Riemannian Tensor Completion with Side Information

Tengfei Zhou  Hui Qian*  Zebang Shen  Congfu Xu
Computer Science Institute, Zhejiang University
{zhoutengfei_zju, qianhui, shenziebang, xucungfu@zju.edu.cn}

February 24, 2017

Abstract

By restricting the iterate on a nonlinear manifold, the recently proposed Riemannian optimization methods prove to be both efficient and effective in low rank tensor completion problems. However, existing methods fail to exploit the easily accessible side information, due to their format mismatch. Consequently, there is still room for improvement in such methods. To fill the gap, in this paper, a novel Riemannian model is proposed to organically integrate the original model and the side information by overcoming their inconsistency. For this particular model, an efficient Riemannian conjugate gradient descent solver is devised based on a new metric that captures the curvature of the objective. Numerical experiments suggest that our solver is more accurate than the state-of-the-art without compromising the efficiency.

I. Introduction

Low Rank Tensor Completion (LRTC) problem, which aims to recover a tensor from its linear measurements, arises naturally in many artificial intelligence applications. In hyperspectral image inpainting, LRTC is applied to interpolate the unknown pixels based on the partial observation Xu et al. (2015). In recommendation tasks, LRTC helps users find interesting items under specific contexts such as locations or time Liu et al. (2015). In computational phenotyping, one adopts LRTC to discovery phenotypes in heterogeneous electronic health records Wang et al. (2015).

Euclidean Models: LRTC can be formulated by a variety of optimization models over the Euclidean space. Amongst them, convex models that encapsulate LRTC as a regression problem penalized by a tensor nuclear norm are the most popular and well-understood Romera-Paredes and Pontil (2013), Zhang et al. (2014). Though most of them have sound theoretical guarantees Chen et al. (2013), Yuan and Zhang (2015), Zhang and Aeron (2016), in general, their solvers are ill-suited for large tensors because these procedures usually involve Singular Value Decomposition (SVD) of huge matrices per iteration Liu et al. (2013). Another class of Euclidean models is formulated as the decomposition problem that factorizes a low rank tensor into small factors Filipović and Jukić (2015), Jain and Oh (2014), Xu et al. (2015). Many solvers for such decomposition based model have been proposed to recover large tensors, and low per-iteration computational cost is illustrated Beutel et al. (2014), Liu et al. (2014), Smith et al. (2016).

Riemannian Models: LRTC can also be modeled by nonconvex optimization constrained on Riemannian manifolds Kasai and Mishra (2016), Kressner et al. (2014), which is easily handled by many manifold based solvers Absil et al. (2009). Empirical comparison has shown that Riemannian solvers use significantly less CPU time to recover the underlying tensor in contrast to the Euclidean solvers Kasai and Mishra (2016). The main reason resides in that such solvers avoid SVD of...
huge matrices by explicitly exploiting the geometrical structure of LRTC, which makes them more suitable for massive problem.

Of all the Riemannian models, two search spaces, fix multi-linear rank manifold [Kressner et al. 2014] and Tucker manifold [Kasai and Mishra 2016], are usually employed. The former is a sub-manifold of Euclidean space, and the latter is a quotient manifold induced by the Tucker decomposition. Generally, quotient manifold based solvers have higher convergence rates because it is usually easier to design a pre-conditioner for them [Kasai and Mishra 2016].

**Side Information:** In the Euclidean models of LRTC, side information is helpful in improving the accuracy [Acar et al. 2011], Beutel et al. [2014], Narita et al. [2011]. One common form of the side information is the feature matrix, which measures the statistical properties of tensor modes [Kolda and Bader 2009]. For example, in Netflix tasks, feature matrix can be built from the demography of users [Bell and Koren 2007]. Another form is the similarity matrix, which quantifies the resemblance between two entities of a tensor mode. For instance, the social network generates the similarity matrix by utilizing the correspondence between users [Rai et al. 2015]. In practice, these two matrices can be transformed to each other, and we only consider the feature matrix case throughout this paper.

However, as far as we know, side information has not been incorporated in any Riemannian model. The first difficulty lies in the model design. Fusing the side information into the Riemannian model inevitably compromises the integrity of the low rank tensor due to the compactness of the manifold. The second difficulty results from the solver design. Incorporating the side information may aggravate the ill-conditioning of LRTC problem and degenerates the convergence significantly.

**Contributions:** To address these difficulties, a novel Riemannian LRTC method is proposed from the perspective of both model and solver designs. By exploring the relation between the subspace spanned by the tensor fibers and the column space of the feature matrix, we explicitly integrate the side information in a compact way. Meanwhile, a first order solver is devised under the manifold spanned by the tensor fibers and the column space of the feature matrix, we explicitly integrate the perspective of both model and solver designs. By exploring the relation between the subspace

II. Notations and Preliminaries

In this paper, we only focus on the 3rd order tensor, but generalizing our method to high order is straightforward. We use the notation $\mathbf{X} \in \mathbb{R}^{n \times m}$ to denote a matrix, and the notation $\mathbf{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ to denote a $d$-th order tensor. We also denote by $\mathbf{X}(i_1, \cdots, i_d)$ the element in position $(i_1, \cdots, i_d)$ of $\mathbf{X}$. For many cases, we use curly braces with indexes to simplify the notation. For example, $\{O_i\}_{i=1}^3$ is used to denote $O_1, O_2, O_3$, and $\{U_iO_i\}_{i=1}^3$ refer to $U_1O_1, U_2O_2, U_3O_3$.

**Mode-$k$ Fiber and matricization:** A fiber of a tensor is obtained by varying one index while fixing the others, i.e. $\mathbf{X}(i_1, \cdots, i_{k-1}, i_{k+1}, \cdots, i_d)$ is the mode-$k$ fiber of a $d$-th order tensor $\mathbf{X}$. Here we use the colon to denote $\{1, \cdots, n_k\}$. A mode-$k$ matricization $\mathbf{X}(k) \in \mathbb{R}^{n_k \times (n_1 \cdots n_{k-1}n_{k+1} \cdots n_d)}$ of a tensor $\mathbf{X}$ is obtained by arranging the mode-$n$ fibers of $\mathbf{X}$ so that each of them is a column of $\mathbf{X}(k)$ [Kolda and Bader 2009].

**Inner product and norm:** The inner product of two tensors with the same size is defined by $\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} \mathbf{X}(i_1, \cdots, i_d)\mathbf{Y}(i_1, \cdots, i_d)$. The Frobenius norm of a tensor $\mathbf{X}$ is defined by $\|\mathbf{X}\|_F = \sqrt{\langle \mathbf{X}, \mathbf{X} \rangle}$.

**Multi-linear rank and Tucker decomposition:** The multi-linear rank rank$^\text{vec}(\mathbf{X})$ of a tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is defined as a vector $(\text{rank}(\mathbf{X}(1)), \text{rank}(\mathbf{X}(2)), \text{rank}(\mathbf{X}(3)))$. If rank$^\text{vec}(\mathbf{X}) = (1, 1, 1)$ then $\mathbf{X}$ is a rank-$1$ tensor.
Specifically, the fix multi-rank manifold, which will be helpful in understanding the whole derivation. Tucker decomposition. In order to lay the ground for Tucker manifold, we first describe its coun-

The Tucker manifold that we used in our Riemannian model is a quotient manifold induced by the 

II.1 Search Space of Riemannian Models

The Tucker manifold is a quotient manifold of the total space (1). We use the abstract quotient 

For simplicity, we denote \[ [\mathcal{G}, \{U_i\}_{i=1}^3] \] by [\mathcal{X}], when \( \mathcal{X} = \mathcal{G} \times_{i=1}^3 U_i \). Usually, [\mathcal{X}] is called the Tucker representation of \( \mathcal{X} \), while \( \mathcal{X} \) is call the tensor representation of [\mathcal{X}]. We also use \( \overline{\mathcal{X}} \) to denote a specific decomposition of \( \mathcal{X} \), additionally \( \overline{\mathcal{X}} \in [\mathcal{X}] \).

II.2 Vanilla Riemannian Tensor Completion

The purest incarnation of Riemannian tensor completion model is the Riemannian model over the 

with \( \mathcal{P}_\Omega \) maps \( \mathcal{X} \) to the sparsified tensor \( \mathcal{P}_\Omega(\mathcal{X}) \), where \( \mathcal{P}_\Omega(\mathcal{X})(i_1, i_2, i_3) = \mathcal{X}(i_1, i_2, i_3) \) if \( (i_1, i_2, i_3) \in \Omega \), and \( \mathcal{P}_\Omega(\mathcal{X})(i_1, i_2, i_3) = 0 \) otherwise.

Another popular model, Tucker model, is based on the quotient manifold \( \mathcal{M}_r/\sim \), which can be expressed as:

\[
\min_{\mathcal{X}} \frac{1}{2} \| \mathcal{P}_\Omega(\mathcal{X} - \mathcal{R}) \|_F^2 \quad \text{s.t.} \quad \mathcal{X} \in \mathcal{F}_r, \tag{4}
\]
with $\rho$ defined in Prop. 1.

Note that since the dawn of Riemannian framework for LRTC, a quandary exists: on one hand, sparse measurement limits the capacity of the solution; on the other hand, rich side information cannot be incorporated into this framework. In many artificial intelligence applications, demands for high accuracy further exacerbates such dilemma.

III. Riemannian Model with Side Information

We focus on the case that the side information is encoded in feature matrices $\mathbf{P}_i \in \mathbb{R}^{n_i \times k_i}$. Suppose $\mathbf{R} \in \mathcal{F}_r$ has tucker factors $(\mathcal{G}, \{\mathbf{U}_i\}_{i=1}^3)$. Without loss of generality, we assume that $k_i \geq r_i$ and $\mathbf{P}_i$ has orthogonal columns.

In the ideal case, we assume that

$$\text{span}(\mathbf{U}_i) \subset \text{span}(\mathbf{P}_i).$$

Such relation means that the feature matrices contain all the information in the latent space of the underlying tensor. Equivalently, there exists a matrix $\mathbf{W}_i$ such that $\mathbf{U}_i = \mathbf{P}_i \mathbf{W}_i$. However, in practice, due to the existence of noise, one can only expect such relation to hold approximately, i.e. $\mathbf{U}_i \approx \mathbf{P}_i \mathbf{W}_i$. Incorporating such relation to a tensor completion model via penalization, we have the following formulation

$$\min_{\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3} L(\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3) + \sum_{i=1}^3 \alpha_i \|\mathbf{U}_i - \mathbf{P}_i \mathbf{W}_i\|_F^2,$$

s.t. $(\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3) \in \mathcal{M}_r$,

where $L(\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3) = \|\mathbf{P}_\Omega(\mathbf{G} \times_3 \mathbf{U}_i - \mathbf{R})\|_F^2/2$. Fixing $\mathbf{G}$ and $\mathbf{U}_i$, with respect to $\mathbf{W}_i$, (6) has a close form solution

$$\mathbf{W}_i = (\mathbf{P}_i^T \mathbf{P}_i)^{-1} \mathbf{P}_i^T \mathbf{U}_i.$$  

Since $\min_{x,y} l(x, y) = \min_x l(x, y(x))$ where $y(x) = \text{arg min}_y l(x, y)$, one can substitute (7) into the above problem and obtain the following equivalence

$$\min_{\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3} L(\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3) + \sum_{i=1}^3 \alpha_i \|\mathbf{U}_i - \mathbf{P}_i \mathbf{W}_i\|_F^2 \text{ trace}(\mathbf{U}_i^T (\mathbf{I} - \mathbf{P}_i \mathbf{P}_i^T) \mathbf{U}_i)$$

$$\triangleq f(\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3)$$

s.t. $(\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3) \in \mathcal{M}_r$.

Although the cost function is already smooth over the total space $\mathcal{M}_r$, due to its invariance over the equivalent class $[\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3]$, there can be infinite local optima, which is extremely undesirable. Indeed, if $(\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3)$ is a local optimal of the objective, then so is every point in the infinite set $[\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3]$. One way to reduce the number of local optima is to mathematically treated the entire set $[\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3]$ as a point. Consequently, we redefine the cost by $\tilde{f}(\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3) = f(\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3)$ and obtain the following Riemannian optimization problem over the quotient manifold $\mathcal{M}_r/\sim$:

$$\min_{[\mathbf{X}]} \tilde{f}(\mathbf{X}) \quad \text{s.t. } [\mathbf{X}] \in \mathcal{M}_r/\sim.$$  

Remark 1. In Riemannian optimization literature, problem (8) is called the lifted representation of problem (9) over the total space [Absil et al. 2009]. This model is closely related to the Laplace regularization model [Narita et al. 2011]. Concretely, they share the same form:

$$\min_{\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3} L(\mathbf{G}, \{\mathbf{U}_i\}_{i=1}^3) + \sum_{i=1}^3 C_i \text{ trace}(\mathbf{U}_i^T \mathbf{L}_i \mathbf{U}_i).$$  

4
Figure 1: Optimization Framework for Quotient Manifold: most Riemannian solvers are based on the iteration formula: \( [x^+] \leftarrow R_{[x]}(t\eta_{[x]}) \), where \( t > 0 \) is the stepsize, \( \eta_{[x]} \) is the search direction picked from current tangent space \( T_{[x]}M/\sim \), and \( R_{[x]}(\cdot) \) is the retraction, i.e. a map from current tangent space to \( M/\sim \). Due to the abstractness of quotient manifold, such iteration is often lifted to (represented in) the total space as \( x = R_x(t\eta_x) \) where \( x \in [x] \), \( \eta_x \) is the horizontal lift of \( \eta_{[x]} \), and \( R_x(\cdot) \) is the lifted retraction. Such representation is possible only if \( M/\sim \) has the structure of Riemannian quotient, that is the total space is endowed with an invariant Riemannian metric.

The difference lies in that \( L_i \) is a projection matrix in our case, while, in the Laplace regularization model, \( L_i \) is a Laplacian matrix.

**Remark 2.** Since each \([X] \in M_r/\sim \) has a unique tensor representation in \( X \in F_r \), we show that the abstract model (9) can be represented as a concrete model over the manifold \( F_r \). Specifically, the following Proposition interprets the proposed model as an optimization problem with a regularizer that encourages the mode-\( i \) space of the estimated tensor close to \( \text{span}(P_i) \).

**Proposition 2.** If \([X] \) is a critical point of problem (9) then its tensor representation \( X \) is a critical point of the following problem.

\[
\min_{X \in F_r} \frac{1}{2} \| P_\Omega (X - \mathcal{R}) \|_F^2 + \sum_{i=1}^{3} \frac{\alpha_i |\Omega|}{2} \text{dist}^2(\text{span}(X(i)), \text{span}(P_i))
\]

where \( \text{dist}(\cdot, \cdot) \) is the Chodal distance [Ye and Lim (2014)] between two subspaces. And vice versa.

**IV. Riemannian Conjugate Gradient Descent**

We depict the optimization framework for quotient manifolds in Fig. 1. Under this framework, we solve the proposed problem (9) by Riemannian Conjugate Gradient descent (CG). With the details specified later, we list our CG solver for problem (9) in Alg. 1, where the CG direction is composed in the Polak-Ribiere+ manner with the momentum weight \( \beta^{(k)} \) computed by Flecher-Reeves formula [Absil et al. (2009)], and \( T_k(\cdot) \) is the projector of horizontal space \( \mathcal{H}_{X(k)} \). To represent Alg. 1 in concrete tensor formulations, four items must be specified: the Riemannian metric \( \langle \cdot, \cdot \rangle_{X} \), the Riemannian gradient \( \text{grad} f(X) \), the retraction \( R_X(\cdot) \), and the projector onto horizontal space \( T_{X} \).
Algorithm 1 CGSI: a Riemannian CG method

Input: Initializer $\mathcal{X}^{(0)} = \{(\mathcal{G}^{(0)}, \{U_i^{(0)}\})_{i=1}^{3}\}$ and tolerance $\epsilon$

1: $k = 0$
2: $\eta^{(-1)} = (0, \{0\})_{i=1}^{3}$
3: repeat
4: compute current Riemannian gradient $\xi^{(k)} = \text{grad} f(\mathcal{X}^{(k)})$
5: compose CG direction $\eta^{(k)} = -\xi^{(k)} + \beta^{(k)}T_k(\eta^{(k-1)})$
6: choose a step size $t_k > 0$
7: update by retraction $\mathcal{X}^{(k+1)} = R_{\mathcal{X}^{(k)}}(t_k\eta^{(k)})$
8: $k = k + 1$
9: until $(\xi^{(k-1)}, \xi^{(k-1)})_{\mathcal{X}^{(k-1)}} \leq \epsilon$
10: return $\mathcal{X}^{(k)}$

IV.1 Metric Tuning

Riemannian metric $\langle \cdot, \cdot \rangle_{\mathcal{X}}$ of $\mathcal{M}_r$ is an inner product defined over each tangent space $T_{\mathcal{X}}\mathcal{M}_r$. A high-quality Riemannian solver for a quotient manifold should be equipped with a well-tuned metric, because (1) the metric determines the differential structure of the quotient manifold, and more importantly (2) it implicitly endows the solver with a preconditioner, which heavily affects the convergent rate [Mishra (2014), Mishra and Sepulchre (2014)].

From the perspective of preconditioning, it seems that the best candidate is the Newton metric $\langle \eta, \xi \rangle_{\mathcal{X}} = D^2 f(\mathcal{X})[\eta]\xi$, where $D^2 f(\mathcal{X})$ is the second order differential of the cost function. However, under such metric, computing the search direction involves solving a large system of linear equations, which precludes the Newton metric from the application to huge datasets. Therefore, we propose to use the following alternative:

$$
\langle \eta_{\mathcal{X}}, \xi_{\mathcal{X}} \rangle_{\mathcal{X}} = D^2 g(\mathcal{X})[\eta_{\mathcal{X}}, \xi_{\mathcal{X}}] \\
= \sum_{i=1}^{3} \langle \eta_i, \xi_i, \mathcal{G}(i) \mathcal{G}^T(i) \rangle + \langle \eta_{\mathcal{X}}, \xi_{\mathcal{X}} \rangle
$$

(11)
in which $g(\mathcal{X})$ is a scaled approximation to the original cost function, that is $g(\mathcal{X}) \triangleq \frac{1}{2} \| \mathcal{G} \times U_i - \mathcal{X} \|^2 + \sum_{i=1}^{3} \frac{1}{2} \alpha_i \text{trace}(U_i^T(I - P_iP_i^T)U_i)$ with $N = n_1n_2n_3$.

Our metric is more scalable than Newton metric. The following Proposition indicates that the scale gradient induced by this metric can be computed with only $O(\sum_{i=1}^{3} n_i r_i + r_i^3)$ additional operations, which is much less than the operations required by Newton metric.

**Proposition 3.** Suppose that the cost function $f(\cdot)$ has Euclidean gradient $\nabla f(\mathcal{X}) = (\nabla g f, \{\nabla u_i f\})_{i=1}^{3}$. Then its scaled gradient $\nabla f(\mathcal{X})$ under the metric (11) can be computed by:

$$
\nabla g f(\mathcal{X}) = \nabla g f(\mathcal{X})
$$

$$
\nabla u_i f(\mathcal{X}) = E_i G_i^{-1} + F_i (G_i + N_\alpha I)^{-1}
$$

where $E_i = P_i P_i^T \nabla u_i f$, $F_i = \nabla u_i f - E_i$, and $G_i = \mathcal{G}(i) \mathcal{G}^T(i)$.

Moreover, the proposed metric contains the curvature information of the cost. It is easy to validate that $D^2 f(\mathcal{X})/|\Omega| \approx D^2 g(\mathcal{X})/N$, since $f(\mathcal{X})/|\Omega| \approx g(\mathcal{X})/N$ if the observed entries are sampled uniformly at random.

The final proposition suggests that the proposed metric makes the representation of solvers in the total space possible.
IV.2 Other Optimization Related Items

**Projectors:** To derive the optimization related items, two orthogonal projectors, \( \Psi_{\mathcal{M}_r}(\cdot) \) and \( \Psi_{\mathcal{T}}(\cdot) \), are required. The former projects a vector onto the tangent space \( T_{\mathcal{M}_r}\mathcal{M}_r \), and the latter is a projector from the tangent space onto the horizontal space \( H_{\mathcal{T}} \). The orthogonality of both projectors is measured by the metric (11). Mathematical derivation of these projectors are given in Sec. VII.2.2 and Sec. VII.2.1.

**Riemannian Gradient:** According to Absil et al. (2009), the Riemannian gradient can be computed by projecting the scaled gradient onto tangent space, specifically

\[
\text{grad } f(\mathcal{X}) = \Psi_{\mathcal{M}_r}(\nabla f(\mathcal{X})).
\]

**Retraction:** We use the retraction defined by

\[
R_{\mathcal{T}}(\eta_{\mathcal{M}_r}) = (\mathcal{G} + \eta_{\mathcal{G}}, \{u_{\mathcal{G}}(U_i + \eta_i)\}_{i=1}^3).
\]

where \( u_{\mathcal{G}}(\cdot) \) extracts the orthogonal component from a matrix. Such retraction is proposed by Kasai and Mishra (2016). We give rigorous analysis to prove that the above retraction is compatible with the proposed metric in Sec. VII.2.3.

V. Experiments

We validate the effectiveness of the proposed solver CGSI by comparing it with the state-of-the-art. The baseline can be partitioned into three classes. The first class contains Riemannian solvers including GeomCG Kressner et al. (2014), FTC Kasai and Mishra (2016), and gHOI Liu et al.
The second class consists of Euclidean solvers that take no account of the side information, including AltMin [Romera-Paredes et al. (2013)] and HalRTC [Liu et al. (2013)]. The third class comprises of two methods that incorporate side information, including RUBIK [Wang et al. (2015)] and TFAI [Narita et al. (2011)]. All the experiments are performed in Matlab on the same machine with 3.0 GHz Intel E5-2690 CPU and 128GB RAM.

All solvers are based on the Tucker decomposition, except that RUBIK is based on the CP decomposition. For fairness, the CP rank of RUBIK is set to $\left\lceil \frac{\sum_{i=1}^{3} n_i r_i + r_1 r_2 r_3}{\sum_{i=1}^{3} n_i} \right\rceil$.

### V.1 Hyperspectral Image Inpainting

A hyperspectral image is a tensor whose the slices are photographs of the same scene under different wavelets. We adopt the dataset provided in Foster et al. (2006) which contains images about eight different rural scenes taken under 33 various wavelets. To make all methods in our baseline applicable to the completion problem, we resize each hyperspectral images to a small dimension such that $n_1 = 306$, $n_2 = 402$, and $n_3 = 33$. Empirically, we treat these graphs as tensors of rank $r = (30, 30, 6)$. The observed pixels, or the training set, are sampled from the tensors uniformly at random. And the sample size is set to $|\Omega| = OS \times p$ in which $OS$ is so-called Over-Sampling ratio and $p = \sum_{i=1}^{3} (n_i r_i - r_i^2) + r_1 r_2 r_3$ is the number of free parameters in a size $n$ tensor with rank $r$. In addition to the observed entries, the mode-1 feature matrix is constructed by extracting the top-$(r_1 + 10)$ singular vectors from a matrix of size $n_1 \times 10 r_1$ whose columns are sampled from the mode-1 fibers of the hyperspectral graphs. The recovery accuracy is measured by Normalized Root Mean Square Error (NRMSE) [Kressner et al. (2014)]. All the compared methods are terminated when the training NRMSE is less than 0.003 or iterate more than 300 epochs. We report the NRMSE and CPU time of the compared methods in Tab. 2. From the table, we can see that the proposed method has much higher accuracy than the other solvers in our baseline. The empirical results also indicate that our method has nearly the same running time with FTC, the fastest tensor completion method. The visual results of the 27th slices of recovered hyperspectral images of scene 7 are illustrated in Fig. 2.

![Visual results of the recovered 27th frame of scene7 when OS is set to 3.](image)

### V.2 Recommender System

In recommendation tasks, two datasets are considered: MovieLens 10M (ML10M) and MovieLens 20M (ML20M). Both datasets contain the rating history of users for items at specific moments. For both datasets, we partition the samples into 731 slices in terms of time stamp. Those slices have the identical time intervals. Accordingly, the completion tasks for the two datasets are of sizes $71567 \times 10681 \times 731$ and $138493 \times 26744 \times 731$ respectively. In addition to the rating history, both datasets contain two extra files: one describes the genres of each movie, and the other contains tags of each movie. We construct a corpus that contains the text description of all movies from the genres descriptions and all the tags. The feature matrix is extracted from the above corpus by the latent semantic analysis (LSA) method. The processing is efficient since LSA is implemented via randomized SVD.
Table 2: Performance of the compared methods on hyperspectral images.

| Data Set | NRMSE (%) | Time (s) | NRMSE (%) | Time (s) | NRMSE (%) | Time (s) | NRMSE (%) | Time (s) | NRMSE (%) | Time (s) | NRMSE (%) | Time (s) |
|----------|------------|----------|------------|----------|------------|----------|------------|----------|------------|----------|------------|----------|
| Scene1   | 0.164      | 153      | 0.091      | 12       | 0.133      | 61       | 0.135      | 66       | 0.096      | 197      | 0.101      | 194      |
| Scene2   | 0.155      | 165      | 0.067      | 70       | 0.077      | 93       | 0.103      | 109      | 0.078      | 177      | 0.085      | 194      |
| Scene3   | 0.156      | 208      | 0.060      | 100      | 0.056      | 124      | 0.092      | 152      | 0.077      | 177      | 0.085      | 193      |
| Scene4   | 0.156      | 550      | 0.046      | 126      | 0.044      | 151      | 0.078      | 195      | 0.077      | 178      | 0.085      | 192      |

... (Continued)
Various empirical studies are conducted to validate the performance of the proposed method. In the first scenario, we record the CPU time and the Root Mean Square Error (RMSE) outputted by the compared algorithms under different choices of multi-linear rank. In this scenario, for both datasets, 80% samples are chosen as training set, and the rest are left for testing. The results are listed in Tab. 3 which suggests that the proposed method outperforms all other solvers in terms of accuracy. For ML10M, our method uses significantly less CPU time than its competitors. In Fig. 4, we report another scenario, in which the percentage of training samples are varied from 10% to 70% and the rank parameter is fixed to (10, 10, 10). Experimental results in this figure indicate that our method has the lowest RMSE.

To show the impact of parameter $\alpha$ on the performance of our method, we depict the relation between RMSE and $\alpha$ in Fig. 3 where the rank parameter is set to (10, 10, 10), and the partitioning scheme for training and testing samples is the same as the first scenario. From this Figure we can see that our method has higher accuracy than the vanilla Riemannian model’s solver FTC for a wide range of parameter choices.

### Table 3: Performance of the compared methods on Recommendation Tasks.

| Dataset | Rank | AltMin | RMSE  | Time(s) | RMSE  | Time(s) | RMSE  | Time(s) | RMSE  | Time(s) | RMSE  | Time(s) | RMSE  | Time(s) |
|---------|------|--------|-------|---------|-------|---------|-------|---------|-------|---------|-------|---------|-------|---------|
| ML10M   | (4,4) | 0.982  | 924   | 0.924   | 236   | 0.836   | 307   | 1.016   | 363   | 0.911   | 240   | 0.823   | 198   |
|         | (6,6) | 0.898  | 1830  | 0.814   | 535   | 0.826   | 675   | 1.062   | 802   | 0.9948  | 342   | 0.814   | 454   |
|         | (8,8) | 1.91   | 3123  | 0.822   | 928   | 0.833   | 1135  | 1.052   | 1115  | 0.993   | 1617  | 0.810   | 754   |
|         | (10,10,10) | 1.147 | 4963  | 0.842   | 1631  | 0.843   | 2220  | 1.094   | 2788  | 0.992   | 2522  | 0.807   | 1067  |
| ML20M   | (4,4) | 0.905  | 600   | 0.922   | 666   | 0.829   | 601   | 1.050   | 918   | 1.092   | 737   | 0.818   | 303   |
|         | (6,6) | 1.080  | 341   | 0.828   | 992   | 0.822   | 1309  | 1.057   | 1869  | 1.008   | 1644  | 0.805   | 1107  |
|         | (8,8) | 1.092  | 5960  | 0.812   | 1725  | 0.828   | 2271  | 1.045   | 3369  | 1.004   | 3144  | 0.804   | 1739  |
|         | (10,10,10) | 1.092 | 9418  | 0.818   | 3161  | 0.834   | 4306  | 1.054   | 5795  | 1.025   | 5394  | 0.799   | 2813  |
VI. Conclusion

In this paper, we exploit the side information to improve the accuracy of Riemannian tensor completion. A novel Riemannian model is proposed. To solve the model efficiently, we design a new Riemannian metric. Such metric will induce an adaptive preconditioner for the solvers of the proposed model. Then, we devise a Riemannian conjugate gradient descent method using the adaptive preconditioner. Empirical results show that our solver outperforms state-of-the-arts.

VII. Appendix

VII.1 Proof of Propositions

Before delve into the proofs of the propositions, we construct the submersion between the total space $M_r$ and fix multilinear rank manifold $F_r$ in the following Lemma.

Lemma 5. Let $\pi : M_r \rightarrow F_r$ be a mapping defined by $\pi(\mathcal{G}, U_1, U_2, U_3) = \mathcal{G} \times^3_{i=1} U_i$. Then it is a submersion from $M_r$ to $F_r$.

Proof. To begin with, we define a function $\pi : M_r \rightarrow F_r$ as follows $\pi(\mathcal{G}, U_1, U_2, U_3) = \mathcal{G} \times^3_{i=1} U_i$.

Note that $\pi()$ is a smooth function over $M_r$, and for all $\mathcal{X} = (\mathcal{G}, U_1, U_2, U_3) \in M_r$, and for all the tangent vectors $\eta_{\mathcal{X}} = (\eta_\mathcal{G}, \eta_1, \eta_2, \eta_3) \in T_{\mathcal{X}}M_r$, the first order derivative of $\pi()$ can be computed as follows:

$$D\pi(\mathcal{X})[\eta_{\mathcal{X}}] = \eta_\mathcal{G} \times^3_{i=1} U_i + \mathcal{G} \times \eta_1 \times^2 U_2 \times^3 U_3 + \mathcal{G} \times \eta_2 \times^3 U_3 + \mathcal{G} \times \eta_3 \times^3 U_3 \times^2 U_2 \times^3 (14)$$

Note that $\eta_\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ and $\eta_i \in T_{U_i} \text{St}(r_i, n_i)$ which means they can be expressed as $\eta_i = U_i \Omega_i + U_i,\perp K_i$ where $\Omega_i \in \mathbb{R}^{r_i \times r_i}$ is a skew matrix, $K_i \in \mathbb{R}^{(n_i-r_i) \times r_i}$, and $U_i,\perp \in \mathbb{R}^{n_i \times (n_i-r_i)}$ is the orthogonal basis, the spanned subspace of which is the orthogonal complement of $\text{span}(U_i)$. Substitute these expressions to equation (14), we have:

$$D\pi(\mathcal{X})[\eta_{\mathcal{X}}] = (\eta_\mathcal{G} + \sum_{i=1}^3 \mathcal{G} \times \Omega_i) \times^3_{i=1} U_i + \sum_{i=1}^3 \mathcal{G} \times U_i \times\perp K_i \times^3_{j \neq i, \leq 3} U_j. \quad (15)$$
Therefore, the range of the map \(D\pi(\mathcal{X})[\cdot]\) over the tangent space \(T\mathcal{X}_r\)

\[
\text{range}(D\pi(\mathcal{X})) = \left\{ \mathcal{H} \times_{i=1}^3 U_i + \sum_{i=1}^3 \mathcal{G} \times_i U_{i,\perp} K_i \times_{j \neq i, 1 \leq j \leq 3} U_j | \mathcal{H} \in \mathbb{R}^{r_1 \times r_2 \times r_3}, K_i \in \mathbb{R}^{(n_i - r_i) \times r_i} \right\}
\]

(16)

Note that the tangent space of fix multilinear rank manifold \(\mathcal{F}_r\) at the point \(\mathcal{X} = \pi(\mathcal{G}, U_1, U_2, U_3)\) is

\[
T\mathcal{X}_r = \left\{ \mathcal{H} \times_{i=1}^3 U_i + \sum_{i=1}^3 \mathcal{G} \times_i U_{i,\perp} V_i \times_{j \neq i, 1 \leq j \leq 3} U_j | \mathcal{H} \in \mathbb{R}^{r_1 \times r_2 \times r_3}, V_i \in \mathbb{R}^{n_i \times r_i} \text{ and } V_i U_i = 0 \right\}.
\]

(17)

Using the fact any matrix \(V_i \in \mathbb{R}^{n_i \times r_i}\) and \(V_i^\top \times U_i = 0\), there exist \(K_i \in \mathbb{R}^{(n_i - r_i) \times r_i}\) such that \(V_i = U_{i,\perp} K_i\), we can infer that

\[
\text{range}(D\pi(\mathcal{X})) = T\mathcal{X}_r.
\]

(18)

As a result, \(\pi(\cdot)\) is a submersion from \(\mathcal{M}_r\) to \(\mathcal{F}_r\).

\[\square\]

VII.1.1 Horizontal Space

Proposition 6. Let \(\mathcal{X} = (\mathcal{G}, U_1, U_2, U_3) \in [\mathcal{X}]\), the horizontal space of \(\mathcal{M}_r\) at point \(\mathcal{X}\) is

\[
\left\{ \eta_{\mathcal{X}} \in T\mathcal{X}_r | \mathcal{V}_i^\top \eta_{\mathcal{G}} U_i + W_i^\top \eta_{\mathcal{G}_\alpha}, \text{is symmetric} | 1 \leq i \leq 3 \right\}
\]

where \(V_i = P_i P_i^\top U_i\), \(W_i = U_i - P_i P_i^\top U_i\), \(G_i = G_{(i)} G_{(i)}^\top\), \(G_{\alpha_i} = N\alpha_i I_i + G_{(i)} G_{(i)}^\top\).

Proof. Let \(\mathcal{X} \in \mathcal{F}_r\) be a tensor with Tucker factorization \(\mathcal{X} = (\mathcal{G}, U_1, U_2, U_3) \in [\mathcal{X}]\). In quotient manifold framework [Absil et al. (2009)], the equivalent class \([\mathcal{X}]\) is called the fiber of total space. The lifted representation of the tangent space \(T\mathcal{X}_r\mathcal{M}_r/\sim\) is identified with a subspace of the tangent space \(T\mathcal{X}_r\mathcal{M}_r\) that does not produce a displacement along the fiber \([\mathcal{X}]\). This is realized by decomposing \(T\mathcal{X}_r\mathcal{M}_r\) into two complementary subspaces, the vertical and horizontal spaces, such that \(T\mathcal{X}_r\mathcal{M}_r = \mathcal{H}_{\mathcal{X}} \oplus \mathcal{V}_{\mathcal{X}}\), where \(\mathcal{H}_{\mathcal{X}}\) is the horizontal space and \(\mathcal{V}_{\mathcal{X}}\) is the vertical space. It should be emphasized that the decomposition is respect to the metric \([11]\). The vertical space \(\mathcal{V}_{\mathcal{X}}\) is the tangent space of the fiber \([\mathcal{X}]\). According to [Kasai and Mishra (2016)], the vertical space can be expressed as follows.

\[
\mathcal{V}_{\mathcal{X}} = \{ (-\sum_{i=1}^3 \mathcal{G} \times_i \Omega_i, U_1 \Omega_1, U_2 \Omega_2, U_3 \Omega_3) | \Omega_i^\top \Omega_i = -\Omega_i \}. \tag{19}
\]

Since horizontal space \(\mathcal{H}_{\mathcal{X}}\) is an orthogonal complement of \(\mathcal{V}_{\mathcal{X}}\) with respect to the Riemannian metric \([11]\), for all horizontal vectors \(\eta_{\mathcal{X}} = (\eta_{\mathcal{G}}, \eta_{\Omega_1}, \eta_{\Omega_2}, \eta_{\Omega_3}) \in \mathcal{H}_{\mathcal{X}}\) we have

\[
\langle \eta_{\mathcal{X}}, \zeta_{\mathcal{X}} \rangle_{\mathcal{X}} = 0, \forall \zeta_{\mathcal{X}} \in \mathcal{V}_{\mathcal{X}}. \tag{20}
\]

Using the expression for the horizontal space, the above equation is equivalent to the following one:

\[
\sum_{i=1}^3 \langle \eta_i, U_i \Omega_i \mathcal{G}_{(i)} \mathcal{G}_{(i)}^\top \rangle + \langle \eta_{\mathcal{G}}, -\sum_{i=1}^3 \mathcal{G} \times_i \Omega_i \rangle
\]

\[
+ \sum_{i=1}^3 N\alpha_i \langle \eta_i, (I_i - P_i P_i^\top) U_i \Omega_i \rangle = 0. \tag{21}
\]
Using the property for the Euclidean inner product that for matrix $A, B, C, D$ we have $(A, BCD) = (B^TAD^T, C)$. And for tensor $A, B$ and matrix $C$ we have $(A, B \times_i C) = (A_{(i)}B_{(i)}^T, C)$. The above equation (22) is equivalent to the following one

$$\sum_{i=1}^{3} (U_i^\top \eta_i g_{(i)}^\top + \eta_i g_{(i)}^\top + N\alpha_i(U_i - P_iP_i^\top U_i)^\top \eta_i, \Omega_i) = 0, \forall \text{skew matrix } \Omega_i \tag{22}$$

Thus we have $\eta_{\mathcal{X}}$ satisfy the following conditions

$$(P_iP_i^\top U_i)^\top \eta_i g_{(i)}^\top + (U_i - P_iP_i^\top U_i)^\top \eta_i(N\alpha_iI + g_{(i)}^\top g_{(i)}) \text{ is a symmetric matrix} \forall i \in \{1, 2, 3\}. \tag{23}$$

Defining $V_i := P_iP_i^\top U_i$, $W_i := U_i - P_iP_i^\top U_i$, $G_i := g_{(i)}^\top g_{(i)}$, $G_{\alpha_i} := N\alpha_iI + g_{(i)}^\top g_{(i)}$, we obtain the formula for the horizontal space:

$$H_{\mathcal{X}} = \{ \eta_{\mathcal{X}} \in T_{\mathcal{X}}M_r | V_i^\top \eta_i G_i + W_i^\top \eta_i G_{\alpha_i} \text{is symmetric} \} \tag{24}$$

### VII.1.2 Proof of Prop. (1)

Suppose $\mathcal{X}$ has tucker factors $(\mathcal{G}, U_1, U_2, U_3)$, then one can certify that:

$$\pi^{-1}(\mathcal{X}) = [\mathcal{G}, U_1, U_2, U_3].$$

And hence the equivalent relationship $\sim$ defined by the equivalent classes $[\mathcal{G}, U_1, U_2, U_3]$ can also be expressed in terms of the map $\pi(\cdot)$:

$$(\mathcal{G}, U_1, U_2, U_3) \sim (\mathcal{H}, V_1, V_2, V_3) \text{ if and only if } \pi(\mathcal{G}, U_1, U_2, U_3) = \pi(\mathcal{H}, V_1, V_2, V_3).$$

Since $\pi(\cdot)$ is a submersion (see Lemma 9), by the submersion theorem (Prop. 3.5.23 of [Abraham et al. 2012]), the equivalent relation $\sim$ defined by the equivalent classes is regular and the quotient manifold $M_r/\sim$ is diffeomorphic to $F_r$. And according to the proof of Prop. 3.5.23 of [Abraham et al. 2012], the mapping $\rho([\mathcal{G}, U_1, U_2, U_3]) = \mathcal{G} \times_{i=1}^{3} U_i$ defines the diffeomorphism from $M_r/\sim$ to $F_r$. Therefore, $\rho(\mathcal{X}) = \rho^{-1}(\mathcal{X}) = [\mathcal{G}, U_1, U_2, U_3]$, where $[\mathcal{G}, U_1, U_2, U_3]$ is the tucker representation of $\mathcal{X}$.

### VII.1.3 Proof of Proposition 2

Let $\overline{\mathcal{X}} = (\mathcal{G}, U_1, U_2, U_3)$ be any tucker factors of tensor $\mathcal{X} \in F_r$. According to the definition of Chordal distance of subspaces of different dimension [Ye and Lim 2014], we have

$$\text{dist}^2(\text{span}(\mathcal{X}_{(i)}), \text{span}(P_i)) = \text{dist}^2(\text{span}(U_i), \text{span}(P_i)) \tag{25}$$

$$= \sum_{i=1}^{r_i} \sin^2(\theta_i) + k_i - r_i \tag{26}$$

$$= \sum_{i=1}^{r_i} (1 - \cos^2(\theta_i)) + k_i - r_i \tag{27}$$

$$= \text{trace}(I) - \|P_i^\top U_i\|^2_F + k_i - r_i \tag{28}$$

$$= \text{trace}(U_i^\top (I - P_i^\top P_i)U_i) + k_i - r_i \tag{29}$$

where in the second equation $\theta_i$ is the $i$-th principal angle between span $U_i$ and span $P_i$, the second equation is derived from the definition of Chordal distance, the fourth equation is derived from the
fact that \(\cos(\theta_i)\) is the \(i\)-th singular value of \(P_i^TQ_i\) due to \(P_i\) and \(Q_i\) are orthogonal bases (see Alg 12.4.3 of [Golub and Van Loan, 2012]). Therefore for all \((\mathcal{G}, U_1, U_2, U_3) \in \mathcal{M}_r\), we have

\[
l(\pi(\mathcal{G}, U_1, U_2, U_3)) = \frac{1}{2}\|P_{12}(\mathcal{G} \times_{i=1}^3 U_i - \mathcal{R})\|_F^2 + \sum_{i=1}^3 (\text{trace}(U_i^T(I - P_i^TP_i)U_i) + k_i - r_i). \tag{30} \]

Which is equivalent to:

\[
l(\pi(\mathcal{G}, U_1, U_2, U_3)) = f(\mathcal{G}, U_1, U_2, U_3) + C \tag{31} \]

where \(C = \sum_{i=1}^3 (k_i - r_i)\) is a constant.

Note that the critical points of a function \(h(x)\) over a smooth manifold \(\mathcal{M}\) are those whose Riemannian gradient vanishing, that is \(\nabla h(x) = 0\). And one can show that:

\[
\nabla h(x) = 0 \text{ if and only if } Dh(x)[\eta_x] = 0 \forall \eta_x \in T_x\mathcal{M}. \tag{32} \]

To prove that \(\mathcal{X}\) is a critical point of \(l(\cdot)\) over \(\mathcal{F}_r\), if and only if \([\mathcal{X}]\) is a critical point of \(\hat{f}(\cdot)\) over \(\mathcal{M}_r/\sim\), we need to prove that

\[
\nabla l(\mathcal{X}) = 0 \text{ if and only if } \nabla \hat{f}(\mathcal{X}) = 0. \tag{33} \]

Note that since \(\nabla f(\mathcal{G}, U_1, U_2, U_3)\) is the horizontal lift of \(\hat{f}([\mathcal{X}])\) for all \((\mathcal{G}, U_1, U_2, U_3) \in [\mathcal{X}]\). We have \(\hat{f}([\mathcal{X}]) = 0\) if and only if \(f(\mathcal{G}, U_1, U_2, U_3) = 0\) for at least one \((\mathcal{G}, U_1, U_2, U_3) \in [\mathcal{X}]\). Thus to prove (33), one only need to certify

\[
\nabla l(\mathcal{X}) = 0 \text{ if and only if } \exists \mathcal{G}, U_1, U_2, U_3 \in [\mathcal{X}] \text{ such that } f(\mathcal{G}, U_1, U_2, U_3) = 0. \tag{34} \]

On one side, suppose \(\nabla l(\mathcal{X}) = 0\), and \(\overline{\mathcal{X}} = (\mathcal{G}, U_1, U_2, U_3) \in [\mathcal{X}]\). Let \(\eta_{\overline{\mathcal{X}}}\) be any tangent vector belonging to \(T_{\overline{\mathcal{X}}}\mathcal{M}_r\). We have:

\[
Df(\overline{\mathcal{X}})[\eta_{\overline{\mathcal{X}}}] = Dl(\pi(\overline{\mathcal{X}}))[D\pi(\overline{\mathcal{X}})[\eta_{\overline{\mathcal{X}}}]] = Dl(\mathcal{X})[D\pi(\overline{\mathcal{X}})[\eta_{\overline{\mathcal{X}}}]] = 0. \tag{35} \]

where the first equation is derived from equation (31) and chain rule of first order derivative; the third equation is due to \(\nabla l(\mathcal{X}) = 0\) and \(D\pi(\overline{\mathcal{X}})[\eta_{\overline{\mathcal{X}}}] \in T_{\overline{\mathcal{X}}}\mathcal{F}_r\) since \(\pi(\cdot)\) is a submersion (See Lemma 5). Because \(\eta_{\overline{\mathcal{X}}}\) is an arbitrary tangent vector, we have

\[
Df(\overline{\mathcal{X}})[\eta_{\overline{\mathcal{X}}}] = 0 \forall \eta_{\overline{\mathcal{X}}} \in T_{\overline{\mathcal{X}}}\mathcal{M}_r. \tag{36} \]

And according to (32) we have \(\nabla f(\overline{\mathcal{X}}) = 0\). Thus, we prove that

\[
\nabla l(\mathcal{X}) = 0 \Rightarrow \nabla \hat{f}(\overline{\mathcal{X}}) = 0. \tag{37} \]

On the other side, suppose \(\overline{\mathcal{X}} = (\mathcal{G}, U_1, U_2, U_3) \in [\mathcal{X}]\) and \(f(\overline{\mathcal{X}}) = 0\). Then for all \(\eta_{\overline{\mathcal{X}}} \in T_{\overline{\mathcal{X}}}\mathcal{F}_r\) we have:

\[
Dl(\mathcal{X})[\eta_{\overline{\mathcal{X}}}] = Dl(\pi(\overline{\mathcal{X}}))[\eta_{\overline{\mathcal{X}}}] = Dl(\pi(\overline{\mathcal{X}}))[D\pi(\overline{\mathcal{X}})[\eta_{\overline{\mathcal{X}}}]] = Df(\overline{\mathcal{X}})[\eta_{\overline{\mathcal{X}}}] = 0. \tag{38} \]

where the second equation is because there exist \(\eta_{\overline{\mathcal{X}}} \in T_{\overline{\mathcal{X}}}\mathcal{M}_r\) such that \(D\pi(\overline{\mathcal{X}})[\eta_{\overline{\mathcal{X}}}] = \eta_{\overline{\mathcal{X}}}\) due to \(\pi(\cdot)\) being a submersion (See Lemma 5); the third equation is derived by the equation (31) and chain rule of first order derivative; the fourth equation is due to \(\nabla f(\overline{\mathcal{X}}) = 0\). Thus we have proved that

\[
\nabla f(\overline{\mathcal{X}}) = 0 \Rightarrow \nabla l(\mathcal{X}) = 0. \tag{39} \]

Since we have proved both (39) and (44), we have (34) holds.
VII.1.4 Proof of Proposition 3

Since the Euclidean ambient space \( \mathbb{R}^{r_1 \times r_2 \times r_3} \times \mathbb{R}^{n_1 \times r_1} \times \mathbb{R}^{n_2 \times r_2} \times \mathbb{R}^{n_3 \times r_3} \) is a special smooth manifold, with tangent space at each point being the ambient space itself [Absil et al. 2009]. Therefore, one can endow the ambient space with a metric, and treat it as a Riemannian manifold. By endowing the ambient space with the same metric with total space, namely:

\[
\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbb{E}} = \sum_{i=1}^{3} \langle \mathbf{X}_{U_i}, \mathbf{Y}_{U_i} \rangle_{(i)} + \sum_{i=1}^{3} \langle \mathbf{X}_{g}, \mathbf{Y}_{g} \rangle_{(i)} + \langle \mathbf{X}_{U_i}, \mathbf{Y}_{U_i} \rangle_{\mathbb{E}}, \forall \mathbf{X}, \mathbf{Y} \in \mathbb{E}
\]

where \( \mathbf{X}, \mathbf{Y}, \mathbf{Z} \) are any ambient vectors, and all of them are tuples like \( (\mathbf{X}_{g}, \mathbf{X}_{U_1}, \mathbf{X}_{U_2}, \mathbf{X}_{U_3}) \).

The scaled Euclidean of the cost \( \mathbf{f}(\mathbf{X}) \) means the ambient vector which satisfies the following condition

\[
\langle \nabla f(\mathbf{X}), \mathbf{Y} \rangle_{\mathbb{E}} = Df(\mathbf{X})[\mathbf{Y}], \forall \mathbf{Y} \in \mathbb{E}
\]

This equation is equivalent to the following:

\[
\sum_{i=1}^{3} \langle \nabla f(\mathbf{X}), \mathbf{Y} \rangle_{(i)} + \langle \mathbf{Y}_{g}, \nabla f(\mathbf{X}) \rangle + \sum_{i=1}^{3} \langle \nabla f(\mathbf{X}), (I_i - P_i P_i^T) \mathbf{Y}_{U_i} \rangle
\]

By taking the partial Euclidean gradient both side of above equation with respect to \( \mathbf{Y}_{g} \) and \( \mathbf{Y}_{U_i} \), one has

\[
\nabla g f(\mathbf{X}) = \nabla g f(\mathbf{X})
\]

\[
\nabla U_i f(\mathbf{X}) = E_i (\mathbf{X}_{g})_{(i)}^{-1} + F_i (N \mathbf{a} + (\mathbf{X}_{g})_{(i)})^{-1}
\]

where \( E_i = P_i P_i^T \nabla U_i, f(\mathbf{X}) \) and \( F_i = \nabla U_i, f(\mathbf{X}) - E_i \).

VII.1.5 Proof of Proposition 4

According to [Absil et al. 2009], to prove \( \mathcal{M}_r \sim \mathcal{M}_r \) has the structure of Riemannian manifolds, one need to show that for all \( \mathcal{X} \in \mathcal{M}_r \sim \mathcal{M}_r \) and for all tangent vectors \( \eta_{\mathcal{X}}, \xi_{\mathcal{X}} \in \mathcal{T}_\mathcal{X} \mathcal{M}_r \sim \mathcal{M}_r \) we have

\[
\langle \eta_{\mathcal{X}}, \xi_{\mathcal{X}} \rangle_{\mathcal{M}_r} = \langle \eta_{\mathcal{X}}, \xi_{\mathcal{X}} \rangle_{\mathcal{M}_r}, \forall \mathcal{X}_{1}, \mathcal{X}_{2} \in [\mathcal{X}]
\]

where \( \eta_{\mathcal{X}}, \xi_{\mathcal{X}} \) are horizontal lift of \( \eta_{\mathcal{X}} \) and \( \xi_{\mathcal{X}} \). To prove that, we first express \( \mathcal{X}_2, \eta_{\mathcal{X}_2}, \xi_{\mathcal{X}_2} \) in terms of \( \mathcal{X}_1, \eta_{\mathcal{X}_1}, \xi_{\mathcal{X}_1} \), then verify the invariant property (49).

Let \( \mathcal{X}_1 = (\mathcal{G}, U_1, U_2, U_3) \). Since \( \mathcal{X}_1, \mathcal{X}_2 \in [\mathcal{X}] \), there exist orthogonal matrices \( O_i \in O(r_i) \) such that

\[
\mathcal{X}_2 = (\mathcal{G} \times_{i=1}^{3} O_i^T, U_1 O_1, U_2 O_2, U_3 O_3).
\]

Let \( \eta_{\mathcal{X}_2} = (\eta_{\mathcal{G}}, \eta_{U_1}, \eta_{U_2}, \eta_{U_3}) \), in this paragraph, we will prove that \( \eta_{\mathcal{X}_2} \) can be expressed by the following formula

\[
\eta_{\mathcal{X}_2} = (\eta_{\mathcal{G}} \times_{i=1}^{3} O_i^T, \eta_{U_1} O_1, \eta_{U_2} O_2, \eta_{U_3} O_3).
\]

Note that \( \eta_{\mathcal{X}_2} \) is the horizontal lift of \( \eta_{\mathcal{X}_1} \), to prove (51), one only need to show that \( (\eta_{\mathcal{G}} \times_{i=1}^{3} O_i^T, \eta_{U_1} O_1, \eta_{U_2} O_2, \eta_{U_3} O_3) \) satisfy the following two conditions (See Sec. 3.6.2 of [Absil et al. 2009])

\[
\zeta = \mathcal{H}_{\mathcal{X}_2}
\]

\[
D\tau(\mathcal{X}_2)[\zeta] = \eta_{\mathcal{X}_1}
\]
where for brevity we denote \((\eta_\theta \times 3_{i=1}^3 O_i \top, \eta_1 O_1, \eta_2 O_2, \eta_3 O_3)\) by \(\zeta\); the \(H_{\mathcal{X}}\) is the horizontal space at \(\mathcal{X}\) (See Lemma \[5\] for its expression); \(\tau(\cdot)\) is the nature mapping from \(\mathcal{M}_r\) to \(\mathcal{M}_r/\sim\) which is defined by

\[
\tau(\mathcal{X}) = [\mathcal{X}]
\]

Note that \(\tau(\cdot)\) is a composition of map \(\rho(\cdot)\) and map \(\pi(\cdot)\) defined in Prop. \[1\] and Lemma \[5\] namely

\[
\tau(\mathcal{X}) = \rho(\pi(\mathcal{X})).
\]

According to Lemma \[6\], \(H_{\mathcal{X}, \eta} = \{\eta_{\mathcal{X}} \in T_{\mathcal{X}, \eta}\mid V_i \top \eta_i G_i + W_i \top \eta_i G_{\alpha_i} \text{ is symmetric}\}\) where \(V_i = P_i P_i \top U_i, W_i = U_i - V_i, G_i = G_{\alpha_i} \top\) and \(G_{\alpha_i} = N\alpha_i I_i + G_{\alpha_i} G_{\alpha_i} \top\). Using the equation \((50)\), we have:

\[
H_{\mathcal{X}_2} = \{\eta_{\mathcal{X}_2} \in T_{\mathcal{X}_2, \eta}\mid O_i \top V_i \top \eta_i O_i, O_i \top W_i \top \eta_i O_i, O_i \top G_{\alpha_i}\text{ is symmetric}\}
\]

(Note that when proving the above equation, we use the equations like: \((G \times 3_{i=1}^3 O_i \top)_{(1)} = O_i \top G_{\alpha_i} G_{\alpha_i} \top\). To prove \(\zeta \in H_{\mathcal{X}_2}\), on one hand we noticed that:

\[
\zeta_i \top U_i O_i + O_i \top U_i \top \eta_i O_i = O_i \top \eta_i U_i O_i + O_i \top U_i \top \eta_i O_i = 0
\]

where the first equation use the fact \(\zeta_i = \eta_i O_i\), the third equation use the fact \(\eta_i \in T_{U_i, St(r_i, n_i)}\) is equivalent to \(\eta_i \top U_i + U_i \top \eta_i = 0\) (See Sec 3.5.7 of \[Abisl et al. (2009)\]). The above equation implies that \(\zeta_i \in T_{U_i, O_i}\) St(r_i, n_i). And hence we have

\[
\zeta \in \mathbb{R}^{r_1 \times r_2 \times r_3} \times T_{U_1, O_1}\text{St}(r_1, n_1) \times T_{U_2, O_2}\text{St}(r_2, n_2) \times T_{U_3, O_3}\text{St}(r_3, n_3) = T_{\mathcal{X}_2, \eta}\text{M}_r.
\]

One the other hand, we have \(O_i \top V_i \top \zeta_i O_i \top G_{\alpha_i}\text{ is symmetric since:}\)

\[
(O_i \top V_i \top \zeta_i O_i \top G_{\alpha_i}, O_i \top W_i \top \zeta_i O_i \top G_{\alpha_i}, O_i) = (O_i \top V_i \top \eta_i O_i \top G_{\alpha_i}, O_i \top W_i \top \eta_i O_i \top G_{\alpha_i}, O_i) = O_i \top (V_i \top \eta_i O_i \top G_{\alpha_i}) O_i = O_i \top V_i \top \zeta_i O_i \top G_{\alpha_i} O_i + O_i \top W_i \top \zeta_i O_i \top G_{\alpha_i} O_i.
\]

Thus, we have proved that \(\zeta \in H_{\mathcal{X}_2}\). The following equations verify \((53)\) holds.

\[
D\tau(\mathcal{X}_2)\{\zeta\} = D\rho(\pi(\mathcal{X}_2))[D\pi(\mathcal{X}_2)\{\zeta\}]
\]

\[
= D\rho(\mathcal{X})\left[\zeta G \times 3_{i=1}^3 U_i O_i + \sum_{i=1}^3 \left(G \times 3_{i=1}^3 O_i \top\right) \times_i \zeta_i \times_{1 \leq j \leq 3, j \neq i} U_i O_i\right]
\]

\[
= D\rho(\mathcal{X})\left[\eta G \times 3_{i=1}^3 U_i O_i + \sum_{i=1}^3 \left(G \times 3_{i=1}^3 O_i \top\right) \times_i \eta_i O_i \times_{1 \leq j \leq 3, j \neq i} U_i O_i\right]
\]

\[
= D\rho(\mathcal{X})\left[\eta G \times 3_{i=1}^3 U_i + \sum_{i=1}^3 \left(G \times 3_{i=1}^3 O_i\right) \times_i \eta_i \times_{1 \leq j \leq 3, j \neq i} U_i\right]
\]

\[
= D\rho(\mathcal{X})\left[D\pi(\mathcal{X}_1)\{\eta_{\mathcal{X}_1}\}\right]
\]

\[
= D\rho(\mathcal{X})\left[D\pi(\mathcal{X}_1)\{\eta_{\mathcal{X}_1}\}\right]
\]

\[
= D\rho(\mathcal{X})\left[D\pi(\mathcal{X}_1)\{\eta_{\mathcal{X}_1}\}\right]
\]

\[
= \eta_{\mathcal{X}_1}
\]
where the first equation is derived by the chain rule of derivative, the second equation is derived by using our definition of $\zeta$, the fourth equation is using the property of tensor matrix product that $A \times_i A \times_j B = A \times_i (BA)$ and $A \times_i A \times_j B = A \times_j B \times_j A \forall j \neq i$ [Kolda and Bader (2009)], the fifth equation result from [14], the eighth equation is because $\eta_{\mathbf{X}}$ is the horizontal lift of $\eta_{\mathbf{X}}$.

By similar arguments of above paragraph, one can verify that
\[
\xi_{\mathbf{X}_2} = (\xi_{\mathbf{g}} \times_{i=1}^3 O_i^T, \xi_1 O_1, \xi_2 O_2, \xi_3 O_3).
\]
Now we have
\[
\langle \eta_{\mathbf{X}_2}, \xi_{\mathbf{X}_2} \rangle_{\mathbf{X}_2} = \sum_{i=1}^3 \langle \eta_i O_i, \xi_i O_i((\mathbf{g} \times_{i=1}^3 O_i^T)_{(i)} (\mathbf{g} \times_{i=1}^3 O_i^T)_T)_{(i)} \rangle + \langle \eta_{\mathbf{g}} \times_{i=1}^3 O_i, \xi_{\mathbf{g}} \times_{i=1}^3 O_i \rangle
\]
\[
+ \sum_{i=1}^3 N \alpha_i \langle \eta_i O_i, (I_i - P_i P_i^T) \xi_i O_i \rangle
\]
\[
= \sum_{i=1}^3 \langle \eta_i O_i, \xi_i O_i O_i^T \mathbf{g}_{(i)}(\mathbf{g})_{(i)}^T O_i \rangle + \langle \eta_{\mathbf{g}} \times_{i=1}^3 O_i, \xi_{\mathbf{g}} \times_{i=1}^3 O_i \rangle
\]
\[
+ \sum_{i=1}^3 N \alpha_i \langle \eta_i O_i, (I_i - P_i P_i^T) \xi_i O_i \rangle
\]
\[
= \sum_{i=1}^3 \langle \eta_i, \xi_i \mathbf{g}_{(i)}(\mathbf{g})_{(i)}^T \rangle + \langle \eta_{\mathbf{g}}, \xi_{\mathbf{g}} \rangle + \sum_{i=1}^3 N \alpha_i \langle \eta_i, (I_i - P_i P_i^T) \xi_i \rangle
\]
\[
= \langle \eta_{\mathbf{X}_1}, \xi_{\mathbf{X}_1} \rangle_{\mathbf{X}_1}
\]
(68)
(69)
(70)
where the first equation use the expressions of $\mathbf{X}_2, \eta_{\mathbf{X}_2}, \chi_{\mathbf{X}_2}$ in terms of $\mathbf{X}_1, \eta_{\mathbf{X}_1}, \chi_{\mathbf{X}_1}$ (see equations (50-51-67)); the second equation is derived by using equations like
\[
(\mathbf{g} \times_{i=1}^3 O_i^T)_{(i)}(\mathbf{g} \times_{i=1}^3 O_i^T)_T = (O_i^T \mathbf{g}_{(i)}(O_i^T \otimes O_i^T)^T)(O_i^T \mathbf{g}_{(i)}(O_i^T \otimes O_i^T)^T)^T = O_i^T \mathbf{g}_{(i)}(\mathbf{g})_{(i)}^T O_i
\]
the third equation is derived from the fact that Euclidean inner product is orthogonal invariant. And the invariant property of the proposed metric is being proved.

VII.2 Derivation of The Expressions of Optimization Related Objects

VII.2.1 Projector from ambient space onto tangent space

We call the Euclidean space
\[
\mathbb{R}^{r_1 \times r_2 \times r_3} \times \mathbb{R}^{n_1 \times r_1} \times \mathbb{R}^{n_2 \times r_2} \times \mathbb{R}^{n_3 \times r_3}
\]
the ambient space. The vector belonging to ambient space is called by ambient vector. One ambient vector is denoted by $(Z_g, Z_1, Z_2, Z_3)$, for brevity the notation may be shorted to $Z$.

**Proposition 7.** Let $M_r$ be the total space, endowed with the Riemannian metric [11]. Let $\mathcal{X} = (\mathcal{G}, U_1, U_2, U_3) \in M M_r$ Then the orthogonal projection of an ambient vector $(Z_g, Z_1, Z_2, Z_3)$ onto the tangent space $T_{\mathcal{X}} M_r$ can be computed by
\[
\Psi_{\mathcal{X}}(Z_g, Z_1, Z_2, Z_3) = \begin{cases}
\begin{aligned}
Z_g &= Z_g - V_1 S_1(\mathbf{g}(1)^T \mathbf{g})_{(1)}^{-1} - W_1 S_1(\mathbf{g}(1)^T \mathbf{g})_{(1)} + \alpha_1 N \mathbf{I}_1^{-1} \\
Z_1 &= Z_1 - V_2 S_2(\mathbf{g}(2)^T \mathbf{g})_{(2)}^{-1} - W_2 S_2(\mathbf{g}(2)^T \mathbf{g})_{(2)} + \alpha_2 N \mathbf{I}_2^{-1} \\
Z_2 &= Z_2 - V_3 S_3(\mathbf{g}(3)^T \mathbf{g})_{(3)}^{-1} - W_3 S_3(\mathbf{g}(3)^T \mathbf{g})_{(3)} + \alpha_3 N \mathbf{I}_3^{-1}
\end{aligned}
\end{cases}
\]
(72)
where $V_i = P_i P_i^T U_i$ and $W_i = U_i - V_i$ and $S_i$ is the solution of the following matrix linear equation

$$\begin{cases} \text{sym}(V_i^T V_i S_i (G(i) G(i)^T)^{-1}) - \text{sym}(U_i^T Z_i) + \text{sym}(W_i^T W_i S_i (N \alpha_i I_i + G(i) G(i)^T)^{-1}) = 0, \\
S_i = S_i^T \end{cases}$$

(73)
in which $\text{sym}(A) = 1/2(A + A^T)$ for all square matrices.

**Proof.** The orthogonal projection of an ambient vector to the tangent space, is computed by subtraction of its component belongs to the normal space. To begin with we derive the normal space $N_{\mathcal{X}}$ which orthogonal complement of $T_{\mathcal{X}} M_r$ with respect to the Riemannian metric (11). Let $\zeta = (\zeta_1, \zeta_2, \zeta_3) \in N_{\mathcal{X}}$ be any vector of the normal space. Then we have

$$\langle \zeta, \eta \rangle_{\mathcal{X}} = 0 \forall \eta \in T_{\mathcal{X}} M_r \quad (74)$$

Since the tangent space of total space can be expressed as

$$T_{\mathcal{X}} M_r = R^{r_1 \times r_2 \times r_3} \times T_{U_i} St(r_1, n_1) \times T_{U_2} St(r_2, n_2) \times T_{U_3} St(r_3, n_3) \quad (75)$$

where the tangent space of Stiefel manifold can be formulated as:

$$T_{U_i} St(r_1, n_1) = \{ U_i \Omega_i + U_{i, \perp} K_i | \Omega_i \in R^{r_1 \times r_i}, K_i \text{ skew and } K_i \in R^{(n_i-r_i) \times r_i} \} \quad (76)$$

and $U_{i, \perp} \text{ is also a matrix with orthogonal columns such that } U_{i, \perp} U_i = 0$. Using the formula (75) and (76), the equation (74) is equivalent to the following formula

$$\sum_{i=1}^3 (U_i \Omega_i + U_{i, \perp} K_i, \zeta_i G(i) G(i)^T) + \langle \eta \varphi, \zeta \varphi \rangle + \sum_{i=1}^3 N \alpha_i (U_i \Omega_i + U_{i, \perp} K_i, (I_i - P_i P_i^T) \zeta_i) = 0 \quad (77)$$

forall $K_i \in R^{r_1 \times (n_i - r_i)}$ skew matrix $\Omega_i \in R^{r_1 \times r_i}$, skew matrix $\varphi \in R^{r_1 \times r_2 \times r_3}$.

Using the fact that the condition $\langle Z, U_i \Omega_i + U_{i, \perp} K_i \rangle = 0 \forall K_i$ and skew matrix $\Omega_i$ is equivalent to that $Z = U_i S_i$, where $S_i$ is any symmetric matrix. The above equation (77) can be simplified as the following conditions

$$\begin{cases} \zeta \varphi = 0, \\
P_i P_i^T \zeta_i = V_i S_i (G(i) G(i)^T)^{-1} + (I_i - P_i P_i^T) \zeta_i (N \alpha_i I_i + G(i) G(i)^T)^{-1} = U_i S_i \forall i \in \{1, 2, 3\} \quad (78) \end{cases}$$

where $S_i$ is a symmetric matrix. Note that the second equation of (78), is equivalent to the following equations:

$$P_i P_i^T \zeta_i = V_i S_i (G(i) G(i)^T)^{-1} \quad (79)$$

where $V_i = P_i P_i^T U_i$ and $W_i = U_i - V_i$, and the first equation is obtained by multiplying the both side of the second formula of (78) by $P_i P_i^T$, the second equation is obtained by multiplying the both side of the second formula of (78) by $I - P_i P_i^T$. The above equation array is further equivalent to

$$\zeta_i = V_i S_i (G(i) G(i)^T)^{-1} + W_i S_i (N \alpha_i I_i + G(i) G(i)^T)^{-1} \quad (80)$$

since one can obtain equation (80) by adding the two equations in (79), and one an obtain the two equations in (79) via multiplying both sides of (80) by $P_i P_i^T$ or $I - P_i P_i^T$. Therefore, the normal space $N_{\mathcal{X}}$ can be expressed as follows.

$$N_{\mathcal{X}} = \{(0, \zeta_1, \zeta_2, \zeta_3) | \zeta_i = V_i S_i (G(i) G(i)^T)^{-1} + W_i S_i (N \alpha_i I_i + G(i) G(i)^T)^{-1}, S_i = S_i^T, 1 \leq i \leq 3 \} \quad (81)$$
Now the projection of an ambient vector can be calculated by subtracting its components in the normal space $N_{\mathcal{X}}$. Specifically, suppose $\Psi_{\mathcal{X}}(Z_{\mathcal{G}}, Z_{1}, Z_{2}, Z_{3}) = (Y_{\mathcal{G}}, Y_{1}, Y_{2}, Y_{3})$, we have $Y_{\mathcal{G}} = Z_{\mathcal{G}}$ and there exist symmetric matrices $S_{i}$ such that

$$Y_{i} = Z_{i} - V_{i}S_{i}(G_{(i)}G_{(i)}^{\top})^{-1} - W_{i}S_{i}(N\alpha_{i}I_{i} + G_{(i)}G_{(i)}^{\top})^{-1}$$  \hspace{1cm} (82)

where $V_{i} = P_{i}P_{i}^{\top}U_{i}$, $W_{i} = U_{i} - V_{i}$ and $1 \leq i \leq 3$. Since $(Y_{\mathcal{G}}, Y_{1}, Y_{2}, Y_{3}) \in T_{\mathcal{X}}\mathcal{M}_{r}$, we have

$$U_{i}^{\top}Y_{i} + Y_{i}^{\top}U_{i} = 0, 1 \leq i \leq 3.$$

By plugging in the equation (82) into the above equation we can obtain the linear equations for the symmetric matrix $S_{i}$:

$$\text{sym}(V_{i}^{\top}V_{i}S_{i}(G_{(i)}G_{(i)}^{\top})^{-1}) - \text{sym}(U_{i}^{\top}Z_{i}) + \text{sym}(W_{i}^{\top}W_{i}S_{i}(N\alpha_{i}I_{i} + G_{(i)}G_{(i)}^{\top})^{-1}) = 0.$$ \hspace{1cm} (84)

\[\square\]

VII.2.2 Projector from Tangent Space onto Horizontal Space

**Proposition 8.** Let $\mathcal{M}_{r}$ be the total space, endowed with the Riemannian metric $\langle \cdot, \cdot \rangle$. Let $\mathcal{X} = (\mathcal{G}, U_{1}, U_{2}, U_{3}) \in M\mathcal{M}_{r}$. Then the orthogonal projector $\Pi_{\mathcal{X}}$ from tangent space $T_{\mathcal{X}}\mathcal{M}_{r}$ to horizontal space $\mathcal{H}_{\mathcal{X}}$ has the following form

$$\Pi_{\mathcal{X}}(\eta_{\mathcal{X}}) = (\eta_{\mathcal{G}} + \sum_{i=1}^{3} \mathcal{G} \times_{i} \Omega_{i}, \eta_{1} - U_{1}\Omega_{1}, \eta_{2} - U_{2}\Omega_{2}, \eta_{3} - U_{3}\Omega_{3})$$  \hspace{1cm} (85)

where $\eta_{\mathcal{X}} = (\eta_{\mathcal{G}}, \eta_{1}, \eta_{2}, \eta_{3})$ is a tangent vector. And $(\Omega_{1}, \Omega_{2}, \Omega_{3})$ is the solution of the following linear matrix equation:

$$\begin{cases}
\text{skw}(V_{i}^{\top}V_{i}R_{i}G_{(i)}G_{(i)}^{\top}) + \text{skw}(G_{(i)}G_{(i)}^{\top}R_{i}) + \text{skw}(W_{i}^{\top}W_{i}R_{i}(N\alpha_{i}I_{i} + G_{(i)}G_{(i)}^{\top})) \\
- G_{(i)}(I_{j_{i}} \otimes \Omega_{k_{i}} + \Omega_{j_{i}} \otimes I_{k_{i}})G_{(i)}^{\top} \\
= \text{skw}[V_{i}^{\top}\eta_{i}G_{(i)}G_{(i)}^{\top} + W_{i}^{\top}\eta_{i}(N\alpha_{i}I_{i} + G_{(i)}G_{(i)}^{\top}) + G_{(i)}(\eta_{\mathcal{G}})^{\top}] \forall i \in \{1, 2, 3\}
\end{cases}$$  \hspace{1cm} (86)

where $j_{i} = \max\{k|k \in \{1, 2, 3\}, k \neq i\}$ and $k_{i} = \min\{k|k \in \{1, 2, 3\}, k \neq i\}$, $V_{i} = P_{i}P_{i}^{\top}U_{i}$ and $W_{i} = U_{i} - V_{i}$.

**Proof.** The projection from tangent space $T_{\mathcal{X}}\mathcal{M}_{r}$ onto the horizontal space $\mathcal{H}_{\mathcal{X}}$ is also derived by subtracting the normal component from the tangent vector. Note that the normal space to $T_{\mathcal{X}}\mathcal{M}_{r}$ is the vertical space $\mathcal{V}_{\mathcal{X}}$ defined in [19]. Then the projection $\Psi(\eta_{\mathcal{X}}) = (\varsigma_{\mathcal{G}}, \varsigma_{1}, \varsigma_{2}, \varsigma_{3})$ have the following form:

$$\begin{cases}
\varsigma_{\mathcal{G}} = \eta_{\mathcal{G}} + \sum_{i=1}^{3} \mathcal{G} \times_{i} \Omega_{i}, \\
\varsigma_{i} = \eta_{i} - U_{i}\Omega_{i}, \forall i \in \{1, 2, 3\}
\end{cases}$$  \hspace{1cm} (87)

where $\Omega_{i}$ is a skew matrix to be determined. Since $(\varsigma_{\mathcal{G}}, \varsigma_{1}, \varsigma_{2}, \varsigma_{3}) \in \mathcal{H}_{\mathcal{X}}$, then according to Prop. 6, it must satisfy that:

$$\text{skw}(V_{i}^{\top}\varsigma_{i}G_{i} + W_{i}^{\top}\varsigma_{i}G_{\alpha_{i}}) = 0 \forall i \leq 3$$  \hspace{1cm} (88)

where $V_{i} = P_{i}P_{i}^{\top}U_{i}$, $W_{i} = U_{i} - P_{i}P_{i}^{\top}U_{i}$, $G_{i} = G_{(i)}G_{(i)}^{\top}$, $G_{\alpha_{i}} = N\alpha_{i}I_{i} + G_{(i)}G_{(i)}^{\top}$, and $\text{skw}(\cdot)$ is a map define on square matrices, $\text{skw}(A) = 1/2(A - A^{\top})$. Doing some algebra, we obtain the linear system (86). \[\square\]
VII.2.3 Retraction

We prove that the retraction is compatible with the metric by showing it induce a retraction over the quotient manifold.

Lemma 9. Let \( R(\cdot) \) be the retraction defined in \([13]\). Then

\[
E_{[\mathcal{X}]}(\eta_{[\mathcal{X}]}):= [R_{\mathcal{X}}(\eta_{\mathcal{X}})]
\]

where \( \mathcal{X} \in [\mathcal{X}] \) and \( \eta_{\mathcal{X}} \) is a horizontal lift of \( \eta_{[\mathcal{X}]} \), defines an retraction over the quotient manifold \( \mathcal{M}_{r/\sim} \).

Proof. Let \( \mathcal{X}_1, \mathcal{X}_2 \) be any tucker factors belonging to equivalent classes \([\mathcal{X}]\). Let \( \eta_{[\mathcal{X}_1]} \) and \( \eta_{[\mathcal{X}_2]} \) are horizontal lifts of \( \eta_{[\mathcal{X}]} \). Suppose \( \mathcal{X}_1 = (\mathcal{G}, U_1, U_2, U_3) \), then we have

\[
[R_{\mathcal{X}_2}(\eta_{[\mathcal{X}_2]})] = \left[ R_{[\mathcal{G} \times \mathcal{O} \times \mathcal{O}]} \right] = \left[ (\mathcal{G} + \eta_{\mathcal{G}}) \times_{i=1}^{3} O_i^\top, \{\eta_{\mathcal{O}}(U_i + \eta_{\mathcal{O}})\}_{i=1}^{3} \right]
\]

where the first equation is because of \([50]\) and \([51]\), the second equation use the definition of retraction \([13]\), the third equation is because \( uf(\mathcal{A}) = uf(\mathcal{A}) \mathcal{O} \) for all orthogonal matrix \( \mathcal{O} \).

Thus, according to Prop 4.1.3 of \( \text{Absil et al.} \) \(2009\), we have that \( E(\cdot) \) is a valid retraction of \( \mathcal{M}_{r/\sim} \).

\( \square \)

VII.2.4 The Euclidean Gradient of the Cost

The Euclidean gradient of the cost \( \nabla f(\mathcal{G}, U_1, U_2, U_3) \) can be decompose as \( \nabla f(\mathcal{G}, U_1, U_2, U_3) = (\nabla_{\mathcal{G}} f, \nabla_{U_1} f, \nabla_{U_2} f, \nabla_{U_3} f) \) where \( \nabla_{\mathcal{G}} f \) and \( \nabla_{U_i} f \) are partial derivatives of the cost with respect to \( \mathcal{G} \) and \( U_i \). By doing some algebra, one has:

\[
\nabla_{\mathcal{G}} f(\mathcal{G}, U_1, U_2, U_3) = \mathcal{S} \times_{i=1}^{3} U_i^\top
\]

\[
\nabla_{U_i} f(\mathcal{G}, U_1, U_2, U_3) = \mathcal{S}_{(i)}(U_{j_i} \otimes U_{k_i}) \mathcal{G}_{(i)} + N \alpha_i \mathcal{W}_i
\]

where

\[
\mathcal{S} = P_{\Omega}(\mathcal{G} \times_{i=1}^{3} U_i - \mathcal{R})
\]

\[
\mathcal{W}_i = U_i - P_i P_i^\top U_i
\]

\( j_i = \max\{k|k \in \{1, 2, 3\}, k \neq i\} \) and \( k_i = \min\{k|k \in \{1, 2, 3\}, k \neq i\} \).

VII.3 More Empirical Results: Simulation

In the simulations, we complete a random tensor \( \mathcal{R} \) whose size is fixed to \( 5000 \times 5000 \times 5000 \) and multilinear rank to \( (10, 10, 10) \). And it is generated by \( \mathcal{R} = \mathcal{A} \times_1 \mathcal{B}_1 \times_2 \mathcal{B}_2 \times_3 \mathcal{B}_3 \) where \( \mathcal{A} \in \mathbb{R}^{10 \times 10 \times 10} \) and \( \mathcal{B}_i \in \mathbb{R}^{5000 \times 10} \) are random (multi-dimensional) arrays with i.i.d standard Gaussian entries. The side informations are encoded in three feature matrices. They are generated by \( \mathcal{F}_i = \mathcal{B}_i + s||\mathcal{N}_i||_F \mathcal{N}_i \), where \( \mathcal{N}_i \) is a noise matrix with entries drew from i.i.d normal distribution. The indices of the observed entries \( \Omega \) are sampled from the full indices set of the \( 5000 \times 5000 \times 5000 \) tensor uniformly at random. Its cardinality \( |\Omega| \) is set to \( OS \times D \) where \( D = 3 \times (5000 \times 10^{-10^2}) + 10^3 \).
is the dimension of the manifolds of $5000 \times 5000 \times 5000$ tensors with multilinear rank $(10, 10, 10)$ and $OS$ is called the Over-Sampling ratio. We compare the five tensor completion solvers under the following four scenarios. In each run the compared solvers are started with the same initializer generated from random, and stopped when either the norm of the gradient is less than $10^{-4}$ or the number of iterations is more than 300. To show the effectiveness of the propose metric, we also implemented an Riemannian CG solver, with the least square metric Kasai and Mishra (2016). And the parameters of $CGSI$ and $FTCSI$ are set to the same values as they solve the same problem.

VII.3.1 Case 1: influence of sampling ratio

We study the number of observed samples on the performance of the compared solvers. We vary the oversampling ratio in the set $OS \in \{0.1, 1, 5\}$ while fixing the noise scale of the feature matrices to $10^{-5}$. Then, run the five solvers on each tasks. For each run, we set $\alpha_i, 1 \leq i \leq 3$ are all set to $10/|\Omega|$ and $\lambda = 0$ for CGSI and FTCSI. The parameters of other baselines are set to the defaults. We report the convergence behavior of the compared solvers in Fig. 5(a-c). Note that in Fig. 5(a) the RMSE curve of FTC coincides with that of GE and in Fig. 5(c) the RMSE curve of FTC coincides with that of FTCSI. From Fig. 5(a) and (b), we can see that only CGSI and FTCSI successfully bring the RMSE down below $10^{-2}$ when OS is smaller than 1. This shows that when the observed entries are scarce, using the side information in the optimization can make a big difference on the accuracy of tensor completion task. And from Fig. 5(a-c), we can see that CGSI converges to the solution faster than FTCSI. This is shows that our proposed metric can indeed accelerate the convergence of Riemannian conjugate gradient descent method.

VII.3.2 Case 2: influence of noisy side information

To study the affect of noisy feature matrix on the performance of the proposed method. We fix the oversampling ratio to $OS = 1$ and vary the noise scale of the feature $c$ matrix in the set $\{10^{-4}, 10^{-3}, 10^{-2}\}$. For CGSI and FTCSI, their parameters $\alpha_i$ are all set to 1 and $\lambda$ is set to 0. The convergence behavior of the compared methods are reported in Fig. 5(d-f). From these figures we can see that when converging, the RMSE of CGSI and FTCSI are similar. This is because they solve the same problem. And even the feature matrices are noisy, the RMSE of CGSI and FTCSI are much better than the other baselines. These figures also show that CGSI is much faster than FTCSI, which is attributed to that CGSI is endowed with a better Riemannian metric.

VII.3.3 Case 3: influence of non-relevant features

We consider the performance of the proposed method, when the provided feature matrices $F_i$ have much more columns than the correct ones $B_i$. The matrices $F_i \in \mathbb{R}^{5000 \times 10(k+1)}$ is generated by augmenting the correct feature matrices $B_i$ with $10k$ randomly generated columns. That is, we set $F_i = [B_i, G_i] + 10^{-5}\|B_i\|E_i$ where $G_i \in \mathbb{R}^{5000 \times 10k}$ and $E_i \in \mathbb{R}^{5000 \times 10(k+1)}$ are random matrices with entries drew from i.i.d standard Gaussian distribution. We fix the oversampling ratio to $OS = 1$, and vary the parameter $k \in \{10, 30, 50\}$. For CGSI and FTCSI, $\alpha_i, 1 \leq i \leq 3$ are set to 0.5 and $\lambda$ is set to 0. The parameters of other baselines are set to the default. We report the convergence behavior of the compared solvers in Fig. 5(g-i). From these figures we can see that both CGSI and FTCSI successfully bring the RMSE down around $10^{-5}$ even when the columns of $F_i$ are 50 times larger than $B_i$. And These figures also shows that the proposed solver CGSI converges much faster than FTCSI, which is attributed to CGSI being endowed with a better Riemannian metric.
VII.3.4 Case 4: influence of noisy samples

We consider the case where the observed entries are noisy by adding a scaled Gaussian noise $\epsilon P_{\Omega}(\mathcal{E})$ to $P_{\Omega}(\mathcal{R})$ where $\mathcal{E}$ is a noise tensor with i.i.d standard Gaussian entries. We fix the oversampling ratio $OS$ to 1, the noise scale $c$ of feature matrices to $10^{-4}$ and vary the noise scale of samples such that $\epsilon \in \{10^{-3}, 10^{-4}, 10^{-2}\}$. For CGSI and FTCSI, their parameters are set as follows. When $\alpha_i = 5, 1 \leq i \leq 3$ and $\lambda = 0$. The parameters of other baselines are set to defaults. We report the performance of the compared solvers in Fig. 5 (j-l). From these figures we can see that only the solvers for the proposed model, that is CGSI and FTCSI, bring the RMSE down to the level of noise $\epsilon$ when converging. This shows that when the observed entries are few, exploiting the side information can significantly improve the RMSE. Also we can see that CGSI converges much faster than FTCSI, this exhibit that the proposed metric (11) is able to accelerate the convergence of Riemannian conjugate gradient descent method.

References

Ralph Abraham, Jerrold E Marsden, and Tudor Ratiu. *Manifolds, tensor analysis, and applications*, volume 75. Springer Science & Business Media, 2012.

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Evrim Acar, Tamara G. Kolda, and Daniel M. Dunlavy. All-at-once optimization for coupled matrix and tensor factorizations. In *MLG’11*, 2011.

Robert M Bell and Yehuda Koren. Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter*, 9(2):75–79, 2007.

Alex Beutel, Partha Pratim Talukdar, Abhimanyu Kumar, Christos Faloutsos, Evangelos E Papalexakis, and Eric P Xing. Flexifact: Scalable flexible factorization of coupled tensors on hadoop. In *ICDM*, pages 109–117. SIAM, 2014.

Shouyuan Chen, Michael R Lyu, Irwin King, and Zenglin Xu. Exact and stable recovery of pairwise interaction tensors. In *NIPS*, 2013.

Marko Filipović and Ante Jukić. Tucker factorization with missing data with application to low-rank tensor completion. *Multidimensional systems and signal processing*, 26(3):677–692, 2015.

David H Foster, Kinjiro Amano, Sérgio MC Nascimento, and Michael J Foster. Frequency of metamerism in natural scenes. *Journal of the Optical Society of America A*, 23:2359–2372, 2006.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

Prateek Jain and Sewoong Oh. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, pages 1431–1439, 2014.

Hiroyuki Kasai and Bamdev Mishra. Low-rank tensor completion: a riemannian manifold preconditioning approach. In *ICML*, 2016.

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

Daniel Kressner, Michael Steinlechner, and Bart Vandereycken. Low-rank tensor completion by riemannian optimization. *BIT Numerical Mathematics*, 54(2):447–468, 2014.
Figure 5: Simulation results of different solvers on the task of tensor completion.
Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013.

Qiang Liu, Shu Wu, and Liang Wang. Cot: Contextual operating tensor for context-aware recommender systems. In *AAAI*, pages 203–209, 2015.

Yuanyuan Liu, Fanhua Shang, Hong Cheng, James Cheng, and Haohang Tong. Factor matrix trace norm minimization for low-rank tensor completion. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 866–874. SIAM, 2014.

Yuanyuan Liu, Fanhua Shang, Wei Fan, James Cheng, and Hong Cheng. Generalized higher order orthogonal iteration for tensor learning and decomposition. *IEEE transactions on neural networks and learning systems*, 27(12):2551–2563, 2016.

Bamdev Mishra. *A Riemannian approach to large-scale constrained least-squares with symmetries*. PhD thesis, Universite de Liege, Liege, Belgique, 2014.

Bamdev Mishra and Rodolphe Sepulchre. R3mc: A riemannian three-factor algorithm for low-rank matrix completion. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pages 1137–1142. IEEE, 2014.

Bamdev Mishra and Rodolphe Sepulchre. Riemannian preconditioning. *SIAM Journal on Optimization*, 26(1):635–660, 2016.

Atsuhiro Narita, Kohei Hayashi, Ryota Tomioka, and Hisashi Kashima. Tensor factorization using auxiliary information. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 501–516. Springer, 2011.

Piyush Rai, Yingjian Wang, and Lawrence Carin. Leveraging features and networks for probabilistic tensor decomposition. In *AAAI*, pages 2942–2948, 2015.

Bernardino Romera-Paredes and Massimiliano Pontil. A new convex relaxation for tensor completion. In *Advances in Neural Information Processing Systems*, pages 2967–2975, 2013.

Bernardino Romera-Paredes, Hane Aung, Nadia Bianchi-Berthouze, and Massimiliano Pontil. Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1444–1452, 2013.

Shaden Smith, Jongsoo Park, and George Karypis. An exploration of optimization algorithms for high performance tensor completion. In *Proceedings of the 2016 ACM/IEEE conference on Supercomputing*, 2016.

Yichen Wang, Robert Chen, Joydeep Ghosh, Joshua C Denny, Abel Kho, You Chen, Bradley A Malin, and Jimeng Sun. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1265–1274. ACM, 2015.

Yangyang Xu, Ruru Hao, Wotao Yin, and Zhixun Su. Parallel matrix factorization for low-rank tensor completion. *Inverse Problems & Imaging*, 9(2), 2015.

Ke Ye and Lek-Heng Lim. Distance between subspaces of different dimensions. *arXiv preprint arXiv:1407.0900*, 2014.

Ming Yuan and Cun-Hui Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, pages 1–38, 2015.
Xiaoqin Zhang, Zhengyuan Zhou, Di Wang, and Yi Ma. Hybrid singular value thresholding for tensor completion. In AAAI, pages 1362–1368, 2014.

Zemin Zhang and Shuchin Aeron. Exact tensor completion using t-svd. *IEEE Transactions on Signal Processing*, 2016.