Guaranteed Noisy CP Tensor Recovery via Riemannian Optimization on the Segre Manifold

Ke Xu
Department of Applied and Computational Mathematics and Statistics
University of Notre Dame
Notre Dame, IN, USA
kxu6@nd.edu

Yuefeng Han
Department of Applied and Computational Mathematics and Statistics
University of Notre Dame
Notre Dame, IN, USA
yuefeng.han@nd.edu

Abstract

Recovering a low-CP-rank tensor from noisy linear measurements is a central challenge in high-dimensional data analysis, with applications spanning tensor PCA, scalar-on-tensor regression, tensor-on-tensor regression, and beyond. We exploit the intrinsic geometry of rank-one tensors by casting the recovery task as an optimization problem over the Segre manifold, the smooth Riemannian manifold of rank-one tensors. This geometric viewpoint yields two powerful algorithms: Riemannian Gradient Descent (RGD) and Riemannian Gauss-Newton (RGN), each of which preserves feasibility at every iteration. Under mild noise assumptions, we prove that RGD converges at a local linear rate, while RGN exhibits an initial local quadratic phase followed by linear refinement to the statistical error. Extensive synthetic experiments validate these convergence guarantees and demonstrate the practical effectiveness of our methods.

1 Introduction

Tensor decomposition, particularly the CP decomposition, has emerged as a powerful tool for analyzing high-dimensional data across diverse domains such as chemometrics, neuroscience, and recommendation systems [3, 17, 28, 49, 57], and tensor-valued time series in economics and finance [8, 9, 21, 22, 23]. Specifically, for an order-d tensor $\mathcal{T} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ with $d \geq 3$, the CP decomposition is a sum of rank-one tensors:

$$\mathcal{T} = \sum_{i=1}^{r} \lambda_i u_{1,i} \otimes u_{2,i} \otimes \cdots \otimes u_{d,i}, \tag{1}$$

where $\otimes$ denotes the tensor product, each factor $u_{k,i} \in \mathbb{R}^{p_k}$ satisfies $\|u_{k,i}\|_2 = 1$, r is the CP rank, and $\lambda_i > 0$ after absorbing each component sign into one factor. Under mild identifiability conditions (e.g., Kruskal's criterion [30]), this representation is essentially unique up to scaling and permutation, making it a widely adopted model in multi-way data analysis.

In practice, one often only observes noisy measurements of $\mathcal{T}$, for example

$$\mathcal{Y} = \mathcal{A}(\mathcal{T}) + \mathcal{E},$$

where $\mathcal{A}$ is a linear observation operator (possibly compressive) and $\mathcal{E}$ denotes additive noise.

In this work, we address the problem of recovering the underlying low-CP-rank tensor $\mathcal{T}$ from noisy measurements. In particular, we perform optimization directly on the *Segre manifold*, a smooth Riemannian manifold composed of rank-one tensors. Utilizing Riemannian optimization techniques ensures that the iterates remain on the manifold, thereby preserving the structure of the CP model and achieving improved convergence properties over traditional Euclidean approaches [5]. Our contributions can be summarized as follows:

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1. We develop **Riemannian Gradient Descent (RGD)** and **Riemannian Gauss-Newton (RGN)** algorithms for noisy low-CP-rank tensor recovery from general linear measurements, including tensor decomposition and tensor-on-tensor regression, by directly optimizing on the Segre manifold and using its established geometry [5, 47].
2. We derive **convergence guarantees** for both RGD and RGN in the noisy case and analyze how the CP geometry affects the convergence. In particular, for CP tensor-on-tensor regression under general linear measurements, we provide algorithm-level local convergence guarantees and nonasymptotic statistical-error guarantees.
3. Numerical experiments demonstrate that our algorithms yield robust factor recovery under noise and achieve competitive convergence and reconstruction accuracy relative to standard methods.

1.1 Related Work

Alternating Least Squares (ALS) [7, 11, 24, 28] is the classical workhorse for CP tensor decomposition, valued for its conceptual simplicity and low per-iteration cost. Despite its widespread use, explicit local convergence rates for ALS on non-orthogonal CP tensors were established only recently [26, 48]. Early theoretical analyses circumvented this difficulty by restricting attention to the orthogonal CP model [1, 52]. Other work has treated the non-orthogonal case under incoherence assumptions: Anandkumar et al. [2] analyzed alternating rank-one updates when the factors are random vectors on the sphere, while Sharan and Valiant [42] proposed orthogonalized ALS, which orthogonalizes the factor estimates before each sweep and provably recovers the factors from a random start.

Low-rank tensor regression offers another important motivation for the general linear measurement formulation. Classical formulations with tensor covariates and scalar responses impose low-rank structure on the coefficient tensor, often through CP-type parameterizations [58]. Tensor-on-tensor regression models predict a tensor-valued response from a tensor-valued covariate. Approaches include a penalized CP estimator for the contracted tensor product model [35], likelihood estimators for several low-rank formats [34], the latent-variable method HOPLS [56], and methods for sparse coefficients or outliers [32, 38]. Tucker-rank RGD and RGN have nonasymptotic recovery and convergence theory [37], and tensor-train IHT and RGD have RIP-based recovery guarantees [40]. Our framework departs from these approaches by casting CP decomposition and regression with a CP coefficient as instances of a common noisy linear inverse problem, optimized over products of Segre manifolds, with algorithm-level convergence to a nonasymptotic statistical error bound.

A complementary line of work leverages manifold optimization for tensor estimation, with notable success on low-rank matrix and Tucker decompositions [4, 36, 37]. Tensors of fixed Tucker rank form a smooth Riemannian manifold whose tangent space admits a simple parameterization [29], providing a natural and computationally tractable framework for optimization. Tangent-space projections on the Tucker manifold also underlie recent debiased inference for low-Tucker-rank regression and PCA [53]. By incorporating this geometry directly into the optimization process, such methods have yielded substantial improvements in tensor recovery, especially under noise.

Extending these ideas to the CP setting, however, presents distinct challenges: the CP model is inherently non-orthogonal, leading to slow convergence, spurious local minima, and higher computational cost. Our work addresses these challenges by explicitly leveraging the geometric structure of the rank-one tensor space. The key observation is that a rank-one tensor can be regarded as a Tucker rank-one tensor, sidestepping the non-orthogonality that complicates direct optimization on the CP manifold. Such greedy rank-one updates are a natural strategy for CP decomposition [55], and linear convergence rates have been established for incoherent CP tensors [2, 46]. Breiding and Vannieuwenhoven [5] use the Cartesian product of r Segre manifolds in a Riemannian Gauss-Newton trust-region method for CP approximation, and Swijsen et al. [47] derive geodesics on the Segre manifold. Building on this geometry and on the projection–retraction analysis of Luo and Zhang [36, 37], we develop algorithms that respect the geometry of the CP model and ensure feasibility at every iteration, and we analyze the coupling of the r components under noise.

1.2 Organization

The remainder of the paper is organized as follows. Section 2 introduces a unified linear observation framework and instantiates it on two core examples: CP tensor decomposition and tensor-on-tensor regression, with scalar-on-tensor regression recovered as the special case. Section 3 develops the proposed Riemannian optimization algorithms in full detail. Section 4 establishes the theoretical

analysis, including local convergence guarantees. Section 5 reports comprehensive numerical experiments, and Section 6 concludes with directions for future work. Computational complexity, additional numerical results, real data analysis, and proofs are deferred to the appendix.

Notation. $[n] = \{1, \dots, n\}$, $\langle \cdot, \cdot \rangle$ is the entrywise inner product, vec is vectorization, $\times_k$ is the mode-k product, $\preceq$ is the Loewner order, and $\widetilde{O}(\cdot)$ hides logarithmic factors. For a linear map $\mathcal{M}$ and a subspace $\mathbb{T}$, $\mathcal{M}^*$, $\|\mathcal{M}\|$ and $\mathcal{M}|_{\mathbb{T}}$ are the adjoint, operator norm and restriction, $\mathcal{P}_{\mathbb{T}}$ is the orthogonal projector onto $\mathbb{T}$, $\mathcal{P}_{\mathbb{T}}^{\perp} = I - \mathcal{P}_{\mathbb{T}}$, and I and Id are identities.

2 Model and Problem Formulation

Our goal is to recover the signal tensor $\mathcal{T}$ admitting the CP decomposition in (1), by solving an optimization problem that exploits the geometry of the Segre manifold. Specifically, we consider

$$\min_{\mathcal{T}_i \in \mathrm{Seg}, i \in [r]} \mathcal{L}(\{\mathcal{T}_i\}_{i=1}^r) = \min_{\mathcal{T}_i \in \mathrm{Seg}, i \in [r]} \frac{1}{2} \Big\| \mathcal{Y} - \sum_{i=1}^r \mathcal{A}(\mathcal{T}_i) \Big\|_{\mathrm{F}}^2, \tag{2}$$

where Seg denotes the Segre manifold of rank-one tensors (Definition 3.1), and $\mathcal{A} : \mathbb{R}^{p_1 \times \cdots \times p_d} \to \mathcal{H}$ is a (possibly random) linear operator from the ambient tensor space to a finite-dimensional Euclidean space $\mathcal{H}$, equipped with the Frobenius norm $\|\cdot\|_{\mathrm{F}}$. This formulation is general enough to accommodate both complete and compressive observations of the tensor. In particular, for tensor decomposition, $\mathcal{H} = \mathbb{R}^{p_1 \times \cdots \times p_d}$, whereas for tensor-on-tensor regression, $\mathcal{H} = (\mathbb{R}^{p_{d_{\mathrm{in}}+1} \times \cdots \times p_d})^n$, with $d_{\mathrm{out}} = d - d_{\mathrm{in}}$ and the scalar-response case corresponding to $d_{\mathrm{out}} = 0$.

Previous work has largely focused on estimating the tensor factors by iteratively optimizing over each tensor mode [7, 42]. In contrast, our formulation directly iterates on the Segre manifold, the smooth Riemannian manifold composed of rank-one tensors. This intrinsic approach leverages the rank-one structure of each component, preserving the CP structure throughout the optimization. This formulation accommodates a variety of applications, two of which we highlight below.

Tensor Decomposition. When the entire signal tensor $\mathcal{T}$ is observed, we set $\mathcal{A} = \mathrm{Id} : \mathbb{R}^{p_1 \times \cdots \times p_d} \to \mathbb{R}^{p_1 \times \cdots \times p_d}$. In this case, the problem in (2) becomes $\min_{(\mathcal{T}_1, \dots, \mathcal{T}_r) \in \mathrm{Seg}^r} \frac{1}{2} \|\mathcal{Y} - \sum_{i=1}^r \mathcal{T}_i\|_{\mathrm{F}}^2$, which is exactly the classical noisy CP decomposition.

Tensor-on-Tensor Regression. Let $d_{\mathrm{in}} \in \{1, \dots, d\}$ denote the number of input modes, and $d_{\mathrm{out}} = d - d_{\mathrm{in}}$ the number of output modes, so that the unknown coefficient tensor $\mathcal{T} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ partitions accordingly. For each $m = 1, \dots, n$, we observe a covariate tensor $\mathcal{X}_m \in \mathbb{R}^{p_1 \times \cdots \times p_{d_{\mathrm{in}}}}$, and a response tensor $\mathcal{Y}_m \in \mathbb{R}^{p_{d_{\mathrm{in}}+1} \times \cdots \times p_d}$ satisfying $\mathcal{Y}_m = \langle \mathcal{X}_m, \mathcal{T} \rangle_{\mathrm{in}} + \mathcal{E}_m$, where $\mathcal{E}_m$ is a noise tensor of the same dimensions as $\mathcal{Y}_m$, and $\langle \mathcal{X}_m, \mathcal{T} \rangle_{\mathrm{in}}$ denotes contraction of $\mathcal{T}$ with $\mathcal{X}_m$ along the first d_{in} modes, namely

$$\big(\langle \mathcal{X}_m, \mathcal{T} \rangle_{\mathrm{in}}\big)_{j_{d_{\mathrm{in}}+1}, \dots, j_d} = \sum_{j_1=1}^{p_1} \cdots \sum_{j_{d_{\mathrm{in}}}=1}^{p_{d_{\mathrm{in}}}} (\mathcal{X}_m)_{j_1, \dots, j_{d_{\mathrm{in}}}} \mathcal{T}_{j_1, \dots, j_{d_{\mathrm{in}}}, j_{d_{\mathrm{in}}+1}, \dots, j_d}.$$

To express this model in (2), we stack and scale the responses and errors as $\mathcal{Y} = n^{-1/2}(\mathcal{Y}_1, \dots, \mathcal{Y}_n)$ and $\mathcal{E} = n^{-1/2}(\mathcal{E}_1, \dots, \mathcal{E}_n)$, with $\mathcal{H}$ equipped with the usual product Frobenius inner product. Then $\mathcal{Y} = \mathcal{A}(\mathcal{T}) + \mathcal{E}$, where $\mathcal{A} : \mathbb{R}^{p_1 \times \cdots \times p_d} \to (\mathbb{R}^{p_{d_{\mathrm{in}}+1} \times \cdots \times p_d})^n$ is defined by $\mathcal{A}(\mathcal{T}) = \frac{1}{\sqrt{n}} (\langle \mathcal{X}_1, \mathcal{T} \rangle_{\mathrm{in}}, \dots, \langle \mathcal{X}_n, \mathcal{T} \rangle_{\mathrm{in}})$, whose adjoint is $\mathcal{A}^*(\mathcal{Z}_1, \dots, \mathcal{Z}_n) = \frac{1}{\sqrt{n}} \sum_{m=1}^n \mathcal{X}_m \otimes \mathcal{Z}_m$, yielding $\mathcal{A}^*\mathcal{A}(\mathcal{T}) = \frac{1}{n} \sum_{m=1}^n \mathcal{X}_m \otimes \big(\langle \mathcal{X}_m, \mathcal{T} \rangle_{\mathrm{in}}\big)$.

3 Method

In this section, we present two algorithms for noisy CP tensor recovery: Riemannian Gradient Descent (RGD) and Riemannian Gauss-Newton (RGN). Rather than optimizing in the full ambient space, both methods sequentially update the r rank-one tensor components on the Segre manifold.

3.1 Background: Geometry of the Segre Manifold

We begin by reviewing the geometric foundations underlying our Riemannian framework [5, 47]. A nonzero *rank-one tensor* in $\mathbb{R}^{p_1 \times p_2 \times \cdots \times p_d}$ takes the form $\mathcal{T} = u_1 \otimes u_2 \otimes \cdots \otimes u_d$, with $u_k \in \mathbb{R}^{p_k} \setminus \{0\}$ for $k = 1, \dots, d$. The set of projective classes of such tensors constitutes the *Segre variety* in algebraic geometry [31]. Equipping the set of nonzero rank-one tensors in the ambient space $\mathbb{R}^{p_1 \times \cdots \times p_d}$ with the Frobenius metric instead yields a smooth Riemannian submanifold, the *Segre manifold* (Seg), whose geometry is detailed in [27].

Definition 3.1 (Segre Manifold)**.** The **Segre manifold** is the set of all nonzero rank-one tensors in the ambient space $\mathbb{R}^{p_1 \times \cdots \times p_d}$,

$$\mathrm{Seg} \;=\; \Big\{ u_1 \otimes u_2 \otimes \cdots \otimes u_d : u_l \in \mathbb{R}^{p_l} \setminus \{0\},\ \forall l \in [d] \Big\}.$$

It is a smooth embedded submanifold of $\mathbb{R}^{\prod_{l=1}^{d} p_l} \setminus \{0\}$ of dimension $\dim(\mathrm{Seg}) = 1 + \sum_{l=1}^{d} (p_l - 1)$.

Remark 3.2. The Segre manifold admits the equivalent parameterization

$$\mathrm{Seg} \;\cong\; \big((0,\infty) \times \mathbb{S}^{p_1 - 1} \times \cdots \times \mathbb{S}^{p_d - 1}\big)/G,$$

where $\mathbb{S}^{p-1}$ is the unit sphere of $\mathbb{R}^p$ and $G = \{(\varepsilon_1, \ldots, \varepsilon_d) \in \{\pm 1\}^d : \prod_{k=1}^{d} \varepsilon_k = 1\}$ acts by simultaneous sign flips. The quotient absorbs the sign ambiguity arising from the fact that different sign patterns of the factor vectors can represent the same tensor.

We therefore optimize over r-tuples of rank-one tensors lying on the Segre manifold (Seg). To ensure these components remain distinguishable, we impose an incoherence condition among them, effectively acting as a soft-orthogonality constraint.

Assumption 3.3. For each mode $l \in [d]$, the factor vectors satisfy the incoherence condition

$$\mu_l = p_l \cdot \max_{i,j \in [r], i \neq j} |\langle u_{l,i}, u_{l,j} \rangle|^2,$$

and we denote the maximum incoherence by $\eta = \max_{l \in [d]} \sqrt{\mu_l / p_l} = \max_l \max_{i \neq j} |\langle u_{l,i}, u_{l,j} \rangle|$.

This incoherence assumption is standard in the CP tensor estimation literature [2, 6]. By Lemma 2 of Anandkumar et al. [2], if $\{u_{l,i}\}_{l \in [d], i \in [r]}$ are drawn i.i.d. from the unit sphere $\mathbb{S}^{p_l - 1}$, then $\max_{i \neq j} \{|\langle u_{l,i}, u_{l,j} \rangle|\} = \widetilde{O}(1/\sqrt{p_l})$ with high probability. Existing analyses typically require η to vanish with the dimension, e.g., $\eta = O(1/\sqrt{\max_{l \in [d]} p_l})$ [2, 6]. Our analysis, by contrast, only requires η to be bounded by a sufficiently small constant, independent of dimension.

Any CP tensor of rank r admits a Tucker representation with multilinear rank bounded componentwise by $(r, \ldots, r)$, and in the rank-one case, the two parameterizations coincide. Optimizing over the product of r rank-one manifolds, rather than over each of the d mode factors of a rank-r tensor, thus fully exploits the intrinsic rank-one structure and handles non-orthogonal factor interactions.

3.2 Riemannian Gradient Descent on the Segre Manifold

Since standard Euclidean gradient descent ignores the underlying manifold structure, we employ Riemannian gradient descent instead. For a current rank-one component $\widehat{\mathcal{T}}_i \in \mathrm{Seg}$ with tangent space $\widehat{\mathbb{T}}_i$, each component update projects the Euclidean gradient onto $\widehat{\mathbb{T}}_i$ and retracts the result back to the manifold,

$$\widehat{\mathcal{T}}_i^{+} = \mathcal{R}_{\widehat{\mathcal{T}}_i}\Big(- \alpha_t \, \mathcal{P}_{\widehat{\mathbb{T}}_i}\Big(\nabla_{\mathcal{T}_i} \mathcal{L}\big(\{\widehat{\mathcal{T}}_j\}_{j=1}^{r}\big)\Big)\Big),$$

where $\alpha_t > 0$ is the step size, $\nabla_{\mathcal{T}_i} \mathcal{L}$ is the partial Euclidean gradient, and $\widehat{\mathcal{T}}_i^{+}$ denotes the updated component.

Tangent Space of the Segre Manifold. The tangent space captures the manifold's local linear structure around a base point. For $\mathcal{T}_i = \lambda_i u_{1,i} \otimes \cdots \otimes u_{d,i}$ with $\lambda_i > 0$ and $\|u_{k,i}\|_2 = 1$, the tangent space $\mathbb{T}_i := T_{\mathcal{T}_i} \mathrm{Seg}$ consists of all first-order variations along the factor directions. Concretely, every $\xi_i \in \mathbb{T}_i$ decomposes as

$$\xi_i = \sum\nolimits_{k=1}^{d} u_{1,i} \otimes \cdots \otimes u_{k-1,i} \otimes h_{k,i} \otimes u_{k+1,i} \otimes \cdots \otimes u_{d,i},$$

where each $h_{k,i} \in \mathbb{R}^{p_k}$ is an infinitesimal perturbation of the k-th factor.

For each mode k, let $\mathcal{P}_{k,i} = u_{k,i} u_{k,i}^{\top}$ denote the orthogonal projector onto the span of $u_{k,i}$, and let $\mathcal{P}_{k,i}^{\perp} = I_{p_k} - u_{k,i} u_{k,i}^{\top}$ denote its complement. Minimizing $\|\widetilde{\mathcal{T}} - \xi_i\|_{\mathrm{F}}^2$ subject to $\xi_i \in \mathbb{T}_i$ yields the projection of an arbitrary tensor $\widetilde{\mathcal{T}}$ onto the tangent space at $\mathcal{T}_i$,

$$\xi_i = \mathcal{P}_{\mathbb{T}_i}(\widetilde{\mathcal{T}}) = \sum\nolimits_{k=1}^{d} \widetilde{\mathcal{T}} \times_k \mathcal{P}_{k,i}^{\perp} \times_{l \neq k} \mathcal{P}_{l,i} + \widetilde{\mathcal{T}} \times_{l \in [d]} \mathcal{P}_{l,i}. \tag{3}$$

Retraction. A tangent-space update generally lands off the manifold, so we apply a retraction to map it back onto the Segre manifold. Common choices include the truncated higher-order singular

value decomposition (T-HOSVD) [13] and its sequential variant (ST-HOSVD) [5, 50], and recent work derives explicit geodesics and the exponential map on the Segre manifold [27, 47]. We refer the reader to Boumal [4] for a comprehensive treatment of these operators. We adopt T-HOSVD: for a tangent increment ξ_i at $\mathcal{T}_i$, $\mathcal{R}_{\mathcal{T}_i}(\xi_i)$ is the rank-$(1,\dots,1)$ T-HOSVD of $\mathcal{T}_i + \xi_i$ [29], and by Lemma E.1 it is a smooth retraction whenever $\|\xi_i\|_{\mathrm{F}} < \|\mathcal{T}_i\|_{\mathrm{F}}/[2(\sqrt{d}+1)]$.

3.3 Riemannian Gauss-Newton on the Segre Manifold

Riemannian gradient descent is conceptually simple but can converge slowly, particularly for large-scale problems or when high accuracy is required. Incorporating second-order information offers a powerful remedy. The Riemannian Gauss-Newton method [37], tailored to nonlinear least-squares, provides an efficient approximation to the full Riemannian Newton step.

At each component update, RGN seeks a tangent-space direction $\Delta_i \in \widehat{\mathbb{T}}_i$ satisfying the Gauss-Newton equation

$$\left(\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}\Big|_{\widehat{\mathbb{T}}_i}\right)\Delta_i = \mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\left\{\mathcal{A}\left(\sum_{j=1}^{r}\widehat{\mathcal{T}}_j\right) - \mathcal{Y}\right\}.$$

which has a unique solution whenever the restricted normal operator on the left is invertible. Replacing the Riemannian Hessian by the normal operator $\mathcal{A}^*\mathcal{A}$ restricted to the current tangent space captures the essential curvature at low cost. To maintain feasibility, RGN then retracts $-\Delta_i$ back onto the Segre manifold. This approach still amounts to a least-squares problem, but in a drastically lower-dimensional space: the tangent space has $1 + \sum_{l\in[d]}(p_l - 1)$ degrees of freedom, compared to $\prod_{l\in[d]} p_l$ parameters in the ambient tensor space $\mathbb{R}^{p_1\times p_2\times\cdots\times p_d}$.

Algorithm 1 Riemannian Gradient Descent for CP Tensor Estimation

Input: Observation $\mathcal{Y} \in \mathcal{H}$, linear operator $\mathcal{A} : \mathbb{R}^{p_1\times\cdots\times p_d} \to \mathcal{H}$, target CP rank r, step size α_t, and initial rank-one tensor estimates $\{\mathcal{T}_i^{(0)}\}_{i=1}^r$.

1: Compute the residual $\mathrm{res} \leftarrow \sum_{j=1}^r \mathcal{A}(\mathcal{T}_j^{(0)}) - \mathcal{Y}$.
2: **for** $t = 0, 1, \dots, t_{\max} - 1$ **do**
3: **for** $i = 1, \dots, r$ **do**
4: **(RGD Update)** With $\mathbb{T}_i^{(t)}$ the tangent space at $\mathcal{T}_i^{(t)}$, update

$$\mathcal{T}_i^{(t+1)} = \mathcal{R}_{\mathcal{T}_i^{(t)}}\left(-\alpha_t\,\mathcal{P}_{\mathbb{T}_i^{(t)}}\mathcal{A}^*(\mathrm{res})\right),$$

5: Update $\mathrm{res} \leftarrow \mathrm{res} + \mathcal{A}(\mathcal{T}_i^{(t+1)} - \mathcal{T}_i^{(t)})$.
6: **end for**
7: **end for**

Output: $\{\mathcal{T}_i^{(t_{\max})}\}_{i=1}^r$.

4 Theoretical Analysis

4.1 Convergence Analysis of Riemannian Optimization

In this subsection, we establish a deterministic convergence theory for both RGD and RGN, stated in Theorems 4.1 and 4.2 for one component update. Even in the presence of noise, these bounds control the distance between each updated component and its true rank-one component by exploiting the intrinsic geometry of the Segre manifold. Appendix G composes them over each sequential pass, and the resulting local convergence guarantees are stated in Section 4.2. Throughout this section, $d \geq 3$ and $\lambda_1 \geq \cdots \geq \lambda_r > 0$.

For a current collection $\{\widehat{\mathcal{T}}_j\}_{j=1}^r$, let $\widehat{\mathbb{T}}_i$ be the tangent space at $\widehat{\mathcal{T}}_i$. Whenever the restricted normal operator $\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}\big|_{\widehat{\mathbb{T}}_i}$ is invertible, define $\widehat{\mathsf{C}}_i := (\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}\big|_{\widehat{\mathbb{T}}_i})^{-1}\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i^\perp}$. This operator captures the *coupling* between the normal space $\widehat{\mathbb{T}}_i^\perp$ and the tangent space $\widehat{\mathbb{T}}_i$ induced by the normal operator $\mathcal{A}^*\mathcal{A}$: it maps an off-tangent perturbation into $\widehat{\mathbb{T}}_i$, normalized by the local action of $\mathcal{A}^*\mathcal{A}$ restricted to the tangent space. To quantify the worst-case strength of this coupling over all rank-one directions (the Segre manifold), we introduce the Segre-restricted operator norm $\|\mathsf{C}\|_{\mathrm{Seg}\to\mathrm{F}} := \sup_{V\in\mathrm{Seg},\|V\|_{\mathrm{F}}=1}\ \|\mathsf{C}V\|_{\mathrm{F}}$.

Algorithm 2 Riemannian Gauss-Newton for CP Tensor Estimation

Input: Observation $\mathcal{Y} \in \mathcal{H}$, linear operator $\mathcal{A} : \mathbb{R}^{p_1 \times \cdots \times p_d} \to \mathcal{H}$, target CP rank r, and initial rank-one tensor estimates $\{\mathcal{T}_i^{(0)}\}_{i=1}^r$.

Compute the residual $\mathrm{res} \leftarrow \sum_{j=1}^r \mathcal{A}(\mathcal{T}_j^{(0)}) - \mathcal{Y}$.
for $t = 0, 1, \ldots, t_{\max} - 1$ **do**
  for $i = 1, \ldots, r$ **do**
    (RGN Update) Solve for $\Delta_i^{(t)} \in \mathbb{T}_i^{(t)}$:

$$\left(\mathcal{P}_{\mathbb{T}_i^{(t)}} \mathcal{A}^* \mathcal{A} \mathcal{P}_{\mathbb{T}_i^{(t)}} \Big|_{\mathbb{T}_i^{(t)}} \right) \Delta_i^{(t)} = \mathcal{P}_{\mathbb{T}_i^{(t)}} \mathcal{A}^*(\mathrm{res}),$$

    (Retraction) $\mathcal{T}_i^{(t+1)} \leftarrow \mathcal{R}_{\mathcal{T}_i^{(t)}}(-\Delta_i^{(t)})$.
    Update $\mathrm{res} \leftarrow \mathrm{res} + \mathcal{A}(\mathcal{T}_i^{(t+1)} - \mathcal{T}_i^{(t)})$.
  end for
end for

Output: $\{\mathcal{T}_i^{(t_{\max})}\}_{i=1}^r$.

For either method, let $\widehat{\xi}_i \in \widehat{\mathbb{T}}_i$ denote its tangent increment, namely $\widehat{\xi}_i = -\alpha_t \mathcal{P}_{\widehat{\mathbb{T}}_i} \mathcal{A}^* \{\mathcal{A}(\sum_{j=1}^r \widehat{\mathcal{T}}_j) - \mathcal{Y}\}$ for RGD and $\widehat{\xi}_i = -\widehat{\Delta}_i$ for RGN, where $\widehat{\Delta}_i$ is the RGN normal-equation solution.

Theorem 4.1 (Local Convergence of RGD)**.** *Consider one component update with step size $\alpha_t > 0$ from the current collection $\{\widehat{\mathcal{T}}_j\}_{j=1}^r$, let $\widehat{\mathcal{T}}_i^+$ denote the updated i-th component, and let $\varepsilon = \max_{j \in [r]} \|\widehat{\mathcal{T}}_j - \mathcal{T}_j\|_{\mathrm{F}} / \lambda_j$ denote the largest relative Frobenius error of the current components, where λ_j is the weight of the j-th CP component. Let $\kappa = \lambda_1/\lambda_r$ be the condition number and η the incoherence parameter from Assumption 3.3. Suppose that $\varepsilon \le 1/(4d)$ and $\max_{i \in [r]} \|\widehat{\xi}_i\|_{\mathrm{F}} / \|\widehat{\mathcal{T}}_i\|_{\mathrm{F}} < 1/\{2(\sqrt{d}+1)\}$. Then, for every $i \in [r]$,*

$$\begin{aligned}
\frac{\|\widehat{\mathcal{T}}_i^+ - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i} \le{}& \underbrace{C_1 \max_{k \in [r]} \left\| \mathcal{P}_{\widehat{\mathbb{T}}_k} (I - \alpha_t \mathcal{A}^* \mathcal{A}) \mathcal{P}_{\widehat{\mathbb{T}}_k} \right\|}_{\text{first-order contraction}} \varepsilon + \underbrace{C_1 r \alpha_t \kappa \max_{\substack{k,j \in [r] \\ k \neq j}} \left\| \mathcal{P}_{\widehat{\mathbb{T}}_k} \mathcal{A}^* \mathcal{A} \mathcal{P}_{\widehat{\mathbb{T}}_k}^{\perp} \mathcal{P}_{\widehat{\mathbb{T}}_j} \right\|}_{\text{first-order contraction}} \varepsilon \\
&+ \underbrace{C_2 \Big[1 + r \alpha_t \kappa \max_{k \in [r]} \sup_{\substack{V \in \mathrm{Seg} \\ \|V\|_{\mathrm{F}} = 1}} \left\| \mathcal{P}_{\widehat{\mathbb{T}}_k} \mathcal{A}^* \mathcal{A} \mathcal{P}_{\widehat{\mathbb{T}}_k}^{\perp} V \right\|_{\mathrm{F}} \Big] \varepsilon^2}_{\text{second-order contraction}} + \underbrace{\frac{\|\widehat{\xi}_i\|_{\mathrm{F}}^2}{\lambda_i \|\widehat{\mathcal{T}}_i\|_{\mathrm{F}}}}_{\text{retraction remainder}} \\
&+ \underbrace{C_2 r \alpha_t \kappa \max_{k \in [r]} \left\| \mathcal{P}_{\widehat{\mathbb{T}}_k} \mathcal{A}^* \mathcal{A} \mathcal{P}_{\widehat{\mathbb{T}}_k} \right\| \big[\eta^{d-2} + \varepsilon \big] \varepsilon}_{\text{cross-component interaction}} + \underbrace{C_1 \alpha_t \max_{k \in [r]} \frac{\left\| \mathcal{P}_{\widehat{\mathbb{T}}_k} (\mathcal{A}^* \mathcal{E}) \right\|_{\mathrm{F}}}{\lambda_k}}_{\text{noise term}},
\end{aligned}$$

where $C_1 = 1$ and $C_2 = 4d$.

Theorem 4.2 (Local Convergence of RGN)**.** *With the notation of Theorem 4.1, suppose that $\mathcal{P}_{\widehat{\mathbb{T}}_k} \mathcal{A}^* \mathcal{A} \mathcal{P}_{\widehat{\mathbb{T}}_k} \big|_{\widehat{\mathbb{T}}_k}$ is invertible for every $k \in [r]$ and that the conditions on ε and $\widehat{\xi}_i$ in Theorem 4.1 hold. Then the RGN update satisfies, for every $i \in [r]$,*

$$\begin{aligned}
\frac{\|\widehat{\mathcal{T}}_i^+ - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i} \le{}& \underbrace{C_1 r \kappa \max_{k \neq j, k, j \in [r]} \left\| \widehat{\mathsf{C}}_k \mathcal{P}_{\widehat{\mathbb{T}}_j} \right\| \varepsilon}_{\text{first-order contraction}} + \underbrace{C_1 \max_{k \in [r]} \frac{\left\| \left(\mathcal{P}_{\widehat{\mathbb{T}}_k} \mathcal{A}^* \mathcal{A} \mathcal{P}_{\widehat{\mathbb{T}}_k} \big|_{\widehat{\mathbb{T}}_k} \right)^{-1} \mathcal{P}_{\widehat{\mathbb{T}}_k} \mathcal{A}^*(\mathcal{E}) \right\|_{\mathrm{F}}}{\lambda_k}}_{\text{noise term}} \\
&+ \underbrace{C_2 r \kappa \varepsilon \left[\eta^{d-2} + \varepsilon \right]}_{\text{cross-component interaction}} + \underbrace{C_2 \left[1 + \{1 + r\kappa\} \max_{k \in [r]} \|\widehat{\mathsf{C}}_k\|_{\mathrm{Seg} \to \mathrm{F}} \right] \varepsilon^2}_{\text{second-order contraction}} + \underbrace{\frac{\|\widehat{\xi}_i\|_{\mathrm{F}}^2}{\lambda_i \|\widehat{\mathcal{T}}_i\|_{\mathrm{F}}}}_{\text{retraction remainder}},
\end{aligned}$$

where $C_1 = 1$ and $C_2 = 4d$.

Remark 4.3. Sign alignment follows from the local error bound. If $\|\widehat{\mathcal{T}}_i - \mathcal{T}_i\|_{\mathrm{F}} \le \varepsilon\lambda_i$ for some $\varepsilon < 1$, then $\|\widehat{\mathcal{T}}_i\|_{\mathrm{F}} \ge (1-\varepsilon)\lambda_i$ and the polarization identity gives

$$2\langle\widehat{\mathcal{T}}_i, \mathcal{T}_i\rangle = \|\widehat{\mathcal{T}}_i\|_{\mathrm{F}}^2 + \lambda_i^2 - \|\widehat{\mathcal{T}}_i - \mathcal{T}_i\|_{\mathrm{F}}^2 \ge 2(1-\varepsilon)\lambda_i^2 > 0.$$

Thus the local condition $\varepsilon \le 1/(4d)$ in Theorems 4.1–4.2 implies sign alignment at every component update. By the preceding display, an even number of factor inner products are negative. Flipping their signs is an element of G in Remark 3.2 and leaves $\widehat{\mathcal{T}}_i$ unchanged, so the factorwise alignment in the proofs is a normalization. The initialization condition of Corollaries 4.5–4.9 does not depend on $p_1, \dots, p_d$, and Appendix E.4 gives conditions under which CPCA satisfies it in CP decomposition.

While both RGD and RGN have a first-order error term proportional to ε, RGN incorporates curvature information and attains a quadratic local phase. The scale at which its quadratic phase ends is governed by η^{d-2} and $\|\widehat{\mathsf{C}}_i\mathcal{P}_{\widehat{\mathbb{T}}_j}\|$, which enter its first-order terms, while $\|\widehat{\mathsf{C}}_i\|_{\mathrm{Seg}\to\mathrm{F}}$ enters its second-order term. In CP decomposition, these coupling norms vanish exactly, so RGN has a quadratic phase above the η^{d-2} scale followed by linear refinement, and it is purely quadratic when $\eta = 0$. In CP tensor-on-tensor regression, they are controlled by the random design: under the Gaussian design of Section 4.2, $\mathcal{A}^*\mathcal{A}$ acts approximately as σ_X^2 times the identity on the relevant low-dimensional tangent and normal subspaces, so the tangent–normal cross interactions remain small with high probability, giving the desired contraction.

Remark 4.4. Recent work has substantially advanced our theoretical understanding of ALS, which remains the standard workhorse for CP decomposition [28]. Hu et al. [26] established a polynomial convergence rate (of order $d-1$ for order d tensors) for standard CP-ALS on orthogonally decomposable tensors, together with linear convergence for incoherently decomposable tensors. Tang et al. [48] derived statistical error bounds for ALS and identified a two-phase trajectory, an initial quadratic regime followed by a linear regime, that mirrors the local behavior of our RGN method. Beyond CP decomposition, existing results for CP regression concern global or penalized estimators [35, 59] or objective-preserving sketches [25], and the convergence of concrete iterates to the statistical error under a general linear operator has remained largely unexplored. By optimizing directly over the Segre manifold, our methods provide such guarantees in a unified geometric framework that accommodates different tasks through the choice of the forward operator.

4.2 Implications in Statistics and Machine Learning

We now specialize the convergence theory to two concrete problems: CP tensor decomposition and CP tensor-on-tensor regression, after matching component labels at initialization. More general versions of the RGD corollaries below, with variable step sizes, are given in Appendix G, and the computational complexity analysis is provided in Appendix B.

Throughout this paper, we set $p^* = \prod_{l=1}^{d} p_l, p_{\mathrm{in}}^* = \prod_{\ell=1}^{d_{\mathrm{in}}} p_\ell, p_{\mathrm{out}}^* = \prod_{\ell=d_{\mathrm{in}}+1}^{d} p_\ell$ with the convention $p_{\mathrm{out}}^* = 1$ when $d_{\mathrm{out}} = 0$. The tangent dimensions of a rank-one CP component are $\mathrm{df} = 1 + \sum_{\ell=1}^{d}(p_\ell - 1)$ and $\mathrm{df}_{\mathrm{in}} = 1 + \sum_{\ell=1}^{d_{\mathrm{in}}}(p_\ell - 1)$. We write $\kappa = \lambda_1/\lambda_r$ for the condition number and, for iterates $\{\mathcal{T}_i^{(t)}\}_{i=1}^r$, $\varepsilon^{(t)} = \max_{i\in[r]} \|\mathcal{T}_i^{(t)} - \mathcal{T}_i\|_{\mathrm{F}}/\lambda_i$.

Tensor Decomposition. Consider the noisy CP decomposition model

$$\mathcal{Y} = \mathcal{T} + \mathcal{E} \ \in \ \mathbb{R}^{p_1\times\cdots\times p_d},$$

where $\mathcal{T} = \sum_{i=1}^{r} \lambda_i u_{1,i} \otimes \cdots \otimes u_{d,i}$ is as defined in (1) and $\mathrm{vec}(\mathcal{E}) \sim \mathcal{N}\big(0, \Sigma_{p^*}\big)$, with noise covariance $\underline{\sigma}^2 I_{p^*} \preceq \Sigma_{p^*} \preceq \overline{\sigma}^2 I_{p^*}$. Under the incoherence assumption and a sufficiently accurate initialization, we obtain the following convergence guarantees.

Corollary 4.5 (Convergence of RGD for CP tensor decomposition)**.** *There exist universal constants $c, C > 0$ such that the following holds. Suppose that $\varepsilon^{(0)} \le c/\{d(1+\kappa r)\}$ and $\overline{\sigma}\sqrt{\mathrm{df}}/\lambda_r \le c/\{d(1+\kappa r)\}$, and that, for every $t \ge 0$, $0 < \alpha_t \le 1$, $|1-\alpha_t| \le c$ and $\alpha_t\eta^{d-2} \le c/(d\kappa r)$. Then, with probability at least $1 - 2\exp(-c\,\mathrm{df})$, simultaneously for all $t \ge 0$,*

$$\varepsilon^{(t)} \le 2^{-t}\varepsilon^{(0)} + C\frac{\overline{\sigma}\sqrt{\mathrm{df}}}{\lambda_r}.$$

Corollary 4.6 (Convergence of RGN for CP tensor decomposition)**.** *Under the assumptions of Corollary 4.5 with $\alpha_t = 1$, the RGN iterates satisfy the same linear bound. If, in addition,*

$\overline{\sigma}\sqrt{\mathrm{df}}/\lambda_r \le cd(1+\kappa r)\eta^{2(d-2)}$, *then, with probability at least* $1-2\exp(-c\,\mathrm{df})$, *RGN has a quadratic-to-linear local trajectory. More precisely, for universal constants* $C_3, C_4 > 0$ *and* $t^* = \inf\{t \ge 0 : \varepsilon^{(t)} \le C_4\eta^{d-2}\}$,

$$\varepsilon^{(t)} \le \begin{cases} \dfrac{\{C_3 d(1+\kappa r)\varepsilon^{(0)}\}^{2^t}}{C_3 d(1+\kappa r)}, & 0 \le t < t^*, \\ 2^{-(t-t^*)}\varepsilon^{(t^*)} + C\dfrac{\overline{\sigma}\sqrt{\mathrm{df}}}{\lambda_r}, & t \ge t^*. \end{cases}$$

Theorems 4.1–4.2 are deterministic. The corollaries impose the stated noise and sample-size conditions to keep their recursions inside the local basin.

Tensor-on-Tensor Regression. For $d_{\mathrm{in}} \in \{1,\dots,d\}$ and $d_{\mathrm{out}} = d - d_{\mathrm{in}}$, we observe

$$\mathcal{Y}_m = \langle \mathcal{X}_m, \mathcal{T}\rangle_{\mathrm{in}} + \mathcal{E}_m, \qquad m = 1,\dots,n,$$

where $\mathcal{X}_m \in \mathbb{R}^{p_1\times\cdots\times p_{d_{\mathrm{in}}}}, \mathcal{Y}_m, \mathcal{E}_m \in \mathbb{R}^{p_{d_{\mathrm{in}}+1}\times\cdots\times p_d}$, with $\langle \mathcal{X}_m, \mathcal{T}\rangle_{\mathrm{in}}$ denoting contraction over the first d_{in} modes of $\mathcal{T}$. We assume that the design tensors $\{\mathcal{X}_m\}_{m=1}^n$ are i.i.d. centered Gaussian with $\mathrm{Cov}(\mathrm{vec}(\mathcal{X}_m)) = \sigma_X^2 I_{p_{\mathrm{in}}^*}$, where $\sigma_X > 0$, and that the noise tensors are mutually independent centered Gaussian, independent of the designs, with covariance satisfying $\underline{\sigma}_\xi^2 I_{p_{\mathrm{out}}^*} \preceq \mathrm{Cov}\{\mathrm{vec}(\mathcal{E}_m)\} \preceq \overline{\sigma}_\xi^2 I_{p_{\mathrm{out}}^*}$. A sub-Gaussian extension requires the corresponding local matrix-deviation and Hanson–Wright bounds. A tensor restricted isometry property [37, Definition 1 and Proposition 1] provides a standard sufficient framework. In the Gaussian corollaries below, $\gamma = C\sqrt{(\mathrm{df}_{\mathrm{in}} + \log r)/n}$ is the restricted-design scale. The projected-noise term depends instead on the full tangent dimension df.

Remark 4.7. Under the Gaussian design, the isotropy assumption $\mathrm{Cov}\{\mathrm{vec}(\mathcal{X}_m)\} = \sigma_X^2 I_{p_{\mathrm{in}}^*}$ is equivalent to requiring the entries of each $\mathcal{X}_m$ to be i.i.d. with variance σ_X^2. For a general covariance $\Sigma_X = \mathrm{Cov}\{\mathrm{vec}(\mathcal{X}_m)\}$, the operators $\mathcal{A}, \mathcal{A}^*$ remain as defined in Section 2. Extending the statistical guarantees requires corresponding restricted conditioning bounds for $\mathcal{A}^*\mathcal{A}$. Estimating or preconditioning Σ_X requires additional structural assumptions and is beyond the present scope.

Corollary 4.8 (Convergence of RGD for CP tensor-on-tensor regression)**.** *There exist universal constants* $c, C > 0$ *such that the following holds. Suppose that* $n \ge C\,\mathrm{df}$, $\gamma \le c/(1+\kappa r)$, $\varepsilon^{(0)} \le c/\{d(1+\kappa r)\}$ *and* $\{\overline{\sigma}_\xi/(\sigma_X\lambda_r)\}\sqrt{\mathrm{df}/n} \le c/\{d(1+\kappa r)\}$, *and that, for every* $t \ge 0$, $0 < \alpha_t\sigma_X^2 \le 1$, $|1-\alpha_t\sigma_X^2| \le c$ *and* $\alpha_t\sigma_X^2\eta^{d-2} \le c/(d\kappa r)$. *Then, with probability at least* $1 - C\exp\{-c(\mathrm{df}_{\mathrm{in}} + \log r)\}$, *simultaneously for all* $t \ge 0$,

$$\varepsilon^{(t)} \le 2^{-t}\varepsilon^{(0)} + C\frac{\overline{\sigma}_\xi}{\sigma_X\lambda_r}\sqrt{\frac{\mathrm{df}}{n}}.$$

Corollary 4.9 (Convergence of RGN for CP tensor-on-tensor regression)**.** *Under the assumptions of Corollary 4.8 with* $\alpha_t\sigma_X^2 = 1$, *the RGN iterates satisfy the same linear bound. Let* $h = \gamma/d + \eta^{d-2}$. *If, in addition,* $\{\overline{\sigma}_\xi/(\sigma_X\lambda_r)\}\sqrt{\mathrm{df}/n} \le cd(1+\kappa r)h^2$, *then, with probability at least* $1 - C\exp\{-c(\mathrm{df}_{\mathrm{in}} + \log r)\}$, *for universal constants* $C_3, C_4 > 0$ *and* $t^* = \inf\{t \ge 0 : \varepsilon^{(t)} \le C_4 h\}$,

$$\varepsilon^{(t)} \le \begin{cases} \dfrac{\{C_3 d(1+\kappa r)\varepsilon^{(0)}\}^{2^t}}{C_3 d(1+\kappa r)}, & 0 \le t < t^*, \\ 2^{-(t-t^*)}\varepsilon^{(t^*)} + C\dfrac{\overline{\sigma}_\xi}{\sigma_X\lambda_r}\sqrt{\dfrac{\mathrm{df}}{n}}, & t \ge t^*. \end{cases}$$

Remark 4.10. The two-phase behavior in Corollaries 4.6 and 4.9 is driven by the normalized error recursion $\varepsilon^{(t+1)} \le Cd(1+\kappa r)\{\varepsilon^{(t)}\}^2 + C\kappa r\{\gamma + d\eta^{d-2}\}\varepsilon^{(t)}$ plus the projected-noise term, where $\gamma = 0$ for decomposition. While $\varepsilon^{(t)}$ remains above the transition scale, of order η^{d-2} for decomposition and $\gamma/d + \eta^{d-2}$ for regression, the quadratic term dominates, yielding local quadratic convergence. Below this scale, the linear term dominates, and the error contracts linearly until it settles at the statistical error. The additional noise conditions in Corollaries 4.6 and 4.9 ensure that the quadratic phase reaches the transition scale. Without them, the same recursion holds but the quadratic phase can end at the statistical error level.

5 Experiments and Results

Simulation. We evaluate the convergence behavior of RGD and RGN under varying incoherence and noise levels in CP decomposition of fourth-order tensors of size $60 \times 30 \times 30 \times 30$ and tensor-on-tensor regression with covariates $\mathcal{X}_m \in \mathbb{R}^{10\times10\times10}$, responses $\mathcal{Y}_m \in \mathbb{R}^{10\times10}$ and $n = 1000$. In both problems $r = 3$, the weights have condition number $\kappa = 30$, and the factors in each mode have maximal pairwise coherence $\rho \in \{0, 0.25, 0.5, 0.75\}$. The noise level is set by the signal-to-noise ratio $s \in \{\infty, 30, 10, 3\}$ of the weakest component. CP-ALS, RGD and RGN share the CPCA initialization, and all results are averaged over 20 replications. For regression we compare with CP-ALS and RRR [35] in the CP format, and with Tucker-RGD [37], Tucker-RGN [36], TT-IHT [40] and HOPLS [56] in other formats. For decomposition we compare with CP-ALS [28], Orth-ALS [42], ICO [20], NLS-GN [45], GN-CG [44], CP-ALS with enhanced line search (CP-ALS+ELS) [41] and Hybrid AMDM [43]. Full experimental details, including the coherence construction and the initialization, are given in Appendix C.

Figure 1 shows that, with noise, CP-ALS, RGD and RGN reach the same error level, whereas RRR, started at random, often does not within 20 iterations. When $\rho \leq 0.5$ and $s \geq 10$, RGN needs no more iterations than CP-ALS and RGD, and at $\rho = 0.75$ the three methods converge linearly, as described in Remark 4.10. The methods in other tensor formats reach higher error levels with noise, and Tucker-RGN converges fastest without noise at $\rho = 0.75$. In CP decomposition with noise, RGN reaches the same error level as CP-ALS (Figure 3), whereas Orth-ALS and ICO do not when $\rho > 0$ and $s \geq 10$. ICO and Hybrid AMDM need the fewest iterations without noise, and NLS-GN and GN-CG need the fewest when $s \geq 10$. RGD and RGN have the lowest cost per iteration in both problems (Table 3).

Real data. Table 1 reports five-fold cross-validation results for a tensor-on-tensor regression of 2019 export outcomes on 2018 trade-cost and market covariates of 80 exporters (Appendix D). All methods agree within one standard error. RGD and TT-IHT have not converged after 30 iterations, and their slightly lower test errors disappear at convergence (Table 10).

6 Discussion and Future Extensions

In this paper, we have proposed a unified framework with local convergence guarantees for noisy low-CP-rank tensor recovery from general linear measurements, encompassing both tensor decomposition and tensor-on-tensor regression. To the best of our knowledge, this is the first framework with algorithm-level guarantees for CP tensor-on-tensor regression. The framework offers several practical advantages: it seamlessly handles a broad class of linear measurement operators, and because each component update acts on a single Segre factor, the algorithms admit streaming implementations well-suited to large-scale or time-evolving tensor data. While we use fixed step sizes here, adaptive schemes or Riemannian momentum could further accelerate convergence.

Several extensions remain open. First, our analysis assumes that the CP rank is known exactly, and addressing rank selection and misspecification is an important next step. Second, each iteration involves retractions and tangent-space projections, which may become costly in ultra-high dimensions, and efficient approximations or randomized updates are a promising direction.

Table 1: Five-fold international-trade regression after 30 iterations: mean $\pm$ SE of the held-out test RMSE and test R^2 over the five folds. Bold: lowest mean test RMSE in each group.

Method	Model	Test RMSE	Test R^2
CP format			
CP-ALS	CP rank $= 3$	2.6747 ± 0.0286	0.1313 ± 0.0390
RRR	CP rank $= 3$	2.6763 ± 0.0279	0.1303 ± 0.0387
RGD	CP rank $= 3$	$\mathbf{2.6657 \pm 0.0278}$	0.1372 ± 0.0379
RGN	CP rank $= 3$	2.6743 ± 0.0286	0.1316 ± 0.0391
Other tensor formats			
Tucker-RGD	Tucker rank $= (3, 3, 3, 3)$	2.6748 ± 0.0269	0.1314 ± 0.0376
Tucker-RGN	Tucker rank $= (3, 3, 3, 3)$	2.6746 ± 0.0276	0.1315 ± 0.0379
TT-IHT	TT rank $= (3, 3, 3)$	$\mathbf{2.6663 \pm 0.0292}$	0.1367 ± 0.0388
HOPLS	3 latent blocks	2.6795 ± 0.0391	0.1297 ± 0.0303

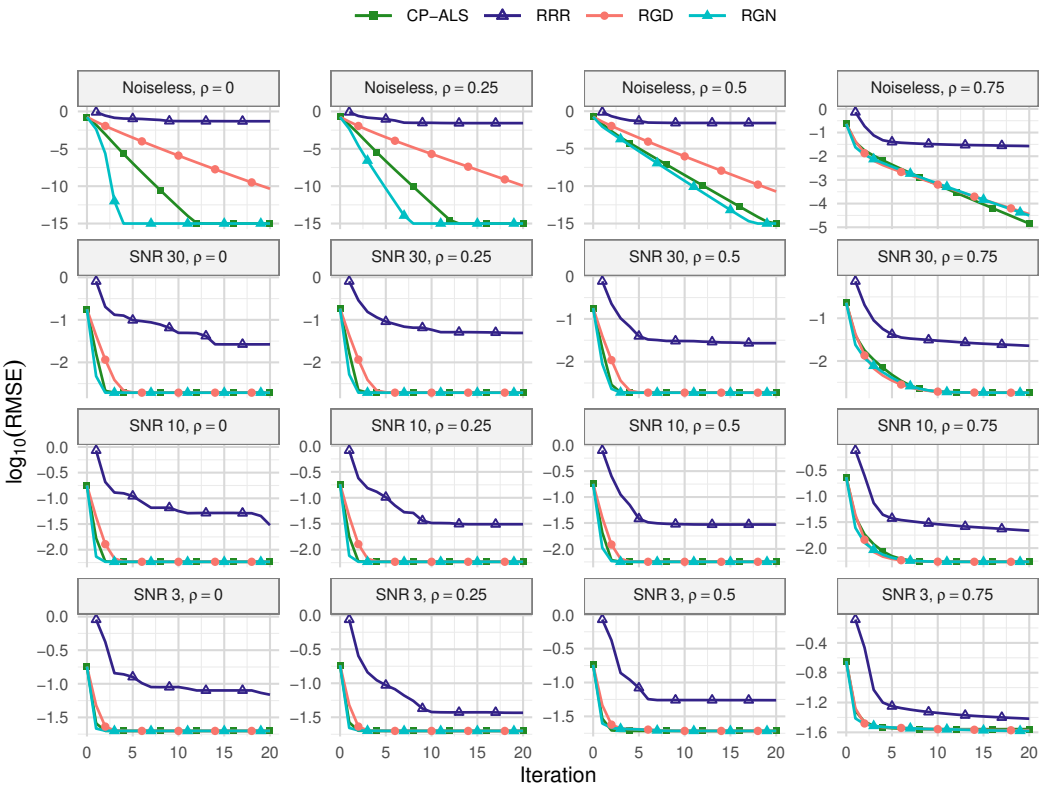


Figure 1: Convergence for CP tensor-on-tensor regression with rank $r = 3$. Each curve is the root mean square, over 20 replications, of the relative Frobenius error. RRR uses a random start, and errors above 1 are omitted. Errors below 10^{-15} are drawn at 10^{-15}. Point symbols are drawn at staggered iterations so that coinciding curves remain distinguishable.

Code and Data Availability

The R package `SegreCP`, which implements RGD and RGN, and the scripts that reproduce all experiments are available at `https://github.com/kxu-stat/SegreCP`. The trade data are compiled from public sources (CEPII BACI and Gravity, World Bank WDI and LPI, and WITS/UNCTAD TRAINS), and the repository includes scripts to download and process them.

Acknowledgments and Disclosure of Funding

We thank the area chair and the anonymous reviewers for helpful suggestions. The work of Yuefeng Han is supported in part by National Science Foundation Grant DMS-2412578. The authors declare no competing interests.

References

[1] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(80):2773–2832, 2014.

[2] Animashree Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180v4*, 2015.

[3] Xuan Bi, Xiwei Tang, Yubai Yuan, Yanqing Zhang, and Annie Qu. Tensors in statistics. *Annual Review of Statistics and Its Application*, 8(1):345–368, 2021.

[4] Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

[5] Paul Breiding and Nick Vannieuwenhoven. A Riemannian trust region method for the canonical tensor rank approximation problem. *SIAM Journal on Optimization*, 28(3):2435–2465, 2018.

[6] Changxiao Cai, H. Vincent Poor, and Yuxin Chen. Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. *IEEE Transactions on Information Theory*, 69(1):407–452, 2023.

[7] J. Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart–Young" decomposition. *Psychometrika*, 35 (3):283–319, 1970.

[8] Bin Chen, Yuefeng Han, and Qiyang Yu. Diffusion index forecasting with tensor data. *Journal of Econometrics*, 254:106204, 2026.

[9] Bin Chen, Yuefeng Han, and Qiyang Yu. Estimation and inference for CP tensor factor models. *Journal of Econometrics*, 253:106167, 2026.

[10] Cody Coleman, Daniel Kang, Deepak Narayanan, Luigi Nardi, Tian Zhao, Jian Zhang, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. Analysis of DAWNBench, a time-to-accuracy machine learning performance benchmark. *ACM SIGOPS Operating Systems Review*, 53(1):14–25, 2019.

[11] Pierre Comon, Xavier Luciani, and André L. F. de Almeida. Tensor decompositions, alternating least squares and other tales. *Journal of Chemometrics*, 23(7–8):393–405, 2009.

[12] Maddalena Conte, Pierre Cotterlaz, and Thierry Mayer. The CEPII Gravity database. Working Paper 2022-05, CEPII, 2022.

[13] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[14] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[15] Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20(53): 1–29, 2015.

[16] Elizabeth D. Dolan and Jorge J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, 2002.

[17] Evgeny Frolov and Ivan Oseledets. Tensor methods and recommender systems. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(3):e1201, 2017.

[18] Guillaume Gaulier and Soledad Zignago. BACI: International trade database at the product-level. The 1994–2007 version. Working Paper 2010-23, CEPII, 2010.

[19] Rungang Han, Rebecca Willett, and Anru R. Zhang. An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics*, 50(1):1–29, 2022.

[20] Yuefeng Han and Cun-Hui Zhang. Tensor principal component analysis in high dimensional CP models. *IEEE Transactions on Information Theory*, 69(2):1147–1167, 2023.

[21] Yuefeng Han, Rong Chen, and Cun-Hui Zhang. Rank determination in tensor factor model. *Electronic Journal of Statistics*, 16(1):1726–1803, 2022.

[22] Yuefeng Han, Rong Chen, Dan Yang, and Cun-Hui Zhang. Tensor factor model estimation by iterative projection. *The Annals of Statistics*, 52(6):2641–2667, 2024.

[23] Yuefeng Han, Dan Yang, Cun-Hui Zhang, and Rong Chen. CP factor model for dynamic tensors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5): 1383–1413, 2024.

[24] Richard A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, 1970.

[25] Jarvis Haupt, Xingguo Li, and David P. Woodruff. Near optimal sketching of low-rank tensor regression. In *Advances in Neural Information Processing Systems*, volume 30, pages 3466–3476, 2017.

[26] Nicholas Hu, Mark A. Iwen, Deanna Needell, and Rongrong Wang. Convergence of the alternating least squares algorithm for CP tensor decompositions. *arXiv preprint arXiv:2505.14037*, 2025.

[27] Simon Jacobsson, Lars Swijsen, Joeri Van der Veken, and Nick Vannieuwenhoven. Warped geometries of Segre–Veronese manifolds. *SIAM Journal on Matrix Analysis and Applications*, 47(3):1551–1577, 2026.

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

[29] Daniel Kressner, Michael Steinlechner, and Bart Vandereycken. Low-rank tensor completion by Riemannian optimization. *BIT Numerical Mathematics*, 54(2):447–468, 2014.

[30] Joseph B. Kruskal. Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2): 95–138, 1977.

[31] Joseph M. Landsberg. *Tensors: Geometry and Applications*, volume 128 of *Graduate Studies in Mathematics*. American Mathematical Society, 2012.

[32] Hung Yi Lee, Mostafa Reisi Gahrooei, Hongcheng Liu, and Massimo Pacella. Robust tensor-on-tensor regression for multidimensional data modeling. *IISE Transactions*, 56(1):43–53, 2024.

[33] James Li, Jacob Bien, and Martin T. Wells. rTensor: An R package for multidimensional array (tensor) unfolding, multiplication, and decomposition. *Journal of Statistical Software*, 87(10): 1–31, 2018.

[34] Carlos Llosa-Vite and Ranjan Maitra. Reduced-rank tensor-on-tensor regression and tensor-variate analysis of variance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2282–2296, 2023.

[35] Eric F. Lock. Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics*, 27(3):638–647, 2018.

[36] Yuetian Luo and Anru R. Zhang. Low-rank tensor estimation via Riemannian Gauss–Newton: Statistical optimality and second-order convergence. *Journal of Machine Learning Research*, 24(381):1–48, 2023.

[37] Yuetian Luo and Anru R. Zhang. Tensor-on-tensor regression: Riemannian optimization, over-parameterization, statistical-computational gap and their interplay. *The Annals of Statistics*, 52 (6):2583–2612, 2024.

[38] Haiyi Mao and Jason Xiaotian Dou. Decomposable sparse tensor on tensor regression. *arXiv preprint arXiv:2212.05024*, 2022.

[39] Jorge J. Moré and Stefan M. Wild. Benchmarking derivative-free optimization algorithms. *SIAM Journal on Optimization*, 20(1):172–191, 2009.

[40] Zhen Qin and Zhihui Zhu. Computational and statistical guarantees for tensor-on-tensor regression with tensor train decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):10577–10587, 2025.

[41] Myriam Rajih, Pierre Comon, and Richard A. Harshman. Enhanced line search: A novel method to accelerate PARAFAC. *SIAM Journal on Matrix Analysis and Applications*, 30(3): 1128–1147, 2008.

[42] Vatsal Sharan and Gregory Valiant. Orthogonalized ALS: A theoretically principled tensor decomposition algorithm for practical use. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3095–3104. PMLR, 2017.

[43] Navjot Singh and Edgar Solomonik. Alternating Mahalanobis distance minimization for accurate and well-conditioned CP decomposition. *SIAM Journal on Scientific Computing*, 45(6): A2781–A2812, 2023.

[44] Navjot Singh, Linjian Ma, Hongru Yang, and Edgar Solomonik. Comparison of accuracy and scalability of Gauss–Newton and alternating least squares for CANDECOMP/PARAFAC decomposition. *SIAM Journal on Scientific Computing*, 43(4):C290–C311, 2021.

[45] Laurent Sorber, Marc Van Barel, and Lieven De Lathauwer. Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank-$(L_r, L_r, 1)$ terms, and a new generalization. *SIAM Journal on Optimization*, 23(2):695–720, 2013.

[46] Will Wei Sun, Junwei Lu, Han Liu, and Guang Cheng. Provable sparse tensor decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):899–916, 2017.

[47] Lars Swijsen, Joeri Van der Veken, and Nick Vannieuwenhoven. Tensor completion using geodesics on Segre manifolds. *Numerical Linear Algebra with Applications*, 29(6):e2446, 2022.

[48] Runshi Tang, Julien Chhor, Olga Klopp, and Anru R. Zhang. Revisit CP tensor decomposition: Statistical optimality and fast convergence. *arXiv preprint arXiv:2505.23046*, 2025.

[49] Xiwei Tang and Lexin Li. Multivariate temporal point process regression. *Journal of the American Statistical Association*, 118(542):830–845, 2023.

[50] Nick Vannieuwenhoven, Raf Vandebril, and Karl Meerbergen. A new truncation strategy for the higher-order singular value decomposition. *SIAM Journal on Scientific Computing*, 34(2): A1027–A1052, 2012.

[51] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2018.

[52] Po-An Wang and Chi-Jen Lu. Tensor decomposition via simultaneous power iteration. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3665–3673. PMLR, 2017.

[53] Ke Xu, Elynn Chen, and Yuefeng Han. Statistical inference for low-rank tensor models. *arXiv preprint arXiv:2501.16223*, 2025.

[54] Anru R. Zhang, Yuetian Luo, Garvesh Raskutti, and Ming Yuan. ISLET: Fast and optimal low-rank tensor regression via importance sketching. *SIAM Journal on Mathematics of Data Science*, 2(2):444–479, 2020.

[55] Tong Zhang and Gene H. Golub. Rank-one approximation to high order tensors. *SIAM Journal on Matrix Analysis and Applications*, 23(2):534–550, 2001.

[56] Qibin Zhao, Cesar F. Caiafa, Danilo P. Mandic, Zenas C. Chao, Yasuo Nagasaka, Naotaka Fujii, Liqing Zhang, and Andrzej Cichocki. Higher order partial least squares (HOPLS): A generalized multilinear regression method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1660–1673, 2013.

[57] Guanhao Zhou, Yuefeng Han, and Xiufan Yu. Factor augmented tensor-on-tensor neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22928–22936, 2025.

[58] Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

[59] Ya Zhou, Raymond K. W. Wong, and Kejun He. Tensor linear regression: Degeneracy and solution. *IEEE Access*, 9:7775–7788, 2021.

Appendices

A Notation

Throughout this paper, we use the following notation and conventions. We use calligraphic uppercase letters (e.g. $\mathcal{T}, \mathcal{X}$) for tensors, uppercase letters (e.g. A, U) for matrices and, in Section 4 and the appendices, also for tensors (e.g. V, W), and lowercase letters (e.g. u, v) for vectors or scalars. For any positive integer m, let $[m] = \{1, 2, \dots, m\}$. We consider order-d tensors with mode dimensions $p_1, p_2, \dots, p_d$, so that $\mathcal{T} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ contains $p^* = \prod_{l=1}^{d} p_l$ total entries. The CP rank is denoted by r, and the sample size in regression contexts is denoted by n.

The vectorization of a tensor $\mathcal{T}$ is denoted by $\mathrm{vec}(\mathcal{T})$. The mode-$k$ unfolding (matricization) is $\mathrm{mat}_k(\mathcal{T}) \in \mathbb{R}^{p_k \times (p^*/p_k)}$. The outer (tensor) product is written $\otimes$. For vectors, it induces a rank-one tensor. The multilinear (Tucker) product of $\mathcal{T}$ with matrices $U_k \in \mathbb{R}^{q_k \times p_k}$ is $\mathcal{T} \times_1 U_1 \times_2 \cdots \times_d U_d$. The k-mode product with U alone is $\mathcal{T} \times_k U$. The tensor inner product is $\langle \mathcal{T}, \mathcal{T}' \rangle = \sum_{i_1,\dots,i_d} \mathcal{T}_{i_1 \dots i_d} \mathcal{T}'_{i_1 \dots i_d}$. The induced Frobenius norm is $\|\mathcal{T}\|_{\mathrm{F}} = \sqrt{\langle \mathcal{T}, \mathcal{T} \rangle}$. For matrices, $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|$ denote the Frobenius and spectral norms. For vectors, $\|\cdot\| = \|\cdot\|_2$ denotes the Euclidean norm. For a linear operator $\mathcal{M}$, denote the operator norm of $\mathcal{M}$ by $\|\mathcal{M}\| := \sup_{\|X\|_{\mathrm{F}}=1} \|\mathcal{M}(X)\|_{\mathrm{F}}$. For a subspace $\mathbb{T}$ of tensors, $\mathcal{P}_{\mathbb{T}}$ is the orthogonal projector onto $\mathbb{T}$, $\mathcal{P}_{\mathbb{T}}^{\perp} = I - \mathcal{P}_{\mathbb{T}}$, and $\mathcal{M}|_{\mathbb{T}}$ is the restriction of $\mathcal{M}$ to $\mathbb{T}$. The adjoint of a linear operator $\mathcal{A}$ is $\mathcal{A}^*$, and I and Id denote identity maps. $\widetilde{O}(\cdot)$ hides logarithmic factors, and $\mathbb{E}$ and $\mathbb{P}$ denote expectation and probability. T-HOSVD is the truncated higher-order singular value decomposition [13]. For a set S and a norm $\|\cdot\|$, $N(\epsilon, S, \|\cdot\|)$ is the ϵ-covering number of S. With df as in Section 4.2, the CP rank satisfies $r \le \min_{k \in [d]} \prod_{\ell \ne k} p_\ell$, so $\log r \le \mathrm{df} - 1$.

Let $\mathbb{S}^{p-1}$ denote the unit sphere in $\mathbb{R}^p$. For a unit vector $u \in \mathbb{R}^p$, define the projector onto its span as $\mathcal{P}_u = uu^\top$ and its orthogonal complement as $\mathcal{P}_u^{\perp} = I_p - uu^\top$. For rank-one tensors, these projections are applied in a modewise manner. The set of nonzero rank-one tensors of the form $u_1 \otimes \cdots \otimes u_d$, where each $u_k \in \mathbb{R}^{p_k} \setminus \{0\}$, forms the Segre manifold. Its geometric structure, including tangent spaces, Riemannian gradients, and retraction maps, is further discussed in Section 3.2.

B Computational Complexity

Denote $p^* = \prod_{l=1}^{d} p_l$ and let r be the CP rank. For tensor-on-tensor regression, define $p_{\mathrm{in}}^* = \prod_{\ell=1}^{d_{\mathrm{in}}} p_\ell$, $p_{\mathrm{out}}^* = \prod_{\ell=d_{\mathrm{in}}+1}^{d} p_\ell$, $\mathrm{df}_{\mathrm{in}} = 1 + \sum_{\ell=1}^{d_{\mathrm{in}}} (p_\ell - 1)$, and $\mathrm{df}_{\mathrm{out}} = 1 + \sum_{\ell=d_{\mathrm{in}}+1}^{d} (p_\ell - 1)$. The rank-one Segre tangent dimension is $\mathrm{df} = \mathrm{df}_{\mathrm{in}} + \mathrm{df}_{\mathrm{out}} - 1 = 1 + \sum_{\ell=1}^{d} (p_\ell - 1)$.

Table 2 summarizes the leading per-iteration costs of all methods in Section 5 for fixed rank and order and, in regression, under the conditions given after the table. One iteration is a full alternating update or one outer Riemannian step. HOPLS instead extracts latent blocks with inner HOOI iterations, so no per-iteration cost is given for it. The costs of RGD and RGN refer to the structure-exploiting implementations described below, and their lower-order terms are given after the table.

The line search of CP-ALS+ELS evaluates $2^d r$ inner products with $\mathcal{Y}$. The RRR entry is the cost of the alternating updates of Lock [35], which is at most $O(dnrp^*)$. Tucker-RGD and TT-IHT compute the full gradient of the least-squares loss, which costs $O(np^*)$, and the cost of Tucker-RGN follows from that of RGN for tensor regression in Luo and Zhang [36, Table 1]. The structured RGD implementation has cost $C_{\mathrm{RGD}} = O(r[n\{d_{\mathrm{in}} p_{\mathrm{in}}^* + p_{\mathrm{out}}^*\} + p_{\mathrm{out}}^* \mathrm{df}_{\mathrm{out}}])$. RGN additionally forms the input-tangent Gram matrix and solves its coupled block, so $C_{\mathrm{RGN}} = C_{\mathrm{RGD}} + O(r\{n\mathrm{df}_{\mathrm{in}}^2 + \mathrm{df}_{\mathrm{in}}^3\})$. Under $n \gtrsim \mathrm{df}_{\mathrm{out}}$, $d_{\mathrm{in}} p_{\mathrm{in}}^* + p_{\mathrm{out}}^* \gtrsim \mathrm{df}_{\mathrm{in}}^2$, and $n(d_{\mathrm{in}} p_{\mathrm{in}}^* + p_{\mathrm{out}}^*) \gtrsim \mathrm{df}_{\mathrm{in}}^3$, these costs reduce to the regression terms in the table. For scalar-on-tensor regression, $p_{\mathrm{out}}^* = \mathrm{df}_{\mathrm{out}} = 1$.

CP decomposition. Classical ALS updates each of the d factor matrices in turn. For a fixed mode ℓ, the matricized tensor times Khatri–Rao product (MTTKRP) costs $O(rp^*)$ for a dense tensor. The Gram matrices and the associated $r \times r$ normal equations contribute $O(r^2 \sum_{\ell=1}^{d} p_\ell)$ per ALS iteration, while the d factorizations of the $r \times r$ normal matrices cost $O(dr^3)$. Summing over all modes gives $O(drp^* + r^2 \sum_{\ell=1}^{d} p_\ell + dr^3)$.

Table 2: Leading per-iteration computational costs. For NLS-GN, m counts the additional residual evaluations. For RRR, k runs over the output modes and ℓ over the input modes.

CP decomposition		CP tensor-on-tensor regression	
Method	Cost	Method	Cost
CP-ALS	$O(drp^*)$	CP-ALS	$O\{dnr(p^*_{\text{in}} + p^*_{\text{out}})\}$
Orth-ALS	$O(drp^*)$	RRR	$O\{nr(p^* \sum_k p_k^{-1} + p^*_{\text{out}} \sum_\ell p_\ell)\}$
ICO	$O(drp^*)$	Tucker-RGD	$O(np^*)$
NLS-GN	$O\{(d+m)rp^*\}$	Tucker-RGN	$O(nrp^*)$
GN-CG	$O(drp^*)$	TT-IHT	$O(np^*)$
CP-ALS+ELS	$O(2^d rp^*)$	HOPLS	N/A
Hybrid AMDM	$O(drp^*)$		
RGD	$O(drp^*)$	RGD	$O\{rn(d_{\text{in}}p^*_{\text{in}} + p^*_{\text{out}})\}$
RGN	$O(drp^*)$	RGN	$O\{rn(d_{\text{in}}p^*_{\text{in}} + p^*_{\text{out}})\}$

For RGD, a fused pass through the residual computes the modewise tangent contractions in $O(dp^*)$ operations per component. The exact rank-$(1, \dots, 1)$ T-HOSVD of the resulting rank-one-plus-tangent proposal is obtained from one 2×2 Gram eigenproblem per mode. Residual and cache updates cost $O(rp^*)$, so the total per-iteration cost is $O(drp^*)$.

RGN for CP decomposition, where the measurement operator $\mathcal{A}$ is the identity mapping, becomes equivalent to an RGD step with a unit step size ($\alpha_t = 1$), and thus has an identical per-iteration cost of $O(rdp^*)$.

CP tensor-on-tensor regression. The regression implementation contracts the design modewise and forms the exact input and output tangent sufficient statistics without constructing an ambient $p^* \times \mathrm{df}$ basis. The input contractions cost $O(nd_{\text{in}}p^*_{\text{in}})$, while the response and output-tangent contractions cost $O(np^*_{\text{out}} + p^*_{\text{out}}\mathrm{df}_{\text{out}})$, per component. These give C_{RGD} above. RGN uses the same contractions, solves the coupled $\mathrm{df}_{\text{in}} \times \mathrm{df}_{\text{in}}$ block, and updates the orthogonal output coordinates by scalar division, yielding C_{RGN}. The exact factorization underlying these costs is given in Appendix E.2. For CP-ALS, updating the factor of one mode contracts the design and the response with the factors of the other modes at cost $O\{nr(p^*_{\text{in}} + p^*_{\text{out}})\}$, so one sweep costs $O\{dnr(p^*_{\text{in}} + p^*_{\text{out}})\}$. The normal equations of an input mode ℓ add $O(nr^2p_\ell^2 + r^3p_\ell^3)$, which is of no larger order for fixed r under the conditions above.

C Simulation Results

Incoherence condition. To explore scenarios with non-orthogonal factors, we generate factor matrices whose columns achieve a prescribed level of pairwise coherence. Specifically, for a given coherence parameter $\rho \in [0, 1)$ and rank r, we first construct the $r \times r$ Gram matrix Γ with entries

$$\Gamma_{ij} = \rho^{|i-j|}, \quad i, j \in [r],$$

which corresponds to an autoregressive correlation structure of order one (AR(1)). We then compute the Cholesky factorization $\Gamma = LL^\top$ and embed $L^\top$ into $\mathbb{R}^p$ by stacking r identity rows on top of $(p - r)$ zero rows, forming the matrix $(I_r, 0)^\top \in \mathbb{R}^{p \times r}$. Multiplying this matrix with $L^\top$ yields vectors with the desired correlation pattern, and each column is normalized to unit length.

Finally, to avoid artificial alignment with the coordinate axes, we multiply by a random orthogonal matrix, the orthogonal factor of the QR decomposition of a $p \times p$ matrix with i.i.d. standard normal entries. The resulting factor matrix thus has columns with controlled coherence. By varying ρ, we tune the similarity ("coherence") between the factors: $\rho = 0$ corresponds to orthogonal columns, whereas ρ close to 1 yields highly coherent columns. The incoherence parameter of Assumption 3.3 is $\eta = \rho$.

Initialization. We employ Composite Principal Component Analysis (CPCA) [20] as a warm start: Algorithm 3 for tensor decomposition and Algorithm 6 for tensor-on-tensor regression. CPCA constructs the initial CP basis vectors through a two-stage spectral procedure, first decomposing a matrix unfolding and then decomposing the refolded singular vectors.

Signal-to-noise ratio. Let $\mathrm{df} = 1 + \sum_{\ell=1}^{d}(p_\ell - 1)$ be the dimension of the tangent space of the Segre manifold. With noise standard deviation σ, we set $s = \lambda_r/(\sigma\sqrt{\mathrm{df}/n})$ for regression and $s = \lambda_r/(\sigma\sqrt{\mathrm{df}})$ for decomposition. Up to constants, $1/s$ is the noise term in the error bounds of Section 4.2.

Tensor-on-tensor regression. The coefficient tensor $\mathcal{T}^\star = \sum_{i=1}^{r} \lambda_i\, u_{1,i} \otimes \cdots \otimes u_{5,i} \in \mathbb{R}^{10\times10\times10\times10\times10}$ has rank $r = 3$ and weights $\lambda_i = \kappa^{1/2-(i-1)/(r-1)}$ with $\kappa = 30$. The covariates have i.i.d. standard normal entries, so $\sigma_X = 1$, and $\mathcal{Y}_m = \langle \mathcal{X}_m, \mathcal{T}^\star\rangle_{\mathrm{in}} + \mathcal{E}_m$, where $\mathcal{E}_m$ has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. RGD uses the fixed step size $\alpha_t = \sigma_X^{-2}$, which satisfies the conditions $0 < \alpha_t\sigma_X^2 \le 1$ and $|1 - \alpha_t\sigma_X^2| \le 1/2$ of Corollary G.4. RRR starts from its own random initialization, Tucker-RGD and Tucker-RGN from the T-HOSVD of $\mathcal{A}^*(\mathcal{Y})$, and TT-IHT from its TT-SVD. All methods run for 20 iterations. The Tucker methods use multilinear rank $(3, \ldots, 3)$, TT-IHT uses TT rank $(3, \ldots, 3)$, and HOPLS extracts three latent blocks. HOPLS is not iterative and is shown as a horizontal line. Figures 1 and 2 show the convergence. Without noise at $\rho = 0$, Tucker-RGN misses the weakest component in three of the 20 replications, which determines its root mean square error.

CP decomposition. The signal tensor $\mathcal{T}^\star$ has the same CP form with dimensions $(60, 30, 30, 30)$, rank $r = 3$, and weights $\lambda_i = \sqrt{r}\,\kappa^{1/2-(i-1)/(r-1)} p_{\max}^{d/4}$ with $\kappa = 30$ and $p_{\max} = 60$. We observe $\mathcal{Y} = \mathcal{T}^\star + \mathcal{E}$, where $\mathcal{E}$ has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. RGD uses step size $\alpha = 0.5$, which satisfies the conditions $0 < \alpha \le 1$ and $|1 - \alpha| \le 1/2$ of Corollary G.3, and RGN uses unit step size (with $\alpha = 1$, RGD coincides with RGN, Appendix B). All methods start from CPCA and run for 100 iterations. Figures 3 and 4 show the first 20 iterations, or all 100 at $\rho = 0.75$, where convergence is slower. With orthogonal factors ($\rho = 0$), CPCA recovers the noiseless tensor exactly, so all methods start close to the error level in that column.

Running time. Table 3 reports the median running time per iteration over all settings, measured on one core of an Intel Xeon 2.1 GHz CPU. All methods are implemented in R. RGD and RGN have the lowest cost per iteration in both problems.

Time to the error level. Tables 4 and 5 report the time to the error level in the runs of 100 iterations (in regression, of CP-ALS, RGD and RGN). In each replication, the level is the lowest error at iteration 100 among CP-ALS, RGD and RGN, adapting the convergence test of Moré and Wild [39], and a method reaches it at the first iteration at which its error is within 1% of it. The time, a time-to-accuracy measure as in Coleman et al. [10], is the median over the 20 replications of this number of iterations times the median time per iteration of the method in the setting, from the timing study of Table 3. In regression, RGD or RGN is the fastest in 11 of the 12 settings, and RGN is within a factor of 1.4 of the fastest method in every setting. In decomposition, RGN is the fastest in 5 of the 11 settings in which the methods do not all start at the level, NLS-GN in 3, CP-ALS in 2 and GN-CG in 1, and RGN is within a factor of 3.1 of the fastest method. RGN has the best average rank over all 12 settings in both problems, 1.4 and 2.0 [14], and is faster than CP-ALS in 10 of the 12 regression and 9 of the 11 decomposition settings, by up to a factor of 2.5.

Performance profiles. Figure 5 summarizes Tables 4 and 5 and the errors at the last iteration of Tables 6 and 7 as performance profiles [16]. For a method and a factor $\tau \ge 1$, the profile is the fraction of the 12 settings in which the time of the method to the level, or its mean error at the last iteration, is at most τ times the best value among the methods shown. The value at $\tau = 1$ is the fraction of settings in which the method is the best, and the smallest τ at which the profile reaches one bounds how far the method is from the best in its worst setting. In regression the time profile of RGN is never below those of the other methods and reaches one at $\tau = 1.4$, and in decomposition it is never below them for $\tau \le 2.3$ and reaches one at $\tau = 3.0$. The error profiles of RGD, RGN and CP-ALS reach one by $\tau = 1.07$ in both problems, whereas those of RRR and the methods in other tensor formats in regression, and of Orth-ALS and ICO in decomposition, stay below one for $\tau \le 2$.

Errors at the last iteration. Each curve in Figures 1–4 is the root mean square, over 20 replications, of the relative Frobenius error $\|\widehat{\mathcal{T}} - \mathcal{T}^\star\|_{\mathrm{F}}/\|\mathcal{T}^\star\|_{\mathrm{F}}$. Tables 6 and 7 report the mean and standard deviation of this error at the last iteration, 20 in regression and 100 in decomposition, in

the settings with noise. In regression, RGN attains the lowest mean error in every setting, in most settings together with RGD and CP-ALS, whereas the methods in other tensor formats have errors 1.5 to 2.7 times larger, and HOPLS and, in 10 of the 12 settings, RRR larger still. In decomposition, RGD, RGN, CP-ALS, NLS-GN, GN-CG, CP-ALS+ELS and Hybrid AMDM attain the same error up to one standard deviation in every setting, whereas Orth-ALS and ICO have larger errors when $\rho > 0$ and $s \geq 10$.

Settings without noise. Without noise the errors fall to machine precision, so we describe these settings instead of tabulating them. In regression, RGN reaches machine precision in fewer iterations than CP-ALS when $\rho \leq 0.5$, and at $\rho = 0.75$ only Tucker-RGN reaches it within 20 iterations. In decomposition with $\rho > 0$, ICO and Hybrid AMDM reach machine precision within two iterations, NLS-GN and GN-CG level off between 10^{-12} and 10^{-7}, and at $\rho = 0.75$ CP-ALS, CP-ALS+ELS and RGN reach 10^{-8} or less within 100 iterations.

Sensitivity to the initialization. Tables 8 and 9 report, in the settings with noise, the errors at iterations 0, 3, 5, 10 and 20 from three starts: CPCA, T-HOSVD, which takes the i-th component from the i-th leading singular vectors of the mode unfoldings, and random unit factors with weights fitted to the data. When $\rho \leq 0.5$, CPCA starts closer to $\mathcal{T}^\star$ than T-HOSVD, and at $\rho = 0.75$ the errors of the two starts are within 4% of each other. When $\rho \leq 0.5$ and $s \geq 10$, RGN started from CPCA is within 1% of its error at iteration 20 after five iterations in both problems. In the same settings, RGN started from T-HOSVD attains this error within ten iterations in decomposition, whereas in regression its error at iteration 10 is 1.2 to 3.5 times that from CPCA. At $\rho = 0.75$, after three iterations of RGN, the error from CPCA is 5 to 19 times smaller than that from T-HOSVD in regression and 1.7 to 3.5 times smaller in decomposition. At iteration 20, some replications in regression started from T-HOSVD miss the weakest component, and when $\rho \leq 0.5$ this happens less often for RGD and RGN than for CP-ALS. In decomposition, the errors at iteration 20 from T-HOSVD are within about one standard deviation (Table 7) of those from CPCA, except at $\rho = 0.75$, where RGD and RGN converge more slowly. From a random start, all three methods often miss components within 20 iterations, RGD and RGN more often than CP-ALS in most settings. This agrees with the local nature of the guarantees in Section 4.

Table 3: Median running time per iteration in milliseconds over all settings.

Tensor-on-tensor regression		CP decomposition	
RGD	15.4	RGD	14.9
RGN	16.8	RGN	14.7
CP-ALS	21.6	CP-ALS	22.4
RRR	338.3	Hybrid AMDM	24.5
TT-IHT	18.6	Orth-ALS	27.2
Tucker-RGD	76.3	ICO	28.6
Tucker-RGN	262.3	NLS-GN	32.6
		GN-CG	41.3
		CP-ALS+ELS	531.8

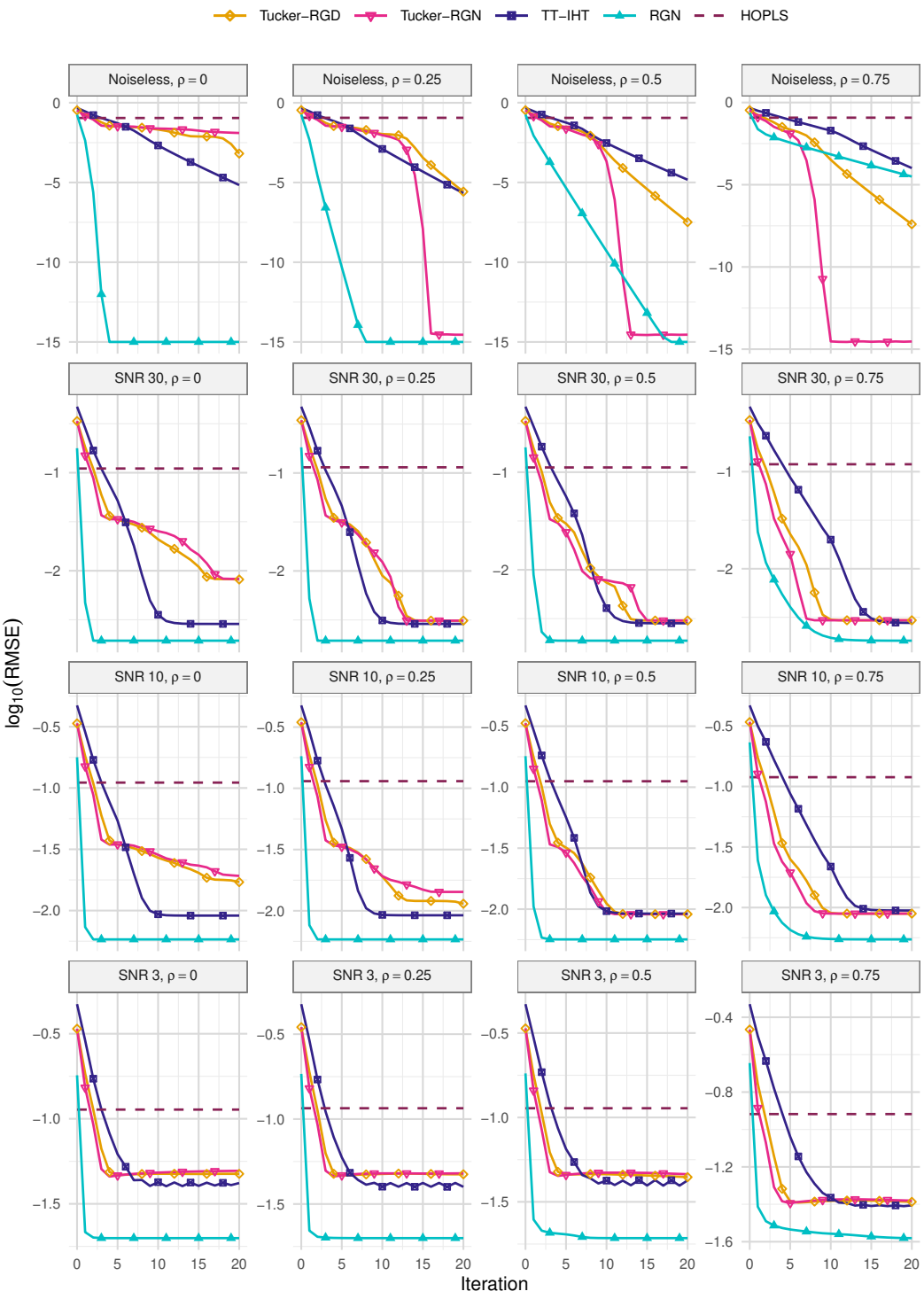


Figure 2: Convergence for tensor-on-tensor regression of methods in other tensor formats, with RGN as reference. HOPLS (dashed) is not iterative. Errors below 10^{-15} are drawn at 10^{-15}. Without noise at $\rho = 0$, Tucker-RGN misses the weakest component in three of the 20 replications, which keeps its root mean square error near 10^{-2}.

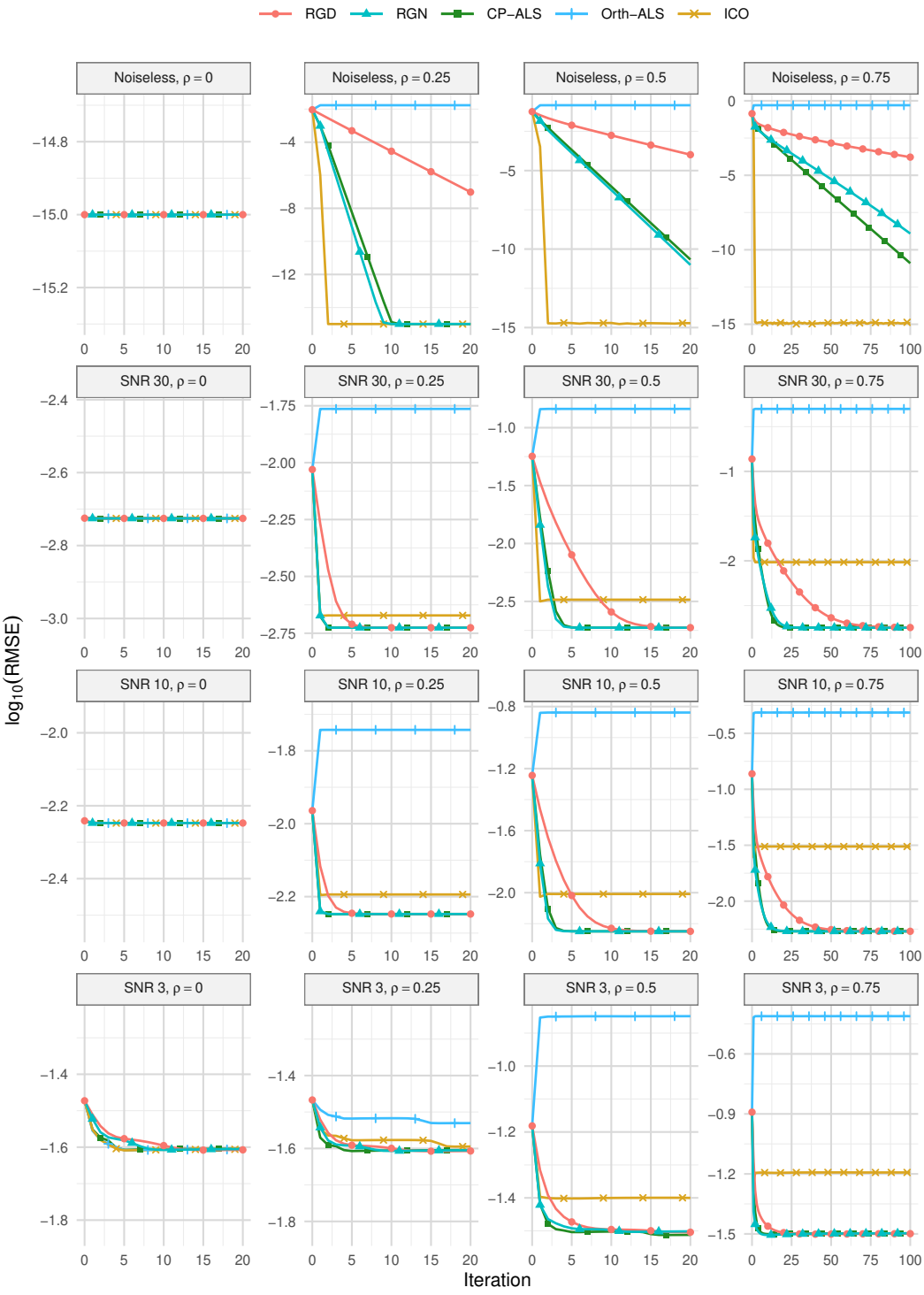


Figure 3: Convergence for CP decomposition of fourth-order tensors. At $\rho = 0.75$ the horizontal axis extends to 100 iterations. Errors below 10^{-15} are drawn at 10^{-15}. At $\rho = 0$, CPCA recovers the noiseless tensor exactly, so without noise all curves stay at 10^{-15}, and with noise they start close to the error level. Orth-ALS orthogonalizes the factor estimates before every sweep, which biases them when the factors are coherent, so for $\rho > 0$ its error levels off above that of the other methods. ICO contracts the data with the pseudo-inverses of the factor matrices, which recovers the noiseless tensor within two iterations but amplifies the noise when $\rho > 0$.

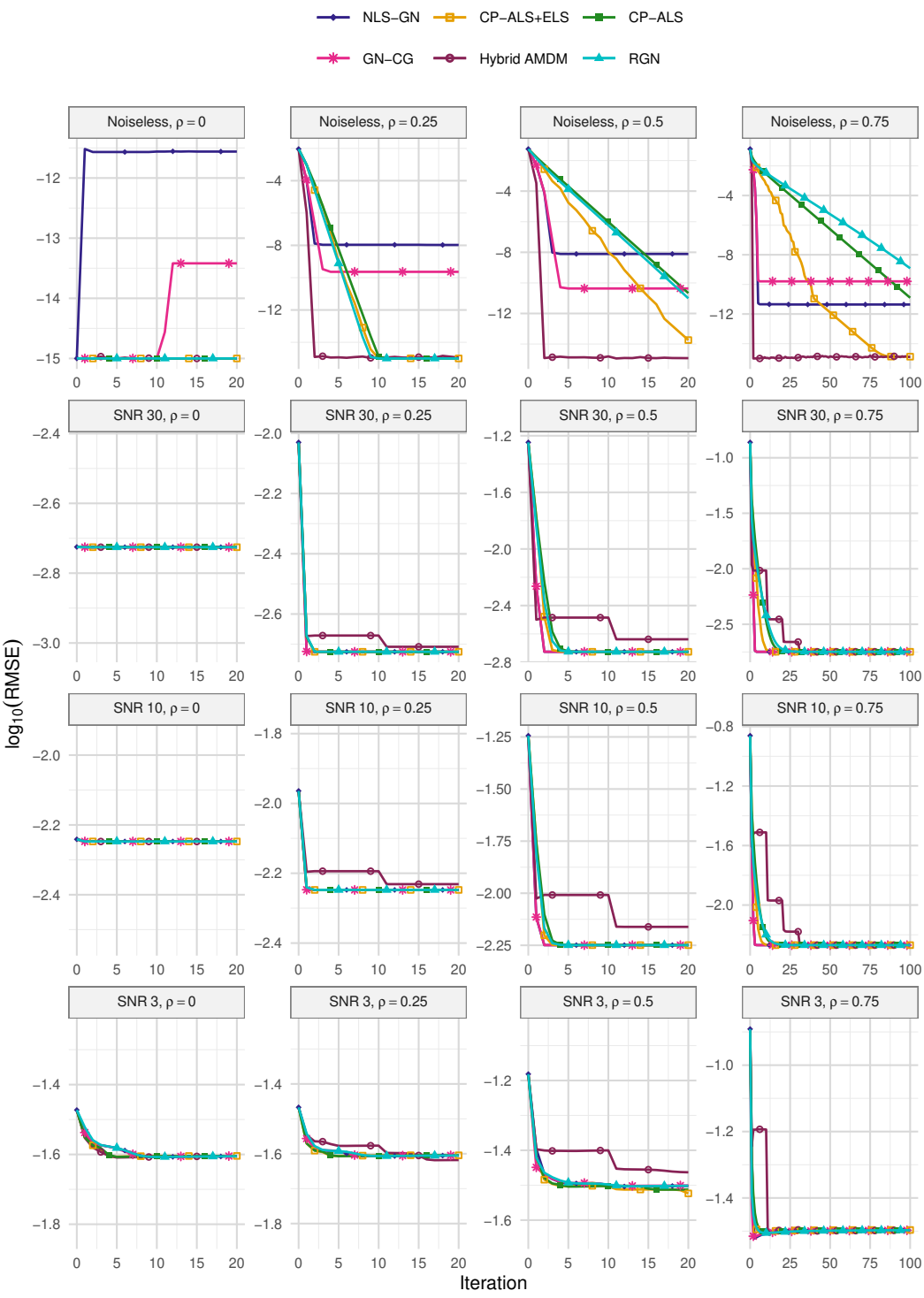


Figure 4: Convergence for CP decomposition: Gauss-Newton and accelerated alternating methods, with CP-ALS and RGN as reference. At $\rho = 0.75$ the horizontal axis extends to 100 iterations. Errors below 10^{-15} are drawn at 10^{-15}, and at $\rho = 0$ the curves start at or close to the error level for the reason given in Figure 3. Without noise at $\rho = 0$, the errors of NLS-GN and GN-CG rise from 10^{-15} to their numerical floors, about $10^{-11.6}$ and $10^{-13.4}$. Hybrid AMDM moves from AMDM toward ALS updates every 10 iterations [43], which produces the steps in its curves.

Table 4: Tensor-on-tensor regression: time in milliseconds to come within 1% of the error level, the lowest error at iteration 100 among CP-ALS, RGD and RGN in the replication, in the settings with noise. The time is the median over the 20 replications of the number of iterations needed, times the median time per iteration of the method in the setting. A replication in which the method does not reach the level counts as infinitely many iterations. Bold: fastest in each setting. RRR, Tucker-RGD, Tucker-RGN, TT-IHT and HOPLS are not shown, as they reach the level in fewer than half of the replications in every setting.

s	Method	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
30	RGD	86	84	74	**196**
	RGN	**32**	**34**	**44**	234
	CP-ALS	69	62	64	264
10	RGD	78	63	79	**135**
	RGN	**38**	**34**	**49**	180
	CP-ALS	48	71	63	213
3	RGD	62	57	67	**148**
	RGN	**35**	49	**56**	174
	CP-ALS	45	**42**	63	150

Table 5: CP decomposition: time in milliseconds to come within 1% of the error level, the lowest error at iteration 100 among CP-ALS, RGD and RGN in the replication, in the settings with noise. The time is the median over the 20 replications of the number of iterations needed, times the median time per iteration of the method in the setting. A replication in which the method does not reach the level counts as infinitely many iterations. Bold: fastest in each setting. At $s = 30$ and $\rho = 0$, every method starts at the level. Orth-ALS and ICO are not shown, as they reach the level in fewer than half of the replications when $\rho > 0$.

s	Method	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
30	RGD	**0**	103	241	1290
	RGN	**0**	**28**	68	391
	CP-ALS	**0**	45	101	543
	NLS-GN	**0**	30	**55**	**130**
	GN-CG	**0**	42	86	174
	CP-ALS+ELS	**0**	1067	2022	6253
	Hybrid AMDM	**0**	746	741	808
10	RGD	**14**	66	187	895
	RGN	**14**	**28**	**56**	293
	CP-ALS	22	41	86	377
	NLS-GN	32	29	67	**108**
	GN-CG	41	38	83	139
	CP-ALS+ELS	538	1017	1594	5274
	Hybrid AMDM	26	746	789	744
3	RGD	107	89	140	359
	RGN	53	**41**	76	68
	CP-ALS	**46**	42	**74**	167
	NLS-GN	108	98	132	103
	GN-CG	103	99	129	**66**
	CP-ALS+ELS	1124	1066	1580	2703
	Hybrid AMDM	54	693	749	295

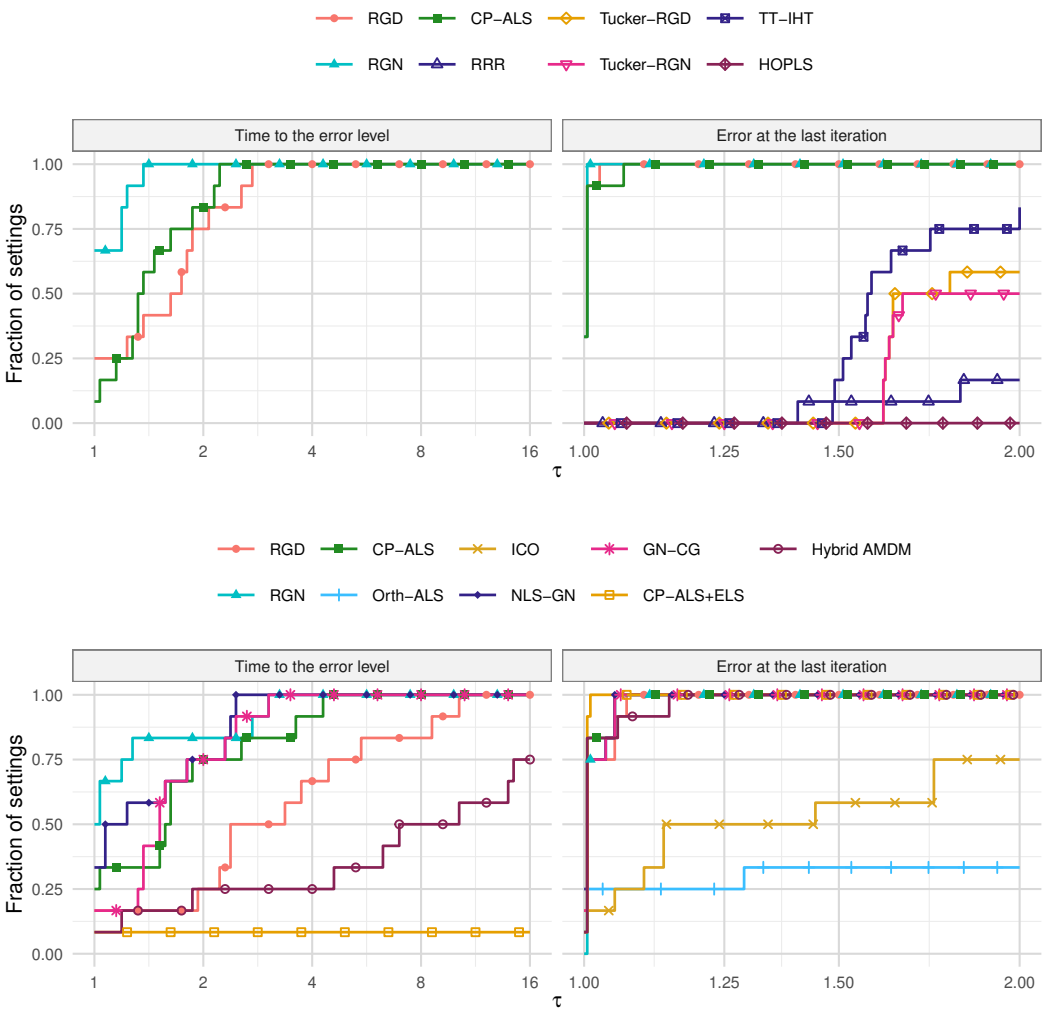


Figure 5: Performance profiles over the 12 settings with noise, for tensor-on-tensor regression (top) and CP decomposition (bottom). Left: the fraction of settings in which the method comes within 1% of the error level in at most τ times the time of the fastest method (Tables 4 and 5); at $s = 30$, $\rho = 0$ in decomposition every method starts at the level. Right: the fraction of settings in which the mean error of the method at the last iteration is at most τ times the lowest among the methods shown (Tables 6 and 7). The horizontal axis is logarithmic.

Table 6: Tensor-on-tensor regression: mean (standard deviation) over 20 replications of the relative Frobenius error at iteration 20, in units of 10^{-3}, in the settings with noise. Bold: lowest mean in each setting. HOPLS is not iterative.

s	Method	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
30	RGD	**1.93** (0.07)	**1.94** (0.16)	**1.88** (0.12)	**1.81** (0.08)
	RGN	**1.93** (0.07)	**1.94** (0.16)	**1.88** (0.12)	**1.81** (0.08)
	CP-ALS	**1.93** (0.07)	**1.94** (0.16)	**1.88** (0.12)	**1.81** (0.08)
	RRR	22.0 (15.1)	32.6 (37.7)	23.2 (14.3)	20.1 (10.9)
	Tucker-RGD	4.58 (6.91)	3.11 (0.15)	3.02 (0.11)	2.94 (0.10)
	Tucker-RGN	4.63 (7.03)	3.11 (0.15)	3.02 (0.11)	2.94 (0.10)
	TT-IHT	2.87 (0.14)	2.88 (0.14)	2.83 (0.12)	2.77 (0.09)
	HOPLS	109 (16)	114 (13)	111 (12)	119 (12)
10	RGD	**5.81** (0.22)	**5.82** (0.49)	**5.64** (0.36)	**5.45** (0.25)
	RGN	**5.81** (0.22)	**5.82** (0.49)	**5.64** (0.36)	**5.45** (0.25)
	CP-ALS	**5.81** (0.22)	**5.82** (0.49)	**5.64** (0.36)	**5.45** (0.25)
	RRR	27.9 (11.4)	29.3 (10.0)	27.3 (11.2)	19.1 (10.5)
	Tucker-RGD	14.0 (10.1)	10.4 (4.8)	9.08 (0.35)	8.90 (0.32)
	Tucker-RGN	15.6 (11.5)	11.9 (8.0)	9.08 (0.35)	8.90 (0.32)
	TT-IHT	9.09 (0.42)	9.18 (0.50)	9.18 (0.47)	9.43 (0.66)
	HOPLS	110 (16)	114 (13)	111 (12)	119 (12)
3	RGD	**19.9** (0.7)	**19.8** (1.7)	**19.2** (1.3)	25.6 (8.0)
	RGN	**19.9** (0.7)	**19.8** (1.7)	**19.2** (1.3)	**25.0** (8.1)
	CP-ALS	**19.9** (0.7)	**19.8** (1.7)	**19.2** (1.3)	26.6 (8.2)
	RRR	52.7 (45.4)	36.0 (7.8)	46.0 (30.6)	35.2 (15.6)
	Tucker-RGD	47.2 (5.4)	47.0 (5.8)	43.8 (6.6)	40.9 (3.7)
	Tucker-RGN	49.5 (0.6)	47.7 (4.5)	45.8 (5.7)	41.5 (3.9)
	TT-IHT	41.6 (5.2)	39.6 (4.8)	42.2 (4.5)	39.3 (1.9)
	HOPLS	112 (16)	115 (13)	113 (12)	120 (11)

Table 7: CP decomposition: mean (standard deviation) over 20 replications of the relative Frobenius error at iteration 100, in units of 10^{-3}, in the settings with noise. Bold: lowest mean in each setting.

s	Method	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
30	RGD	**1.88** (0.06)	**1.88** (0.05)	**1.87** (0.07)	**1.79** (0.05)
	RGN	**1.88** (0.06)	**1.88** (0.05)	**1.87** (0.07)	**1.79** (0.06)
	CP-ALS	**1.88** (0.06)	**1.88** (0.05)	**1.87** (0.07)	**1.79** (0.06)
	Orth-ALS	**1.88** (0.06)	17.2 (0.1)	145 (0)	500 (2)
	ICO	**1.88** (0.06)	2.13 (0.07)	3.27 (0.12)	9.56 (1.33)
	NLS-GN	**1.88** (0.06)	**1.88** (0.05)	**1.87** (0.07)	**1.79** (0.06)
	GN-CG	**1.88** (0.06)	**1.88** (0.05)	**1.87** (0.07)	**1.79** (0.06)
	CP-ALS+ELS	**1.88** (0.06)	**1.88** (0.05)	**1.87** (0.07)	**1.79** (0.06)
	Hybrid AMDM	**1.88** (0.06)	**1.88** (0.05)	**1.87** (0.07)	**1.79** (0.06)
10	RGD	**5.65** (0.17)	**5.65** (0.16)	**5.63** (0.20)	5.38 (0.17)
	RGN	**5.65** (0.17)	**5.65** (0.16)	**5.63** (0.20)	**5.37** (0.17)
	CP-ALS	**5.65** (0.17)	**5.65** (0.16)	**5.63** (0.20)	**5.37** (0.17)
	Orth-ALS	**5.65** (0.17)	18.1 (0.4)	145 (1)	485 (7)
	ICO	**5.65** (0.17)	6.39 (0.20)	9.80 (0.36)	30.6 (3.5)
	NLS-GN	**5.65** (0.17)	**5.65** (0.16)	**5.63** (0.20)	**5.37** (0.17)
	GN-CG	**5.65** (0.17)	**5.65** (0.16)	**5.63** (0.20)	**5.37** (0.17)
	CP-ALS+ELS	**5.65** (0.17)	**5.65** (0.16)	**5.63** (0.20)	**5.37** (0.17)
	Hybrid AMDM	**5.65** (0.17)	**5.65** (0.16)	**5.63** (0.20)	**5.37** (0.17)
3	RGD	23.5 (8.5)	23.4 (8.6)	29.0 (10.4)	**31.7** (0.3)
	RGN	23.5 (8.6)	23.5 (8.7)	28.1 (10.4)	31.8 (0.3)
	CP-ALS	23.5 (8.5)	23.5 (8.7)	**27.2** (10.2)	31.9 (0.3)
	Orth-ALS	**22.4** (7.6)	28.9 (6.2)	142 (2)	388 (6)
	ICO	23.4 (8.6)	24.6 (7.0)	39.2 (5.2)	63.9 (5.6)
	NLS-GN	23.5 (8.5)	23.5 (8.7)	28.1 (10.4)	31.8 (0.3)
	GN-CG	23.5 (8.5)	23.5 (8.7)	28.1 (10.4)	31.8 (0.3)
	CP-ALS+ELS	**22.4** (7.7)	**22.4** (7.7)	**27.2** (10.2)	31.9 (0.3)
	Hybrid AMDM	23.5 (8.6)	**22.4** (7.7)	31.1 (10.2)	31.8 (0.3)

Table 8: Sensitivity to the initialization in tensor-on-tensor regression: mean over 20 replications of the relative Frobenius error, in units of 10^{-3}, at iterations $t = 0, 3, 5, 10, 20$ from the CPCA, T-HOSVD and random starts, in the settings with noise. At $t = 0$ the entry is the error of the start, the same for all three methods.

			CPCA					T-HOSVD					Random				
s	ρ	Method	0	3	5	10	20	0	3	5	10	20	0	3	5	10	20
30	0	RGD	176	3.78	1.97	1.93	1.93	248	33.2	13.1	5.75	1.93	1000	886	332	174	144
		RGN	176	1.93	1.93	1.93	1.93	248	28.6	11.4	6.75	1.93	1000	887	311	166	136
		CP-ALS	176	1.94	1.93	1.93	1.93	248	17.7	14.1	9.68	8.14	1000	168	152	93.2	48.4
	0.25	RGD	181	3.74	1.97	1.94	1.94	254	38.5	11.9	3.54	3.54	1000	971	622	114	73.6
		RGN	181	1.94	1.94	1.94	1.94	254	33.8	9.67	5.14	3.54	1000	968	597	122	80.7
		CP-ALS	181	1.94	1.94	1.94	1.94	254	17.1	11.3	9.69	8.14	1000	161	55.6	35.4	23.6
	0.5	RGD	178	3.48	1.90	1.88	1.88	250	97.0	31.2	2.35	1.88	1000	920	450	106	55.2
		RGN	178	1.89	1.88	1.88	1.88	250	94.3	30.7	2.61	1.88	1000	915	433	113	37.2
		CP-ALS	178	1.90	1.88	1.88	1.88	250	19.8	15.3	10.8	10.3	1000	126	56.0	30.9	26.1
	0.75	RGD	230	6.56	3.41	1.92	1.81	231	147	99.5	18.8	8.42	1000	931	369	67.0	35.2
		RGN	230	7.69	4.03	1.94	1.81	231	147	96.4	14.5	7.60	1000	931	359	60.3	35.5
		CP-ALS	230	10.4	4.66	1.92	1.81	231	38.7	20.7	13.4	7.46	1000	77.7	46.7	30.4	20.1
10	0	RGD	177	6.72	5.83	5.81	5.81	248	34.3	16.6	8.58	5.81	1000	886	333	176	157
		RGN	177	5.81	5.81	5.81	5.81	248	30.0	17.0	9.41	7.09	1000	887	317	171	130
		CP-ALS	177	5.81	5.81	5.81	5.81	248	26.1	18.4	15.5	12.8	1000	166	125	116	54.9
	0.25	RGD	182	6.67	5.83	5.82	5.82	254	39.4	14.9	8.72	5.97	1000	971	631	110	70.1
		RGN	182	5.82	5.82	5.82	5.82	254	35.2	13.9	8.49	5.82	1000	969	612	101	60.7
		CP-ALS	182	5.82	5.82	5.82	5.82	254	23.9	17.0	12.8	11.4	1000	129	64.5	38.3	29.3
	0.5	RGD	179	6.41	5.65	5.64	5.64	250	98.0	36.6	8.92	5.64	1000	921	463	120	37.5
		RGN	179	5.65	5.64	5.64	5.64	250	95.0	30.4	7.04	5.64	1000	916	441	113	36.9
		CP-ALS	179	5.65	5.64	5.64	5.64	250	24.8	16.0	12.1	9.75	1000	132	60.7	34.0	28.8
	0.75	RGD	230	8.36	6.18	5.48	5.45	231	148	106	20.1	9.75	1000	934	372	64.8	33.9
		RGN	230	9.25	6.52	5.49	5.45	231	147	103	17.2	10.6	1000	934	359	58.3	30.4
		CP-ALS	230	11.7	6.99	5.49	5.45	231	36.8	20.7	13.9	11.6	1000	77.1	48.9	32.7	22.9
3	0	RGD	179	20.2	19.9	19.9	19.9	249	43.4	34.1	27.3	27.1	1000	895	376	161	121
		RGN	179	19.9	19.9	19.9	19.9	249	40.0	34.0	28.2	28.1	1000	898	365	162	111
		CP-ALS	179	19.9	19.9	19.9	19.9	249	39.1	38.8	37.1	35.1	1000	155	130	90.1	75.7
	0.25	RGD	183	20.2	19.9	19.8	19.8	254	49.5	35.1	27.5	25.0	1000	974	691	127	79.8
		RGN	183	20.0	19.9	19.8	19.8	254	47.1	35.6	27.7	25.2	1000	973	680	122	75.3
		CP-ALS	183	19.9	19.8	19.8	19.8	254	37.9	37.4	35.3	35.5	1000	138	88.3	46.7	47.2
	0.5	RGD	181	20.6	20.1	19.3	19.2	250	106	53.5	34.0	30.6	1000	932	549	128	39.6
		RGN	181	20.4	20.0	19.3	19.2	250	104	50.9	33.4	30.9	1000	927	532	119	45.0
		CP-ALS	181	19.3	19.2	19.2	19.2	250	38.8	35.9	35.6	33.9	1000	134	81.9	57.0	46.9
	0.75	RGD	225	30.1	28.5	26.7	25.6	231	150	123	51.0	31.0	1000	950	440	76.1	51.0
		RGN	225	30.3	28.6	26.7	25.0	231	150	121	54.8	37.7	1000	949	420	75.7	37.5
		CP-ALS	225	30.3	27.5	26.5	26.6	231	70.4	42.1	33.7	30.9	1000	79.6	51.6	40.3	33.2

Table 9: Sensitivity to the initialization in CP decomposition, in the format of Table 8.

			CPCA					T-HOSVD					Random				
s	ρ	Method	0	3	5	10	20	0	3	5	10	20	0	3	5	10	20
30	0	RGD	1.88	1.88	1.88	1.88	1.88	1.96	1.88	1.88	1.88	1.88	1000	494	182	176	175
		RGN	1.88	1.88	1.88	1.88	1.88	1.96	1.88	1.88	1.88	1.88	1000	292	174	139	138
		CP-ALS	1.88	1.88	1.88	1.88	1.88	1.96	1.88	1.88	1.88	1.88	1000	95.1	85.2	85.2	70.3
	0.25	RGD	9.33	2.46	1.95	1.88	1.88	91.7	14.2	4.64	1.90	1.88	1000	465	176	153	144
		RGN	9.33	1.88	1.88	1.88	1.88	91.7	1.88	1.88	1.88	1.88	1000	215	164	48.1	36.8
		CP-ALS	9.33	1.88	1.88	1.88	1.88	91.7	1.88	1.88	1.88	1.88	1000	53.9	40.0	35.2	29.7
	0.5	RGD	56.7	15.1	7.98	2.56	1.88	152	41.4	20.4	4.67	1.89	1000	508	167	109	88.9
		RGN	56.7	2.23	1.88	1.87	1.87	152	9.63	2.07	1.87	1.87	1000	296	147	57.2	42.2
		CP-ALS	56.7	2.60	1.89	1.87	1.87	152	3.56	1.91	1.87	1.87	1000	72.6	47.1	35.0	27.1
	0.75	RGD	138	33.8	25.3	15.7	7.72	133	65.7	45.8	24.3	10.7	1000	484	134	68.4	37.1
		RGN	138	14.2	8.98	3.80	1.90	133	49.6	20.4	6.34	2.06	1000	222	98.8	30.6	22.4
		CP-ALS	138	18.7	10.2	3.38	1.81	133	22.1	11.5	3.65	1.82	1000	51.8	40.2	32.7	26.8
10	0	RGD	5.73	5.66	5.65	5.65	5.65	7.80	5.69	5.66	5.65	5.65	1000	519	190	182	175
		RGN	5.73	5.65	5.65	5.65	5.65	7.80	5.65	5.65	5.65	5.65	1000	311	182	112	100
		CP-ALS	5.73	5.65	5.65	5.65	5.65	7.80	5.65	5.65	5.65	5.65	1000	109	85.7	85.6	78.0
	0.25	RGD	10.9	5.89	5.68	5.65	5.65	92.0	15.2	7.08	5.66	5.65	1000	480	181	132	108
		RGN	10.9	5.65	5.65	5.65	5.65	92.0	5.65	5.65	5.65	5.65	1000	223	172	51.7	31.9
		CP-ALS	10.9	5.65	5.65	5.65	5.65	92.0	5.65	5.65	5.65	5.65	1000	81.7	34.5	33.3	30.6
	0.5	RGD	57.0	16.1	9.59	5.89	5.63	152	41.7	21.1	7.06	5.63	1000	528	177	104	76.1
		RGN	57.0	5.76	5.63	5.63	5.63	152	11.0	5.70	5.63	5.63	1000	310	157	54.4	34.2
		CP-ALS	57.0	5.92	5.64	5.63	5.63	152	6.41	5.64	5.63	5.63	1000	70.9	44.5	36.5	30.2
	0.75	RGD	138	34.4	25.9	16.5	9.23	133	66.0	46.2	24.9	11.8	1000	509	145	65.2	37.4
		RGN	138	15.1	10.3	6.33	5.41	133	50.2	21.1	8.11	5.47	1000	248	110	35.5	22.0
		CP-ALS	138	19.4	11.4	6.10	5.38	133	22.7	12.6	6.25	5.39	1000	45.8	32.5	26.2	21.6
3	0	RGD	33.5	26.4	25.2	24.1	23.3	37.8	33.2	28.7	23.5	21.4	1000	621	238	170	150
		RGN	33.5	25.4	24.8	23.3	23.4	37.8	33.2	27.8	22.4	21.3	1000	440	228	143	96.3
		CP-ALS	33.5	24.8	23.3	23.4	23.4	37.8	30.5	27.7	25.5	25.5	1000	155	140	82.8	75.8
	0.25	RGD	33.9	25.4	24.3	23.8	23.3	97.7	35.7	30.4	26.3	24.5	1000	535	202	133	85.1
		RGN	33.9	24.4	24.1	23.3	23.4	97.7	32.6	29.7	25.3	24.4	1000	294	182	102	61.3
		CP-ALS	33.9	24.0	23.3	23.4	23.4	97.7	29.7	27.3	26.5	26.6	1000	121	91.6	60.3	60.2
	0.5	RGD	65.8	36.5	32.9	30.5	29.7	154	54.7	41.6	35.9	35.5	1000	592	206	125	99.5
		RGN	65.8	32.5	30.9	30.3	29.9	154	37.6	36.5	35.5	35.4	1000	392	177	100	68.5
		CP-ALS	65.8	30.6	29.8	29.9	29.0	154	36.9	35.1	32.9	32.1	1000	122	77.9	53.4	41.6
	0.75	RGD	128	44.6	39.2	34.6	32.1	133	71.6	54.8	39.5	33.0	1000	611	191	105	68.8
		RGN	128	33.5	31.9	31.2	31.4	133	57.3	36.4	31.5	31.3	1000	397	147	98.5	64.3
		CP-ALS	128	35.9	32.6	31.6	31.6	133	37.8	33.0	31.4	31.5	1000	78.6	54.6	44.2	35.7

D Real Data Analysis: International Trade Tensor Regression

We evaluate the proposed methods on a tensor-on-tensor regression problem constructed from international trade data. Each observational unit is an exporter country. The lagged design uses 2018 covariates and trade-exposure weights to predict a 2019 tensor of export outcomes indexed by importer partner, product section, and response channel. The scientific question is whether a low-dimensional summary of bilateral trade costs, market size, and policy variables can predict this structured outcome. This setting is substantially more heterogeneous than the Gaussian designs used in the simulation study: trade outcomes are sparse, strongly right-skewed before transformation, and organized by both geography and product composition.

Data sources and tensor construction. The raw data are assembled from public international databases. Exporter–importer–HS6 trade values and quantities are downloaded from CEPII BACI HS17 [18]. Bilateral distance, geography, and trade-agreement covariates are from the CEPII Grav-

ity database [12]. Most-favored-nation (MFN) tariff schedules are obtained from World Bank WITS, which disseminates UNCTAD TRAINS/WTO tariff data. The logistics-friction proxy is based on the World Bank Logistics Performance Index. GDP, GDP per capita, and population are from the World Bank World Development Indicators.

We aggregate HS6 products to the 21 standard HS1 sections and retain the top 25 importer partners by 2018 import volume after requiring coverage for gravity, macroeconomic, tariff, and logistics inputs. For exporter m, the 2019 response tensor is

$$\mathcal{Y}_m \in \mathbb{R}^{25\times 21\times 8},$$

where the three modes correspond to the importer partner, the HS1 product section, and the response channel. The eight response channels are log export value, log quantity, log unit value, exporter share, partner market share, Balassa revealed comparative advantage, extensive margin, and intensive margin. The raw 2018 covariate tensor contains six channels: MFN tariff, logistics-friction proxy, partner log GDP, log bilateral distance, regional-trade-agreement scope, and partner world demand. Agreement scope is derived from CEPII's coverage field and is coded as 0 for no recorded agreement, 1 for goods only, and 2 for goods and services. For the tensor-on-tensor regression operator used here, these cell-level covariates are summarized for each exporter by a vector $\mathcal{X}_m \in \mathbb{R}^6$ with weights given by the square root of the exporter's 2018 trade value in each cell, while the full $25 \times 21 \times 8$ response tensor is retained. Under the input-first convention of Section 2, the resulting coefficient tensor has dimension

$$\mathcal{T}^\star \in \mathbb{R}^{6\times 25\times 21\times 8}.$$

We retain the 80 largest exporters in 2018 that meet the coverage requirements in Appendix E.3.

Methods and evaluation protocol. We compare eight estimators. In the CP format, these are RGD and RGN, CP-ALS [28] and RRR [35], all with CP rank 3. In other tensor formats, Tucker-RGD [37] and Tucker-RGN [36] use multilinear rank $(3, 3, 3, 3)$, TT-IHT [40] uses TT rank $(3, 3, 3)$, and HOPLS [56] extracts three latent blocks. These are rank-three specifications within each method's native format, not capacity-matched models. In the notation of Appendix B, one rank-one CP component has $\mathrm{df} = 1 + (25-1) + (21-1) + (8-1) + (6-1) = 57$, so the CP-rank-3 model has 171 intrinsic degrees of freedom, and the TT-rank-$(3, 3, 3)$ model, with the modes ordered as $(25, 21, 8, 6)$ in the TT-IHT implementation, has $25(3) + 3(21)(3) + 3(8)(3) + 3(6) - 3(3^2) = 327$. CP-ALS, RGD and RGN share the CPCA initialization within each fold. RGD uses the fixed step size $1/\lambda_{\max}(X^\top X/n)$ for the standardized design X, which equals σ_X^{-2} when $X^\top X/n = \sigma_X^2 I$, and Table 1 reports the fits after 30 iterations.

Prediction accuracy is evaluated by five-fold cross-validation over exporters, with folds stratified by quintiles of 2018 trade volume and shared by all methods. Table 1 reports the mean and standard error of the fold-specific test RMSE and test R^2, computed on the centered response scale.

Optimization paths. Table 10 reports the held-out test RMSE after 3, 5, 10, 30 and 300 iterations. HOPLS is reported after one, two and three latent blocks, because adding a block changes the fitted model. CP-ALS, RGN and Tucker-RGN are close to their limits after a few iterations. RGD and TT-IHT take steps set by the largest eigenvalue of the design Gram matrix, so they approach their limits slowly along the other directions. Their test RMSE is smallest at about 10 iterations and increases to the level of the other methods by 300 iterations. Their lower test RMSE in Table 1 therefore comes from stopping before convergence.

E Additional Details on Algorithms

In this section, we provide additional details on the implementation of the algorithms.

E.1 Local T-HOSVD Retraction

Lemma E.1 (Local T-HOSVD retraction)**.** *Let $\mathcal{T} \in \mathrm{Seg}$, let $\mathbb{T}_\mathcal{T}$ be its tangent space, and let $\xi \in \mathbb{T}_\mathcal{T}$. If*

$$\|\xi\|_{\mathrm{F}} < \frac{\|\mathcal{T}\|_{\mathrm{F}}}{2(\sqrt{d}+1)}, \tag{4}$$

Table 10: Mean held-out test RMSE over the five folds along the fitting paths. Panel A reports iterations, with the CP-format methods first, and Panel B reports HOPLS latent blocks.

	Panel A: iteration				
Method	3	5	10	30	300
CP-ALS	2.6763	2.6756	2.6743	2.6747	2.6749
RRR	2.6955	2.6867	2.6776	2.6763	2.6749
RGD	2.6688	2.6646	2.6613	2.6657	2.6746
RGN	2.6732	2.6736	2.6736	2.6743	2.6749
Tucker-RGD	2.7087	2.6970	2.6822	2.6748	2.6743
Tucker-RGN	2.6712	2.6710	2.6716	2.6746	2.6743
TT-IHT	2.6669	2.6653	2.6630	2.6663	2.6772
	Panel B: HOPLS latent blocks				
Method	1	2	3		
HOPLS	2.7093	2.6965	2.6795		

then

$$\mathcal{R}_{\mathcal{T}}(\xi) := \text{T-HOSVD}_{(1,\dots,1)}(\mathcal{T}+\xi)$$

is a nonzero rank-one tensor, and the displayed map is smooth jointly in $(\mathcal{T},\xi)$ *on the tangent-bundle neighborhood defined by* (4)*. Moreover, for every* $\zeta \in \mathbb{T}_{\mathcal{T}}$*,*

$$\mathcal{R}_{\mathcal{T}}(0) = \mathcal{T}, \qquad D\mathcal{R}_{\mathcal{T}}(0)[\zeta] = \zeta. \tag{5}$$

Furthermore,

$$\|\mathcal{R}_{\mathcal{T}}(\xi) - (\mathcal{T}+\xi)\|_{\mathrm{F}} \le \frac{\|\xi\|_{\mathrm{F}}^2}{\|\mathcal{T}\|_{\mathrm{F}}}. \tag{6}$$

Proof. Set $\lambda = \|\mathcal{T}\|_{\mathrm{F}}$ and $\mathcal{Z} = \mathcal{T}+\xi$. For each $k \in [d]$, Weyl's inequality gives, with singular values extended by zeros,

$$\sigma_1\{\text{mat}_k(\mathcal{Z})\} \ge \lambda - \|\xi\|_{\mathrm{F}}, \qquad \sigma_2\{\text{mat}_k(\mathcal{Z})\} \le \|\xi\|_{\mathrm{F}}.$$

Under (4), $\|\xi\|_{\mathrm{F}} < \lambda/2$, so the preceding display gives

$$\sigma_1\{\text{mat}_k(\mathcal{Z})\} \ge \lambda - \|\xi\|_{\mathrm{F}} > \|\xi\|_{\mathrm{F}} \ge \sigma_2\{\text{mat}_k(\mathcal{Z})\}, \qquad k \in [d].$$

Hence the projector $P_k(\mathcal{Z})$ onto the leading left singular vector of $\text{mat}_k(\mathcal{Z})$ is uniquely defined and smooth in $\mathcal{Z}$, and

$$\mathcal{R}_{\mathcal{T}}(\xi) = \mathcal{Z} \times_1 P_1(\mathcal{Z}) \times_2 \cdots \times_d P_d(\mathcal{Z})$$

is smooth jointly in $(\mathcal{T},\xi)$ on the displayed domain. The T-HOSVD quasi-best approximation bound [13, 29] yields

$$\|\mathcal{Z} - \mathcal{R}_{\mathcal{T}}(\xi)\|_{\mathrm{F}} \le \sqrt{d} \inf_{V \in \mathrm{Seg}} \|\mathcal{Z} - V\|_{\mathrm{F}} \le \sqrt{d}\|\xi\|_{\mathrm{F}}. \tag{7}$$

Consequently,

$$\|\mathcal{R}_{\mathcal{T}}(\xi) - \mathcal{T}\|_{\mathrm{F}} \le (\sqrt{d}+1)\|\xi\|_{\mathrm{F}} < \frac{\lambda}{2},$$

so $\mathcal{R}_{\mathcal{T}}(\xi) \neq 0$ and therefore belongs to Seg.

Write $\mathcal{T} = \lambda u_1 \otimes \cdots \otimes u_d$, where $\|u_k\| = 1$ for every $k \in [d]$. Since $\xi \in \mathbb{T}_{\mathcal{T}}$,

$$\mathcal{Z} = a\, u_1 \otimes \cdots \otimes u_d + \sum_{k=1}^{d} u_1 \otimes \cdots \otimes w_k \otimes \cdots \otimes u_d, \qquad w_k \perp u_k.$$

Set $\psi = \sum_{k=1}^{d} \|w_k\|^2$. Orthogonality gives $\|\xi\|_{\mathrm{F}}^2 = (a-\lambda)^2 + \psi$ and $a \ge \lambda - \|\xi\|_{\mathrm{F}} > 0$. For $w_k \neq 0$, the matrix $\text{mat}_k(\mathcal{Z})\text{mat}_k(\mathcal{Z})^\top$ is supported on $\mathrm{span}(u_k, w_k)$ and has representation

$$\begin{pmatrix} a^2 + \psi - \|w_k\|^2 & a\|w_k\| \\ a\|w_k\| & \|w_k\|^2 \end{pmatrix} \tag{8}$$

in the orthonormal basis $(u_k, w_k/\|w_k\|)$. Its trace is $a^2+\psi$, and its determinant is $\|w_k\|^2(\psi - \|w_k\|^2) \le \psi^2/4$. Consequently,

$$\sigma_1^2\{\mathrm{mat}_k(\mathcal{Z})\} \ge \frac{a^2+\psi+a\sqrt{a^2+2\psi}}{2} = \frac{(a+\sqrt{a^2+2\psi})^2}{4},$$

and

$$\|\mathcal{Z} - \mathcal{Z}\times_k P_k(\mathcal{Z})\|_{\mathrm{F}}^2 = \sigma_2^2\{\mathrm{mat}_k(\mathcal{Z})\} \le \frac{4\|w_k\|^2(\psi-\|w_k\|^2)}{(a+\sqrt{a^2+2\psi})^2}.$$

When $w_k = 0$, both sides of the last inequality are zero. The T-HOSVD error bound gives

$$\begin{aligned}\|\mathcal{Z} - \mathcal{R}_{\mathcal{T}}(\xi)\|_{\mathrm{F}}^2 &\le \sum_{k=1}^{d} \|\mathcal{Z} - \mathcal{Z}\times_k P_k(\mathcal{Z})\|_{\mathrm{F}}^2 \\ &\le \frac{4\psi^2}{(a+\sqrt{a^2+2\psi})^2} = \left(\sqrt{a^2+2\psi}-a\right)^2.\end{aligned}$$

Moreover,

$$\left(a+\frac{\|\xi\|_{\mathrm{F}}^2}{\lambda}\right)^2 - (a^2+2\psi) = 2\left(a-\lambda+\frac{\|\xi\|_{\mathrm{F}}^2}{2\lambda}\right)^2 + \frac{\|\xi\|_{\mathrm{F}}^4}{2\lambda^2} \ge 0.$$

Since $a>0$, these inequalities prove (6).

At $\xi = 0$, each $P_k(\mathcal{T})$ fixes the corresponding mode factor, so $\mathcal{R}_{\mathcal{T}}(0) = \mathcal{T}$. For every $\zeta \in \mathbb{T}_{\mathcal{T}}$ and all sufficiently small $|t|$, (6) gives

$$\|\mathcal{R}_{\mathcal{T}}(t\zeta) - (\mathcal{T}+t\zeta)\|_{\mathrm{F}} \le \frac{t^2\|\zeta\|_{\mathrm{F}}^2}{\|\mathcal{T}\|_{\mathrm{F}}},$$

which proves $D\mathcal{R}_{\mathcal{T}}(0)[\zeta] = \zeta$. □

E.2 Code Implementation Details

Exact structured T-HOSVD step. The mode Gram matrices in (8) reduce the T-HOSVD of $\mathcal{Z} = \mathcal{T}+\xi$ to at most d two-dimensional symmetric eigenproblems. When $w_k = 0$, the leading mode vector is u_k. Contracting $\mathcal{Z}$ with the resulting vectors gives the T-HOSVD weight. This evaluates the rank-$(1,\dots,1)$ T-HOSVD without forming $\mathcal{Z}$ or any full unfolding.

Tensor-on-tensor regression. Let $X \in \mathbb{R}^{n\times p_{\mathrm{in}}^*}$ have mth row $\mathrm{vec}(\mathcal{X}_m)^\top/\sqrt{n}$. For component i, let $u_{\mathrm{in},i}$ and $u_{\mathrm{out},i}$ be the Kronecker products of its input and output factors, let $V_{\mathrm{in},i} \in \mathbb{R}^{p_{\mathrm{in}}^*\times \mathrm{df}_{\mathrm{in}}}$ be an orthonormal input-tangent basis whose first column is $u_{\mathrm{in},i}$, and let $V_{\mathrm{out},i}^{\perp} \in \mathbb{R}^{p_{\mathrm{out}}^*\times(\mathrm{df}_{\mathrm{out}}-1)}$ have orthonormal columns spanning the output-tangent subspace orthogonal to $u_{\mathrm{out},i}$. The implementation stores the output modes first. Under this fixed mode permutation, the observation operator and its adjoint, tangent projection, and T-HOSVD step yield exactly the permuted RGD/RGN iterates. In these coordinates, the projected design is

$$\begin{aligned}\mathsf{D}_i &= \left[u_{\mathrm{out},i}\otimes(XV_{\mathrm{in},i}),\quad V_{\mathrm{out},i}^{\perp}\otimes(Xu_{\mathrm{in},i})\right], \\ \mathsf{D}_i^\top\mathsf{D}_i &= \mathrm{diag}\left\{(XV_{\mathrm{in},i})^\top XV_{\mathrm{in},i},\ \ \|Xu_{\mathrm{in},i}\|^2 I_{\mathrm{df}_{\mathrm{out}}-1}\right\}.\end{aligned}$$

In these coordinates, $\mathsf{D}_i^\top\mathsf{D}_i$ represents the restricted normal operator $\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}\big|_{\widehat{\mathbb{T}}_i}$. The implementation forms this Gram matrix and the corresponding right-hand side by modewise contractions, without constructing D_i or the ambient tangent basis. These identities are exact: fixed-step RGD uses the right-hand side, whereas RGN solves the $\mathrm{df}_{\mathrm{in}}\times\mathrm{df}_{\mathrm{in}}$ block and divides each output coordinate by $\|Xu_{\mathrm{in},i}\|^2$. The structured T-HOSVD step above is then applied directly in tangent coordinates.

For RGN, whenever the restricted normal operator is invertible, $\widehat{\mathcal{T}}_i \in \widehat{\mathbb{T}}_i$ implies

$$\operatorname*{argmin}_{Z\in\widehat{\mathbb{T}}_i}\frac{1}{2}\left\|\mathcal{A}(Z) - \left\{\mathcal{Y} - \sum_{j\ne i}\mathcal{A}(\widehat{\mathcal{T}}_j)\right\}\right\|_{\mathrm{F}}^2 = \widehat{\mathcal{T}}_i - \Delta_i,$$

where Δ_i is the RGN direction in Section 3.3 and Algorithm 2. Under (4), applying the structured T-HOSVD to this tangent-space least-squares solution is exactly $\mathcal{R}_{\widehat{\mathcal{T}}_i}(-\Delta_i)$.

Baselines. The competing methods are not part of `SegreCP` and are provided in the code repository. In regression, RRR is the reduced-rank regression of Lock [35], Tucker-RGD and Tucker-RGN are Riemannian gradient descent and Gauss-Newton on the manifold of fixed Tucker rank [36, 37], TT-IHT is iterative hard thresholding in the tensor-train format [40], and HOPLS is higher-order partial least squares [56]. In decomposition, Orth-ALS is orthogonalized ALS [42], ICO is iterative concurrent orthogonalization [20], NLS-GN and GN-CG are Gauss-Newton methods for the nonlinear least-squares formulation [44, 45], CP-ALS+ELS is CP-ALS with enhanced line search [41], and Hybrid AMDM moves from Mahalanobis distance minimization to ALS by inverting one fewer singular value of each factor every ten iterations [43]. Orth-ALS follows the MATLAB implementation of Sharan and Valiant [42], CP-ALS follows the `cp` function of the `rTensor` package [33], RRR uses the `MultiwayRegression` package of Lock [35], and Tucker-RGN follows the code of Luo and Zhang [36]. The other baselines are our R implementations of the cited algorithms.

Compute resources. All experiments were run in R on one machine with an Intel Xeon 2.1 GHz CPU and 8 GB of memory, with one core per fit.

E.3 Additional Details on Real Data Analysis

Raw inputs and preprocessing. The pipeline uses 2018 to construct the covariates and lagged trade-exposure weights and 2019 to construct the response tensor. HS6 products are mapped to the 21 HS1 sections. Trade values and quantities are summed within each exporter–importer–HS1 cell. MFN tariffs are aggregated to importer–HS1 cells using import-value-weighted averages with a simple-mean fallback, and 2018 world demand is computed as the log total imports of each importer in each HS1 section. The share and revealed-comparative-advantage response channels use totals from the full 2019 BACI cross-section as their denominators.

Sample construction. The partner mode consists of the top 25 importer partners by 2018 aggregate import volume subject to macroeconomic, tariff, logistics, and gravity-data availability. Exporters must trade positively with at least 13 selected partners and at least 11 HS1 sections and have complete required macroeconomic and selected-partner distance information. The top 80 eligible exporters by 2018 export value are retained. The final covariate and response tensors have dimensions $80 \times 25 \times 21 \times 6$ and $80 \times 25 \times 21 \times 8$.

Regression representation. For each exporter m, the raw 2018 covariate tensor records six channels over partner and HS1 section: MFN tariff, logistics-friction proxy, partner log GDP, log distance, regional-trade-agreement scope, and world demand. The agreement-scope channel is coded as 0 for no recorded agreement, 1 for goods only, and 2 for goods and services. Because the solver uses one input mode, the cell-level covariates are collapsed into a six-dimensional vector by averaging over partner–section cells with weights given by the square root of the exporter's 2018 trade value in each cell. Within each cross-validation split, feature standardization and response-channel centering are estimated from the training exporters and then applied to the held-out exporters. The response tensor keeps the full partner–section–channel resolution. The resulting tensor-on-tensor model is

$$\mathcal{Y}_m = \langle \mathcal{X}_m, \mathcal{T}^\star \rangle_{\mathrm{in}} + \mathcal{E}_m, \qquad \mathcal{X}_m \in \mathbb{R}^6, \qquad \mathcal{T}^\star \in \mathbb{R}^{6\times 25\times 21\times 8},$$

where the first mode of $\mathcal{T}^\star$ indexes input features and the remaining modes index response coordinates.

E.4 Tensor CP Decomposition

Initialization for CP decomposition. Our convergence results apply to any initialization satisfying the stated componentwise local condition. We use CPCA (factor construction: [20, Algorithm 3]) because Han and Zhang [20] establish componentwise recovery guarantees for noisy non-orthogonal CP decomposition under mild incoherence. In their numerical settings, CPCA also yields smaller initialization error than T-HOSVD [13]. With $\mathrm{vec}(\mathcal{E}) \sim \mathcal{N}(0, \sigma^2 I_{p^*})$, Theorem 5 of Han and Zhang [20] shows that, with high probability, the CPCA factors satisfy $\max_{i\in[r],k\in[d]} \|\widehat{u}_{k,i}^{\mathrm{cpca}} \widehat{u}_{k,i}^{\mathrm{cpca}\top} - u_{k,i} u_{k,i}^\top\| \le \tau$, where τ is given there. When $p_S = \sqrt{p^*}$ in Algorithm 3 and the factors are nearly orthogonal, τ is of order $\sigma (p^*)^{1/4} / \min_{i\in[r]}(\lambda_i - \lambda_{i+1})$, with $\lambda_{r+1} = 0$. Write the CPCA initialization as $\mathcal{T}_i^{(0)} = \langle \mathcal{Y}, \widehat{U}_i \rangle \widehat{U}_i$, where $\widehat{U}_i = \widehat{u}_{1,i}^{\mathrm{cpca}} \otimes \cdots \otimes \widehat{u}_{d,i}^{\mathrm{cpca}}$. If

$d^2(1+\kappa r)\tau \le c$, expanding $\langle \mathcal{Y}, \widehat{U}_i \rangle$ and bounding $\langle \mathcal{E}, \widehat{U}_i \rangle$ by Lemma G.2 give, with high probability,

$$\varepsilon^{(0)} \le \sqrt{2} d\tau + d\tau^2 + (r-1)\kappa(\eta + \sqrt{2}\tau)^d + C\frac{\sigma\sqrt{\mathrm{df}}}{\lambda_r}.$$

Hence CPCA satisfies the initialization condition of Corollaries 4.5 and 4.6 if, in addition, $d\kappa r(1+\kappa r)(\eta+\sqrt{2}\tau)^d \le c$ and $\sigma\sqrt{\mathrm{df}}/\lambda_r \le c/\{d(1+\kappa r)\}$, where $c > 0$ is a sufficiently small universal constant.

Algorithm 3 Composite PCA (CPCA) initialization used in the experiments [20]

Input: Noisy tensor $\mathcal{Y}$, target CP rank r, and an optional nonempty proper subset $S \subsetneq [d]$

1: **if** S is not supplied **then**
2: Choose S to minimize $|\prod_{k\in S} p_k - \sqrt{p^*}|$ over nonempty proper subsets of $[d]$, where $p^* = \prod_{k=1}^d p_k$.
3: **end if**
4: Set $S^c = [d] \setminus S$ and $p_S = \prod_{k\in S} p_k$, with $r \le \min\{p_S, p^*/p_S\}$.
5: Compute the top-r singular triplets $\{(\widehat{s}_i, \widehat{u}_i, \widehat{v}_i)\}_{i=1}^r$ of the $p_S \times (p^*/p_S)$ matricization $\mathrm{mat}_S(\mathcal{Y})$, whose rows and columns index the modes in S and S^c, respectively.
6: **for** $i = 1, \dots, r$ **do**
7: Refold $\widehat{u}_i$ over the modes in S and $\widehat{v}_i$ over the modes in S^c, and extract their leading modewise left singular vectors $\{\widehat{u}_{k,i}^{\mathrm{cpca}}\}_{k=1}^d$.
8: Choose their signs so that $\langle \widehat{u}_i, \bigotimes_{k\in S} \widehat{u}_{k,i}^{\mathrm{cpca}} \rangle \ge 0$ and $\langle \widehat{v}_i, \bigotimes_{k\in S^c} \widehat{u}_{k,i}^{\mathrm{cpca}} \rangle \ge 0$.
9: Set $\widetilde{\lambda}_i = \langle \mathcal{Y}, \bigotimes_{k=1}^d \widehat{u}_{k,i}^{\mathrm{cpca}} \rangle$. If $\widetilde{\lambda}_i < 0$, replace $\widehat{u}_{1,i}^{\mathrm{cpca}}$ by $-\widehat{u}_{1,i}^{\mathrm{cpca}}$. Set $\widehat{\lambda}_i^{\mathrm{cpca}} = |\widetilde{\lambda}_i|$.
10: **end for**

Output: $\{\widehat{\lambda}_i^{\mathrm{cpca}}, \widehat{u}_{k,i}^{\mathrm{cpca}}\}_{i\in[r], k\in[d]}$.

Algorithm 4 Riemannian Gradient Descent for CP Tensor Decomposition

Input: Noisy tensor $\mathcal{Y}$, target CP rank r, step size α_t, and initial rank-one tensors $\left\{\mathcal{T}_i^{(0)}\right\}_{i=1}^r$.

1: Compute the residual $\mathrm{res} \leftarrow \sum_{j=1}^r \mathcal{T}_j^{(0)} - \mathcal{Y}$.
2: **for** $t = 0, 1, \dots, t_{\max} - 1$ **do**
3: **for** $i = 1, \dots, r$ **do**
4: **(RGD Update)** Compute

$$\mathcal{T}_i^{(t+1)} = \mathcal{R}_{\mathcal{T}_i^{(t)}}\left(-\alpha_t \mathcal{P}_{\mathbb{T}_i^{(t)}} \mathrm{res}\right).$$

5: Update $\mathrm{res} \leftarrow \mathrm{res} + \mathcal{T}_i^{(t+1)} - \mathcal{T}_i^{(t)}$.
6: **end for**
7: **end for**

Output: $\left\{\mathcal{T}_i^{(t_{\max})}\right\}_{i=1}^r$.

Riemannian Gradient Descent for tensor decomposition. In the CP decomposition experiments, RGD uses the constant step size $\alpha_t \equiv 0.5$. It satisfies $0 < \alpha_t \le 1$ and $|1 - \alpha_t| \le 1/2$ in Corollary G.3, and the factor $1 - 3\alpha_t/4$ in that corollary equals $5/8$.

Riemannian Gauss-Newton for tensor decomposition. For tensor decomposition, Riemannian Gauss-Newton is Riemannian gradient descent with step size $\alpha_t \equiv 1$.

E.5 Tensor-on-Tensor Regression

Initialization for tensor-on-tensor regression. To estimate the low-rank tensor coefficient in a regression setting, we adopt an initialization strategy based on the adjoint operator of the linear map

Algorithm 5 Riemannian Gauss-Newton for CP Tensor Decomposition

Input: Noisy tensor $\mathcal{Y}$, target CP rank r, and initial rank-one tensors $\left\{\mathcal{T}_i^{(0)}\right\}_{i=1}^r$.

1: Compute the residual $\mathrm{res} \leftarrow \sum_{j=1}^r \mathcal{T}_j^{(0)} - \mathcal{Y}$.
2: **for** $t = 0, 1, \ldots, t_{\max} - 1$ **do**
3: **for** $i = 1, \ldots, r$ **do**
4: **(RGN Update)**

$$\mathcal{T}_i^{(t+1)} = \mathcal{R}_{\mathcal{T}_i^{(t)}}\left(-\mathcal{P}_{\mathbb{T}_i^{(t)}}\mathrm{res}\right).$$

5: Update $\mathrm{res} \leftarrow \mathrm{res} + \mathcal{T}_i^{(t+1)} - \mathcal{T}_i^{(t)}$.
6: **end for**
7: **end for**

Output: $\left\{\mathcal{T}_i^{(t_{\max})}\right\}_{i=1}^r$.

$\mathcal{A}$ induced by the covariates $\{\mathcal{X}_m\}_{m=1}^n$. Specifically, we use the scaled empirical cross-moment $\widetilde{\mathcal{T}} := \sigma_X^{-2}\mathcal{A}^*(\mathcal{Y})$, which satisfies $\mathbb{E}(\widetilde{\mathcal{T}}) = \mathcal{T}^\star$ under the isotropic design and independent mean-zero noise assumed above [19, 54]. We then apply CPCA [20] to $\widetilde{\mathcal{T}}$ to obtain the initial component weights and mode factors.

Algorithm 6 Initialization of Low-Rank Tensor-on-Tensor Regression

Input: Observations $\{\mathcal{X}_m, \mathcal{Y}_m\}_{m=1}^n$, design scale σ_X, and target CP rank r

1: Compute

$$\widetilde{\mathcal{T}} = \sigma_X^{-2}\mathcal{A}^*(\mathcal{Y}) = \frac{1}{n\sigma_X^2}\sum_{m=1}^n \mathcal{X}_m \otimes \mathcal{Y}_m$$

2: Compute $\{\widehat{\lambda}_i^{\mathrm{cpca}}, \widehat{u}_{k,i}^{\mathrm{cpca}}\}_{i\in[r],k\in[d]} \leftarrow \mathrm{CPCA}(\widetilde{\mathcal{T}})$, where CPCA is given in Algorithm 3 [20].
3: Set $\mathcal{T}_i^{(0)} = \widehat{\lambda}_i^{\mathrm{cpca}}\widehat{u}_{1,i}^{\mathrm{cpca}} \otimes \cdots \otimes \widehat{u}_{d,i}^{\mathrm{cpca}}$ for every $i \in [r]$.

Output: $\left\{\mathcal{T}_i^{(0)}\right\}_{i=1}^r$

Riemannian Gradient Descent for tensor-on-tensor regression. Using the normalized response $\mathcal{Y}$ and the operators $\mathcal{A}, \mathcal{A}^*$ defined in Section 2, RGD updates each rank-one component by tangent projection and retraction.

Riemannian Gauss-Newton for tensor-on-tensor regression. An orthonormal basis of $\mathbb{T}_i^{(t)}$ reduces the displayed normal equation to a $\mathrm{df} \times \mathrm{df}$ linear system, where $\mathrm{df} = 1 + \sum_{l=1}^d (p_l - 1)$.

F Proof of Main Theorems

In this section, we provide the proofs of error bounds incurred by Riemannian updates.

Proof of Theorem 4.1. Fix $i \in [r]$. The condition $\varepsilon \le 1/(4d)$ and Remark 4.3 permit Lemma H.1 to be applied to $\{\widehat{\mathcal{T}}_j\}_{j=1}^r$. Define

$$\widehat{G}_i = \mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\left\{\mathcal{A}\left(\sum_{j=1}^r \widehat{\mathcal{T}}_j\right) - \mathcal{Y}\right\}, \qquad \widehat{W}_i = \widehat{\mathcal{T}}_i - \alpha_t\widehat{G}_i, \qquad \widehat{\mathcal{T}}_i^{+} = \mathcal{R}_{\widehat{\mathcal{T}}_i}(-\alpha_t\widehat{G}_i).$$

Algorithm 7 Riemannian Gradient Descent for CP Tensor-on-Tensor Regression

Input: Normalized response $\mathcal{Y} \in \mathcal{H}$, linear operator $\mathcal{A}$ defined in Section 2, target CP rank r, step size α_t, and initial rank-one tensors $\left\{\mathcal{T}_i^{(0)}\right\}_{i=1}^r$.

Compute the residual $\mathrm{res} \leftarrow \mathcal{A}\big(\sum_{j=1}^r \mathcal{T}_j^{(0)}\big) - \mathcal{Y}$.
for $t = 0, 1, \ldots, t_{\max} - 1$ **do**
  for $i = 1, \ldots, r$ **do**
    (RGD Update) Compute

$$\mathcal{T}_i^{(t+1)} = \mathcal{R}_{\mathcal{T}_i^{(t)}}\left(-\alpha_t \mathcal{P}_{\mathbb{T}_i^{(t)}} \mathcal{A}^*\big(\mathrm{res}\big)\right).$$

    Update $\mathrm{res} \leftarrow \mathrm{res} + \mathcal{A}\big(\mathcal{T}_i^{(t+1)} - \mathcal{T}_i^{(t)}\big)$.
  end for
end for

Output: $\left\{\mathcal{T}_i^{(t_{\max})}\right\}_{i=1}^r$.

Algorithm 8 Riemannian Gauss-Newton for CP Tensor-on-Tensor Regression

Input: Normalized response $\mathcal{Y} \in \mathcal{H}$, linear operator $\mathcal{A}$ defined in Section 2, target CP rank r, and initial rank-one tensors $\left\{\mathcal{T}_i^{(0)}\right\}_{i=1}^r$.

Compute the residual $\mathrm{res} \leftarrow \mathcal{A}\big(\sum_{j=1}^r \mathcal{T}_j^{(0)}\big) - \mathcal{Y}$.
for $t = 0, 1, \ldots, t_{\max} - 1$ **do**
  for $i = 1, \ldots, r$ **do**
    (RGN Update) Whenever the restricted normal operator below is invertible, solve for the unique $\Delta_i^{(t)} \in \mathbb{T}_i^{(t)}$ satisfying

$$\left(\mathcal{P}_{\mathbb{T}_i^{(t)}} \mathcal{A}^* \mathcal{A} \mathcal{P}_{\mathbb{T}_i^{(t)}}\Big|_{\mathbb{T}_i^{(t)}}\right) \Delta_i^{(t)} = \mathcal{P}_{\mathbb{T}_i^{(t)}} \mathcal{A}^*\big(\mathrm{res}\big).$$

    Update $\mathcal{T}_i^{(t+1)} \leftarrow \mathcal{R}_{\mathcal{T}_i^{(t)}}\big(-\Delta_i^{(t)}\big)$.
    Update $\mathrm{res} \leftarrow \mathrm{res} + \mathcal{A}\big(\mathcal{T}_i^{(t+1)} - \mathcal{T}_i^{(t)}\big)$.
  end for
end for

Output: $\left\{\mathcal{T}_i^{(t_{\max})}\right\}_{i=1}^r$.

Since $\mathcal{Y} = \mathcal{A}(\sum_{j=1}^r \mathcal{T}_j) + \mathcal{E}$ and $\widehat{\xi}_i = -\alpha_t \widehat{G}_i$, (6) gives

$$\begin{aligned}
\|\widehat{\mathcal{T}}_i^+ - \mathcal{T}_i\|_{\mathrm{F}} &\le \|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}} + \frac{\|\widehat{\xi}_i\|_{\mathrm{F}}^2}{\|\widehat{\mathcal{T}}_i\|_{\mathrm{F}}} \\
&\le \left\| (\widehat{\mathcal{T}}_i - \mathcal{T}_i) - \alpha_t \mathcal{P}_{\widehat{\mathbb{T}}_i} \mathcal{A}^* \mathcal{A} \sum_{j=1}^r (\widehat{\mathcal{T}}_j - \mathcal{T}_j) \right\|_{\mathrm{F}} \\
&\quad + \alpha_t \|\mathcal{P}_{\widehat{\mathbb{T}}_i} \mathcal{A}^* \mathcal{E}\|_{\mathrm{F}} + \frac{\|\widehat{\xi}_i\|_{\mathrm{F}}^2}{\|\widehat{\mathcal{T}}_i\|_{\mathrm{F}}}. \qquad (9)
\end{aligned}$$

Since $\mathcal{T}_i \in \mathrm{Seg}$, the quasi-best approximation bound (7) gives $\|\widehat{\mathcal{T}}_i^{+} - \widehat{W}_i\|_{\mathrm{F}} \le \sqrt{d}\|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}}$, so

$$\begin{aligned}\|\widehat{\mathcal{T}}_i^{+} - \mathcal{T}_i\|_{\mathrm{F}} &\le (\sqrt{d}+1)\|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}} \\ &\le (\sqrt{d}+1)\left\|(\widehat{\mathcal{T}}_i - \mathcal{T}_i) - \alpha_t \mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\sum_{j=1}^{r}(\widehat{\mathcal{T}}_j - \mathcal{T}_j)\right\|_{\mathrm{F}} \\ &\quad + (\sqrt{d}+1)\alpha_t\|\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{E}\|_{\mathrm{F}}. \end{aligned} \tag{10}$$

Since $\mathcal{P}_{\widehat{\mathbb{T}}_i}^2 = \mathcal{P}_{\widehat{\mathbb{T}}_i}$ and $\mathcal{P}_{\widehat{\mathbb{T}}_j} + \mathcal{P}_{\widehat{\mathbb{T}}_j}^{\perp} = I$ for $j \in [r]$,

$$\begin{aligned}&(\widehat{\mathcal{T}}_i - \mathcal{T}_i) - \alpha_t \mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\sum_{j=1}^{r}(\widehat{\mathcal{T}}_j - \mathcal{T}_j) \\ &= \mathcal{P}_{\widehat{\mathbb{T}}_i}(I - \alpha_t\mathcal{A}^*\mathcal{A})\mathcal{P}_{\widehat{\mathbb{T}}_i}(\widehat{\mathcal{T}}_i - \mathcal{T}_i) + \mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}(\widehat{\mathcal{T}}_i - \mathcal{T}_i) - \alpha_t\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}(\widehat{\mathcal{T}}_i - \mathcal{T}_i) \\ &\quad - \alpha_t\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}\sum_{j\neq i}(\widehat{\mathcal{T}}_j - \mathcal{T}_j) - \alpha_t\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}\sum_{j\neq i}\mathcal{P}_{\widehat{\mathbb{T}}_j}(\widehat{\mathcal{T}}_j - \mathcal{T}_j) \\ &\quad - \alpha_t\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}\sum_{j\neq i}\mathcal{P}_{\widehat{\mathbb{T}}_j}^{\perp}(\widehat{\mathcal{T}}_j - \mathcal{T}_j),\end{aligned}$$

so the triangle inequality bounds the norm in (10) by $\mathrm{I} + \mathrm{II} + \alpha_t(\mathrm{III} + \mathrm{IV} + \mathrm{V} + \mathrm{VI})$, where

$$\begin{aligned}\mathrm{I} &= \left\|\mathcal{P}_{\widehat{\mathbb{T}}_i}(I - \alpha_t\mathcal{A}^*\mathcal{A})\mathcal{P}_{\widehat{\mathbb{T}}_i}(\widehat{\mathcal{T}}_i - \mathcal{T}_i)\right\|_{\mathrm{F}}, \\ \mathrm{II} &= \|\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}(\widehat{\mathcal{T}}_i - \mathcal{T}_i)\|_{\mathrm{F}}, \\ \mathrm{III} &= \|\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}(\widehat{\mathcal{T}}_i - \mathcal{T}_i)\|_{\mathrm{F}}, \\ \mathrm{IV} &= \left\|\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}\sum_{j\neq i}(\widehat{\mathcal{T}}_j - \mathcal{T}_j)\right\|_{\mathrm{F}}, \\ \mathrm{V} &= \left\|\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}\sum_{j\neq i}\mathcal{P}_{\widehat{\mathbb{T}}_j}(\widehat{\mathcal{T}}_j - \mathcal{T}_j)\right\|_{\mathrm{F}}, \\ \mathrm{VI} &= \left\|\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}\sum_{j\neq i}\mathcal{P}_{\widehat{\mathbb{T}}_j}^{\perp}(\widehat{\mathcal{T}}_j - \mathcal{T}_j)\right\|_{\mathrm{F}}.\end{aligned}$$

The first two terms satisfy

$$\mathrm{I} \le \|\mathcal{P}_{\widehat{\mathbb{T}}_i}(I - \alpha_t\mathcal{A}^*\mathcal{A})\mathcal{P}_{\widehat{\mathbb{T}}_i}\|\|\widehat{\mathcal{T}}_i - \mathcal{T}_i\|_{\mathrm{F}}, \tag{11}$$

$$\mathrm{II} = \|\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}\mathcal{T}_i\|_{\mathrm{F}} \le 3d\frac{\|\widehat{\mathcal{T}}_i - \mathcal{T}_i\|_{\mathrm{F}}^2}{\lambda_i}, \tag{12}$$

where $\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}\widehat{\mathcal{T}}_i = 0$ and (66) were used in (12).

Let $\widehat{M}_i = \mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}$. For every tensor W, with the Segre atomic norm $\|\cdot\|_{\mathrm{Seg},*}$ of Lemma H.1,

$$\|\widehat{M}_i W\|_{\mathrm{F}} \le \sup_{\substack{V\in\mathrm{Seg}\\ \|V\|_{\mathrm{F}}=1}} \|\widehat{M}_i V\|_{\mathrm{F}}\, \|W\|_{\mathrm{Seg},*}.$$

Consequently, since $\widehat{M}_i = \widehat{M}_i\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}$, (67) gives

$$\mathrm{III} \le 4d \sup_{\substack{V\in\mathrm{Seg}\\ \|V\|_{\mathrm{F}}=1}} \|\widehat{M}_i V\|_{\mathrm{F}}\frac{\|\widehat{\mathcal{T}}_i - \mathcal{T}_i\|_{\mathrm{F}}^2}{\lambda_i}, \tag{13}$$

$$\mathrm{VI} \le 4d \sup_{\substack{V\in\mathrm{Seg}\\ \|V\|_{\mathrm{F}}=1}} \|\widehat{M}_i V\|_{\mathrm{F}}\sum_{j\neq i}\frac{\|\widehat{\mathcal{T}}_j - \mathcal{T}_j\|_{\mathrm{F}}^2}{\lambda_j}. \tag{14}$$

After division by λ_i, (14) and $\lambda_j/\lambda_i \le \kappa$ imply

$$\frac{\mathrm{VI}}{\lambda_i} \le 4d(r-1)\kappa \sup_{\substack{V\in\mathrm{Seg}\\ \|V\|_{\mathrm{F}}=1}} \|\widehat{M}_i V\|_{\mathrm{F}}\, \varepsilon^2. \tag{15}$$

For the remaining cross-component terms, (68) yields

$$\frac{\mathrm{IV}}{\lambda_i} \le 4d(r-1)\kappa \|\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}\| \{\eta^{d-2}+\varepsilon\}\varepsilon, \tag{16}$$

$$\frac{\mathrm{V}}{\lambda_i} \le (r-1)\kappa \max_{j\ne i} \|\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}\mathcal{P}_{\widehat{\mathbb{T}}_j}\|\, \varepsilon. \tag{17}$$

Since $\widehat{W}_i - \mathcal{T}_i = (\widehat{\mathcal{T}}_i - \mathcal{T}_i) - \alpha_t \mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\sum_{j=1}^r(\widehat{\mathcal{T}}_j - \mathcal{T}_j) + \alpha_t \mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{E}$, the decomposition above gives $\|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}} \le \mathrm{I}+\mathrm{II}+\alpha_t(\mathrm{III}+\mathrm{IV}+\mathrm{V}+\mathrm{VI}) + \alpha_t\|\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{E}\|_{\mathrm{F}}$. Dividing by λ_i and inserting (11)–(13) and (15)–(17) give

$$\begin{aligned}
\frac{\|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i} \le{}& \big\|\mathcal{P}_{\widehat{\mathbb{T}}_i}(I-\alpha_t\mathcal{A}^*\mathcal{A})\mathcal{P}_{\widehat{\mathbb{T}}_i}\big\|\,\varepsilon + (r-1)\alpha_t\kappa \max_{j\ne i}\big\|\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}\mathcal{P}_{\widehat{\mathbb{T}}_j}\big\|\,\varepsilon \\
&+ \Big[3d + 4d\{1+(r-1)\kappa\}\,\alpha_t \sup_{\substack{V\in\mathrm{Seg}\\ \|V\|_{\mathrm{F}}=1}} \|\widehat{M}_i V\|_{\mathrm{F}}\Big]\varepsilon^2 \\
&+ 4d(r-1)\alpha_t\kappa\big\|\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}\big\|\{\eta^{d-2}+\varepsilon\}\varepsilon + \alpha_t\frac{\|\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{E}\|_{\mathrm{F}}}{\lambda_i}.
\end{aligned} \tag{18}$$

Since $r-1 \le r$, $3d \le 4d$ and $1+(r-1)\kappa \le r\kappa$, bounding each i-dependent norm in (18) by its maximum over $i \in [r]$ and adding the remainder in (9) prove the theorem with $C_1 = 1$ and $C_2 = 4d$. By (10), $\|\widehat{\mathcal{T}}_i^{+} - \mathcal{T}_i\|_{\mathrm{F}} \le (\sqrt{d}+1)\|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}}$, so the same argument gives the bound without the retraction remainder, with $C_1 = \sqrt{d}+1$ and $C_2 = 4d(\sqrt{d}+1)$. □

Proof of Theorem 4.2. Fix $i \in [r]$. The condition $\varepsilon \le 1/(4d)$ and Remark 4.3 permit Lemma H.1 to be applied to $\{\widehat{\mathcal{T}}_j\}_{j=1}^r$.

The current RGN direction and updated component are

$$\widehat{\Delta}_i = \Big(\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}\big|_{\widehat{\mathbb{T}}_i}\Big)^{-1}\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\left[\mathcal{A}\left\{\sum_{j=1}^r(\widehat{\mathcal{T}}_j - \mathcal{T}_j)\right\} - \mathcal{E}\right],$$

$$\widehat{\mathcal{T}}_i^{+} = \mathcal{R}_{\widehat{\mathcal{T}}_i}(-\widehat{\Delta}_i).$$

Define

$$\widehat{\mathsf{B}}_i = \Big(\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}\big|_{\widehat{\mathbb{T}}_i}\Big)^{-1}\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A},$$

$$\widehat{\mathsf{H}}_i = \Big(\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}\big|_{\widehat{\mathbb{T}}_i}\Big)^{-1}\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*.$$

The inverse on $\widehat{\mathbb{T}}_i$ gives

$$\widehat{\mathsf{B}}_i\mathcal{P}_{\widehat{\mathbb{T}}_i} = \mathcal{P}_{\widehat{\mathbb{T}}_i}, \qquad \widehat{\mathsf{B}}_i\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp} = \widehat{\mathsf{C}}_i, \qquad \widehat{\mathsf{C}}_i\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp} = \widehat{\mathsf{C}}_i. \tag{19}$$

Since $\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}\widehat{\mathcal{T}}_i = 0$,

$$\begin{aligned}
\widehat{\mathcal{T}}_i - \widehat{\Delta}_i - \mathcal{T}_i &= (\widehat{\mathcal{T}}_i - \mathcal{T}_i) - \widehat{\mathsf{B}}_i\sum_{j=1}^r(\widehat{\mathcal{T}}_j - \mathcal{T}_j) + \widehat{\mathsf{H}}_i\mathcal{E} \\
&= \mathcal{P}_{\widehat{\mathbb{T}}_i}\sum_{j\ne i}(\mathcal{T}_j - \widehat{\mathcal{T}}_j) + \widehat{\mathsf{C}}_i\mathcal{T}_i \\
&\quad + \widehat{\mathsf{C}}_i\sum_{j\ne i}\mathcal{P}_{\widehat{\mathbb{T}}_j}(\mathcal{T}_j - \widehat{\mathcal{T}}_j) + \widehat{\mathsf{C}}_i\sum_{j\ne i}\mathcal{P}_{\widehat{\mathbb{T}}_j}^{\perp}(\mathcal{T}_j - \widehat{\mathcal{T}}_j) \\
&\quad - \mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}\mathcal{T}_i + \widehat{\mathsf{H}}_i\mathcal{E}.
\end{aligned} \tag{20}$$

Let $\widehat{W}_i = \widehat{\mathcal{T}}_i - \widehat{\Delta}_i$. Since $\widehat{\xi}_i = -\widehat{\Delta}_i \in \widehat{\mathbb{T}}_i$, (6) gives

$$\|\widehat{\mathcal{T}}_i^{+} - \mathcal{T}_i\|_{\mathrm{F}} \le \|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}} + \frac{\|\widehat{\xi}_i\|_{\mathrm{F}}^2}{\|\widehat{\mathcal{T}}_i\|_{\mathrm{F}}}. \tag{21}$$

Since $\mathcal{T}_i \in \mathrm{Seg}$, (7) also yields

$$\begin{aligned}\|\widehat{\mathcal{T}}_i^{+} - \mathcal{T}_i\|_{\mathrm{F}} &\le \|\widehat{\mathcal{T}}_i^{+} - \widehat{W}_i\|_{\mathrm{F}} + \|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}} \\ &\le (\sqrt{d}+1)\|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}}.\end{aligned} \tag{22}$$

Let $\mathrm{I}, \dots, \mathrm{VI}$ denote the Frobenius norms of the six terms on the right-hand side of (20), in their displayed order.

By (68),

$$\frac{\mathrm{I}}{\lambda_i} \le 4d(r-1)\kappa\{\eta^{d-2} + \varepsilon\}\varepsilon. \tag{23}$$

For every tensor W, with the Segre atomic norm of Lemma H.1, $\|\widehat{\mathsf{C}}_i W\|_{\mathrm{F}} \le \|\widehat{\mathsf{C}}_i\|_{\mathrm{Seg}\to\mathrm{F}}\|W\|_{\mathrm{Seg},*}$. For II, (19), $\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}\widehat{\mathcal{T}}_i = 0$, and (67) give

$$\begin{aligned}\mathrm{II} &= \left\|\widehat{\mathsf{C}}_i \mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}(\mathcal{T}_i - \widehat{\mathcal{T}}_i)\right\|_{\mathrm{F}} \\ &\le 4d\|\widehat{\mathsf{C}}_i\|_{\mathrm{Seg}\to\mathrm{F}} \frac{\|\widehat{\mathcal{T}}_i - \mathcal{T}_i\|_{\mathrm{F}}^2}{\lambda_i}.\end{aligned} \tag{24}$$

Since $\|\widehat{\mathcal{T}}_j - \mathcal{T}_j\|_{\mathrm{F}} \le \varepsilon\lambda_j \le \kappa\varepsilon\lambda_i$, we have

$$\frac{\mathrm{III}}{\lambda_i} \le (r-1)\kappa \max_{j\neq i} \|\widehat{\mathsf{C}}_i \mathcal{P}_{\widehat{\mathbb{T}}_j}\|\,\varepsilon. \tag{25}$$

Applying (67) to each $\mathcal{P}_{\widehat{\mathbb{T}}_j}^{\perp}(\mathcal{T}_j - \widehat{\mathcal{T}}_j)$ in IV gives

$$\frac{\mathrm{IV}}{\lambda_i} \le 4d(r-1)\kappa\|\widehat{\mathsf{C}}_i\|_{\mathrm{Seg}\to\mathrm{F}}\,\varepsilon^2. \tag{26}$$

Finally, (66) and the definition of $\widehat{\mathsf{H}}_i$ imply

$$\frac{\mathrm{V}}{\lambda_i} \le 3d\varepsilon^2, \tag{27}$$

$$\frac{\mathrm{VI}}{\lambda_i} = \frac{1}{\lambda_i}\left\|\left(\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}\big|_{\widehat{\mathbb{T}}_i}\right)^{-1}\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{E}\right\|_{\mathrm{F}}. \tag{28}$$

Since $\widehat{W}_i - \mathcal{T}_i$ is the left-hand side of (20), $\|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}} \le \mathrm{I} + \cdots + \mathrm{VI}$, and (23)–(28) give

$$\begin{aligned}\frac{\|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i} &\le (r-1)\kappa \max_{j\neq i} \|\widehat{\mathsf{C}}_i \mathcal{P}_{\widehat{\mathbb{T}}_j}\|\,\varepsilon + \frac{\|\widehat{\mathsf{H}}_i\mathcal{E}\|_{\mathrm{F}}}{\lambda_i} + 4d(r-1)\kappa\{\eta^{d-2}+\varepsilon\}\varepsilon \\ &\quad + \left[3d + 4d\{1+(r-1)\kappa\}\|\widehat{\mathsf{C}}_i\|_{\mathrm{Seg}\to\mathrm{F}}\right]\varepsilon^2.\end{aligned} \tag{29}$$

Since $r-1 \le r$, $3d \le 4d$ and $1+(r-1)\kappa \le 1+r\kappa$, bounding each i-dependent norm in (29) by its maximum over $i \in [r]$ and adding the remainder in (21) prove the theorem with $C_1 = 1$ and $C_2 = 4d$. By (22), the same argument gives the bound without the retraction remainder, with $C_1 = \sqrt{d}+1$ and $C_2 = 4d(\sqrt{d}+1)$.

□

G Proof of Corollaries

This appendix proves variable-step versions of the RGD corollaries and the RGN corollaries in Section 4.2. Throughout, c and C denote universal constants whose values may change from line to line, and c_0, C_0, C_3, C_4 denote universal constants fixed within each proof. For a current collection $\{\widehat{\mathcal{T}}_j\}_{j=1}^r \subset \mathrm{Seg}$, $\widehat{\mathbb{T}}_j$ is the tangent space at $\widehat{\mathcal{T}}_j$, $D_j = \widehat{\mathcal{T}}_j - \mathcal{T}_j$, $\widehat{\xi}_i \in \widehat{\mathbb{T}}_i$ is the tangent increment of the method under consideration, $\widehat{W}_i = \widehat{\mathcal{T}}_i + \widehat{\xi}_i$, and $\widehat{\mathcal{T}}_i^{+} = \mathcal{R}_{\widehat{\mathcal{T}}_i}(\widehat{\xi}_i)$ is the updated component.

Lemma G.1 (Composition of sequential component updates)**.** *Let $\Phi_t : [0, R] \to [0, \infty)$ be nondecreasing. Suppose that, for every current collection $\{\widehat{\mathcal{T}}_j\}_{j=1}^r$ satisfying*

$$\max_{j\in[r]} \frac{\|\widehat{\mathcal{T}}_j - \mathcal{T}_j\|_{\mathrm{F}}}{\lambda_j} \le R,$$

the update of component i satisfies

$$\frac{\|\widehat{\mathcal{T}}_i^{+} - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i} \le \Phi_t\left(\max_{j\in[r]} \frac{\|\widehat{\mathcal{T}}_j - \mathcal{T}_j\|_{\mathrm{F}}}{\lambda_j}\right), \qquad i \in [r].$$

Set $\varepsilon^{(t)} = \max_{j\in[r]} \|\mathcal{T}_j^{(t)} - \mathcal{T}_j\|_{\mathrm{F}}/\lambda_j$. If $\varepsilon^{(t)} \le \nu \le R$ and $\Phi_t(\nu) \le \nu$, then every intermediate collection generated by the sequential updates $i = 1, \dots, r$ has relative error at most ν, and

$$\varepsilon^{(t+1)} \le \Phi_t(\nu). \tag{30}$$

Proof. For $s = 0, \dots, r$, let

$$\mathcal{T}_j^{(t,s)} = \begin{cases} \mathcal{T}_j^{(t+1)}, & j \le s, \\ \mathcal{T}_j^{(t)}, & j > s, \end{cases} \qquad \varepsilon^{(t,s)} = \max_{j\in[r]} \frac{\|\mathcal{T}_j^{(t,s)} - \mathcal{T}_j\|_{\mathrm{F}}}{\lambda_j},$$

so that $\varepsilon^{(t,0)} = \varepsilon^{(t)}$ and $\varepsilon^{(t,r)} = \varepsilon^{(t+1)}$. The s-th update changes component s only. If $\varepsilon^{(t,s-1)} \le \nu$, the hypothesis and the monotonicity of Φ_t give

$$\frac{\|\mathcal{T}_s^{(t+1)} - \mathcal{T}_s\|_{\mathrm{F}}}{\lambda_s} \le \Phi_t(\varepsilon^{(t,s-1)}) \le \Phi_t(\nu) \le \nu, \qquad \varepsilon^{(t,s)} \le \nu.$$

Induction over $s = 1, \dots, r$ yields $\max_{0\le s\le r} \varepsilon^{(t,s)} \le \nu$ and $\|\mathcal{T}_s^{(t+1)} - \mathcal{T}_s\|_{\mathrm{F}}/\lambda_s \le \Phi_t(\nu)$ for every $s \in [r]$, which is (30). □

Lemma G.2 (Local Gaussian tangent projection)**.** *Let $\mathcal{T}_i = \lambda_i u_{1,i} \otimes \cdots \otimes u_{d,i}$, $i \in [r]$, be the components of the fixed rank-r tensor in* (1)*, and let $\mathrm{vec}(\mathcal{E}) \sim \mathcal{N}(0, \Sigma_{p^*})$ with $\Sigma_{p^*} \preceq \overline{\sigma}^2 I_{p^*}$. There exist universal constants $c, C > 0$ such that, for every $0 < \nu \le 1/(2\sqrt{d})$, with probability at least $1 - 2\exp(-c\,\mathrm{df})$,*

$$\max_{i\in[r]} \sup_{\substack{\widetilde{\mathcal{T}}\in\mathrm{Seg} \\ \|\widetilde{\mathcal{T}}-\mathcal{T}_i\|_{\mathrm{F}}\le\nu\lambda_i}} \|\mathcal{P}_{\mathbb{T}_{\widetilde{\mathcal{T}}}}\mathcal{E}\|_{\mathrm{F}} \le C\overline{\sigma}\sqrt{\mathrm{df}}. \tag{31}$$

Proof. Fix $i \in [r]$ and $\widetilde{\mathcal{T}} \in \mathrm{Seg}$ with $\|\widetilde{\mathcal{T}} - \mathcal{T}_i\|_{\mathrm{F}} \le \nu\lambda_i$. The hypothesis $\nu \le 1/(2\sqrt{d})$ gives $\nu \le 1/2$ and $\nu\sqrt{d} \le 1$. Write $\mathcal{T}_i = \lambda_i U_i$ and $\widetilde{\mathcal{T}} = \widetilde{\lambda}\widetilde{U}$ with $U_i = u_{1,i} \otimes \cdots \otimes u_{d,i}$, $\widetilde{U} = \widetilde{u}_1 \otimes \cdots \otimes \widetilde{u}_d$, unit factor vectors, and $\widetilde{\lambda} > 0$. By polarization,

$$2\langle \widetilde{\mathcal{T}}, \mathcal{T}_i\rangle = \|\widetilde{\mathcal{T}}\|_{\mathrm{F}}^2 + \lambda_i^2 - \|\widetilde{\mathcal{T}} - \mathcal{T}_i\|_{\mathrm{F}}^2 \ge 2(1-\nu)\lambda_i^2 > 0,$$

so the factor signs can be chosen with $\langle u_{\ell,i}, \widetilde{u}_\ell\rangle = \cos\theta_\ell \ge 0$ for angles $\theta_\ell \in [0, \pi/2]$, $\ell \in [d]$. Then

$$\nu^2 \ge \inf_{a\in\mathbb{R}} \|a\widetilde{U} - U_i\|_{\mathrm{F}}^2 = 1 - \langle \widetilde{U}, U_i\rangle^2 = 1 - \prod_{\ell=1}^d \cos^2\theta_\ell,$$

$$\sum_{\ell=1}^d \|\mathcal{P}_{u_{\ell,i}}^{\perp}\widetilde{u}_\ell\|_2^2 = \sum_{\ell=1}^d \sin^2\theta_\ell \le -\log\left(\prod_{\ell=1}^d \cos^2\theta_\ell\right) \le -\log(1-\nu^2) \le 2\nu^2. \tag{32}$$

With $K_\ell = \widetilde{u}_\ell u_{\ell,i}^{\top} - u_{\ell,i}\widetilde{u}_\ell^{\top}$, let

$$Q_\ell = I + K_\ell + \frac{K_\ell^2}{1 + u_{\ell,i}^{\top}\widetilde{u}_\ell}, \qquad Q = Q_1 \otimes \cdots \otimes Q_d.$$

Each Q_ℓ is the orthogonal rotation of $\mathrm{span}\{u_{\ell,i}, \widetilde{u}_\ell\}$ that maps $u_{\ell,i}$ to $\widetilde{u}_\ell$, extended by the identity on the orthogonal complement, and the denominator is at least one. Hence

$$\mathbb{T}_{\widetilde{\mathcal{T}}} = Q\mathbb{T}_{\mathcal{T}_i}. \tag{33}$$

Step 1 (Entropy bound). Since $\|\widetilde{u}_\ell\|_2 = 1$ and $\cos\theta_\ell \geq 0$,

$$\widetilde{u}_\ell = \left(1 - \|\mathcal{P}^{\perp}_{u_{\ell,i}} \widetilde{u}_\ell\|_2^2\right)^{1/2} u_{\ell,i} + \mathcal{P}^{\perp}_{u_{\ell,i}} \widetilde{u}_\ell.$$

For two factor tuples $(\widetilde{u}_\ell)_{\ell=1}^d$ and $(\widetilde{u}'_\ell)_{\ell=1}^d$ of this form satisfying (32),

$$\sum_{\ell=1}^d \|\widetilde{u}_\ell - \widetilde{u}'_\ell\|_2^2 \leq \frac{1}{1-2\nu^2} \sum_{\ell=1}^d \|\mathcal{P}^{\perp}_{u_{\ell,i}}(\widetilde{u}_\ell - \widetilde{u}'_\ell)\|_2^2 \leq 2 \sum_{\ell=1}^d \|\mathcal{P}^{\perp}_{u_{\ell,i}}(\widetilde{u}_\ell - \widetilde{u}'_\ell)\|_2^2.$$

The tuples $(\mathcal{P}^{\perp}_{u_{\ell,i}} \widetilde{u}_\ell)_{\ell=1}^d$ lie in the Euclidean ball of radius $\sqrt{2}\nu$ of the $(\mathrm{df}-1)$-dimensional space $\prod_{\ell=1}^d u_{\ell,i}^{\perp}$. Equip the factor tuples of the tensors $\widetilde{\mathcal{T}} \in \mathrm{Seg}$ satisfying $\|\widetilde{\mathcal{T}} - \mathcal{T}_i\|_{\mathrm{F}} \leq \nu\lambda_i$ with the distance $(\sum_{\ell=1}^d \|\widetilde{u}_\ell - \widetilde{u}'_\ell\|_2^2)^{1/2}$. By the preceding display, an ϵ-separated set of factor tuples has $\epsilon/\sqrt{2}$-separated projections, so it has at most $(1 + 4\nu/\epsilon)^{\mathrm{df}-1}$ elements by the volume comparison in the proof of Vershynin [51, Proposition 4.2.12], and a maximal ϵ-separated set is an ϵ-net [51, Lemma 4.2.6].

Let Q'_ℓ and $Q' = \bigotimes_{\ell=1}^d Q'_\ell$ be defined from $(\widetilde{u}'_\ell)_{\ell=1}^d$ as Q_ℓ and Q are defined from $(\widetilde{u}_\ell)_{\ell=1}^d$. Since the denominators are at least one and $\|K_\ell\| \leq 1$, and since the factors Q_ℓ, Q'_ℓ are orthogonal,

$$\|Q_\ell - Q'_\ell\| \leq 3\|(\widetilde{u}_\ell - \widetilde{u}'_\ell)u_{\ell,i}^\top - u_{\ell,i}(\widetilde{u}_\ell - \widetilde{u}'_\ell)^\top\| + \|\widetilde{u}_\ell - \widetilde{u}'_\ell\|_2 \leq 7\|\widetilde{u}_\ell - \widetilde{u}'_\ell\|_2,$$

$$\|Q - Q'\| \leq \sum_{\ell=1}^d \|Q_\ell - Q'_\ell\| \leq 7\sqrt{d}\left(\sum_{\ell=1}^d \|\widetilde{u}_\ell - \widetilde{u}'_\ell\|_2^2\right)^{1/2}.$$

Consequently, for $Z, Z' \in \mathbb{T}_{\mathcal{T}_i}$ with $\|Z\|_{\mathrm{F}} = \|Z'\|_{\mathrm{F}} = 1$,

$$\|QZ - Q'Z'\|_{\mathrm{F}} \leq \|Z - Z'\|_{\mathrm{F}} + 7\sqrt{d}\left(\sum_{\ell=1}^d \|\widetilde{u}_\ell - \widetilde{u}'_\ell\|_2^2\right)^{1/2}.$$

By (33), every unit vector of $\mathbb{T}_{\widetilde{\mathcal{T}}}$ equals QZ for a unit vector $Z \in \mathbb{T}_{\mathcal{T}_i}$. Combining an $\epsilon/2$-net of the unit sphere of the df-dimensional space $\mathbb{T}_{\mathcal{T}_i}$, of cardinality at most $(6/\epsilon)^{\mathrm{df}}$ [51, Corollary 4.2.13], with an $\epsilon/(14\sqrt{d})$-net of the factor tuples gives, for $0 < \epsilon < 1$ and $\nu\sqrt{d} \leq 1$,

$$\log N\left(\epsilon, \bigcup_{\substack{\widetilde{\mathcal{T}} \in \mathrm{Seg} \\ \|\widetilde{\mathcal{T}} - \mathcal{T}_i\|_{\mathrm{F}} \leq \nu\lambda_i}} \{Z \in \mathbb{T}_{\widetilde{\mathcal{T}}} : \|Z\|_{\mathrm{F}} = 1\}, \|\cdot\|_{\mathrm{F}}\right) \leq \mathrm{df} \log\frac{6}{\epsilon} + (\mathrm{df}-1)\log\left(1 + \frac{56\nu\sqrt{d}}{\epsilon}\right)$$

$$\leq C\,\mathrm{df}\log\frac{C}{\epsilon}. \tag{34}$$

Since $\log r \leq \mathrm{df} - 1$ (Appendix A), the union of these sets over $i \in [r]$ satisfies (34) with a larger C.

Step 2 (Chaining and concentration). Let Z range over the union over $i \in [r]$ of the sets in (34). The covariance assumption gives

$$\mathbb{E}\langle Z - Z', \mathcal{E}\rangle^2 = \mathrm{vec}(Z - Z')^\top \Sigma_{p^*} \mathrm{vec}(Z - Z') \leq \overline{\sigma}^2 \|Z - Z'\|_{\mathrm{F}}^2.$$

The index set is symmetric with Frobenius diameter at most two, so $\sup_Z |\langle Z, \mathcal{E}\rangle| = \sup_Z \langle Z, \mathcal{E}\rangle$, and the increments $\langle Z - Z', \mathcal{E}\rangle$ are centered Gaussian. Dudley's inequality [51, Theorem 8.1.3] and (34) give

$$\mathbb{E}\sup_Z |\langle Z, \mathcal{E}\rangle| \leq C\overline{\sigma}\int_0^2 \sqrt{\mathrm{df}\log(1 + C/\epsilon)}\, d\epsilon \leq C\overline{\sigma}\sqrt{\mathrm{df}}.$$

Write $\mathrm{vec}(\mathcal{E}) = \Sigma_{p^*}^{1/2} g$ with $g \sim \mathcal{N}(0, I_{p^*})$. The map $g \mapsto \sup_Z |\langle \mathrm{vec}(Z), \Sigma_{p^*}^{1/2} g\rangle|$ is $\overline{\sigma}$-Lipschitz because $\|\Sigma_{p^*}^{1/2}\| \leq \overline{\sigma}$, so Gaussian concentration [51, Theorem 5.2.2] gives

$$\mathbb{P}\left\{\sup_Z |\langle Z, \mathcal{E}\rangle| > C\overline{\sigma}\sqrt{\mathrm{df}}\right\} \leq 2\exp(-c\,\mathrm{df}).$$

Finally, $\|\mathcal{P}_{\mathbb{T}_{\widetilde{\mathcal{T}}}}\mathcal{E}\|_{\mathrm{F}} = \sup\{|\langle Z, \mathcal{E}\rangle| : Z \in \mathbb{T}_{\widetilde{\mathcal{T}}},\ \|Z\|_{\mathrm{F}} = 1\}$, which proves (31). □

Corollary G.3 (Convergence of RGD for CP tensor decomposition)**.** *Let* $\varepsilon^{(t)} = \max_{i\in[r]} \|\mathcal{T}_i^{(t)} - \mathcal{T}_i\|_{\mathrm{F}}/\lambda_i$*. There exist universal constants* $c, C > 0$ *such that the following holds. Suppose, for every* $t \geq 0$*,*

$$0 < \alpha_t \leq 1, \qquad |1-\alpha_t| \leq \frac{1}{2}, \qquad \alpha_t \eta^{d-2} \leq \frac{c}{d\kappa r},$$

and

$$\varepsilon^{(0)} \leq \frac{c}{d(1+\kappa r)}, \qquad \frac{\overline{\sigma}\sqrt{\mathrm{df}}}{\lambda_r} \leq \frac{c}{d(1+\kappa r)}.$$

Then, with probability at least $1 - 2\exp(-c\,\mathrm{df})$*, simultaneously for all* $t \geq 0$*,*

$$\varepsilon^{(t)} \leq \left\{\prod_{s=0}^{t-1}\left(1 - \frac{3\alpha_s}{4}\right)\right\}\varepsilon^{(0)} + C\frac{\overline{\sigma}\sqrt{\mathrm{df}}}{\lambda_r}.$$

Proof. Let

$$R = \frac{c_0}{d(1+\kappa r)}, \qquad b = C_0\frac{\overline{\sigma}\sqrt{\mathrm{df}}}{\lambda_r},$$

where C_0 is eight times the constant C in (31) and c_0 is so small that Lemma G.2 applies with $\nu = R$ and

$$R \leq \min\left\{\frac{1}{4d}, \frac{1}{24}\right\}, \qquad 2R < \frac{1}{2(\sqrt{d}+1)}.$$

The constant c in the corollary is then chosen so small that $\varepsilon^{(0)} \leq R$, $b \leq R/4$, and the inequalities below hold. All statements are on the event of Lemma G.2.

Step 1 (One component update). Let $\{\widehat{\mathcal{T}}_j\}_{j=1}^r \subset \mathrm{Seg}$ satisfy $\varepsilon = \max_j \|D_j\|_{\mathrm{F}}/\lambda_j \leq R$. Since $\mathcal{A}^*\mathcal{A} = I$ and $\mathcal{Y} - \sum_{j=1}^r \mathcal{T}_j = \mathcal{E}$,

$$\widehat{\xi}_i = -\alpha_t \mathcal{P}_{\widehat{\mathbb{T}}_i}\left(\sum_{j=1}^r D_j - \mathcal{E}\right),$$
$$\widehat{W}_i - \mathcal{T}_i = (1-\alpha_t)D_i + \alpha_t \mathcal{P}^{\perp}_{\widehat{\mathbb{T}}_i} D_i - \alpha_t \sum_{j\neq i} \mathcal{P}_{\widehat{\mathbb{T}}_i} D_j + \alpha_t \mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{E}.$$

By (66), (68) with $\|D_j\|_{\mathrm{F}} \leq \varepsilon\lambda_j \leq \kappa\varepsilon\lambda_i$, (31) with $\nu = R$, and $0 < \alpha_t \leq 1$,

$$\frac{\|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i} \leq (1-\alpha_t)\varepsilon + \alpha_t\Big[3d\varepsilon^2 + (r-1)\kappa\big\{(d+1)\eta^{d-2} + (2d+4)\varepsilon\big\}\varepsilon + \frac{b}{8}\Big],$$
$$\frac{\|\widehat{\xi}_i\|_{\mathrm{F}}}{\lambda_i} \leq \alpha_t\Big[\varepsilon + (r-1)\kappa\big\{(d+1)\eta^{d-2} + (2d+4)\varepsilon\big\}\varepsilon + \frac{b}{8}\Big].$$

Since $2d+4 \leq 4d$ and $3d + (2d+4)(r-1)\kappa \leq 4d(1+r\kappa)$,

$$\begin{aligned}\frac{\|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i} &\leq \varphi_t(\varepsilon) := (1-\alpha_t)\varepsilon + \alpha_t\left\{(d+1)r\kappa\eta^{d-2}\varepsilon + 4d(1+r\kappa)\varepsilon^2 + \frac{b}{8}\right\},\\ \frac{\|\widehat{\xi}_i\|_{\mathrm{F}}}{\lambda_i} &\leq \alpha_t\left\{\varepsilon + (d+1)r\kappa\eta^{d-2}\varepsilon + 4dr\kappa\varepsilon^2 + \frac{b}{8}\right\}.\end{aligned} \tag{35}$$

The step-size conditions give $\alpha_t \geq 1/2$ and $\eta^{d-2} \leq 2c/(d\kappa r)$. Since $d+1 \leq 4d/3$,

$$(d+1)r\kappa\eta^{d-2} + 4d(1+r\kappa)R \leq \frac{8}{3}c + 4c_0 \leq \frac{1}{8},$$

and therefore, for $0 \leq \varepsilon \leq R$,

$$\varphi_t(\varepsilon) \leq \left(1 - \frac{7\alpha_t}{8}\right)\varepsilon + \frac{\alpha_t b}{8}, \qquad \frac{\|\widehat{\xi}_i\|_{\mathrm{F}}}{\lambda_i} \leq \frac{\alpha_t(9\varepsilon + b)}{8}. \tag{36}$$

Since $\|\widehat{\mathcal{T}}_i\|_{\mathrm{F}} \geq (1-\varepsilon)\lambda_i$, $R \leq 1/24$, and $b \leq R/4$,

$$\begin{aligned} \frac{\|\widehat{\xi}_i\|_{\mathrm{F}}}{\|\widehat{\mathcal{T}}_i\|_{\mathrm{F}}} &\leq \frac{\alpha_t(9\varepsilon+b)}{8(1-\varepsilon)} \leq \frac{9R+b}{8(1-R)} \leq 2R < \frac{1}{2(\sqrt{d}+1)}, \\ \frac{\|\widehat{\xi}_i\|_{\mathrm{F}}^2}{\lambda_i\|\widehat{\mathcal{T}}_i\|_{\mathrm{F}}} &\leq \frac{\alpha_t^2(9\varepsilon+b)^2}{64(1-\varepsilon)} \leq \alpha_t(3\varepsilon^2+b^2) \leq \frac{\alpha_t\varepsilon}{8} + \frac{7\alpha_t b}{8}. \end{aligned} \tag{37}$$

Hence (6) applies to $\widehat{\mathcal{T}}_i^+ = \mathcal{R}_{\widehat{\mathcal{T}}_i}(\widehat{\xi}_i)$, and (36)–(37) give

$$\frac{\|\widehat{\mathcal{T}}_i^+ - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i} \leq \varphi_t(\varepsilon) + \frac{\|\widehat{\xi}_i\|_{\mathrm{F}}^2}{\lambda_i\|\widehat{\mathcal{T}}_i\|_{\mathrm{F}}} \leq \left(1 - \frac{3\alpha_t}{4}\right)\varepsilon + \alpha_t b, \qquad 0 \leq \varepsilon \leq R. \tag{38}$$

Step 2 (Composition). Suppose $\varepsilon^{(t)} \leq R$ and set $\nu_t = \max\{\varepsilon^{(t)}, 2b\} \leq R$. The right-hand side of (38) is nondecreasing in ε and satisfies $(1-3\alpha_t/4)\nu_t + \alpha_t b \leq \nu_t$, so Lemma G.1 gives $\varepsilon^{(t+1)} \leq \nu_t \leq R$ and

$$\varepsilon^{(t+1)} \leq \left(1 - \frac{3\alpha_t}{4}\right)\nu_t + \alpha_t b \leq 2b + \left(1 - \frac{3\alpha_t}{4}\right)\{\varepsilon^{(t)} - 2b\}_+.$$

Since $\varepsilon^{(0)} \leq R$, induction over t yields

$$\varepsilon^{(t)} \leq 2b + \left\{\prod_{s=0}^{t-1}\left(1 - \frac{3\alpha_s}{4}\right)\right\}\{\varepsilon^{(0)} - 2b\}_+ \leq \left\{\prod_{s=0}^{t-1}\left(1 - \frac{3\alpha_s}{4}\right)\right\}\varepsilon^{(0)} + 2b, \qquad t \geq 0.$$

This proves the corollary with $C = 2C_0$. Under the conditions of Corollary 4.5, $c \leq 1/3$ gives $\alpha_t \geq 2/3$ and $1 - 3\alpha_t/4 \leq 1/2$, which proves that corollary.

Step 3 (RGN). For $\mathcal{A} = \mathrm{Id}$, the RGN direction is $\widehat{\Delta}_i = \mathcal{P}_{\widehat{\mathbb{T}}_i}(\sum_{j=1}^r D_j - \mathcal{E})$, so the RGN update is the RGD update with $\alpha_t = 1$. Taking $\alpha_t = 1$ in (35) and (37) and using $b/8 + b^2 \leq b$ gives, for $0 \leq \varepsilon \leq R$,

$$\frac{\|\widehat{\mathcal{T}}_i^+ - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i} \leq \Psi(\varepsilon) := \{4d(1+r\kappa)+5\}\varepsilon^2 + (d+1)r\kappa\eta^{d-2}\varepsilon + b \leq \frac{\varepsilon}{2} + b. \tag{39}$$

Step 2 with Ψ in place of the right-hand side of (38) proves the linear assertion of Corollary 4.6.

For the two-phase assertion, let $A = C_3 d(1+\kappa r)$ with a sufficiently large universal constant C_3, and fix $C_4 \geq 4$. By (39),

$$\Psi(\varepsilon) \leq \frac{A}{2}\{\varepsilon^2 + \eta^{d-2}\varepsilon\} + b, \qquad 0 \leq \varepsilon \leq R. \tag{40}$$

After fixing C_3 and C_4, decrease c_0 and then c, if necessary, so that

$$AR \leq \frac{1}{2}, \qquad A(C_4+1)\eta^{d-2} \leq 1, \qquad b \leq \frac{A\eta^{2(d-2)}}{8},$$

where the last inequality follows from the additional noise condition in Corollary 4.6. If $C_4\eta^{d-2} < \varepsilon \leq R$, then $\eta^{d-2}\varepsilon \leq \varepsilon^2/C_4$ and $b \leq A\varepsilon^2/(8C_4^2)$. If $0 \leq \varepsilon \leq \min\{R, C_4\eta^{d-2}\}$, then $A(\varepsilon + \eta^{d-2}) \leq 1$. Hence (40) gives

$$\Psi(\varepsilon) \leq A\varepsilon^2 \leq \varepsilon, \qquad C_4\eta^{d-2} < \varepsilon \leq R, \tag{41}$$

$$\Psi(\varepsilon) \leq \frac{\varepsilon}{2} + b, \qquad 0 \leq \varepsilon \leq \min\{R, C_4\eta^{d-2}\}, \tag{42}$$

and $2b \leq A\eta^{2(d-2)}/4 \leq C_4\eta^{d-2}$.

For $t < t^*$, $\varepsilon^{(t)} > C_4\eta^{d-2}$. If also $\varepsilon^{(t)} \leq R$, Lemma G.1 with $\nu = \varepsilon^{(t)}$ and (41) give $\varepsilon^{(t+1)} \leq A\{\varepsilon^{(t)}\}^2 \leq \varepsilon^{(t)} \leq R$. Induction from $\varepsilon^{(0)} \leq R$ yields

$$\varepsilon^{(t)} \leq \frac{\{A\varepsilon^{(0)}\}^{2^t}}{A}, \qquad 0 \leq t < t^*,$$

which is the quadratic-phase bound in Corollary 4.6. If $t^* < \infty$, then $\varepsilon^{(t^*)} \le \min\{R, C_4\eta^{d-2}\}$. For $t \ge t^*$ with $\varepsilon^{(t)} \le \min\{R, C_4\eta^{d-2}\}$, the number $\nu_t = \max\{\varepsilon^{(t)}, 2b\}$ satisfies $\nu_t \le \min\{R, C_4\eta^{d-2}\}$ and, by (42), $\Psi(\nu_t) \le \nu_t/2 + b \le \nu_t$. Lemma G.1 gives $\varepsilon^{(t+1)} \le \nu_t \le \min\{R, C_4\eta^{d-2}\}$ and $\varepsilon^{(t+1)} \le 2b + \{\varepsilon^{(t)} - 2b\}_+/2$, and induction yields

$$\varepsilon^{(t)} \le 2^{-(t-t^*)}\varepsilon^{(t^*)} + 2b, \qquad t \ge t^*.$$

Enlarging the universal constant multiplying the statistical error proves Corollary 4.6. □

Corollary G.4 (Convergence of RGD for CP tensor-on-tensor regression)**.** *Let* $\varepsilon^{(t)} = \max_{i\in[r]} \|\mathcal{T}_i^{(t)} - \mathcal{T}_i\|_{\mathrm{F}}/\lambda_i$. *There exist universal constants* $c, C > 0$ *such that the following holds. Suppose* $n \ge C\,\mathrm{df}$ *and*

$$\gamma \le \frac{c}{1+\kappa r}, \qquad \varepsilon^{(0)} \le \frac{c}{d(1+\kappa r)},$$

and, for every $t \ge 0$,

$$0 < \alpha_t\sigma_X^2 \le 1, \qquad |1 - \alpha_t\sigma_X^2| \le \frac{1}{2}, \qquad \alpha_t\sigma_X^2\eta^{d-2} \le \frac{c}{d\kappa r}.$$

Assume also that

$$\frac{\overline{\sigma}_\xi}{\sigma_X\lambda_r}\sqrt{\frac{\mathrm{df}}{n}} \le \frac{c}{d(1+\kappa r)}.$$

Then, with probability at least $1 - C\exp\{-c(\mathrm{df}_{\mathrm{in}} + \log r)\}$, *simultaneously for all* $t \ge 0$,

$$\varepsilon^{(t)} \le \left\{\prod_{s=0}^{t-1}\left(1 - \frac{3\alpha_s\sigma_X^2}{4}\right)\right\}\varepsilon^{(0)} + C\frac{\overline{\sigma}_\xi}{\sigma_X\lambda_r}\sqrt{\frac{\mathrm{df}}{n}}.$$

Proof. Write $\gamma = C_0\sqrt{(\mathrm{df}_{\mathrm{in}} + \log r)/n}$ with a sufficiently large universal constant $C_0 > 1$, so that $n \ge C\mathrm{df}$ and $\mathrm{df}_{\mathrm{in}} + \log r \le 2\mathrm{df}$ give $\gamma \le 1/2$. Let $x_m = \mathrm{vec}(\mathcal{X}_m) \in \mathbb{R}^{p_{\mathrm{in}}^*}$, $\widehat{\Sigma}_X = n^{-1}\sum_{m=1}^n x_m x_m^\top$, and let $\mathrm{mat}_{\mathrm{in}}(\mathcal{T}) \in \mathbb{R}^{p_{\mathrm{in}}^* \times p_{\mathrm{out}}^*}$ be the unfolding of $\mathcal{T}$ with the input modes as rows. By the definition of $\mathcal{A}$ in Section 2, for every tensor W,

$$\mathrm{mat}_{\mathrm{in}}\{\mathcal{A}^*\mathcal{A}(\mathcal{T})\} = \widehat{\Sigma}_X\,\mathrm{mat}_{\mathrm{in}}(\mathcal{T}), \qquad \langle W, \mathcal{A}^*\mathcal{E}\rangle = \frac{1}{n}\sum_{m=1}^n x_m^\top\,\mathrm{mat}_{\mathrm{in}}(W)\,\mathrm{vec}(\mathcal{E}_m). \tag{43}$$

Let

$$\mathfrak{B}_{\mathrm{loc}} = \bigcup_{i=1}^r \Big\{\widetilde{\mathcal{T}} \in \mathrm{Seg} : \|\widetilde{\mathcal{T}} - \mathcal{T}_i\|_{\mathrm{F}} \le \frac{\lambda_i}{4d}\Big\}.$$

Step 1 (Restricted design bounds). For a nonzero $\widetilde{\mathcal{T}} = \widetilde{\lambda}\widetilde{u}_1 \otimes \cdots \otimes \widetilde{u}_d \in \mathrm{Seg}$ with unit factor vectors, the tangent space of the Segre manifold in $\mathbb{R}^{p_1\times\cdots\times p_{d_{\mathrm{in}}}}$ at $\widetilde{u}_1 \otimes \cdots \otimes \widetilde{u}_{d_{\mathrm{in}}}$ is

$$\mathbb{T}_{\widetilde{u}_1\otimes\cdots\otimes\widetilde{u}_{d_{\mathrm{in}}}} = \mathrm{span}\{\widetilde{u}_1 \otimes \cdots \otimes \widetilde{u}_{d_{\mathrm{in}}}\} + \sum_{\ell=1}^{d_{\mathrm{in}}}\{\widetilde{u}_1 \otimes \cdots \otimes v_\ell \otimes \cdots \otimes \widetilde{u}_{d_{\mathrm{in}}} : v_\ell \perp \widetilde{u}_\ell\}.$$

It has dimension $\mathrm{df}_{\mathrm{in}}$, and by (3) every column of $\mathrm{mat}_{\mathrm{in}}(Z)$ belongs to it whenever $Z \in \mathbb{T}_{\widetilde{\mathcal{T}}}$. Let $\mathfrak{U}_{\mathrm{loc}}$ be the union of the unit spheres of $\mathbb{T}_{\widetilde{u}_1\otimes\cdots\otimes\widetilde{u}_{d_{\mathrm{in}}}}$ over $\widetilde{\mathcal{T}} \in \mathfrak{B}_{\mathrm{loc}}$. For $\|\widetilde{\mathcal{T}} - \mathcal{T}_i\|_{\mathrm{F}} \le \lambda_i/(4d)$, (32) with $\nu = 1/(4d)$ gives $\sum_{\ell=1}^{d_{\mathrm{in}}}\|\mathcal{P}_{u_{\ell,i}}^\perp\widetilde{u}_\ell\|_2^2 \le 1/(8d^2)$, and the proof of (34), applied to the first d_{in} factors with (d, df) replaced by $(d_{\mathrm{in}}, \mathrm{df}_{\mathrm{in}})$, gives

$$\log N(\epsilon, \mathfrak{U}_{\mathrm{loc}}, \|\cdot\|_2) \le C\mathrm{df}_{\mathrm{in}}\log(C/\epsilon) + \log r \le C(\mathrm{df}_{\mathrm{in}} + \log r)\log(C/\epsilon), \qquad 0 < \epsilon < 1. \tag{44}$$

Let $\mathbb{X} \in \mathbb{R}^{n\times p_{\mathrm{in}}^*}$ have rows $\sigma_X^{-1}x_m^\top$, so that $\mathbb{X}$ has independent $\mathcal{N}(0,1)$ entries. Since $\mathfrak{U}_{\mathrm{loc}} = -\mathfrak{U}_{\mathrm{loc}}$, the set $\mathfrak{U}_{\mathrm{loc}} + \mathfrak{U}_{\mathrm{loc}}$ contains $y \pm y'$ for all $y, y' \in \mathfrak{U}_{\mathrm{loc}}$ and has radius at most two, and by Dudley's inequality [51, Theorem 8.1.3] and (44) its Gaussian complexity is at most $C\sqrt{\mathrm{df}_{\mathrm{in}} + \log r}$. The matrix deviation inequality [51, Theorem 9.1.1] gives

$$\mathbb{E}\sup_{z\in\mathfrak{U}_{\mathrm{loc}}+\mathfrak{U}_{\mathrm{loc}}}\big|\|\mathbb{X}z\|_2 - \sqrt{n}\|z\|_2\big| \le C\sqrt{\mathrm{df}_{\mathrm{in}} + \log r}.$$

The supremum is a 2-Lipschitz function of $\mathbb{X}$ in the Frobenius norm, so Gaussian concentration [51, Theorem 5.2.2] gives, with probability at least $1 - 2\exp\{-c(\mathrm{df}_{\mathrm{in}} + \log r)\}$,

$$\sup_{z \in \mathfrak{U}_{\mathrm{loc}} + \mathfrak{U}_{\mathrm{loc}}} \big| \|\mathbb{X}z\|_2 - \sqrt{n}\|z\|_2 \big| \le C\sqrt{\mathrm{df}_{\mathrm{in}} + \log r} \le \frac{\sqrt{n}\,\gamma}{32}$$

for C_0 sufficiently large. On this event, since $\|\mathbb{X}z\|_2^2/n = z^\top \sigma_X^{-2}\widehat{\Sigma}_X z$, every $z = y \pm y'$ with $y, y' \in \mathfrak{U}_{\mathrm{loc}}$ satisfies

$$\big|(z^\top \sigma_X^{-2}\widehat{\Sigma}_X z)^{1/2} - \|z\|_2\big| \le \frac{\gamma}{32},$$
$$\big|z^\top(\sigma_X^{-2}\widehat{\Sigma}_X - I)z\big| \le \frac{\gamma}{32}\left(2\|z\|_2 + \frac{\gamma}{32}\right) \le \frac{\gamma}{8} + \frac{\gamma^2}{1024},$$

and polarization with $\gamma \le 1/2$ yields

$$\sup_{y, y' \in \mathfrak{U}_{\mathrm{loc}}} \big|y^\top(\sigma_X^{-2}\widehat{\Sigma}_X - I)y'\big| \le \frac{1}{2}\left(\frac{\gamma}{8} + \frac{\gamma^2}{1024}\right) \le \frac{\gamma}{4}. \tag{45}$$

For $\widetilde{\mathcal{T}}, \widetilde{\mathcal{T}}' \in \mathfrak{B}_{\mathrm{loc}}$, $Y \in \mathbb{T}_{\widetilde{\mathcal{T}}}$, and $Z \in \mathbb{T}_{\widetilde{\mathcal{T}}'}$, let y_k and z_k be the columns of $\mathrm{mat}_{\mathrm{in}}(Y)$ and $\mathrm{mat}_{\mathrm{in}}(Z)$. Then (43), (45), and the Cauchy–Schwarz inequality give

$$\big|\langle Y, (\sigma_X^{-2}\mathcal{A}^*\mathcal{A} - I)Z\rangle\big| = \Big|\sum_k y_k^\top(\sigma_X^{-2}\widehat{\Sigma}_X - I)z_k\Big| \le \frac{\gamma}{4}\sum_k \|y_k\|_2\|z_k\|_2 \le \frac{\gamma}{4}\|Y\|_{\mathrm{F}}\|Z\|_{\mathrm{F}}. \tag{46}$$

For $Y \in \mathbb{T}_{\widetilde{\mathcal{T}}}$ and an arbitrary tensor Z, since $\langle Y, \mathcal{P}_{\mathbb{T}_{\widetilde{\mathcal{T}}}}^{\perp} W\rangle = 0$ for every tensor W,

$$\big\langle Y, \sigma_X^{-2}\mathcal{A}^*\mathcal{A}\,\mathcal{P}_{\mathbb{T}_{\widetilde{\mathcal{T}}}}^{\perp}\mathcal{P}_{\mathbb{T}_{\widetilde{\mathcal{T}}'}} Z\big\rangle = \big\langle Y, (\sigma_X^{-2}\mathcal{A}^*\mathcal{A} - I)\mathcal{P}_{\mathbb{T}_{\widetilde{\mathcal{T}}'}} Z\big\rangle - \big\langle Y, (\sigma_X^{-2}\mathcal{A}^*\mathcal{A} - I)\mathcal{P}_{\mathbb{T}_{\widetilde{\mathcal{T}}}}\mathcal{P}_{\mathbb{T}_{\widetilde{\mathcal{T}}'}} Z\big\rangle,$$

and (46) bounds the right-hand side by $(\gamma/2)\|Y\|_{\mathrm{F}}\|Z\|_{\mathrm{F}}$. Consequently, for all $\widetilde{\mathcal{T}}, \widetilde{\mathcal{T}}' \in \mathfrak{B}_{\mathrm{loc}}$,

$$\big\|\mathcal{P}_{\mathbb{T}_{\widetilde{\mathcal{T}}}}(\sigma_X^{-2}\mathcal{A}^*\mathcal{A} - I)\mathcal{P}_{\mathbb{T}_{\widetilde{\mathcal{T}}}}\big\| \le \frac{\gamma}{4}, \tag{47}$$

$$\Big\|\mathcal{P}_{\mathbb{T}_{\widetilde{\mathcal{T}}}}\sigma_X^{-2}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\mathbb{T}_{\widetilde{\mathcal{T}}}}^{\perp}\mathcal{P}_{\mathbb{T}_{\widetilde{\mathcal{T}}'}}\Big\| \le \gamma. \tag{48}$$

Step 2 (Normal chord bound). Fix $i, j \in [r]$ and $\widehat{\mathcal{T}}_i, \widehat{\mathcal{T}}_j \in \mathrm{Seg}$ with $\|\widehat{\mathcal{T}}_k - \mathcal{T}_k\|_{\mathrm{F}} \le \lambda_k/(4d)$ for $k \in \{i, j\}$, and let $\delta_j = \|D_j\|_{\mathrm{F}}/\lambda_j$. We show that, with probability at least $1 - e^{-\mathrm{df}}$, simultaneously for all such $i, j, \widehat{\mathcal{T}}_i, \widehat{\mathcal{T}}_j$,

$$\Big\|\mathcal{P}_{\widehat{\mathbb{T}}_i}(\sigma_X^{-2}\mathcal{A}^*\mathcal{A} - I)\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}\mathcal{P}_{\widehat{\mathbb{T}}_j}^{\perp} D_j\Big\|_{\mathrm{F}} \le C\left\{\sqrt{\frac{d\,\mathrm{df}}{n}} + \frac{d\,\mathrm{df}}{n}\right\}\frac{\|D_j\|_{\mathrm{F}}^2}{\lambda_j} \le Cd\frac{\|D_j\|_{\mathrm{F}}^2}{\lambda_j}. \tag{49}$$

Both sides vanish when $\delta_j = 0$, so let $\delta_j > 0$. By (69) and the orthogonality argument following it, applied with $(\mathcal{T}_i, \mathcal{T}_i^{(t)})$ replaced by $(\mathcal{T}_j, \widehat{\mathcal{T}}_j)$, there are orthonormal rank-one tensors V_S, $S \subseteq [d]$, and numbers $s_\ell \in [0, 1]$, $c_\ell = (1 - s_\ell^2)^{1/2}$, such that

$$\mathcal{P}_{\widehat{\mathbb{T}}_j}^{\perp} D_j = -\lambda_j \sum_{|S| \ge 2}\left(\prod_{\ell \in S} s_\ell \prod_{\ell \notin S} c_\ell\right) V_S, \qquad \sum_{\ell=1}^{d} s_\ell^2 \le 2\delta_j^2, \tag{50}$$

where the second bound is the inequality $q \le 2\{\delta_i^{(t)}\}^2$ in the proof of Lemma H.1, applied to $(\mathcal{T}_j, \widehat{\mathcal{T}}_j)$. Since every S with $|S| \ge 2$ contains a pair $\{k, \ell\}$, all summands are nonnegative, and $s_m^2 + c_m^2 = 1$,

$$\begin{aligned}\frac{\|\mathcal{P}_{\widehat{\mathbb{T}}_j}^{\perp} D_j\|_{\mathrm{F}}^2}{\lambda_j^2} &= \sum_{|S| \ge 2}\prod_{\ell \in S} s_\ell^2 \prod_{\ell \notin S} c_\ell^2 \le \sum_{1 \le k < \ell \le d} s_k^2 s_\ell^2 \sum_{S \supseteq \{k,\ell\}} \prod_{m \in S \setminus \{k,\ell\}} s_m^2 \prod_{m \notin S} c_m^2 \\ &= \sum_{1 \le k < \ell \le d} s_k^2 s_\ell^2 \le \frac{1}{2}\left(\sum_{\ell=1}^{d} s_\ell^2\right)^2 \le 2\delta_j^4.\end{aligned} \tag{51}$$

Set $\chi_\ell = s_\ell/\delta_j$, so that $\|\chi\|_2 \le \sqrt{2}$, $\|\chi\|_1 \le \sqrt{2d}$, and $\delta_j|\chi_\ell| \le \sqrt{2}/(4d)$. Dividing (50) by $\lambda_j\delta_j^2$ gives

$$\frac{\mathcal{P}_{\widehat{\mathbb{T}}_j}^{\perp} D_j}{\lambda_j \delta_j^2} = -\sum_{|S|\ge 2} a_S(\delta_j, \chi) V_S, \qquad a_S(\delta,\chi) = \delta^{|S|-2}\left(\prod_{\ell\in S}\chi_\ell\right)\prod_{\ell\notin S}\sqrt{1-\delta^2\chi_\ell^2},$$

which extends continuously to $\delta_j = 0$. On the domain $0 \le \delta \le 1/(4d)$, $\|\chi\|_2 \le \sqrt{2}$, every square-root factor is at least $1/2$ and $\delta\|\chi\|_1 \le 1$. Bounding the elementary symmetric polynomials of $|\chi_1|, \dots, |\chi_d|$ by $\|\chi\|_1^k/k!$ gives

$$\sum_{|S|\ge 2} |a_S(\delta,\chi)| \le \sum_{k\ge 2} \frac{\delta^{k-2}\|\chi\|_1^k}{k!} \le \frac{\|\chi\|_1^2}{2}\exp(\delta\|\chi\|_1) \le Cd. \tag{52}$$

Differentiating the products in a_S and summing in the same way, the terms obtained from $\delta^{|S|-2}$ are at most $C\|\chi\|_1^3\exp(\delta\|\chi\|_1)$, those obtained from the square-root factors are at most $C\delta\|\chi\|_1^2\|\chi\|_2^2\exp(\delta\|\chi\|_1)$, and the derivatives in χ are at most $Cd\|\chi\|_1\exp(\delta\|\chi\|_1)$, so that

$$\sum_{|S|\ge 2}\{|\partial_\delta a_S(\delta,\chi)| + \|\nabla_\chi a_S(\delta,\chi)\|_1\} \le Cd^2. \tag{53}$$

Let $\mathfrak{H}$ be the Frobenius-norm closure of the set consisting of the zero tensor and all tensors

$$H = Z \pm \frac{\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}\mathcal{P}_{\widehat{\mathbb{T}}_j}^{\perp} D_j}{\lambda_j\delta_j^2}, \qquad Z \in \widehat{\mathbb{T}}_i, \qquad \|Z\|_{\mathrm{F}} = 1,$$

over $i, j \in [r]$ and all $\widehat{\mathcal{T}}_i, \widehat{\mathcal{T}}_j$ as above with $\delta_j > 0$. By (51), $\sup_{H\in\mathfrak{H}}\|H\|_{\mathrm{F}} \le 1+\sqrt{2}$. For fixed (i,j) and sign, $H = H(\vartheta)$ is a function of

$$\vartheta = \big(\delta_j, \chi, (\widehat{u}_{\ell,j}, w_\ell)_\ell, (\widehat{u}_{\ell,i})_\ell, \beta\big),$$

where ℓ ranges over the modes with $p_\ell \ge 2$, $(\widehat{u}_{\ell,j}, w_\ell)$ are the orthonormal pairs that define V_S in (69), $(\widehat{u}_{\ell,i})_\ell$ are the unit factors of $\widehat{\mathcal{T}}_i$, and $\beta \in \mathbb{S}^{\mathrm{df}-1}$ gives $Z = Q\sum_{k=1}^{\mathrm{df}}\beta_k E_k$ for a fixed orthonormal basis $E_1, \dots, E_{\mathrm{df}}$ of $\mathbb{T}_{\mathcal{T}_i}$ and the rotation Q in (33) with $\widetilde{\mathcal{T}} = \widehat{\mathcal{T}}_i$. If $p_\ell = 1$, the sign convention gives $\widehat{u}_{\ell,i} = u_{\ell,i}$, $\widehat{u}_{\ell,j} = u_{\ell,j}$, and $s_\ell = \chi_\ell = 0$, so mode ℓ carries no parameter, and if $\chi_\ell = 0$, then w_ℓ may be any unit vector orthogonal to $\widehat{u}_{\ell,j}$. There are at most $\mathrm{df}-1$ modes with $p_\ell \ge 2$, and $\sum_\ell p_\ell \le 2(\mathrm{df}-1)$ over them, so

$$\|\vartheta\|_2^2 \le \frac{1}{16} + 2 + 3d + 1 \le 9d, \qquad \dim(\vartheta) \le 1 + (\mathrm{df}-1) + 6(\mathrm{df}-1) + \mathrm{df} \le 8\,\mathrm{df}.$$

For unit vectors u_ℓ, u'_ℓ, $\ell \in [d]$, by (3),

$$\Big\|\bigotimes_{\ell=1}^d u_\ell - \bigotimes_{\ell=1}^d u'_\ell\Big\|_{\mathrm{F}} \le \sum_{\ell=1}^d \|u_\ell - u'_\ell\|_2,$$

$$\|\mathcal{P}_{\mathbb{T}_{u_1\otimes\cdots\otimes u_d}} - \mathcal{P}_{\mathbb{T}_{u'_1\otimes\cdots\otimes u'_d}}\| \le 2(d+1)\sum_{\ell=1}^d \|u_\ell - u'_\ell\|_2,$$

$$\mathcal{P}_{\mathbb{T}_{u_1\otimes\cdots\otimes u_d}} = \bigotimes_{\ell=1}^d u_\ell u_\ell^\top + \sum_{k=1}^d (I - u_k u_k^\top)\otimes\bigotimes_{\ell\ne k} u_\ell u_\ell^\top.$$

Together with the bound $\|Q - Q'\| \le 7\sqrt{d}(\sum_\ell \|\widehat{u}_{\ell,i} - \widehat{u}'_{\ell,i}\|_2^2)^{1/2}$ in the proof of Lemma G.2, (52), and (53), these inequalities give

$$\|H(\vartheta) - H(\vartheta')\|_{\mathrm{F}} \le Cd^2\|\vartheta - \vartheta'\|_2.$$

Since the parameter domain is compact and H is continuous, every nonzero element of $\mathfrak{H}$ equals $H(\vartheta)$ for some (i,j), sign, and ϑ. By Vershynin [51, Lemma 4.2.8], the volume comparison in the

proof of [51, Proposition 4.2.12], the Lipschitz bound, and $2\log r \le 2(\mathrm{df}-1)$ for the union over (i,j),

$$\log N(\epsilon, \mathfrak{H}, \|\cdot\|_{\mathrm{F}}) \le \log 3 + 2\log r + 8\,\mathrm{df}\log\left(1+\frac{6Cd^{5/2}}{\epsilon}\right) \le Cd\,\mathrm{df}\log(C/\epsilon), \qquad 0<\epsilon<1. \tag{54}$$

For $H \in \mathfrak{H}$, let $B_H = n^{-1/2}\operatorname{diag}(\mathrm{mat}_{\mathrm{in}}(H)^\top, \dots, \mathrm{mat}_{\mathrm{in}}(H)^\top)$ with n diagonal blocks and $g = \sigma_X^{-1}(x_1^\top, \dots, x_n^\top)^\top \sim \mathcal{N}(0, I)$. By (43),

$$\begin{aligned} \Xi_H &:= \|B_H g\|_2^2 - \mathbb{E}\|B_H g\|_2^2 = \left\langle H, (\sigma_X^{-2}\mathcal{A}^*\mathcal{A} - I)H\right\rangle, \\ \|B_H\|_{\mathrm{F}} &= \|H\|_{\mathrm{F}}, \qquad \|B_H\| \le \frac{\|H\|_{\mathrm{F}}}{\sqrt{n}}, \\ \|B_H - B_{H'}\|_{\mathrm{F}} &= \|H-H'\|_{\mathrm{F}}, \qquad \|B_H - B_{H'}\| \le \frac{\|H-H'\|_{\mathrm{F}}}{\sqrt{n}}. \end{aligned} \tag{55}$$

Hence $\|B_H^\top B_H - B_{H'}^\top B_{H'}\|_{\mathrm{F}} \le Cn^{-1/2}\|H-H'\|_{\mathrm{F}}$ and $\|B_H^\top B_H - B_{H'}^\top B_{H'}\| \le Cn^{-1}\|H-H'\|_{\mathrm{F}}$, and the Gaussian Hanson–Wright inequality [51, Theorem 6.2.1] gives, for every $u>0$,

$$\mathbb{P}\left\{|\Xi_H - \Xi_{H'}| > C\|H-H'\|_{\mathrm{F}}\left(\sqrt{u/n}+u/n\right)\right\} \le 2e^{-u}.$$

Thus $(\Xi_H)_{H\in\mathfrak{H}}$ satisfies the mixed-tail condition (3.8) of Dirksen [15] with $d_2(H,H') = C\|H-H'\|_{\mathrm{F}}/\sqrt{n}$ and $d_1(H,H') = C\|H-H'\|_{\mathrm{F}}/n$. Let $F \subset \mathfrak{H}$ be finite with $0 \in F$. Since $\Xi_0 = 0$, $N(\epsilon, F, \|\cdot\|_{\mathrm{F}}) \le N(\epsilon/2, \mathfrak{H}, \|\cdot\|_{\mathrm{F}})$ [51, Lemma 4.2.8], and $\sup_{H\in\mathfrak{H}}\|H\|_{\mathrm{F}} \le 1+\sqrt{2}$, Theorem 3.5 of Dirksen [15] with $t_0 = 0$ and $u = \mathrm{df}$, the entropy bound (2.3) there, and (54) give, with probability at least $1-e^{-\mathrm{df}}$,

$$\begin{aligned} \max_{H\in F}|\Xi_H| &\le \frac{C}{\sqrt{n}}\int_0^\infty \sqrt{\log N(\epsilon, F, \|\cdot\|_{\mathrm{F}})}\,d\epsilon + \frac{C}{n}\int_0^\infty \log N(\epsilon, F, \|\cdot\|_{\mathrm{F}})\,d\epsilon \\ &\quad + C\left(\sqrt{\frac{\mathrm{df}}{n}} + \frac{\mathrm{df}}{n}\right) \le C\left\{\sqrt{\frac{d\,\mathrm{df}}{n}} + \frac{d\,\mathrm{df}}{n}\right\}, \end{aligned}$$

where C does not depend on F. Letting F increase along finite sets whose union is dense in $\mathfrak{H}$ and using the continuity of $H \mapsto \Xi_H$,

$$\sup_{H\in\mathfrak{H}}|\Xi_H| \le C\left\{\sqrt{\frac{d\,\mathrm{df}}{n}} + \frac{d\,\mathrm{df}}{n}\right\} \tag{56}$$

with probability at least $1-e^{-\mathrm{df}}$. For $W = \mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}\mathcal{P}_{\widehat{\mathbb{T}}_j}^{\perp}D_j/(\lambda_j\delta_j^2)$ and unit $Z \in \widehat{\mathbb{T}}_i$, polarization in (55) gives

$$4\left\langle Z, (\sigma_X^{-2}\mathcal{A}^*\mathcal{A} - I)W\right\rangle = \Xi_{Z+W} - \Xi_{Z-W}.$$

Taking the supremum over Z and multiplying by $\lambda_j\delta_j^2 = \|D_j\|_{\mathrm{F}}^2/\lambda_j$ proves (49) on the event of (56), where the last inequality in (49) uses $n \ge C\mathrm{df}$.

Step 3 (Projected noise). Let

$$\mathfrak{Z}_{\mathrm{loc}} = \bigcup_{\widetilde{\mathcal{T}}\in\mathfrak{B}_{\mathrm{loc}}}\{Z \in \mathbb{T}_{\widetilde{\mathcal{T}}} : \|Z\|_{\mathrm{F}} = 1\}, \qquad \sup_{Z\in\mathfrak{Z}_{\mathrm{loc}}}|\langle Z, \mathcal{A}^*\mathcal{E}\rangle| = \sup_{\widetilde{\mathcal{T}}\in\mathfrak{B}_{\mathrm{loc}}}\|\mathcal{P}_{\mathbb{T}_{\widetilde{\mathcal{T}}}}\mathcal{A}^*\mathcal{E}\|_{\mathrm{F}}.$$

By (34) with $\nu = 1/(4d)$ and $\log r \le \mathrm{df}-1$,

$$\log N(\epsilon, \mathfrak{Z}_{\mathrm{loc}}, \|\cdot\|_{\mathrm{F}}) \le C\mathrm{df}\log(C/\epsilon), \qquad 0<\epsilon<1. \tag{57}$$

Fix a tensor W. Whitening x_m and $\mathrm{vec}(\mathcal{E}_m)$ and taking a singular value decomposition of $\sigma_X\,\mathrm{mat}_{\mathrm{in}}(W)\,\mathrm{Cov}\{\mathrm{vec}(\mathcal{E}_m)\}^{1/2}$, with singular values $s_{m,k}$, give

$$\begin{aligned} x_m^\top \mathrm{mat}_{\mathrm{in}}(W)\,\mathrm{vec}(\mathcal{E}_m) &\overset{\mathrm{d}}{=} \sum_k s_{m,k}g_{m,k}g'_{m,k}, \\ \sum_k s_{m,k}^2 \le \sigma_X^2\overline{\sigma}_\xi^2\|W\|_{\mathrm{F}}^2, &\qquad \max_k s_{m,k} \le \sigma_X\overline{\sigma}_\xi\|W\|_{\mathrm{F}}, \end{aligned}$$

with mutually independent standard Gaussian variables $g_{m,k}, g'_{m,k}$. For $|\omega| \le \{2\sigma_X\overline{\sigma}_\xi\|W\|_{\mathrm{F}}\}^{-1}$,

$$\log \mathbb{E}\exp\{\omega x_m^\top \mathrm{mat}_{\mathrm{in}}(W)\,\mathrm{vec}(\mathcal{E}_m)\} = -\frac{1}{2}\sum_k \log(1-\omega^2 s_{m,k}^2) \le \omega^2\sigma_X^2\overline{\sigma}_\xi^2\|W\|_{\mathrm{F}}^2.$$

Hence the summands in the second identity of (43) are independent, centered, and sub-exponential with parameter at most $C\sigma_X\overline{\sigma}_\xi\|W\|_{\mathrm{F}}$ [51, Proposition 2.7.1], and Bernstein's inequality [51, Corollary 2.8.3] gives, for every $u > 0$,

$$\mathbb{P}\left\{|\langle W, \mathcal{A}^*\mathcal{E}\rangle| > C\sigma_X\overline{\sigma}_\xi\|W\|_{\mathrm{F}}\left(\sqrt{u/n} + u/n\right)\right\} \le 2e^{-u}. \tag{58}$$

With $W = Z - Z'$, the process $(\langle Z, \mathcal{A}^*\mathcal{E}\rangle)_{Z\in\mathfrak{Z}_{\mathrm{loc}}}$ satisfies (3.8) of Dirksen [15] with $d_2(Z,Z') = C\sigma_X\overline{\sigma}_\xi\|Z-Z'\|_{\mathrm{F}}/\sqrt{n}$ and $d_1(Z,Z') = C\sigma_X\overline{\sigma}_\xi\|Z-Z'\|_{\mathrm{F}}/n$. Fix $Z_0 \in \mathfrak{Z}_{\mathrm{loc}}$. The diameter of $\mathfrak{Z}_{\mathrm{loc}}$ is at most two, so the chaining argument of Step 2, with Theorem 3.5 of Dirksen [15] applied at $t_0 = Z_0$ and $u = \mathrm{df}$ and with (57) in place of (54), together with (58) for $W = Z_0$ and $u = \mathrm{df}$, gives, with probability at least $1 - 3e^{-\mathrm{df}}$,

$$\begin{aligned}\sup_{\widetilde{\mathcal{T}}\in\mathfrak{B}_{\mathrm{loc}}} \|\mathcal{P}_{\mathbb{T}_{\widetilde{\mathcal{T}}}}\mathcal{A}^*\mathcal{E}\|_{\mathrm{F}} &\le |\langle Z_0, \mathcal{A}^*\mathcal{E}\rangle| + \sup_{Z\in\mathfrak{Z}_{\mathrm{loc}}} |\langle Z - Z_0, \mathcal{A}^*\mathcal{E}\rangle| \\ &\le C\sigma_X\overline{\sigma}_\xi\left(\sqrt{\frac{\mathrm{df}}{n}} + \frac{\mathrm{df}}{n}\right) \le C_0\sigma_X\overline{\sigma}_\xi\sqrt{\frac{\mathrm{df}}{n}},\end{aligned} \tag{59}$$

where the last inequality uses $n \ge C\mathrm{df}$ and C_0 sufficiently large.

Since $\mathrm{df}_{\mathrm{in}} + \log r \le 2\mathrm{df}$, the events of Steps 1–3 hold simultaneously with probability at least

$$1 - 2\exp\{-c(\mathrm{df}_{\mathrm{in}} + \log r)\} - 4e^{-\mathrm{df}} \ge 1 - C\exp\{-c(\mathrm{df}_{\mathrm{in}} + \log r)\}.$$

The remaining steps are on this event.

Step 4 (One component update). Let $\{\widehat{\mathcal{T}}_j\}_{j=1}^r \subset \mathrm{Seg}$ satisfy $\varepsilon = \max_j \|D_j\|_{\mathrm{F}}/\lambda_j \le 1/(4d)$, so that $\widehat{\mathcal{T}}_j \in \mathfrak{B}_{\mathrm{loc}}$ for every $j \in [r]$. Equations (47), (48), and (59) give

$$\begin{aligned}&\max_i \|\mathcal{P}_{\widehat{\mathbb{T}}_i}(I - \alpha_t\mathcal{A}^*\mathcal{A})\mathcal{P}_{\widehat{\mathbb{T}}_i}\| \le |1-\alpha_t\sigma_X^2| + \alpha_t\sigma_X^2\gamma, \qquad \max_i \|\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}\| \le \sigma_X^2(1+\gamma), \\ &\max_{i\ne j} \|\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}\mathcal{P}_{\widehat{\mathbb{T}}_j}\| \le \sigma_X^2\gamma, \qquad \max_i \frac{\|\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{E}\|_{\mathrm{F}}}{\lambda_i} \le C_0\frac{\sigma_X\overline{\sigma}_\xi}{\lambda_r}\sqrt{\frac{\mathrm{df}}{n}}.\end{aligned} \tag{60}$$

For RGD, $\widehat{\xi}_i = -\alpha_t\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\sum_{j=1}^r D_j + \alpha_t\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{E}$ and $\widehat{W}_i - \mathcal{T}_i = D_i + \widehat{\xi}_i$. With $\mathrm{II}, \ldots, \mathrm{VI}$ as in the proof of Theorem 4.1,

$$\begin{aligned}D_i - \alpha_t\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}D_i = (1-\alpha_t\sigma_X^2)D_i + \alpha_t\sigma_X^2\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}D_i \\ - \alpha_t\mathcal{P}_{\widehat{\mathbb{T}}_i}(\mathcal{A}^*\mathcal{A} - \sigma_X^2 I)\mathcal{P}_{\widehat{\mathbb{T}}_i}D_i - \alpha_t\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp}D_i,\end{aligned}$$

where the third term has norm at most $\alpha_t\sigma_X^2\gamma\lambda_i\varepsilon$ by (47) and the last term has norm $\alpha_t\mathrm{III}$. Since $\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{P}_{\widehat{\mathbb{T}}_i}^{\perp} = 0$, (49) remains valid with $\sigma_X^{-2}\mathcal{A}^*\mathcal{A} - I$ replaced by $\sigma_X^{-2}\mathcal{A}^*\mathcal{A}$. Applying it with $j = i$ and, after the triangle inequality, with $j \ne i$, and using (66), (68), and (60), we obtain

$$\begin{aligned}&\frac{\mathrm{II}}{\lambda_i} \le 3d\varepsilon^2, \qquad \frac{\mathrm{III}}{\lambda_i} \le Cd\sigma_X^2\varepsilon^2, \qquad \frac{\mathrm{IV}}{\lambda_i} \le r\kappa\sigma_X^2(1+\gamma)\{(d+1)\eta^{d-2} + 4d\varepsilon\}\varepsilon, \\ &\frac{\mathrm{V}}{\lambda_i} \le (r-1)\kappa\sigma_X^2\gamma\varepsilon, \qquad \frac{\mathrm{VI}}{\lambda_i} \le Cd(r-1)\kappa\sigma_X^2\varepsilon^2.\end{aligned}$$

Since $\|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}} \le \|D_i - \alpha_t \mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}D_i\|_{\mathrm{F}} + \alpha_t(\mathrm{IV}+\mathrm{V}+\mathrm{VI}) + \alpha_t\|\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{E}\|_{\mathrm{F}}$, $0 < \alpha_t\sigma_X^2 \le 1$, and $\gamma \le 1/2$,

$$\begin{aligned}\frac{\|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i} &\le (1-\alpha_t\sigma_X^2)\varepsilon + \alpha_t\sigma_X^2\big\{(1+r\kappa)\gamma + (d+1)r\kappa(1+\gamma)\eta^{d-2}\big\}\varepsilon \\ &\quad + Cd\alpha_t\sigma_X^2(1+r\kappa)\varepsilon^2 + C_0\alpha_t\frac{\sigma_X\overline{\sigma}_\xi}{\lambda_r}\sqrt{\frac{\mathrm{df}}{n}}, \qquad (61)\\ \frac{\|\widehat{\xi}_i\|_{\mathrm{F}}}{\lambda_i} &\le \alpha_t\sigma_X^2\varepsilon + \alpha_t\sigma_X^2\big\{(1+r\kappa)\gamma + (d+1)r\kappa(1+\gamma)\eta^{d-2}\big\}\varepsilon \\ &\quad + Cd\alpha_t\sigma_X^2(1+r\kappa)\varepsilon^2 + C_0\alpha_t\frac{\sigma_X\overline{\sigma}_\xi}{\lambda_r}\sqrt{\frac{\mathrm{df}}{n}}, \qquad (62)\end{aligned}$$

where the second bound uses $\widehat{\xi}_i = (D_i - \alpha_t\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}D_i) - D_i - \alpha_t\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\sum_{j\ne i}D_j + \alpha_t\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{E}$ and $\|(1-\alpha_t\sigma_X^2)D_i + \alpha_t\sigma_X^2\mathcal{P}^{\perp}_{\widehat{\mathbb{T}}_i}D_i - D_i\|_{\mathrm{F}} = \alpha_t\sigma_X^2\|\mathcal{P}_{\widehat{\mathbb{T}}_i}D_i\|_{\mathrm{F}} \le \alpha_t\sigma_X^2\lambda_i\varepsilon$.

For RGN, with $\widehat{\mathsf{H}}_i$ as in the proof of Theorem 4.2, (47), $\gamma \le 1/2$, and (60) give

$$\begin{gathered}\max_i \left\| \left(\mathcal{P}_{\widehat{\mathbb{T}}_i}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\widehat{\mathbb{T}}_i}|_{\widehat{\mathbb{T}}_i}\right)^{-1}\right\| \le \frac{2}{\sigma_X^2}, \qquad \max_{i\ne j}\|\widehat{\mathsf{C}}_i\mathcal{P}_{\widehat{\mathbb{T}}_j}\| \le 2\gamma, \\ \max_i \frac{\|\widehat{\mathsf{H}}_i\mathcal{E}\|_{\mathrm{F}}}{\lambda_i} \le 2C_0\frac{\overline{\sigma}_\xi}{\sigma_X\lambda_r}\sqrt{\frac{\mathrm{df}}{n}}.\end{gathered} \qquad (63)$$

With $\mathrm{I},\dots,\mathrm{VI}$ as in the proof of Theorem 4.2, (63), (49), (66), and (68) give

$$\begin{gathered}\frac{\mathrm{I}}{\lambda_i} \le r\kappa\{(d+1)\eta^{d-2} + 4d\varepsilon\}\varepsilon, \qquad \frac{\mathrm{II}}{\lambda_i} \le Cd\varepsilon^2, \qquad \frac{\mathrm{III}}{\lambda_i} \le 2(r-1)\kappa\gamma\varepsilon, \\ \frac{\mathrm{IV}}{\lambda_i} \le Cd(r-1)\kappa\varepsilon^2, \qquad \frac{\mathrm{V}}{\lambda_i} \le 3d\varepsilon^2,\end{gathered}$$

and therefore

$$\frac{\|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i} \le \{2r\kappa\gamma + (d+1)r\kappa\eta^{d-2}\}\varepsilon + Cd(1+r\kappa)\varepsilon^2 + 2C_0\frac{\overline{\sigma}_\xi}{\sigma_X\lambda_r}\sqrt{\frac{\mathrm{df}}{n}}. \qquad (64)$$

Step 5 (Composition). Let

$$R = \frac{c_0}{d(1+\kappa r)}, \qquad b = 8C_0\frac{\overline{\sigma}_\xi}{\sigma_X\lambda_r}\sqrt{\frac{\mathrm{df}}{n}},$$

with c_0 so small that $R \le \min\{1/(4d), 1/24\}$ and $2R < 1/\{2(\sqrt{d}+1)\}$, and with the constant c of the corollary chosen afterwards so small that $\varepsilon^{(0)} \le R$, $b \le R/4$, and the inequalities below hold. Since $\alpha_t\sigma_X^2 \ge 1/2$, the conditions of the corollary give $\eta^{d-2} \le 2c/(d\kappa r)$ and $(1+r\kappa)\gamma \le c$, so that

$$(1+r\kappa)\gamma + (d+1)r\kappa(1+\gamma)\eta^{d-2} + Cd(1+r\kappa)R \le 5c + Cc_0 \le \frac{1}{8}.$$

Hence (61) and (62) give, for every current collection with $\varepsilon = \max_j\|D_j\|_{\mathrm{F}}/\lambda_j \le R$,

$$\frac{\|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i} \le (1-\alpha_t\sigma_X^2)\varepsilon + \frac{\alpha_t\sigma_X^2}{8}(\varepsilon+b), \qquad \frac{\|\widehat{\xi}_i\|_{\mathrm{F}}}{\lambda_i} \le \frac{\alpha_t\sigma_X^2}{8}(9\varepsilon+b),$$

which is (36) with $\alpha_t\sigma_X^2$ in place of α_t. The derivation of (37)–(38) and Step 2 of the proof of Corollary G.3 apply verbatim with this replacement and give

$$\frac{\|\widehat{\xi}_i\|_{\mathrm{F}}}{\|\widehat{\mathcal{T}}_i\|_{\mathrm{F}}} < \frac{1}{2(\sqrt{d}+1)}, \qquad \frac{\|\widehat{\mathcal{T}}_i^{+} - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i} \le \left(1 - \frac{3\alpha_t\sigma_X^2}{4}\right)\varepsilon + \alpha_t\sigma_X^2 b, \qquad 0 \le \varepsilon \le R,$$

and then

$$\varepsilon^{(t)} \le \left\{\prod_{s=0}^{t-1}\left(1 - \frac{3\alpha_s\sigma_X^2}{4}\right)\right\}\varepsilon^{(0)} + 2b, \qquad t \ge 0.$$

This proves the corollary with $C = 16C_0$. Under the assumptions of Corollary 4.8, $|1 - \alpha_t\sigma_X^2| \le c \le 1/3$ gives $1 - 3\alpha_t\sigma_X^2/4 \le 1/2$, which proves that corollary.

For RGN, $\alpha_t\sigma_X^2 = 1$ and the conditions of the corollary give $2r\kappa\gamma+(d+1)r\kappa\eta^{d-2}+Cd(1+r\kappa)R \le (10/3)c + Cc_0 \le 1/4$, so (64) gives, for every current collection with $\varepsilon = \max_j \|D_j\|_{\mathrm{F}}/\lambda_j \le R$,

$$\frac{\|\widehat{W}_i - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i} \le \frac{\varepsilon}{4} + \frac{b}{4}, \qquad \frac{\|\widehat{\xi}_i\|_{\mathrm{F}}}{\lambda_i} \le \frac{5\varepsilon}{4} + \frac{b}{4},$$
$$\frac{\|\widehat{\xi}_i\|_{\mathrm{F}}}{\|\widehat{\mathcal{T}}_i\|_{\mathrm{F}}} \le \frac{5R/4 + b/4}{1-R} < 2R < \frac{1}{2(\sqrt{d}+1)}, \qquad \frac{\|\widehat{\xi}_i\|_{\mathrm{F}}^2}{\lambda_i\|\widehat{\mathcal{T}}_i\|_{\mathrm{F}}} \le \frac{(5\varepsilon/4 + b/4)^2}{1-\varepsilon} \le 5\varepsilon^2 + b^2.$$

Hence (6) applies, and $b/4 + b^2 \le b$ gives, after enlarging C,

$$\frac{\|\widehat{\mathcal{T}}_i^{+} - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i} \le \Psi(\varepsilon) := \{2r\kappa\gamma + (d+1)r\kappa\eta^{d-2}\}\varepsilon + Cd(1+r\kappa)\varepsilon^2 + b \le \frac{\varepsilon}{2} + b, \qquad 0 \le \varepsilon \le R. \tag{65}$$

Step 2 of the proof of Corollary G.3 with Ψ in place of the right-hand side of (38) proves the linear assertion of Corollary 4.9.

For the two-phase assertion, let $h = \gamma/d + \eta^{d-2}$ and $A = C_3 d(1 + \kappa r)$ with a sufficiently large universal constant C_3, and fix $C_4 \ge 4$. By (65),

$$\Psi(\varepsilon) \le \frac{A}{2}\{\varepsilon^2 + h\varepsilon\} + b, \qquad 0 \le \varepsilon \le R,$$

and $AR = C_3c_0$, $Ah = C_3(1+\kappa r)\{\gamma + d\eta^{d-2}\} \le 3C_3c$. After fixing C_3 and C_4, decrease c_0 and then c, if necessary, so that

$$AR \le \frac{1}{2}, \qquad A(C_4+1)h \le 1, \qquad b \le \frac{Ah^2}{8},$$

where the last inequality follows from the additional noise condition in Corollary 4.9. With h in place of η^{d-2}, the argument following (40) gives

$$\Psi(\varepsilon) \le A\varepsilon^2 \le \varepsilon, \qquad C_4h < \varepsilon \le R,$$
$$\Psi(\varepsilon) \le \frac{\varepsilon}{2} + b, \qquad 0 \le \varepsilon \le \min\{R, C_4h\},$$

$2b \le C_4h$, and then the quadratic-phase and linear-phase bounds of Corollary 4.9 for $t^* = \inf\{t \ge 0 : \varepsilon^{(t)} \le C_4h\}$. Enlarging the universal constant multiplying the statistical error completes the proof. □

H Proof of Lemmas

This section proves Lemma H.1, which underpins the convergence analysis.

Lemma H.1. *Suppose $d \ge 3$. Let $\mathcal{T}_i = \lambda_i u_{1,i} \otimes \cdots \otimes u_{d,i}$, $i \in [r]$, be nonzero rank-one tensors with $\lambda_i > 0$ and unit factor vectors satisfying Assumption 3.3. Let $\mathcal{T}_i^{(t)} \in \mathrm{Seg}$ have tangent space $\mathbb{T}_i^{(t)}$, and set*

$$\delta_i^{(t)} = \frac{\|\mathcal{T}_i^{(t)} - \mathcal{T}_i\|_{\mathrm{F}}}{\lambda_i}.$$

If $\max_{i\in[r]} \delta_i^{(t)} \le 1/(4d)$, then, for every $i \in [r]$,

$$\left\|\mathcal{P}_{\mathbb{T}_i^{(t)}}^{\perp}(\mathcal{T}_i)\right\|_{\mathrm{F}} \le 3d \cdot \frac{\left\|\mathcal{T}_i^{(t)} - \mathcal{T}_i\right\|_{\mathrm{F}}^2}{\lambda_i}. \tag{66}$$

Define the Segre atomic norm by

$$\|W\|_{\mathrm{Seg},*} = \inf\left\{\sum_{k=1}^{m} |a_k| : m \in \mathbb{N}, \quad W = \sum_{k=1}^{m} a_k V_k, \quad a_k \in \mathbb{R}, \quad V_k \in \mathrm{Seg}, \quad \|V_k\|_{\mathrm{F}} = 1\right\}.$$

Then

$$\left\|\mathcal{P}_{\mathbb{T}_i^{(t)}}^{\perp}(\mathcal{T}_i^{(t)}-\mathcal{T}_i)\right\|_{\mathrm{Seg},*}=\left\|\mathcal{P}_{\mathbb{T}_i^{(t)}}^{\perp}(\mathcal{T}_i)\right\|_{\mathrm{Seg},*}\le 4d\,\frac{\|\mathcal{T}_i^{(t)}-\mathcal{T}_i\|_{\mathrm{F}}^2}{\lambda_i}. \tag{67}$$

For every pair $i\neq j$,

$$\left\|\mathcal{P}_{\mathbb{T}_i^{(t)}}\left(\mathcal{T}_j^{(t)}-\mathcal{T}_j\right)\right\|_{\mathrm{F}}\le\left\{(d+1)\eta^{d-2}+2d\delta_i^{(t)}+4\delta_j^{(t)}\right\}\left\|\mathcal{T}_j^{(t)}-\mathcal{T}_j\right\|_{\mathrm{F}}. \tag{68}$$

Proof. Fix $i\in[r]$. Write

$$\mathcal{T}_i=\lambda_i u_{1,i}\otimes\cdots\otimes u_{d,i},\qquad \mathcal{T}_i^{(t)}=\lambda_i^{(t)}u_{1,i}^{(t)}\otimes\cdots\otimes u_{d,i}^{(t)},$$

where $\lambda_i^{(t)}=\|\mathcal{T}_i^{(t)}\|_{\mathrm{F}}>0$ and all factor vectors have unit Euclidean norm. Since $\|\mathcal{T}_i^{(t)}\|_{\mathrm{F}}\ge(1-\delta_i^{(t)})\lambda_i$, the polarization identity gives $2\langle\mathcal{T}_i^{(t)},\mathcal{T}_i\rangle\ge 2(1-\delta_i^{(t)})\lambda_i^2>0$. We therefore choose the factor signs so that

$$c_\ell=\langle u_{\ell,i},u_{\ell,i}^{(t)}\rangle\ge 0,\qquad \ell=1,\dots,d.$$

Let

$$s_\ell=(1-c_\ell^2)^{1/2}.$$

If $s_\ell>0$, write

$$u_{\ell,i}=c_\ell u_{\ell,i}^{(t)}+s_\ell w_\ell,\qquad w_\ell\perp u_{\ell,i}^{(t)},\qquad \|w_\ell\|_2=1.$$

If $s_\ell=0$, set $w_\ell=0$. Terms containing this zero coefficient are omitted from the expansion below.

Let

$$U_i^{(t)}=u_{1,i}^{(t)}\otimes\cdots\otimes u_{d,i}^{(t)},\qquad U_i=u_{1,i}\otimes\cdots\otimes u_{d,i}.$$

Since $\mathcal{T}_i^{(t)}=\lambda_i^{(t)}U_i^{(t)}\in\mathbb{T}_i^{(t)}$, we have

$$\mathcal{P}_{\mathbb{T}_i^{(t)}}^{\perp}(\mathcal{T}_i^{(t)}-\mathcal{T}_i)=-\lambda_i\mathcal{P}_{\mathbb{T}_i^{(t)}}^{\perp}U_i.$$

Expanding U_i around $U_i^{(t)}$ gives

$$U_i=\sum_{S\subseteq[d]}\left(\prod_{\ell\in S}s_\ell\prod_{\ell\notin S}c_\ell\right)V_S, \tag{69}$$

where

$$V_S=\bigotimes_{\ell=1}^{d}v_{\ell,S},\qquad v_{\ell,S}=\begin{cases}w_\ell, & \ell\in S,\\ u_{\ell,i}^{(t)}, & \ell\notin S.\end{cases}$$

Each retained V_S has unit Frobenius norm. For $S\neq S'$ and $\ell\in S\triangle S'$, the product $\langle V_S,V_{S'}\rangle=\prod_{m=1}^{d}\langle v_{m,S},v_{m,S'}\rangle$ contains the factor $\langle w_\ell,u_{\ell,i}^{(t)}\rangle=0$, so the retained V_S are orthonormal. By (3), $\mathbb{T}_i^{(t)}$ is spanned by $U_i^{(t)}=V_\emptyset$ and the tensors $u_{1,i}^{(t)}\otimes\cdots\otimes h_k\otimes\cdots\otimes u_{d,i}^{(t)}$ with $h_k\perp u_{k,i}^{(t)}$. Hence $V_S\in\mathbb{T}_i^{(t)}$ if $|S|\le 1$. If $|S|\ge 2$, then for every $k\in[d]$ some $\ell\in S\setminus\{k\}$ gives a factor $\langle w_\ell,u_{\ell,i}^{(t)}\rangle=0$ in

$$\langle V_S,U_i^{(t)}\rangle\quad\text{and}\quad\langle V_S,u_{1,i}^{(t)}\otimes\cdots\otimes h_k\otimes\cdots\otimes u_{d,i}^{(t)}\rangle=\langle v_{k,S},h_k\rangle\prod_{m\neq k}\langle v_{m,S},u_{m,i}^{(t)}\rangle,$$

so $V_S\in\mathbb{T}_i^{(t),\perp}$. Therefore,

$$\mathcal{P}_{\mathbb{T}_i^{(t)}}^{\perp}U_i=\sum_{|S|\ge 2}\left(\prod_{\ell\in S}s_\ell\prod_{\ell\notin S}c_\ell\right)V_S.$$

By the definition of the Segre atomic norm and $0\le c_\ell\le 1$,

$$\left\|\mathcal{P}_{\mathbb{T}_i^{(t)}}^{\perp}(\mathcal{T}_i^{(t)}-\mathcal{T}_i)\right\|_{\mathrm{Seg},*}\le\lambda_i\sum_{|S|\ge 2}\prod_{\ell\in S}s_\ell.$$

It remains to bound the last sum. Let $q = \sum_{\ell=1}^{d} s_\ell^2$, $C_\Pi = \prod_{\ell=1}^{d} c_\ell$. Since $\langle \mathcal{T}_i^{(t)}, \mathcal{T}_i \rangle = \lambda_i^{(t)} \lambda_i C_\Pi$, we have

$$\left(\delta_i^{(t)}\right)^2 = \frac{\|\mathcal{T}_i^{(t)} - \mathcal{T}_i\|_{\mathrm{F}}^2}{\lambda_i^2} = \frac{\{\lambda_i^{(t)}\}^2}{\lambda_i^2} + 1 - \frac{2\lambda_i^{(t)}}{\lambda_i} C_\Pi = \left(\frac{\lambda_i^{(t)}}{\lambda_i} - C_\Pi\right)^2 + 1 - C_\Pi^2 \geq 1 - C_\Pi^2.$$

Since $C_\Pi^2 = \prod_{\ell=1}^{d}(1 - s_\ell^2) \leq \exp(-q)$, we obtain $\left(\delta_i^{(t)}\right)^2 \geq 1 - \exp(-q)$. Because $\delta_i^{(t)} \leq 1/(4d) \leq 1/2$, it follows that $q \leq -\log\left(1 - \left(\delta_i^{(t)}\right)^2\right) \leq 2\left(\delta_i^{(t)}\right)^2$. By the Cauchy–Schwarz inequality,

$$\sum_{\ell=1}^{d} s_\ell \leq \sqrt{dq} \leq \sqrt{2d}\, \delta_i^{(t)} \leq \frac{1}{2}.$$

Hence

$$\sum_{|S|\geq 2} \prod_{\ell \in S} s_\ell \leq \exp\Big(\sum_{\ell=1}^{d} s_\ell\Big) - 1 - \sum_{\ell=1}^{d} s_\ell \leq \Big(\sum_{\ell=1}^{d} s_\ell\Big)^2 \leq dq \leq 2d\left(\delta_i^{(t)}\right)^2.$$

Consequently,

$$\left\|\mathcal{P}_{\mathbb{T}_i^{(t)}}^{\perp}(\mathcal{T}_i^{(t)} - \mathcal{T}_i)\right\|_{\mathrm{Seg},*} \leq 2d\lambda_i \left(\delta_i^{(t)}\right)^2 = 2d\frac{\|\mathcal{T}_i^{(t)} - \mathcal{T}_i\|_{\mathrm{F}}^2}{\lambda_i}.$$

Since $\mathcal{P}_{\mathbb{T}_i^{(t)}}^{\perp} \mathcal{T}_i^{(t)} = 0$ and $\|W\|_{\mathrm{F}} \leq \|W\|_{\mathrm{Seg},*}$ for every tensor W, the preceding display gives

$$\|\mathcal{P}_{\mathbb{T}_i^{(t)}}^{\perp} \mathcal{T}_i\|_{\mathrm{F}} \leq \|\mathcal{P}_{\mathbb{T}_i^{(t)}}^{\perp} \mathcal{T}_i\|_{\mathrm{Seg},*} = \|\mathcal{P}_{\mathbb{T}_i^{(t)}}^{\perp}(\mathcal{T}_i^{(t)} - \mathcal{T}_i)\|_{\mathrm{Seg},*} \leq 2d\frac{\|\mathcal{T}_i^{(t)} - \mathcal{T}_i\|_{\mathrm{F}}^2}{\lambda_i},$$

which proves (66) and (67) because $2d \leq 3d \leq 4d$.

Fix distinct $i, j \in [r]$ and write $D_j = \mathcal{T}_j^{(t)} - \mathcal{T}_j$. Let $\mathbb{T}_i$ and $\mathbb{T}_j$ be the tangent spaces at $\mathcal{T}_i$ and $\mathcal{T}_j$, respectively. For $k \in [d]$, let $\mathcal{P}_{k,i} = u_{k,i} u_{k,i}^{\top}$ and $\mathcal{P}_{k,i}^{\perp} = I_{p_k} - \mathcal{P}_{k,i}$, and define the mutually orthogonal projectors

$$\mathsf{P}_{0,i} = \bigotimes_{\ell=1}^{d} \mathcal{P}_{\ell,i}, \qquad \mathsf{P}_{k,i} = \mathcal{P}_{k,i}^{\perp} \bigotimes_{\ell \neq k} \mathcal{P}_{\ell,i}, \qquad k \in [d].$$

Then $\mathcal{P}_{\mathbb{T}_i} = \sum_{k=0}^{d} \mathsf{P}_{k,i}$, and the analogous decomposition holds for $\mathcal{P}_{\mathbb{T}_j}$. Assumption 3.3 and $0 \leq \eta \leq 1$ imply, for $k, k' \in \{0, \ldots, d\}$,

$$\|\mathsf{P}_{k,i}\mathsf{P}_{k',j}\| \leq \begin{cases} \eta^d, & k = k' = 0, \\ \eta^{d-1}, & k = 0 < k' \text{ or } k' = 0 < k, \\ \eta^{d-1}, & k = k' \geq 1, \\ \eta^{d-2}, & k, k' \geq 1,\ k \neq k', \end{cases} \leq \eta^{d-2}.$$

For $Z \in \mathbb{T}_j$, write $Z = \sum_{k'=0}^{d} Z_{k'}$ with $Z_{k'} = \mathsf{P}_{k',j} Z$. Orthogonality and Cauchy–Schwarz give

$$\|\mathcal{P}_{\mathbb{T}_i} Z\|_{\mathrm{F}}^2 = \sum_{k=0}^{d} \left\| \sum_{k'=0}^{d} \mathsf{P}_{k,i} Z_{k'} \right\|_{\mathrm{F}}^2 \leq (d+1)^2 \eta^{2(d-2)} \sum_{k'=0}^{d} \|Z_{k'}\|_{\mathrm{F}}^2.$$

Consequently,

$$\|\mathcal{P}_{\mathbb{T}_i} \mathcal{P}_{\mathbb{T}_j}\| \leq (d+1)\eta^{d-2}. \tag{70}$$

Since $\mathcal{P}_{\mathbb{T}_j}^{\perp} \mathcal{T}_j = 0$, we have $\mathcal{P}_{\mathbb{T}_j}^{\perp} D_j = \mathcal{P}_{\mathbb{T}_j}^{\perp} \mathcal{T}_j^{(t)}$. Let $\theta_{\ell,j} \in [0, \pi/2]$ be the angle between $u_{\ell,j}$ and $u_{\ell,j}^{(t)}$, and let $q_j = \sum_{\ell=1}^{d} \sin^2 \theta_{\ell,j}$. Applying (69) and the orthogonality argument following it with the roles of $\mathcal{T}_j$ and $\mathcal{T}_j^{(t)}$ interchanged, we obtain

$$\frac{\|\mathcal{P}_{\mathbb{T}_j}^{\perp} D_j\|_{\mathrm{F}}^2}{\|\mathcal{T}_j^{(t)}\|_{\mathrm{F}}^2} = \sum_{|S|\geq 2} \prod_{\ell \in S} \sin^2 \theta_{\ell,j} \prod_{\ell \notin S} \cos^2 \theta_{\ell,j} \leq \sum_{m \geq 2} \frac{q_j^m}{m!} = \exp(q_j) - 1 - q_j.$$

The bound on q derived above depends only on these angles, so $q_j \le 2\{\delta_j^{(t)}\}^2 \le 1/2$. Since also $\|\mathcal{T}_j^{(t)}\|_{\mathrm{F}} \le (1+\delta_j^{(t)})\lambda_j \le 2\lambda_j$,

$$\|\mathcal{P}_{\mathbb{T}_j}^{\perp} D_j\|_{\mathrm{F}} \le \|\mathcal{T}_j^{(t)}\|_{\mathrm{F}} \left\{\exp(q_j) - 1 - q_j\right\}^{1/2} \le 4\lambda_j \{\delta_j^{(t)}\}^2 = 4\delta_j^{(t)} \|D_j\|_{\mathrm{F}}. \tag{71}$$

Here the second inequality uses $\exp(q_j) - 1 - q_j \le q_j^2$ for $0 \le q_j \le 1/2$.

For component i, choose the factor signs so that the angle $\theta_{\ell,i}$ between $u_{\ell,i}$ and $u_{\ell,i}^{(t)}$ belongs to $[0, \pi/2]$. Let $Q_{\ell,i}$ be the planar orthogonal rotation mapping $u_{\ell,i}$ to $u_{\ell,i}^{(t)}$, extended by the identity on the orthogonal complement of their span, and set $Q_i = Q_{1,i} \otimes \cdots \otimes Q_{d,i}$. Then $\mathbb{T}_i^{(t)} = Q_i \mathbb{T}_i$, so $\mathcal{P}_{\mathbb{T}_i^{(t)}} = Q_i \mathcal{P}_{\mathbb{T}_i} Q_i^{\top}$ and $\mathcal{P}_{\mathbb{T}_i}$ have equal rank. The next display uses $\|P - P'\| = \|P^{\perp} P'\|$ for orthogonal projectors P, P' of equal rank, the identities

$$\mathcal{P}_{\mathbb{T}_i}^{\perp} \mathcal{P}_{\mathbb{T}_i^{(t)}} = \mathcal{P}_{\mathbb{T}_i}^{\perp} (Q_i - I) \mathcal{P}_{\mathbb{T}_i} Q_i^{\top},$$

$$Q_i - I = \sum_{k=1}^{d} \Big(\bigotimes_{\ell<k} Q_{\ell,i} \Big) \otimes (Q_{k,i} - I) \otimes \Big(\bigotimes_{\ell>k} I_{p_\ell} \Big),$$

$\|Q_{\ell,i}\| = 1$, and $\|Q_{\ell,i} - I\| = 2\sin(\theta_{\ell,i}/2) \le \theta_{\ell,i}$:

$$\begin{aligned} \|\mathcal{P}_{\mathbb{T}_i^{(t)}} - \mathcal{P}_{\mathbb{T}_i}\| = \|\mathcal{P}_{\mathbb{T}_i}^{\perp} \mathcal{P}_{\mathbb{T}_i^{(t)}}\| &\le \|Q_i - I\| \le \sum_{\ell=1}^{d} \|Q_{\ell,i} - I\| \le \sum_{\ell=1}^{d} \theta_{\ell,i} \\ &\le \frac{\pi}{2} \sqrt{d \sum_{\ell=1}^{d} \sin^2 \theta_{\ell,i}} \le \pi \sqrt{\frac{d}{2}}\, \delta_i^{(t)} \le 2d\delta_i^{(t)}. \end{aligned} \tag{72}$$

The fourth inequality uses $\theta \le (\pi/2)\sin\theta$ for $\theta \in [0, \pi/2]$ and the Cauchy–Schwarz inequality. The fifth uses $\sum_\ell \sin^2 \theta_{\ell,i} = q \le 2\{\delta_i^{(t)}\}^2$ from the proof of (67), and the last uses $d \ge 3$.

Finally, (70)–(72) give

$$\begin{aligned} \|\mathcal{P}_{\mathbb{T}_i^{(t)}} D_j\|_{\mathrm{F}} &\le \|\mathcal{P}_{\mathbb{T}_i} \mathcal{P}_{\mathbb{T}_j} D_j\|_{\mathrm{F}} + \|\mathcal{P}_{\mathbb{T}_i} \mathcal{P}_{\mathbb{T}_j}^{\perp} D_j\|_{\mathrm{F}} + \|(\mathcal{P}_{\mathbb{T}_i^{(t)}} - \mathcal{P}_{\mathbb{T}_i}) D_j\|_{\mathrm{F}} \\ &\le \left\{ (d+1)\eta^{d-2} + 4\delta_j^{(t)} + 2d\delta_i^{(t)} \right\} \|D_j\|_{\mathrm{F}}. \end{aligned}$$

This proves (68).

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions and scope, including the unified CP tensor recovery framework, the RGD/RGN algorithms, local convergence guarantees, and empirical validation. The claims are supported by the theory in Section 4 and the experiments in Section 5 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses key limitations and future directions in Section 6, including known-rank assumptions, fixed step sizes, and the computational cost of tangent-space projections and retractions in very high dimensions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The main assumptions are stated in the theoretical results, including incoherence, local initialization, noise, and regression design conditions. Complete proofs of the main theorems, corollaries, and supporting lemmas are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper reports the simulation settings, real-data construction, hyperparameters, folds, evaluation metrics, and compute environment in Section 5 and the appendix. The code repository provides the scripts that reproduce every table and figure.

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code and the scripts that download and process the public trade data are available at `https://github.com/kxu-stat/SegreCP` (see Code and Data Availability), and the simulation results and the prepared trade data are in the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings are specified in Section 5, Appendix C, Appendix D, and Appendix E.3, including ranks, tensor dimensions, noise/coherence grids, replications, step sizes, cross-validation splits, and baseline methods.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Synthetic results are summarized over 20 independent replications, and Tables 6 and 7 report the mean and standard deviation of the error at the last iteration in each setting with noise. The real-data comparison reports five-fold cross-validation means and standard errors of the fold-specific test RMSE (Table 1).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix C and Appendix E.2 report the hardware, software and running times.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public aggregate trade and development data and synthetic simulations, and we have reviewed the NeurIPS Code of Ethics. We do not release private, sensitive, or personally identifiable data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This work is primarily methodological and studies tensor recovery algorithms. It does not introduce a deployed system or application with a direct societal intervention. The empirical trade analysis uses public aggregate data for evaluation.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release high-risk models, generative systems, scraped personal datasets, or tools with obvious dual-use misuse risk. The released materials are code, synthetic simulation results, and public aggregate-data analysis files.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the software packages, baselines and data sources. The R packages rTensor 1.5.0 and MultiwayRegression 1.2 are released under the MIT and GPL-3 licenses. CEPII BACI (release V202601) and Gravity (release V202211) are released under the Etalab Open License 2.0, World Bank WDI and LPI under CC BY 4.0, and WITS/UNCTAD TRAINS tariffs under the WITS terms of use, so the raw WITS files are not redistributed.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The R package `SegreCP` is a new asset. It is documented in its README and reference manual and released under the MIT license. The simulation results and the prepared trade data in the supplementary material are documented in its README.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or experiments with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subject experiments, interventions, or participant data collection. It uses public aggregate datasets and synthetic simulations.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology, theory, algorithms, and experiments do not rely on LLMs as a component of the research method.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.