Computation over Tensor Stiefel Manifold: A Preliminary Study

Xianpeng Mao∗, Ying Wang†, Yuning Yang‡†

June 20, 2022

Abstract

Let * denote the t-product [25] between two third-order tensors. The purpose of this work is to study fundamental computation over the set \( \mathbb{St}(n, p, l) := \{X \in \mathbb{R}^{n \times p \times l} \mid X^T * X = I\} \), where \( X \) is a third-order tensor of size \( n \times p \times l \) (\( n \geq p \)) and \( I \) is the identity tensor. It is first verified that \( \mathbb{St}(n, p, l) \) endowed with the usual Frobenius norm forms a Riemannian manifold, which is termed as the (third-order) tensor Stiefel manifold in this work. We then derive the tangent space, Riemannian gradient, and Riemannian Hessian on \( \mathbb{St}(n, p, l) \). In addition, formulas of various retractions based on t-QR, t-polar decomposition, t-Cayley transform, and t-exponential, as well as vector transports, are presented. It is expected that analogous to their matrix counterparts, the formulas derived in this study may serve as building blocks for analyzing optimization problems over the tensor Stiefel manifold and designing Riemannian algorithms for them.

Keywords: tensor; t-product; Stiefel manifold; retraction; vector transport; manifold optimization

AMS subject classifications. 90C26, 15A69, 41A50, 65K05, 90C30

1 Introduction

Higher-order tensors play important roles in linear and multilinear algebra, statistics, optimization, machine learning, and engineering [8,9,27,41]. However, the notion of multiplication between tensors was unclear based on the traditional tensor computation; this prevents the extensions of several matrix operations to higher-order tensors. Such a problem was addressed by Kilmer, Martin, Braman, and their coauthors, who proposed a type of multiplication, termed the t-product, between third-order tensors [4,23,25]. The t-product allows the possibility of usual notions and properties of matrices living in the tensor world. For example, the authors of [23,25] also defined notions such as inverse tensors, orthogonal tensors, tensor transpose, and proposed t-SVD and t-QR decomposition. [32] proposed tensor spectral norm, nuclear norm in the sense of the t-product, and presented an efficient way for computing t-SVD; the authors used these tools to develop tensor robust PCA models. Based on the t-product, t-Jordan canonical form and t-Drazin inverse were generalized to third-order tensors [36]; the tensor t-functions were established in [33,35]; the t-eigenvalues and related properties were studied in [30]. Recently, the concepts of t-Hessian tensor, t-convexity, and t-(semi)definiteness were defined in [53]. The t-SVD was further investigated in [39]. Compared with other tensor (decomposition) models, the t-product based one allows us to deal with tensors quite similar to their matrix counterparts; moreover, most basic operations can be efficiently implemented via FFT [25].

∗School of Physical Science and Technology, Guangxi University, Nanning 530004, China
†College of Mathematics and Information Science, Guangxi University, Nanning, 530004, China
‡Corresponding author: Yuning Yang, yyang@gxu.edu.cn
Let $*$ denote the t-product and $\text{St}((n,p,l))$ the set of partially orthogonal tensors:

$$\text{St}((n,p,l)) := \{ \mathcal{X} \in \mathbb{R}^{n \times p \times l} | \mathcal{X}^\top \ast \mathcal{X} = \mathcal{I}, n \geq p \},$$

(1.1)

where $\mathcal{X}$ is a third-order tensor of size $n \times p \times l$ ($n \geq p$), $\ast$ denotes the transpose, and $\mathcal{I}$ represents the identity tensor that will be detailed later. Several (potential) tensor problems take the form:

$$\min_{\mathcal{X} \in \mathbb{R}^{n \times p \times l}} f(\mathcal{X}) \text{ s.t. } \mathcal{X} \in \text{St}((n,p,l)),$$

(1.2)

such as tensor approximation (with missing entries), joint diagonalization, joint t-SVD, (sparse) tensor PCA, and beyond; these will be introduced in Sect. 5. In fact, when $l = 1$, (1.2) boils down to optimization over the orthogonal matrix constraint, namely, the Stiefel manifold, which is a special Riemannian manifold. In recent years, Riemannian manifold optimization has drawn much attention; see, e.g., [7,10,16,17,19,20]; fundamental concepts, tools, and algorithms can be found in [1,3,44]. Classical methods in the Euclidean space, including the gradient descent/conjugate gradient/(quasi-)Newton’s method/trust region method, have been generalized to Riemannian manifolds. It is known that the orthogonal projection operator, Riemannian gradient and Riemannian Hessian, the retraction, and the vector transport are fundamental tools for Riemannian manifold optimization.

Riemannian structure and computation have also been studied in the context of tensors. For example, [45] investigated the geometry of the hierarchical Tucker format of tensors; [14] considered the manifold of tensors of tensor-train (TT) format of fixed TT-rank; a Riemannian conjugate gradient was developed in [28] for tensor completion of Tucker format of fixed multilinear-rank; for the same task and the same format, [13] proposed a Riemannian trust-region method, while in the TT format of fixed rank, [43] proposed a Riemannian conjugate gradient; such a method was also developed for the canonical polyadic approximation [5]; just to name a few. On the other hand, in the t-product sense, [11] recently proposed a Grassmannian optimization based approach for online tensor completion and tracking, and [42] devised a Riemannian conjugate gradient descent over the manifold of fixed transformed multi-rank tensors for tensor completion.

However, although optimization over the (matrix) Stiefel manifold develops rapidly, over the set $\text{St}((n,p,l))$ in (1.1), it has not been studied yet. In view of the aforementioned progress on t-product based tensor theory and Riemannian optimization, as well as the real-world demand, this work intends to make a study concerning $\text{St}((n,p,l))$. Specifically, our progress is:

1. We first show that $\text{St}((n,p,l))$ endowed with the Frobenius norm forms a Riemannian manifold, which is termed as the (third-order) tensor Stiefel manifold in this paper;

2. The parameterized form of the tangent space of the tensor Stiefel manifold is established. Furthermore, the orthogonal projector operator is studied, based on which we deduce the Riemannian gradient and Riemannian Hessian of an objective function over $\text{St}((n,p,l))$ from the Euclidean gradient and Hessian of an extended objective function on the ambient Euclidean space;

3. Several retractions based on different tensor decompositions, such as t-QR and t-polar decomposition, t-Cayley transform, and the geodesic based on t-exponential, which map points from the tangent space of $\text{St}((n,p,l))$ to $\text{St}((n,p,l))$, are derived. The vector transports upon various retractions, which compare tangent vectors at distinct points on the manifold, are also obtained;

4. As byproducts, we define skew-symmetric tensors, t-polar decomposition, and related properties; the analytical solution of the tensor Sylvester equation is derived.

Owing to the nice properties of the t-product, the derived formulas have similar forms as their matrix counterparts. It is expected that these formulas can be useful for analyzing optimization problems over the tensor Stiefel manifold and designing Riemannian algorithms for them. In particular, as these formulas are consistent with their matrix counterparts, the recently developed algorithms over the matrix Stiefel manifold, such as [7,10,16], might be parallely transplanted to the tensor setting without many modifications.
The rest of this work is organized as follows. In Sect. 2, we summarize preliminaries on t-product, t-exponential and tensor decompositions, such as t-polar decomposition and t-QR decomposition, which are used throughout this paper, while preliminaries on Riemannian manifold are introduced in Sect. 3. In Sect. 4, we study the tensor Stiefel manifold (1.1), the tangent space of (1.1), the orthogonal projector operator, various retractions and vector transports. In Sect. 5, several examples of (1.2) and related optimization problems are presented. In Sect. 6, we conducted preliminary numerical experiments to verify the derived formulas. Finally, some concluding remarks are given in Sect. 7.

2 t-Product based Tensor Computation

Notation. Throughout this paper, scalars are written as small letters (a, b, · · · ), vectors are written as boldface lowercase letters (x, y, · · · ), matrices correspond to italic capitals (A, B, · · · ), tensors are written as calligraphic capitals (A, B, · · · ), and manifolds are written as Ralph Smith’s formal script font (Α, Ω, · · · ). \( \mathbb{R}^{n \times p \times l}(\mathbb{C}^{n \times p \times l}) \) denotes the space of \( n \times p \times l \) real (complex) tensors. The \((i, j, k)\) entry of \( A \) is denoted as \( a_{ijk} \).

For any positive integer \( l \), denote \([l] := \{1, 2, \cdots, l\}\). For a third-order tensor \( A \in \mathbb{R}^{n \times p \times l} \), \( A(i) := A(:,:,i) \), \( i \in [l] \) represents each frontal slice, which is defined by fixing the third index and varying the first two. The inner product \( (A, B) \) between two real tensors \( A \) and \( B \) of the same size is the sum of entry-wise product and the Frobenius norm \( ||A||_F = (A, A)^{1/2} \). \( I_p(O_p) \) denotes a unit (zero) matrix of dimension \( p \times p \). \( A^T(A^H) \), \( \text{conj}(A) \), and \( A^\dagger \) represent transpose, conjugate transpose, conjugate, and Moore-Penrose generalized inverse of the matrix \( A \), respectively. The diagonal of \( A \in \mathbb{R}^{n \times n} \) is defined as \( \text{diag}(A) \in \mathbb{R}^{n \times n} \), with all the non-diagonal entries of \( A \) zeroed out. \( \odot \) means the Kronecker product of matrices. \( \odot \) represents the Khatri-Rao product for partitioned matrices [52].

A is called f-square if \( n = p \). A tensor \( A \in \mathbb{C}^{n \times p \times l} \) with \( n \geq p \) is called “f-full rank \( p \)” if each frontal slice of \( A \) is of full rank \( p \). The sets of f-full rank \( p \) complex tensors are denoted as \( \mathbb{C}^{n \times p \times l} \). A tensor \( A \in \mathbb{R}^{n \times p \times l}(\mathbb{C}^{n \times p \times l}) \) is called “f-diagonal” or “f-upper triangular”, if each frontal slice of \( A \) is diagonal or upper triangular, respectively. The sets of f-upper triangular real tensors (with strictly positive diagonal elements) are denoted as \( \mathbb{R}^{n \times n \times l}(\mathbb{R}_{upp}^{n \times n \times l}) \). The sets of f-upper triangular complex tensors (with strictly positive diagonal elements) are denoted as \( \mathbb{C}^{n \times n \times l}(\mathbb{C}_{upp}^{n \times n \times l}) \).

2.1 t-Product for Third-Order Tensors

Before giving the definition of the t-product, some preparations are needed first.

Definition 2.1. [25] The “unfold” command is anchored to the frontal slices of the tensor \( A \in \mathbb{R}^{n \times p \times l} \), i.e.,

\[
\text{unfold}(A) := \begin{bmatrix} A^{(1)} \\ \vdots \\ A^{(l)} \end{bmatrix} \in \mathbb{R}^{nl \times p}.
\]

And the operation takes “unfold” back to tensor form is the fold command: \( \text{fold}(\text{unfold}(A)) = A \).

Definition 2.2. [25] Let \( A \in \mathbb{R}^{n \times p \times l} \); then its circulant matrices is

\[
\text{bcirc}(A) := \begin{bmatrix} A^{(1)} & A^{(2)} & \cdots & A^{(2)} \\ A^{(2)} & A^{(3)} & \cdots & A^{(3)} \\ \vdots & \vdots & \ddots & \vdots \\ A^{(l)} & \cdots & \cdots & A^{(l)} \end{bmatrix} \in \mathbb{R}^{nl \times pl}.
\]

Definition 2.3. [25] The t-product between \( A \in \mathbb{R}^{n \times p \times l} \) and \( B \in \mathbb{R}^{p \times m \times l} \) is defined as

\[
A \ast B := \text{fold}(\text{bcirc}(A) \cdot \text{unfold}(B)) \in \mathbb{R}^{n \times m \times l}.
\]
Remark 2.1. [25] Throughout this paper, the notation $i$ referred to the $i$-th frontal slice of $A$, and sometimes, they are abbreviated as $\text{Diag}(\mathcal{A},i)$. Using this notation, $A = \text{fold}((A(i) : i \in [l]))$ is sometimes abbreviated as $A = \text{fold}(A(i) : i \in [l])$.

Definition 2.5. [25] The identity tensor $\mathcal{I} \in \mathbb{R}^{n \times n \times l}$ is the tensor whose first frontal slice is the $n \times n$ identity matrix and other frontal faces are zero. For a $f$-square tensor $A \in \mathbb{R}^{n \times n \times l}$, an inverse $B$ exists if it satisfies $A \ast B = B \ast A = \mathcal{I}$. $B$ is denoted as $A^{-1}$.

2.2 Fourier domain representation

The t-product based computation can be efficiently implemented by using fast Fourier transform (FFT) instead of directly computing $(\mathcal{A}(\mathcal{T}))$. Before that, the following notations are introduced first.

For any $\mathcal{A}_i \in \mathbb{R}^{m_i \times n_i \times p}$ and $\mathcal{V}_i \in \mathbb{R}^{m_i \times n_i \times p}$ with $i \in [l]$, we denote

$$\text{Diag} (\mathcal{A}_1, \cdots, \mathcal{A}_l) := \begin{bmatrix} \mathcal{A}_1 \\ \vdots \\ \mathcal{A}_l \end{bmatrix}, \quad \text{Vec} (\mathcal{V}_1, \cdots, \mathcal{V}_l) := \begin{bmatrix} \mathcal{V}_1 \\ \vdots \\ \mathcal{V}_l \end{bmatrix},$$

and sometimes, they are abbreviated as $\text{Diag}(\mathcal{A}_i : i \in [l])$ and $\text{Vec}(\mathcal{V}_i : i \in [l])$, respectively. When all $\mathcal{A}_i (\mathcal{V}_i)$ become matrices (or vectors or scalars), similar symbols are also used.

Using this notation, $A = \text{fold}((A(i) : i \in [l]))$ is sometimes abbreviated as $A = \text{fold}(A(i) : i \in [l])$.

Definition 2.6. [25] For $A \in \mathbb{R}^{n \times p \times l}$, the discrete Fourier transform (DFT) of $A$ is defined as

$$\hat{A} := \text{fold} \left( \sqrt{l} (F_{l} \otimes I_n) \text{ unfold}(A) \right) = \text{fold} \left( \sqrt{l} (F_{l} \otimes I_n) \text{ Vec} \left( A(i) : i \in [l] \right) \right) \in \mathbb{C}^{n \times p \times l},$$

(4.4)

where $F_l$ is the normalized Fourier operator and its $(i,j)$-th entry is $(F_{l})_{ij} = \frac{1}{\sqrt{l}} \omega^{(i-1)(j-1)}$ and $\omega = e^{-\frac{2 \pi i}{l}}$ is the primitive $l$-th root of unity and $i^2 = -1$. There hold $\hat{A}^{(1)} \in \mathbb{R}^{n \times p}, \hat{A}^{(i)} \in \mathbb{C}^{n \times p}, \hat{A}^{(i)} = \text{conj}(\hat{A}^{(l+2-i)}), i \in [l] \setminus \{1\}$.

Throughout this paper, we represent the linear transform (4.4) as $\hat{A} = L(A)$, i.e., we will use the notation $L(\cdot)$ to denote the DFT in (4.4).

The inverse discrete Fourier transform (IDFT) of $\hat{A}$ is defined as

$$A = \text{fold} \left( \frac{1}{\sqrt{l}} (F_{l}^{H} \otimes I_n) \text{ unfold}(\hat{A}) \right) = \text{fold} \left( \frac{1}{\sqrt{l}} (F_{l}^{H} \otimes I_n) \text{ Vec} \left( \hat{A}(i) : i \in [l] \right) \right),$$

(5.5)

where $F_{l}^{H} = F_{l}^{-1}$.

Likewise, we will use $A = L^{-1}(\hat{A})$ to represent the IDFT in (5.5).

Remark 2.1. It follows from [25] that obtaining $\hat{A}$ can be efficiently done by using FFT in Matlab: $\hat{A} = \text{fft}(A, [], 3)$. Similarly, one can compute $A$ from $\hat{A}$ using the command $A = \text{ifft}(\hat{A}, [], 3)$.

Remark 2.2. Throughout this paper, the notation $\hat{A}$ is always referred to the DFT of $A$, and $\hat{A}(i)$ is always referred to the $i$-th frontal slice of $\hat{A}$.

Remark 2.3. [53] pointed out the following relations between $A$ and $\hat{A}$:

$$\begin{align*}
\hat{A}^{(k)} &= \omega^{(k-1) \cdot 0} \hat{A}^{(1)} + \omega^{(k-1) \cdot 1} \hat{A}^{(2)} + \cdots + \omega^{(k-1) \cdot (l-1)} \hat{A}^{(l)} \\
A^{(k)} &= \frac{1}{l} \left( \bar{\omega}^{(k-1) \cdot 0} \hat{A}^{(1)} + \bar{\omega}^{(k-1) \cdot 1} \hat{A}^{(2)} + \cdots + \bar{\omega}^{(k-1) \cdot (l-1)} \hat{A}^{(l)} \right),
\end{align*}$$

(6.6)

where $k \in [l]$ and $\bar{\omega} = \omega^{-1}$. 


Proposition 2.1. (cf. [32, Sect. 2.3]) For any $A, B \in \mathbb{R}^{n \times p \times l}$,

1. $(F_i \otimes I_n) \text{bcirc}(A)(F_i^H \otimes I_p) = \text{Diag} \left( \hat{A}^{(i)} : i \in [l] \right) =: \hat{A}$;

2. $C = A * B \iff \text{Diag} \left( \hat{C}^{(i)} : i \in [l] \right) = \text{Diag} \left( \hat{A}^{(i)} \hat{B}^{(i)} : i \in [l] \right) \iff \hat{C}^{(i)} = \hat{A}^{(i)} \hat{B}^{(i)}, i \in [l]$.

Proposition 2.2. Let $A \in \mathbb{R}^{n \times p \times l}$; then $\sqrt{I}(F_i \otimes I_p) \text{unfold}(A^T) = \text{Vec} \left( (\hat{A}^{(i)})^H : i \in [l] \right)$, which means $L(A^T) = \text{fold} \left( (\hat{A}^{(i)})^H : i \in [l] \right)$.

The proof of Proposition 2.2 is left to Appendix A.1.

Remark 2.4. Combing Proposition 2.2 and item 2 of Proposition 2.1 leads to

$$C = A^T * B \iff \hat{C}^{(i)} = (\hat{A}^{(i)})^H \hat{B}^{(i)}, i \in [l].$$

2.3 Trace, t-positive (semi)definiteness, skew-symmetric, and orthogonality

Definition 2.7. (cf. [53, Def. 7], [53, Prop. 1 (b)]) Let $A \in \mathbb{R}^{n \times n \times l}$. The trace of $A$, denoted by $\text{tr}(A)$, is defined as $\text{tr}(A) := \sum_{i=1}^l \text{tr} \left( \hat{A}^{(i)} \right) = \text{tr} (\text{bcirc}(A))$.

A symmetric version of the following relation was given in [53, Rmk. 9]. Here we need a nonsymmetric version.

Proposition 2.3. For any $A, B \in \mathbb{R}^{n \times p \times l}$, $\text{tr} \left( (A^T) * B \right) = l(A, B)$.

The proof of Proposition 2.3 is left to Appendix A.4. Proposition 2.3 immediately gives that:

Proposition 2.4. For any $A, B \in \mathbb{R}^{n \times p \times l}$,

1. $\text{tr} \left( (A^T) * B \right) = \text{tr} \left( A * B^T \right) = \text{tr} \left( B^T * A \right) = \text{tr} \left( B * A^T \right)$;

2. $\langle A * B, C \rangle = \langle A, C * B^T \rangle = \langle B, A^T * C \rangle$;

3. $\langle \hat{A}, \hat{B} \rangle = l(A, B)$.

Definition 2.8. (c.f. [23, 53]) A tensor $A \in \mathbb{R}^{n \times n \times l}$ is called symmetric if $A = A^T$. The set of symmetric tensors is denoted as $\text{Sym} (\mathbb{R}^{n \times n \times l})$. $A \in \text{Sym} (\mathbb{R}^{n \times n \times l})$ is called symmetric t-positive (semi)definite if $\langle X, A * X \rangle \geq 0$ for any $X \in \mathbb{R}^{n \times 1 \times l} \setminus \{0\}$ ($X \in \mathbb{R}^{n \times n \times l}$). The sets of symmetric t-positive (semi)definite tensors are denoted as $\text{Sym} (\mathbb{R}^{n \times n \times l}_+) \setminus (\text{Sym} (\mathbb{R}^{n \times n \times l})))$.

Remark 2.5. $A \in \text{Sym} (\mathbb{R}^{n \times n \times l}_+) \setminus (\text{Sym} (\mathbb{R}^{n \times n \times l}))$ if and only if every frontal slice $\hat{A}^{(i)}$ of $\hat{A}$ in the Fourier domain is Hermitian positive (semi)definite.

Next, similar to skew-matrices, we define skew-symmetric tensors. Skew-symmetric tensors are important in deriving retractions.

Definition 2.9. A tensor $A \in \mathbb{R}^{n \times n \times l}$ is called skew-symmetric if $A = -A^T$. The set of skew-symmetric tensors are denoted as $\text{Skew} (\mathbb{R}^{n \times n \times l})$.

Lemma 2.1. $I + V^T * V \in \text{Sym} (\mathbb{R}^{n \times n \times l}_+)$ for all $V \in \mathbb{R}^{n \times p \times l}$.

Lemma 2.2. The orthogonal complement of $\text{Skew} (\mathbb{R}^{p \times p \times l})$ is $\text{Sym} (\mathbb{R}^{p \times p \times l})$. 


The proofs of Lemma 2.1 and 2.2 are left to Appendix A.2 – A.3, respectively.

**Definition 2.10.** ([25]) A tensor $\mathcal{X} \in \mathbb{R}^{n \times n \times l}$ is orthogonal if $\mathcal{X}^T \ast \mathcal{X} = \mathcal{X} \ast \mathcal{X}^T = I \in \mathbb{R}^{n \times n \times l}$. And $\mathcal{X} \in \mathbb{R}^{n \times p \times l}$ ($n \geq p$) is partially orthogonal if $\mathcal{X}^T \ast \mathcal{X} = I \in \mathbb{R}^{p \times p \times l}$. As noted in the introduction, the set of partially orthogonal tensors of size $n \times p \times l$ is denoted as $\text{St}(n,p,l)$.

$\mathcal{X}$ being orthogonal implies that every $\hat{X}^{(i)}$ in the Fourier domain is unitary [25].

Based on Proposition 2.3 and the definition of the trace, if $\mathcal{X} \in \text{St}(n,p,l)$ where $n \geq p$, then $\|\mathcal{X}\|_F$ is a constant over $\text{St}(n,p,l)$. This means the following:

**Proposition 2.5.** For any given $\mathcal{A} \in \mathbb{R}^{n \times p \times l}$ with $n \geq q$, $\min_{\mathcal{X} \in \text{St}(n,p,l)} \|\mathcal{A} - \mathcal{X}\|_F^2$ and $\max_{\mathcal{X} \in \text{St}(n,p,l)} \langle \mathcal{A}, \mathcal{X} \rangle$ are equivalent.

### 2.4 t-SVD

SVd in the t-product sense is important.

**Theorem 2.1.** (t-SVD, [25, Thm. 4.1], [32, Thm. 2.2]) Let $\mathcal{A} \in \mathbb{R}^{n \times p \times l}$. Then it can be factorized as $\mathcal{A} = \mathcal{U} \ast \mathcal{S} \ast \mathcal{V}^T$, where $\mathcal{U} \in \mathbb{R}^{n \times n \times l}$ and $\mathcal{V} \in \mathbb{R}^{p \times p \times l}$ are orthogonal tensors, and $\mathcal{S} \in \mathbb{R}^{n \times p \times l}$ is a f-diagonal tensor.

**Remark 2.6.** [32] pointed out that to keep $\mathcal{U}, \mathcal{S}$ and $\mathcal{V}$ to be real, one need to perform [32, Alg. 2]. Specifically, for $i = 1, \cdots, \lceil \frac{l+1}{2} \rceil$, let $\hat{A}^{(i)} = \hat{U}^{(i)} \hat{S}^{(i)} (\hat{V}^{(i)})^H$ be the SVD of $\hat{A}^{(i)}$, where $\lfloor x \rfloor$ is denoted as the nearest integer greater than or equal to $x$. For $i = 1 + \lfloor \frac{l+1}{2} \rfloor, \cdots, l$, $\hat{A}^{(i)} = \text{conj} \left( \hat{A}^{(l+2-i)} \right), \hat{U}^{(i)} = \text{conj} \left( \hat{U}^{(l+2-i)} \right), \hat{V}^{(i)} = \text{conj} \left( \hat{V}^{(l+2-i)} \right), \hat{S}^{(i)} = \hat{S}^{(l+2-i)}$.

As in the matrix case, the frontal slices $\hat{S}^{(i)}$ of $\hat{S}$ are all diagonal, and their diagonal entries can be chosen nonnegative. The compact t-SVD was mentioned in [25,32,35]. We formally present it here for later use.

**Theorem 2.2.** [25,32,35] Let $\mathcal{A} \in \mathbb{R}^{n \times p \times l}$ with $n \geq q$. Then it can be factorized as $\mathcal{A} = \mathcal{U} \ast \mathcal{S} \ast \mathcal{V}^T$, where $\mathcal{U} \in \mathbb{R}^{n \times p \times l}$ is partially orthogonal, $\mathcal{V} \in \mathbb{R}^{p \times p \times l}$ is orthogonal, and $\mathcal{S} \in \mathbb{R}^{n \times p \times l}$ is f-diagonal.

### 2.5 t-QR decomposition

**Theorem 2.3** (t-QR, [23]). Let $\mathcal{A} \in \mathbb{R}^{n \times p \times l}$ with $n \geq p$. Then $\mathcal{A}$ can be written as

$$\mathcal{A} = \mathcal{Q} \ast \mathcal{R},$$

where $\mathcal{Q} \in \text{St}(n,p,l)$ and $\mathcal{R} \in \mathbb{R}^{p \times p \times l}_{\text{upp}}$. If $\mathcal{A} \in L^{-1} \left( \mathbb{C}^{n \times p \times l}_{\text{upp}} \right)$, then the decomposition (2.7) is unique. Factorization of the form (2.7) is called the t-QR decomposition (t-QR for short).

**Lemma 2.3.** $L^{-1} \left( \mathbb{C}^{p \times p \times l}_{\text{upp}} \right)$ is isomorphic to $\mathbb{R}^{\frac{2l+1}{2}}$ if $l$ is odd, and $L^{-1} \left( \mathbb{C}^{p \times p \times l}_{\text{upp}} \right)$ is isomorphic to $\mathbb{R}^{\frac{2l+2}{2}+p}$ if $l$ is even.

The proofs of Theorem 2.3 and Lemma 2.3 are left to Appendix A.5 and A.6, respectively.

### 2.6 t-polar decomposition

Similar to the matrix counterpart, based on t-SVD, we can define t-polar decomposition (t-PD for short).
Theorem 2.4 (t-PD). Let $\mathbf{A} \in \mathbb{R}^{n \times p \times l}$ with $n \geq p$. Then $\mathbf{A}$ can be written as

$$\mathbf{A} = \mathcal{P} \ast \mathcal{H},$$

(2.8)

where $\mathcal{P} \in \text{St}(n, p, l)$ and $\mathcal{H} \in \text{Sym}(\mathbb{R}_+^{p \times p \times l})$. $\mathcal{H}$ is unique. Furthermore, if $\mathbf{A}^\top \ast \mathbf{A} \in \text{Sym}(\mathbb{R}_+^{p \times p \times l})$, then $\mathcal{P}$ is unique and $\mathcal{H} \in \text{Sym}(\mathbb{R}_+^{p \times p \times l})$.

Proposition 2.6. If $\mathbf{A}^\top \ast \mathbf{A} \in \text{Sym}(\mathbb{R}_+^{p \times p \times l})$, then $\mathcal{P}$ and $\mathcal{H}$ defined in Theorem 2.4 are given by

$$\mathcal{P} = \mathbf{A} \ast (\mathbf{A}^\top \ast \mathbf{A})^{-\frac{1}{2}} \quad \text{and} \quad \mathcal{H} = (\mathbf{A}^\top \ast \mathbf{A})^{\frac{1}{2}},$$

where the square root notation on tensors was defined in [53, Sect. 4.4].

Proposition 2.7. If $\mathbf{A} \in L^{-1}(\mathbb{C}_+^{n \times p \times l})$, then $\mathbf{A}^\top \ast \mathbf{A} \in \text{Sym}(\mathbb{R}_+^{p \times p \times l})$.

Theorem 2.5. Let $\mathbf{A} \in \mathbb{R}^{n \times p \times l}$ with $n \geq q$, admit the compact t-SVD $\mathbf{A} = \mathbf{U} \ast \mathcal{S} \ast \mathcal{V}^\top$. Then the optimal solution to $\max_{\mathcal{P} \in \text{St}(n, p, l)} \langle \mathbf{A}, \mathcal{P} \rangle$ is given by the t-PD of $\mathbf{A}$, namely, $\mathcal{P} = \mathbf{U} \ast \mathcal{V}^\top$.

The proofs of Theorem 2.4, Proposition 2.6, and Theorem 2.5 are left to Appendix A.7 − A.10, respectively.

2.7 t-exponential

In [33], the author defined tensor t-functions for a third-order f-square tensor $\mathbf{A} \in \mathbb{R}^{n \times n \times l}$ based on t-product. In particular, the exponential of a third-order tensor $\mathbf{A} \in \mathbb{R}^{n \times n \times l}$ is defined as follows:

$$\exp [\mathbf{A}] = \text{fold} (\text{exp} [\text{bcirc}(\mathbf{A})]) \text{unfold} (\mathcal{I}),$$

(2.9)

In [35], the authors extended the definition of tensor t-functions to arbitrary third-order tensors (not necessarily f-square). We present an equivalent definition of the t-exponential of third-order tensors for later use.

Definition 2.11. The exponential of tensor $\mathbf{A} \in \mathbb{R}^{n \times n \times l}$ based on t-product (t-exponential for short) is

$$\exp [\mathbf{A}] := \sum_{k=0}^{\infty} \frac{1}{k!} \mathbf{A}^k,$$

(2.10)

where $\mathbf{A}^k = \mathbf{A} \ast \mathbf{A} \ast \cdots \ast \mathbf{A}$ (k copies) with the convention that $\mathbf{A}^0 = \mathcal{I}$.

Remark 2.7. This is well defined, i.e., the series is convergent. The proof of well-defined property of (2.10) is left to Appendix A.11. According to the proof, (2.10) can be further written as

$$\exp [\mathbf{A}] = L^{-1} \left(\text{fold} \left(\exp \left[\mathbf{A}^{(i)}\right] : i \in [l]\right)\right) = L^{-1} \left(\text{fold} \left(\exp \left[(L(A))^{(i)}\right] : i \in [l]\right)\right).$$

(2.11)

The proof of equivalence of (2.11) and (2.9) is left to Appendix A.12.

Properties of the matrix exponential, such as those mentioned above, can be extended to the t-exponential. We list those that are needed later.

Proposition 2.8. The exponential map $\exp : \mathbb{R}^{n \times n \times l} \to \mathbb{R}^{n \times n \times l}, \mathbf{A} \mapsto \exp [\mathbf{A}]$ is smooth.

Proposition 2.9. Let $\mathbf{A} \in \mathbb{R}^{n \times p \times l}$. Then $\frac{\partial}{\partial t} \exp [t \mathbf{A}] = \exp [t \mathbf{A}] \ast \mathbf{A} = \mathbf{A} \ast \exp [t \mathbf{A}]$.

Proposition 2.10. Consider $\mathbf{A} \in \mathbb{R}^{n \times n \times l}$ and $\mathcal{X} \in \text{St}(m, n, l)$ with $m \geq n$. Then

$$\exp [\mathcal{X} \ast \mathbf{A} \ast \mathcal{X}^\top] = \mathcal{X} \ast \exp [\mathbf{A}] \ast \mathcal{X}^\top.$$

Proposition 2.11. Let $\mathbf{D}_j \in \mathbb{R}^{m \times n \times l}, j \in [p]$. Then

$$\exp [\text{Diag} (\mathbf{D}_j : j \in [p])] = \text{Diag} (\exp [\mathbf{D}_j] : j \in [p]) .$$

Proposition 2.12. Let $\mathbf{A} \in \mathbb{R}^{n \times p \times l}$. Then $(\exp [\mathbf{A}])^\top = \exp [\mathbf{A}^\top]$.

Proposition 2.13. If $\mathbf{A} \ast \mathbf{B} = \mathbf{B} \ast \mathbf{A}$, then $\exp [\mathbf{A}] \ast \exp [\mathbf{B}] = \exp [\mathbf{A} + \mathbf{B}]$.

The proofs of Proposition 2.8 − 2.13 are left to Appendix A.13 − A.18, respectively.
3 Preliminaries on Riemannian Manifold

Basic definitions and properties concerning the Riemannian manifold can be found in the books [1, 3, 44]. To be more convenient and to make the paper self-contained, we summarize the necessary ones in this section.

Definition 3.1. [44] A topological space $\mathcal{M}$ is locally Euclidean dimension $n$ if every point $p$ in $\mathcal{M}$ has a neighborhood $U$ such that there is a homeomorphism $\phi$ from $U$ onto an open subset of $\mathbb{R}^n$. A topological manifold of dimension $n$ is a Hausdorff, second countable, locally Euclidean dimension $n$ space. Especially, every vector space is a linear manifold.

Definition 3.2. [1] Let $\mathcal{N}$ be a submanifold of $\mathcal{M}$. If the manifold topology of $\mathcal{N}$ coincides with its subspace topology induced from the topological space $\mathcal{M}$, then $\mathcal{N}$ is called an embedded submanifold of the manifold $\mathcal{M}$.

Definition 3.3. [1] A tangent vector $\xi_x$ to $\mathcal{M}$ at $x$ is a mapping such that there exists a curve $\gamma$ on $\mathcal{M}$ with $\gamma(0) = x$, satisfying

$$\xi_x f := \frac{\gamma'(0) f}{\gamma'(t)} \bigg|_{t=0}, \quad \forall f \in \mathfrak{F}_x(\mathcal{M}),$$

where $\mathfrak{F}_x(\mathcal{M})$ is the set of all real-valued functions $f$ defined in a neighborhood of $x$ in $\mathcal{M}$. The tangent space $T_x \mathcal{M}$ to $\mathcal{M}$ is defined as the set of all tangent vectors to $\mathcal{M}$ at $x$. $T \mathcal{M} := \bigcup_{x \in \mathcal{M}} T_x \mathcal{M}$ is called the tangent bundle of the manifold.

Definition 3.4. [1] The differential of $F : \mathcal{M} \to \mathcal{N}$ at $x$ is a linear operator $DF(x) : T_x \mathcal{M} \to T_{F(x)} \mathcal{N}$ defined by:

$$DF(x)[v] := \frac{d}{dt} F(\gamma(t)) \bigg|_{t=0},$$

where $\gamma(t)$ is any curve on the manifold that satisfies $\gamma(0) = x$ and $\gamma'(0) = v$.

Definition 3.5. [1] A Riemannian metric $g$ is defined on each tangent space of $x$ as an inner product $g_x : T_x \mathcal{M} \times T_x \mathcal{M} \to \mathbb{R}$. $g_x(\eta, \xi) = \langle \eta, \xi \rangle_x$ where $\eta, \xi \in T_x \mathcal{M}$ and the $x$ is dropped when context permits. A Riemannian manifold is the combination $(\mathcal{M}, g)$.

Definition 3.6. [1] A smooth scalar field on a manifold $\mathcal{M}$ is a smooth function $f : \mathcal{M} \to \mathbb{R}$. The set of smooth scalar fields on $\mathcal{M}$ is denoted by $\mathfrak{F}(\mathcal{M})$. A smooth vector field $\xi$ on a manifold $\mathcal{M}$ is a smooth function from $\mathcal{M}$ to the tangent bundle $T \mathcal{M}$ that assigns to each point $x \in \mathcal{M}$ a tangent vector $\xi_x \in T_x \mathcal{M}$. Let $\mathfrak{X}(\mathcal{M})$ denote the set of smooth vector fields on $\mathcal{M}$.

Definition 3.7. [1] An affine connection $\nabla$ on a manifold $\mathcal{M}$ is a mapping $\nabla : \mathfrak{X}(\mathcal{M}) \times \mathfrak{X}(\mathcal{M}) \to \mathfrak{X}(\mathcal{M})$, which is denoted by $\nabla_{\xi} \eta$ and satisfies the following properties. For $\eta, \xi, \zeta \in \mathfrak{X}(\mathcal{M})$, $f, g \in \mathfrak{F}(\mathcal{M})$, and $a, b \in \mathbb{R}$: (i) $\nabla_{f\xi + g\eta} \zeta = f\nabla_{\xi} \zeta + g\nabla_{\eta} \zeta$, (ii) $\nabla_{(a\xi + b\zeta)} \eta = a\nabla_{\xi} \eta + b\nabla_{\zeta} \eta$, and (iii) $\nabla_{\eta}(fg \xi) = (\eta f) \xi + f\nabla_{\eta} \xi$.

The vector field $\nabla_{\eta} \zeta$ is called the covariant derivative of $\zeta$ with respect to $\eta$ for the affine connection $\nabla$. For a Riemannian manifold, one of the affine connections, called Riemannian connection or Levi-Civita connection, uniquely satisfies the following two additional conditions: (i) $(\nabla_{\eta} \xi - \nabla_{\xi} \eta) f = \eta(\xi f) - \xi(\eta f)$, and (ii) $\zeta(\eta, \xi) = \langle \nabla_{\eta} \xi, \zeta \rangle + \langle \eta, \nabla_{\xi} \zeta \rangle$.

Definition 3.8. Let $c : I \to \mathcal{M}$ be a smooth curve on a manifold equipped with a connection $\nabla$. There exists a unique operator $\frac{\partial}{\partial t} : \mathfrak{X}(c) \to \mathfrak{X}(c)$ which satisfies the following properties for all $Y, Z \in \mathfrak{X}(c), U \in \mathfrak{X}(\mathcal{M}), g \in \mathfrak{F}(I)$, and $a, b \in \mathbb{R}$: (i) $\mathbb{R}$-linearity: $\frac{\partial}{\partial t}(aY + bZ) = a\frac{\partial}{\partial t} Y + b\frac{\partial}{\partial t} Z$; (ii) Leibniz rule: $\frac{\partial}{\partial t}(gZ) = g'Z + g\frac{\partial}{\partial t} Z$; (iii) Chain rule: $(\frac{\partial}{\partial t}(U \circ c))(t) = \nabla_{c(t)} U$ for all $t \in I$. $\frac{\partial}{\partial t}$ is called the induced covariant derivative.

Definition 3.9. [18] The geodesic $\gamma(t)$ defined by an affine connection is a curve that satisfies

$$\ddot{\gamma}(t) := \nabla_{\dot{\gamma}(t)} \dot{\gamma}(t) := \frac{\partial^2}{\partial t^2} \gamma(t) := \frac{d}{dt} \dot{\gamma}(t) = 0.$$
Definition 3.10. [1] The Riemannian gradient \( \text{grad} f(x) \) of a function \( f \) at \( x \) is an unique vector in \( T_xM \) satisfying \( \langle \text{grad} f(x), \xi \rangle_x = Df(x)[\xi], \quad \forall \xi \in T_xM. \)

Definition 3.11. [1] The Riemannian Hessian \( \text{Hess} f(x) \) is a mapping from the tangent space \( T_xM \) to the tangent space \( T_xM : \text{Hess} f(x)[\xi] := \nabla_\xi \text{grad} f(x), \) where \( \nabla \) is the Riemannian connection on \( M. \)

Lemma 3.1. [3] For a function \( f \) defined on a submanifold \( M \) with the Euclidean metric on its tangent space, if it can be extended to the ambient Euclidean space denoted as \( \bar{f}, \) one has

\[
\text{grad} f(x) = P_x(\text{grad} \bar{f}(x)),
\]

\[
\text{Hess} f(x)[u] = P_x(D\bar{G}(x)[u]), u \in T_xM,
\]

where \( D \) is the Euclidean derivative, \( P_x \) is the orthogonal projection operator from Euclidean space \( \mathcal{E} \) to \( T_xM, \) and \( \bar{G}(x) \) denote a smooth extension of the grad \( f(x) \) to a neighborhood of \( M \) in the ambient Euclidean space.

Retraction provides a method to map the tangent vector to the next iterate on the manifold.

Definition 3.12. (cf. [1, Def. 4.1.1]) A retraction on a manifold \( M \) is a smooth mapping \( R \) from the tangent bundle \( TM \) onto \( M \) with the following properties. Let \( R_x \) denote the restriction of \( R \) to \( T_xM, \)

(i) \( R_x(0_x) = x, \) where \( 0_x \) denotes the zero element of \( T_xM, \) and (ii) \( DR_x(0_x) : T_xM \to T_xM \) is the identity map: \( DR_x(0_x)[v] = v. \)

For the embedded submanifold of a vector space, there is a simple way to construct retractions, as specified in the following lemma.

Lemma 3.2. (cf. [1, Prop. 4.1.2]) Let \( \mathcal{M} \) be an embedded manifold of a vector space \( \mathcal{E} \) and let \( \mathcal{N} \) be an abstract manifold such that \( \dim \mathcal{M} + \dim \mathcal{N} = \dim \mathcal{E}. \) Assume that there is a diffeomorphism \( \phi : \mathcal{M} \times \mathcal{N} \to \mathcal{E}, \) \( (F,G) \to \phi(F,G), \) where \( \mathcal{E} \) is an open subset of \( \mathcal{E} \) (thus \( \mathcal{E} \) is an open submanifold of \( \mathcal{E} \)), with a neutral element \( I \in \mathcal{N} \) satisfying \( \phi(I,F) = F, \forall F \in \mathcal{M}. \) Then the retraction is \( R_X(\xi) = \pi_1(\phi^{-1}(X + \xi)), \) where \( \pi_1 : \mathcal{M} \times \mathcal{N} \to \mathcal{M} : (F,G) \to F \) is the projection onto the first component, defines a retraction on \( \mathcal{M}. \)

To compare tangent vectors at distinct points on the manifold, the vector transport upon retraction \( R \) gives us a way to transport a tangent vector \( \xi \in T_xM \) to the tangent space \( T_{R_x(0)}(0) \) for some \( \eta \in T_xM. \)

Definition 3.13. (cf. [1, Def. 8.1.1]) A vector transport \( \mathcal{T} : T_xM \oplus T_0M \to T_xM : (\eta_x, \xi_x) \to \mathcal{T}_{\eta_x, \xi_x} \) associated with a retraction \( R \) is a smooth mapping satisfying the following properties for all \( x \in \mathcal{M}: \)

(i) \( \mathcal{T}_{\eta_x, \xi_x} \in T_{R_x(\eta_x)}M, \)

(ii) \( \mathcal{T}_{0_x, \xi_x} = \xi_x \) for all \( \xi_x \in T_xM, \)

and (iii) \( \mathcal{T}_{\eta_x}(a\xi_x + b\zeta) = a\mathcal{T}_{\eta_x, \xi_x} + b\mathcal{T}_{\eta_x, \zeta}. \) Vector transport by differentiated retraction is defined as

\[
\mathcal{T}_{\eta_x, \xi_x} := DR_x(\eta_x)[\xi_x] = \frac{d}{dt}R_x(\eta_x + t\xi_x)|_{t=0}.
\]

Lemma 3.3. (cf. [1, Sect. 8.1.3]) A vector transport on \( M \) associated with a retraction \( R \) is given by the orthogonal projection onto the tangent space, i.e., \( \mathcal{T}_{\eta_x, \xi_x} = P_{R_x(\eta_x)}\xi_x, \) where \( P_x \) denotes the orthogonal projector onto \( T_xM. \)

Definition 3.14. A vector transport \( \mathcal{T} \) is called isometric if it satisfies \( \langle \mathcal{T}_{\eta}(\xi), \mathcal{T}_{\eta}(\zeta) \rangle_{R_x(\eta)} = \langle \xi, \zeta \rangle_x \) for all \( \eta, \xi \in T_xM, \) where \( R \) is the retraction associated with \( \mathcal{T}. \)

4 Computation over Tensor Stiefel Manifold

In this Section, we first prove that \( \text{St} (n,p,l) \) with \( n \geq p \) is a manifold and establish the parameterized form of its tangent space in Sect. 4.1. Next, in Sect. 4.2 we further show that \( \text{St} (n,p,l) \) is a Riemannian manifold and
obtain the Riemannian gradient and Riemannian Hessian by the orthogonal projector operator. In Sect. 4.3, the geodesic and some retractions based on various tensor decompositions are derived. Finally, we construct various vector transports based on the projector operator and by differentiated the retraction in Sec. 4.4.

4.1 Manifold setting

The following theorem shows that the set \( St(n,p,l) \) with \( n \geq p \) is indeed a manifold.

**Theorem 4.1** (Manifold). For \( n \geq p \), let \( St(n,p,l) = \{ \mathcal{X} \in \mathbb{R}^{n \times p \times l} | \mathcal{X}^\top \mathcal{X} = \mathcal{I} \} \). Then \( St(n,p,l) \) is an embedded submanifold of \( \mathbb{R}^{n \times p \times l} \) of dimension

\[
\dim(St(n,p,l)) = p \left( nl - \frac{pl}{2} - \frac{1}{2|\sin(\frac{l}{2})|} \right).
\]

**Proof.** It follows from [53, Prop. 2] that

\[
\dim(\text{Sym}(\mathbb{R}^{p \times p \times l})) = \begin{cases} 
(p(pk + 1), & l = 2k \\
p(pk + \frac{h}{2} + \frac{k}{2}), & l = 2k + 1 
\end{cases} = p \left( \frac{pl}{2} + \frac{1}{2|\sin(\frac{l}{2})|} \right).
\]

(4.13)

Consider the following function:

\[
h : \mathbb{R}^{n \times p \times l} \to \text{Sym}(\mathbb{R}^{p \times p \times l}) : \mathcal{X} \mapsto h(\mathcal{X}) = \mathcal{X}^\top \mathcal{X} - \mathcal{I}.
\]

Obviously, \( h \) is smooth and the zero level set \( h^{-1}(\mathcal{O}) = St(n,p,l) \). Note that \( \text{Sym}(\mathbb{R}^{p \times p \times l}) \) is a vector space and so a linear manifold. Since \( h \) is a smooth map of manifolds, [44, Thm. 9.9] shows that if the level set \( h^{-1}(\mathcal{O}) \) is a regular level set of \( h \), then \( St(n,p,l) \) is an embedded submanifold of \( \mathbb{R}^{n \times p \times l} \) of dimension equal to \( \dim(\mathbb{R}^{n \times p \times l}) - \dim(\text{Sym}(\mathbb{R}^{p \times p \times l})) \). Recall that \( h^{-1}(\mathcal{O}) \) is a regular level set of \( h \) if only if the differential of \( h \) is surjective (cf. [44, Sect. 9.2]). To this end, for all \( \mathcal{X} \in St(n,p,l) \), consider \( Dh(\mathcal{X}) : \mathbb{R}^{n \times p \times l} \to \text{Sym}(\mathbb{R}^{p \times p \times l}) : \mathcal{V} \mapsto \frac{h(\mathcal{X} + t\mathcal{V}) - h(\mathcal{X})}{t} \).

For \( \mathcal{V} = \frac{1}{2} \mathcal{X} \ast \mathcal{A} \) with \( \mathcal{A} \in \text{Sym}(\mathbb{R}^{p \times p \times l}) \) arbitrary, there holds

\[
Dh(\mathcal{X})[\mathcal{V}] = \frac{1}{2} \mathcal{X}^\top \ast \mathcal{X} \ast \mathcal{A} + \frac{1}{2} \mathcal{A}^\top \ast \mathcal{X} \ast \mathcal{X} = \mathcal{A}.
\]

In other words, for any tensor \( \mathcal{A} \in \text{Sym}(\mathbb{R}^{p \times p \times l}) \), there exists a tensor \( \mathcal{V} \in \mathbb{R}^{n \times p \times l} \) such that \( Dh(\mathcal{X})[\mathcal{V}] = \mathcal{A} \). This confirms that the range of \( Dh(\mathcal{X}) \) is \( \text{Sym}(\mathbb{R}^{p \times p \times l}) \). Thus, \( h^{-1}(\mathcal{O}) = St(n,p,l) \) is the a nonempty regular level set, making \( St(n,p,l) \) an embedded submanifold of \( \mathbb{R}^{n \times p \times l} \) of dimension

\[
\dim(St(n,p,l)) = \dim(\mathbb{R}^{n \times p \times l}) - \dim(\text{Sym}(\mathbb{R}^{p \times p \times l})) = p \left( nl - \frac{pl}{2} - \frac{1}{2|\sin(\frac{l}{2})|} \right).
\]

\[\square\]

To apply optimization algorithms based on line search, we must consider a direction on a manifold, that is the tangent vector. The tangent space of \( St(n,p,l) \) can be parametrized as follows.

**Theorem 4.2** (Tangent space). The tangent space \( T_{\mathcal{X}} St(n,p,l) \) is a subspace of \( \mathbb{R}^{n \times p \times l} \):

\[
T_{\mathcal{X}} St(n,p,l) = \left\{ \mathcal{X} \ast \mathcal{W} + \mathcal{X}_\perp \ast \mathcal{B} \in \mathbb{R}^{n \times p \times l} \mid \mathcal{W} \in \text{Skew}(\mathbb{R}^{p \times p \times l}), \mathcal{B} \in \mathbb{R}^{(n-p) \times p \times l} \right\},
\]

where \( \mathcal{X}_\perp = L^{-1}(\mathcal{X}_\perp) \), \( \mathcal{X} = L(\mathcal{X}) \) and \( \mathcal{X}_\perp^{(i)} \in \mathbb{C}^{n \times (n-p)} \) is any matrix such that \( \text{span}(\mathcal{X}_\perp^{(i)}) = \{ \mathcal{X}_\perp^{(i)} \alpha | \alpha \in \mathbb{C}^{n-p} \} \) is the orthogonal complement of \( \text{span}(\mathcal{X}^{(i)}) = \{ \mathcal{X}^{(i)} \beta | \beta \in \mathbb{C}^p \} \).
Proof. The rank of $h$ at $\mathcal{X}$ is defined as the dimension of the range of the differential $Dh(\mathcal{X})$ (cf. [44, Sect. 8.9]), which is equal to $\dim(\Sym(\mathbb{R}^{p \times p}))$ given in (4.13). Since $\St(n,p,l)$ is defined as a level set of a constant-rank function $h : \mathbb{R}^{n \times p \times l} \rightarrow \Sym(\mathbb{R}^{p \times p \times l}) : \mathcal{X} \mapsto h(\mathcal{X}) = \mathcal{X}^\top \ast \mathcal{X} - \mathcal{I}$, it follows from [1, Sect. 3.5.7] that 

$$T_X \St(n,p,l) = \ker(Dh(\mathcal{X})) = \{ \mathcal{V} \in \mathbb{R}^{n \times p \times l} | \mathcal{X}^\top \ast \mathcal{V} + \mathcal{V}^\top \ast \mathcal{X} = \mathcal{O} \}.$$ 

Using Remark 2.4, we get 

$$\mathcal{X}^\top \ast \mathcal{V} + \mathcal{V}^\top \ast \mathcal{X} = \mathcal{O} \iff (\mathbf{\hat{X}}^{(i)})^H \mathbf{\hat{V}}^{(i)} + (\mathbf{\hat{V}}^{(i)})^H \mathbf{\hat{X}}^{(i)} = O_p, i \in [l].$$

(4.14)

Similarly, $\mathcal{X}^\top \ast \mathcal{X} = \mathcal{I}$ leads to $(\mathbf{\hat{X}}^{(i)})^H \mathbf{\hat{X}}^{(i)} = I_p, i \in [l]$ which means $\mathbf{\hat{X}}^{(i)}$ is full rank $p$. Let $\mathbf{\hat{X}}_\perp^{(i)}$ be any $n \times (n-p)$ complex matrix such that $\text{span}(\mathbf{\hat{X}}_\perp^{(i)}) = \{ \mathbf{\hat{X}}_\perp^{(i)} \alpha | \alpha \in \mathbb{C}^{n-p} \}$ is the orthogonal complement of $\text{span}(\mathbf{\hat{X}}^{(i)}) = \{ \mathbf{\hat{X}}^{(i)} \beta | \beta \in \mathbb{C}^p \}$. By the definition, $\left( \mathbf{\hat{X}}^{(i)} \beta, \mathbf{\hat{X}}_\perp^{(i)} \alpha \right) = 0$ for any $\alpha \in \mathbb{C}^{n-p}$ and $\beta \in \mathbb{C}^p$, leading to 

$$(\mathbf{\hat{X}}^{(i)})^H \mathbf{\hat{X}}_\perp^{(i)} = O_{p \times (n-p)}.$$ 

(4.15)

Since $[\mathbf{\hat{X}}^{(i)}, \mathbf{\hat{X}}_\perp^{(i)}] \in \mathbb{C}^{n \times n}$ is invertible, any matrix $\mathbf{\hat{V}}^{(i)}$ can be written as 

$$\mathbf{\hat{V}}^{(i)} = [\mathbf{\hat{X}}^{(i)}, \mathbf{\hat{X}}_\perp^{(i)}] \begin{bmatrix} \mathbf{\hat{W}}^{(i)} \\ \mathbf{\hat{B}}^{(i)} \end{bmatrix} = \mathbf{\hat{X}}^{(i)} \mathbf{\hat{W}}^{(i)} + \mathbf{\hat{X}}_\perp^{(i)} \mathbf{\hat{B}}^{(i)} ,$$

(4.16)

for a unique choice of $\mathbf{\hat{W}}^{(i)} \in \mathbb{C}^{p \times p}$ and $\mathbf{\hat{B}}^{(i)} \in \mathbb{C}^{(n-p) \times p}$. Combining equation (4.14) and (4.16) yields 

$$O_p = (\mathbf{\hat{X}}^{(i)})^H (\mathbf{\hat{X}}^{(i)} \mathbf{\hat{W}}^{(i)} + \mathbf{\hat{X}}_\perp^{(i)} \mathbf{\hat{B}}^{(i)}) + (\mathbf{\hat{X}}^{(i)} \mathbf{\hat{W}}^{(i)} + \mathbf{\hat{X}}_\perp^{(i)} \mathbf{\hat{B}}^{(i)})^H \mathbf{\hat{X}}^{(i)} = \mathbf{\hat{W}}^{(i)} + (\mathbf{\hat{W}}^{(i)})^H, i \in [l],$$

(4.17)

which together with Remark 2.4 gives 

$$\mathcal{W} + \mathcal{W}^\top = \mathcal{O}.$$ 

(4.18)

Remark 2.4 shows that (4.15) is equivalent to 

$$\mathcal{X}^\top \ast \mathcal{X}_\perp = \mathcal{O},$$

where $\mathcal{X}_\perp = L^{-1} \left( \mathbf{\hat{X}}_\perp \right), \mathbf{\hat{X}} = L \left( \mathcal{X} \right)$ and $\mathbf{\hat{X}}_\perp^{(i)} \in \mathbb{C}^{n \times (n-p)}$ are any matrix such that $\text{span}(\mathbf{\hat{X}}_\perp^{(i)}) = \{ \mathbf{\hat{X}}_\perp^{(i)} \alpha | \alpha \in \mathbb{C}^{(n-p)} \}$ is the orthogonal complement of $\text{span}(\mathbf{\hat{X}}^{(i)}) = \{ \mathbf{\hat{X}}^{(i)} \beta | \beta \in \mathbb{C}^p \}$. Thus it holds that $\mathcal{V} = \mathcal{X} \ast \mathcal{W} + \mathcal{X}_\perp \ast \mathcal{B} \in \mathbb{R}^{n \times p \times l}$ where $\mathcal{W} \in \text{Skew}(\mathbb{R}^{p \times p \times l})$ by (4.18), and so 

$$T_X \St(n,p,l) = \ker(Dh(\mathcal{X})) = \left\{ \mathcal{X} \ast \mathcal{W} + \mathcal{X}_\perp \ast \mathcal{B} \in \mathbb{R}^{n \times p \times l} \middle| \mathcal{W} \in \text{Skew}(\mathbb{R}^{p \times p \times l}), \mathcal{B} \in \mathbb{R}^{(n-p) \times p \times l} \right\}.$$ 

4.2 Riemannian metric, gradient and Hessian on $\St(n,p,l)$

For the smooth manifold $\St(n,p,l)$, the Riemannian metric $g_X$ is defined as 

$$g_X(\mathcal{V}, \mathcal{U}) := \langle \mathcal{V}, \mathcal{U} \rangle_X = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{l} v_{ijk} \cdot u_{ijk} \quad \text{with} \quad \mathcal{X} \in \St(n,p,l) \quad \text{and} \quad \mathcal{V}, \mathcal{U} \in T_X \St(n,p,l),$$

(4.19)

which is indeed the Euclidean metric from the embedded space induced by the Frobenius norm. Equipped with this metric, $(\St(n,p,l), g_X)$ becomes a Riemannian manifold. In the sequel, we write $\St(n,p,l)$ as a Riemannian manifold for simplicity. The norm induced by the Riemannian metric $g_X$ is defined as $\|\mathcal{V}\|_X := \langle \mathcal{V}, \mathcal{V} \rangle_X$. 

11
Theorem 4.3 (Orthogonal projector operator). The orthogonal projector operator from Euclidean space \( \mathbb{R}^{n \times p \times l} \) to \( T_X \text{St} (n, p, l) \) is
\[
P_X(U) = U - X^T (U^T + U^T * X) = U - X^T \text{sym}(X^T * U) = (I - X^T * X^T) * U + X^T \text{skew}(X^T * U),
\]
where \( \text{sym}(A) = \frac{A + A^T}{2} \) and \( \text{skew}(A) = \frac{A - A^T}{2} \).

Proof. From Lemma 3.1, there holds \( \langle U - P_X(U), V \rangle = 0 \) for all \( V \in T_X \text{St} (n, p, l) \) and \( U \in \mathbb{R}^{n \times p \times l} \). Hence we consider the normal space of \( T_X \text{St} (n, p, l) \) as follows:
\[
N_X \text{St} (n, p, l) = \left\{ N \in \mathbb{R}^{n \times p \times l} \mid \langle N, V \rangle = O \quad \text{for all} \quad V \in T_X \text{St} (n, p, l) \right\} = \left\{ N \in \mathbb{R}^{n \times p \times l} \mid \langle N, X^T * W + X^T * B \rangle = 0, \quad \forall W \in \text{Skew}(\mathbb{R}^{p \times p \times l}), B \in \mathbb{R}^{(n-p) \times p \times l} \right\},
\]
where we use the parameter expression of \( T_X \text{St} (n, p, l) \) in Theorem 4.2. Since \( \hat{X}^{(i)}_{\perp} \in \mathbb{C}^{n \times (n-p)} \) is any matrix such that \( \text{span}(\hat{X}^{(i)}_{\perp}) = \{ \hat{X}^{(i)}_{\perp} \alpha \mid \alpha \in \mathbb{C}^{n-p} \} \) is the orthogonal complement of \( \text{span}(\hat{X}^{(i)}) = \{ \hat{X}^{(i)} \beta \mid \beta \in \mathbb{C}^p \} \), we can expand the DFT of normal vectors as \( \hat{N}^{(i)} = \hat{X}^{(i)} \hat{A}^{(i)} + \hat{X}^{(i)} \hat{C}^{(i)} \) with \( \hat{A}^{(i)} \in \mathbb{C}^{p \times p} \) and \( \hat{C}^{(i)} \in \mathbb{C}^{(n-p) \times p} \), which together with Remark 2.4 leads to \( N = X^{(i)} A + X^{(i)} C \). Then it holds that
\[
N_X \text{St} (n, p, l) = \left\{ N \in \mathbb{R}^{n \times p \times l} \mid \langle X^{(i)} * A + X^{(i)} * C, X^{(i)} * W + X^{(i)} * B \rangle = 0, \quad \forall W \in \text{Skew}(\mathbb{R}^{p \times p \times l}), B \in \mathbb{R}^{(n-p) \times p \times l} \right\} = \left\{ N \in \mathbb{R}^{n \times p \times l} \mid \langle X^{(i)} * A, W \rangle + \langle X^{(i)} * C, B \rangle = 0, \quad \forall W \in \text{Skew}(\mathbb{R}^{p \times p \times l}), B \in \mathbb{R}^{(n-p) \times p \times l} \right\} = \left\{ X^{(i)} * A \in \mathbb{R}^{n \times p \times l} \mid A \in \text{Sym}(\mathbb{R}^{p \times p \times l}) \right\},
\]
where the last equation comes from Lemma 2.2. Thus, the orthogonal projector obeys:
\[
P_X(U) = U - N = U - X^T A
\]
for some symmetric tensor \( A \in \text{Sym}(\mathbb{R}^{p \times p \times l}) \). Since the projected tensor must lie in \( T_X \text{St} (n, p, l) \), we have
\[
X^T * (U - X^T A) + (U - X^T A)^T * X = 0,
\]
that is, \( X^T * U + U^T * X = A + A^T = 2A \). Thus, \( A = \frac{1}{2} (X^T * U + U^T * X) \) and
\[
P_X(U) = U - X^T \text{sym}(X^T * U) = (I - X^T * X^T) * U + X^T \text{skew}(X^T * U).
\]

Lemma 3.1 tells us that the orthogonal projector yields a convenient formula for the Riemannian gradient and Riemannian Hessian.

Proposition 4.1 (Gradient). The gradient of smooth functions \( f \) defined on \( \text{St} (n, p, l) \) is
\[
\text{grad} f(X) = P_X(\text{grad} \bar{f}(X)) = \text{grad} \bar{f}(X) - X^T \text{sym}(X^T * \text{grad} \bar{f}(X)),
\]
where \( \bar{f} \) defined on \( \mathbb{R}^{n \times p \times l} \) coincides with \( f \) on \( \text{St} (n, p, l) : f = \bar{f}|_{\text{St}(n,p,l)} \).

The definitions and computation of the Euclidean gradient \( \text{grad} f(X) \) above and the Euclidean directional derivative \( Df(X)[H] \) that will appear in the following are given in Appendix A.20.

Theorem 4.4 (Hessian). The Riemannian Hessian of a real-valued function \( f \) at a point \( X \) on \( \text{St} (n, p, l) \) is
\[
\text{Hess} f(X)[\mathcal{V}] = P_X(\text{Hess} \bar{f}(X)[\mathcal{V}]) - \mathcal{V}^T \text{sym}(X^T * \text{grad} \bar{f}(X)).
\]
Proof. Let $\hat{G}(X) = \grad f(X) - X \ast \sym(X^\top \ast \grad f(X))$ denote a smooth extension of $\grad f(X)$ to a neighborhood of $St(n,p,l)$ in $\mathbb{R}^{n \times p \times l}$.

$$D\hat{G}(X)[V] = \lim_{t \to 0} \frac{1}{t} \left[ \hat{G}(X + tv) - \hat{G}(X) \right]$$

$$= \Hess \hat{f}(X)[V] - \lim_{t \to 0} \frac{1}{t} \left[ (X + tv) \ast \sym((X + tv)^\top \ast \grad f(X + tv)) - X \ast \sym(X^\top \ast \grad f(X)) \right]$$

$$= \Hess \hat{f}(X)[V] - V \ast \sym(X^\top \ast \grad \hat{f}(X))$$

$$= \Hess \hat{f}(X)[V] - V \ast \sym(X^\top \ast \grad \hat{f}(X)) - X \ast \sym(V^\top \ast \grad \hat{f}(X) + X^\top \Hess \hat{f}(X)[V]).$$

$St(n,p,l)$ is a Riemannian submanifold of $\mathbb{R}^{n \times p \times l}$, and its Riemannian Hessian is

$$\Hess f(X)[V] = P_X \left(D\hat{G}(X)[V]\right)$$

$$= P_X \left(\Hess \hat{f}(X)[V] - V \ast \sym(X^\top \ast \grad \hat{f}(X)) - X \ast S\right)$$

$$= P_X \left(\Hess \hat{f}(X)[V] - V \ast \sym(X^\top \ast \grad \hat{f}(X)),ight.$$\)

where $S = \sym(V^\top \ast \grad \hat{f}(X) + X^\top \Hess \hat{f}(X)[V])$ and $X \ast S \in N_X St(n,p,l)$ vanishes through $P_X$. \hfill \Box

4.3 Retraction on $St(n,p,l)$

According to Lemma 3.2, we can construct different retractions based on various tensor decompositions. Based on t-QR in Subsection 2.5, we get the following retraction.

**Theorem 4.5** (t-QR based retraction). The retraction on $St(n,p,l)$ based on the t-QR decomposition is

$$R(X) = qf(X + V), \quad (4.21)$$

where $X \in St(n,p,l), V \in T_X St(n,p,l)$, and $qf(A)$ denotes the $Q$ factor of the t-QR decomposition of $A \in L^{-1}(C_{\ast}^{n \times p \times l})$ as $A = Q \ast R$, where $Q \in St(n,p,l)$ and $R \in L^{-1}(C_{\ast}^{p \times p \times l}).$

**Proof.** According to Theorem 2.3, if $A \in L^{-1}(C_{\ast}^{n \times p \times l})$, then t-QR decomposition of $A$ is unique. Hence the inverse of t-QR decomposition of $A \in L^{-1}(C_{\ast}^{n \times p \times l})$ is a one-to-one mapping

$$\phi : St(n,p,l) \times L^{-1}(C_{\ast}^{p \times p \times l}) \to L^{-1}(C_{\ast}^{n \times p \times l}) : (Q,R) \mapsto Q \ast R = A.$$ 

where $L^{-1}(C_{\ast}^{p \times p \times l})$ represents the set of $R$ factor of t-QR decomposition of $A \in \mathbb{R}^{n \times p \times l}$. Since $L$ is a continuous function and $C_{\ast}^{n \times p \times l}$ is an open set in $C^{n \times p \times l}$, it follows that the preimage $L^{-1}(C_{\ast}^{n \times p \times l})$ is open in $\mathbb{R}^{n \times p \times l}$. Combining Theorem 4.1 and Lemma 2.3 gives rise to

$$\dim(St(n,p,l)) + \dim \left(L^{-1}(C_{\ast}^{p \times p \times l})\right) = \dim(\mathbb{R}^{n \times p \times l}) - \dim(\text{Sym}(\mathbb{R}^{p \times p \times l})) + \dim \left(L^{-1}(C_{\ast}^{n \times p \times l})\right) = \dim(\mathbb{R}^{n \times p \times l}).$$

The identity tensor is the neutral element. The mapping $\phi$ is smooth since it is the restriction of a smooth map (tensor product) to a submanifold. For $\phi^{-1}$, notice that its first tensor component $Q$ is obtained by the Gram-Schmidt process and the inverse Fourier transform according to [23, Alg. 1], which are $C^\infty$. Since the second component $R$ is obtained as $Q^{-1} \ast A$, it follows that $\phi^{-1}$ is $C^\infty$. Thus $\phi$ is a diffeomorphism. From Lemma 3.2, we have

$$R_X(V) = \pi_1 \left(\phi^{-1}(X + V)\right) = qf(X + V),$$

where $qf(A) = \pi_1 \circ \phi^{-1}$ denotes the mapping that sends a tensor to the $Q$ factor of its t-QR decomposition. This is well defined, i.e., the t-QR decomposition $X + V$ is unique. To see this, it follows from Remark 2.4 that

$$A^\top \ast A = (X + V)^\top \ast (X + V) = I + V^\top \ast V \iff (\hat{A}^{(i)})^H \hat{A}^{(i)} = I_p + (\hat{V}^{(i)})^H \hat{V}^{(i)}, i \in [l],$$
confirming that $\mathcal{A} = \mathcal{X} + \mathcal{V} \in L^{-1} \left( \mathbb{C}^{n \times p \times l}_+ \right)$.

In the same vein, we get the following retraction based on t-PD in Subsection 2.6.

**Theorem 4.6** (t-PD based retraction). The retraction on $\text{St}(n, p, l)$ based on t-PD is

\[
R_X(\mathcal{V}) = (\mathcal{X} + \mathcal{V}) * (I + \mathcal{V}^\top * \mathcal{V})^{-\frac{1}{2}},
\]

(4.22)

where $\mathcal{X} \in \text{St}(n, p, l)$ and $\mathcal{V} \in T_X\text{St}(n, p, l)$.

**Proof.** Proposition 2.7 shows that if $\mathcal{A} \in L^{-1} \left( \mathbb{C}^{n \times p \times l}_+ \right)$, then $\mathcal{A}^\top * \mathcal{A} \in \text{Sym}(\mathbb{R}^{p \times p \times l}_+)$, hence the t-PD of $\mathcal{A} \in L^{-1} \left( \mathbb{C}^{n \times p \times l}_+ \right)$ is unique due to Theorem 2.4. The inverse of t-PD is a mapping

\[
\phi : \text{St}(n, p, l) \times \text{Sym}(\mathbb{R}^{p \times p \times l}_+) \rightarrow L^{-1}(\mathbb{C}^{n \times p \times l}_+) : (\mathcal{A}, \mathcal{H}) \mapsto \mathcal{X} * \mathcal{H} = \mathcal{A}.
\]

Since $L$ is a continuous function and $\mathbb{C}^{n \times p \times l}_+$ is an open set in $\mathbb{C}^{n \times p \times l}$, it follows that the preimage $L^{-1}(\mathbb{C}^{n \times p \times l}_+)$ is open in $\mathbb{R}^{n \times p \times l}$. Theorem 4.1 shows that

\[
\text{dim}(\text{St}(n, p, l)) + \text{dim}(\text{Sym}(\mathbb{R}^{p \times p \times l}_+)) = \text{dim}(\mathbb{R}^{n \times p \times l}) - \text{dim}(\text{Sym}(\mathbb{R}^{p \times p \times l}_+)) + \text{dim}(\text{Sym}(\mathbb{R}^{p \times p \times l}_+)) = \text{dim}(\mathbb{R}^{n \times p \times l}).
\]

The identity tensor is the neutral element. By Proposition 2.6, if $\mathcal{A}^\top * \mathcal{A} \in \text{Sym}(\mathbb{R}^{p \times p \times l}_+)$, then $\phi^{-1}(\mathcal{A}) = (\mathcal{A} * (\mathcal{A}^\top * \mathcal{A})^{-\frac{1}{2}}, (\mathcal{A}^\top * \mathcal{A})^{\frac{1}{2}})$ which shows that $\phi$ is a diffeomorphism, and thus

\[
R_X(\mathcal{V}) = \pi_1 \left( \phi^{-1}(\mathcal{X} + \mathcal{V}) \right)
= (\mathcal{X} + \mathcal{V}) * ((\mathcal{X} + \mathcal{V})^\top * (\mathcal{X} + \mathcal{V}))^{-\frac{1}{2}}
= (\mathcal{X} + \mathcal{V}) * (I + \mathcal{V}^\top * \mathcal{V} + \mathcal{X}^\top * \mathcal{V} + \mathcal{V}^\top * \mathcal{X})^{-\frac{1}{2}}
= (\mathcal{X} + \mathcal{V}) * (I + \mathcal{V}^\top * \mathcal{V})^{-\frac{1}{2}}.
\]

This is well defined. To see this, it follows from 2.4 that

\[
\mathcal{A}^\top * \mathcal{A} = (\mathcal{X} + \mathcal{V})^\top * (\mathcal{X} + \mathcal{V}) = I + \mathcal{V}^\top * \mathcal{V} \Leftrightarrow (\hat{A}(i))^H \hat{A}(i) = I_p + (\hat{V}(i))^H \hat{V}(i), i \in [l],
\]

confirming that $\mathcal{A} = \mathcal{X} + \mathcal{V} \in L^{-1} \left( \mathbb{C}^{n \times p \times l}_+ \right)$.

We can also construct retractions directly from the Definition 3.12.

**Theorem 4.7** (t-Cayley based retraction). The retraction on $\text{St}(n, p, l)$ based on t-Cayley transform is

\[
R_X(\mathcal{V}) = \left( I - \frac{1}{2} \mathcal{W}_\mathcal{V} \right)^{-1} * \left( I + \frac{1}{2} \mathcal{W}_\mathcal{V} \right) * \mathcal{X},
\]

(4.23)

where $\mathcal{X} \in \text{St}(n, p, l), \mathcal{V} \in T_X\text{St}(n, p, l), \mathcal{W}_\mathcal{V} = \mathcal{P} * \mathcal{X}^\top - \mathcal{X} * \mathcal{V}^\top * \mathcal{P} \in \text{Skew}(\mathbb{R}^{n \times n \times l})$ and $\mathcal{P} = I - \frac{1}{2} \mathcal{X} * \mathcal{X}^\top$.

**Proof.** According to Lemma 2.1, for any $\mathcal{W}_\mathcal{V} \in \text{Skew}(\mathbb{R}^{n \times n \times l})$, it holds that

\[
\left( I - \frac{1}{2} \mathcal{W}_\mathcal{V} \right)^\top * \left( I - \frac{1}{2} \mathcal{W}_\mathcal{V} \right) = I + \frac{1}{4} \mathcal{W}_\mathcal{V}^\top * \mathcal{W}_\mathcal{V} \in \text{Sym}(\mathbb{R}^{n \times n \times l}_+),
\]

namely, $I - \frac{1}{2} \mathcal{W}_\mathcal{V}$ is invertible. Since $\mathcal{W}_\mathcal{V}$ is skew-symmetric and $(I - \mathcal{A}) * (I + \mathcal{B}) = (I + \mathcal{B}) * (I - \mathcal{A})$ for all $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{n \times p \times l}$, we obtain

\[
\left( I - \frac{1}{2} \mathcal{W}_\mathcal{V} \right)^{-1} * \left( I + \frac{1}{2} \mathcal{W}_\mathcal{V} \right) \in \text{St}(n, n, l),
\]

and thus $R_X(\mathcal{V}) \in \text{St}(n, p, l)$. Consider the following curve on $\text{St}(n, p, l)$:

\[
\mathcal{F}(t) = R_X(t \mathcal{V}) = \left( I - \frac{1}{2} \mathcal{W}_\mathcal{V} \right)^{-1} * \left( I + \frac{1}{2} \mathcal{W}_\mathcal{V} \right) * \mathcal{X},
\]

where $t \in [0, 1]$. Therefore, $\mathcal{F}(1) = \mathcal{F}(0) = \mathcal{X}$, and $\mathcal{F}(t)$ is a curve from $\mathcal{X}$ to $\mathcal{V}$ within $\text{St}(n, p, l)$ with $\mathcal{F}(t) \in \text{St}(n, p, l)$ for all $t \in [0, 1]$. This concludes the proof of Theorem 4.7.
with $F(0) = X$. Differentiating both sides of the following equation

$$
\left( I - \frac{t}{2} W_V \right) * F(t) = \left( I + \frac{t}{2} W_V \right) * X
$$

with respect to $t$ gives

$$
\dot{F}(0) = W_V * X = V - \frac{1}{2} X * \left( X^T * V + V^T * X \right) = V
$$

for all $V \in T_X St(n, p, l)$. Hence $R_X(O_X) = X$ and

$$
DR_X(O_X)[V] = \frac{d}{dt} R_X(O_X + tV)|_{t=0} = V,
$$

which shows that $R_X(V)$ is a retraction. □

Based on $t$-exponential introduced in Subsection 2.7, we get the following retraction.

**Theorem 4.8** ($t$-exponential based retraction). The retraction on $St(n, p, l)$ based on $t$-exponential is

$$
R_X(V) = (X * Q) * \exp \left( \left[ \left( X^T * V - R^T \right) \right] * \left( \frac{2}{t} \right) \right),
$$

where $Q * R = C$ is the $t$-QR decomposition of $C = (I - X * X^T) * V$ if $\hat{R} \in \mathbb{C}^{n \times p \times l}$; alternatively, if $\hat{C} \notin \mathbb{C}^{n \times p \times l}$, then $Q \in \mathbb{R}^{\times n \times (n-p) \times l}, R \in \mathbb{R}^{(n-p) \times p \times l}$, in such a way that $Q$ is partially orthogonal and $X^T * Q = O$.

**Proof.** When $\hat{C} \in \mathbb{C}^{n \times p \times l}, \hat{R}(i)$ are invertible and therefore $R$ is invertible and

$$
X^T * Q = X^T * C * R^{-1} = O,
$$

since $X^T * C = X^T * (V - X * X^T + V) = X^T * V - X^T * X^T * X^T * V = O$.

Now consider the case that when $\hat{C}$ is not of f-full rank $p$. Such a choice is always available as we show below. Recall that every $V \in T_X St(n, p, l)$ can be written in the form $V = X * A + X_\perp * B$, where $X_\perp \in \mathbb{R}^{n \times (n-p) \times l}$ is partially orthogonal and $X^T * X_\perp = O$. According to Remark 2.4, for each $i \in [l]$, the complex matrix $\hat{X}_\perp^{(i)}$ is partially unitary and $(\hat{X}^{(i)})^H \hat{X}_\perp^{(i)} = O_{p \times (n-p)}$.

Note that

$$
C = V - X * X^T * V = V - X^T * X^T * X_\perp * (X * A + X_\perp * B) = V - X * A = X_\perp * B.
$$

Therefore, using Remark 2.4 again, for $i \in [l]$, we have $\hat{C}^{(i)} = \hat{X}_\perp^{(i)} \hat{B}^{(i)}$ and consequently span $\hat{C}^{(i)} \subset$ span $\hat{X}_\perp^{(i)}$, with $\text{dim} \left( \text{span} \hat{X}_\perp^{(i)} \right) = n - p$. Pick an orthonormal basis of span $\hat{X}_\perp^{(i)}$ and form the partially unitary matrix $\hat{Q}^{(i)} \in \mathbb{C}^{n \times (n-p)}$. Then we have $(\hat{X}^{(i)})^H \hat{Q}^{(i)} = O$ since the column vectors of $\hat{X}_\perp$ are orthogonal to any vector in span $\hat{X}_\perp$. Since span $\hat{C}_\perp \subset$ span $\hat{X}_\perp$, we have $\hat{C}_\perp = \hat{Q}^{(i)} \hat{R}^{(i)}$ for some $\hat{R}^{(i)} \in \mathbb{C}^{(n-p) \times p}$. Applying inverse DFT to the third-order tensors fold $\hat{Q}^{(i)} : i \in [l]$ and fold $\hat{R}^{(i)} : i \in [l]$, we obtain a partially orthogonal tensor $Q \in \mathbb{R}^{\times n \times (n-p) \times l}$ with $X^T * Q = O$ and a tensor $R \in \mathbb{R}^{(n-p) \times p \times l}$ with $Q * R = C$. Indeed we can simply take $Q = X_\perp$ and $R = B$.

In conclusion, there always holds that $X^T * Q = O$ regardless of whether $\hat{C}$ is f-full rank or not. Now we prove that the exponential retraction on $St(n, p, l)$ as defined above is indeed a retraction.

First of all, we prove that $R_X(V) \in St(n, p, l)$. Denotes $A = \left( \frac{X^T * V - R^T}{O} \right)$. Since $A$ is a skew-symmetric tensor, it follows from Proposition 2.13 that $\exp [A^T] * \exp [A] = \exp [-A] * \exp [A] = \exp [-A + A] = I$. Thus we have

$$
R_X(V)^T * R_X(V) = \left( I \ O \right) * \exp [A]^T * \left( \frac{X^T}{Q} \right) * (X * Q) * \exp [A] * \left( \frac{2}{t} \right) = \left( I \ O \right) * \exp [A]^T * \left( \frac{X^T * X * Q}{Q^T * X^T * Q} \right) * \exp [A] * \left( \frac{2}{t} \right) = \left( I \ O \right) * \exp [A]^T * \exp [A] * \left( \frac{2}{t} \right) = I,
$$

15
where the second equality comes from Proposition 2.12. We now derive an equivalent formula for the exponential retraction as follows:

\[
R_X(V) = (X \cdot Q) \ast \exp [A] \ast \left( \begin{array}{c}
\frac{\partial}{\partial t} \\
\frac{\partial}{\partial t}
\end{array} \right)
\]

which is due to Proposition 2.10. This equivalent expression for \( R : T\text{St}(n,p,l) \to \text{St}(n,p,l) \) proves its smoothness since it involves only the exponential and the t-product of \( X \) and \( V \), which from Proposition 2.8 are both smooth operations.

We then verify that \( R_X(O) = X \). Note that in this case \( (V = O) \), we have \( X^T \ast V = O \) and \( Q \ast R = (I - X \ast X^T) \ast V = O \). Therefore \( R = Q \ast R \ast R = O \). Then it holds that

\[
R_X(O) = (X \cdot Q) \ast \exp \left[ \left( \begin{array}{c}
\frac{\partial}{\partial t} \\
\frac{\partial}{\partial t}
\end{array} \right) \right] = (X \cdot Q) \ast \left( \begin{array}{c}
\frac{\partial}{\partial t} \\
\frac{\partial}{\partial t}
\end{array} \right) = X.
\]

Finally we show that the derivative \( DR_X(O) : T_X\text{St}(n,p,l) \to T_X\text{St}(n,p,l) \) is the identity mapping. For any \( V \in T_X\text{St}(n,p,l) \), we have

\[
DR_X(O)[V] = \left. \frac{d}{dt}(R_X(tV)) \right|_{t=0} = \left. \left\{ \left( X \cdot Q \right) \ast \exp \left[ t \left( \begin{array}{cc}
X^T \ast V & -R^T \\
R & O
\end{array} \right) \right] \ast \left( X^T \ast V & -R^T \\
R & O
\end{array} \right) \left( \begin{array}{c}
\frac{\partial}{\partial t} \\
\frac{\partial}{\partial t}
\end{array} \right) \right\} \right|_{t=0}
\]

where the second equality is due to Proposition 2.9.

**Remark 4.1.** It follows from this formula that when \( n = p \), we have the simpler formula for the exponential retraction on the group \( \text{St}(n,n,l) \) of orthogonal tensors: \( R_X(V) = \exp \left[ V \ast X^T \right] \ast X = X \ast \exp \left[ X^T \ast V \right] \).

The following proposition indicates that the retraction based on t-exponential is actually a geodesic on \( \text{St}(n,p,l) \).

**Proposition 4.2** (Geodesic). The geodesic on \( \text{St}(n,p,l) \) emanating from \( X \) in direction \( V \) is given by the curve \( \mathcal{C}(t) = R_X(tV) = (X \cdot Q) \ast \exp \left[ tA \right] \ast \left( \begin{array}{c}
\frac{\partial}{\partial t} \\
\frac{\partial}{\partial t}
\end{array} \right) \), where \( A = \left( \begin{array}{cc}
X^T \ast V & -R^T \\
R & O
\end{array} \right) \) is a skew-symmetric tensor.

**Proof.** The proof of Theorem 4.8 shows that \( \mathcal{C}(0) = X \) and \( \dot{\mathcal{C}}(0) = V \). Since \( \text{St}(n,p,l) \) is a Riemannian submanifold of \( \mathbb{R}^{n \times p \times l} \), [3, Sect. 5.8] shows that the acceleration of \( \mathcal{C}(t) \) is

\[
\ddot{\mathcal{C}}(t) = \mathbf{P}_{\mathcal{C}(t)} \left( \frac{d^2}{dt^2} \mathcal{C}(t) \right),
\]

where \( \mathbf{P}_{\mathcal{C}(t)} = \mathcal{U} - X \ast \text{sym}(X^T \ast \mathcal{U}) \) and \( \text{sym}(A) = \frac{A + A^T}{2} \). Hence

\[
\ddot{\mathcal{C}}(t) = \frac{d^2}{dt^2} \mathcal{C}(t) - \mathcal{C}(t) \ast \text{sym} \left( \mathcal{C}^T(t) \ast \frac{d^2}{dt^2} \mathcal{C}(t) \right),
\]

where

\[
\mathcal{U} = \left( \begin{array}{cc}
X^T \ast V & -R^T \\
R & O
\end{array} \right).
\]
where \( \frac{d^2}{dt^2} C(t) = (X \cdot Q) \star \exp [tA] \cdot A^2 \star \left( \frac{\partial}{\partial t} \right) \). Since \( A \) is a skew-symmetric tensor, it is easy to check \( \dot{C}(t) = \mathcal{O} \), which means \( C(t) \) is the geodesic on \( St(n, p, l) \).

\[ \square \]

### 4.4 Vector transport on \( St(n, p, l) \)

By Lemma 3.3 and the orthogonal projector operator \( (4.20) \), we obtain a series of vector transports as follows.

**Theorem 4.9** (Orthogonal projector based vector transport). The vector transport on \( St(n, p, l) \) is

\[
T_{rV} = P_{R_X(U)} V = (I - Y \ast Y^T) \ast V + Y \ast \text{skew}(Y^T \ast V)
\]

\[
= V - Y \ast \text{sym}(Y^T \ast V) \in T_Y St(n, p, l),
\]

(4.24)

where \( Y = R_X(U) \) is any retraction on \( St(n, p, l) \).

To derive the vector transport by differentiated retraction based on \( t \)-\( QR \) decomposition, the following two lemmas are necessary.

**Lemma 4.1.** \( \mathbb{C}^{p \times p \times l}_{upp} \) is an open submanifold of linear manifold \( \mathbb{C}^{p \times p \times l} \) with real diagonal elements, and its tangent space at any point \( Y \in \mathbb{C}^{p \times p \times l}_{upp} \) is just \( \mathbb{C}^{p \times p \times l} \) with real diagonal elements.

**Proof.** Note that \( \mathbb{C}^{p \times p \times l} \) with real diagonal elements is a vector space and so a linear manifold. Since \( \mathbb{R}^+ \) is open in \( \mathbb{R} \), it follows that \( \mathbb{C}^{p \times p \times l}_{upp} \) is open in \( \mathbb{C}^{p \times p \times l} \) with real diagonal elements. Then it follows from [1, Sect. 3.5.2] that \( \mathbb{C}^{p \times p \times l}_{upp} \) is a manifold and its tangent space at any point \( Y \in \mathbb{C}^{p \times p \times l}_{upp} \) is just \( \mathbb{C}^{p \times p \times l} \) with real diagonal elements. \( \square \)

**Lemma 4.2.** Let \( C(t) = A(t) \ast B(t) \in \mathbb{R}^{m \times n \times l} \). Then the tangent vector to the curve \( C(t) \) is

\[
\dot{C}(t) = \dot{A}(t) \ast B(t) + A(t) \ast \dot{B}(t).
\]

**Proof.** By Definition 2.6, there holds

\[
\text{unfold}(C(t)) = \text{unfold}(A(t) \ast B(t)) = \text{bcirc}(A(t)) \cdot \text{unfold}(B(t)).
\]

Hence it follows that

\[
C^{(k)}(t) = \sum_{i=1}^{l} A^{(h_i)}(t) B^{(i)}(t), \quad k \in [l],
\]

where

\[
h_i = \begin{cases} 
  l + k + 1 - i, & i > k \\
  k + 1 - i, & i \leq k
\end{cases}
\]

Using the corresponding property of the matrix case [1], we obtain

\[
\dot{C}^{(k)}(t) = \sum_{i=1}^{l} \left( \dot{A}^{(h_i)}(t) B^{(i)}(t) + A^{(h_i)}(t) \dot{B}^{(i)}(t) \right),
\]

which means \( \text{unfold}(\dot{C}(t)) = \text{bcirc}(\dot{A}(t)) \cdot \text{unfold}(B(t)) + \text{bcirc}(A(t)) \cdot \text{unfold}(\dot{B}(t)) \). Thus it follows that

\[
\dot{C}(t) = \dot{A}(t) \ast B(t) + A(t) \ast \dot{B}(t).
\]

\( \square \)

We are now in a position to derive computational formulae from (3.12) for the vector transport as the differentiated retraction \( R_X(U) = qf(A + U) \).
Theorem 4.10 (t-QR based vector transport). The vector transport on $\text{St}(n,p,l)$ by differentiated retraction $R_X(U) = qf(X + U)$ is
\[
T_d V = Q \ast L^{-1} \left( \text{fold} \left( \text{P}_{\text{skew}} \left( \hat{B}^{(i)} \right) : i \in [l] \right) \right) + (I - Q \ast Q^T) \ast C,
\]
where $Q = R_X(U)$, $C = V \ast (Q^T \ast (X + U))^{-1}$, $B = Q^T \ast C$ and $\text{P}_{\text{skew}}(\hat{B}^{(i)})$ denotes the skew-symmetric term of the decomposition of the complex matrix $\hat{B}^{(i)} = (L(B))^{(i)}$ into the sum of a skew-symmetric term and an upper triangular term with real diagonal elements. Specifically,
\[
(\text{P}_{\text{skew}}(\hat{B}^{(i)}))_{mn} = \begin{cases} 
-\text{conj}(\hat{B}^{(i)}_{mn}), & m < n \\
\text{Im}(\hat{B}^{(i)}_{mn}), & m = n \\
\hat{B}^{(i)}_{mn}, & m > n
\end{cases}
\]
(4.25)
where $\text{Im}(c)$ represents the imaginary part of the complex number $c$.

Proof. From (3.12) and Theorem 4.5, for $U, V \in T_X \text{St}(n,p,l)$, we have
\[
T_d V = DR_X(U)[V] = Dqf(X + U)[V] = \frac{d}{dt}qf(X + U + t V)|_{t=0}.
\]
This is well defined, i.e., the t-QR decomposition of $W(t) := G + tV$ is unique, where $G := X + U$. To see this, it follows from Remark 2.4 that
\[
W(t)^T \ast W(t) = I + (U + tV)^T \ast (U + tV) \Leftrightarrow (\hat{W}(t)) = I + \left( \hat{U}(t) + t \hat{V}(t) \right)^H \left( \hat{U}(t) + t \hat{V}(t) \right), i \in [l],
\]
showing that $\hat{W}(t) \in C_{n \times p \times l}$, which together with Theorem 4.5 gives the desired result. Hence $W(t) = G + tV$ is a curve on $L^{-1} \left( C_{n \times p \times l} \right)$ with $W(0) = G$ and $W(0) = V$. Let $W(t) = Q(t) \ast R(t)$ denote the t-QR decomposition of $W(t)$, where $Q(t) \in \text{St}(n,p,l)$ and $R(t) \in L^{-1} \left( C_{p \times p \times l} \right)$. Hence $Q(0) = qf(G)$, $R(0) = qf(G)^T \ast G$. Our task now is to compute $T_d V = \frac{d}{dt}Q(t)|_{t=0} = \dot{Q}(0)$. Since $\dot{Q} = Q(t) \ast Q(t)^T + (I - Q(t) \ast Q(t)^T)$, we have the decomposition
\[
\dot{Q}(t) = Q(t) \ast Q(t)^T \ast \dot{Q}(t) + (I - Q(t) \ast Q(t)^T) \ast \dot{Q}(t).
\]
(4.26)
It follows from Lemma 4.2 that
\[
\dot{W}(t) = Q(t) \ast R(t) + Q(t) \ast \dot{R}(t).
\]
(4.27)
Since $\dot{R}(t) \in C_{p \times p \times l}$ which means $\dot{R}^{(i)}(t)$ are invertible, it follows from Remark 2.4 and Definition 2.5 that $R(t)$ are invertible. Multiplying (4.27) by $I - Q(t) \ast Q(t)^T$ on the left and by $(R(t))^{-1}$ on the right yields
\[
(I - Q(t) \ast Q(t)^T) \ast \dot{Q}(t) = (I - Q(t) \ast Q(t)^T) \ast \dot{W}(t) \ast (R(t))^{-1}.
\]
(4.28)
which is the second term of (4.26). It remains to derive the computational formulae for the first term of (4.26). Since $\dot{Q}(t)$ is a tangent vector at the point $Q(t)$, it follows that (4.26) satisfies the form:
\[
T_{Q(t)} \text{St}(n,p,l) = \left\{ Q(t) \ast W(t) + \tilde{Q}(t) \ast B(t) \in \mathbb{R}^{n \times p \times l} \mid W(t) \in \text{Skew} \left( \mathbb{R}^{p \times p \times l} \right), B(t) \in \mathbb{R}^{(n-p) \times p \times l} \right\},
\]
where $\tilde{Q}^{(i)}(t) \in C_{n \times (n-p)}$ is any matrix such that $\text{span}(\tilde{Q}^{(i)}(t)) = \{ \tilde{Q}^{(i)}(t) \alpha \mid \alpha \in C^{n-p} \}$ is the orthogonal complement of $\text{span}(\tilde{Q}^{(i)}(t)) = \{ \tilde{Q}^{(i)}(t) \beta \mid \beta \in C^p \}$. It is easy to check that $\left\langle \tilde{Q}^{(i)}(t) \beta, (I - \tilde{Q}^{(i)}(t)('tilde{Q}^{(i)}(t))^H \right\rangle \alpha = 0$ for any $\alpha \in C^{n-p}$ and $\beta \in C^p$, which means matrix $(I - \tilde{Q}^{(i)}(t)('tilde{Q}^{(i)}(t))^H)$ is a choice of matrix $\tilde{Q}^{(i)}(t)$, hence the term $Q(t)^T \ast \dot{Q}(t) \in \text{Skew} \left( \mathbb{R}^{p \times p \times l} \right)$. Next it is sufficient for us to obtain the formula for $Q(t)^T \ast \dot{Q}(t)$.

Multiplying (4.27) on the left by $Q(t)^T$ and on the right by $(R(t))^{-1}$ leads to
\[
B(t) := Q(t)^T \ast \dot{W}(t) \ast (R(t))^{-1} = Q(t)^T \ast \dot{Q}(t) + \dot{R}(t) \ast (R(t))^{-1}.
\]
(4.29)
Since $\hat{R}(t) \in \mathbb{C}^{p \times p \times l}_{\text{app}+}$, Lemma 4.1 shows that $\dot{\hat{R}}(t) \in \mathbb{C}^{p \times p \times l}_{\text{app}+}$ with real diagonal elements. Note that $\dot{\hat{R}}(t) \in \mathbb{C}^{p \times p}_{\text{app}+}$ are invertible, hence $\left(\dot{\hat{R}}(t)\right)^{-1} \in \mathbb{C}^{p \times p}_{\text{app}+}$. Thus $\dot{\hat{R}}(t) \left(\dot{\hat{R}}(t)^{-1}\right)^{-1}$ is upper triangular complex matrix with real diagonal elements. Using Remark 2.4 again, we get

$$
\dot{\hat{B}}^{(i)} = \left(\dot{\hat{Q}}^{(i)}(0)\right)^{H} \hat{W}^{(i)}(0) \left(\dot{\hat{R}}^{(i)}(0)^{-1}\right)^{-1} = \left(\dot{\hat{Q}}^{(i)}(0)\right)^{H} \dot{\hat{Q}}^{(i)}(0) + \dot{\hat{R}}^{(i)}(0) \left(\dot{\hat{R}}^{(i)}(0)^{-1}\right)^{-1}.
$$

(4.30)

Recalling that $\left(\dot{\hat{Q}}^{(i)}(t)\right)^{H} \dot{\hat{Q}}^{(i)}(t)$ is skew-symmetric and applying the operator $P_{\text{skew}}$ which denotes the skew-symmetric term of the decomposition of the complex matrix into the sum of a skew-symmetric term and an upper triangular term with real diagonal elements, we obtain

$$\left(L \left(Q(0)\right)^{T} \ast \hat{Q}(0)\right)^{(i)} = \left(\dot{\hat{Q}}^{(i)}(t)\right)^{H} \dot{\hat{Q}}^{(i)}(t) = P_{\text{skew}} \left(\dot{\hat{B}}^{(i)}\right)
$$

(4.31)

And Remark 2.4 tell us that (4.31) can be equivalently rewritten as

$$Q(0)^{T} \ast \hat{Q}(0) = L^{-1} \left(\text{fold} \left(P_{\text{skew}} \left(\dot{\hat{B}}^{(i)}\right) : i \in \left[l\right]\right)\right),
$$

(4.32)

where $\mathcal{B} = Q(0)^{T} \ast \hat{W}(0) \ast (\mathcal{R}(0))^{-1}$. Replacing (4.32) and (4.28) in (4.26) gives

$$\dot{\hat{Q}}(0) = Q(0) \ast L^{-1} \left(\text{fold} \left(P_{\text{skew}} \left(\dot{\hat{B}}^{(i)}\right) : i \in \left[l\right]\right)\right) + (I - Q(0) \ast Q(0)^{T}) \ast \hat{W}(0) \ast (\mathcal{R}(0))^{-1}
$$

$$= qf(\mathcal{G}) \ast L^{-1} \left(\text{fold} \left(P_{\text{skew}} \left(\dot{\hat{B}}^{(i)}\right) : i \in \left[l\right]\right)\right) + (I - qf(\mathcal{G}) \ast qf(\mathcal{G})^{T}) \ast \mathcal{V} \ast (qf(\mathcal{G})^{T} \ast \mathcal{G})^{-1},
$$

where $\mathcal{B} = qf(\mathcal{G})^{T} \ast \mathcal{V} \ast (qf(\mathcal{G})^{T} \ast \mathcal{G})^{-1}$. Finally, we have for $\mathcal{U}, \mathcal{V} \in T_{\mathcal{X}} \text{St}(n, p, l)$,

$$\mathcal{T}_{\mathcal{U}} \mathcal{V} = \dot{\hat{Q}}(0) = Q \ast L^{-1} \left(\text{fold} \left(P_{\text{skew}} \left(\dot{\hat{B}}^{(i)}\right) : i \in \left[l\right]\right)\right) + (I - Q \ast Q^{T}) \ast \mathcal{C},
$$

(4.33)

where $Q = R_{\mathcal{X}}(\mathcal{U})$, $\mathcal{V} = \mathcal{V} \ast (\mathcal{Q}^{T} \ast (\mathcal{X} + \mathcal{U}))^{-1}$ and $\mathcal{B} = Q^{T} \ast \mathcal{C}$.

To derive the vector transport by differentiated retraction $R_{\mathcal{X}}(\mathcal{U})$ based on t-PD decomposition, we then need the circulant matrices of $\mathcal{A}$ in another order as follows.

**Definition 4.1.** Let $\mathcal{A} \in \mathbb{R}^{n \times p \times l}$; then the circulant matrices in another order is defined as

$$\widetilde{\text{bcirc}}(\mathcal{A}) := \begin{bmatrix}
A^{(1)} & A^{(2)} & \cdots & A^{(t)} \\
A^{(0)} & A^{(1)} & \cdots & A^{(t-1)} \\
\vdots & \vdots & \ddots & \vdots \\
A^{(p-1)} & \cdots & \cdots & A^{(0)}
\end{bmatrix} \in \mathbb{R}^{nI \times pl}.
$$

**Theorem 4.11** (t-Sylvester equation). Let $\mathcal{A} \in \mathbb{R}^{n \times n \times l}$, $\mathcal{X} \in \mathbb{R}^{n \times p \times l}$, $\mathcal{B} \in \mathbb{R}^{p \times p \times l}$. The analytical solution of the $t$-Sylvester equation

$$\mathcal{A} \ast \mathcal{X} \ast \mathcal{X} \ast \mathcal{B} = \mathcal{C},
$$

(4.34)

is $\text{vec}(\mathcal{C}) = \left(\text{vec}(\mathcal{B})^{T} \otimes I_{n} + [I_{p} I_{l}] \otimes \text{vec}(\mathcal{A})\right)^{\dagger}$, where $[I_{p} I_{l}] \in \mathbb{R}^{pl \times pl}$ is a block matrix whose $(i, j)$ submatrix is $I_{p}$, $i \in [l], j \in [l]$, and vec($\mathcal{C}$) = $\text{vec}(\mathcal{C})$ denotes the vectorized of $\mathcal{C}$ in the meaning of lexicographical ordering.

Necessary lemmas for proving Theorem 4.11 are provided in Appendix A.19. The proof of Theorem 4.11 is also left to Appendix A.19.

**Lemma 4.3.** Sym($\mathbb{R}^{n \times p \times l}_{++}$) is an open submanifold of linear manifold Sym($\mathbb{R}^{n \times p \times l}$) and its tangent space at any point $\mathcal{Y} \in \text{Sym}(\mathbb{R}^{n \times p \times l}_{++})$ is just Sym($\mathbb{R}^{n \times p \times l}$).

**Proof.** Note that Sym($\mathbb{R}^{p \times p \times l}$) is a vector space thus a linear manifold. [53, Rmk. 10] shows that Sym($\mathbb{R}^{n \times p \times l}$) is an open subset of Sym($\mathbb{R}^{n \times p \times l}$) thus an open submanifold of Sym($\mathbb{R}^{n \times p \times l}$). Then it follows from [1, Sect. 3.5.2] that $T_{\mathcal{Y}} \text{Sym}(\mathbb{R}^{n \times p \times l}_{++}) = \text{Sym}(\mathbb{R}^{n \times p \times l})$. □
With the help of Theorem 4.11 and Lemma 4.3, we can now derive the following t-PD based vector transport.

**Theorem 4.12** (t-PD based vector transport). *The vector transport on St\((n, p, l)\) by differentiated retraction based on t-PD is*

\[
\mathcal{T}_t \mathcal{V} = \mathcal{Y} \ast \mathcal{S} + (\mathcal{I} - \mathcal{Y} \ast \mathcal{Y}^\top) \ast \mathcal{V} \ast (\mathcal{Y}^\top \ast (\mathcal{X} + \mathcal{U}))^{-1},
\]

where \(\mathcal{Y} = R_X(\mathcal{U}) = (\mathcal{X} + \mathcal{U}) \ast \mathcal{P}^{-1}\), vec(\(\mathcal{S}\)) = \(\left(\text{bcirc}(\mathcal{P})^\top \otimes I_p + [I_p]_{l \times l} \circ \text{bcirc}(\mathcal{P})\right)^\dagger \cdot \text{vec}(\mathcal{Y}^\top \ast \mathcal{V} - \mathcal{V}^\top \ast \mathcal{Y})\) and \(\mathcal{P} = (\mathcal{I} + \mathcal{U}^\top \ast \mathcal{U})^{\frac{1}{2}}\).

**Proof.** From (3.12) and Theorem 4.6, for \(\mathcal{U}, \mathcal{V} \in T_X\text{St}\((n, p, l)\)\), we have

\[
\mathcal{T}_t \mathcal{V} = DR_X(\mathcal{U})[\mathcal{V}] = \frac{d}{dt} R_X(\mathcal{U} + t\mathcal{V})|_{t=0}.
\]

This is well defined, i.e., the t-PD decomposition of \(\mathcal{W}(t) := \mathcal{X} + \mathcal{U} + t\mathcal{V}\) is unique. To see this, it follows from Remark 2.4 that

\[
\mathcal{W}(t)^\top \ast \mathcal{W}(t) = \mathcal{I} + (\mathcal{U} + t\mathcal{V})^\top \ast (\mathcal{U} + t\mathcal{V}) \iff (\mathcal{W}(t)^{(i)}H \dot{\mathcal{W}}(t)^{(i)}) = I_p + \left(\dot{U}^{(i)} + t \dot{V}^{(i)}\right)H \left(\dot{U}^{(i)} + t \dot{V}^{(i)}\right), i \in [l],
\]

showing that \(\dot{\mathcal{W}}(t) \in \mathbb{C}_a^{n \times p \times l}\), which together with Theorem 4.5 gives the desired result. Hence \(\mathcal{W}(t)\) is a curve on \(L^{-1}\left(\mathbb{C}_a^{n \times p \times l}\right)\) with \(\mathcal{W}(0) = \mathcal{X} + \mathcal{U}\) and \(\mathcal{W}(0) = \mathcal{V}\). Let \(\mathcal{W}(t) = \mathcal{Y}(t) \ast \mathcal{P}(t)\) denote the t-PD of \(\mathcal{W}(t)\). Theorem 2.4 shows that \(\mathcal{Y}(t) \in \text{St}\((n, p, l)\)\) and \(\mathcal{P}(t) \in \text{Sym}\left(\mathbb{R}^{n \times p \times l}_+\right)\). Hence \(\mathcal{Y}(0) = R_X(\mathcal{U}), \mathcal{P}(0) = R_X(\mathcal{U})^\top \ast (\mathcal{X} + \mathcal{U})\). Our task now is to compute \(\mathcal{T}_t \mathcal{V} = \frac{d}{dt} \mathcal{Y}(0) = 0\). Since \(\mathcal{I} = \mathcal{Y}(t) \ast \mathcal{Y}(t)^\top + (\mathcal{I} - \mathcal{Y}(t) \ast \mathcal{Y}(t)^\top)\), we have the decomposition

\[
\dot{\mathcal{Y}}(t) = \mathcal{Y}(t) \ast \mathcal{Y}(t)^\top \ast \dot{\mathcal{Y}}(t) + (\mathcal{I} - \mathcal{Y}(t) \ast \mathcal{Y}(t)^\top) \ast \dot{\mathcal{Y}}(t).
\]

(4.35)

It follows from Lemma 4.2 that

\[
\dot{\mathcal{W}}(t) = \dot{\mathcal{Y}}(t) \ast \mathcal{P}(t) + \mathcal{Y}(t) \ast \dot{\mathcal{P}}(t).
\]

(4.36)

Multiplying (4.36) by \(\mathcal{I} - \mathcal{Y}(t) \ast \mathcal{Y}(t)^\top\) on the left and \(\mathcal{P}(t)^{-1}\) on the right yields

\[
(\mathcal{I} - \mathcal{Y}(t) \ast \mathcal{Y}(t)^\top) \ast \dot{\mathcal{Y}}(t) = (\mathcal{I} - \mathcal{Y}(t) \ast \mathcal{Y}(t)^\top) \ast \dot{\mathcal{W}}(t) \ast (\mathcal{P}(t)^{-1}),
\]

(4.37)

which is the second term of (4.35). It remains to derive the computational formulae for the first term of (4.35). Since \(\dot{\mathcal{Y}}(t)\) is a tangent vector at the point \(\mathcal{Y}(t), (4.35)\) satisfies the form:

\[
T_{Y(t)}\text{St}\((n, p, l)\) = \left\{\dot{\mathcal{Y}}(t) \ast \mathcal{W}(t) + \mathcal{Y}_\perp(t) \ast \mathcal{B}(t) \in \mathbb{R}^{n \times p \times l} \bigg| \mathcal{W}(t) \in \text{Skew}(\mathbb{R}^{p \times p \times l}), \mathcal{B}(t) \in \mathbb{R}^{(n - p) \times p \times l}\right\},
\]

where \(\dot{\mathcal{Y}}(i)(t) \in \mathbb{C}^{n \times (n-p)}\) is any matrix such that \(\text{span}(\dot{\mathcal{Y}}(i)(t)) = \{\dot{\mathcal{Y}}(i)(t)\alpha | \alpha \in \mathbb{C}^{n-p}\}\) is the orthogonal complement of \(\text{span}(\dot{\mathcal{Y}}(i)(t)) = \{\dot{\mathcal{Y}}(i)(t)\beta | \beta \in \mathbb{C}^p\}\). It is easy to check that \(\left(\dot{\mathcal{Y}}(i)(t)\beta, \left(I - \dot{\mathcal{Y}}(i)(t)(\dot{\mathcal{Y}}(i)(t))^H\right)\alpha\right) = 0\) for any \(\alpha \in \mathbb{C}^{n-p}\) and \(\beta \in \mathbb{C}^p\), which means matrix \(I - \dot{\mathcal{Y}}(i)(t)(\dot{\mathcal{Y}}(i)(t))^H\) is a choice of matrix \(\dot{\mathcal{Y}}(i)(t)\), hence the term \(S(t) := \mathcal{Y}(t)^\top \ast \dot{\mathcal{Y}}(t) \in \text{Skew}(\mathbb{R}^{p \times p \times l})\). Next it is sufficient to obtain the formula for \(S(t)\).

Multiplying (4.36) by \(\mathcal{Y}(t)^\top\) on the left gives

\[
\mathcal{Y}(t)^\top \ast \dot{\mathcal{W}}(t) = \mathcal{Y}(t)^\top \ast \dot{\mathcal{Y}}(t) \ast \mathcal{P}(t) + \dot{\mathcal{P}}(t) = S(t) \ast \mathcal{P}(t) + \dot{\mathcal{P}}(t).
\]

(4.38)

From Lemma 4.3, \(\dot{\mathcal{P}}(t) \in T_X\text{Sym}(\mathbb{R}^{n \times p \times l}_+)\) for any \(\mathcal{X} \in \text{Sym}(\mathbb{R}^{n \times p \times l}_+)\), hence

\[
\mathcal{Y}(t)^\top \ast \dot{\mathcal{W}}(t) - \dot{\mathcal{W}}(t)^\top \ast \mathcal{Y}(t) = S(t) \ast \mathcal{P}(t) - \mathcal{P}(t)^\top \ast S(t)^\top = S(t) \ast \mathcal{P}(t) + \mathcal{P}(t) \ast S(t).
\]

(4.39)

Therefore, according to Theorem 4.11, a analytical solution for \(S(t)\) exists and is given by

\[
\text{vec}(S(t)) = \left(\text{bcirc}(\mathcal{P}(t))^\top \otimes I_p + [I_p]_{l \times l} \circ \text{bcirc}(\mathcal{P}(t))\right)^\dagger \cdot \text{vec}(\mathcal{Y}(t)^\top \ast \dot{\mathcal{W}}(t) - \dot{\mathcal{W}}(t)^\top \ast \mathcal{Y}(t)).
\]

(4.40)
By substitute (4.40) and (4.37) into (4.35), we obtain
\[
\dot{Y}(t) = Y(t) * S(t) + (I - Y(t) * Y(t)^T) * \dot{W}(t) * (P(t))^{-1}.
\]

Thus, there holds
\[
\mathcal{T}_t \mathcal{V} = \dot{Y}(0) = Y * S + (I - Y * Y^T) * \mathcal{V} * (Y^T * (X + U))^T,
\]
where $Y = R_X(U)$, vec$(S) = (\text{be circ}(P)^T \otimes I_p + [I_p]_{I \times I} \otimes \text{be circ}(P))^T \cdot \text{vec}(Y^T * \mathcal{V} - \mathcal{V}^T * Y)$ and $P = R_X(U)^T * (X + U) = (I + U^T * U)^{2}$. \hfill \Box

Next we derive the vector transport as the differentiated retraction based on t-Cayley transform.

**Theorem 4.13 (t-Cayley based vector transport).** The vector transport on $\text{St} (n, p, l)$ by differentiated retraction based on t-Cayley transform is
\[
\mathcal{T}_t \mathcal{V} = \left(I - \frac{1}{2}W_{\mathcal{V}}\right)^{-1} * \mathcal{V} * \left(I - \frac{1}{2}W_{\mathcal{V}}\right)^{-1} * \mathcal{X},
\]
where $\mathcal{X} \in \text{St} (n, p, l)$, $\mathcal{U}, \mathcal{V} \in T_X \text{St} (n, p, l)$, $W_{\mathcal{V}} = P * \mathcal{U} * X^T - \mathcal{X} * U^T * P \in \text{Skew}(R^{n \times n \times l})$, $W_{\mathcal{V}} = P * \mathcal{V} * X^T - \mathcal{X} * V^T * P \in \text{Skew}(R^{n \times n \times l})$, and $P = I - \frac{1}{2}X * X^T$.

**Proof.** Consider the following retraction based on t-Cayley transform:
\[
R_X(U + t\mathcal{V}) = \left(I - \frac{1}{2}W_{\mathcal{V}} - \frac{t}{2}W_{\mathcal{V}}\right)^{-1} * \left(I + \frac{1}{2}W_{\mathcal{V}} + \frac{t}{2}W_{\mathcal{V}}\right) * \mathcal{X}.
\]
Differentiating both sides of
\[
\left(I - \frac{1}{2}W_{\mathcal{V}} - \frac{t}{2}W_{\mathcal{V}}\right) * R_X(U + t\mathcal{V}) = \left(I + \frac{1}{2}W_{\mathcal{V}} + \frac{t}{2}W_{\mathcal{V}}\right) * \mathcal{X}
\]
with respect to $t$, we have
\[
-\frac{1}{2}W_{\mathcal{V}} * R_X(U + t\mathcal{V}) + \left(I - \frac{1}{2}W_{\mathcal{V}} - \frac{t}{2}W_{\mathcal{V}}\right) * \frac{d}{dt} R_X(U + t\mathcal{V}) = \frac{1}{2}W_{\mathcal{V}} * \mathcal{X}.
\]
According to (3.12), the vector transport by differentiated retraction is
\[
\mathcal{T}_t \mathcal{V} = \frac{d}{dt} R_X(U + t\mathcal{V})\big|_{t=0}
\]
\[
= \frac{1}{2} \left(I - \frac{1}{2}W_{\mathcal{V}}\right)^{-1} * \mathcal{V} * (\mathcal{X} + R_X(U))
\]
\[
= \frac{1}{2} \left(I - \frac{1}{2}W_{\mathcal{V}}\right)^{-1} * \mathcal{V} * \left(I + \frac{1}{2}W_{\mathcal{V}}\right)^{-1} * \left(I + \frac{1}{2}W_{\mathcal{V}}\right) * \mathcal{X}
\]
\[
= \left(I - \frac{1}{2}W_{\mathcal{V}}\right)^{-1} * \mathcal{V} * \left(I - \frac{1}{2}W_{\mathcal{V}}\right)^{-1} * \mathcal{X},
\]
where the last equality comes from $W_{\mathcal{V}} * \mathcal{X} = U$. \hfill \Box

Finally, we introduce an isometric vector transport.

**Theorem 4.14 (isometric vector transport).** The following formulae is an isometric vector transport on $\text{St} (n, p, l)$
\[
\mathcal{T}_t \mathcal{V} = \left(I - \frac{1}{2}W_{\mathcal{V}}\right)^{-1} * \left(I + \frac{1}{2}W_{\mathcal{V}}\right) * \mathcal{V},
\]
(4.41)
where $\mathcal{X} \in \text{St} (n, p, l)$, $\mathcal{U}, \mathcal{V} \in T_X \text{St} (n, p, l)$, $W_{\mathcal{V}} = P * U * X^T - \mathcal{X} * U^T * P \in \text{Skew}(R^{n \times n \times l})$ and $P = I - \frac{1}{2}X * X^T$. 

21
Proof. Consider the following retraction based on t-Cayley transform:
\[
R_X(U) = \left( I - \frac{1}{2} W_U \right)^{-1} * \left( I + \frac{1}{2} W_U \right) * X.
\]
Since \( W_U \in \text{Skew}(\mathbb{R}^{n \times n \times l}) \) and \((I - A) * (I + B) = (I + B) * (I - A)\) for all \(A, B \in \mathbb{R}^{n \times p \times l}\), we have
\[
\mathcal{T}_U V^T * R_X(U) + R_X(U)^T * \mathcal{T}_U V = V^T * X + X^T * V = O,
\]
which combines Theorem 4.2 lead to \( \mathcal{T}_U V \in T_{R_X(U)} \text{St}(n, p, l) \). It is easy to check that \( \mathcal{T}_O V = V \) and for \( V_1, V_2 \in T_X \text{St}(n, p, l) \). The smoothness follows immediately from (4.41). According to Definition 3.13, \( T \) is indeed a vector transport on \( \text{St}(n, p, l) \). It follows from the skew-symmetry of \( W_U \) that \( \langle \mathcal{T}_U(V), \mathcal{T}_U(V) \rangle_{R_X(U)} = \langle V, V \rangle_X \) for all \( U, V \in T_X \text{St}(n, p, l) \).

5 Examples of (1.2)

We present some (potential) examples of (1.2) and related problems in this section.

**Best approximation.** Given \( A \in \mathbb{R}^{n \times p \times l} \), its best \( k \)-term approximation was given in [25, Thm. 4.3]. Such a problem can also be formulated as \((k \leq \min(n, p)) \): \( \min_{U \in \text{St}(n, k, l), S \in \mathbb{R}^{k \times k \times l}, V \in \text{St}(p, k, l)} \| A - U * S * V \|^2_F \).

By using Proposition 2.5 to eliminate the variable \( S \), such a problem is equivalent to
\[
\max_{U \in \text{St}(n, k, l), V \in \text{St}(p, k, l)} \| U^T * A * V \|^2_F. \tag{5.42}
\]

By denoting \( W := \frac{O_U}{O} \in \mathbb{R}^{(n+p) \times 2k \times l} \) with \( O \) the zero tensor of proper size and reformulating the objective function accordingly, such a problem is of the form (1.2).

Given \( A \in \mathbb{R}^{n \times n \times l} \) which is symmetric, its eigenvalue decomposition was given in [53]. Correspondingly, one can define its best \( k \)-term symmetric approximation as \((k \leq n) \): \( \min_{U \in \text{St}(n, k, l), S \in \mathbb{R}^{n \times n \times l}} \| A - U * S * U^T \|^2_F \).

Similarly, the variable \( S \) can be eliminated and the problem is equivalent to
\[
\max_{U \in \text{St}(n, k, l)} \| U^T * A * U \|^2_F, \tag{5.43}
\]
which is of the form (1.2). Similar to their matrix counterparts, (5.42) and (5.43) are also equivalent to
\[
\min_{U \in \text{St}(n, k, l), V \in \text{St}(p, k, l)} -\text{tr} (U^T * A * V) \quad \text{and} \quad \min_{U \in \text{St}(n, k, l)} -\text{tr} (U^T * A * U). \tag{5.44}
\]

**Best approximation with missing entries.** In real-world applications, we are sometimes faced the situation that part of the observation data is missing; this troubles the approximation problem. Similar to tensor completion, one can formulate the problem as
\[
\min_{U \in \text{St}(n, k, l), S \in \mathbb{R}^{k \times k \times l}, V \in \text{St}(p, k, l)} \| O \odot (A - U * S * V^T) \|^2_F,
\]
where \( \odot \) denotes the Hadamard operator and \( O \in \mathbb{R}^{n \times p \times l} \) is a 0-1 tensor whose entries take 1 if the associated entries of \( A \) are available and 0 otherwise. The variable \( S \) cannot be eliminated and such a problem is a variant of (1.2) that can be possibly solved in an alternating fashion.

In the symmetric tensors setting, similar troubles might occur. Such a problem is thus formulated as
\[
\min_{U \in \text{St}(n, k, l), S \in \mathbb{R}^{k \times k \times l}} \| O \odot (A - U * S * U^T) \|^2_F. \tag{5.45}
\]
Joint f-diagonalization. The connection of simultaneous f-diagonalization to commutative tensors was discovered theoretically in [30,36]. For more than two tensors, the joint f-diagonalization is difficult in theory. In the matrix case, however, this can be resolved numerically by formulating the problem as optimization models over orthogonal or non-orthogonal constraints; see, e.g., [6]. Similarly, for joint f-diagonalization of more than two tensors of size $n \times n \times l$, one can consider the following optimization models:

$$
\min_{U \in \text{St}(n,k,l)} \sum_{i=1}^{N} \text{off}(U^\top \ast A_i \ast U),
$$

(5.46)

where $A_i \in \mathbb{R}^{n \times n \times l}$, $i = 1, \ldots, N$, and $\text{off}(\mathcal{X}) = \sum_{i=1}^N \sum_{1 \leq i \neq j \leq k} (x_{i_1,i_2,i_3})^2$ or $= \sum_{i=1}^N \sum_{1 \leq i \neq j \leq k} |x_{i_1,i_2,i_3}|$, which is analogous to its matrix counterpart.

Joint t-SVD. We first consider the matrix cases. Assume that $N$ matrices $A_1, \ldots, A_N \in \mathbb{R}^{n \times p}$ are given, which are regarded as $N$ samples. The joint SVD is to find common orthogonal matrices $U \in \mathbb{R}^{n \times k}$, $V \in \mathbb{R}^{p \times l}$, $k \leq \min\{n,p\}$, such that $U^\top A_i V$ are as diagonal as possible. The joint SVD is useful in image representation and dimension reduction; see, e.g., [38,40]. Now assume that the samples are third-order tensors $A_1, \ldots, A_N \in \mathbb{R}^{n \times p \times l}$. To perform dimension reduction on these samples, it is quite natural to extend such an idea to obtain the following joint t-SVD models:

$$
\min_{U \in \text{St}(n,k,l), V \in \text{St}(p,k,l)} \sum_{i=1}^{N} \text{off}(U^\top \ast A_i \ast V).
$$

Sparse tensor PCA. Sparse tensor PCA was introduced in [2] and then its solution methods were further studied in [34,49]. The purpose is to find sparse principal components for higher-order data. The model of [2] is based on the canonical polyadic format and is approximately solved by a deflation approach, where the orthogonality cannot be assured. However, by using t-product, it is more natural to directly extend the sparse matrix PCA to the third-order tensor setting, leading to the following model:

$$
\min_{U \in \text{St}(n,k,l)} \text{tr}(U^\top \ast A \ast A^\top \ast U) + \rho \|U\|_1,
$$

(5.47)

where $A \in \mathbb{R}^{n \times p \times l}$ is the data tensor consisting of samples, $\|U\|_1 = \sum_{i=1}^{n\times k\times l} |u_{i_1,i_2,i_3}|$, and $\rho > 0$. Based on the formulas derived in this work, it is expected that the recently developed Riemannian algorithms, such as [7,20], can be applied to the above problem without many modifications.

Besides the above basic examples, some applications have been or can be formulated as optimization over the tensor Stiefel manifold in the literature; see, e.g., [15,29,37,50,51]. For instance, [29] proposed a tensor subspace representation method for hyperspectral image denoising, where the model is exactly an optimization over the tensor Stiefel manifold. To save space, we do not introduce them in detail here; interested readers can be referred to them.

We make the following remark to end this section.

Remark 5.1. Considering the relation (2.5), one may wonder whether (1.2) can be solved in the Fourier domain. That is to say, since $X = L^{-1}L(X)$ is equivalent to minimizing $f(L^{-1}(\hat{X}))$ subject to $(\hat{X}^{(i)})^H \hat{X}^{(i)} = I_p$, $i = 1, \ldots, \lfloor \frac{l+1}{2} \rfloor$, $\hat{X}^{(i)} = \text{conj}\left(\hat{X}^{(i+2-i)}\right)$, $i = 1 + \lfloor \frac{l+1}{2} \rfloor, \ldots, l$, which is minimizing a real-valued function over the product of $k$ real matrix Stiefel manifolds and $\frac{l+1}{2}$ complex matrix Stiefel manifolds (when $l$ is even, $k = 2$; when $l$ is odd, $k = 1$). Comparing with (1.2), solving this problem may have some drawbacks. Firstly, it is in the complex field which is more complicated to analyze than (1.2); secondly, the introduction of $L^{-1}$ in the problem might destroy certain structure of the problem; thirdly, if $f$ is nonsmooth, such as $f$ involves the $\ell_1$ norm as that in the sparse tensor PCA model (5.47), then as far as we know, no Riemannian algorithms can handle problems involving the term $\|L^{-1}(\hat{X})\|_1$. 

23
6 Preliminary Numerical Examples

We conducted preliminary numerical experiments to verify the derived formulas. To this end, we applied a Riemannian CG algorithm summarized in Algorithm 1 with various retractions to the four problems introduced in Section 5, namely, (5.44), (5.45), (5.46), and (5.47). The considered retractions were the t-QR based one (4.21), the t-PD based one (4.22), and the t-Cayley transform based one (4.23), while the vector transports were performed by the orthogonal projector (4.24) associated with the various retractions mentioned above. All the experiments were conducted on an Intel i7 CPU desktop computer with 16 GB of RAM. The supporting software is Matlab R2022a. Tensorlab [48] and Tensor-Tensor Product Toolbox [31] were employed for tensor operations.

Algorithm 1 was modified from [54] to the t-product setting. In the algorithm, we set $\beta_{k+1} = \min\{\beta_{k+1}^{\text{FR}}, \beta_{k+1}^{\text{D}}\}$ where

$$
\beta_{k+1}^{\text{D}} = \frac{\|\nabla f(X_{k+1})\|_F^2}{\max\{\langle \nabla f(X_{k+1}), T_{\alpha_k}z_k(Z_k)\rangle_{X_{k+1}} - \langle \nabla f(X_k), Z_k/X_k \rangle, \left(\langle \nabla f(X_k), Z_k/X_k \rangle\right)^2\}}
$$

(6.48)

is a generalization of Dai’s nonmonotone parameter [54] and

$$
\beta_{k+1}^{\text{FR}} = \|\nabla f(X_{k+1})\|_F\|\nabla f(X_k)\|_F^2
$$

(6.49)

is the Fletcher–Reeves parameter. The steplength $\alpha_{k+1} = \max\{\alpha_{k+1}^{\text{BB}}, \alpha_{\max}\}$, where \(\alpha_{k+1}^{\text{BB}} = \langle S_k, S_k/X_k \rangle / \langle S_k, V_k/X_k \rangle \),

(6.50)

with $S_k = -\alpha_k T_{\alpha_k}z_k(\nabla f(X_k))$, $V_k = \nabla f(X_{k+1}) + \alpha_k^{-1}S_k$, which is a Riemannian generalization of the Barzilai-Borwein steplength [21]. The iterative algorithms were stopped if $\|X_{k+1} - X_k\|_F/\sqrt{n} < 10^{-6}$ or $|f(X_{k+1}) - f(X_k)|/(1 + |f(X_k)|) < 10^{-12}$ or $k > 1000$. In the algorithm, the initial steplength $\alpha_0 = 10^{-3}$, $\alpha_{\min}, \alpha_{\max} = (10^{-20}, 1); \lambda = 0.2$, and $\delta = 10^{-4}$. We remark that in the problem (5.47), for simplicity, we just used the subgradient of the objective function in our computation; in (5.45), $U$ and $S$ are computed in an alternating fashion, where $U$ is computed by one step of Algorithm 1.

In the experiments, we set $(n, p, l) = (50, 10, 8)$. The data tensors in (5.44) is set to be $A = V^T \ast V$, where $V = \text{randn}(n, n, l)$. In (5.45), $A = X^T \ast W \ast X^T$ where $X \in \text{St}(n, p, l)$ and the $f$-diagonal tensor $W \in \mathbb{R}^{p \times p \times l}$ were both randomly generated; the entries were randomly missing with missing ratio being 30%. In (5.46), the tensors were constructed as $A_i = X^T C_i X^T + \rho \frac{E_i}{\|E_i\|_F}$, $i \in [N]$, where the number of samples $N = 3$, $X \in \text{St}(n, p, l)$, the $f$-diagonal tensors $C_i \in \mathbb{R}^{p \times p \times l}$, and the noise term $E_i \in \mathbb{R}^{n \times n \times l}$ were all randomly generated with the noise level $\rho = 0.1$. In (5.47), $A = \text{randn}(n, p, l)$ and the parameter $\rho = 0.1$. All the initial points for the algorithm $X_0$ were randomly generated feasible points. For each case, we randomly generated 50 instances, and the averaged results are presented. Figure 1 shows the curves of the objective values of the four test problems versus iterations, whose colors are respectively green (t-QR based retraction), blue (t-PD based retraction), cyan (t-Cayley transform based retraction).

The performance and the feasibility of Riemannian CG on the problems with various retractions are reported in Tables 1 and 2, respectively. “obj.” stands for the objective value, “feasi.” specifies the feasibility $\|X^T X - I\|_F$, “iter.” means the iterations, “time.” represents the CPU time where the unit is second, and “re.” means the relative error: specifically, “re.” $= |\|X^T W X^T - U_{\text{out}} \ast S_{\text{out}} \ast U_{\text{out}}^T\|_F/\|V\|_F|$ in (5.45) and “re.” $= \left(\sum_{i=1}^{N} |\|X^T C_i X^T - U^T_{\text{out}} \ast A_i \ast U_{\text{out}} + U_{\text{out}}^T \ast A_i \ast U_{\text{out}}^T\|_F/\|C_i\|_F\right)/N$ in (5.46), where $U_{\text{out}}$ is generated by the Algorithm 1 and $S_{\text{out}}$ is generated by the nonmonotone gradient method with Barzilai-borwein step size on $\mathbb{R}^{p \times p \times l}$ in an alternating fashion. Empirically, we can observe that the algorithm converges in all the examples, indicating that the derived formulas (using orthogonal projector based vector transport) are correct.
Figure 1: Objective values on four test problem versus iterations with different retractions

Table 1: The performance of Riemannian CG method on four test problem with various retractions

| Retraction | Best approximation | with missing entries | Joint f-diagonalization | Sparse tensor PCA |
|------------|--------------------|----------------------|-------------------------|-------------------|
|            | obj. | iter. | time. | re. | iter. | time. | re. | iter. | time. | obj. | iter. | time. |
| t-QR       | -8.37E+04 | 6.00E+01 | 5.39E-01 | 1.59E-01 | 1.72E+01 | 2.24E-03 | 2.80E+01 | 2.31E+00 | -3.22E+04 | 3.63E+02 | 2.70E+00 |
| t-PD       | -8.37E+04 | 5.10E+01 | 4.98E-01 | 1.62E-02 | 6.62E+02 | 3.38E+01 | 5.10E+01 | 4.18E+00 | -3.22E+04 | 1.00E+03 | 8.24E+00 |
| t-Cayley   | -8.37E+04 | 5.50E+01 | 1.04E+00 | 6.83E-03 | 5.06E+02 | 3.08E+01 | 5.70E+01 | 5.19E+00 | -3.22E+04 | 1.00E+03 | 1.78E+01 |

Algorithm 1 A Riemannian nonmonotone conjugate gradient method on St (n, p, l)

Input: X₀ ∈ St (n, p, l), α₀ = 10⁻³, (α_min, α_max) = (10⁻²₀, 1).
Output: {X_k}, {f(X_k)} and {grad f(X_k)}.

1: Set k = 0, Z₀ = −grad f(X₀).
2: while ∥X_{k+1} − X_k∥P/√n > 10⁻⁶ or |f(X_{k+1}) − f(X_k)|/(1 + |f(X_k)|) > 10⁻¹² or k < 1000 do
3: if f(R_{X_k}(α_k Z_k)) ≤ max{f(X_k), f(X_{k-1})} + 10⁻⁴α_k ⟨grad f(X_k), Z_k⟩X_k then
4:    Set X_{k+1} = R_{X_k}(α_k Z_k).
5: else
6:    Set α_k ← 0.2α_k and go to line 3.
7: end if
8: Compute Z_{k+1} = −grad f(X_{k+1}) + β_{k+1} T_{α_k Z_k}(Z_k), where β_{k+1} = min{β_F, β_D}.
9: Compute α_{k+1} = max{min{α_{BB}, α_max}, α_min}.
10: Set k ← k + 1.
11: end while
Table 2: The feasibility of Riemannian CG method on four test problem with various retractions

| Best approximation with missing entries | Joint f-diagonalization | Sparse tensor PCA |
|----------------------------------------|------------------------|-------------------|
| t-QR                                   | t-PD                   | t-Cayley          |
| 1.47E-15                               | 7.47E-15               | 2.26E-14          |
| 1.29E-15                               | 5.06E-15               | 1.29E-14          |
| 1.67E-13                               | 5.47E-15               | 2.09E-14          |
| 1.06E-15                               | 3.95E-13               | 3.95E-13          |

7 Concluding Remarks

Optimization over the matrix Stiefel manifold draws much attention in recent years. With the properties and decompositions built upon the t-product of third-order tensors, we study computation over the tensor Stiefel manifold $\text{St}(n,p,l) = \{X \in \mathbb{R}^{n \times p \times l} \mid X^\top \ast X = I\}$ in this work. Firstly, it is shown that $\text{St}(n,p,l)$ endowed with the Frobenius norm is known to admit a Riemannian manifold structure; then, explicit expressions over $\text{St}(n,p,l)$, such as the tangent space, Riemannian gradient, Riemannian Hessian, several retractions, and vector transports, are derived, which may serve as building blocks for designing and analyzing Riemannian algorithms over the tensor Stiefel manifold. As byproducts, we define the skew tensors, t-polar decomposition, and obtain the analytical solution to the tensor Sylvester equation in the t-product sense.

We also remark that although this work is focused on the tensor Stiefel manifold in the sense of t-product, it is straightforward to derive similar results in the sense of the more general tensor-tensor product [24].

In the future, it would be necessary to further study properties and computation concerning the tensor Stiefel manifold, and it would be interesting to find more instances of the form (1.2). In particular, we prefer to systematically study the sparse tensor PCA model (5.47) and algorithms in our future work.

Acknowledgment

This work was supported by National Natural Science Foundation of China Grant 12171105, Fok Ying Tong Education Foundation Grant 171094, and the special foundation for Guangxi Ba Gui Scholars. All authors equally contributed to this research.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] G. I. Allen. Sparse higher-order principal components analysis. In *International Conference on Machine Learning*, pages 27–36, April 2012.

[3] N. Boumal. An introduction to optimization on smooth manifolds. To appear with Cambridge University Press, Jan 2022.

[4] K. Braman. Third-order tensors as linear operators on a space of matrices. *Linear Algebra Appl.*, 433(7):1241–1253, 2010.

[5] P. Breiding and N. Vannieuwenhoven. A Riemannian trust region method for the canonical tensor rank approximation problem. *SIAM J. Optim.*, 28(3):2435–2465, 2018.
[6] A. Bunse-Gerstner, R. Byers, and V. Mehrmann. Numerical methods for simultaneous diagonalization. \textit{SIAM J. Matrix Anal. Appl.}, 14(4):927–949, 1993.

[7] S. Chen, S. Ma, A. M.-C. So, and T. Zhang. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. \textit{SIAM J. Optim.}, 30(1):210–239, 2020.

[8] A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. A. Phan. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. \textit{IEEE Signal Process. Mag.}, 32(2):145–163, 2015.

[9] P. Comon. Tensors: a brief introduction. \textit{IEEE Signal Process. Mag.}, 31(3):44–53, 2014.

[10] B. Gao, X. Liu, X. Chen, and Y.-X. Yuan. A new first-order algorithmic framework for optimization problems with orthogonality constraints. \textit{SIAM J. Optim.}, 28(1):302–332, 2018.

[11] K. Gilman, D. A. Tarzanagh, and L. Balzano. Grassmannian optimization for online tensor completion and tracking with the t-SVD. \textit{IEEE Trans. Signal Process.}, 70:2152–2167, 2022.

[12] B. C. Hall. \textit{Lie groups, Lie Algebras, and representations}. Springer, 2015.

[13] G. Heidel and V. Schulz. A Riemannian trust-region method for low-rank tensor completion. \textit{Numer. Linear Algebra Appl.}, 25(6):e2175, 2018.

[14] S. Holtz, T. Rohwedder, and R. Schneider. On manifolds of tensors of fixed TT-rank. \textit{Numer. Math.}, 120(4):701–731, 2012.

[15] R. C. Hoover, K. Caudle, and K. Braman. Multilinear discriminant analysis through tensor-tensor eigen-decomposition. In \textit{2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)}, pages 578–584. IEEE, 2018.

[16] J. Hu, B. Jiang, L. Lin, Z. Wen, and Y.-X. Yuan. Structured quasi-Newton methods for optimization with orthogonality constraints. \textit{SIAM J. Sci. Comput.}, 41(4):A2239–A2269, 2019.

[17] J. Hu, X. Liu, Z.-W. Wen, and Y.-X. Yuan. A brief introduction to manifold optimization. \textit{J. Oper. Res. Soc. China}, 8(2):199–248, 2020.

[18] W. Huang. \textit{Optimization algorithms on Riemannian manifolds with applications}. PhD thesis, The Florida State University, 2013.

[19] W. Huang, P.-A. Absil, and K. A. Gallivan. A Riemannian BFGS method without differentiated retraction for nonconvex optimization problems. \textit{SIAM J. Optim.}, 28(1):470–495, 2018.

[20] W. Huang and K. Wei. Riemannian proximal gradient methods. \textit{Math. Program.}, https://doi.org/10.1007/s10107-021-01632-3, 2021.

[21] B. Iannazzo and M. Porcelli. The Riemannian Barzilai–Borwein method with nonmonotone line search and the matrix geometric mean computation. \textit{IMA J. Numer. Anal.}, 38(1):495–517, 2018.

[22] E. Kernfeld, M. Kilmer, and S. Aeron. Tensor–tensor products with invertible linear transforms. \textit{Linear Alg. Appl.}, 485:545–570, 2015.

[23] M. E. Kilmer, K. Braman, N. Hao, and R. C. Hoover. Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging. \textit{SIAM J. Matrix Anal. Appl.}, 34(1):148–172, 2013.

[24] M. E. Kilmer, L. Horesh, H. Avron, and E. Newman. Tensor-tensor algebra for optimal representation and compression of multiway data. \textit{Proc. Natl. Acad. Sci. U. S. A.}, 118(28):e2015851118, 2021.
[25] M. E. Kilmer and C. D. Martin. Factorization strategies for third-order tensors. *Linear Alg. Appl.*, 435(3):641–658, 2011.

[26] T. G. Kolda. Multilinear operators for higher-order decompositions. Technical report, Citeseer, 2006.

[27] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Rev.*, 51:455–500, 2009.

[28] D. Kressner, M. Steinlechner, and B. Vandereycken. Low-rank tensor completion by Riemannian optimization. *BIT Numer. Math.*, 54(2):447–468, 2014.

[29] J. Lin, T.-Z. Huang, X.-L. Zhao, T.-X. Jiang, and L. Zhuang. A tensor subspace representation-based method for hyperspectral image denoising. *IEEE Tran. Geosci. Remote Sens.*, 59(9):7739–7757, 2020.

[30] W.-H. Liu and X.-Q. Jin. A study on T-eigenvalues of third-order tensors. *Linear Alg. Appl.*, 2020.

[31] C. Lu. *Tensor-Tensor Product Toolbox*. Carnegie Mellon University, June 2018. 
https://github.com/canyilu/tproduct.

[32] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(4):925–938, 2019.

[33] K. Lund. The tensor t-function: A definition for functions of third-order tensors. *Numer. Linear Algebr. Appl.*, 27(3):e2288, 2020.

[34] X. Mao and Y. Yang. Several approximation algorithms for sparse best rank-1 approximation to higher-order tensors. *J. Glob. Optim.*, https://doi.org/10.1007/s10898-022-01140-4, 2022.

[35] Y. Miao, L. Qi, and Y. Wei. Generalized tensor function via the tensor singular value decomposition based on the T-product. *Linear Alg. Appl.*, 590:258–303, 2020.

[36] Y. Miao, L. Qi, and Y. Wei. T-Jordan canonical form and t-Drazin inverse based on the t-product. *Commun. Appl. Math. Comput. Sci.*, 3(2):201–220, 2021.

[37] C. Ozdemir, R. C. Hoover, K. Caudle, and K. Braman. High-order multilinear discriminant analysis via order-n tensor eigendecomposition. *arXiv preprint arXiv:2205.09191*, 2022.

[38] J.-C. Pesquet-Popescu, B. and Pesquet and A. P. Petropulu. Joint singular value decomposition—a new tool for separable representation of images. In *ICIP*, volume 2, pages 569–572. IEEE, 2001.

[39] L. Qi and Z. Luo. Tubal matrix. *arXiv preprint arXiv:2105.00793*, 2021.

[40] A. Shashua and A. Levin. Linear image coding for regression and classification using the tensor-rank principle. In *CVPR*, volume 1, pages I–I. IEEE, 2001.

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

[42] G.-J. Song, X.-Z. Wang, and M. K. Ng. Riemannian conjugate gradient descent method for third-order tensor completion. *arXiv preprint arXiv:2011.11417*, 2020.

[43] M. Steinlechner. Riemannian optimization for high-dimensional tensor completion. *SIAM J. Sci. Comput.*, 38(5):S461–S484, 2016.

[44] L. W. Tu. *An Introduction to Manifolds*. Universitext. Springer-Verlag New York, second edition, 2011.

[45] A. Uschmajew and B. Vandereycken. The geometry of algorithms using hierarchical tensors. *Linear Alg. Appl.*, 439(1):133–166, 2013.
A Appendix

A.1 Proof of Proposition 2.2

Proof. Taking the conjugate transpose of both sides of the equation in item 1 of Proposition 2.1, then multiplying both sides by \((F_l \otimes I_n)\), we get

\[(F_l \otimes I_p) \text{bcirc}(A^\top) = \text{Diag}\left(\hat{A}(i)^H : i \in [l]\right) (F_l \otimes I_n).\]

Taking the first column of block matrix on both sides of the above equation yields

\[(F_l \otimes I_p) \text{unfold}(A^\top) = \frac{1}{\sqrt{l}} \text{Diag}\left(\hat{A}(i)^H : i \in [l]\right) \text{Vec}\left(I_n : i \in [l]\right),\]

which combing with Definition 2.6 and (A.51) gives \(L(A^\top) = \text{fold}\left(\hat{A}(i)^H : i \in [l]\right)\). \(\square\)

A.2 Proof of Lemma 2.1

Proof. According to [53, Thm. 5], \(I + V^\top V \in \text{Sym}(R_{++}^{p \times p \times l})\) if only if \(I_p + (\hat{V}(i))^H \hat{V}(i), i \in [l]\) are Hermitian positive definite. \(\square\)
A.3 The proof of Lemma 2.2

Proof. For any tensor $A \in (\text{Skew}(\mathbb{R}^{p \times p \times l}))^\perp$, there holds
\[
\langle A - A^T, A - A^T \rangle = \langle A - A^T, A \rangle - \langle A - A^T, A^T \rangle = \langle A - A^T, A \rangle - \langle A^T - A, A \rangle = 0,
\]
where the last equation follows from $A - A^T \in \text{Skew}(\mathbb{R}^{p \times p \times l})$. Thus $A \in \text{Sym}(\mathbb{R}^{p \times p \times l})$. \qed

A.4 Proof of Proposition 2.3

Proof. We denote $\mathcal{D} := \hat{A}^T \in \mathbb{R}^{p \times n \times l}$ and $\mathcal{C} := \mathcal{D} \ast \mathcal{B} \in \mathbb{R}^{p \times p \times l}$. Then according to Definition 2.7 and by the definition of $\hat{C}$ in item 1 of Proposition 2.1, we get
\[
\text{tr} (\mathcal{C}) = \sum_{i=1}^{l} \text{tr} (\hat{C}_i) = \text{tr} (\hat{C}) .
\]
Using item 1 of Proposition 2.1 again, we have
\[
\text{tr} (\hat{C}) = \text{tr} (\hat{D} \hat{B}) = \text{tr} (\left((F_i \otimes I_p) \ast \text{bcirc}(\mathcal{D}) \ast (F_i^H \otimes I_n) \ast \text{bcirc}(\mathcal{B}) \ast (F_i^H \otimes I_p)\right)) = \text{tr} (\left(\text{bcirc}(\hat{A}^T) \ast \text{bcirc}(\mathcal{B})\right)) = l \langle \hat{A}, \mathcal{B} \rangle ,
\]
where the third line uses [36, Lem. 3] and the last equation comes from [53, Rmk. 9]. \qed

A.5 Proof of Theorem 2.3

Proof. t-QR was proposed in [23, Sect. 5]. Similar to t-SVD, to compute t-QR, we would only need to compute individual matrix QR’s for about half the frontal slices of $\hat{A}$ and the remaining part is obtained by the conjugate symmetry of the Fourier transform. Specifically, for $i = 1, \ldots, \left\lceil \frac{l+1}{2} \right\rceil$, let $\hat{A}^{(i)} = \hat{Q}^{(i)} \ast \hat{R}^{(i)}$ be the QR decomposition of $\hat{A}^{(i)} \in \mathbb{C}^{n \times p}$ \footnote{For QR factorization of complex matrices, we can choose that $R$ factor is upper triangular with real nonzero diagonal elements.} where $\hat{Q}^{(i)} \in \mathbb{C}^{n \times p}$, $(\hat{Q}^{(i)})^H$, $\hat{U}^{(i)} = I_p$, $\hat{R}^{(i)} \in \mathbb{C}_{\text{upp}}^{p \times p}$ and $\text{diag}(\hat{R}^{(i)}) \in \mathbb{R}^{p \times p}$, namely, the diagonal entries of $\hat{R}^{(i)}$ are real. For $i = 1 + \left\lceil \frac{l+1}{2} \right\rceil, \ldots, l$, $\hat{A}^{(i)} = \text{conj} \left(\hat{A}^{(i-1)}\right)$, $\hat{Q}^{(i)} = \text{conj} \left(\hat{U}^{(l+2-i)}\right)$, $\hat{R}^{(i)} = \text{conj} \left(\hat{R}^{(l+2-i)}\right)$. It follows from Remark 2.3 and Remark 2.4 that $Q \in \mathbb{R}^{n \times p \times l} \in \text{St} (n, p, l)$ and $\mathcal{R} \in \mathbb{R}_{\text{upp}}^{p \times p \times l}$. Here $\mathcal{R}$ to be real is because of Remark 2.3 and direct computation. Using Remark 2.3 again, further we have $\hat{A}^{(1)} \in \mathbb{R}^{n \times p}, \hat{Q}^{(1)} \in \mathbb{R}^{n \times p}, \hat{R}^{(1)} \in \mathbb{R}^{p \times p}$.

We then show the uniqueness of the decomposition. As we know, for QR decomposition of a matrix $\hat{A}^{(i)} \in \mathbb{C}^{n \times p}$ with $n \geq p$, if $A^{(i)}$ is of full rank $p$, then the QR decomposition $\hat{A}^{(i)} = \hat{Q}^{(i)} \ast \hat{R}^{(i)}$ is unique if we require that the diagonal entries of $\hat{R}^{(i)}$ are all positive, i.e., $\mathcal{R} \in \mathbb{R}_{\text{upp}}^{p \times p \times l}$. Since the Fourier transform is bijective, the uniqueness of the matrix QR decomposition leads to the uniqueness of the t-QR decomposition. \qed

A.6 Proof of Lemma 2.3

Proof. Theorem 2.3 shows that $L^{-1} \left( \mathbb{C}_{\text{upp}}^{p \times p \times l} \right)$ is isomorphic to
\[
\mathbb{C}_{\text{upp}}^{p \times p \times l} = \left\{ R \right| \hat{R}^{(1)} \in \mathbb{R}_{\text{upp}}^{p \times p}, \hat{R}^{(i)} \in \mathbb{C}_{\text{upp}}^{p \times p}, \text{diag}(\hat{R}^{(i)}) \in \mathbb{R}^{p \times p}, i = \{i\} \setminus \{1\}, \hat{R}^{(i)} = \text{conj}(\hat{R}^{(l+2-i)})$, $i = 2, \ldots, \left\lceil \frac{l+1}{2} \right\rceil \right\} .
\]
If \( l \) is even, then it holds that
\[
\mathcal{C}^{p \times p \times l}_{\text{upp} +} = \left\{ \tilde{R} \left| \tilde{R}^{(1)} \in \mathbb{R}^{p \times p}, \tilde{R}^{(i)} \in \mathbb{C}^{p \times p}, \text{diag}(\tilde{R}^{(i)}) \in \mathbb{R}^{p \times p}, i = [l] \setminus \{1\}, \tilde{R}^{(i)} = \text{conj}(\tilde{R}^{(l+2-i)}), i = 2, \ldots, \frac{l}{2} \right\}.
\]

There are two real upper triangular \( p \times p \) matrices, whose dimensions are \( \frac{(1+p)p}{2} \); there are \( \frac{l-2}{2} \) pairs of complex upper triangular \( p \times p \) matrices with positive diagonal elements, whose dimensions are \( \frac{(p-1)p}{2} \times 2 + p \). Hence the dimension of \( \mathcal{C}^{p \times p \times l}_{\text{upp} +} \) is \( 2 \times \frac{(1+p)p}{2} + \frac{l-2}{2} \times \left( \frac{(p-1)p}{2} \times 2 + p \right) = \frac{p^2 l + p}{2} \).

If \( l \) is odd, then it holds that
\[
\mathcal{C}^{p \times p \times l}_{\text{upp} +} = \left\{ \tilde{R} \left| \tilde{R}^{(1)} \in \mathbb{R}^{p \times p}, \tilde{R}^{(i)} \in \mathbb{C}^{p \times p}, \text{diag}(\tilde{R}^{(i)}) \in \mathbb{R}^{p \times p}, i = [l] \setminus \{1\}, \tilde{R}^{(i)} = \text{conj}(\tilde{R}^{(l+2-i)}), i = 2, \ldots, \frac{l+1}{2} \right\}.
\]

There is one real upper triangular \( p \times p \) matrix, whose dimensions is \( \frac{(1+p)p}{2} \); there are \( \frac{l-1}{2} \) pairs of complex upper triangular \( p \times p \) matrices with positive diagonal elements, whose dimensions are \( \frac{(p-1)p}{2} \times 2 + p \). Hence the dimension of \( \mathcal{C}^{p \times p \times l}_{\text{upp} +} \) is \( \frac{(1+p)p}{2} + \frac{l-1}{2} \times \left( \frac{(p-1)p}{2} \times 2 + p \right) = \frac{p^2 l + p}{2} \). \( \square \)

### A.7 Proof of Theorem 2.4

**Proof.** Let the compact t-SVD of \( A = U \ast S \ast V^T \). Let \( \mathcal{P} := U \ast V^T \) and \( \mathcal{H} := V \ast S \ast V^T \). Then it is clear that (2.8) is satisfied. To see that \( \mathcal{H} \in \text{Sym}(\mathbb{R}^{p \times p \times l}_+) \), first we show that \( S \in \text{Sym}(\mathbb{R}^{p \times p \times l}_+) \). This is obvious, as each \( \hat{S}^{(i)} \) is diagonal with nonnegative entries, and so \( S \in \text{Sym}(\mathbb{R}^{p \times p \times l}_+) \), according to Remark 2.5. By [53, Thm. 7], there is a unique \( T \) such that \( T \ast T^T = S \). Then \( \mathcal{H} \) can be written as \( \mathcal{H} = V \ast T \ast (V \ast T)^T \), which together with [53, Thm. 8] shows that \( \mathcal{H} \in \text{Sym}(\mathbb{R}^{p \times p \times l}_+) \).

To show the uniqueness of \( \mathcal{H} \), note that \( A^T \ast A = \mathcal{H} \ast \mathcal{H} \), which by [53, Thm. 8] is clearly symmetric positive semidefinite. Revoking again [53, Thm. 7] gives the uniqueness of \( \mathcal{H} \).

If \( A^T \ast A \in \text{Sym}(\mathbb{R}^{p \times p \times l}_+) \), [53, Thm. 8] shows that \( \mathcal{H} \) is nonsingular (invertable, Def. 2.5), and so \( \mathcal{P} = A \ast \mathcal{H}^{-1} \), which is unique. \( \square \)

**Remark A.1.** The proof of Theorem 2.4 gives the way to obtain t-PD from the compact t-SVD. This is analogous to the matrix case.

### A.8 Proof of Proposition 2.6

**Proof.** This can be easily derived from the proof of Theorem 2.4. Here the root of a symmetric positive definite tensor was defined in [53, Thm. 7]. \( \square \)

### A.9 Proof of Proposition 2.7

**Proof.** If \( \hat{A} \in \mathbb{C}^{p \times p \times l}_+ \), then \( (\hat{A}^{(i)})^H \hat{A}^{(i)}, i \in [l] \) are Hermitian positive definite. Note that [53, Thm. 5] shows that \( (\hat{A}^{(i)})^H \hat{A}^{(i)}, i \in [l] \) are Hermitian positive definite if only if \( A^T \ast A \in \text{Sym}(\mathbb{R}^{p \times p \times l}_+) \). \( \square \)
A.10 Proof of Theorem 2.5

Proof. Let \( D := U^\top \in \mathbb{R}^{p \times n \times l} \). Then for any \( P \in \text{St} (n, p, l) \),

\[
l (A, P) = \text{tr} (A^\top * P) = \text{tr} (V * S * U^\top * P) = \text{tr} (S * D * P * V)
\]

\[
= \text{tr} (\hat{S} \hat{D} \hat{P} \hat{V}) = \sum_{i=1}^l \text{tr} (\hat{S}^i \hat{D}^i \hat{P}^i \hat{V}^i) = \sum_{i=1}^l \text{tr} (\hat{S}^i \hat{W}^i),
\]

where we let \( \hat{W}^i := \hat{D}^i \hat{P}^i \hat{V}^i \in \mathbb{C}^{p \times p} \). Note that \( \hat{D}^i (\hat{D}^i)^H = I, (\hat{P}^i)^H \hat{P}^i = I, (\hat{V}^i)^H \hat{V}^i = I \).

Thus \( |(\hat{W}^i)_{jj}| \leq 1, i \in [l], j \in [p] \). Therefore,

\[
\sum_{i=1}^l \text{tr} (\hat{S}^i \hat{W}^i) = \sum_{i=1}^l \sum_{j=1}^p (\hat{S}^i)_{jj} (\hat{W}^i)_{jj} \leq \sum_{i=1}^l \sum_{j=1}^p (\hat{S}^i)_{jj} |(\hat{W}^i)_{jj}| \leq \sum_{i=1}^l \text{tr} (\hat{S}^i) = \text{tr} (\hat{S}),
\]

where \( \hat{S}^i \geq 0 \). On the other hand, take \( P := U * V^\top \). It is easy to see that

\[
l (A, P) = \text{tr} (\hat{S}),
\]

namely, the upper bound is tight, which is achieved when \( P = U * V^\top \). This gives the desired result. \( \square \)

A.11 Proof of the well-defined property of (2.10)

Proof. To be convenient, we will use the notation \( \Delta \) as the frontal-slice-wise product (cf. [22, Def. 2.1]) between two tensors in the Fourier domain, i.e., if \( \hat{C}^i = \hat{A}^i \hat{B}^i, i \in [l] \), then it holds that \( L(A) \Delta L(B) = \text{fold} \left( \hat{A}^i \hat{B}^i : i \in [l] \right) \); in other words,

\[
L(C) = L(A) \Delta L(B) \Leftrightarrow \hat{C}^i = \hat{A}^i \hat{B}^i, i \in [l] \Leftrightarrow C = A \ast B.
\]

(A.51)

Using this notation, we have

\[
A^k = L^{-1} \left( L(A) \Delta \cdots \Delta L(A) \right) = L^{-1} \left( \text{fold} \left( \hat{A}^i : i \in [l] \right) \right).
\]

Thus for any \( N \),

\[
\sum_{k=0}^N \frac{1}{k!} A^k = \sum_{k=0}^N \frac{1}{k!} L^{-1} \left( \text{fold} \left( \hat{A}^i : i \in [l] \right) \right) = L^{-1} \left( \text{fold} \left( \sum_{k=0}^N \frac{1}{k!} (\hat{A}^i)^k : i \in [l] \right) \right).
\]

Let \( N \to \infty \), it holds that

\[
\exp [A] = L^{-1} \left( \text{fold} \left( \exp \left( \hat{A}^i \right) : i \in [l] \right) \right) = L^{-1} \left( \text{fold} \left( \exp \left( L(A)^i \right) : i \in [l] \right) \right),
\]

(A.52)

since the series defining the matrix exponential is convergent [12, Prop. 2.1]. \( \square \)
A.12 Proof of equivalence of (2.11) and (2.9)

Proof. Using (2.9) and item 1 of Proposition 2.1, we have

\[
\exp [A] = \text{fold} \left( \exp \left[ \text{bcirc}(A) \right] \text{unfold}(I) \right) \\
= \text{fold} \left( \exp \left[ (F_t^H \otimes I_n) \text{Diag} \left( A^{(i)} : i \in [l] \right) (F_t \otimes I_n) \right] \text{unfold}(I) \right) \\
= \text{fold} \left( (F_t^H \otimes I_n) \exp \left[ \text{Diag} \left( A^{(i)} : i \in [l] \right) \right] (F_t \otimes I_n) \text{unfold}(I) \right) \\
= \text{fold} \left( (F_t^H \otimes I_n) \exp \left[ \text{Diag} \left( A^{(i)} : i \in [l] \right) \right] \frac{1}{\sqrt{l}} \text{Vec} \left( I_n : i \in [l] \right) \right) \\
= \text{fold} \left( \frac{1}{\sqrt{l}} (F_t^H \otimes I_n) \text{Diag} \left( \exp \left[ A^{(i)} \right] : i \in [l] \right) \text{Vec} \left( I_n : i \in [l] \right) \right) \\
= \text{fold} \left( \frac{1}{\sqrt{l}} (F_t^H \otimes I_n) \text{Vec} \left( \exp \left[ A^{(i)} \right] : i \in [l] \right) \right) \\
= L^{-1} \left( \text{fold} \left( \exp \left[ A^{(i)} \right] : i \in [l] \right) \right),
\]

where the third equality is due to the following property of the matrix exponential ( [12, Prop. 2.3, 6]): If \( X^T X = I \), then \( \exp \left[ XAX^T \right] = X \exp [A] X^T \), and the fifth equality comes from the following formula which follows immediately from definition: \( \exp \left[ \text{Diag} \left( D_i : i \in [l] \right) \right] = \text{Diag} \left( \exp \left[ D_i \right] : i \in [l] \right) \), and the last equality comes from (2.5) and (2.11), and the fact that \( \exp \left[ A^{(i)} \right] = \left( \exp [A] \right)^{(i)} \) gives the penultimate equation. \( \square \)

A.13 Proof of Proposition 2.8

Proof. Since the t-exponential mapping

\[
\exp [A] = L^{-1} \left( \text{fold} \left( \exp \left[ L(A)^{(i)} \right] : i \in [l] \right) \right)
\]

is the composite of the matrix exponential mapping and linear mappings and the matrix exponential is smooth ( [12, Prop. 2.16]), we conclude that the t-exponential mapping is smooth. \( \square \)

A.14 The proof of Proposition 2.9

Proof. Using the corresponding property of the matrix exponential [12, Prop. 2.4], we obtain

\[
\frac{d}{dt} \exp [tA] = \frac{d}{dt} L^{-1} \left( \text{fold} \left( \exp \left[ tA^{(i)} \right] : i \in [l] \right) \right) \\
= \frac{d}{dt} L^{-1} \left( \text{fold} \left( \frac{d}{dt} \exp \left[ tA^{(i)} \right] : i \in [l] \right) \right) \\
= \frac{d}{dt} L^{-1} \left( \text{fold} \left( \exp \left[ tA^{(1)} \right] A^{(i)} : i \in [l] \right) \right) \\
= L^{-1} \left( L(\exp [tA]) \Delta L(A) \right) = \exp [tA] * \mathcal{A},
\]

where the first equality comes from (2.11), while (A.51) gives the last two equality. Similarly we can show that \( \frac{d}{dt} \exp [tA] = \mathcal{A} * \exp [tA] \). \( \square \)
A.15 Proof of Proposition 2.10

Proof. Applying the corresponding property in the matrix case [12, Prop. 2.3, 6] and (A.51), it follows that

\[
\exp \left[ \mathcal{X} \ast \mathcal{A} \ast \mathcal{X}^\top \right] = L^{-1} \left( \text{fold} \left( \exp \left[ (L(\mathcal{X} \ast \mathcal{A} \ast \mathcal{X}^\top))^\ast : i \in [l] \right] \right) \right) \\
= L^{-1} \left( \text{fold} \left( \exp \left[ (L(\mathcal{X}^\ast)\Delta L(\mathcal{A})\Delta L(\mathcal{X}^\top))^\ast : i \in [l] \right] \right) \right) \\
= L^{-1} \left( \text{fold} \left( \exp \left[ \hat{\mathcal{X}}(i)\hat{\mathcal{A}}(i)\hat{\mathcal{X}}(i)^\ast : i \in [l] \right] \right) \right) \\
= L^{-1} \left( \text{fold} \left( \hat{\mathcal{X}}(i)\exp \left[ \hat{\mathcal{A}}(i)\hat{\mathcal{X}}(i)^\ast : i \in [l] \right] \right) \right) \\
= L^{-1} \left( L(\mathcal{X}^\ast)\Delta L(\exp [\mathcal{A}]\Delta L(\mathcal{X}^\ast)) \right) = \mathcal{X} \ast \exp [\mathcal{A}] \ast \mathcal{X}^\top,
\]

where the first equality comes from (2.11).

A.16 Proof of Proposition 2.11

Proof. Denotes \( \mathcal{A} = \text{Diag} (\mathcal{D}_j : j \in [p]) \) and \( \mathcal{B} = \text{Diag} (\exp [\mathcal{D}_j] : j \in [p]) \). Applying (2.11), we get

\[
\exp [\mathcal{A}] = L^{-1} \left( \text{fold} \left( \exp \left[ \hat{\mathcal{A}}(i) : i \in [l] \right] \right) \right) \\
= L^{-1} \left( \text{fold} \left( \exp \left[ \text{Diag} \left( \hat{\mathcal{D}}_j(i) : j \in [p] \right) : i \in [l] \right] \right) \right) \\
= L^{-1} \left( \text{fold} \left( \text{Diag} \left( \exp \left[ \hat{\mathcal{D}}_j(i) : j \in [p] \right] : i \in [l] \right) \right) \right) \\
= L^{-1} \left( \text{fold} \left( \hat{\mathcal{B}}(i) : i \in [l] \right) \right) = \mathcal{B},
\]

where the third equality is due to the property of the matrix exponential [46]: \( \exp [\text{Diag} (\mathcal{C}_i : i \in [l])] = \text{Diag} (\exp [\mathcal{C}_i] : i \in [l]) \).

A.17 Proof of Proposition 2.12

Proof.

\[
(\exp [\mathcal{A}])^\top = L^{-1} \left( \text{fold} \left( \exp \left[ \hat{\mathcal{A}}(i)^\top : i \in [l] \right] \right) \right) \\
= L^{-1} \left( \text{fold} \left( \exp \left[ \hat{\mathcal{A}}(i)^\ast : i \in [l] \right] \right) \right) = \exp [\mathcal{A}^\top],
\]

where Proposition 2.2 gives the first equality, while the second equality comes from the corresponding property in the matrix case [12, Prop. 2.3, 2].

A.18 Proof of Proposition 2.13

Proof. Using (A.51), we have

\[
\exp [\mathcal{A}] \ast \exp [\mathcal{B}] = L^{-1} (L(\exp [\mathcal{A}]\Delta L(\exp [\mathcal{B}])) \\
= L^{-1} \left( \text{fold} \left( \exp \left[ \hat{\mathcal{A}}(i) \exp \left[ \hat{\mathcal{B}}(i) : i \in [l] \right] \right) \right) \right) \\
= L^{-1} \left( \text{fold} \left( \exp \left[ \hat{\mathcal{A}}(i) + \hat{\mathcal{B}}(i) : i \in [l] \right] \right) \right) = \exp [\mathcal{A} + \mathcal{B}],
\]

where the second equality comes from the property in the matrix exponential [12, Prop. 2.3, 5].
Lemma A.1. Let $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times k}$. Then

$$(I_{kl} \otimes [A^{(1)}, \ldots, A^{(l)}]) \text{vec}(\text{bcirc}(B)) = ([I_k]_{l \times l} \otimes \text{bcirc}(A)) \text{vec}(B).$$

Proof. By definition, the left hand side part is

$$LHS = \begin{bmatrix} [A^{(1)}, \ldots, A^{(l)}] & \cdots & [A^{(1)}, \ldots, A^{(l)}] \\ \vdots & \ddots & \vdots \\ [A^{(1)}, \ldots, A^{(l)}] & \cdots & [A^{(1)}, \ldots, A^{(l)}] \end{bmatrix}_{jk}$$

where $B_{ij}^{(l)}$ is the $j$th column of $B^{(l)}$, $i \in [l]$ and the right hand side part is

$$RHS = \begin{bmatrix} [A^{(1)}] & \cdots & [A^{(1)}] \\ \vdots & \ddots & \vdots \\ [A^{(1)}] & \cdots & [A^{(1)}] \end{bmatrix}_{jk} \begin{bmatrix} [B^{(1)}] & \cdots & [B^{(1)}] \\ \vdots & \ddots & \vdots \\ [B^{(1)}] & \cdots & [B^{(1)}] \end{bmatrix}_{jk}.$$

We observe that the $(q, 1)$--th block of partitioned matrix on LHS is

$$\sum_{i=1}^{l} A^{(i)} B_{ij}^{(h_{i})}, \quad q = (p - 1)k + j \in [kl], \; j \in [k], \; p \in [l], \quad (A.53)$$

where

$$h_{i} = \begin{cases} l + p + 1 - i, & i > p \\ p + 1 - i, & i \leq p. \end{cases}$$

While the $(q, 1)$--th block of partitioned matrix on RHS is $\sum_{i=1}^{l} A^{(h_{i})} B_{ij}^{(l)}$, which is equal to (A.53).

Lemma A.2. [47] Let $C \in \mathbb{R}^{m \times n}$, $X \in \mathbb{R}^{n \times p}$, $B \in \mathbb{R}^{k \times p}$. Then

$$Y = CXB^T \iff \text{vec}(Y) = (B \otimes C) \text{vec}(X).$$

Lemma A.3. Let $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times k}$, $C \in \mathbb{R}^{m \times k}$. Then

$$C = A \ast B \iff \text{vec}(C) = \text{vec}(A) \otimes I_{m} \text{vec}(B) = ([I_k]_{l \times l} \otimes \text{bcirc}(A)) \text{vec}(B).$$
Proof. We observe that \(\text{vec}(C) = \text{vec}([C^{(1)}, \ldots, C^{(l)}])\). Since \(\text{unfold}(C) = \text{bcirc}(A)\ \text{unfold}(B)\), i.e., \([C^{(1)}, \ldots, C^{(l)}] = [A^{(1)}, \ldots, A^{(l)}] \text{bcirc}(B)\), we have

\[
\begin{align*}
\text{vec}(C) &= \text{vec}([C^{(1)}, \ldots, C^{(l)}]) \\
&= \text{vec}([A^{(1)}, \ldots, A^{(l)}] \text{bcirc}(B)) \\
&= (\text{bcirc}(B)^\top \otimes I_m) \text{vec}([A^{(1)}, \ldots, A^{(l)}]) \\
&= (\text{bcirc}(B)^\top \otimes I_m) \text{vec}(A),
\end{align*}
\]

where the third equation comes from Lemma A.2. Similarly, by lemma A.1, there holds

\[
\begin{align*}
\text{vec}(C) &= \text{vec}([C^{(1)}, \ldots, C^{(l)}]) \\
&= \text{vec}([A^{(1)}, \ldots, A^{(l)}] \text{bcirc}(B)) \\
&= (I_{kl} \otimes [A^{(1)}, \ldots, A^{(l)}]) \text{vec}(\text{bcirc}(B)) \\
&= ([I_k]_{kl} \otimes \text{bcirc}(A)) \text{vec}(B),
\end{align*}
\]

where the third equation follows from Lemma A.2. \(\blacksquare\)

Proof. Applying lemma A.3, the tensor Sylvester equation (4.34) can be rewritten in the form

\[
\text{vec}(C) = \left(\text{bcirc}(B)^\top \otimes I_k + [I_k]_{kl} \otimes \text{bcirc}(A)\right) \text{vec}(X).
\] (A.54)

\(\blacksquare\)

A.20 The Euclidean gradient \(\text{grad} f(X)\) and the Euclidean directional derivative \(Df(X)[H]\) in subsection 4.2

Similar to [53, Def. 4], for third-order tensor \(X \in \mathbb{R}^{n \times p \times l}\), we can also introduce the definition of the Euclidean gradient \(\text{grad} f(X)\) and the Euclidean Hessian \(\text{Hess} f(X)\) from the Fréchet differentiable.

**Definition A.1.** Let \(f : \mathcal{U} \subseteq \mathbb{R}^{n \times p \times l} \to \mathbb{R}\) be a continuous map. Then, we say \(f\) is \(t\)-differentiable at \(X \in \mathcal{U}\) if and only if there exists a third-order tensor \(\text{grad} f(X) \in \mathbb{R}^{n \times p \times l}\) such that

\[
\lim_{\|H\|_F \to 0} \frac{\|f(X + H) - f(X) - \langle \text{grad} f(X), H \rangle \|_F}{\|H\|_F} = 0,
\]

where \(\text{grad} f(X)\) is called the gradient of \(f\) at \(X\) and \(Df(X)[H] = \langle \text{grad} f(X), H \rangle\) called the directional derivative of \(f\) at \(X\) along \(H\). And we say \(f\) is twice \(t\)-differentiable at \(X \in \mathcal{U}\) if and only if \(f\) is continuously \(t\)-differentiable and there exists a bounded linear operator \(\text{Hess} f(X) : \mathbb{R}^{n \times p \times l} \to \mathbb{R}^{n \times p \times l}\) such that

\[
\lim_{\|H\|_F \to 0} \frac{\|\text{grad} f(X + H) - \text{grad} f(X) - \text{Hess} f(X)[H]\|_F}{\|H\|_F} = 0.
\]

Furthermore, we say \(f\) is \(t\)-differentiable (twice \(t\)-differentiable) on \(\mathcal{U}\) if and only if \(f\) is \(t\)-differentiable (twice \(t\)-differentiable) at every \(X \in \mathcal{U}\).

**Theorem A.1.** Let \(f\) be a continuous map from \(\mathcal{U} \subseteq \mathbb{R}^{n \times p \times l}\) to \(\mathbb{R}\). Then \(f\) is \(t\)-differentiable on \(\mathcal{U}\) if and only if \(\frac{\partial f(X)}{\partial \text{vec}(X)}\) exists for every \(X \in \mathcal{U}\), where \(\frac{\partial f(X)}{\partial \text{vec}(X)}\) is a vector in \(\mathbb{R}^{npl}\) with \(\left(\frac{\partial f(X)}{\partial \text{vec}(X)}\right)_i = \frac{\partial f(X)}{\partial (\text{vec}(X))_i}\) for any \(i \in [npl]\). Especially, for any \(X \in \mathcal{U}\),

\[
\text{grad} f(X) = \text{vec}^{-1}\left(\frac{\partial f(X)}{\partial \text{vec}(X)}\right),
\] (A.55)

where \(\text{vec}(X)\) denotes the vectorized tensor of \(X\) and \(\text{vec}^{-1}(v) = A\) represents the operator that converts a vector \(v\) back to a tensor \(A\), which can all be implemented with functions \texttt{reshape, permute} and \texttt{ipermute} of Matlab (cf. [26]).

36
Proof. The proof is similar to that of [53, Thm. 1] and is omitted.