Open Problem: How much overparametrization is needed for ALS in tensor decomposition?

Dionysis Arvanitakis DIONARVA@U.NORTHWESTERN.EDU
Northwestern University
Vaidehi Srinivas VAIDEHI@U.NORTHWESTERN.EDU
Northwestern University
Aravindan Vijayaraghavan ARAVINDV@NORTHWESTERN.EDU
Northwestern University

Editors: Steve Hanneke and Tor Lattimore

Abstract

We ask how much overparameterization is needed for simple iterative methods such as alternating least squares (ALS) and gradient descent to decompose a third-order tensor. This question can be viewed as a basic setting to study feature learning: when a rank-r tensor in ambient dimension n has $r \ll n$, the latent rank-one components are the features, and k is the amount of overparameterization used by the algorithm. For rank r tensors, recent work shows that overparametrized rank $k = O(r^2)$ suffices for the popular ALS heuristic (with random initialization) to converge to a global optima. Is the quadratic dependence on r an inherent barrier for ALS-like methods? We pose the open problem of proving convergence to the global optimum for $k = o(r^2)$, or proving that a lower bound on the overparametrized rank of $k = \Omega(r^{1+c})$ for some absolute constant $c > 0$ is necessary.

Keywords: Tensor decomposition, Iterative methods, Non-convex optimization, Alternating least squares, Gradient descent

1. Introduction

Tensor decomposition is a canonical non-convex optimization problem. Given a third-order tensor $T \in \mathbb{R}^{n\times n\times n}$, the tensor decomposition (also called CP decomposition) problem asks for a representation

$$T = \sum_{i=1}^{r} a_i \otimes b_i \otimes c_i, \tag{1}$$

where $a_i, b_i, c_i \in \mathbb{R}^n$. Tensor methods have played a central role in algorithms for latent variable models, mixtures, topic models, independent component analysis, and related problems; see, e.g., Kolda and Bader (2009); Moitra (2018). In the worst case, tensor rank and tensor decomposition are computationally intractable (Håstad, 1990; Hillar and Lim, 2013). Nevertheless, under suitable genericity, incoherence, orthogonality, or smoothed-analysis assumptions, there are polynomial-time algorithms based on diagonalization, tensor power methods, spectral methods, sum-of-squares, and algebraic techniques (Harshman, 1970; Leurgans et al., 1993; Bhaskara et al., 2014a; Ma et al., 2016; Sharan and Valiant, 2017; Vijayaraghavan; Kothari et al., 2025).

In practice, however, the workhorse algorithms are much simpler iterative methods: alternating least squares (ALS), alternating minimization, and gradient descent Kolda and Bader (2009). These methods minimize the least-squares objective

$$\min_{x_i,y_i,z_i\in\mathbb{R}^n,\ i\in[k]} \left\| T - \sum_{i=1}^{k} x_i \otimes y_i \otimes z_i \right\|_F^2. \tag{2}$$

When $k = r$, the parameterization matches the target rank. In the overparameterized regime, the algorithm is allowed to output a decomposition with $k > r$ terms. Overparameterization is known empirically to make non-convex optimization easier, but it is still poorly understood even in structured problems such as tensor decomposition. Moreover, this also makes tensor decomposition a basic setting to understand *feature learning* in machine learning. An algorithm that finds a decomposition of rank $k \ll n$ from random initialization learns a simple feature representation of the data, e.g., in the ground-truth rank-r decomposition, the unknown rank-one components are low-dimensional latent features.

Prior work on ALS and Gradient Descent. Observe that if $\{y_i\}_{i=1}^k$ and $\{z_i\}_{i=1}^k$ are fixed then optimizing over $\{x_i\}_{i=1}^k$ constitutes a simple linear least squares problem that can be solved efficiently (the same of course holds for the other two modes as well). The algorithm that we are mostly interested in for our questions, called alternating least squares (ALS), iteratively fixes two of the three modes and optimizes over the third. The method continues by alternating between modes until convergence.

A recent result of Arvanitakis et al. (2025) gives the first global convergence guarantee for ALS with moderate overparameterization. Informally, if T has a mildly conditioned rank-r decomposition and n is sufficiently large compared to r, then a parallel variant of ALS initialized randomly and run with rank $k = O(r^2)$ converges with high probability to a global optimum of (2). This improves over earlier guarantees of Wang et al. (2020) for a variant of gradient descent requiring $k = O(r^{7.5} \log n)$, and it is essentially independent of the ambient dimension n.

The proof of Arvanitakis et al. (2025) reveals a natural r^2-dimensional bottleneck. The analysis succeeds by showing that, after random initialization, the algorithmic Khatri–Rao features $\{x_i \odot y_i\}_{i=1}^k$ span the column span of all the pairwise products in the Kronecker product $A \otimes B$. This span-containment argument requires $k = \Omega(r^2)$ and does not extend below the quadratic threshold. The central question is whether this is an artifact of the analysis, or a genuine limitation of ALS and related iterative algorithms. Experiments in Arvanitakis et al. (2025) suggest that superlinear overparameterization is indeed necessary.

2. Open Problems

We now formally pose our question. Let

$$T = \sum_{j=1}^{r} a_j \otimes b_j \otimes c_j, \tag{3}$$

where the factor matrices are $A = [a_1 \ \cdots \ a_r]$, $B = [b_1 \ \cdots \ b_r]$, and $C = [c_1 \ \cdots \ c_r]$ and $r \ll n$. We are most interested in the setting where the factors are "typical", and not worst-case. We will state our problems in the *smoothed analysis* setting where the factors are drawn *generically* i.e., they are obtained by perturbing arbitrary well-conditioned factors by small independent Gaussian noise, as described below.

Smoothed Analysis Setting. Let $\overline{A}, \overline{B}, \overline{C}$ be arbitrary $n \times r$ matrices. For each $i \in [r]$, the factor a_i is a random perturbation of $\overline{a}_i$ i.e., $a_i - \overline{a}_i \sim \mathcal{N}\big(0, \frac{\rho^2}{n}\big)^n$; similarly $b_i - \overline{b}_i$, and $c_i - \overline{c}_i$ are also drawn i.i.d. from the Gaussian with covariance $\frac{\rho^2}{n} I$. We will think of $\rho \approx 1/\operatorname{poly}(r)$, so the random perturbations all have average length $1/\operatorname{poly}(r)$.

Given as input T, we run an iterative algorithm such as ALS or gradient descent on (2) with algorithmic rank k and random initialization.

Open Problem 1 (Broad open question) *What is the smallest value of $k = k(r)$ so that ALS (or other iterative algorithms like GD) on smoothed instances (under random initialization) converges to the global optimum with high probability (over the smoothed instance)?*

We split this up into two questions— one algorithmic, and the other a lower bound question, which together aim to make progress towards the above question beyond the current state-of-the-art.

Open Problem 2 (Subquadratic overparameterization for iterative tensor decomposition) *Given as input a smoothed tensor $T \in \mathbb{R}^{n\times n\times n}$ as described above with $\rho = 1/\operatorname{poly}(r)$, does overparametrization of $k = o(r^2)$ suffice for ALS or gradient descent (or other iterative algorithms) to find with high probability (over the randomness in the instance) a rank-k decomposition in time $\operatorname{poly}(n, r, \log(1/\epsilon))$ satisfying*

$$\left\| T - \sum_{i=1}^{k} x_i \otimes y_i \otimes z_i \right\|_F \leqslant \varepsilon \|T\|_F?$$

A positive resolution to the above question entails an algorithmic guarantee for an iterative algorithm like ALS or gradient descent. While the above question was phrased for smoothed instances, it is conceivable that the result holds for any instance with a well-conditioned distribution i.e., the condition number $\sigma_1(A)/\sigma_r(A)$ is upper bounded by $\operatorname{poly}(r)$ (and similarly for B, C). We also remark that the high probability is over the randomness in smoothed analysis (since the input is only drawn once). However for a fixed input, a success probability of a constant (or even inverse polynomial) over the random initialization would be sufficient for our purposes— we can then try out many random initializations to successfully decompose the tensor.

We also remark that there exist more sophisticated algebraic methods e.g., simultaneous diagonalization that recover the ground-truth decomposition for generic/ smoothed tensors (in fact, there is a unique rank-r decomposition for a generic tensor) Harshman (1972); Bhaskara et al. (2014a); Koiran (2024); Kothari et al. (2025) with polynomial time guarantees in the setting of Open Problem 2. However, the focus of our open question is to understand the performance of iterative methods. The above algorithmic problem is also interesting to explore for random instances (where the factors are Gaussian or random unit vectors), where one may hope for even stronger guarantees on k or a wider range of r e.g., the overcomplete setting with $r \gg n$ (Ge and Ma, 2017).

A negative resolution would be just as interesting. In fact, resolving the following open question would represent a striking lower bound against iterative methods.

Open Problem 3 (Polynomial lower bound) *Can one prove that there exists a universal constant $c > 0$ such that ALS and gradient descent (and other iterative methods) require overparametrization of $k = r^{1+c}$ i.e., given a smoothed instance as described earlier and overparametrized rank $k \leqslant r^{1+c}$, the algorithm converges to a solution of positive objective value with constant probability?*

We suspect that if such a lower bound holds, it likely holds even for random tensors of this rank. Moreover, it is even conceivable that the above lower bound holds even for rank $k = r^{2-c}$ for a small constant $c > 0$. Providing an explicit tensor (with well-conditioned factors) for which ALS fails unless $k = \Theta(r^2)$ would be interesting, but the most compelling lower bound would not rely on worst-case pathological tensors. Instead, it would show that for a random rank-r tensor, or a smoothed tensor an overparameterized rank below r^2 leads to a genuine obstruction. We are particularly interested in the lower bound for random tensors, where experimental results show that superlinear overparameterization is indeed necessary (see Arvanitakis et al. (2025)).

3. Why this question is interesting

Tensor decomposition provides a compelling testbed for understanding *overparameterization and feature learning* in non-convex optimization. In many learning problems the ambient dimension n is very large, while the intrinsic number of latent features is much smaller, say $r \ll n$. Tensor decomposition is one of the simplest settings where this structure is explicit: the unknown factors $\{a_j, b_j, c_j\}_{j=1}^r$ are the latent features, and the optimization algorithm must discover a low-dimensional feature space containing them from random initialization. Thus, as k gets closer to r, the algorithm must learn the true feature structure using only mild overparameterization.

This setting is attractive because it is algebraically structured enough that sharp statements may be possible, while the objective (2) already contains many difficulties from modern machine learning: non-convexity, spurious stationary points, and feature discovery from random initialization. A positive answer would give a clean example of feature learning in a nonlinear, non-convex model beyond the lazy-training or kernel regime. A negative answer would be equally informative: it would suggest that even when the statistical model has only r latent degrees of freedom, local dynamics may require many more algorithmic features to reliably find them. Moreover, although more sophisticated polynomial-time algorithms are known in this regime, the methods of choice in practice are scalable iterative heuristics such as ALS and gradient descent (Ballard and Kolda, 2025); despite their wide use, their behavior is still poorly understood.

Resolving either open problem would be technically interesting for the following reasons.

- A positive resolution of Open Problem 2 would show that simple iterative methods can exploit low-rank tensor structure much more efficiently than current analyses suggest. Proving guarantees below rank r^2 appears to require a more global analysis of how random initialization, alternating updates, and nonlinear feature interactions gradually align with the true components, going beyond the current proof strategy based on the r^2-dimensional Kronecker lift of the factor matrices. This could provide a new mechanism for understanding feature learning under mild overparameterization, distinct from analyses based on very large overparameterization or near-linearized dynamics.
- Conversely, a positive resolution of Open Problem 3 would identify an inherent optimization barrier: although the target tensor has only r components and is statistically well-conditioned, local iterative dynamics may require about r^2 algorithmic components to escape poor alignment. Such a lower bound for random or smoothed instances would highlight a fundamental difference between practical instances and average-case models, and would motivate more refined beyond-worst-case assumptions for tensor inputs.
- Finally there appears to be an algorithmic phase transition at overparameterization rank of r^2: while finding a decomposition of rank r^2 is easy, as it reduces to a matrix problem (see e.g. Bhaskara et al. (2014b)), all known algorithms that can decompose a tensor using less than r^2 seem to use more sophisticated ideas. In fact, we are not aware of any algorithm that decomposes a tensor with overparameterization less than r^2 without also recovering the unique decomposition of rank r. It would be very interesting to understand where iterative methods lie: can they go beyond matrices and exploit the tensor structure of the problem?

Prize. As a lower bound (Problem 3) is established as $\Omega(r^{(1+\alpha)})$ and an upper bound (Problem 2) is established as $O(r^{(1+\beta)})$, we (the authors) will donate $(1-(\beta-\alpha)) \times \$300$ to research in treating Amytrophic Lateral Sclerosis (ALS), totaling $\$300$ when the upper and lower bounds meet. In addition, the first authors to publish an improvement to either bound will receive coffee from Metropolis coffee, the best coffee roaster in Chicago, perhaps the world.

References

Dionysis Arvanitakis, Vaidehi Srinivas, and Aravindan Vijayaraghavan. Guarantees for alternating least squares in overparameterized tensor decompositions. *Advances in Neural Information Processing Systems*, 38:78062–78112, 2025.

Grey Ballard and Tamara G. Kolda. *Tensor Decompositions for Data Science*. Cambridge University Press, 2025.

Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Symposium on the Theory of Computing (STOC)*, 2014a.

Aditya Bhaskara, Moses Charikar, and Aravindan Vijayaraghavan. Uniqueness of tensor decompositions with applications to polynomial identifiability. *Conference on Learning Theory*, 2014b.

Rong Ge and Tengyu Ma. On the optimization landscape of tensor decompositions. In *Advances in Neural Information Processing Systems 30*. 2017.

Richard A. Harshman. Foundations of the parafac procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, 1970.

Richard A. Harshman. Determination and proof of minimum uniqueness conditions for PARAFAC1. *UCLA Working Papers in Phonetics*, 22:111–117, 1972. URL https://psychology.uwo.ca/faculty/harshman/wpppfac1.pdf.

Johan Håstad. Tensor rank is NP-complete. *Journal of Algorithms*, 11(4), 1990.

Christopher J. Hillar and Lek-Heng Lim. Most tensor problems are NP-hard. *Journal of the American Mathematical Society*, 60, 2013.

Pascal Koiran. On the uniqueness and computation of commuting extensions. *Linear Algebra and its Applications*, 703:645–666, 2024. ISSN 0024-3795. doi: https://doi.org/10.1016/j.laa.2024.10.004. URL https://www.sciencedirect.com/science/article/pii/S0024379524003835.

Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM review*, 51 (3):455–500, 2009.

Pravesh K Kothari, Ankur Moitra, and Alexander S Wein. Overcomplete tensor decomposition via koszul-young flattenings. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1871–1882. IEEE, 2025.

S. E. Leurgans, R. T. Ross, and R. B. Abel. A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083, 1993. doi: 10.1137/0614071. URL https://doi.org/10.1137/0614071.

Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *Proceedings of the 57th Annual IEEE Symposium on the Foundations of Computer Science (FOCS)*, 2016.

Ankur Moitra. *Algorithmic Aspects of Machine Learning*. Cambridge University Press, 2018. doi: 10.1017/9781316882177.

Vatsal Sharan and Gregory Valiant. Orthogonalized als: A theoretically principled tensor decomposition algorithm for practical use. In *International Conference on Machine Learning*, pages 3095–3104. PMLR, 2017.

Aravindan Vijayaraghavan. chapter 19 Efficient Tensor Decomposition. URL https://arxiv.org/abs/2007.15589.

Xiang Wang, Chenwei Wu, Jason D Lee, Tengyu Ma, and Rong Ge. Beyond lazy training for over-parameterized tensor decomposition. *Advances in Neural Information Processing Systems*, 33: 21934–21944, 2020.