Mathematical Programming
https://doi.org/10.1007/s10107-024-02165-1

FULL LENGTH PAPER

Series A

Quantifying low rank approximations of third order symmetric tensors

Shenglong Hu[1] · **Defeng Sun**[2] · **Kim-Chuan Toh**[3]

Received: 21 July 2023 / Accepted: 23 October 2024

Abstract
In this paper, we present a method to certify the approximation quality of a low rank tensor to a given third order symmetric tensor. Under mild assumptions, best low rank approximation is attained if a control parameter is zero or quantified quasi-optimal low rank approximation is obtained if the control parameter is positive. This is based on a primal-dual method for computing a low rank approximation for a given tensor. The certification is derived from the global optimality of the primal and dual problems, and is characterized by easily checkable relations between the primal and the dual solutions together with another rank condition. The theory is verified theoretically for orthogonally decomposable tensors as well as numerically through examples in the general case.

Keywords Tensor · Low rank approximation · Quasi-optimal · Polynomial · SDP relaxation · Moment · Rank constraint · Duality · Optimality · Orthogonally decomposable tensor · Projection

Mathematics Subject Classification 90C26 · 15A69 · 65F18 · 90C46

✉ Defeng Sun
defeng.sun@polyu.edu.hk

Shenglong Hu
hushenglong@nudt.edu.cn

Kim-Chuan Toh
mattohkc@nus.edu.sg

[1] Department of Mathematics, College of Science, National University of Defense Technology, Changsha 410072, Hunan, China

[2] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

[3] Department of Mathematics, National University of Singapore, 10 Lower Kent Ridge Road, Singapore, Singapore

Published online: 28 November 2024

1 Introduction

Let $\mathbb{R}$ be the field of real numbers and $\mathrm{S}^m(\mathbb{R}^n)$ be the space of symmetric tensors with real entries of order m and dimension n for positive integers m and n. When $m = 2$, $\mathrm{S}^2(\mathbb{R}^n)$ is the space of symmetric $n \times n$ real matrices. A tensor $\mathcal{A} \in \mathrm{S}^m(\mathbb{R}^n)$ can be represented by its entries $a_{i_1 \ldots i_m}$ with $i_j \in \{1, \ldots, n\}$ for all $j \in \{1, \ldots, m\}$. In this paper, we will focus on third order symmetric tensors, i.e., $m = 3$, which is the most important case in several applications [36]. In the set $\mathrm{S}^3(\mathbb{R}^n)$, there is the fundamental set of *decomposable tensors*, i.e., tensors of the form $\mathbf{x}^{\otimes 3}$ for a vector $\mathbf{x} \in \mathbb{R}^n$, where $\mathbf{x}^{\otimes 3}$ is a short hand for

$$\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x} \in \mathrm{S}^3(\mathbb{R}^n)$$

whose (i_1, i_2, i_3)-entry is $x_{i_1} x_{i_2} x_{i_3}$ for all $i_1, i_2, i_3 \in \{1, \ldots, n\}$. A third order symmetric tensor $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$ has (real symmetric) *rank* r if it can be represented as (cf. [7])

$$\mathcal{A} = \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3} \text{ for some } \lambda_i > 0 \text{ and } \|\mathbf{x}_i\| = 1 \text{ with } i = 1, \ldots, r \tag{1}$$

and a decomposition of the form (1) with a summand strictly smaller than r does not exist. The rank of a tensor $\mathcal{A}$ is denoted as $\mathrm{rank}(\mathcal{A})$. A rank one decomposition of a given $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$ is a decomposition as in (1), and it becomes a *rank decomposition* if $r = \mathrm{rank}(\mathcal{A})$. It is a fact that each tensor $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$ has a rank decomposition as (1) [7].

In this paper, we consider the following problem, which is termed as the *best rank-r approximation problem*:

$$\inf \left\{ \frac{1}{2} \|\mathcal{A} - \mathcal{B}\|^2 : \mathrm{rank}(\mathcal{B}) \leq r \right\} \tag{2}$$

for a given tensor $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$ and a positive integer r. Usually, r is assumed to be small, and thus it is also called the *best low rank approximation problem*. The prefix "inf" in the optimization problem (2) instead of "min" is due to the fact that the constraint set is not necessarily closed and hence the optimizer may not exist, which is emphasized by De Silva and Lim [12]. Therefore, the following assumption is necessary.

Assumption 1 *There exists a best rank-r approximation for the given tensor $\mathcal{A}$, i.e., problem (2) has an optimal solution.*

Assumption 1 is assumed throughout this paper. Under Assumption 1, the "inf" in problem (2) can be strengthened as "min".

Tensors (a.k.a. hypermatrices) have become a standard tool in a wide variety of applications and mathematical sciences [28, 31, 35, 36]. Particularly, it is one of the foundations in multilinear algebra. A fundamental topic in this area is tensor decomposition, which finds a decomposition of the given tensor via rank one tensors with the smallest possible length (cf. (1)). A companion but with the same importance is the best low rank approximation for a given tensor, which is a cornerstone in several

disciplines, especially in applications where data is collected with noise and thus approximation is prevalent [16]. The best low rank approximation problem for a given tensor has a rich literature, see [6, 7, 11, 16, 24, 28, 31] and references therein.

In this paper, we will focus on symmetric tensors. A second order symmetric tensor is the symmetric matrix in the usual sense. However, many crucial properties change dramatically if we switch from second order symmetric tensors to higher orders, e.g., the possible nonexistence of optimizers for (2) [12]. Behind this unfavorable phenomena is that the rank of a tensor has more complicated but rich features, such as the NP-hardness to compute it as established by Håstad [17], the nonsymmetric rank differs from the rank as noted by Shitov [48]; we refer the reader to [7, 31] and references therein for more information. However, despite of these mysteries, many favourable properties, such as the celebrated Alexander–Hirschowitz theorem [1] which identifies all the generic ranks, and Kruskal's theorem on the uniqueness of the rank decomposition [30], make methods based on symmetric tensor approximation/decomposition unbelievably appealing in a wide range of applications, such as blind source separation [8].

The developments of the symmetric tensor approximation problem are based on the decomposition problem, cf. (2). The symmetric tensor decomposition problem, also known as the *Waring decomposition* in the algebraic geometry literature [52], has attracted considerable attentions even very recently, such as Balllico and Bernardi [2], Bernardi, Gimigliano and Idà [3], Brachat, Comon, Mourrain and Tsigaridas [5], Comon and Mourrian [9], Nie [41], and Oeding and Ottaviani [45], etc.

The other side of the coin is the symmetric tensor best low rank approximation problem. It is particularly preferable in applications, see [16, 28]. For example, a low rank approximation can reduce the complexity of manipulating an m-th order symmetric tensor of dimension n from $O(n^m)$ to $O(n)$, which is important and indispensable for applications where n is large and computational cost such as $O(n^m)$ is prohibitive. The matrix case (i.e., $m = 2$) is resolved by well-developed techniques in both theory and algorithms [15]. However, in contrast to the matrix counterpart, the higher order equivalents are still under investigation and several key problems must be resolved first. Besides the possible nonexistence of optimal solutions [12], there are challenging issues such as the NP-hardness even just to obtain the best rank one approximation [20], etc. Therefore, it is an important research topic in the community. For the case $r = 1$, i.e., the best rank one approximation, there are extensive research done, such as Kofidis and Regalia [26], Kolda and Mayo [29], Nie and Wang [44], Qi [46], Zhang, Ling and Qi [55], etc. For general r, alternating minimization techniques are adopted to solve (2), see the survey [6, 16] and references therein. While very good numerical performances were observed, *there is a lack of theoretical justification on what is found, such as global optimality certification, approximation quality guarantees, etc*, which are long standing fundamental questions for best low rank approximation. A very recent progress is made by Nie [42], in which a method is proposed using generating polynomials and it can find a quasi-optimal solution if the given tensor is sufficiently close to the targeted low rank one.

In this paper, we will study the fundamental question in symmetric tensor best low rank approximation:

Certify a candidate as being globally optimal or as a quantified quasi-optimal solution.

For the rank one case (i.e., $r = 1$), Nie and Wang [44] proposed a semidefinite relaxation method utilizing Reznick's Positivstellensatz for forms. It is shown that if the computed moment matrix has rank one, then certification of global optimality can be derived. A theoretical result which guarantees the rank one property is given under mild assumptions. Numerical results show that global optimality can always be certified. Moreover, the nonsymmetric cases are studied and similar conclusions are achieved as well in [44]. The rank one approximation under nonnegativity is studied by the authors in [23]. This paper considers the case $r \geq 1$.

Contributions. Our main theorem can be stated as follows, in which we defer the exact meaning of the notation in the content for clarity.

Theorem 1 (Rank-r Approximation). *Let $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$, $k \geq 2$, $\sigma \geq 0$ and $r \geq 1$. Given a tensor $\mathcal{B} = \sum_{i=1}^r \lambda_i \mathbf{x}_i^{\otimes 3}$. Let $\mathbf{y}$ be the moment sequence defined by the measure $\sum_{i=1}^r \lambda_i \delta_{\mathbf{x}_i}$. If there exists a triplet (U, W, Z) such that*

1. *the feasibility holds*

$$\mathcal{M}_k^*(Z) + \mathcal{P}^*(U) - \mathcal{L}_k^*(W) = \sigma \mathcal{M}_k^*(E_0), \text{ and } Z \succeq 0;$$

2. *the optimality holds*

$$\langle Z, \mathcal{M}_k(\mathbf{y}) \rangle = 0, \text{ and } \mathcal{P}(\mathbf{y}) \in \Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - U).$$

Then, it holds

1. *if $r = 1$ and $\sigma < \rho(\mathcal{A})$, then $(\lambda_1 + \sigma)\mathbf{x}_1^{\otimes 3}$ is a best rank one approximation of $\mathcal{A}$;*
2. *if $\sigma < \frac{\tau(\mathcal{A})\rho(\mathcal{A})}{2r}$ and Assumption 2 holds when $\sigma > 0$, then $\mathcal{B}$ is a $2\sqrt{\frac{r}{\tau(\mathcal{A})}}\left(\left(1 - \sqrt{\frac{\tau(\mathcal{A})}{r}}\right)\|\mathcal{A}\| + 2\sigma\right)\sigma$-quasi-optimal rank-$r$ approximation of $\mathcal{A}$.*

Note that when $\sigma = 0$, we get best rank-r approximants. We also note that $\tau(\mathcal{A})$ is an intrinsic number determined by a rank decomposition of a best rank-r approximant of $\mathcal{A}$, which depends on r, and $\tau(\mathcal{A}) = 1$ when $r = 1$.

Theorem 1 is proved by first establishing a primal-dual method (11)–(12) for solving a relaxation of the best low rank approximation problem (2), and then an approximation quality analysis of the primal problem (11) to (2).

More precisely, we apply the semidefinite relaxation for moment problems and a rank characterization to reformulate the third order symmetric tensor best low rank approximation problem as a rank constrained nonlinear matrix optimization problem. Advantages of nonsmooth analysis on low rank projection of matrices are explored to propose a relaxation for the reformulation, which is the *primal* problem. Then, the *dual* problem is explicitly given. A theorem on the certification of the global optimality of a candidate solution for the primal and dual problem is given. One ingredient is that the global optimality certification conditions are explicitly given and easily checkable

once a feasible solution for the primal problem, together with a Lagrange multiplier which is always available from a primal-dual algorithm, is given. This is the result of our careful design of the primal problem (11), while keeping in mind of a target dual certificate. Approximation quality of problem (11) to the original best low rank approximation problem (2) is then established. It depends on a control parameter σ. If $\sigma = 0$, it relates to the best low rank approximation, and if $\sigma > 0$, a quantified quasi-optimal low rank approximation is given. Positive σ is preferable to ensure the strict feasibility of the dual problem, see Lemma 2. While, both $\sigma = 0$ and $\sigma > 0$ are allowable for approximation quality certifications, see Theorem 1. The validation of this approach is verified for orthogonally decomposable tensors from a theoretical perspective, and several examples numerically.

Our method employs recent advances from the semidefinite relaxation for the moment problems, duality theory in low rank matrix optimization and nonsmooth analysis for low rank matrix projection. In particular, it is directly motivated by the works of Gao and Sun on low rank matrix optimization problems and the duality theory [13, 14], of Nie on semidefinite relaxation of moment problems [38, 40, 43], and of Tang and Shah on the tensor decomposition method [50], in which a theoretical justification for the decomposition of orthogonally decomposable tensors is established.

Our method extends Nie and Wang's important result on global optimality certification of best rank one approximation in [44], and Nie's interesting investigation on quasi-optimality of best low rank approximation in [42]. For global optimality, the rank one case is studied in [44], and Theorem 1 extends it to best rank-r approximation with $r \geq 1$. Moreover, if global optimality cannot be certified, then Theorem 1 gives a quantification on quasi-optimality. In [42], the best rank-r approximation with $r > 1$ is studied and qualitative quasi-optimality is shown under the assumption that the given data is sufficiently close to the best rank-r approximation. In Theorem 1, the restriction on the distance between the approximant and the given data is removed, and either global optimality can be certified or a quantitative estimation on quasi-optimality can be derived under mild assumptions. It is also worth mentioning that the approach in this paper gives an alternative to those in [42, 44]. In particular, we explicitly involve in our formulation the rank constraint and put emphasis on the dual problem.

Contents. The approximation problem and related preliminaries are given in Sect. 2. The nonlinear matrix optimization reformation is given in Sect. 3. In order to keep the main theme of this paper on the tensor best low rank approximation, some notations and supporting techniques for the nonsmooth analysis of low rank matrix projections and others are put in Appendix A and Appendix B. Section 4 studies the approximation quality of the problem proposed in Sect. 3, and presents the theoretical certification. The numerical illustration is given in Sect. 5. Some final remarks are given in Sect. 6.

2 Preliminaries

In this paper, we will focus on third order symmetric tensors. Given a positive integer n, a third order symmetric tensor $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$ is a collection of n^3 scalars $a_{i_1 i_2 i_3}$, termed the entries of $\mathcal{A}$, for all $i_j \in \{1, \ldots, n\}$ and $j \in \{1, 2, 3\}$. As the case of symmetric

matrices, the number of independent entries are smaller than, but in the same order of, n^3 due to the symmetry. There are all together $\binom{n+2}{n-1}$ independent entries encoded by a third order n dimensional symmetric tensor. A symmetric rank one tensor in $\mathrm{S}^3(\mathbb{R}^n)$ is an element $\mathbf{x}^{\otimes 3}$ for some vector $\mathbf{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$. We refer the readers to [35] and references herein for basic notions on tensors.

The notation $\|\cdot\|$ represents the *Euclidean norm* for a vector, the *Frobenius norm* for a matrix [21], and the *Hilbert-Schmidt norm* for a tensor [35], defined as

$$\|\mathcal{A}\| := \Big(\sum_{i,j,k=1}^{n} a_{ijk}^2 \Big)^{\frac{1}{2}} \text{ for all } \mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$$

with the corresponding inner product defined as

$$\langle \mathcal{A}, \mathcal{B} \rangle := \sum_{i,j,k=1}^{n} a_{ijk} b_{ijk}.$$

It can be shown that

$$\rho(\mathcal{A}) := \max\{\langle \mathcal{A}, \mathbf{x}^{\otimes 3} \rangle : \mathbf{x}^{\mathsf{T}}\mathbf{x} = 1\}$$

defines a norm on $\mathrm{S}^3(\mathbb{R}^n)$ [46]. As for the matrix case, $\rho(\mathcal{A})$ is called the *spectral radius* of $\mathcal{A}$. It can be shown that

$$\rho(\mathcal{A}) \le \|\mathcal{A}\|.$$

Define

$$\zeta(n, s) := \binom{n+s-1}{n-1}.$$

Further basic notations are put in Appendix A, such as several bases of monomials, moment sequences and matrices, localizations, and flatness, etc.

Given a third order symmetric tensor $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$, we will identify it with its flattening matrix $M(\mathcal{A}) \in \mathbb{R}^{n \times n^2}$ via

$$M(\mathcal{A})_{i,(j-1)*n+k} := a_{ijk} \text{ for all } i, j, k \in \{1, \ldots, n\}. \tag{3}$$

Here $M(\mathcal{A})$ is different from the Catalecticant matrix of $\mathcal{A}$ [31]. In view of the identification (3), we will interchangeably refer to a given tensor by $\mathcal{A}$ and $M(\mathcal{A})$. It is straightforward to check that $\|\mathcal{A}\|^2 = \|M(\mathcal{A})\|^2$.

2.1 Polynomial identification

The problem (2) can be parameterized as

$$\min \left\{ \frac{1}{2} \Big\| \mathcal{A} - \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3} \Big\|^2 : \lambda_i \ge 0,\ \|\mathbf{x}_i\| = 1 \text{ for all } i = 1, \ldots, r \right\}. \tag{4}$$

Note that the constraint $\|\mathbf{x}_i\| = 1$ is added to remove some ambiguity, since the unitary scaling between $\mathbf{x}_i$ and λ_i gives the same approximate tensor. Problem (4) can be studied by applying techniques directly from polynomial optimization [33]. However, this may not be the best way from the computational perspective. Take the case $n = 10$ and $r = 2$ for example, the standard Lasserre relaxation will give an SDP with matrix size around $\zeta(25, 4) = 12650$ and number of equations around $\zeta(25, 4)^2/2$ [33]. The current SDP solvers have limited ability to solve such instances [49, 51, 53, 56], let alone certifying global optimality of (2). In this paper, we will present a method which gives a nonlinear matrix optimization problem with matrix size around $\zeta(n + 1, 2) = \zeta(11, 2) = 55$.

For easy reference, in the following, we will denote the $(n-1)$-dimensional sphere in $\mathbb{R}^n$ as $\mathbb{S}^{n-1}$, i.e.,

$$\mathbb{S}^{n-1} := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}^{\mathsf{T}}\mathbf{x} = 1\}.$$

Let $\mathcal{M}(\mathbb{S}^{n-1})$ be the set of Borel measures on $\mathbb{S}^{n-1}$. The *support* of a Borel measure $\mu \in \mathcal{M}(\mathbb{S}^{n-1})$ is denoted as $\mathrm{supp}(\mu)$, which is defined as the smallest closed set $S \subseteq \mathbb{S}^{n-1}$ such that $\mu(\mathbb{S}^{n-1} \backslash S) = 0$. Given a measure $\mu \in \mathcal{M}(\mathbb{S}^{n-1})$, we denote by $\|\mu\|_0$ the cardinality of its support $\mathrm{supp}(\mu)$. Whenever μ is finitely supported, $\|\mu\|_0$ counts the number of points in the support; otherwise $\|\mu\|_0 := +\infty$. Let $\mathbf{x} \in \mathbb{S}^{n-1}$, then $\delta_{\mathbf{x}} \in \mathcal{M}(\mathbb{S}^{n-1})$ is the *Dirac measure* at $\mathbf{x}$, with support $\{\mathbf{x}\}$ and having mass 1 at $\mathbf{x}$ and mass 0 elsewhere. If a measure $\mu \in \mathcal{M}(\mathbb{S}^{n-1})$ has finite support, whose cardinality is $r \geq 0$, then it can be represented as

$$\mu = \sum_{i=1}^{r} \lambda_i \delta_{\mathbf{x}_i} \tag{5}$$

for some $\mathbf{x}_i \in \mathbb{S}^{n-1}$ and $\lambda_i > 0$ with $i = 1, \ldots, r$. In this case, it is called an *r-atomic* measure.

A given third order symmetric tensor $\mathcal{A}$ can be uniquely decoded via

$$a_{i_1 i_2 i_3} \leftrightarrow a_\alpha \text{ with } \alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n_{=3} \text{ via } \mathbf{x}^\alpha := \prod_{i=1}^{n} x_i^{\alpha_i} = x_{i_1} x_{i_2} x_{i_3}. \tag{6}$$

It is easy to see that this correspondence is one to one. With this correspondence, a third order symmetric tensor $\mathcal{A}$ can be interpreted as a *truncated moment sequence* (abbreviated as *tms*), which is defined as a vector

$$(a_\alpha)_{:\, \alpha \in \mathbb{N}^n_{=3}} \in \mathbb{R}^{\binom{n+2}{3}}.$$

Note that only the independent elements of $\mathcal{A}$ are coded in this vector. More precisely, a third order symmetric tensor $\mathcal{A}$ with a rank-r decomposition as (1) can be naturally restated as a truncated moment sequence of a finite (r-atomic) Borel measure (5) on

$\mathbb{S}^{n-1}$. Actually, with $\mu = \sum_{i=1}^{r} \lambda_i \delta_{\mathbf{x}_i}$, it follows from (1) and (6) that

$$a_{i_1 i_2 i_3} = a_\alpha = \sum_{i=1}^{r} \lambda_i (\mathbf{x}_i)_{i_1} (\mathbf{x}_i)_{i_2} (\mathbf{x}_i)_{i_3} = \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^\alpha = \int_{\mathbb{S}^{n-1}} \mathbf{x}^\alpha \, \mathrm{d}\, \mu(\mathbf{x})$$

for all $i_1, i_2, i_3 \in \{1, \ldots, n\}$ and the corresponding α such that $\prod_{i=1}^{n} x_i^{\alpha_i} = x_{i_1} x_{i_2} x_{i_3}$. More concisely, (1) can be written as

$$\mathcal{A} = \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3} \simeq \int_{\mathbb{S}^{n-1}} \mathbf{x}^{\otimes 3} \, \mathrm{d}\, \mu(\mathbf{x}), \tag{7}$$

where the symbol "$\simeq$" is understood as the obvious correspondence between the tensor $\mathcal{A}$ and the vector on the right hand side. Therefore, from this perspective, each third order symmetric tensor can be regarded as a truncated moment sequence of total degree three and vice versa.

With the moment representation (7), the third order symmetric tensor best rank-r approximation problem (2) can be reformulated as a quadratic moment optimization problem with support constraint as follows.

Proposition 2 (Moment Reformulation). *For any given third order symmetric tensor $\mathcal{A}$ and any nonnegative integer r, the best rank-r tensor approximation problem (2) is equivalent to the following moment optimization problem*

$$\begin{array}{l} \min \ \frac{1}{2}\|\mathcal{A} - \mathcal{B}\|^2 \\ s.t. \ \mathcal{B} \simeq \int_{\mathbb{S}^{n-1}} \mathbf{x}^{\otimes 3} \, \mathrm{d}\, \mu(\mathbf{x}), \\ \quad \|\mu\|_0 \le r, \\ \quad \mu \in \mathcal{M}(\mathbb{S}^{n-1}) \end{array} \tag{8}$$

in the sense that $(\mathcal{B} = \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3}, \mu = \sum_{i=1}^{r} \lambda_i \delta_{\mathbf{x}_i})$ is an optimal solution of (8) whenever $\sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3}$ forms an optimal solution of (2) and vice verse.

Proof It follows from the preceding discussions. □

Note that if r is chosen as $\mathrm{rank}(\mathcal{A})$, then (8) becomes the tensor rank decomposition problem. While $\mathrm{rank}(\mathcal{A})$ is difficult to find [17], an upper bound r is usually given, which then makes (8) a tensor decomposition problem.

There are two difficult issues in solving (8). The first one is the constraint $\|\mu\|_0 \le r$, and the other one is the characterization for the set $\mathcal{M}(\mathbb{S}^{n-1})$. For the latter, there are standard positive semidefinite relaxation schemes for approximating the set $\mathcal{M}(\mathbb{S}^{n-1})$ exteriorly [32, 33, 38, 40, 43]. For the former, we will develop a dual certification technique from optimization. Combing these two techniques, we will present a possibility, for the first time, to certify the global optimality of the third order symmetric tensor best rank-r approximation problem (2). In the following, for simplicity, we will say $B := M(\mathcal{B})$ is a feasible (an optimal) solution of (2) in view of the equivalence in (3).

2.2 Semidefinite relaxation for the measure

A positive semidefinite symmetric matrix $A \in \mathrm{S}^2(\mathbb{R}^n)$ is written as $A \succeq 0$ or $A \in \mathrm{S}^n_+$. Let k be a positive integer, and $\mathbf{y} \in \mathbb{R}^{\zeta(n+1,2k)}$ be the *moment sequence* of a given measure $\mu \in \mathcal{M}(\mathbb{S}^{n-1})$ up to degree $2k$, i.e.,

$$\mathbf{y} = \int_{\mathbb{S}^{n-1}} \mathbf{x}^{\circ 2k} \, \mathrm{d}\, \mu(\mathbf{x}).$$

It is well-known that if the measure is finitely supported with $\|\mu\|_0 = r$, then the rank of the truncated moment matrix $\mathrm{rank}(\mathcal{M}_k(\mathbf{y})) \leq r$ for every positive integer k. More details on moment sequences and matrices are collected in Appendix A. We refer the reader to [33, 43] for more basic notions and concepts on semidefinite relaxation hierarchy of polynomial optimization.

With the observation on the rank constraint, we consider the following problem:

$$\begin{aligned} \min\ & \tfrac{1}{2}\|\mathcal{A} - \mathcal{B}\|^2 \\ \text{s.t. } & \mathcal{B} \simeq \mathbf{y}|_{\mathbb{N}^n_{=3}}, \\ & \mathrm{rank}(\mathcal{M}_k(\mathbf{y})) \leq r, \\ & \mathcal{M}_k(\mathbf{y}) \succeq 0, \\ & \mathcal{L}_k(\mathbf{y}) = 0, \\ & \mathbf{y} \in \mathbb{R}^{\zeta(n+1,2k)}, \end{aligned} \tag{9}$$

where $k \geq 2$, $\mathcal{M}_k(\mathbf{y}) \in \mathrm{S}^2(\mathbb{R}^{\zeta(n+1,k)})$ represents the k-th moment matrix of the moment sequence $\mathbf{y}$, $\mathcal{L}_k(\mathbf{y}) \in \mathrm{S}^2(\mathbb{R}^{\zeta(n+1,k-1)})$ represents the $(k-1)$-th localizing matrix of $1 - \mathbf{x}^{\mathsf{T}}\mathbf{x}$ at $\mathbf{y}$.

The next result is a basis for the subsequent analysis.

Proposition 3 (Exact Relaxation). *Let $k \geq 2$ and $\mathbf{y}^* \in \mathbb{R}^{\zeta(n+1,2k)}$ be an optimal solution of* (9)*. If the k-th flatness condition holds (cf.* Appendix A*), i.e.,*

$$\mathrm{rank}(\mathcal{M}_k(\mathbf{y}^*)) = \mathrm{rank}(\mathcal{M}_{k-1}(\mathbf{y}^*)), \tag{10}$$

then $\mathbf{y}^*$ is the truncated moment sequence of a unique $\mathrm{rank}(\mathcal{M}_k(\mathbf{y}^*))$-atomic measure and $\mathcal{B}^*$ is a best rank-r approximation of the given tensor $\mathcal{A}$, i.e., a global minimizer of (2)*.*

Proof First of all, problem (9) is a relaxation of the problem (8). Thus, if there is an optimal solution $\mathcal{B}$ of (9) such that it is a feasible solution of (8), then it must be an optimal solution of the problem (8) and hence a global minimizer of (2) by Proposition 2.

Since the sequence $\mathbf{y}^*$ satisfies the flatness condition, by a well-known result of Curto and Fialkow [10], it follows that the sequence $\mathbf{y}^*$ admits a unique measure supported by $\mathbb{S}^{n-1}$, which is $\mathrm{rank}(\mathcal{M}_k(\mathbf{y}^*))$-atomic. This, together with the rank constraint $\mathrm{rank}(\mathcal{M}_k(\mathbf{y})) \leq r$, implies that $\mathcal{B}^* \simeq \mathbf{y}^*|_{\mathbb{N}^n_{=3}}$ is a tensor with rank at most r, which is thus a feasible solution of (8) with exactly the same objective function value as that of (9). □

There is a method to extract the support of a measure whenever the flatness condition is satisfied [18, 25, 33]. Thus, a rank decomposition of an optimal solution for (2) which is a tensor of rank at most r can be computed, if an optimal solution of (9) satisfying the flatness condition (10) was found. In our numerical computation, the method in [18] is adopted. One could also apply the robust method in [25] for extraction when the flatness condition is approximately satisfied.

3 Rank constrained matrix optimization

The problem (9) is a nonconvex optimization problem with rank constraint, which is NP-hard in general. Nonetheless, a much harder part is to certify the global optimality for a candidate of (9). Thus, Proposition 3 can merely be utilized in few peculiar scenarios. In order to employ a dual certificate for global optimality for a wider class of problems, we will propose a carefully designed variation for it.

3.1 Reformulation

We consider the following optimization problem

$$\begin{aligned} \min\ & \psi(B, X) := \tfrac{1}{2}\|M(\mathcal{A}) - B\|^2 + \sigma\langle E_0, X\rangle \\ \text{s.t. } & B - \mathcal{P}(\mathbf{y}) = 0, \\ & X - \mathcal{M}_k(\mathbf{y}) = 0, \\ & \mathcal{L}_k(\mathbf{y}) = 0, \\ & \operatorname{rank}(B) \le r, \\ & X \succeq 0, \end{aligned} \tag{11}$$

where $\sigma \ge 0$ is a *control parameter*, $\mathcal{P}(\mathbf{y})$ is the matrix generated by $\mathbf{y}$ as (A4), and $E_0 \in \mathrm{S}^2(\mathbb{R}^{\zeta(n+1,k)})$ is the matrix with all entries being zero except that $E_0(1, 1) = 1$.

Compared with (9), both the parameter σ and the rank constraint on B instead of $\mathcal{M}_k(\mathbf{y})$ are for dual certificate reasons which will be addressed later. The parameter σ is also for numerical considerations (cf. Lemma 2). In general, both $\sigma > 0$ and $\sigma = 0$ are allowable in problem (11). Actually, if $\sigma = 0$ is chosen, then (11) is (9) except that the constraint $\operatorname{rank}(B) \le r$ is employed instead of the constraint $\operatorname{rank}(\mathcal{M}_k(\mathbf{y})) \le r$. Since the matrix B in (11) is essentially a block sub-matrix of $\mathcal{M}_k(\mathbf{y})$ in the sense of Lemma 6, problem (11) is a relaxation of (9). Nevertheless, exact relaxation results will be shown in Theorem 9. For general $\sigma > 0$, quantified quasi-optimal approximation results will be given in Sect. 4.2.

Problem (11) is called the *k-th relaxation* of problem (2). Since $k \ge 2$, the second relaxation is called the *basic relaxation*. For the optimization problem (11), X and B are determined once $\mathbf{y}$ was given. Thus, for simplicity, unless otherwise stated, only the variable $\mathbf{y}$ is referred when we talk about feasible or optimal solutions. Note that the feasible set of (11) is closed and the objective function is a polynomial which is bounded from below on the feasible set. Moreover, we can prove the following result.

Proposition 4 (Solvability). *Each level set of the feasible set of the optimization problem* (11) *is bounded for* $\sigma > 0$*, and there is an optimizer of* (11) *for each* $\sigma \geq 0$*.*

Proof If $\sigma > 0$, we have that in each level set of the feasible set of problem (11), both B and $y_{\mathbf{0}}$ must be bounded. It follows from the constraint $\mathcal{L}_k(\mathbf{y}) = 0$ that

$$y_{\mathbf{0}} = \sum_{i=1}^{n} y_{2\mathbf{e}_i}.$$

It then follows from the constraint $X \succeq 0$ that all $y_{2\mathbf{e}_i}$'s are nonnegative and bounded. In turn, it follows that all y_α with $|\alpha| = 2$ are bounded by the positive semidefiniteness of X. The boundedness of y_α with $|\alpha| = 1$ follows from the boundedness of the matrix B and the constraint $\mathcal{L}_k(\mathbf{y}) = 0$. In the following, we show that each $y_{2\mathbf{e}_i+2\mathbf{e}_j}$ is bounded, which will imply the boundedness of all y_α with $|\alpha| = 4$ by the positive semidefinitenss of X. Since it follows from $\mathcal{L}_k(\mathbf{y}) = 0$ that

$$y_{2\mathbf{e}_i} = \sum_{j=1}^{n} y_{2\mathbf{e}_i+2\mathbf{e}_j},$$

the boundedness of the left hand side and the positive semidefiniteness of X imply the desired result. This proves the boundedness for the case $k = 2$.

The general case for $k > 2$ follows from a similar argument through induction. We omit the details. The solvability for the case $\sigma > 0$ then follows immediately.

The solvability for the case $\sigma = 0$ has a different argument. It is clear that the projection of the level set of the feasible set onto the B part is bounded. In the following, we show that it is also closed. Then the conclusion follows.

For each given B in the closure of this given projection, it corresponds to a third order symmetric tensor $\mathcal{B}$. Each third order symmetric tensor has a rank decomposition as (1), corresponding to a finite measure $\mu = \sum_{i=1}^{s} \lambda_i \delta_{\mathbf{x}_i}$ for some s which could be different from r. Let $\bar{\mathbf{y}}$ be the moment sequence generated by μ and $(\bar{X}, \bar{\mathbf{y}}, \bar{B})$ the defined point by the first two constraints in (11). It is clear that all the constraints are satisfied. Moreover, the resulting feasible point $(\bar{X}, \bar{\mathbf{y}}, \bar{B})$ has the same objective function value as the given B, since $B = \bar{B}$. Thus, the point $(\bar{X}, \bar{\mathbf{y}}, \bar{B})$ is in the level set of the feasible set from which the projection is performed. Consequently, B is in the projection and the projected set is closed as desired. The proof is completed. □

3.2 Duality and feasibility

In this section, we discuss the dual problem of (11). To that end, the projection of a given matrix onto the set of matrices of rank at most r is involved. Let $\mathrm{R}(r) \subseteq \mathbb{R}^{m\times n}$ be the set of matrices in $\mathbb{R}^{m\times n}$ with rank at most $r \leq \min\{m, n\}$. For notational simplicity, the dependence on m and n is omitted in $\mathrm{R}(r)$ and it will be clear from the context.

We will use $\Pi_{\mathrm{R}(r)}(A) \subset \mathrm{R}(r)$ to denote the optimal solution set for the problem

$$\begin{aligned} &\min_X \ \tfrac{1}{2}\|A - X\|^2 \\ &\text{s.t. } \operatorname{rank}(X) \le r. \end{aligned}$$

Here $\Pi_{\mathrm{R}(r)}(A)$ can be a set with infinitely many elements, but each element $X \in \Pi_{\mathrm{R}(r)}(A)$ has the same norm $\|X\|$. Thus, $\|\Pi_{\mathrm{R}(r)}(A)\|$ can be used to define this common constant. We refer the reader to Appendix B for more details and the necessary nonsmooth analysis for this projection.

Proposition 5 (Lagrangian Dual Problem). *The Lagrangian dual problem of* (11) *is*

$$\begin{aligned} &\min \ \tfrac{1}{2}\|\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - U)\|^2 \\ &\text{s.t. } \mathcal{M}_k^*(Z) + \mathcal{P}^*(U) - \mathcal{L}_k^*(W) = \sigma \mathcal{M}_k^*(E_0), \\ &\quad Z \succeq 0, \end{aligned} \tag{12}$$

where $\mathcal{L}_k^*$, $\mathcal{M}_k^*$ and $\mathcal{P}^*$ are the adjoint operators of $\mathcal{L}_k$, $\mathcal{M}_k$ and $\mathcal{P}$ respectively.

Proof The Lagrangian function of problem (11) is

$$\begin{aligned} L(B, X, \mathbf{y}; U, V, W) := \frac{1}{2}\|M(\mathcal{A}) - B\|^2 + \sigma\langle E_0, X\rangle \\ + \langle U, B - \mathcal{P}(\mathbf{y})\rangle + \langle V, X - \mathcal{M}_k(\mathbf{y})\rangle + \langle W, \mathcal{L}_k(\mathbf{y})\rangle \end{aligned}$$

and problem (11) can be equivalently written as

$$\min_{X\succeq 0,\ \operatorname{rank}(B)\le r,\ \mathbf{y}} \quad \max_{U,V,W} L(B, X, \mathbf{y}; U, V, W).$$

Let

$$g(U, V, W) := \min_{X\succeq 0,\ \operatorname{rank}(B)\le r,\ \mathbf{y}} L(B, X, \mathbf{y}; U, V, W).$$

The Lagrangian dual problem of (11) is then (cf. [4])

$$\max_{U,V,W} g(U, V, W).$$

By a direct calculation, we have

$$\begin{aligned} &\min_{X\succeq 0,\ \operatorname{rank}(B)\le r,\ \mathbf{y}} L(B, X, \mathbf{y}; U, V, W) \\ &= \min_{\operatorname{rank}(B)\le r} \Big\{\frac{1}{2}\|M(\mathcal{A}) - B\|^2 + \langle U, B\rangle\Big\} \\ &\quad + \min_{X\succeq 0} \big\{\langle \sigma E_0 + V, X\rangle\big\} + \min_{\mathbf{y}} \big\{\langle W, \mathcal{L}_k(\mathbf{y})\rangle - \langle V, \mathcal{M}_k(\mathbf{y})\rangle - \langle U, \mathcal{P}(\mathbf{y})\rangle\big\} \\ &= \min_{\operatorname{rank}(B)\le r} \Big\{\frac{1}{2}\big(\|B - (M(\mathcal{A}) - U)\|^2 - \|M(\mathcal{A}) - U\|^2 + \|M(\mathcal{A})\|^2\big)\Big\} \end{aligned}$$

$$
\begin{aligned}
&+ \min_{X \succeq 0} \left\{ \langle \sigma E_0 + V, X \rangle \right\} - \chi_{\{\mathbf{0}\}}(\mathcal{L}_k^*(W) - \mathcal{M}_k^*(V) - \mathcal{P}^*(U)) \\
&= \frac{1}{2}\left(\|M(\mathcal{A})\|^2 - \|\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - U)\|^2 \right) \\
&- \chi_{\mathrm{S}_+^{\zeta(n+1,k)}}(\sigma E_0 + V) - \chi_{\{\mathbf{0}\}}(\mathcal{L}_k^*(W) - \mathcal{M}_k^*(V) - \mathcal{P}^*(U)),
\end{aligned}
$$

where χ_S is the indicator function of a given set S, i.e., $\chi_S(\mathbf{x}) = 0$ if $\mathbf{x} \in S$, and ∞ otherwise. Therefore, we have the dual problem

$$
\begin{aligned}
\max\ & \tfrac{1}{2}\|M(\mathcal{A})\|^2 - \tfrac{1}{2}\|\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - U)\|^2 \\
\text{s.t. } & \mathcal{L}_k^*(W) - \mathcal{M}_k^*(V) - \mathcal{P}^*(U) = \mathbf{0}, \\
& \sigma E_0 + V \succeq 0.
\end{aligned}
$$

In a more concise form, it is

$$
\begin{aligned}
\max\ & \phi(U) := \tfrac{1}{2}\|M(\mathcal{A})\|^2 - \tfrac{1}{2}\|\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - U)\|^2 \\
\text{s.t. } & \mathcal{M}_k^*(Z) + \mathcal{P}^*(U) - \mathcal{L}_k^*(W) = \sigma \mathcal{M}_k^*(E_0), \\
& Z \succeq 0.
\end{aligned} \tag{13}
$$

We see that (12) is actually the minimization formulation of (13). □

Problem (13) (equivalently (12)) is a convex optimization problem, as expected, but with a nonsmooth objective function $\phi(U)$. One advantage of the formulation (12) is that we can interpret the constraint via sums of squares of polynomials.

Lemma 1 (Dual Feasibility). *A triplet (U, W, Z) is a feasible solution of problem (12) if and only if there exist a homogeneous polynomial $u(\mathbf{x})$, and polynomials $w(\mathbf{x})$ and $z(\mathbf{x})$ with*

$$
\deg(w(\mathbf{x})) \le 2k - 2,\ \deg(u(\mathbf{x})) = 3,\ \textit{and}\ \deg(z(\mathbf{x})) \le 2k
$$

such that

$$
z(\mathbf{x}) = (1 - \mathbf{x}^{\mathsf{T}}\mathbf{x})w(\mathbf{x}) - u(\mathbf{x}) + \sigma \tag{14}
$$

is a sum of squares of polynomials. Moreover, there is a correspondence between the triplets (U, W, Z) and $(u(\mathbf{x}), w(\mathbf{x}), z(\mathbf{x}))$ as indicated in Appendix A.

Proof Recall the sizes of the triplet (U, W, Z), which are

$$
U \in \mathbb{R}^{n \times n^2},\ W \in \mathrm{S}^2(\mathbb{R}^{\zeta(n+1,k-1)}) \text{ and } Z \in \mathrm{S}^2(\mathbb{R}^{\zeta(n+1,k)}).
$$

Let $\mathbf{x}^{\circ k}$ be the monomial basis up to order k defined as in (A1) and let $\mathbf{x}^{\otimes 2}$ be the extended monomial basis of order 2 defined as in (A2). Let

$$
w(\mathbf{x}) := (\mathbf{x}^{\circ(k-1)})^{\mathsf{T}} W \mathbf{x}^{\circ(k-1)},\ u(\mathbf{x}) := \mathbf{x}^{\mathsf{T}} U \mathbf{x}^{\otimes 2},\ \text{and } z(\mathbf{x}) := (\mathbf{x}^{\circ k})^{\mathsf{T}} Z \mathbf{x}^{\circ k}.
$$

Then by the feasibility of (U, W, Z), we have that

$$(1 - \mathbf{x}^{\mathsf{T}}\mathbf{x})w(\mathbf{x}) + \sigma - u(\mathbf{x}) - z(\mathbf{x}) = 0.$$

Since Z is a positive semidefinite matrix, which is equivalent to having the polynomial $z(\mathbf{x})$ being a sum of squares of polynomials [33], the conclusion follows. □

Lemma 2 (Strict Feasibility). *For any positive $\sigma > 0$ and integer $k \geq 2$, problem (12) is strictly feasible, i.e., there exists a triplet (U, W, Z) with $Z \succ 0$ such that $\mathcal{M}_k^*(Z) + \mathcal{P}^*(U) - \mathcal{L}_k^*(W) = \sigma \mathcal{M}_k^*(E_0)$.*

Proof First note that for any given positive scalars μ_i $(i = 0, \ldots, k)$, we can find a positive definite diagonal matrix A such that

$$(\mathbf{x}^{\circ k})^{\mathsf{T}} A \mathbf{x}^{\circ k} = \sum_{i=0}^{k} \mu_i (\mathbf{x}^{\mathsf{T}}\mathbf{x})^i.$$

Thus, the conclusion will follow, by Lemma 1, if we can find polynomials $w(\mathbf{x})$ and $u(\mathbf{x})$ such that

$$(1 - \mathbf{x}^{\mathsf{T}}\mathbf{x})w(\mathbf{x}) + \sigma - u(\mathbf{x}) = \sum_{i=0}^{k} \mu_i (\mathbf{x}^{\mathsf{T}}\mathbf{x})^i$$

for positive μ_i's. This can be fulfilled by taking $u(\mathbf{x}) = 0$ and

$$w(\mathbf{x}) := \sum_{i=0}^{k-1} \lambda_i (\mathbf{x}^{\mathsf{T}}\mathbf{x})^i$$

for any choices of λ_i's such that $-\sigma < \lambda_0 < \lambda_1 < \cdots < \lambda_{k-1} < 0$. In this case,

$$\mu_0 = \sigma + \lambda_0 > 0,\ \mu_i = \lambda_i - \lambda_{i-1} > 0 \text{ for } i = 1, \ldots, k-1 \text{ and } \mu_k = -\lambda_{k-1} > 0.$$

This completes the proof. □

The strict feasibility does not hold for $\sigma = 0$. The cubic form $u(\mathbf{x})$ either is zero or takes positive value on the sphere. Thus, $z(\mathbf{x}) = (1 - \mathbf{x}^{\mathsf{T}}\mathbf{x})w(\mathbf{x}) - u(\mathbf{x})$ either is identically zero or takes negative function value on the sphere. Consequently, if $\sigma = 0$, the feasibility condition (14) forces $U = 0$, and hence problem (12) becomes a feasibility problem with constant objective function. Then the possibility for strong duality between (11) and (13) is weakened. As a result, it is necessary for numerical reasons to impose positive σ and therefore the relationship between the optimal solutions for (11) and those for the original best approximation problem (2) should be established.

3.3 Optimality

The following conclusion is classical, which follows from the saddle point theorem [4].

Proposition 6 (Lagrangian Duality). *Let* $(B, X, \mathbf{y})$ *and* (U, V, W) *be feasible solutions of problems* (11) *and* (12) *respectively. Then we have*

$$\psi(B, X) \geq \phi(U).$$

If $\psi(B, X) = \phi(U)$*, then both* $(B, X, \mathbf{y})$ *and* (U, V, W) *are optimal solutions of problems* (11) *and* (12) *respectively.*

While the primal problem (11) is a nonlinear semidefinite matrix optimization problem which is nonconvex due to the rank constraint, the dual problem (12) is a nonlinear convex semidefinite matrix optimization problem. By the feasibility characterization in Lemma 1, the optimality of (12) can be concisely determined with the help of convex analysis [47].

Proposition 7 (Optimality). *We have that a feasible solution* $(\bar{U}, \bar{W}, \bar{Z})$ *of* (12) *is an optimal solution if there exists a vector* $\bar{\mathbf{y}} \in \mathbb{R}^{\zeta(n+1,2k)}$ *such that*

$$\begin{aligned} &\mathcal{L}_k(\bar{\mathbf{y}}) = 0,\ \mathcal{M}_k(\bar{\mathbf{y}}) \succeq 0,\ \langle \bar{Z}, \mathcal{M}_k(\bar{\mathbf{y}})\rangle = 0,\ \textit{and} \\ &\mathcal{P}(\bar{\mathbf{y}}) \in \operatorname{conv}\left(\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - \bar{U})\right). \end{aligned} \tag{15}$$

It becomes also a necessary condition if $\sigma > 0$.

Proof This follows from Lemma 2, Lemma 7 in Appendix B and standard convex analysis [47, Theorem 27.4].

The sufficiency is important for our subsequent analysis and we give a proof by the following direct calculation. For any feasible (U, W, Z) of (12), we have

$$\begin{aligned} \frac{1}{2}\|\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - U)\|^2 &\geq \frac{1}{2}\|\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - \bar{U})\|^2 + \langle -\mathcal{P}(\bar{\mathbf{y}}), U - \bar{U}\rangle \\ &= \frac{1}{2}\|\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - \bar{U})\|^2 - \langle \bar{\mathbf{y}}, \mathcal{P}^*(U) - \mathcal{P}^*(\bar{U})\rangle \\ &= \frac{1}{2}\|\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - \bar{U})\|^2 - \langle \bar{\mathbf{y}}, \mathcal{L}_k^*(W) - \mathcal{L}_k^*(\bar{W})\rangle \\ &\quad + \langle \bar{\mathbf{y}}, \mathcal{M}_k^*(Z) - \mathcal{M}_k^*(\bar{Z})\rangle \\ &= \frac{1}{2}\|\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - \bar{U})\|^2 + \langle \mathcal{M}_k(\bar{\mathbf{y}}), Z - \bar{Z}\rangle \\ &= \frac{1}{2}\|\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - \bar{U})\|^2 + \langle \mathcal{M}_k(\bar{\mathbf{y}}), Z\rangle \\ &\geq \frac{1}{2}\|\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - \bar{U})\|^2, \end{aligned}$$

where the first inequality follows from Lemma 7 and (15), the second equality from the feasibility, and the rest all follow from (15). □

We are in the position to present one of our main results.

Theorem 8 (Dual Certification). *If there exist a triplet* $(\bar{U}, \bar{W}, \bar{Z})$ *such that the feasibility of problem* (12) *is satisfied and a vector* $\bar{\mathbf{y}}$ *such that the optimality condition* (15) *is satisfied, and*

$$\operatorname{rank}(\mathcal{P}(\bar{\mathbf{y}})) \leq r, \tag{16}$$

then strong duality holds for problems (11) *and* (12)*, and* $\bar{\mathbf{y}}$ *gives an optimal solution to problem* (11)*.*

Proof Let

$$\bar{B} := \mathcal{P}(\bar{\mathbf{y}}) \text{ and } \bar{X} := \mathcal{M}_k(\bar{\mathbf{y}}).$$

Then, $(\bar{B}, \bar{X}, \bar{y})$ is a feasible solution for (11) by (15) and (16). Moreover, $\bar{B} \in \Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - \bar{U})$ by Lemma 8 and (16). Thus,

$$\begin{aligned}
\phi(\bar{U}) &= \frac{1}{2}\|M(\mathcal{A})\|^2 - \frac{1}{2}\|\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - \bar{U})\|^2 \\
&= \frac{1}{2}\|M(\mathcal{A})\|^2 + \frac{1}{2}\|M(\mathcal{A}) - \bar{U} - \Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - \bar{U})\|^2 - \frac{1}{2}\|M(\mathcal{A}) - \bar{U}\|^2 \\
&= \frac{1}{2}\|M(\mathcal{A})\|^2 + \frac{1}{2}\|M(\mathcal{A}) - \bar{U} - \bar{B}\|^2 - \frac{1}{2}\|M(\mathcal{A}) - \bar{U}\|^2 \\
&= \frac{1}{2}\|M(\mathcal{A}) - \bar{B}\|^2 + \langle \bar{U}, \bar{B} \rangle.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\phi(\bar{U}) &= \frac{1}{2}\|M(\mathcal{A}) - \bar{B}\|^2 + \langle \bar{U}, \bar{B} \rangle \\
&= \frac{1}{2}\|M(\mathcal{A}) - \bar{B}\|^2 + \langle \bar{U}, \mathcal{P}(\bar{\mathbf{y}}) \rangle \\
&= \frac{1}{2}\|M(\mathcal{A}) - \bar{B}\|^2 + \langle \mathcal{P}^*(\bar{U}), \bar{\mathbf{y}} \rangle \\
&= \frac{1}{2}\|M(\mathcal{A}) - \bar{B}\|^2 + \langle \mathcal{L}_k^*(\bar{W}) - \mathcal{M}_k^*(\bar{Z}) + \sigma \mathcal{M}_k^*(E_0), \bar{\mathbf{y}} \rangle \\
&= \frac{1}{2}\|M(\mathcal{A}) - \bar{B}\|^2 + \langle \bar{W}, \mathcal{L}_k(\bar{\mathbf{y}}) \rangle - \langle \bar{Z}, \mathcal{M}_k(\bar{\mathbf{y}}) \rangle + \sigma \langle E_0, \mathcal{M}_k(\bar{\mathbf{y}}) \rangle \\
&= \frac{1}{2}\|M(\mathcal{A}) - \bar{B}\|^2 + \sigma \langle E_0, \bar{X} \rangle
\end{aligned}$$

where the fourth equality follows from the feasibility of (12) and the last equality follows from (15). Thus, by (11),

$$\psi(\bar{B}, \bar{X}) = \phi(\bar{U}).$$

By Lagrangian duality (cf. Proposition 6), we get the conclusion. □

4 Quality of approximation

In Theorem 8, we established a certification for global optimality of problem (11). In this section, we will give the relationships between the global optimal solutions of (11) and the original best low rank approximation problem (2). The discussions are divided into two subsections based on the control parameter σ: the case $\sigma = 0$ and the case $\sigma > 0$.

4.1 Best low rank approximation

In this subsection, we consider the case $\sigma = 0$ in problem (11), which is related to the best low rank approximation.

Let $\mathbb{O}(n) \subset \mathbb{R}^{n\times n}$ be the group of orthogonal matrices and $Q \in \mathbb{O}(n)$. For a given $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$ with $\mathcal{A} = (a_{ijk})$, the *matrix-tensor multiplication* $(Q, Q, Q)\cdot\mathcal{A}$ is defined as a tensor in $\mathrm{S}^3(\mathbb{R}^n)$ with its components being

$$\big[(Q, Q, Q)\cdot\mathcal{A}\big]_{rst} = \sum_{i,j,k=1}^{n} q_{ri}q_{sj}q_{tk}a_{ijk} \text{ for all } r, s, t \in \{1, \ldots, n\}. \tag{17}$$

This multiplication can be extended for general cases in an obvious way.

Lemma 3 *If $\{\mathbf{x}_1, \ldots, \mathbf{x}_r\} \subset \mathbb{R}^n$ has rank exactly r, then*

$$\operatorname{rank}\left(\sum_{i=1}^{r+1}\lambda_i\mathbf{x}_i(\mathbf{x}_i^{\otimes 2})^{\mathsf{T}}\right) \le r$$

for any $\mathbf{x}_{r+1} \in \operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$ and $\lambda_i \ge 0$.

Proof Let $\mathbf{x}_{r+1} = \sum_{j=1}^{r}\alpha_j\mathbf{x}_j$ for coefficients α_j's. We have

$$\begin{aligned}\sum_{i=1}^{r+1}\lambda_i\mathbf{x}_i(\mathbf{x}_i^{\otimes 2})^{\mathsf{T}} &= \sum_{i=1}^{r}\lambda_i\mathbf{x}_i(\mathbf{x}_i^{\otimes 2})^{\mathsf{T}} + \lambda_{r+1}\left(\sum_{j=1}^{r}\alpha_j\mathbf{x}_j\right)(\mathbf{x}_{r+1}^{\otimes 2})^{\mathsf{T}}\\ &= \sum_{i=1}^{r}\mathbf{x}_i(\lambda_i\mathbf{x}_i^{\otimes 2} + \lambda_{r+1}\alpha_i\mathbf{x}_{r+1}^{\otimes 2})^{\mathsf{T}},\end{aligned}$$

which is a rank one decomposition with rank at most r. □

Lemma 4 *Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_s\} \subset \mathbb{R}^n$ be a set of nonzero vectors of rank $s-1$. Then, the matrix $B := \sum_{i=1}^{s}\mathbf{x}_i(\mathbf{x}_i^{\otimes 2})^{\mathsf{T}}$ has rank at least $s-2$, and B has rank $s-2$ if and only if $\mathbf{x}_i = -\mathbf{x}_j$ for some $i \ne j$.*

Proof Without loss of generality, we assume that $\mathbf{x}_1, \ldots, \mathbf{x}_{s-1}$ are linearly independent. There exists a nonsingular matrix $P \in \mathbb{R}^{n\times n}$ such that

$$P\mathbf{x}_i = \mathbf{e}_i \text{ for all } i = 1, \ldots, s-1.$$

We have

$$PB(P \otimes P)^{\mathsf{T}} = \sum_{i=1}^{s} (P\mathbf{x}_i)[(P\mathbf{x}_i)^{\otimes 2}]^{\mathsf{T}}.$$

Thus, without loss of generality, we assume that

$$\mathbf{x}_i = \mathbf{e}_i \text{ for all } i = 1, \ldots, s-1 \text{ and } \mathbf{x}_s = \sum_{i=1}^{s-1} \alpha_i \mathbf{e}_i.$$

Let $Q \in \mathbb{R}^{n^2 \times n^2}$ be the permutation matrix such that

$$\sum_{i=1}^{s-1} \mathbf{e}_i (\mathbf{e}_i^{\otimes 2})^{\mathsf{T}} Q = \begin{bmatrix} I_{s-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

Thus,

$$BQ = \begin{bmatrix} I_{s-1} & 0 \\ 0 & 0 \end{bmatrix} + \left(\sum_{i=1}^{s-1} \alpha_i \mathbf{e}_i \right) (\mathbf{x}_s^{\otimes 2})^{\mathsf{T}} Q.$$

As the right most term is rank one, we have $\operatorname{rank}(B) = \operatorname{rank}(BQ) \geq s-2$. If $\operatorname{rank}(B) = s-2$, we must have that the $(s-1) \times (s-1)$ leading principal submatrix of BQ, which is

$$I_{s-1} + \tilde{\mathbf{x}}_s [(\mathbf{x}_s^{\otimes 2})^{\mathsf{T}} Q]_{1:s-1},$$

has rank $s-2$, where $\tilde{\mathbf{x}}_s = (\mathbf{x}_s)_{1:s-1}$. Without loss of generality, assume that $\{1, \ldots, p\}$ are exactly the nonzero components of $\mathbf{x}_s$. Then, BQ has the following block form

$$\begin{bmatrix} I_p + \mathbf{u}\mathbf{v}^{\mathsf{T}} & 0 & \mathbf{u}\mathbf{w}^{\mathsf{T}} & 0 \\ 0 & I_{s-1-p} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{18}$$

where $\mathbf{u} := (\mathbf{x}_s)_{1:p}$ and $\mathbf{v} := [(\mathbf{x}_s^{\otimes 2})^{\mathsf{T}} Q]_{1:p}$, and the third block columns corresponding to the square-free elements of $\mathbf{x}_{1:p}^{\otimes 2}$. If $p = 1$, then $\mathbf{w}$ is vacuous, and if $p > 1$, then $\mathbf{w}$ is a nonzero vector. We know that the matrix $I_p + \mathbf{u}\mathbf{v}^{\mathsf{T}}$ is singular if and only if $\langle \mathbf{u}, \mathbf{v} \rangle = -1$. Thus, when BQ has rank $s-2$, we have that $\langle \mathbf{u}, \mathbf{v} \rangle = -1$. Therefore, the vector $\mathbf{v}$, together with a basis for the orthogonal complement of $\mathbf{u}$, gives a basis for the whole space $\mathbb{R}^p$. Thus, each vector $\mathbf{x} \in \mathbb{R}^p$ can be written uniquely in the form $\mathbf{x} = \mathbf{u}_0 + \mu\mathbf{v}$ for a vector $\mathbf{u}_0$ orthogonal to $\mathbf{u}$ and $\mu \in \mathbb{R}$. Thus,

$$\mathbf{x}^{\mathsf{T}} \begin{bmatrix} I_p + \mathbf{u}\mathbf{v}^{\mathsf{T}} & 0 & \mathbf{u}\mathbf{w}^{\mathsf{T}} & 0 \end{bmatrix} = (\mathbf{u}_0^{\mathsf{T}}, \mathbf{0}, -\mu\mathbf{w}^{\mathsf{T}}, \mathbf{0}). \tag{19}$$

Since $\mathbf{w}$ is a nonzero vector, the resulting vector in (19) is nonzero whenever $\mathbf{x} \neq \mathbf{0}$. Consequently, the left null space of the matrix $\begin{bmatrix} I_p + \mathbf{u}\mathbf{v}^{\mathsf{T}} & 0 & \mathbf{u}\mathbf{w}^{\mathsf{T}} & 0 \end{bmatrix}$ is trivial and this matrix has full rank. Hence, the matrix in (18) has full rank $s-1$. As a result, we

must have that $p = 1$. In this case, $\langle \mathbf{u}, \mathbf{v} \rangle = -1$ implies that $\mathbf{x}_s = -\mathbf{e}_1$, and hence for the general case we have

$$\mathbf{x}_s = -\mathbf{e}_i \text{ for some } i \in \{1, \ldots, s-1\}.$$

The sufficiency is clear. Consequently, the conclusion follows. □

Proposition 3 indicates that (9) is an exact relaxation of the best low rank approximation problem (2). The following result characterizes the approximation quality of the further relaxation (11) to (2).

Theorem 9 *Let* $\mathbf{y}$ *be an optimizer of problem* (11) *with* $k = 2$ *and* $\sigma = 0$ *satisfying* $\operatorname{rank}(\mathcal{M}_1(\mathbf{y})) = \operatorname{rank}(\mathcal{M}_2(\mathbf{y}))$. *Then*

$$\operatorname{rank}(\mathbf{y}) = \operatorname{rank}(\mathcal{M}_2(\mathbf{y})) \leq r + 2. \tag{20}$$

Moreover,

1. *if* $\operatorname{rank}(\mathcal{M}_2(\mathbf{y})) \leq r$ *or* $\operatorname{rank}(\mathcal{M}_2(\mathbf{y})) = r + 2$ *or* $r = 1$, *then* $\mathcal{P}(\mathbf{y})$ *is an optimal solution for* (2)*;*
2. *if* $\operatorname{rank}(\mathcal{M}_2(\mathbf{y})) = r + 1$ *and* $r > 1$, *then a feasible solution* $\mathcal{B}$ *of* (2) *can be constructed from* $\mathbf{y}$ *such that*

$$\|\mathcal{A} - \mathcal{B}\|^2 \leq \|\mathcal{A} - \mathcal{B}^*\|^2 + \rho(\mathcal{A} - \mathcal{B})^2$$

where $\mathcal{B}^*$ *is an optimal solution of* (2).

Proof By Proposition 3, for the conclusion (20), it is sufficient to show that $\operatorname{rank}(\mathcal{M}_2(\mathbf{y})) \leq r + 2$.

Let $\operatorname{rank}(\mathcal{M}_2(\mathbf{y})) = s$ for some integer $s \geq 0$. In the following, we consider the case $s > r$, since when $s \leq r$, the optimal $B = \mathcal{P}(\mathbf{y})$ corresponds to a tensor of rank at most r. Thus, it gives an optimal solution for (2).

By the fact that $\mathbf{y}$ satisfies the second flatness condition, we have that there exist $\lambda_i > 0$ and $\mathbf{x}_i \in \mathbb{S}^{n-1}$ for all $i = 1, \ldots, s$ such that

$$\mathbf{y} = \int_{\mathbb{S}^{n-1}} \mathbf{x}^{\circ 4} \, \mathrm{d}\, \mu(\mathbf{x})$$

with $\mu := \sum_{i=1}^{s} \lambda_i \delta_{\mathbf{x}_i}$. It then follows from the definition that

$$\mathcal{M}_1(\mathbf{y}) = \sum_{i=1}^{s} \lambda_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^{\mathsf{T}}$$

with

$$\bar{\mathbf{x}}_i := \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix} \text{ for all } i = 1, \ldots, s.$$

By the flatness condition, $\operatorname{rank}(\mathcal{M}_1(\mathbf{y})) = s > 0$. Thus, the set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_s\}$ has rank at least $s - 1$. On the other hand, by the feasibility of $\mathbf{y}$ for (11), it follows that the matrix

$$\bar{B} := \sum_{i=1}^{s} \lambda_i \mathbf{x}_i (\mathbf{x}_i^{\otimes 2})^{\mathsf{T}}$$

is of rank not greater than r. If the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_s$ are linearly independent, then the corresponding matrix $\bar{B}$ must have rank $s > r$, which is a contradiction. Thus, the set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_s\}$ has rank $s - 1$. Consequently, by Lemma 4, we have that $r < s \leq r + 2$. The conclusion (20) then follows.

Moreover, by Lemma 4 again, we have that $s = r + 2$ exactly when $\sqrt[3]{\lambda_i}\mathbf{x}_i = -\sqrt[3]{\lambda_j}\mathbf{x}_j$ for a pair $i \neq j$. But in this case, the terms $\lambda_i \mathbf{x}_i^{\otimes 3}$ and $\lambda_j \mathbf{x}_j^{\otimes 3}$ combined into a zero sum. Consequently,

$$\bar{B} = \sum_{k \in \{1,\ldots,s\} \setminus \{i,j\}} \lambda_k \mathbf{x}_k (\mathbf{x}_k^{\otimes 2})^{\mathsf{T}},$$

which corresponds to a tensor of rank $s - 2 = r$. Similar to the case $s = r$, an optimal solution for (2) is found.

In the following, we consider the case $s = r + 1$.

We assume, without loss of generality, that the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_r$ are linear independent and $\mathbf{x}_{r+1} \in \operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$. If $r = 1$, then $\mathbf{x}_{r+1} = \mathbf{x}_2 = \pm\mathbf{x}_1$ and $B = \mathcal{P}(\mathbf{y})$ corresponds to a tensor of rank at most one. Hence, it is an optimal solution.

In the following, we assume that $r > 1$. By the equivalence of the matrix flattening, we have that

$$\|\mathcal{A} - \sum_{i=1}^{r+1} \lambda_i \mathbf{x}_i^{\otimes 3}\|^2 \leq \|\mathcal{A} - \mathcal{B}^*\|^2, \tag{21}$$

where $\mathcal{B}^*$ is an optimal solution to problem (2), since (11) is a relaxation of (2). It follows from the definition that (cf. (17))

$$\|(P, P, P) \cdot \mathcal{A}\| = \|\mathcal{A}\|$$

for an orthogonal matrix $P \in \mathbb{O}(n)$. Thus, we can assume without loss of generality that

$$\operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_r\} = \operatorname{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_r\}, \tag{22}$$

where $\mathbf{e}_i \in \mathbb{R}^n$ is the i-th column vector of the identity matrix of matching size for all $i \in \{1, \ldots, r\}$.

On the other hand, by Lemma 3, it holds that

$$B := \sum_{i=1}^{r} \lambda_i \mathbf{x}_i (\mathbf{x}_i^{\otimes 2})^{\mathsf{T}} + \beta \mathbf{x} (\mathbf{x}^{\otimes 2})^{\mathsf{T}} \tag{23}$$

has rank at most r for any $\mathbf{x} \in \operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$ and $\beta \geq 0$. Hence, every β and $\mathbf{x} \in \operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$, together with $\sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3}$, give a feasible solution for problem

(11) with the corresponding B as (23). Note that $\mathbf{x}_{r+1} \in \mathrm{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$ and (22) holds. Thus, by the global optimality of $\sum_{i=1}^{r+1} \lambda_i \mathbf{x}_i^{\otimes 3}$ and the fact that

$$\begin{aligned}\frac{1}{2}\|M(\mathcal{A}) - B\|^2 &= \frac{1}{2}\left\|\left(\mathcal{A} - \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3} - \beta \mathbf{x}^{\otimes 3}\right)_{1:r}\right\|^2 + c \\ &= \frac{1}{2}\left\|\left(\mathcal{A} - \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3}\right)_{1:r} - \beta\left(\mathbf{x}^{\otimes 3}\right)_{1:r}\right\|^2 + c\end{aligned}$$

for a constant c, we can conclude that $\lambda_{r+1}((\mathbf{x}_{r+1})_{1:r})^{\otimes 3}$ is a best rank one approximation of the sub-tensor $(\mathcal{A} - \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3})_{1:r}$. Here $\mathbf{u}_{1:r} \in \mathbb{R}^r$ is the sub-vector of $\mathbf{u} \in \mathbb{R}^n$ formed by the first r entries $u_1, \ldots, u_r$, and $\mathcal{U}_{1:r} \in \mathrm{S}^3(\mathbb{R}^r)$ is the sub-tensor of $\mathcal{U} \in \mathrm{S}^3(\mathbb{R}^n)$ formed by the entries $u_{i_1 i_2 i_3}$ with $i_1, i_2, i_3 \in \{1, \ldots, r\}$. Moreover, if $\mathbf{x}_{r+1}$ was determined, by expanding the quadratic objective $\frac{1}{2}\left\|\left(\mathcal{A} - \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3}\right)_{1:r} - \beta\left(\mathbf{x}_{r+1}^{\otimes 3}\right)_{1:r}\right\|^2$ with respect to β, the optimal $\beta = \lambda_{r+1}$ should be $\left\langle\left(\mathcal{A} - \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3}\right)_{1:r}, \left(\mathbf{x}_{r+1}^{\otimes 3}\right)_{1:r}\right\rangle$. Thus, the global optimality implies that

$$\lambda_{r+1}^2 = \rho\left((\mathcal{A} - \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3})_{1:r}\right)^2 \le \rho(\mathcal{A} - \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3})^2 \tag{24}$$

and

$$\left\|\mathcal{A} - \sum_{i=1}^{r+1} \lambda_i \mathbf{x}_i^{\otimes 3}\right\|^2 = \left\|\mathcal{A} - \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3}\right\|^2 - \lambda_{r+1}^2. \tag{25}$$

Therefore, with $\mathcal{B} := \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3}$, (21), (24) and (25), we have

$$\|\mathcal{A} - \mathcal{B}\|^2 \le \|\mathcal{A} - \mathcal{B}^*\|^2 + \rho(\mathcal{A} - \mathcal{B})^2.$$

This completes the proof. □

It would be interesting to study the counterparts of Theorem 9 for the case $k > 2$. One difficulty on extending the above analysis is the rank estimation of the k-th order moment matrix from that of the matrix B. While if $\mathrm{rank}(\mathbf{y}) \le r$, we see that $\mathcal{P}(\mathbf{y})$ is an optimal solution by Proposition 3.

In the sequel, we show the exact relaxation of (11) when the given tensor is orthogonally decomposable. To do this, recall that a third order symmetric tensor $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$ is called *orthogonally decomposable* (cf. [27, 54][1]) if there exist an orthonormal matrix

$$A = [\mathbf{a}_1, \ldots, \mathbf{a}_r] \in \mathbb{R}^{n \times r}$$

[1] In [27], this notion was referred as completely orthogonally decomposable tensors by Kolda.

and positive numbers $\lambda_i \in \mathbb{R}$ for $i = 1, \ldots, r$ such that

$$\mathcal{A} = \sum_{i=1}^{r} \lambda_i \mathbf{a}_i^{\otimes 3}. \tag{26}$$

The number r is the rank of the tensor $\mathcal{A}$. It is well-known that the rank one decomposition (26) of an orthogonally decomposable tensor $\mathcal{A}$ is unique [54].

Theorem 10 (Exact Relaxation). *Let $k \geq 2$. If $\mathcal{A}$ is an orthogonally decomposable tensor with rank $s \leq n$, then for $\sigma = 0$ and $r \leq s$, (11) is an exact relaxation of the best rank-r approximation problem.*

Proof Suppose that

$$\mathcal{A} = \sum_{i=1}^{s} \lambda_i \mathbf{x}_i^{\otimes 3}$$

be an orthogonal decomposition of $\mathcal{A}$ with $\lambda_1 \geq \cdots \geq \lambda_s > 0$. Let

$$\mu := \sum_{i=1}^{r} \lambda_i \delta_{\mathbf{x}_i}$$

and $\bar{\mathbf{y}}$ the corresponding moment sequence generated by the r-atomic measure μ. It is immediate to see that $\bar{\mathbf{y}}$ is a feasible solution of problem (11).

A dual feasible solution for (13) is $(U, V, W) = (0, 0, 0)$. Obviously,

$$\operatorname{rank}(\mathcal{P}(\bar{\mathbf{y}})) \leq r,$$

since

$$\mathcal{P}(\bar{\mathbf{y}}) = \sum_{i=1}^{r} \lambda_i \mathbf{x}_i (\mathbf{x}_i^{\otimes 2})^{\mathsf{T}}.$$

If $\mathcal{P}(\bar{\mathbf{y}}) \in \operatorname{conv}(\Pi_{\mathrm{R}(r)}(M(\mathcal{A})))$, then by Proposition 7 and Theorem 8, we can conclude that $\bar{\mathbf{y}}$ is an optimal solution of problem (11). Consequently, it is an optimizer of the best rank-r approximation problem by Theorem 9, since the flatness is satisfied.

Note that

$$M(\mathcal{A}) = \sum_{i=1}^{s} \lambda_i \mathbf{x}_i (\mathbf{x}_i^{\otimes 2})^{\mathsf{T}} \tag{27}$$

and

$$\left[\mathbf{x}_1, \ldots, \mathbf{x}_s\right]$$

is an orthonormal matrix. Likewise, the matrix

$$\left[\mathbf{x}_1^{\otimes 2}, \ldots, \mathbf{x}_s^{\otimes 2}\right]$$

is also orthonormal. Thus (27) is a singular value decomposition (a.k.a. SVD [15]) of the matrix $M(\mathcal{A})$. By the classical Eckart–Young–Mirsky theorem (cf. [21]), the

truncated SVD is an optimizer of the best rank-r approximation problem for the matrix $M(\mathcal{A})$. Thus, $\mathcal{P}(\bar{\mathbf{y}}) \in \mathrm{conv}(\Pi_{\mathrm{R}(r)}(M(\mathcal{A})))$. The result then follows. □

4.2 Quasi-optimality

In this section, we discuss the case when $\sigma > 0$ in problem (11), which is related to quasi-optimal low rank approximations of the given tensor.

Definition 1 Given a nonzero tensor $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$ and a positive integer r, let $\overline{\mathcal{B}}$ be a best rank-r approximation of $\mathcal{A}$ and $\alpha \geq 0$. A tensor $\mathcal{B} \in \mathrm{S}^3(\mathbb{R}^n)$ is called a α-quasi-optimal rank-r approximation of $\mathcal{A}$ if

$$\|\mathcal{A} - \overline{\mathcal{B}}\|^2 \leq \|\mathcal{A} - \mathcal{B}\|^2 \leq \|\mathcal{A} - \overline{\mathcal{B}}\|^2 + \alpha.$$

We first show that optimal solutions of (11) can actually give best rank one approximants.

Proposition 11 *Let $k \geq 2$. Let $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$ be nonzero, $r = 1$, $(B, \mathbf{y}, X)$ be an optimal solution of (11) with $\sigma > 0$. Then, we have $B = X(1,1)\mathbf{x}(\mathbf{x}^{\otimes 2})^{\mathsf{T}}$ for some $\mathbf{x} \in \mathbb{S}^{n-1}$. If $\sigma < \rho(\mathcal{A})$, then $(X(1,1) + \sigma)\mathbf{x}^{\otimes 3}$ is a best rank one approximation of $\mathcal{A}$.*

Proof Note that $B = \mathcal{P}(\mathbf{y})$ has rank at most one by the feasibility, then it follows that

$$B = \lambda \mathbf{x}(\mathbf{x}^{\otimes 2})^{\mathsf{T}}$$

for some unit vector $\mathbf{x} \in \mathbb{S}^{n-1}$ and $\lambda \geq 0$, since a third order symmetric tensor $\mathcal{U}$ has rank one if and only if its flattening matrix $M(\mathcal{U})$ has rank one [31]. First of all, we will derive from the fact that the matrix $X = \mathcal{M}_k(\mathbf{y})$ is positive semidefinite and $\mathcal{L}_k(\mathbf{y}) = 0$ that $X(1,1) \geq \lambda$. Actually, it follows from $\mathcal{L}_k(\mathbf{y}) = 0$ that

$$\mathcal{M}_1(\mathbf{y}) = \begin{bmatrix} \beta & \lambda \mathbf{x}^{\mathsf{T}} \\ \lambda \mathbf{x} & A \end{bmatrix}$$

for some positive semidefinite matrix $A \in \mathrm{S}^2(\mathbb{R}^n)$. The case when $\lambda = 0$ is trivial. In the following, we assume that $\lambda > 0$ and thus $\beta > 0$ by the positive semidefiniteness of $\mathcal{M}_1(\mathbf{y})$. By Schur's complement theory [21], we have that $A \succeq \frac{\lambda^2}{\beta}\mathbf{x}\mathbf{x}^{\mathsf{T}}$. While, it follows from $\mathcal{L}_k(\mathbf{y}) = 0$ that

$$\beta = \mathrm{tr}(A) \geq \frac{\lambda^2}{\beta}.$$

Thus $X(1,1) = \beta \geq \lambda$. The result then follows.

As a result, the optimal X of (11) must have the smallest possible $X(1,1)$, which is λ. Thus, X must be of rank one and $\|B\| = \lambda = X(1,1)$. Let the optimal solution be $\lambda \mathbf{x}^{\otimes 3}$. We must have

$$\frac{1}{2}\|M(\mathcal{A}) - \lambda \mathbf{x}(\mathbf{x}^{\otimes 2})^{\mathsf{T}}\|^2 + \sigma\lambda \leq \min_{\mu \geq 0, \mathbf{z} \in \mathbb{S}^{n-1}} \left\{ \frac{1}{2}\|M(\mathcal{A}) - \mu \mathbf{z}(\mathbf{z}^{\otimes 2})^{\mathsf{T}}\|^2 + \sigma\mu \right\}$$

by the optimality. If the optimal $\mathbf{z}$ was determined as $\mathbf{x}$, then the optimal μ can be computed explicitly since it is a univariate quadratic minimization over the nonnegative orthant. Expanding the objective function, we get

$$\frac{1}{2}\mu^2 + \mu(\sigma - \langle \mathcal{A}, \mathbf{x}^{\otimes 3}\rangle) + \frac{1}{2}\|\mathcal{A}\|^2.$$

Thus, the optimal $\mu \geq 0$ is given by

$$\lambda = \max\{\langle \mathcal{A}, \mathbf{x}^{\otimes 3}\rangle - \sigma, 0\}.$$

Since $\sigma < \rho(\mathcal{A})$, there is a $\mathbf{z}$ such that $\langle \mathcal{A}, \mathbf{z}^{\otimes 3}\rangle - \sigma > 0$. Thus, by the global optimality, $\lambda = \langle \mathcal{A}, \mathbf{x}^{\otimes 3}\rangle - \sigma$ and

$$\frac{1}{2}\|M(\mathcal{A}) - \lambda \mathbf{x}(\mathbf{x}^{\otimes 2})^{\mathsf{T}}\|^2 + \sigma\lambda = \frac{1}{2}\|\mathcal{A}\|^2 - \frac{1}{2}\lambda^2.$$

Note that for the best rank one approximation objective $\frac{1}{2}\|\mathcal{A} - \nu\mathbf{z}^{\otimes 3}\|^2$, where the variables are ν and $\mathbf{z}$, we have that the final $\nu = \langle \mathcal{A}, \mathbf{z}^{\otimes 3}\rangle$ by the optimality with respect to ν. Thus, $\frac{1}{2}\|\mathcal{A} - \nu\mathbf{z}^{\otimes 3}\|^2 = \frac{1}{2}\|\mathcal{A}\|^2 - \frac{1}{2}\nu^2 = \frac{1}{2}\|\mathcal{A}\|^2 - \frac{1}{2}\langle \mathcal{A}, \mathbf{z}^{\otimes 3}\rangle^2$, and hence the best rank one approximation problem is equivalent to solve $\rho(\mathcal{A}) = \max_{\mathbf{z}\in\mathbb{S}^{n-1}} \langle \mathcal{A}, \mathbf{z}^{\otimes 3}\rangle$. It further follows from the optimality that

$$\lambda = \langle \mathcal{A}, \mathbf{x}^{\otimes 3}\rangle - \sigma = \max\{\langle \mathcal{A}, \mathbf{z}^{\otimes 3}\rangle - \sigma : \mathbf{z} \in \mathbb{S}^{n-1}\} = \rho(\mathcal{A}) - \sigma.$$

Hence, $(\lambda + \sigma)\mathbf{x}^{\otimes 3} = \rho(\mathcal{A})\mathbf{x}^{\otimes 3}$ is a best rank one approximation of $\mathcal{A}$. The conclusion then follows. □

The coherence of a matrix $A = [\mathbf{x}_1, \ldots, \mathbf{x}_r]$, denoted by $\mu(A)$, is defined as (cf. [37])

$$\mu(A) := \max_{i \neq j} |\langle \mathbf{x}_i, \mathbf{x}_j\rangle|.$$

It follows that $\mu(A)$ is the maximum absolute value of the off-diagonal elements of $A^{\mathsf{T}}A$.

Lemma 5 *Let $A = [\mathbf{x}_1, \ldots, \mathbf{x}_r]$ be a given matrix with unit columns, and $C = A^{\mathsf{T}}A \circ A^{\mathsf{T}}A \circ A^{\mathsf{T}}A$, where $\circ$ is the Hadamard product. Then,*

1. *if the r-th singular value of A is not smaller than a constant $\kappa > 0$, then C is positive definite with its smallest eigenvalue not smaller than κ^6, and*
2. *if $r > 1$ and the coherence $\mu(A) < \sqrt[3]{\frac{1}{r-1}}$, then C is positive definite with its smallest eigenvalue not smaller than $1 - (r-1)\mu(A)^3$.*

Proof For the first one, since the r-th singular value of A is not smaller than $\kappa > 0$, we see that $A^{\mathsf{T}}A$ is positive definite with the smallest eigenvalue being no smaller than κ^2. It follows from [19] that the smallest eigenvalue of C is not smaller than the smallest eigenvalue of $(A^{\mathsf{T}}A)^3$, which is lower bounded by κ^6.

The second one follows from a similar proof as that for [37, Theorem 25]. The conclusion follows. □

For $r > 1$, the set of tensors with rank at most r has complicated geometry [12, 31]. It may happen that the matrix of factor vectors $[\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_r^{(k)}]$ for a best rank-r approximation tensor sequence approaches to the boundary of the set of matrices with rank at most r. Thus, in addition to the existence Assumption 1, we should also make a well-conditioned assumption on a best rank-r approximation for a given tensor $\mathcal{A}$.

Assumption 2 *For a given tensor $\mathcal{A}$, there is a best rank-r approximation $\mathcal{B} = \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3}$ such that*

1. *either the r-th singular value of the factor matrix $A := [\mathbf{x}_1, \ldots, \mathbf{x}_r]$ is greater than a constant $\kappa(\mathcal{A}) > 0$;*
2. *or the coherence $\mu(A) < \sqrt[3]{\frac{1}{r-1}}$ and $r > 1$.*

Assumption 2 actually indicates that a rank one decomposition of a best rank-r approximant of $\mathcal{A}$ is well-conditioned. It only needs the existence of such a best approximant, and do not require that all best approximants satisfy this condition. If Assumption 2 is satisfied, we define

$$\tau(\mathcal{A}) := \begin{cases} \max\{\kappa(\mathcal{A})^6, 1-(r-1)\mu(A)^3\} & \text{if both (1) and (2) hold,} \\ \kappa(\mathcal{A})^6 & \text{if only (1) holds,} \\ 1-(r-1)\mu(A)^3 & \text{if only (2) holds.} \end{cases} \tag{28}$$

Note that Proposition 11 does not require the flatness condition. However, if $r > 1$, then we need to assume the flatness condition.

Proposition 12 *Let $k \geq 2$, $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$ be nonzero and have rank greater than two, $r \geq 2$, $\overline{\mathcal{B}}$ be a best rank-r approximant of $\mathcal{A}$ satisfying Assumption 2, and $(B, \mathbf{y}, X)$ be an optimal solution of (11) with $\sigma \in (0, \frac{\tau(\mathcal{A})\rho(\mathcal{A})}{2r})$. Suppose that $\mathbf{y}$ satisfies $\mathrm{rank}(\mathcal{M}_{k-1}(\mathbf{y})) = \mathrm{rank}(\mathcal{M}_k(\mathbf{y})) \leq r$. Then $\mathcal{B}$ gives a α-quasi-optimal rank-r approximation of $\mathcal{A}$ with α given by*

$$\alpha := 2\sqrt{\frac{r}{\tau(\mathcal{A})}}\left(\left(1-\sqrt{\frac{\tau(\mathcal{A})}{r}}\right)\|\mathcal{A}\| + 2\sigma\right)\sigma.$$

Proof By the flatness hypothesis, we know that $\mathbf{y}$ has an atomic measure with rank s at most r. Let it be $\mu = \sum_{i=1}^{s} \lambda_i \delta_{\mathbf{x}_i}$. Then $B = \sum_{i=1}^{s} \lambda_i \mathbf{x}_i \left(\mathbf{x}_i^{\otimes 2}\right)^{\mathsf{T}}$, and we have that

$$\frac{1}{2}\|M(\mathcal{A}) - B\|^2 + \sigma \mathbf{e}^{\mathsf{T}}\lambda \leq \frac{1}{2}\|M(\mathcal{A}) - \overline{B}\|^2 + \sigma \mathbf{e}^{\mathsf{T}}\overline{\lambda}, \tag{29}$$

where $\mathbf{e}$ is the vector of all ones, $\overline{B} = \sum_{i=1}^{r} \overline{\lambda}_i \overline{\mathbf{x}}_i (\overline{\mathbf{x}}_i^{\otimes 2})^{\top}$ (with $\overline{\lambda}_i > 0$ and $\|\overline{\mathbf{x}}_i\| = 1$ for all $i = 1, \ldots, r$) corresponds to the best rank-r approximation $\overline{\mathcal{B}}$ of $\mathcal{A}$ and $\overline{\lambda}$ is the

corresponding vector of positive coefficients. Then, we have

$$\frac{1}{2}\|M(\mathcal{A}) - B\|^2 \leq \frac{1}{2}\|M(\mathcal{A}) - \overline{B}\|^2 + \sigma\left(\mathbf{e}^{\mathsf{T}}\overline{\lambda} - \mathbf{e}^{\mathsf{T}}\lambda\right).$$

We note that possibly, λ and $\overline{\lambda}$ have different lengths. An observation is that $\mathbf{e}^{\mathsf{T}}\overline{\lambda}$ for a best rank-r approximation is not smaller than the largest $\mathbf{e}^{\mathsf{T}}\lambda$ over all optimal solutions of (11) satisfying the flatness condition, since otherwise the optimality is violated.

By the optimality of the best rank-r approximation, we have

$$\frac{1}{2}\|M(\mathcal{A}) - \overline{B}\|^2 = \frac{1}{2}\|M(\mathcal{A})\|^2 - \frac{1}{2}\overline{\lambda}^{\mathsf{T}}\overline{C\lambda},$$

where $\overline{C} := \overline{A}^{\mathsf{T}}\overline{A} \circ \overline{A}^{\mathsf{T}}\overline{A} \circ \overline{A}^{\mathsf{T}}\overline{A}$ with $\overline{A} := [\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_r]$. Thus,

$$\overline{\lambda}^{\mathsf{T}}\overline{C\lambda} \leq \|\mathcal{A}\|^2.$$

By Lemma 5 and the hypothesis, we know that the smallest eigenvalue of the positive definite matrix $\overline{C}$ is lower bounded by $\tau(\mathcal{A})$ given by (28). Therefore,

$$\|\overline{\lambda}\|_1 \;=\; \mathbf{e}^{\mathsf{T}}\overline{\lambda} \leq \sqrt{r}\|\overline{\lambda}\| \leq \frac{\sqrt{r}}{\sqrt{\tau(\mathcal{A})}}\|\mathcal{A}\|. \tag{30}$$

On the other hand, we have

$$\rho(\mathcal{A})^2 < \overline{\lambda}^{\mathsf{T}}\overline{C\lambda} \leq \sum_{i=1}^{r}\sum_{j=1}^{r}\overline{\lambda}_i\overline{\lambda}_j|\overline{C}_{ij}| \leq \sum_{i=1}^{r}\sum_{j=1}^{r}\overline{\lambda}_i\overline{\lambda}_j = \|\overline{\lambda}\|_1^2,$$

where the strict inequality follows from the fact that a best rank $r \geq 2$ approximation must be strictly better than the best rank one approximation, and the second inequality from $|\overline{C}_{ij}| = |(\overline{A}^{\top}\overline{A})_{ij}|^3 = |\overline{\mathbf{x}}_i^{\top}\overline{\mathbf{x}}_j|^3 \leq 1$. Thus, $\|\overline{\lambda}\|_1 \geq \rho(\mathcal{A})$. Similarly, we also have $\lambda^{\top}C\lambda \leq \|\lambda\|_1^2$, where $C := A^{\mathsf{T}}A \circ A^{\mathsf{T}}A \circ A^{\mathsf{T}}A$ with $A := [\mathbf{x}_1, \ldots, \mathbf{x}_s]$.

Expanding the left hand side of the inequality (29), it becomes

$$\frac{1}{2}\|M(\mathcal{A})\|^2 - \sum_{i=1}^{s}\lambda_i\langle\mathcal{A}, \mathbf{x}_i^{\otimes 3}\rangle + \frac{1}{2}\lambda^{\mathsf{T}}C\lambda + \sigma\mathbf{e}^{\mathsf{T}}\lambda. \tag{31}$$

Note that each $\lambda_i > 0$, and thus by the optimality of λ (i.e., setting the derivative of the above expression with respect to λ to zero), we must have

$$\sum_{i=1}^{s}\lambda_i\langle\mathcal{A}, \mathbf{x}_i^{\otimes 3}\rangle - \sigma\mathbf{e}^{\mathsf{T}}\lambda = \lambda^{\mathsf{T}}C\lambda.$$

Consequently, we have

$$\frac{1}{2}\|M(\mathcal{A})-B\|^2+\sigma\mathbf{e}^{\mathsf{T}}\lambda=\frac{1}{2}\|M(\mathcal{A})\|^2-\frac{1}{2}\lambda^{\mathsf{T}}C\lambda.$$

Thus, from (29), we have

$$-\frac{1}{2}\lambda^{\mathsf{T}}C\lambda\le-\frac{1}{2}\overline{\lambda}^{\mathsf{T}}\overline{C\lambda}+\sigma\mathbf{e}^{\mathsf{T}}\overline{\lambda}.$$

This, together with the fact that $\lambda^{\top}C\lambda\le\|\lambda\|_1^2$, implies

$$\begin{aligned}\|\lambda\|_1^2&\ge\lambda^{\mathsf{T}}C\lambda\ge\overline{\lambda}^{\mathsf{T}}\overline{C\lambda}-2\sigma\mathbf{e}^{\mathsf{T}}\overline{\lambda}\\&\ge\tau(\mathcal{A})\|\overline{\lambda}\|^2-2\sigma\mathbf{e}^{\mathsf{T}}\overline{\lambda}\ge\frac{\tau(\mathcal{A})}{r}\|\overline{\lambda}\|_1^2-2\sigma\|\overline{\lambda}\|_1\\&=\frac{\tau(\mathcal{A})}{r}\Big(\|\overline{\lambda}\|_1^2-4\frac{r\sigma}{\tau(\mathcal{A})}\|\overline{\lambda}\|_1+\big(\frac{2r\sigma}{\tau(\mathcal{A})}\big)^2+2\frac{r\sigma}{\tau(\mathcal{A})}\|\overline{\lambda}\|_1-\big(2\frac{r\sigma}{\tau(\mathcal{A})}\big)^2\Big),\end{aligned}\tag{32}$$

where the third inequality follows from Lemma 5 and (28). This, together with (30) and $\|\overline{\lambda}\|_1\ge\rho(\mathcal{A})$, implies that when $\sigma\le\frac{\tau(\mathcal{A})\rho(\mathcal{A})}{2r}$, we have

$$\|\lambda\|_1\ge\sqrt{\frac{\tau(\mathcal{A})}{r}}\Big(\|\overline{\lambda}\|_1-2\frac{r}{\tau(\mathcal{A})}\sigma\Big).\tag{33}$$

It then follows from (29), (30), and (33) that

$$\|M(\mathcal{A})-B\|^2\le\|M(\mathcal{A})-\overline{B}\|^2+2\sigma\left(\Big(1-\sqrt{\frac{\tau(\mathcal{A})}{r}}\Big)\sqrt{\frac{r}{\tau(\mathcal{A})}}\|\mathcal{A}\|+2\sqrt{\frac{r}{\tau(\mathcal{A})}}\sigma\right).\tag{34}$$

This completes the proof. □

We can apply a similar refinement technique as Proposition 11 to improve the quality of B given in Proposition 12. Actually, by the optimality of λ in (31), we have that $\lambda=C^{\dagger}(\mathbf{u}-\sigma\mathbf{e})$ and hence

$$\|M(\mathcal{A})-B\|^2=\|\mathcal{A}\|^2-2\lambda^{\mathsf{T}}\mathbf{u}+\lambda^{\mathsf{T}}C\lambda=\|\mathcal{A}\|^2-\mathbf{u}^{\mathsf{T}}C^{\dagger}\mathbf{u}+\sigma^2\mathbf{e}^{\mathsf{T}}C^{\dagger}\mathbf{e},$$

where $\mathbf{u}:=(\langle\mathcal{A},\mathbf{x}_1^{\otimes3}\rangle,\ldots,\langle\mathcal{A},\mathbf{x}_s^{\otimes3}\rangle)^{\mathsf{T}}$. Given the vectors $\mathbf{x}_1,\ldots,\mathbf{x}_s$, we can optimize their coefficients to an approximant $\mathcal{B}':=\sum_{i=1}^s\mu_i\mathbf{x}_i^{\otimes3}$ such that

$$\|\mathcal{A}-\mathcal{B}'\|^2=\|\mathcal{A}\|^2-\mathbf{u}^{\mathsf{T}}C^{\dagger}\mathbf{u}.$$

This can reduce at least the amount of σ^2 from (34) since $\mathbf{e}^{\mathsf{T}}C^{\dagger}\mathbf{e}\ge1$. Note that when $r=1$, we can take $\tau(\mathcal{A})=\kappa(\mathcal{A})=1$, and we thus get a $4\sigma^2$-quasi-optimality

estimation from (34). While, even with the above refinement, we can only get $3\sigma^2$-quasi-optimality; but we know that a best rank one approximant can be recovered by Proposition 11. This follows largely by the estimations from (32) and (33), and it indicates that there are some room for improvement. Nevertheless, the next result shows that the estimation of σ^2 in Proposition 12 cannot be eliminated.

Proposition 13 *Let $k \geq 2$. If $\mathcal{A}$ is an orthogonally decomposable tensor with rank $s \leq n$, then for $\sigma \in [0, \lambda_r - \lambda_{r+1}]$[2] with $r \leq s$, $\mu := \sum_{i=1}^{r}(\lambda_i - \sigma)\delta_{\mathbf{x}_i}$ gives a global optimizer of problem (11) and a $r\sigma^2$-quasi-optimal rank-r approximation of $\mathcal{A}$.*

Proof Let $\mathcal{A} = \sum_{i=1}^{s} \lambda_i \mathbf{x}_i^{\otimes 3}$ be an orthogonal decomposition. Let

$$u(\mathbf{x}) := \sum_{i=1}^{r} \sigma(\mathbf{x}_i^{\mathsf{T}}\mathbf{x})^3.$$

Let $\mu := \sum_{i=1}^{r}(\lambda_i - \sigma)\delta_{\mathbf{x}_i}$ be the r-atomic measure and $\bar{\mathbf{y}}$ be the moment sequence defined by the measure μ. Let U be the corresponding matrix for the cubic polynomial $u(\mathbf{x})$. Since $\lambda_r - \sigma \geq \lambda_{r+1}$, we have that $\mathcal{P}(\bar{\mathbf{y}})$ gives a best rank-r approximation of the matrix $M(\mathcal{A}) - U$. Let

$$w(\mathbf{x}) := \frac{3\sigma}{2}\mathbf{x}^{\mathsf{T}}\mathbf{x}.$$

Let $\bar{Z}$ be the moment matrix defined by the following polynomial:

$$z(\mathbf{x}) := \sigma - u(\mathbf{x}) + w(\mathbf{x})(\|\mathbf{x}\|^2 - 1).$$

Obviously,

$$z(\mathbf{x}_i) = 0 \text{ for all } i = 1, \ldots, r,$$

so the complementarity between $\bar{Z}$ and $\mathcal{M}_k(\bar{\mathbf{y}})$ is fulfilled. In the following, by Theorem 8, we only need to check that the polynomial $w(\mathbf{x})$ satisfies the fact that the polynomial $z(\mathbf{x})$ is a sum of squares of polynomials.

Applying an orthogonal transformation if necessary, we can assume without loss of generality that $\mathbf{x}_i = \mathbf{e}_i$ (the i-th column vector of the identity matrix) for all $i = 1, \ldots, r$. Let the resulting polynomial be $\hat{z}(\mathbf{x})$. We then have

$$\frac{\hat{z}(\mathbf{x})}{\sigma} = 1 - \sum_{i=1}^{r} x_i^3 + \frac{3}{2}(\mathbf{x}^{\mathsf{T}}\mathbf{x})(\mathbf{x}^{\mathsf{T}}\mathbf{x} - 1).$$

Thus, it follows from [50, Section 7.3 (in Supplementary)] that the polynomial $\frac{\hat{z}(\mathbf{x})}{\sigma}$ and hence $z(\mathbf{x})$ is a sum of squares.

By Theorem 8, $\bar{\mathbf{y}}$ is a global minimizer of (11). The approximation error to the best rank-r approximation $\mathbf{y}^*$ generated by $\sum_{i=1}^{r} \lambda_i \delta_{\mathbf{x}_i}$ is given by

$$\|\mathcal{A} - \mathcal{B}\|^2 = \|\mathcal{A} - \mathcal{B}^*\|^2 + r\sigma^2.$$

[2] We let $\lambda_{s+1} = 0$ if needed.

The conclusion then follows. □

Of course, a refinement as in the preceding analysis will give the global optimal solution in the scenario of Proposition 13. But a generic tensor does not necessarily have an orthogonal decomposition [31], which implies that the linear term over σ in (34) is probably essential.

From the proofs of Theorem 10 and Proposition 13, we get the following result.

Proposition 14 *Let $k \geq 2$. If $\mathcal{A}$ is an orthogonally decomposable tensor with rank $s \leq n$, then for $\sigma \in [0, \lambda_r - \lambda_{r+1}]$ with $r \leq s$, strong duality holds for problems (11) and (12).*

As the tensor rank can be larger than the flattening matrix rank, it could happen that $\operatorname{rank}(\mathcal{M}_{k-1}(\mathbf{y})) = \operatorname{rank}(\mathcal{M}_k(\mathbf{y})) = r + 1$. Actually, by the analysis of Theorem 9, this would be the case if σ is small and if the approximation residual (24) is large. Thus, we include the following corollary to address this case.

Corollary 1 *Let $k \geq 2$, $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$ be nonzero and have rank greater than two, $r \geq 2$, $\overline{\mathcal{B}}$ be a best rank-r approximant of $\mathcal{A}$ satisfying Assumption 2, and $(B, \mathbf{y}, X)$ be an optimal solution of (11) with $\sigma \in (0, \frac{\tau(\mathcal{A})\rho(\mathcal{A})}{2r})$. Suppose that $\mathbf{y}$ satisfies $\operatorname{rank}(\mathcal{M}_{k-1}(\mathbf{y})) = \operatorname{rank}(\mathcal{M}_k(\mathbf{y})) \leq r + 1$. Then a candidate $\mathcal{B}'$ can be constructed from B and $\mathbf{y}$ such that it gives a α-quasi-optimal rank-r approximation of $\mathcal{A}$ with α given by*

$$\alpha := 2\sqrt{\frac{r}{\tau(\mathcal{A})}}\left(\left(1 - \sqrt{\frac{\tau(\mathcal{A})}{r}}\right)\|\mathcal{A}\| + 2\sigma\right)\sigma + \rho(\mathcal{A} - \mathcal{B}')^2.$$

Proof The proof is a combination of those of Theorem 9 and Proposition 12. We omit the tedious details. □

4.3 Optimality

In this section, we summarize the established results into certifications on best approximations and quasi-optimal approximations of a given tensor. The next result is for the basic relaxation, i.e., $k = 2$, higher order relaxations can be stated similarly.

Theorem 15 (Rank-r Approximation). *Suppose that $\sigma \geq 0$ and $k = 2$ are chosen in (11). If there exist a triplet $(\bar{U}, \bar{W}, \bar{Z})$ such that the feasibility (13) is satisfied and a vector $\bar{\mathbf{y}}$ such that all the optimality condition (15), $\operatorname{rank}(\mathcal{M}_1(\bar{\mathbf{y}})) = \operatorname{rank}(\mathcal{M}_2(\bar{\mathbf{y}})) \neq r + 1$, and*

$$\operatorname{rank}(\mathcal{P}(\bar{\mathbf{y}})) \leq r$$

are satisfied, then $\bar{\mathbf{y}}$ gives an optimal solution for (11) and

1. *if $\sigma = 0$, or $r = 1$ and $\sigma < \rho(\mathcal{A})$, then $\mathcal{P}(\bar{\mathbf{y}})$ gives a best rank-r approximation of $\mathcal{A}$;*

2. *if* $\sigma \in (0, \frac{\tau(\mathcal{A})\rho(\mathcal{A})}{2r})$, $r > 1$, $\mathrm{rank}(\mathcal{M}_1(\bar{\mathbf{y}})) \leq r$, *and Assumption 2 is satisfied, then* $\mathcal{P}(\bar{\mathbf{y}})$ *gives a* $2\sqrt{\frac{r}{\tau(\mathcal{A})}}\left(\left(1-\sqrt{\frac{\tau(\mathcal{A})}{r}}\right)\|\mathcal{A}\|+2\sigma\right)\sigma$*-quasi-optimal rank-r approximation of* $\mathcal{A}$.

Proof It follows from Theorem 8, Theorem 9, and Proposition 12. □

The case of the best rank one approximation is more clear.

Theorem 16 (Rank One Approximation). *Let* $k \geq 2$. *Suppose that* $\sigma < \rho(\mathcal{A})$ *is chosen in* (11) *and* $r = 1$. *If there exist a triplet* $(\bar{U}, \bar{W}, \bar{Z})$ *such that the feasibility* (13) *is satisfied and a vector* $\bar{\mathbf{y}}$ *such that the optimality condition* (15) *is satisfied, and* $\mathrm{rank}(\mathcal{P}(\bar{\mathbf{y}})) \leq 1$, *then* $\bar{\mathbf{y}}$ *gives a best rank one approximation of* $\mathcal{A}$.

Proof This follows from the fact that (11) is a relaxation of (2) and $\mathcal{B}$ is a tensor of rank at most one if and only if the corresponding matrix B has rank at most one [31]. Thus, the flatness condition in Theorem 15 is not needed in this case. The result follows from Proposition 11 and Theorem 9. □

Note that for the rank one case, if $\sigma > 0$ in Theorems 15 and 16, a simple refinement as in Proposition 11 to get a best approximant is necessary.

Actually, the vector $\bar{\mathbf{y}}$ in both Theorems 15 and 16 need not be computed for a candidate solution to problem (11). We can have such a moment sequence by other means or methods, and we can also check the optimality for it using these theorems. This is exactly Theorem 1 when a candidate tensor is available.

5 Numerical illustration

In this section, we present some numerical examples to illustrate the usefulness of the theoretical results presented so far.

The emphasis is put on certifying the global optimality for the best low rank tensor computed, which is achieved by solving the dual problem (12) and employing Theorem 8 to check the optimality. Note that the dual problem (12) is not easy to solve, due to the particular nonsmooth objective function (cf. (13)). The design of a highly efficient numerical algorithm for solving this problem will be addressed in another paper. We will apply existing methods for (12) in the current paper. For the sake of not lengthening the paper or taking us far afield, we do not include the full details, but just give a brief description of the implementation here. We will apply a proximal sGS-ADMM to solve the dual problem (12).

Note that there are some benefits for solving the dual problem (12) instead of the primal problem (11): (i) The dual problem (12) is convex while the primal problem (11) is nonconvex. For smaller n and r, solving the primal problem (11) works as well in our experiments. While, for larger n and r, it is hard to get a global optimal solution of (11) and consequently the theoretical results established for quantification cannot be verified. (ii) The dual problem has simpler constraint which is linear and has a separable structure, while the primal problem has a complicated rank constraint. Moreover, there are well-developed numerical methods for solving problems of the form as (12).

5.1 Algorithmic rationale

In the following, for simplicity, we omit the subscript k for the order of relaxation in (12). In this section, $k = 2$ is applied. For problem (12), we first rewrite it as

$$\begin{aligned} \min\ & \tfrac{1}{2}\|\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - U)\|^2 + \chi_{\mathrm{S}_+^{\zeta(n+1,2)}}(Z) \\ \text{s.t. } & \mathcal{M}^*(V) + \mathcal{P}^*(U) - \mathcal{L}^*(W) = \sigma\mathcal{M}^*(E_0), \\ & V - Z = 0, \end{aligned} \tag{35}$$

where V is an auxiliary variable.

Problem (35) is a linearly constrained convex matrix optimization problem with a nonsmooth objective function. The variables can be grouped into two sets $\{W, U\}$ and $\{Z, V\}$. Corresponding to each set, the objective function has a nonsmooth part. We apply the proximal symmetric Gauss-Siedel alternating direction method of multipliers (proximal sGS-ADMM) [34] to solve (35) based on the grouping of the above two sets of variables.

The augmented Lagrangian function of (35) is

$$\begin{aligned} L_\beta(U, V, W, Z; \mathbf{y}, X) := & \frac{1}{2}\|\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - U)\|^2 + \chi_{\mathrm{S}_+^{\zeta(n+1,2)}}(Z) \\ & + \langle \mathbf{y}, \mathcal{M}^*(V) + \mathcal{P}^*(U) - \mathcal{L}^*(W) \\ & - \sigma\mathcal{M}^*(E_0)\rangle + \langle X, V - Z\rangle \\ & + \frac{\beta}{2}\|\mathcal{M}^*(V) + \mathcal{P}^*(U) - \mathcal{L}^*(W) \\ & - \sigma\mathcal{M}^*(E_0)\|^2 + \frac{\beta}{2}\|V - Z\|^2, \end{aligned}$$

where $\beta > 0$ is the Lagrange penalty parameter. The main loop of the algorithm is described in Algorithm 1.

In the algorithmic description, $\eta > 0$ is a proximal parameter, τ is a steplength parameter, and $\mathcal{Q}$ is a positive semidefinite operator $\mathcal{Q} : \mathbb{R}^{n\times n^2} \to \mathbb{R}^{n\times n^2}$ defined as

$$\mathcal{Q}(U) := \gamma U - \beta\mathcal{P}\mathcal{P}^*(U) \text{ for all } U$$

with an appropriately chosen $\gamma > 0$. In Algorithm 1, there are closed formulae for the subproblems of W, V and Z respectively. However, solving the subproblem for U is not that straightforward. With the choice of the operator $\mathcal{Q}$, we are giving the problem

$$\min_U\ \frac{1}{2}\left\|\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - U)\right\|^2 + \frac{\gamma}{2}\left\|U - C^i\right\|^2 \tag{36}$$

with

$$C^i := \frac{1}{\gamma}\Big(\mathcal{Q}(U^i) - \beta\mathcal{P}(\mathcal{M}^*(V^i) - \mathcal{L}^*(W^{i+\frac{1}{2}}) - \sigma\mathcal{M}^*(E_0)) - \mathcal{P}(\mathbf{y}^i)\Big)$$

Algorithm 1 proximal sGS-ADMM

1: Step 1: Compute (U^{i+1}, W^{i+1}) via the following steps.

2: Substep 1: compute $W^{i+\frac{1}{2}}$ via

$$W^{i+\frac{1}{2}} \in \arg\min L_\beta(U^i, V^i, W, Z^i; \mathbf{y}^i, X^i) + \frac{\eta}{2}\|W - W^i\|^2.$$

3: Substep 2: compute U^{i+1} via

$$U^{i+1} \in \arg\min L_\beta(U, V^i, W^{i+\frac{1}{2}}, Z^i; \mathbf{y}^i, X^i) + \frac{1}{2}\|U - U^i\|_{\mathcal{Q}}^2.$$

4: Substep 3: compute W^{i+1} via

$$W^{i+1} \in \arg\min L_\beta(U^{i+1}, V^i, W, Z^i; \mathbf{y}^i, X^i) + \frac{\eta}{2}\|W - W^{i+\frac{1}{2}}\|^2.$$

5: Step 2: Compute (V^{i+1}, Z^{i+1}) via the following steps.

6: Substep 1: compute $Z^{i+\frac{1}{2}}$ via

$$Z^{i+\frac{1}{2}} \in \arg\min L_\beta(U^{i+1}, V^i, W^{i+1}, Z; \mathbf{y}^i, X^i).$$

7: Substep 2: compute V^{i+1} via

$$V^{i+1} \in \arg\min L_\beta(U^{i+1}, V, W^{i+1}, Z^{i+\frac{1}{2}}; \mathbf{y}^i, X^i).$$

8: Substep 3: compute Z^{i+1} via

$$Z^{i+1} \in \arg\min L_\beta(U^{i+1}, V^{i+1}, W^{i+1}, Z; \mathbf{y}^i, X^i).$$

9: Step 3: Update the multipliers via

$$\begin{aligned}\mathbf{y}^{i+1} &:= \mathbf{y}^i + \tau\beta(\mathcal{M}^*(V^{i+1}) + \mathcal{P}^*(U^{i+1}) - \mathcal{L}^*(W^{i+1}) - \sigma\mathcal{M}^*(E_0)),\\ X^{i+1} &:= X^i + \tau\beta(V^{i+1} - Z^{i+1}).\end{aligned}$$

10: Step 4: If a termination is not reached, go back to Step 1.

$$= U^i - \frac{\beta}{\gamma}\mathcal{P}\big(\mathcal{P}^*(U^i) + \mathcal{M}^*(V^i) - \mathcal{L}^*(W^{i+\frac{1}{2}}) - \sigma\mathcal{M}^*(E_0)\big) - \frac{1}{\gamma}\mathcal{P}(\mathbf{y}^i).$$

Thus,

$$U^{i+1} := M(\mathcal{A}) - \mathrm{prox}_{\Pi^2_{\mathrm{R}(r)}, \frac{1}{\gamma}}(M(\mathcal{A}) - C^i),$$

where $\mathrm{prox}_{\Pi^2_{\mathrm{R}(r)}, \frac{1}{\gamma}}$ is the proximal operator of the function $\Pi^2_{\mathrm{R}(r)}$, i.e., $\mathrm{prox}_{\Pi^2_{\mathrm{R}(r)}, \frac{1}{\gamma}}(Y)$ represents the optimizer of the following problem

$$\min_X \ \frac{1}{2}\left\|\Pi_{\mathrm{R}(r)}(X)\right\|^2 + \frac{\gamma}{2}\|X - Y\|^2.$$

In the following, we briefly describe this proximal operator's calculation. Suppose that $n \leq m$ and $Y \in \mathbb{R}^{n\times m}$. Let the SVD of Y be

$$Y = U\Sigma V^{\mathsf{T}} \text{ with } \Sigma = \operatorname{diag}\{y_1, \ldots, y_n\},$$

where $y_1 \geq \cdots \geq y_n$. Then, by the unitary invariant property of the objective function, the solution X must have the form

$$X = U \operatorname{diag}(\mathbf{x}) V^{\mathsf{T}}$$

for a nonnegative vector $\mathbf{x}$. Since the low rank projection is permutation invariant and $\mathbf{y}$ is ordered nonincreasingly, we must have that $\mathbf{x}$ is ordered nonincreasingly as well. The objective function value is then

$$\frac{1}{2}\sum_{i=1}^{r} x_i^2 + \frac{\gamma}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

If $\frac{\gamma}{1+\gamma} y_r \geq y_{r+1}$, then the optimal solution is given by

$$\mathbf{x}_{1:r} := \frac{\gamma}{1+\gamma}\mathbf{y}_{1:r} \text{ and } \mathbf{x}_{r+1:n} := \mathbf{y}_{r+1:n}.$$

Otherwise, suppose that for a pair (s, p) with $s < r$ and $p \geq r+1$, and some $\kappa > 0$, the optimizer is

$$\mathbf{x}_{1:s} := \frac{\gamma}{1+\gamma}\mathbf{y}_{1:s},\ \mathbf{x}_{s+1:p} := \kappa,\ \text{and } \mathbf{x}_{p+1:n} := \mathbf{y}_{p+1:n}.$$

By the optimality, we must have

$$\kappa = \frac{\gamma \sum_{i=s+1}^{p} y_i}{(r-s) + \gamma(p-s)}.$$

A necessary condition for the optimality is that

$$\frac{\gamma}{1+\gamma} y_s \geq \kappa \text{ or } s = 0,\ \text{and } \kappa \geq y_{p+1} \text{ or } p = n.$$

A method can be designed for finding such a pair (s, p), and the detail is omitted in this paper.

For the optimality criteria, by a direct calculation, the optimality condition of (35) is

$$\begin{gathered} 0 \preceq Z \perp \mathcal{M}(\mathbf{y}) \succeq 0, \\ \mathcal{P}(\mathbf{y}) \in \operatorname{conv}(\Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - U)), \\ \mathcal{L}^*(W) - \mathcal{P}^*(U) - \mathcal{M}^*(Z) + \sigma\mathcal{M}^*(E_0) = 0, \\ Z - V = 0,\ \text{and } \mathcal{L}(\mathbf{y}) = 0. \end{gathered} \tag{37}$$

If $\text{rank}(\mathcal{P}(\mathbf{y})) \leq r$, then by Lemma 8 the second condition is equivalent to

$$\mathcal{P}(\mathbf{y}) \in \Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - U).$$

In this case, by Łojasiewicz's inequality, this condition can be measured by

$$\left| \|M(\mathcal{A}) - U - \Pi_{\mathrm{R}(r)}(M(\mathcal{A}) - U)\| - \|M(\mathcal{A}) - U - \mathcal{P}(\mathbf{y})\| \right|.$$

5.2 Illustrative examples

All the tests were conducted on a Lenovo laptop with 128GB RAM and 2.8GHz E-2276M CPU running 64bit Windows operation system. All codes were written in MATLAB. The default parameters are chosen as $\tau = 1.25$, $\beta = 100$, $\eta = 10^{-5}$, $\sigma = 10^{-5}$, $\gamma = 10^3$, where τ is the steplength in sGS-ADMM, β is the penalty parameter for the augmented Lagrangian function of (35), η is a proximal parameter, and γ is the proximal parameter for the subproblem of U. The computed multiplier $\mathbf{y}$ is used to generate a feasible solution of problem (11). In the examples, the *duality gap* refers to the difference $\psi(B, X) - \phi(U)$ as defined respectively in (11) and (13); the *feasibility* refers to the maximum of the primal feasibility and the dual feasibility violations; the *psd residual* refers to the violation of the first condition in (37); and the *projection residual* refers to the violation of the second condition in (37). The algorithm is terminated whenever either the residual of the system (37) is smaller than 10^{-10} or the number of iterations is over 2×10^5.

We see that when all the duality gap, feasibility violation, psd residual and projection residual are small, and the rank of B is bounded by r, then both the primal problem (11) and the dual problem (12) are solved globally. If furthermore the flatness condition is satisfied or $\text{rank}(B) = 1$, then the original best rank-r approximation problem is solved globally with good quality (cf. Theorem 15, Theorem 16 and Corollary 1), and the *certification* is met. If furthermore the flatness condition is satisfied with rank $\neq r + 1$ or $\text{rank}(B) = 1$ (cf. Theorem 15, Theorem 16), then we say that *strong certification* is met.

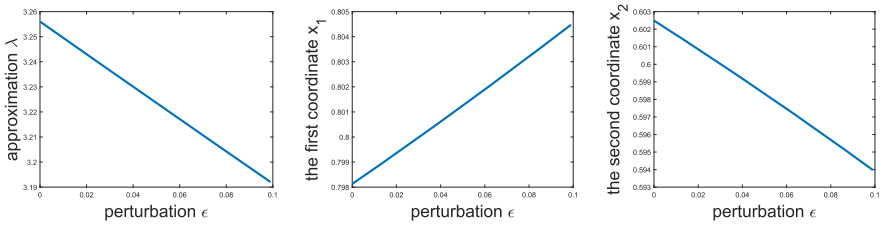
Example 1 This example is taken from De Lathauwer, De Moor and Vandewalle [11, Example 5]. It is a tensor in $\mathrm{S}^3(\mathbb{R}^2)$ with the independent elements being

$$a_{111} = 2,\ a_{112} = 1,\ a_{122} = 1,\ \text{and}\ a_{222} = 1.$$

The best rank one approximation computed is

$$\lambda = 3.2560 \text{ with } \mathbf{x} = (0.7981, 0.6025)^{\mathsf{T}},$$

which is exactly the one given in [11]. The global optimality is certified with the duality gap $= 2.9 \times 10^{-11}$, feasibility $= 6.6 \times 10^{-15}$, psd residual $= 1.7 \times 10^{-13}$, projection residual $= 3.8 \times 10^{-11}$, and the computed matrix B having rank one. The approximation residual is $\|\mathcal{A} - \mathcal{B}\| = 0.6310$.



(a) The norm λ (b) The first coordinate x_1 (c) The second coordinate x_2

Fig. 1 The computed best rank one tensors along the perturbations

Example 2 This example tests a set of perturbed versions of Example 1. They are tensors in $S^3(\mathbb{R}^2)$ with the independent elements being

$$a_{111} = 2,\ a_{112} = 1,\ a_{122} = 1 - \epsilon,\ \text{and}\ a_{222} = 1 + \epsilon,$$

where $\epsilon > 0$ is a perturbation in $[10^{-6}, 10^{-1}]$. We tested 100 instances, each taking an $\epsilon \in [10^{-6}, 10^{-1}]$, starting from 10^{-6} with an equal difference 10^{-3}. In each case, the method successfully computed the best rank one approximation, together with a global optimality certification as in Example 1. We do not present the similar but tedious data, while show the computed λ, and the coordinates of the vector $\mathbf{x}$ in Fig. 1, from which we can see the evolution of the optimal solutions along the perturbations.

Example 3 This tensor is taken from Qi [46, Example 2] as well as Nie and Wang [44, Example 3.3]. This is a tensor in $S^3(\mathbb{R}^3)$ with the independent elements being

$$\begin{aligned} &a_{111} = 0.0517,\ a_{112} = 0.3579,\ a_{113} = 0.5298,\ a_{122} = 0.7544,\ a_{123} = 0.2156,\\ &a_{133} = 0.3612,\ a_{222} = 0.3943,\ a_{223} = 0.0146,\ a_{233} = 0.6718,\ a_{333} = 0.9723. \end{aligned}$$

The best rank one approximation computed is

$$\lambda = 2.1110 \text{ with } \mathbf{x} = (0.5204, 0.5113, 0.6839)^{\mathsf{T}},$$

which is certified with the duality gap $= 5.1 \times 10^{-11}$, feasibility $= 7.8 \times 10^{-15}$, psd residual $= 5.9 \times 10^{-13}$, projection residual $= 3.8 \times 10^{-11}$, and the computed matrix B having rank one. The result agrees with that in [44, Example 3.3]. The approximation residual is 1.2672.

Example 4 This is a tensor in $S^3(\mathbb{R}^3)$ with the independent elements being

$$\begin{aligned} &a_{111} = 0.7239,\ a_{112} = 0.1505,\ a_{113} = 0.0199,\ a_{122} = 0.0266,\ a_{123} = 0.1232,\\ &a_{133} = 0.5176,\ a_{222} = 0.0835,\ a_{223} = 0.0833,\ a_{233} = 0.0801,\ a_{333} = -0.1353. \end{aligned}$$

This tensor has rank three and has a rank decomposition given by

$$0.8768 \times \begin{bmatrix} 0.7015 \\ -0.0770 \\ -0.7132 \end{bmatrix}^{\otimes 3} + 0.7093 \times \begin{bmatrix} 0.8494 \\ 0.3156 \\ 0.6092 \end{bmatrix}^{\otimes 3} + 0.6065 \times \begin{bmatrix} -0.2804 \\ 0.4666 \\ 0.3328 \end{bmatrix}^{\otimes 3}.$$

The computed best rank two approximation tensor $\mathcal{B}$ has independent elements

$$a_{111} = 0.7173,\ a_{112} = 0.1546,\ a_{113} = 0.0178,\ a_{122} = 0.0609,\ a_{123} = 0.1235,$$
$$a_{133} = 0.5202,\ a_{222} = 0.0230,\ a_{223} = 0.0442,\ a_{233} = 0.0797,\ a_{333} = -0.1280.$$

It is a rank two tensor, and has a rank decomposition given by

$$0.8379 \times \begin{bmatrix} 0.7066 \\ -0.0139 \\ -0.7039 \end{bmatrix}^{\otimes 3} + 0.9140 \times \begin{bmatrix} 0.7728 \\ 0.2923 \\ 0.5605 \end{bmatrix}^{\otimes 3}.$$

The approximation residual is 0.1092. The duality gap is 4.2×10^{-12} with the dual objective function value $= 0.0060$. The feasibility $= 4.5 \times 10^{-15}$, psd residual $= 2.4 \times 10^{-14}$, and projection residual $= 3.0 \times 10^{-11}$. The computed moment vector $\mathbf{y}$ is

$$\begin{aligned}(&1.7519, 1.2984, 0.2572, -0.0660, 0.9641, 0.1994, -0.0123, 0.0788, 0.1606,\\ &0.7090, 0.7173, 0.1546, 0.0178, 0.0609, 0.1235, 0.5202, 0.0230, 0.0442,\\ &0.0797, -0.1280, 0.5347, 0.1198, 0.0331, 0.0470, 0.0950, 0.3824, 0.0177,\\ &0.0342, 0.0619, -0.0795, 0.0067, 0.0130, 0.0251, 0.0527, 0.3014)^{\mathsf{T}},\end{aligned}$$

from which we can see that the (numerically) nonzero eigenvalues of the first moment matrix $\mathcal{M}_1(\mathbf{y})$ are 0.7486, 2.7552, and those of the second moment matrix $\mathcal{M}_2(\mathbf{y})$ are 1.1870, 3.6142. Therefore, the flatness condition is satisfied and hence the quantified optimality is certified.

Example 5 This is a tensor in $\mathrm{S}^3(\mathbb{R}^4)$ with the independent elements in lexicographic order as follows

$$\begin{aligned}&0.4287, -0.1614, -0.0696, -0.2829, 0.0404, -0.0544, 0.0888,\\ &-0.0715, 0.0159, 0.2633, -0.0979, -0.0933, -0.0524, -0.1570,\\ &-0.0081, -0.0644, -0.2576, -0.0222, -0.0299, -0.3253.\end{aligned}$$

This tensor has rank four and has a rank decomposition given by

$$0.7868 \times \begin{bmatrix} 0.5581 \\ -0.0331 \\ -0.1593 \\ -0.1216 \end{bmatrix}^{\otimes 3} + 0.3361 \times \begin{bmatrix} 0.5549 \\ -0.0319 \\ -0.1092 \\ -0.9176 \end{bmatrix}^{\otimes 3} + 0.6221 \times \begin{bmatrix} 0.7320 \\ -0.3972 \\ 0.0278 \\ -0.4689 \end{bmatrix}^{\otimes 3} + 0.4636 \times \begin{bmatrix} -0.2744 \\ -0.5026 \\ -0.8182 \\ -0.0511 \end{bmatrix}^{\otimes 3}.$$

The computed best rank three approximation tensor $\mathcal{B}$ has independent elements in lexicographic order as

$$\begin{aligned}&0.4128, -0.1766, -0.0546, -0.2891, 0.0324, -0.0443, 0.0873,\\ &-0.0829, 0.0190, 0.2617, -0.0856, -0.0989, -0.0405, -0.1565,\end{aligned}$$

$$-0.0152, -0.0605, -0.2542, -0.0188, -0.0318, -0.3250.$$

It is a rank three tensor, and has a rank decomposition given as

$$0.6365 \times \begin{bmatrix} 0.8303 \\ -0.3470 \\ -0.0381 \\ -0.4343 \end{bmatrix}^{\otimes 3} + 0.4372 \times \begin{bmatrix} 0.5097 \\ -0.0577 \\ -0.0819 \\ -0.8552 \end{bmatrix}^{\otimes 3} + 0.4635 \times \begin{bmatrix} -0.2746 \\ -0.5027 \\ -0.8182 \\ -0.0512 \end{bmatrix}^{\otimes 3}.$$

The approximation residual is 0.0639. The duality gap is 1.4×10^{-12} with the dual objective function value $= 0.0021$. The feasibility $= 2.1 \times 10^{-15}$, psd residual $= 3.2 \times 10^{-14}$, and projection residual $= 1.8 \times 10^{-11}$. The computed $\mathbf{y} \in \mathbb{R}^{70}$, from which we get the (numerically) nonzero eigenvalues of the first moment matrix $\mathcal{M}_1(\mathbf{y})$ are 0.0948, 0.5875, 2.3921, and those of the second moment matrix $\mathcal{M}_2(\mathbf{y})$ are 0.2449, 0.9786, 3.0383. Therefore, the flatness condition is satisfied and hence the quantified optimality is certified.

Example 6 This is a tensor $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^2)$ with the independent elements being

$$a_{111} = 0.5662,\ a_{112} = -0.0971,\ a_{122} = 0.0713,\ \text{and}\ a_{222} = 0.2664.$$

This tensor is an orthogonally decomposable tensor with rank two. The computed best rank two approximation is given by

$$\mathcal{B} = 0.5950 \times \begin{bmatrix} 0.9826 \\ -0.1859 \end{bmatrix}^{\otimes 3} + 0.2848 \times \begin{bmatrix} 0.1859 \\ 0.9826 \end{bmatrix}^{\otimes 3}$$

with the approximation residual being $\|\mathcal{A} - \mathcal{B}\| = 1.4 \times 10^{-5}$. We see that $\mathcal{B}$ is an orthogonally decomposable tensor. The duality gap $= 3.9 \times 10^{-13}$, feasibility $= 2.7 \times 10^{-13}$, psd residual $= 2.2 \times 10^{-12}$, and projection residual $= 5.3 \times 10^{-11}$. The approximation quality is consistent with the theoretical bound given in Proposition 13.

We also tested its best rank one approximation. The best rank one approximation computed is

$$\mathcal{B} = 0.5949 \times \mathbf{x}^{\otimes 3} \text{ with } \mathbf{x} = (0.9826, -0.1859)^{\mathsf{T}}$$

with the approximation residual being $\|\mathcal{A} - \mathcal{B}\| = 0.2848$. The duality gap $= 1.4 \times 10^{-13}$, feasibility $= 5.6 \times 10^{-15}$, psd residual $= 2.8 \times 10^{-11}$, and projection residual $= 5.1 \times 10^{-13}$. We see that the approximation quality is very good.

Since the given tensor is orthogonally decomposable, we know all the local minimizers [22]. By a DCA method for the primal problem (11), a local minimizer is found as

$$\mathcal{B} = 0.2848 \times \mathbf{x}^{\otimes 3} \text{ with } \mathbf{x} = (0.1861, 0.9825)^{\mathsf{T}}$$

with the approximation residual being $\|\mathcal{A} - \mathcal{B}\| = 0.5949$. The duality gap $= 0.1429$, feasibility $= 5.2 \times 10^{-9}$, psd residual $= 2.0 \times 10^{-9}$, and projection residual $= 0.3209$. Thus, it cannot be certified as a global optimizer by the theory established in this paper, which agrees with the observed fact.

Table 1 Performance of the perturbed orthogonally decomposable tensors

ϵ \ Paras	$\|B\|$	Num	Max-gap	Mean-gap	Max-res	Mean-res
10^{-1}	0.66623	100	1.5417×10^{-11}	1.5378×10^{-11}	0.47144	0.47144
10^{-2}	0.60013	100	2.2139×10^{-12}	2.2082×10^{-12}	0.30137	0.30137
10^{-3}	0.59543	100	1.7190×10^{-13}	1.6918×10^{-13}	0.28645	0.28645
10^{-4}	0.59498	100	1.0934×10^{-12}	8.6368×10^{-13}	0.28501	0.28501
10^{-5}	0.59493	100	1.8477×10^{-13}	1.1607×10^{-13}	0.28486	0.28486
10^{-6}	0.59493	100	1.4534×10^{-13}	1.2692×10^{-13}	0.28485	0.28485

Example 7 Tensors in this example are perturbed variations of the tensor in Example 6. We tested six variations, by adding to each component of the tensor in Example 6 with $\epsilon = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$ and 10^{-6} respectively. For each case, the algorithm is excuted 100 times with random initialization. The results are summarized in Table 1. In this table, "$\|B\|$" represents the norm of the best rank one approximation tensor found by the algorithm, "Num" represents the number of strong certification, "max-gap" and "mean-gap" represent the maximum duality gap and the mean duality gap between the primal and the dual problems respectively, and "max-res" and "mean-res" represent the maximum approximation residual and the mean approximation residual respectively. We see from Table 1 that the computation is very stable, and in all cases global optimal solutions are found. From the last column of this table, we see that the approximation quality is consistent with our theory.

Example 8 Tensors in this example are randomly generated with each element in [0, 1]. The best rank one approximation is computed, i.e., $r = 1$. Examples with different dimensions n are simulated. For each $n \in \{2, \ldots, 10\}$, 100 randomly generated instances are tested. The results are summarized in Table 2. In this table, "Num" refers to the number of strong certification; "max-gap" and "mean-gap" are as Table 1, for the certified simulations; "mean-psd", "mean-feas" and "mean-proj" are the mean residuals for the psd residual, the feasibility, and the projection residual respectively. We see from Table 2 that the performance is quite promising.

Example 9 Tensors in this example are in the following form

$$\mathcal{A} = \sum_{i=1}^{r+2} \lambda_i \mathbf{x}_i^{\otimes 3},$$

where λ_i and each component of $\mathbf{x}_i \in \mathbb{R}^n$ are randomly generated from [0, 1]. The best rank-r approximation is computed. Examples with different pairs (n, r) of dimension n and approximation rank r are simulated. For each pair, 100 randomly generated instances are tested. The computational results are summarized in Table 3. In this table, "c" and "sc" refer to "certification" and "strong certification" respectively. Whenever

Table 2 Performance of randomly generated tensors for $r = 1$

n \ Paras	Num	Max-gap	Mean-gap	Mean-psd	Mean-feas	Mean-proj
2	98	5.0×10^{-11}	1.5×10^{-11}	2.4×10^{-14}	2.5×10^{-15}	2.8×10^{-11}
3	100	7.3×10^{-11}	3.8×10^{-11}	3.7×10^{-13}	5.8×10^{-15}	2.9×10^{-11}
4	100	9.3×10^{-11}	5.1×10^{-11}	5.3×10^{-12}	1.0×10^{-14}	2.3×10^{-11}
5	100	8.9×10^{-11}	3.2×10^{-11}	2.2×10^{-11}	1.4×10^{-14}	1.0×10^{-11}
6	100	6.9×10^{-11}	2.4×10^{-11}	2.3×10^{-11}	1.9×10^{-14}	5.6×10^{-12}
7	100	7.3×10^{-11}	1.7×10^{-11}	1.8×10^{-11}	2.4×10^{-14}	3.2×10^{-12}
8	100	4.3×10^{-11}	1.3×10^{-11}	1.3×10^{-11}	3.0×10^{-14}	2.0×10^{-12}
9	100	3.9×10^{-11}	9.5×10^{-12}	1.0×10^{-11}	3.9×10^{-14}	1.2×10^{-12}
10	100	3.5×10^{-11}	8.5×10^{-12}	9.5×10^{-12}	4.4×10^{-14}	9.3×10^{-13}

it appears "70(92)", the number in the bracket means that there are extra 22 simulations for which no duality gap between the primal and the dual problem is observed but the flatness condition fails, thus we cannot get a certification. The other parameters are as those in Table 2, for the certified cases. The performance is also promising. We remark that for the case $(n, r) = (6, 5)$, standard Lasserre's relaxation to problem (4) gives SDP with matrix size $\zeta(38, 4) = 73815$ and number of equations around $\zeta(38, 4)^2/2$. From this perspective, essentially, it is a very hard problem.

Example 10 Tensors in this example are principal component tensors with perturbations, i.e., $\mathcal{A} \in \mathrm{S}^3(\mathbb{R}^n)$ in the form

$$\mathcal{A} = \sum_{i=1}^{r} \lambda_i \mathbf{x}_i^{\otimes 3} + \epsilon\mathcal{E},$$

where λ_i and $\mathbf{x}_i$'s are as Example 9, $\mathcal{E} \in \mathrm{S}^3(\mathbb{R}^n)$ is the tensor of all ones, and $\epsilon \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. For each pair of (n, r), 100 randomly generated instances are tested for each ϵ. The computational results are summarized in Table 4. The parameters are the same as Table 3. Data in bold are extra experiments under higher accuracy (tolerance lower than 10^{-12} or number of iterations upto 3×10^5). The performance is similar as Example 9. Higher accuracy promotes performance. Note that when $\epsilon = 0.1$, the principal components are merged by the perturbation.

6 Conclusions

In this paper, we presented a method for computing the best low rank approximation for a given third order symmetric tensor. It is shown that this method can certify the

Table 3 Performance of randomly generated tensors for $r > 1$

Paras / (n, r)	c/sc	Max-gap	Mean-gap	Mean-psd	Mean-feas	Mean-proj
(3, 2)	93/93	1.1×10^{-11}	1.4×10^{-12}	1.1×10^{-12}	3.7×10^{-15}	2.7×10^{-11}
(4, 2)	95/95	1.5×10^{-11}	3.4×10^{-12}	6.0×10^{-13}	5.4×10^{-15}	2.4×10^{-11}
(4, 3)	70(92)/19	8.3×10^{-12}	1.5×10^{-12}	5.2×10^{-12}	2.2×10^{-13}	2.6×10^{-11}
(5, 2)	87/87	3.4×10^{-11}	5.6×10^{-12}	1.5×10^{-12}	7.7×10^{-15}	2.3×10^{-11}
(5, 3)	66(85)/11	1.4×10^{-11}	3.1×10^{-12}	5.3×10^{-12}	2.3×10^{-13}	2.5×10^{-11}
(5, 4)	60(91)/7	5.7×10^{-12}	9.9×10^{-13}	3.9×10^{-11}	5.9×10^{-14}	2.4×10^{-11}
(6, 2)	88/88	3.8×10^{-11}	8.3×10^{-12}	3.1×10^{-12}	9.8×10^{-15}	2.1×10^{-11}
(6, 3)	60(79)/10	1.8×10^{-11}	4.6×10^{-12}	4.0×10^{-11}	3.3×10^{-13}	1.9×10^{-11}
(6, 4)	44(85)/5	3.0×10^{-11}	2.4×10^{-12}	5.8×10^{-11}	5.8×10^{-11}	1.6×10^{-11}
(6, 5)	19(73)/2	6.7×10^{-11}	7.5×10^{-12}	1.8×10^{-9}	4.9×10^{-11}	2.0×10^{-10}

Table 4 Performance of randomly generated principal component tensors

ϵ / Type / (n, r)	10^{-1} c/sc	10^{-2} c/sc	10^{-3} c/sc	10^{-4} c/sc	10^{-5} c/sc	10^{-6} c/sc
(3, 2)	93/39	87/72	95/92	99/99	100/100	100/100
(4, 2)	95/60	86/ 80	100/ 100	99/99	100/100	100/100
(4, 3)	54(91)/3	55(80) /14	86(89)/47	98(99)/91	99(100)/99	100/100
(5, 2)	94/66	79/73	99/99	100/100	100/100	100/100
(5, 3)	50(86)/3	49(70)/11	96(98)/55	99/96	99(100)/99	100/100
(5, 4)	10(90)/1	14(78)/2	65(91)/30	98/84	100/100	100/100
	32(93)/11	**62(80)/18**				
(6, 2)	96/72	77/72	98/98	100/100	100/100	100/100
(6, 3)	63(82)/6	53(64)/12	92(93)/72	100/100	100/100	100/100
(6, 4)	13(90)/3	15(67)/1	73(84)/33	100/94	100/100	100/100
	39(82)/13	**63(77)/21**				
(6, 5)	5(87)/0	6(77)/2	58(88)/17	95(96)/82	99(100)/99	99(100)/98
	34(90)/8	**44(82)/11**				

global optimality or quantified quasi-optimality under mild assumptions by employing techniques from polynomial optimization, matrix optimization, duality theory, and nonsmooth analysis. The applicability of the theory is verified by several numerical examples.

The emphasis of this paper is on the global optimality and quantified quasi-optimality certification of the best low rank approximation. Numerical illustration is presented for the validation of the theory as well. However, more carefully and wisely designed numerical methods should be investigated in our future research for solving the hard optimization problems involved in the theory. In particular, the method employed in this paper for the problem (35) is a first order method, which typically has a slow convergence and it is difficult to get a high accuracy solution for large scale problems. In order to facilitate the global optimality certification, high accuracy solution for the dual problem (12) is necessary (cf. Table 4). While problem (12) has a complicated nonsmooth convex objective function, it is a challenging problem to be solved, especially at degenerate solutions. On the other hand, note that we used the approximate solution (the multipler for (35)) for the dual problem of (35) to generate a candidate for the solution of (11). The polynomial optimization ingredients in our reformulation (11) require a high accuracy solution; otherwise, the flatness condition is impossible to be satisfied. Hence, methods and theory for solving (35) and its dual problem targeted with high accuracy and fast convergent properties should be developed. In particular, properties of the conjugate function and the proximal mapping of the squared low rank projection function involved in (36) should be investigated.

Nevertheless, the numerical examples in Sect. 5 as well as the theoretical results in Sect. 4 on the global optimality certification convinced us that this approach is quite promising.

Appendix A Basics on moments

Moment matrices are useful tools in the study of polynomial optimization, we refer to [32, 33, 38–40, 43] and references therein for basic notions and advances of polynomial optimization. Let

$$\mathbf{x}^{\circ s} := (1, x_1, \ldots, x_n, x_1^2, \ldots, x_n^s)^{\mathsf{T}} \tag{A1}$$

be the vector of monomials up to degree s in the n variables $x_1, \ldots, x_n$ ordered lexicographically. The dimension of $\mathbf{x}^{\circ s}$ is $\binom{n+s}{n}$. Let

$$\mathbf{x}^{[s]} = (x_1^s, x_1^{s-1}x_2, \ldots, x_n^s)^{\mathsf{T}}$$

be the sub-vector of $\mathbf{x}^{\circ s}$ corresponding to the monomials of degree exactly s. The dimension of $\mathbf{x}^{[s]}$ is

$$\zeta(n, s) := \binom{n+s-1}{n-1}.$$

Define the extended monomial basis of order 2 as

$$\mathbf{x}^{\otimes 2} := (x_1^2, x_1x_2, \ldots, x_1x_n, x_2x_1, x_2^2, \ldots, x_2x_n, \ldots, x_nx_1, \ldots, x_n^2)^{\mathsf{T}}, \tag{A2}$$

and generalize to $\mathbf{x}^{\otimes s}$ for $s \geq 3$ in a straightforward way. Note that $\mathbf{x}^{\otimes s}$ is used to refer to the symmetric rank one tensor of order s generated by $\mathbf{x}$ as well. We hope that this abuse of notation will not bring confusion, since (A2) is consistent with the classical meaning and it is actually a vectorization of the rank one tensor $\mathbf{x}^{\otimes s}$ in the lexicographic order. The exact meaning would be clear from the context.

There is a natural one to one correspondence between a monomial $\mathbf{x}^\alpha$ and a vector in $\mathbb{N}^n$. The relation is indicated directly by the exponent vector α of the given monomial. Let

$$\mathbb{N}^n_{\leq s} := \{\alpha \in \mathbb{N}^n : |\alpha| := \alpha_1 + \cdots + \alpha_n \leq s\},$$
$$\mathbb{N}^n_{= s} := \{\alpha \in \mathbb{N}^n : |\alpha| := \alpha_1 + \cdots + \alpha_n = s\}.$$

Note that for each given integer $s \geq 0$, there exists a set of mutually orthogonal symmetric matrices $A_\alpha \in \mathrm{S}^2(\{0, 1\}^{\zeta(n+1,s)})$ such that

$$\mathbf{x}^{\circ s}(\mathbf{x}^{\circ s})^{\mathsf{T}} = \sum_{\alpha \in \mathbb{N}^n_{\leq 2s}} \mathbf{x}^\alpha A_\alpha.$$

In the classical analysis of polynomial optimization, a moment matrix of order s is a matrix $M \in \mathrm{S}^2(\mathbb{R}^{\zeta(n+1,s)})$ in the form

$$M := \sum_{\alpha \in \mathbb{N}^n_{\leq 2s}} y_\alpha A_\alpha \tag{A3}$$

for a vector $\mathbf{y} \in \mathbb{R}^{\zeta(n+1,2s)}$ which is indexed by the vector of exponents of monomials in $\mathbf{x}^{\circ 2s}$. The matrix M in (A3) is denoted as $\mathcal{M}_s(\mathbf{y})$, and it is known as the *s-th order moment matrix generated by* $\mathbf{y}$.

Let $s \geq 2$ be a given integer. Suppose that vectors in $\mathbb{R}^{\zeta(n+1,2s)}$ are indexed by $\mathbb{N}^n_{\leq 2s}$ as above. Define an operator $\mathcal{P} : \mathbb{R}^{\zeta(n+1,2s)} \to \mathbb{R}^{n \times n^2}$ as follows: given a vector $\mathbf{y} \in \mathbb{R}^{\zeta(n+1,2s)}$, $\mathcal{P}(\mathbf{y})$ is defined component-wisely as

$$(\mathcal{P}(\mathbf{y}))_{i,(j-1)*n+k} := y_{\mathbf{e}_i+\mathbf{e}_j+\mathbf{e}_k} \text{ for all } i, j, k \in \{1, \ldots, n\}. \tag{A4}$$

To certain extent, the operator $\mathcal{P}$ is independent of s. Thus, for simplicity, we omit this nominal dependence in the notation.

Lemma 6 *Let positive integer s be given and $\mathbf{y} \in \mathbb{R}^{\zeta(n+1,2s)}$. Then there exists a nonsingular matrix $P \in \mathbb{R}^{n^2 \times n^2}$ such that*

$$\mathcal{P}(\mathbf{y})P = \begin{bmatrix} M & 0 \end{bmatrix},$$

where M is the submatrix of $\mathcal{M}_s(\mathbf{y})$ corresponding to the $\mathbb{N}^n_{=1} \times \mathbb{N}^n_{=2}$ block. Therefore, $\operatorname{rank}(\mathcal{P}(\mathbf{y})) \leq \operatorname{rank}(\mathcal{M}_s(\mathbf{y}))$.

Proof By (A4), the matrix $\mathcal{P}(\mathbf{y})$ is almost the target $\mathbb{N}^n_{=1} \times \mathbb{N}^n_{=2}$ block of $\mathcal{M}_s(\mathbf{y})$, but with some repeated columns. These columns can be eliminated, and the result follows then. □

Let $\mathbf{y} \in \mathbb{R}^{\mathbb{N}^n}$ be a moment sequence. The (infinite) *moment matrix* $M(\mathbf{y})$ is defined element-wisely as

$$(M(\mathbf{y}))_{\alpha,\beta} := y_{\alpha+\beta}.$$

We see that the moment matrix of order s defined by (A3) is actually the leading $|\mathbb{N}^n_{\leq s}| \times |\mathbb{N}^n_{\leq s}|$ principal sub-matrix of the moment matrix $M(\mathbf{y})$. Likewise, the *moment tensor* $\mathcal{T}(\mathbf{y})$ is defined element-wisely as

$$\big(\mathcal{T}(\mathbf{y})\big)_{\alpha,\beta,\gamma} := y_{\alpha+\beta+\gamma}.$$

The *moment tensor of order* p, q, r, denoted as $\mathcal{T}_{p,q,r}(\mathbf{y})$, is then defined as the leading $|\mathbb{N}^n_{\leq p}| \times |\mathbb{N}^n_{\leq q}| \times |\mathbb{N}^n_{\leq r}|$ principal sub-tensor of the moment tensor $\mathcal{T}(\mathbf{y})$. Given a polynomial $g(\mathbf{x}) \in \mathbb{R}[\mathbf{x}]_r$, the *localizing matrix* $L^k_g(\mathbf{y})$ of order k is given by

$$\mathbf{p}^{\mathsf{T}} L^k_g(\mathbf{y})\mathbf{p} = \langle \mathbf{y}, p^2 g \rangle \text{ for all } p(\mathbf{x}) \in \mathbb{R}[\mathbf{x}]_k.$$

Proposition 17 (Localizing Matrix via Moment Tensor). *For any given polynomial $g(\mathbf{x}) \in \mathbb{R}[\mathbf{x}]_r$, it always holds that*

$$L^k_g(\mathbf{y}) = \langle \mathcal{T}_{k,k,r}(\mathbf{y}), \mathbf{g} \rangle_{3:1} \in \mathbb{R}^{|\mathbb{N}^n_{\leq k}| \times |\mathbb{N}^n_{\leq k}|},$$

where $\big(\langle \mathcal{T}_{k,k,r}(\mathbf{y}), \mathbf{g} \rangle_{3:1}\big)_{\alpha,\beta} := \sum_{\gamma \in \mathbb{N}^n_{\leq r}} \big(\mathcal{T}_{k,k,r}(\mathbf{y})\big)_{\alpha,\beta,\gamma} g_\gamma$ *for all* α, β.

Proof It follows that

$$\langle \mathbf{y}, p^2\, g\rangle = \sum_{\alpha,\beta,\gamma} y_{\alpha+\beta+\gamma}\, p_\alpha p_\beta g_\gamma = \langle \mathcal{T}_{k,k,r}(\mathbf{y}), \mathbf{p}\otimes\mathbf{p}\otimes\mathbf{g}\rangle.$$

The result thus follows. □

In the following, we review basic facts about flatness of truncated moment sequence over the unit sphere $\mathbb{S}^{n-1}$ (abbreviated as *utms*). For $k \geq 2$, a utms $\mathbf{y} \in \mathbb{R}^{\zeta(n+1,2k)}$ is *flat* if (cf. [38])

$$\mathcal{M}_k(\mathbf{y}) \succeq 0,\ \mathcal{L}_k(\mathbf{y}) = 0,\ \text{and}\ \operatorname{rank}(\mathcal{M}_k(\mathbf{y})) = \operatorname{rank}(\mathcal{M}_{k-1}(\mathbf{y})), \tag{A5}$$

where $\mathcal{L}_k(\mathbf{y}) := L^{k-1}_{1-\mathbf{x}^{\mathsf{T}}\mathbf{x}}(\mathbf{y})$ is the $(k-1)$-th localizing matrix of the polynomial $1-\mathbf{x}^{\mathsf{T}}\mathbf{x}$.

To be more precise, the condition (A5) is called *the k-th flatness condition* for the utms. If $\mathbf{y}$ satisfies the k-th flatness condition, then $\mathbf{y}$ can be represented as a unique measure which is $\operatorname{rank}(\mathcal{M}_k(\mathbf{y}))$-atomic [10, 38]. We will call the cardinality of the support of this unique measure the *rank of the utms*, denoted as $\operatorname{rank}(\mathbf{y})$. Thus, in this case, $\operatorname{rank}(\mathbf{y}) = \operatorname{rank}(\mathcal{M}_k(\mathbf{y}))$. If $\mathbf{y}$ does not satisfy the k-th flatness condition but some extension $\mathbf{z}$ of $\mathbf{y}$ satisfies the s-th flatness condition with $s > k$, then $\mathbf{y}$ can also be represented as a unique measure which is $\operatorname{rank}(\mathcal{M}_s(\mathbf{z}))$-atomic, and $\operatorname{rank}(\mathbf{y}) := \operatorname{rank}(\mathbf{z})$. Since $\mathcal{M}_k(\mathbf{y})$ is a principal sub-matrix of $\mathcal{M}_s(\mathbf{z})$, it may happen that $\operatorname{rank}(\mathbf{y}) = \operatorname{rank}(\mathbf{z}) = \operatorname{rank}(\mathcal{M}_s(\mathbf{z})) > \operatorname{rank}(\mathcal{M}_k(\mathbf{y}))$.

Appendix B Nonsmooth analysis of matrix low rank projection

Let positive integers $m \leq n$. Given a matrix $X \in \mathbb{R}^{m\times n}$ and a positive integer $r \leq m$, we consider the following problem on projection of X onto the set $\mathrm{R}(r)$ of matrices of rank at most r in the ambient space $\mathbb{R}^{m\times n}$, i.e.,

$$\begin{array}{l} \min\ \frac{1}{2}\|Y - X\|^2 \\ \text{s.t. } \operatorname{rank}(Y) \leq r, \\ \quad Y \in \mathbb{R}^{m\times n}. \end{array} \tag{B6}$$

It is well-known that an optimizer of (B6) can be computed via *singular value decomposition* by Eckart–Young–Mirsky's theorem [15]. Actually, let

$$X = P\Sigma(X)Q^{\mathsf{T}}$$

be the singular value decomposition of X with an orthogonal matrix $P \in \mathbb{R}^{m\times m}$ and an orthonormal $Q \in \mathbb{R}^{n\times m}$, and a diagonal matrix $\Sigma = \operatorname{diag}\{\sigma_1, \ldots, \sigma_m\}$ with the singular values being ordered nonincreasingly. In the sequel, we follow [13, 14] for the nonsmooth analysis of the matrix low rank projection. We can partition the index set as

$$\alpha := \{i : \sigma_i > \sigma_r\},\ \beta := \{i : \sigma_i = \sigma_r\}\ \text{and}\ \gamma := \{i : \sigma_i < \sigma_r\}. \tag{B7}$$

Let $\Pi_{\mathrm{R}(r)}(X)$ be the set of optimizers of problem (B6). In generic case, the set $\Pi_{\mathrm{R}(r)}(X)$ is a singleton, while in some cases, it is a smooth manifold of dimension greater than one. Nevertheless, each optimizer of (B6) can be written as

$$\begin{bmatrix} P_\alpha & P_\beta U_\beta \end{bmatrix} \mathrm{diag}(\mathbf{v}) \begin{bmatrix} Q_\alpha & Q_\beta U_\beta \end{bmatrix}^{\mathsf{T}}$$

with an orthogonal matrix $U_\beta \in \mathbb{O}(|\beta|)$ and $\mathbf{v} \in V$ with

$$V := \left\{ \mathbf{v} \in \mathbb{R}^{|\alpha|+|\beta|} \colon v_i = \begin{cases} \sigma_i & \text{for all } i \in \alpha \cup \beta^* \text{ with } \beta^* \subseteq \beta \\ & \text{and } |\beta^*| = r - |\alpha|, \\ 0 & \text{for the others} \end{cases} \right\}. \tag{B8}$$

It is a direct calculation to check that

$$\|X - Y\|^2 = \sum_{i=r+1}^{m} \sigma_i^2$$

is a constant for all $Y \in \Pi_{\mathrm{R}(r)}(X)$, if it is not a singleton. Thus, we will (in some sense abuse of notation) use

$$\|X - \Pi_{\mathrm{R}(r)}(X)\|^2$$

to denote the above constant. Similar convention is taken in some other situation as well. Let

$$\Theta_r(X) := \frac{1}{2}\|\Pi_{\mathrm{R}(r)}(X)\|^2.$$

We then have

$$\begin{aligned} \Theta_r(X) &= \frac{1}{2}\|X\|^2 - \frac{1}{2}\|X - \Pi_{\mathrm{R}(r)}(X)\|^2 \\ &= \frac{1}{2}\|X\|^2 - \min_{Y \in \mathrm{R}(r)} \frac{1}{2}\|X - Y\|^2 \\ &= \max_{Y \in \mathrm{R}(r)} \left\{ \frac{1}{2}\|X\|^2 - \frac{1}{2}\|X - Y\|^2 \right\} \\ &= \max_{Y \in \mathrm{R}(r)} \left\{ \langle X, Y \rangle - \frac{1}{2}\|Y\|^2 \right\}, \end{aligned}$$

which shows that Θ_r is a convex function. As a convex function, we can compute its subdifferentials [47]. Given a subset S, $\mathrm{conv}(S)$ denotes its convex hull in the ambient space. The next result follows from [13, Proposition 2.16].

Lemma 7 *We have*

$$\partial \Theta_r(X) = \mathrm{conv}\left(\Pi_{\mathrm{R}(r)}(X)\right). \tag{B9}$$

Lemma 8 *Given a matrix $X \in \mathbb{R}^{m \times n}$ and positive integer $r \leq m$. If $Y \in \mathrm{conv}(\Pi_{\mathrm{R}(r)}(X))$, then*

$$Y \in \Pi_{\mathrm{R}(r)}(X) \textit{ if and only if } \mathrm{rank}(Y) \leq r.$$

Proof Let

$$X = P\Sigma(X)Q^{\mathsf{T}}$$

be the singular value decomposition of X with $\Sigma = \mathrm{diag}\{\sigma_1, \ldots, \sigma_m\}$ consisting of nonincreasingly ordered singular values. We can partition the index set as (B7). Each matrix $Z \in \Pi_{\mathrm{R}(r)}(X)$ takes the following form

$$\begin{bmatrix} P_\alpha & P_\beta U_\beta \end{bmatrix} \mathrm{diag}(\mathbf{v}) \begin{bmatrix} Q_\alpha & Q_\beta U_\beta \end{bmatrix}^{\mathsf{T}}$$

with $\mathbf{v} \in V$ and V defined as in (B8). More concretely, it can be written as

$$P^{\mathsf{T}} Z \tilde{Q} = \begin{bmatrix} \mathrm{diag}(\Sigma_\alpha) & 0 & 0 \\ 0 & \sigma_r U_{\beta^*} U_{\beta^*}^{\mathsf{T}} & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where $U_{\beta^*} \in \mathbb{R}^{|\beta| \times |\beta^*|}$ is formed by the columns of U_β indexed by β^*, and $\tilde{Q}$ is an orthogonal matrix formed as $[Q \ \bar{Q}]$. Let $Y \in \mathrm{conv}(\Pi_{\mathrm{R}(r)}(X))$. By Carathéodory's theorem [47], we can write

$$Y = \sum_{s=1}^{S} \mu_s Z_s$$

as a convex combination of $Z_s \in \Pi_{\mathrm{R}(r)}(X)$. We thus have

$$P^{\mathsf{T}} Y \tilde{Q} = \begin{bmatrix} \mathrm{diag}(\Sigma_\alpha) & 0 & 0 \\ 0 & \sigma_r \sum_{s=1}^{S} \mu_s U^s_{(\beta^*)_s} (U^s_{(\beta^*)_s})^{\mathsf{T}} & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{B10}$$

for orthogonal matrices U^s and index sets $(\beta^*)_s$ with $s \in \{1, \ldots, S\}$. The case when $\sigma_r = 0$ is trivial. In the following, we assume that $\sigma_r > 0$.

Let

$$p := |(\beta^*)_s| \text{ for all } s = 1, \ldots, S$$

and $p = r - |\alpha|$. Then

$$\mathrm{rank}((U^s_{(\beta^*)_s})(U^s_{(\beta^*)_s})^{\mathsf{T}}) = p, \quad \forall\, s = 1, \ldots, S.$$

By the assumption, we have

$$\mathrm{rank}(Y) \le r = |\alpha| + p.$$

Therefore, we have from (B10) that

$$\mathrm{rank}\left(\sum_{s=1}^{S} \mu_s (U^s_{(\beta^*)_s})(U^s_{(\beta^*)_s})^{\mathsf{T}} \right) = \mathrm{rank}((U^s_{(\beta^*)_s})(U^s_{(\beta^*)_s})^{\mathsf{T}}) = p$$

for each $s = 1, \ldots, S$. In fact, since each component matrix in the summation

$$\sum_{s=1}^{S} \mu_s (U^s_{(\beta^*)_s})(U^s_{(\beta^*)_s})^{\mathsf{T}}$$

is positive semidefinite and has all the eigenvalues being 0 or μ_s, we must have the component matrices are the same. Actually, we have

$$\ker\left(\sum_{s=1}^{S} \mu_s (U^s_{(\beta^*)_s})(U^s_{(\beta^*)_s})^{\mathsf{T}}\right) \subseteq \ker\left((U^1_{(\beta^*)_1})(U^1_{(\beta^*)_1})^{\mathsf{T}}\right)$$

whose dimensions equal to $|\beta| - p$, and thus the two kernels are equal to each other. Consequently, all the kernels

$$\ker((U^s_{(\beta^*)_s})(U^s_{(\beta^*)_s})^{\mathsf{T}}) \text{ for all } s = 1, \ldots, S$$

are the same. Let $W \in \mathbb{R}^{|\beta| \times (|\beta| - p)}$ be a matrix with orthonormal columns which form a basis for the common kernel. Then, we have for all $s = 1, \ldots, S$

$$\left[(U^s_{(\beta^*)_s})\ W\right]$$

is an orthogonal matrix. Therefore, we have

$$(U^s_{(\beta^*)_s})(U^s_{(\beta^*)_s})^{\mathsf{T}} = I - WW^{\mathsf{T}} \text{ for all } s = 1, \ldots, S,$$

which implies that all the matrices in the convex combination are the same. Therefore,

$$Z_1 = \cdots = Z_S.$$

The conclusion then follows. □

By the proof, we see that the extreme points of $\mathrm{conv}(\Pi_{\mathrm{R}(r)}(X))$ are those in the set $\Pi_{\mathrm{R}(r)}(X)$, and can be characterized by the rank function. A similar result for symmetric matrices can be proved similarly, we state it here for its independent interest. Let $\mathrm{S}(r)$ be the set of symmetric matrices of rank at most r.

Proposition 18 *Given a matrix $X \in \mathrm{S}^2(\mathbb{R}^n)$ and a positive integer $r \leq n$. If $Y \in \mathrm{conv}(\Pi_{\mathrm{S}(r)}(X))$, then*

$$Y \in \Pi_{\mathrm{S}(r)}(X) \textit{ if and only if } \mathrm{rank}(Y) \leq r.$$

Acknowledgements We are grateful to the anonymous referees for their valuable comments and suggestions that have helped to improve this paper. Shenglong Hu is supported by the National Science Foundation of China under Grant 12171128 and the Natural Science Foundation of Zhejiang Province, China, under Grant LY22A010022. Defeng Sun is supported in part by RGC Senior Research Fellow Scheme (SRFS) under SRFS2223-5S02 and the Research Center for Intelligent Operations Research at The Hong Kong Polytechnic University. Kim-Chuan Toh is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 3 Grant call (MOE-2019-T3-1-010).

Funding Open access funding provided by The Hong Kong Polytechnic University.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Alexander, J., Hirschowitz, A.: Polynomial interpolation in several variables. J. Algebraic Geom. **4**, 201–222 (1995)
2. Ballico, E., Bernardi, A.: Decomposition of homogeneous polynomials with low rank. Math. Z. **271**(3–4), 1141–1149 (2012)
3. Bernardi, A., Gimigliano, A., Idà, M.: Computing symmetric rank for symmetric tensors. J. Symb. Comput. **46**(1), 34–53 (2011)
4. Bertsekas, D.: Nonlinear Programming, 2nd edn. Athena Scientiffc, Belmont, MA (1999)
5. Brachat, J., Comon, P., Mourrain, B., Tsigaridas, E.: Symmetric tensor decomposition. Linear Algebra Appl. **433**(11–12), 1851–1872 (2010)
6. Comon, P.: Tensor decompostions: state of the art and applications. Institute of Mathematics and its Applications conference series, vol. 71, pp. 1–24. Oxford (2002)
7. Comon, P., Golub, G., Lim, L.-H., Mourrain, B.: Symmetric tensors and symmetric tensor rank. SIAM J. Matrix Anal. Appl. **30**(3), 1254–1279 (2008)
8. Comon, P., Jutten, C.: Handbook of Blind Source Separation: Independent Component Analysis and Applications. Academic Press, Burlington, MA (2010)
9. Comon, P., Mourrain, B.: Decomposition of quantics in sums of powers of linear forms. Signal Process. **53**(2–3), 93–107 (1996)
10. Curto, R.E., Fialkow, L.A.: Truncated K-moment problems in several variables. J. Oper. Theory **54**(1), 189–226 (2005)
11. De Lathauwer, L., De Moor, B., Vandewalle, J.: On the best rank-1 and rank-(R_1, R_2, ..., R_N) approximation of higher-order tensors. SIAM J. Matrix Anal. Appl. **21**(4), 1324–1342 (2001)
12. De Silva, V., Lim, L.-H.: Tensor rank and the ill-posedness of the best low-rank approximation problem. SIAM J. Matrix Anal. Appl. **30**(3), 1084–1127 (2008)
13. Gao, Y.: Structured low rank matrix optimization problems: A penalty approach. Ph.D. thesis, National University of Singapore (2010)
14. Gao, Y., Sun, D.: A majorized penalty approach for calibrating rank constrained correlation matrix problems. Preprint http://www.math.nus.edu.sg/matsundf/MajorPenMay5.pdf (2010)
15. Golub, G.H., Van Loan, C.F.: Matrix Computations, 4th edn. Johns Hopkins University Press, Baltimore (2012)

16. Grasedyck, L., Kressner, D., Tobler, C.: A literature survey of low-rank tensor approximation techniques. GAMM-Mitteilungen **36**(1), 53–78 (2013)
17. Håstad, J.: Tensor rank is NP-complete. J. Algorithms **11**(4), 644–654 (1990)
18. Henrion, D., Lasserre, J.-B.: Detecting global optimality and extracting solutions in GloptiPoly. In: Positive Polynomials in Control, pp. 293–310. Springer, New York (2005)
19. Hiai, F., Lin, M.: On an eigenvalue inequality involving the Hadamard product. Linear Algebra Appl. **515**, 313–320 (2017)
20. Hillar, C.J., Lim, L.-H.: Most tensor problems are NP-hard. J. ACM **60**(6), 1–39 (2013)
21. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, New York (2012)
22. Hu, S., Li, G.: Convergence rate analysis for the higher order power method in best rank one approximations of tensors. Numer. Math. **140**(4), 993–1031 (2018)
23. Hu, S., Sun, D., Toh, K.-C.: Best nonnegative rank-one approximations of tensors. SIAM J. Matrix Anal. Appl. **40**(4), 1527–1554 (2019)
24. Hu, S., Ye, K.: Linear convergence of an alternating polar decomposition method for low rank orthogonal tensor approximations. Math. Program. **199**(1–2), 1305–1364 (2023)
25. Klep, I., Povh, J., Volčič, J.: Minimizer extraction in polynomial optimization is robust. SIAM J. Optim. **28**(4), 3177–3207 (2018)
26. Koffdis, E., Regalia, P.A.: On the best rank-1 approximation of higher-order supersymmetric tensors. SIAM J. Matrix Anal. Appl. **23**(3), 863–884 (2002)
27. Kolda, T.G.: Orthogonal tensor decompositions. SIAM J. Matrix Anal. Appl. **23**(1), 243–255 (2001)
28. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Rev. **51**(3), 455–500 (2009)
29. Kolda, T.G., Mayo, J.R.: Shifted power method for computing tensor eigenpairs. SIAM J. Matrix Anal. Appl. **32**(4), 1095–1124 (2011)
30. 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 (1977)
31. Landsberg, J.: Tensors: Geometry and Applications, vol. 128. Graduate Studies in Mathematics. AMS, Providence, RI (2012)
32. Lasserre, J.B.: Global optimization with polynomials and the problem of moments. SIAM J. Optim. **11**(3), 796–817 (2001)
33. Laurent, M.: Sums of squares, moment matrices and optimization over polynomials. In: Emerging Applications of Algebraic Geometry, pp. 157–270. Springer, New York (2009)
34. Li, X., Sun, D., Toh, K.-C.: A Schur complement based semi-proximal ADMM for convex quadratic conic programming and extensions. Math. Program. **155**(1–2), 333–373 (2016)
35. Lim, L.-H.: Tensors and hypermatrices. In: Handbook of Linear Algebra, pp. 231–260 (2013)
36. Lim, L.-H.: Tensors in computations. Acta Numer. **30**, 555–764 (2021)
37. Lim, L.-H., Comon, P.: Blind multilinear identiffcation. IEEE Trans. Inf. Theory **60**(2), 1260–1280 (2013)
38. Nie, J.: The $\mathcal{A}$-truncated K-moment problem. Found. Comput. Math. **14**, 1243–1276 (2014)
39. Nie, J.: Optimality conditions and finite convergence of Lasserre's hierarchy. Math. Program. **146**(1–2), 97–121 (2014)
40. Nie, J.: Linear optimization with cones of moments and nonnegative polynomials. Math. Program. **153**(1), 247–274 (2015)
41. Nie, J.: Generating polynomials and symmetric tensor decompositions. Found. Comput. Math. **17**(2), 423–465 (2017)
42. Nie, J.: Low rank symmetric tensor approximations. SIAM J. Matrix Anal. Appl. **38**(4), 1517–1540 (2017)
43. Nie, J.: Moment and Polynomial Optimization. SIAM, Philadelphia (2023)
44. Nie, J., Wang, L.: Semideffnite relaxations for best rank-1 tensor approximations. SIAM J. Matrix Anal. Appl. **35**(3), 1155–1179 (2014)
45. Oeding, L., Ottaviani, G.: Eigenvectors of tensors and algorithms for Waring decomposition. J. Symb. Comput. **54**, 9–35 (2013)
46. Qi, L.: The best rank-one approximation ratio of a tensor space. SIAM J. Matrix Anal. Appl. **32**(2), 430–442 (2011)
47. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
48. Shitov, Y.: A counterexample to Comon's conjecture. SIAM J. Appl. Algebra Geom. **2**(3), 428–443 (2018)

49. Sturm, J.: SeDuMi 1.02: A Matlab toolbox for optimization over symmetric cones. Optim. Methods Softw. **11**(12), 625–653 (1999)
50. Tang, G., Shah, P.: Guaranteed tensor decomposition: a moment approach. In: International Conference on Machine Learning, pp. 1491–1500 (2015)
51. Toh, K.-C., Todd, M., Tutuncu, R.: SDPT3: a Matlab software package for semideffnite programming. Optim. Methods Softw. **11**(12), 545–581 (1999)
52. Vaughan, R.C., Wooley, T.D.: Waring's problem: a survey. In: Number Theory for the Millennium, vol. III, 301–340 (2002)
53. Yang, L., Sun, D., Toh, K.-C.: SDPNAL+: a majorized semismooth Newton-CG augmented Lagrangian method for semideffnite programming with nonnegative constraints. Math. Program. Comput. **7**(3), 331–366 (2015)
54. Zhang, T., Golub, G.H.: Rank-one approximation to high order tensors. SIAM J. Matrix Anal. Appl. **23**(2), 534–550 (2001)
55. Zhang, X., Ling, C., Qi, L.: The best rank-1 approximation of a symmetric tensor and related spherical optimization problems. SIAM J. Matrix Anal. Appl. **33**(3), 806–821 (2012)
56. Zhao, X.-Y., Sun, D., Toh, K.-C.: A Newton-CG augmented Lagrangian method for semideffnite programming. SIAM J. Optim. **20**(4), 1737–1765 (2010)