



A new Riemannian Multiplicative Updates for Chordal Nonnegative Matrix Factorization

Flavia Esposito¹ · Andersen Ang²

Received: 11 July 2024 / Accepted: 29 September 2025
© The Author(s) 2025

Abstract

Nonnegative Matrix Factorization (NMF) is the problem of approximating a given non-negative matrix $\mathbf{M}$ through the product of two nonnegative low-rank matrices $\mathbf{W}$ and $\mathbf{H}$. 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 (F-norm). In this work, we argue that the F-norm, as the “point-to-point” distance, may not always be appropriate. Viewing from the perspective of cone, NMF may not naturally align with F-norm. So, a ray-to-ray chordal distance is proposed as an alternative way of measuring the quality of the approximation. As this measure corresponds to the Euclidean distance on the sphere, it motivates the use of manifold optimization techniques. We apply Riemannian optimization technique to solve chordal-NMF by casting it on a manifold. Unlike works on Riemannian optimization that require the manifold to be smooth, the nonnegativity in chordal-NMF defines a non-differentiable manifold. We propose a Riemannian Multiplicative Update (RMU), and showcase the effectiveness of the chordal-NMF on synthetic and real-world datasets.

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

Mathematics Subject Classification 15A23 · 78M50 · 49Q99 · 90C26 · 90C30

1 Introduction

Given a nonnegative matrix $\mathbf{M} \in \mathbb{R}_+^{m \times n}$ and a rank $r \leq \min\{m, n\}$, the goal of 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 $\mathbf{M} \approx \mathbf{WH}$ [1]. NMF is commonly achieved by minimizing the Frobenius norm (F-

F. Esposito and A. Ang: These authors contributed equally to this work.

✉ Flavia Esposito
flavia.esposito@uniba.it

¹ Department of Mathematics, Università degli Studi di Bari Aldo Moro, Bari, Italy

² School of Electronics and Computer Science, University of Southampton, Southampton, United Kingdom

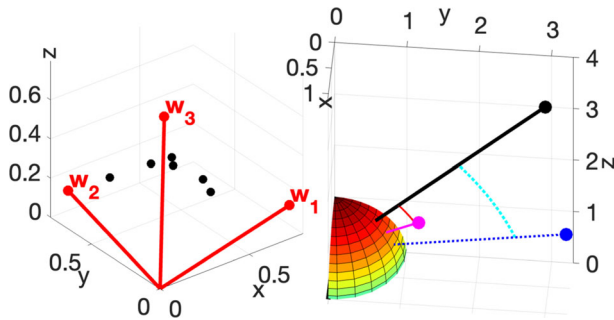


Fig. 1 Left: Picture of a rank-3 NMF. Data points (in black) that represent the columns of $\mathbf{M}$ encapsulated by a polyhedral cone generated by $r = 3$ columns $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ (in red). Right: The sphere $\mathbb{S}_+^2$ in the nonnegative orthant $\mathbb{R}_+^3$. The 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

norm) $\|\mathbf{M} - \mathbf{WH}\|_F = \sqrt{\sum_{ij} (M_{ij} - (WH)_{ij})^2}$ (where M_{ij} is the (i, j) th-entry of $\mathbf{M}$) that measures the quality of the approximation.

For $\mathbf{M} \approx \mathbf{WH}$ with nonnegativity constraints, the point cloud $\mathbf{M}$ 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 $M_{ij} - (WH)_{ij}$ in the F-norm does not naturally fit NMF. In this work, we propose to replace F-norm by a ray-to-ray distance that we call *chordal distance*, detailed in the next section. We are interested in solving

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

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 $\mathbf{WH}$ 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 nonnegativity constraints $\geq \mathbf{0}$ is element-wise, where $\mathbf{0}$ is zero matrix of the appropriate size. For simplicity, we assume $\mathbf{M} \in \mathbb{R}_+^{m \times n}$ has no zero column, and $\mathbf{M}$ is normalized that $\|\mathbf{m}_{:j}\|_2 = 1$ in (Chordal-NMF). We give the motivation of using ray-to-ray distance in §1.1.

Contribution. Our contributions are 3-folds.

1. We propose a new model (Chordal-NMF). To the best of our knowledge, such problem is new and has not been studied in the NMF literature.
2. Solving Chordal-NMF is nontrivial, 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 derive the Riemannian gradient of the objective function F .
 - The nonnegativity constraints in Chordal-NMF introduce nondifferentiability in the manifold and make some manifold techniques ineffective. We propose RMU to solve the nonsmoothness issue in the 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 in §4.

Paper organization. In the remaining of this section, we describe the overall algorithmic framework on solving Chordal-NMF, as well as the motivation of the chordal distance. In §2 we review Riemannian optimization techniques. In §3.1 and §3.2 we discuss how we update the variable $\mathbf{H}$ and $\mathbf{W}$, respectively. §4 contains the experiments and §5 sketches the conclusion.

Notation. We use {italic, bold italic, bold italic capital} letters to denote {scalar, vector, matrix} resp.. For a matrix $\mathbf{A}$, we denote $\mathbf{a}_{:j}$ its j th column, $\mathbf{a}_j$: its j th row, and $\mathbf{A}^\top$ its transpose. We denote $(\mathbf{M})_{ij}$ or M_{ij} the ij component of $\mathbf{M}$. The notation $\langle \xi, \zeta \rangle$ denotes the Euclidean inner product in the standard basis, and $\|\xi\|_2$ denotes the Euclidean norm of ξ . For a vector $\mathbf{v} \neq \mathbf{0}$, we denote $\hat{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|_2$ the unit vector of $\mathbf{v}$. If $\mathbf{v} = \mathbf{0}$ we define $\hat{\mathbf{v}}$ to be any unit vector. We use $[\theta]_+ := \max\{0, \theta\}$ element-wise.

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

Proposition 1 Let $f(\mathbf{x}) = \langle \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{c} \rangle / \|\mathbf{D}\mathbf{x} + \mathbf{e}\|_2$ with $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{c} \in \mathbb{R}^m$, $\mathbf{D} \in \mathbb{R}^{p \times n}$, $\mathbf{e} \in \mathbb{R}^p$ and $\mathbf{D}\mathbf{x} + \mathbf{e} \neq \mathbf{0}$, the Euclidean gradient ∇f , 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.

We make use of tensor product in this work. Let V and W be two vector spaces with inner product. Let $\mathbf{v} \otimes \mathbf{w} \in V \otimes W$ be the tensor product of $\mathbf{v} \in V$, $\mathbf{w} \in W$, then

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

Block Coordinate Descent (BCD). We use BCD to solve the Chordal-NMF, see Algorithm 1, where we also hide the factor $1/n$.

Algorithm 1 BCD for Chordal-NMF, with a starting point $(\mathbf{W}_0, \mathbf{H}_0)$

```

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

Chordal-NMF is asymmetric. Usually considering that NMF in F-norm is symmetric: $\|\mathbf{M} - \mathbf{W}\mathbf{H}\|_F = \|\mathbf{M}^\top - \mathbf{H}^\top \mathbf{W}^\top\|_F$, we can use the same procedure updating $\mathbf{H}$ to update $\mathbf{W}$ (with a transpose). Chordal-NMF is asymmetric: it measures the cosine distance between $\mathbf{m}_{:j}$ and $\mathbf{W}\mathbf{h}_{:j}$, not between the rows $\mathbf{m}_i$: and $\mathbf{w}_{i,:}$: $\mathbf{H}$. This leads to the W-subproblem and H-subproblem have different structure, and we use different approaches to solve the subproblems. We solve H-subproblem column-wise as in line 3 in Algorithm 1, to be discussed in §3.1. We solve W-subproblem matrix-wise, to be discussed in §3.2.

1.1 Deriving and motivating chordal distance

From the conic view of NMF (see introduction), the geometry of chordal distance can be seen as the Euclidean distance between unit vectors on the sphere. Writing $\|\mathbf{M} - \mathbf{W}\mathbf{H}\|_F^2 = \sum_j \|\mathbf{m}_{:j} - \mathbf{W}\mathbf{h}_{:j}\|_2^2$, if both $\mathbf{m}_{:j}$ and $\mathbf{W}\mathbf{h}_{:j}$ have unit ℓ_2 -norm, then by the Euclidean inner product $\langle \xi, \zeta \rangle = \|\xi\|_2 \|\zeta\|_2 \cos \theta(\xi, \zeta)$ with $\theta(\xi, \zeta)$ the angle between the vectors ξ, ζ , we have

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

While the normalization of $\mathbf{m}_{:j}$ is assumed, the normalization of $\mathbf{W}\mathbf{h}_{:j}$ is important here. Note that enforcing $\mathbf{W}\mathbf{h}_{:j}$ to have unit ℓ_2 -norm forms a nonconvex constraint.

The expression (2) tells that the normalized Euclidean distance is the cosine distance. Cosine distance is insensitive to the length of vector, see Fig. 1 for an example.

- **Quotient space interpretation.** From (2), we define a new objective in (Chordal-NMF) for measuring the distance between $\mathbf{m}_{:j}$ and $\mathbf{W}\mathbf{h}_{:j}$ by angle. The division of the norm $\|\mathbf{W}\mathbf{h}_{:j}\|_2$ in (Chordal-NMF) collapses all vectors $\mathbf{x}$ that are proportional to $\mathbf{W}\mathbf{h}$ onto a single point on the unit sphere. In this way, the chordal distance becomes purely angle-based, since the length information is ignored.
- **Haversine interpretation.** Navy navigation [2] makes use of the haversine $\text{hav}(\theta) := (1 - \cos \theta)/2$, where $\theta = d/r$ is 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 $\text{hav}(\theta)$ with $r = 1$.
- **Sphere interpretation.** 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 map $\xi \mapsto \xi/\|\xi\|_2$ in (Chordal-NMF) sends nonzero vector $\xi \in \mathbb{R}^m$ to the (nonconvex) unit sphere $\mathbb{S}^{m-1} := \{\xi \in \mathbb{R}^m \mid \|\xi\|_2 = \sqrt{\langle \xi, \xi \rangle} = 1\}$. Considering $\mathbf{u} \neq \mathbf{0}, \mathbf{v} \neq \mathbf{0}$, the following function

$$\tilde{f}_{\text{chord}} : \mathbb{R}^m \times \mathbb{R}^m \rightarrow [0, 2] \cup \{\infty\} \cup \{\pm\sqrt{-1}\infty\} \quad \text{as} \quad \mathbf{u}, \mathbf{v} \mapsto \sqrt{2 - 2 \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2}},$$

can be seen as the Euclidean distance between two unit vectors $\hat{\mathbf{u}}, \hat{\mathbf{v}}$ on $\mathbb{S}^{m-1}$, since

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

For this function it should be noted that:

- $\tilde{f}_{\text{chord}}$ is undefined at $\mathbf{u} = \mathbf{0}$ and/or $\mathbf{v} = \mathbf{0}$, it take ∞ or complex values $\pm\sqrt{-1}\infty$.
- $\tilde{f}_{\text{chord}}$ is nondifferentiable with respect to (wrt.) parallel unit vectors as $\tilde{f}_{\text{chord}}(\hat{\mathbf{u}}, \hat{\mathbf{v}}) := \sqrt{2 - 2} = 0$.
- In the Euclidean case, $\tilde{f}_{\text{chord}}$ is non-convex wrt. $\mathbf{u}, \mathbf{v}$ and the pair $(\mathbf{u}, \mathbf{v})$.

These undesirable properties lead to choice of chordal distance below.

Definition 1 On $\mathbb{S}^{m-1}$ with inner product $\langle \mathbf{u}, \mathbf{v} \rangle$, we define

$$f_{\text{sq-chord}} : \mathbb{S}^{m-1} \times \mathbb{S}^{m-1} \rightarrow [0, 4] : \mathbf{u}, \mathbf{v} \mapsto 2 - 2\langle \mathbf{u}, \mathbf{v} \rangle.$$

Given the above discussion and the fact that the chordal distance is non-Euclidean, we employ Riemannian optimization methods in this work.

2 Background of optimization on manifold

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

Let E be 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 differentiable function $f : E \rightarrow \mathbb{R}$ and the nonnegative orthant $\mathcal{X} := \{x \in E \mid x \geq \mathbf{0}\}$, consider the minimization problems

$$\mathcal{P}_0 : \operatorname{argmin}_x f(x) \text{ s.t. } x \in \mathcal{M}, x \in \mathcal{X} \xleftrightarrow{\text{restriction}} \mathcal{P}_1 : \operatorname{argmin}_{x \in \mathcal{M}} f(x) \text{ s.t. } x \in \mathcal{X},$$

with the constraint set $\mathcal{M} := \{x \in E \mid h(x) = 0\}$ under the defining function $h(x)$. We focus on convex compact set $\mathcal{M}$ being a smooth embedded submanifold of $\mathbb{R}^n$, where $h(x)$ is many-times continuously differentiable.

In Euclidean optimization, $x \in \mathcal{M}$ is a constraint in $(\mathcal{P}_0)$. We “remove” such constraint using a restricted function $f|_{\mathcal{M}} : \mathcal{M} \rightarrow \mathbb{R}$. Here f is the *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 a constrained manifold optimization $(\mathcal{P}_1)$, which can further be converted into an unconstrained problem using indicator $\iota(x) : E \rightarrow \mathbb{R} \cup \{+\infty\}$ that $\iota(x) = 0$ if $x \in \mathcal{X}$ and $\iota(x) = +\infty$ if $x \notin \mathcal{X}$. The problems are related, solving one of them will help solving the others. Below we review approaches in manifold optimization on solving these problems.

2.1 Nonnegative-constrained manifold optimization

Solving $(\mathcal{P}_0)$ directly by manifold techniques is nontrivial. Manifold optimization refers to an optimization on a smooth (differentiable) manifold. The set $\mathcal{X}$ is nonsmooth, same for $\mathcal{M}_+ := \mathcal{M} \cap \mathcal{X}^1$ (assumed nonempty). Thus, several manifold optimization techniques do not have convergence guarantee for solving $(\mathcal{P}_0)$. The indicator ι is a nonsmooth convex lower semicontinuous (l.s.c) function [3], making $(\mathcal{P}_1)$ written in indicator form not satisfying the assumptions in some existing works in nonsmooth Riemannian optimization, e.g. Projected Gradient Descent on Riemannian Manifolds [4], Riemannian proximal gradient [5], and Manifold proximal gradient [6].

Dual methods: not strictly feasible. We can also solve $(\mathcal{P}_0)$ by Riemannian Augmented Lagrangian multiplier (RALM) [7] or solve $(\mathcal{P}_1)$ using Riemannian Alternating Direction Method of Multipliers (RADMM) [8]. The sequence generated by these methods is not strictly-primal-feasible and it is dangerous to use an infeasible variable to update the other block of variables in the BCD framework. Therefore they are not feasible for our application.

Euclidean projected gradient is infeasible. We can solve $(\mathcal{P}_0)$ by Euclidean projected gradient descent (EPG). The projection $\operatorname{proj}_{\mathcal{M}_+}(z) = \operatorname{argmin}_{x \in \mathcal{M}_+} \frac{1}{2} \|x - z\|_E^2$ takes an alternating

projection method such as the Dykstra’s algorithm [9] to solve, which can be expensive for our purpose.

Manifold projection-free method is expensive. Projection-free method, such as the Riemannian Frank-Wolfe (RFW) [10], has an expensive subproblem that requires the computation of exponential map and geodesic.

Our solution: Riemannian multiplicative update (RMU). In this work we provide a projection-free method that only requires the computation of the Riemannian gradient part

¹ For any point x in the boundary of $\mathcal{M}_+$, there does not exist a $\mathbb{R}^n$ -homeomorphic neighborhood in $\mathcal{M}$.

$\text{grad} f$. To tackle these technical issues, we propose a cost-effective Riemannian multiplicative update (RMU) for solving Problem (P_1) . Our approach is motivated by the research of NMF [1, §8.2]. The advantages of RMU are

- The expensive projection $\text{proj}_{\mathcal{M}_+}$ 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.

We introduce RMU in §2.3, before that we review Riemannian optimization below.

2.2 Background of Riemannian optimization

Riemannian optimization, or manifold optimization, has a long history [11–18]. So we give the minimum material on manifold optimization for the paper.

Submanifold and ambient space. Let E be an Euclidean space with $\dim(E) > k$, with inner product $\langle \cdot, \cdot \rangle$, and induced norm $\| \cdot \|$. Considering the smooth function $h : E \rightarrow \mathbb{R}^k$, with $Dh(\mathbf{x})[\mathbf{v}] = \langle \text{grad} h(\mathbf{x}), \mathbf{v} \rangle_E$ its differential with full rank k for all $\mathbf{x}$ such that $h(\mathbf{x}) = \mathbf{0}_k$, the set $\mathcal{M} := \{\mathbf{x} \in E \mid h(\mathbf{x}) = \mathbf{0}_k\}$ is an embedded submanifold of E of dimension $\dim E - k$ (see [18, Ch7.7]). In this work we focus on h in the form $h(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x} \rangle_E - 1$. For a manifold $\mathcal{M}$ in E , we call E the ambient space of $\mathcal{M}$. In this work, all ambient spaces are some specific linear vector space E , such as $\mathbb{R}^r$ and $\mathbb{R}^{m \times r}$.

- In the h -subproblem, $\mathcal{M}$ is the “shell” of an ellipsoid, it has 1 dimension lower than its ambient Euclidean space $\mathbb{R}^r$.
- In the W -subproblem, $\mathcal{M}$ is the “shell” of a twisted spectrahedron.

Tangent space and projection. Let $\ker$ be the kernel of a matrix, 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., $T_{\mathbf{x}}\mathcal{M} := \ker Dh(\mathbf{x}) = \{\mathbf{v} \in E \mid Dh(\mathbf{x})[\mathbf{v}] = 0\}$. $T_{\mathbf{x}}\mathcal{M}$ collects vectors $\mathbf{v} \in E$ that is tangent to $\mathcal{M}$ at $\mathbf{x}$. The orthogonal projection to $T_{\mathbf{x}}\mathcal{M}$ is the linear map $\text{proj}_{T_{\mathbf{x}}\mathcal{M}} : \mathcal{M} \rightarrow T_{\mathbf{x}}\mathcal{M}$ characterized by the following properties [18, Ch3, Def3.60]:

- Range: $\text{Im}(\text{proj}_{T_{\mathbf{x}}\mathcal{M}}) = T_{\mathbf{x}}\mathcal{M}$;
- Projection: $\text{proj}_{T_{\mathbf{x}}\mathcal{M}} \circ \text{proj}_{T_{\mathbf{x}}\mathcal{M}} = \text{proj}_{T_{\mathbf{x}}\mathcal{M}}$;
- Orthogonality: $\langle \mathbf{u} - \text{proj}_{T_{\mathbf{x}}\mathcal{M}}(\mathbf{u}), \mathbf{y} \rangle = 0$ for all $\mathbf{y} \in T_{\mathbf{x}}\mathcal{M}$ and $\mathbf{u} \in \mathcal{M}$.

In this work, we have manifolds defined by smooth functions h [18, §7.7], we use the orthogonal projection $\text{proj}_{T_{\mathbf{x}}\mathcal{M}} : E \rightarrow T_{\mathbf{x}}\mathcal{M}$, 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 (technically called covector) as the solution of $\arg\min_{\alpha \in \mathbb{R}} \|\mathbf{v} - Dh(\mathbf{x})^*[\alpha]\|_E^2 = (Dh(\mathbf{x})^*)^\dagger[\mathbf{v}]$, where $\dagger$ is pseudo-inverse. This gives $\text{proj}_{T_{\mathbf{x}}\mathcal{M}}(\mathbf{v}) = \mathbf{v} - Dh(\mathbf{x})^*[(Dh(\mathbf{x})^*)^\dagger[\mathbf{v}]]$. Note that if the tangent space $\text{proj}_{T_{\mathbf{x}}\mathcal{M}}$ equals the ambient space, then the projection is not necessary.

Remark 1 (Notational difference) In [18, Ch3, Def3.60] $\text{proj}_{\mathbf{x}}$ denotes orthogonal projection, here for indicating orthogonal projection from E to $T_{\mathbf{x}}\mathcal{M}$, we use $\text{proj}_{T_{\mathbf{x}}\mathcal{M}}$.

Retraction. A point $\mathbf{x}$ that is originally sitting on a manifold $\mathcal{M}$ may go outside $\mathcal{M}$ after a gradient operation, so we need to pull it back onto $\mathcal{M}$. This can be achieved by a particular smooth map, known as the retraction $\mathcal{R} : T_{\mathbf{x}}\mathcal{M} \rightarrow \mathcal{M}$ with $\mathbf{v} \mapsto \mathcal{R}_{\mathbf{x}}(\mathbf{v})$. A retraction maps a point on the tangent space $T_{\mathbf{x}}\mathcal{M}$ onto $\mathcal{M}$ such that each curve $c(t) = \mathcal{R}_{\mathbf{x}}(t\mathbf{v})$ satisfies $c(0) = \mathbf{x}$ and $c'(0) = \mathbf{v}$ [18, Ch3, Def3.47]. There are different ways to perform retractions,

such as exponential map and metric retractions [18, §3.6]. However, in general, to obtain the exponential map of a manifold, one requires to solve a differential equation. Thus, for computational efficiency, we consider the metric retraction $\mathcal{R}_x(\mathbf{v}) = \operatorname{argmin}_{\mathbf{y} \in E} \|\mathbf{x} + \mathbf{v} - \mathbf{y}\|_E^2$ s.t. $h(\mathbf{y}) = 0$ [18, §7.7]. Note that $\mathcal{R}_x(\mathbf{v})$ is possibly non-unique and hard to compute. Among all the retractions we focus on

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

Riemannian and Euclidean gradient. Given a (restricted) function $f|_{\mathcal{M}} : \mathcal{M} \rightarrow \mathbb{R}$ with its smooth extension $f : E \rightarrow \mathbb{R}$, let $\nabla f(\mathbf{x})$ be 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 $\operatorname{grad} f|_{\mathcal{M}}$, at a point $\mathbf{z}$, wrt. the reference point $\mathbf{x}$, is the Euclidean gradient $\nabla f(\mathbf{z})$ projected onto $T_{\mathbf{x}}\mathcal{M}$. That is $\operatorname{grad} f|_{\mathcal{M}}(\mathbf{z}) = \operatorname{proj}_{T_{\mathbf{x}}\mathcal{M}} \nabla f(\mathbf{z})$. For the sake of cheap computational cost, we endow the manifold with the standard Euclidean metric². To ease the notation, we write $f|_{\mathcal{M}}$ as f taking the natural extension.

2.3 Riemannian Multiplicative Update (RMU)

In this work, we propose a RMU for solving Chordal-NMF in the form of $(\mathcal{P}_1)$. RMU is a modified Riemannian gradient descent (RGD) update: $\mathbf{x}_{k+1} = \mathcal{R}_{\mathbf{x}_k}(\alpha \mathbf{v}_k)$, where $\mathbf{v}_k = -\operatorname{grad} f(\mathbf{x}_k)$ is the (negative) Riemannian gradient of f at $\mathbf{x}_k$. RMU is projection-free, and it guarantees the nonnegativity of $\mathbf{x}$ by selecting a special element-wise stepsize such that $\mathbf{x}_{k+1}$ stays within $\mathcal{X}$. Below we review Euclidean Multiplicative Update (EMU) and then we generalize it to the Riemannian case.

Euclidean MU (EMU). EMU was first proposed in [19–21], and gained popularity in NMF [1]. EMU can be done via a “element-wise sign decomposition” of the Euclidean gradient $\nabla f(\mathbf{x}) = \nabla^+ f(\mathbf{x}) - \nabla^- f(\mathbf{x})$, where $\nabla^+ f(\mathbf{x}) \geq \mathbf{0}$ and $\nabla^- f(\mathbf{x}) \geq \mathbf{0}$. Let denote $\odot$ the element-wise product and $\oslash$ the element-wise division, for solving a nonnegative-constrained Euclidean optimization problem, the EMU step has the form

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

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 [22] for the derivation³.

RMU. Now we generalize EMU. First, a Riemannian gradient $\operatorname{grad} f(\mathbf{x})$ always admits element-wise sign decomposition⁴ $\operatorname{grad} f(\mathbf{x}) = \operatorname{grad}^+ f(\mathbf{x}) - \operatorname{grad}^- f(\mathbf{x})$.

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

Proposition 2 (RMU) *Denote $\mathbf{v}_k = -\operatorname{grad} f(\mathbf{x}_k)$ the anti-parallel direction of the Riemannian gradient of a manifold $\mathcal{M}$ at $\mathbf{x}_k$, and let $\mathcal{R}_{\mathbf{x}_k}$ be the metric retraction onto $\mathcal{M}$.*

² 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 g , then $\operatorname{grad} f|_{\mathcal{M}}(\mathbf{z}) = G_{\mathbf{x}}^{-1} \operatorname{proj}_{T_{\mathbf{x}}\mathcal{M}} \nabla f(\mathbf{z})$.

³ In the literature (EMU) is 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.

⁴ 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 $\mathbf{b}^- = \max(\mathbf{0}, -\mathbf{b})$.

If a nonnegative $\mathbf{x}_k$ is updated by $\mathbf{x}_{k+1} = \mathcal{R}_{\mathbf{x}_k}(\boldsymbol{\alpha} \odot \mathbf{v}_k)$ with an element-wise stepsize $\boldsymbol{\alpha} = \mathbf{x}_k \oslash \text{grad}^+ f(\mathbf{x}_k) \in E$, then $\mathbf{x}_{k+1} \geq \mathbf{0}$ and is on $\mathcal{M}$.

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

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

The element-wise operations as linear transformations in general do not preserve the tangent condition of a manifold into a point. However here we can still apply (Metric retraction) on $-\boldsymbol{\alpha} \odot \text{grad} f(\mathbf{x}_k)$, the $\mathbf{x}_k$ terms got canceled and gives

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

The ratio between $\text{grad}^+ f$ and $\text{grad}^- f$ is nonnegative as its parts are nonnegative by definition, the denominator is nonnegative, therefore $\mathbf{x}_{k+1}$ is nonnegative if $\mathbf{x}_k$ is nonnegative. $\square$

Proposition 2 also gives a convergence condition to a critical point:

- If the solution lies strictly inside the nonnegative orthant, this corresponds to $\text{grad} f(\mathbf{x}_k) = \mathbf{0}$, which is equivalent to $\text{grad}^- f(\mathbf{x}_k) = \text{grad}^+ f(\mathbf{x}_k)$. That is, the convergence of the sequence $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$ can be easily checked by comparing $\text{grad}^- f(\mathbf{x}_k)$ and $\text{grad}^+ f(\mathbf{x}_k)$.
- Otherwise, in the constrained setting, convergence is characterized by the KKT conditions on $\mathcal{M} \cap \mathbb{R}_+^n$.

As RMU is projection-free, it 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 $\mathbf{W}$.

3 Subproblems in the algorithm

In this section, we discuss how to solve the subproblems in Algorithm 1.

3.1 Column-wise h-subproblem over ellipsoid

Here we discuss how to solve the h-subproblem. We re-formulate the h-subproblem on $\mathbf{h}$ over the ellipsoid $\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}$ as

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

where we assume $\mathbf{W}$ has full rank, the ellipsoid $\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1} \subset \mathbb{R}^r$ defined by a Positive Definite (PD) matrix $\mathbf{W}^\top \mathbf{W} \in \mathbb{R}^{r \times r}$ is

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

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 $\|\boldsymbol{\xi}\|_{\mathcal{E}}$. We remark that:

Table 1 Summary of objects. Top: for $\mathbf{h}$ and $\mathbf{A} = \mathbf{W}^\top \mathbf{W}$. Bottom: For $\mathbf{W}$ and $\mathbf{A}_j = \mathbf{h}_{\cdot j} \mathbf{h}_{\cdot j}^\top$

Name / Reference	Definition / expression
Manifold of $\mathbf{h}$ (Ellipsoid)	$\mathcal{E}_A^{r-1} := \{\xi \in \mathbb{R}^r \mid \langle \mathbf{A}\xi, \xi \rangle = 1\}$
Tangent space of $\mathcal{E}_A^{r-1}$ at ζ (3)	$T_\zeta \mathcal{E}_A^{r-1} := \{\xi \in \mathbb{R}^r \mid \langle \mathbf{A}\xi, \zeta \rangle = 0\}$
Project ξ onto $T_\zeta \mathcal{E}_A^{r-1}$ (Proposition 3)	$\text{proj}_{T_\zeta \mathcal{E}_A^{r-1}}(\xi) = \left(\mathbf{I}_r - \frac{(\mathbf{A}\zeta)(\mathbf{A}\zeta)^\top}{\ \mathbf{A}\zeta\ _2^2} \right) \xi$
Retraction (Proposition 4)	$\mathcal{R}_\zeta(\xi) = \frac{\zeta + \xi}{\ \zeta + \xi\ _\mathcal{E}}$
Manifold of $\mathbf{W}$ (Shell of twisted spectrahedron)	$\mathcal{M}_j := \{\mathbf{W} \in \mathbb{R}^{m \times r} \mid \langle \mathbf{W}, \mathbf{W} \rangle_{\mathbf{A}_j^{1/2}} = 1\}$
Tangent space of $\mathcal{M}_j$ at $\mathbf{Z}$ (6)	$T_{\mathbf{Z}} \mathcal{M}_j := \{\mathbf{W} \in \mathbb{R}^{m \times r} \mid \langle \mathbf{W}, \mathbf{Z} \mathbf{A}_j \rangle_{\mathbf{F}} = 0\}$
Project $\mathbf{W}$ onto $T_{\mathbf{Z}} \mathcal{M}_j$ (Proposition 7)	$\text{proj}_{T_{\mathbf{Z}} \mathcal{M}_j}(\mathbf{W}) = \left(\mathbf{I} - \frac{(\mathbf{Z} \mathbf{A}_j)^{\otimes 2}}{\ \mathbf{Z} \mathbf{A}_j\ _{\mathbf{F}}^2} \right) \mathbf{W}$
Retraction (Proposition 8)	$\mathcal{R}_{\mathbf{Z}}(\mathbf{W}) = \frac{\mathbf{Z} + \mathbf{W}}{\ \mathbf{Z} + \mathbf{W}\ _{\mathbf{A}_j^{1/2}}}$

- As a subset of $\mathbb{R}^r$, the set $\mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}$ is a smooth submanifold [18, Def3.10] and we can show that the inner product $\langle \xi, \zeta \rangle_\mathcal{E}$ is a Riemannian metric [18, Prop3.54].
- The set $\mathbb{R}_+^r \cap \mathcal{E}_{\mathbf{W}^\top \mathbf{W}}^{r-1}$ is a nonsmooth manifold: $\mathbb{R}_+^r$ has sharp corners at the boundary. See §2 for discussion on the issues caused by such non-smoothness.

Tools on ellipsoid manifold. The tools to solve (h-manifold-subproblem) by RMU is summarized in Table 1 and detail in the following paragraphs.

Tangent space. Let $\mathbf{A} := \mathbf{W}^\top \mathbf{W}$ and the function $h(\xi) = \langle \mathbf{A}\xi, \xi \rangle - 1$, we denote $\mathcal{E}_A^{r-1}$ as the manifold and its differential $Dh(\xi)[\zeta] = 2\langle \mathbf{A}\xi, \zeta \rangle$. The tangent space of $\mathcal{E}_A^{r-1}$ at a reference point $\zeta \in \mathcal{E}_A^{r-1}$, is defined as the set

$$T_\zeta \mathcal{E}_A^{r-1} := \{\xi \in \mathbb{R}^r \mid \langle \mathbf{A}\xi, \zeta \rangle = 0\} = \text{Ker } Dh(\zeta). \quad (3)$$

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

Proposition 3 The map $\text{proj}_{T_\zeta \mathcal{E}_A^{r-1}} : \mathbb{R}^r \rightarrow T_\zeta \mathcal{E}_A^{r-1}$ that bring ξ onto the tangent space at the reference point, ζ , defined as

$$\text{proj}_{T_\zeta \mathcal{E}_A^{r-1}}(\xi) = \xi - \frac{\langle \mathbf{A}\zeta, \xi \rangle}{\|\mathbf{A}\zeta\|_2^2} \mathbf{A}\zeta, \quad (\text{proj-ellips})$$

is an orthogonal projector.

Proof Using formula (1), we have (proj-ellips) equals to the following matrix form

$$\xi - \frac{\langle A\xi, \xi \rangle}{\|A\xi\|_2^2} A\xi = \left(I_r - \frac{(A\xi)^{\otimes 2}}{\|A\xi\|_2^2} \right) \xi =: P\xi.$$

To show that (proj-ellips) is a projector onto the tangent space of the reference point ξ we use the definition of orthogonal projector in §2.2.

First, the range of $\text{proj}_{T_\xi \mathcal{E}_A^{r-1}}$, denoted as $\text{Im}(\text{proj}_{T_\xi \mathcal{E}_A^{r-1}})$, is exactly $T_\xi \mathcal{E}_A^{r-1}$, since

$$\xi = \xi - \frac{\langle A\xi, \xi \rangle}{\|A\xi\|_2^2} A\xi \iff \frac{\langle A\xi, \xi \rangle}{\|A\xi\|_2^2} A\xi = \mathbf{0} \xLeftrightarrow{A\xi \neq \mathbf{0}} \xi \in T_\xi \mathcal{E}_A^{r-1}.$$

Second, we show that $\text{proj}_{T_\xi \mathcal{E}_A^{r-1}}$ is idempotent by showing P is idempotent

$$P^2 = I_r - \frac{(A\xi)^{\otimes 2}}{\|A\xi\|_2^2} - \underbrace{\frac{(A\xi)^{\otimes 2}}{\|A\xi\|_2^2} + \frac{(A\xi)^{\otimes 2}}{\|A\xi\|_2^2} \frac{(A\xi)^{\otimes 2}}{\|A\xi\|_2^2}}_{=0} = P.$$

We show $\text{proj}_{T_\xi \mathcal{E}_A^{r-1}}$ is orthogonal by showing P is orthogonal. As A is symmetric, so $P^\top = P$. Being symmetric and idempotent, P is orthogonal [23, §0.9.13]. $\square$

Proposition 4 (metric retraction) *The map $\mathcal{R}_\xi : T_\xi \mathcal{E}_A^{r-1} \rightarrow \mathcal{E}_A^{r-1}$ defined as*

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

is a metric retraction for the manifold $\mathcal{E}_A^{r-1}$.

We give the proof of the proposition in Appendix B.

Riemannian Gradient. We solve (h-manifold-subproblem) using RMU discussed in §2.3. The Riemannian gradient $\text{grad}\phi(\mathbf{h}_k) = \text{proj}_{T_{\mathbf{h}_k} \mathcal{E}} \nabla \bar{\phi}_\mathcal{E}(\mathbf{h}_k)$, is

$$\text{grad}\phi(\mathbf{h}_k) = \left(I_r - \frac{(A\mathbf{h}_k)^{\otimes 2}}{\|A\mathbf{h}_k\|_2^2} \right) \left(-W^\top \mathbf{m}_{:j} \right) = \underbrace{\frac{(A\mathbf{h}_k)^{\otimes 2}}{\|A\mathbf{h}_k\|_2^2} W^\top \mathbf{m}_{:j}}_{\text{grad}^+ \phi} - \underbrace{W^\top \mathbf{m}_{:j}}_{\text{grad}^- \phi},$$

where we used both: (proj-ellips), (1), the fact that $\mathbf{h}_k \in \mathcal{E}_A^{r-1}$, and the Euclidean gradient of $\mathbf{h}_k$ through Proposition 1.

Then, the RMU under metric retraction discussed in §2.3 is

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

3.2 The finite sum matrix-wise W-subproblem

Now we discuss how to solve the W-subproblem in Algorithm 1. Let $\langle \cdot, \cdot \rangle_F = \text{Tr}(\cdot^\top \cdot)$ be the Frobenius inner product, $\mathbf{B}_j := \mathbf{m}_{:j} \mathbf{h}_{:j}^\top$ and $\mathbf{A}_j := \mathbf{h}_{:j} \mathbf{h}_{:j}^\top$ (not the A, B defined in §3.1), we rewrite the W-subproblem as

$$\arg\max_{W \geq 0} \left\{ F(W) = \sum_{j=1}^n \frac{\langle \mathbf{B}_j, W \rangle_F}{\sqrt{\langle W, W \mathbf{A}_j \rangle_F}} \right\}. \quad (\text{W-subproblem})$$

Note that if a column $\mathbf{h}_j$ is zero, the corresponding term is removed in the sum.

To solve (W-subproblem), we first consider $n = 1$. We have a manifold which we name “shell of a single twisted spectrahedron”. Table 1 summarizes the results of RMU on such manifold. See Appendix C for the derivation of manifold $\mathcal{M}_j$, the tangent space $T_{\mathbf{Z}}\mathcal{M}_j$, the projection and the retraction.

Lemma 5 (RMU update on $\mathbf{W}$) *If $n = 1$, we update $\mathbf{W}$ as*

$$\mathbf{W}_{k+1} = \frac{\mathbf{W}_k \odot \text{grad}^- f_j(\mathbf{W}_k)[\mathbf{W}_k] \odot \text{grad}^+ f_j(\mathbf{W}_k)[\mathbf{W}_k]}{\left\| \mathbf{W}_k \odot \text{grad}^- f_j(\mathbf{W}_k)[\mathbf{W}_k] \odot \text{grad}^+ f_j(\mathbf{W}_k)[\mathbf{W}_k] \right\|_{\mathbf{A}_j^{1/2}}}, \quad (4)$$

$$\text{where } \text{grad} f_j(\mathbf{W}) = \underbrace{\frac{(\mathbf{Z}\mathbf{A}_j)^{\otimes 2}}{\|\mathbf{Z}\mathbf{A}_j\|_F^2} \mathbf{B}_j}_{\text{grad}^+ f_j(\mathbf{W})} - \underbrace{\mathbf{B}_j}_{\text{grad}^- f_j(\mathbf{W})}.$$

For (W-subproblem) with $n \geq 2$, having multiple fractions makes the problem hard to solve. First, there is no a single manifold to guarantee $\langle \mathbf{W}, \mathbf{W}\mathbf{A}_j \rangle_F = 1$ for all the denominators. Second, the computation cost of manifold approach for $n \geq 2$ is high: it takes $\mathcal{O}(2m^2)$ cost to compute the adjoint, with a projection step that has no closed-form solution. Details are reported in §D of the Appendix.

We do not solve (W-subproblem) with $n \geq 2$ using RMU (which is designed for $n = 1$). For illustration, we also considered a consensus-averaging approach between all the manifolds, which we call avgRMU. The motivation was that the manifolds can conflict with each other in RMU, so averaging provides a natural “relaxation” to fit RMU in the intersection of the manifolds. We remark that later we found that experiments show that this approach performs worse than the Euclidean method, and we do not use it as our main solution.

Furthermore, we try two alternative methods.

EPG for (W-subproblem) $\mathbf{W}_{k+1} = [\mathbf{W}_k - \eta \nabla F(\mathbf{W}_k)]_+$ with a stepsize $\eta \geq 0$. In the update, we compute the gradient by quotient rule, $\nabla \langle \mathbf{W}, \mathbf{W}\mathbf{A}_j \rangle_F = \mathbf{W}\mathbf{A}_j + \mathbf{W}\mathbf{A}_j^\top$, $\mathbf{A}_j^\top = \mathbf{A}_j$ and Proposition 1 as in the following proposition.

Proposition 6 *The Euclidean gradient for $F(\mathbf{W})$ in (W-subproblem) is*

$$\nabla F(\mathbf{W}) = \sum_{j=1}^n \frac{\mathbf{B}_j}{\langle \mathbf{W}, \mathbf{W}\mathbf{A}_j \rangle_F^{1/2}} - \frac{\langle \mathbf{B}_j, \mathbf{W} \rangle_F \mathbf{W}\mathbf{A}_j}{\langle \mathbf{W}, \mathbf{W}\mathbf{A}_j \rangle_F^{3/2}}. \quad (5)$$

Fractional Programming (FP). We solve (W-subproblem) using FP by the Dinkelbach transform [24] and Jagannathan’s iteration [25] with the EPG.

4 Experiment

In this section we first perform testing on verifying the effectiveness of RMU on the sub-problems, then we test the performance of Chordal-NMF on synthetic and real datasets.

4.1 RMU on h-subproblem and W-subproblem

Test on updating H. We compare RMU and methods mentioned in §2.1 on synthetic datasets randomly generated under normal distribution $\mathcal{N}(\mathbf{0}_r, \mathbf{I}_r)$ with negative entries replaced by

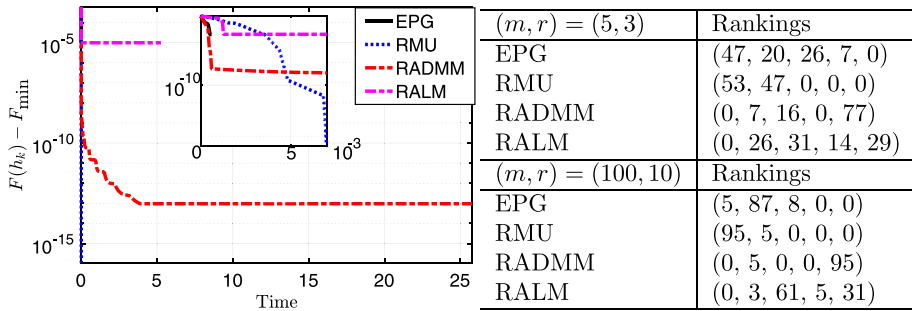


Fig. 2 Left: A typical convergence of the four methods: EPG, RMU, RADMM and RALM. In the y-axis, $F_{\min}$ refers to the smallest achieved F value across the methods. The smaller subplot is zooming the first 7×10^{-3} seconds. In this test, we have $(m, r) = (5, 3)$ and RMU is the fastest methods with strict feasibility. Right: the results on different settings of (m, r) . A ranking vector $(x, y, z, \dots)$ means the method has x times out of 100 runs having the best performance, y times out of 100 runs having the 2nd performance, and so on. There are four methods, and the fifth number indicate how many times the method produces an infeasible solution (having negative values) in the end

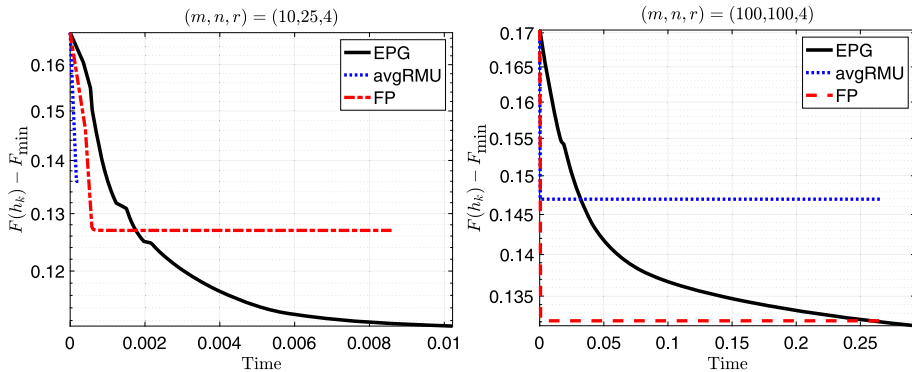


Fig. 3 Typical convergence of three methods: EPG, avgRMU and FP

zero. We performed Monte Carlo trials on datasets with different settings of (m, r) . Fig. 2 shows a typical convergence of the objective function. RMU is among the fastest methods with strict feasibility. RADMM (Riemannian ADMM in §2.1) is also fast, while EPG and RALM (Riemannian Augmented Lagrangian multiplier method) has the worst convergence. In the test, we run at most 10^5 iterations, or stop the algorithm when the $\|\mathbf{h}_{j,k} - \mathbf{h}_{j,k-1}\| \leq 10^{-12}$. **Test on updating $\mathbf{W}$.** As RMU for the whole sum $F(\mathbf{W})$ in (W-subproblem) is computationally infeasible, we consider averaging the RMU for each manifold. We compare the performance of three methods: EPG, avgRMU and FP. We run experiments 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 $(m, n, r) = (10, 25, 4)$ and $(m, n, r) = (100, 100, 4)$. Fig. 3 shows the typical convergence of three methods. Generally EPG is the fastest method. For $(m, n, r) = (10, 25, 4)$, 99 times out of 100 MC runs EPG is the fastest and 96 times out of 100 runs FP is the 2nd fastest. For $(m, n, r) = (100, 100, 4)$, 93 times out of 100 MC runs EPG is the fastest and 93 times out of 100 runs FP is the 2nd fastest. avgRMU is the slowest in all the 100 runs for both cases.

4.2 Testing Chordal-NMF on synthetic dataset

How Chordal-NMF is applied. Given a data matrix $\mathbf{M}$, we remove all zero columns, then we normalize each column by its ℓ_2 -norm. After that we run Chordal-NMF on $\mathbf{M}$ and get the decomposition $\mathbf{W}\mathbf{H}$. According to the preliminary test results, for the h-subproblem, we perform RMU (§3.1), and for the W-subproblem, we perform EPG (§3.2).

Benchmark. We compare Chordal-NMF with the classical F-norm NMF (FroNMF) based on the algorithm HALS [1, Ch8.3.3]. We set the rank of Chordal-NMF equal to the one in FroNMF. In the experiments, all the methods start with the same initialization, where elements in $\mathbf{W}_0$, $\mathbf{H}_0$ are generated under uniform distribution $\mathcal{U}[0, 1]$. The experiments were conducted in MATLAB 2023a⁵ in a machine with OS Windows 11 Pro on a Intel Core 12 gen. CPU 2.20GHz and 16GB RAM.

Synthetic dataset. We use a dataset $\mathbf{M}^{\epsilon, \delta} = \mathbf{W}_{\text{true}} \mathbf{H}_{\text{true}}^{\epsilon, \delta}$ with $\epsilon > 0$, $\delta > 0$ as

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

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 ϵ and an attenuation δ . The perturbation ϵ represents how much the columns of $\mathbf{M}^{\epsilon, \delta}$ deviate from the columns of $\mathbf{W}_{\text{true}}$, while the attenuation δ 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 δ 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 δ . See Fig. 4 for an illustration.

We construct $\mathbf{M}^{\epsilon, \delta}$ across different values of (ϵ, δ) , 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 heat-map of the results across different values of (ϵ, δ) , showing that Chordal-NMF on average has a better recovery of $\mathbf{H}$ than FroNMF, especially for the case when δ is small. We also compute the relative error on the matrix $\mathbf{W}$, where we found that the result on Chordal-NMF gives exactly the same performance as the one obtained from the FroNMF⁶.

4.3 Testing Chordal-NMF on real-world data

NMF finds applications in real-world data analysis [1]. An example is Earth Observation (EO) in remote sensing. EO typically involves the utilization of hyperspectral images, which are stored as matrices and can be effectively analyzed through NMF for unmixing purposes [26]. A conventional approach in this context involves addressing the standard NMF problem by minimizing the point-to-point F-norm [27]. However, in this section, we introduce a comparative analysis between the performance of FroNMF and the Chordal-NMF proposed in this work.

The EO Dataset. We chose a cloudy multispectral image due to its relevance in highlighting the advantages of employing Chordal-NMF compared to standard FroNMF for analyzing

⁵ The MATLAB code is available at https://github.com/flaespo/Chordal_NMF.

⁶ Both Chordal-NMF and FroNMF gives a 4.02% relative error on $\mathbf{W}$.

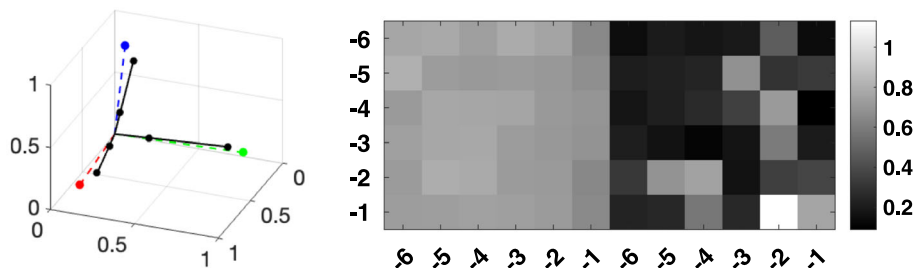


Fig. 4 **Left:** the plot of W_{true} (the red, blue, green ray) and $M(0.1, 0.3)$ (the black rays). **Right:** The relative error $\|H - H^{\epsilon, \delta}_{\text{true}}\|_F / \|H^{\epsilon, \delta}_{\text{true}}\|_F$ for the two methods, where left grid (the first six columns) is the result from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

multispectral images under various cloudiness conditions. Our goal is to provide an illustration of how Chordal-NMF can better manage the presence of cloud conditions, thereby presenting itself as a useful alternative for conducting image analysis for EO. Here the matrix M is a cloudy multispectral image in an area of Apulia region in Italy, from the Copernicus space ecosystem⁷. We use a reference image obtained in cloudless conditions as the ground truth⁸.

The reconstruction. We run FroNMF and Chordal-NMF on the cloudy data M with the same initialization W_0, H_0 obtained from random uniform distribution $\mathcal{U}[0, 1]$. We run Chordal-NMF and FroNMF with 5000 iterations. For FroNMF, we run the implementation from [1]. For Chordal-NMF, we run BCD with 25 iterations for each column on $h_{:,j}$ to update H , and 1 iteration of matrix-wise EPG update for W . The values of the Chordal function for the methods are: 0.1906100 (Initial Value), 0.0014398 (Chordal-NMF), 0.001279 (FroNMF). It took ChordalNMF 2775 seconds (about 0.5 second per iteration), while it took FroNMF 2 seconds. Fig. 5 shows an RGB representation of the ground truth (cloudless reference image), the cloudy data M , and the reconstruction obtained from Chordal-NMF and FroNMF on M .

Further analysis on the three selected areas. In Fig. 5, 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 Fig. 6. We numerically compared these spectral profile vectors by two criteria:

1. **SID-SAM between the spectral profile vectors.** Given two spectral profile vectors $t = [t_1, \dots, t_N]$ and $r = [r_1, \dots, r_N]$, we compute $p = t / \|t\|_2$, $q = r / \|r\|_2$, and then we compute the SID-SAM [28] of the spectral information divergence (SID) and the spectral angle mapper (SAM), defined as $\text{SID-SAM} = \text{SID} \times \tan \alpha$, where

$$\text{SID} = \sum_{i=1}^N p_i \log \frac{p_i}{q_i} + \sum_{i=1}^N q_i \log \frac{q_i}{p_i}, \quad \alpha = \cos^{-1} \left(\frac{\sum_{i=1}^N t_i r_i}{\sqrt{\left(\sum_{i=1}^N t_i^2 \right) \left(\sum_{i=1}^N r_i^2 \right)}} \right).$$

2. **ℓ_2 -norm of the difference between the spectral profile vectors.** That is, $\|t - t_{\text{ChordalNMF}}\|_2$ and $\|t - t_{\text{FroNMF}}\|_2$.

⁷ Pixel size rows:150, cols:290 and bands:12. Data from <https://dataspace.copernicus.eu/>

⁸ The cloudy image and the cloudless image are obtained in the same condition (except the cloudiness) with a temporal difference of 5 days.

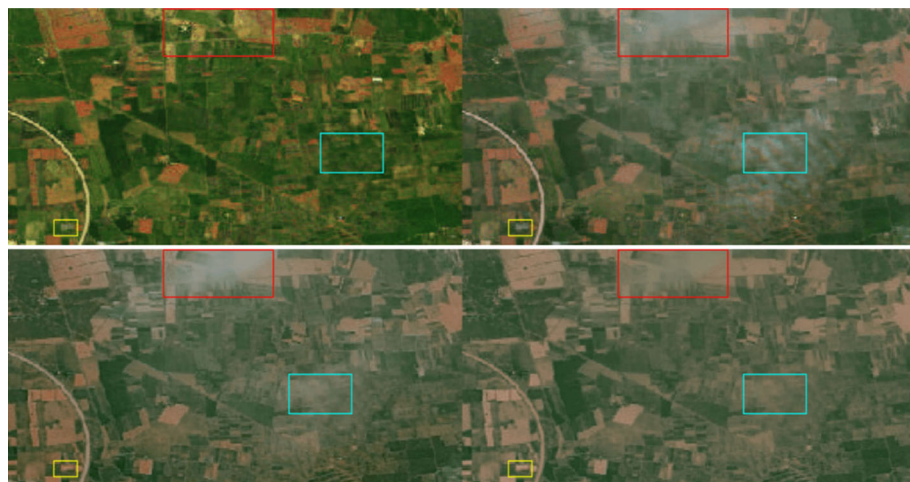


Fig. 5 The RGB image of a scene. Top left: the cloudless reference image. Top right: the cloudy data image. Bottom left: the reconstruction obtained by FroNMF. Bottom right: the reconstruction obtained by Chordal-NMF. In the images, the three color-boxes are selected areas under different cloudiness in which we perform further quantitative analysis

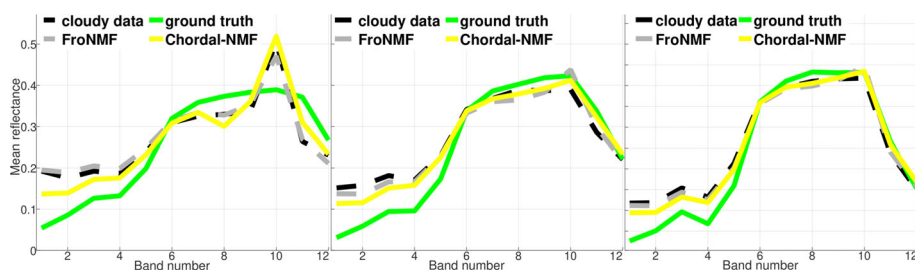


Fig. 6 From left to right: the spectral signatures of the pixels in the red/cyan/yellow box in Fig. 5. In all the three cases, the spectral profile of obtained by Chordal-NMF is on average closer to the cloudless reference spectral profile

Table 2 The numerics of the pixels in the three boxes

Compared with cloudless	Red box		Cyan box		Yellow box	
	SID-SAM	ℓ_2	SID-SAM	ℓ_2	SID-SAM	ℓ_2
Cloudy data	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

Results are reported in Table 2 for the three areas, in which Chordal-NMF achieves a better performance.

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 ℓ_2 -norm value. For example, in the red box, Chordal-NMF seems to have a better recovery of the region under the cloud.

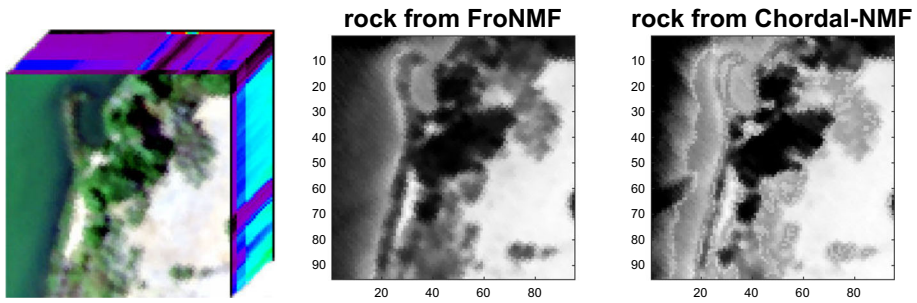


Fig. 7 The Samson dataset and the rock abundances map of the decompositions. From left to right: the data, the abundance maps obtained from FroNMF and Chordal-NMF

Performance on Samson dataset We test our approach also on a benchmark dataset for EO applications: Samson dataset [1]. In this image, there are 95×95 pixels, each pixel is recorded at 156 channels. This dataset is not challenging as many analyses have been already carried out [27]. 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. We can see that the Chordal-NMF is able to recover rock under shadow water regions near the coast better than FroNMF.

5 Conclusion

In this work, we introduced a NMF model called Chordal-NMF, which is different from the classical NMF using a point-to-point Euclidean distance, Chordal-NMF uses a ray-to-ray distance. 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 projection-free algorithm, called Riemannian Multiplicative Update to solve the Chordal-NMF, where Riemannian optimization technique is used. We showcase the effectiveness of the Chordal-NMF on the synthetic dataset and real-world multispectral images.

A The proof of Proposition 1

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

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

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

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

Equate (chain-rule) and (grad-h) gives $\nabla f(x) = \frac{\|Dx + e\|_2}{\langle Ax + b, c \rangle} \nabla \left(\frac{(\langle Ax + b, c \rangle)^2}{2\|Dx + e\|_2^2} \right)$. By assumption $Dx + e \neq 0$, we use quotient rule $\nabla \frac{f}{g} = \frac{g \nabla f - f \nabla g}{g^2}$ to arrive at

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

□

B The proof of Proposition 4

In this appendix, we give the proof of Proposition 4 for the metric retraction over the ellipsoid.

Proof To show that ($\mathcal{R}$ -ellipsoid) is a retraction, we use the definition of retraction in §2.2. Consider a continuous curve $c : \mathbb{R} \rightarrow \mathcal{E}_A^{r-1}$, that is smooth for all variables $(\zeta, \xi) \in \mathbb{R} \times \mathbb{R}^r$, defined as $c(t) = \mathcal{R}_\zeta(t\xi) = (\zeta + t\xi) / \sqrt{1 + t^2 \|\xi\|_\mathcal{E}^2}$. Then $c(0) = \zeta$ and

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

So $\mathcal{R}$ is a retraction for $\mathcal{E}_A^{r-1}$.

□

C Derivation of content in Table 1

Consider ($\mathcal{W}$ -subproblem) with $n = 1$. Following the discussion in §3.1, we get grid of the denominator $\langle W, WA_1 \rangle_F$ in by introducing a constrained problem $\arg\max_{W \geq 0} \langle B_j, W \rangle_F$ s.t. $\langle W, WA_1 \rangle_F = 1$. We note that 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, A_1 has its square-root $A_1^{1/2}$, so $\langle W^\top W, A_1 \rangle_F = \langle W^\top W, A_1^{1/2} A_1^{1/2} \rangle_F = \langle WA_1^{1/2}, WA_1^{1/2} \rangle_F = \langle W, W \rangle_{A_1^{1/2}}$ allows us to define the manifold $\mathcal{M}_1$ as

$$\mathcal{M}_1 = \left\{ W \in \mathbb{R}^{m \times r} \mid \langle W, W \rangle_{A_1^{1/2}} - 1 = 0 \right\}. \quad (\text{Shell of twisted spectrahedron})$$

The term spectrahedron [29] refers to the eigen-spectrum of a matrix behaves like a polyhedron, while the word “twisted” refers to the linear transformation A_1 .

Tangent space. On a single $\mathcal{M}_j$, let $h(W) = \langle W, WA_j \rangle_F - 1$ define $\mathcal{M}_j$, its differential is $Dh(Z)[Z] = 2\langle WA_j, Z \rangle_F$. The tangent space of $\mathcal{M}_j$ at a reference point $Z \in \mathcal{M}_j$, denoted

as $T_Z \mathcal{M}_j$, is

$$T_Z \mathcal{M}_j := \left\{ \mathbf{W} \in \mathbb{R}^{m \times r} \mid \langle \mathbf{Z} \mathbf{A}_j, \mathbf{W} \rangle_F = 0 \right\} = \text{Ker} Dh(\mathbf{W}). \quad (6)$$

Projection onto tangent space, and Retraction. We define the adjoint operator of $Dh(\mathbf{W})[\mathbf{Z}]$ as $Dh(\mathbf{W})^*[\alpha] = 2\alpha \mathbf{W} \mathbf{A}_j$, where α takes the value $\bar{\alpha} = \frac{1}{2} \frac{\langle \mathbf{W}, \mathbf{Z} \mathbf{A}_j \rangle_F}{\|\mathbf{Z} \mathbf{A}_j\|_F^2}$ obtained by solving the least squares problem $\bar{\alpha} = \underset{\alpha \in \mathbb{R}}{\text{argmin}} \|\mathbf{W} - Dh(\mathbf{Z})^*[\alpha]\|_F^2$. Now by $Dh(\mathbf{W})^*[\alpha]$, $\bar{\alpha}$, the definition of projector onto tangent space [18, Def3.60], and the property of orthogonal projector [18, Eq.7.74], we have the projector $\text{proj}_{T_Z \mathcal{M}_j}(\mathbf{W}) = \mathbf{W} - Dh(\mathbf{Z})^*[\bar{\alpha}]$ as shown in the following proposition.

Proposition 7 The map $\text{proj}_{T_Z \mathcal{M}_j} : \mathbb{R}^{m \times r} \rightarrow T_Z \mathcal{M}_j$ that brings $\mathbf{W}$ onto the tangent space at the reference point $\mathbf{Z}$ defined as

$$\text{proj}_{T_Z \mathcal{M}_j}(\mathbf{W}) = \mathbf{W} - \frac{\langle \mathbf{Z} \mathbf{A}_j, \mathbf{W} \rangle_F}{\|\mathbf{Z} \mathbf{A}_j\|_F^2} \mathbf{Z} \mathbf{A}_j \quad (7)$$

is an orthogonal projector.

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

Proposition 8 (Retraction) The map $\mathcal{R}_Z : T_Z \mathcal{M}_j \rightarrow \mathcal{M}_j$ defined as

$$\mathcal{R}_Z(\mathbf{W}) = \frac{\mathbf{W} + \mathbf{Z}}{\|\mathbf{Z} + \mathbf{W}\|_{A_j^{1/2}}} \stackrel{T_Z \mathcal{M}_j}{=} \frac{\mathbf{Z} + \mathbf{W}}{\sqrt{1 + \|\mathbf{W}\|_{A_j^{1/2}}^2}}, \quad (\mathcal{R})$$

is a metric retraction for the manifold $\mathcal{M}_j$.

Proof For a continuous curve $c : \mathbb{R} \rightarrow \mathcal{M}_j$ that is smooth for all $(\mathbf{W}, \mathbf{Z}) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{m \times r}$, defined as

$$c(t) = \mathcal{R}_Z(t\mathbf{W}) = \frac{\mathbf{Z} + t\mathbf{W}}{\sqrt{1 + t^2 \|\mathbf{W}\|_{A_j^{1/2}}^2}}.$$

We have $c(0) = \mathbf{Z}$ and the following so $(\mathcal{R})$ is a retraction for $\mathcal{M}_j$:

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

□

D Details of manifold approach for W-subproblem with $n = 2$

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

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

We call the manifold $\mathcal{M}^{[2]}$ “spectrahedra” as it is constructed by spectrahedron. We now compute its tangent space. Following arguments in §3.1 and §C, we have the following results that we hide the proofs. The function h in (Spectrahedra) has the differential $Dh(\mathbf{W})[\mathbf{Z}] = 2 \begin{bmatrix} \langle \mathbf{W} \mathbf{A}_1, \mathbf{Z} \rangle_{\mathbf{F}} \\ \langle \mathbf{W} \mathbf{A}_2, \mathbf{Z} \rangle_{\mathbf{F}} \end{bmatrix}$. The tangent space of $\mathcal{M}^{[2]}$ is the set $T_{\mathbf{Z}}\mathcal{M}^{[2]} := \left\{ \mathbf{W} \mid \langle \mathbf{W} \mathbf{A}_1, \mathbf{Z} \rangle_{\mathbf{F}} = \langle \mathbf{W} \mathbf{A}_2, \mathbf{Z} \rangle_{\mathbf{F}} = 0 \right\}$, 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 vec be vectorization and let $\otimes_{\mathbb{K}}$ be the Kronecker product, now the $\boldsymbol{\alpha}$ that minimizes $\|\mathbf{W} - Dh(\mathbf{Z})^*[\boldsymbol{\alpha}]\|_{\mathbf{F}}^2$ also minimizes

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

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

$$\text{proj}_{T_{\mathbf{Z}}\mathcal{M}^{[2]}}(\mathbf{W}) = \mathbf{W} - 2\mathbf{Z} \sum_{j=1}^2 \frac{1}{2} \left(\mathbf{S}_{\mathbf{Z}}^{\dagger} \text{vec} \mathbf{W} \right)_j \mathbf{A}_j = \mathbf{W} - 2\mathbf{Z}(\alpha_1^* \mathbf{A}_1 + \alpha_2^* \mathbf{A}_2). \quad (8)$$

The Riemannian gradient is then

$$\text{grad} F(\mathbf{W}) = 2 \text{proj}_{T_{\mathbf{Z}}\mathcal{M}^{[2]}}(\nabla F(\mathbf{W})) \stackrel{(8)}{=} \nabla F(\mathbf{W}) - 2\mathbf{Z}(\alpha_1^* \mathbf{A}_1 + \alpha_2^* \mathbf{A}_2). \quad (9)$$

Note that the value of α_1^* and α_2^* in (9) is an implicit function of $\mathbf{Z}$ and $\nabla F(\mathbf{W})$ as we have

that $\begin{bmatrix} \alpha_1^* \\ \alpha_2^* \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \left(\mathbf{S}_{\mathbf{Z}}^{\dagger} \text{vec} \nabla F(\mathbf{W}) \right)_1 \\ \left(\mathbf{S}_{\mathbf{Z}}^{\dagger} \text{vec} \nabla F(\mathbf{W}) \right)_2 \end{bmatrix}$. So 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 to run RMU on $\mathcal{M}^{[2]}$, there are 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 $\mathcal{M}^{[2]}$ itself is a difficult problem.

Acknowledgements FE is member of the Gruppo Nazionale Calcolo Scientifico - Istituto Nazionale di Alta Matematica (GNCS-INdAM) and she would like to thank Prof. Nicoletta Del Buono from University of Bari Aldo Moro for scientific discussion during this project.

Author Contributions All authors contribute equally to this work.

Funding Open access funding provided by Università degli Studi di Bari Aldo Moro within the CRUI-CARE Agreement. 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, by INdAM - GNCS Project “Ottimizzazione e Algebra Lineare Randomizzata” (CUP E53C24001950001).

Data Availability MATLAB code is available at https://github.com/flaespo/Chordal_NMF

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Gillis, N.: Nonnegative Matrix Factorization. SIAM, Philadelphia, USA (2020)
2. Robusto, C.C.: The cosine-haversine formula. *Am. Math. Mon.* **64**(1), 38–40 (1957)
3. Hosseini, S.: Characterization of lower semicontinuous convex functions on riemannian manifolds. *Set-Valued and Var. Anal.* **25**, 297–311 (2017)
4. Hauswirth, A., Bolognani, S., Hug, G., Dörfler, F.: Projected Gradient Descent on Riemannian Manifolds with Applications to Online Power System Optimization. In: 2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 225–232 (2016). IEEE
5. Huang, W., Wei, K.: Riemannian proximal gradient methods. *Math. Program.* **194**(1–2), 371–413 (2022)
6. Chen, S., Ma, S., Man-Cho So, A., Zhang, T.: Proximal gradient method for nonsmooth optimization over the stiefel manifold. *SIAM J. Optim.* **30**(1), 210–239 (2020)
7. Liu, C., Boumal, N.: Simple algorithms for optimization on riemannian manifolds with constraints. *Appl. Math. Optim.* **82**, 949–981 (2020)
8. Kovnatsky, A., Glashoff, K., Bronstein, M.M.: Madmm: a generic algorithm for non-smooth optimization on manifolds. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14, pp. 680–696. Springer (2016)
9. Boyle, J.P., Dykstra, R.L.: A method for finding projections onto the intersection of convex sets in hilbert spaces. In: Advances in Order Restricted Statistical Inference: Proceedings of the Symposium on Order Restricted Statistical Inference Held in Iowa City, Iowa, September 11–13, 1985, pp. 28–47. Springer (1986)
10. Weber, M., Sra, S.: Riemannian optimization via frank-wolfe methods. *Math. Program.* **199**(1–2), 525–556 (2023)
11. Luenberger, D.G.: The gradient projection method along geodesics. *Manage. Sci.* **18**(11), 620–631 (1972)
12. Gabay, D.: Minimizing a differentiable function over a differential manifold. *J. Optim. Theory Appl.* **37**, 177–219 (1982)
13. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20**(2), 303–353 (1998)
14. Absil, P.-A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton, USA (2008)
15. Udriste, C.: Convex Functions and Optimization Methods on Riemannian Manifolds, vol. 297. Springer, Berlin, Germany (2013)
16. Rapsák, T.: Smooth Nonlinear Optimization in R^n , vol. 19. Springer, US (2013)

17. Bento, G.C., Ferreira, O.P., Melo, J.G.: Iteration-complexity of gradient, subgradient and proximal point methods on riemannian manifolds. *J. Optim. Theory Appl.* **173**(2), 548–562 (2017)
18. Boumal, N.: *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, Cambridge, England (2023)
19. Richardson, W.H.: Bayesian-based iterative method of image restoration. *J. Opt. Soc. Am.* **62**(1), 55–59 (1972)
20. Lucy, L.B.: An iterative technique for the rectification of observed distributions. *Astron. J.* **79**, 745 (1974)
21. Daube-Witherspoon, M.E., Muehllehner, G.: An iterative image space reconstruction algorithm suitable for volume ect. *IEEE Trans. Med. Imaging* **5**(2), 61–66 (1986)
22. Yoo, J., Choi, S.: Orthogonal nonnegative matrix factorization: multiplicative updates on stiefel manifolds. In: *International Conference on Intelligent Data Engineering and Automated Learning*, pp. 140–147. Springer (2008)
23. Horn, R.A., Johnson, C.R.: *Matrix Analysis*. Cambridge University Press (2012)
24. Dinkelbach, W.: On nonlinear fractional programming. *Manage. Sci.* **13**(7), 492–498 (1967)
25. Jagannathan, R.: On some properties of programming problems in parametric form pertaining to fractional programming. *Manage. Sci.* **12**(7), 609–615 (1966)
26. Feng, X.-R., Li, H.-C., Wang, R., Du, Q., Jia, X., Plaza, A.: Hyperspectral unmixing based on nonnegative matrix factorization: a comprehensive review. *IEEE J. of Sel. Topics in Appl. Earth Obs. and Remote Sens.* **15**, 4414–4436 (2022)
27. Ang, A.M.S., Gillis, N.: Algorithms and comparisons of nonnegative matrix factorizations with volume regularization for hyperspectral unmixing. *IEEE J. of Sel. Topics in Appl. Earth Obs. and Remote Sens.* **12**(12), 4843–4853 (2019)
28. Du, Y., Chang, C.-I., Ren, H., Chang, C.-C., Jensen, J.O., D’Amico, F.M.: New hyperspectral discrimination measure for spectral characterization. *Opt. Eng.* **43**(8), 1777–1786 (2004)
29. Ramana, M., Goldman, A.J.: Some geometric results in semidefinite programming. *J. Global Optim.* **7**(1), 33–50 (1995)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.