Landscape analysis of an improved power method for tensor decomposition

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this work, we consider the optimization formulation for symmetric tensor decomposition recently introduced in the Subspace Power Method (SPM) of Kileel & Pereira (2019). Unlike popular alternative functionals for tensor decomposition, the SPM objective function has the desirable property that its global optima are *exactly* the rank-1 components of the tensor (in the case of a sufficiently low-rank tensor). We analyze the non-convex optimization landscape associated with the SPM objective. Specifically, we derive quantitative bounds such that any second-order critical point with SPM objective value exceeding the bound must equal a tensor component in the noiseless case, and must approximate a tensor component in the noisy case. We obtain, for decomposing tensors of size $D^{\times m}$, a "near-global" guarantee up to rank $\widetilde{o}(D^{\lfloor m/2 \rfloor})$ under a random tensor model, and a global guarantee up to rank $\mathcal{O}(D)$ assuming deterministic frame conditions. Our analysis accounts for working with noisy tensors, and implies that SPM with suitable initialization is a provable, efficient, robust algorithm for low-rank symmetric tensor decomposition. Further, we include numerics to show a distinct practical preferability for using the SPM functional over a more established counterpart.

1 Introduction

From applied mathematics [24] to machine learning [31] and multivariate statistics [26] to many-body quantum systems [35], high-dimensional data sets arise that are naturally organized into higher-order arrays. Frequently these arrays, known as hypermatrices or *tensors*, are decomposed into low-rank representations. In particular, the real symmetric CANDECOMP/PARAFAC (CP) decomposition [10] is often appropriate:

$$T = \sum_{i=1}^{K} \lambda_i a_i^{\otimes m}. \tag{1}$$

In this decomposition, we are given T, a tensor (real symmetric) of size $D^{\times m}$. The goal is to expand T as a sum of K rank-1 terms, coming from scalar/vector pairs $(\lambda_i, a_i) \in \mathbb{R} \times \mathbb{R}^D$. Importantly, the number of terms K must be minimal possible for the given T. In this case, K is the *rank* of T. By fundamental results in algebraic geometry [7], when $m > 2$ and $K = \mathcal{O}(D^m)$, CP decompositions are generically unique (up to permutation and scaling). An actionable interpretation [3] is that

CP decomposition infers (well-defined) latent variables encoded by T.

Indeed in learning applications, symmetric tensors T are formed from statistical moments (higher-order covariances) or multivariate derivatives (higher-order Hessians). Here CP decomposition has enabled parameter estimation for mixtures of Gaussians [17, 30], generalized linear models [29], shallow neural networks [16, 21, 36], deeper networks [14, 15, 27] and hidden Markov models [5].

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

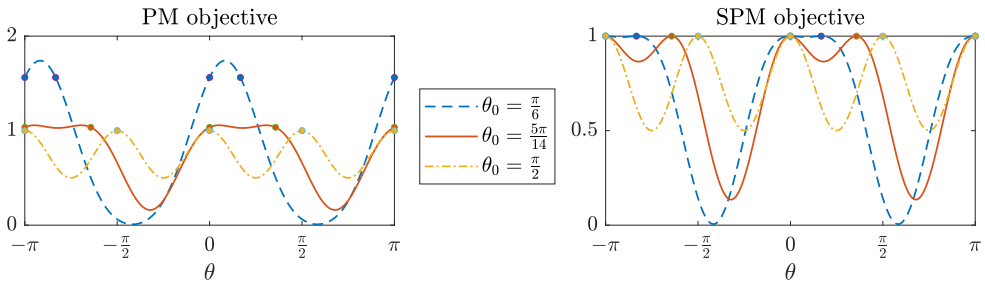


Figure 1: Illustration of **PM-P** objective (left) and **SPM-P** objective (right) for $D = K = 2$ and $m = 4$, where $\mathbb{S}^1$ is mapped to $(-\pi, \pi]$. We set $a_1 = (1, 0)$, $a_2 = (\cos\theta_0, \sin\theta_0)$ for different values of θ_0. If a_1 and a_2 are not orthogonal ($\theta_0 \neq \frac{\pi}{2}$), then a_1 and a_2 do not coincide with global or local maximizers of **PM-P**. Too much correlation ($\theta_0 \leq \frac{\pi}{3}$) leads to a collapsing of two distinct maxima into one isolated maximum for **PM-P**. On the other hand, **SPM-P** is not affected in this way by correlation: local maxima coincide with global maxima coincide with tensor components a_1, a_2.

In the worst case, symmetric CP decomposition is NP-hard [20]. In fact, it is believed to possess a *computational-to-statistical gap* [6], with efficient algorithms expected to exist for random tensors only up to rank $K = \mathcal{O}(D^{m/2})$. For this reason, and due to its non-convexity and high-dimensionality, CP decomposition has served theoretically as a test case for better understanding mysteries of non-convex optimization landscapes. To date, this focus has been on the non-convex program

$$\max_{x \in \mathbb{R}^D : \|x\|_2 = 1} \left\langle T, x^{\otimes m} \right\rangle . \tag{PM-P}$$

The problem is labeled **PM-P**, because projected gradient ascent corresponds to the Shifted Symmetric Higher-Order Power Method of Kolda & Mayo (2011) [25]. Landmark analyses of **PM-P** include Ge & Ma (2020) [18] as well as [4].

In the present paper, we perform an analysis of the non-convex optimization landscape associated with the Subspace Power Method (SPM) of Kileel & Pereira (2019) for computing symmetric tensor decompositions [22]. SPM is based on the following non-convex program:

$$\max_{x \in \mathbb{R}^D : \|x\|_2 = 1} F_{\mathcal{A}}(x), \tag{SPM-P}$$

where $F_{\mathcal{A}}(x) := \|P_{\mathcal{A}}(x^{\otimes n})\|^2, \quad n := \lceil m/2 \rceil, \quad \mathcal{A} := \mathrm{Span}\{a_1^{\otimes n}, \ldots, a_K^{\otimes n}\},$

and $P_{\mathcal{A}} : (\mathbb{R}^D)^{\otimes n} \to \mathcal{A}$ is orthogonal projection with respect to Frobenius inner product.

Section 3 discusses this functional, as well as the overall method of [22].

The optimization landscape of **SPM-P** is compelling to study for two reasons. First, the authors of [22] argued through numerical experiments their algorithm outperforms state-of-the-art tensor decomposition methods within its applicable rank range of $K = \mathcal{O}(D^{\lfloor m/2 \rfloor})$, giving a one order of magnitude speed-up over methods in [23] as implemented in Matlab's Tensorlab [34] while still matching the numerical stability of FOOBI [11]. Thus, SPM is a *practical* algorithm. Second, from a *theory* standpoint, the program **SPM-P** has certain desirable properties that **PM-P** lacks. In the setting of a rank K tensor $T = \sum_{i=1}^K \lambda_i a_i^{\otimes m}$ with $K \lesssim D^{\lfloor m/2 \rfloor}$ and $\{(\lambda_i, a_i)\}_{i=1}^K$ Zariski-generic,[1]

- each $\pm a_i$ is *exactly* a global maximum for **SPM-P**, and there are *no other* global maxima;
- the maximum value for **SPM-P** is exactly 1, so the objective value in **SPM-P** provides a *certificate* for global optimality (in particular, non-global critical points can be discarded).

These properties were shown for **SPM-P** in [22], but both fail for **PM-P** (Figure 1). Therefore, **SPM-P** is more relevant than **PM-P** as a test problem in the context of non-convex tensor decompositions.

Previously in [22], it was proven that projected gradient ascent, initialized at almost any starting point, with a constant explicitly bounded step size, must converge to a second-order critical point of **SPM-P**

[1] *Zariski-generic* means that the failure set can be described by the vanishing non-zero polynomials [19], so in particular, has Lebesgue measure 0.

at a power rate or faster. This left open the possible existence of *spurious* second-order critical points, i.e., second-order points with reasonably high objective value that are not global maxima (unequal to, possibly even distant from, each $\pm a_i$). Such critical points would pose a challenge to the successful optimization of **SPM-P**. All theory for **SPM-P** in [22], as well as the landscape analysis for **PM-P** in [18], was restricted to the *clean case*: when the tensor T is *exactly* of a sufficiently low rank K.

Main contributions. We perform a landscape analysis of **SPM-P** by characterizing all second-order critical points, using suitable assumptions about $a_1, \ldots, a_K$. Under deterministic frame conditions, which are satisfied by mutually incoherent vectors, near-orthonormal systems, and random ensembles of size $K = \mathcal{O}(D)$, Theorem 6 shows that all second-order critical points of **SPM-P** coincide exactly with $\pm a_i$. Theorem 15 shows the same result for overcomplete random ensembles of size $K = \tilde{o}(D^{\lfloor m/2 \rfloor})$, however requiring an additional superlevel set condition to exclude maximizers with vanishing objective values. Both results extend to noisy tensor decomposition, where SPM is applied to a perturbation $\hat{T} \approx T$. In this setting, non-degenerate local maximizer with objective values exceeding $\mathcal{O}(\|\hat{T} - T\|)$ are $\mathcal{O}(\|\hat{T} - T\|)$-near to one of the components $\pm a_i$. We also show in Lemma 8 that spurious local maximizers do exist in the noisy case.

The results imply a clear separation of the function landscape between near global maximizers, with objective values close to 1, and spurious local maximizers with small objective value. Hence, the **SPM-P** objective can be used to validate a final iterate of projected gradient ascent.

Finally, we expose the relation between **PM-P** and **SPM-P**. Precisely, **SPM-P** can be expressed as **PM-P** with appropriate insertion of the inverse of the Grammian $(G_n)_{ij} := \langle a_i, a_j \rangle^n$. The resulting de-biasing effect on the local maximizers with respect to the components $\pm a_i$ (Figure 1) is responsible for many advantages of **SPM-P**. As a by-product, we state a conjecture about the minimal eigenvalue of G_n when a_i are i.i.d. uniformly randomly on the sphere, which may be of independent interest.

2 Notation

Vectors and matrices. When x is a vector, $\|x\|_p$ is the ℓ_p-norm ($p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$). For $x, y \in \mathbb{R}^D$, the entrywise (or Hadamard) product is $x \odot y \in \mathbb{R}^D$, and the entrywise power is $x^{\odot s} := x \odot \ldots \odot x \in \mathbb{R}^D$ ($s \in \mathbb{N}$). When M is a matrix, $\|M\|_2$ is the spectral norm. If M is real symmetric, $\mu_j(M)$ is the eigenvalue of M that is the j-th largest in absolute value. The identity matrix is $\mathsf{Id}_D \in \mathbb{R}^{D \times D}$.

Tensors. A real tensor of length D and order m is an array of size $D \times D \times \ldots \times D$ (m times) of real numbers. Use $\mathcal{T}_D^m := \left(\mathbb{R}^D\right)^{\otimes m} \cong \mathbb{R}^{D^m}$ for the space of tensors of size $D^{\times m}$. Meanwhile, $\mathrm{Sym}(\mathcal{T}_D^m) \subseteq \mathcal{T}_D^m$ is the subspace of symmetric tensors (i.e., tensors unchanged by any permutation of indices). The notation $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ shall refer to the Frobenius inner product and norm, respectively. Given a linear subspace of tensors $\mathcal{B} \subseteq \mathcal{T}_D^m$, let $P_{\mathcal{B}} : \mathcal{T}_D^m \to \mathcal{B}$ denote the orthogonal projector onto $\mathcal{B}$ with respect to $\langle \cdot, \cdot \rangle$. In the case $\mathcal{B} = \mathrm{Sym}(\mathcal{T}_D^m)$, the projector $P_{\mathrm{Sym}(\mathcal{T}_D^m)}$ is the symmetrization operator, $\mathrm{sym} : \mathcal{T}_D^m \to \mathrm{Sym}(\mathcal{T}_D^m)$. Given $T \in \mathcal{T}_D^{m_1}$ and $S \in \mathcal{T}_D^{m_2}$, the tensor (or outer) product is $T \otimes S \in \mathcal{T}_D^{m_1+m_2}$, defined by $(T \otimes S)_{i_1,\ldots,i_{m_1+m_2}} := T_{i_1,\ldots,i_{m_1}} S_{i_{m_1+1},\ldots,i_{m_2}}$. For $T \in \mathcal{T}_D^m$ and $s \in \mathbb{N}$, the tensor power is $T^{\otimes s} := T \otimes \ldots \otimes T \in \mathcal{T}_D^{sm}$. For $T \in \mathcal{T}_D^{m_1}$, $S \in \mathcal{T}_D^{m_2}$ with $m_1 \geq m_2$, the contraction $T \cdot S \in \mathcal{T}_D^{m_1-m_2}$ is defined by $(T \cdot S)_{i_1,\ldots,i_{m_1-m_2}} := \sum_{j_1,\ldots,j_{m_2}} T_{i_1,\ldots,i_{m_1-m_2},j_1,\ldots,j_{m_2}} S_{j_1,\ldots,j_{m_2}}$. Define $\mathrm{reshape}(T, [d_1, \ldots, d_\ell])$ as the function that reshapes the tensor T to have dimensions $d_1, \ldots, d_\ell$, as in corresponding Matlab/NumPy commands.

Other. The unit sphere in $\mathbb{R}^D$ is $\mathbb{S}^{D-1}$, and $\mathrm{Unif}(\mathbb{S}^{D-1})$ is the associated uniform probability distribution. Given a function $f : \mathbb{R}^D \to \mathbb{R}$, the Euclidean gradient and Hessian matrix at $x \in \mathbb{R}^D$ are $\nabla f(x) \in \mathbb{R}^D$ and $\nabla^2 f(x) \in \mathrm{Sym}(\mathcal{T}_D^2)$. The Riemannian gradient and Hessian with respect to $\mathbb{S}^{D-1}$ at $x \in \mathbb{S}^{D-1}$ are $\nabla_{\mathbb{S}^{D-1}} f(x)$ and $\nabla^2_{\mathbb{S}^{D-1}} f(x)$ (see [1]). Write Span for linear span, $[K] := \{1, \ldots, K\}$, and $|A|$ for the cardinality of a finite set A. Lastly, we use asymptotic notation freely.

3 Symmetric tensor decomposition via Subspace Power Method

In this section, we outline the tensor decomposition method of [22], and provide the first insights to the program **SPM-P**. Throughout we assume $m \geq 3$ is an integer and define $n := \lceil m/2 \rceil$.

SPM algorithm. The input is a tensor $\hat{T} \in \mathrm{Sym}(\mathcal{T}_D^m)$, with the promise that $\hat{T} \approx T = \sum_{i=1}^K \lambda_i a_i^{\otimes m}$ for $\{(\lambda_i, a_i)\}_{i=1}^K$ Zariski-generic and $K \leq \binom{D+n-1}{n} - D$ (if m is even) and $K \leq D^n$ (if m is odd). As a first step, SPM computes the orthogonal projection $P_{\hat{\mathcal{A}}} : \mathcal{T}_D^m \to \hat{\mathcal{A}}$, projecting onto the column span of $\mathrm{reshape}(\hat{T}, [D^n, D^{m-n}])$, by using matrix singular value decomposition. Provided that $\hat{T} \approx T$, the associated subspace approximation error $\hat{\mathcal{A}} \approx \mathcal{A} = \mathrm{Span}\left\{a_1^{\otimes n}, \ldots, a_K^{\otimes n}\right\}$, defined by

$$\Delta_{\hat{\mathcal{A}}} := \left\|P_{\mathcal{A}} - P_{\hat{\mathcal{A}}}\right\|_{F\to F} = \sup_{T \in \mathcal{T}_D^n,\ \|T\|_F = 1} \left\|P_{\mathcal{A}}(T) - P_{\hat{\mathcal{A}}}(T)\right\|_F \tag{2}$$

can be bounded as follows (stated only for even m).

Lemma 1 (Error in the subspace)**.** *Let $T \in \mathrm{Sym}(\mathcal{T}_D^{2n})$ and assume that $M := \mathrm{reshape}(T, [D^n, D^n])$ has K nonzero eigenvalues. Let $\hat{T} \in \mathrm{Sym}(\mathcal{T}_D^{2n})$, $\hat{M} := \mathrm{mat}(\hat{T})$, and assume $2\Delta_T := \|T - \hat{T}\| < |\mu_K(M)|$. We have*

$$\left\|P_{\mathrm{Im}(M)} - P_{\mathrm{Im}_K(\hat{M})}\right\|_{F\to F} \leq \frac{\pi}{\sqrt{2}} \frac{\Delta_T}{|\mu_K(M)| - \Delta_T}, \tag{3}$$

where $\mathrm{Im}_K(\hat{M})$ denotes the subspace spanned by the K leading eigenvectors of $\hat{M}$. In particular, if $T = \sum_{i=1}^K \lambda_i a_i^{\otimes 2n}$, $\mathcal{A} = \mathrm{Span}\left\{a_i^{\otimes n} : i \in [K]\right\}$, and $\hat{\mathcal{A}}$ is the subspace spanned by K leading tensorized eigenvectors of $\hat{M}$, then the right hand side of (3) *bounds $\Delta_{\mathcal{A}}$.*

Given $\hat{\mathcal{A}}$, SPM next attempts to find a tensor component a_i by solving a noisy **SPM-P**, given by

$$\max_{x \in \mathbb{S}^{D-1}} F_{\hat{\mathcal{A}}}(x), \quad \text{where} \quad F_{\hat{\mathcal{A}}}(x) := \left\|P_{\hat{\mathcal{A}}}(x^{\otimes n})\right\|^2. \tag{\textbf{nSPM-P}}$$

Starting from a random initial point $x_0 \sim \mathrm{Unif}(\mathbb{S}^{D-1})$, the projected gradient ascent iteration

$$x \leftarrow P_{\hat{\mathcal{A}}}(x^{\otimes n}) \cdot x^{\otimes(n-1)} + \gamma x, \tag{4}$$

with step size γ (see [22] for an upper bound) is guaranteed to converge to constrained second order critical points almost surely [22]. We denote the final iterate as $\hat{a}_i$ in what follows.

Given access to an approximate tensor component $\hat{a}_i$, the third step in SPM uses a deflation formula based on Wedderburn rank reduction [9] to compute the corresponding weight $\hat{\lambda}_i$. Then we update the tensor $\hat{T} \leftarrow \hat{T} - \hat{\lambda}_i \hat{a}_i^{\otimes m}$, and repeat the second and third step K times to compute all tensor components $\{(\hat{\lambda}_i, \hat{a}_i)\}_{i=1}^K$.

Preparatory material about nSPM-P. Our goal is to show that second-order critical points of **nSPM-P** with reasonable function value must be near the global maximizers $\pm a_1, \ldots, \pm a_K$, of **SPM-P** under suitable incoherence assumptions about the rank-one components $a_1, \ldots, a_K$. Naturally, the optimality conditions of **nSPM-P** play an important part in the analysis.

Proposition 2 (Optimality conditions)**.** *Let $x, z \in \mathbb{S}^{D-1}$ with $z \perp x$. The stationary point and second order optimality conditions of* **nSPM-P** *imply*

$$P_{\hat{\mathcal{A}}}(x^{\otimes n}) \cdot x^{\otimes(n-1)} = F_{\hat{\mathcal{A}}}(x) x \tag{5}$$

$$F_{\hat{\mathcal{A}}}(x) \geq n \|P_{\hat{\mathcal{A}}}(x^{\otimes(n-1)} \otimes z)\|^2 + (n-1)\langle P_{\hat{\mathcal{A}}}(x^{\otimes n}), x^{\otimes(n-2)} \otimes z^{\otimes 2}\rangle. \tag{6}$$

If the vectors $a_1^{\otimes n}, \ldots, a_K^{\otimes n}$ were linearly dependent, we can still span $\mathcal{A}$ after removing one of the tensor components from the system. In particular, recovering all components is not possible if $a_1^{\otimes n}, \ldots, a_K^{\otimes n}$ are linearly dependent, and we thus exclude these cases in the following.

In the analysis later, we make frequent use of expressing the objective $F_{\mathcal{A}}(x)$ using the Gram matrix $(G_n)_{ij} := \langle a_i^{\otimes n}, a_j^{\otimes n}\rangle = \langle a_i, a_j\rangle^n$. Under linear independence, the inverse G_n^{-1} exists and the noiseless program **SPM-P** can be expressed as follows.

Lemma 3. *Let $A := [a_1 | \ldots | a_K] \in \mathbb{R}^{D\times K}$ and $\{a_i^{\otimes n} : i \in [K]\}$ be linearly independent. We have*

$$F_{\mathcal{A}}(x) = \left\|P_{\mathcal{A}}(x^{\otimes n})\right\|^2 = \left((A^\top x)^{\odot(n)}\right)^\top G_n^{-1} \left((A^\top x)^{\odot(n)}\right). \tag{7}$$

Lemma 3 clarifies the relation between **SPM-P** and **PM-P**. While **PM-P** can be rewritten as $\langle (A^\top x)^{\odot n}, (A^\top x)^{\odot n}\rangle$, **SPM-P** takes into account correlations among the tensors $a_1^{\otimes n}, \dots, a_K^{\otimes n}$ and inserts the Grammian G_n^{-1} into the quadratic form (7). As a consequence, correlations among the tensor components are considered in **SPM-P**, despite not knowing the tensors $a_1^{\otimes n}, \dots, a_K^{\otimes n}$ *a priori*.

In the special case of orthonormal systems, or more generally, systems that resemble equiangular tight frames [13, 32], the SPM and PM objectives coincide up to shift and scaling.

Lemma 4. *Assume there exist $\rho \in [0,1]$ and $M \in \mathbb{R}_{\geq 1}$ such that $\langle a_i, a_j\rangle^n = \rho$ for all $i \neq j$ and $\sum_{i\in[K]} \langle x, a_i\rangle^n = M$ for all $x \in \mathbb{S}^{D-1}$. Denote $A = [a_1|\dots|a_K] \in \mathbb{R}^{D\times K}$. Then*

$$F_{\mathcal{A}}(x) = (1-\rho)^{-1}\|A^\top x\|_{2n}^{2n} - \big(1 + K\rho(1-\rho)^{-1}\big)^{-1}\rho M^2.$$

4 Main results

In this section, we present the main results about local maximizers of the **nSPM-P** program. Section 4.1 is tailored to low-rank tensor models with $K = \mathcal{O}(D)$ components that satisfy certain deterministic frame conditions. Section 4.2 then considers the overcomplete case $K = \widetilde{o}(D^{\lfloor m/2\rfloor})$ in an average case scenario, where $a_1, \dots, a_K$ are modeled as independent copies of an isotropic random vector.

4.1 Low-rank tensors under deterministic frame conditions

Motivated by frame constants in frame theory [8], we measure the incoherence of the ensemble $a_1, \dots, a_K$ by scalars $\rho_s \in \mathbb{R}_{>0}$, which are defined via

$$\rho_s := \sup_{x\in\mathbb{S}^{D-1}} \sum_{i=1}^{K} |\langle x, a_i\rangle|^s - 1. \tag{8}$$

They satisfy the order relation $\rho_s \leq \rho_{s'}$ for $s' \leq s$, due to $\|a_i\|_2 = 1$, and can be related to extremal eigenvalues of Grammians G_s and $G_{\lfloor s/2\rfloor}$ as shown in the following result.

Lemma 5. *Let $\{a_i : i \in [K]\} \subseteq \mathbb{S}^{D-1}$ and $(G_s)_{ij} := \langle a_i, a_j\rangle^s$ for $s \in \mathbb{N}$. Then we have*

$$1 - \rho_s \leq \mu_K(G_s) \leq \mu_1(G_s) \leq 1 + \rho_s \leq \mu_1(G_{\lfloor s/2\rfloor}). \tag{9}$$

The characterization in Lemma 5 allows to compute bounds for ρ_s for low-rank tensors with mutually incoherent components or rank $\mathcal{O}(D)$ tensor with random components. We provide details on this in Remark 7 below, but first state the main guarantee about local maximizer using ρ_2 and ρ_n.

Theorem 6 (Main deterministic result)**.** *Assume there exists $\tau \in (0,1)$ such that ρ_2, ρ_n satisfy*

$$8n^2\rho_2 + 4n(2n+3)\frac{\rho_n}{1-\rho_n} \leq 1 - \tau. \tag{10}$$

The program **nSPM-P** *has exactly $2K$ local maximizers in the level set $F_{\hat{\mathcal{A}}}(x) \geq 4n(3n+4)\Delta_{\mathcal{A}}/\tau$, and each of these local maximizers satisfies*

$$\min_{i\in[K],\ s\in\{-1,1\}} \|x - sa_i\|_2^2 \leq \frac{2\Delta_{\mathcal{A}}}{n}. \tag{11}$$

Remark 7. Using Lemma 5 we identify two classes, where ρ_2 and ρ_n can be bounded from above.

1. Mutually incoherent ensembles. Let $a_1, \dots, a_K$ have mutual incoherence $\rho := \max_{i\neq j} |\langle a_i, a_j\rangle|$. Using Gershgorin's circle theorem, we obtain

$$\rho_s \leq \mu_1(G_{\lfloor s/2\rfloor}) - 1 \leq \max_{i\in[K]} \sum_{j\neq i} |\langle a_i, a_j\rangle|^{\lfloor s/2\rfloor} \leq (K-1)\rho^{\lfloor s/2\rfloor} \quad \text{for any } s \in \mathbb{N}_{\geq 2}.$$

2. Low-rank random ensembles. Let $a_1, \dots, a_K$ be independent copies of an isotropic unit norm random vector a with sub-gaussian norm $\mathcal{O}(1/\sqrt{D})$. With high probability, the ensemble can achieve arbitrarily small ρ_2, provided $K \leq CD$ for small enough $C > 0$ (more details in Remark 16).

In the noiseless, where $\Delta_{\mathcal{A}} = 0$, Theorem 6 shows that all local maximizers coincide with global maximizers $\pm a_1, \dots, \pm a_K$, provided that the incoherence condition (10) is satisfied for some $\tau \in (0, 1)$. In the noisy case, all local maximizers with objective values $F_{\hat{\mathcal{A}}}(x) \geq C(n, \tau)\Delta_{\mathcal{A}}$ are $\mathcal{O}(\Delta_{\mathcal{A}})$-close to the global optimizers of the noiseless objective $F_{\mathcal{A}}$, where the constant $C(n, \tau)$ increases as the incoherence of the vectors $a_1, \dots, a_K$ shrinks. Unfortunately, the presence of degenerate local maximizers with small objective values cannot be avoided under our noise model.

Lemma 8. *Fix $\delta \in (0, 1)$. Let $a \in \mathbb{S}^{D-1}$, $\mathcal{A} = \operatorname{Span}\{a^{\otimes n}\}$. There exists a subspace $\hat{\mathcal{A}} \subseteq \operatorname{Sym}(\mathcal{T}_D^n)$ with $\|P_{\mathcal{A}} - P_{\hat{\mathcal{A}}}\|_{F \to F} = \delta$ so that* **nSPM-P** *has a local maximizer with objective value δ^2.*

4.2 Average case analysis of overcomplete tensors

The overcomplete case with $K = \widetilde{o}(D^{\lfloor m/2 \rfloor})$ is not in range of Theorem 6 in the previous section, because the vectors $a_1, \dots, a_K$ cannot satisfy (10) for $K \gg D$. Instead, the analysis of the overcomplete case relies on $A = [a_1 | \dots | a_K] \in \mathbb{R}^{D \times K}$ satisfying the (δ, p)-restricted isometry property (RIP) for $p = \mathcal{O}(D/\log(K))$.

Definition 9. Let $H \in \mathbb{R}^{D \times K}$, $1 \leq p \leq K$ be an integer, and $\delta \in (0, 1)$. We say that H is (p, δ)-RIP if every $D \times p$ submatrix H_p of H satisfies $\|H_p^\top H_p - \mathsf{Id}_p\|_2 \leq \delta$.

A consequence of the RIP, which is particularly useful in the analysis of overcomplete tensor models, is that the correlation coefficients $\{\langle a_i, x \rangle : i \in [K]\}$ can naturally be split into two groups.

Proposition 10. *Suppose that $A = [a_1 | \dots | a_K] \in \mathbb{R}^{D \times K}$ satisfies the (δ, p)-RIP for $p = \lceil 1 + \delta) D/\log(K) \rceil$. For all $x \in \mathbb{S}^{D-1}$ there is a subset of indices $\mathcal{I}(x) \subseteq [K]$ with cardinality p such that*

$$\sum_{i \in \mathcal{I}(x)} \langle a_i, x \rangle^2 \leq 1 + \delta \quad \textit{and} \quad \langle a_i, x \rangle^2 \leq \frac{\log(K)}{D} \textit{ for } i \notin \mathcal{I}(x). \tag{12}$$

Let us now state all assumptions used to analyse the overcomplete case.

Definition 11 (Assumptions)**.** Let $A := [a_1 | \dots | a_K]$. We require the following assumptions:

- **A1** A is $((1 + \delta)D/\log(K), \delta)$-RIP
- **A2** There exists $c_1 > 0$, independent of K, D, so that $\max_{i \neq j} \langle a_i, a_j \rangle^2 \leq c_1 \log(K)/D$
- **A3** There exists $c_2 > 0$, independent of K, D, so that $\left\|G_n^{-1}\right\|_2 \leq c_2$

To the best of our knowledge, there are no deterministically constructed systems of size $K = \widetilde{o}(D^{\lfloor m/2 \rfloor})$, which satisfy **A1** - **A3** for small $\delta \ll 1$ and constants c_1, c_2 that do not depend on K, D. However, by modeling components via a sufficiently spread random vector, i.e., by considering an average case scenario with $a_i \sim \operatorname{Unif}(\mathbb{S}^{D-1})$, they hold with high probability.

Lemma 12. *Let $a \sim \operatorname{Unif}(\mathbb{S}^{D-1})$ and assume $a_1, \dots, a_K$ are K independent copies of a. Fix an arbitrary constant $\delta \in (0, 1)$. There exists a universal constant $C > 0$ and $D_0 \in \mathbb{N}$ depending only on δ such that for all $D \geq D_0$, and with probability at least $1 - D^{-1} - 2\exp(-C\delta^2 D)$, conditions* **A1** *and* **A2** *(with $c_1 \leq C$) hold. Furthermore, if $n = 2$ and $K = o(D^2)$,* **A3** *holds with probability $1 - C\exp(-D)$ for $c_2 \leq C$.*

Remark 13 (**A3** when $n > 2$)**.** Following [2, 12], we give a self-contained proof for **A3** in the case $n = 2$ in the supplementary material. We are currently not able to extend the proof to $n > 2$, because some technical tools such as the Hanson-Wright inequality [28] and an extension of [2, Theorem 3.3], which ensures the RIP for matrices whose columns have sub-exponential tails, have not yet been fully developed for random vectors with dependent entries and heavier tails (so-called α-sub-exponential tails or sub-Weibull tails). We strongly believe that **A3** holds for $n > 2$ and $K = o(D^n)$. For now, we formulate this as a conjecture and supply numerical evidence in the supplementary material.

Conjecture 14 (Grammians of independent symmetric rank-one tensors)**.** *Let $a_1, \dots, a_K$ be independent copies of the random vector $a \sim \operatorname{Unif}(\mathbb{S}^{D-1})$ and fix $\epsilon > 0$ arbitrarily. There exists some*

constant $\kappa_n > 0$ depending only on n, and an increasing function $\gamma_n : (0, \kappa_n) \to (0, 1)$, such that if $K \leq \kappa D^n$ for some $\kappa < \kappa_n$, we have

$$\mathbb{P}(\left\|G_n^{-1}\right\|_2 \geq \gamma_n(\kappa) - \epsilon) \to 1 \quad \textit{as} \quad D \to \infty. \tag{13}$$

Moreover, $\gamma_n(\kappa) \to 1$ as $\kappa \to 0$, and $\|G_n^{-1}\|_2 \to 1$ if $K = o(D^n)$.

We now give our main theorem about local maximizers of **nSPM-P** in the overcomplete case.

Theorem 15 (Main random overcomplete result)**.** *Let $K, D \in \mathbb{N}$, define $\varepsilon_K := K \log^n(K)/D^n$, and suppose that $\lim_{D\to\infty} \varepsilon_K = 0$. Assume $a_1, \ldots, a_K \in \mathbb{R}^D$ satisfy* **A1** - **A3** *for some $\delta, c_1, c_2 > 0$. There exists $\delta_0, D_0, \Delta_0, C$, which depend only on n, c_1, and c_2, such that if $\delta < \delta_0$, $D > D_0$, and $\Delta_{\mathcal{A}} \leq \Delta_0$, the program* **nSPM-P** *has exactly $2K$ local maximizers in the superlevel set*

$$\left\{x \in \mathbb{S}^{D-1} : F_{\hat{\mathcal{A}}}(x) \geq C\varepsilon_K + 13\Delta_{\mathcal{A}}\right\}. \tag{14}$$

Furthermore, each of these local maximizers satisfies

$$\min_{i\in[K], s\in\{-1,1\}} \|x - sa_i\|_2^2 \leq \frac{2\Delta_{\mathcal{A}}}{n}. \tag{15}$$

By Theorem 15, all non-degenerate local maximizers in the superlevel set (14) are close to global maximizers of the noiseless objective $F_{\mathcal{A}}$. In the noiseless case $\Delta_{\mathcal{A}} = 0$, they coincide with global maximizers and the right hand side in (14) tends to 0 as $D \to \infty$. Hence, we have a clear separation between the global maximizers $\pm a_1, \ldots, \pm a_K$ and degenerate local maximizers with vanishing objective value, so that the objective value acts as a certificate for the validity of an identified maximizer.

Theorem 15 is not a global guarantee because a a random starting point $x_0 \sim \text{Unif}(\mathbb{S}^{D-1})$, used for starting the projected gradient ascent iteration (4), is, with high probability, not contained in (14). In practice, we never experienced convergence to degenerate local maximizers in the noiseless case and under suitably randomized tensor component models. Hence, it might be possible to weaken the condition in (14) further; see, e.g., the **PM-P** analysis in [18].

Remark 16 (Low-rank random tensors)**.** If $K \leq CD$ and $a_1, \ldots, a_K$ are independent copies of an isotropic unit norm random vector with $\mathcal{O}(1/\sqrt{D})$ sub-Gaussian norm (e.g., $a_i \sim \text{Unif}(\mathbb{S}^{D-1})$), the result [33, Theorem 5.65] implies that $A = [a_1|\ldots|a_K]$ is (K, δ)-RIP with high probability and arbitrarily small $\delta > 0$ (provided D is large enough and C is small enough). By (12), this implies $\rho_2 \leq \delta$. Hence, Theorem 6 globally characterizes local maximizers of **nSPM-P** in this case.

Remark 17 (Noiseless deflation)**.** The deflation procedure, which was introduced in [22], updates $T \leftarrow T - \hat{\lambda}_i(\hat{a}_i)^{\otimes m}$, where $\hat{a}_i$ is a local maximizer of **nSPM-P**. In the noiseless case, Theorem 15 guarantees $\hat{a}_i = a_i$ (and as a by-product $\hat{\lambda}_i = \lambda_i$ [22]), provided we filter out spurious local maximizer by checking the objective of the final iterate. In that case the deflation is exact and there is no error accumulation over the successive iterations needed to recover all tensor components.

Remark 18 (Noisy deflation)**.** In the noisy case, we have $\|\hat{a}_i - a_i\|_2^2 = \mathcal{O}(\Delta_{\mathcal{A}})$, suggesting that errors can accumulate over successive iterates. To reduce error propagation, we recommend the following modification. Assuming $\ell < K$ components have been approximated so far, we first compute a local maximizer $\tilde{a}_{\ell+1}$ of **nSPM-P** with subspace $\mathcal{A}$ induced by the current deflated tensor $\hat{T}_\ell := \hat{T} - \hat{\lambda}_1\hat{a}_1^{\otimes m} - \ldots - \hat{\lambda}_\ell\hat{a}_\ell^{\otimes m}$. Then, instead of deflating, we use $\tilde{a}_{\ell+1}$ as a starting point in projected gradient ascent for computing a local maximizer $\hat{a}_{\ell+1}$ of **nSPM-P** based on the original (non-deflated) $\hat{T}$. Provided $\tilde{a}_{\ell+1}$ is contained in the level set (14), the vector $\hat{a}_{\ell+1}$, which is used for the next deflation, is an $\mathcal{O}(\Delta_{\mathcal{A}})$-approximation to a true tensor component. The effectiveness of the approach is illustrated in the numerical section in Figure 2c, where error propagation is completely avoided. Unfortunately, we do not yet have a theoretical understanding of the phenomenon.

5 Proof ingredients

We now briefly outline the main steps in the proofs of Theorems 6 and 15, while deferring the details to the supplementary material. We recall the definition $A := [a_1|\ldots|a_K] \in \mathbb{R}^{D\times K}$.

The proofs consists of three steps. We begin by combining the optimality conditions (5) and (6) of **nSPM-P** to get, for any $y \in \mathbb{S}^{D-1}$, the inequality

$$(1+2(n-1)\langle x, y\rangle^2)F_{\hat{\mathcal{A}}}(x) \geq n\|P_{\hat{\mathcal{A}}}(x^{\otimes(n-1)} \otimes y)\|_F^2 + (n-1)\langle P_{\hat{\mathcal{A}}}(x^{\otimes n}), x^{\otimes(n-2)} \otimes y^{\otimes 2}\rangle. \tag{16}$$

Plugging in $y = a_i$, where i is any index with $|\langle a_i, x\rangle| = \|A^\top x\|_\infty$, and using either incoherence constants ρ_s and ρ_n or Assumptions **A1** - **A3**, a suitable re-arrangement of (16) immediately gives

$$\|A^\top x\|_\infty^2 \geq \begin{cases} 1 - 2\rho_2 - 2\frac{\rho_n}{1-2\rho_n} - 3\frac{\Delta_\mathcal{A}}{F_\mathcal{A}(x)}, & \text{for Theorem 6,} \\ 1 - 2\delta - \bar{C}\left(\frac{\varepsilon_K}{F_\mathcal{A}(x)}\right)^{\frac{1}{n}} - 3\frac{\Delta_\mathcal{A}}{F_\mathcal{A}(x)}, & \text{for Theorem 15,} \end{cases} \tag{17}$$

where the constant $\bar{C}$ in the second case depends on n, c_1 and c_2. For Theorem 6, this immediately suggests the largest correlation coefficient satisfies $\|A^\top x\|_\infty \geq 1 - \mathcal{O}(\rho_2 + \rho_n + \Delta_\mathcal{A})$, provided $F_\mathcal{A}(x) \gg \Delta_\mathcal{A}$ as assumed in the statement. In the case of Theorem 15, and within the level set $F_{\hat{\mathcal{A}}}(x) \geq C\varepsilon_K + 13\Delta_\mathcal{A}$, (17) first implies that $\|A^\top x\|_\infty$ is bounded from below by a constant that is independent of D, K. However, by $f(x) \geq \|A^\top x\|_\infty^{2n}$, we can update the lower bound on $F_\mathcal{A}(x)$ and reuse it again in (16). The resulting bootstrap argument implies $\|H^\top x\|_\infty \geq 1 - \mathcal{O}(\varepsilon_K^{1/n} + \Delta_\mathcal{A})$, i.e., there must be at least one large correlation coefficient.

The second step is to show that the Riemannian Hessian $\nabla^2_{\mathbb{S}^{D-1}} F_{\hat{\mathcal{A}}}$ is negative definite on spherical caps around the tensor components $\pm a_i$. We prove, uniformly for all unit norm $z \perp x$, the bound

$$\frac{1}{2n} z^\top \nabla^2_{\mathbb{S}^{D-1}} \hat{f}(x) z \leq \begin{cases} -1 + n\frac{2-\rho_n}{1-\rho_n}\eta + \frac{3n-2}{1-\rho_n}\rho_n + 4\Delta_\mathcal{A}, & \text{for Theorem 6,} \\ -1 + \bar{C}\sqrt{\eta} + \tilde{C}(\delta + \varepsilon_K)^{\frac{n-1}{n}} + 4\Delta_\mathcal{A}, & \text{for Theorem 15,} \end{cases} \tag{18}$$

where $\eta := 1 - \max_{i\in[K]} \langle a_i, x\rangle^{2n}$, and $\bar{C}, \tilde{C}$ depend only on n, c_1, c_2. Hence, the Riemannian Hessian is negative definite and $F_{\hat{\mathcal{A}}}$ is strictly concave in spherical caps around $\pm a_i$.

In the third step, we consider a spherical cap B_{sa_i} ($s \in \{-1, 1\}$) around sa_i. By compactness, a local maximizer x^* of $F_{\hat{\mathcal{A}}}$ must exist within the cap B_{sa_i} and by strict concavity it must be unique. Moreover, its objective is bounded from below $F_{\hat{\mathcal{A}}}(x^*) \geq F_{\hat{\mathcal{A}}}(a_i) \geq 1 - \Delta_\mathcal{A}$. Using a Taylor expansion of the function $g(t) = F_{\hat{\mathcal{A}}}(\gamma(t))$ along the geodesic path $\gamma(t)$ from x^* to a_i, and combining it with the upper bound in (18), we obtain $\|x^* - sa_i\|_2^2 \leq 2\Delta_\mathcal{A}/n$. Since we can start the argument with a spherical cap, which is strictly larger than $\{y \in \mathbb{S}^{D-1} : \langle y, sa_i\rangle \geq 1 - \Delta_\mathcal{A}/n\}$, it follows that x^* is not on the boundary of B_{sa_i} and thus x^* is a local maximizer of **nSPM-P**.

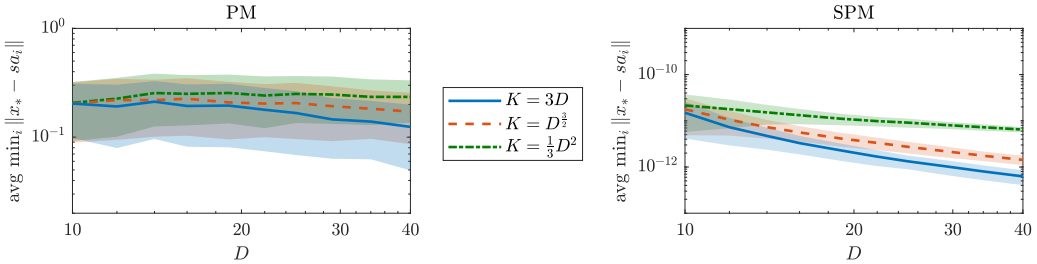
We complete the proof by noting that all local maximizers of **nSPM-P** contained in the superlevel set, which is specified in the respective theorem, satisfies $\|A^\top x\|_\infty \geq 1 - \epsilon$, where ϵ depends on $\Delta_\mathcal{A}$ and ρ_2, ρ_n, or ε_K. In particular, they are contained in spherical caps B_{sa_i} constructed in the third step.

6 Numerical experiments

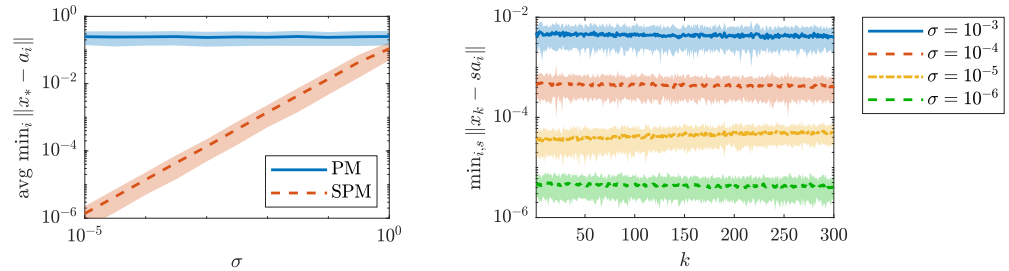
In this section we present numerical experiments that corroborate the theoretical findings of Section 4. We illustrate that **SPM-P** identifies exact tensor components in the noiseless case, which is in contrast to **PM-P**, and that **nSPM-P** behaves robustly in noisy tensor decomposition. We also refer to extensive numerical experiments in [22], which test **SPM-P** in a variety of scenarios, and justify it as a possible replacement to state-of-the-art methods like FOOBI [11] or Tensorlab [34]. We use the implementation of SPM provided by Kileel and Pereira, available on `https://github.com/joaompereira/SPM`, which is licensed under the MIT license. All experiments can be conducted on a standard laptop computer within a few hours.

Global optimizers and noise robustness. In the first set of experiments we are interested in the recovery of individual tensor components a_i for different D, K from noisy, approximately low rank tensors. We fix the tensor order $m = 4$ and create noiseless tensors by sampling vectors $\tilde{a}_i \sim \mathcal{N}(0, \mathsf{Id}_D)$ and using $a_i = \tilde{a}_i / \|\tilde{a}_i\|_2$ (i.e., $a_i \sim \text{Unif}(\mathbb{S}^{D-1})$) as the tensor components. The tensor weights are given by $\lambda_i = \|\tilde{a}_i\|_2^m / \sqrt{K}$, so that the variance of each tensor element is about 1. We construct noisy tensors $\hat{T} \approx T$ by adding independent copies of $\epsilon \sim \mathcal{N}(0, \sigma^2)$ to each entry of the tensor and then project onto $\text{Sym}(\mathcal{T}_D^m)$.

After constructing the tensor, we sample 10 points from $\text{Unif}(\mathbb{S}^{D-1})$ as starting points for the projected gradient ascent applied to **PM-P** and **nSPM-P**. We iterate until convergence and record the distance between the final iterate and the closest $\pm a_i$ among the tensor components. Averages and standard deviations (represented by shaded areas) over 1000 distances (10 distances per tensor, 100



(a) Recovering individual tensor components a_i using **PM-P** (left) and **nSPM-P** (right) in the noiseless case;



(b) **PM-P** and **nSPM-P** applied to noisy tensors; (c) Error of indiv. recovery with **nSPM-P** after k deflations.

Figure 2: Tensor decomposition with **PM-P** and **nSPM-P**. Shaded areas indicate plus/minus one standard deviation. See Section 6 for more details.

tensors) are depicted in Figures 2a and 2b. In Figure 2a, we use noiseless tensors with $\sigma = 0$, while in Figure 2b we set $D = 30$, $K = 300$ and σ ranges from 10^{-5} to 1.

Figure 2a illustrates that projected gradient ascent applied to **nSPM-P** always converges to the global maxima in the noiseless case (up to numerical error), provided the scaling is $K = \mathcal{O}(D^{\lfloor m/2 \rfloor}) = \mathcal{O}(D^2)$ and the constant adheres to the rank constraints pointed out in [22, Proposition 4.2]. This is in agreement with Theorem 15, which shows local maxima with sufficiently large objective have to coincide with global maximizers. In the noisy case, Figure 2b illustrates the robustness of **nSPM-P**, giving an error of $\mathcal{O}(\sigma)$. In contrast to **nSPM-P**, the **PM-P** objective suffers from a large bias and does not recover exact tensor components. The effect of the bias even dominates errors induced by moderately sized entrywise noise.

Deflation with SPM. In the second experiment we recover complete tensors by using the modified deflation procedure in Remark 18. We are mostly interested in the noisy case, since deflation with exact a_i's, as identified by **nSPM-P** in the noiseless case, does not induce an additional error [22].

We fix $D = 40, K = 300$, vary the noise level σ, and construct random tensors as in the previous experiment. Figure 2c plots the average error over 100 repetitions of successively recovered tensor components, where the x-axis ranges from the first recovered component at 1 to the last component at index 300. The figure illustrates that **nSPM-P** combined with modified deflation allows for recovering all tensor components up to an error of $\mathcal{O}(\sigma)$. Surprisingly, we do not observe error propagation, despite the fact that noisy recovered tensor components are being used within each deflation step.

7 Conclusion

We have presented a satisfying picture for the optimization landscape of a recent formulation of the non-convex, high-dimensional problem of symmetric tensor decomposition. In a quantitative analysis, we identified different regimes, “frame conditions+low-rank" and “random+overcomplete", in which second-order critical points with reasonable quality (sufficiently high function value) are stable to perturbations in the tensor entries and tightly approximate the desired tensor components. The essential novelty of the SPM functional is the “debiasing" role played by the inverse of a certain Grammian matrix recording correlations between rank-1 tensors. Our findings demonstrate a theoretical and numerical preferability for using the SPM functional over its more popular PM counterpart.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

[2] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1):61–88, 2011.

[3] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(80):2773–2832, 2014.

[4] A. Anandkumar, R. Ge, and M. Janzamin. Analyzing tensor power method dynamics in overcomplete regime. *The Journal of Machine Learning Research*, 18(1):752–791, 2017.

[5] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden markov models. In *Conference on Learning Theory*, pages 33–1. JMLR Workshop and Conference Proceedings, 2012.

[6] A. S. Bandeira, A. Perry, and A. S. Wein. Notes on computational-to-statistical gaps: predictions using statistical physics. *Portugaliae Mathematica*, 75(2):159–186, 2018.

[7] L. Chiantini and C. Ciliberto. Weakly defective varieties. *Transactions of the American Mathematical Society*, 354(1):151–178, 2002.

[8] O. Christensen et al. *An Introduction to Frames and Riesz Bases*, volume 7. Springer, 2003.

[9] M. T. Chu, R. E. Funderlic, and G. H. Golub. A rank–one reduction formula and its applications to matrix factorizations. *SIAM Review*, 37(4):512–530, 1995.

[10] P. Comon, G. Golub, L.-H. Lim, and B. Mourrain. Symmetric tensors and symmetric tensor rank. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1254–1279, 2008.

[11] L. De Lathauwer, J. Castaing, and J.-F. Cardoso. Fourth-order cumulant-based blind identification of underdetermined mixtures. *IEEE Transactions on Signal Processing*, 55(6):2965–2973, 2007.

[12] A. Fengler and P. Jung. On the restricted isometry property of centered self Khatri-Rao products. *arXiv preprint arXiv:1905.09245*, 2019.

[13] M. Fickus, D. G. Mixon, and J. C. Tremain. Steiner equiangular tight frames. *Linear Algebra and its Applications*, 436(5):1014–1027, 2012.

[14] C. Fiedler, M. Fornasier, T. Klock, and M. Rauchensteiner. Stable recovery of entangled weights: Towards robust identification of deep neural networks from minimal samples. *arXiv preprint arXiv:2101.07150*, 2021.

[15] M. Fornasier, T. Klock, and M. Rauchensteiner. Robust and resource efficient identification of two hidden layer neural networks. *Construction Approximation*, to appear, 2021.

[16] M. Fornasier, J. Vybíral, and I. Daubechies. Robust and resource efficient identification of shallow neural networks by fewest samples. *Information and Inference: A Journal of the IMA*, 2021.

[17] R. Ge, Q. Huang, and S. M. Kakade. Learning mixtures of gaussians in high dimensions. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, pages 761–770, 2015.

[18] R. Ge and T. Ma. On the optimization landscape of tensor decompositions. *Mathematical Programming*, pages 1–47, 2020.

[19] J. Harris. *Algebraic Geometry: A First Course*, volume 133. Springer Science & Business Media, 2013.

[20] C. J. Hillar and L.-H. Lim. Most tensor problems are NP-hard. *Journal of the ACM (JACM)*, 60(6):45:1–45:39, 2013.

[21] M. Janzamin, H. Sedghi, and A. Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

[22] J. Kileel and J. M. Pereira. Subspace power method for symmetric tensor decomposition and generalized PCA. *arXiv preprint arXiv:1912.04007*, 2019.

[23] T. G. Kolda. Numerical optimization for symmetric tensor decomposition. *Mathematical Programming*, 151(1):225–248, 2015.

[24] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[25] T. G. Kolda and J. R. Mayo. An adaptive shifted power method for computing generalized tensor eigenpairs. *SIAM Journal on Matrix Analysis and Applications*, 35(4):1563–1581, 2014.

[26] P. McCullagh. *Tensor Methods in Statistics*. Courier Dover Publications, 2018.

[27] S. Oymak and M. Soltanolkotabi. Learning a deep convolutional neural network via tensor decomposition. *Information and Inference: A Journal of the IMA*, 2021.

[28] M. Rudelson, R. Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[29] H. Sedghi, M. Janzamin, and A. Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics*, pages 1223–1231. PMLR, 2016.

[30] S. Sherman and T. G. Kolda. Estimating higher-order moments using symmetric tensor decomposition. *SIAM Journal on Matrix Analysis and Applications*, 41(3):1369–1387, 2020.

[31] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

[32] M. A. Sustik, J. A. Tropp, I. S. Dhillon, and R. W. Heath Jr. On the existence of equiangular tight frames. *Linear Algebra and its applications*, 426(2-3):619–635, 2007.

[33] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[34] N. Vervliet, O. Debals, L. Sorber, M. Van Barel, and L. De Lathauwer. Tensorlab 3.0. Available online. *URL: http://www.tensorlab.net*, 2016.

[35] S. R. White. Density matrix formulation for quantum renormalization groups. *Physical Review Letters*, 69(19):2863, 1992.

[36] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *International Conference on Machine Learning*, pages 4140–4149. PMLR, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] The main limitations of our work are an completely global understanding of **nSPM-P** in the overcomplete case and a vanishing error propagation after using deflation in random noise models. These points are discussed after Theorem 15, respectively, in Remark 18. Moreover, we state a conjecture with supplied numerical evidence in the supplementary for a missing piece in the theory for higher order tensors ($m > 4$), see Conjecture 14.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Due to space constraints, most proofs are included in the supplementary material. A proof sketch for the main theoretical results is given in Section 5.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We will upload the Matlab code for conducting our experiments.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We run 100 repititions of each conducted experiment and report mean, 1/4, and 3/4 quantiles.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] All experiments in the main submission can be conducted on a personal laptop with a Intel® Core™ i7-7700HQ CPU and 16.0GB of RAM in less than an hour.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We use the implementation of the SPM algorithm by Kileel and Pereira and properly cite their work.

(b) Did you mention the license of the assets? [Yes]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]