within $[x_f,\beta]$. □

We recognize that

$$f(\theta)\leq 1-\frac{3}{20}\theta^2,\ \text{for}\theta\in[0,\pi/4] \tag{G.2}$$

and $g(\theta) := 1 - \frac{3}{20}\theta^2$ is strictly deceasing $[0, \pi/4]$. Clearly,

$$\delta_b := \sqrt{\frac{80\tau(\log n)}{3}} n^{-0.5r_c}$$

is the root of $g(\theta) = 1 - 4\tau(\log n)n^{-r_c}$ over the interval $[0, \frac{\pi}{4}]$. By Lemma G.1, $\delta_b \geq \varpi$. Therefore, (G.1) holds on $(\delta_b, \frac{\pi}{2} - \delta_b) \cup (\frac{\pi}{2} + \delta_b, \pi - \delta_b)$. By (5.28), we obtain $F(\theta_1, \theta_2, \theta_3) < 1$ for $(\theta_1, \theta_2, \theta_3) \in \mathbb{N}_b(\delta_b)$.

G.1 *Proof of (G2)*

Showing (G.2) is equivalent to showing

$$\sin^3(x) + \cos^3(x) \leq 1 - \frac{3}{20}x^2, \ \forall x \in [0, \pi/4] \tag{G.3}$$

since $\sin(x), \cos(x) > 0$ for $x \in [0, \pi/4]$. Before moving on, we need the following lemma to prove (G.3).

Lemma G.2. The following inequality

$$\frac{(3^{2n-1} - 3)}{4 \cdot (2n-1)!}x^{2n-1} + \frac{(3^{2n} + 3)}{4 \cdot (2n)!}x^{2n} - \frac{(3^{2n+1} - 3)}{4 \cdot (2n+1)!}x^{2n+1} - \frac{(3^{2n+2} + 3)}{4 \cdot (2n+2)!}x^{2n+2} \geq 0 \tag{G.4}$$

holds for all $x \in [0, \pi/4]$ and $n \geq 2$,

Proof. Let p equal the expression on the left side of Equation (G.4). A simplification on p yields

$$p(x) = q_1(x)\frac{x^{2n-1}}{4(2n-1)!} + q_2(x)\frac{x^{2n+2}}{4(2n)!},$$

where $q_1(x) = (3^{2n-1} - 3) - \frac{3^{2n+1}-3}{2n(2n+1)}x^2$ and $q_2(x) = (3^{2n} + 3) - \frac{3^{2n+2}+3}{(2n+1)(2n+2)}x^2$.

As functions of x, q_1 and q_2 have roots at

$$\pm\sqrt{\frac{2n(2n+1)(3^{2n-1} - 3)}{3^{2n+1} - 3}} \quad \text{and} \quad \pm\sqrt{\frac{(2n+1)(2n+2)(3^{2n} + 3)}{3^{2n+2} + 3}},$$

respectively, provided $n \geq 1$.

Since

$$10(3^{2n-1} - 3) \geq 3^{2n+1} - 3, \ \text{for all} n \geq 2,$$

$$9(3^{2n} + 3) > (3^{2n+2} + 3), \ \text{for all} n \geq 2,$$

it follows that the positive root of q_1 satisfies

$$\sqrt{\frac{2n(2n+1)(3^{2n-1} - 3)}{3^{2n+1} - 3}} \geq \sqrt{\frac{2n(2n+1)}{10}} > \sqrt{2} > \frac{\pi}{4}, \ \text{for} n \geq 2,$$

and the positive root of q_2 satisfies

$$\sqrt{\frac{(2n+1)(2n+2)(3^{2n} + 3)}{3^{2n+2} + 3}} > \sqrt{\frac{(2n+1)(2n+2)}{9}} > \sqrt{\frac{10}{3}} > \frac{\pi}{4}, \ \text{for} n \geq 2.$$

Therefore, both q_1 and q_2 are positive on $[0, \pi/4]$ for all $n \geq 2$, and Equation (G.4) holds. □

LEMMA G.3. The following statement

$$\sin^3(x) + \cos^3(x) \leq 1 - \frac{3}{20}x^2$$

holds for all $x \in [0, \frac{\pi}{4}]$.

Proof. Recall that

$$\sin^3(x) = \frac{1}{4}\left(3\sin(x) - \sin(3x)\right),$$

$$\cos^3(x) = \frac{1}{4}\left(3\cos(x) + \cos(3x)\right).$$

Therefore,

$$\sin^3(x) = x^3 + \sum_{n=5}^{\infty}(-1)^n \frac{3^{2n-1} - 3}{4(2n-1)!}x^{2n-1},$$

$$\cos^3(x) = 1 - \frac{3}{2}x^2 + \frac{7}{8}x^4 + \sum_{n=3}^{\infty}(-1)^n \frac{3^{2n} + 3}{4(2n)!}x^{2n}.$$

Thus,

$$\sin^3(x) + \cos^3(x) \leq 1 - \frac{3}{2}x^2 + x^3 + \frac{7}{8}x^4,$$

for all $x \in [0, \pi/4]$ since by Lemma G.2

$$\begin{aligned}
&\sum_{n=3}^{\infty}(-1)^n \frac{3^{2n-1} - 3}{4(2n-1)!}x^{2n-1} + \sum_{n=3}^{\infty}(-1)^n \frac{3^{2n} + 3}{4(2n)!}x^{2n} \\
&= -\sum_{n=3,\ n\ \text{odd}}^{\infty}\left(\frac{3^{2n-1} - 3}{4(2n-1)!}x^{2n-1} + \frac{3^{2n} + 3}{4(2n)!}x^{2n} - \frac{3^{2n+1} - 3}{4(2n+1)!}x^{2n+1} - \frac{3^{2n+2}}{4(2n+2)!}x^{2n+2}\right) \\
&\leq 0.
\end{aligned}$$

Finally, note that

$$1 - \frac{3}{2}x^2 + x^3 + \frac{7}{8}x^4 = 1 - \frac{3}{20}x^2 + x^2 h(x)$$

with

$$h(x) = -\frac{27}{20} + x + \frac{7}{8}x^2 \geq 0 \text{ when} x \in [0, \pi/4].$$

□