Noname manuscript No.
(will be inserted by the editor)

Optimization Landscape of Tucker Decomposition

Abraham Frandsen · Rong Ge

Abstract Tucker decomposition is a popular technique for many data analysis and machine learning applications. Finding a Tucker decomposition is a nonconvex optimization problem. As the scale of the problems increases, local search algorithms such as stochastic gradient descent have become popular in practice. In this paper, we characterize the optimization landscape of the Tucker decomposition problem. In particular, we show that if the tensor has an exact Tucker decomposition, for a standard nonconvex objective of Tucker decomposition, all local minima are also globally optimal. We also give a local search algorithm that can find an approximate local (and global) optimal solution in polynomial time.

Keywords tensor decomposition, non-convex optimization, local search, optimization landscape

Mathematics Subject Classification (2010) 90C26 · 15A99

This work was supported by NSF Awards CCF-1704656 and CCF-1845171 (CAREER); Sloan Fellowship; and Google Faculty Research Award.

Abraham Frandsen
Duke University, Durham NC
E-mail: abef@cs.duke.edu

Rong Ge
Duke University, Durham, NC
E-mail: rongge@cs.duke.edu

1 Introduction

Tensor decompositions have been widely applied in data analysis and machine learning. In this paper we focus on Tucker decomposition [16,29]. Tucker decomposition has been applied to TensorFaces [30], data compression [31], handwritten digits [25] and more recently to word embeddings [10].

Unlike CP/PARAFAC [5,15] decomposition, Tucker decomposition can be computed efficiently if the original tensor has low rank. For example, this can be done by high-order SVD [7]. Many other algorithms have also been proposed for tensor Tucker decomposition, see for example [8,9,23].

In modern applications, the dimension of the tensor and the amount of data available are often quite large. In practice, simple local search algorithms such as stochastic gradient descent are often used. Even for matrix problems where exact solutions can be computed, local search algorithms are often applied directly to a nonconvex objective [18,24]. Recently, a line of work [11,3, 27,13,28,2] showed that although these problems have nonconvex objectives, they can still be solved by local search algorithms, because they have a simple *optimization landscape*. In particular, for matrix problems such as matrix sensing [3,22] and matrix completion [13,12], it was shown that all local minima are globally optimal. Similar results were also known for special cases of tensor CP decomposition [11].

In this paper, we prove similar results for Tucker decomposition. Given a tensor $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{d\times d\times d}$ with multilinear rank (r,r,r), the Tucker decomposition of the tensor T has the form

$$\boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{S}}^*(\boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{C}^*),$$

where $\boldsymbol{\mathcal{S}}^* \in \mathbb{R}^{r\times r\times r}$ is a core tensor, $\boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{C}^* \in \mathbb{R}^{r\times d}$ are three components (factor matrices). The notation $\boldsymbol{\mathcal{S}}^*(\boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{C}^*)$ is a multilinear form defined in Section 2.1.

To find a Tucker decomposition by local search, the most straight-forward idea is to directly optimize the following nonconvex objective:

$$L(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \|\boldsymbol{\mathcal{T}} - \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F^2.$$

Clearly, $(\boldsymbol{\mathcal{S}}^*, \boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{C}^*)$ is a global minimizer. However, since the optimization problem is nonconvex, it is unclear whether any local search algorithm can efficiently find a globally optimal solution. Our first result (Theorem 1) shows that with an appropriate regularizer (designed in Section 2.2), all local minima of Tucker decomposition are globally optimal.

The main difficulty of analyzing the optimization landscape of Tucker decomposition comes from the existence of *high order saddle points*. For example, when $\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are all equal to 0, any local movement of norm ϵ will only change the objective by at most $O(\epsilon^4)$. Characterizing the possible locations of such high order saddle points, and showing that they cannot become local minima is one of the major technical contributions of this paper.

In general, even if all local minima are globally optimal, a local search algorithm may still fail to find a global optimal solution due to high order saddle

points. In the worst case it is known that 3rd order saddle points can be handled efficiently, while 4th order saddle point are hard to escape from [1]. The objective L has 4th order saddle points. However, our next result (Theorem 2) shows that there is a specifically designed local search algorithm that can find an approximate global optimal solution in polynomial time.

2 Preliminaries

2.1 Tensor Notation and Basic Facts

We use bold lower-case letters like $\boldsymbol{u}$ to denote vectors, bold upper-case letters like $\boldsymbol{A}$ to denote matrices, and bold caligraphic upper-case letters like $\boldsymbol{\mathcal{T}}$ to denote tensors. We reserve the symbol $\boldsymbol{I}$ to denote the identity matrix; its particular dimension will be clear from context. Given a third order tensor $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ and matrices $\boldsymbol{A} \in \mathbb{R}^{r_1 \times d_1}$, $\boldsymbol{B} \in \mathbb{R}^{r_2 \times d_2}$, $\boldsymbol{C} \in \mathbb{R}^{r_3 \times d_3}$, we define $\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ by

$$[\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})]_{ijk} = \sum_{xyz} \boldsymbol{\mathcal{S}}_{xyz} \boldsymbol{A}_{xi} \boldsymbol{B}_{yj} \boldsymbol{C}_{zk}.$$

In the special case where one or more of r_1, r_2, r_3 equals 1, we view $\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ appropriately as a matrix, column vector, or scalar. We equip $\mathbb{R}^{d_1 \times d_2 \times d_3}$ with the Frobenius inner product $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \|_F$ given by

$$\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \rangle = \sum_{i,j,k=1}^{d_1,d_2,d_3} \boldsymbol{\mathcal{X}}_{ijk} \boldsymbol{\mathcal{Y}}_{ijk} \qquad \|\boldsymbol{\mathcal{X}}\|_F = \sqrt{\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}} \rangle}$$

We also define the operator 2-norm $\| \cdot \|_2$ (i.e. the spectral norm) by

$$\|\boldsymbol{\mathcal{X}}\|_2 = \sup \{ \boldsymbol{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \, : \, \|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = \|\boldsymbol{w}\|_2 = 1 \}$$

These two norms are related as follows [32]:

$$\left(\frac{\max(d_1, d_2, d_2)}{d_1 d_2 d_3} \right)^{1/2} \|\boldsymbol{\mathcal{X}}\|_F \leq \|\boldsymbol{\mathcal{X}}\|_2 \leq \|\boldsymbol{\mathcal{X}}\|_F.$$

In the special case of $d_1 = d_2 = d_3 = d$, we have $\|\boldsymbol{\mathcal{X}}\|_F \leq d \|\boldsymbol{\mathcal{X}}\|_2$. Another important fact is that for $\sigma = \|\boldsymbol{\mathcal{X}}\|_2$, there exist unit vectors $\boldsymbol{u} \in \mathbb{R}^{d_1}$, $\boldsymbol{v} \in \mathbb{R}^{d_2}$, and $\boldsymbol{w} \in \mathbb{R}^{d_3}$ such that the following hold [20]:

$$\boldsymbol{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \sigma \quad \boldsymbol{\mathcal{X}}(\boldsymbol{I}, \boldsymbol{v}, \boldsymbol{w}) = \sigma \boldsymbol{u} \quad \boldsymbol{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{I}, \boldsymbol{w}) = \sigma \boldsymbol{v} \quad \boldsymbol{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{I}) = \sigma \boldsymbol{w}$$

Let $\boldsymbol{\mathcal{X}}_{(i)} \in \mathbb{R}^{d_i \times \Pi_{j \neq i} d_j}$ denote the factor-i flattening of $\boldsymbol{\mathcal{X}}$ (for $i = 1, 2, 3$). We say $\boldsymbol{\mathcal{X}}$ has multilinear rank (r_1, r_2, r_3) if the matrix $\boldsymbol{\mathcal{X}}_{(i)}$ has rank r_i; in this case, there exists a tensor $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ and matrices $\boldsymbol{A} \in \mathbb{R}^{r_1 \times d_1}$, $\boldsymbol{B} \in \mathbb{R}^{r_2 \times d_2}$, and $\boldsymbol{C} \in \mathbb{R}^{r_3 \times d_3}$ such that $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$. The tuple $(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ gives a *Tucker decomposition* of $\boldsymbol{\mathcal{X}}$. Note that $\boldsymbol{\mathcal{X}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})_{(1)} = \boldsymbol{A}^\top \boldsymbol{\mathcal{X}}_{(1)} (\boldsymbol{B} \otimes \boldsymbol{C})$ where $\otimes$ denotes the Kronecker product of matrices.

The space of parameters for our objective function is $\mathbb{R}^{r_1 \times r_2 \times r_3} \times \mathbb{R}^{r_1 \times d_1} \times \mathbb{R}^{r_2 \times d_2} \times \mathbb{R}^{r_3 \times d_3}$. We write a point in this space as $(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$, and equip it with inner product $\langle (\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}), (\boldsymbol{\mathcal{S}}', \boldsymbol{A}', \boldsymbol{B}', \boldsymbol{C}') \rangle = \langle \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{S}}' \rangle + \langle \boldsymbol{A}, \boldsymbol{A}' \rangle + \langle \boldsymbol{B}, \boldsymbol{B}' \rangle + \langle \boldsymbol{C}, \boldsymbol{C}' \rangle$ and associated norm

$$\|(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F = \sqrt{\|\boldsymbol{\mathcal{S}}\|_F^2 + \|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2 + \|\boldsymbol{C}\|_F^2}.$$

2.2 Optimization Problem

For simplicity, in this paper we assume $r_1 = r_2 = r_3 = r$, and $d_1 = d_2 = d_3 = d$. It is easy to generalize the result to the case with different r_i's and d_i's. Let $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{d \times d \times d}$ be a fixed third order tensor with multilinear rank (r, r, r) for $r < d$. A simple objective for tensor decomposition can be defined as:

$$L(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \|\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}}\|_F^2. \tag{1}$$

Suppose $\boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{S}}^*(\boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{C}^*)$, then Equation (1) has a global minimum at $(\boldsymbol{\mathcal{S}}^*, \boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{C}^*)$ with the minimum possible L value 0. In fact, due to symmetry, we know there are many more global minimizers of L: for any invertible matrices $\boldsymbol{Q_A}, \boldsymbol{Q_B}, \boldsymbol{Q_C} \in \mathbb{R}^{r \times r}$, let $\boldsymbol{S} = \boldsymbol{\mathcal{S}^*}(\boldsymbol{Q_A}, \boldsymbol{Q_B}, \boldsymbol{Q_C})$, and $\boldsymbol{A} = \boldsymbol{Q_A^{-1}} \boldsymbol{A}^*$, $\boldsymbol{B} = \boldsymbol{Q_B^{-1}} \boldsymbol{B}^*$ and $\boldsymbol{C} = \boldsymbol{Q_C^{-1}} \boldsymbol{C}^*$, then we also have $\boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$. Therefore, the loss L has infinitely many global optimal solutions.

The existence of many equivalent global optimal solutions causes problems for local search algorithms, especially simpler ones like gradient descent. The reason is that if we scale $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ with a large constant c, and scale $\boldsymbol{\mathcal{S}}$ with $1/c^3$, the tensor $\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ does not change. However, after this scaling the partial gradient of $\boldsymbol{\mathcal{S}}$ is multiplied by c^3, while the partial gradients of $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are multiplied by $1/c$. When c is large one has to choose a very small step size for gradient descent, and this results in very slow convergence.

We address the problem of scaling by introducing a regularizer $l(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ given by

$$\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{\mathcal{S}}_{(1)} \boldsymbol{\mathcal{S}}_{(1)}^\top\|_F^2 + \|\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{\mathcal{S}}_{(2)} \boldsymbol{\mathcal{S}}_{(2)}^\top\|_F^2 + \|\boldsymbol{C}\boldsymbol{C}^\top - \boldsymbol{\mathcal{S}}_{(3)} \boldsymbol{\mathcal{S}}_{(3)}^\top\|_F^2. \tag{2}$$

Intuitively, the three terms in the regularizer ensure that $\boldsymbol{A}$ and $\boldsymbol{\mathcal{S}}$ (similarly, $\boldsymbol{B}, \boldsymbol{C}$ and $\boldsymbol{\mathcal{S}}$) have similar norms. Similar regularizers were used for analyzing the optimization landscape of asymmetric matrix problems[22], where the same scaling problem exists. However, to the best of our knowledge we have not seen this regularizer used for Tucker decomposition.

For technical reasons that will become clear in Section 3 (especially in Lemma 4), we actually use $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = l(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})^2$ as the regularizer with weight $\lambda > 0$, so the final optimization problem we consider is:

$$\min_{\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}} L(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) + \lambda R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}). \tag{3}$$

Note that even for Equation (3), there are still infinitely many global minimizers. In particular, one can rotate $\boldsymbol{A}$ and $\boldsymbol{\mathcal{S}}$ (similarly, $\boldsymbol{B}, \boldsymbol{C}$ and $\boldsymbol{\mathcal{S}}$) simultaneously to get equivalent solutions. A priori it is unclear whether there always exists a global minimizer that achieves 0 loss for Equation (3). Our proof in Section 3 implicitly shows that such a solution must exist.

3 Characterization of Optimization Landscape

In this section, we analyze the optimization landscape for the objective (3) for Tucker decomposition. In particular, we establish the following result.

Theorem 1. *For any fixed $\lambda > 0$, all local minima of the objective function $f = L + \lambda R$ as in Equation* (3) *have loss 0.*

Note that the theorem would not hold for $\lambda = 0$ (when there is no regularizer). A counter-example is when $\boldsymbol{\mathcal{T}} = \boldsymbol{a}^* \otimes \boldsymbol{b}^* \otimes \boldsymbol{c}^*$ for some unit vectors $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$, and $\boldsymbol{\mathcal{S}} = \boldsymbol{0}, \boldsymbol{A} = \boldsymbol{a}^\top, \boldsymbol{B} = \boldsymbol{b}^\top, \boldsymbol{C} = \boldsymbol{c}^\top$ where $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ are unit vectors that are orthogonal to $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$ respectively. A local change will have no effect if the new $\boldsymbol{S}$ is still $\boldsymbol{0}$, and will make the objective function larger if the new $\boldsymbol{S}$ is nonzero.

In order to prove this theorem, we demonstrate a direction of improvement for all points $(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ that don't achieve the global optimum. A direction of improvement is a tuple $(\Delta\boldsymbol{\mathcal{S}}, \Delta\boldsymbol{A}, \Delta\boldsymbol{B}, \Delta\boldsymbol{C})$ such that

$$f(\boldsymbol{\mathcal{S}} + \epsilon\Delta\boldsymbol{\mathcal{S}}, \boldsymbol{A} + \epsilon\Delta\boldsymbol{A}, \boldsymbol{B} + \epsilon\Delta\boldsymbol{B}, \boldsymbol{C} + \epsilon\Delta\boldsymbol{C}) < f(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$$

for all sufficiently small $\epsilon > 0$. Clearly, if a point $(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ has a direction of improvement, then it cannot be a local minimum.

Throughout the section, let $\boldsymbol{P}_1$ ($\boldsymbol{P}_2, \boldsymbol{P}_3$) be the projection onto the column span of $\boldsymbol{\mathcal{T}}_{(1)}$ ($\boldsymbol{\mathcal{T}}_{(2)}, \boldsymbol{\mathcal{T}}_{(3)}$). Let $\boldsymbol{A}_1 = \boldsymbol{A}\boldsymbol{P}_1$ and $\boldsymbol{A}_2 = \boldsymbol{A}(\boldsymbol{I} - \boldsymbol{P}_1)$ (similarly for $\boldsymbol{B}, \boldsymbol{C}$). The proof works in the following 4 steps:

Bounding the regularizer First we show that when $\nabla f = \boldsymbol{0}$, the regularizer R must be equal to 0 (Lemm 1 in Section 3.1). At a high level, this is because the gradient of regularizer R is always orthogonal to the gradient of main term L. Therefore if the gradient of the entire objective is $\boldsymbol{0}$, the gradient of R must also be $\boldsymbol{0}$. We complete the proof by showing that $\nabla R = \boldsymbol{0}$ implies $R = 0$.

Removing extraneous directions Next, we show that when $\nabla f = \boldsymbol{0}$, the projection in the wrong subspaces $\boldsymbol{A}_2, \boldsymbol{B}_2, \boldsymbol{C}_2$ are all equal to $\boldsymbol{0}$. This is because the direction of directly removing the projection in the wrong subspace $\boldsymbol{A}_2$ is a direction of improvement (see Lemma 5).

Adding missing directions After the previous steps, we know that the rows of $\boldsymbol{A}$ are in the column span of $\boldsymbol{\mathcal{T}}_{(1)}$. However, the row span of $\boldsymbol{A}$ might be smaller. In this case, there exist directions $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ such that $\boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) > 0$, and $\boldsymbol{A}\boldsymbol{a} = \boldsymbol{0}$. We will show that in this case we can always add the missing

directions into $\boldsymbol{A}$ and $\boldsymbol{\mathcal{S}}$. This is the most technical part of our proof, and high order stationary points may appear when $\boldsymbol{Bb}$ or $\boldsymbol{Cc}$ are also $\mathbf{0}$. See Sections 3.3.

Fixing $\boldsymbol{\mathcal{S}}$ Finally, we know that the components $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ must span the correct subspaces. Our final step shows that in this case, if $L > 0$ then it is easy to find a direction of improvement, see Section 3.4.

3.1 Direction of Improvement for Points with Nonzero Regularizer

We show any point with nonzero regularizer must also have a nonzero gradient, therefore the (negative) gradient itself is a direction of improvement.

Lemma 1. *For any $\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, if $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) > 0$ then $\|\nabla f\| > 0$.*

To prove this, we first show the gradient of the regularizer is always orthogonal to the gradient of the main term (i.e. the tensor loss L).

Lemma 2. *For any $\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, $\langle \nabla L(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}), \nabla R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \rangle = 0$.*

Proof. We start by calculating the partial gradients for L and r. We have

$$\nabla_{\boldsymbol{A}} L = 2\boldsymbol{\mathcal{S}}_{(1)}(\boldsymbol{B} \otimes \boldsymbol{C})(\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}})_{(1)}^\top \quad \nabla_{\boldsymbol{A}} l = 4(\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top)\boldsymbol{A}$$
$$\nabla_{\boldsymbol{B}} L = 2\boldsymbol{\mathcal{S}}_{(2)}(\boldsymbol{A} \otimes \boldsymbol{C})(\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}})_{(2)}^\top \quad \nabla_{\boldsymbol{B}} l = 4(\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{\mathcal{S}}_{(2)}\boldsymbol{\mathcal{S}}_{(2)}^\top)\boldsymbol{B}$$
$$\nabla_{\boldsymbol{C}} L = 2\boldsymbol{\mathcal{S}}_{(3)}(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}})_{(3)}^\top \quad \nabla_{\boldsymbol{C}} l = 4(\boldsymbol{C}\boldsymbol{C}^\top - \boldsymbol{\mathcal{S}}_{(3)}\boldsymbol{\mathcal{S}}_{(3)}^\top)\boldsymbol{C}$$
$$\nabla_{\boldsymbol{\mathcal{S}}} L = 2(\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}})(\boldsymbol{A}^\top, \boldsymbol{B}^\top, \boldsymbol{C}^\top)$$
$$\nabla_{\boldsymbol{\mathcal{S}}} l = 4\boldsymbol{\mathcal{S}}(\boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top - \boldsymbol{A}\boldsymbol{A}^\top, \boldsymbol{I}, \boldsymbol{I}) + 4\boldsymbol{\mathcal{S}}(\boldsymbol{I}, \boldsymbol{\mathcal{S}}_{(2)}\boldsymbol{\mathcal{S}}_{(2)}^\top - \boldsymbol{B}\boldsymbol{B}^\top, \boldsymbol{I})$$
$$+ 4\boldsymbol{\mathcal{S}}(\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{\mathcal{S}}_{(3)}\boldsymbol{\mathcal{S}}_{(3)}^\top - \boldsymbol{C}\boldsymbol{C}^\top)$$

We now compute the following:

$$\begin{aligned}\langle \nabla_{\boldsymbol{A}} L, \nabla_{\boldsymbol{A}} l \rangle &= 8\langle \boldsymbol{\mathcal{S}}_{(1)}(\boldsymbol{B} \otimes \boldsymbol{C})(\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}})_{(1)}^\top, (\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top)\boldsymbol{A} \rangle \\ &= 8\langle (\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}})(\boldsymbol{A}^\top, \boldsymbol{B}^\top, \boldsymbol{C}^\top), \boldsymbol{\mathcal{S}}(\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top, \boldsymbol{I}, \boldsymbol{I}) \rangle\end{aligned}$$

From here it is easy to see that $\langle \nabla_{\boldsymbol{\mathcal{S}}} L, \nabla_{\boldsymbol{\mathcal{S}}} l \rangle = -\langle \nabla_{\boldsymbol{A}} L, \nabla_{\boldsymbol{A}} l \rangle - \langle \nabla_{\boldsymbol{B}} L, \nabla_{\boldsymbol{B}} l \rangle - \langle \nabla_{\boldsymbol{C}} L, \nabla_{\boldsymbol{C}} l \rangle$, therefore $\langle \nabla L, \nabla l \rangle = 0$. Since $\nabla R = 2l\nabla l$, the result follows. □

We next show that if the regularizer is nonzero, then its gradient is nonzero.

Lemma 3. *The function l satisfies*

$$4l(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \langle \nabla_{\boldsymbol{A}} l, \boldsymbol{A} \rangle + \langle \nabla_{\boldsymbol{B}} l, \boldsymbol{B} \rangle + \langle \nabla_{\boldsymbol{C}} l, \boldsymbol{C} \rangle + \langle \nabla_{\boldsymbol{\mathcal{S}}} l, \boldsymbol{\mathcal{S}} \rangle$$

Proof. Note the following calculations:

$$\begin{aligned}\langle \nabla_{\boldsymbol{A}} l, \boldsymbol{A} \rangle &= \langle 4(\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top)\boldsymbol{A}, \boldsymbol{A} \rangle \\ &= 4\langle \boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top, \boldsymbol{A}\boldsymbol{A}^\top \rangle \\ \langle 4\boldsymbol{\mathcal{S}}(\boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top - \boldsymbol{A}\boldsymbol{A}^\top, \boldsymbol{I}, \boldsymbol{I}), \boldsymbol{\mathcal{S}} \rangle &= -4\langle \boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top, \boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top \rangle\end{aligned}$$

The left-hand side above is one of the terms in $\nabla_{\boldsymbol{\mathcal{S}}} l$. Doing the same calculation for the other modes and then adding everything together yields the result. □

Now we are ready to prove Lemma 1:

Proof. By Lemma 2, we know $\|\nabla f\|_F^2 = \|\nabla L\|_F^2 + \|\nabla R\|_F^2$. On the other hand, by Lemma 3 and an application of the Cauchy-Schwarz inequality, we see that

$$\|\nabla l\|_F \|(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F \geq 4l(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}),$$

which means that $\|\nabla l\|_F > 0$ whenever $R = l^2 > 0$. But $\nabla R = 2l\nabla l$, so we have that $\|\nabla R\|_F > 0$, whence $\nabla f \neq \mathbf{0}$. □

To facilitate later proofs, we will also show a fact that if one perturbs a solution with 0 regularizer, then the regularizer remains very small.

Lemma 4. *If $R = 0$, and $\|\Delta \boldsymbol{A}\|_F + \|\Delta \boldsymbol{B}\|_F + \|\Delta \boldsymbol{C}\|_F + \|\Delta \boldsymbol{\mathcal{S}}\|_F \leq O(1)$, then $R(\boldsymbol{\mathcal{S}} + \epsilon\Delta\boldsymbol{\mathcal{S}}, \boldsymbol{A} + \epsilon\Delta\boldsymbol{A}, \boldsymbol{B} + \epsilon\Delta\boldsymbol{B}, \boldsymbol{C} + \epsilon\Delta\boldsymbol{C}) = O(\epsilon^4)$ for sufficiently small ϵ.*

Proof. It suffices to check that the term $\|(\boldsymbol{A} + \epsilon\Delta\boldsymbol{A})(\boldsymbol{A} + \epsilon\Delta\boldsymbol{A})^\top - (\boldsymbol{\mathcal{S}} + \epsilon\Delta\boldsymbol{\mathcal{S}})_{(1)}(\boldsymbol{\mathcal{S}} + \epsilon\Delta\boldsymbol{\mathcal{S}})_{(1)}^\top\|_F = O(\epsilon)$, as other terms are symmetric, and the final R is degree 4 over these terms. This is clear as we know $\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top\|_F = 0$ because $R = 0$, and all the remaining terms are bounded by $O(\epsilon)$. □

3.2 Removing Extraneous Directions

In this section, we show that if $\boldsymbol{A}$ (respectively $\boldsymbol{B}$, $\boldsymbol{C}$) has a direction in its row-space that is perpendicular to the column-space of $\boldsymbol{\mathcal{T}}_{(1)}$ (respectively $\boldsymbol{\mathcal{T}}_{(2)}$, $\boldsymbol{\mathcal{T}}_{(3)}$), then we have a direction of improvement. In particular, our goal is to show $\boldsymbol{A}_2 = \mathbf{0}$ for all local minima (symmetric arguments will then show $\boldsymbol{B}_2 = \boldsymbol{C}_2 = \mathbf{0}$). We first show that $\boldsymbol{\mathcal{S}}(\boldsymbol{A}_2, \boldsymbol{B}, \boldsymbol{C}) = \mathbf{0}$.

Lemma 5. *Assume that $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = 0$. If $\boldsymbol{\mathcal{S}}(\boldsymbol{A}_2, \boldsymbol{B}, \boldsymbol{C}) \neq \mathbf{0}$, then $\Delta\boldsymbol{A} = -\boldsymbol{A}_2$ is a direction of improvement.*

Proof. Set $\Delta\boldsymbol{A} = -\boldsymbol{A}_2$. Then for $\epsilon > 0$

$$\begin{aligned} L(\boldsymbol{\mathcal{S}}, \boldsymbol{A} + \epsilon\Delta\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) &= \|\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}} + \epsilon\boldsymbol{\mathcal{S}}(\Delta\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F^2 \\ &= L(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - 2\epsilon\|\boldsymbol{\mathcal{S}}(\boldsymbol{A}_2, \boldsymbol{B}, \boldsymbol{C})\|_F^2 + O(\epsilon^2), \end{aligned}$$

since $\langle \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{S}}(\boldsymbol{A}_2, \boldsymbol{B}, \boldsymbol{C})\rangle = \langle \boldsymbol{\mathcal{S}}(\boldsymbol{A}_2, \boldsymbol{B}, \boldsymbol{C}), \boldsymbol{\mathcal{S}}(\boldsymbol{A}_2, \boldsymbol{B}, \boldsymbol{C})\rangle$. Hence, for all sufficiently small ϵ, $L(\boldsymbol{\mathcal{S}}, \boldsymbol{A}+\epsilon\Delta\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) < L(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$. By Lemma 4 we know $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A} + \epsilon\Delta\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = O(\epsilon^4)$. Hence, for sufficiently small ϵ, the decrease in L will exceed any increase in R. This shows that $\Delta\boldsymbol{A}$ is a direction of improvement. □

We next establish that $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = 0$ and $\boldsymbol{\mathcal{S}}(\boldsymbol{A}_2, \boldsymbol{B}, \boldsymbol{C}) = \mathbf{0}$ together imply that $\boldsymbol{A}_2 = \mathbf{0}$.

Lemma 6. *If $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = 0$ and $\boldsymbol{\mathcal{S}}(\boldsymbol{A}_2, \boldsymbol{B}, \boldsymbol{C}) = \mathbf{0}$, then $\boldsymbol{A}_2 = \mathbf{0}$.*

Proof. Since $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = 0$, we have $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{\mathcal{S}}_{(2)}\boldsymbol{\mathcal{S}}_{(2)}^\top$ and $\boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{\mathcal{S}}_{(3)}\boldsymbol{\mathcal{S}}_{(3)}^\top$. This means the column span of $\boldsymbol{\mathcal{S}}_{(2)}$ ($\boldsymbol{\mathcal{S}}_{(3)}$) is the same as column span of $\boldsymbol{B}$ ($\boldsymbol{C}$). Let $\boldsymbol{B}^+$ and $\boldsymbol{C}^+$ denote the pseudoinverses of $\boldsymbol{B}$ and $\boldsymbol{C}$. Note that the orthogonal projections onto the column-space of $\boldsymbol{B}$ and $\boldsymbol{C}$ are given by $\boldsymbol{P_B} := \boldsymbol{B}\boldsymbol{B}^+$ and $\boldsymbol{P_C} := \boldsymbol{C}\boldsymbol{C}^+$, respectively. Using these facts along with $\boldsymbol{\mathcal{S}}(\boldsymbol{A}_2, \boldsymbol{B}, \boldsymbol{C}) = \boldsymbol{0}$, we have

$$\boldsymbol{0} = \boldsymbol{\mathcal{S}}(\boldsymbol{A}_2, \boldsymbol{B}, \boldsymbol{C})(\boldsymbol{I}, \boldsymbol{B}^+, \boldsymbol{C}^+) = \boldsymbol{\mathcal{S}}(\boldsymbol{A}_2, \boldsymbol{P_B}, \boldsymbol{P_C}) = \boldsymbol{\mathcal{S}}(\boldsymbol{A}_2, \boldsymbol{I}, \boldsymbol{I}).$$

Using the fact that $\boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top = \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{A}_1\boldsymbol{A}_1^\top + \boldsymbol{A}_2\boldsymbol{A}_2^\top$, we have

$$\|\boldsymbol{A}_2\boldsymbol{A}_2^\top\|_F^2 \leq \langle \boldsymbol{A}_2\boldsymbol{A}_2^\top, \boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top \rangle = \|\boldsymbol{\mathcal{S}}(\boldsymbol{A}_2, \boldsymbol{I}, \boldsymbol{I})\|_F^2 = 0,$$

which, in particular, means that $\boldsymbol{A}_2 = \boldsymbol{0}$. □

3.3 Adding Missing Directions

We now consider the case where the row-spans of $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are not equal to the column-spans of $\boldsymbol{\mathcal{T}}_{(1)}$,$\boldsymbol{\mathcal{T}}_{(2)}$, and $\boldsymbol{\mathcal{T}}_{(3)}$, respectively. Again by symmetry, we focus on the case when row-span of $\boldsymbol{A}$ is not equal to column-span of $\boldsymbol{\mathcal{T}}_{(1)}$.

Lemma 7. *If the row-span of $\boldsymbol{A}$ is a strict subset of the column-span of $\boldsymbol{\mathcal{T}}_{(1)}$ and $R = 0$, then there is a direction of improvement.*

Proof. If the row-span of $\boldsymbol{A}$ is a strict subset of column-span of $\boldsymbol{\mathcal{T}}_{(1)}$, we must have a vector $\boldsymbol{a}$ that is in the column-span of $\boldsymbol{\mathcal{T}}_{(1)}$, but $\boldsymbol{A}\boldsymbol{a} = \boldsymbol{0}$. For this vector we know $\boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{I}, \boldsymbol{I}) \neq \boldsymbol{0}$, therefore there must exist vectors $\boldsymbol{b}, \boldsymbol{c}$ such that $\boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) > 0$. Notice that if we let $\boldsymbol{b}_1$ be the projection of $\boldsymbol{b}$ onto the row-span of $\boldsymbol{B}$ and $\boldsymbol{c}_1$ be the projection of $\boldsymbol{c}$ onto the row-span of $\boldsymbol{C}$, and set $\boldsymbol{b}_2 = \boldsymbol{b} - \boldsymbol{b}_1$ and $\boldsymbol{c}_2 = \boldsymbol{c} - \boldsymbol{c}_1$, then we have $\boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \sum_{i,j \in \{1,2\}} \boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{b}_i, \boldsymbol{c}_j)$. Hence, $\boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{b}_i, \boldsymbol{c}_j) > 0$ for some choice of $i, j \in \{1, 2\}$. Therefore, without loss of generality, we can restrict $\boldsymbol{b}$ to be either in the row-span of $\boldsymbol{B}$ or to satisfy $\boldsymbol{B}\boldsymbol{b} = \boldsymbol{0}$ (and similarly for $\boldsymbol{c}$), while retaining the property that $\boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) > 0$. We proceed by considering the following cases.

One missing direction In this case $\boldsymbol{b}$ and $\boldsymbol{c}$ are in row span of $\boldsymbol{B}, \boldsymbol{C}$ respectively. Choose unit vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^r$ such that $\boldsymbol{A}^\top\boldsymbol{u} = \boldsymbol{0}$, $\boldsymbol{B}^\top\boldsymbol{v} = \alpha_1\boldsymbol{b}$, and $\boldsymbol{C}^\top\boldsymbol{w} = \alpha_2\boldsymbol{c}$, where α_1 and α_2 are positive real numbers. Consider the directions $\varDelta\boldsymbol{A} = \boldsymbol{u}\boldsymbol{a}^\top$, $\varDelta\boldsymbol{\mathcal{S}} = \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}$. Observe that $\varDelta\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \boldsymbol{A}^\top\boldsymbol{u} \otimes \boldsymbol{B}^\top\boldsymbol{v} \otimes \boldsymbol{C}^\top\boldsymbol{w} = \boldsymbol{0}$ and $\boldsymbol{\mathcal{S}}(\varDelta\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \boldsymbol{0}$ since the column-space of $\boldsymbol{S}_{(1)}$ is equal to the column-space of $\boldsymbol{A}$. Moreover, $\varDelta\boldsymbol{\mathcal{S}}(\varDelta\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \boldsymbol{a} \otimes \boldsymbol{B}^\top\boldsymbol{v} \otimes \boldsymbol{C}^\top\boldsymbol{w} = \alpha_1\alpha_2\boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}$. Hence, for $\epsilon > 0$, we have

$$\begin{aligned} L(\boldsymbol{\mathcal{S}} + \epsilon\varDelta\boldsymbol{\mathcal{S}}, \boldsymbol{A} + \epsilon\varDelta\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) &= \|\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}} + \epsilon^2\varDelta\boldsymbol{\mathcal{S}}(\varDelta\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F^2 \\ &= L(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - 2\epsilon^2\alpha_1\alpha_2\boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) + O(\epsilon^4). \end{aligned}$$

On the other hand, by Lemma 4 $R(\boldsymbol{\mathcal{S}} + \epsilon\Delta\boldsymbol{\mathcal{S}}, \boldsymbol{A} + \epsilon\Delta\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = O(\epsilon^4)$ since $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = 0$. Hence, for small enough ϵ, the improvement in the tensor loss dominates all other perturbations, so we have a direction of improvement.

Two missing directions Now assume that $\boldsymbol{Aa} = \boldsymbol{Bb} = \boldsymbol{0}$, and $\boldsymbol{c}$ is in the row span of $\boldsymbol{C}$. Choose unit vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^r$ such that $\boldsymbol{A}^\top \boldsymbol{u} = \boldsymbol{B}^\top \boldsymbol{v} = \boldsymbol{0}$ and $\boldsymbol{C}^\top \boldsymbol{w} = \alpha \boldsymbol{c}$ where $\alpha > 0$. Consider the directions $\Delta\boldsymbol{A} = \boldsymbol{u}\boldsymbol{a}^\top$, $\Delta\boldsymbol{B} = \boldsymbol{v}\boldsymbol{b}^\top$, $\Delta\boldsymbol{\mathcal{S}} = \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}$. Through a very similar calculation as in the previous case,

$$L(\boldsymbol{\mathcal{S}} + \epsilon\Delta\boldsymbol{\mathcal{S}}, \boldsymbol{A} + \epsilon\Delta\boldsymbol{A}, \boldsymbol{B} + \epsilon\Delta\boldsymbol{B}, \boldsymbol{C}) = L(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - 2\epsilon^3\alpha\boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) + \epsilon^6\alpha^2.$$

As before, by Lemma 4 $R(\boldsymbol{\mathcal{S}} + \epsilon\Delta\boldsymbol{\mathcal{S}}, \boldsymbol{A} + \epsilon\Delta\boldsymbol{A}, \boldsymbol{B} + \epsilon\Delta\boldsymbol{B}, \boldsymbol{C}) = O(\epsilon^4)$. Hence, the decrease in the tensor loss dominates all other perturbations for sufficiently small ϵ, and so this is a direction of improvement. Note that in this case the amount of improvement is $\Theta(\epsilon^3)$, so the point is a 3rd order saddle point.

The case where $\boldsymbol{Cc} = \boldsymbol{0}$ and $\boldsymbol{b}$ is in the row-span of $\boldsymbol{B}$ is similar, and likewise yields a direction of improvement.

Three missing directions Now assume that $\boldsymbol{Aa} = \boldsymbol{Bb} = \boldsymbol{Cc} = \boldsymbol{0}$, and choose unit vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^r$ such that $\boldsymbol{A}^\top \boldsymbol{a} = \boldsymbol{B}^\top \boldsymbol{v} = \boldsymbol{C}^\top \boldsymbol{w} = \boldsymbol{0}$. Consider the directions $\Delta\boldsymbol{A} = \boldsymbol{u}\boldsymbol{a}^\top$, $\Delta\boldsymbol{B} = \boldsymbol{v}\boldsymbol{b}^\top$, $\Delta\boldsymbol{C} = \boldsymbol{w}\boldsymbol{c}^\top$, and $\Delta\boldsymbol{\mathcal{S}} = \boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w}$. Once again, most perturbations in the tensor loss vanish, and we have

$$L(\boldsymbol{\mathcal{S}}+\epsilon\Delta\boldsymbol{\mathcal{S}}, \boldsymbol{A}+\epsilon\Delta\boldsymbol{A}, \boldsymbol{B}+\epsilon\Delta\boldsymbol{B}, \boldsymbol{C}+\epsilon\Delta\boldsymbol{C}) = L(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - 2\epsilon^4\boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) + \epsilon^8.$$

In this case, the regularizer doesn't change at all, since $\Delta\boldsymbol{\mathcal{S}}_{(i)}\boldsymbol{\mathcal{S}}_{(i)}^\top = \boldsymbol{0}$ for $i = 1, 2, 3$, $\Delta\boldsymbol{A}\boldsymbol{A}^\top = \Delta\boldsymbol{B}\boldsymbol{B}^\top = \Delta\boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{0}$, and $\Delta\boldsymbol{A}\Delta\boldsymbol{A}^\top - \Delta\boldsymbol{\mathcal{S}}_{(1)}\Delta\boldsymbol{\mathcal{S}}_{(1)}^\top = \boldsymbol{u}\boldsymbol{u}^\top - \boldsymbol{u}\boldsymbol{u}^\top = \boldsymbol{0}$ (and the two other analogous terms likewise vanish). Hence, for sufficiently small ϵ, the objective function decreases, so this is a direction of improvement. This point is a 4th order saddle point.

□

3.4 Improving the core tensor

We finally consider the case where the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ have the correct row-spaces but $\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \neq \boldsymbol{\mathcal{T}}$. In this situation, we can make progress by changing only $\boldsymbol{\mathcal{S}}$.

Lemma 8. *If $R = 0$, row spans of $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are equal to column span of $\boldsymbol{\mathcal{T}}_{(1)}, \boldsymbol{\mathcal{T}}_{(2)}, \boldsymbol{\mathcal{T}}_{(3)}$ respectively, but $L > 0$, then there exists a direction of improvement.*

Proof. Since the spans of $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are already correct, let $\boldsymbol{A}^+$ be the pseudoinverse of $\boldsymbol{A}$, then if we let $\boldsymbol{\mathcal{S}}' = \boldsymbol{\mathcal{T}}(\boldsymbol{A}^+, \boldsymbol{B}^+, \boldsymbol{C}^+)$, we have $\boldsymbol{\mathcal{S}}'(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \boldsymbol{\mathcal{T}}$. Consider the direction $\Delta\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{S}}' - \boldsymbol{\mathcal{S}}$.

$$\begin{aligned} L(\boldsymbol{\mathcal{S}} + \epsilon\Delta\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) &= \|(1-\epsilon)\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - (1-\epsilon)\boldsymbol{\mathcal{T}}\|_F^2 \\ &= (1-\epsilon)^2 L(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}). \end{aligned}$$

For the regularizer R, again by Lemma 4 we have $R(\boldsymbol{\mathcal{S}} + \epsilon\Delta\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = O(\epsilon^4)$. Hence, this is a direction of improvement. □

3.5 Proof of Main Theorem

Now with all the lemmas we are ready to prove the main theorem:

Proof of Theorem 1. The Theorem follows immediately from the sequence of lemmas.

First, by Lemma 1, we know any local minima must satisfy $R = 0$. Next, by Lemma 6 and Lemma 5, we know the row spans of $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ must be subsets of column spans of $\boldsymbol{\mathcal{T}}_{(1)}, \boldsymbol{\mathcal{T}}_{(2)}, \boldsymbol{\mathcal{T}}_{(3)}$ respectively. In the third step, by Lemma 7, we further show that the row spans of $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ must be exactly equal to column spans of $\boldsymbol{\mathcal{T}}_{(1)}, \boldsymbol{\mathcal{T}}_{(2)}, \boldsymbol{\mathcal{T}}_{(3)}$ respectively. Finally, by Lemma 8 we know the loss function must be equal to 0. □

4 Escaping from High Order Saddle Points for Tucker Decomposition

As we discussed before, since our objective $f = L + \lambda R$ as in (3) may have high order saddle points, standard local search algorithms may not be able to find a local minimum. However, in this section we show that the high order saddle points of f are *benign*: there is a polynomial time local search algorithm that can find an approximate local and global minimum of f.

We will first review the guarantees of standard local search algorithms, and then describe how to escape from high order saddle points.

4.1 Local search algorithms for second order stationary points

For a general function $f(\boldsymbol{x})$ whose first two derivatives exist, we say a point $\boldsymbol{x}$ is a (τ_1, τ_2)-second order stationary point if

$$\|\nabla f(\boldsymbol{x})\| \leq \tau_1, \quad \lambda_{min}(\nabla^2 f(\boldsymbol{x})) \geq -\tau_2.$$

If the function $f(\boldsymbol{x})$ satisfies the gradient and Hessian Lipschitz conditions

$$\forall \boldsymbol{x}, \boldsymbol{y} \quad \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq \rho_1 \|\boldsymbol{x} - \boldsymbol{y}\|_2,$$
$$\forall \boldsymbol{x}, \boldsymbol{y} \quad \|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \leq \rho_2 \|\boldsymbol{x} - \boldsymbol{y}\|_2,$$

there are many local search algorithms that can find (τ_1, τ_2)-second order stationary points in polynomial time. This includes traditional second order algorithms such as cubic regularization[21], and more recently first order algorithms such as perturbed gradient descent[17].

Of course, these guarantees are not enough for our objective f, as it has higher order saddle points. The main theorem in this section shows that there is an efficient local search algorithm that can optimize f.

Symbol	Definition	Note
λ	$\frac{1}{16r^4}$	weight for regularizer
K	$O^*(1)$	universal bound for norms of $\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\mathcal{T}}$
τ	< 1	bound on $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$
γ	$\Theta^*(\tau^{1/48})$	bound on the norm of $\boldsymbol{A}_3, \boldsymbol{B}_3, \boldsymbol{C}_3$, introduced in Lemma 14
σ	$\sqrt{\gamma}$	singular value threshold for $\boldsymbol{A}_1, \boldsymbol{B}_1, \boldsymbol{C}_1$
κ_0	$\sqrt{\gamma}$	max error in $\boldsymbol{\mathcal{T}}_{1,1,1}$, introduced in Lemma 17
κ_1	$2K\sigma^{3/4}$	max error in $\boldsymbol{\mathcal{T}}_{2,1,1}$, introduced in Lemma 18
κ_2	$2K\sigma^{1/8}$	max error in $\boldsymbol{\mathcal{T}}_{2,2,1}$, introduced in Lemma 20
κ_3	$2K\sigma^{1/2}$	max error in $\boldsymbol{\mathcal{T}}_{2,2,2}$, introduced in Lemma 21

Table 1 Notation and definitions used in Section 4

Theorem 2. *Let $\lambda = 1/16r^4$, assume wlog. that $\|\boldsymbol{\mathcal{T}}\|_F = 1$ and the initial point satisfies $f = L + \lambda R = O(1)$[1]. Then there is a local search algorithm that in poly$(d, r, 1/\epsilon)$ time finds a point $(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ such that $f(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \le \epsilon$.*

The algorithm that we will design is just a proof of concept: although its running time is polynomial, it is far from practical. We have not attempted to improve the dependencies on $d, r, 1/\epsilon$. Local search algorithms seem to perform much better for Tucker decomposition in practice, and understanding that is an interesting open problem.

To prove Theorem 2, we will first show that sublevel sets of f are all bounded (Section 4.2). This allows us to bound the gradient and Hessian Lipschitz constants ρ_1 and ρ_2, so we can use any of the previous local search algorithm to find a (τ_1, τ_2)-second order stationary point.

Next, we follow the steps of Theorem 1, but we do it much more carefully to show that as long as the objective is larger than ϵ, then either the point has a large gradient or a negative eigenvalue in Hessian, or there is a way to construct a direction of improvement. This is captured in our main Lemma 12.

Finally we give a sketch of the algorithm and show that these local improvements are enough to guarantee the convergence in Section 4.8.

Throughout the section, we use $O^*(\cdot)$, $\Omega^*(\cdot)$ and $\Theta^*(\cdot)$ to hide polynomial factors of r and d. We will introduce several numerical quantities in the remainder of this section; we list the most important ones in Table 1.

4.2 Bounded Sublevel Sets

We first establish the boundedness of sublevel sets of the objective function. Our local search algorithm will guarantee the function value decreases in every iteration, so the trajectory of the algorithm will remain in a sublevel set. As a result, we know that the parameters remain bounded in norm at each step by some constant, say K.

Lemma 9. *For all $\Gamma \ge 0$, the set of points $(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ with $f(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \le \Gamma$ satisfy that $\|\boldsymbol{\mathcal{S}}\|_F, \|\boldsymbol{A}\|_F, \|\boldsymbol{B}\|_F, \|\boldsymbol{C}\|_F \le K$ where $K = O^*((\Gamma + 1)^{1/8})$.*

[1] This can be achieved by initializing at 0, or any point with norm $O(1)$.

To prove this lemma, we will first state some tools that we need.

Lemma 10. *For any parameter tuple $(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$, we have*

$$\|\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F \leq \|\boldsymbol{\mathcal{S}}\|_F \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_2 \|\boldsymbol{C}\|_2.$$

Proof. This follows from the fact that $\|\cdot\|_F$ is invariant to matricization, and the fact that $\|\boldsymbol{PQ}\|_F \leq \|\boldsymbol{P}\|_2 \|\boldsymbol{Q}\|_F$. Observe that

$$\|\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F = \|\boldsymbol{A}^\top \boldsymbol{\mathcal{S}}(\boldsymbol{I}, \boldsymbol{B}, \boldsymbol{C})_{(1)}\|_F \leq \|\boldsymbol{A}\|_2 \|\boldsymbol{\mathcal{S}}(\boldsymbol{I}, \boldsymbol{B}, \boldsymbol{C})\|_F,$$

and then repeat the argument for the other modes. □

Lemma 11. *For any $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{r \times r \times r}$, it holds that*

$$\|\boldsymbol{\mathcal{S}}(\boldsymbol{\mathcal{S}}_{(1)}, \boldsymbol{\mathcal{S}}_{(2)}, \boldsymbol{\mathcal{S}}_{(3)})\|_F \geq \frac{1}{r^4} \|\boldsymbol{\mathcal{S}}\|_F^4$$

Proof. Let $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^r$ be unit vectors such that $\boldsymbol{\mathcal{S}}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \|\boldsymbol{\mathcal{S}}\|_2$. Then

$$\begin{aligned}
\boldsymbol{\mathcal{S}}_{(3)} \mathrm{vec}(\boldsymbol{u} \otimes \boldsymbol{v}) &= \boldsymbol{\mathcal{S}}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{I}) = \|\boldsymbol{\mathcal{S}}\|_2 \boldsymbol{w}, \\
\boldsymbol{\mathcal{S}}_{(2)} \mathrm{vec}(\boldsymbol{u} \otimes \boldsymbol{w}) &= \boldsymbol{\mathcal{S}}(\boldsymbol{u}, \boldsymbol{I}, \boldsymbol{w}) = \|\boldsymbol{\mathcal{S}}\|_2 \boldsymbol{v}, \\
\boldsymbol{\mathcal{S}}_{(1)} \mathrm{vec}(\boldsymbol{v} \otimes \boldsymbol{w}) &= \boldsymbol{\mathcal{S}}(\boldsymbol{I}, \boldsymbol{v}, \boldsymbol{w}) = \|\boldsymbol{\mathcal{S}}\|_2 \boldsymbol{u}.
\end{aligned}$$

Then

$$\|\boldsymbol{\mathcal{S}}(\boldsymbol{\mathcal{S}}_{(1)}, \boldsymbol{\mathcal{S}}_{(2)}, \boldsymbol{\mathcal{S}}_{(3)})\|_2 \geq \boldsymbol{\mathcal{S}}(\|\boldsymbol{\mathcal{S}}\|_2 \boldsymbol{u}, \|\boldsymbol{\mathcal{S}}\|_2 \boldsymbol{v}, \|\boldsymbol{\mathcal{S}}\|_2 \boldsymbol{w}) = \|\boldsymbol{\mathcal{S}}\|_2^3 \boldsymbol{\mathcal{S}}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \|\boldsymbol{\mathcal{S}}\|_2^4$$

The result then follows from the norm inequality $\|\boldsymbol{\mathcal{S}}\|_F \leq r \|\boldsymbol{\mathcal{S}}\|_2$. □

Now we are ready to prove Lemma 9:

Proof of Lemma 9. Assume that $\Gamma \geq f(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$. From L, we have

$$\begin{aligned}
\sqrt{\Gamma} &\geq \|\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}}\|_F \\
&\geq \|\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F - \|\boldsymbol{\mathcal{T}}\|_F,
\end{aligned}$$

so that $\|\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F \leq \sqrt{\Gamma} + \|\boldsymbol{\mathcal{T}}\|_F$. Next, define the following for $i = 1, 2, 3$: $d_i(\boldsymbol{X}, \boldsymbol{\mathcal{S}}) = \boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\mathcal{S}}_{(i)} \boldsymbol{\mathcal{S}}_{(i)}^\top$. Note that $d_1(\boldsymbol{A}, \boldsymbol{\mathcal{S}}), d_2(\boldsymbol{B}, \boldsymbol{\mathcal{S}}), d_3(\boldsymbol{C}, \boldsymbol{\mathcal{S}})$ are each bounded above in norm by $\Gamma^{1/4}$. We have

$$\begin{aligned}
\|\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F^2 &= \langle \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}), \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \rangle \\
&= \langle \boldsymbol{\mathcal{S}}(\boldsymbol{A}\boldsymbol{A}^\top, \boldsymbol{B}\boldsymbol{B}^\top, \boldsymbol{C}\boldsymbol{C}^\top), \boldsymbol{\mathcal{S}} \rangle \\
&= \langle \boldsymbol{\mathcal{S}}(\boldsymbol{\mathcal{S}}_{(1)} \boldsymbol{\mathcal{S}}_{(1)}^\top, \boldsymbol{\mathcal{S}}_{(2)} \boldsymbol{\mathcal{S}}_{(2)}^\top, \boldsymbol{\mathcal{S}}_{(3)} \boldsymbol{\mathcal{S}}_{(3)}^\top), \boldsymbol{\mathcal{S}} \rangle + g(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \\
&= \|\boldsymbol{\mathcal{S}}(\boldsymbol{\mathcal{S}}_{(1)}, \boldsymbol{\mathcal{S}}_{(2)}, \boldsymbol{\mathcal{S}}_{(3)})\|_F^2 + g(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}),
\end{aligned}$$

where $g(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ is a sum of the seven remainder terms of the form

$$\langle \boldsymbol{\mathcal{S}}(d_1(\boldsymbol{A}, \boldsymbol{\mathcal{S}}), \boldsymbol{\mathcal{S}}_{(2)} \boldsymbol{\mathcal{S}}_{(2)}^\top, \boldsymbol{\mathcal{S}}_{(3)} \boldsymbol{\mathcal{S}}_{(3)}^\top), \boldsymbol{\mathcal{S}} \rangle \tag{4}$$

$$\langle \boldsymbol{\mathcal{S}}(d_1(\boldsymbol{A}, \boldsymbol{\mathcal{S}}), d_2(\boldsymbol{B}, \boldsymbol{\mathcal{S}}), \boldsymbol{\mathcal{S}}_{(3)} \boldsymbol{\mathcal{S}}_{(3)}^\top), \boldsymbol{\mathcal{S}} \rangle \tag{5}$$

$$\langle \boldsymbol{\mathcal{S}}(d_1(\boldsymbol{A}, \boldsymbol{\mathcal{S}}), d_2(\boldsymbol{B}, \boldsymbol{\mathcal{S}}), d_3(\boldsymbol{C}, \boldsymbol{\mathcal{S}})), \boldsymbol{\mathcal{S}} \rangle \tag{6}$$

There are three terms of type (4), and each can be bounded below using Cauchy-Schwarz and Lemma 10 as follows:

$$\langle \boldsymbol{\mathcal{S}}(d_1(\boldsymbol{A},\boldsymbol{\mathcal{S}}), \boldsymbol{\mathcal{S}}_{(2)}\boldsymbol{\mathcal{S}}_{(2)}^\top, \boldsymbol{\mathcal{S}}_{(3)}\boldsymbol{\mathcal{S}}_{(3)}^\top), \boldsymbol{\mathcal{S}}\rangle \geq -\|\boldsymbol{\mathcal{S}}\|_F^6 \|d_1(\boldsymbol{A},\boldsymbol{\mathcal{S}})\|_F \geq -\Gamma^{1/4}\|\boldsymbol{\mathcal{S}}\|_F^6.$$

Similarly, we have

$$\langle \boldsymbol{\mathcal{S}}(d_1(\boldsymbol{A},\boldsymbol{\mathcal{S}}), d_2(\boldsymbol{B},\boldsymbol{\mathcal{S}}), \boldsymbol{\mathcal{S}}_{(3)}\boldsymbol{\mathcal{S}}_{(3)}^\top), \boldsymbol{\mathcal{S}}\rangle \geq -\Gamma^{1/2}\|\boldsymbol{\mathcal{S}}\|_F^4,$$
$$\langle \boldsymbol{\mathcal{S}}(d_1(\boldsymbol{A},\boldsymbol{\mathcal{S}}), d_2(\boldsymbol{B},\boldsymbol{\mathcal{S}}), d_3(\boldsymbol{C},\boldsymbol{\mathcal{S}})), \boldsymbol{\mathcal{S}}\rangle \geq -\Gamma^{3/4}\|\boldsymbol{\mathcal{S}}\|_F^2.$$

Putting this together and applying Lemma 11, we have that

$$\frac{1}{r^8}\|\boldsymbol{\mathcal{S}}\|_F^8 - 3\Gamma^{1/4}\|\boldsymbol{\mathcal{S}}\|_F^6 - 3\Gamma^{1/2}\|\boldsymbol{\mathcal{S}}\|_F^4 - \Gamma^{3/4}\|\boldsymbol{\mathcal{S}}\|_F^2 \leq (\sqrt{\Gamma} + \|\boldsymbol{\mathcal{T}}\|_F)^2,$$

which means that $\|\boldsymbol{\mathcal{S}}\|_F$ must be bounded by $O^*((\Gamma+1)^{1/8})$. From R, we have

$$\left(\frac{\Gamma}{\lambda}\right)^{1/4} + \|\boldsymbol{\mathcal{S}}\|_F^2 \geq \|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top\|_F + \|\boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top\|_F \geq \|\boldsymbol{A}\boldsymbol{A}^\top\|_F,$$

so $\|\boldsymbol{A}\|_F$ is bounded by $O^*((\Gamma+1)^{1/8})$. We bound $\boldsymbol{B}$ and $\boldsymbol{C}$ similarly. □

4.3 Main step: making local improvements

In order to prove Theorem 2, we rely on the following main lemma:

Lemma 12. *In the same setting as Theorem 2, there exist positive constants $q_1, q_2, \tau_1 = \Theta^*(\epsilon^{q_1})$, $\tau_2 = \Theta^*(\epsilon^{q_2})$, such that for any point $\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ where $\epsilon < f(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) < O(1)$, one of the following is true:*

1. $\|\nabla f(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\| \geq \tau_1$,
2. $\lambda_{min}(\nabla^2 f(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})) \leq -\tau_2$,
3. *With constant probability, Algorithm 1 constructs a direction of improvement that improves the function value by poly(ϵ).*

Algorithm 1 uses notation that we specify in the paragraphs below. The

Algorithm 1 Sampling algorithm for adding missing directions

Require: matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, threshold σ, subspace indicator $(i,j,k) \in \{1,2\}^3$
Compute the subspaces $U_{1,i}, U_{2,j}, U_{3,k}, V_{1,i}, V_{2,j}, V_{3,k}$
Sample unit vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ uniformly from $U_{1,i}, U_{2,j}, U_{3,k}$
if $i = 1$ **then** $\boldsymbol{u}' = (\boldsymbol{A}_1^\top)^+\boldsymbol{a}$; **else** Randomly sample nonzero $\boldsymbol{u}' \in V_{1,2}$
if $j = 1$ **then** $\boldsymbol{v}' = (\boldsymbol{B}_1^\top)^+\boldsymbol{b}$; **else** Randomly sample nonzero $\boldsymbol{v}' \in V_{2,2}$
if $k = 1$ **then** $\boldsymbol{w}' = (\boldsymbol{C}_1^\top)^+\boldsymbol{c}$; **else** Randomly sample nonzero $\boldsymbol{w}' \in V_{3,2}$
Return $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{u}'/\|\boldsymbol{u}'\|_2, \boldsymbol{v}'/\|\boldsymbol{v}'\|_2, \boldsymbol{w}'/\|\boldsymbol{w}'\|_2$

proof of this lemma has similar steps to the proof of Theorem 1. However, it is more complicated because we are not looking at exact local minima. We

give the details of these steps in the following subsections. A key parameter that we will use is a bound τ on the regularizer. We will consider different cases when $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \geq \tau$ and when $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \leq \tau$. All of our other parameters (including τ_1, τ_2, ϵ) will be polynomials in τ.

For the analysis, it is useful to consider $\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ and $\boldsymbol{\mathcal{T}}$ projected onto various subspaces of interest. To this end, we introduce the following notation. Let $\sigma > 0$ be a threshold that we will specify later. For matrix $\boldsymbol{A}$, we let $V_{1,1}$ and $U_{1,1}$ denote the spaces spanned by the left/right singular vectors of $\boldsymbol{A}$ with singular value greater than σ, and let $V_{1,2} = V_{1,1}^{\perp}$, $U_{1,2} = U_{1,1}^{\perp}$. We can then write $\boldsymbol{A} = \boldsymbol{A}_1 + \boldsymbol{A}_2$, where $\boldsymbol{A}_1 = \mathrm{Proj}_{V_{1,1}} \boldsymbol{A}$ contains the larger singular vectors and $\boldsymbol{A}_2 = \mathrm{Proj}_{V_{1,2}} \boldsymbol{A}$ contains the smaller singular vectors. Let $\boldsymbol{P}_1$ be the orthogonal projection onto the column-space of $\boldsymbol{\mathcal{T}}_{(1)}$ and define $\boldsymbol{A}_3 = \boldsymbol{A}(\boldsymbol{I} - \boldsymbol{P}_1)$, the projection of $\boldsymbol{A}$ onto directions that are unrelated to the true tensor. Similarly, we define $U_{2,1}, U_{2,2}, V_{2,1}, V_{2,2}, \boldsymbol{P}_2, \boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3$ for matrix $\boldsymbol{B}$ and $U_{3,1}, U_{3,2}, V_{3,1}, V_{3,2}, \boldsymbol{P}_3, \boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{C}_3$ for matrix $\boldsymbol{C}$. Define $\boldsymbol{\mathcal{S}}_{i,j,k} = \boldsymbol{\mathcal{S}}(\mathrm{Proj}_{V_{1,i}}, \mathrm{Proj}_{V_{2,j}}, \mathrm{Proj}_{V_{3,k}})$ and $\boldsymbol{\mathcal{T}}_{i,j,k} = \boldsymbol{\mathcal{T}}(\mathrm{Proj}_{U_{1,i}}, \mathrm{Proj}_{U_{2,j}}, \mathrm{Proj}_{U_{3,k}})$. We can decompose the tensor loss as

$$\|\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}}\|_F^2 = \sum_{i,j,k \in \{1,2\}} \|\boldsymbol{\mathcal{S}}_{i,j,k}(\boldsymbol{A}_i, \boldsymbol{B}_j, \boldsymbol{C}_k) - \boldsymbol{\mathcal{T}}_{i,j,k}\|_F^2.$$

Our analysis shows how to decrease the objective function if the regularizer or any one of the terms in the right-hand sum is sufficiently large. In particular, after finding a second-order stationary point, the only terms in this sum that may be large are when at least two of i, j, k are equal to 2. In this case, Algorithm 1 can be used to make further progress toward a local minimum.

4.4 Decreasing the Regularizer

We first show if the regularizer is large, then the gradient is large. This is very similar to Lemma 1.

Lemma 13. *If $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \geq \tau$, then $\|\nabla f(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F \geq 4\lambda\tau/K$.*

Proof. By assumption, $l(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \geq \tau^{1/2}$, and we have$\|\nabla R\|_F = \|2l\nabla l\|_F \geq 2\tau^{1/2}\|\nabla l\|_F$. By Lemma 3 and the Cauchy-Schwarz inequality, we have that

$$\|\nabla l\|_F \geq \frac{1}{2K}\|\nabla l\|_F \|(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F \geq \frac{1}{2K}\langle \nabla l, (\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\rangle = \frac{2l}{K}.$$

Then $\|\nabla R\|_F \geq 4\tau/K$. Since $\|\nabla f\|_F^2 = \|\lambda \nabla R\|_F^2 + \|\nabla L\|_F^2$, we are done. □

4.5 Removing Extraneous Directions

We show that if the projection $\boldsymbol{A}_3$ in the incorrect subspace is large, the gradient must be large so the point cannot be a local minimum.

Lemma 14. *Let $\gamma = \Theta^*(\tau^{1/48})$. If $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) < \tau$ and $\|\boldsymbol{A}_3\|_F \geq \gamma$, then*

$$\|\nabla f(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F = \Omega^*(\tau^{1/6}).$$

Proof. Set $\gamma = C\tau^{1/48}$, where we choose C to be a constant such that

$$\gamma^2 \geq \max\left(r(\tau^{1/24} + \tau^{1/4}), r^4(4K^6\tau^{1/8} + K^4\tau^{3/8})\right).$$

This particular definition allows us to simplify inequality (10) below. Consider the direction $\Delta \boldsymbol{A} = -\boldsymbol{A}_3$. When we step in this direction, the first-order perturbation of $L(\boldsymbol{\mathcal{S}}, \boldsymbol{A} + \epsilon\Delta\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ is $-2\epsilon\|\boldsymbol{\mathcal{S}}(\boldsymbol{A}_3, \boldsymbol{B}, \boldsymbol{C})\|_F^2$ (a simple calculation). For the regularizer, observe that since $\boldsymbol{A}_3 = \boldsymbol{A}(\boldsymbol{I} - \boldsymbol{P}_3)$, we have $(\boldsymbol{A} + \epsilon\Delta\boldsymbol{A})(\boldsymbol{A} + \epsilon\Delta\boldsymbol{A})^\top = \boldsymbol{A}\boldsymbol{A}^\top - (2\epsilon - \epsilon^2)\boldsymbol{A}_3\boldsymbol{A}_3^\top$. Hence the first-order perturbation of R is

$$8\epsilon\lambda l(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\langle \boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top, \boldsymbol{A}_3\boldsymbol{A}_3^\top\rangle \leq 8\epsilon\lambda\tau^{3/4}\|\boldsymbol{A}_3\|_F^2.$$

Intuitively, we will show that the first-order decrease in L is greater than the first-order increase in R, so that $\Delta\boldsymbol{A}$ is aligned negatively with $\nabla_{\boldsymbol{A}} f$.

First, through very similar arguments to those found in the proof of Lemmas 9 and 11, we have that

$$\|\boldsymbol{\mathcal{S}}(\boldsymbol{A}_3, \boldsymbol{B}, \boldsymbol{C})\|_F^2 \geq \|\boldsymbol{\mathcal{S}}(\boldsymbol{A}_3, \boldsymbol{\mathcal{S}}_{(2)}, \boldsymbol{\mathcal{S}}_{(3)})\|_F^2 - 2\tau^{1/4}K^6 - \tau^{1/2}K^4 \tag{7}$$

and if we set $\boldsymbol{u}$ to be the top left singular vector of $\boldsymbol{A}_3$,

$$\|\boldsymbol{\mathcal{S}}(\boldsymbol{A}_3, \boldsymbol{\mathcal{S}}_{(2)}, \boldsymbol{\mathcal{S}}_{(3)})\|_F \geq \frac{1}{r^2}\|\boldsymbol{\mathcal{S}}(\boldsymbol{u}, \boldsymbol{I}, \boldsymbol{I})\|_F^3\|\boldsymbol{A}_3\|_F \tag{8}$$

$$\begin{aligned}\|\boldsymbol{\mathcal{S}}(\boldsymbol{u}, \boldsymbol{I}, \boldsymbol{I})\|_F^2 &= \boldsymbol{u}^\top\boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top\boldsymbol{u} \\ &= \boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{u} + \boldsymbol{u}^\top(\boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)} - \boldsymbol{A}\boldsymbol{A}^\top)\boldsymbol{u} \\ &\geq \frac{1}{r}\|\boldsymbol{A}_3\|_F^2 - \tau^{1/4}. \end{aligned} \tag{9}$$

Combining inequalities (7), (8), and (9), we have

$$\|\boldsymbol{\mathcal{S}}(\boldsymbol{A}_3, \boldsymbol{B}, \boldsymbol{C})\|_F^2 \geq \frac{1}{r^4}\|\boldsymbol{A}_3\|_F^2\left(\frac{1}{r}\|\boldsymbol{A}_3\|_F^2 - \tau^{1/4}\right)^3 - 2\tau^{1/4}K^6 - \tau^{1/2}K^4 \tag{10}$$

Using the assumption $\|\boldsymbol{A}_3\|_F \geq \gamma$ and the choice of γ, we can simplify inequality (10) to $\|\boldsymbol{\mathcal{S}}(\boldsymbol{A}_3, \boldsymbol{B}, \boldsymbol{C})\|_F^2 \geq \frac{\tau^{1/8}}{2r^4}\|\boldsymbol{A}_3\|_F^2$. Now using the fact that $\lambda = 1/16r^4$ and $\tau^{1/8} > \tau^{3/4}$, we have $\frac{\tau^{1/8}}{2r^4}\|\boldsymbol{A}_3\|_F^2 > 8\lambda\tau^{3/4}\|\boldsymbol{A}_3\|_F^2$. Thus, we see that the first-order decrease in L is greater than the first-order increase in R, and the overall first-order decrease in f is $\Omega^*(\tau^{1/8}\|A_3\|_F^2/2r^4)$. The Taylor expansion of f implies that $|\langle\Delta\boldsymbol{A}, \nabla_{\boldsymbol{A}} f\rangle| \geq \frac{\tau^{1/8}}{2r^4}\|\boldsymbol{A}_3\|_F^2$, so that

$$\|\nabla f\|_F \geq |\langle\Delta\boldsymbol{A}, \nabla_{\boldsymbol{A}} f\rangle|/\|\Delta\boldsymbol{A}\|_F = \Omega^*(\tau^{1/8}\gamma) = \Omega^*(\tau^{1/6}),$$

which provides the desired bound on $\|\nabla f\|_F$. □

From this point forward, set $\sigma = \sqrt{\gamma}$. An important consequence of the fact that $\boldsymbol{A}_3$ is small is that if $\boldsymbol{\mathcal{T}}_{2,1,1}$ is large enough, then $\boldsymbol{A}_1$ must be rank deficient. This rank deficiency allows us to readily contruct a direction of improvement when we are near a saddle point corresponding to a single missing direction. This is also true when we are near saddle points corresponding to two or three missing directions; see section 4.7.

To prove the rank deficiency, we use subspace perturbation bounds. The technical tool we use here is Wedin's Theorem [33, 26].

Theorem (Wedin's Theorem, adapted from [26])**.** *Let $\tilde{\boldsymbol{A}}, \boldsymbol{A}, \boldsymbol{E} \in \mathbb{R}^{d\times r}$ with $d \geq r$ and $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{E}$. Write the singular value decompositions*

$$\boldsymbol{A} = \begin{pmatrix}\boldsymbol{U}_1 \; \boldsymbol{U}_2\end{pmatrix} \begin{pmatrix}\boldsymbol{\Sigma} \\ \boldsymbol{0}\end{pmatrix} \boldsymbol{V}^\top \qquad \tilde{\boldsymbol{A}} = \begin{pmatrix}\tilde{\boldsymbol{U}_1} \; \tilde{\boldsymbol{U}_2}\end{pmatrix} \begin{pmatrix}\tilde{\boldsymbol{\Sigma}} \\ \boldsymbol{0}\end{pmatrix} \tilde{\boldsymbol{V}}^\top$$

Let $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ denote the matrices of principal angles between the column spans of $\boldsymbol{U}_1, \tilde{\boldsymbol{U}}_1$ and $\boldsymbol{V}, \tilde{\boldsymbol{V}}$, respectively. If there exists some $\delta > 0$ such that $\min \sigma(\tilde{\boldsymbol{\Sigma}}) \geq \delta$, then

$$\sqrt{\|\sin \boldsymbol{\Theta}\|_F^2 + \|\sin \boldsymbol{\Phi}\|_F^2} \leq \frac{\sqrt{2}\|\boldsymbol{E}\|_F}{\delta}.$$

Lemma 15. *Let $\boldsymbol{M} \in \mathbb{R}^{r\times d}$, and let $\boldsymbol{M} = \boldsymbol{M}_1 + \boldsymbol{M}_2$, where rank$(\boldsymbol{M}) =$ rank$(\boldsymbol{M}_1) = r$. Let $\boldsymbol{P}, \boldsymbol{P}_1 \in \mathbb{R}^{d\times d}$ be the orthogonal projections onto the row spans of $\boldsymbol{M}$ and $\boldsymbol{M}_1$, respectively. Let σ be the smallest nontrivial singular value of $\boldsymbol{M}$. Then*

$$\|\boldsymbol{P} - \boldsymbol{P}_1\|_F \leq \frac{2\|\boldsymbol{M}_2\|_F}{\sigma}.$$

Proof. This is a corollary of Wedin's Theorem. Set $\boldsymbol{A} = \boldsymbol{M}_1^\top$, $\tilde{\boldsymbol{A}} = \boldsymbol{M}^\top$, and $\boldsymbol{E} = \boldsymbol{M}_2^\top$. Note that $\boldsymbol{A}$ and $\tilde{\boldsymbol{A}}$ have full row rank, so we have the SVDs

$$\boldsymbol{A} = \begin{pmatrix}\boldsymbol{U}_1 \; \boldsymbol{U}_2\end{pmatrix} \begin{pmatrix}\boldsymbol{\Sigma} \\ \boldsymbol{0}\end{pmatrix} \boldsymbol{V}^\top \qquad \tilde{\boldsymbol{A}} = \begin{pmatrix}\tilde{\boldsymbol{U}_1} \; \tilde{\boldsymbol{U}_2}\end{pmatrix} \begin{pmatrix}\tilde{\boldsymbol{\Sigma}} \\ \boldsymbol{0}\end{pmatrix} \tilde{\boldsymbol{V}}^\top$$

where $\boldsymbol{V}, \tilde{\boldsymbol{V}}$ are $r \times r$ orthogonal matrices, $\boldsymbol{\Sigma}, \tilde{\boldsymbol{\Sigma}}$ are $r \times r$, $\boldsymbol{U}_1, \tilde{\boldsymbol{U}}_1$ are $d \times r$, and $\boldsymbol{U}_2, \tilde{\boldsymbol{U}}_2$ are $d \times (d - r)$. Since $\boldsymbol{V}$ and $\tilde{\boldsymbol{V}}$ have the same column spans, we have that $\sin \boldsymbol{\Phi} = \boldsymbol{0}$. Further, it is a fact that $\|\boldsymbol{P} - \boldsymbol{P}_1\|_F = \sqrt{2}\|\sin \boldsymbol{\Theta}\|_F$. By assumption, $\min \sigma(\tilde{\boldsymbol{\Sigma}}) = \sigma$. Then Wedin's Theorem states that

$$\sqrt{\|\sin \boldsymbol{\Theta}\|_F^2 + \|\sin \boldsymbol{\Phi}\|_F^2} \leq \frac{\sqrt{2}\|\boldsymbol{E}\|_F}{\sigma},$$

and our result follows immediately. □

Lemma 16. *Let $\boldsymbol{P}$ be the orthogonal projection onto the row-span of $\boldsymbol{A}_1$. If rank$(\boldsymbol{A}_1) = r$ and $\|\boldsymbol{A}_3\|_F \leq \gamma$, then $\|\boldsymbol{\mathcal{T}}(\boldsymbol{I} - \boldsymbol{P}, \boldsymbol{I}, \boldsymbol{I})\|_F < 2K\sqrt{\gamma}$. In particular, if any of the $T_{1,j,k}$ $(j, k = 1, 2)$ is large, the rank of $\boldsymbol{A}_1$ must be less than r.*

Proof. Recall we set $\sigma = \sqrt{\gamma}$. Let $\boldsymbol{P}_1$ be the orthogonal projection onto the column-span of $\boldsymbol{\mathcal{T}}_{(1)}$. Write $\boldsymbol{A}_{1,1} = \boldsymbol{A}_1\boldsymbol{P}_1$, $\boldsymbol{A}_{1,2} = \boldsymbol{A}_1(\boldsymbol{I} - \boldsymbol{P}_1)$. Observe that $\|\boldsymbol{A}_{1,2}\|_F \leq \|\boldsymbol{A}_3\|_F \leq \gamma < \sigma$. Note that $\|\boldsymbol{A}_1 - \boldsymbol{A}_{1,1}\|_F = \|\boldsymbol{A}_{1,2}\|_F < \sigma$, which means that $\text{rank}(\boldsymbol{A}_{1,1}) = r$, since $\boldsymbol{A}_1$ has distance at least σ to the closest lower-rank matrix.

Since $\boldsymbol{A}_{1,1}$ has rank r, its rows form a basis for the column-span of $\boldsymbol{\mathcal{T}}_{(1)}$, and so $\boldsymbol{P}_1$ is also the orthogonal projection onto the row-span of $\boldsymbol{A}_{1,1}$. Then

$$\begin{aligned}\|\boldsymbol{\mathcal{T}}(\boldsymbol{I} - \boldsymbol{P}, \boldsymbol{I}, \boldsymbol{I})\|_F &= \|\boldsymbol{\mathcal{T}}(\boldsymbol{P}_1 - \boldsymbol{P}, \boldsymbol{I}, \boldsymbol{I})\|_F \\ &\leq \|\boldsymbol{\mathcal{T}}\|_F\|\boldsymbol{P}_1 - \boldsymbol{P}\|_F \\ &\leq K\frac{2\|\boldsymbol{A}_{1,2}\|_F}{\sigma} \\ &< 2K\sqrt{\gamma},\end{aligned}$$

where the penultimate line follows from Lemma 15. □

4.6 Improving S

Unlike the proof of Theorem 1, we will first focus on the simple case of improving the core tensor $\boldsymbol{\mathcal{S}}$. Note that here we only try to make sure we get close to $\boldsymbol{\mathcal{T}}_{1,1,1}$ as the components $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ may still be missing directions.

Lemma 17. *Set $\kappa_0 = \sqrt{\gamma}$. Assume $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) < \tau$. Then*

$$\|\boldsymbol{\mathcal{T}}_{1,1,1} - \boldsymbol{\mathcal{S}}(\boldsymbol{A}_1, \boldsymbol{B}_1, \boldsymbol{C}_1)\|_F > \kappa_0 \;\Rightarrow\; \|\nabla f(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F = \Omega(\gamma^{2.5}).$$

Proof. Define $\boldsymbol{\mathcal{S}}^* = \boldsymbol{\mathcal{T}}(\boldsymbol{A}_1^+, \boldsymbol{B}_1^+, \boldsymbol{C}_1^+)$, so that $\boldsymbol{\mathcal{S}}^*(\boldsymbol{A}_1, \boldsymbol{B}_1, \boldsymbol{C}_1) = \boldsymbol{\mathcal{T}}_{1,1,1}$. We consider the direction $\Delta\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{S}}^* - \boldsymbol{\mathcal{S}}_{1,1,1}$. Observe that $\Delta\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \boldsymbol{\mathcal{T}}_{1,1,1} - \boldsymbol{\mathcal{S}}(\boldsymbol{A}_1, \boldsymbol{B}_1, \boldsymbol{C}_1)$. We can write

$$\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}} = \sum_{i,j,k \in \{1,2\}} \boldsymbol{\mathcal{S}}(\boldsymbol{A}_i, \boldsymbol{B}_j, \boldsymbol{C}_k) - \boldsymbol{\mathcal{T}}_{i,j,k},$$

and this is a sum of mutually orthogonal tensors. Hence, the the first-order perturbation of $L(\boldsymbol{\mathcal{S}} + \epsilon\Delta\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ is

$$2\langle \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}}, \Delta\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\rangle = -2\|\Delta\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F^2.$$

The first-order perturbation in the regularizer $\langle \nabla_{\boldsymbol{\mathcal{S}}} R, \Delta\boldsymbol{\mathcal{S}}\rangle$ is bounded by $O(\tau^{3/4}\sigma^{-3}) = o(\sigma)$, since $\|\boldsymbol{\mathcal{S}}^*\|_F = O(\sigma^{-3})$. Therefore, the decrease in the tensor loss dominates all other first-order perturbations, so we have a viable direction of improvement. In particular, by moving in direction $\epsilon\Delta\boldsymbol{\mathcal{S}}$, we decrease the objective function by

$$\epsilon \cdot \Omega(\|\boldsymbol{\mathcal{T}}_{1,1,1} - \boldsymbol{\mathcal{S}}(\boldsymbol{A}_1, \boldsymbol{B}_2, \boldsymbol{C}_1)\|_F^2) = \Omega(\epsilon\kappa_0^2).$$

The direction of movement has norm bounded by

$$\|\boldsymbol{\mathcal{T}}_{1,1,1}\|_F\|\boldsymbol{A}_1^+\|_2\|\boldsymbol{B}_1^+\|_2\|\boldsymbol{C}_1^+\|_2 \leq K\sigma^{-3}.$$

By Cauchy-Schwarz, the gradient has norm at least $\Omega(\kappa_0^2) \times \sigma^3 = \Omega(\gamma^{5/2})$. □

4.7 Adding missing directions

Finally, we try to add missing directions to $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$. As before we separate the cases into missing 1, 2 and 3 directions. This first case (missing one direction) is easy as it is a normal saddle point with negative Hessian.

Lemma 18. *Set $\kappa_1 = 2K\sigma^{3/4}$. Assume $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) < \tau$ and $\|\boldsymbol{A}_3\|_F$, $\|\boldsymbol{B}_3\|_F$, and $\|\boldsymbol{C}_3\|_F$ are all less than γ. If $\|\boldsymbol{\mathcal{T}}_{2,1,1}\|_2 \geq \kappa_1$, then $\nabla^2 f$ has a negative eigenvalue of at most $-\Omega(\sigma^{15/4})$.*

Proof. Since $\kappa_1 > 2K\sqrt{\gamma}$, by Lemma 16, we have $\operatorname{rank}(\boldsymbol{A}_1) < r$.

By assumption, there exist unit vectors $\boldsymbol{a} \in U_{1,2}$, $\boldsymbol{b} \in U_{2,1}$, and $\boldsymbol{c} \in U_{3,1}$ such that $\boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \geq \kappa_1$. Take unit vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^r$ such that $\boldsymbol{A}_1^\top \boldsymbol{u} = \boldsymbol{0}$, $\boldsymbol{B}_1^\top \boldsymbol{v} = \alpha_1 \boldsymbol{b}$, $\boldsymbol{B}_2^\top \boldsymbol{v} = \boldsymbol{0}$, $\boldsymbol{C}_1^\top \boldsymbol{w} = \alpha_2 \boldsymbol{c}$, and $\boldsymbol{C}_2^\top \boldsymbol{w} = \boldsymbol{0}$, where $\alpha_i \geq \sigma$ for $i = 1, 2$. In this situation, we are near a second-order saddle point, so we seek to demonstrate a direction with sufficient negative curvature in the objective function. To this end, define

$$\Delta \boldsymbol{A} = \sigma \boldsymbol{u} \boldsymbol{a}^\top \qquad \Delta \boldsymbol{\mathcal{S}} = \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}.$$

For a step size $\epsilon > 0$, our source of improvement in the tensor loss comes from the second-order perturbation of L in this direction. We aim to compare the second-order decrease in L against the second-order increases in L and R. The second-order perturbation in the tensor loss $\nabla^2 L$ applied to $(\Delta \boldsymbol{\mathcal{S}}, \Delta \boldsymbol{A}, \boldsymbol{0}, \boldsymbol{0})$ is

$$2\langle \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}}, \Delta \boldsymbol{\mathcal{S}}(\Delta \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \rangle + \|\Delta \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) + \boldsymbol{\mathcal{S}}(\Delta \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|^2$$

The magnitude of *decrease* in this perturbation is given by

$$\begin{aligned} \langle \boldsymbol{\mathcal{T}}, \Delta \boldsymbol{\mathcal{S}}(\Delta \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \rangle &= \sigma \boldsymbol{\mathcal{T}}(\boldsymbol{a}, \alpha_1 \boldsymbol{b}, \alpha_2 \boldsymbol{c}) \\ &= \sigma \alpha_1 \alpha_2 \boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \\ &\geq \sigma \alpha_1 \alpha_2 \kappa_1. \end{aligned}$$

To bound the magnitude of the *increase*, observe that

$$\|\boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{v}\|_F = \|\alpha_1 \boldsymbol{B}_1 \boldsymbol{b}\|_F \leq \alpha_1 K; \quad \|\boldsymbol{C}\boldsymbol{C}^\top \boldsymbol{w}\|_F \leq \alpha_2 K$$

Then we have

$$\begin{aligned} \langle \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}), \Delta \boldsymbol{\mathcal{S}}(\Delta \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \rangle &= \langle \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}), \sigma \boldsymbol{a} \otimes \boldsymbol{B}^\top \boldsymbol{v} \otimes \boldsymbol{C}^\top \boldsymbol{w} \rangle \\ &= \sigma \boldsymbol{\mathcal{S}}(\boldsymbol{A}\boldsymbol{a}, \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{v}, \boldsymbol{C}\boldsymbol{C}^\top \boldsymbol{w}) \\ &\leq \sigma^2 \alpha_1 \alpha_2 K^3 \end{aligned} \tag{11}$$

Additionally,

$$\begin{aligned}
\|\Delta\boldsymbol{\mathcal{S}}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C})\|_F^2 &= \|\boldsymbol{A}_2^\top \boldsymbol{u}\otimes\alpha_1\boldsymbol{b}\otimes\alpha_2\boldsymbol{c}\|_F^2\\
&\le \sigma^2\alpha_1^2\alpha_2^2\\
\|\boldsymbol{\mathcal{S}}_{(1)}^\top\boldsymbol{u}\|_F^2 &= \boldsymbol{u}^\top(\boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top-\boldsymbol{A}\boldsymbol{A}^\top)\boldsymbol{u}+\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{u}\\
&\le \tau^{1/4}+\sigma^2\\
\|\boldsymbol{\mathcal{S}}(\Delta\boldsymbol{A},\boldsymbol{B},\boldsymbol{C})\|_F^2 &= \sigma^2\|\boldsymbol{a}\boldsymbol{u}^\top\boldsymbol{\mathcal{S}}_{(1)}(\boldsymbol{B}\otimes\boldsymbol{C})\|_F^2\\
&\le \sigma^2\|\boldsymbol{\mathcal{S}}_{(1)}^\top\boldsymbol{u}\|_F^2\|\boldsymbol{B}\|_F^2\|\boldsymbol{C}\|_F^2\\
&\le \sigma^2K^4(\tau^{1/4}+\sigma^2)
\end{aligned}$$

Putting this together, we bound $\|\Delta\boldsymbol{\mathcal{S}}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C})+\boldsymbol{\mathcal{S}}(\Delta\boldsymbol{A},\boldsymbol{B},\boldsymbol{C})\|_F^2$ above by

$$\left(\sigma\alpha_1\alpha_2+\sigma K^2\sqrt{\tau^{1/4}+\sigma^2}\right)^2 \tag{12}$$

In light of the definition of κ_1 and inequalities (11) and (12), the second-order perturbation in L is $-\Omega(\sigma\alpha_1\alpha_2\kappa_1)$, i.e. L decreases to second-order in this direction.

Now we turn our attention to the regularizer. We need to show that the second-order increase in the regularizer doesn't overwhelm the decrease in L. Note that the regularizer is degree 4 with respect to $\|\boldsymbol{A}\boldsymbol{A}^\top-\boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top\|_F$ (and same terms for $\boldsymbol{B}$ and $\boldsymbol{C}$), so the second order derivatives have a quadratic term in $\|\boldsymbol{A}\boldsymbol{A}^\top-\boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top\|_F$, which is $O(\tau^{1/4})=o(\sigma\alpha_1\alpha_2\kappa_1)$; higher-order terms are negligible in comparison.

We've shown that the loss function decreases by at least $\Omega(\sigma\alpha_1\alpha_2\kappa_1)\cdot\epsilon^2$. Since our direction of improvement has constant norm, this implies that $\nabla^2 f$ has an eigenvalue that is smaller than $-\Omega(\sigma\alpha_1\alpha_2\kappa_1)=-\Omega(\sigma^{15/4})$. □

Next, we need to deal with the high order saddle points. Here our main observation is that if we choose directions randomly in the correct subspace, then the perturbation is going to have a reasonable correlation with the residual tensor with constant probability. This is captured by the following anti-concentration property:

Lemma 19. *Let $\boldsymbol{\mathcal{X}}\in\mathbb{R}^{d_1\times d_2\times d_3}$, and let $\boldsymbol{a}\in\mathbb{R}^{d_1},\boldsymbol{b}\in\mathbb{R}^{d_2},\boldsymbol{c}\in\mathbb{R}^{d_3}$ be independent, uniformly distributed unit vectors. There exist positive numbers $C_1=\Omega(1/\sqrt{d_1d_2d_3})=\Omega^*(1), C_2=\Omega(1)$ such that*

$$Pr[|\boldsymbol{\mathcal{X}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})|\ge C_1\|\boldsymbol{\mathcal{X}}\|_F]>C_2.$$

Proof. Although our Algorithm 1 for sampling missing directions only requires uniform unit vectors (from appropriate subspaces), we construct these vectors as normalized Gaussian vectors for this lemma in order to apply a Gaussian polynomial anti-concentration result (Theorem 8 in [4]). As such, let $\boldsymbol{a}',\boldsymbol{b}',\boldsymbol{c}'$ be independent standard Gaussian random vectors (of appropriate dimension) and set $\boldsymbol{a}=\boldsymbol{a}'/\|\boldsymbol{a}'\|_2$, $\boldsymbol{b}=\boldsymbol{b}'/\|\boldsymbol{b}'\|_2$, and $\boldsymbol{c}=\boldsymbol{c}'/\|\boldsymbol{c}'\|_2$. Note that there exists

some constant $p > 0$ such that $\|\boldsymbol{a}'\|_2 \leq 2\sqrt{d_1}$, $\|\boldsymbol{b}'\|_2 \leq 2\sqrt{d_2}$, and $\|\boldsymbol{c}'\| \leq 2\sqrt{d_3}$ with probability at least p. Next, note that $\mathbb{E}\boldsymbol{\mathcal{X}}(\boldsymbol{a}', \boldsymbol{b}', \boldsymbol{c}') = 0$ and

$$\begin{aligned}\mathrm{Var}[\boldsymbol{\mathcal{X}}(\boldsymbol{a}', \boldsymbol{b}', \boldsymbol{c}')] &= \mathbb{E}(\boldsymbol{\mathcal{X}}(\boldsymbol{a}', \boldsymbol{b}', \boldsymbol{c}')^2) \\ &= \mathbb{E}\langle \boldsymbol{\mathcal{X}}(\boldsymbol{a}'\boldsymbol{a}'^\top, \boldsymbol{b}'\boldsymbol{b}'^\top, \boldsymbol{c}'\boldsymbol{c}'^\top), \boldsymbol{\mathcal{X}}\rangle \\ &= \langle \boldsymbol{\mathcal{X}}(\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{I}), \boldsymbol{\mathcal{X}}\rangle \\ &= \|\boldsymbol{\mathcal{X}}\|_F^2.\end{aligned}$$

Now $\boldsymbol{\mathcal{X}}(\boldsymbol{a}', \boldsymbol{b}', \boldsymbol{c}')/\|\boldsymbol{\mathcal{X}}\|_F$ is a degree three polynomial function with unit variance, so the anti-concentration inequality implies that for any $\epsilon > 0$,

$$\Pr[|\boldsymbol{\mathcal{X}}(\boldsymbol{a}', \boldsymbol{b}', \boldsymbol{c}')/\|\boldsymbol{\mathcal{X}}\|_F| \leq \epsilon] \leq O(1)\epsilon^{1/3}.$$

Simply choosing a constant ϵ and re-arranging terms completes the proof. □

Using this idea, when $\boldsymbol{\mathcal{T}}_{2,2,1}$ is large, we show how to get a direction of improvement.

Lemma 20. *Set $\kappa_2 = 2K\sigma^{1/8}$. Assume $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) < \tau$ and $\|\boldsymbol{A}_3\|_F$, $\|\boldsymbol{B}_3\|_F$, and $\|\boldsymbol{C}_3\|_F$ are all less than γ. Further assume that $\|\boldsymbol{\mathcal{T}}_{2,1,1}\|_2$, $\|\boldsymbol{\mathcal{T}}_{1,2,1}\|_2$, and $\|\boldsymbol{\mathcal{T}}_{1,1,2}\|_2$ are each less than κ_1. Let $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ be the output of Algorithm 1 given the input $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \sigma, (2, 2, 1)$. Define the directions $\Delta\boldsymbol{A} = \boldsymbol{u}\boldsymbol{a}^\top$, $\Delta\boldsymbol{B} = \boldsymbol{v}\boldsymbol{b}^\top$, $\Delta\boldsymbol{\mathcal{S}} = \boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w}$. If $\|\boldsymbol{\mathcal{T}}_{2,2,1}\|_2 \geq \kappa_2$, then with constant probability, a step in these directions decreases the objective function by $\Omega^*(\sigma^{15/8})$.*

Proof. First, observe that $\kappa_2 \geq 2K\sqrt{\gamma}$, which implies that $\mathrm{rank}(\boldsymbol{A}_1) < r$ and $\mathrm{rank}(\boldsymbol{B}_1) < r$ by Lemma 16. Set α such that $\alpha\boldsymbol{c} = \boldsymbol{C}_1^\top\boldsymbol{w}$, and note that $\alpha \geq \sigma$. Per lemma 19, with constant probability we have $|\boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})|$ is with some constant factor of $\|\boldsymbol{\mathcal{T}}_{2,2,1}\|_2$. Therefore, with constant probability, $|\boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})| \geq C\kappa_2$ for some positive constant C.

Observe that $p(\delta) := f(\boldsymbol{\mathcal{S}}+\delta\Delta\boldsymbol{\mathcal{S}}, \boldsymbol{A}+\delta\Delta\boldsymbol{A}, \boldsymbol{B}+\delta\Delta\boldsymbol{B}, \boldsymbol{C})$ defines a degree 8 polynomial in δ. Set $\delta = \sigma^{1/4}$. For convenience, define the following expressions related to the perturbations of L:

$$\begin{aligned}L_0 &= \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}} \\ L_1 &= \Delta\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) + \boldsymbol{\mathcal{S}}(\Delta\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) + \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \Delta\boldsymbol{B}, \boldsymbol{C}) \\ L_2 &= \Delta\boldsymbol{\mathcal{S}}(\Delta\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) + \Delta\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \Delta\boldsymbol{B}, \boldsymbol{C}) + \boldsymbol{\mathcal{S}}(\Delta\boldsymbol{A}, \Delta\boldsymbol{B}, \boldsymbol{C}) \\ L_3 &= \Delta\boldsymbol{\mathcal{S}}(\Delta\boldsymbol{A}, \Delta\boldsymbol{B}, \boldsymbol{C})\end{aligned}$$

We can upper bound each of these terms in norm, e.g.

$$\begin{aligned}\|L_1\|_F &= \|\boldsymbol{A}^\top\boldsymbol{u} \otimes \boldsymbol{B}^\top\boldsymbol{v} \otimes \boldsymbol{C}^\top\boldsymbol{w} + \boldsymbol{\mathcal{S}}(\boldsymbol{u}\boldsymbol{a}^\top, \boldsymbol{B}, \boldsymbol{C}) + \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{v}\boldsymbol{b}^\top, \boldsymbol{C})\|_F \\ &\leq \sigma^2\alpha + 2K^2\sqrt{\tau^{1/4} + \sigma^2} \\ &= O(\alpha\sigma^2 + \sigma).\end{aligned}$$

Through similar calculations, we have $\|L_2\|_F = O(\alpha\sigma+\sigma)$ and $\|L_3\|_F = O(\alpha)$. The perturbation in the tensor loss is then

$$\delta^3\langle L_0, L_3\rangle + \delta\langle L_0, L_1\rangle + \delta^2\langle L_0, L_2\rangle + \|\delta L_1 + \delta^2 L_2 + \delta^3 L_3\|_F^2. \tag{13}$$

Here the first term is responsible for the decrease in tensor loss:

$$\begin{aligned}\delta^3\langle L_0, L_3\rangle &= \delta^3\alpha\langle \boldsymbol{\mathcal{S}}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}) - \boldsymbol{\mathcal{T}}, \boldsymbol{a}\otimes\boldsymbol{b}\otimes\boldsymbol{c}\rangle \\ &\le \delta^3\alpha(K^2\sigma^2 - \boldsymbol{\mathcal{T}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})) \\ &= -\delta^3\alpha\Omega(\kappa_2).\end{aligned}$$

For the other terms, we show that they are small enough so they will not cancel this improvement. Observe that

$$\begin{aligned}\langle L_0, L_1\rangle &= \langle L_0, \Delta\boldsymbol{\mathcal{S}}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C})\rangle + \langle L_0, \boldsymbol{\mathcal{S}}(\boldsymbol{u}\boldsymbol{a}^\top,\boldsymbol{B},\boldsymbol{C}) + \boldsymbol{\mathcal{S}}(\boldsymbol{A},\boldsymbol{v}\boldsymbol{b}^\top,\boldsymbol{C})\rangle \\ &= O(\alpha\sigma^2) + O(\kappa_1\sigma).\end{aligned}$$

The $O(\kappa_1\sigma)$ term appears because $\|\boldsymbol{\mathcal{S}}_{2,1,1}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}) - \boldsymbol{\mathcal{T}}_{2,1,1}\|_F = O(\kappa_1)$, $\|\boldsymbol{\mathcal{S}}_{1,2,1}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}) - \boldsymbol{\mathcal{T}}_{1,2,1}\|_F = O(\kappa_1)$.

For the next term, we note that $\langle L_0, L_2\rangle$ is a sum of three inner products, any two of which we can make nonpositive by flipping the sign of $\Delta\boldsymbol{\mathcal{S}}$ and one of $\Delta\boldsymbol{A}$, $\Delta\boldsymbol{B}$ (doing so doesn't change the amount by which the tensor loss decreases). Hence, by design of Algorithm 1, with constant probability we know that $\langle L_0, L_2\rangle \le 0$. As a result, we know (13) is at most $-\delta^3\alpha\Omega(\kappa_2)$.

We now consider the perturbations of the regularizer. Define the following terms:

$$\begin{aligned}l_{0,1} &= \boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top, \quad l_{0,2} = \boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{\mathcal{S}}_{(2)}\boldsymbol{\mathcal{S}}_{(2)}^\top, \quad l_{0,3} = \boldsymbol{C}\boldsymbol{C}^\top - \boldsymbol{\mathcal{S}}_{(3)}\boldsymbol{\mathcal{S}}_{(3)}^\top \\ l_{1,1} &= \boldsymbol{A}\Delta\boldsymbol{A}^\top + \Delta\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{\mathcal{S}}_{(1)}(\Delta\boldsymbol{\mathcal{S}})_{(1)}^\top - (\Delta\boldsymbol{\mathcal{S}})_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top \\ l_{1,2} &= \boldsymbol{B}\Delta\boldsymbol{B}^\top + \Delta\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{\mathcal{S}}_{(2)}(\Delta\boldsymbol{\mathcal{S}})_{2)}^\top - (\Delta\boldsymbol{\mathcal{S}})_{(2)}\boldsymbol{\mathcal{S}}_{(2)}^\top \\ l_{1,3} &= -\boldsymbol{\mathcal{S}}_{(3)}(\Delta\boldsymbol{\mathcal{S}})_{(3)}^\top - (\Delta\boldsymbol{\mathcal{S}})_{(3)}\boldsymbol{\mathcal{S}}_{(3)}^\top, \quad l_{2,3} = -(\Delta\boldsymbol{\mathcal{S}})_{(3)}(\Delta\boldsymbol{\mathcal{S}})_{(3)}^\top\end{aligned}$$

We bound these terms in norm as follows:

$$\begin{aligned}\|l_{1,1}\|_F &= \|\boldsymbol{A}\boldsymbol{a}\boldsymbol{u}^\top + \boldsymbol{u}\boldsymbol{a}^\top\boldsymbol{A}^\top - \boldsymbol{u}\boldsymbol{\mathcal{S}}(\boldsymbol{I},\boldsymbol{v},\boldsymbol{w})^\top - \boldsymbol{\mathcal{S}}(\boldsymbol{I},\boldsymbol{v},\boldsymbol{w})\boldsymbol{u}^\top\|_F \\ &\le 2\|\boldsymbol{A}_2\|_F + 2\|\boldsymbol{\mathcal{S}}(\boldsymbol{I},\boldsymbol{v},\boldsymbol{w})\|_F \\ &\le 2\sigma + 2\sqrt{\tau^{1/4}+\sigma^2} \\ &= O(\sigma).\end{aligned}$$

Likewise, $\|l_{1,2}\|_F = O(\sigma)$, $\|l_{1,3}\|_F = O(\sigma)$, and $\|l_{2,3}\| = O(1)$. Also note that $\|l_{0,i}\|_F \le \tau^{1/4}$ for $i = 1, 2, 3$. Using this, we have

$$\begin{aligned}\|l_{0,1} + \delta l_{1,1}\|_F &\le O(\tau^{1/4}) + O(\delta\sigma) \\ \|l_{0,2} + \delta l_{1,2}\|_F &\le O(\tau^{1/4}) + O(\delta\sigma) \\ \|l_{0,3} + \delta l_{1,3} + \delta^2 l_{2,3}\|_F &\le O(\tau^{1/4}) + O(\delta\sigma) + O(\delta^2).\end{aligned}$$

All of these terms are dominated by $O(\delta^2)$, and so the perturbed regularizer is bounded by $O(\delta^8) = O(\sigma^2) = o(\sigma^{15/8})$. Hence, we see that the decrease in the tensor loss dominates the increase in the regularizer, as desired. □

We next address the case where $\boldsymbol{\mathcal{T}}_{2,2,2}$ is large, which corresponds to $\boldsymbol{A}_1, \boldsymbol{B}_1, \boldsymbol{C}_1$ being rank deficient.

Lemma 21. *Set $\kappa_3 = 2K\sigma^{1/2}$. Assume $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) < \tau$ and $\|\boldsymbol{A}_3\|_F$, $\|\boldsymbol{B}_3\|_F$, and $\|\boldsymbol{C}_3\|_F$ are all less than γ. Further assume that $\|\boldsymbol{\mathcal{T}}_{2,1,1}\|_2$, $\|\boldsymbol{\mathcal{T}}_{1,2,1}\|_2$, and $\|\boldsymbol{\mathcal{T}}_{1,1,2}\|_2$ and each less than κ_1, while $\|\boldsymbol{\mathcal{T}}_{2,2,1}\|_2$, $\|\boldsymbol{\mathcal{T}}_{2,1,2}\|_2$, and $\|\boldsymbol{\mathcal{T}}_{1,2,2}\|_2$ are each less than κ_2. Let $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ be the output of Algorithm 1 with input $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \sigma, (2, 2, 2)$. Define the directions $\Delta\boldsymbol{A} = \boldsymbol{u}\boldsymbol{a}^\top$, $\Delta\boldsymbol{B} = \boldsymbol{v}\boldsymbol{b}^\top$, $\Delta\boldsymbol{C} = \boldsymbol{w}\boldsymbol{c}^\top$, $\Delta\boldsymbol{\mathcal{S}} = \boldsymbol{u}\otimes\boldsymbol{v}\otimes\boldsymbol{w}$. If $\|\boldsymbol{\mathcal{T}}_{2,2,2}\|_2 \geq \kappa_3$, then with constant probability, a step in these directions decreases the objective function by $\Omega^*(\sigma^{3/4})$.*

Proof. First observe that $\kappa_3 \geq 2K\sqrt{\gamma}$, which by Lemma 16 means that $\operatorname{rank}(\boldsymbol{A}_1)$, $\operatorname{rank}(\boldsymbol{B}_1)$, and $\operatorname{rank}(\boldsymbol{C}_1)$ are all strictly less than r. Then $\boldsymbol{A}_1$, $\boldsymbol{B}_1$, and $\boldsymbol{C}_1$ are all missing directions from the relevant subspaces of $\boldsymbol{\mathcal{T}}$, and we are near a fourth-order saddle point. By lemma 19, with constant probability, $|\boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})| > C\kappa_3$ for some positive constant C.

Again let $p(\delta) = f(\boldsymbol{\mathcal{S}} + \delta\Delta\boldsymbol{\mathcal{S}}, \boldsymbol{A} + \delta\Delta\boldsymbol{A}, \boldsymbol{B} + \delta\Delta\boldsymbol{B}, \boldsymbol{C} + \delta\Delta\boldsymbol{C})$, and set $\delta = \sigma^{1/8}$. As in the proof of lemma 20, let for $i = 0, \ldots, 4$, let L_i denote the i-th order perturbation term in

$$(\boldsymbol{\mathcal{S}} + \Delta\boldsymbol{\mathcal{S}})(\boldsymbol{A} + \Delta\boldsymbol{A}, \boldsymbol{B} + \Delta\boldsymbol{B}, \boldsymbol{C} + \Delta\boldsymbol{C}) - \boldsymbol{\mathcal{T}}.$$

We can upper bound each of these terms in norm, e.g.

$$\begin{aligned}
\|L_1\|_F &= \|\boldsymbol{A}^\top\boldsymbol{u} \otimes \boldsymbol{B}^\top\boldsymbol{v} \otimes \boldsymbol{C}^\top\boldsymbol{w} + \boldsymbol{\mathcal{S}}(\boldsymbol{u}\boldsymbol{a}^\top, \boldsymbol{B}, \boldsymbol{C}) + \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{v}\boldsymbol{b}^\top, \boldsymbol{C}) \\
&\quad + \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{w}\boldsymbol{c}^\top)\|_F \\
&\leq 8\sigma^3 + K^2(\|\boldsymbol{\mathcal{S}}(\boldsymbol{u}, \boldsymbol{I}, \boldsymbol{I})\|_F + \|\boldsymbol{\mathcal{S}}(\boldsymbol{I}, \boldsymbol{v}, \boldsymbol{I})\|_F + \|\boldsymbol{\mathcal{S}}(\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{w})\|_F) \\
&\leq 8\sigma^3 + 3K^2\sqrt{\tau^{1/4} + 2\sigma^2} \\
&= O(\sigma).
\end{aligned}$$

Through similar calculations, we have $\|L_2\| = O(\sigma)$ and $\|L_3\| = O(\sigma)$. On the other hand, $\|L_4\| \leq 1$ and $\|L_0\| \leq 2K$. The perturbation in the tensor loss is then

$$\sum_{i,j=0}^{4} \langle L_i, L_j\rangle\delta^{i+j} \tag{14}$$

The decrease in the tensor loss is due to the following term:

$$\begin{aligned}
\delta^4\langle L_0, L_4\rangle &= \delta^4\langle\boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) - \boldsymbol{\mathcal{T}}, \boldsymbol{a}\otimes\boldsymbol{b}\otimes\boldsymbol{c}\rangle \\
&\leq \delta^4(K\sigma^3 - \boldsymbol{\mathcal{T}}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})) \\
&= -\delta^4\Omega(\kappa_3) \\
&= -\Omega(\sigma^{3/4}).
\end{aligned}$$

By a simple Cauchy-Schwarz bound, the other perturbation terms in (14) are all bounded by $O(\sigma + \delta^8) = O(\sigma) = o(\sigma^{3/4})$.

Now we analyze the perturbations of the regularizer. As before, define the terms

$$l_{0,1} = \boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{\mathcal{S}}_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top, \quad l_{0,2} = \boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{\mathcal{S}}_{(2)}\boldsymbol{\mathcal{S}}_{(2)}^\top, \quad l_{0,3} = \boldsymbol{C}\boldsymbol{C}^\top - \boldsymbol{\mathcal{S}}_{(3)}\boldsymbol{\mathcal{S}}_{(3)}^\top$$
$$l_{1,1} = \boldsymbol{A}\Delta\boldsymbol{A}^\top + \Delta\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{\mathcal{S}}_{(1)}(\Delta\boldsymbol{\mathcal{S}})_{(1)}^\top - (\Delta\boldsymbol{\mathcal{S}})_{(1)}\boldsymbol{\mathcal{S}}_{(1)}^\top$$
$$l_{1,2} = \boldsymbol{B}\Delta\boldsymbol{B}^\top + \Delta\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{\mathcal{S}}_{(2)}(\Delta\boldsymbol{\mathcal{S}})_{2)}^\top - (\Delta\boldsymbol{\mathcal{S}})_{(2)}\boldsymbol{\mathcal{S}}_{(2)}^\top$$
$$l_{1,3} = \boldsymbol{C}\Delta\boldsymbol{C}^\top + \Delta\boldsymbol{C}\boldsymbol{C}^\top - \boldsymbol{\mathcal{S}}_{(3)}(\Delta\boldsymbol{\mathcal{S}})_{(3)}^\top - (\Delta\boldsymbol{\mathcal{S}})_{(3)}\boldsymbol{\mathcal{S}}_{(3)}^\top$$

We bound these terms in norm as follows:

$$\begin{aligned}\|l_{1,1}\|_F &= \|\boldsymbol{A}\boldsymbol{a}\boldsymbol{u}^\top + \boldsymbol{u}\boldsymbol{a}^\top\boldsymbol{A}^\top - \boldsymbol{u}\boldsymbol{\mathcal{S}}(\boldsymbol{I},\boldsymbol{v},\boldsymbol{w})^\top - \boldsymbol{\mathcal{S}}(\boldsymbol{I},\boldsymbol{v},\boldsymbol{w})\boldsymbol{u}^\top\|_F \\ &\le 2\|\boldsymbol{A}_2\|_F + 2\|\boldsymbol{\mathcal{S}}(\boldsymbol{I},\boldsymbol{v},\boldsymbol{w})\|_F \\ &\le 2\sigma + 2\sqrt{\tau^{1/4} + \sigma^2} \\ &= O(\sigma).\end{aligned}$$

Likewise, $\|l_{1,i}\|_F = O(\sigma)$ for $i = 2, 3$, and of course $\|l_{0,i}\|_F \le \tau^{1/4}$ for $i = 1, 2, 3$. Again,

$$\|l_{0,i} + \delta l_{1,i}\|_F \le O(\tau^{1/4}) + O(\delta\sigma)$$

and using this, we can bound the perturbed regularizer as $O(\delta^4\sigma^4) = o(\sigma^{3/4})$. Hence, the decrease in the tensor loss dominates all other perturbations, and we improve the objective function by $\Omega(\sigma^{3/4})$. □

4.8 Algorithm Description and Proof of Main Theorem

Before sketching the algorithm we will first prove Lemma 12 and explain some of the parameter choices. Refer to Table 1 for a list of the numerical quantities that were introduced.

Proof of Lemma 12. Set τ small enough so that κ_0, $d\kappa_1 + K^3\sigma$, $d\kappa_2 + K^2\sigma^2$, $d\kappa_3 + K\sigma^3 < \sqrt{\epsilon}/4$ and $\tau < \epsilon/2$. We then set $\tau_1 = \Theta^*(\tau)$ from Lemma 13 and $\tau_2 = \Theta(\sigma^{15/4})$ from Lemma 18.

Now assume that conditions (1), (2), and (3) from the statement of the Lemma fail to hold. We seek to show that $f(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) < \epsilon$. By Lemma 13 and our choice of τ_1, we have that $R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) < \tau \le \epsilon/2$. By Lemma 14, we have that $\|\boldsymbol{A}_3\|_F, \|\boldsymbol{B}_3\|_F, \|\boldsymbol{C}_3\|_F$ are all less than γ. By Lemma 17, we have that $\|\boldsymbol{\mathcal{T}}_{1,1,1} - \boldsymbol{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\|_F \le \kappa_0$. By Lemma 18, we have that $\|\boldsymbol{\mathcal{T}}_{i,j,k}\|_2 < \kappa_1$ for $(i,j,k) \in \{(2,1,1), (1,2,1), (1,1,2)\}$. By Lemma 20, we have that $\|\boldsymbol{\mathcal{T}}_{i,j,k}\|_2 < \kappa_2$ for $(i,j,k) \in \{(2,2,1), (2,1,2), (1,2,2)\}$. By Lemma 21, we have that $\|\boldsymbol{\mathcal{T}}_{2,2,2}\|_2 < \kappa_3$.

Combining all of these bounds, we have

$$\begin{aligned} f(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) &= R(\boldsymbol{\mathcal{S}}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) + \sum_{i,j,k} \|\boldsymbol{\mathcal{S}}_{i,j,k}(\boldsymbol{A}_i, \boldsymbol{B}_j, \boldsymbol{C}_k) - \boldsymbol{\mathcal{T}}_{i,j,k}\|_F^2 \\ &< \epsilon/2 + \kappa_0^2 + 3(K^3\sigma + d\kappa_1)^2 + 3(K^2\sigma^2 + d\kappa_2)^2 + (K\sigma^3 + d\kappa_3)^2 \\ &< \epsilon/2 + \epsilon/2, \end{aligned}$$

as desired.

□

We now sketch our algorithm in Algorithm 2. The algorithm basically tries to follow the main Lemma 12. If the point has large gradient or negative eigenvalue in Hessian, we can just use any standard local search algorithm. When the point is a higher order saddle point, we use Algorithm 1 as in Lemma 20 or Lemma 21 to generate directions of improvements.

Algorithm 2 Local search algorithm for Tucker decomposition

Require: tensor $\boldsymbol{\mathcal{T}}$, error threshold ϵ
  Choose thresholds τ_1, τ_2 according to Lemma 12.
  repeat
    Run a local search algorithm to find (τ_1, τ_2)-second order stationary point.
    Call Algorithm 1 for $i, j, k = 1, 2$ to generate improvement directions, repeat for $O(\log 1/\epsilon)$ times.
    if any of the generated directions improve the function value by at least $\Omega^*(\sigma^{15/8})$ **then**
      Move in the direction.
      Break.
    end if
  until no direction of improvement can be found

Now we are ready to prove Theorem 2

Proof of Theorem 2. By Lemma 12, for any (τ_1, τ_2)-second order stationary point, if $f \geq \epsilon$ Lemma 20 and Lemma 21 will be able to generate a direction of improvement that improves the function value by at least $\Omega^*(\sigma^{15/8})$ with constant probability. Since the initial point has constant loss, if a direction of improvement is found for more than $O^*(1/\sigma^{15/8})$ iterations, then the function value must already be smaller than ϵ.

After the repetition, the probability that we find a direction of improvement is at least $1 - o(\sigma)$. By union bound, we know that with high probability for all the iterations we can find a direction of improvement. □

5 Conclusion

In this paper we showed that the standard nonconvex objective for Tucker decomposition with appropriate regularization does not have any spurious local

minima. We further give a local search algorithm that can optimize a regularized version of the objective in polynomial time. Our algorithm includes additional steps that try to escape from high order saddle points. However, in practice simple local search algorithms like gradient descent rarely encounter saddle points. This phenomenon was analyzed in the matrix setting [19,14], and proving similar things for Tucker decomposition is an open problem. Another set of interesting open problems come from applications where the low rank tensor $\boldsymbol{\mathcal{T}}$ is not known exactly. We either have significant additive noise $\boldsymbol{\mathcal{T}}+\boldsymbol{\mathcal{E}}$, or observe only a subset of entries of $\boldsymbol{\mathcal{T}}$ (tensor completion). Local search algorithms on the nonconvex objective are able to handle similar settings for matrices [6]. We hope our techniques in this paper can be extended to give stronger guarantees for noisy Tucker decomposition and tensor completion.

References

1. Anandkumar, A., Ge, R.: Efficient approaches for escaping higher order saddle points in non-convex optimization. In: Conference on learning theory, pp. 81–102 (2016)
2. Bandeira, A.S., Boumal, N., Voroninski, V.: On the low-rank approach for semidefinite programs arising in synchronization and community detection. arXiv preprint arXiv:1602.04426 (2016)
3. Bhojanapalli, S., Neyshabur, B., Srebro, N.: Global optimality of local search for low rank matrix recovery. In: Advances in Neural Information Processing Systems, pp. 3873–3881 (2016)
4. Carbery, A., Wright, J.: Distributional and l^ q norm inequalities for polynomials over convex bodies in r^ n. Mathematical research letters **8**(3), 233–248 (2001)
5. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. Psychometrika **35**(3), 283–319 (1970)
6. Chi, Y., Lu, Y.M., Chen, Y.: Nonconvex optimization meets low-rank matrix factorization: An overview. IEEE Transactions on Signal Processing **67**(20), 5239–5269 (2019)
7. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. SIAM journal on Matrix Analysis and Applications **21**(4), 1253–1278 (2000)
8. De Lathauwer, L., De Moor, B., Vandewalle, J.: On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. SIAM journal on Matrix Analysis and Applications **21**(4), 1324–1342 (2000)
9. Eldén, L., Savas, B.: A newton–grassmann method for computing the best multilinear rank-(r_1, r_2, r_3) approximation of a tensor. SIAM Journal on Matrix Analysis and applications **31**(2), 248–271 (2009)
10. Frandsen, A., Ge, R.: Understanding composition of word embeddings via tensor decomposition. In: International Conference on Learning Representations (2019)
11. Ge, R., Huang, F., Jin, C., Yuan, Y.: Escaping from saddle points—online stochastic gradient for tensor decomposition. In: Conference on Learning Theory, pp. 797–842 (2015)
12. Ge, R., Jin, C., Zheng, Y.: No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In: International Conference on Machine Learning (2017)
13. Ge, R., Lee, J.D., Ma, T.: Matrix completion has no spurious local minimum. In: Advances in Neural Information Processing Systems, pp. 2973–2981 (2016)
14. Gunasekar, S., Woodworth, B.E., Bhojanapalli, S., Neyshabur, B., Srebro, N.: Implicit regularization in matrix factorization. In: Advances in Neural Information Processing Systems, pp. 6151–6159 (2017)
15. Harshman, R.A., et al.: Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis (1970)
16. Hitchcock, F.L.: The expression of a tensor or a polyadic as a sum of products. Journal of Mathematics and Physics **6**(1-4), 164–189 (1927)

17. Jin, C., Ge, R., Netrapalli, P., Kakade, S.M., Jordan, M.I.: How to escape saddle points efficiently. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1724–1732. JMLR. org (2017)
18. Koren, Y.: The bellkor solution to the netflix grand prize. Netflix prize documentation **81**(2009), 1–10 (2009)
19. Li, Y., Ma, T., Zhang, H.: Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. arXiv preprint arXiv:1712.09203 (2017)
20. Lim, L.H.: Singular values and eigenvalues of tensors: a variational approach. In: 1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2005., pp. 129–132. IEEE (2005)
21. Nesterov, Y., Polyak, B.T.: Cubic regularization of newton method and its global performance. Mathematical Programming **108**(1), 177–205 (2006)
22. Park, D., Kyrillidis, A., Caramanis, C., Sanghavi, S.: Non-square matrix sensing without spurious local minima via the burer-monteiro approach. arXiv preprint arXiv:1609.03240 (2016)
23. Phan, A.H., Cichocki, A., Tichavskỳ, P.: On fast algorithms for orthogonal tucker decomposition. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6766–6770. IEEE (2014)
24. Recht, B., Ré, C.: Parallel stochastic gradient algorithms for large-scale matrix completion. Mathematical Programming Computation **5**(2), 201–226 (2013)
25. Savas, B., Eldén, L.: Handwritten digit classification using higher order singular value decomposition. Pattern recognition **40**(3), 993–1003 (2007)
26. Stewart, G.W.: Perturbation theory for the singular value decomposition. Tech. rep. (1998)
27. Sun, J., Qu, Q., Wright, J.: Complete dictionary recovery over the sphere i: Overview and the geometric picture. IEEE Transactions on Information Theory (2016)
28. Sun, J., Qu, Q., Wright, J.: A geometric analysis of phase retrieval. In: Information Theory (ISIT), 2016 IEEE International Symposium on, pp. 2379–2383. IEEE (2016)
29. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika **31**(3), 279–311 (1966)
30. Vasilescu, M.A.O., Terzopoulos, D.: Multilinear analysis of image ensembles: Tensorfaces. In: European Conference on Computer Vision, pp. 447–460. Springer (2002)
31. Wang, H., Ahuja, N.: Compact representation of multidimensional data using tensor rank-one decomposition. vectors **1**, 5 (2004)
32. Wang, M., Duc, K.D., Fischer, J., Song, Y.S.: Operator norm inequalities between tensor unfoldings on the partition lattice. Linear algebra and its applications **520**, 44–66 (2017)
33. Wedin, P.Å.: Perturbation bounds in connection with singular value decomposition. BIT Numerical Mathematics **12**(1), 99–111 (1972)