Riemannian block SPD coupling manifold and its application to optimal transport

Andi Han[1] · **Bamdev Mishra[2]** · **Pratik Jawanpuria[2]** · **Junbin Gao[1]**

Abstract

In this work, we study the optimal transport (OT) problem between symmetric positive definite (SPD) matrix-valued measures. We formulate the above as a generalized optimal transport problem where the cost, the marginals, and the coupling are represented as block matrices and each component block is a SPD matrix. The summation of row blocks and column blocks in the coupling matrix are constrained by the given block-SPD marginals. We endow the set of such block-coupling matrices with a novel Riemannian manifold structure. This allows to exploit the versatile Riemannian optimization framework to solve generic SPD matrix-valued OT problems. We illustrate the usefulness of the proposed approach in several applications.

1 Introduction

Optimal transport (OT) offers a systematic approach to compare probability distributions by finding a transport plan (coupling) that minimizes the cost of transporting mass from one distribution to another. It has been successfully applied in a wide range of fields, such as computer graphics (Solomon et al., 2015, 2014), graph representation learning (Chen

---

✉ Andi Han
andi.han@sydney.edu.au

Bamdev Mishra
bamdevm@microsoft.com

Pratik Jawanpuria
pratik.jawanpuria@microsoft.com

Junbin Gao
junbin.gao@sydney.edu.au

[1] The University of Sydney, Camperdown, Australia

[2] Microsoft, Hyderabad, India

et al., 2020; Maretic et al., 2019), text classification (Yurochkin et al., 2019), and domain adaptation (Courty et al., 2016, 2014), to name a few.

Despite the popularity of OT, existing OT formulations are mostly limited to scalar-valued distributions. On the other hand, many applications involve symmetric positive definite (SPD) matrix-valued distributions. In diffusion tensor imaging (Le Bihan et al., 2001), the local diffusion of water molecules in human brain are encoded in fields of SPD matrices (Assaf and Pasternak, 2008). In image processing, region information of an image can be effectively captured through several SPD covariance descriptors (Tuzel et al., 2006). For the application of image set/video classification, each set of images/frames can be represented by its covariance matrix, which has shown great promise in modelling the intra-set variations (Huang et al., 2015; Harandi et al., 2014). In addition, fields of SPD matrices are also important in computer graphics for anisotropic diffusion (Weickert, 1998), remeshing (Alliez et al., 2003) and texture synthesis (Galerne et al., 2010), just to name a few. In all such cases, being able to compare fields represented by SPD matrices is crucial. This, however, requires a nontrivial generalization of existing (scalar-valued) optimal transport framework with careful construction of cost and transport plan.

In the quantum mechanics setting, existing works (Jiang et al., 2012; Carlen and Maas, 2014; Chen et al., 2017, 2018) have explored geodesic formulation of the Wasserstein distance between vector and matrix-valued densities. In (Ning, 2013; Ning et al., 2014), the Monge-Kantorovich optimal mass transport problem has been studied for comparing matrix-valued power spectra measures. Recently, Peyré et al. (2019a) proposed to solve an unbalanced optimal transport problem for SPD-valued distributions of unequal masses.

In this paper, we propose a general framework for solving the balanced OT problem between SPD-valued distributions, where the cost and the coupling are represented as block SPD matrices. We discuss a Riemannian manifold structure for the set of such block coupling matrices, and we are able to use the Riemannian optimization framework (Absil et al., 2008; Boumal, 2020) to solve various generalized OT problems. Specifically, our contributions are as follows.

1. We introduce the general SPD matrix-valued balanced OT problem for SPD matrix-valued marginals and study its metric properties for a specific setting.
2. We propose a novel manifold structure for the set of block matrix coupling matrices, which generalizes the manifold structures studied in (Douik and Hassibi, 2019; Shi et al., 2021; Mishra et al., 2021, 2019). We discuss optimization-related ingredients like Riemannian metric, Riemannian gradient, Hessian, and retraction.
3. We extend our SPD-valued balanced OT formulation to block SPD Wasserstein barycenter and Gromov-Wasserstein OT.
4. We empirically illustrate the benefit of the proposed framework in domain adaptation, tensor-valued shape interpolation, and displacement interpolation between tensor fields.

Organizations. We start with a brief review of Riemannian optimization and SPD matrix-valued optimal transport problem in Sect. 2. In Sect. 3, we introduce the generalized SPD matrix-valued OT problem and define the proposed block SPD coupling manifold. Sect. 4 discusses the Riemannian structure of the proposed manifold and derives the necessary optimization-related ingredients. Sect. 5 presents two additional OT related applications of the proposed Block SPD coupling manifold. In Sect. 6, we empirically evaluate the proposed approach in various applications. Sect. 7 concludes the paper. In the appendix sections, we provide the proofs and present additional experiments.

2 Preliminaries

2.1 Riemannian optimization

A matrix manifold $\mathcal{M}$ is a smooth subset of the ambient vector space $\mathcal{V}$ with local bijectivity to the Euclidean space. A Riemannian manifold is a manifold endowed with a Riemannian metric (a smooth, symmetric positive definite inner product structure $\langle \cdot, \cdot \rangle_x$) on every tangent space $T_x\mathcal{M}$. The induced norm on the tangent space is thus $\|u\|_x = \sqrt{\langle u, u \rangle_x}$.

The orthogonal projection operation for an embedded matrix manifold $P_x : \mathcal{V} \to T_x\mathcal{M}$ is a projection that is orthogonal with respect to the Riemannian metric $\langle \cdot, \cdot \rangle_x$. Retraction is a smooth map from tangent space to the manifold That is, for any $x \in \mathcal{M}$, retraction $R_x : T_x\mathcal{M} \to \mathcal{M}$ such that 1) $R_x(0) = x$ and 2) $DR_x(0)[u] = u$, where $Df(x)[u]$ is the derivative of a function at x along direction u.

The Riemannian gradient of a function $F : \mathcal{M} \to \mathbb{R}$ at x, denoted as $\mathrm{grad}F(x)$, generalizes the notion of the Euclidean gradient $\nabla F(x)$. It is defined as the unique tangent vector satisfying $\langle \mathrm{grad}F(x), u \rangle_x = DF(x)[u] = \langle \nabla F(x), u \rangle_2$ for any $u \in T_x\mathcal{M}$, where $\langle \cdot, \cdot \rangle_2$ denotes the Euclidean inner product. To minimize the function, Riemannian gradient descent (Absil et al., 2008) and other first-order solvers apply retraction to update the iterates along the direction of negative Riemannian gradient while staying on the manifold, i.e., $x_{t+1} = R_{x_t}(-\eta \, \mathrm{grad}F(x_t))$, where η is the step size. Similarly, the Riemannian Hessian $\mathrm{Hess}F(x) : T_x\mathcal{M} \to T_x\mathcal{M}$ is defined as the covariant derivative of Riemannian gradient. Popular second-order methods, such as trust regions and cubic regularized Newton's methods have been adapted to Riemannian optimization (Absil et al., 2007; Agarwal et al., 2018).

2.2 Scalar-valued optimal transport

Consider two discrete measures supported on $\mathbb{R}^d$, $\mu = \sum_{i=1}^m p_i \delta_{\mathbf{x}_i}$, $\nu = \sum_{j=1}^n q_j \delta_{\mathbf{y}_j}$, where $\mathbf{x}_i, \mathbf{y}_j \in \mathbb{R}^d$ and $\delta_{\mathbf{x}}$ is the Dirac at $\mathbf{x}$. The weights $\mathbf{p} \in \Sigma_m, \mathbf{q} \in \Sigma_n$ are in probability simplex where $\Sigma_k := \{\mathbf{p} \in \mathbb{R}^k : p_i \geq 0, \sum_i p_i = 1\}$. The 2-Wasserstein distance between μ, ν is given by solving the Monge-Kantorovich optimal transport problem:

$$W_2^2(\mathbf{p}, \mathbf{q}) = \min_{\gamma \in \Pi(\mathbf{p}, \mathbf{q})} \sum_{i,j} \|\mathbf{x}_i - \mathbf{y}_j\|^2 \gamma_{i,j}, \tag{1}$$

where $\Pi(\mathbf{p}, \mathbf{q}) := \{\gamma \in \mathbb{R}^{m \times n} : \gamma \geq 0, \gamma \mathbf{1} = \mathbf{p}, \gamma^\top \mathbf{1} = \mathbf{q}\}$ is the space of joint distribution between the source and the target marginals. An optimal solution of (1) is referred to as an optimal transport plan (or coupling). Recently, Cuturi (2013) proposed the Sinkhorn-Knopp algorithm (Sinkhorn, 1964; Knight, 2008) for entropy-regularized OT formulation. In case μ and ν are measures (i.e., the setting is not restricted to probability measures), it may happen that they are of unequal masses. OT in this case is termed as unbalanced optimal transport (Chizat et al., 2018; Liero et al., 2018). For a recent survey of OT literature and related machine learning applications, please refer to (Peyré et al., 2019b).

2.3 SPD matrix-valued optimal transport

A SPD matrix-valued measure is a generalization of the (scalar-valued) probability measure (discussed in Sect. 2.2). Let us consider a SPD matrix-valued measure M and a scalar-valued measure μ defined on a space $\mathcal{X}$. Let A be a measurable subset of $\mathcal{X}$. Then, while $\mu(A)$ is a non-negative scalar, the "mass" $M(A) \in \mathbb{S}_+^d$, where $\mathbb{S}_+^d$ denotes the set of $d \times d$ positive semi-definite matrices. SPD matrix-valued measures have been employed in applications such as diffusion tensor imaging (Le Bihan et al., 2001), image set classification (Huang et al., 2015; Harandi et al., 2014), anisotropic diffusion (Weickert, 1998), and brain imaging (Assaf and Pasternak, 2008), to name a few.

Recent works (Carlen and Maas, 2014; Chen et al., 2017; Ryu et al., 2018; Peyré et al., 2019a) have explored optimal transport formulations for SPD matrix-valued measures. While the works (Carlen and Maas, 2014; Chen et al., 2017; Ryu et al., 2018) discuss dynamical (geodesic) OT framework, (Peyré et al., 2019a) studies the "static" OT formulation that learns a suitable joint coupling between the input SPD matrix-valued measures. However, Peyré et al. (2019a) explore an unbalanced OT setup for SPD matrix-valued measures and term it as quantum optimal transport (QOT). Thus, the marginals of the (learned) joint coupling in QOT is not equal to the input SPD matrix-valued measures. As in case of unbalanced (scalar-valued) OT (Chizat et al., 2018; Liero et al., 2018), the discrepancy between marginals of the joint and the input measures in QOT is penalized via the Kulback-Leibler divergence (for SPD matrix-valued measures).

3 Block SPD optimal transport

In this section, we study a balanced OT formulation for SPD matrix-valued measures. Consider $\mathbf{P}$ and $\mathbf{Q}$ to be (d-dimensional) SPD matrix-valued input measures. Let $\mathbf{P} := \{[\mathbf{P}_i]_{m \times 1} : \mathbf{P}_i \in \mathbb{S}_{++}^d\}$ and $\mathbf{Q} := \{[\mathbf{Q}_j]_{n \times 1} : \mathbf{Q}_j \in \mathbb{S}_{++}^d\}$ and $\mathbf{P}$ and $\mathbf{Q}$ have the same total mass. Without loss of generality, we assume $\sum_i \mathbf{P}_i = \sum_j \mathbf{Q}_j = \mathbf{I}$. Here, $[\cdot]_{m \times n}$ denotes a collection of mn matrices organized as a block matrix and $\mathbf{I}$ represents the identity matrix. The cost of transporting a positive definite matrix-valued mass $\mathbf{A}$ from position $\mathbf{x}_i$ (in source space) to $\mathbf{y}_j$ (in target space) is parameterized by a (given) positive semi-definite matrix $\mathbf{C}_{i,j}$ and is computed as $\text{tr}(\mathbf{C}_{i,j}\mathbf{A})$. Under this setting, we propose the block SPD matrix-valued balanced OT problem as

$$\text{MW}^2(\mathbf{P}, \mathbf{Q}) := \min_{\mathbf{\Gamma} \in \mathbf{\Pi}(m,n,d,\mathbf{P},\mathbf{Q})} \sum_{i,j} \text{tr}(\mathbf{C}_{i,j}\mathbf{\Gamma}_{i,j}), \tag{2}$$

where $\mathbf{\Gamma} = [\mathbf{\Gamma}_{i,j}]_{m \times n}$ is a block-matrix coupling of size $m \times n$ and the set of such couplings are defined as $\mathbf{\Pi}(m,n,d,\mathbf{P},\mathbf{Q}) := \{[\mathbf{\Gamma}_{i,j}]_{m \times n} : \mathbf{\Gamma}_{i,j} \in \mathbb{S}_+^d, \sum_j \mathbf{\Gamma}_{i,j} = \mathbf{P}_i, \sum_i \mathbf{\Gamma}_{i,j} = \mathbf{Q}_j, \forall i \in [m], j \in [n]\}$. Here $\mathbb{S}_+^d$ is used to denote the set of $d \times d$ positive semi-definite matrices and $\text{tr}(\cdot)$ is the matrix trace. The problem is well-defined provided that the corresponding coupling constraint set $\mathbf{\Pi}(m,n,d,\mathbf{P},\mathbf{Q})$ is non-empty. For arbitrary SPD marginals $\mathbf{P}, \mathbf{Q}$, there is no guarantee that the set $\mathbf{\Pi}(m,n,d,\mathbf{P},\mathbf{Q})$ defined in (2) is not empty (Ning et al., 2014). Hence, in this work, we assume that the given marginals $\mathbf{P}$ and $\mathbf{Q}$ are such that $\mathbf{\Pi}(m,n,d,\mathbf{P},\mathbf{Q})$ is not empty. In Sect. 4.3 later, we discuss a block matrix balancing algorithm which can be used to check whether $\mathbf{\Pi}(m,n,d,\mathbf{P},\mathbf{Q})$ is empty or not for given marginals $\mathbf{P}$ and $\mathbf{Q}$.

3.1 Metric properties of MW(**P**, **Q**)

In the following result, we show that MW(**P**, **Q**) is a valid distance metric for a special case of block SPD marginals.

Proposition 1 *Suppose the input SPD matrix-valued marginals have the same support size n and the costs $\{\mathbf{C}_{i,j}\}_{i,j=1}^{n}$ satisfy*

1. $\mathbf{C}_{i,j} = \mathbf{C}_{j,i}$ *and*
2. $\mathbf{C}_{i,j} \succ \mathbf{0}$ *for $i \neq j$ and $\mathbf{C}_{i,j} = \mathbf{0}$ for $i = j$,*
3. $\forall (i,j,k) \in [n]^3, \text{and } \mathbf{A} \succeq \mathbf{0}, \sqrt{\text{tr}(\mathbf{C}_{i,j}\mathbf{A})} \leq \sqrt{\text{tr}(\mathbf{C}_{i,k}\mathbf{A})} + \sqrt{\text{tr}(\mathbf{C}_{j,k}\mathbf{A})}.$

Then, MW(**P**, **Q**) is a metric between the SPD matrix-valued marginals **P** and **Q** defined as $\mathbf{P} := \{[\mathbf{P}_i]_{m\times 1} : \mathbf{P}_i = p_i\mathbf{I}\}$ and $\mathbf{Q} := \{[\mathbf{Q}_j]_{n\times 1} : \mathbf{Q}_j = q_j\mathbf{I}\}$, where $\mathbf{p}, \mathbf{q} \in \Sigma_n$ and **I** is the $d \times d$ identity matrix.

We remark that the conditions on $\mathbf{C}_{i,j}$ in Proposition 1 generalize the conditions required for $\text{W}_2(\mathbf{p}, \mathbf{q})$ in (1) to be a metric. See for example (Peyré et al. 2019b, Proposition 2.2). In "Appendix B", we discuss some particular constructions of the cost that satisfy the conditions.

3.2 Manifold structure for the coupling set Π(*m*, *n*, *d*, **P**, **Q**)

We next analyze the coupling constraint set $\mathbf{\Pi}(m, n, d, \mathbf{P}, \mathbf{Q})$ and show that it can be endowed with a manifold structure. This allows to exploit the versatile Riemannian optimization framework to solve (2) and any more general problem (Absil et al., 2008).

We propose the following manifold structure, termed as the block SPD coupling manifold,

$$\mathcal{M}_{m,n}^{d}(\mathbf{P}, \mathbf{Q}) := \{\mathbf{\Gamma} : \mathbf{\Gamma}_{i,j} \in \mathbb{S}_{++}^{d}, \sum_{j} \mathbf{\Gamma}_{i,j} = \mathbf{P}_i, \sum_{i} \mathbf{\Gamma}_{i,j} = \mathbf{Q}_j\}, \tag{3}$$

where $\sum_i \mathbf{P}_i = \sum_j \mathbf{Q}_j = \mathbf{I}$. Particularly, we restrict $\mathbf{P}_i, \mathbf{Q}_j, \mathbf{\Gamma}_{i,j} \in \mathbb{S}_{++}^{d}$, the set of SPD matrices. This ensures that the proposed manifold $\mathcal{M}_{m,n}^{d}(\mathbf{P}, \mathbf{Q})$ in (3) is the interior of the set $\mathbf{\Pi}(m, n, d, \mathbf{P}, \mathbf{Q})$.

As discussed earlier $\mathbf{\Pi}(m, n, d, \mathbf{P}, \mathbf{Q})$ is not guaranteed to be non-empty for arbitrary choices of block SPD marginals **P** and **Q** (Ning, 2013). To this end, we assume that the marginals **P** and **Q** that are given ensure feasibility of the set $\mathbf{\Pi}(m, n, d, \mathbf{P}, \mathbf{Q})$. In particular, the manifold $\mathcal{M}_{m,n}^{d}(\mathbf{P}, \mathbf{Q})$ inherits the following assumption.

Assumption 1 In this work, we consider block-SPD marginals **P** and **Q** such that the set $\mathcal{M}_{m,n}^{d}(\mathbf{P}, \mathbf{Q})$ is not empty.

It should be noted that Assumption 1 is trivially satisfied for diagonal SPD marginals, i.e., when $\mathbf{P}_i$ and $\mathbf{Q}_j$ are diagonal. However, non-diagonal SPD marginals may also satisfy Assumption 1 for many problem instances. In Sect. 6, we discuss empirical

settings where non-diagonal SPD marginals satisfying Assumption 1 are considered. The following proposition implies that we can endow $\mathcal{M}^d_{m,n}(\mathbf{P}, \mathbf{Q})$ with a differentiable structure.

Proposition 2 *Under Assumption* 1, *the set* $\mathcal{M}^d_{m,n}(\mathbf{P}, \mathbf{Q})$ *is smooth, i.e., differentiable.*

It should be emphasized that the proposed manifold $\mathcal{M}^d_{m,n}(\mathbf{P}, \mathbf{Q})$ can be regarded as a generalization to existing manifold structures. For example, when $d = 1$ and either $m = 1$ or $n = 1$, $\mathcal{M}^d_{m,n}(\mathbf{P}, \mathbf{Q})$ reduces to the multinomial manifold of probability simplex (Sun et al., 2015). When $d = 1$ and $m, n \neq 1$, it reduces the so-called doubly stochastic manifold (Douik and Hassibi, 2019) with uniform marginals or the more general matrix coupling manifold (Shi et al., 2021). When $d > 1$ and either $m = 1$ or $n = 1$, our proposed manifold simplifies to the simplex manifold of SPD matrices (Mishra et al., 2019).

In the next section, we derive various optimization-related ingredients on $\mathcal{M}^d_{m,n}(\mathbf{P}, \mathbf{Q})$ that allow optimization of an arbitrary differentiable objective function on the manifold. In particular, we propose a Riemannian optimization approach following the general treatment by (Absil et al., 2008; Boumal, 2020). It allows employing the proposed approach not only for (2) but also for other OT problems as discussed in Sect. 5.

Algorithm 1 Riemannian optimization for solving (4).

1: Initialize a feasible $\mathbf{\Gamma}_0 \in \mathcal{M}^d_{m,n}$.
2: $\mathbf{\Gamma} = \mathbf{\Gamma}_0$.
3: **while** not converging **do**
4:     Compute Riemannian gradient (and Hessian) at $\mathbf{\Gamma}$.
5:     Compute the update step $\boldsymbol{\xi} \in T_{\mathbf{\Gamma}}\mathcal{M}^d_{m,n}$.
6:     Update $\mathbf{\Gamma} \leftarrow R_{\mathbf{\Gamma}}(\boldsymbol{\xi})$.
7: **end while**
8: **Output:** $\mathbf{\Gamma}$.

---

4 Riemannian geometry and optimization over $\mathcal{M}^d_{m,n}(\mathbf{P}, \mathbf{Q})$

We consider the general optimization problem

$$\min_{\mathbf{\Gamma} \in \mathcal{M}^d_{m,n}(\mathbf{P}, \mathbf{Q})} F(\mathbf{\Gamma}), \tag{4}$$

where $F : \mathcal{M}^d_{m,n}(\mathbf{P}, \mathbf{Q}) \to \mathbb{R}$ is a differentiable objective function. The proposed manifold $\mathcal{M}^d_{m,n}(\mathbf{P}, \mathbf{Q})$ can be endowed with a smooth Riemannian manifold structure (Absil et al., 2008; Boumal, 2020). Consequently, (4) is an optimization problem on a Riemannian manifold. We solve the problem via the Riemannian optimization framework. It provides a principled class of optimization methods and computational tools for manifolds, both first order and second order, as long as the ingredients such as Riemannian metric, orthogonal projection, retraction, and Riemannian gradient (and Hessian) of a function are defined (Absil et al., 2008; Boumal et al., 2014; Boumal, 2020). Conceptually, the Riemannian

optimization framework treats (4) as an "unconstrained" optimization problem over the constraint manifold $\mathcal{M}_{m,n}^d$ (omitted marginals **P**, **Q** for clarity).

In Algorithm 1, we outline the skeletal steps involved in optimization over $\mathcal{M}_{m,n}^d$, where the step ξ can be computed from different Riemannian methods. In Riemannian steepest descent, $\xi = -\eta \operatorname{grad} F(\mathbf{\Gamma})$, where $\operatorname{grad} F(\mathbf{\Gamma})$ is the Riemannian gradient at $\mathbf{\Gamma}$. Also, ξ is given by the "conjugate" direction of $\operatorname{grad} F(\mathbf{\Gamma})$ in the Riemannian conjugate gradient method. And, for the Riemannian trust-region method, ξ computation involves minimizing a second-order approximation of the objective function in a trust-region ball (Absil et al., 2008). Below, we show the computations of these ingredients.

4.1 Riemannian metric

The manifold $\mathcal{M}_{m,n}^d$ is a submanifold of the Cartesian product of $m \times n$ SPD manifold of size $d \times d$, which we denote as $\times_{m,n} \mathbb{S}_{++}^d$. The dimension of the manifold $\mathcal{M}_{m,n}^d$ is $(m-1)(n-1)d(d+1)/2$. The tangent space characterization of $\mathcal{M}_{m,n}^d$ at $\mathbf{\Gamma}$ is obtained as

$$T_{\mathbf{\Gamma}}\mathcal{M}_{m,n}^d = \left\{ [\mathbf{U}_{i,j}]_{m \times n} : \mathbf{U}_{i,j} \in \mathbb{S}^d, \sum_j \mathbf{U}_{i,j} = \mathbf{0}, \sum_i \mathbf{U}_{i,j} = \mathbf{0} \right\},$$

where $\mathbb{S}^d$ is the set of $d \times d$ symmetric matrices. The expression for the tangent space is obtained by linearizing the constraints. We endow each SPD manifold with the affine-invariant Riemannian metric (Bhatia, 2009), which induces a Riemannian metric for the product manifold $\mathcal{M}_{m,n}^d$ as

$$\langle \mathbf{U}, \mathbf{V} \rangle_{\mathbf{\Gamma}} = \sum_{i,j} \operatorname{tr}(\mathbf{\Gamma}_{i,j}^{-1} \mathbf{U}_{i,j} \mathbf{\Gamma}_{i,j}^{-1} \mathbf{V}_{i,j}), \tag{5}$$

for any $\mathbf{U}, \mathbf{V} \in T_{\mathbf{\Gamma}}\mathcal{M}_{m,n}^d$.

4.2 Orthogonal projection, riemannian gradient, and riemannian Hessian

As an embedded submanifold, the orthogonal projection plays a crucial role in deriving the Riemannian gradient (as orthogonal projection of the Euclidean gradient in the ambient space).

Proposition 3 *The orthogonal projection of any* $\mathbf{S} \in \times_{m,n} \mathbb{S}^d$ *to* $T_{\mathbf{\Gamma}}\mathcal{M}_{m,n}^d$ *with respect to the Riemannian metric* (5) *is given by*

$$P_{\mathbf{\Gamma}}(\mathbf{S}) = \mathbf{U}, \text{ with } \mathbf{U}_{i,j} = \mathbf{S}_{i,j} + \mathbf{\Gamma}_{i,j}(\mathbf{\Lambda}_i + \mathbf{\Theta}_j)\mathbf{\Gamma}_{i,j},$$

where auxiliary variables $\mathbf{\Lambda}_i, \mathbf{\Theta}_j$ *are solved from the system of matrix linear equations*:

$$\begin{cases} -\sum_i \mathbf{S}_{i,j} = \sum_i \mathbf{\Gamma}_{i,j}(\mathbf{\Lambda}_i + \mathbf{\Theta}_j)\mathbf{\Gamma}_{i,j}, \ \forall j \\ -\sum_j \mathbf{S}_{i,j} = \sum_j \mathbf{\Gamma}_{i,j}(\mathbf{\Lambda}_i + \mathbf{\Theta}_j)\mathbf{\Gamma}_{i,j}, \ \forall i. \end{cases}$$

Subsequently, the Riemannian gradient and Hessian are derived as the orthogonal projection of the gradient and Hessian from the ambient space.

Proposition 4 *The Riemannian gradient and Hessian of $F : \mathcal{M}_{m \times n}^d \to \mathbb{R}$ are derived as*

$$\mathrm{grad}F(\mathbf{\Gamma}) = P_{\mathbf{\Gamma}}([\mathbf{\Gamma}_{i,j}\{\nabla F(\mathbf{\Gamma}_{i,j})\}_S \mathbf{\Gamma}_{i,j}]_{m \times n}),$$

$$\mathrm{Hess}F(\mathbf{\Gamma})[\mathbf{U}] = P_{\mathbf{\Gamma}}([\mathrm{Dgrad}F(\mathbf{\Gamma}_{i,j})[\mathbf{U}_{i,j}] - \{\mathbf{U}_{i,j}\mathbf{\Gamma}_{i,j}^{-1}\mathrm{grad}F(\mathbf{\Gamma}_{i,j})\}_S]_{m \times n}),$$

where $\mathbf{U} \in T_{\mathbf{\Gamma}}\mathcal{M}_{m,n}^d$ and $\nabla F(\mathbf{\Gamma}_{i,j})$ is the block partial derivative of F with respect to $\mathbf{\Gamma}_{i,j}$. Here, $\mathrm{Dgrad}F(\mathbf{\Gamma}_{i,j})[\mathbf{U}_{i,j}]$ denotes the directional derivative of the Riemannian gradient $\mathrm{grad}F$ along $\mathbf{U}$ and $\{\mathbf{A}\}_S := (\mathbf{A} + \mathbf{A}^\top)/2$.

4.3 Retraction and block matrix balancing algorithm

The retraction operation on $\mathcal{M}_{m,n}^d$ is given by a composition of two operations. The first operation is to ensure positive definiteness of the blocks in the coupling matrix. In particular, we use the exponential map associated with the affine-invariant metric on the SPD manifold $\mathbb{S}_{++}^d$ (Bhatia, 2009). The second operation is to ensure that the summation of the row blocks and column blocks respect the block-SPD marginals. Given an initialized block SPD matrix $[\mathbf{A}_{i,j}]_{m \times n}$, where $\mathbf{A}_{i,j} \in \mathbb{S}_{++}^d$, the goal is to find a 'closest' block SPD coupling matrix $\mathbf{B} \in \mathcal{M}_{m,n}^d$. This is achieved by alternatively normalizing the row and column blocks to the corresponding marginals. The procedure is outlined in Algorithm 2. The solution for the row and column normalization factors $\mathbf{R}_j, \mathbf{L}_i$, which are SPD matrices, are computed by solving the Riccati equation $\mathbf{T}\mathbf{X}\mathbf{T} = \mathbf{Y}$ for given $\mathbf{X}, \mathbf{Y} \in \mathbb{S}_{++}^d$. Here, $\mathbf{T}$ admits a unique solution (Bhatia, 2009; Malagò et al., 2018). Different from the scalar marginals case where the scaling can be expressed as a diagonal matrix, we need to symmetrically normalize each SPD block matrix. Algorithm 2 is a generalization of the RAS algorithm for balancing non-negative matrices (Sinkhorn, 1967), which is related to the popular Sinkhorn-Knopp algorithm (Sinkhorn, 1964; Knight, 2008). We also use Algorithm 2 to test feasibility of the set $\mathcal{M}_{m,n}^d$ by checking whether Algorithm 2 outputs a balanced block SPD matrix for a random block SPD matrix $\mathbf{A}$.

It should be noted that a similar matrix balancing algorithm has been introduced for positive operators (Gurvits, 2004; Georgiou and Pavon, 2015), where the convergence is only established in limited cases. Algorithm 2 is different from the quantum Sinkhorn algorithm proposed in (Peyré et al., 2019a) that applies to the unbalanced setting. Although we do not provide a theoretical convergence analysis for Algorithm 2, we empirically observe quick convergence of this algorithm in various settings (see "Appendix A").

Algorithm 2 Block matrix balancing algorithm

1: **Input:** $[\mathbf{A}_{i,j}]_{m \times n}$, where $\mathbf{A}_{i,j} \in \mathbb{S}^d_{++}$ and block SPD marginals $\mathbf{P}$ and $\mathbf{Q}$.
2: Initialize $\mathbf{B} = [\mathbf{A}_{i,j}]_{m \times n}$.
3: **while** not converging **do**
4:     Find $\mathbf{R}_j \in \mathbb{S}^d_{++}$ such that $\sum_i \mathbf{R}_j \mathbf{B}_{i,j} \mathbf{R}_j = \mathbf{Q}_j$, $\forall j$.
5:     Update $\mathbf{B}_{i,j} \leftarrow \mathbf{R}_j \mathbf{B}_{i,j} \mathbf{R}_j$, $\forall j$.
6:     Find $\mathbf{L}_i \in \mathbb{S}^d_{++}$ such that $\sum_j \mathbf{L}_i \mathbf{B}_{i,j} \mathbf{L}_i = \mathbf{P}_i$, $\forall i$.
7:     Update $\mathbf{B}_{i,j} \leftarrow \mathbf{L}_i \mathbf{B}_{i,j} \mathbf{L}_i$, $\forall i$.
8: **end while**
9: **Output:** The balanced matrix $\mathbf{B}$.

Based on Algorithm 2, we define a retraction $R_{\boldsymbol{\Gamma}}(\mathbf{U})$ at $\boldsymbol{\Gamma} \in \mathcal{M}^d_{m,n}$ for any $\mathbf{U} \in T_{\boldsymbol{\Gamma}}\mathcal{M}^d_{m,n}$ as

$$R_{\boldsymbol{\Gamma}}(\mathbf{U}) = \mathrm{MBalance}\left([\boldsymbol{\Gamma}_{i,j}\exp(\boldsymbol{\Gamma}_{i,j}^{-1}\mathbf{U}_{i,j})]_{m \times n} \right), \tag{6}$$

where MBalance calls the matrix balancing procedure in Algorithm 2 and $\exp(\cdot)$ denotes the matrix exponential. The retraction proposed in (6) is valid (i.e., satisfy the two conditions) for diagonal marginals and empirically we also see the retraction is well-defined for arbitrary block-SPD marginals. See "Appendix A" for more details.

4.4 Convergence and computational complexity

Convergence of riemannian optimization. Similar to Euclidean optimization, the necessary first-order optimality condition for any differentiable F on $\mathcal{M}^d_{m,n}$ is $\mathrm{grad}F(\boldsymbol{\Gamma}^*) = 0$, i.e., where the Riemannian gradient vanishes. We call such $\boldsymbol{\Gamma}^*$ the stationary point. The Riemannian methods are known to converge to a stationary point (Absil et al., 2008; Boumal, 2020) under standard assumptions. Additionally, we show the following.

Theorem 2 *Suppose the objective function of the problem* $\min_{\boldsymbol{\Gamma} \in \boldsymbol{\Pi}(m,n,d,\mathbf{P},\mathbf{Q})} F(\boldsymbol{\Gamma})$ *is strictly convex and the optimal solution* $\boldsymbol{\Gamma}^*$ *is positive definite, i.e., it lies in the interior of* $\boldsymbol{\Pi}(m, n, d, \mathbf{P}, \mathbf{Q})$. *Then, Riemannian optimization (Algorithm 1) for (4) converges to the same global optimal solution* $\boldsymbol{\Gamma}^*$.

Theorem 2 guarantees the quality of the solution obtained by Riemannian optimization for a class of objective functions which includes the SPD matrix-valued OT problem with convex regularization.

Computational complexity. The complexity of each iteration of the Riemannian optimization algorithm is dominated by the computations of retraction, the Riemannian gradient, the Riemannian Hessian. These also make use of the orthogonal projection operation. All these operations cost $O(mnd^3)$. Since the number of parameters to be learned is $N = mnd^2$ (size of the coupling block SPD matrix $\boldsymbol{\Gamma}$), the above cost is almost linear in N.

5 Applications of block SPD coupling manifold

As discussed earlier, we employ the proposed block SPD coupling manifold optimization approach to solve the block SPD matrix valued balanced OT problem (2). We now present two other OT related applications of the block SPD coupling manifold: learning Wasserstein barycenters and the Gromov-Wasserstein averaging of distance matrices.

5.1 Block SPD Wasserstein barycenter learning

We consider the problem of computing the Wasserstein barycenter of a set of block SPD matrix-valued measures. Let $\Delta_n(\mathbb{S}_{++}^d) := \{\mathbf{P} = [\mathbf{P}_i]_{n \times 1} : \mathbf{P}_i \in \mathbb{S}_{++}^d, \sum_i \mathbf{P}_i = \mathbf{I}\}$ denotes the space of $n \times 1$ block SPD marginals. Then, the Wasserstein barycenter $\bar{\mathbf{P}}$ of a set $\mathbf{P}^\ell \in \Delta_{n_\ell}(\mathbb{S}_{++}^d)$ for all $\ell = \{1, \dots, K\}$ is computed as follows:

$$\bar{\mathbf{P}} = \underset{\mathbf{P} \in \Delta_n(\mathbb{S}_{++}^d)}{\arg\min} \sum_{\ell=1}^{K} \omega_\ell \mathrm{MW}_\epsilon^2(\mathbf{P}, \mathbf{P}^\ell), \tag{7}$$

where the given non-negative weights satisfy $\sum_\ell \omega_\ell = 1$. It should be noted that we employ a regularized version of the proposed block SPD OT problem (2) to ensure the differentiability of the objective function near boundary in (7). The regularized block SPD OT problem is defined as

$$\mathrm{MW}_\epsilon^2(\mathbf{P}, \mathbf{Q}) := \min_{\Gamma \in \mathcal{M}_{m,n}^d(\mathbf{P}, \mathbf{Q})} \sum_{i,j} \left(\mathrm{tr}(\mathbf{C}_{i,j} \Gamma_{i,j}) + \epsilon\, \Omega(\Gamma_{i,j}) \right), \tag{8}$$

where $\epsilon > 0$ is the regularization parameter and $\Omega(\cdot)$ is a strictly convex regularization (e.g., entropic regularization) on the block SPD coupling matrices.

To solve for $\bar{\mathbf{P}}$ in (7), we consider Riemannian optimization on $\Delta_n(\mathbb{S}_{++}^d)$, which has recently been studied in (Mishra et al., 2019). The following result provides an expression for the Euclidean gradient of the objective function in problem (7).

Proposition 5 *The Euclidean gradient of* (7) *with respect to* $\mathbf{P}_i$, *for* $i \in [n]$ *is*

$$\sum_{\ell=1}^{K} \omega_\ell \nabla_{\mathbf{P}_i} \mathrm{MW}_\epsilon(\mathbf{P}, \mathbf{P}^\ell) = -\sum_{\ell=1}^{K} \omega_\ell (\Lambda_i^\ell)^*,$$

where $(\Lambda_i^\ell)^*$ *is given by evaluating the orthogonal projection* $\mathrm{P}_{(\Gamma^\ell)^*}(\nabla_{(\Gamma^\ell)^*} \mathrm{MW}_\epsilon)$, *where* $\nabla_{(\Gamma_{i,j}^\ell)^*} \mathrm{MW}_\epsilon = \mathbf{C}_{i,j}^\ell + \epsilon \nabla\Omega((\Gamma_{i,j}^\ell)^*)$ *and* $(\Gamma^\ell)^*$ *is the optimal coupling for* $\mathbf{P}^\ell$. *That is,* $(\Lambda_i^\ell)^*$ *is the auxiliary variable obtained during the solving of the system of matrix linear equations in Proposition* 3.

The complete algorithm for computing the barycenter in (7) is outlined in Algorithm 3 ("Appendix E").

5.2 Block SPD Gromov-Wasserstein discrepancy

The Gromov-Wasserstein (GW) distance (Mémoli, 2011) generalizes the optimal transport to the case where the measures are supported on possibly different metric spaces $\mathcal{X}$ and $\mathcal{Y}$.

Let $\mathbf{D}^x \in \mathbb{R}^{m \times m}$ and $\mathbf{D}^y \in \mathbb{R}^{n \times n}$ represent the similarity (or distance) between elements in metric spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively. Let $\mathbf{p} \in \Sigma_m$ and $\mathbf{q} \in \Sigma_n$ be the marginals corresponding to the elements in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Then, the GW discrepancy between the two distance-marginal pairs $(\mathbf{D}^x, \mathbf{p})$ and $(\mathbf{D}^y, \mathbf{q})$ is defined as

$$\text{GW}((\mathbf{D}^x, \mathbf{p}), (\mathbf{D}^y, \mathbf{q})) := \min_{\gamma \in \Pi(\mathbf{p}, \mathbf{q})} \sum_{i,i',j,j'} \mathcal{L}(D_{i,i'}^x, D_{j,j'}^y) \gamma_{i,j} \gamma_{i',j'},$$

where $D_{k,l}$ denotes the (k, l)-th element in the matrix $\mathbf{D}$ and $\mathcal{L}$ is a loss between the distance pairs. Common choices of $\mathcal{L}$ include the L_2 distance and the KL divergence.

We now generalize the GW framework to our setting where the marginals are SPD matrix-valued measures. Let $(\mathbf{D}^x, \mathbf{P})$ and $(\mathbf{D}^y, \mathbf{Q})$ be two distance-marginal pairs, where the Dirac measures are given by $\sum_i \mathbf{P}_i \delta_{x_i}$, $\sum_j \mathbf{Q}_j \delta_{y_j}$ respectively, for $\{x_i\}_{i \in [m]} \subset \mathcal{X}$, $\{y_j\}_{j \in [n]} \subset \mathcal{Y}$. The marginals are tensor-valued with $\mathbf{P} \in \Delta_m(\mathbb{S}_{++}^d)$, $\mathbf{Q} \in \Delta_n(\mathbb{S}_{++}^d)$. We define the SPD generalized GW discrepancy as

$$\text{MGW}((\mathbf{D}^x, \mathbf{P}), (\mathbf{D}^y, \mathbf{Q})) := \min_{\mathbf{\Gamma} \in \mathcal{M}_{m \times n}^d} \sum_{i,i',j,j'} \mathcal{L}\left(D_{i,i'}^x, D_{j,j'}^y\right) \text{tr}(\mathbf{\Gamma}_{i,j} \mathbf{\Gamma}_{i',j'}), \tag{9}$$

where we use Riemannian optimization (Algorithm 1) to solve problem (9).

Gromov-Wasserstein averaging of distance matrices. The GW formulation with scalar-valued probability measures has been used for averaging distance matrices (Peyré et al., 2016). Building on (9), we consider the problem of averaging distance matrices where the marginals are SPD-valued. Let $\{(\mathbf{D}^\ell, \mathbf{P}^\ell)\}_{\ell=1}^K$ with $\mathbf{P}^\ell \in \Delta_{n_\ell}(\mathbb{S}_{++}^d)$, be a set of distance-marginal pairs on K incomparable domains. Suppose the barycenter marginals $\bar{\mathbf{P}} \in \Delta_n(\mathbb{S}_{++}^d)$ are given, the goal is to find the average distance matrix $\bar{\mathbf{D}}$ by solving

$$\bar{\mathbf{D}} = \arg\min_{\mathbf{D} \in \mathbb{S}^n : D_{i,j} \geq 0} \sum_{\ell=1}^K \omega_\ell \, \text{MGW}\left((\mathbf{D}, \bar{\mathbf{P}}), (\mathbf{D}^\ell, \mathbf{P}^\ell)\right), \tag{10}$$

where the given weights satisfy $\sum_\ell \omega_\ell = 1$. Problem (10) can be solved via a block coordinate descent method, that iteratively updates the couplings $\{\mathbf{\Gamma}^\ell\}_{\ell=1}^K$ and the distance matrix $\bar{\mathbf{D}}$. The update of the coupling is performed via Algorithm 1. For the update of the distance matrix, we show when the loss $\mathcal{L}$ is decomposable, including the case of L_2 distance or the KL divergence, the optimal $\bar{\mathbf{D}}$ admits a closed-form solution. This is a generalization of the result (Peyré et al. 2016, Proposition 3) to SPD-valued marginals.

Proposition 6 *Suppose the loss $\mathcal{L}$ can be decomposed as $\mathcal{L}(a, b) = f_1(a) + f_2(b) - h_1(a)h_2(b)$ with f_1'/h_1' invertible, then (10) has a closed form solution given by $\bar{D}_{i,i'} = \left(\frac{f_1'}{h_1'}\right)^{-1}(h_{i,i'})$ with*

$$h_{i,i'} = \left(\frac{\sum_{\ell=1}^K \omega_\ell \text{tr}\left(\sum_j \mathbf{\Gamma}_{i,j}^\ell \sum_{j'} h_2(D_{j,j'}^\ell) \mathbf{\Gamma}_{i',j'}^\ell\right)}{\text{tr}(\bar{\mathbf{P}}_i \bar{\mathbf{P}}_{i'})} \right).$$

6 Experiments

In this section, we show the utility of the proposed framework in a number of applications. For empirical comparisons, we refer to our approaches, block SPD OT (2), the corresponding Wasserstein barycenter (7), and block SPD Gromov-Wasserstein OT (9) & (10), collectively as RMOT (Riemannian optimized Matrix Optimal Transport). For all the experiments, we use the Riemannian steepest descent method using the Manopt toolbox (Boumal et al., 2014) for implementing Algorithm 1. The codes are available at https://github.com/andyjm3/BlockSPDOT.

6.1 Domain adaptation

We apply our OT framework to the application of unsupervised domain adaptation where the goal is to align the distribution of the source with the target for subsequent tasks.

Suppose we are given the source $\mathbf{p} \in \Sigma_m$ and target marginals $\mathbf{q} \in \Sigma_n$, along with samples $\{\mathbf{X}_i\}_{i=1}^m$, $\{\mathbf{Y}_j\}_{j=1}^n$ from the source and target distributions. The samples are matrix-valued, i.e., $\mathbf{X}_i, \mathbf{Y}_j \in \mathbb{R}^{d \times s}$. We define the cost as $\mathbf{C}_{i,j} = (\mathbf{X}_i - \mathbf{Y}_j)(\mathbf{X}_i - \mathbf{Y}_j)^\top$. It should be noted that $\mathrm{tr}(\mathbf{C}_{i,j}) = \|\mathbf{X}_i - \mathbf{Y}_j\|_F^2$ is the cost function under the 2-Wasserstein OT setting (1).

For domain adaptation, we first learn an optimal coupling between the source and target samples by solving the proposed OT problem (2) with marginals $\mathbf{P}, \mathbf{Q}$ constructed as $\mathbf{P} := \{[\mathbf{P}_i]_{m \times 1} : \mathbf{P}_i = p_i \mathbf{I}\}$ and $\mathbf{Q} := \{[\mathbf{Q}_j]_{n \times 1} : \mathbf{Q}_j = q_j \mathbf{I}\}$. Finally, the source samples are projected to the target domain via barycentric projection. Once the optimal couplings $[\mathbf{\Gamma}_{i,j}^*]_{m \times n}$, the barycentric projection of a source sample $\mathbf{X}_i$ is computed as

$$\hat{\mathbf{X}}_i = \underset{\mathbf{X}_i \in \mathbb{R}^{d \times s}}{\arg\min} \sum_{i,j} \mathrm{tr}((\mathbf{X}_i - \mathbf{Y}_j)(\mathbf{X}_i - \mathbf{Y}_j)^\top \mathbf{\Gamma}_{i,j}^*) = \mathbf{P}_i^{-1}\left(\sum_j \mathbf{\Gamma}_{i,j}^* \mathbf{Y}_j\right). \tag{11}$$

The above approach also works for structured samples. For instance, when the samples are SPD, i.e., $\mathbf{X}_i, \mathbf{Y}_j \in \mathbb{S}_{++}^d$, the projected source sample $\hat{\mathbf{X}}_i$ is now the solution to the matrix Lyapunov equation: $\{\mathbf{P}_i \hat{\mathbf{X}}_i\}_S = \{\sum_j \mathbf{\Gamma}_{i,j}^* \mathbf{Y}_j\}_S$. Here, $\{\mathbf{A}\}_S = (\mathbf{A} + \mathbf{A}^\top)/2$.

For the scalar-valued OT case, discussed in Sect. 2.2, the barycentric projection of a source sample $\mathbf{X}_i$ is computed as

$$\hat{\mathbf{X}}_i = \underset{\mathbf{X}_i \in \mathbb{R}^{d \times s}}{\arg\min} \sum_{i,j} \|\mathbf{X}_i - \mathbf{Y}_j\|_F^2 \gamma_{i,j}^* = p_i^{-1}\left(\sum_j \gamma_{i,j}^* \mathbf{Y}_j\right), \tag{12}$$

where $\boldsymbol{\gamma}^* = [\gamma_{i,j}^*]$ is the optimal coupling matrix of size $m \times n$ for the scalar-valued OT problem.

Contrasting the barycentric projection operations (11) with (12), we observe that (11) allows to capture feature-specific correlations more appropriately. The benefit of the matrix-valued OT modeling over the scalar-valued OT modeling is reflected in the experiments below.

Experimental setup. We employ domain adaptation to classify the test sets (target) of multiclass image datasets, where the training sets (source) have a different class distribution than the test sets. Suppose we are given a training set $\{\mathbf{X}_i\}_{i=1}^m$ and a test set $\{\mathbf{Y}_j\}_{j=1}^n$ where $\mathbf{X}_i, \mathbf{Y}_j \in \mathbb{R}^{d \times s}$ are s (normalized) image samples of the same class in d dimension for each image set i, j. Instead of constructing the cost directly on the input
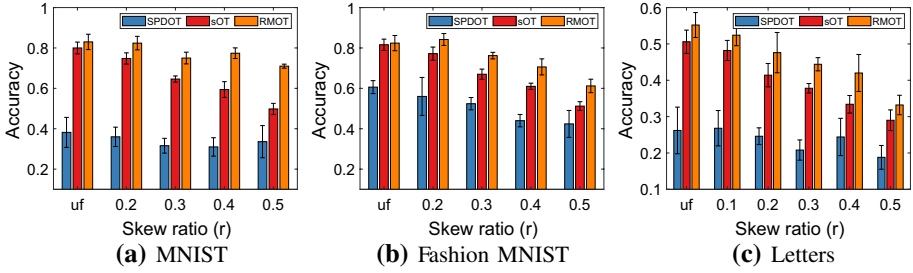
Fig. 1 Domain adaptation and classification results for three datasets: MNIST 1a, Fashion MNIST 1b and Letters 1c. The skew ratio increases from uniform (uf) to $r = 0.5$. For MNIST and Fashion MNIST, uf $= 0.1$ and for Letters, uf $= 1/26$. We observe that the proposed RMOT performs significantly better than the baselines

space, which are not permutation-invariant, we first compute the sample covariances $\mathbf{S}_{x_i} = \mathbf{X}_i\mathbf{X}_i^\top/s$ and $\mathbf{S}_{y_j} = \mathbf{Y}_j\mathbf{Y}_j^\top/s$, $\forall i,j$. Now the cost between i, j is given by $\mathbf{C}_{i,j} = (\mathbf{S}_{x_i} - \mathbf{S}_{y_j})(\mathbf{S}_{x_i} - \mathbf{S}_{y_j})^\top$. Once the block SPD matrix coupling is learnt, the $\mathbf{S}_{x_i}$ covarinaces are projected using the barycerntric projection to obtain $\hat{\mathbf{S}}_{x_i}, i \in [m]$. This is followed by nearest neighbour classification of j based on the Frobenius distance $\|\hat{\mathbf{S}}_{x_i} - \mathbf{S}_{y_j}\|_F \forall i,j$.

We compare the proposed RMOT (2) with the following baselines: (i) sOT: the 2-Wasserstein OT (1) with the cost $c_{i,j} = \mathrm{tr}(\mathbf{C}_{i,j}) = \|\mathbf{S}_{x_i} - \mathbf{S}_{y_j}\|_F^2$ (Courty et al., 2016), and (ii) SPDOT: the 2-Wasserstein OT (1) with the cost as the squared Riemannian geodesic distance between the SPD matrices $\mathbf{S}_{x_i}$ and $\mathbf{S}_{y_j}$ (Yair et al., 2019).

Datasets. We experiment on three multiclass image datasets - handwritten letters (Frey and Slate, 1991), MNIST (LeCun et al., 1998) and Fashion MNIST (Xiao et al., 2017) - with various skewed distributions for the training set. MNIST and Fashion MNIST have 10 classes, while Letters has 26 classes. Specifically, we fix the distribution of the test set to be
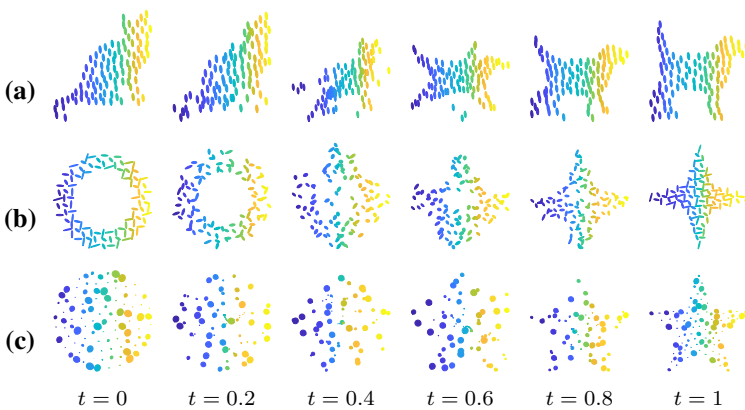

Fig. 2 Tensor-valued shape interpolation obtained using the proposed Gromov-Wasserstein RMOT formulation (Sect. 5.2). We note that each shape consists of several SPD-matrix/tensor valued fields (displayed using ellipses). In (a), the tensors follow a uniform distribution. In (b), the tensors are generated with multiple orientation and in (c), tensors vary smoothly in both size and orientation. The proposed approach takes the anisotropy and orientation of tensor fields into account while interpolating shapes

uniform (with the same number of image sets per class). We increase the proportion of the a randomly chosen class in the training set to the ratio r, where $r = \{uf, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and uf is the ratio corresponding to the uniform distribution of all classes. We reduce the dimension of MNIST, fashion MNIST, and Letters by PCA to $d = 5$ features. We set $s = d$, $m = 250$, and $n = 100$ for each dataset.

Results. Figs. 1a-c shows the classification accuracy on the three datasets. We observe that the proposed RMOT outperforms sOT and SPDOT, especially in more challenging domain adaptation settings, i.e., higher skew ratios. This implies the usefulness of the non-trivial correlations learned by the SPD matrix valued couplings of RMOT.

6.2 Tensor Gromov-Wasserstein distance averaging for shape interpolation

We consider an application of the proposed block SPD Gromov-Wasserstein OT formulation (Sect. 5.2) for interpolating tensor-valued shapes. We are given two distance-marginal pairs $(\mathbf{D}^0, \mathbf{P}^0), (\mathbf{D}^1, \mathbf{P}^1)$ where $\mathbf{D}^0, \mathbf{D}^1 \in \mathbb{R}^{n \times n}$ are distance matrices computed from the shapes and $\mathbf{P}^0, \mathbf{P}^1$ are given tensor fields. The aim is to interpolate between the distance matrices with weights $\boldsymbol{\omega} = (t, 1 - t), t \in [0, 1]$. The interpolated distance matrix $\mathbf{D}^t$ is computed by solving (10) via Riemannian optimization and Proposition 6, with the barycenter tensor fields $\mathbf{P}^t$ given. Finally, the shape is recovered by performing multi-dimensional scaling to the distance matrix.

Figure 2 presents the interpolated shapes with $n = 100$ sample points for the input shapes. The matrices $\mathbf{D}^0, \mathbf{D}^1$ are given by the Euclidean distance and we consider L_2 loss for $\mathcal{L}$. The input tensor fields $\mathbf{P}^0, \mathbf{P}^1$ are generated as uniformly random in (a), cross-oriented in (b) and smoothly varying in (c). For simplicity, we consider the barycenter tensor fields given by the linear interpolation of the inputs, i.e., $\mathbf{P}^t = (1 - t)\mathbf{P}^0 + t\mathbf{P}^1$. In Peyré et al. (2016), we highlight that the marginals are scalar-valued and fixed to be uniform. Here, on the other hand, the marginals are tensor-valued and the resulting distance matrix interpolation would be affected by the relative mass of the tensors, as shown by Proposition 6. The results show the proposed Riemannian optimization approach (Sect. 4) converges to reasonable stationary solutions for non-convex OT problems.

6.3 Tensor field optimal transport mass interpolation

We consider performing optimal transport and displacement interpolation between two tensor fields supported on regular 1-d (or 2-d) grids (Peyré et al., 2019a). We consider a common domain $\mathcal{D} = [0, 1]$ (or $[0, 1]^2$) with the cost defined as $\mathbf{C}_{i,j} = \|\mathbf{x}_i - \mathbf{y}_j\|^2 \mathbf{I}$ for $\mathbf{x}_i, \mathbf{y}_j \in \mathcal{D}$. The marginals $\mathbf{P}, \mathbf{Q}$ are given tensor fields. We first compute the balanced coupling $\boldsymbol{\Gamma}$ by solving an entropy regularized OT problem (8):

$$\min_{\boldsymbol{\Gamma} \in \mathcal{M}_{m \times n}^d(\mathbf{P}, \mathbf{Q})} \sum_{i,j} \left(\mathrm{tr}(\mathbf{C}_{i,j} \boldsymbol{\Gamma}_{i,j}) - \epsilon\, H(\boldsymbol{\Gamma}_{i,j}) \right),$$

where the quantum entropy is defined as $H(\boldsymbol{\Gamma}_{i,j}) := -\mathrm{tr}(\boldsymbol{\Gamma}_{i,j} \log(\boldsymbol{\Gamma}_{i,j}) - \boldsymbol{\Gamma}_{i,j})$. Then, the coupling is used to interpolate between the two tensor fields by generalizing the displacement interpolation (McCann, 1997) to SPD-valued marginals. Please refer to (Peyré et al. 2019a, Sect. 2.2) for more details. It should be noted that due to the balanced nature of our
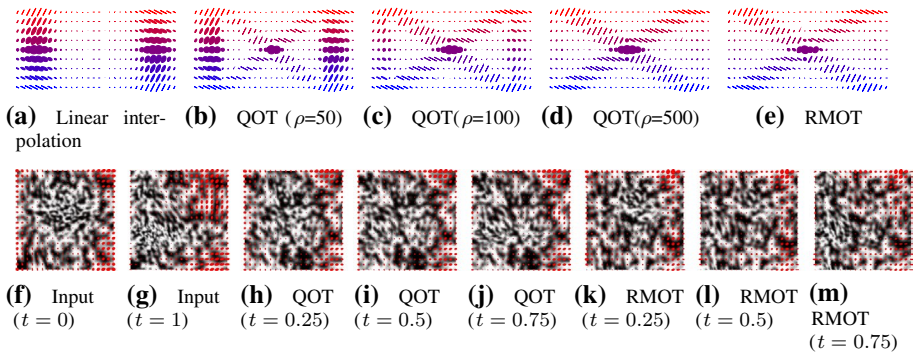
(a) Linear interpolation **(b)** QOT (ρ=50) **(c)** QOT(ρ=100) **(d)** QOT(ρ=500) **(e)** RMOT

(f) Input ($t = 0$) **(g)** Input ($t = 1$) **(h)** QOT ($t = 0.25$) **(i)** QOT ($t = 0.5$) **(j)** QOT ($t = 0.75$) **(k)** RMOT ($t = 0.25$) **(l)** RMOT ($t = 0.5$) **(m)** RMOT ($t = 0.75$)

Fig. 3 Tensor field mass interpolation on 1-d (top) and 2-d (bottom) grids. On the top, each row corresponds to an interpolation where we show 7 evenly-spaced interpolated tensor fields. On the bottom, the inputs are given in (f) and (g). We set $\rho = 100$ for QOT and show 3 evenly-spaced interpolated tensor fields

Fig. 4 Barycenter interpolation. From left to right $t = 0$ (input), $t = 0.25, 0.5, 0.75$ (barycenters), $t = 1$ (input). The top row is QOT and the bottom is RMOT

formulation, we do not need to adjust the couplings after matching as required in (Peyré et al., 2019a).

We compare interpolation results of the proposed (balanced) RMOT with both linear interpolation $(1 − t)\mathbf{P} + t\mathbf{Q}$ for $t \in [0, 1]$ and the *unbalanced* quantum OT (QOT) of (Peyré et al., 2019a). The QOT solves the following problem with quantum KL regularization, i.e.,

$$\min_{\boldsymbol{\Gamma}} \sum_{i,j} \left(\text{tr}(\mathbf{C}_{i,j}\boldsymbol{\Gamma}_{i,j}) - \epsilon\, H(\boldsymbol{\Gamma}_{i,j}) + \rho\, \text{KL}(\boldsymbol{\Gamma}\mathbb{1}|\mathbf{P}) + \rho\, \text{KL}\boldsymbol{\Gamma}^{\top}\mathbb{1}|\mathbf{Q}) \right),$$

where $\text{KL}(\mathbf{P}|\mathbf{Q}) := \sum_i \text{tr}\left(\mathbf{P}_i \log(\mathbf{P}_i) - \mathbf{P}_i \log(\mathbf{Q}_i) - \mathbf{P}_i + \mathbf{Q}_i\right)$ and $\boldsymbol{\Gamma}\mathbb{1} := [\sum_j(\boldsymbol{\Gamma}_{i,j})]_{m \times 1}$ and $\boldsymbol{\Gamma}^{\top}\mathbb{1} := [\sum_i(\boldsymbol{\Gamma}_{i,j})]_{n \times 1}$. For comparability, we set the same ϵ for both QOT and RMOT.

Figure 3 compares the mass interpolation for both 1-d (top) and 2-d (bottom) grids. For the 2-d tensor fields, we further render the tensor fields via a background texture where we perform anisotropic smoothing determined by the tensor direction. To be specific, we follow the procedures in Peyré et al. (2019a) by applying the tensor to the gradient vector of the textures on the grid such that the texture is stretched in the main eigenvector directions of the tensor. In both the settings, we observe that the tensor fields generated from RMOT respect the marginal constraints more closely.

6.4 Tensor field Wasserstein barycenter

We also analyze the Wasserstein barycenters learned by the proposed RMOT approach and qualitatively compare with QOT barycenter (Peyré et al. 2019a, Section 4.1). We test on two tensor fields ($n = 4$) supported 2-d grids.
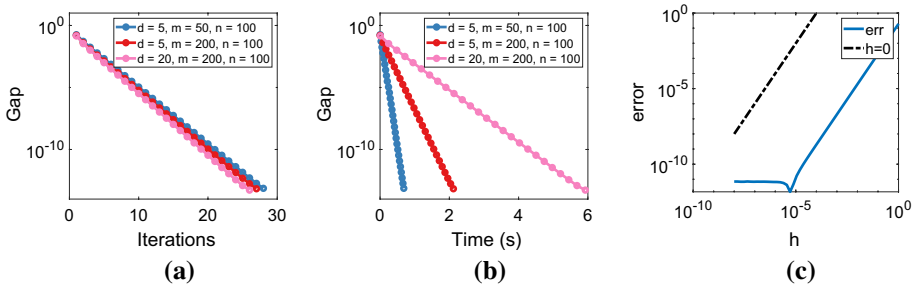
Fig. 5 Convergence of Algorithm 2 in terms of iterations (a), runtime (b), and validity test for retraction (c). For the retraction to be valid, the slope of the continuous line should match the dotted line (which represents the line $h = 0$) and also start from 0 when h tends to 0

Figure 4 compares barycenter from QOT (top) and RMOT (bottom) initialized from the normalized solution of QOT. We observe that the QOT solution is not optimal when the marginal constraint is enforced and the barycenter obtained does not lie in the simplex of tensors. Such a claim is strengthened by comparing the objective value versus the optimal value, obtained by the CVX toolbox (Grant and Boyd, 2014). The objective can be further decreased when initialized from the (normalized) QOT solution, see more discussions in "Appendix C".

7 Conclusion

In this paper, we have discussed the balanced optimal transport (OT) problem involving SPD matrix-valued measures. For the SPD matrix-valued OT problem, the coupling matrix is a block matrix where each block is a symmetric positive definite matrix. The set of such coupling matrices can be endowed with Riemannian geometry, which enables optimization both linear and non-linear objective functions. We have also shown how the SPD-valued OT setup extend many optimal transport problems to general SPD-valued marginals, including the Wasserstein barycenter and the Gromov-Wasserstein (GW) discrepancy. Experiments in a number of applications confirm the benefit of our approach.

Appendix A Convergence of block matrix balancing algorithm and validity of retraction

In Sect. 4, we generalize the matrix scaling algorithm to block matrix cases, which is essential to derive the retraction for the manifold $\mathcal{M}_{m,n}^d$. Here, we empirically show that the algorithm quickly converges and the proposed retraction is valid and satisfies the two conditions: 1) $R_x(0) = x$ and 2) $DR_x(0)[u] = u$, where $Df(x)[u]$ is the derivative of a function at x along direction u.

Convergence. We show in Fig. 5 the convergence of the proposed block matrix balancing procedure in Algorithm 2. We generate the marginals as random SPD matrices for different dimensions d and size m, n. The convergence is measured as the relative gap to satisfy the constraints. We observe that the number of iterations for convergence are similar with different parameters while the runtime increases by increasing the dimension and size.

Validity of retraction. The first condition of retraction is easily satisfied as $R_{\mathbf{\Gamma}}(\mathbf{0}) = \mathrm{MBalance}(\mathbf{\Gamma}) = \mathbf{\Gamma}$. For the second one, we have for any $\mathbf{\Gamma} \in \mathcal{M}_{m,n}^d$ and $\mathbf{U} \in T_{\mathbf{\Gamma}}\mathcal{M}_{m,n}^d$,

$$\mathrm{D}R_{\mathbf{\Gamma}}(\mathbf{0})[\mathbf{U}] = \lim_{h \to 0} \frac{R_{\mathbf{\Gamma}}(h\mathbf{U}) - R_{\mathbf{\Gamma}}(\mathbf{0})}{h}.$$

Hence, we need to numerically verify $R_{\mathbf{\Gamma}}(h\mathbf{U}) - \mathbf{\Gamma} = O(h)\mathbf{U}$ for any $\mathbf{\Gamma}, \mathbf{U}$. We compute an approximation error in terms of the inner product on the tangent space $T_{\mathbf{\Gamma}}\mathcal{M}_{m,n}^d$ as

$$\varepsilon = \left| \langle \mathrm{P}_{\mathbf{\Gamma}}(R_{\mathbf{\Gamma}}(h\mathbf{U}) - \mathbf{\Gamma}), \mathbf{V} \rangle_{\mathbf{\Gamma}} - \langle h\mathbf{U}, \mathbf{V} \rangle_{\mathbf{\Gamma}} \right|,$$

for any $\mathbf{V} \in T_{\mathbf{\Gamma}}\mathcal{M}_{m,n}^d$ different from $\mathbf{U}$. In Fig. 5(c), we show that the slope of the approximation error (as a function of h) matches the dotted line $h = 0$, which suggests hat the error $\varepsilon = O(1)$, thereby indicating that the retraction is valid.

Appendix B Discussion on construction of matrix-valued cost

As highlighted in Proposition 1 for MW$(\mathbf{P}, \mathbf{Q})$ to be a metric for probability measures there are some conditions for the cost $[\mathbf{C}_{i,j}]_{m \times n}$ to satisfy. In the following, we give some examples of how such costs are constructed:

1. Let the samples are given by $\{\mathbf{X}_i\}_{i \in [m]}$, $\{\mathbf{Y}_j\}_{j \in [n]}$, where $\mathbf{X}_i, \mathbf{Y}_j \in \mathbb{R}^{d \times s}$. Define $\mathbf{C}_{i,j} = d(\mathbf{X}_i, \mathbf{Y}_j)^2 \mathbf{I}$, where $d : \mathbb{R}^{d \times s} \times \mathbb{R}^{d \times s} \to \mathbb{R}_+$ is a distance function.
2. Let the samples are given by $\{\mathbf{X}_i\}_{i \in [m]}$, $\{\mathbf{Y}_j\}_{j \in [n]}$, where $\mathbf{X}_i, \mathbf{Y}_j \in \mathbb{R}^{d \times s}$, where $s \geq d$. Assume the matrix $\mathbf{X}_i - \mathbf{Y}_j$ has column full rank. Define $\mathbf{C}_{i,j} = (\mathbf{X}_i - \mathbf{Y}_j)(\mathbf{X}_i - \mathbf{Y}_j)^\top$.

Proof

(1) The first definition of cost trivially satisfies all the conditions due to the metric properties of a well-defined scalar-valued distance.
(2) For the second definition of cost, The first two conditions, i.e., symmetric and positive definite conditions are easily satisfied and we only need to verify the third condition in Proposition 1. The third condition is also satisfied due to the triangle inequality of Mahalanobis distance metric in the vectorized form. That is, for any $\mathbf{A} \succeq \mathbf{0}$, we consider three sets of samples $\{\mathbf{X}_i\}, \{\mathbf{Y}_k\}, \{\mathbf{Z}_j\} \subset \mathbb{R}^{d \times s}$. Then, we have

$$\sqrt{\mathrm{tr}(\mathbf{C}_{i,j}\mathbf{A})} = \sqrt{\mathrm{tr}((\mathbf{X}_i - \mathbf{Z}_j)^\top \mathbf{A}(\mathbf{X}_i - \mathbf{Z}_j))}$$

$$= \sqrt{(\mathrm{vec}(\mathbf{X}_i) - \mathrm{vec}(\mathbf{Z}_j))^\top (\mathbf{I} \otimes \mathbf{A})(\mathrm{vec}(\mathbf{X}_i) - \mathrm{vec}(\mathbf{Z}_j))}$$

$$\leq \sqrt{(\mathrm{vec}(\mathbf{X}_i) - \mathrm{vec}(\mathbf{Y}_k))^\top (\mathbf{I} \otimes \mathbf{A})(\mathrm{vec}(\mathbf{X}_i) - \mathrm{vec}(\mathbf{Y}_k))}$$

$$+ \sqrt{(\mathrm{vec}(\mathbf{Y}_k) - \mathrm{vec}(\mathbf{Z}_j))^\top (\mathbf{I} \otimes \mathbf{A})(\mathrm{vec}(Y_k) - \mathrm{vec}(\mathbf{Z}_j))}$$

$$= \sqrt{\mathrm{tr}((\mathbf{X}_i - \mathbf{Y}_k)^\top \mathbf{A}(\mathbf{X}_i - \mathbf{Y}_k))} + \sqrt{\mathrm{tr}((\mathbf{Y}_k - \mathbf{Z}_j)^\top \mathbf{A}(\mathbf{Y}_k - \mathbf{Z}_j))}$$

$$= \sqrt{\mathrm{tr}(\mathbf{C}_{i,k}\mathbf{A})} + \sqrt{\mathrm{tr}(\mathbf{C}_{k,j}\mathbf{A})},$$

where $\mathrm{vec}(\mathbf{C})$ denotes the vectorization of matrix $\mathbf{C}$ by stacking the columns.
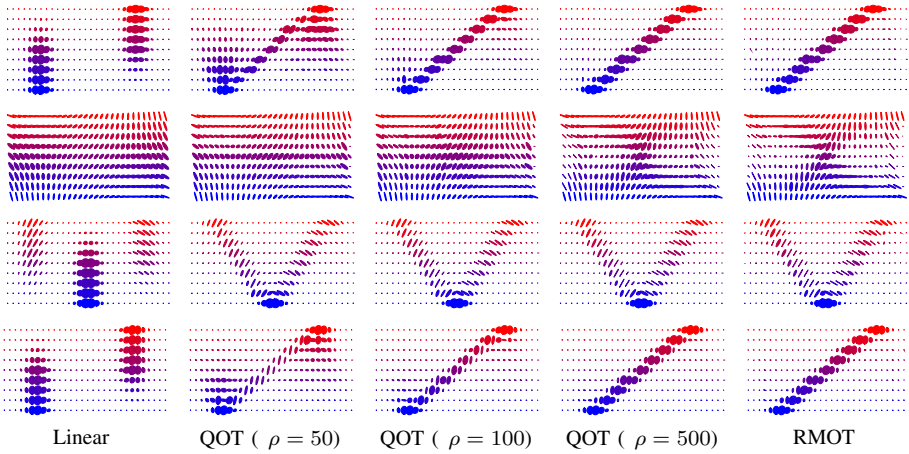
Fig. 6 1-*d* tensor fields mass interpolation. Tensor fields are generated as cross-oriented (first row), multi-oriented (second row), split (third row) and iso-oriented (fourth row)

□

Appendix C Additional experiments

In this section, we give additional experiments to further substantiate the claims made in the main text.

C.1 Tensor field optimal transport mass interpolation

We first provide more details on displacement interpolation considered in the experiment. After we obtain the optimal $\mathbf{\Gamma}^*$, for $t \in [0, 1]$, we compute the interpolated measure at t as

$$\sum_{i,j} ((1 - t)\mathbf{P}_i + t\mathbf{Q}_j)\mathbf{\Gamma}_{i,j}\delta_{x_{i,j}^t},$$

where $x_{i,j}^t$ is the interpolated location on the 2-d grid.

In addition to the experiments presented in the main texts, we also show other examples of tensor fields mass interpolation in Figs. 6 and 7. In Fig. 6, the inputs are given as 1-d tensor fields, which are the first and last row for each subfigure. We compare the
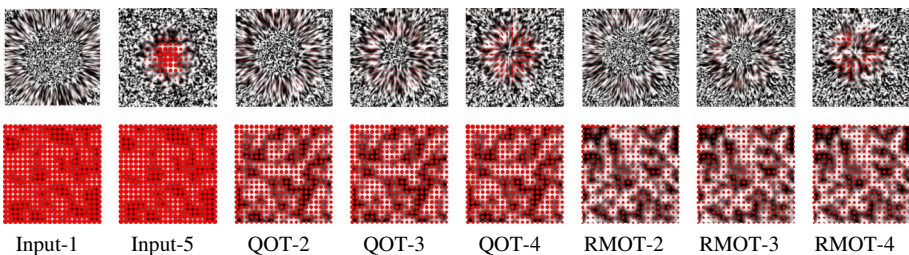

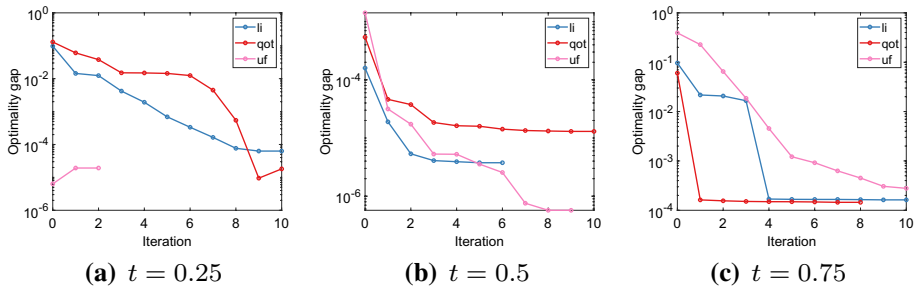**Fig. 7** 2-*d* tensor fields mass interpolation

Fig. 8 Convergence of ($n = 4$) barycenter update initialized from linear interpolation (li), QOT (qot), and uniform identity (uf). Irrespective of the initialization, RMOT continue to achieve better optimality gap (to the CVX optimal solution) with iterations. As the initial optimality gap is high for all the cases, it shows that the linear interpolation (li) and QOT (qot) solutions are not optimal
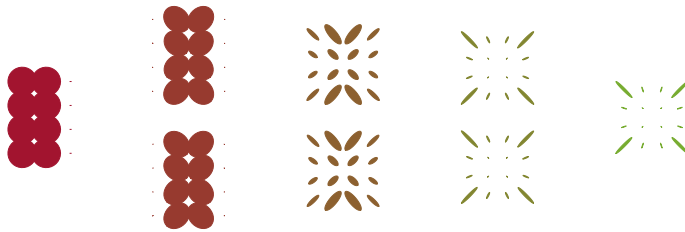

Fig. 9 Tensor field Wasserstein barycenter ($n = 16$). The barycetners are shown in (a) to (e). From left to right $t = 0$ (input), $t = 0.25, 0.5, 0.75$ (barycenters), $t = 1$ (input). The top row is QOT and the bottom is RMOT

interpolation given by the linear interpolation (first column), QOT with different values of ρ and RMOT (last column). In Fig. 7, Input-1 and Input-5 are with $t = 0$ and $t = 1$, respectively. QOT-2 and RMOT-2 are with $t = 0.25$. QOT-3 and RMOT-3 are with $t = 0.5$. QOT-4 and RMOT-4 are with $t = 0.75$.

C.2 Tensor field Wasserstein barycenter

We first show how both linear interpolation and QOT solutions are not optimal. We initialize our Riemannian optimizers for $\bar{\mathbf{P}}$ from the linear interpolation and (normalized) QOT. We also include uniform initialization as a benchmark.

We compare the objective value of $\sum_\ell \omega_\ell \mathrm{MW}_\epsilon(\bar{\mathbf{P}}, \mathbf{P}^\ell)$ against the optimal objective value obtained from the CVX toolbox (Grant and Boyd, 2014). This allows to compute the optimality gap.

In Fig. 8, we see that the optimality gap keeps reducing with iterations even after properly normalizing the barycenter from linear interpolation and (normalized) QOT. This shows that linear interpolation and (normalized) QOT solutions are not optimal. Also, the performance of RMOT with uniform initialization is competitive to that initialized with linear interpolation and (normalized) QOT, implying that RMOT is a competitive solver in itself and obtains better solutions.

Additionally, we show the barycenter results for $n = 16$ along with convergence of RMOT in Fig. 9 and 10. From Fig. 9, we see visually no difference in the solutions obtained by QOT and RMOT, which suggests the solution by QOT (with normalization) is close to optimal. This observation is further validated in Fig. 10 where we see the objective value is already quite small when initialized from the QOT solution.

C.3 Additional experiments on domain adaptation

Here, we perform the experiments of domain adaptation on more challenging tasks, including video based face recognition with YouTube Celebrities (YTC) dataset (Kim et al., 2008) and texture classification via Dynamic Texture (DynTex) (Ghanem and
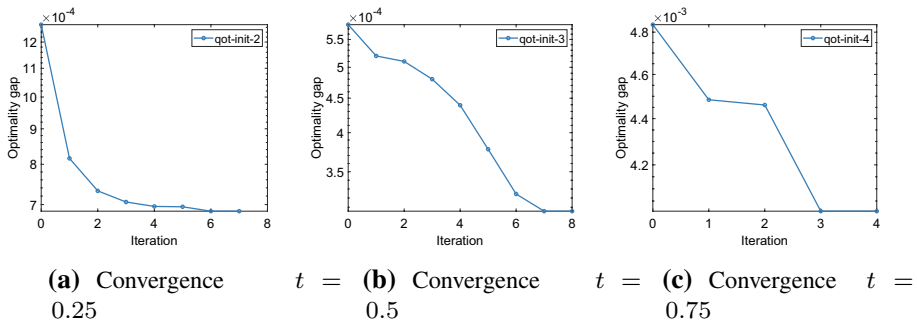

(a) Convergence $t = 0.25$　　**(b)** Convergence $t = 0.5$　　**(c)** Convergence $t = 0.75$

Fig. 10 The convergence plots for RMOT initialized from (normalized) QOT solution. For this setting, the solution from QOT is close to optima

Ahuja, 2010) dataset, where covariance representation learning has shown great promise Huang et al. (2015); Harandi et al. (2014).

Datasets and experimental setup. YTC (Kim et al., 2008) comprises of 1910 low-resolution videos of 47 celebrities from YouTube. Here we only select 9 persons with video size larger than 15. Following standard preprocessing techniques (Huang et al., 2015), we first crop the frames of each video to the detected face regions and resize into 10×10 intensity images. Then we construct the covariance representation for each video, which is a 100×100 SPD matrix. We then apply the geometry-aware principal component analysis for SPD manifold (Horev et al., 2016) via the Bures-Wasserstein Riemannian metric (Bhatia et al., 2019; Han et al., 2021b, a) to reduce the dimensionality to $d = 5$. Finally, we obtain a collection of 194 SPD covariance matrices of size 5×5, each representing one video. Given the relatively small sample size, we select 8 videos per class as the test data and the rest are treated as the training data. Different to the settings in Sect. 6.1, we skew the selected class by sub-selecting a ratio α of the samples in the training set, where $\alpha = 0.2, 0.4, 0.6, 0.8, 1.0$. This is again due to the small data size. To further test the robustness of the algorithms, we then randomly truncate the training size to 100. This results in a training set of 100 videos against a test set of 72 videos. Such randomization of process is repeated 5 times.

DynTex (Ghanem and Ahuja, 2010) collects video sequences of 36 moving scenes, such as sea waves, fire, clouds. For our experiment, we choose 10 classes, each with 20 videos. The subsequent processing steps are the same as for YTC dataset.
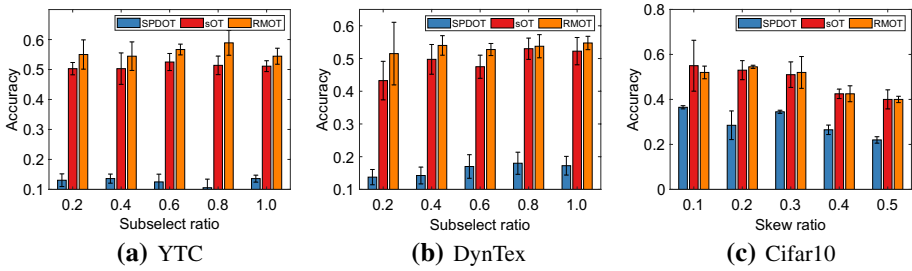
Fig. 11 Additional results on domain adaptation

Finally, we also test on Cifar10 (Krizhevsky et al., 2009) under the same settings as in Sect. 6.1 in the main text. However, because when $d = 5$, much information is lost for this complex dataset, we choose $d = 17$, which captures 70% of the variance in the samples.

Results. The final results are shown in Fig. 11 where we observe consistent good performance of the proposed RMOT compared to both sOT and SPDOT. This strengthens the findings that matrix-valued OT is able to explore more variations in the dataset compared to scalar-valued OT.

Appendix D Proofs

Proof of Proposition 1 For simplicity, we assume $\mathbf{p}, \mathbf{q} > 0$. Otherwise, we can follow (Peyré et al., 2019b) to define $\tilde{p}_j = p_j$ if $p_j > 0$ and 1 otherwise.

We note that $\mathbf{P}$ and $\mathbf{Q}$ are defined as $\mathbf{P} := \{[\mathbf{P}_i]_{m \times 1} : \mathbf{P}_i = p_i \mathbf{I}\}$ and $\mathbf{Q} := \{[\mathbf{Q}_j]_{n \times 1} : \mathbf{Q}_j = q_j \mathbf{I}\}$, where $\mathbf{I}$ is the $d \times d$ identity matrix. With a slight abuse of notation and for simplicity, we define $\mathrm{MW}(\mathbf{p}, \mathbf{q}) := \mathrm{MW}(\mathbf{P}, \mathbf{Q})$.

First, it is easy to verify the symmetry property, i.e., $\mathrm{MW}(\mathbf{p}, \mathbf{q}) = \mathrm{MW}(\mathbf{q}, \mathbf{p})$. For the definiteness, when $\mathbf{p} = \mathbf{q}$, we have $\mathbf{C}_{i,i} = \mathbf{0}$ and $\mathbf{C}_{i,j} > \mathbf{0}$ for $i \neq j$. Hence the optimal coupling is a block diagonal matrix with $\mathbf{\Gamma}_{i,i} = p_i \mathbf{I}$. Hence $\mathrm{MW}(\mathbf{p}, \mathbf{q}) = 0$. For the opposite direction, if $\mathrm{MW}(\mathbf{p}, \mathbf{q}) = 0$, we always need to have $\mathbf{\Gamma}_{i,j} = \mathbf{0}$, for $i \neq j$ because $\mathrm{tr}(\mathbf{C}_{i,j}\mathbf{\Gamma}_{i,j}) > 0$ for any $\mathbf{C}_{i,j} > \mathbf{0}$ and $i \neq j$. Thus, $\mathbf{\Gamma}_{i,i} \neq \mathbf{0}$, which gives $\mathbf{C}_{i,i} = \mathbf{0}$ and $\mathbf{p} = \mathbf{q}$.

Finally, for triangle inequality, given $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \Sigma_n$, and optimal matrix coupling $\mathbf{\Gamma}, \mathbf{\Delta}$ between $(\mathbf{a}, \mathbf{b})$ and $(\mathbf{b}, \mathbf{c})$, respectively. That is, $\sum_j \mathbf{\Gamma}_{i,j} = a_i \mathbf{I}, \sum_i \mathbf{\Gamma}_{i,j} = b_j \mathbf{I}$ and similarly $\sum_j \mathbf{\Delta}_{i,j} = b_i \mathbf{I}, \sum_i \mathbf{\Delta}_{i,j} = c_j \mathbf{I}$. We now follow the same strategy by gluing the coupling $\mathbf{\Gamma}, \mathbf{\Delta}$ in (Peyré et al., 2019b; Villani, 2021). That is, we define a coupling $\mathbf{T}$ as

$$\mathbf{T}_{i,j} = \sum_k \frac{1}{2b_k} (\mathbf{\Gamma}_{i,k}\mathbf{\Delta}_{k,j} + \mathbf{\Delta}_{k,j}\mathbf{\Gamma}_{i,k}), \quad \forall i, j.$$

We can verify $\mathbf{T}_{i,j} \in \mathbb{S}_+^d$, given $\mathbf{\Gamma}_{i,j}, \mathbf{\Delta}_{i,j} \in \mathbb{S}_+^d$. Furthermore, we have $\forall i, j$,

$$\sum_j \mathbf{T}_{i,j} = \sum_k \frac{1}{2b_k} \left(\mathbf{\Gamma}_{i,k} \sum_j \mathbf{\Delta}_{k,j} + \sum_j \mathbf{\Delta}_{k,j}\mathbf{\Gamma}_{i,k} \right) = \sum_k \mathbf{\Gamma}_{i,k} = a_i \mathbf{I},$$

$$\sum_i \mathbf{T}_{i,j} = \sum_k \frac{1}{2b_k} \left(\sum_i \mathbf{\Gamma}_{i,k}\mathbf{\Delta}_{k,j} + \mathbf{\Delta}_{k,j} \sum_i \mathbf{\Gamma}_{i,k} \right) = \sum_k \mathbf{\Delta}_{k,j} = c_j \mathbf{I}.$$

Hence, $[\mathbf{T}_{i,j}]_{m\times n}$ is a valid coupling between $(\mathbf{a}, \mathbf{c})$. Let $\mathbf{P}_i = a_i\mathbf{I}, \mathbf{Q}_j = c_j\mathbf{I}$ and the corresponding samples as $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ for measures $\mathbf{a}, \mathbf{b}, \mathbf{c}$ respectively. Then,

$$
\begin{aligned}
\mathrm{MW}(\mathbf{a}, \mathbf{c}) &= \left(\min_{\mathbf{A}\in\Pi(n,n,d,\mathbf{P},\mathbf{Q})} \sum_{i,j} \mathrm{tr}\left(\mathbf{C}_{i,j}\mathbf{A}_{i,j}\right) \right)^{1/2} \leq \left(\sum_{i,j} \mathrm{tr}\left(\mathbf{C}_{i,j}\mathbf{T}_{i,j}\right) \right)^{1/2} \\
&= \left(\sum_{i,j,k} \frac{1}{2b_k} \mathrm{tr}\left(\mathbf{C}_{i,j}(\mathbf{\Gamma}_{i,k}\mathbf{\Delta}_{k,j} + \mathbf{\Delta}_{k,j}\mathbf{\Gamma}_{i,k})\right) \right)^{1/2} \\
&\leq \left(\sum_{i,j,k} \frac{1}{2b_k} \left(\sqrt{\mathrm{tr}\left(\mathbf{C}_{i,k}(\mathbf{\Gamma}_{i,k}\mathbf{\Delta}_{k,j} + \mathbf{\Delta}_{k,j}\mathbf{\Gamma}_{i,k})\right)} + \sqrt{\mathrm{tr}\left(\mathbf{C}_{k,j}(\mathbf{\Gamma}_{i,k}\mathbf{\Delta}_{k,j} + \mathbf{\Delta}_{k,j}\mathbf{\Gamma}_{i,k})\right)} \right)^2 \right)^{1/2} \\
&\leq \left(\sum_{i,j,k} \frac{1}{2b_k} \mathrm{tr}\left(\mathbf{C}_{i,k}(\mathbf{\Gamma}_{i,k}\mathbf{\Delta}_{k,j} + \mathbf{\Delta}_{k,j}\mathbf{\Gamma}_{i,k})\right) \right)^{1/2} \\
&\quad + \left(\sum_{i,j,k} \frac{1}{2b_k} \mathrm{tr}\left(\mathbf{C}_{k,j}(\mathbf{\Gamma}_{i,k}\mathbf{\Delta}_{k,j} + \mathbf{\Delta}_{k,j}\mathbf{\Gamma}_{i,k})\right) \right)^{1/2} \\
&= \left(\sum_{i,k} \mathrm{tr}\left(\mathbf{C}_{i,k}\mathbf{\Gamma}_{i,k}\right) \right)^{1/2} + \left(\sum_{k,j} \mathrm{tr}\left(\mathbf{C}_{k,j}\mathbf{\Delta}_{k,j}\right) \right)^{1/2} \\
&= \mathrm{MW}(\mathbf{a}, \mathbf{b}) + \mathrm{MW}(\mathbf{b}, \mathbf{c}),
\end{aligned}
$$

where the second inequality is by assumption (iii) of the proposition and the third inequality is due to the Minkowski inequality. This completes the proof. $\qquad\square$

Proof of Proposition 2 For a given feasible element $\mathbf{\Gamma} \in \mathcal{M}_{m,n}^d(\mathbf{P}, \mathbf{Q})$, we can construct a family of feasible elements. For example, choose $0 \leq \zeta < \min_{i,j}\{\lambda_{\min}(\mathbf{\Gamma}_{i,j})\}$. Then, we can add/subtract the equal number of $\zeta\mathbf{I}$ and the result is still feasible. In other words, the set is smooth in a ball around the element $\mathbf{\Gamma}$ of radius ζ. $\qquad\square$

Proof of Proposition 3 Following (Mishra and Sepulchre, 2016), the projection is derived orthogonal to the Riemannian metric (5) as

$$
P_{\mathbf{\Gamma}}(\mathbf{S}) = \arg\min_{\mathbf{U}\in T_{\mathbf{\Gamma}}\mathcal{M}_{m,n}^d} f(\mathbf{U}) = -g_{\mathbf{\Gamma}}(\mathbf{U}, \mathbf{S}) + \frac{1}{2}g_{\mathbf{\Gamma}}(\mathbf{U}, \mathbf{U}). \tag{13}
$$

The Lagrangian of problem (13) is

$$
f(\mathbf{U}) - \mathrm{tr}\left(\mathbf{\Lambda}_i \sum_j \mathbf{U}_{i,j} \right) - \mathrm{tr}\left(\mathbf{\Theta}_j \sum_i \mathbf{U}_{i,j} \right), \tag{14}
$$

where $\mathbf{\Lambda}_i, \mathbf{\Theta}_j$ are dual variables for $i \in [m], j \in [n]$. The orthogonal projection follows from the stationary conditions of (14). $\qquad\square$

Proof of Proposition 4 Given the manifold $\mathcal{M}_{m,n}^d$ is a submanifold of $\times_{m,n}\mathbb{S}_{++}^d$ with affine-invariant (AI) Riemannian metric, the Riemannian gradient is given by

$$
\mathrm{grad}F(\mathbf{\Gamma}) = P_{\mathbf{\Gamma}}([\mathrm{grad}_{ai}F(\mathbf{\Gamma}_{i,j})]_{m\times n}) = P_{\mathbf{\Gamma}}([\mathbf{\Gamma}_{i,j}\{\nabla F(\mathbf{\Gamma}_{i,j})\}_S\mathbf{\Gamma}_{i,j}]_{m\times n}),
$$

where $\mathrm{grad}_{ai}F(\mathbf{X})$ is the Riemannian gradient of $\mathbf{X} \in \mathbb{S}_{++}^d$ with AI metric. Similarly, the Riemannian Hessian $\mathrm{Hess}F(\mathbf{\Gamma})[\mathbf{U}] = \nabla_{\mathbf{U}}\mathrm{grad}F(\mathbf{\Gamma})$ where ∇ denotes the Riemannian

connection. For submanifolds, the connection $\nabla_{\mathbf{U}} \text{grad} F(\mathbf{\Gamma}) = P_{\mathbf{\Gamma}}([\tilde{\nabla}_{\mathbf{U}_{i,j}}(\text{grad} F(\mathbf{\Gamma}_{i,j}))]_{m \times n})$, where $\tilde{\nabla}$ represents the connection of $\mathbb{S}_{++}^d$. From (Sra and Hosseini, 2015), $\tilde{\nabla}_{\mathbf{U}_{i,j}} \text{grad} F(\mathbf{\Gamma}_{i,j}) = \text{Dgrad} F(\mathbf{\Gamma}_{i,j})[\mathbf{U}_{i,j}] - \{\mathbf{U}_{i,j} \mathbf{\Gamma}_{i,j}^{-1} \text{grad} F(\mathbf{\Gamma}_{i,j})\}_{\mathbf{S}}$. Hence, the proof is complete. $\qquad\square$

Proof of Theorem 2 We first write the Lagrange dual function as

$$g(\mathbf{\Lambda}, \mathbf{\Theta}, \mathbf{\Psi}) = \min_{\mathbf{\Gamma} = [\mathbf{\Gamma}_{i,j}]_{m \times n}} F(\mathbf{\Gamma}) + \sum_i \text{tr}\left(\mathbf{\Lambda}_i\left(\sum_j \mathbf{\Gamma}_{i,j} - \mathbf{P}_i\right)\right)$$
$$+ \sum_j \text{tr}\left(\mathbf{\Theta}_j\left(\sum_i \mathbf{\Gamma}_{i,j} - \mathbf{Q}_j\right)\right) - \sum_{i,j} \text{tr}\left(\mathbf{\Psi}_{i,j} \mathbf{\Gamma}_{i,j}\right),$$

where we relax the SPD constraint on $\mathbf{\Gamma}_{i,j}$ to the semidefinite constraint, i.e. $\mathbf{\Gamma}_{i,j} \succeq \mathbf{0}$, for some dual variable $\mathbf{\Lambda}_i, \mathbf{\Theta}_j \in \mathbb{S}^d$ and $\mathbf{\Psi}_{i,j} \succeq \mathbf{0}$. Given the function F is convex with nonempty constraint set, by Slater's condition, strong duality holds and the primal and dual variables should jointly satisfy the KKT conditions.

First, we notice by complementary slackness, $\text{tr}(\mathbf{\Psi}_{i,j}^* \mathbf{\Gamma}_{i,j}^*) = 0$ for $\mathbf{\Gamma}_{i,j}^* \succ \mathbf{0}$. This implies that $\mathbf{\Psi}_{i,j}^* = \mathbf{0}$ since $\mathbf{\Psi}_{i,j}^* \succeq \mathbf{0}$. Note that in some cases $\mathbf{\Gamma}_{i,j}^*$ may be rank-deficient (i.e., some eigenvalues are close to zero), which gives rise to non-zero $\mathbf{\Psi}_{i,j}^*$. Regardless, from the optimality condition, it always satisfies for optimal $\mathbf{\Gamma}_{i,j}^*, \mathbf{\Lambda}_i^*, \mathbf{\Theta}_j^*$,

$$\mathbf{\Gamma}_{i,j}^*(\nabla F(\mathbf{\Gamma}_{i,j}^*) + \mathbf{\Lambda}_i^* + \mathbf{\Theta}_j^*)\mathbf{\Gamma}_{i,j}^* = \mathbf{0}, \tag{15}$$

due to that $\mathbf{\Gamma}_{i,j}^*$ is orthogonal to $\mathbf{\Psi}_{i,j}^*$. $\nabla F(\mathbf{\Gamma}_{i,j}^*)$ denotes the block partial derivative of F with respect to $\mathbf{\Gamma}_{i,j}$ at optimality. On the other hand, to perform Riemannian optimization, the Riemannian gradient is first computed for the primal objective F as

$$\text{grad} F(\mathbf{\Gamma}) = P_{\mathbf{\Gamma}}([\mathbf{\Gamma}_{i,j}\{\nabla F(\mathbf{\Gamma}_{i,j})\}_{\mathbf{S}} \mathbf{\Gamma}_{i,j}]_{m \times n}),$$

which from the definition of orthogonal projection, gives

$$\text{grad} F(\mathbf{\Gamma}_{i,j}) = \mathbf{\Gamma}_{i,j}\left(\nabla F(\mathbf{\Gamma}_{i,j}) + \tilde{\mathbf{\Lambda}}_i + \tilde{\mathbf{\Theta}}_j\right)\mathbf{\Gamma}_{i,j},$$

where $\text{grad} F(\mathbf{\Gamma}_{i,j})$ represents the Riemannian partial derivative and $\tilde{\mathbf{\Lambda}}_i, \tilde{\mathbf{\Theta}}_j \in \mathbb{S}^d$ are computed such that

$$\begin{cases} \sum_i \mathbf{\Gamma}_{i,j}(\mathbf{C}_{i,j} + \epsilon \nabla \Omega(\mathbf{\Gamma}_{i,j}) + \tilde{\mathbf{\Lambda}}_i + \tilde{\mathbf{\Theta}}_j)\mathbf{\Gamma}_{i,j} = \mathbf{0}, \ \forall j \\ \sum_j \mathbf{\Gamma}_{i,j}(\mathbf{C}_{i,j} + \epsilon \nabla \Omega(\mathbf{\Gamma}_{i,j}) + \tilde{\mathbf{\Lambda}}_i + \tilde{\mathbf{\Theta}}_j)\mathbf{\Gamma}_{i,j} = \mathbf{0}, \ \forall i. \end{cases} \tag{16}$$

Comparing (16) to (15), we see that at optimality, there exists $\mathbf{\Lambda}_i^*, \mathbf{\Theta}_j^*$ such that for all i, j, the conditions (16) are satisfied, with $\tilde{\mathbf{\Lambda}}_i = \mathbf{\Lambda}_i^* + \Delta, \tilde{\mathbf{\Theta}}_j = \mathbf{\Theta}_j^* - \Delta$, for any symmetric matrix Δ, i.e., the Riemannian gradient $\text{grad} F(\mathbf{\Gamma}_{i,j}^*) = \mathbf{0}$, thus completing the proof. $\qquad\square$

Proof of Proposition 5 For each regularized OT problem, we consider the Lagrange dual problem of $\min_{\mathbf{\Gamma} \in \mathcal{M}_{m,n}^d} \text{MW}_\epsilon(\bar{\mathbf{P}}, \mathbf{P}^\ell)$, which is given as

$$\mathcal{L}_{\mathrm{MW}_\epsilon} = \max_{\substack{\Lambda_i^\ell, \Theta_j^\ell \in \mathbb{S}^d \\ \Psi_{i,j}^\ell \succeq 0}} \min_{\Gamma_{i,j}^\ell} \sum_{i,j} \left(\mathrm{tr}(\mathbf{C}_{i,j}^\ell \Gamma_{i,j}^\ell) + \epsilon\, \Omega(\Gamma_{i,j}^\ell) \right) + \sum_i \mathrm{tr}\left(\Lambda_i^\ell \left(\sum_j \Gamma_{i,j}^\ell - \bar{\mathbf{P}}_i \right) \right)$$

$$+ \sum_j \mathrm{tr}\left(\Theta_j^\ell \left(\sum_i \Gamma_{i,j}^\ell - \mathbf{P}_j^\ell \right) \right) - \sum_{i,j} \mathrm{tr}\left(\Psi_{i,j}^\ell \Gamma_{i,j}^\ell \right).$$

$$(17)$$

From the Lagrangian (17), it is easy to see the Euclidean gradient of the barycenter problem with respect to $\bar{\mathbf{P}}_i$ is $-\sum_\ell \Lambda_i^\ell$ with the dual optimal Λ_i^ℓ for problem (17). The proof is complete by substituting the objective $F(\Gamma) = \sum_{i,j} \left(\mathrm{tr}(\mathbf{C}_{i,j} \Gamma_{i,j} + \epsilon\Omega(\Gamma_{i,j})) \right)$ as in Theorem 2. $\qquad\square$

Proof of Proposition 6 First we rewrite SPD matrix-valued GW discrepancy as

$$\mathrm{MGW}\left((\bar{\mathbf{D}}, \bar{\mathbf{P}}), (\mathbf{D}^\ell, \mathbf{P}^\ell) \right)$$

$$= \sum_{i,i',j,j'} \left(f_1(\bar{D}_{i,i'}) + f_2(D_{j,j'}^\ell) - h_1(\bar{D}_{i,i'}) h_2(D_{j,j'}^\ell) \right) \mathrm{tr}(\Gamma_{i,j}^\ell \Gamma_{i',j'}^\ell)$$

$$= \sum_{i,j} \mathrm{tr}\left(\Gamma_{i,j} \left(\sum_{i'} f_1(\bar{D}_{i,i'}) \bar{\mathbf{P}}_{i'} + \sum_{j'} f_2(D_{j,j'}^\ell) \mathbf{P}_{j'}^\ell - \sum_{i'} h_1(\bar{D}_{i,i'}) \sum_{j'} h_2(D_{j,j'}^\ell) \Gamma_{i',j'}^\ell \right) \right),$$

where we use the fact that $\Gamma_{i,j}^\ell$ are optimal and satisfy the constraints $\sum_j \Gamma_{i,j}^\ell = \bar{\mathbf{P}}_i$ and $\sum_i \Gamma_{i,j}^\ell = \mathbf{P}_j^\ell$. By the first order condition, we have

$$\mathrm{tr}\left(\bar{\mathbf{P}}_i f_1'(\bar{D}_{i,i'}) \bar{\mathbf{P}}_{i'} - h_1'(\bar{D}_{i,i'}) \sum_j \Gamma_{i,j} \left(\sum_{j'} h_2(D_{j,j'}^\ell) \Gamma_{i',j'}^\ell \right) \right) = 0, \quad \forall i, i' \in [m],$$

which gives the desired result. $\qquad\square$

Appendix E Riemannian geometry for block SPD Wasserstein barycenter

Riemannian geometry of $\Delta_n(\mathbb{S}_{++}^d)$**.** In (Mishra et al., 2019), the authors endow a Riemannian manifold structure for the set $\Delta_n(\mathbb{S}_{++}^d) := \{ \mathbf{P} = [\mathbf{P}_i]_{n\times 1} : \sum_i \mathbf{P}_i = \mathbf{I} \}$. Its tangent space is given by $T_\mathbf{P}\Delta_n(\mathbb{S}_{++}^d) = \{ (\mathbf{U}_1, ..., \mathbf{U}_n) : \mathbf{U}_i \in \mathbb{S}^d, \sum_i \mathbf{U}_i = \mathbf{0} \}$. By introducing the affine-invariant metric $\langle \mathbf{U}, \mathbf{V} \rangle_\mathbf{P} = \sum_i \mathrm{tr}(\mathbf{P}_i^{-1} \mathbf{U}_i \mathbf{P}_i^{-1} \mathbf{V}_i)$, $\Delta_n(\mathbb{S}_{++}^d)$ has a submanifold structure. The retraction from the tangent space to the manifold is derived as

$$R_\mathbf{P}(\mathbf{U}) = (\hat{\mathbf{P}}_{\mathrm{sum}}^{-1/2} \hat{\mathbf{P}}_1 \hat{\mathbf{P}}_{\mathrm{sum}}^{-1/2}, ..., \hat{\mathbf{P}}_{\mathrm{sum}}^{-1/2} \hat{\mathbf{P}}_n \hat{\mathbf{P}}_{\mathrm{sum}}^{-1/2}),$$

where $\hat{\mathbf{P}}_i = \mathbf{P}_i(\exp(\mathbf{P}_i^{-1}\mathbf{U}_i))$ and $\hat{\mathbf{P}}_{\mathrm{sum}} = \sum_i \hat{\mathbf{P}}_i$.

The Riemannian gradient of a function $F : \Delta_n(\mathbb{S}_{++}^d) \to \mathbb{R}$ is computed as

$$\mathrm{grad}F(\mathbf{P}) = \mathrm{Proj}_\mathbf{P}\left(\left(\mathbf{P}_1\{\nabla F(\mathbf{P}_1)\}_\mathrm{S} \mathbf{P}_1, ..., \mathbf{P}_n\{\nabla F(\mathbf{P}_n)\}_\mathrm{S} \mathbf{P}_n \right) \right),$$

where the orthogonal projection $\mathrm{P}_\mathbf{P}$ of a of $\mathbf{S} = (\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_n)$ such that $\mathbf{S}_i \in \mathbb{S}^d$ is

$$\text{Proj}_{\mathbf{P}}(\mathbf{S}) = (\mathbf{S}_1 + \mathbf{P}_1\mathbf{\Lambda}\mathbf{P}_1, ..., \mathbf{S}_n + \mathbf{P}_n\mathbf{\Lambda}\mathbf{P}_n),$$

where $\mathbf{\Lambda} \in \mathbb{S}^d$ is the solution to the linear equation $\sum_i \mathbf{P}_i\mathbf{\Lambda}\mathbf{P}_i = -\sum_i \mathbf{S}_i$.

Optimization for Wasserstein barycenter. With the Riemannian geometry defined for the simplex of SPD matrices, we can update the barycenter by Riemannian optimization as shown in Algorithm 3.

Algorithm 3 Block-SPD Wasserstein barycenter

1: **Input:** block-SPD marginals $\{\mathbf{P}^\ell\}_{\ell=1}^K$, cost matrices $\{\mathbf{C}^\ell\}_{\ell=1}^K$.
2: Initialize $\bar{\mathbf{P}}^0$ as uniform distribution.
3: **for** $t = 1, ..., T$ **do**
4:    **for** $\ell = 1, ..., K$ **do**
5:       Compute $(\mathbf{\Lambda}^\ell)^*$ as in Proposition 5.
6:    **end for**
7:    Compute the Riemannian gradient of (7) by orthogonally projecting $\sum_\ell \omega_\ell(\mathbf{\Lambda}^\ell)^*$ onto the tangent space of $\Delta_n(\mathbb{S}_{++}^d)$.
8:    Update $\bar{\mathbf{P}}^t$ by retracting on $\Delta_n(\mathbb{S}_{++}^d)$ with a step size.
9: **end for**
10: **Output:** Barycenter $\bar{\mathbf{P}}^T$.

Author contributions AH, BM, PJ, and JG contributed to the analysis and developed the theoretical parts. AH, BM, and PJ conceived and designed the experiments. AH, BM, and PJ performed the experiments. AH, BM, PJ, and JG wrote the paper.

Availability of data and material Link to data and code is included in the manuscript (Sect. 6)

Code availability Link to data and code is included in the manuscript (Sect. 6)

Declarations

Conflict of interest The university of Sydney and Microsoft India

Consent to participate Not Applicable

Consent for publication Not Applicable

Ethics approval Not Applicable

References

Absil, P.A., Mahony, R., & Sepulchre, R. (2008). *Optimization algorithms on matrix manifolds*. Princeton University Press.

Absil, P.-A., Baker, Christopher G., & Gallivan, Kyle A. (2007). Trust-region methods on riemannian manifolds. *Foundations of Computational Mathematics, 7*(3), 303–330.

Agarwal, N., Boumal, N., Bullins, B., & Cartis, C. (2018). Adaptive regularization with cubics on manifolds. arXiv:1806.00065.

Alliez, Pierre, Cohen-Steiner, David, Devillers, Olivier, Lévy, Bruno, & Desbrun, Mathieu. (2003). Anisotropic polygonal remeshing. *ACM Transactions on Graphics, 22*(3), 485–493.

Assaf, Yaniv, & Pasternak, Ofer. (2008). Diffusion tensor imaging (DTI)-based white matter mapping in brain research: a review. *Journal of Molecular Neuroscience, 34*(1), 51–61.

Bhatia, R. (2009). *Positive definite matrices*. Princeton University Press.

Bhatia, Rajendra, Jain, Tanvi, & Lim, Yongdo. (2019). On the Bures-Wasserstein distance between positive definite matrices. *Expositiones Mathematicae, 37*(2), 165–191.

Boumal, N. (Aug 2020). An introduction to optimization on smooth manifolds. Available online. URL http://www.nicolasboumal.net/book.

Boumal, Nicolas, Mishra, Bamdev, Absil, P.-A., & Sepulchre, Rodolphe. (2014). Manopt, a matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research, 15*(1), 1455–1459.

Carlen, Eric A., & Maas, Jan. (2014). An analog of the 2-Wasserstein metric in non-commutative probability under which the Fermionic Fokker-Planck equation is gradient flow for the entropy. *Communications in Mathematical Physics, 331*(3), 887–926.

Chen, L., Gan, Z., Cheng, Y., Li, L., Carin, L., & Liu, J. (2020). Graph optimal transport for cross-domain alignment. In International Conference on Machine Learning, pages 1542–1553. PMLR.

Chen, Yongxin, Georgiou, Tryphon T., & Tannenbaum, Allen. (2017). Matrix optimal mass transport: a quantum mechanical approach. *IEEE Transactions on Automatic Control, 63*(8), 2612–2619.

Chen, Yongxin, Georgiou, Tryphon T., & Tannenbaum, Allen. (2018). Vector-valued optimal mass transport. *SIAM Journal on Applied Mathematics, 78*(3), 1682–1696.

Chizat, L., Peyre, G., Schmitzer, B., & Vialard, F.-X. (2018). Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis, 274*(11), 3090–3123.

Courty, N., Flamary, Rémi, & Tuia, D. (2014). Domain adaptation with regularized optimal transport. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 274–289. Springer.

Courty, Nicolas, Flamary, Rémi., Tuia, Devis, & Rakotomamonjy, Alain. (2016). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(9), 1853–1865.

Cuturi, Marco. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems, 26,* 2292–2300.

Douik, Ahmed, & Hassibi, Babak. (2019). Manifold optimization over the set of doubly stochastic matrices: A second-order geometry. *IEEE Transactions on Signal Processing, 67*(22), 5761–5774.

Frey, Peter W., & Slate, David J. (1991). Letter recognition using Holland-style adaptive classifiers. *Machine Learning, 6*(2), 161–182.

Galerne, Bruno, Gousseau, Yann, & Morel, Jean-Michel. (2010). Random phase textures: Theory and synthesis. *IEEE Transactions on Image Processing, 20*(1), 257–267.

Georgiou, Tryphon T., & Pavon, Michele. (2015). Positive contraction mappings for classical and quantum Schrödinger systems. *Journal of Mathematical Physics, 56*(3), 033301.

Ghanem, B., & Ahuja, N. (2010). Maximum margin distance learning for dynamic texture recognition. In European Conference on Computer Vision, pages 223–236. Springer.

Grant, M., & Boyd, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1.

Gurvits, Leonid. (2004). Classical complexity and quantum entanglement. *Journal of Computer and System Sciences, 69*(3), 448–484.

Han, A., Mishra, B., Jawanpuria, P., & Gao, J. (2021a). Generalized Bures-Wasserstein geometry for positive definite matrices. arXiv:2110.10464.

Han, Andi, Mishra, Bamdev, Jawanpuria, Pratik Kumar, & Gao, Junbin. (2021b). On Riemannian optimization over positive definite matrices with the Bures-Wasserstein geometry. *Advances in Neural Information Processing Systems, 34,* 8940–8953.

Harandi, MT., Salzmann, M. & Hartley, R. (2014). From manifold to manifold: Geometry-aware dimensionality reduction for SPD matrices. In European Conference on Computer Vision, pages 17–32. Springer.

Horev, I., Yger, F., & Sugiyama, M. (2016). Geometry-aware principal component analysis for symmetric positive definite matrices. In Asian Conference on Machine Learning, pages 1–16. PMLR.

Huang, Z., Wang, R., Shan, S., Li, X., & Chen, X. (2015). Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification. In International Conference on Machine Learning, pages 720–729. PMLR.

Jiang, Xianhua, Ning, Lipeng, & Georgiou, Tryphon T. (2012). Distances and Riemannian metrics for multivariate spectral densities. *IEEE Transactions on Automatic Control, 57*(7), 1723–1735.

Kim, M., Kumar, S., Pavlovic, V., & Rowley, H. (2008). Face tracking and recognition with visual constraints in real-world videos. In Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE.

Knight, Philip A. (2008). The Sinkhorn-Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications, 30*(1), 261–275.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Le Bihan, Denis, Mangin, Jean-François., Poupon, Cyril, Clark, Chris A., Pappata, Sabina, Molko, Nicolas, & Chabriat, Hughes. (2001). Diffusion tensor imaging: concepts and applications. *Journal of Magnetic Resonance Imaging, 13*(4), 534–546.

LeCun, Yann, Bottou, Léon., Bengio, Yoshua, & Haffner, Patrick. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.

Liero, Matthias, Mielke, Alexander, & Savaré, Giuseppe. (2018). Optimal entropy-transport problems and a new hellinger-kantorovich distance between positive measures. *Inventiones mathematicae, 211*(3), 969–1117.

Malagò, Luigi, Montrucchio, Luigi, & Pistone, Giovanni. (2018). Wasserstein Riemannian geometry of Gaussian densities. *Information Geometry, 1*(2), 137–179.

Maretic, Hermina Petric, El Gheche, Mireille, Chierchia, Giovanni, & Frossard, Pascal. (2019). GOT: An optimal transport framework for graph comparison. *Advances in Neural Information Processing Systems, 32,* 13876–13887.

McCann, Robert J. (1997). A convexity principle for interacting gases. *Advances in Mathematics, 128*(1), 153–179.

Mémoli, Facundo. (2011). Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics, 11*(4), 417–487.

Mishra, B., Kasai, H., & Jawanpuria, P. (2019). Riemannian optimization on the simplex of positive definite matrices. arXiv:1906.10436.

Mishra, B., Satyadev, N.T.V., Kasai, H., & Jawanpuria, P. (2021). Manifold optimization for non-linear optimal transport problems. arXiv:2103.00902.

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

Ning, L. (2013). Matrix-valued optimal mass transportation and its applications. PhD thesis, University of Minnesota.

Ning, Lipeng, Georgiou, Tryphon T., & Tannenbaum, Allen. (2014). On matrix-valued Monge-Kantorovich optimal mass transport. *IEEE Transactions on Automatic Control, 60*(2), 373–382.

Peyré, G., Cuturi, M, & Solomon, J. (2016). Gromov–Wasserstein averaging of kernel and distance matrices. In International Conference on Machine Learning, pages 2664–2672. PMLR.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning,11*(5–6), 355–607.

Peyré, Gabriel, Chizat, Lenaic, Vialard, François-Xavier., & Solomon, Justin. (2019a). Quantum entropic regularization of matrix-valued optimal transport. *European Journal of Applied Mathematics, 30*(6), 1079–1102.

Ryu, Ernest K., Chen, Yongxin, Li, Wuchen, & Osher, Stanley. (2018). Vector and matrix optimal mass transport: Theory, algorithm, and applications. *SIAM Journal on Scientific Computing, 40*(5), A3675–A3698.

Shi, Dai, Gao, Junbin, Hong, Xia, Boris Choy, S. T., & Wang, Zhiyong. (2021). Coupling matrix manifolds assisted optimization for optimal transport problems. *Machine Learning, 110*(3), 533–558.

Sinkhorn, Richard. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics, 35*(2), 876–879.

Sinkhorn, Richard. (1967). Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly, 74*(4), 402–405.

Solomon, Justin, De Goes, Fernando, Peyré, Gabriel, Cuturi, Marco, Butscher, Adrian, Nguyen, Andy, et al. (2015). Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG), 34*(4), 1–11.

Solomon, Justin, Rustamov, Raif, Guibas, Leonidas, & Butscher, Adrian. (2014). Earth mover's distances on discrete surfaces. *ACM Transactions on Graphics (TOG), 33*(4), 1–12.

Sra, Suvrit, & Hosseini, Reshad. (2015). Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization, 25*(1), 713–739.

Sun, Y., Gao, J., Hong, X., Mishra, B., & Yin, B. (2015). Heterogeneous tensor decomposition for clustering via manifold optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(3), 476–489.

Tuzel, O., Porikli, F., & Meer, P. (2006). Region covariance: A fast descriptor for detection and classification. In European Conference on Computer Cision, pages 589–600. Springer.

Villani, Cédric. (2021). *Topics in optimal transportation* (Vol. 58). American Mathematical Soc.

Weickert, J. (1998). *Anisotropic diffusion in image processing* (Vol. 1). Teubner Stuttgart.

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747.

Yair, Or., Ben-Chen, Mirela, & Talmon, Ronen. (2019). Parallel transport on the cone manifold of spd matrices for domain adaptation. *IEEE Transactions on Signal Processing, 67*(7), 1797–1811.

Yurochkin, Mikhail, Claici, Sebastian, Chien, Edward, Mirzazadeh, Farzaneh, & Solomon, Justin M. (2019). Hierarchical optimal transport for document representation. *Advances in Neural Information Processing Systems, 32,* 1601–1611.