

## Chordal-NMF with Riemannian Multiplicative Update

Flavia Esposito · Andersen Ang

Received: date / Accepted: date

**Abstract** Nonnegative Matrix Factorization (NMF) is the problem of approximating a given nonnegative matrix  $M$  through the conic combination of two nonnegative low-rank matrices  $W$  and  $H$ . Traditionally NMF is tackled by optimizing a specific objective function evaluating the quality of the approximation. This assessment is often done based on the Frobenius norm. In this study, we argue that the Frobenius norm as the “point-to-point” distance may not always be appropriate. Due to the nonnegative combination resulting in a polyhedral cone, this conic perspective of NMF may not naturally align with conventional point-to-point distance measures. Hence, a ray-to-ray chordal distance is proposed as an alternative way of measuring the discrepancy between  $M$  and  $WH$ . This measure is related to the Euclidean distance on the unit sphere, motivating us to employ nonsmooth manifold optimization approaches.

We apply Riemannian optimization technique to solve chordal-NMF by casting it on a manifold. Unlike existing works on Riemannian optimization that require the manifold to be smooth, the nonnegativity in chordal-NMF is a non-differentiable manifold. We propose a Riemannian Multiplicative Update (RMU) that preserves the convergence properties of Riemannian gradient descent without breaking the smoothness condition on the manifold.

We showcase the effectiveness of the Chordal-NMF on synthetic datasets as well as real-world multispectral images.

---

The authors have equal contributions.

Flavia Esposito  
Department of Mathematics,  
University of Bari Aldo Moro, Italy E-mail: flavia.esposito@uniba.it

Andersen Ang  
School of Electronics and Computer Science  
University of Southampton, UK E-mail: andersen.ang@soton.ac.uk

**Keywords** Nonnegative Matrix Factorization · Manifold · Chordal distance · Nonconvex Optimization · Multiplicative Update · Riemannian gradient

**Mathematics Subject Classification (2020)** MSC 15A23 · MSC 78M50 · MSC 49Q99 · MSC 90C26 · MSC 90C30

## 1 Introduction

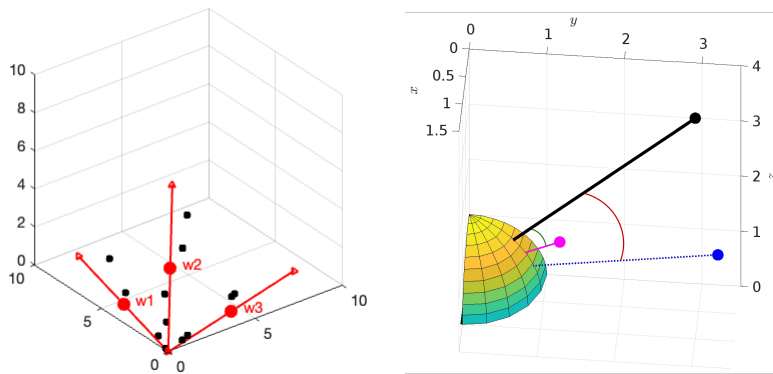
Given a nonnegative matrix  $\mathbf{M} \in \mathbb{R}_+^{m \times n}$  and a rank  $r \leq \min\{m, n\}$ , Nonnegative Matrix Factorization (NMF) is to find factor matrices  $\mathbf{W} \in \mathbb{R}_+^{m \times r}$  and  $\mathbf{H} \in \mathbb{R}_+^{r \times n}$  such that  $\mathbf{M} \approx \mathbf{W}\mathbf{H}$  [1]. NMF is commonly achieved by minimizing the Frobenius norm  $\|\mathbf{M} - \mathbf{W}\mathbf{H}\|_F = \sqrt{\sum_{ij} (M_{ij} - (\mathbf{W}\mathbf{H})_{ij})^2}$  (where  $M_{ij}$  is the  $(i, j)$ th-entry of  $\mathbf{M}$ ) that measures the quality of the approximation. Note that the nonnegativity constraints in  $\mathbf{W}$  and  $\mathbf{H}$  restrict linear combination in  $\mathbf{W}\mathbf{H}$  to conic combination, thus  $\mathbf{M} \approx \mathbf{W}\mathbf{H}$  is saying that  $\mathbf{M}$ , which represents a point cloud, is contained within a polyhedral cone generated by the  $r$  columns of  $\mathbf{W}$  with nonnegative weights encoded in  $\mathbf{H}$ , see Fig. 1. This conic view of NMF suggests that the point-to-point distance  $M_{ij} - (\mathbf{W}\mathbf{H})_{ij}$  in the Frobenius norm does not naturally fit NMF. In this work, we propose to measure the discrepancy between  $\mathbf{M}$  and  $\mathbf{W}\mathbf{H}$  using a ray-to-ray distance that we call *chordal distance* that we defer the background to the next section. In this work, we are interested in solving

$$\begin{aligned} & \underset{\mathbf{W} \geq \mathbf{0}, \mathbf{H} \geq \mathbf{0}}{\operatorname{argmin}} \left\{ F(\mathbf{W}, \mathbf{H}) := \frac{1}{n} \sum_{j=1}^n \left( 1 - \frac{\langle \mathbf{m}_{:j}, \mathbf{W}\mathbf{h}_{:j} \rangle}{\|\mathbf{m}_{:j}\|_2 \|\mathbf{W}\mathbf{h}_{:j}\|_2} \right) \right\} \\ & = \underset{\mathbf{W} \geq \mathbf{0}, \mathbf{H} \geq \mathbf{0}}{\operatorname{argmax}} \sum_{j=1}^n \frac{\langle \mathbf{m}_{:j}, \mathbf{W}\mathbf{h}_{:j} \rangle}{\|\mathbf{m}_{:j}\|_2 \|\mathbf{W}\mathbf{h}_{:j}\|_2}, \end{aligned} \quad (\text{Chordal-NMF})$$

where the objective function  $F : \mathbb{R}^{m \times r} \times \mathbb{R}^{r \times n} \rightarrow \mathbb{R}$  is defined as the chordal distance between  $\mathbf{W}\mathbf{H}$  and  $\mathbf{M}$ ,  $\|\cdot\|_2$  is the Euclidean norm,  $\langle \cdot, \cdot \rangle$  is the Euclidean inner product, and  $\mathbf{m}_{:j}$  is the  $j$ th column of  $\mathbf{M}$ . The constraints  $\geq \mathbf{0}$  denote element-wise nonnegativity, where  $\mathbf{0}$  is zero matrix of the appropriate size.

*Contribution.* Our contributions are 3-folds.

1. We propose a new model (Chordal-NMF) with motivation. To the best of our knowledge, such problem is new and has not been studied in the NMF literature.
2. Solving Chordal-NMF is not trivial: it is a nonsmooth nonconvex and block-nonconvex problem. We propose a Block Coordinate Descent (BCD) algorithm with Riemannian Multiplicative Update (RMU) for solving Chordal-NMF.
  - We provide a theory on deriving the Riemannian gradient of the objective function  $F$ .



**Fig. 1** Left: Picture of a rank-3 NMF. Data points (in black) that represent the columns of  $M$  encapsulated by a polyhedral cone generated by  $r = 3$  columns  $w_1, w_2, w_3$  (in red). Right: The sphere  $\mathbb{S}_+^2$  in the nonnegative orthant  $\mathbb{R}_+^3$ . The black dot is closer to the blue dot in Euclidean distance, but closer to the pink dot in chordal distance, which is equivalent to the geodesic arc length on the sphere.

- The nonnegativity constraints in Chordal-NMF introduce nondifferentiability in the manifold and make some existing manifold techniques infeasible or ineffective. We propose RMU to solve the nonsmoothness issue in the manifold optimization. In particular, we show that, if the initial variable in the algorithm is feasible, the whole sequence is guaranteed to be feasible.
3. We showcase the effectiveness of the Chordal-NMF on synthetic datasets as well as real-world multispectral images.

*Literature review.* NMF has a rich literature, for details we refer the reader to the book [1]. We review the literature related to chordal distance in Section 2 and Riemannian optimization in Section 3. As we will see in the next section, chordal distance is linked to the cosine similarity, commonly used in face recognition task [2, 3] and in deep learning as normalization [4]. Note that these works transformed the data to another space (known as Metric Learning) to improve the use of cosine similarity, in this work we do not perform data transformation. The chordal distance is also linked to the Hamming distance on binary code, which is in fact a form of cosine similarity [5]. Chordal distance is related to the Euclidean distance on unit sphere [6], in which manifold optimization technique can be applied.

*Paper organization and notation.* In the remaining of this section, we describe the overall algorithmic framework on solving Chordal-NMF. In Section 2 we discuss the motivation and the details of the chordal distance. In Section 3 we review Riemannian optimization techniques. In Section 4 and Section 5 we discuss how do we update the variable  $H$  and  $W$ , respectively. Section 6 contains the experiments and Section 7 sketches the conclusion.

*Notation.* We use {italic, bold italic, bold italic capital} letters to denote {scalar, vector, matrix} resp.. Given a matrix  $\mathbf{A}$ , we denote  $\mathbf{a}_{:j}$  the  $j$ th column of  $\mathbf{A}$  and  $\mathbf{a}_j$  the  $j$ th row of  $\mathbf{A}$ . We denote by  $(\mathbf{M})_{ij}$  or  $M_{ij}$  the  $ij$  component of matrix  $\mathbf{M}$ . The notation  $\langle \boldsymbol{\xi}, \boldsymbol{\zeta} \rangle$  denotes the Euclidean inner product in the standard basis, and the norm  $\|\boldsymbol{\xi}\|_2$  denotes the Euclidean norm of  $\boldsymbol{\xi}$ . Given a vector  $\mathbf{v} \neq \mathbf{0}$ , we denote  $\hat{\mathbf{v}}$  the unit vector of  $\mathbf{v}$ , i.e.,  $\hat{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|_2$ . If  $\mathbf{v} = \mathbf{0}$  then we define  $\hat{\mathbf{v}}$  to be any unit vector. We denote  $[\theta]_+ = \max\{0, \theta\}$  element-wise. The symbol  $k$  denotes iteration counter, and  $i, j, p, q, s, t$  denote column and/or row indices. We use  $\boldsymbol{\xi}, \boldsymbol{\zeta}$  to denote dummy variable, such as  $\min_{\boldsymbol{\xi}} f(\boldsymbol{\xi})$ .

We introduce new notation when needed.

*Useful tools.* We list two useful tools. The first one is useful for deriving the Euclidean gradient of the objective function.

**Proposition 1** *Let  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $\mathbf{b} \in \mathbb{R}^m$ ,  $\mathbf{c} \in \mathbb{R}^m$ ,  $\mathbf{D} \in \mathbb{R}^{p \times n}$ ,  $\mathbf{e} \in \mathbb{R}^p$  and  $f(\mathbf{x}) = \langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle / \|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2$ . The Euclidean gradient of  $f$  in  $\mathbb{R}^n$ , denoted as  $\nabla f$ , for  $\mathbf{D}\mathbf{x} + \mathbf{e} \neq \mathbf{0}$ , is*

$$\nabla f(\mathbf{x}) = \frac{\|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2^2 \mathbf{A}^\top \mathbf{c} - \langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle \mathbf{D}^\top (\mathbf{D}\mathbf{x} + \mathbf{e})}{\|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2^3}.$$

We put the proof in the appendix.

Next, in this paper, we make use of a useful tensor product formula. Let  $V$  and  $W$  be two vector spaces with inner product. Denote the tensor product of  $\mathbf{v} \in V$  and  $\mathbf{w} \in W$  as  $\mathbf{v} \otimes \mathbf{w} \in V \otimes W$ . Then we have

$$\langle \mathbf{v}, \mathbf{x} \rangle \mathbf{w} = (\mathbf{w} \otimes \mathbf{v})(\mathbf{x}). \quad (1)$$

*Block Coordinate Descent framework.* To solve the Chordal-NMF, we use Block Coordinate Descent as shown in Algorithm 1 with starting point  $(\mathbf{W}_0, \mathbf{H}_0)$ .

To simplify the discussion, we assume  $\mathbf{M} \in \mathbb{R}_+^{m \times n}$  contains no zero columns (zero columns provide no information so we discard them). Next, we assume  $\mathbf{M}$  is pre-processed as  $\mathbf{m}_{:j} = \mathbf{m}_{:j} / \|\mathbf{m}_{:j}\|_2$  so all columns  $\mathbf{m}_{:j}$  have unit  $\ell_2$ -norm and we hide  $\|\mathbf{m}_{:j}\|_2$  in (Chordal-NMF). For compactness we hide the factor  $1/n$  in the subproblem formulation inside Algorithm 1.

*Chordal-NMF is not symmetric.* Classical NMF in Frobenious norm is symmetric:  $\|\mathbf{M} - \mathbf{W}\mathbf{H}\|_F = \|\mathbf{M}^\top - \mathbf{H}^\top \mathbf{W}^\top\|_F$  so we can use the same procedure updating  $\mathbf{H}$  to update  $\mathbf{W}$  (subject to transpose). Chordal-NMF measures the cosine distance between  $\mathbf{m}_{:j}$  and  $\mathbf{W}\mathbf{h}_{:j}$ , not between the rows  $\mathbf{m}_{i:}$  and  $\mathbf{w}_{i:}\mathbf{H}$ , so chordal-NMF is not symmetric. The asymmetric leads to the  $\mathbf{W}$ -subproblem and  $\mathbf{H}$ -subproblem have different structure, and we use different approaches to solve the subproblems. We solve  $\mathbf{H}$ -subproblem column-wise as in lines 3-4 in Algorithm 1, to be discussed in Section 4. We solve  $\mathbf{W}$ -subproblem matrix-wise, to be discussed in Section 5.

**Algorithm 1:** Block Coordinate Descent (BCD) for Chordal-NMF

---

```

1 for  $k = 1, 2, \dots$  do
2   H-subproblem Update  $\mathbf{H}_{k+1}$  using column-wise update
3   for  $j = 1, 2, \dots, n$  do
4     h-subproblem Update  $\mathbf{h}_{:j,k+1} = \operatorname{argmax}_{\mathbf{h} \geq \mathbf{0}} \frac{\langle \mathbf{m}_{:j}, \mathbf{W}\mathbf{h} \rangle}{\|\mathbf{W}\mathbf{h}\|_2}$ .
5   W-subproblem Update
       $\mathbf{W}_{k+1} = \operatorname{argmax}_{\mathbf{W} \geq \mathbf{0}} \left\{ F(\mathbf{W}; \mathbf{H}) = \sum_{j=1}^n \frac{\langle \mathbf{m}_{:j}, \mathbf{W}\mathbf{h}_{:j} \rangle}{\|\mathbf{W}\mathbf{h}_{:j}\|_2} \right\} \Big|_{\mathbf{H}=\mathbf{H}_{k+1}}$ .

```

---

**2 The chordal distance**

In this section, we discuss the motivation and the details of Chordal-NMF.

*Deriving Chordal-NMF from Frobenius NMF.* From the conic view of NMF discussed in the introduction, the geometry of chordal distance can be seen as the Euclidean distance between unit vectors on the unit sphere. Expanding the squared-Frobenius objective  $\|\mathbf{M} - \mathbf{W}\mathbf{H}\|_F^2 = \sum_j \|\mathbf{m}_{:j} - \mathbf{W}\mathbf{h}_{:j}\|_2^2$  gives  $\sum_j \|\mathbf{m}_{:j}\|_2^2 - 2\langle \mathbf{m}_{:j}, \mathbf{W}\mathbf{h}_{:j} \rangle + \|\mathbf{W}\mathbf{h}_{:j}\|_2^2$ , which tells that the pairwise (squared-)Euclidean distance between  $\mathbf{m}_{:j}$  and  $\mathbf{W}\mathbf{h}_{:j}$ , denoted by  $\|\mathbf{m}_{:j} - \mathbf{W}\mathbf{h}_{:j}\|_2^2$ , is contributed by three parts: the inner product  $\langle \mathbf{m}_{:j}, \mathbf{W}\mathbf{h}_{:j} \rangle$  and the sizes  $\|\mathbf{m}_{:j}\|_2$ ,  $\|\mathbf{W}\mathbf{h}_{:j}\|_2$ . If  $\mathbf{m}_{:j}$  and  $\mathbf{W}\mathbf{h}_{:j}$  in this expression have unit  $\ell_2$ -norm, the terms  $\|\mathbf{m}_{:j}\|_2$ ,  $\|\mathbf{W}\mathbf{h}_{:j}\|_2$  vanishes and gives

$$\begin{aligned}
\frac{1}{2}\|\mathbf{M} - \mathbf{W}\mathbf{H}\|_F^2 &= \sum_{j=1}^n \left(1 - \langle \mathbf{m}_{:j}, \mathbf{W}\mathbf{h}_{:j} \rangle\right) \\
&= \sum_{j=1}^n \left(1 - \cos \theta(\mathbf{m}_{:j}, \mathbf{W}\mathbf{h}_{:j})\right),
\end{aligned} \tag{2}$$

where we have used for the Euclidean inner product  $\langle \boldsymbol{\xi}, \boldsymbol{\zeta} \rangle = \|\boldsymbol{\xi}\|_2 \|\boldsymbol{\zeta}\|_2 \cos \theta(\boldsymbol{\xi}, \boldsymbol{\zeta})$  with  $\theta(\boldsymbol{\xi}, \boldsymbol{\zeta})$  denoting the angle between the vectors  $\boldsymbol{\xi}, \boldsymbol{\zeta}$ .

*Quotient space interpretation.* From (2), we define a new objective function as in (Chordal-NMF) that we want to use for measuring the distance between  $\mathbf{m}_{:j}$  and  $\mathbf{W}\mathbf{h}_{:j}$  purely by the angle in-between and disregarding the sizes  $\|\mathbf{m}_{:j}\|_2$ ,  $\|\mathbf{W}\mathbf{h}_{:j}\|_2$ . This is done by the division of  $\|\mathbf{m}_{:j}\|_2 \|\mathbf{W}\mathbf{h}_{:j}\|_2$  in (Chordal-NMF), which can be interpreted by the notion of quotient space. The division of the norm  $\|\mathbf{W}\mathbf{h}_{:j}\|_2$  collapses all the elements in the set  $\{\mathbf{x} : \mathbf{x} = \mathbf{W}\mathbf{h}\}$  to a single point, leading to the chordal distance a purely angle-based distance by ignoring the length information of the vector.

*Haversine interpretation.* The haversine function in navy navigation [7] is defined as  $\text{hav}(\theta) := (1 - \cos \theta)/2$ , where  $\theta = d/r$  is called the central angle, defined as the distance  $d$  between two points along a great circle of the sphere, normalized by the radius of the sphere  $r$ . The expression  $1 - \cos \theta$  in (2) is the haversine distance with  $r = 1$ .

*Sphere interpretation.* The map  $\boldsymbol{\xi} \mapsto \boldsymbol{\xi}/\|\boldsymbol{\xi}\|_2$  in (Chordal-NMF) sends nonzero vector  $\boldsymbol{\xi} \in \mathbb{R}^m$  to the unit sphere in  $\mathbb{R}^m$ . In  $\mathbb{R}^m$  with Euclidean inner product  $\langle \mathbf{u}, \mathbf{v} \rangle$  and norm  $\|\mathbf{u}\|_2 = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$ , the unit sphere and unit ball are

$$\begin{aligned} \mathbb{S}^{m-1} &:= \{\boldsymbol{\xi} \in \mathbb{R}^m \mid \|\boldsymbol{\xi}\|_2 = \sqrt{\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle} = 1\}, \\ \mathbb{B}^m &:= \{\boldsymbol{\xi} \in \mathbb{R}^m \mid \|\boldsymbol{\xi}\|_2 = \sqrt{\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle} \leq 1\}, \end{aligned}$$

resp., which both are subsets of  $\mathbb{R}^m$ . We remark that  $\mathbb{S}^{m-1}$  is nonconvex while  $\mathbb{B}^m$  is convex. Now we define four functions

$$\begin{aligned} \bar{f}_{\text{chord}} &: \mathbb{R}^m \times \mathbb{R}^m \rightarrow [0, 2] \cup \{+\infty\} \cup \{\pm\sqrt{-1}\infty\} \\ &: \mathbf{u}, \mathbf{v} \mapsto \sqrt{2 - 2\langle \mathbf{u}, \mathbf{v} \rangle / (\|\mathbf{u}\|_2 \|\mathbf{v}\|_2)}, \\ \bar{f}_{\text{sq-chord}} &: \mathbb{R}^m \times \mathbb{R}^m \rightarrow [0, 4] \cup \{\pm\infty\} \\ &: \mathbf{u}, \mathbf{v} \mapsto 2 - 2\langle \mathbf{u}, \mathbf{v} \rangle / (\|\mathbf{u}\|_2 \|\mathbf{v}\|_2), \\ f_{\text{chord}} &: \mathbb{S}^{m-1} \times \mathbb{S}^{m-1} \rightarrow [0, 2] \\ &: \mathbf{u}, \mathbf{v} \mapsto \sqrt{2 - 2\langle \mathbf{u}, \mathbf{v} \rangle}, \\ f_{\text{sq-chord}} &: \mathbb{S}^{m-1} \times \mathbb{S}^{m-1} \rightarrow [0, 4] \\ &: \mathbf{u}, \mathbf{v} \mapsto 2 - 2\langle \mathbf{u}, \mathbf{v} \rangle. \end{aligned} \tag{3}$$

Let  $\mathbf{u} \neq \mathbf{0}, \mathbf{v} \neq \mathbf{0}$ . The function  $\bar{f}_{\text{chord}}(\mathbf{u}, \mathbf{v})$  can be viewed as the Euclidean distance between two unit vectors on  $\mathbb{S}^{m-1}$ , see Fig. 1 for the case for  $n = 3$ . Let  $\hat{\mathbf{u}}$  be the unit vector of  $\mathbf{u}$ , obtained by normalizing  $\mathbf{u}$  to unit  $\ell_2$ -norm (known as *pullback*), we have

$$\begin{aligned} \bar{f}_{\text{chord}}(\mathbf{u}, \mathbf{v}) &= \sqrt{\left\langle \frac{\mathbf{u}}{\|\mathbf{u}\|_2} - \frac{\mathbf{v}}{\|\mathbf{v}\|_2}, \frac{\mathbf{u}}{\|\mathbf{u}\|_2} - \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right\rangle} \\ &= \left\| \frac{\mathbf{u}}{\|\mathbf{u}\|_2} - \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right\|_2 = \|\hat{\mathbf{u}} - \hat{\mathbf{v}}\|_2. \end{aligned}$$

There are several undesirable properties for  $\bar{f}_{\text{chord}}$ :

1. At  $\mathbf{u} = \mathbf{0}$  and/or  $\mathbf{v} = \mathbf{0}$ , the function  $\bar{f}_{\text{chord}}$  is undefined, it can take  $\infty$  or even the complex values  $\pm\sqrt{-1}\infty$ . If we restrict the domain of  $\bar{f}_{\text{chord}}$  from  $\mathbb{R}^m$  to  $\mathbb{R}_{++}^m$ , the domain becomes an open set. Optimization problem on open set has no solution.
2. The function  $\bar{f}_{\text{chord}}$  is not differentiable with respect to (wrt.)  $\mathbf{u}, \mathbf{v}$  being unit parallel vector as  $\bar{f}_{\text{chord}}(\hat{\mathbf{u}}, \hat{\mathbf{v}}) := \sqrt{2 - 2} = |0|$ , and absolute value is not differentiable at zero.

3. In the Euclidean case,  $\bar{f}_{\text{chord}}$  is a non-convex function wrt. the variables  $\mathbf{u}$ ,  $\mathbf{v}$  and the pair  $(\mathbf{u}, \mathbf{v})$ .

These undesirable properties lead us to define chordal distance on the unit sphere  $\mathbb{S}^{m-1}$  as follows.

**Definition 1** On  $\mathbb{S}^{m-1}$  with inner product  $\langle \mathbf{u}, \mathbf{v} \rangle$ , we define  $f_{\text{chord}}$  and  $f_{\text{sq-chord}}$  as in (3).

For example, if  $\mathbf{u}$  and  $\mathbf{v}$  are {parallel, anti-parallel, perpendicular}, then  $f_{\text{chord}}(\mathbf{u}, \mathbf{v})$  gives  $\{0, 2, \sqrt{2}\}$ , resp., and  $f_{\text{sq-chord}}(\mathbf{u}, \mathbf{v})$  gives  $\{0, 4, 2\}$  resp..

### 3 Background of optimization on manifold

In this section, we first discuss about nonnegative-constrained optimization problems, then we review manifold optimization. Lastly, we present the Riemannian Multiplicative Update with some analysis.

*Nonnegative-constrained optimization and variants.* Denote  $E$  a linear space (e.g.,  $\mathbb{R}^n, \mathbb{R}^{m \times r}$ ) with an inner product  $\langle \cdot, \cdot \rangle_E$  and an induced norm  $\| \cdot \|_E$ . Given a cost function  $f : E \rightarrow \mathbb{R}$ , consider the minimization problems

$$\begin{aligned} \mathcal{P}_0 &: \underset{\mathbf{x}}{\operatorname{argmin}} f(\mathbf{x}) \text{ s.t. } \mathbf{x} \in \mathcal{M}, \mathbf{x} \in \mathcal{X} \\ \mathcal{P}_1 &: \underset{\mathbf{x} \in \mathcal{M}}{\operatorname{argmin}} f(\mathbf{x}) \text{ s.t. } \mathbf{x} \in \mathcal{X}, & \mathcal{P}_2 &: \underset{\mathbf{x} \in \mathcal{M}}{\operatorname{argmin}} f(\mathbf{x}) + \iota(\mathbf{x}), \\ \mathcal{P}_3 &: \underset{\mathbf{x} \in \mathcal{M}_+}{\operatorname{argmin}} f(\mathbf{x}), & \mathcal{P}_4 &: \underset{\mathbf{x} \in \mathcal{M}}{\operatorname{argmin}} f(\mathbf{x}) + \gamma p(\mathbf{x}). \end{aligned}$$

---


$$\begin{array}{ccccc} \mathcal{P}_4 & \xleftrightarrow{\gamma \text{ large}} & \mathcal{P}_0 & \xleftrightarrow{\text{restriction}} & \mathcal{P}_1 & \xleftrightarrow{\text{indicator}} & \mathcal{P}_2 \\ & & & & \updownarrow & & \\ & & & & \mathcal{M}_+ & & \\ & & & & \mathcal{P}_3 & & \end{array}$$

where the sets  $\mathcal{M} := \{\mathbf{x} \in E \mid h(\mathbf{x}) = 0\}$  and  $\mathcal{X} := \{\mathbf{x} \in E \mid \mathbf{x} \geq \mathbf{0}\}$  are constraints. Here  $\mathcal{X}$  is the nonnegative orthant. The function  $h(\mathbf{x})$  is called the defining function of  $\mathcal{M}$ . We focus on convex compact set  $\mathcal{M}$  being a smooth embedded submanifold of  $\mathbb{R}^n$ , where  $h(\mathbf{x})$  is many-times continuously differentiable. We also assume  $f$  is differentiable.

In Euclidean optimization, we treat  $\mathbf{x} \in \mathcal{M}$  as a constraint in  $(\mathcal{P}_0)$ . We “remove” such constraint by defining a function  $f|_{\mathcal{M}} : \mathcal{M} \rightarrow \mathbb{R}$  as the restriction of  $f$  that the domain of  $f|_{\mathcal{M}}$  is  $\mathcal{M}$ , and here  $f$  is called the *smooth extension* of  $f|_{\mathcal{M}}$  that extends  $\operatorname{dom} f|_{\mathcal{M}}$  from  $\mathcal{M}$  to  $E$ . In this way,  $(\mathcal{P}_0)$  can be written as  $(\mathcal{P}_1)$ , a constrained manifold optimization. Problem  $(\mathcal{P}_1)$  can be converted into a constraint-free problem using indicator. Let  $\iota(\mathbf{x}) : E \rightarrow \mathbb{R} \cup \{+\infty\}$  be the indicator function of the set  $\mathcal{X}$  that, if  $\mathbf{x} \in \mathcal{X}$  then  $\iota(\mathbf{x}) = 0$  and if  $\mathbf{x} \notin \mathcal{X}$  then  $\iota(\mathbf{x}) = +\infty$ , then using  $\iota$  we have  $(\mathcal{P}_2)$ . We can also rewrite  $(\mathcal{P}_1)$  using

$\mathcal{M}_+ := \mathcal{M} \cap \mathcal{X}$  (assumed nonempty), then  $(\mathcal{P}_1)$  becomes  $(\mathcal{P}_3)$ . Lastly, using a (possibly nondifferentiable) penalty function  $p(\mathbf{x}) : E \rightarrow \mathbb{R}$  with a penalty parameter  $\gamma \geq 0$ , we can rewrite  $(\mathcal{P}_0)$  as  $(\mathcal{P}_4)$ . These problems are equivalent (under certain conditions), and solving any one of them will solve the others. We show their relationships next to the problem definition.

### 3.1 Nonnegative-constrained manifold optimization

We note that solving  $\mathcal{P}_0$  directly by manifold techniques is not trivial. In the following paragraphs, we discuss several key points related to this idea. This section also serves as a literature review on manifold optimization.

*Violating the smoothness condition of manifold.* We refer to the term “manifold optimization” as an optimization problem on a smooth (differentiable) manifold. The nonnegative orthant  $\mathcal{X}$ , as a cone, is not a smooth manifold. Furthermore, the intersection  $\mathcal{M}_+ := \mathcal{M} \cap \mathcal{X}$  for  $\mathcal{X}$  as a nonsmooth manifold, where  $\mathcal{M}_+$  is assumed to be nonempty, is also not a smooth manifold, because, for any point  $\mathbf{x}$  in the boundary (the “corner”) of  $\mathcal{M}_+$ , there does not exist a  $\mathbb{R}^n$ -homeomorphic neighborhood in  $\mathcal{M}$ . As a result, several techniques in manifold optimization do not have convergence guarantee for solving problems like  $(\mathcal{P}_0)$  and  $(\mathcal{P}_2)$  as these problems do not belong to manifold optimization.

As a remark, a naive fix of the smoothness issue is to replace the closed set  $\mathcal{X}$  by the open set  $\mathcal{Y} := \mathcal{X} \cap \mathbb{R}_{++}^n$ , i.e., take variable strictly positive instead of nonnegative. See [8, Ch11.6] for some discussion. However, we remark that the optimization problem on the open set has no solution in general.

*Existing methods not applicable to indicator functions.* With variable restricted to be nonnegative, the indicator function  $\iota$  is a nonsmooth (non-differentiable) function from  $E$  to the extended real  $\mathbb{R} \cup \{+\infty\}$ , and such function belongs to the class of convex lower semicontinuous (l.s.c) function [9], making  $(\mathcal{P}_2)$  not satisfying the assumptions in many existing works in nonsmooth Riemannian optimization, e.g. Projected Gradient Descent on Riemannian Manifolds [10], Riemannian proximal gradient [11], and Manifold proximal gradient (ManPG) [12].

*Penalty methods will possibly produce non-strictly feasible points.* There are several ways to choose the penalty function  $p(\mathbf{x})$  for nonnegativity constraints.

- **Finite nonsmooth penalty.** Instead of the possibly infinite-valued indicator function in  $(\mathcal{P}_2)$ , it is possible to use a finite nonsmooth penalty  $p(\mathbf{x}) = \|\max\{\mathbf{0}, -\mathbf{x}\}\|_1$  where  $\max$  is taken element-wise and  $\|\cdot\|_1$  is  $\ell_1$ -norm. In this way, we can apply ManPG for solving  $(\mathcal{P}_4)$ . However, ManPG has the following subproblem [12, Eq.4.3]

$$\mathbf{v}_{k+1} = \underset{\mathbf{v}}{\operatorname{argmin}} \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle_E + \frac{\|\mathbf{v}\|_E^2}{2t} + \gamma p(\mathbf{x}_k + \mathbf{v}) \quad \text{s.t.} \quad \mathbf{v} \in T_{\mathbf{x}_k} \mathcal{M}, \quad (4)$$



where  $T_{\mathbf{x}_k} \mathcal{M}$  denotes the tangent space of  $\mathcal{M}$  at point  $\mathbf{x}_k$ , to be explained later in this section. In short, since solving (4) is possibly expensive, thus the per-iteration cost of ManPG is high and therefore ManPG is not suitable for Chordal-NMF.

- **Smoothing.** We can choose a smooth penalty based on smoothing such as  $p(\mathbf{x}) = \|(\max\{\mathbf{0}, -\mathbf{x}\})^2\|_1$ , where  $(\cdot)^2$  is taken element-wise. In this way, we can use Riemannian optimization techniques to solve  $(\mathcal{P}_4)$ . However, we do not consider smoothing in this paper due to speed issues.
- **Nonsmooth penalty with IRLS.** Iterative reweighted least squares (IRLS) [13] is another approach to deal with nonsmooth penalty term such as  $p(\mathbf{x}) = \|\max\{\mathbf{0}, -\mathbf{x}\}\|_1$ . This is based on the variational inequality of absolute value that  $|w| = \underset{\eta \geq 0}{\operatorname{argmin}} (w^2/\eta + \eta)/2$ . Hence, we have the following equivalent expression of the nonsmooth penalty

$$p(\mathbf{x}) = \|\max\{\mathbf{0}, -\mathbf{x}\}\|_1 = \sum_i \frac{|x_i| - x_i}{2} = \underset{\eta_i \geq 0}{\operatorname{argmin}} \frac{1}{2} \sum_i \left( \frac{1}{2} \left( \frac{x_i^2}{\eta_i} + \eta_i \right) - x_i \right).$$

However, IRLS suffers from the same disadvantage of smoothing that the resultant update often has a slower convergence than directly updating the optimization variable.

*Euclidean projected gradient descent is infeasible.* A naive idea to deal with the optimization problem is to ignore the manifold and solve  $\mathcal{P}_0$  as it is using Euclidean optimization methods such as the projected gradient descent. However such approach is infeasible: in the update, after we performed a gradient descent step, we project the variable onto the intersection set  $\mathcal{M}_+ := \mathcal{X} \cap \mathcal{M}$  as  $\mathbf{x}_{k+1} = \operatorname{proj}_{\mathcal{M}_+}(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))$ , in which the project subproblem  $\operatorname{proj}_{\mathcal{M}_+}(\mathbf{z}) = \underset{\mathbf{x} \in \mathcal{M}_+}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|_E^2$  is general hard that there is no closed-form solution, or the problem is expensive to solve. Generally  $\operatorname{proj}_{\mathcal{M}_+}$  is a composition

$$\operatorname{proj}_{\mathcal{M}_+} = \underbrace{\operatorname{proj}_{\mathcal{M}} \circ \operatorname{proj}_{\mathcal{X}} \circ \operatorname{proj}_{\mathcal{M}} \circ \cdots \circ \operatorname{proj}_{\mathcal{X}}}_{\text{possibly many}},$$

thus Euclidean projected gradient descent is infeasible due to a very high per-iteration cost.

*Projection-free method is expensive.* The in-feasibility of  $\operatorname{proj}_{\mathcal{M}_+}$  motivates the use of projection-free method, such as the Riemannian Frank-Wolfe (RFW) [14]. The core of RFW is to solve a subproblem, expensive for our purpose, that is in the form  $\underset{\mathbf{z} \in \mathcal{M}_+}{\operatorname{argmin}} \left\langle \operatorname{grad} f(\mathbf{x}_k), \operatorname{Exp}_{\mathbf{x}_k}^{-1}(\mathbf{z}) \right\rangle_E$ , since it requires the evaluation of an exponential map, followed by a computation on the geodesic. In this paper we provide a projection-free method that only requires the computation of the Riemannian gradient part  $\operatorname{grad} f$  without the exponential map.

*Dual methods are slow and primal sequence is not strictly feasible.* It is natural to solve  $(\mathcal{P}_0)$  by Riemannian Augmented Lagrangian multiplier [15] or to solve  $(\mathcal{P}_2)$  using Riemannian ADMM. However, dual approaches have drawbacks. First, they have additional parameters to tune. Second, the sequence generated by these methods is not strictly-primal-feasible: the primal variable is only feasible when a sufficiently large number of iterations is carried out (in principle, after infinitely many iterations). Since dual approaches are well-known for having a slow convergence, this makes their per-iteration cost very high and therefore they are not feasible for our application. Furthermore, it is dangerous to use an infeasible variable to update the other block of variables in the BCD framework.

*On fractional programming.* Fractional Programming (FP) [16,17] (see the monograph [18] for a modern treatment) can also be used to solve the sub-problems in Chordal-NMF. The idea of FP is to convert the optimization problem in the form  $\max_x f(x)/g(x)$  on the fraction of two functions  $f, g$ , to the “linearized form”  $\max_{x,y} f(x) - yg(x)$ . We do not consider fractional program in this work because (i) it introduces another variable  $y$ , and (ii) the function  $f(x) - yg(x)$  conversion for our application (Chordal-NMF) is non-convex.

*Our solution: Riemannian multiplicative update.* To tackle these technical issues, we propose a cost-effective Riemannian gradient descent (RGD) method for solving Problem  $(\mathcal{P}_1)$ . The idea is to perform a RGD with a special stepsize that guarantees the feasibility of the variable. Our approach is motivated by the research of NMF on multiplicative update, see [1, Section 8.2] for details. There are advantages of this method:

- The expensive projection  $\text{proj}_{\mathcal{M}_+}$  discussed above is not required.
- Unlike the dual approaches, if the initial variable in the algorithm is feasible, the whole sequence is guaranteed to be feasible, see Proposition 2.
- RMU allows the tools from Riemannian optimization to be utilized, there is no violation of the smoothness in the method.

We introduce RMU in Section 3.3, before that we review the background of Riemannian optimization below.

### 3.2 Background of Riemannian optimization

Riemannian optimization, or manifold optimization, has a long history [19, 20, 21, 22, 23, 24, 25, 8]. Such rich literature makes it impossible to review all the content. Hence we give the absolute minimum prerequisite on manifold optimization for the paper.

*Submanifold and ambient space.* For smooth function  $h : E \rightarrow \mathbb{R}^k$ , the set  $\mathcal{M} := \{\mathbf{x} \in E \mid h(\mathbf{x}) = \mathbf{0}_k\}$  is an embedded submanifold of  $E$  of dimension  $\dim E - k$ . In this paper we focus on  $h$  in the form  $h(\mathbf{x}) = \langle \cdot, \cdot \rangle_E - k$ . As a preview, in the  $h$ -subproblem,  $\mathcal{M}$  is the “shell” of an ellipsoid, it has  $k(=1)$  dimension lower than its ambient Euclidean space (which is  $\mathbb{R}^r$ ). In the  $W$ -subproblem,  $\mathcal{M}$  is the “shell” of a twisted spectrahedron. For a manifold  $\mathcal{M}$  in  $E$ , we call  $E$  the ambient space of  $\mathcal{M}$ . In this work, all ambient spaces of  $\mathcal{M}$  are some specific linear vector space  $E$ , such as  $\mathbb{R}^r$  (for the  $h$  subproblem) and  $\mathbb{R}^{m \times r}$  (for the  $W$  subproblem).

*Tangent space and projection.* Let the differential of  $h$   $Dh(\mathbf{x})[\mathbf{v}]$  defined as  $Dh(\mathbf{x})[\mathbf{v}] = \langle \text{grad}h(\mathbf{x}), \mathbf{v} \rangle_E$ . Let  $\ker$  denote the kernel of a matrix in linear algebra, the tangent space of  $\mathcal{M}$  at a reference point  $\mathbf{x}$ , denoted as  $T_{\mathbf{x}}\mathcal{M}$ , is defined as the kernel of  $Dh(\mathbf{x})$ , i.e.,

$$T_{\mathbf{x}}\mathcal{M} := \ker Dh(\mathbf{x}) = \{\mathbf{v} \in E \mid \langle \text{grad}h(\mathbf{x}), \mathbf{v} \rangle_E = 0\}.$$

The set  $T_{\mathbf{x}}\mathcal{M}$  refers to the collection of vectors  $\mathbf{v} \in E$  that is tangent to  $\mathcal{M}$  at  $\mathbf{x}$ . The orthogonal projection  $\text{proj}_{T_{\mathbf{x}}\mathcal{M}} : E \rightarrow T_{\mathbf{x}}\mathcal{M}$ , is defined based on orthogonal decomposition of a vector space as  $\mathbf{v} = \text{proj}_{T_{\mathbf{x}}\mathcal{M}}(\mathbf{v}) + Dh(\mathbf{x})^*[\alpha]$ , where  $Dh(\mathbf{x})^*[\alpha]$  is the adjoint of  $Dh(\mathbf{x})[\mathbf{v}]$ , and  $\alpha \in \mathbb{R}$  plays the role of the dual variable (with a technical name covector) as the unique solution to  $\alpha = \text{argmin}_{\alpha \in \mathbb{R}} \|\mathbf{v} - Dh(\mathbf{x})^*[\alpha]\|_E^2 = (Dh(\mathbf{x})^*)^\dagger[\mathbf{v}]$ , where  $\dagger$  is pseudo-inverse. This gives an explicit expression  $\text{proj}_{T_{\mathbf{x}}\mathcal{M}}(\mathbf{v}) = \mathbf{v} - Dh(\mathbf{x})^*[(Dh(\mathbf{x})^*)^\dagger[\mathbf{v}]]$ . We remark that if the tangent space  $\text{proj}_{T_{\mathbf{x}}\mathcal{M}}$  is equals to the ambient space, then the projection is not necessary.

*Retraction.* A point  $\mathbf{x}$  that originally sitting on a manifold  $\mathcal{M}$  may leave outside  $\mathcal{M}$  after a gradient operation, hence certain operation is required to pull the point back onto  $\mathcal{M}$ . This can be achieved by a particular smooth map, known as the retraction  $\mathcal{R} : T_{\mathbf{x}}\mathcal{M} \rightarrow \mathcal{M}$  that maps a point on the tangent space  $T_{\mathbf{x}}\mathcal{M}$  onto  $\mathcal{M}$ . Different retractions can be found in the literature, such as exponential map and metric retractions [8, Section 3.6]. However, in the general case, to obtain the exponential map of a manifold  $\mathcal{M}$  one requires to solve a differential equation. Therefore, for computational efficiency reasons, we consider another approach to pull the point back to the manifold  $\mathcal{M}$  defined by  $h$ , known as the metric retraction  $\mathcal{R}_{\mathbf{x}}(\mathbf{v}) = \text{argmin}_{\mathbf{y} \in E} \|\mathbf{x} + \mathbf{v} - \mathbf{y}\|_E^2$  s.t.  $h(\mathbf{y}) = 0$ .

Note that  $\mathcal{R}_{\mathbf{x}}(\mathbf{v})$  is possibly non-unique and possibly hard to compute. Among all the possible retractions we focus on the following *metric retraction*:

$$\mathcal{R}_{\mathbf{x}}(\mathbf{v}) := \frac{\mathbf{x} + \mathbf{v}}{\|\mathbf{x} + \mathbf{v}\|_E}. \quad (\text{Metric retraction})$$

*Restriction and smooth extension.* Given a function  $f : E \rightarrow \mathbb{R}$ , the function  $f|_{\mathcal{M}} : \mathcal{M} \rightarrow \mathbb{R}$  is called the *restriction* of  $f$ . The function  $f|_{\mathcal{M}}$  is obtained by restricting the domain of  $f$  from  $\mathbb{R}^n$  to  $\mathcal{M}$ . Given a function  $g : \mathcal{M} \rightarrow \mathbb{R}$ , the function  $\bar{g} : \mathcal{U} \rightarrow \mathbb{R}$  is called the *smooth extension* of  $g$ , where  $\mathcal{U} \subset E$ . The function  $\bar{g}$  is obtained by defining  $g$  in a neighborhood  $\mathcal{U}$  of  $\mathcal{M}$ . We can see that restriction and extension are “inverse” of each other.

*Riemannian gradient and Euclidean gradient.* Given a function  $f|_{\mathcal{M}} : \mathcal{M} \rightarrow \mathbb{R}$  with its smooth extension  $f : E \rightarrow \mathbb{R}$ , let  $\nabla f(\mathbf{x})$  denotes the Euclidean gradient of  $f$  in the standard Euclidean basis at a point  $\mathbf{x} \in E$ , the Riemannian gradient of  $f|_{\mathcal{M}}$  defined on  $\mathcal{M}$ , denoted as  $\text{grad}f|_{\mathcal{M}}$ , at a point  $\mathbf{z}$ , wrt. the reference point  $\mathbf{x}$ , is the Euclidean gradient  $\nabla f(\mathbf{z})$  projected onto the tangent space  $T_{\mathbf{x}}\mathcal{M}$ . That is  $\text{grad}f|_{\mathcal{M}}(\mathbf{z}) = \text{proj}_{T_{\mathbf{x}}\mathcal{M}}\nabla f(\mathbf{z})$ . We note that the “complete” Riemannian gradient is computed using metric tensor. Let  $g$  be the metric tensor of  $\mathcal{M}$  and let  $G_{\mathbf{x}}$  denotes the matrix representation of the  $g$  in coordinates, then  $\text{grad}f|_{\mathcal{M}}(\mathbf{z}) = G_{\mathbf{x}}^{-1}\text{proj}_{T_{\mathbf{x}}\mathcal{M}}\nabla f(\mathbf{z})$ . In this work we do not consider metric tensor for the sake of cheap computational cost.

For simplifying the notation, we usually write  $f|_{\mathcal{M}}$  as  $f$  taking the natural extension.

### 3.3 Riemannian Multiplicative Update (RMU)

In this work, we propose a Riemannian Multiplicative Update (RMU) for solving Chordal-NMF in the form of  $(\mathcal{P}_3)$ . At the core, RMU is a special kind of Riemannian gradient descent (RGD) with the update  $\mathbf{x}_{k+1} = \mathcal{R}_{\mathbf{x}_k}(\alpha\mathbf{v}_k)$ , where  $\mathbf{v}_k = -\text{grad}f(\mathbf{x}_k)$  is the (negative) Riemannian gradient of  $f$  at  $\mathbf{x}_k$ . It is important to note that RMU is a projection-free method for  $(\mathcal{P}_3)$  which contains the nonnegativity constraint. RMU guarantees the nonnegativity of  $\mathbf{x}$  by selecting a stepsize  $\alpha$  such that  $\mathcal{R}_{\mathbf{x}_k}(\alpha\mathbf{v}_k)$  stays within  $\mathcal{X}$ . In the following, we first review euclidean multiplicative update (MU) and then we generalize its variant in the Riemannian case.

*Euclidean MU.* MU was first proposed in [26,27,28], and during years has gained popularity in NMF [1]. MU can be done via a “element-wise sign decomposition”: Euclidean gradient can be written as  $\nabla f(\mathbf{x}) = \nabla^+ f(\mathbf{x}) - \nabla^- f(\mathbf{x})$ , where  $\nabla^+ f(\mathbf{x}) \geq \mathbf{0}$  and  $\nabla^- f(\mathbf{x}) \geq \mathbf{0}$ . Let  $\odot$  be element-wise product and  $\oslash$  be element-wise division, for solving a nonnegative-constrained Euclidean optimization problem, the MU step has the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k \odot \nabla^- f(\mathbf{x}_k) \oslash \nabla^+ f(\mathbf{x}_k), \quad (5)$$

which is obtained by choosing an element-wise stepsize  $\alpha = \mathbf{x}_k \oslash \nabla^+ f(\mathbf{x}_k)$  in the Euclidean gradient descent step  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$ , see [29] for the derivation. We remark that in the literature the MU update (5) is frequently written as  $\mathbf{x}_{k+1} = \mathbf{x}_k \frac{\nabla^- f(\mathbf{x}_k)}{\nabla^+ f(\mathbf{x}_k)}$ . We do not use this convention because it confuses with metric retraction for our purpose.

*RMU.* Now we generalize MU to the Riemannian case. First, a Riemannian gradient  $\text{grad}f(\mathbf{x})$  always admits the element-wise sign decomposition  $\text{grad}f(\mathbf{x}) = \text{grad}^+f(\mathbf{x}) - \text{grad}^-f(\mathbf{x})$ , where  $\text{grad}^+f(\mathbf{x}_k)$  is the positive part of the Riemannian gradient of  $f$ . This sign decomposition holds since for any object  $\mathbf{b}$ , it can be written as  $\mathbf{b} = \mathbf{b}^+ - \mathbf{b}^-$  with  $\mathbf{b}^+ = \max(\mathbf{0}, \mathbf{b})$  and  $\mathbf{b}^- = \max(\mathbf{0}, -\mathbf{b})$ .

We generalize (5) to the Riemannian case as follows, where we prove that the nonnegativity of the update is preserved by RMU without projection (Proposition 2). The key of the proof is metric retraction and a particular choice of the element-wise step-size  $\alpha$  that acts component-wise on the update direction  $\mathbf{v}_k$ .

**Proposition 2 (Riemannian Multiplicative Update)** *Denote  $\mathbf{v}_k$  the anti-parallel direction of the Riemannian gradient of a manifold  $\mathcal{M}$  at  $\mathbf{x}_k$  by the expression  $\mathbf{v}_k = -\text{grad}f(\mathbf{x}_k)$ , and let  $\mathcal{R}_{\mathbf{x}_k}$  the metric retraction onto  $\mathcal{M}$ . If a nonnegative  $\mathbf{x}_k$  is updated by RGD step  $\mathbf{x}_{k+1} = \mathcal{R}_{\mathbf{x}_k}(\alpha \odot \mathbf{v}_k)$  with an element-wise stepsize  $\alpha \in E$  defined as  $\alpha = \mathbf{x}_k \oslash \text{grad}^+f(\mathbf{x}_k)$ , then  $\mathbf{x}_{k+1}$  is nonnegative and is on  $\mathcal{M}$ .*

*Proof* The term  $\alpha \odot (-\text{grad}f(\mathbf{x}_k))$  with  $\alpha = \mathbf{x}_k \oslash \text{grad}^+f(\mathbf{x}_k)$  can be computed as

$$\begin{aligned} -\alpha \odot \text{grad}f(\mathbf{x}_k) &= -(\mathbf{x}_k \oslash \text{grad}^+f(\mathbf{x}_k)) \odot \text{grad}f(\mathbf{x}_k) \\ &= (\mathbf{x}_k \oslash \text{grad}^+f(\mathbf{x}_k)) \odot (\text{grad}^-f(\mathbf{x}_k) - \text{grad}^+f(\mathbf{x}_k)) \\ &= \mathbf{x}_k \odot \text{grad}^-f(\mathbf{x}_k) \oslash \text{grad}^+f(\mathbf{x}_k) - \mathbf{x}_k. \end{aligned}$$

The element-wise operations as linear transformations preserve the tangent condition of a manifold into a point [30]. Now apply (Metric retraction) on  $-\alpha \odot \text{grad}f(\mathbf{x}_k)$ , the  $\mathbf{x}_k$  terms got canceled and gives

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathcal{R}_{\mathbf{x}_k}(-\alpha \odot \text{grad}f(\mathbf{x}_k)) \\ &= \frac{\mathbf{x}_k - \alpha \odot \text{grad}f(\mathbf{x}_k)}{\|\mathbf{x}_k - \alpha \odot \text{grad}f(\mathbf{x}_k)\|_E} \\ &= \frac{\mathbf{x}_k \odot \text{grad}^-f(\mathbf{x}_k) \oslash \text{grad}^+f(\mathbf{x}_k)}{\|\mathbf{x}_k \odot \text{grad}^-f(\mathbf{x}_k) \oslash \text{grad}^+f(\mathbf{x}_k)\|_E}. \end{aligned}$$

The numerator is nonnegative since  $\text{grad}^+f$  and  $\text{grad}^-f$  are nonnegative by definition, the denominator is nonnegative and  $\mathbf{x}_k$  is nonnegative by assumption, therefore the updated point  $\mathbf{x}_{k+1}$  is nonnegative.

Proposition 2 besides telling that RMU is projection-free for nonnegative-constrained manifold optimization, it also gives a convergence condition for  $\text{grad}f(\mathbf{x}_k) = \mathbf{0}$ , which is equivalent to  $\text{grad}^-f(\mathbf{x}_k) = \text{grad}^+f(\mathbf{x}_k)$ . That is, we have a simple way to check the convergence of the sequence  $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$  by comparing  $\text{grad}^-f(\mathbf{x}_k)$ ,  $\text{grad}^+f(\mathbf{x}_k)$  if the explicit gradient expressions are available.

As a remark, note that RMU is a projection-free method that does not require the additional computation of the exponential map and the subsequent steps in RFW that we reviewed in Section 3.1, and therefore has a lower per-iteration computational cost, making it suitable for Chordal-NMF.

We are now ready to move on to the explicit update of  $\mathbf{H}$  and  $\mathbf{W}$  in the next two sections.

#### 4 Column-wise $\mathbf{H}$ -subproblem over ellipsoid

In this section, we discuss how to solve (h-manifold-subproblem) in Algorithm 1. We consider the argmin form as

$$\mathbf{h}_{:j,k+1} = \underset{\mathbf{h}}{\operatorname{argmin}} \left\{ \phi(\mathbf{h}) := \left( 1 - \frac{\langle \mathbf{m}_{:j}, \mathbf{W}\mathbf{h} \rangle}{\|\mathbf{W}\mathbf{h}\|_2} \right) \right\} \text{ s.t. } \mathbf{h} \geq \mathbf{0}. \quad (\text{h-subproblem})$$

The denominator in  $\phi$  is problematic if  $\mathbf{W}\mathbf{h} = \mathbf{0}$ . We get rid of this denominator based on the discussion in Section 2 as follows. We introduce a constraint  $\mathbf{W}\mathbf{h} \in \mathbb{S}^{m-1}$  and a function  $\phi|_{\mathbf{W}\mathbf{h} \in \mathbb{S}^{m-1}}$ . This gives an equivalent problem

$$\underset{\mathbf{h}}{\operatorname{argmin}} \left\{ \phi|_{\mathbf{W}\mathbf{h} \in \mathbb{S}^{m-1}}(\mathbf{W}\mathbf{h}) := 1 - \langle \mathbf{m}_{:j}, \mathbf{W}\mathbf{h} \rangle \right\} \text{ s.t. } \mathbf{h} \geq \mathbf{0}.$$

*From sphere to ellipsoid* Instead of working on  $\phi|_{\mathbf{W}\mathbf{h} \in \mathbb{S}^{m-1}}$  that is defined on  $\mathbf{W}\mathbf{h} \in \mathbb{S}^{m-1}$ , we consider working on the problem over ellipsoid. Let  $\mathbf{W}$  has full rank, we define an ellipsoid  $\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1} \subset \mathbb{R}^r$  by a Positive Definite (PD) matrix  $\mathbf{W}^\top \mathbf{W} \in \mathbb{R}^{r \times r}$  as

$$\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1} := \left\{ \boldsymbol{\xi} \in \mathbb{R}^r \mid \langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_{\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}} := \langle \mathbf{W}^\top \mathbf{W} \boldsymbol{\xi}, \boldsymbol{\xi} \rangle = 1 \right\} \subset \mathbb{R}^r, \quad (\text{Ellipsoid Manifold})$$

where the weighted inner product is defined as  $\langle \boldsymbol{\xi}, \boldsymbol{\zeta} \rangle_{\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}} := \langle \mathbf{W}^\top \mathbf{W} \boldsymbol{\xi}, \boldsymbol{\zeta} \rangle$  and the induced norm  $\|\boldsymbol{\xi}\|_{\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}}^2 = \langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_{\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}}$ .

Let  $R$  be the length of the semi-axes of  $\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}$ , then the principal axes of  $\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}$  and the value  $1/R^2$  are given by the eigenvectors and the corresponding eigenvalues of  $\mathbf{W}^\top \mathbf{W}$ , resp. [31]. Now we arrive at a problem wrt.  $\mathbf{h}$  on the ellipsoid  $\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}$ .

To ease notation, sometimes we write  $\langle \boldsymbol{\xi}, \boldsymbol{\zeta} \rangle_{\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}}$  as  $\langle \boldsymbol{\xi}, \boldsymbol{\zeta} \rangle_{\mathcal{E}}$  and  $\|\boldsymbol{\xi}\|_{\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}}$  as  $\|\boldsymbol{\xi}\|_{\mathcal{E}}$ .

*Optimization over manifold.* We consider the following re-formulation of the h-subproblem

$$\begin{aligned} & \underset{\mathbf{h}}{\operatorname{argmin}} \left\{ \phi(\mathbf{h}) := 1 - \langle \mathbf{m}_{:j}, \mathbf{W}\mathbf{h} \rangle \right\} \\ & \text{s.t. } \mathbf{h} \in \mathcal{H} := \mathbb{R}_+^r \cap \mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1} \subset \mathcal{M} := \mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}. \end{aligned} \quad (\text{h-manifold-subproblem})$$

Here the optimization variable  $\mathbf{h}$  is constrained to be inside the set  $\mathcal{H}$  which takes the nonnegativity and the ellipsoid  $\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}$  into account. We remark that:

- If we perform the smooth extension of  $\phi$  from  $\mathcal{M}$  to  $\mathbb{R}^r$ , and take  $\mathbf{h} \geq \mathbf{0}$ , we go back to (h-subproblem) in Algorithm 1, under a sign change.
- Being a subset of  $\mathbb{R}^r$ , the set  $\mathcal{M} := \mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}$  is a smooth submanifold [8, Def 3.10] and we can show that the inner product  $\langle \boldsymbol{\xi}, \boldsymbol{\zeta} \rangle_{\mathcal{E}}$  is a Riemannian metric [8, Proposition 3.54].
- The subset  $\mathcal{H}$  of  $\mathcal{M}$  is a *nsmooth* manifold. The set  $\mathcal{H}$  is constructed as the intersection of the closed set  $\mathbb{R}_+^r$  with  $\mathcal{M}$ . The set  $\mathbb{R}_+^r$  has sharp corners at the boundary and thus it is not smooth. See Section 3 for our discussion on the issues caused by such non-smoothness.
- Strictly speaking, the function  $\phi$  in the h-subproblem and the function  $\phi$  in (h-manifold-subproblem) have different domains and therefore they are not the same function, here we have abused the notation that if  $\mathbf{h} \in \mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}$  then the denominator  $\|\mathbf{W}\mathbf{h}\|_2 = 1$  disappears.

#### 4.1 Tools on ellipsoid manifold

In this section, we summarize the tools to solve (h-manifold-subproblem) by RMU, collected in Table 1 and detailed in the following paragraphs.

**Table 1** Summary of mathematical objects for Riemannian optimization

Name / Reference	Definition / expression
Ellipsoid manifold of $\mathbf{h}$ (Ellipsoid Manifold)	$\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1} := \{ \boldsymbol{\xi} \in \mathbb{R}^r \mid \langle \mathbf{W}^\top \mathbf{W} \boldsymbol{\xi}, \boldsymbol{\xi} \rangle = 1 \}$
Tangent space of $\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}$ at $\boldsymbol{\zeta}$ Definition 2	$T_{\boldsymbol{\zeta}} \mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1} := \{ \boldsymbol{\xi} \in \mathbb{R}^r \mid \langle \mathbf{W}^\top \mathbf{W} \boldsymbol{\xi}, \boldsymbol{\zeta} \rangle = 0 \}$
Project $\boldsymbol{\xi}$ onto $T_{\boldsymbol{\zeta}} \mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}$ Proposition 3	$\text{proj}_{T_{\boldsymbol{\zeta}} \mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}}(\boldsymbol{\xi}) = \left( \mathbf{I}_r - \frac{(\mathbf{W}^\top \mathbf{W} \boldsymbol{\zeta})^{\otimes 2}}{\ \mathbf{W}^\top \mathbf{W} \boldsymbol{\zeta}\ _2^2} \right) \boldsymbol{\xi}$
Retraction Proposition 4	$\mathcal{R}_{\boldsymbol{\zeta}}(\boldsymbol{\xi}) = (\boldsymbol{\zeta} + \boldsymbol{\xi}) / \ \boldsymbol{\zeta} + \boldsymbol{\xi}\ _{\mathcal{E}}$

*Tangent space.* For compact notation, we use  $\mathbf{A} := \mathbf{W}^\top \mathbf{W}$ . On  $\mathcal{E}_{\mathbf{A}}$ , we define the tangent space as follows.

**Definition 2** Let function  $h(\boldsymbol{\xi}) = \langle \mathbf{A} \boldsymbol{\xi}, \boldsymbol{\xi} \rangle - 1$  defines  $\mathcal{E}_{\mathbf{A}}^{r-1}$ , its differential is  $Dh(\boldsymbol{\xi})[\boldsymbol{\zeta}] = 2\langle \mathbf{A} \boldsymbol{\xi}, \boldsymbol{\zeta} \rangle$ . The tangent space of  $\mathcal{E}_{\mathbf{A}}^{r-1}$  at a reference point  $\boldsymbol{\zeta} \in \mathcal{E}_{\mathbf{A}}^{r-1}$ , denoted as  $T_{\boldsymbol{\zeta}} \mathcal{E}_{\mathbf{A}}^{r-1}$ , is  $T_{\boldsymbol{\zeta}} \mathcal{E}_{\mathbf{A}}^{r-1} := \{ \boldsymbol{\xi} \in \mathbb{R}^r \mid \langle \mathbf{A} \boldsymbol{\zeta}, \boldsymbol{\xi} \rangle = 0 \} = \text{Ker } Dh(\boldsymbol{\zeta})$ .

*Projection onto tangent space.* We define the adjoint operator of  $Dh(\boldsymbol{\xi})[\zeta]$  as  $Dh(\boldsymbol{\xi})^*[\alpha] = 2\alpha\mathbf{A}\boldsymbol{\zeta}$ , where  $\alpha$  takes the value  $\bar{\alpha} = \frac{1}{2} \frac{\langle \mathbf{A}\boldsymbol{\zeta}, \boldsymbol{\xi} \rangle}{\|\mathbf{A}\boldsymbol{\zeta}\|_2^2}$  obtained by solving the least squares  $\bar{\alpha} = \operatorname{argmin}_{\alpha \in \mathbb{R}} \|\boldsymbol{\xi} - Dh(\boldsymbol{\zeta})^*[\alpha]\|_2^2$ . Now by  $Dh(\boldsymbol{\xi})^*[\alpha]$ ,  $\bar{\alpha}$ , we have the projector  $\operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}}(\boldsymbol{\xi}) = \boldsymbol{\xi} - Dh(\boldsymbol{\zeta})^*[\bar{\alpha}]$  as shown in the following proposition.

**Proposition 3** *The projector  $\operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}} : \mathbb{R}^r \rightarrow T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}$  of  $\boldsymbol{\xi}$  onto the tangent space at the reference point,  $\boldsymbol{\zeta}$  is*

$$\begin{aligned} \operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}}(\boldsymbol{\xi}) &= \boldsymbol{\xi} - \frac{\langle \mathbf{A}\boldsymbol{\zeta}, \boldsymbol{\xi} \rangle}{\|\mathbf{A}\boldsymbol{\zeta}\|_2^2} \mathbf{A}\boldsymbol{\zeta} \\ &= \left( \mathbf{I}_r - \frac{(\mathbf{A}\boldsymbol{\zeta}) \otimes (\mathbf{A}\boldsymbol{\zeta})}{\|\mathbf{A}\boldsymbol{\zeta}\|_2^2} \right) \boldsymbol{\xi} = \left( \mathbf{I}_r - \frac{(\mathbf{A}\boldsymbol{\zeta})^{\otimes 2}}{\|\mathbf{A}\boldsymbol{\zeta}\|_2^2} \right) \boldsymbol{\xi}. \end{aligned} \quad (\text{proj-ellips})$$

*Proof* In (proj-ellips), the first equality is by definition of orthogonal projection, the last two equalities are based on tensor product  $\otimes$  and (1), where  $\mathbf{x}^{\otimes 2}$  denotes  $\mathbf{x} \otimes \mathbf{x}$ . Now we show (proj-ellips) satisfies the three conditions in [8, Def. 3.60]. First, the range of  $\operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}}$ , denoted as  $\operatorname{Im}(\operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}})$ , is exactly  $T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}$ , since

$$\boldsymbol{\xi} = \left( \mathbf{I}_r - \frac{(\mathbf{A}\boldsymbol{\zeta})^{\otimes 2}}{\|\mathbf{A}\boldsymbol{\zeta}\|_2^2} \right) \boldsymbol{\xi} \iff -\frac{\langle \mathbf{A}\boldsymbol{\zeta}, \boldsymbol{\xi} \rangle}{\|\mathbf{A}\boldsymbol{\zeta}\|_2^2} \mathbf{A}\boldsymbol{\zeta} = \mathbf{0} \iff \boldsymbol{\xi} \in T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}.$$

Next, we show the orthogonality of  $\operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}}$ . First, we have

$$\langle \boldsymbol{\xi} - \operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}}(\boldsymbol{\xi}), \boldsymbol{\xi} \rangle \stackrel{(\text{proj-ellips})}{=} \left\langle \boldsymbol{\xi} - \left( \mathbf{I}_r - \frac{(\mathbf{A}\boldsymbol{\zeta})^{\otimes 2}}{\|\mathbf{A}\boldsymbol{\zeta}\|_2^2} \right) \boldsymbol{\xi}, \boldsymbol{\xi} \right\rangle,$$

cancelling  $\boldsymbol{\xi}$  gives  $\frac{1}{\|\mathbf{A}\boldsymbol{\zeta}\|_2^2} \langle (\mathbf{A}\boldsymbol{\zeta})^{\otimes 2} \boldsymbol{\xi}, \boldsymbol{\xi} \rangle$ . Now apply a tensor product trick  $\langle \mathbf{a}, (\mathbf{b} \otimes \mathbf{b}) \mathbf{c} \rangle \stackrel{(1)}{=} \langle \mathbf{a}, \langle \mathbf{b}, \mathbf{c} \rangle \mathbf{b} \rangle = \langle \mathbf{b}, \mathbf{c} \rangle \langle \mathbf{a}, \mathbf{b} \rangle$  gives  $\frac{1}{\|\mathbf{A}\boldsymbol{\zeta}\|_2^2} \langle \mathbf{A}\boldsymbol{\zeta}, \boldsymbol{\xi} \rangle \langle \mathbf{A}\boldsymbol{\zeta}, \boldsymbol{\xi} \rangle \stackrel{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}}{=} 0$  so the proof of orthogonality is finished.

Lastly, we show that  $\operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}}$  is idempotent, i.e.,

$$\operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}} \left( \operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}}(\boldsymbol{\xi}) \right) = \operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}}(\boldsymbol{\xi}).$$

First  $\operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}} \left( \operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}}(\boldsymbol{\xi}) \right) \stackrel{(\text{proj-ellips})}{=} \operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}} \left( \left( \mathbf{I}_r - \frac{(\mathbf{A}\boldsymbol{\zeta})^{\otimes 2}}{\|\mathbf{A}\boldsymbol{\zeta}\|_2^2} \right) \boldsymbol{\xi} \right)$ , giving  $\operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}} \left( \boldsymbol{\xi} - \frac{(\mathbf{A}\boldsymbol{\zeta})^{\otimes 2}}{\|\mathbf{A}\boldsymbol{\zeta}\|_2^2} \boldsymbol{\xi} \right)$ . By the same tensor trick again, we have that  $\operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}}(\boldsymbol{\xi}) - \frac{\langle \mathbf{A}\boldsymbol{\zeta}, \boldsymbol{\xi} \rangle}{\|\mathbf{A}\boldsymbol{\zeta}\|_2^2} \operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}}(\mathbf{A}\boldsymbol{\zeta})$ . Expand the term  $\operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}}(\mathbf{A}\boldsymbol{\zeta})$  we get  $\operatorname{proj}_{T_{\boldsymbol{\zeta}}\mathcal{E}_A^{r-1}}(\boldsymbol{\xi}) - \frac{\langle \mathbf{A}\boldsymbol{\zeta}, \boldsymbol{\xi} \rangle}{\|\mathbf{A}\boldsymbol{\zeta}\|_2^2} \underbrace{\left( \mathbf{A}\boldsymbol{\zeta} - \frac{\langle \mathbf{A}\boldsymbol{\zeta}, \mathbf{A}\boldsymbol{\zeta} \rangle}{\|\mathbf{A}\boldsymbol{\zeta}\|_2^2} \mathbf{A}\boldsymbol{\zeta} \right)}_{=0}$ , and the proof is completed.



*Retraction.* We define retraction on the ellipsoid  $\mathcal{R}_\zeta : T_\zeta \mathcal{E}_A^{r-1} \rightarrow \mathcal{E}_A^{r-1}$  as follows.

**Proposition 4 (metric retraction)** *A way to compute  $\mathcal{R}_\zeta : T_\zeta \mathcal{E}_A^{r-1} \rightarrow \mathcal{E}_A^{r-1}$  for the manifold  $\mathcal{E}_A^{r-1}$  is*

$$\mathcal{R}_\zeta(\xi) = \frac{\zeta + \xi}{\|\zeta + \xi\|_\mathcal{E}} \stackrel{T_\zeta \mathcal{E}_A^{r-1}}{=} \frac{\zeta + \xi}{\sqrt{1 + \|\xi\|_\mathcal{E}^2}}. \quad (\text{retract-ellipsoid})$$

*Proof* Consider a continuous curve  $c : \mathbb{R} \rightarrow \mathcal{E}_A^{r-1}$  defined by the following expression  $c(t) = \mathcal{R}_\zeta(t\xi) = (\zeta + t\xi) / \sqrt{1 + t^2 \|\xi\|_\mathcal{E}^2}$ , then  $c$  is a smooth (differentiable) since  $(\zeta + t\xi) / \sqrt{1 + t^2 \|\xi\|_\mathcal{E}^2}$  is a smooth function for all  $(\zeta, \xi) \in \mathbb{R}^r \times \mathbb{R}^r$ . Next, we have  $c(0) = \zeta$  and

$$c'(0) := \left. \frac{dc(t)}{dt} \right|_{t=0} = \left. \frac{\xi - t\xi \|\xi\|_\mathcal{E}^2}{(1 + t^2 \|\xi\|_\mathcal{E}^2) \sqrt{1 + t^2 \|\xi\|_\mathcal{E}^2}} \right|_{t=0} = \xi.$$

Hence  $\mathcal{R}$  is a retraction for  $\mathcal{E}_A^{r-1}$  by definition [8, Definition 3.47].

*Riemannian Gradient Descent.* We solve (h-manifold-subproblem) using RMU discussed in Section 3.3. Based on the discussion above, the Riemannian gradient  $\text{grad}\phi(\xi) = \text{proj}_{T_\zeta \mathcal{E}_A^{r-1}} \nabla \bar{\phi}_\mathcal{E}(\xi)$  can be split as

$$\begin{aligned} \text{grad}\phi(\xi) &\stackrel{(\text{proj-ellips})}{=} \left( I_r - \frac{(\mathcal{A}\zeta)^{\otimes 2}}{\|\mathcal{A}\zeta\|_2^2} \right) \nabla \bar{\phi}_\mathcal{E}(\xi) \\ &\stackrel{(6)}{=} \left( I_r - \frac{(\mathcal{A}\zeta)^{\otimes 2}}{\|\mathcal{A}\zeta\|_2^2} \right) \frac{\mathbf{W}^\top}{\|\mathbf{W}\xi\|_2} \left( \frac{(\mathbf{W}\xi)^{\otimes 2}}{\|\mathbf{W}\xi\|_2^2} - I_m \right) \mathbf{m}_{:j} \\ &= \underbrace{\left( \frac{\mathbf{W}^\top (\mathbf{W}\xi)^{\otimes 2}}{\|\mathbf{W}\xi\|_2 \|\mathbf{W}\xi\|_2^2} + \frac{(\mathcal{A}\zeta)^{\otimes 2} \mathbf{W}^\top}{\|\mathcal{A}\zeta\|_2^2 \|\mathbf{W}\xi\|_2} \right)}_{\text{grad}^+ \phi} \mathbf{m}_{:j} \\ &\quad - \underbrace{\left( \frac{\mathbf{W}^\top}{\|\mathbf{W}\xi\|_2} + \frac{(\mathcal{A}\zeta)^{\otimes 2} \mathbf{W}^\top (\mathbf{W}\xi)^{\otimes 2}}{\|\mathcal{A}\zeta\|_2^2 \|\mathbf{W}\xi\|_2 \|\mathbf{W}\xi\|_2^2} \right)}_{\text{grad}^- \phi} \mathbf{m}_{:j}, \end{aligned}$$

where we make use of Proposition 1 to compute the Euclidean gradient of  $\mathbf{h}$ , denoted as  $\nabla \bar{\phi}_\mathcal{E}(\mathbf{h})$ , as follows

$$\frac{\langle \mathbf{W}\mathbf{h}, \mathbf{m}_{:j} \rangle \mathbf{W}^\top (\mathbf{W}\mathbf{h})}{\|\mathbf{W}\mathbf{h}\|_2^3} - \frac{\mathbf{W}^\top \mathbf{m}_{:j}}{\|\mathbf{W}\mathbf{h}\|_2} \stackrel{(1)}{=} \frac{\mathbf{W}^\top}{\|\mathbf{W}\mathbf{h}\|_2} \left( \frac{(\mathbf{W}\mathbf{h})^{\otimes 2}}{\|\mathbf{W}\mathbf{h}\|_2^2} - I_m \right) \mathbf{m}_{:j}. \quad (6)$$

Finally, to update  $\mathbf{h}$ , we put  $\xi = \zeta = \mathbf{h}_k$  in  $\text{grad}\phi$ , perform the RMU under metric retraction discussed in Section 3.3 as

$$\mathbf{h}_{:j,k+1} = \frac{\mathbf{h}_{:j,k} \odot \text{grad}^- \phi(\mathbf{h}_{:j,k})[\mathbf{h}_{:j,k}] \odot \text{grad}^+ \phi(\mathbf{h}_{:j,k})[\mathbf{h}_{:j,k}]}{\left\| \mathbf{h}_{:j,k} \odot \text{grad}^- \phi(\mathbf{h}_{:j,k})[\mathbf{h}_{:j,k}] \odot \text{grad}^+ \phi(\mathbf{h}_{:j,k})[\mathbf{h}_{:j,k}] \right\|_{\mathcal{E}_A^{r-1}}}.$$

Since RMU is a special kind of Riemannian gradient method, by the theory of Riemannian gradient method [8], we have convergence rate  $\mathcal{O}(\frac{1}{\sqrt{k}})$  for the h-subproblem.

*Efficient implementation.* Pre-compute  $\mathbf{B} := \mathbf{W}^\top \mathbf{M}$  and  $\mathbf{c} = \mathbf{A}\mathbf{h}$  gives

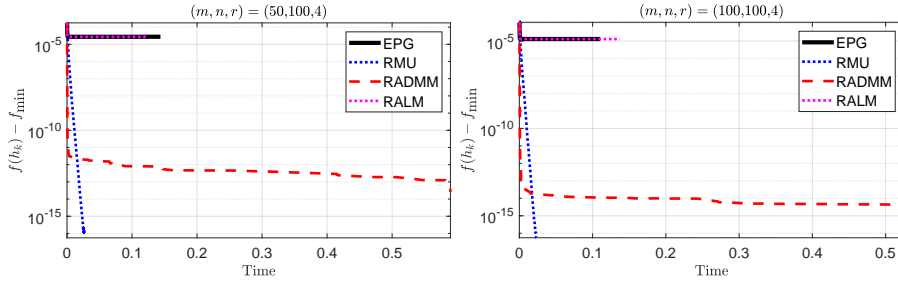
$$\text{grad}\phi(\mathbf{h}) = \frac{1}{\langle \mathbf{h}, \mathbf{c} \rangle^{1/2}} \left( \left( \frac{\langle \mathbf{B}_{:j}, \mathbf{h} \rangle \mathbf{c}}{\langle \mathbf{h}, \mathbf{c} \rangle} + \frac{\mathbf{c}^{\otimes 2}}{\langle \mathbf{c}, \mathbf{c} \rangle} \mathbf{B}_{:j} \right) - \left( \mathbf{B}_{:j} + \frac{\mathbf{c}^{\otimes 2}}{\langle \mathbf{c}, \mathbf{c} \rangle} \frac{\langle \mathbf{B}_{:j}, \mathbf{h} \rangle \mathbf{c}}{\langle \mathbf{h}, \mathbf{c} \rangle} \right) \right).$$

Using (1) we bypass tensor product and compute  $\text{grad}\phi(\mathbf{h})$  involving only four inner products

$$\text{grad}\phi(\mathbf{h}) = \frac{1}{\langle \mathbf{h}, \mathbf{c} \rangle^{1/2}} \left( \left( \frac{\langle \mathbf{B}_{:j}, \mathbf{h} \rangle}{\langle \mathbf{h}, \mathbf{c} \rangle} + \frac{\langle \mathbf{B}_{:j}, \mathbf{c} \rangle}{\langle \mathbf{c}, \mathbf{c} \rangle} \right) \mathbf{c} - \left( \mathbf{B}_{:j} + \frac{\langle \mathbf{B}_{:j}, \mathbf{h} \rangle \mathbf{c}}{\langle \mathbf{h}, \mathbf{c} \rangle} \right) \right).$$

The overall cost of computing  $\text{grad}\phi(\mathbf{h})$  is thus in  $\mathcal{O}(4r)$ , cheaper than other updates.

*Numerical performance.* We compare RMU and methods mentioned in Section 3.1 on two synthetic datasets randomly generated under zero-mean unit variance Gaussian distribution with negative entries replaced by zero. We performed 100 Monte Carlo trials on two datasets sizing  $(m, n, r) = (50, 100, 4)$  and  $(m, n, r) = (100, 100, 4)$ . Fig. 2 shows the median of the objective function values. For the case,  $(m, n, r) = (50, 100, 4)$ , RMU achieved the best performance (fastest convergence to the minimum achievable function value) 71 times, and RADMM achieved the best performance 41 times, and for the case  $(m, n, r) = (100, 100, 4)$ , RMU achieved the best performance 97 times and RADMM achieved the best performance 27 times.



**Fig. 2** Median convergence of the 100 random trials. RMU is among the fastest methods with strict feasibility. RADMM has the best convergence in the early stage of the iteration but the primal variable sequence is possibly not feasible.

## 5 Matrix-wise $\mathbf{W}$ -subproblem as a finite sum minimization

In this section we discuss how to solve the  $\mathbf{W}$ -subproblem in Algorithm 1.

*Notation simplification.* We ease the notation. Let  $\langle \cdot, \cdot \rangle_{\text{Fro}} = \text{Tr}(\cdot^\top \cdot)$  be the Frobenius inner product, from  $\langle \mathbf{m}_{:j}, \mathbf{W} \mathbf{h}_{:j} \rangle = \langle \mathbf{m}_{:j} \mathbf{h}_{:j}^\top, \mathbf{W} \rangle_{\text{Fro}}$  we define the matrices  $\mathbf{B}_j := \mathbf{m}_{:j} \mathbf{h}_{:j}^\top = \mathbf{m}_{:j} \otimes \mathbf{h}_{:j}$  and  $\mathbf{A}_j := \mathbf{h}_{:j} \mathbf{h}_{:j}^\top = \mathbf{h}_{:j} \otimes \mathbf{h}_{:j} = \mathbf{h}_{:j}^{\otimes 2}$ . We remark not to confuse matrices  $\mathbf{A}, \mathbf{B}$  here with the  $\mathbf{A}, \mathbf{B}$  defined in the last paragraph of Section 4. We rewrite the  $\mathbf{W}$ -subproblem in Algorithm 1 using  $\mathbf{A}_j, \mathbf{B}_j$  in the argmin form

$$\begin{aligned} \operatorname{argmin}_{\mathbf{W} \geq \mathbf{0}} \left\{ F(\mathbf{W}) = \frac{1}{n} \sum_{j=1}^n \bar{f}_j(\mathbf{W}) \right\}, \\ \text{where } \bar{f}_j(\mathbf{W}) = 1 - \frac{\langle \mathbf{B}_j, \mathbf{W} \rangle_{\text{Fro}}}{\langle \mathbf{W}, \mathbf{W} \mathbf{A}_j \rangle_{\text{Fro}}^{1/2}}. \end{aligned} \quad (\text{W-subproblem})$$

In  $\bar{f}_j$ , the subscript  $j$  means  $\bar{f}_j$  is defined by the  $j$ th column of  $\mathbf{H}$  and  $\mathbf{M}$ , and the over-line in  $\bar{f}$  emphasizes that we are taking smooth extension of a restricted function of  $\mathbf{W}$  from a manifold  $\mathcal{M}$  of  $\mathbf{W}$  (to be defined later) to the ambient Euclidean space  $\mathbb{R}^{m \times r}$ , assuming the denominator is nonzero. In practice if a column  $\mathbf{h}_j$  is zero, then the term  $\bar{f}_j$  is removed in the sum from the very beginning.

*RMU is inefficient on  $\mathbf{W}$ -subproblem.* We have Theorem 1 shows that RMU approach on the finite-sum (W-subproblem) is inefficient.

**Theorem 1** *It is computationally inefficient to perform RMU on  $\mathbf{W}$  to solve (W-subproblem) for a large  $n$ .*

To illustrate and to make Theorem 1 more accessible, we first consider the following three lemmas.

**Lemma 1** *For  $n = 1$  in (W-subproblem), we have a manifold which we name “shell of a single twisted spectrahedron”. Table 2 summarizes the results of RMU on such manifold.*

**Table 2** Summary of mathematical objects for manifold optimization on the shell of a single twisted spectrahedron

Name / Reference	Definition / expression
Manifold of $\mathbf{W}$ (Shell of twisted spectrahedron)	$\mathcal{M}_j := \left\{ \mathbf{W} \in \mathbb{R}^{m \times r} \mid \langle \mathbf{W}, \mathbf{W} \rangle_{\mathbf{A}_j^{1/2}} = 1 \right\}$
Tangent space of $\mathcal{M}_j$ at $\mathbf{Z}$ Definition 3	$T_{\mathbf{Z}} \mathcal{M}_j := \left\{ \mathbf{W} \in \mathbb{R}^{m \times r} \mid \langle \mathbf{W}, \mathbf{Z} \mathbf{A}_j \rangle_{\text{Fro}} = 0 \right\}$
Project $\mathbf{W}$ onto $T_{\mathbf{Z}} \mathcal{M}_j$ Proposition 6	$\text{proj}_{T_{\mathbf{Z}} \mathcal{M}_j}(\mathbf{W}) = \left( \mathbf{I} - \frac{(\mathbf{Z} \mathbf{A}_j)^{\otimes 2}}{\ \mathbf{Z} \mathbf{A}_j\ _{\text{Fro}}^2} \right) \mathbf{W}$
Retraction Proposition 7	$\mathcal{R}_{\mathbf{W}}(\mathbf{Z}) = (\mathbf{W} + \mathbf{Z}) / \ \mathbf{W} + \mathbf{Z}\ _{\mathbf{A}_j^{1/2}}$ .

*Proof* We put the proof of these results in Section 7.2 in the Appendix.

**Lemma 2 (RMU update on  $\mathbf{W}$  for single twisted spectrahedron)** *If  $n = 1$ , we can solve (W-subproblem) using RMU. The Riemannian gradient  $\text{grad}f_j(\mathbf{W}) = \text{proj}_{T_{\mathbf{Z}\mathcal{M}_j}}\nabla f_j(\mathbf{W})$  is*

$$\text{grad}f_j(\mathbf{W}) = \underbrace{\left( \frac{\langle \mathbf{B}_j, \mathbf{W} \rangle_{\text{Fro}} \mathbf{W} \mathbf{A}_j}{\langle \mathbf{W}, \mathbf{W} \mathbf{A}_j \rangle_{\text{Fro}}^{3/2}} + \frac{(\mathbf{Z} \mathbf{A}_j)^{\otimes 2}}{\|\mathbf{Z} \mathbf{A}_j\|_{\text{Fro}}^2} \frac{\mathbf{B}_j}{\langle \mathbf{W}, \mathbf{W} \mathbf{A}_j \rangle_{\text{Fro}}^{1/2}} \right)}_{\text{grad}^+ f_j(\mathbf{W})} - \underbrace{\left( \frac{\mathbf{B}_j}{\langle \mathbf{W}, \mathbf{W} \mathbf{A}_j \rangle_{\text{Fro}}^{1/2}} + \frac{(\mathbf{Z} \mathbf{A}_j)^{\otimes 2}}{\|\mathbf{Z} \mathbf{A}_j\|_{\text{Fro}}^2} \frac{\langle \mathbf{B}_j, \mathbf{W} \rangle_{\text{Fro}} \mathbf{W} \mathbf{A}_j}{\langle \mathbf{W}, \mathbf{W} \mathbf{A}_j \rangle_{\text{Fro}}^{3/2}} \right)}_{\text{grad}^- f_j(\mathbf{W})}.$$

With  $\mathbf{Z} = \mathbf{W} = \mathbf{W}_k$ , we arrive at the RMU update of  $\mathbf{W}$  under metric retraction (see Section 3.3) as

$$\mathbf{W}_{k+1} = \frac{\mathbf{W}_k \odot \text{grad}^- f_j(\mathbf{W}_k)[\mathbf{W}_k] \odot \text{grad}^+ f_j(\mathbf{W}_k)[\mathbf{W}_k]}{\left\| \mathbf{W}_k \odot \text{grad}^- f_j(\mathbf{W}_k)[\mathbf{W}_k] \odot \text{grad}^+ f_j(\mathbf{W}_k)[\mathbf{W}_k] \right\|_{\mathbf{A}_j^{1/2}}}. \quad (7)$$

*Proof* The derivation is based on the Proposition 5 on the Euclidean gradient  $\nabla F(\mathbf{W})$ , and the Proposition 7 in Section 7.2 in the Appendix.

We give an efficient implementation of the Riemannian gradient  $\text{grad}f_j(\mathbf{W})$  in Section 7.3 in the Appendix.

**Lemma 3** *For (W-subproblem) with  $n = 2$ , we have a manifold which we name “spectrahedra”. The computation of the mathematical objects for the manifold optimization on such spectrahedra have a high per-iteration cost. Precisely, the computation of RMU involves a pseudo-inverse with a cost about  $\mathcal{O}(2m^2)$  for computing the adjoint operator, and a possibly expensive metric projection step that potentially has no closed-form solution.*

*Proof* See Section 7.4 in the Appendix.

The conclusion of Lemma 3 generalizes to  $n > 2$ , hence we are now ready to prove Theorem 1.

*Proof* (The proof of Theorem 1) Based on Lemma 3, the computation of the mathematical objects for the manifold optimization on the generalized spectrahedra (with  $n \gg 2$ ) have a high per-iteration cost. Precisely, the computation of RMU involves a pseudo-inverse with a cost about  $\mathcal{O}(nm^2)$  for computing the adjoint operator, and a possibly expensive metric projection step that potentially has no closed-form solution.

*Euclidean projected gradient descent (EPG).* Ignoring the notion of manifold, (W-subproblem) can be solved by the Euclidean projected gradient update  $\mathbf{W}_{k+1} = \text{proj}_+(\mathbf{W}_k - \alpha \nabla F(\mathbf{W}_k))$  with a stepsize  $\alpha \geq 0$ . We compute  $\nabla F(\mathbf{W}_k)$  as follows.

**Proposition 5** For  $F(\mathbf{W})$  in (W-subproblem), its Euclidean gradient is

$$\nabla F(\mathbf{W}) = \frac{1}{n} \sum_{j=1}^n \nabla \bar{f}_j = \frac{1}{n} \sum_{j=1}^n \left( \frac{\langle \mathbf{B}_j, \mathbf{W} \rangle_{\text{Fro}} \mathbf{W} \mathbf{A}_j}{\langle \mathbf{W}, \mathbf{W} \mathbf{A}_j \rangle_{\text{Fro}}^{3/2}} - \frac{\mathbf{B}_j}{\langle \mathbf{W}, \mathbf{W} \mathbf{A}_j \rangle_{\text{Fro}}^{1/2}} \right). \quad (8)$$

*Proof* Quotient rule of gradient gives

$$\nabla \bar{f}_j = - \frac{\langle \mathbf{W}, \mathbf{W} \mathbf{A}_j \rangle_{\text{Fro}}^{1/2} \left( \nabla \langle \mathbf{B}_j, \mathbf{W} \rangle_{\text{Fro}} \right) - \langle \mathbf{B}_j, \mathbf{W} \rangle_{\text{Fro}} \left( \nabla \langle \mathbf{W}, \mathbf{W} \mathbf{A}_j \rangle_{\text{Fro}}^{1/2} \right)}{\langle \mathbf{W}, \mathbf{W} \mathbf{A}_j \rangle_{\text{Fro}}}.$$

By  $\mathbf{A}_j^\top = \mathbf{A}_j$  and  $\nabla \langle \mathbf{W}, \mathbf{W} \mathbf{A}_j \rangle_{\text{Fro}} = \mathbf{W} \mathbf{A}_j + \mathbf{W} \mathbf{A}_j^\top$  we arrive at (8).

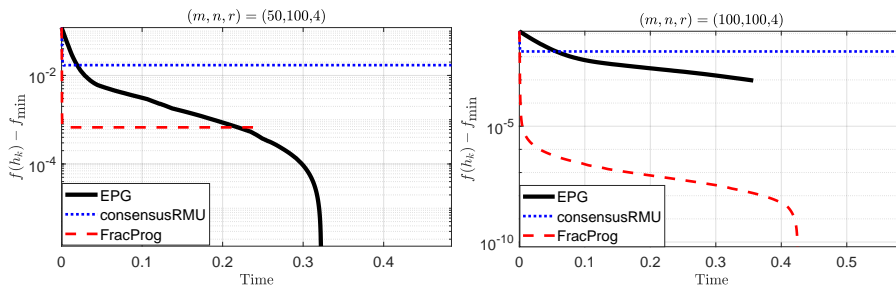
*Fractional programming.* Fractional program (FP) is another approach to handle (W-subproblem). Below we derive the FP-formulation of (W-subproblem). Apply the Dinkelbach transform [17] on (W-subproblem) by introducing variables  $\lambda_1, \lambda_2, \dots, \lambda_n$  gives

$$\underset{\mathbf{W} \geq \mathbf{0}, \lambda_1, \dots, \lambda_n}{\text{argmin}} \sum_{j=1}^n \lambda_j \langle \mathbf{W}, \mathbf{W} \mathbf{A}_j \rangle_{\text{Fro}}^{1/2} - \langle \mathbf{B}_j, \mathbf{W} \rangle_{\text{Fro}}. \quad (\text{FP-formulation})$$

Following Jagannathan's method [16], we have the following algorithm

- At each iteration  $k$ , set  $\lambda_j = \frac{\langle \mathbf{m}_{:j} \mathbf{h}_{:j}^\top, \mathbf{W} \rangle}{\langle \mathbf{W}, \mathbf{W} \mathbf{h}_{:j} \mathbf{h}_{:j}^\top \rangle^{1/2}}$  using the most recent version of  $\mathbf{W}$  and  $\mathbf{H}$ .
- Perform Euclidean gradient descent (with Nesterov's acceleration) on (FP-formulation).

*Numerical performance.* As RMU for the whole sum  $F(\mathbf{W})$  in (W-subproblem) is computationally infeasible, we consider consensus setup [32] on each manifold. We numerically compare the performance of three methods: EPG, consensus RMU and Fractional Programming, on two synthetic datasets randomly generated under zero-mean unit variance Gaussian distribution with negative entries replaced by zero. We performed 100 Monte Carlo runs on two datasets sizing  $(m, n, r) = (50, 100, 4)$  and  $(m, n, r) = (100, 100, 4)$ . Fig. 3 shows the median of the objective function values. For the case  $(m, n, r) = (50, 100, 4)$ , EPG achieved the fastest convergence 100% of the time, and for the case  $(m, n, r) = (100, 100, 4)$ , FracProg achieved the fastest convergence 100% of the time.



**Fig. 3** Median convergence comparison, among 100 Monte Carlo runs, of consensus RMU with EPG (Euclidean Projected Gradient) and Fractional Programming. EPG and Fractional Programming approaches are the fastest methods.

## 6 Experiment

In this section we report the numerical results concerning the performance of Chordal-NMF. We report results on synthetic data in Section 6.1, and then on real-world data in Section 6.2. Below we give the description of the experimental setup.

*How Chordal-NMF is applied.* We briefly describe how Chordal-NMF is applied on a data matrix  $\mathbf{M}$ . First we remove all zero columns in  $\mathbf{M}$ , then we normalize each column of  $\mathbf{M}$  by its  $\ell_2$  norm, where we also record the value of the norm. After that we run Chordal-NMF on the normalized  $\mathbf{M}$  and get the decomposition  $\mathbf{WH}$ . Lastly, we multiply the column norm on  $\mathbf{M}$  before the normalization back to each column  $\mathbf{h}_j$ .

*Implementation of Chordal-NMF.* We implement Chordal-NMF using the BCD structure discussed in the introduction. Based on the preliminary tests on different solvers for the sub-problems (see the end of Section 4 and Section 5), we implement Chordal-NMF as follows: for the H-subproblem, we perform RMU (Riemannian Multiplicative Update) as described in Section 4. For the W-subproblem, we perform EPG (Euclidean Projected Gradient) as described in Section 5.

*Benchmark.* We compare the result of the Chordal-NMF with that of the classical Frobenius norm NMF (FroNMF) based on the algorithm HALS [1, Chapter 8.3.3]. In all the experiments, all the methods start with the same random initialization, where all elements in  $\mathbf{W}_0, \mathbf{H}_0$  are generated under uniform distribution  $\mathcal{U}[0, 1]$ .

*Code.* The experiments were conducted in MATLAB 2023a<sup>1</sup> in a machine with OS Windows 11 Pro on a Intel Core 12 gen. CPU 2.20GHz and 16GB RAM.

<sup>1</sup> The MATLAB code is available at [https://github.com/flaespo/Chordal\\_NMF](https://github.com/flaespo/Chordal_NMF).

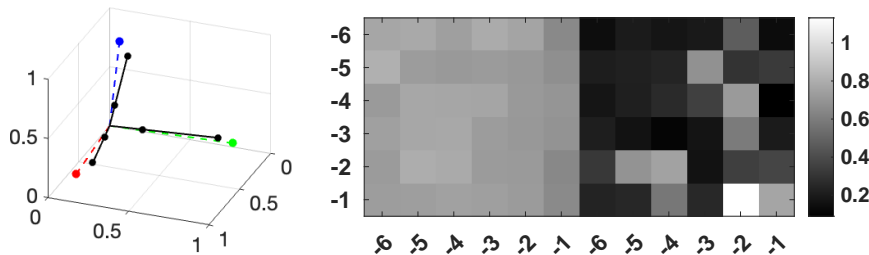
## 6.1 Synthetic Dataset

*Dataset description.* We use the specific dataset  $\mathbf{M}^{\epsilon,\delta} = \mathbf{W}_{\text{true}}\mathbf{H}_{\text{true}}^{\epsilon,\delta}$  with  $\epsilon > 0$  and  $\delta > 0$  as follows.

$$\mathbf{W}_{\text{true}} = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}, \mathbf{H}_{\text{true}}^{\epsilon,\delta} = \begin{bmatrix} 1 - \epsilon & \delta(1 - \epsilon) & \epsilon & \delta\epsilon & \epsilon & \delta\epsilon \\ \epsilon & \delta\epsilon & 1 - \epsilon & \delta(1 - \epsilon) & \epsilon & \delta\epsilon \\ \epsilon & \delta\epsilon & \epsilon & \delta\epsilon & 1 - \epsilon & \delta(1 - \epsilon) \end{bmatrix}.$$

We explain the meaning and the purpose of such setup. The matrix  $\mathbf{W}_{\text{true}}$  represents a cone in  $\mathbb{R}^3$  (see Fig. 4). The matrix  $\mathbf{H}_{\text{true}}^{\epsilon,\delta}$  represents how we generate the six columns of  $\mathbf{M}^{\epsilon,\delta}$  by conic combination of columns of  $\mathbf{W}_{\text{true}}$  under a small perturbation  $\epsilon$  and an attenuation  $\delta$ . The perturbation  $\epsilon$  represents how much the columns of  $\mathbf{M}^{\epsilon,\delta}$  deviate from the columns of  $\mathbf{W}_{\text{true}}$ , while the attenuation  $\delta$  represents how much the columns  $\mathbf{M}_{:2}, \mathbf{M}_{:4}, \mathbf{M}_{:6}$  in  $\mathbf{M}^{\epsilon,\delta}$  have their norm scaled downward. For  $\delta$  getting smaller, it is getting harder for FroNMF to recover  $\mathbf{H}_{\text{true}}$ , while it is less a problem for Chordal-NMF since the angle between the data columns are invariant to the attenuation  $\delta$ . See Fig. 4 for an illustration.

We construct  $\mathbf{M}^{\epsilon,\delta}$  across different values of  $(\epsilon, \delta)$ , and run Chordal-NMF and FroNMF on each instance of  $\mathbf{M}^{\epsilon,\delta}$  generated. Then we extract the matrix  $\mathbf{H}$  produced by the last iteration of the each method, and calculate the relative error between  $\mathbf{H}$  and  $\mathbf{H}_{\text{true}}^{\epsilon,\delta}$  as  $\|\mathbf{H} - \mathbf{H}_{\text{true}}^{\epsilon,\delta}\|_F / \|\mathbf{H}_{\text{true}}^{\epsilon,\delta}\|_F$ . Fig. 4 shows the heatmap of the results across different values of  $(\epsilon, \delta)$ , showing that Chordal-NMF on average has a better recovery of  $\mathbf{H}$  than FroNMF, especially for the case when  $\delta$  is small.



**Fig. 4** Left subplot: the plot of  $\mathbf{W}_{\text{true}}$  (the red, blue, green rays) and  $\mathbf{M}(0.1, 0.3)$  (the black rays). Right subplot: The relative error  $\|\mathbf{H} - \mathbf{H}_{\text{true}}^{\epsilon,\delta}\|_F / \|\mathbf{H}_{\text{true}}^{\epsilon,\delta}\|_F$  for the two methods, where left grid (the first six columns) is the result from FroNMF, and the right grid (the last six columns) is the result from Chordal-NMF. In the grid, the x-axis is the value of  $\delta$  (in log-scale) and the y-axis is the value of  $\epsilon$  (in log-scale) .

## 6.2 On real-world dataset

NMF finds extensive use across various application domains in real-world data analysis [1]. A prevalent application domain is within the realm of Earth Ob-

ervation (EO), particularly in the analysis of remote sensing data. EO applications typically involve the utilization of multispectral or hyperspectral images, which are stored as matrices and can be effectively analyzed through NMF for unmixing purposes [33]. A conventional approach in this context involves addressing the standard NMF problem by minimizing the point-to-point Frobenius norm [34]. However, in this section, we introduce a comparative analysis between the performance of FroNMF and the novel Chordal-NMF method proposed in this work.

*Dataset description.* We chose a cloudy multispectral image due to its relevance in highlighting the advantages of employing Chordal-NMF compared to standard FroNMF for analyzing multispectral images under various cloudiness conditions. Our primary goal here is to provide a comprehensive illustration of how Chordal-NMF can better manage the presence of different cloud conditions, thereby presenting itself as a useful alternative for conducting image analysis for EO applications. Specifically, the matrix  $\mathbf{M}$  is a cloudy multispectral image (with a pixel size rows:150, cols:290 and bands:12), in a selected area of Apulia region in Italy, from the Copernicus data space ecosystem<sup>2</sup>. We use a reference image obtained in cloudless conditions as the ground truth<sup>3</sup>. We benchmark Chordal-NMF with FroNMF, where both methods start with the same (random) initialization.

*The reconstruction.* We run FroNMF and Chordal-NMF on the cloudy data  $\mathbf{M}$  with the same initialization  $\mathbf{W}_0, \mathbf{H}_0$  obtained from random uniform distribution  $\mathcal{U}[0, 1]$ . The initial objective function value  $F(\mathbf{W}_0, \mathbf{H}_0) = 0.1906100$ . The reconstruction given by  $\mathbf{W}_{\text{FroNMF}}, \mathbf{H}_{\text{FroNMF}}$  from FroNMF gives a chordal function value 0.0012792. The reconstruction obtained by  $\mathbf{W}_{\text{Ch}}, \mathbf{H}_{\text{Ch}}$  from Chordal-NMF gives a chordal function value 0.0014398. We consider the two methods achieve a similar chordal function value.

From a qualitative point of view, Fig. 5 shows an RGB representation of the ground truth (cloudless reference image), the cloudy data  $\mathbf{M}$ , and the reconstruction obtained from FroNMF and Chordal-NMF on  $\mathbf{M}$ .



**Fig. 5** The RGB image of a scene. From left to right, the images are: the ground truth (the cloudless reference image), the cloudy data image  $\mathbf{M}$ , the reconstruction obtained by FroNMF on  $\mathbf{M}$ , and the reconstruction obtained by Chordal-NMF on  $\mathbf{M}$ . In the image, the three color-boxes are selected areas under different cloudiness in which we perform further quantitative analysis.

<sup>2</sup> <https://dataspace.copernicus.eu/>

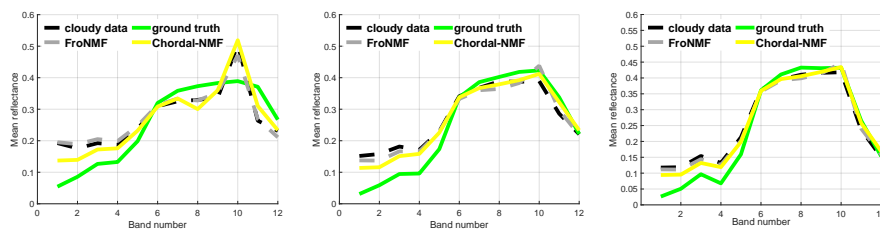
<sup>3</sup> The cloudy image and the cloudless image are obtained in the same condition (except the cloudiness) with a temporal difference of 5 days.



*Further analysis on the selected areas.* We now present the analysis of the three selected areas. The three boxes were chosen to represent three distinct cloud conditions: the red box is a cloudy area, the cyan box is a less-cloudy areas, and the yellow box is a cloudless area. We compute the spectral signatures of the pixels in these areas, and their mean behavior is plotted in Figs. 6. We also numerically compared these spectral profile vectors by two criteria:

1. the SID-SAM between the spectral profile vectors, where we consider the hybrid SID-SAM [35] of the spectral information divergence (SID) and spectral angle mapper (SAM),
2. the  $\ell_2$ -norm of the difference between the spectral profile vectors.

Results are reported in Table 3 for the three areas, in which Chordal-NMF achieves a better performance.



**Fig. 6** From left to right: the spectral signatures of the pixels in the red/cyan/yellow box in Fig.5. In all the three cases, the spectral profile of obtained by Chordal-NMF is on average closer to the ground truth spectral profile (measurement obtained in cloudless condition).

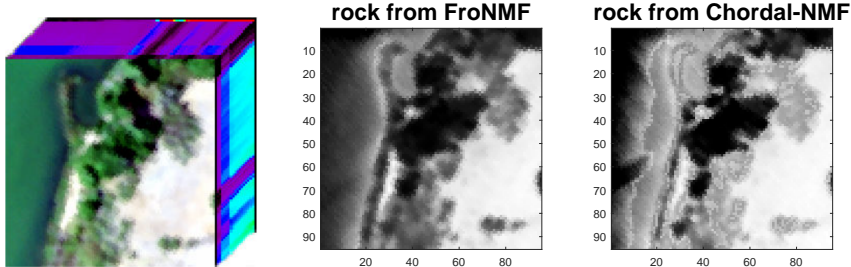
**Table 3** The numerics of the pixels in the three boxes. Here “w GT” stands for “with ground truth”.

Compared w GT	Red		Cyan		Yellow	
	SID-SAM	$\ell_2$	SID-SAM	$\ell_2$	SID-SAM	$\ell_2$
Cloudy dataset	2.2559	0.2655	0.0633	0.2154	0.0807	0.1594
FroNMF	2.0214	0.2650	0.5518	0.1901	0.3574	0.1458
Chordal-NMF	0.2090	0.2116	0.1486	0.1487	0.1161	0.1161

On the cloudy image recovery, we see that Chordal-NMF is always better than FroNMF regardless of the cloudiness of the image by achieving a lower SID-SAM value and a lower  $\ell_2$ -norm value. For example, in the red box, Chordal-NMF seems to have a better recovery of the region under the cloud.

*On Samson dataset.* We test our approach also on a benchmark dataset for EO applications: Samson dataset [1]. In this image, there are  $95 \times 95$  pixels, each pixel is recorded at 156 channels. This dataset is not challenging since many analyses have been already carried out [34], however we found it interesting to report some results on it. In fact, even if it is well known from the

literature that there are three targets in the image, we want to highlight how Chordal-NMF is better at extracting the rock/soil component. Fig. 7 shows the abundance map from the FroNMF and Chordal-NMF. From this, we can see the Chordal-NMF is able to recover rock under shadow water regions near the coast better than standard FroNMF.



**Fig. 7** The Samson dataset and the rock abundances map of the decompositions. From left to right: the data, the abundance maps obtained from FroNMF and Chordal-NMF.

## 7 Conclusion

In this paper, we introduced a NMF model called Chordal-NMF, which is different from the classical NMF that the objective function is a point-to-point Euclidean distance, where in Chordal-NMF a ray-to-ray distance is used. Based on the geometric interpretation that NMF describes a cone, we argued that chordal distance, which measures the angle between two vector in the nonnegative orthant, is more suitable than the Euclidean distance for NMF.

Under a BCD framework, we developed a new algorithm to solve the Chordal-NMF, where Riemannian optimization technique is used to solve the H-subproblem. To be exact we proposed a Riemannian Multiplicative Update (RMU) that preserves the convergence properties of Riemannian gradient descent without breaking the smoothness condition on the manifold.

We showcase the effectiveness of the Chordal-NMF on the synthetic dataset as well as real-world multispectral images.

## Appendix

### 7.1 The proof of Proposition 1

*Proof* Let  $h(\mathbf{x}) = (g \circ f)(\mathbf{x}) = (f(\mathbf{x}))^2$  where function  $g : \mathbb{R} \rightarrow \mathbb{R} : x \mapsto x^2$  is convex, increasing (on  $\mathbb{R}_+$ ) and differentiable, then by chain rule we have  $\partial(g \circ f)(\mathbf{x}) = g'(f(\mathbf{x}))\partial f(\mathbf{x})$ , which gives

$$\nabla h(\mathbf{x}) = 2f(\mathbf{x})\nabla f(\mathbf{x}) = \frac{\langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle}{2\|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2} \nabla f(\mathbf{x}) \quad (\text{chain-rule})$$

Now by the definition  $h(\mathbf{x}) = (f(\mathbf{x}))^2$ , so

$$\nabla h(\mathbf{x}) = \nabla \left( \frac{(\langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle)^2}{\|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2^2} \right) \quad (\text{grad-h})$$

Equate (chain-rule) and (grad-h) gives

$$\nabla f(\mathbf{x}) = \frac{\|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2}{\langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle} \nabla \left( \frac{(\langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle)^2}{2\|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2^2} \right)$$

By assumption  $\mathbf{D}\mathbf{x} + \mathbf{e} \neq \mathbf{0}$ , we make use of quotient rule  $\nabla \frac{f}{g} = \frac{g\nabla f - f\nabla g}{g^2}$  to arrive at

$$\begin{aligned} & \nabla f(\mathbf{x}) \\ &= \frac{\|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2}{\langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle} \frac{\|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2^2 \nabla (\langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle)^2 - (\langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle)^2 \nabla \|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2^2}{2\|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2^4} \\ &= \frac{\|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2^2 \nabla (\langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle)^2 - (\langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle)^2 \nabla \|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2^2}{2\langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle \|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2^3} \\ &= \frac{\|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2^2 2\langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle \mathbf{A}^\top \mathbf{c} - (\langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle)^2 2\mathbf{D}^\top (\mathbf{D}\mathbf{x} + \mathbf{e})}{2\langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle \|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2^3} \\ &= \frac{\|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2^2 \mathbf{A}^\top \mathbf{c} - \langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle \mathbf{D}^\top (\mathbf{D}\mathbf{x} + \mathbf{e})}{\|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2^3} \end{aligned}$$

## 7.2 The proof of Lemma 1 on a single shell of twisted spectrahedra

Consider (W-subproblem) with  $n = 1$ . Following the discussion in Section 4, we get grid of the denominator  $\langle \mathbf{W}, \mathbf{W}\mathbf{A}_1 \rangle_{\text{Fro}}$  in  $\bar{f}_1$  by introducing a constrained problem  $\operatorname{argmax}_{\mathbf{W} \geq \mathbf{0}} \langle \mathbf{B}_j, \mathbf{W} \rangle_{\text{Fro}}$  s.t.  $\langle \mathbf{W}, \mathbf{W}\mathbf{A}_1 \rangle_{\text{Fro}} = 1$ . We note that  $\mathbf{A}_1$  is a rank-1 symmetric positive semi-definite (PSD) matrix with two eigenvalues: a single positive eigenvalue and 0 with multiplicity  $r - 1$ . Moreover,  $\mathbf{A}_1$  has its square-root  $\mathbf{A}_j^{1/2}$ , so  $\langle \mathbf{W}^\top \mathbf{W}, \mathbf{A}_1 \rangle_{\text{Fro}} = \langle \mathbf{W}^\top \mathbf{W}, \mathbf{A}_1^{1/2} \mathbf{A}_1^{1/2} \rangle_{\text{Fro}} = \langle \mathbf{W}\mathbf{A}_1^{1/2}, \mathbf{W}\mathbf{A}_1^{1/2} \rangle_{\text{Fro}} = \langle \mathbf{W}, \mathbf{W} \rangle_{\mathbf{A}_1^{1/2}}$  allows us to define the manifold  $\mathcal{M}_1$  as

$$\mathcal{M}_1 = \left\{ \mathbf{W} \in \mathbb{R}^{m \times r} \mid \langle \mathbf{W}, \mathbf{W} \rangle_{\mathbf{A}_1^{1/2}} - 1 = 0 \iff \operatorname{Tr}(\mathbf{A}_1 \mathbf{W}^\top \mathbf{W}) = 1 \right\},$$

(Shell of twisted spectrahedron)

which we interpret  $\mathcal{M}_1$  as the shell of a twisted spectrahedron. The term spectrahedron [36] refers to the eigen-spectrum of a matrix behaves like a polyhedron, while the word ‘‘twisted’’ refers to the linear transformation  $\mathbf{A}_1$ .

Now we mirror Section 3.2 and Section 4.1 on deriving the tools on manifold of  $\mathbf{W}$ , summarized in Table 2, where consider a particular  $j$  for  $\mathcal{M}_j$ .

*Tangent space.* On the shell of a single twisted spectrahedron  $\mathcal{M}_j$  we define the tangent space as follows.

**Definition 3** Let  $h(\mathbf{W}) = \langle \mathbf{W}, \mathbf{W}\mathbf{A}_j \rangle_{\text{Fro}} - 1$  defines  $\mathcal{M}_j$ , its differential is  $Dh(\mathbf{Z})[\mathbf{Z}] = 2\langle \mathbf{W}\mathbf{A}_j, \mathbf{Z} \rangle_{\text{Fro}}$ . The tangent space of  $\mathcal{M}_j$  at a reference point  $\mathbf{Z} \in \mathcal{M}_j$ , denoted as  $T_{\mathbf{Z}}\mathcal{M}_j$ , is

$$T_{\mathbf{Z}}\mathcal{M}_j := \left\{ \mathbf{W} \in \mathbb{R}^{m \times r} \mid \langle \mathbf{Z}\mathbf{A}_j, \mathbf{W} \rangle_{\text{Fro}} = 0 \right\} = \text{Ker} Dh(\mathbf{W}).$$

*Projection onto tangent space.* We define the adjoint operator of  $Dh(\mathbf{W})[\mathbf{Z}]$  as  $Dh(\mathbf{W})^*[\alpha] = 2\alpha\mathbf{W}\mathbf{A}_j$ , where  $\alpha$  takes the value  $\bar{\alpha} = \frac{1}{2} \frac{\langle \mathbf{W}, \mathbf{Z}\mathbf{A}_j \rangle_{\text{Fro}}}{\|\mathbf{Z}\mathbf{A}_j\|_{\text{Fro}}^2}$  obtained by solving the least squares  $\bar{\alpha} = \underset{\alpha \in \mathbb{R}}{\text{argmin}} \|\mathbf{W} - Dh(\mathbf{Z})^*[\alpha]\|_{\text{Fro}}^2$ . Now by  $Dh(\mathbf{W})^*[\alpha]$ ,  $\bar{\alpha}$ , the definition of projector onto tangent space [8, Definition 3.60], and the property of orthogonal projector [8, Equation 7.74], we have the projector  $\text{proj}_{T_{\mathbf{Z}}\mathcal{M}_j}(\mathbf{W}) = \mathbf{W} - Dh(\mathbf{Z})^*[\bar{\alpha}]$  as shown in the following proposition.

**Proposition 6** *The projector  $\text{proj}_{T_{\mathbf{Z}}\mathcal{M}_j} : \mathbb{R}^{m \times r} \rightarrow T_{\mathbf{Z}}\mathcal{M}_j$  of  $\mathbf{W}$  onto the tangent space at the reference point  $\mathbf{Z}$  is*

$$\begin{aligned} \text{proj}_{T_{\mathbf{Z}}\mathcal{M}_j}(\mathbf{W}) &= \mathbf{W} - \frac{\langle \mathbf{Z}\mathbf{A}_j, \mathbf{W} \rangle_{\text{Fro}}}{\|\mathbf{Z}\mathbf{A}_j\|_{\text{Fro}}^2} \mathbf{Z}\mathbf{A}_j \\ &= \left( \mathbf{I} - \frac{\mathbf{Z}\mathbf{A}_j \otimes \mathbf{Z}\mathbf{A}_j}{\|\mathbf{Z}\mathbf{A}_j\|_{\text{Fro}}^2} \right) \mathbf{W} = \left( \mathbf{I} - \frac{(\mathbf{Z}\mathbf{A}_j)^{\otimes 2}}{\|\mathbf{Z}\mathbf{A}_j\|_{\text{Fro}}^2} \right) \mathbf{W}. \end{aligned} \quad (9)$$

*Proof* The proof follows the same arguments used in the proof of Proposition 3.

*Retraction.* We define the retraction on  $\mathcal{M}_j$  as  $\mathcal{R}_{\mathbf{W}} : T_{\mathbf{Z}}\mathcal{M}_j \rightarrow \mathcal{M}_j$  as follows.

**Proposition 7 (Retraction)** *A way to compute the retraction*

$\mathcal{R}_{\mathbf{W}} : T_{\mathbf{Z}}\mathcal{M}_j \rightarrow \mathcal{M}_j$  *for the manifold  $\mathcal{M}_j$  is*

$$\mathcal{R}_{\mathbf{W}}(\mathbf{Z}) = \frac{\mathbf{W} + \mathbf{Z}}{\|\mathbf{W} + \mathbf{Z}\|_{\mathbf{A}_j^{1/2}}} \stackrel{T_{\mathbf{Z}}\mathcal{M}_j}{=} \frac{\mathbf{W} + \mathbf{Z}}{\sqrt{1 + \|\mathbf{W}\|_{\mathbf{A}_j^{1/2}}^2}}. \quad (\text{retract-spectrahedron})$$

*Proof* Consider a continuous curve  $c : \mathbb{R} \rightarrow \mathcal{M}_j$  defined by the following expression  $c(t) = \mathcal{R}_{\mathbf{W}}(t\mathbf{Z}) = (\mathbf{W} + t\mathbf{Z}) / \sqrt{1 + t^2\|\mathbf{W}\|_{\mathbf{A}_j^{1/2}}^2}$ . Then  $c$  is a smooth (differentiable) since  $(\mathbf{W} + t\mathbf{Z}) / \sqrt{1 + t^2\|\mathbf{W}\|_{\mathbf{A}_j^{1/2}}^2}$  is a smooth function for all  $(\mathbf{W}, \mathbf{Z}) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{m \times r}$ . Next, we have  $c(0) = \mathbf{W}$  and

$$c'(0) := \left. \frac{dc(t)}{dt} \right|_{t=0} = \left. \frac{\mathbf{Z} - t\|\mathbf{Z}\|_{\mathbf{A}_j^{1/2}}^2 \mathbf{W}}{(1 + t^2\|\mathbf{W}\|_{\mathbf{A}_j^{1/2}}^2) \sqrt{1 + t^2\|\mathbf{W}\|_{\mathbf{A}_j^{1/2}}^2}} \right|_{t=0} = \mathbf{Z}.$$

Hence (retract-spectrahedron) is a retraction for  $\mathcal{M}_j$  by definition [8, Definition 3.47].

### 7.3 Efficient implementation of the Riemannian gradient in Lemma 2

Consider Lemma 2, let  $\mathbf{C} = \mathbf{W}_k \mathbf{A}_j$  in  $\text{grad}f_j(\mathbf{W}_k)[\mathbf{W}_k]$  gives

$$\begin{aligned} \text{grad}f_j(\mathbf{W}_k)[\mathbf{W}_k] &= \frac{1}{\langle \mathbf{W}, \mathbf{C} \rangle_{\text{Fro}}^{1/2}} \left( \frac{\langle \mathbf{B}_j, \mathbf{W} \rangle_{\text{Fro}} \mathbf{C}}{\langle \mathbf{W}, \mathbf{C} \rangle_{\text{Fro}}} + \frac{(\mathbf{C})^{\otimes 2}}{\|\mathbf{C}\|_{\text{Fro}}^2} \mathbf{B}_j \right) \\ &\quad - \left( \mathbf{B}_j + \frac{(\mathbf{C})^{\otimes 2}}{\|\mathbf{C}\|_{\text{Fro}}^2} \frac{\langle \mathbf{B}_j, \mathbf{W} \rangle_{\text{Fro}} \mathbf{C}}{\langle \mathbf{W}, \mathbf{C} \rangle_{\text{Fro}}} \right). \end{aligned}$$

We remark that the use of tensor product is purely for the ease of deriving the expression. The tensor product between two matrices is likely to be very expensive. Hence, we avoid the expensive tensor product and compute  $\text{grad}f_j(\mathbf{W})$  as

$$\begin{aligned} \text{grad}f_j(\mathbf{W}) &= \frac{1}{\langle \mathbf{W}, \mathbf{C} \rangle_{\text{Fro}}^{1/2}} \left( \left( \frac{\langle \mathbf{B}_j, \mathbf{W} \rangle_{\text{Fro}}}{\langle \mathbf{W}, \mathbf{C} \rangle_{\text{Fro}}} + \frac{\langle \mathbf{B}_j, \mathbf{C} \rangle_{\text{Fro}}}{\|\mathbf{C}\|_{\text{Fro}}^2} \right) \mathbf{C} \right. \\ &\quad \left. - \left( \mathbf{B}_j + \frac{\langle \mathbf{B}_j, \mathbf{W} \rangle_{\text{Fro}}}{\langle \mathbf{W}, \mathbf{C} \rangle_{\text{Fro}}} \mathbf{C} \right) \right). \end{aligned}$$

### 7.4 The proof of Lemma 3 for $n = 2$

Consider  $n = 2$  in (W-subproblem). By the fact that the Cartesian products of manifolds are manifolds, we consider product space  $\mathcal{M}_1 \times \mathcal{M}_2$ . Define

$$\begin{aligned} \mathcal{M}^{[2]} &= \{ \mathbf{W} \in \mathbb{R}^{m \times r} \mid h(\mathbf{W}) = \mathbf{0}_2 \}, \\ \text{where } h : \mathbb{R}^{m \times r} &\rightarrow \mathbb{R}^2 : \mathbf{W} \mapsto \begin{bmatrix} \langle \mathbf{W}, \mathbf{W} \mathbf{A}_1 \rangle_{\text{Fro}} - 1 \\ \langle \mathbf{W}, \mathbf{W} \mathbf{A}_2 \rangle_{\text{Fro}} - 1 \end{bmatrix}. \end{aligned} \quad (\text{Spectrahedra})$$

We call the manifold  $\mathcal{M}^{[2]}$  ‘‘spectrahedra’’ as it is constructed by spectrahedron. We now compute its tangent space. Following arguments in Section 4.1 and Section 7.2, we have the following results that we hide the proofs. The function  $h$  in (Spectrahedra) has the differential  $Dh(\mathbf{W})[\mathbf{Z}] = 2 \begin{bmatrix} \langle \mathbf{W} \mathbf{A}_1, \mathbf{Z} \rangle_{\text{Fro}} \\ \langle \mathbf{W} \mathbf{A}_2, \mathbf{Z} \rangle_{\text{Fro}} \end{bmatrix}$ .

The tangent space of  $\mathcal{M}^{[2]}$  is the set

$$T_{\mathbf{Z}} \mathcal{M}^{[2]} := \left\{ \mathbf{W} \mid \langle \mathbf{W} \mathbf{A}_1, \mathbf{Z} \rangle_{\text{Fro}} = \langle \mathbf{W} \mathbf{A}_2, \mathbf{Z} \rangle_{\text{Fro}} = 0 \right\},$$

its adjoint is  $Dh(\mathbf{W})^*[\boldsymbol{\alpha}] = 2\alpha_1 \mathbf{W} \mathbf{A}_1 + 2\alpha_2 \mathbf{W} \mathbf{A}_2 = 2\mathbf{W}(\alpha_1 \mathbf{A}_1 + \alpha_2 \mathbf{A}_2)$ . Let  $\text{vec}$  be vectorization and let  $\otimes_{\mathbb{K}}$  be the Kronecker product, now the  $\boldsymbol{\alpha}$  that

minimizes  $\|\mathbf{W} - Dh(\mathbf{Z})^*[\boldsymbol{\alpha}]\|_{\text{Fro}}^2$  also minimizes

$$\begin{aligned} & \left\| \text{vec}\left(\mathbf{W} - 2\mathbf{Z}(\alpha_1\mathbf{A}_1 + \alpha_2\mathbf{A}_2)\right) \right\|_2^2 \\ &= \left\| \text{vec}(\mathbf{W}) - 2(\mathbf{I} \otimes_{\mathbb{K}} \mathbf{Z})\text{vec}\left(\alpha_1\mathbf{A}_1 + \alpha_2\mathbf{A}_2\right) \right\|_2^2 \\ &= \left\| \text{vec}(\mathbf{W}) - 2(\mathbf{I} \otimes_{\mathbb{K}} \mathbf{Z}) \begin{bmatrix} \text{vec}\mathbf{A}_1 & \text{vec}\mathbf{A}_2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \right\|_2^2. \end{aligned}$$

To simplify the notation, let  $\mathbf{S}_{\mathbf{Z}} = (\mathbf{I} \otimes_{\mathbb{K}} \mathbf{Z})[\text{vec}\mathbf{A}_1 \text{ vec}\mathbf{A}_2]$ , where the subscript  $\mathbf{Z}$  indicates the dependence of  $\mathbf{Z}$ . Now  $\boldsymbol{\alpha}$  is the root of the following normal equation

$$\mathbf{S}_{\mathbf{Z}}^\top \mathbf{S}_{\mathbf{Z}} \begin{bmatrix} \alpha_1^* \\ \alpha_2^* \end{bmatrix} = \frac{1}{2} \mathbf{S}_{\mathbf{Z}}^\top \text{vec}\mathbf{W} \implies \boldsymbol{\alpha}^* = \frac{1}{2} (\mathbf{S}_{\mathbf{Z}}^\top \mathbf{S}_{\mathbf{Z}})^{-1} \mathbf{S}_{\mathbf{Z}}^\top \text{vec}\mathbf{W} = \frac{1}{2} \mathbf{S}_{\mathbf{Z}}^\dagger \text{vec}\mathbf{W}.$$

Now the orthogonal projector  $\text{proj}_{T_{\mathbf{Z}}\mathcal{M}^{[2]}} : \mathbb{R}^{m \times n} \mapsto \mathcal{M}^{[2]}$  is

$$\text{proj}_{T_{\mathbf{Z}}\mathcal{M}^{[2]}}(\mathbf{W}) = \mathbf{W} - 2\mathbf{Z} \sum_{j=1}^2 \frac{1}{2} \left( \mathbf{S}_{\mathbf{Z}}^\dagger \text{vec}\mathbf{W} \right)_j \mathbf{A}_j = \mathbf{W} - 2\mathbf{Z}(\alpha_1^* \mathbf{A}_1 + \alpha_2^* \mathbf{A}_2). \quad (10)$$

The Riemannian gradient is then

$$\text{grad}F(\mathbf{W}) = 2\text{proj}_{T_{\mathbf{Z}}\mathcal{M}^{[2]}}\left(\nabla F(\mathbf{W})\right) \stackrel{(10)}{=} \nabla F(\mathbf{W}) - 2\mathbf{Z}(\alpha_1^* \mathbf{A}_1 + \alpha_2^* \mathbf{A}_2). \quad (11)$$

Note that the value of  $\alpha_1^*$  and  $\alpha_2^*$  in (11) is an implicit function of  $\mathbf{Z}$  and  $\nabla F(\mathbf{W})$  as  $\begin{bmatrix} \alpha_1^* \\ \alpha_2^* \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \left( \mathbf{S}_{\mathbf{Z}}^\dagger \text{vec}\nabla F(\mathbf{W}) \right)_1 \\ \left( \mathbf{S}_{\mathbf{Z}}^\dagger \text{vec}\nabla F(\mathbf{W}) \right)_2 \end{bmatrix}$ . Hence the explicit expression of the Riemannian gradient is

$$\text{grad}F(\mathbf{W}) = \nabla F(\mathbf{W}) - \mathbf{Z} \left( \mathbf{S}_{\mathbf{Z}}^\dagger \text{vec}\nabla F(\mathbf{W}) \right)_1 \mathbf{A}_1 - \mathbf{Z} \left( \mathbf{S}_{\mathbf{Z}}^\dagger \text{vec}\nabla F(\mathbf{W}) \right)_2 \mathbf{A}_2.$$

Then we compute  $\text{grad}^+ F = \max\{\text{grad}F, \mathbf{0}\}$  and  $\text{grad}^- F = \max\{-\text{grad}F, \mathbf{0}\}$  to proceed with RMU.

In conclusion, we can see that in order to run RMU on  $\mathcal{M}^{[2]}$ , there are several challenges:

- the computation of  $\boldsymbol{\alpha}^*$ , which includes the computation of  $\mathbf{S}_{\mathbf{Z}} \in \mathbb{R}^{m^2 \times 2}$ ,  $\mathbf{S}_{\mathbf{Z}}^\dagger \in \mathbb{R}^{2 \times 2}$ , in which all these terms have to be re-computed in each iteration.
- the computation of the metric projection onto  $\mathcal{M}^{[2]}$ , which itself is a difficult problem.

**Acknowledgements** F.E. is supported by ERC Seeds Uniba project “Biomes Data Integration with Low-Rank Models” (CUP H93C23000720001), Piano Nazionale di Ripresa e Resilienza (PNRR), Missione 4 “Istruzione e Ricerca”-Componente C2 Investimento 1.1, “Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale”, Progetto PRIN-2022 PNRR, P2022BLN38, Computational approaches for the integration of multi-omics data. CUP: H53D23008870001, is member of the Gruppo Nazionale Calcolo Scientifico - Istituto Nazionale di Alta Matematica (GNCS-INdAM).

FE would like to thank Prof. Nicoletta Del Buono from University of Bari Aldo Moro for scientific discussion during this project.

## Conflict of interest

The authors declare that they have no conflict of interest.

## References

1. N. Gillis, *Nonnegative Matrix Factorization* (SIAM, 2020)
2. H.V. Nguyen, L. Bai, in *Asian Conference on Computer Vision* (Springer, 2010), pp. 709–720
3. P. Xia, L. Zhang, F. Li, *Information sciences* **307**, 39 (2015)
4. C. Luo, J. Zhan, X. Xue, L. Wang, R. Ren, Q. Yang, in *Artificial Neural Networks and Machine Learning-ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I 27* (Springer, 2018), pp. 382–391
5. J.T. Hoe, K.W. Ng, T. Zhang, C.S. Chan, Y.Z. Song, T. Xiang, *Advances in Neural Information Processing Systems* **34**, 24286 (2021)
6. O. Ferreira, A. Iusem, S. Németh, *Top* **22**, 1148 (2014)
7. C.C. Robusto, *The American Mathematical Monthly* **64**(1), 38 (1957)
8. N. Boumal, *An Introduction to Optimization on Smooth Manifolds* (Cambridge University Press, 2023)
9. S. Hosseini, *Set-Valued and Variational Analysis* **25**, 297 (2017)
10. A. Hauswirth, S. Bolognani, G. Hug, F. Dörfler, in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (IEEE, 2016), pp. 225–232
11. W. Huang, K. Wei, *Mathematical Programming* **194**(1-2), 371 (2022)
12. S. Chen, S. Ma, A. Man-Cho So, T. Zhang, *SIAM Journal on Optimization* **30**(1), 210 (2020)
13. C.L. Lawson, Ph. D. dissertation. University of California, Los Angeles (1961)
14. M. Weber, S. Sra, *Mathematical Programming* **199**(1-2), 525 (2023)
15. C. Liu, N. Boumal, *Applied Mathematics & Optimization* **82**, 949 (2020)
16. R. Jagannathan, *Management Science* **12**(7), 609 (1966)
17. W. Dinkelbach, *Management science* **13**(7), 492 (1967)
18. I.M. Stancu-Minasian, *Fractional Programming: Theory, Methods and Applications*, vol. 409 (Springer Science & Business Media, 2012)
19. D.G. Luenberger, *Management Science* **18**(11), 620 (1972)
20. D. Gabay, *Journal of Optimization Theory and Applications* **37**, 177 (1982)
21. A. Edelman, T.A. Arias, S.T. Smith, *SIAM journal on Matrix Analysis and Applications* **20**(2), 303 (1998)
22. P.A. Absil, R. Mahony, R. Sepulchre, *Optimization Algorithms on Matrix Manifolds* (Princeton University Press, 2008)
23. C. Udriste, *Convex functions and optimization methods on Riemannian manifolds*, vol. 297 (Springer Science & Business Media, 2013)
24. T. Rapcsák, *Smooth Nonlinear Optimization in  $R^n$* , vol. 19 (Springer Science & Business Media, 2013)
25. G.C. Bento, O.P. Ferreira, J.G. Melo, *Journal of Optimization Theory and Applications* **173**(2), 548 (2017)

26. W.H. Richardson, *Journal of the Optical Society of America* **62**(1), 55 (1972)
27. L.B. Lucy, *Astronomical Journal*, Vol. 79, p. 745 (1974) **79**, 745 (1974)
28. M.E. Daube-Witherspoon, G. Muehllehner, *IEEE transactions on medical imaging* **5**(2), 61 (1986)
29. J. Yoo, S. Choi, in *International conference on intelligent data engineering and automated learning* (Springer, 2008), pp. 140–147
30. J.M. Lee, *Introduction to Riemannian Manifolds*, vol. 2 (Springer, 2018)
31. C.D. Meyer, I. Stewart, *Matrix Analysis and Applied Linear Algebra* (SIAM, 2023)
32. S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *Foundations and Trends® in Machine learning* **3**(1), 1 (2011)
33. X.R. Feng, H.C. Li, R. Wang, Q. Du, X. Jia, A. Plaza, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **15**, 4414 (2022)
34. A.M.S. Ang, N. Gillis, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **12**(12), 4843 (2019)
35. Y. Du, C.I. Chang, H. Ren, C.C. Chang, J.O. Jensen, F.M. D'Amico, *Optical engineering* **43**(8), 1777 (2004)
36. M. Ramana, A.J. Goldman, *Journal of Global Optimization* **7**(1), 33 (1995)