Beyond Lazy Training for Over-parameterized Tensor Decomposition

Xiang Wang* Duke University xwang@cs.duke.edu

Chenwei Wu* Duke University cwwu@cs.duke.edu

Jason D. Lee Princeton University jasonlee@princeton.edu

Tengyu Ma Stanford University tengyuma@stanford.edu

Rong Ge Duke University rongge@cs.duke.edu

Abstract

Over-parametrization is an important technique in training neural networks. In both theory and practice, training a larger network allows the optimization algorithm to avoid bad local optimal solutions. In this paper we study a closely related tensor decomposition problem: given an l-th order tensor in $(\mathbb{R}^d)^{\otimes l}$ of rank r (where $r \ll d$), can variants of gradient descent find a rank m decomposition where $m > r$? We show that in a lazy training regime (similar to the NTK regime for neural networks) one needs at least $m = \Omega(d^{l-1})$, while a variant of gradient descent can find an approximate tensor when $m = O^*(r^{2.5l} \log d)$. Our results show that gradient descent on over-parametrized objective could go beyond the lazy training regime and utilize certain low-rank structure in the data.

1 Introduction

The success of training neural networks has sparked theoretical research in understanding non-convex optimization. Over-parametrization – using more neurons than the number of training data or than what is necessary for expressivity – is crucial to the success of optimizing neural networks (Livni et al., 2014; Jacot et al., 2018; Mei et al., 2018). The idea of over-parametrization also applies to other related or simplified problems, such as matrix factorization and tensor decomposition, which are of their own interests and also serve as testbeds for analysis techniques of non-convex optimization. We focus on over-parameterized tensor decomposition in this paper (which is closely connected to over-parameterized neural networks (Ge et al., 2018)).

Concretely, given an order-l symmetric tensor T^* in $(\mathbb{R}^d)^{\otimes l}$ with rank r, we aim to decompose it into a sum of rank-1 tensors with as few components as possible. Finding the low-rank decomposition with the smallest possible rank r is known to be NP-hard (Hillar and Lim, 2013). The problem becomes easier if we relax the goal to finding a decomposition with m components where m is allowed to be larger than r. The natural approach is to optimize the following objective using gradient descent

$$\min_{u_i \in \mathbb{R}^d, c_i \in \mathbb{R}} \left\| \sum_{i=1}^{m} c_i u_i^{\otimes l} - \sum_{i=1}^{r} c_i^* [u_i^*]^{\otimes l} \right\|_F^2. \tag{1}$$

When $m = r$, gradient descent on the objective above will empirically get stuck at a bad local minimum even for orthogonal tensors (Ge et al., 2015). On the other hand, when $m = \Omega(d^{l-1})$, gradient descent provably converges to a global minimum near the initialization. This result follows straightforwardly from

*Equal contribution.

the Neural Tangent Kernel (NTK) technique (Jacot et al., 2018), which was originally developed to analyze neural network training, and is referred to as the "lazy training" regime because essentially the algorithm is optimizing a convex function near the initialization (Chizat and Bach, 2018a).

The main goal of this paper is to understand whether we can go beyond the lazy training regime for the tensor decomposition problem via better algorithm design and analysis. In other words, we aim to use a much milder over-parametrization than $m = \Omega(d^{l-1})$ and still converge to the global minimum of objective (1). We view the problem as an important first step towards analyzing neural network training beyond the lazy training regime.

We build upon the technical framework of mean-field analysis (Mei et al., 2018), which was developed to analyze overparameterized neural networks. It allows the parameters to move far away from the initialization and therefore has the potential to capture the realistic training regime of neural networks. However, to date, all the provable optimization results with mean-field analysis essentially operate in the infinite or exponential overparameterization regime (Chizat and Bach, 2018b; Wei et al., 2019), and applying these techniques to our problem naively would require m to be exponentially large in d, which is even worse than the NTK result. The exponential dependency is *not surprising* because the mean-field analyses in (Chizat and Bach, 2018b; Wei et al., 2019) do not leverage or assume any particular structures of the data so they fail to produce polynomial-time guarantees on the worst-case data. Motivated by identifying problem structure that allows for polynomial-time guarantees, we study the mean-field analysis applied to tensor decomposition.

The main contribution of this paper is to attain nearly dimension-independent over-parametrization for the mean-field analysis in Wei et al. (2019) by leveraging the particular structure of the tensor decomposition problem, and to show that with $m = O^*(r^{2.5l} \log d)$, a modified version of gradient descent on a variant of objective (1) converges to the global minimum and recovers the ground-truth tensor. This is a significant improvement over the NTK requirement of $m = \Omega(d^{l-1})$ and an exponential improvement upon the existing mean-field analysis that requires $m = \exp(d)$. Our analysis shows that unlike the lazy training regime, gradient descent with small initialization and appropriate regularizer can identify the subspace that the ground-truth components lie in, and automatically exploit such structure to reduce the number of necessary components. As shown in Ge et al. (2018), the population-level objective of two-layer networks is a mixture of tensor decomposition objectives with different orders, so our analysis may be extendable to improve the over-parametrization necessary in analysis of two-layer networks.

1.1 Related work

Neural Tangent Kernel There has been a recent surge of research on connecting neural networks trained via gradient descent with the neural tangent kernel (NTK) (Jacot et al., 2018; Du et al., 2018a,b; Chizat and Bach, 2018a; Allen-Zhu et al., 2018; Arora et al., 2019a,b; Zou and Gu, 2019; Oymak and Soltanolkotabi, 2020). This line of analysis proceeds by coupling the training dynamics of the nonlinear network with the training dynamics of its linearization in a local neighborhood of the initialization, and then analyzing the optimization dynamics of the linearization which is convex.

Though powerful and applicable to any function class including tensor decomposition, NTK is not yet a completely satisfying theory for explaining the success of over-parametrization in deep learning. Neural tangent kernel analysis is essentially dataset independent and requires at least number of neurons $m \geq \frac{n}{d} = d^{l-1}$ to find a global optimum (Zou and Gu, 2019; Daniely, 2019)[1].

Beyond NTK approach The gap between linearized models and the full neural network has been established in theory by (Wei et al., 2019; Allen-Zhu and Li, 2019; Yehudai and Shamir, 2019; Ghorbani et al., 2019; Dyer and Gur-Ari, 2019; Woodworth et al., 2020) and observed in practice (Chizat and Bach, 2018a; Arora et al., 2019a; Lee et al., 2019). Higher-order approximations of the gradient dynamics such as Taylorized Training (Bai and Lee, 2019; Bai et al., 2020) and the Neural Tangent Hierarchy (Huang and Yau, 2019) have been recently proposed towards closing this gap. Unlike this paper, existing results mostly try to improve the sample complexity instead of the level of over-parametrization for the NTK approach.

[1]Here n is the number of samples which is effectively $\Theta(d^l)$ in our setting.

Mean field approach For two-layer networks, a series of works used the mean field approach to establish the evolution of the network parameters (Mei et al., 2018; Chizat and Bach, 2018b; Wei et al., 2018; Rotskoff and Vanden-Eijnden, 2018; Sirignano and Spiliopoulos, 2018). In the mean field regime, the parameters move significantly from their initialization, unlike NTK regime, so it is *a priori* possible for the mean field approach to exploit data-dependent structure to utilize fewer neurons. However the current analysis techniques for mean field approach need either exponential in dimension or exponential in time number of neurons to attain small training error and do not exploit any data structure. One of the main contributions of our work is to show that gradient descent can benefit from the low-rank structure in T^*.

Tensor decomposition Tensor decomposition is in general an NP-hard problem (Hillar and Lim, 2013). There are many algorithms that find the exact decomposition (when $m = r$) under various assumptions. In particular Jenrich's algorithm (Harshman, 1970) works when $r \leq d$ and the components are linearly independent. In our setting, the components may not be linearly independent, this is similar to the overcomplete tensor decomposition problem. Although there are some algorithms for overcomplete tensor decomposition (e.g., Cardoso (1991); Ma et al. (2016)), they require nondegeneracy conditions which we are not assuming. When the number of components m is allowed to be larger than r, one can use spectral algorithms to find a decomposition where $m = \Theta(r^{l-1})$. In this paper our focus is to achieve similar guarantees using a direct optimization approach.

Neural network with polynomial activations Another model that sits between tensor decomposition and standard ReLU neural network is neural network with polynomial activations. Livni et al. (2013) gave an algorithm for training network with quadratic activations with specific algorithm. Andoni et al. (2014) gave a way to learn degree l polynomials over d variables using $\Omega(d^l)$ neurons, which is similar to the guarantee of (much later) NTK approach.

2 Notations

We use $[n]$ as a shorthand for $\{1, 2, \cdots, n\}$. We use $O(\cdot), \Omega(\cdot)$ to hide constant factor dependencies. We use $O^*(\cdot)$ to hide constant factors and also the dependency on accuracy ϵ. We use $\mathrm{poly}(\cdot)$ to represent a polynomial on the relevant parameters with constant degree.

Tensor notations: We use $\otimes$ as the tensor product (outer product). An l-th order d-dimensional tensor is defined as an element in space $\mathbb{R}^d \otimes \cdots \otimes \mathbb{R}^d$, succinctly denoted as $(\mathbb{R}^d)^{\otimes l}$. For any $i_1, \cdots, i_l \in [d]$, we use $T_{i_1,\cdots,i_l}$ to refer to the $(i_1, \cdots, i_l)$-th entry of $T \in (\mathbb{R}^d)^{\otimes l}$ with respect to the canonical basis. For a vector $v \in \mathbb{R}^d$, we define $v^{\otimes l}$ as a tensor in $(\mathbb{R}^d)^{\otimes l}$ such that $\left(v^{\otimes l}\right)_{i_1,\cdots,i_l} = v_{i_1} v_{i_2} \cdots v_{i_l}$. A tensor is symmetric if the entry values remain unchanged for any permutation of its indices. We define $\mathrm{vec}(\cdot)$ to be the vectorize operator for tensors, mapping a tensor in $(\mathbb{R}^d)^{\otimes l}$ to a vector in $\mathbb{R}^{d^l}$: $\mathrm{vec}(T)_{(i_1-1)d^{l-1}+(i_2-1)d^{l-2}+\cdots+(i_{l-1}-1)d+i_l} := T_{i_1,i_2,\cdots,i_l}$.

A tensor $T \in (\mathbb{R}^d)^{\otimes l}$ is rank-1 if it can be written as $T = w \cdot v_1 \otimes v_2 \otimes \cdots \otimes v_l$ for some $w \in \mathbb{R}$ and $v_1, \cdots, v_l \in \mathbb{R}^d$, and the rank of a tensor is defined as the minimum integer k such that this tensor equals the sum of k rank-1 tensors.

Norm and inner product: We use $\|v\|$ to denote the ℓ_2 norm of a vector v. For l-th order tensors $T, T' \in (\mathbb{R}^d)^{\otimes l}$ (vectors and matrices can be viewed as tensors with order 1 and 2, respectively), we define the inner product as $\langle T, T' \rangle := \sum_{i_1,\cdots,i_l \in [d]} T_{i_1,\cdots,i_l} T'_{i_1,\cdots,i_l}$ and the Frobenius norm as $\|T\|_F = \sqrt{\sum_{i_1,\cdots,i_l \in [d]} T^2_{i_1,\cdots,i_l}}$.

3 Problem setup and challenges

In this section we discuss the objective for over-parameterized tensor decomposition and explain the challenges in optimizing this objective.

We consider tensor decomposition problems with general order $l \geq 3$. Throughout the paper we consider l as a constant. Specifically, we assume that the ground-truth tensor is T^* of rank at most r:

$$T^* := \sum_{i=1}^{r} c_i^* [u_i^*]^{\otimes l},$$

where $\forall i \in [r], c_i^* \in \mathbb{R}$ and $u_i^* \in \mathbb{R}^d$. Without loss of generality, we assume that $\|T^*\|_F = 1$. We focus on the low rank setting where r is much smaller than d. Note we don't assume that u_i^*'s are linearly independent.

The vanilla over-parameterized model we use consists of m components (where $m \geq r$):

$$T_v := \sum_{i=1}^{m} c_i u_i^{\otimes l},$$

where $\forall i \in [m], c_i \in \mathbb{R}$ and $u_i \in \mathbb{R}^d$. We use $U \in \mathbb{R}^{d \times m}$ to denote the matrix whose i-th column is u_i, and denote $C \in \mathbb{R}^{m \times m}$ as the diagonal matrix with $C_{i,i} = c_i$. The vanilla loss function we are considering is the square loss:

$$f_v(U, C) = \frac{1}{2} \|T_v - T^*\|_F^2 = \frac{1}{2} \left\| \sum_{i=1}^{m} c_i u_i^{\otimes l} - \sum_{i=1}^{r} c_i^* [u_i^*]^{\otimes l} \right\|_F^2. \tag{2}$$

In other words, we are looking for a rank m approximation to a rank r tensor. When $m = r$, the problem of finding a decomposition is known to be NP-hard. Therefore, our goal is to get a small objective value with small m (which corresponds to the rank of T_v). In the following sub-sections, we will see that there are many challenges to directly optimize the vanilla over-parameterized model over the vanilla objective, so we will need to modify the parametrization of the tensor T_v and the optimization algorithm to overcome them.

3.1 Challenge 0: lazy training requires immense over-parameterization

We show lazy training requires $\Omega(d^{l-1})$ components to fit a rank-one tensor in the following theorem.

Theorem 1. *Suppose the ground truth tensor $T^* = [u^*]^{\otimes l}$, where u^* is uniformly sampled from the unit sphere $\mathbb{S}^{d-1}$. Lazy training (defined as below) requires $\Omega(d^{l-1})$ components to achieve $o(1)$ error in expectation.*

In the lazy training regime, all the u_i's stay very close to the initialization. Assuming the final u_i' is equal to $u_i + \delta_i$, all the higher-order terms in δ_i can be ignored. Therefore, the model can only capture tensors in the linear subspace $S_U = \text{span}\{P_{sym}\text{vec}(u_i^{\otimes l-1} \otimes \delta_i)\}_{i=1}^m$ (here P_{sym} is the projection to the space of vectorized symmetric tensors, u_i's are the initialization and δ_i's are arbitrary vectors in $\mathbb{R}^d$). The dimension of this subspace is upperbounded by dm. Let W_l be the space of all vectorized symmetric tensors in $(\mathbb{R}^d)^{\otimes l}$ (with dimension $\Omega(d^l)$), and $S_U^\perp$ be the subspace of W_l orthogonal to S_U. We show that for a random rank-1 tensor T^*, it will often have a large projection in $S_U^\perp$ unless $m = \Omega(d^{l-1})$. Basically, the subspace S_U has to cover the whole space W_l to approximate a random rank-1 tensor. The proof of Theorem 1 is in Appendix A.

3.2 Challenge 1: zero is a high-order saddle point for vanilla objective

As Chizat and Bach (2018a) pointed out, lazy training regime corresponds to the case where the initialization has large norm. A natural way to get around lazy training is to use a much smaller initialization. However, for the vanilla objective, 0 will be a saddle point of order l on the loss landscape. This makes gradient descent really slow at the beginning. In Section 4, we fix this issue by re-parameterizing the model into a 2-homogeneous model.

3.3 Challenge 2: existence of bad local minima far away from 0

It was shown that no bad local minima exist in matrix decomposition problems (Ge et al., 2016). Therefore, (stochastic) gradient descent is guaranteed to find a global minimum. In this section, we show that in contrary tensor decomposition problems with order at least 3 have bad local minima.

Theorem 2. *Let $f_v(U, C)$ be as defined in Equation 2. Assume $l \geq 3, d > r \geq 1$ and $m \geq r(l+1)+1$. There exists a symmetric ground truth tensor T^* with rank at most $r(l+1)+1$ such that a local minimum with function value $l(l-1)r/4$ exists while the global minimum has function value zero.*

In the construction, we set all the u_i's to be $e_1/m^{1/l}$ so that $T = e_1^{\otimes l}$. We define the ground truth tensor by setting the residual $T - T^*$ to be $\sum_{j=2}^{r+1} e_j^{\otimes 2} \otimes e_1^{\otimes l-2}$ plus its $\binom{l}{2}$ permutations. At this point, the gradient equals zero, so there is no first order change to the function value. Furthermore, we show if any component moves in one of the missing direction e_j for $2 \leq j \leq r+1$, it will incur a second order function value increase. So the tensor can only moves along e_1 direction, which cannot further decrease the function value because $e_1^{\otimes l}$ is orthogonal with the residual. Note this is a bad local min but not a strict bad local min because we can shrink one component to zero and meanwhile increase another component so that the tensor does not change. When we have a zero component (it's a saddle point), we can add a missing direction to decrease the function value.

In Appendix B, we prove Theorem 2 and also construct a bad local minimum for 2-homogeneous model defined in Section 4. To escape these spurious local minima, our algorithm re-initializes one component after a fixed number of iterations.

4 Algorithms and main results

In this section, we introduce our main algorithm, a modified version of gradient descent on a non-convex objective, and state our main results.

To address the high-order saddle point issue in Section 3.2, we introduce a new variant of the parameterized models.

$$T := \sum_{i=1}^{m} a_i c_i^{l-2} u_i^{\otimes l},$$

where $\forall i \in [m], a_i \in \{-1, 1\}, c_i = \frac{1}{\|u_i\|}$ and $u_i \in \mathbb{R}^d$.

Note that since $u_i^{\otimes l}$ is homogeneous, there is a redundancy in the vanilla parametrization between the coefficient and the norm of $\|u\|$. Here we do the rescaling to make sure that the model T is a 2-homogeneous function of u_i's. Using the new formulation of T, 0 will no longer be a high order saddle point.

Recall that we use $U \in \mathbb{R}^{d\times m}$ to denote the matrix whose i-th column is u_i. We use $C, A \in \mathbb{R}^{m\times m}$ to denote the diagonal matrices with $C_{ii} = c_i, A_{ii} = a_i$. The loss function we are considering is the square loss plus a regularization term:

$$f(U, C, \hat{C}, A) \triangleq \frac{1}{2} \left\| \sum_{i=1}^{m} a_i c_i^{l-2} u_i^{\otimes l} - T^* \right\|_F^2 + \lambda \sum_{i=1}^{m} \hat{c}_i^{l-2} \|u_i\|^l,$$

where $\forall i \in [m], \hat{c}_i \in \mathbb{R}^+$ and we use $\hat{C} \in \mathbb{R}^{m\times m}$ to denote the diagonal matrix with $\hat{C}_{ii} = \hat{c}_i$. For simplicity, we use $\bar{C}$ to denote the tuple $(C, \hat{C}, A)$. Therefore, we can write $f(U, C, \hat{C}, A)$ as $f(U, \bar{C})$.

The algorithm contains K epochs, where each epoch includes H iterations. At the initialization, we independently sample each u_i from $\delta\text{Unif}(\mathbb{S}^{d-1})$, where the radius δ will be set to be $\text{poly}(\epsilon, 1/d)$.

Denote the subspace of $\text{span}\{u_i^*\}$ as S and its orthogonal subspace in $\mathbb{R}^d$ as B. Let P_S, P_B be the projection matrices onto subspace S and B, respectively. Since the components of the ground-truth tensor lies in the subspace S, ideally we want to make sure that the components of tensor T lies in the same subspace S. We also want to make sure c_i roughly equals $1/\|P_S u_i\|$ to ensure the improvement in S subspace is large enough. However, the algorithm does not know the subspace S. We address this problem

using the observation that $\|P_S u_i\| \approx \frac{\sqrt{r}}{\sqrt{d}}\|u_i\|$ at initialization; and $\|P_S u_i\| \approx \|u_i\|$ if norm of u_i is large, but its projection in B has not grown larger. In our algorithm we introduce a "scalar mode switch" step between these two regimes by the separation between C and $\hat{C}$: For the i-th component, the coefficients c_i and $\hat{c}_i$ are initialized as $\sqrt{d(m+K)}/\|u_i\|$ and $1/\|u_i\|$, respectively, and we reduce c_i by a factor of $\sqrt{d(m+K)}$ (c_i will be equal to $\hat{c}_i$ afterwards) when $\|u_i\|$ exceeds $2\sqrt{m+K}\delta$ for the first time. For each $i \in [m], a_i$ is i.i.d. sampled from Unif$\{1, -1\}$.

We also re-initialize one component at the beginning of each epoch. At each iteration, we first update U by gradient descent: $U' \leftarrow U - \eta \nabla_U f(U, \bar{C})$. Note that when taking the gradient over U, we treat c_i's and $\hat{c}_i$'s as constants. Then we update each c_i and $\hat{c}_i$ using the updated value of u_i to preserve 2-homogeneity, i.e., $c_i' = \frac{\|u_i\|}{\|u_i'\|} c_i$ and $\hat{c}_i' = \frac{\|u_i\|}{\|u_i'\|}\hat{c}_i$. We fix a_i's during the algorithm except for the initialization and re-initialization steps.

The pseudocode is given in Algorithm 1. Using this variant of gradient descent, we can recover the ground truth tensor T^* with high probability using only $O\left(\frac{r^{2.5l}}{\epsilon^5}\log(d/\epsilon)\right)$ number of components. The formal theorem is stated below.

Algorithm 1 Variant of Gradient Descent for Tensor Decomposition

Input: number of epochs K, number of iterations in one epoch H, initialization size δ, step size η.

For each $i \in [m]$, initialize u_i i.i.d. from δUnif$(\mathbb{S}^{d-1})$; initialize a_i i.i.d. from Unif$\{1, -1\}$; initialize c_i as $\frac{\sqrt{d(m+K)}}{\|u_i\|}$ and $\hat{c}_i$ as $\frac{1}{\|u_i\|}$.

for epoch $k := 1$ to K **do**

  Let u_j be any vector with the smallest ℓ_2 norm among all columns of U.

  Re-initialize u_j from δUnif$(\mathbb{S}^{d-1})$, re-initialize a_j from Unif$\{1, -1\}$ and set $c_j = \frac{\sqrt{d(m+K)}}{\|u_j\|}, \hat{c}_j = \frac{1}{\|u_j\|}$.

  for iteration $t := 1$ to H **do**

    $U' \leftarrow U - \eta \nabla_U f(U, \bar{C})$.

    for $i := 1$ to m **do**

      $c_i' \leftarrow \frac{\|u_i\|}{\|u_i'\|} c_i$; $\hat{c}_i' \leftarrow \frac{\|u_i\|}{\|u_i'\|}\hat{c}_i$.

      if $\|u_i\| \le 2\sqrt{m+K}\delta < \|u_i'\|$ holds for the first time since it was (re)-initialized **then**

        $c_i' \leftarrow \frac{c_i'}{\sqrt{d(m+K)}}$. ▷ Scalar Mode Switch

    $U \leftarrow U', C \leftarrow C', \hat{C} \leftarrow \hat{C}$.

Output: $T := \sum_{i=1}^m a_i c_i^{l-2} u_i^{\otimes l}$.

Theorem 3. *Given any target accuracy $\epsilon > 0$, there exists $m = O\left(\frac{r^{2.5l}}{\epsilon^5}\log(d/\epsilon)\right), \lambda = O\left(\frac{\epsilon}{r^{0.5l}}\right), \delta = \text{poly}(\epsilon, 1/d), \eta = \text{poly}(\epsilon, 1/d), H = \text{poly}(1/\epsilon, d)$ such that with probability at least 0.99, our algorithm finds a tensor T satisfying*

$$\|T - T^*\|_F \le \epsilon,$$

within $K = O\left(\frac{r^{2l}}{\epsilon^4}\log(d/\epsilon)\right)$ epochs.

5 Summary of our techniques

In this section, we discuss the high-level ideas that we need to prove Theorem 3. The full proof is deferred into Appendix C.

Generally, doing gradient descent never increases the objective value (though this is not obvious for our algorithm as it is slightly different in handling the normalization $c_i, \hat{c}_i$'s). Our main concern is to address Challenge 2, namely, the algorithm might get stuck at a bad local minimum. We will show that this cannot happen with the re-initialization procedure.

More precisely, we rely on the following main lemma to show that as long as the objective is large, there is at least a constant probability to improve the objective within one epoch.

Lemma 1. *In the setting of Theorem 3, let $(U_0', \bar{C}_0')$ and $(U_H, \bar{C}_H)$ be the parameters at the beginning of an epoch and the parameters at the end of the same epoch. Assume $\|T_0' - T^*\|_F \geq \epsilon$, where T_0' is tensor with parameters $(U_0', \bar{C}_0')$. Then with probability at least $1/6$, we have*

$$f(U_H, \bar{C}_H) - f(U_0', \bar{C}_0') = -\Omega\left(\frac{\epsilon^4}{r^{2l}\log(d/\epsilon)}\right).$$

We complement this lemma by showing that even if an epoch does not decrease the objective, it will not overly increase the objective.

Lemma 2. *In the setting of Theorem 3, let $(U_0', \bar{C}_0')$ and $(U_H, \bar{C}_H)$ be the parameters at the beginning of an epoch and the parameters at the end of the same epoch. Assume $f(U_0', \bar{C}_0') \geq \epsilon^2$, where ϵ is the target accuracy in Theorem 3. Then, we have $f(U_H, \bar{C}_H) - f(U_0', \bar{C}_0') = O(\frac{1}{\lambda m})$.*

From these two lemmas, we know that in each epoch, the loss function can decrease by $\Omega\left(\frac{\epsilon^4}{r^{2l}\log(d/\epsilon)}\right)$ with probability at least $\frac{1}{6}$, and even if we fail to decrease the function value, the increase of function value is at most $O\left(\frac{1}{\lambda m}\right)$. By our choice of parameters in Theorem 3, $m = \Theta\left(\frac{r^{2.5l}}{\epsilon^5}\log(d/\epsilon)\right), \lambda = \Theta\left(\frac{\epsilon}{r^{0.5l}}\right)$ and then $O(\frac{1}{\lambda m}) = O(\frac{\epsilon^4}{r^{2l}\log(d/\epsilon)})$. Choosing a large constant factor in m, we can ensure that the function value decrease will dominate the increase. This allows us to prove Theorem 3.

In the next two subsections, we will discuss how to prove Lemma 1 and Lemma 2, respectively.

5.1 Proof sketch for Lemma 2 - upper bound on function increase

To prove the increase of f is bounded in one epoch, we identify all the possible ways that the loss can increase and upper bound each of them. We first show that a normal step (without scalar mode switch) of the algorithm will not increase the objective function

Lemma 3. *In the setting of Theorem 3, let $(U, \bar{C})$ be the parameters at the beginning of one iteration and let $U', \bar{C}'$ be the updated parameters (before potential scalar mode switch). Assuming $f(U, \bar{C}) \leq 10$, we have $f(U', \bar{C}') - f(U, \bar{C}) \leq -\frac{\eta}{l}\left\|\nabla_U f(U, \bar{C})\right\|_F^2$.*

Note that we treat C and $\hat{C}$ as constants when taking gradient with respect to U and then update C and $\hat{C}$ according to the updated value of U, so this lemma does not directly follow from standard optimization theory. The gradient descent on U decreases the function value when the step size is small enough while updating $C, \hat{C}$ can potentially increase the function value. In order to show that overall the function value decreases, we need to bound the function value increase due to updating $C, \hat{C}$. We are able to do this because of the special regularizer we choose. In particular, our regularizer guarantees that the change introduced by updating C and $\hat{C}$ is proportional to the change of the gradient step, and is smaller in scale. Therefore we maintain the decrease in the gradient step.

Since we already know that the function value cannot increase in a normal iteration (before potential scalar mode switch), the only causes of the function value increase are the re-initialization or scalar mode switches. According to the algorithm, we only switch the scalar mode when the norm of a component reaches $2\sqrt{m+K}\delta$ for the first time, so the number of scalar mode switches in each epoch is at most m. Choosing δ to be small enough, the effects of scalar mode switches should be negligible. In the re-intialization, we remove the component with smallest ℓ_2 norm, which can increase the function value by at most $O(\frac{1}{\lambda m})$. This is proved in Lemma 4.

Lemma 4. *In the setting of Theorem 3, let $(U_0', \bar{C}_0')$ and $(U_0, \bar{C}_0)$ be the parameters before and after the reinitialization step, respectively. Assume $f(U_0', \bar{C}_0') \geq \epsilon^2$, where ϵ comes from Theorem 3. Then, we have $f(U_0, \bar{C}_0) - f(U_0', \bar{C}_0') = O(\frac{1}{\lambda m})$.*

In the proof, we can show the function value is at most a constant and then $\sum_{i=1}^{m} \|u_i\|^2 = O(1/\lambda)$ due to the regularizer. Since we choose the reinitialized component u as one of the component with smallest ℓ_2 norm, we know $\|u\|^2 = O(\frac{1}{\lambda m})$. This then allows us to bound the function value change from reinitialization by $O(\frac{1}{\lambda m})$. Lemma 2 follows from Lemma 3 and Lemma 4.

5.2 Proof sketch for Lemma 1 - escaping local minima

In this section, we will show how we can escape local minima by re-initialization. Intuitively, we will show that when a component is randomly re-initialized, it has a positive probability of having a good correlation with the current residual $T - T^*$. However, there is a major obstacle here: because the component is re-initialized in the full d-dimensional space, the correlation of this new component with $T - T^*$ is going to be of the order $d^{-l/2}$. If every epoch can only improve the objective function by $d^{-l/2}$ we would need a much larger number of epochs and components.

We solve this problem by observing that both T and T^* are almost entirely in the subspace S. If we only care about the projection in S, the random component will have a correlation of $r^{-l/2}$ with the residual $T - T^*$. We will show that such a correlation will keep increasing until the norm of the new component is large enough, therefore decreasing the objective function.

First of all, we need to show that the influence coming from the subspace B (the orthogonal subspace of the span of $\{u_i^*\}$) is small enough so that it can be ignored.

Lemma 5. *In the setting of Theorem 3, we have $\|P_B U\|_F^2 \le (m+K)\delta^2$ throughout the algorithm.*

We prove Lemma 5 by showing the norm of $P_B U$ only increases at the (re-)initializations, so it will stay small throughout this algorithm. This lemma is also the motivation of our algorithm, i.e., we treat C and $\hat{C}$ as constants when taking the gradient so that the gradient of U will never have negative correlation with $P_B U$.

Now let us focus on the subspace S. We denote the re-initialized vector at t-th step as u_t, and its sign as $a \in \{\pm 1\}$, and we will take a look at the change of $P_S u_t$. Our analysis focuses on the correlation between $P_S u_t$ and the residual tensor: $\langle (P_{S^{\otimes l}} T_t - T^*), a(\overline{P_S u_t}^{\otimes l}) \rangle$. Here $\overline{P_S u_t}$ is the normalized version $P_S u_t$. We will show that the norm of u_t will blow up exponentially if this correlation is significantly negative at every iteration.

Towards this goal, first we will show that the initial point $P_S u_0$ has a large negative correlation with the residual. We lower bound this correlation by anti-concentration of Gaussian polynomials:

Lemma 6. *Suppose the residual at the beginning of one epoch is $T_0' - T^*$. Suppose $a c_0^{l-2} u_0^{\otimes l}$ is the reinitialized component. With probability at least $1/5$,*

$$\left\langle P_{S^{\otimes l}} T_0' - T^*, a\overline{P_S u_0}^{\otimes l} \right\rangle \le -\Omega\left(\frac{1}{r^{0.5l}}\right) \left\| P_{S^{\otimes l}} T_0' - P_{S^{\otimes l}} T^* \right\|_F,$$

where $\overline{P_S u_0} = P_S u_0 / \|P_S u_0\|$.

Our next step argues that if this negative correlation is large in every step, then the norm of u_t blows up exponentially:

Lemma 7. *In the setting of Theorem 3, within one epoch, let T_0 be the tensor after the reinitilization and let T_τ be the tensor at the end of the τ-th iteration. Assume $\|P_S u_0\| \ge \Omega(\delta/\sqrt{d})$. For any $t \ge 1$, as long as $\left\langle P_{S^{\otimes l}} T_\tau - T^*, a\overline{P_S u_\tau}^{\otimes l} \right\rangle \le -\Omega\left(\frac{\epsilon}{r^{0.5l}}\right)$ for all $t-1 \ge \tau \ge 0$, we have*

$$\|P_S u_t\|^2 \ge \left(1 + \Omega\left(\frac{\eta\epsilon}{r^{0.5l}}\right)\right)^t \|P_S u_0\|^2.$$

Therefore the final step is to show that $P_S u_t$ always have a large negative correlation with $T_t - T^*$, unless the function value has already decreased. The difficulty here is that both the current reinitialized component u_t and other components are moving, therefore T_t is also changing.

We can bound the change of the correlation by separating it into two terms, which are the change of the re-initialized component and the change of the residual:

$$\begin{aligned}&\left\langle P_{S^{\otimes l}}T_t - T^*, a\overline{P_S u_t}^{\otimes l}\right\rangle - \left\langle P_{S^{\otimes l}}T_0 - T^*, a\overline{P_S u_0}^{\otimes l}\right\rangle \\ \leq &\sum_{\tau=1}^{t}\left(\left\langle P_{S^{\otimes l}}T_{\tau-1} - T^*, \overline{P_S u_\tau}^{\otimes l}\right\rangle - \left\langle P_{S^{\otimes l}}T_{\tau-1} - T^*, \overline{P_S u_{\tau-1}}^{\otimes l}\right\rangle\right) + \sum_{\tau=1}^{t}\left\|T_\tau - T_{\tau-1}\right\|_F.\end{aligned}$$

The change of the re-initialized component has a small effect on the correlation because the change in S subspace can only improve the correlation, and the influence of the B subspace can be bounded. This is formally proved in the following lemma.

Lemma 8. *In the setting of Theorem 3, suppose at the beginning of one iteration, the tensor T has parameters $(U, \bar{C})$. Suppose u is one column vector in U with $\|P_S u\| = \Omega(\frac{\delta}{\sqrt{d}})$ and $u' = u - \eta\nabla_u f(U, \bar{C})$. We have*

$$\left\langle P_{S^{\otimes l}}T - T^*, a\overline{P_S u'}^{\otimes l}\right\rangle \leq \left\langle P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*, a\overline{P_S u}^{\otimes l}\right\rangle + \eta\delta poly(d),$$

where poly(d) does not hide any dependency on η, δ.

Therefore, the only way to change the residual term by a lot must be changing the tensor T, and the accumulated change of T is strongly correlated with the decrease of f. This is similar to the technique of bounding the function value decrease in Wei et al. (2019). The connection between them are formalized in the following lemma:

Lemma 9. *In the same setting of Lemma 7, within one epoch, let $(U_0, \bar{C}_0)$ be the parameters after the reinitialization step and let $(U_H, \bar{C}_H)$ be the parameters at the end of this epoch. We have*

$$\sum_{\tau=1}^{H}\left\|T_\tau - T_{\tau-1}\right\|_F \leq O\left(\sqrt{\frac{\eta H}{\lambda}}\right)\sqrt{f(U_0, \bar{C}_0) - f(U_H, \bar{C}_H) + \delta^2 poly(d)} + \delta^2 poly(d),$$

where poly(d) does not hide dependency on δ.

Intuitively, Lemma 9 is true because a large accumulated change of T indicates large gradients along the trajectory, which suggests a large decrease in the function value. In fact, we choose the parameters such that $\lambda = \Theta(\frac{\epsilon}{r^{0.5l}}), \eta H = \Theta(\frac{r^{0.5l}}{\epsilon}\log(d/\epsilon))$. If the accumulated change of T is larger than $\Omega(\frac{\epsilon}{r^{0.5l}})$, the function value decreases by at least $\Omega(\frac{\epsilon^4}{r^{2l}\log(d/\epsilon)})$, as stated in Lemma 1.

Combining all the steps, we show that either the function value has already decreased (by Lemma 9), or the correlation remains negative and the norm $\|P_S u_t\|$ blows up exponentially (by Lemma 7). The norm cannot grow exponentially because of the regularizer, so the function value must eventually decrease. This finishes the proof of Lemma 1.

6 Conclusion

In this paper we show that for an over-parameterized tensor decomposition problem, a variant of gradient descent can learn a rank r tensor using $O^*(r^{2.5l}\log(d/\epsilon))$ components. The result shows that gradient-based methods are capable of leveraging low-rank structure in the input data to achieve lower level of over-parametrization. There are still many open problems, in particular extending our result to a mixture of tensors of different orders which would have implications for two-layer neural network with ReLU activations. We hope this serves as a first step towards understanding what structures can help gradient descent to learn efficient representations.

Acknowledgements

RG acknowledges support from NSF CCF1704656, NSF CCF-1845171 (CAREER), NSF CCF-1934964 (TRIPODS), NSF-Simons Research Collaborations on the Mathematical and Scientific Foundations of Deep Learning, a Sloan Fellowship, and a Google Faculty Research Award.

JDL acknowledges support of the ARO under MURI Award W911NF-11-1-0303, the Sloan Research Fellowship, and NSF CCF 2002272. This is part of the collaboration between US DOD, UK MOD and UK Engineering and Physical Research Council (EPSRC) under the Multidisciplinary University Research Initiative.

TM acknowledges support of Google Faculty Award. The work is also partially supported by SDSI and SAIL at Stanford.

References

Allen-Zhu, Z. and Li, Y. (2019). What can resnet learn efficiently, going beyond kernels? *arXiv preprint arXiv:1905.10337*.

Allen-Zhu, Z., Li, Y., and Liang, Y. (2018). Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*.

Andoni, A., Panigrahy, R., Valiant, G., and Zhang, L. (2014). Learning polynomials with neural networks. In *International conference on machine learning*, pages 1908–1916.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. (2019a). On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*.

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. (2019b). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*.

Bai, Y., Krause, B., Wang, H., Xiong, C., and Socher, R. (2020). Taylorized training: Towards better approximation of neural network training at finite width. *arXiv preprint arXiv:2002.04010*.

Bai, Y. and Lee, J. D. (2019). Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*.

Carbery, A. and Wright, J. (2001). Distributional and l^q norm inequalities for polynomials over convex bodies in r^n. *Mathematical research letters*, 8(3):233–248.

Cardoso, J.-F. (1991). Super-symmetric decomposition of the fourth-order cumulant tensor. blind identification of more sources than sensors. In *International Conference on Acoustics, Speech, & Signal Processing, Icassp*.

Chizat, L. and Bach, F. (2018a). A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 8.

Chizat, L. and Bach, F. (2018b). On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3040–3050.

Daniely, A. (2019). Neural networks learning and memorization with (almost) no over-parameterization. *arXiv preprint arXiv:1911.09873*.

Dasgupta, S. and Gupta, A. (2003). An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. (2018a). Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. (2018b). Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*.

Dyer, E. and Gur-Ari, G. (2019). Asymptotics of wide networks from feynman diagrams. *arXiv preprint arXiv:1909.11304*.

Ge, R., Huang, F., Jin, C., and Yuan, Y. (2015). Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842.

Ge, R., Lee, J. D., and Ma, T. (2016). Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981.

Ge, R., Lee, J. D., and Ma, T. (2018). Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. (2019). Limitations of lazy training of two-layers neural networks. *arXiv preprint arXiv:1906.08899*.

Harshman, R. (1970). Foundations of the parafac procedure: Model and conditions for an explanatory factor analysis. *Technical Report UCLA Working Papers in Phonetics 16, University of California, Los Angeles, Los Angeles, CA*.

Hillar, C. J. and Lim, L.-H. (2013). Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):1–39.

Huang, J. and Yau, H.-T. (2019). Dynamics of deep neural networks and neural tangent hierarchy. *arXiv preprint arXiv:1909.08156*.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*.

Livni, R., Shalev-Shwartz, S., and Shamir, O. (2013). An algorithm for training polynomial networks. *arXiv preprint arXiv:1304.7045*.

Livni, R., Shalev-Shwartz, S., and Shamir, O. (2014). On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863.

Ma, T., Shi, J., and Steurer, D. (2016). Polynomial-time tensor decompositions with sum-of-squares. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 438–446. IEEE.

Mei, S., Montanari, A., and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671.

Oymak, S. and Soltanolkotabi, M. (2020). Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*.

Rotskoff, G. M. and Vanden-Eijnden, E. (2018). Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *arXiv preprint arXiv:1805.00915*.

Sirignano, J. and Spiliopoulos, K. (2018). Mean field analysis of neural networks. *arXiv preprint arXiv:1805.01053*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

Vignat, C. and Bhatnagar, S. (2008). An extension of wick's theorem. *Statistics & probability letters*, 78(15):2404–2407.

Wei, C., Lee, J. D., Liu, Q., and Ma, T. (2018). On the margin theory of feedforward neural networks. *arXiv preprint arXiv:1810.05369*.

Wei, C., Lee Jason, D., Liu, Q., and Ma, T. (2019). Regularization matters: Generalization and optimization of neural nets vs their induced kernel. *arXiv preprint arXiv:1810.05369*.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. (2020). Kernel and rich regimes in overparametrized models. *arXiv preprint arXiv:2002.09277*.

Yehudai, G. and Shamir, O. (2019). On the power and limitations of random features for understanding neural networks. *arXiv preprint arXiv:1904.00687*.

Zou, D. and Gu, Q. (2019). An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2053–2062.

Notations

Besides the notations defined in Section 2, we also use the following notations in the proofs.

We use $\odot$ for the Khatri-Rao product. We denote e_i as the i-th basis vector in $\mathbb{R}^d$.

We define $\mathrm{mat}(\cdot)$ to be the matrixize operator for tensors, mapping a tensor in $(\mathbb{R}^d)^{\otimes l}$ to a matrix in $\mathbb{R} \times \mathbb{R}^{d^{l-1}}$: $\mathrm{mat}(T)_{i_1,(i_2-1)d^{l-2}+\cdots+(i_{l-1}-1)d+i_l} := T_{i_1,i_2,\cdots,i_l}$ for any $i_1, i_2, \cdots, i_l \in [d]$.

We view a tensor $T \in (\mathbb{R}^d)^{\otimes l}$ as a multilinear form. For matrices $M_1 \in \mathbb{R}^{d\times k_1}, \cdots, M_l \in \mathbb{R}^{d\times k_l}$, the tensor $T(M_1, M_2, \cdots, M_l) \in \mathbb{R}^{k_1\times k_2\times\cdots\times k_l}$ is defined such that $T(M_1, M_2, \cdots, M_l)_{j_1,\cdots,j_l} := \sum_{i_1,\cdots,i_l\in[d]} T_{i_1,\cdots,i_l}(M_1)_{i_1,j_1}\cdots(M_l)_{i_l,j_l}$, for any $j_1 \in [k_1], \cdots, j_l \in [k_l]$. For notation simplicity, we use $T(M^{\otimes k}, M_{k+1}, \cdots, M_l)$ to denote $T(M, M, \cdots, M, M_{k+1}, \cdots, M_l)$. In particular, for any $v \in \mathbb{R}^d$, $T(v^{\otimes l})$ is a scalar equals to $\left\langle T, v^{\otimes l}\right\rangle = \sum_{i_1,\cdots,i_l\in[d]} T_{i_1,\cdots,i_l} v_{i_1} v_{i_2}\cdots v_{i_l}$.

A Lower Bound for the Number of Components Needed for Kernels

In this section, we will prove that a lazy training model requires $\Omega(d^{l-1})$ components to fit a random rank-one tensor with $o(1)$ loss. Recall Theorem 1 as follows.

Theorem 1. *Suppose the ground truth tensor $T^* = [u^*]^{\otimes l}$, where u^* is uniformly sampled from the unit sphere $\mathbb{S}^{d-1}$. Lazy training (defined as below) requires $\Omega(d^{l-1})$ components to achieve $o(1)$ error in expectation.*

Recall that in our definition, a lazy training model can only capture tensors in the linear subspace $S_U = \mathrm{span}\{P_{sym}\mathrm{vec}(u_i^{\otimes l-1} \otimes \delta_i)\}_{i=1}^m$ (here P_{sym} is the projection to the space of vectorized symmetric tensors, δ_i's are arbitrary vectors in $\mathbb{R}^d$). The dimension of this subspace is upperbounded by dm. Let W_l be the space of all vectorized symmetric tensors in $(\mathbb{R}^d)^{\otimes l}$, and $S_U^\perp$ be the subspace of W_l orthogonal to S_U. We only need to show that for a random rank-one tensor, in expectation its projection on the orthogonal subspace $S_U^\perp$ is at least a constant unless $m = \Omega(d^{l-1})$. In the following lemma, we first lower bound the projection of the ground truth tensor on a fixed direction. The proof of Lemma 10 is deferred into Section A.2.

Lemma 10. *Let $u \in \mathbb{R}^d$ be a vector sampled uniformly on the unit sphere $\mathbb{S}^{d-1}$. For any vectorized symmetric l-th order tensor $b \in \mathbb{R}^{d^l}$ with unit ℓ_2 norm, we have*

$$b^\top \mathbb{E}[vec(u^{\otimes l})vec(u^{\otimes l})^\top]b \geq \frac{\Gamma\left(\frac{d}{2}\right)}{2^l\Gamma\left(l+\frac{d}{2}\right)} l!,$$

where $\Gamma(\cdot)$ is the Gamma function.

Next, we lower bound the projection of $\mathrm{vec}(T^*)$ on subspace $S_U^\perp$ by summation up the projections on the subspace bases, each of which can be bounded by Lemma 10. We give the proof of Theorem 1 as follows.

Proof of Theorem 1. Recall that W_l is the space of all vectorized symmetric tensors in $(\mathbb{R}^d)^{\otimes l}$. Due to the symmetry, the dimension of W_l is $\binom{d+l-1}{l}$. Since the dimension of S_U is at most dm, we know that the dimension of $S_U^\perp$ is at least $\binom{d+l-1}{l} - dm$. Assuming $S_U^\perp$ is an $\bar{m}$-dimensional space, we have $\bar{m} \geq \binom{d+l-1}{l} - dm \geq \frac{d^l}{l!} - dm$. Let $\{e_1, \cdots, e_{\bar{m}}\}$ be a set of orthonormal bases of $S_U^\perp$, and $\Pi_U^\perp$ be the projection matrix from $\mathbb{R}^{d^l}$ onto $S_U^\perp$, then we know that the smallest possible error that we can get given U is

$$\frac{1}{2}\mathbb{E}_{u^*}\left[\left\|\Pi_U^\perp \mathrm{vec}(T^*)\right\|_F^2\right] = \frac{1}{2}\mathbb{E}_{u^*}\left[\sum_{i=1}^{\bar{m}} \langle \mathrm{vec}(T^*), e_i\rangle^2\right] = \frac{1}{2}\sum_{i=1}^{\bar{m}} \mathbb{E}_{u^*}\left[\langle \mathrm{vec}(T^*), e_i\rangle^2\right],$$

where the expectation is taken over $u^* \sim \mathrm{Unif}(\mathbb{S}^{d-1})$.

By Lemma 10, we know that for any $i \in [m]$,

$$\begin{aligned}\mathbb{E}_{u^*}\left[\langle \mathrm{vec}(T^*), e_i\rangle^2\right] =& e_i^\top \mathbb{E}_{u^*}[(\mathrm{vec}([u^*]^{\otimes l})\mathrm{vec}([u^*]^{\otimes l})^\top]e_i \\ \geq& \frac{\Gamma\left(\frac{d}{2}\right)}{2^l\Gamma\left(l+\frac{d}{2}\right)} l! \geq \mu\frac{l!}{d^l},\end{aligned}$$

where μ is a positive constant only related to l.

Therefore,

$$\frac{1}{2}\mathbb{E}_{u^*}\left[\left\|\Pi_U^\perp T^*\right\|_F^2\right] = \frac{1}{2}\sum_{i=1}^{\bar{m}}\mathbb{E}_{u^*}\left[\langle \mathrm{vec}(T^*), e_i\rangle^2\right] \geq \left(\frac{d^l}{l!} - dm\right)\frac{\mu l!}{2d^l} = \frac{\mu}{2} - \frac{\mu l!}{2}\cdot\frac{m}{d^{l-1}}.$$

Note that we assume l is a constant. If $m = o(d^{l-1})$, i.e., $\frac{m}{d^{l-1}} = o(1)$, then the expectation of the smallest possible error is at least some constant. Thus, if we want the error to be $o(1)$, we must have $m = \Omega(d^{l-1})$. This finishes the proof of Theorem 1. □

A.1 Numerical verification of the lower bound

In this section, we plot the projection of the ground truth tensor on the orthogonal subspace $\mathbb{E}_{u^*}\left\|\Pi_U^\perp T^*\right\|_F^2$ under different dimension d and number of components m. For convenience, we only plot the lower bound for the projection that is $\left(\binom{d+l-1}{l} - dm\right)\frac{\Gamma\left(\frac{d}{2}\right)}{2^l\Gamma\left(l+\frac{d}{2}\right)} l!$ as we derived previously.

Figure 1 shows that under different dimensions, $\mathbb{E}_{u^*}\left\|\Pi_U^\perp T^*\right\|_F^2$ is at least a constant until $\log_d m$ gets close to $l-1=3$. As dimension d increases, the threshold when the orthogonal projection significantly drops becomes closer to 3.

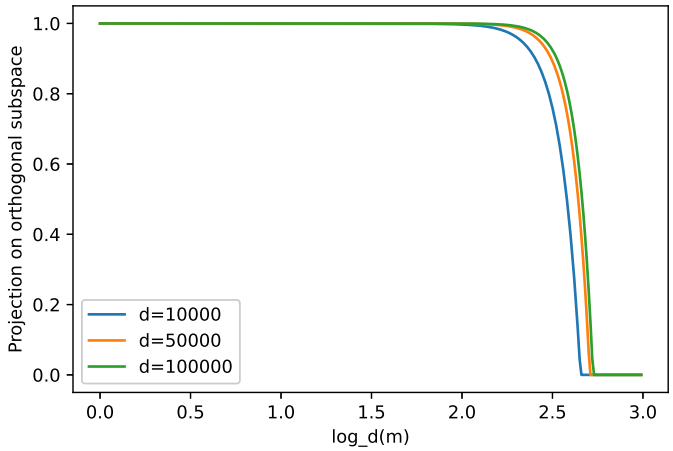

Figure 1: The projection of the ground truth tensor on the orthogonal subspace when $l = 4$.

A.2 Proofs of technical lemmas

To prove Lemma 10, we need another technical lemma, which is stated and proved below.

Lemma 11. *Let $u \in \mathbb{R}^d$ be a standard normal vector. For any vectorized symmetric l-th order tensor $b \in \mathbb{R}^{d^l}$ with unit norm, we have*

$$b^\top \mathbb{E}[vec(u^{\otimes l})vec(u^{\otimes l})^\top]b \geq l!.$$

Proof of Lemma 11. First we define the notion of symmetry: For a vector $x \in \mathbb{R}^{d^B}$, where $B \in \mathbb{N}^*$, if for all permutation σ of $[d]$, and for all $i_1, \cdots, i_B \in [d]$, $x_{i_1 i_2 \cdots i_B} = x_{\sigma(i_1)\sigma(i_2)\cdots\sigma(i_B)}$, then we say vector x is symmetric.

Besides, for a vector $v \in \mathbb{R}^{d^l}$, we use $v_{i_1,i_2,\cdots,i_l}$ to refer to $\text{Tensor}(v)_{i_1,i_2,\cdots,i_l}$ where Tensor is the inverse translation of vec, i.e., Tensor translates a $\mathbb{R}^{d^l}$ vector back to a $\mathbb{R}^{d^{\otimes l}}$ tensor. In other words, we use $v_{i_1,i_2,\cdots,i_l}$ to refer to the entry $v_{(i_1-1)d^{l-1}+(i_2-1)d^{l-2}+\cdots+(i_{l-1}-1)d+i_l}$. Similarly, for a $d^l \times d^l$ matrix M, we use $M_{i_1,i_2,\cdots,i_l,j_1,j_2,\cdots,j_l}$ to denote $\text{Tensor}(M)_{i_1,i_2,\cdots,i_l,j_1,j_2,\cdots,j_l}$, or in other words,
$M_{(i_1-1)d^{l-1}+(i_2-1)d^{l-2}+\cdots+(i_{l-1}-1)d+i_l,(j_1-1)d^{l-1}+(j_2-1)d^{l-2}+\cdots+(j_{l-1}-1)d+j_l}$.

Assume that $v = u^{\otimes l}$, then $v \in \mathbb{R}^{d^{\otimes l}}$, and v is a symmetric tensor. Note that

$$\forall i_1, \cdots, i_l \in [d], v_{i_1,\cdots,i_l} = u_{i_1} \cdots u_{i_l}.$$

Define $M \triangleq \text{vec}(v)\text{vec}(v)^\top = \text{vec}(u^{\otimes l})\text{vec}(u^{\otimes l})^\top$, then

$$M_{i_1,\cdots,i_l,j_1,\cdots,j_l} = u_{i_1} \cdots u_{i_l} \cdot u_{j_1} \cdots u_{j_l}.$$

By Wick's theorem, we know that

$$\mathbb{E}[M_{i_1,\cdots,i_l,j_1,\cdots,j_l}] = \sum_{\sigma \in P} \prod_{t \in [l]} \mathbb{E}[u_{\sigma(2t-1)} u_{\sigma(2t)}],$$

where P is the set that containing all distinct partitions of $S = \{i_1, \cdots, i_l, j_1, \cdots, j_l\}$ into l pairs. Each two variables in S are considered different even if the values of them are the same, e.g., $\{(i_1, i_2), (j_1, j_2)\}$ and $\{(i_1, j_2), (j_1, i_2)\}$ are different partitions even if $i_2 = j_2$. In other words, the partition is independent of the value of those variables. Thus, we can decompose matrix M into the sum of $(2l-1)!!$ matrices, i.e.,

$$M = \sum_{\sigma \in P} M_\sigma.$$

Assume σ_1 is the partition $\{(i_1, j_1), \cdots, (i_l, j_l)\}$, so

$$\mathbb{E}[M_{\sigma_1}] = \prod_{t \in [l]} \mathbb{E}[u_{i_t} u_{j_t}].$$

Since $\mathbb{E}[u_{i_t} u_{j_t}] = \mathbb{I}\{i_t = j_t\}$($\mathbb{I}$ is the indicator function), we know that all elements on the diagonal of $\mathbb{E}[M_{\sigma_1}]$ are 1, and all other elements are 0, which means that $\mathbb{E}[M_{\sigma_1}]$ is the identity matrix. Hence,

$$b^\top \mathbb{E}[M_{\sigma_1}] b = 1.$$

Note that b is a symmetric vector, meaning $b^\top \mathbb{E}[M_{\sigma_1}] b$ doesn't change if we permute $\{i_1, \cdots, i_l\}$. Thus, as long as each i is paired with a j in σ, we will have $b^\top \mathbb{E}[M_{\sigma_1}] b = b^\top \mathbb{E}[M_\sigma] b$. There are $n!$ such partitions, so summing them up gives us $l!$..

For any other partition σ, we can always permute $\{i_1, \cdots, i_l\}$ and $\{j_1, \cdots, j_l\}$ such that the partition becomes $\{(i_1, i_2), \cdots, (i_{2t-1}, i_{2t}), (j_1, j_2), \cdots, (j_{2t-1}, j_{2t}), (i_{2t+1}, j_{2t+1}), \cdots, (i_l, j_l)\}$. Then

$$\mathbb{E}[M_\sigma] = I_{d^{l-2t}} \otimes (ww^\top),$$

where $w \in \mathbb{R}^{d^{2t}}$ and $w_{i_1,\cdots,i_{2t}} = \mathbb{I}\{i_1 = i_2\} \cdots \mathbb{I}\{i_{2t-1} = i_{2t}\}$. Therefore, $\mathbb{E}[M_\sigma]$ is a positive semi-definite matrix, i.e., $b^\top \mathbb{E}[M_\sigma] b \geq 0$.

In a word, we can divide $\mathbb{E}[M]$ into the sum of two sets of matrices. In the symmetric sense, the first set of matrices are equivalent to identity matrices while the second set of matrices are equivalent to some semi-definite matrices. Therefore,

$$b^\top \mathbb{E}[\text{vec}(u^{\otimes l})\text{vec}(u^{\otimes l})^\top] b \geq l!.$$

□

Proof of Lemma 10. Let $u \in \mathbb{R}^d$ be a standard normal vector, i.e., $u \sim \mathcal{N}(0, I_d)$, then from Theorem 2 in Vignat and Bhatnagar (2008) we know that

$$b^\top \mathbb{E}\left[\operatorname{vec}\left((u/\|u\|)^{\otimes l}\right)\operatorname{vec}\left((u/\|u\|)^{\otimes l}\right)^\top\right] b = \frac{\Gamma\left(\frac{d}{2}\right)}{2^l\Gamma\left(l+\frac{d}{2}\right)} b^\top \mathbb{E}[\operatorname{vec}(u^{\otimes l})\operatorname{vec}(u^{\otimes l})^\top] b.$$

Furthermore, from Lemma 11 we know that

$$^\top \mathbb{E}[\operatorname{vec}(u^{\otimes l})\operatorname{vec}(u^{\otimes l})^\top] b \geq l!.$$

Note that $u/\|u\|$ is distributed as a uniform vector from the unit sphere $\mathbb{S}^{d-1}$. Combining the above equality and inequality, we finish the proof of this lemma. □

B Construction of Bad Local Minimum

In this section, we construct a bad local min for the vanilla loss function with vanilla parameterization of T. That is, $T := \sum_{i=1}^m c_i u_i^{\otimes l}$ and $f_v(U, C) = 1/2 \|T - T^*\|_F^2$. Recall Theorem 2 as follows.

Theorem 2. *Let $f_v(U, C)$ be as defined in Equation 2. Assume $l \geq 3, d > r \geq 1$ and $m \geq r(l+1)+1$. There exists a symmetric ground truth tensor T^* with rank at most $r(l+1)+1$ such that a local minimum with function value $l(l-1)r/4$ exists while the global minimum has function value zero.*

In our example, the model fits one direction in T^* but misses all the other directions. Moving any component towards one of the missing directions would actually make the approximation worse because of the cross terms. The proof of Theorem 2 is in Section B.1.

We also extend the local min to the vanilla loss function with 2-homogeneous parameterization of T. That is, $T := \sum_{i=1}^m a_i c_i^{l-2} u_i^{\otimes l}$ and $f(U, C) = 1/2 \|T - T^*\|_F^2$. For simplicity, we assume half of the a_i's are 1's.

Theorem 4. *Let $f(U, C) := 1/2 \left\|\sum_{i=1}^m a_i c_i^{l-2} u_i^{\otimes l} - T^*\right\|_F^2$. Assume $\lfloor m/2 \rfloor$ of a_i's are 1's and the remaining are -1's. Assume $l \geq 3, d-2 \geq r \geq 1$ and $m \geq 4r(l+1)+2$. There exists a symmetric ground truth tensor T^* with rank at most $2r(l+1)+2$ such that a local minimum with function value $l(l-1)r/2$ exists while the global minimum has function value zero.*

In the above Theorem, we treat c_i's as separate variables from u_i's. That is, at a local min (U, C), we allow arbitrary perturbations to c_i's regardless of the perturbations to u_i's and show none of these perturbations can decrease the function value. Note our result trivially extends to the case when $c_i = 1/\|u_i\|$ since the coupling between c_i and u_i only restricts the set of possible perturbations to (U, C). The detailed proof of Theorem 4 is in Section B.1.

B.1 Detailed Proofs

Proof of Theorem 2. In this proof, we first construct a ground truth tensor and a local min with non-zero loss. To further prove this local min is indeed spurious, we show there exists a global min with zero loss under the same ground truth tensor.

We first define the local min. For every $i \in [m]$, let c_i be 1 and u_i be $e_1/m^{\frac{1}{l}}$. Then, we know at this point $T = e_1^{\otimes l}$.

We define the ground truth tensor T^* by defining the residual $R := T - T^*$. The residual R is defined as the summation of $\hat{R}$ and all its permutation. We define $\hat{R}$ as follows,

$$\hat{R} := \sum_{j=2}^{r+1} e_j^{\otimes 2} \otimes e_1^{\otimes l-2}.$$

Then, R is defined as the summation of all $\binom{l}{2}$ permutations of $\hat{R}$. It's clear that R is symmetric and therefore T^* is symmetric.

Let U be a $d \times m$ matrix whose i-th row is u_i, and C be an $m \times m$ diagonal matrix with $C_{ii} = c_i, \forall i \in [m]$. Suppose we perform a local change to U and C such that $U' = U + \Delta U, C' = C + \Delta C$ and $\|\Delta U\|_F, \|\Delta C\|_F \to 0$. We prove that for any $\Delta U, \Delta C$, we have $f(U', C') \geq f(U, C)$. Let's first show that the gradient at U is zero, which means there is no locally first-order change on the function value.

First-order Change Let's first show the gradient of f_v w.r.t. U and C at (U, C) is zero. Here we first compute the gradient in terms of one column u_i,

$$\forall i \in [m], \nabla_{u_i} f_v(U, C) = lR(u_i^{\otimes l-1}, I)c_i = \frac{l}{m^{\frac{l-1}{l}}} R(e_1^{\otimes l-1}, I).$$
$$\forall i \in [m], \nabla_{c_i} f_v(U, C) = R(u_i^{\otimes l}) = \frac{1}{m} R(e_1^{\otimes l}).$$

In order to compute $R(e_1^{\otimes l-1}, I)$, we first consider $\hat{R}(e_1^{\otimes l-1}, I)$. We have

$$\hat{R}(e_1^{\otimes l-1}, I) = \sum_{j=2}^{r+1} e_j \langle e_j, e_1 \rangle \langle e_1, e_1 \rangle^{l-2} = 0.$$

Similarly,

$$\hat{R}(e_1^{\otimes l}) = \sum_{j=2}^{r+1} \langle e_j, e_1 \rangle^2 \langle e_1, e_1 \rangle^{l-2} = 0.$$

The computation for other permutations of $\hat{R}$ is the same. Overall, we have $\nabla f_v(U, C) = 0$.

Second-order change The second order change of $f_v(U, C)$ is as follows,

$$\frac{1}{2} \left\| \sum_{i=1}^{m} \left(c_i(\Delta u_i) \otimes u_i^{\otimes l-1} + c_i u_i \otimes (\Delta u_i) \otimes u_i^{\otimes l-2} + \cdots + c_i u_i^{\otimes l-1} \otimes (\Delta u_i) \right) + (\Delta c_i) u_i^{\otimes l} \right\|_F^2$$
$$+ \sum_{i=1}^{m} \left(l(l-1) R(u_i^{\otimes l-2}, \Delta u_i, \Delta u_i) c_i + lR(u_i^{\otimes l-1}, \Delta u_i) \Delta c_i \right).$$

The first term is always non-negative, and the second term can be further computed as follows:

$$l(l-1) \sum_{i=1}^{m} R(u_i^{l-2}, \Delta u_i, \Delta u_i) = l(l-1) \frac{1}{m^{(l-2)/l}} \sum_{i=1}^{m} R(e_1^{\otimes l-2}, \Delta u_i, \Delta u_i)$$
$$= l(l-1) \frac{1}{m^{(l-2)/l}} \sum_{i=1}^{m} \sum_{j=2}^{r+1} [\Delta u_i]_j^2.$$

Similar to the computations in the first-order change part, we know $R(u_i^{\otimes l-1}, \Delta u_i) = 0$. Therefore, the second-order change of $f(U, C)$ can be lower bounded by $l(l-1) \frac{1}{m^{(l-2)/l}} \sum_{i=1}^{m} \sum_{j=2}^{r+1} [\Delta u_i]_j^2$.

For any $\Delta U, \Delta C$, if there exists $i \in [m], 2 \leq j \leq r+1$ such that $[\Delta u_i]_j \neq 0$, we know the second order change is positive. Combining with the fact that the gradient is zero at (U, C), this implies the function value increases.

Otherwise, if $[\Delta u_i]_j = 0$ for all $i \in [m]$ and all $2 \leq j \leq r+1$, we know $\Delta u_i \in B$ for all $i \in [m]$ where B is the span of $\{e_1, e_k | r+2 \leq k \leq d\}$. Let the perturbed tensor be T', we know $T' - T$ lies in the $B^{\otimes l}$ subspace. Note perturbing c_i introduces changes in $e_1^{\otimes l}$ direction that is also in the $B^{\otimes l}$ subspace. This type of perturbation cannot decrease the function value because the residual R is orthogonal with $B^{\otimes l}$ subspace.

Overall, we have proved that (U, C) is a local minimizer. Notice that residual R contains $r\binom{l}{2}$ orthogonal components with unit norm. Therefore, the function value at (U, C) is $f(U, C) = \frac{1}{2} \|R\|_F^2 = \frac{1}{2} \times r \times \binom{l}{2} = \frac{l(l-1)r}{4}$.

Construction of global minimizer: Next we will show that when $m \geq r(l+1)+1$, there exists U and C such that $f(U,C)=0$. Therefore, the local minimizer we found above must be a spurious local minimizer. We only need to show that T^* can be expressed as the summation of $r(l+1)+1$ rank-one symmetric tensors.

Define $\hat{R}_j := e_j^{\otimes 2} \otimes e_1^{\otimes l-2}$, and define R_j to be the sum of all $\binom{l}{2}$ permutations of $\hat{R}_j$. Then we can write T^* as

$$T^* = e_1^{\otimes l} + \sum_{j=2}^{r+1} R_j.$$

Note that R_j is a symmetric tensor with entries equal to 1 if the index of the entry has 2 j's and $(l-2)$ 1's, and entries equal to 0 otherwise. Define $v_{i,j} := e_1 + b_{i,j}e_j$ where $b_{i,j} \in \mathbb{R}$, and consider the tensor $\bar{T}_j := \sum_{i=1}^{l+1} \bar{b}_{i,j} v_{i,j}^{\otimes l}$. Then we know that $\bar{T}_j$ is also a symmetric tensor with entries equal to $\sum_{i=1}^{l+1} \bar{b}_{i,j} b_{i,j}^k$ if the index of the entry has k j's and $(l-k)$ 1's, and entries equal to 0 otherwise. Therefore, if $\forall k \in \{0,1,\cdots,l\}\backslash\{2\}$, $\sum_{i=1}^{l+1} \bar{b}_{i,j} b_{i,j}^k = 0$ and $\sum_{i=1}^{l+1} \bar{b}_{i,j} b_{i,j}^2 = 1$, then $\bar{T}_j = R_j$. In other words, we want

$$\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ b_{1,j} & b_{2,j} & \cdots & b_{l+1,j} \\ b_{1,j}^2 & b_{2,j}^2 & \cdots & b_{l+1,j}^2 \\ \vdots & \vdots & \vdots & \vdots \\ b_{1,j}^l & b_{2,j}^l & \cdots & b_{l+1,j}^l \end{pmatrix} \begin{pmatrix} \bar{b}_{1,j} \\ \bar{b}_{2,j} \\ \vdots \\ \bar{b}_{l+1,j} \end{pmatrix}.$$

Denote the matrix in the middle by M_j, then

$$|M_j| = \prod_{1 \leq s < t \leq l+1} (b_{s,j} - b_{t,j}).$$

Thus, as long as the $b_{i,j}$'s are mutually different, the matrix M_j will be full rank, and there must exist a set of $\bar{b}_{i,j}$'s such that the equation above holds. In other words, we have shown that there exists such $b_{i,j}$'s and $\bar{b}_{i,j}$'s that $\forall 2 \leq j \leq r+1$, $\bar{T}_j = R_j$. Therefore, we know each R_j can be expressed as the summation of $(l+1)$ rank-one symmetric tensors.

To summarize, when $m \geq r(l+1)+1$, we can construct T^* such that there exists a local minimum with function value $\frac{l(l-1)r}{4}$ while the global minimum has function value zero.

□

Proof of Theorem 4. The proof is very similar as the proof of Theorem 2. The only difference is that in the 2-homogeneous, all the c_i's are positive and we need to rely on positive and negative a_i's to fit the ground truth tensors. We need to define a slightly different ground truth tensor and bad local min.

Same as in the proof of Theorem 2, we first construct a ground truth tensor and a local min with non-zero loss. To further prove this local min is indeed spurious, we show there exists a global min with zero loss under the same ground truth tensor.

We first define the local min. Let $m' = \lfloor m/2 \rfloor$. Without loss of generality, assume $a_i = 1$ for all $i \in [m']$ and $a_i = -1$ for all $i \in [m'+1, m]$. For any $i \in [m']$, let $u_i = \sqrt{1/m'} e_1$ and $c_i = 1/\left\|u_i\right\|$. For any $i \in [m'+1, m]$, let $u_i = \sqrt{1/(m-m')} e_d$ and $c_i = 1/\left\|u_i\right\|$. With this choice of parameters, it's not hard to verify that $T = e_1^{\otimes l} - e_d^{\otimes l}$.

We define the ground truth tensor T^* by defining the residual $R := T - T^*$. The residual R is defined as the summation of $\hat{R}$ and all its permutation, where $\hat{R}$ is defined as:

$$\hat{R} := \sum_{j=2}^{r+1} \left(e_j^{\otimes 2} \otimes e_1^{\otimes l-2} - e_j^{\otimes 2} \otimes e_d^{\otimes l-2} \right).$$

Since we assume $r \leq d-2$, we know $r+1 \leq d-1$ and e_j is orthogonal with e_1, e_d for all $2 \leq j \leq r+1$. Then, R is defined as the summation of all $\binom{l}{2}$ permutations of $\hat{R}$. It's clear that R is symmetric and T^* is also symmetric.

Suppose we perform a local change to U and C such that $U' = U+\Delta U, C' = C+\Delta C$ and $\|\Delta U\|_F, \|\Delta C\|_F \to 0$, where ΔC is a diagonal matrix. We prove that for all possible $\Delta U, \Delta C$, $f(U', C') \geq f(U, C)$.

First-order Change Let's first show the derivative of f in terms of all u_i's and c_i's at (U, C) is zero. This means there is no first order decrease direction at (U, C).

For any $i \in [m']$, we can compute the derivative in terms of u_i and c_i:

$$\begin{aligned}\nabla_{u_i} f(U,C) =& lR(u_i^{\otimes l-1}, I)c_i^{l-2} = \frac{l}{\sqrt{m'}} R(e_1^{\otimes l-1}, I),\\ \nabla_{c_i} f(U,C) =& (l-2)R(u_i^{\otimes l})c_i^{l-3} = \frac{l-2}{(m')^{3/2}} R(e_1^{\otimes l}).\end{aligned}$$

It's not hard to verify that $R(e_1^{\otimes l-1}, I) = 0$ and $R(e_1^{\otimes l}) = 0$ using the orthogonality between e_1 and e_j for all $2 \leq j \leq r+1$. For the same reason, we also have $\nabla_{u_i} f(U,C) = 0, \nabla_{c_i} f(U,C) = 0$ for all $i \in [m'+1, 2m]$.

Second-order change The second order change of $f(U', C')$ compared with $f(U, C)$ is as follows,

$$\begin{aligned}&\frac{1}{2}\left\|\sum_{i=1}^{m} a_i \left(c_i^{l-2}\left((\Delta u_i)\otimes u_i^{\otimes l-1} + u_i \otimes (\Delta u_i) \otimes u_i^{\otimes l-2} + \cdots + u_i^{\otimes l-1}\otimes(\Delta u_i)\right) + (l-2)c_i^{l-3}(\Delta c_i)u_i^{\otimes l}\right)\right\|_F^2\\ &+ \sum_{i=1}^{m} a_i \left(l(l-1)R(u_i^{\otimes l-2}, \Delta u_i, \Delta u_i)c_i^{l-2} + l(l-2)R(u_i^{\otimes l-1}, \Delta u_i)c_i^{l-3}\Delta c_i + (l-2)(l-3)R(u_i^{\otimes l})c_i^{l-4}(\Delta c_i)^2\right).\end{aligned}$$

(When $l = 3$, we do not have the c_i^{l-4} term)

The first term is always non-negative. By the previous argument, we also know $R(u_i^{\otimes l-1}, \Delta u_i) = 0$ and $R(u_i^{\otimes l}) = 0$. Therefore, the second order change is lower bounded by $\sum_{i=1}^{m} a_i l(l-1)R(u_i^{\otimes l-2}, \Delta u_i, \Delta u_i)c_i^{l-2}$. Let's first consider the components from 1 to m' for which $a_i = 1$,

$$\begin{aligned}l(l-1)a_i \sum_{i=1}^{m'} R(u_i^{\otimes l-2}, \Delta u_i, \Delta u_i)c_i^{l-2} =& l(l-1)\sum_{i=1}^{m'} R(e_1^{\otimes l-2}, \Delta u_i, \Delta u_i)\\ =& l(l-1)\sum_{i=1}^{m'}\sum_{j=2}^{r+1}[\Delta u_i]_j^2.\end{aligned}$$

For the components from $m'+1$ to m, we have

$$\begin{aligned}l(l-1)a_i \sum_{i=m'+1}^{m} R(u_i^{\otimes l-2}, \Delta u_i, \Delta u_i)c_i^{l-2} =& l(l-1)\sum_{i=m'+1}^{m} -R(e_d^{\otimes l-2}, \Delta u_i, \Delta u_i)\\ =& l(l-1)\sum_{i=m'+1}^{m}\sum_{j=2}^{r+1}[\Delta u_i]_j^2.\end{aligned}$$

Overall, we have

$$\sum_{i=1}^{m} a_i l(l-1)R(u_i^{\otimes l-2}, \Delta u_i, \Delta u_i)c_i^{l-2} \geq l(l-1)\sum_{i=1}^{m}\sum_{j=2}^{r+1}[\Delta u_i]_j^2.$$

For any $\Delta U, \Delta C$, if there exists $i \in [m], 2 \leq j \leq r+1$ such that $[\Delta u_i]_j \neq 0$, we know the second order change is positive. Combining with the fact that the gradient is zero at (U, C), this implies the function value increases.

Otherwise, if $[\Delta u_i]_j = 0$ for all $i \in [m]$ and all $2 \leq j \leq r+1$, we know $\Delta u_i \in B$ for all $i \in [m]$ where B is the span of $\{e_1, e_k | r+2 \leq k \leq d\}$. Let the perturbed tensor be T', we know $T' - T$ lies in the $B^{\otimes l}$ subspace.

Note perturbing c_i introduces changes in $e_1^{\otimes l}$ direction or $e_d^{\otimes l}$ direction that are also in the $B^{\otimes l}$ subspace. This type of perturbation cannot decrease the function value because the residual R is orthogonal with $B^{\otimes l}$ subspace.

Overall, we have proved that (U, C) is a local minimizer. Notice that residual R contains $2r\binom{l}{2}$ orthogonal components with unit norm. Therefore, the function value at (U, C) is $f(U, C) = \frac{1}{2}\|R\|_F^2 = \frac{1}{2} \times 2r \times \binom{l}{2} = \frac{l(l-1)r}{2}$.

Construction of global minimizer: Next, we show as long as $m \geq 4r(l+1)+2$, there exists parameters (U, C) such that $f(U, C) = 0$. To prove this, we first write T^* as summation of rank-one symmetric tensors.

For any $2 \leq j \leq r+1$, define $\hat{R}_{j,1} := e_j^{\otimes 2} \otimes e_1^{\otimes l-2}$ and $\hat{R}_{j,d} = e_j^{\otimes 2} \otimes e_d^{\otimes l-2}$, and define $R_{j,1}, R_{j,d}$ to be the sum of all $\binom{l}{2}$ permutations of $\hat{R}_{j,1}$ and $\hat{R}_{j,d}$ respectively. Then we can write T^* as

$$T^* = e_1^{\otimes l} - e_d^{\otimes l} - \sum_{j=2}^{r+1} R_{j,1} + \sum_{j=2}^{r+1} R_{j,d}.$$

Same as in the proof of Theorem 2, we can show each $R_{j,1}$ or $R_{j,d}$ can be written as the sum of $(l+1)$ rank-one symmetric tensors.

Therefore, we know the ground truth T^* can be expressed as the summation of $2 + 2r(l+1)$ rank-one symmetric tensors. For each component, we can re-scale it to make it fit the form of $a_i c_i^{l-2} u_i^{\otimes l}$, with $a_i = \pm 1$ and $c_i = 1/\|u_i\|$. In these rank-one tensors, at most $2r(l+1)+1$ has positive (negative) a_i. So, as long as $m' \geq 2r(l+1)+1$ or $m \geq 4r(l+1)+2$ our model is able to fit this ground truth tensor and achieve zero loss.

To summarize, when $m \geq 4r(l+1)+2$, we can construct T^* with rank at most $2r(l+1)+2$ such that there exists a local minimum with function value $\frac{l(l-1)r}{2}$ while the global minimum has function value zero. □

C Detailed Proofs of Theorem 3

In this section, we give the proof of Theorem 3. We first state a formal version of Theorem 3.

Theorem 5. *Given any target accuracy $\epsilon > 0$, there exists $m = O\left(\frac{r^{2.5l}}{\epsilon^5}\log(d/\epsilon)\right)$, $\lambda = O\left(\frac{\epsilon}{r^{0.5l}}\right)$, $\delta = O\left(\frac{\epsilon^{5l-1.5}}{d^{l-1.5}(\log(d/\epsilon))^{l+0.5}r^{2.5l^2-0.75l}}\right)$, $\eta = O\left(\frac{\epsilon^{15l-4.5}}{d^{3l-4.5}(\log(d/\epsilon))^{3l+1.5}r^{7.5l^2-2.25l}}\right)$ and $H = O\left(\frac{d^{3l-4.5}(\log(d/\epsilon))^{3l+2.5}r^{7.5l^2-1.75l}}{\epsilon^{15l-3.5}}\right)$ such that with probability at least 0.99, our algorithm finds a tensor T satisfying*

$$\|T - T^*\|_F \leq \epsilon,$$

within $K = O\left(\frac{r^{2l}}{\epsilon^4}\log(d/\epsilon)\right)$ epochs.

We follow the proof strategy outlined in Section 5.

As we discussed in Challenge 2, bad local minima exist for our loss function. Therefore, gradient descent might get stuck at a bad local minima. This issue is fixed in our algorithm by re-initializing one component at the beginning of each epoch. In Lemma 1, we show as long as the objective is large, there is at least a constant probability to improve the objective within one epoch. We state the formal version of Lemma 1 as follows. The proof of Lemma 12 is in Section C.2.

Lemma 12. *Let $(U_0', \bar{C}_0')$ and $(U_H, \bar{C}_H)$ be the parameters at the beginning of an epoch and the parameters at the end of the same epoch. For the target accuracy $\epsilon > 0$ in Theorem 5, assume $K \leq \frac{\lambda m}{14}$ and $\|T_0' - T^*\|_F \geq \epsilon$ where T_0' is the tensor with parameters $(U_0', \bar{C}_0')$. There exists $m = O\left(\frac{r^{2.5l}}{\epsilon^5}\log(d/\epsilon)\right)$, $\lambda = O\left(\frac{\epsilon}{r^{0.5l}}\right)$, $\delta =$*

$O\left(\frac{\epsilon^{5l-1.5}}{d^{l-1.5}(\log(d/\epsilon))^{l+0.5}r^{2.5l^2-0.75l}}\right)$, $\eta = O\left(\frac{\epsilon^{15l-4.5}}{d^{3l-4.5}(\log(d/\epsilon))^{3l+1.5}r^{7.5l^2-2.25l}}\right)$, $H = O\left(\frac{d^{3l-4.5}(\log(d/\epsilon))^{3l+2.5}r^{7.5l^2-1.75l}}{\epsilon^{15l-3.5}}\right)$, *such that with probability at least* $\frac{1}{6}$*, we have*

$$f(U_H, \bar{C}_H) - f(U_0', \bar{C}_0') \leq -\Omega\left(\frac{\epsilon^4}{r^{2l}\log(d/\epsilon)}\right).$$

We compliment this lemma by showing that even if an epoch does not improve the objective, it will not increase the function value by too much. The formal version of Lemma 2 is as follows. We prove Lemma 13 in Section C.1.

Lemma 13. *Assume* $K \leq \frac{\lambda m}{14}$, $\delta \leq \frac{\mu_1 \epsilon}{m^{\frac{3}{4}} d^{\frac{l-2}{2}} (m+K)^{\frac{l-1}{2}} \lambda^{\frac{1}{2}}}$, *and* $\eta \leq \frac{\mu_2 \lambda}{m^{\frac{1}{2}} d^{\frac{l-1}{2}} (m+K)^{\frac{l-2}{2}}}$ *for some constants* μ_1, μ_2, *and* $\frac{10}{m} \leq \lambda \leq 1$. *Let* $(U_0', \bar{C}_0')$ *and* $(U_H, \bar{C}_H)$ *be the parameters at the beginning of an epoch and the parameters at the end of the same epoch. Assume* $f(U_0', \bar{C}_0') \geq \epsilon^2$, *where* ϵ *is the target accuracy in Theorem 5. Then we have*

$$f(U_H, \bar{C}_H) - f(U_0', \bar{C}_0') \leq O\left(\frac{1}{\lambda m}\right).$$

From these two lemmas, we know that in each epoch, the loss function can decrease by $\Omega\left(\frac{\epsilon^4}{r^{2l}\log(d/\epsilon)}\right)$ with probability at least $\frac{1}{6}$, and even if we fail to decrease the function value, the increase of function value is at most $O\left(\frac{1}{\lambda m}\right)$. Therefore, choosing a large enough m, the function value decrease will dominate the increase. This allows us to prove Theorem 5.

Proof of Theorem 5. We use a contradiction proof to show that with high probability our algorithm finds a tensor T satisfying $\|T - T^*\|_F \leq \epsilon$ within K epochs. For the sake of contradiction, we assume $\|T - T^*\|_F > \epsilon$ through the first K epochs. Under this assumption, we show with high probability the function value will decrease below zero.

Note that under the choice of parameters of this theorem, all the conditions of Lemma 12 and Lemma 13 are satisfied. By Lemma 12, we know that with probability at least 1/6, the function value decreases by at least $\Lambda := \Omega\left(\frac{\epsilon^4}{r^{2l}\log(d/\epsilon)}\right)$ in each epoch. By Lemma 13, we show that the function value at most increases by $\Lambda' := O(\frac{1}{\lambda m})$ in each epoch. Using our choice of the parameters in Theorem 5, we know that $O(\frac{1}{\lambda m}) = O\left(\frac{\epsilon^4}{r^{2l}\log(d/\epsilon)}\right)$. Choosing a large enough constant factor for m ensures that $\Lambda' \leq \frac{\Lambda}{10}$.

For each $1 \leq k \leq K$, let $\mathcal{E}_k$ be the event that in the beginning of the k-th epoch, the reinitialized component $ac_0^{l-2}u_0^l$ has good correlation with the residual (see Lemma 18) and $\|P_S u_0\| \geq \frac{\mu\delta}{\sqrt{d}}$, where μ is some constant. We know $\mathcal{E}_k$'s are independent with each other and $\Pr[\mathcal{E}_k] \geq 1/6$. By Hoeffding's inequality, we know as long as $K \geq \mu'$ for certain constant μ', we have $\sum_{k=1}^K \mathbb{1}_{\mathcal{E}_k} \geq K/7$ with probability at least 0.99, where $\mathbb{1}_{\mathcal{E}_k}$ is the indicator function of event $\mathcal{E}_k$.

By the proof of Lemma 12, we know conditioning on $\mathcal{E}_k$, the function value decreases by at least Λ in the k-th epoch. Since $\sum_{k=1}^K \mathbb{1}_{\mathcal{E}_k} \geq K/7$, we know the total function value decrease is at least $K\Lambda/7 - K\Lambda/10 = K\Omega\left(\frac{\epsilon^4}{r^{2l}\log(d/\epsilon)}\right)$. Therefore, there exists $K = O\left(\frac{r^{2l}\log(d/\epsilon)}{\epsilon^4}\right)$ such that $K\Lambda/7 - K\Lambda/10 \geq 4$.

By the analysis in Lemma 16, the function value is upper bounded by 3 at initialization. However, with probability at least 0.99, the decrease of the function value is at least 4, meaning that the function value must be negative, which is a contradiction. Therefore, we know that with probability at least 0.99, our algorithm finds a tensor T satisfying $\|T - T^*\|_F \leq \epsilon$ within $K = O\left(\frac{r^{2l}\log(d/\epsilon)}{\epsilon^4}\right)$ epochs. □

C.1 Upper bound on function increase

In this section, we prove Lemma 13.

To prove the increase of f is bounded in one epoch, we identify all the possible ways that the loss can increase and upper bound each of them. We first show that a normal step (without re-initialization or scalar

mode switch) of the algorithm will not increase the objective function. Note that many parts of our proofs rely on an upperbound on function value. To get such a bound the proof includes an induction component: when we prove Lemma 14 and Lemma 15, we assume that the function value is upper bounded by a constant, and we will inductively prove that these conditions are satisfied in Lemma 16. This induction ensures that the conclusions of all the lemmas in this section hold throughout the entire algorithm.

The following lemma is a formal version of Lemma 3 in the main text.

Lemma 14. *Let $(U, \bar{C})$ be the parameters at the beginning of one iteration and let $U', \bar{C}'$ be the updated parameters (before potential scalar mode switch). Assuming $f(U, \bar{C}) \leq 10$, $\lambda \leq 1$, there exists constants μ_1, μ_2 such that*

$$f(U', \bar{C}') - f(U, \bar{C}) \leq -\frac{\eta}{l} \left\| \nabla_U f(U, \bar{C}) \right\|_F^2$$

as long as $\delta \leq \frac{\mu_1}{m^{\frac{1}{4}} \sqrt{\lambda} d^{\frac{l-2}{2}} (m+K)^{\frac{l-1}{2}}}, \eta \leq \frac{\mu_2 \lambda}{m^{\frac{1}{2}} d^{\frac{l-1}{2}} (m+K)^{\frac{l-2}{2}}}$.

Recall that in an iteration, we first update U by gradient descent, then update C and $\hat{C}$ by the updated value of U. The gradient descent step on U cannot increase the function value as long as the step size is small enough. The update on C and $\hat{C}$ can potentially increase the function value. In the proof of Lemma 14, we show the increase due to updating C and $\hat{C}$ is proportional to the decrease by updating U and smaller in scale.

Proof of Lemma 14.

According to the algorithm, each iteration contains two steps: update U as $U' \leftarrow U - \eta \nabla_U f(U, \bar{C})$; update c_i and $\hat{c}_i$ as $c_i' = c_i \frac{\|u_i\|}{\|u_i'\|}$ and $\hat{c}_i' = \hat{c}_i \frac{\|u_i\|}{\|u_i'\|}$. We can divide the function value change into these two steps: $f(U', \bar{C}') - f(U, \bar{C}) = (f(U', \bar{C}) - f(U, \bar{C})) + (f(U', \bar{C}') - f(U', \bar{C}))$. We will show that the function value decrease in the first step and does not increase by too much in the second step. At the end, we will combine them to show that overall the function value decreases.

Since we assume $f(U, \bar{C}) \leq 10$. According to the definition of the loss function, we know $\|T - T^*\|_F \leq \sqrt{20}, \sum_{i=1}^m \|u_i\|^2 \leq \frac{10}{\lambda}$. We also know that $\sum_{i=1}^m \|u_i\|^4 \leq \left(\sum_{i=1}^m \|u_i\|^2\right)^2 \leq \frac{100}{\lambda^2}$. For convenience, denote $\Gamma = 10, M_4^2 := \frac{100}{\lambda^2}$ and $M_2^2 := \frac{10}{\lambda}$.

$f(U', \bar{C}) - f(U, \bar{C})$ is negative: In the first step, we update U by gradient descent, which should decrease the function value as long as we choose the step size to be small enough. To prove that an inverse polynomially step size suffices, we need to bound the second derivative of f in terms of U at $(U'', \bar{C})$ for any $U'' \in \{(1-\theta)U + \theta U' | 0 \leq \theta \leq 1\}$. Let $\mathcal{H}''$ be the Hessian of f in terms of U at $(U'', \bar{C})$. We will bound the Frobenius norm of $\mathcal{H}''$.

Let's first show that $\|u_i''\| \leq (1 + 1/(4l)) \|u_i\|$ when η is small enough. Recall the derivative in u_i is,

$$\nabla_{u_i} f(U, \bar{C}) = l(T - T^*)(u_i^{\otimes(l-1)}, I) c_i^{l-2} a_i + \lambda l u_i.$$

Therefore, we can bound the derivative as

$$\left\| \nabla_{u_i} f(U, \bar{C}) \right\| \leq \left(l\sqrt{2\Gamma} (\sqrt{d(m+K)})^{l-2} + \lambda l \right) \|u_i\|.$$

Thus, as long as $\eta \leq \frac{1}{4l^2 (\sqrt{2\Gamma} (\sqrt{d(m+K)})^{l-2} + \lambda)}$, we have

$$\eta \left\| \nabla_{u_i} f(U, \bar{C}) \right\| \leq \eta \left(l\sqrt{2\Gamma} (\sqrt{d(m+K)})^{l-2} + \lambda l \right) \|u_i\| \leq \frac{1}{4l} \|u_i\|.$$

Since $u_i'' = u_i - \theta \eta \nabla_{u_i} f(U, \bar{C})$ for $0 \leq \theta \leq 1$, we know that

$$\begin{aligned} \|u_i''\| \leq & \|u_i\| + \eta \left\| \nabla_{u_i} f(U, \bar{C}) \right\| \\ \leq & (1 + \frac{1}{4l}) \|u_i\|, \end{aligned}$$

Let T'' be the tensor parameterized by $(U'', \bar{C})$. We can bound $\|T'' - T^*\|_F$ as follows,

$$\begin{aligned}
\|T'' - T^*\|_F \leq & \|T - T^*\|_F + \sum_{i=1}^{m}\sum_{k=1}^{l}\binom{l}{k}\|u_i\|^{l-k}\left\|\eta\nabla_{u_i} f(U, \bar{C})\right\|^k c_i^{l-2} \\
\leq & \|T - T^*\|_F + \sum_{i=1}^{m}\sum_{k=1}^{l}\binom{l}{k}\frac{1}{l^k}\|u_i\|^l c_i^{l-2} \\
\leq & \|T - T^*\|_F + l\sum_{i=1}^{m}\left(4(\sqrt{d(m+K)})^{l-2}(m+K)\delta^2 + \|u_i\|^2\right) \\
\leq & \|T - T^*\|_F + 4lm(\sqrt{d(m+K)})^{l-2}(m+K)\delta^2 + lM_2^2 \\
\leq & \sqrt{2\Gamma} + 2lM_2^2,
\end{aligned}$$

where the last inequality assumes $\delta \leq \frac{M_2}{\sqrt{4m(\sqrt{d(m+K)})^{l-2}(m+K)}}$. For convenience, denote $\beta := \sqrt{2\Gamma} + 2lM_2^2$.

With the bound on $\|T'' - T^*\|$ and $\|u_i''\|$, we are ready to bound the Frobenius norm of $\mathcal{H}''$. For each $i \in [m]$, we have

$$\frac{\partial}{\partial u_i} f(U'', \bar{C}) = l(T'' - T^*)((u_i'')^{l-1}, I)c_i^{l-2}a_i + \lambda l\left(\frac{\|u_i''\|}{\|u_i\|}\right)^{l-2} u_i''. \tag{3}$$

We know $\mathcal{H}''$ is a $dm \times dm$ matrix that contains $m \times m$ block matrices with dimension $d \times d$. Each block corresponds to the second-order derivative of $f(U'', \bar{C})$ in terms of u_i, u_j. We will bound the Frobenius norm of $\mathcal{H}''$ by bounding the Frobenius norm of each block.

For each i, we can compute $\frac{\partial^2}{\partial u_i \partial u_i} f(U'', \bar{C})$ as follows,

$$\begin{aligned}
\frac{\partial^2}{\partial u_i \partial u_i} f(U'', \bar{C}) = & l(l-1)(T'' - T^*)((u_i'')^{l-2}, I, I)c_i^{l-2}a_i + l^2 c_i^{2l-4}\|u_i''\|^{2l-4} u_i'' \otimes u_i'' \\
& + \lambda l\left(\frac{\|u_i''\|}{\|u_i\|}\right)^{l-2} I + \lambda l(l-2)\frac{\|u_i''\|^{l-4} u_i'' \otimes u_i''}{\|u_i\|^{l-2}}.
\end{aligned}$$

For the first term, we have $l(l-1)\left\|(T'' - T^*)((u_i'')^{l-2}, I, I)c_i^{l-2}\right\|_F \leq l^2\sqrt{e}\beta(\sqrt{d(m+K)})^{l-2}$ since $\|u_i''\| / \|u_i\| \leq (1 + 1/(4l))$.

For the second term, we have

$$l^2\left\|c_i^{2l-4}\|u_i''\|^{2l-4} u_i'' \otimes u_i''\right\|_F \leq l^2\sqrt{e}\max\{4(d(m+K))^{l-2}(m+K)\delta^2, \|u_i\|^2\}.$$

For the third term, we have

$$\left\|\lambda l\left(\frac{\|u_i''\|}{\|u_i\|}\right)^{l-2}\mathrm{vec}(I)\right\|_F \leq \lambda l\sqrt{e}\sqrt{d}.$$

For the fourth term, we have

$$\left\|\lambda l(l-2)\frac{\|u_i''\|^{l-4} u_i'' \otimes u_i''}{\|u_i\|^{l-2}}\right\|_F \leq \lambda l^2\sqrt{e}.$$

Combing the bounds on these terms and assuming $\lambda \leq 1$, we have

$$\left\|\frac{\partial^2}{\partial u_i \partial u_i} f(U'', \bar{C})\right\|_F \leq l^2\sqrt{e}\beta(\sqrt{d(m+K)})^{l-2} + l^2\sqrt{e}\max\{4d^{l-2}(m+K)^{l-1}\delta^2, \|u_i\|^2\} + 2l^2\sqrt{e}\sqrt{d}.$$

Thus,

$$\begin{aligned}\left\|\frac{\partial^2}{\partial u_i \partial u_i} f(U'', \bar{C})\right\|_F^2 \leq & 3el^4\beta^2 d^{l-2}(m+K)^{l-2} + 3el^4 \max\{16d^{2l-4}(m+K)^{2l-2}\delta^4, \|u_i\|^4\} + 12el^4 d \\ \leq & 15el^4\beta^2 d^{l-2}(m+K)^{l-2} + 3el^4 \max\{16d^{2l-4}(m+K)^{2l-2}\delta^4, \|u_i\|^4\}.\end{aligned}$$

For each pair of $i \neq j$, we can compute $\frac{\partial^2}{\partial u_i \partial u_j} f(U'', \bar{C})$ as follows

$$\frac{\partial^2}{\partial u_i \partial u_j} f(U'', C) = l^2 a_i a_j c_i^{l-2} c_j^{l-2} \left\langle u_i'', u_j'' \right\rangle^{l-2} u_i'' \otimes u_j''.$$

The Frobenius norm square can be bounded as

$$\begin{aligned}\left\|\frac{\partial^2}{\partial u_i \partial u_j} f(U'', \bar{C})\right\|_F^2 \leq & el^4 \max\{4d^{l-2}(m+K)^{l-1}\delta^2, \|u_i\|^2\} \max\{4d^{l-2}(m+K)^{l-1}\delta^2, \|u_j\|^2\} \\ \leq & el^4 \max\{\max\{\|u_i\|^2, \|u_j\|^2\}^2, 16d^{2l-4}(m+K)^{2l-2}\delta^4\} \\ \leq & el^4 \left(\|u_i\|^4 + \|u_j\|^4 + 16d^{2l-4}(m+K)^{2l-2}\delta^4\right)\end{aligned}$$

Summing over the bounds on blocks, we can bound the Frobeneius norm of $\mathcal{H}''$,

$$\begin{aligned}\|\mathcal{H}''\|_F^2 = & \sum_{i,j} \left\|\frac{\partial^2}{\partial u_i \partial u_j} f(U'', \bar{C})\right\|_F^2 \\ \leq & 15eml^4\beta^2 d^{l-2}(m+K)^{l-2} + 48eml^4 d^{2l-4}(m+K)^{2l-2}\delta^4 + 3el^4 \sum_{i=1}^m \|u_i\|^4 \\ & + (m-1)el^4 \sum_{i=1}^m \|u_i\|^4 + 16el^4 m(m-1)d^{2l-4}(m+K)^{2l-2}\delta^4 \\ = & 15eml^4\beta^2 d^{l-2}(m+K)^{l-2} + 16el^4 m(m+2)d^{2l-4}(m+K)^{2l-2}\delta^4 + (m+2)el^4 \sum_{i=1}^m \|u_i\|^4 \\ \leq & 15eml^4\beta^2 d^{l-2}(m+K)^{l-2} + 2(m+2)el^4 M_4^2\end{aligned}$$

where the last inequality assumes $\delta \leq \left(\frac{M_4^2}{16md^{2l-4}(m+K)^{2l-2}}\right)^{1/4}$.

Denoting $L_1 := \sqrt{15eml^4\beta^2 d^{l-2}(m+K)^{l-2} + 2(m+2)el^4 M_4^2}$, we have

$$f(U', \bar{C}) - f(U, \bar{C}) \leq -\eta \left\|\nabla_U f(U, \bar{C})\right\|_F^2 + \frac{\eta^2 L_1}{2} \left\|\nabla_U f(U, \bar{C})\right\|_F^2$$

$f(U', \bar{C}') - f(U', \bar{C})$ is bounded: Next, we show that setting c_i' as $c_i \frac{\|u_i\|}{\|u_i'\|}$ and $\hat{c}_i'$ as $\hat{c}_i \frac{\|u_i\|}{\|u_i'\|}$ does not increase the function value by too much. We use $\nabla_{\hat{u}_i} f$ to denote the gradient of u_i through c_i and $\hat{c}_i$, which means

$$\nabla_{\hat{u}_i} f = \frac{\partial f}{\partial c_i}\frac{\partial c_i}{\partial u_i} + \frac{\partial f}{\partial \hat{c}_i}\frac{\partial \hat{c}_i}{\partial u_i}.$$

In the following we first bound the Frobenius norm of the Hessian of f in terms of $\hat{U}$ evaluated at $(U', \bar{C}'')$ for any $C'' \in \{\text{diag}(c_1'', \cdots, c_m'') | c_i'' = c_i \frac{\|u_i\|}{\left\|(1-\theta)u_i + \theta u_i'\right\|}, 0 \leq \theta \leq 1\}$ and $\hat{C}'' \in \{\text{diag}(\hat{c}_1'', \cdots, \hat{c}_m'') | \hat{c}_i'' = \hat{c}_i \frac{\|u_i\|}{\left\|(1-\theta)u_i + \theta u_i'\right\|}, 0 \leq \theta \leq 1\}$. We denote the Hessian at $(U', \bar{C}'')$ as $\hat{\mathcal{H}}''$, which is a $md \times md$ matrix.

Hessian $\hat{\mathcal{H}}''$ contains $m \times m$ blocks with dimension $d \times d$, each of which corresponds to $\frac{\partial^2}{\partial \hat{u}_i \partial \hat{u}_j} f(U', \bar{C}'')$ for some $(i, j) \in [m] \times [m]$.

Note that $\frac{\|u_i'\|}{\|u_i''\|} \leq 1 + 1/l$ since $\|u_i' - u_i\| \leq 1/(4l) \|u_i\|$. Let T''' be the tensor corresponds to $(U', \bar{C}'')$, we can bound $\|T''' - T^*\|_F$ as follows,

$$\begin{aligned}
\|T''' - T^*\|_F \leq & \left\| \sum_{i=1}^m a_i (c_i'')^{l-2} (u_i')^{\otimes l} \right\|_F + \|T^*\|_F \\
\leq & e \left(\sum_{i=1}^m \|u_i\|^2 + 4(\sqrt{d(m+K)})^{l-2} m(m+K)\delta^2 \right) + 1 \\
\leq & 2eM_2^2 + 1,
\end{aligned}$$

where the last step assumes $\delta \leq \frac{M_2}{\sqrt{4(\sqrt{d(m+K)})^{l-2} m(m+K)}}$. For convenience, denote $\alpha := 2eM_2^2 + 1$.

Let's first compute the derivative of f in terms of $\hat{u}_i$,

$$\begin{aligned}
& \frac{\partial}{\partial \hat{u}_i} f(U', \bar{C}'') \qquad (4) \\
= & -(l-2) a_i (T''' - T^*)((u_i')^{\otimes l}) \frac{u_i''}{\|u_i''\|^l} \left((\sqrt{d(m+K)})^{l-2} \mathbb{1}_{c_i'' = \sqrt{d(m+K)}/\|u_i''\|} + \mathbb{1}_{c_i'' = 1/\|u_i''\|} \right) \\
& - \lambda (l-2) \frac{u_i''}{\|u_i''\|^l} \|u_i'\|^l . \qquad (5)
\end{aligned}$$

For each i, we have

$$\begin{aligned}
& \frac{\partial^2}{\partial \hat{u}_i \partial \hat{u}_i} f(U', \bar{C}'') \\
= & -(l-2) a_i (T''' - T^*)((u_i')^{\otimes l}) \frac{I}{\|u_i''\|^l} \left((\sqrt{d(m+K)})^{l-2} \mathbb{1}_{c_i'' = \sqrt{d(m+K)}/\|u_i''\|} + \mathbb{1}_{c_i'' = 1/\|u_i''\|} \right) \\
& + l(l-2) a_i (T''' - T^*)((u_i')^{\otimes l}) \frac{u_i''(u_i'')^\top}{\|u_i''\|^{l+2}} \left((\sqrt{d(m+K)})^{l-2} \mathbb{1}_{c_i'' = \sqrt{d(m+K)}/\|u_i''\|} + \mathbb{1}_{c_i'' = 1/\|u_i''\|} \right) \\
& + (l-2)^2 \|u_i'\|^{2l} \frac{u_i''(u_i'')^\top}{\|u_i''\|^{2l}} \left(d^{l-2} (m+K)^{l-2} \mathbb{1}_{c_i'' = \sqrt{d(m+K)}/\|u_i''\|} + \mathbb{1}_{c_i'' = 1/\|u_i''\|} \right) \\
& - \lambda (l-2) \frac{I}{\|u_i''\|^l} \|u_i'\|^l \\
& + \lambda l (l-2) \frac{u_i''(u_i'')^\top}{\|u_i''\|^{l+2}} \|u_i'\|^l .
\end{aligned}$$

We bound its Frobenius norm square by

$$\begin{aligned}
& \left\| \frac{\partial^2}{\partial \hat{u}_i \partial \hat{u}_i} f(U', \bar{C}'') \right\|_F^2 \\
\leq & 5 \left(l^2 \alpha^2 e^2 d^{l-1} (m+K)^{l-2} + l^4 \alpha^2 e^2 d^{l-2} (m+K)^{l-2} \right) \\
& + 5 \left(l^4 e^4 \max\{\|u_i\|^4, 16(m+K)^{2l-2} \delta^4 d^{2l-4}\} + \lambda^2 l^2 d e^2 + \lambda^2 l^4 e^2 \right) \\
\leq & 20 e^2 \alpha^2 l^4 d^{l-1} (m+K)^{l-2} + 5 e^4 l^4 \left(\|u_i\|^4 + 16(m+K)^{2l-2} \delta^4 d^{2l-4} \right),
\end{aligned}$$

where we assume that $\lambda \leq 1$.

For $i \neq j$, we have

$$\begin{aligned}
&\frac{\partial^2}{\partial \hat{u}_i \partial \hat{u}_j} f(U', \bar{C}'') \\
=&(l-2)^2 (\langle u_i', u_j' \rangle^l) \frac{u_i''(u_j'')^\top}{\|u_i''\|^l \left\|u_j''\right\|^l} \\
&\cdot \left((\sqrt{d(m+K)})^{l-2} \mathbb{1}_{c_i'' = \sqrt{d(m+K)}/\|u_i''\|} + \mathbb{1}_{c_i''=1/\|u_i''\|} \right) \\
&\cdot \left((\sqrt{d(m+K)})^{l-2} \mathbb{1}_{c_j'' = \sqrt{d(m+K)}/\|u_j''\|} + \mathbb{1}_{c_j''=1/\|u_j''\|} \right)
\end{aligned}$$

We can bound its Frobenius norm by

$$\begin{aligned}
\left\| \frac{\partial^2}{\partial \hat{u}_i \partial \hat{u}_j} f(U', \bar{C}'') \right\|_F^2 \leq& l^4 e^4 \max\{4d^{l-2}(m+K)^{l-1}\delta^2, \|u_i\|^2\} \max\{4d^{l-2}(m+K)^{l-1}\delta^2, \|u_j\|^2\} \\
\leq& l^4 e^4 \left(\|u_i\|^4 + \|u_j\|^4 + 16(m+K)^{2l-2}\delta^4 d^{2l-4} \right)
\end{aligned}$$

Combing the bounds on all blocks, we have

$$\begin{aligned}
\left\| \hat{\mathcal{H}}'' \right\|_F^2 \leq& \sum_{i=1}^m \left(20e^2\alpha^2 l^4 d^{l-1}(m+K)^{l-2} + 5e^4 l^4 \left(\|u_i\|^4 + 16(m+K)^{2l-2}\delta^4 d^{2l-4} \right) \right) \\
&+ \sum_{\substack{i,j \in [m] \\ i \neq j}} l^4 e^4 \left(\|u_i\|^4 + \|u_j\|^4 + 16(m+K)^{2l-2}\delta^4 d^{2l-4} \right) \\
\leq& 20me^2\alpha^2 l^4 d^{l-1}(m+K)^{l-2} + 80ml^4 e^4 (m+K)^{2l-2}\delta^4 d^{2l-4} + 5l^4 e^4 \sum_{i=1}^m \|u_i\|^4 \\
&+ 16m^2 l^4 e^4 (m+K)^{2l-2}\delta^4 d^{2l-4} + 2l^4 e^4 m \sum_{i=1}^m \|u_i\|^4 \\
\leq& 20me^2\alpha^2 l^4 d^{l-1}(m+K)^{l-2} + 7l^4 e^4 m M_4^2 + 96m^2 l^4 e^4 (m+K)^{2l-2}\delta^4 d^{2l-4} \\
\leq& 20me^2\alpha^2 l^4 d^{l-1}(m+K)^{l-2} + 8l^4 e^4 m M_4^2,
\end{aligned}$$

where the last inequality assumes $\delta \leq \left(\frac{M_4^2}{96m(m+K)^{2l-2}d^{2l-4}} \right)^{1/4}$.

Denote $L_2 := \sqrt{20me^2\alpha^2 l^4 d^{l-1}(m+K)^{l-2} + 8l^4 e^4 m M_4^2}$. Then, we have

$$f(U', \bar{C}') - f(U', \bar{C}) \leq \left\langle \nabla_{\hat{U}} f(U', \bar{C}), -\eta \nabla_U f(U, \bar{C}) \right\rangle + \frac{\eta^2 L_2}{2} \left\| \nabla_U f(U, \bar{C}) \right\|_F^2.$$

From equation 3 and 5 we know that $\forall i \in [m]$, $\nabla_{\hat{u}_i} f(U, \bar{C}) = -\frac{l-2}{l} \cdot \frac{u_i u_i^\top (\nabla_{u_i} f(U, \bar{C}))}{\|u_i\|^2}$. Thus,

$$\left\| \nabla_{\hat{U}} f(U, \bar{C}) \right\|_F \leq \frac{l-2}{l} \left\| \nabla_U f(U, \bar{C}) \right\|_F.$$

In order to bound $\left\| \nabla_{\hat{U}} f(U', \bar{C}) \right\|_F$, we still need to show that $\nabla_{\hat{U}} f(U', \bar{C})$ is close to $\nabla_{\hat{U}} f(U, \bar{C})$.

Bounding $\left\| \nabla_{\hat{U}} f(U', \bar{C}) - \nabla_{\hat{U}} f(U, \bar{C}) \right\|_F$: Define U'' as $(1-\theta)U + \theta U'$ for all $0 \leq \theta \leq 1$. We will show that the derivative of $\nabla_{\hat{U}} f$ in U evaluated $(U'', \bar{C})$ is bounded. We denote this derivative as $\tilde{\mathcal{H}}''$ that is a

$dm \times dm$ matrix. Matrix $\tilde{\mathcal{H}}''$ contains $m \times m$ blocks each of which has dimension $d \times d$ and corresponds to $\frac{\partial^2}{\partial \hat{u}_i \partial u_j} f(U'', \bar{C})$. Denote T'' as the tensor parameterized by $(U'', \bar{C})$. Recall that,

$$\begin{aligned}
&\frac{\partial}{\partial \hat{u}_i} f(U'', \bar{C}) \\
=& -(l-2) a_i (T'' - T^*)((u_i'')^{\otimes l}) \frac{u_i}{\|u_i\|^l} \left((\sqrt{d(m+K)})^{l-2} \mathbb{1}_{c_i = \sqrt{d(m+K)}/\|u_i\|} + \mathbb{1}_{c_i = 1/\|u_i\|} \right) \\
&- \lambda (l-2) \frac{u_i}{\|u_i\|^l} \|u_i''\|^l .
\end{aligned}$$

For any $i \in [m]$, we have

$$\begin{aligned}
&\frac{\partial^2}{\partial u_i \partial \hat{u}_i} f(U'', \bar{C}) \\
=& -l(l-2) a_i (T'' - T^*)((u_i'')^{\otimes l-1}, I) \frac{u_i^\top}{\|u_i\|^l} \left((\sqrt{d(m+K)})^{l-2} \mathbb{1}_{c_i = \sqrt{d(m+K)}/\|u_i\|} + \mathbb{1}_{c_i = 1/\|u_i\|} \right) \\
&- l(l-2) c_i^{l-2} \|u_i''\|^{2l-2} \frac{u_i (u_i'')^\top}{\|u_i\|^l} \left((\sqrt{d(m+K)})^{l-2} \mathbb{1}_{c_i = \sqrt{d(m+K)}/\|u_i\|} + \mathbb{1}_{c_i = 1/\|u_i\|} \right) \\
&- \lambda l(l-2) \frac{u_i (u_i'')^\top}{\|u_i\|^l} \|u_i''\|^{l-2} .
\end{aligned}$$

The Frobenius norm of $\frac{\partial^2}{\partial u_i \partial \hat{u}_i} f(U'', \bar{C})$ can be bounded as follows,

$$\begin{aligned}
\left\| \frac{\partial^2}{\partial u_i \partial \hat{u}_i} f(U'', \bar{C}) \right\|_F \leq& \sqrt{e} l^2 \beta (\sqrt{d(m+K)})^{l-2} + \sqrt{e} l^2 \max \left(\|u_i\|^2, 4 d^{l-2} (m+K)^{l-1} \delta^2 \right) + \sqrt{e} \lambda l^2 \\
\leq& 2\sqrt{e} l^2 \beta (\sqrt{d(m+K)})^{l-2} + \sqrt{e} l^2 \max \left(\|u_i\|^2, 4 d^{l-2} (m+K)^{l-1} \delta^2 \right),
\end{aligned}$$

where the last inequality assumes $\lambda \leq 1$. Therefore,

$$\left\| \frac{\partial^2}{\partial u_i \partial \hat{u}_i} f(U'', \bar{C}) \right\|_F^2 \leq 8 e l^4 \beta^2 d^{l-2} (m+K)^{l-2} + 2 e l^4 \max \left(\|u_i\|^4, 16 d^{2l-4} (m+K)^{2l-2} \delta^4 \right)$$

For $i \neq j$, we have

$$\begin{aligned}
&\frac{\partial^2}{\partial u_j \partial \hat{u}_i} f(U'', \bar{C}) \\
=& -l(l-2) a_i a_j c_j^{l-2} \left\langle u_i'', u_j'' \right\rangle^{l-1} \frac{u_i (u_i'')^\top}{\|u_i\|^l} \left((\sqrt{d(m+K)})^{l-2} \mathbb{1}_{c_i = \sqrt{d(m+K)}/\|u_i\|} + \mathbb{1}_{c_i = 1/\|u_i\|} \right)
\end{aligned}$$

The Frobenius norm square can be bounded as

$$\begin{aligned}
\left\| \frac{\partial^2}{\partial \hat{u}_i \partial u_j} f(U'', C) \right\|_F^2 \leq& e l^4 \max\{4 d^{l-2} (m+K)^{l-1} \delta^2, \|u_i\|^2\} \max\{4 d^{l-2} (m+K)^{l-1} \delta^2, \|u_j\|^2\} \\
\leq& e l^4 \left(\|u_i\|^4 + \|u_j\|^4 + 16 (m+K)^{2l-2} \delta^4 d^{2l-4} \right).
\end{aligned}$$

Summing over the bounds on blocks, we can bound the Frobeneius norm of $\tilde{\mathcal{H}}''$,

$$\begin{aligned}
&\left\|\tilde{\mathcal{H}}''\right\|_F^2 \\
=&\sum_{i,j}\left\|\frac{\partial^2}{\partial u_j \partial \hat{u}_i} f(U'', \bar{C})\right\|_F^2 \\
\leq& 8mel^4\beta^2 d^{l-2}(m+K)^{l-2} + 2el^4\sum_{i=1}^m \|u_i\|^4 + 32mel^4 d^{2l-4}(m+K)^{2l-2}\delta^4 \\
&+ 2mel^4\sum_{i=1}^m \|u_i\|^4 + 16m^2el^4(m+K)^{2l-2}\delta^4 d^{2l-4} \\
\leq& 8mel^4\beta^2 d^{l-2}(m+K)^{l-2} + 48m^2el^4 d^{2l-4}(m+K)^{2l-2}\delta^4 + 3el^4 mM_4^2 \\
\leq& 8mel^4\beta^2 d^{l-2}(m+K)^{l-2} + 4el^4 mM_4^2,
\end{aligned}$$

where the last inequality assumes $\delta \leq \left(\frac{M_4^2}{48md^{2l-4}(m+K)^{2l-2}}\right)^{1/4}$.

Denoting $L_3 := \sqrt{8mel^4\beta^2 d^{l-2}(m+K)^{l-2} + 4el^4 mM_4^2}$, we have

$$\left\|\nabla_{\hat{U}} f(U', \bar{C}) - \nabla_{\hat{U}} f(U, \bar{C})\right\|_F \leq L_3 \left\|\eta \nabla_U f(U, \bar{C})\right\|_F \leq \frac{1}{3l}\left\|\nabla_U f(U, \bar{C})\right\|_F,$$

where the second inequality assumes $\eta \leq \frac{1}{3L_3 l}$. Therefore, we have

$$\begin{aligned}
\left\|\nabla_{\hat{U}} f(U', \bar{C})\right\|_F \leq \left\|\nabla_{\hat{U}} f(U, \bar{C})\right\|_F + \frac{1}{3l}\left\|\nabla_U f(U, \bar{C})\right\|_F &\leq \left(\frac{l-2}{l} + \frac{1}{3l}\right)\left\|\nabla_U f(U, \bar{C})\right\|_F \\
&\leq \left(1 - \frac{5}{3l}\right)\left\|\nabla_U f(U, \bar{C})\right\|_F.
\end{aligned}$$

Overall, we have proved that as long as η is small enough,

$$\begin{aligned}
f(U', \bar{C}') - f(U, \bar{C}) =& f(U', \bar{C}) - f(U, \bar{C}) + f(U', \bar{C}') - f(U', \bar{C}) \\
\leq& -\eta\left\|\nabla_U f(U, \bar{C})\right\|_F^2 + \frac{\eta^2 L_1}{2}\left\|\nabla_U f(U, \bar{C})\right\|_F^2 \\
&+ \eta\left\|\nabla_{\hat{U}} f(U', \bar{C})\right\|_F^2 + \frac{\eta^2 L_2}{2}\left\|\nabla_{\hat{U}} f(U', \bar{C})\right\|_F^2 \\
\leq& -\eta\left\|\nabla_U f(U, \bar{C})\right\|_F^2 + \frac{\eta^2 L_1}{2}\left\|\nabla_U f(U, \bar{C})\right\|_F^2 \\
&+ \eta\left(1 - \frac{5}{3l}\right)\left\|\nabla_U f(U, \bar{C})\right\|_F^2 + \frac{\eta^2 L_2}{2}\left\|\nabla_U f(U, \bar{C})\right\|_F^2 \\
\leq& -\frac{\eta}{l}\left\|\nabla_U f(U, \bar{C})\right\|_F^2,
\end{aligned}$$

where the last inequality assumes $\eta \leq \frac{4}{3l(L_1+L_2)}$. Combining all the bounds on δ, η, we know there exists constant μ_1, μ_2 such that as long as $\delta \leq \frac{\mu_1}{m^{\frac{1}{4}}\sqrt{\lambda} d^{\frac{l-2}{2}}(m+K)^{\frac{l-1}{2}}}, \eta \leq \frac{\mu_2 \lambda}{m^{\frac{1}{2}} d^{\frac{l-1}{2}}(m+K)^{\frac{l-2}{2}}}$, we have

$$f(U', \bar{C}') - f(U, \bar{C}) \leq -\frac{\eta}{l}\left\|\nabla_U f(U, \bar{C})\right\|_F^2.$$

□

Then, we know in an epoch, the function value can only increase because of the initialization and the scalar mode switches. In Lemma 15, we show these operations cannot increase the function by too much. Note that Lemma 15 is the formal version of Lemma 4 together with the bound for scalar mode switches in the main text (these two arguments in the main text correspond to the two claims in Lemma 15).

Lemma 15. *Assume* $f(U_0', \bar{C}_0') \le \tilde{\Gamma} \le 10$ *at the beginning of an epoch,* $\delta \le \frac{\mu_1\sqrt{\tilde{\Gamma}}}{m^{\frac{3}{4}} d^{\frac{l-2}{2}} (m+K)^{\frac{l-1}{2}} \lambda^{\frac{1}{2}}}$, *and* $\eta \le \frac{\mu_2 \lambda}{m^{\frac{1}{2}} d^{\frac{l-1}{2}} (m+K)^{\frac{l-2}{2}}}$ *for some constants* μ_1, μ_2. *Also assume that* $\lambda m \ge 10$. *Denote the parameters at the end of this epoch as* $(U_H, \bar{C}_H)$, *then*

$$f(U_H, \bar{C}_H) \le \exp\left(\frac{14}{\lambda m}\right) \tilde{\Gamma}.$$

Proof of Lemma 15. By Lemma 14, we know the function value does not increase in any iteration (before potential scalar mode switch) as long as the initial function value is at most 10 and $\delta \le \frac{\mu_1}{m^{\frac{1}{4}} \sqrt{\lambda} d^{\frac{l-2}{2}} (m+K)^{\frac{l-1}{2}}}, \eta \le \frac{\mu_2 \lambda}{m^{\frac{1}{2}} d^{\frac{l-1}{2}} (m+K)^{\frac{l-2}{2}}}$ for some constants μ_1, μ_2. Thus, the function value can only increase when we reinitialize a component or when we switch the scaling from $\sqrt{d(m+K)}/\left\|u_i\right\|$ to $1/\left\|u_i\right\|$. In the following, we first show that reinitializing a component can only increase the function value by a small factor.

Claim 1. *Suppose* $f(U, \bar{C}) \le \hat{\Gamma} \le 10$. *Reinitialize any vector with the smallest* ℓ_2 *norm among all columns of* U, *and let the updated parameters be* $(U', \bar{C}')$, *then*

$$f(U', \bar{C}') \le \left(1 + \frac{13}{\lambda m}\right) \hat{\Gamma}.$$

According to the definition of the function value, we know $\left\|T - T^*\right\|_F \le \sqrt{2\hat{\Gamma}} \le \sqrt{20}$, $\sum_{j=1}^m \left\|u_j\right\|^2 \le \frac{\hat{\Gamma}}{\lambda}$, and $\sum_{j=1}^m \left\|u_j\right\|^4 \le \left(\sum_{j=1}^m \left\|u_j\right\|^2\right)^2 \le \frac{\hat{\Gamma}^2}{\lambda^2}$. Suppose u_i is one of the vectors in U with the smallest ℓ_2 norm, then $\left\|u_i\right\|^2 \le \frac{\hat{\Gamma}}{\lambda m}$. Suppose $u_i', c_i', \hat{c}_i', a_i'$ are the corresponding reinitialized vector and coefficients, and we have

$$\begin{aligned}
&\left\|a_i c_i^{l-2} u_i^{\otimes l} - a_i'(c_i')^{l-2}(u_i')^{\otimes l}\right\|_F \\
\le& \left\|a_i c_i^{l-2} u_i^{\otimes l}\right\|_F + \left\|a_i'(c_i')^{l-2}(u_i')^{\otimes l}\right\|_F \\
\le& \max\left(\left\|u_i\right\|^2, (\sqrt{d(m+K)})^{l-2} 4(m+K)\delta^2\right) + (\sqrt{d(m+K)})^{l-2}\delta^2 \\
\le& \frac{\hat{\Gamma}}{\lambda m} + (\sqrt{d(m+K)})^{l-2} 4(m+K)\delta^2 + (\sqrt{d(m+K)})^{l-2}\delta^2 \le \frac{2\hat{\Gamma}}{\lambda m},
\end{aligned}$$

where the last inequality assumes $\delta^2 \le \frac{\hat{\Gamma}}{5\lambda m (m+K)^{\frac{l}{2}} d^{\frac{l-2}{2}}}$. Therefore, we can bound $f(U', \bar{C}')$ as

$$\begin{aligned}
&f(U', \bar{C}') \\
=& \frac{1}{2}\left\|T - T^* - a_i c_i^{l-2} u_i^{\otimes l} + a_i'(c_i')^{l-2}(u_i')^{\otimes l}\right\|_F^2 + \lambda \sum_{j=1}^m \hat{c}_j^{l-2}\left\|u_j\right\|^l - \lambda \hat{c}_i^{l-2}\left\|u_i\right\|^{2l} + \lambda(\hat{c}_i')^{l-2}\left\|u_i'\right\|^l \\
\le& f(U, C) + \left\|T - T^*\right\|_F \left\|a_i c_i^{l-2} u_i^{\otimes l} - a_i'(c_i')^{l-2}(u_i')^{\otimes l}\right\|_F + \frac{1}{2}\left\|a_i c_i^{l-2} u_i^l - a_i'(c_i')^{l-2}(u_i')^l\right\|_F^2 + \lambda(\hat{c}_i')^{l-2}\left\|u_i'\right\|^l \\
\le& f(U, C) + \sqrt{20} \cdot \frac{2\hat{\Gamma}}{\lambda m} + \frac{1}{2}\left(2\frac{\hat{\Gamma}}{\lambda m}\right)^2 + \lambda\delta^2 \\
\le& \left(1 + \frac{12}{\lambda m}\right)\hat{\Gamma} + \lambda\delta^2 \\
\le& \left(1 + \frac{13}{\lambda m}\right)\hat{\Gamma},
\end{aligned}$$

where the second last inequality assumes $\lambda m \ge \hat{\Gamma}$ and the last inequality assumes $\delta^2 \le \frac{\hat{\Gamma}}{\lambda^2 m}$.

Switching the scaling from $\sqrt{d(m+K)}/\left\|u_i\right\|$ to $1/\left\|u_i\right\|$ can also potentially increase the function value. In the following, we show that the function value increase is small because we only switch the scaling mode when $\|u_i\| \leq 2\sqrt{m+K}\delta$.

Claim 2. *Assume $f(U', \bar{C}') \leq \bar{\Gamma}$. Suppose at this iteration we switch the scaling of c_i', i.e., we set c_i' as $c_i'/\sqrt{d(m+K)}$. Let the updated parameters be $(U', C'', \hat{C}', A')$, we have*

$$f(U', C'', \hat{C}', A') \leq \left(1 + \frac{1}{\lambda m^2}\right)\bar{\Gamma}.$$

Suppose u_i is the parameter which is one step of gradient descent before u_i'. According to the algorithm, we know $\|u_i\| \leq 2\sqrt{m+K}\delta$. According to the proof in Lemma 14 where we bound the derivative of f with respect to u_i, we know that as long as $\eta \leq \frac{1}{l(\sqrt{2\bar{\Gamma}}(\sqrt{d(m+K)})^{l-2}+\lambda)}$, we have $\|u_i'\| \leq 2\left\|u_i\right\| \leq 4\sqrt{m+K}\delta$. Therefore,

$$\begin{aligned}\left\|(c_i')^{l-2}(u_i')^{\otimes l} - (c_i'')^{l-2}(u_i')^{\otimes l}\right\|_F =& \left\|(c_i')^{l-2}(u_i')^{\otimes l} - \frac{1}{(\sqrt{d(m+K)})^{l-2}}(c_i')^{l-2}(u_i')^{\otimes l}\right\|_F \\ \leq& \left\|(c_i')^{l-2}(u_i')^{\otimes l}\right\|_F \leq 16(m+K)\delta^2(\sqrt{d(m+K)})^{l-2}.\end{aligned}$$

Suppose the tensor at $(U', \bar{C}')$ is T', then $\left\|T' - T^*\right\|_F \leq \sqrt{2\bar{\Gamma}}$.

Thus, we can bound $f(U', C'', \hat{C}', A')$ as follows:

$$\begin{aligned}&f(U', C'', \hat{C}', A') \\ \leq& f(U', \bar{C}'') + \left\|T' - T^*\right\|_F \left\|(c_i')^{l-2}(u_i')^{\otimes l} - (c_i'')^{l-2}(u_i')^{\otimes l}\right\|_F \\ &+ \frac{1}{2}\left\|(c_i')^{l-2}(u_i')^{\otimes l} - (c_i'')^{l-2}(u_i')^{\otimes l}\right\|_F^2 \\ \leq& \bar{\Gamma} + \sqrt{2\bar{\Gamma}}\left(16(m+K)\delta^2(\sqrt{d(m+K)})^{l-2}\right) + \frac{1}{2}\left(16(m+K)\delta^2(\sqrt{d(m+K)})^{l-2}\right)^2 \\ \leq& \left(1 + \frac{1}{\lambda m^2}\right)\bar{\Gamma},\end{aligned}$$

where the last inequality assumes $\delta^2 \leq \frac{\sqrt{\bar{\Gamma}}}{32\sqrt{2}\lambda m^2(m+K)^{\frac{l}{2}}d^{\frac{l-2}{2}}}$ and $\lambda m^2 \geq 1$.

We are now ready to bound the increase of the function value during this epoch. According to the algorithm, each epoch contains at most m scaling mode switches. Therefore, following Claim 2, all the scaling switches in one epoch can increase the upper bound of the function value by at most a factor of $\left(1 + \frac{1}{\lambda m^2}\right)^m \leq \exp\left(\frac{1}{\lambda m}\right)$. Combining with Claim 1 which considers the re-initialization, we know that in each epoch, the upper bound of the function value increase by at most a factor of $\exp\left(\frac{1}{\lambda m}\right)\left(1 + \frac{13}{\lambda m}\right) \leq \exp\left(\frac{14}{\lambda m}\right)$. □

Following Lemma 14 and Lemma 15, we are ready to show that the function value is upper bounded by a constant by an induction proof. At the beginning of our algorithm, the function value is bounded by a constant as long as the initialization radius δ is small enough. According to Lemma 15, the increase in each epoch is bounded by a factor of $O(\frac{1}{\lambda m})$. Therefore, as long as the total number of epochs do not exceed $O(\lambda m)$, f will always be bounded by a constant. As a consequence, the Frobenius norm square of U must be bounded by $O(\frac{1}{\lambda})$ due to the design of the regularizer. These results are summarized into Lemma 16.

Lemma 16. *Assume $\delta \leq \frac{\mu_1}{m^{\frac{3}{4}}d^{\frac{l-2}{2}}(m+K)^{\frac{l-1}{2}}\lambda^{\frac{1}{2}}}$, and $\eta \leq \frac{\mu_2\lambda}{m^{\frac{1}{2}}l^4d^{\frac{l-1}{2}}(m+K)^{\frac{l-2}{2}}}$ for some constants μ_1, μ_2. Also assume $K \leq \frac{\lambda m}{14}$ and $\frac{10}{m} \leq \lambda \leq 1$. We know throughout the algorithm*

$$f(U, \bar{C}) \leq 10 \text{ and } \sum_{i=1}^m \left\|u_i\right\|^2 \leq \frac{10}{\lambda}.$$

Proof of Lemma 16. Let's first show that the function value is bounded at the initialization if we choose δ to be small enough. At initialization, we have

$$\begin{aligned} f(U,\bar{C}) =& \frac{1}{2}\left\|T-T^*\right\|_F^2 + \lambda\sum_{i=1}^m \hat{c}_i^{l-2}\left\|u_i\right\|^l \\ \leq& \frac{1}{2}\left(\sum_{i=1}^m \left\|c_i^{l-2}u_i^{\otimes l}\right\|_F + \left\|T^*\right\|_F\right)^2 + \lambda\sum_{i=1}^m \left\|u_i\right\|^2 \\ \leq& \frac{1}{2}\left(m(\sqrt{d(m+K)})^{l-2}\delta^2+1\right)^2 + \lambda m\delta^2 \\ \leq& m^2 d^{l-2}(m+K)^{l-2}\delta^4 + 1 + \lambda m\delta^2 \leq 3, \end{aligned}$$

where the last inequality assumes $\delta^4 \leq \frac{1}{m^2 d^{l-2}(m+K)^{l-2}}$ and $\lambda \leq 1$.

We use an inductive proof to prove that the function value at the end of the k-th ($k \leq K$) iteration is at most $3\exp(\frac{14k}{\lambda m})$: At the initialization, the function value is at most 3. For every epoch, assume that our induction hypothesis is true, then at each step (a step can be a re-initialization, a gradient descent update, or a scalar mode switch), from Lemma 15 we know that the function value is upper bounded by $3\exp(\frac{14k}{\lambda m}) \leq 10$, so at this step Lemma 14 is correct, meaning that Lemma 15 is still correct at the next step.

Therefore, throughout the algorithm, we have

$$f(U,C) \leq 3\exp\left(\frac{14K}{\lambda m}\right) \leq 10,$$

where we assume $K \leq \frac{\lambda m}{14}$. This immediately implies that $\sum_i \left\|u_i\right\|^2 \leq \frac{10}{\lambda}$ by the design of our regularizer. □

Now we are ready to prove Lemma 13.

Proof of Lemma 13. From Lemma 16, we know that the function value is upper bounded by 10 throughout our algorithm. Besides, from Lemma 15 we know that the function value increase is at most $(\exp(\frac{14}{\lambda m})-1)$ times the function value at the beginning of this epoch. Choosing $\tilde{\Gamma} = f(U_0', \bar{C}_0')$, we know the function value increase at each epoch cannot exceed $10(\exp(\frac{14}{\lambda m})-1) = O(\frac{1}{\lambda m})$, which finishes the proof of Lemma 13. □

C.2 Escaping local minima

In this section, we will give a formal proof of Lemma 12. We again follow the proof ideas outlined in Section 5.2. Recall that the proof goes in the following steps:

1. We first show that the projection of U in the B subspace must be very small, therefore the influence from incorrect subspace B is small (Lemma 17).
2. We then focus on the correlation in the correct subspace S. First we show that the correlation can be significantly negative at re-initialization with constant probability (Lemma 18).
3. If the correlation is always significantly negative, then the re-initialized component will grow exponentially and eventually decrease the function value (Lemma 19).
4. If the correlation changes significantly, the function value must also decrease (Lemma 20 and Lemma 21).

First of all, we need to show that the influence coming from B is small enough so that it can be ignored. The following lemma is the formal version of Lemma 5. Note that the assumption $\left\|U\right\|_F \leq \sqrt{\frac{10}{\lambda}}$ has been verified in Lemma 16 so Lemma 17 holds for the entire algorithm.

Lemma 17. *Assume $\left\|U\right\|_F \leq \sqrt{\frac{10}{\lambda}}$ throughout the algorithm. Assume $\lambda \leq \sqrt{10}, \delta \leq \frac{\sqrt{10}}{2\sqrt{\lambda m d^{l-2}(m+K)^{l-1}}}$ and $\eta \leq \frac{\lambda}{20l}$. Then, we know $\left\|P_B U\right\|_F^2 \leq (m+K)\delta^2$ throughout the algorithm.*

Proof of Lemma 17. At the initialization,

$$\|P_B U\|_F^2 \le \|U\|_F^2 = m\delta^2.$$

At the beginning of each epoch, we re-initialize one column of U, which at most increases $\|P_B U\|_F^2$ by δ^2. Thus, the total increase due to the re-initialization process is at most $K\delta^2$.

Then, we only need to show that running gradient descent does not increase the norm of $P_B U$. Suppose at the beginning of one iteration, the tensor T is parameterized by $(U, \bar{C})$. Let U' be the updated parameter, which means $U' = U - \eta \nabla_U f(U, \bar{C})$. We have,

$$\begin{aligned}
&\|P_B U'\|_F^2 - \|P_B U\|_F^2 \\
=& \left\|P_B(U - \eta \nabla_U f(U, \bar{C}))\right\|_F^2 - \|P_B U\|_F^2 \\
=& -2\eta \left\langle P_B U, P_B \nabla_U f(U, \bar{C}) \right\rangle + \eta^2 \left\|P_B \nabla_U f(U, \bar{C})\right\|_F^2 . \qquad (6)
\end{aligned}$$

We will show that the first term is negative and dominates the second term when η is small enough, which implies that gradient descent never increases the norm of $P_B U$. We first compute the gradient as follows,

$$\nabla_U f(U, \bar{C}) = l\mathrm{mat}(T - T^*) U^{\odot l-1} C^{l-2} A + l\lambda U.$$

where $\mathrm{mat}(T - T^*) = UC^{l-2}A(U^{\odot l-1})^\top - U^* C^* [(U^*)^{\odot l-1}]^\top$ is a $d \times d^{l-1}$ matrix. Therefore, the projection of the gradient on B subspace is

$$P_B \nabla_U f(U, \bar{C}) = l P_B \mathrm{mat}(T) U^{\odot l-1} C^{l-2} A + l\lambda P_B U.$$

Now, we show that the first term in (6) is negative.

$$\begin{aligned}
-2\eta \left\langle P_B U, P_B \nabla_U f(U, \bar{C}) \right\rangle =& -2l\eta \left\langle P_B U, P_B \mathrm{mat}(T) U^{\odot l-1} C^{l-2} A + \lambda P_B U \right\rangle \\
=& -2l\eta \left\langle P_B U C^{l-2} A [U^{\odot l-1}]^\top, P_B \mathrm{mat}(T) \right\rangle - 2l\eta \left\langle P_B U, \lambda P_B U \right\rangle \\
=& -2l\eta \left\|P_B \mathrm{mat}(T)\right\|_F^2 - 2l\lambda\eta \left\|P_B U\right\|_F^2 .
\end{aligned}$$

Next, we show the second term in (6) is bounded. We have,

$$\begin{aligned}
\eta^2 \left\|P_B \nabla_U f(U, \bar{C})\right\|_F^2 =& \eta^2 \left\|l P_B \mathrm{mat}(T) U^{\odot l-1} C^{l-2} A + l\lambda P_B U\right\|_F^2 \\
\le& 2\eta^2 l^2 \left(\left\|P_B \mathrm{mat}(T) U^{\odot l-1} C^{l-2} A\right\|_F^2 + \lambda^2 \left\|P_B U\right\|_F^2 \right)
\end{aligned}$$

Recall that $M_2 = \sqrt{\frac{10}{\lambda}}$. Note that

$$\begin{aligned}
\left\|P_B \mathrm{mat}(T) U^{\odot l-1} C^{l-2} A\right\|_F^2 &\le \left\|P_B \mathrm{mat}(T)\right\|_F^2 \left\|U^{\odot l-1} C^{l-2}\right\|_F^2 \\
&= \left\|P_B \mathrm{mat}(T)\right\|_F^2 \sum_{i=1}^m c_i^{2l-4} \left\|u_i\right\|^{2l-2} \\
&\le \left\|P_B \mathrm{mat}(T)\right\|_F^2 \sum_{i=1}^m \max\{4(d(m+K))^{l-2}(m+K)\delta^2, \left\|u_i\right\|^2\} \\
&\le M_2^2 \left\|P_B \mathrm{mat}(T)\right\|_F^2
\end{aligned}$$

where the second inequality holds since $c_i = \frac{\sqrt{d(m+K)}}{\|u_i\|}$ only when $\|u_i\| \le 2\sqrt{m+K}\delta$, and otherwise $c_i = \frac{1}{\|u_i\|}$. The last inequality assumes $\delta \le \frac{M_2}{2\sqrt{md^{l-2}(m+K)^{l-1}}}$.

Overall, we have

$$\begin{aligned}
&\|P_B U'\|_F^2 - \|P_B U\|_F^2 \\
=& -2\eta \left\langle P_B U, P_B \nabla_U f(U, \bar{C}) \right\rangle + \eta^2 \left\| P_B \nabla_U f(U, \bar{C}) \right\|_F^2 \\
\le& -2l\eta \left(\|P_B \mathrm{mat}(T)\|_F^2 + \lambda \|P_B U\|_F^2 \right) + 2\eta^2 l^2 \left(M_2^2 \|P_B \mathrm{mat}(T)\|_F^2 + \lambda^2 \|P_B U\|_F^2 \right) \\
\le& -l\eta \left(\|P_B \mathrm{mat}(T)\|_F^2 + \lambda \|P_B U\|_F^2 \right),
\end{aligned}$$

where the last inequality assumes $\lambda \le M_2^2$ and $\eta \le \frac{1}{2lM_2^2}$. □

Lemma 17 shows that the norm of $P_B U$ only increases at the (re-)initializations, so it will stay small throughout this algorithm. This allows us to bound the influence to our algorithm from the orthogonal subspace B and only focus on subspace S. We denote the re-initialized vector at t-th step as u_t, and its sign as $a \in \{\pm 1\}$. Our analysis focuses on the correlation between $P_S u_t$ and the residual tensor

$$\langle P_{S^{\otimes l}} T_t - T^*, a\overline{P_S u_t}^{\otimes l} \rangle.$$

Here $\overline{P_S u_t}$ is the normalized version $P_S u_t$. We will show that if this correlation is significantly negative at every iteration the norm of u_t will blow up exponentially.

Towards this goal, first we will show that the initial point $P_S u_0$ has a large negative correlation with the residual. We will lower bound this correlation by anti-concentration of Gaussian polynomials, and the following lemma is the formal version of Lemma 6. Note in our notation, we have $\langle P_{S^{\otimes l}} T_t - T^*, a\overline{P_S u_t}^{\otimes l} \rangle = a \left(P_{S^{\otimes l}} T_t - T^* \right) \left(\overline{P_S u_t}^{\otimes l} \right)$.

Lemma 18. *Suppose the residual at the beginning of one epoch is $T_0' - T^*$. Suppose $a c_0^{l-2} u_0^{\otimes l}$ is the reinitialized component. There exists absolute constant μ such that with probability at least $1/5$,*

$$\left\langle P_{S^{\otimes l}} T_0' - T^*, a\overline{P_S u_0}^{\otimes l} \right\rangle \le -\frac{1}{(\mu r l)^{l/2}} \left\| P_{S^{\otimes l}} T_0' - P_{S^{\otimes l}} T^* \right\|_F,$$

where $\overline{P_S u_0} = P_S u_0 / \|P_S u_0\|$.

Proof of Lemma 18. Let's restrict into the r^l-dimensional space $S^{\otimes l}$, and let $P_{S^{\otimes l}}$ be the projection operator that projects a d^l-dimensional tensor to the r^l-dimensional space $S^{\otimes l}$. Then, we can think of $(P_{S^{\otimes l}} T - P_{S^{\otimes l}} T^*)$ as an r^l dimensional vector, and $\overline{P_S u}$ comes from uniform distribution on $\mathbb{S}^{r-1}$. Let v be an r-dimensional standard normal vector, then $a(P_{S^{\otimes l}} T - P_{S^{\otimes l}} T^*)(\overline{P_S u}^{\otimes l})$ has the same distribution as $a(P_{S^{\otimes l}} T - P_{S^{\otimes l}} T^*)(v^l) \frac{1}{\|v\|^l}$.

Let's first show that the variance of $a(P_{S^{\otimes l}} T - P_{S^{\otimes l}} T^*)(v^l)$ is large:

$$\begin{aligned}
\mathrm{Var}\left[a(P_{S^{\otimes l}} T - P_{S^{\otimes l}} T^*)(v^l) \right] =& \mathbb{E} \left| a(P_{S^{\otimes l}} T - P_{S^{\otimes l}} T^*)(v^l) \right|^2 \\
\ge& l! \left\| P_{S^{\otimes l}} T - P_{S^{\otimes l}} T^* \right\|_F^2,
\end{aligned}$$

where the equality holds because $\mathbb{E}\left[a(P_{S^{\otimes l}} T - P_{S^{\otimes l}} T^*)(v^l) \right] = 0$ and the inequality follows from Lemma 11. It's not hard to see that $a(P_{S^{\otimes l}} T - P_{S^{\otimes l}} T^*)(v^l)$ is an l-th order polynomial of standard Gaussian vectors. By anti-concentration inequality of Gaussian polynomials (Lemma 25), we know there exists constant κ such that

$$\Pr\left[\left| a(P_{S^{\otimes l}} T - P_{S^{\otimes l}} T^*)(v^l) \right| \le \epsilon \sqrt{l!} \left\| P_{S^{\otimes l}} T - P_{S^{\otimes l}} T^* \right\|_F \right] \le \kappa l \epsilon^{1/l}.$$

Choosing $\epsilon = \frac{1}{2^l \kappa^l l^l}$, we know with probability at least half,

$$\begin{aligned}\left|a(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)(v^l)\right| \geq & \frac{1}{2^l\kappa^l l^l}\sqrt{l!}\left\|P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*\right\|_F \\ \geq & \frac{1}{2^l\kappa^l l^l}\left(\frac{l}{e}\right)^{l/2}\left\|P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*\right\|_F \\ = & \frac{1}{2^l e^{l/2}\kappa^l l^{l/2}}\left\|P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*\right\|_F.\end{aligned}$$

Since the distribution of $a(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)(v^l)$ is symmetric, we know with probability at least $1/4$,

$$a(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)(v^l) \leq -\frac{1}{2^l e^{l/2}\kappa^l l^{l/2}}\left\|P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*\right\|_F.$$

According to Lemma 24, we know that with probability at least $19/20$,

$$\|v\| \leq \kappa'\sqrt{r},$$

where κ' is some constant. This further implies that with probability at least $1/5$,

$$a(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)(v^l)\frac{1}{\|v\|^l} \leq -\frac{1}{2^l e^{l/2}(\kappa\kappa')^l (rl)^{l/2}}\left\|P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*\right\|_F.$$

Choosing $\mu = 4e\kappa^2(\kappa')^2$ finishes the proof. □

Our next step argues that if this negative correlation is large in every step, then the norm of u_t blows up exponentially. Intuitively, this is due to the fact that the correlation is basically the dominating term in the gradient, so when it is significantly negative the vector u_t behaves similar to a vector doing matrix power method (here it is important that our model is 2-homogeneous so the behavior of power method is similar to the matrix setting). Below is the formal version of Lemma 7.

Lemma 19. *In the setting of Theorem 5, within one epoch, let T_0 be the tensor after the reinitilization and let T_τ be the tensor at the end of the τ-th iteration. Assume $\|P_S u_0\| \geq \frac{\mu_2\delta}{\sqrt{d}}$ for some constant $\mu_2 \in (0,1)$. For any $H \geq t \geq 1$, as long as $\left\langle P_{S^{\otimes l}}T_\tau - T^*, a\overline{P_S u_\tau}^{\otimes l}\right\rangle \leq \frac{-\epsilon}{5(\mu_1 rl)^{l/2}}$ for some constant μ_1 for all $t-1 \geq \tau \geq 0$, we have*

$$\|P_S u_t\|^2 \geq \left(1 + \eta\left(\frac{\mu_2}{2}\right)^{l-2}\frac{\epsilon}{10(\mu_1 rl)^{\frac{l}{2}}}\right)^t \|P_S u_0\|^2.$$

Proof of Lemma 19. We will use inductive proof for this lemma. At the first step, we have

$$\begin{aligned}\|P_S u_1\|^2 = & \left\|P_S u_0 - \eta P_S \nabla_u f(U_0, \bar{C}_0)\right\|^2 \\ = & \|P_S u_0\|^2 - \eta\left\langle P_S u_0, P_S\nabla_u f(U_0,\bar{C}_0)\right\rangle + \eta^2\left\|P_S\nabla_u f(U_0,\bar{C}_0)\right\|^2 \\ \geq & \|P_S u_0\|^2 - \eta\left\langle P_S u_0, P_S\nabla_u f(U_0,\bar{C}_0)\right\rangle.\end{aligned}$$

We can write down the $P_S\nabla_u f(U_0,\bar{C}_0)$ as follows,

$$P_S\nabla_u f(U_0,\bar{C}_0) = al(T_0 - T^*)(u_0^{\otimes(l-1)}, P_S)c_0^{l-2} + \lambda l P_S u_0.$$

Let's first consider $al(T_0 - T^*)((u_0)^{\otimes(l-1)}, P_S)c_0^{l-2}$. We can decompose u_0 into $P_S u_0$ and $P_B u_0$, so we can divide $al(T_0 - T^*)(u_0^{\otimes(l-1)}, P_S)c_0^{l-2}$ into 2^{l-1} terms, each of which corresponds to the projection of $u_0^{\otimes(l-1)}$ on a subspace in $\{S,B\}^{\otimes l-1}$. For subspace $S^{\otimes(l-1)}$, the projection is $al(P_{S^{\otimes l}}T_0 - P_{S^{\otimes l}}T^*)((P_S u_0)^{\otimes l-1}, P_S)c_0^{l-2}$. Its inner product with $-P_S u_0$ is

$$\begin{aligned}& \left\langle -P_S u_0, al(P_{S^{\otimes l}}T_0 - P_{S^{\otimes l}}T^*)((P_S u_0)^{\otimes l-1}, P_S)c_0^{l-2}\right\rangle \\ = & -al(P_{S^{\otimes l}}T_0 - P_{S^{\otimes l}}T^*)((P_S u_0)^{\otimes l})c_0^{l-2} \\ = & -al(P_{S^{\otimes l}}T_0 - P_{S^{\otimes l}}T^*)((\overline{P_S u_0})^{\otimes l})\left(\|P_S u_0\| c_0\right)^{l-2}\|P_S u_0\|^2.\end{aligned}$$

Now, we only need to show that $\|P_S u_0\| c_0$ is lower bounded. We have

$$\|P_S u_0\| c_0 = \frac{\sqrt{d(m+K)}\,\|P_S u_0\|}{\|u_0\|} \geq \frac{\sqrt{d(m+K)}\mu_2\delta/\sqrt{d}}{\delta} \geq \frac{\mu_2}{2},$$

where the first inequality uses $\|P_S u_0\| \geq \mu_2\delta/\sqrt{d}$. Therefore,

$$\left\langle -P_S u_0, al(P_{S^{\otimes l}}T_0 - P_{S^{\otimes l}}T^*)((P_S u_0)^{\otimes l-1}, P_S)c_0^{l-2}\right\rangle \geq \left(\frac{\mu_2}{2}\right)^{l-2}\frac{\epsilon}{5(\mu_1 r l)^{l/2}}\|P_S u_0\|^2.$$

We then bound the remaining terms in $al(T_0 - T^*)((u_0)^{\otimes l-1}, P_S)c_0^{l-2}$: For any $l-1 \geq k \geq 1$, we consider the subspace $B^{\otimes k}\otimes S^{\otimes l-1-k}$ and all of its permutations, we bound the norm of $al(T_0-T^*)((P_B u_0)^{\otimes k}, (P_S u_0)^{\otimes l-1-k}, P_S)c_0^{l-2}$ as follows.

$$\begin{aligned}
&\left\|al(T_0 - T^*)\left((P_B u_0)^{\otimes k}, (P_S u_0)^{\otimes l-1-k}, P_S\right)c_0^{l-2}\right\| \\
=&l\left\|T_0((P_B u_0)^{\otimes k}, (P_S u_0)^{\otimes l-1-k}, P_S)c_0^{l-2}\right\| \\
\leq& l\sum_{i=1}^m c_{0,i}^{l-2}\|P_B u_{0,i}\|^k\|P_S u_{0,i}\|^{l-k}\|P_B u_0\|^k\|P_S u_0\|^{l-1-k}c_0^{l-2} \\
\leq& l\sum_{i=1}^m c_{0,i}^{l-2}\|P_B u_{0,i}\|\|u_{0,i}\|^{l-1}\|P_B u_0\|\|u_0\|^{l-2}c_0^{l-2} \\
\leq& l\sum_{i=1}^m d^{l-2}(m+K)^{l-1}\delta^2\|u_{0,i}\| \\
\leq& l\sqrt{m}d^{l-2}(m+K)^{l-1}\delta^2 M_2,
\end{aligned}$$

where $M_2 = \sqrt{\frac{10}{\lambda}}$ is the upper bound of $\|U\|_F$. Denote R_0 as the summation of terms in all subspaces except for $S^{\otimes l-1}$. We have $\|R_0\| \leq (2^{l-1}-1)l\sqrt{m}d^{l-2}(m+K)^{l-1}\delta^2 M_2$. Therefore, we have

$$\begin{aligned}
|\langle P_S u_0, R_0\rangle| \leq& \|P_S u_0\|\|R_0\| \leq 2^l l\sqrt{m}d^{l-2}(m+K)^{l-1}\delta^2 M_2\|P_S u_0\| \\
\leq& \left(\frac{\mu_2}{2}\right)^{l-2}\frac{\epsilon}{20(\mu_1 r l)^{l/2}}\|P_S u_0\|^2
\end{aligned}$$

where the last inequality uses $\|P_S u_0\| \geq \frac{\mu_2\delta}{\sqrt{d}}$ and assumes $\delta \leq \frac{1}{2^l l\sqrt{m}d^{l-2}(m+K)^{l-1}M_2}\cdot\frac{\mu_2}{\sqrt{d}}\left(\frac{\mu_2}{2}\right)^{l-2}\frac{\epsilon}{20(\mu_1 r l)^{l/2}}$.

Next, let's analyze the regularizer $\lambda l P_S u_0$. Its norm can be bounded as follows,

$$\|\lambda l P_S u_0\| \leq \left(\frac{\mu_2}{2}\right)^{l-2}\frac{\epsilon}{20(\mu_1 r l)^{l/2}}\|P_S u_0\|,$$

where we assume $\lambda \leq \frac{1}{l}\cdot\left(\frac{\mu_2}{2}\right)^{l-2}\frac{\epsilon}{20(\mu_1 r l)^{l/2}}$.

Overall, we have

$$\begin{aligned}
\|P_S u_1\|^2 \geq& \|P_S u_0\|^2 - \eta\left\langle P_S u_0, \nabla_u f(U_0, \bar{C}_0)\right\rangle \\
\geq& \|P_S u_0\|^2 + \eta\left(\frac{\mu_2}{2}\right)^{l-2}\left(\frac{\epsilon}{5(\mu_1 r l)^{l/2}} - \frac{\epsilon}{20(\mu_1 r l)^{l/2}} - \frac{\epsilon}{20(\mu_1 r l)^{l/2}}\right)\|P_S u_0\|^2 \\
=& \left(1 + \eta\left(\frac{\mu_2}{2}\right)^{l-2}\frac{\epsilon}{10(\mu_1 r l)^{l/2}}\right)\|P_S u_0\|^2.
\end{aligned}$$

Induction Step: Suppose $\|P_S u_t\|^2 \geq \left(1+\eta\left(\frac{\mu_2}{2}\right)^{l-2}\frac{\epsilon}{10(\mu_1 r l)^{l/2}}\right)^t \|P_S u_0\|^2$, we will prove that $\|P_S u_{t+1}\|^2 \geq \left(1+\eta\left(\frac{\mu_2}{2}\right)^{l-2}\frac{\epsilon}{10(\mu_1 r l)^{l/2}}\right)^{t+1}\|P_S u_0\|^2$. Actually, we have assumed that $a(P_{S^{\otimes l}}T_t - P_{S^{\otimes l}}T^*)(\overline{P_S u_t}^{\otimes l}) \leq -\frac{\epsilon}{5(\mu_1 r l)^{l/2}}$, so we only need to show that $c_t\|P_S u_t\| \geq \frac{\mu_2}{2}$. Based on these two properties, the remaining proofs are exactly the same as that for $t=0$.

The latter property is not hard to verify:

If $\|u_t\| > 2\sqrt{m+K}\delta$, we know $c_t = 1/\|u_t\|$. Then, we have $\|P_S u_t\| c_t = \frac{\|P_S u_t\|}{\|u_t\|} \geq \frac{\|u_t\|-\|P_B u_t\|}{\|u_t\|} \geq \frac{1}{2}$, where we use $\|P_B u_t\| \leq \sqrt{m+K}\delta$.

If $\|u_t\| \leq 2\sqrt{m+K}\delta$, we do not necessarily have $c_t = \frac{\sqrt{d(m+K)}}{\|u_t\|}$ because the norm of $\|u_t\|$ might first exceed the threshold and then drop below the threshold later. Note, by the induction proof, we only know the norm of $P_S u_t$ monotonically increase, which does not imply that $\|u_t\|$ monotonically increases. So, we have to consider both cases here. If $c_t = \frac{\sqrt{d(m+K)}}{\|u_t\|}$, we have $\|P_S u_t\| c_t = \frac{\sqrt{d(m+K)}\|P_S u_t\|}{\|u_t\|} \geq \frac{\mu_2}{2}$, which is because $\|P_S u_t\| \geq \|P_S u_0\| \geq \mu_2\delta/\sqrt{d}$. If $c_t = 1/\|u_t\|$, we know there exists $\tau \leq t$ such that $\|u_\tau\| > 2\sqrt{m+K}\delta$. Since $\|P_B u_\tau\| \leq \sqrt{m+K}\delta$, we know $\|P_S u_\tau\| \geq \sqrt{m+K}\delta$. By the induction proof, we know $\|P_S u_t\| \geq \|P_S u_\tau\| \geq \sqrt{m+K}\delta$. Then, we have $\|P_S u_t\| c_t = \frac{\|P_S u_t\|}{\|u_t\|} \geq \frac{1}{2}$.

This finishes the proof of Lemma 19. □

Therefore the final step is to show that $a\overline{P_S u_t}^{\otimes l}$ always has a large negative correlation with $P_{S^{\otimes l}}T_t - P_{S^{\otimes l}}T^*$, unless the function value has already decreased. The difficulty here is that both the current reinitialized component u_t and other components are moving, therefore T_t is also changing.

We can bound the change of $T - T^*$ by separating it into two terms, which are the change of the re-initialized component and the change of the residual:

$$\begin{aligned}
&\left|a(P_{S^{\otimes l}}T_t - P_{S^{\otimes l}}T^*)(\overline{P_S u_t}^{\otimes l}) - a(P_{S^{\otimes l}}T_0 - P_{S^{\otimes l}}T^*)(\overline{P_S u_0}^{\otimes l})\right| \\
\leq &\left|\sum_{\tau=1}^{t}\left((P_{S^{\otimes l}}T_{\tau-1} - P_{S^{\otimes l}}T^*)(\overline{P_S u_\tau}^{\otimes l}) - (P_{S^{\otimes l}}T_{\tau-1} - P_{S^{\otimes l}}T^*)(\overline{P_S u_{\tau-1}}^{\otimes l})\right)\right| \\
&+ \sum_{\tau=1}^{t}\|T_\tau - T_{\tau-1}\|_F .
\end{aligned}$$

The change of the re-initialized component has a small effect on the correlation because the change in S subspace can only improve the correlation, and the influence of the B subspace can be bounded. This is formally proved in the following lemma, which is the formal version of Lemma 8.

Lemma 20. *Assume $\delta \leq \frac{\mu_1}{m^{\frac{3}{4}}\sqrt{\lambda}d^{\frac{l-2}{2}}(m+K)^{\frac{l-1}{2}}}, \eta \leq \frac{\mu_2}{\lambda^{\frac{3}{2}}m^{\frac{9}{4}}d^{\frac{3l-6}{2}}(m+K)^{\frac{3l-3}{2}}}$ for some constants μ_1, μ_2. Assume $K \leq \frac{\lambda m}{14}$ and $\frac{10}{m} \leq \lambda \leq 1$. Suppose at the beginning of one iteration, the tensor T is parameterized by $(U, \bar{C})$. Suppose u is one column vector in U with $\|P_S u\| \geq \frac{\mu_3\delta}{\sqrt{d}}$ where μ_3 is a constant. Suppose u' is u after one step of gradient descent: $u' = u - \eta\nabla_u f(U, \bar{C})$. We have*

$$a(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)(\overline{P_S u'}^{\otimes l}) \leq a(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)(\overline{P_S u}^{\otimes l}) + \mu l^4 2^l d^{l-1.5}m^{1/2}(m+K)^{l-1}\eta\delta\lambda,$$

where μ is some constant.

Proof of Lemma 20. Define $g(u) := a(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)(\overline{P_S u}^{\otimes l})$. Note that in function $g(u)$, we view T as fixed. We will show that the change of g is bounded when the input changes from u to u'.

Bounding first order change: Let's first compute the gradient of g at u.

$$\begin{aligned}
\nabla g(u) =& al(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)((P_Su)^{\otimes l-1}, P_S)\frac{1}{\|P_Su\|^l} - al(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)((P_Su)^{\otimes l})\frac{P_Su}{\|P_Su\|^{l+2}}\\
=& al(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)((P_Su)^{\otimes l-1}, P_S)\frac{1}{\|P_Su\|^l} - al(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)((P_Su)^{\otimes l-1}, \overline{P_Su})\frac{\overline{P_Su}}{\|P_Su\|^l}\\
=& al(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)((P_Su)^{\otimes l-1}, P_S - \overline{P_Su}\cdot\overline{P_Su}^\top)\frac{1}{\|P_Su\|^l}\\
=& al(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)((P_Su)^{\otimes l-1}, P_D)\frac{1}{\|P_Su\|^l},
\end{aligned}$$

where P_D is the projection matrix on the span of $S\setminus\{u\}$. We can also compute the projection of $\nabla_u f(U,\bar{C})$ on D as follows,

$$P_D\nabla_u f(U,\bar{C}) = al(T-T^*)(u^{\otimes l-1}, P_D)c^{l-2}.$$

We can divide $l(T-T^*)(u^{\otimes l-1}, P_D)c^{l-2}$ into 2^{l-1} terms, each of which corresponds to the projection of u^{l-1} on a subspace. For subspace $S^{\otimes l-1}$, we have

$$al(T-T^*)((P_Su)^{\otimes l-1}, P_D)c^{l-2} = al(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)((P_Su)^{\otimes l-1}, P_D)c^{l-2},$$

which has non-negative inner product with $\nabla g(u)$. We can bound the norm of all the other terms. For any $l-1\geq k\geq 1$, consider subspace $B^{\otimes k}\otimes S^{\otimes(l-1-k)}$, we can bound the norm of $al(T-T^*)((P_Bu)^{\otimes k}, (P_Su)^{\otimes(l-1-k)}, P_D)c^{l-2}$ as follows:

$$\begin{aligned}
&\left\|al(T-T^*)((P_Bu)^{\otimes k}, (P_Su)^{\otimes l-1-k}, P_D)c^{l-2}\right\|\\
=& l\left\|T((P_Bu)^{\otimes k}, (P_Su)^{\otimes l-1-k}, P_D)c^{l-2}\right\|\\
\leq& l\sum_{i=1}^m c_i^{l-2}\left\|P_Bu_i\right\|^k\left\|P_Su_i\right\|^{l-k}\left\|P_Bu\right\|^k\left\|P_Su\right\|^{l-1-k}c^{l-2}\\
\leq& l\sum_{i=1}^m c_i^{l-2}\left\|P_Bu_i\right\|\left\|u_i\right\|^{l-1}\left\|P_Bu\right\|\left\|u\right\|^{l-2}c^{l-2}\\
\leq& l\sum_{i=1}^m d^{l-2}(m+K)^{l-1}\delta^2\left\|u_i\right\|\\
\leq& l\sqrt{m}d^{l-2}(m+K)^{l-1}\delta^2 M_2,
\end{aligned}$$

where the second last inequality comes from Lemma 17.

Denote R as the summation of terms in all subspaces except for $S^{\otimes l-1}$, then

$$\|R\|\leq(2^{l-1}-1)l\sqrt{m}d^{l-2}(m+K)^{l-1}\delta^2 M_2.$$

Therefore, the first order change of g can be bounded as follows,

$$
\begin{aligned}
&\left\langle \nabla g(u), -\eta \nabla_u f(U, \bar{C}) \right\rangle \\
=&\left\langle al(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)((P_S u)^{\otimes l-1}, P_D)\frac{1}{\|P_S u\|^l}, -\eta al(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)((P_S u)^{\otimes l-1}, P_D)c_u^{l-2} - \eta R \right\rangle \\
\leq&\eta \left\| l(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)((P_S u)^{\otimes l-1}, P_D)\frac{1}{\|P_S u\|^l} \right\| \|R\| \\
\leq&\eta l\sqrt{20}\frac{\sqrt{d}}{\mu_3\delta} \cdot (2^{l-1}-1)l\sqrt{m}d^{l-2}(m+K)^{l-1}\delta^2 M_2 \\
\leq&\frac{10l^2 2^l}{\mu_3}\eta d^{l-1.5}\sqrt{m}(m+K)^{l-1}\delta M_2,
\end{aligned}
$$

where the second last inequality assumes $||P_S u|| \geq \frac{\mu_3\delta}{\sqrt{d}}$.

Bounding higher order change: For all $u'' = (1-\theta)u + \theta u'$ with $0 \leq \theta \leq 1$, we prove a uniform upper bound for $\left\|\nabla^2 g(u'')\right\|_F$. Recall the gradient of g at u'',

$$
\nabla g(u'') = al(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)((P_S u'')^{\otimes l-1}, P_D'')\frac{1}{\|P_S u''\|^l},
$$

where P_D'' is the projection matrix to $S \setminus \{u''\}$. We compute $\left\|\nabla^2 g(u'')\right\|$ as follows,

$$
\begin{aligned}
\nabla^2 g(u'') =& al(l-1)(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)((P_S u'')^{\otimes l-2}, P_S, P_D'')\frac{1}{\|P_S u''\|^l} \\
&- al^2(P_{S^{\otimes l}}T - P_{S^{\otimes l}}T^*)((P_S u'')^{\otimes l-1}, P_D'') \otimes \frac{P_S u''}{\|P_S u''\|^{l+2}}.
\end{aligned}
$$

Therefore,

$$
\left\|\nabla^2 g(u'')\right\|_F \leq 2l^2\sqrt{20}\frac{1}{\|P_S u''\|^2}.
$$

Assume that $\eta \leq \frac{\mu_3\delta\sqrt{\lambda}}{2\sqrt{10d}\left(\sqrt{20}l(\sqrt{d(m+K)})^{l-2}+\lambda l\right)}$ and from the proof of Lemma 14 where we bound the gradient, we know that

$$
\left\|\eta\nabla_u f(U, \bar{C})\right\| \leq \eta\left(l\sqrt{20}(\sqrt{d(m+K)})^{l-2} + \lambda l\right)\|u\| \leq \frac{\sqrt{\frac{\lambda}{10}}\mu_3\delta}{2\sqrt{d}}\sqrt{\frac{10}{\lambda}} = \frac{\mu_3\delta}{2\sqrt{d}}.
$$

Thus,

$$
\begin{aligned}
\|P_S u''\| &\geq \|P_S u\| - \|P_S u'' - P_S u\| \\
&\geq \|P_S u\| - \left\|\theta\eta\nabla_u f(U, \bar{C})\right\| \\
&\geq \frac{\mu_3\delta}{\sqrt{d}} - \frac{\mu_3\delta}{2\sqrt{d}} = \frac{\mu_3\delta}{2\sqrt{d}}.
\end{aligned}
$$

Therefore,

$$
\left\|\nabla^2 g(u'')\right\|_F \leq 2l^2\sqrt{20}\frac{4d}{\mu_3^2\delta^2}.
$$

Overall, we have

$$
g(u') - g(u) \leq \left\langle \nabla g(u), -\eta\nabla_u f(U, \bar{C})\right\rangle + \frac{\eta^2}{2}2l^2\sqrt{20}\frac{4d}{\mu_3^2\delta^2}\left\|\nabla_u f(U, \bar{C})\right\|^2.
$$

Recall that,

$$\nabla_u f(U, \bar{C}) = al(T - T^*)(u^{\otimes l-1}, I)c^{l-2} + \lambda l u,$$

we have

$$\begin{aligned}
\left\|\nabla_u f(U, \bar{C})\right\|_F \leq & l\sqrt{20}\max\left(\|u\|, 2\sqrt{m+K}\delta(\sqrt{d(m+K)})^{l-2}\right) + \lambda l \|u\| \\
\leq & l\sqrt{20}\max\left(M_2, 2\sqrt{m+K}\delta(\sqrt{d(m+K)})^{l-2}\right) + \lambda l M_2 \\
\leq & l\sqrt{20}M_2 + \lambda l M_2,
\end{aligned}$$

where the last inequality assumes $\delta \leq \frac{M_2}{2\sqrt{m+K}(\sqrt{d(m+K)})^{l-2}}$.

Finally, we have

$$\begin{aligned}
g(u') - g(u) \leq & \left\langle \nabla g(u), -\eta \nabla_u f(U, \bar{C}) \right\rangle + \frac{\eta^2}{2} 2l^2\sqrt{20}\frac{4d}{\mu_3^2\delta^2}\left\|\nabla_u f(U, \bar{C})\right\|^2 \\
\leq & \frac{10l^2 2^l}{\mu_3}\eta d^{l-1.5}\sqrt{m}(m+K)^{l-1}\delta M_2 + \frac{\eta^2}{2} 2l^2\sqrt{20}\frac{4d}{\mu_3^2\delta^2}\left(l\sqrt{20}M_2 + \lambda l M_2\right)^2 \\
\leq & \frac{10l^2 2^l}{\mu_3}\eta d^{l-1.5}\sqrt{m}(m+K)^{l-1}\sqrt{\frac{10}{\lambda}}\delta + \sqrt{20}\eta^2 l^2\frac{4d}{\mu_3^2\delta^2}\left(40l^2\frac{10}{\lambda} + 2\lambda^2 l^2\frac{10}{\lambda}\right) \\
\leq & \mu l^4 2^l d^{l-1.5} m^{1/2}(m+K)^{l-1}\eta\delta\lambda,
\end{aligned}$$

where the last inequality assumes $l \geq 3$, $\eta \leq \delta^3$ and μ is some constant. □

Therefore, the only way to change the residual term by a lot must be changing the tensor T, and the accumulated change of T is strongly correlated with the decrease of f. This is similar to the technique of bounding the function value decrease in Wei et al. (2019). The connection between them are formalized in the following lemma, which is the formal version of Lemma 9:

Lemma 21. *Assume that $\delta \leq \frac{\mu_1}{m^{\frac{3}{4}}\sqrt{\lambda}d^{\frac{l-2}{2}}(m+K)^{\frac{l-1}{2}}}, \eta \leq \frac{\mu_2\lambda}{m^{\frac{1}{2}}l^4 d^{\frac{l-1}{2}}(m+K)^{\frac{l-2}{2}}}$ for some constants μ_1, μ_2, and $\frac{10}{m} \leq \lambda \leq 1$. Within one epoch, let T_0 be the tensor after reinitialization, and let T_t be the tensor at the end of the t-th iteration. Let $(U_0, \bar{C}_0)$ be the parameters after the reinitialization step and let $(U_H, \bar{C}_H)$ be the parameters at the end of this epoch. We have*

$$\begin{aligned}
& \sum_{\tau=1}^{H} \|T_\tau - T_{\tau-1}\|_F \\
\leq & 200l^{2.5}\sqrt{\frac{1}{\lambda}}\sqrt{\eta H}\sqrt{f(U_0, \bar{C}_0) - f(U_H, \bar{C}_H) + 160m(m+K)\delta^2(\sqrt{d(m+K)})^{l-2}} \\
& + 16m(m+K)\delta^2(\sqrt{d(m+K)})^{l-2}.
\end{aligned}$$

Intuitively, if we are doing a standard gradient descent, at each step the change in function value is going to be proportional to the square of the change in the tensor T, and the guarantee similar to the Lemma above can be proved by applying Cauchy-Schwartz. However, in our setting the proof becomes more complicated because of the normalization steps and in particular the scalar mode switch.

Before proving Lemma 21, we first prove the following lemma which guarantees the function value decrease in one step (without scalar mode switch):

Lemma 22. *Assume $\delta \leq \frac{\mu_1}{m^{\frac{3}{4}}\sqrt{\lambda}d^{\frac{l-2}{2}}(m+K)^{\frac{l-1}{2}}}, \eta \leq \frac{\mu_2\lambda}{m^{\frac{1}{2}}l^4 d^{\frac{l-1}{2}}(m+K)^{\frac{l-2}{2}}}$ for some constants μ_1, μ_2, and $\eta \leq \delta^3$. Assume $K \leq \frac{\lambda m}{14}$. Starting from T parameterized by $(U, \bar{C})$, suppose after one iteration (before potential scalar mode switch) the tensor becomes T' parameterized by $(U', \bar{C}')$. We have*

$$\|T' - T\|_F \leq 200l^2\sqrt{\frac{1}{\lambda}}\left\|\eta\nabla_U f(U, \bar{C})\right\|_F.$$

Proof of Lemma 22. According to the algorithm, we know each iteration is composed of two steps: update U by gradient descent ($U' = U - \eta\nabla_U f(U, \bar{C})$) and update C and $\hat{C}$ according to U'. Let $\hat{T}$ be the intermediate tensor parameterized by $(U', \bar{C})$. We will bound $\|T' - T\|_F$ by bounding $\left\|\hat{T} - T\right\|_F$ and $\left\|T' - \hat{T}\right\|_F$ separately.

According to Lemma 16, we know $\sum_{i=1}^m \|u_i\|^2, \sum_{i=1}^m \|u_i'\|^2 \le \frac{10}{\lambda}$. Denote $M_2^2 = \frac{10}{\lambda}$.

Bounding $\left\|\hat{T} - T\right\|_F$: From T to $\hat{T}$, U is updated to $U' = U - \eta\nabla_U f(U, \bar{C})$ while C and $\hat{C}$ remains the same. Therefore,

$$\begin{aligned}\left\|\hat{T} - T\right\|_F &= \left\|\sum_{i=1}^m a_i c_i^{l-2} \left(u_i - \eta\nabla_{u_i} f(U, \bar{C})\right)^{\otimes l} - \sum_{i=1}^m a_i c_i^{l-2} u_i^{\otimes l}\right\|_F \\ &\le \sum_{i=1}^m l \|u_i\|^{l-1} \left\|\eta\nabla_{u_i} f(U, \bar{C})\right\| c_i^{l-2} + \sum_{i=1}^m \sum_{k=2}^l \binom{l}{k} \|u_i\|^{l-k} \left\|\eta\nabla_{u_i} f(U, \bar{C})\right\|^k c_i^{l-2}.\end{aligned}$$

We can further bound the linear term as follows:

$$\begin{aligned}\sum_{i=1}^m l \|u_i\|^{l-1} \left\|\eta\nabla_{u_i} f(U, \bar{C})\right\| c_i^{l-2} \le& l \sum_{i=1}^m \left\|\eta\nabla_{u_i} f(U, \bar{C})\right\| \max(\|u_i\|, 2\sqrt{m+K}\delta(\sqrt{d(m+K)})^{l-2}) \\ \le& l \sqrt{\sum_{i=1}^m \left\|\eta\nabla_{u_i} f(U, \bar{C})\right\|^2} \sqrt{\sum_{i=1}^m \max(\|u_i\|^2, 4(m+K)^{l-1}\delta^2 d^{l-2})} \\ \le& \sqrt{2} l M_2 \eta \left\|\nabla_U f(U, \bar{C})\right\|_F,\end{aligned}$$

where the last inequality assumes $\delta^2 \le \frac{M_2^2}{4m(m+K)^{l-1}d^{l-2}}$.

According to the proof in Lemma 14, we know $\left\|\eta\nabla_{u_i} f(U, \bar{C})\right\| \le \frac{1}{l}\|u_i\|$. Therefore, for the higher order terms, for each $k \ge 2$,

$$\begin{aligned}\sum_{i=1}^m \binom{l}{k} \|u_i\|^{l-k} \left\|\eta\nabla_{u_i} f(U, \bar{C})\right\|^k c_i^{l-2} \le& \sum_{i=1}^m l^k \|u_i\|^{l-k} \frac{\|u_i\|^{k-1}}{l^{k-1}} \left\|\eta\nabla_{u_i} f(U, \bar{C})\right\| c_i^{l-2} \\ \le& \sum_{i=1}^m l \|u_i\|^{l-1} \left\|\eta\nabla_{u_i} f(U, \bar{C})\right\| c_i^{l-2} \\ \le& \sqrt{2} l M_2 \eta \left\|\nabla_U f(U, \bar{C})\right\|_F.\end{aligned}$$

Overall, we have

$$\left\|\hat{T} - T\right\|_F \le 2\sqrt{2} l^2 M_2 \eta \left\|\nabla_U f(U, \bar{C})\right\|_F.$$

Bounding $\left\|T' - \hat{T}\right\|_F$: From $\hat{T}$ to T', we update C to C' and $\hat{C}$ to $\hat{C}'$ such that $\forall i \in [m], c_i' = c_i \frac{\|u_i\|}{\|u_i'\|}$ and $\hat{c}_i' = \hat{c}_i \frac{\|u_i\|}{\|u_i'\|}$. Thus,

$$\begin{aligned}\left\|T' - \hat{T}\right\|_F &= \left\|\sum_{i=1}^m a_i (c_i')^{l-2} (u_i')^{\otimes l} - \sum_{i=1}^m a_i c_i^{l-2} (u_i')^{\otimes l}\right\|_F \\ &\le \sum_{i=1}^m \left|(c_i')^{l-2} - c_i^{l-2}\right| \|u_i'\|^l.\end{aligned}$$

Now, let's focus on the change in c_i^{l-2}. Define $g(u) = \frac{1}{\|u\|^{l-2}}$. We have,

$$\nabla g(u) = -(l-2)\frac{u}{\|u\|^l} \text{ and } \nabla^2 g(u) = -(l-2)\frac{I}{\|u\|^l} + l(l-2)\frac{uu^\top}{\|u\|^{l+2}}.$$

Therefore, the spectral norm of $\nabla^2 g(u)$ is bounded by $l^2/\|u\|^l$.

For any $i \in [m]$, let u_i'' be any point on the line segment between u_i and u_i', then $\left\|\nabla^2 g(u_i'')\right\|_2 \le l^2/\|u_i''\|^l$. If $c_i = 1/\|u_i\|$, we have

$$\begin{aligned}\left|(c_i')^{l-2} - c_i^{l-2}\right| \le& \|\nabla g(u_i)\| \left\|\eta \nabla_{u_i} f(U, \bar{C})\right\| + \frac{1}{2}\max_{u_i''}\left\|\nabla^2 g(u_i'')\right\|_2 \left\|\eta \nabla_{u_i} f(U, \bar{C})\right\|^2 \\ \le& \frac{l-2}{\|u_i\|^{l-1}}\left\|\eta \nabla_{u_i} f(U, \bar{C})\right\| + \frac{1}{2}\max_{u_i''}\frac{l\|u_i\|}{\|u_i''\|^l}\left\|\eta \nabla_{u_i} f(U, \bar{C})\right\|.\end{aligned}$$

If $c_i = \sqrt{d(m+K)}/\|u_i\|$, we have

$$\left|(c_i')^{l-2} - c_i^{l-2}\right| \le \frac{l-2}{\|u_i\|^{l-1}}\left\|\eta \nabla_{u_i} f(U, \bar{C})\right\| (\sqrt{d(m+K)})^{l-2} + \frac{1}{2}\max_{u_i''}\frac{l\|u_i\|}{\|u_i''\|^l}\left\|\eta \nabla_{u_i} f(U, \bar{C})\right\| (\sqrt{d(m+K)})^{l-2}.$$

Therefore, we have

$$\begin{aligned}\left\|T' - \hat{T}\right\|_F \le& 2el\sum_{i=1}^m \left\|\eta \nabla_{u_i} f(U, \bar{C})\right\| \max(\|u_i\|, 2\sqrt{m+K}\delta(\sqrt{d(m+K)})^{l-2}) \\ \le& 2\sqrt{2}elM_2\left\|\eta \nabla_U f(U, \bar{C})\right\|_F,\end{aligned}$$

where the first inequality holds because $\|u_i'\| \le \left(1 + \frac{1}{l}\right)\|u_i\|$ and the second inequality assumes $\delta^2 \le \frac{M_2^2}{4m(m+K)^{l-1}d^{l-2}}$.

Overall, combing the bounds on $\left\|\hat{T} - T\right\|_F$ and $\left\|T' - \hat{T}\right\|_F$, we have

$$\|T' - T\|_F \le 200l^2\sqrt{\frac{1}{\lambda}}\left\|\eta \nabla_U f(U, \bar{C})\right\|_F.$$

□

Now we are ready to prove Lemma 21.

Proof of Lemma 21. Let's first bound the tensor change and function value change due to scalar mode switches. Following the proof of Claim 2 in Lemma 15, setting $\tilde{\Gamma} = 10$ and assuming $\eta \le \frac{1}{l(\sqrt{2\Gamma}(\sqrt{d(m+K)})^{l-2}+\lambda)}$, we know each scalar mode switch can at most change the tensor Frobenius norm by $16(m+K)\delta^2(\sqrt{d(m+K)})^{l-2}$. Furthermore, using the same argument as Claim 2, the function value can increase by at most $\sqrt{20}\left(16(m+K)\delta^2(\sqrt{d(m+K)})\right.$ $\frac{1}{2}\left(16(m+K)\delta^2(\sqrt{d(m+K)})^{l-2}\right)^2 \le 160(m+K)\delta^2(\sqrt{d(m+K)})^{l-2}$, where we assume $\delta^2 \le \frac{5}{8(m+K)(\sqrt{d(m+K)})^{l-2}}$.

According to the algorithm, we know each epoch contains at most m scalar mode switches. Suppose T_τ' be the tensor before potential scalar mode switch in the τ-th iteration. Then, we have

$$\begin{aligned}\sum_{\tau=1}^t \|T_\tau - T_{\tau-1}\|_F \le& \sum_{\tau=1}^t \|T_\tau' - T_{\tau-1}\|_F + \sum_{\tau=1}^t \|T_\tau - T_\tau'\|_F \\ \le& \sum_{\tau=1}^t \|T_\tau' - T_{\tau-1}\|_F + 16m(m+K)\delta^2(\sqrt{d(m+K)})^{l-2}.\end{aligned}$$

According to Lemma 22, we know

$$\|T'_\tau - T_{\tau-1}\|_F \le 200l^2\sqrt{\frac{1}{\lambda}}\left\|\eta\nabla_U f(U_{\tau-1},\bar{C}_{\tau-1})\right\|_F.$$

Therefore, we have

$$\begin{aligned}\sum_{\tau=1}^{t}\|T'_\tau - T_{\tau-1}\|_F \le& 200l^2\sqrt{\frac{1}{\lambda}}\sum_{\tau=1}^{t}\left\|\eta\nabla_U f(U_{\tau-1},\bar{C}_{\tau-1})\right\|_F\\ \le& 200l^2\sqrt{\frac{1}{\lambda}}\sqrt{t}\sqrt{\sum_{\tau=1}^{t}\left\|\eta\nabla_U f(U_{\tau-1},\bar{C}_{\tau-1})\right\|_F^2}.\end{aligned}$$

According to Lemma 14, we know $f(U'_\tau,\bar{C}'_\tau) - f(U_{\tau-1},\bar{C}_{\tau-1}) \le -\frac{\eta}{l}\left\|\nabla_U f(U_{\tau-1},\bar{C}_{\tau-1})\right\|_F^2$. Therefore, we have

$$\sum_{\tau=1}^{t}\|T'_\tau - T_{\tau-1}\|_F \le 200l^2\sqrt{\frac{1}{\lambda}}\sqrt{t}\sqrt{\sum_{\tau=1}^{t}\eta l\left(f(U_{\tau-1},\bar{C}_{\tau-1}) - f(U'_\tau,\bar{C}'_\tau)\right)}.$$

Since scalar mode switches in total change the function value by at most $160m(m+K)\delta^2(\sqrt{d(m+K)})^{l-2}$, we know

$$\begin{aligned}&\sum_{\tau=1}^{t}\left(f(U_{\tau-1},\bar{C}_{\tau-1}) - f(U'_\tau,\bar{C}'_\tau)\right)\\ \le& f(U_0,\bar{C}_0) - f(U_t,\bar{C}_t) + 160m(m+K)\delta^2(\sqrt{d(m+K)})^{l-2}.\end{aligned}$$

Overall, we have

$$\begin{aligned}&\sum_{\tau=1}^{t}\|T_\tau - T_{\tau-1}\|_F\\ \le& 200l^{2.5}\sqrt{\frac{1}{\lambda}}\sqrt{\eta H}\sqrt{f(U_0,\bar{C}_0) - f(U_t,\bar{C}_t) + 160m(m+K)\delta^2(\sqrt{d(m+K)})^{l-2}}\\ &+ 16m(m+K)\delta^2(\sqrt{d(m+K)})^{l-2}.\end{aligned}$$

□

Combining all the steps above, we are now ready to prove Lemma 12.

Proof of Lemma 12. Let u_0 be the reinitialized vector. According to Lemma 18, we know with probability at least 1/5,

$$a(P_{S^{\otimes l}}T'_0 - P_{S^{\otimes l}}T^*)(\overline{P_S u_0}^{\otimes l}) \le \frac{-1}{(\mu_1 rl)^{l/2}}\left\|P_{S^{\otimes l}}T'_0 - P_{S^{\otimes l}}T^*\right\|_F \le -\frac{\epsilon}{(\mu_1 rl)^{l/2}},$$

where μ_1 is some constant. According to Lemma 23, we know with probability at least $1-1/30$, $\|P_S u_0\| \ge \frac{\mu_2\delta}{\sqrt{d}}$ for some constant $\mu_2 < 1$. Taking a union bound, we know both properties hold with probability at least 1/6.

Conditioning on both properties, we will prove that

$$f(U_0,\bar{C}_0) - f(U_H,\bar{C}_H) \ge \frac{\lambda}{32000000(\mu_1 rl)^l\eta H l^5}\epsilon^2.$$

For the sake of contradiction, assume that $f(U_0, \bar{C}_0) - f(U_H, \bar{C}_H) \leq \frac{\lambda}{32000000(\mu_1 rl)^l \eta H l^5}\epsilon^2$. According to Lemma 21, we know

$$\sum_{\tau=1}^{H} \|T_\tau - T_{\tau-1}\|_F \leq \frac{\epsilon}{10(\mu_1 rl)^{l/2}}$$

as long as $\delta^2 \leq \frac{\epsilon}{320(\mu_1 rl)^{l/2} m(m+K)^{\frac{l}{2}} d^{\frac{l-2}{2}}}$ and $\delta^2 \leq \frac{\lambda\epsilon^2}{32000000(\mu_1 rl)^l \eta H l^5 \cdot 160 m(m+K)^{\frac{l}{2}} d^{\frac{l-2}{2}}}$.

We will prove that $a(P_{S^{\otimes l}} T_t - P_{S^{\otimes l}} T^*)(\overline{P_S u_t}^{\otimes l}) \leq -\frac{\epsilon}{5(C_1 rl)^{l/2}}$ for all $0 \leq t \leq H-1$, so from Lemma 19 we know that the norm of $P_S u_t$ must increase exponentially.

Let's first prove the case at the beginning of an epoch: Let T_0 be the tensor after reinitialization. According to the proof of Claim 1 in Lemma 15, we know

$$\|T_0 - T_0'\|_F \leq 2\sqrt{\frac{10}{\lambda m}} \leq \frac{\epsilon}{2(\mu_1 rl)^{l/2}},$$

where the last inequality assumes $\lambda m \geq \frac{160(\mu_1 rl)^l}{\epsilon^2}$. This implies that

$$a(P_{S^{\otimes l}} T_0 - P_{S^{\otimes l}} T^*)(\overline{P_S u_0}^{\otimes l}) \leq a(P_{S^{\otimes l}} T_0' - P_{S^{\otimes l}} T^*)(\overline{P_S u_0}^{\otimes l}) + \|T_0 - T_0'\|_F \leq -\frac{\epsilon}{2(\mu_1 rl)^{l/2}}.$$

For later steps, we will show that $a(P_{S^{\otimes l}} T_t - P_{S^{\otimes l}} T^*)(\overline{P_S u_t}^{\otimes l})$ is close to $a(P_{S^{\otimes l}} T_0 - P_{S^{\otimes l}} T^*)(\overline{P_S u_0}^{\otimes l})$. Actually,

$$\begin{aligned}
&\left|a(P_{S^{\otimes l}} T_t - P_{S^{\otimes l}} T^*)(\overline{P_S u_t}^{\otimes l}) - a(P_{S^{\otimes l}} T_0 - P_{S^{\otimes l}} T^*)(\overline{P_S u_0}^{\otimes l})\right| \\
\leq &\left|\sum_{\tau=1}^{t} \left((P_{S^{\otimes l}} T_{\tau-1} - P_{S^{\otimes l}} T^*)(\overline{P_S u_\tau}^{\otimes l}) - (P_{S^{\otimes l}} T_{\tau-1} - P_{S^{\otimes l}} T^*)(\overline{P_S u_{\tau-1}}^{\otimes l})\right)\right| \\
&+ \left|\sum_{\tau=1}^{t} \left((P_{S^{\otimes l}} T_\tau - P_{S^{\otimes l}} T^*)(\overline{P_S u_\tau}^{\otimes l}) - (P_{S^{\otimes l}} T_{\tau-1} - P_{S^{\otimes l}} T^*)(\overline{P_S u_\tau}^{\otimes l})\right)\right| \\
\leq &H\mu l^4 2^l d^{l-1.5} m^{1/2} (m+K)^{l-1} \eta\delta\lambda + \sum_{\tau=1}^{t} \|T_\tau - T_{\tau-1}\| \\
\leq &H\mu l^4 2^l d^{l-1.5} m^{1/2} (m+K)^{l-1} \eta\delta\lambda + \frac{\epsilon}{10(\mu_1 rl)^{l/2}} \leq \frac{\epsilon}{5(\mu_1 rl)^{l/2}}.
\end{aligned}$$

The second inequality above comes from Lemma 20, and the last inequality assumes $\delta \leq \frac{1}{\mu l^4 2^l d^{l-1.5} m^{1/2} (m+K)^{l-1} \eta\lambda H} \cdot \frac{\epsilon}{10(\mu_1 rl)^{l/2}}$.

This then implies that for all $0 \leq t \leq H-1$,

$$a(P_{S^{\otimes l}} T_t - P_{S^{\otimes l}} T^*)(\overline{P_S u_t}^{\otimes l}) \leq -\frac{\epsilon}{2(\mu_1 rl)^{l/2}} + \frac{\epsilon}{5(\mu_1 rl)^{l/2}} \leq -\frac{\epsilon}{5(\mu_1 rl)^{l/2}}.$$

Then according to Lemma 19,

$$\begin{aligned}
\|P_S u_H\|^2 &\geq \left(1 + \eta\left(\frac{\mu_2}{2}\right)^{l-2} \frac{\epsilon}{10(\mu_1 rl)^{l/2}}\right)^H \|P_S u_0\|^2 \\
&\geq \left(1 + \eta\left(\frac{\mu_2}{2}\right)^{l-2} \frac{\epsilon}{10(\mu_1 rl)^{l/2}}\right)^H \frac{\mu_2^2 \delta^2}{d} \\
&\geq \exp\left(\frac{1}{2}\eta H \left(\frac{\mu_2}{2}\right)^{l-2} \frac{\epsilon}{10(\mu_1 rl)^{l/2}}\right) \frac{\mu_2^2 \delta^2}{d},
\end{aligned}$$

where the last inequality assumes $\eta \leq \left(\frac{2}{\mu_2}\right)^{l-2} \frac{10(\mu_1 rl)^{l/2}}{\epsilon}$. Therefore, $\|P_S u_H\|^2$ exceeds M_2 as long as $\eta H \geq 2\left(\frac{2}{\mu_2}\right)^{l-2} \frac{10(\mu_1 rl)^{l/2}}{\epsilon} \log\left(\frac{dM_2}{\mu_2^2 \delta^2}\right)$. Since $M_2 = \sqrt{\frac{10}{\lambda}}$ is the upper bound of $\|U\|_F$, this finishes the contradiction proof.

We have shown that

$$f(U_0, \bar{C}_0) - f(U_H, \bar{C}_H) \geq \frac{\lambda}{32000000(\mu_1 rl)^l \eta H l^5} \epsilon^2.$$

In order to show $f(U_0', \bar{C}_0') - f(U_H, \bar{C}_H)$ is large, we still need to bound $|f(U_0', \bar{C}_0') - f(U_0, \bar{C}_0)|$ that comes from reinitialization. According to Lemma 15, we know

$$|f(U_0', \bar{C}_0') - f(U_0, \bar{C}_0)| \leq \frac{200}{\lambda m} \leq \frac{\lambda}{64000000(\mu_1 rl)^l \eta H l^5} \epsilon^2,$$

where the second inequality assumes $\lambda^2 m \geq 1.28 \times 10^{11} (\mu_1 rl)^l \eta H l^5$. Therefore,

$$\begin{aligned}
& f(U_H, \bar{C}_H) - f(U_0', \bar{C}_0') \\
\leq & \big(f(U_0, \bar{C}_0) - f(U_H, \bar{C}_H)\big) + |f(U_0, \bar{C}_0) - f(U_0', \bar{C}_0')| \\
\leq & -\frac{\lambda}{3.2 \times 10^7 (\mu_1 rl)^l \eta H l^5} \epsilon^2 + \frac{\lambda}{6.4 \times 10^8 (\mu_1 rl)^l \eta H l^5} \epsilon^2 \\
\leq & -\frac{\lambda}{6.4 \times 10^7 (\mu_1 rl)^l \eta H l^5} \epsilon^2.
\end{aligned}$$

We choose $m = O\left(\frac{r^{2.5l}}{\epsilon^5} \log(d/\epsilon)\right)$, $\lambda = O\left(\frac{\epsilon}{r^{0.5l}}\right)$, $\delta = O\left(\frac{\epsilon^{5l-1.5}}{d^{l-1.5}(\log(d/\epsilon))^{l+0.5} r^{2.5l^2-0.75l}}\right)$, $\eta = O\left(\frac{\epsilon^{15l-4.5}}{d^{3l-4.5}(\log(d/\epsilon))^{3l+1.5} r^{7.5l^2-2.25l}}\right)$ $H = O\left(\frac{d^{3l-4.5}(\log(d/\epsilon))^{3l+2.5} r^{7.5l^2-1.75l}}{\epsilon^{15l-3.5}}\right)$ and $K = O\left(\frac{r^{2l}}{\epsilon^4} \log(d/\epsilon)\right)$ such that all the conditions are satisfied and the function value decreases by $\Omega(\frac{\epsilon^4}{r^{2l}\log(d/\epsilon)})$ in each epoch. Note that there does exist some circular dependency between the parameters. This turns out to be not an issue in our proof because for example δ depends on $\frac{1}{\eta H}$ while ηH only depends logarithmically on $1/\delta$. Other circular dependencies can be resolved in the same manner. □

D Tools

In this section, we give the technical lemmas we use in the proof.

D.1 Random projection on a subspace

We use the following lemma to show that with good probability, the projection of the reinitialized component on the good subspace is lower bounded.

Lemma 23 (Lemma 2.2 in Dasgupta and Gupta (2003))**.** *Let Y be a d-dimensional vector uniformly sampled from sphere $\mathbb{S}^{d-1}$. Let $Z \in \mathbb{R}^k$ be the projection of Y onto its first k coordinates ($k < d$). For any $\beta < 1$, we have*

$$\Pr\left[\|Z\|^2 \leq \frac{\beta k}{d}\right] \leq \exp\left(\frac{k}{2}(1 - \beta + \ln \beta)\right).$$

D.2 Norm of random Gaussian vectors

The following lemma gives the concentration of ℓ_2 norm of a random Gaussian vector.

Lemma 24 (Theorem 3.1.1 in Vershynin (2018)). *Let $X = (X_1, X_2, \cdots, X_n) \in \mathbb{R}^n$ be a random vector with each entry independently sampled from $\mathcal{N}(0,1)$. Then*

$$\Pr\left[\left|\|x\| - \sqrt{n}\right| \geq t\right] \leq 2\exp\left(-t^2/C^2\right),$$

where C is an absolute constant.

D.3 Anti-concentration of Gaussian polynomials

We use anti-concentration of Gaussian polynomials to argue that a randomly initialized component has good correlation with the residual.

Lemma 25 (Theorem 8 in Carbery and Wright (2001)). *Let $x \in \mathbb{R}^n$ be a Gaussian variable $x \in N(0, I)$, for any polynomial $p(x)$ of degree l, there exists a constant κ such that*

$$\Pr\left[|p(x)| \leq \epsilon\sqrt{Var[p(x)]}\right] \leq \kappa l \epsilon^{1/l}.$$