Transfer Learning on Riemannian Manifolds

Tran Tien Tam, Ines Adouani, C. Samir

▶ **To cite this version:**

Tran Tien Tam, Ines Adouani, C. Samir. Transfer Learning on Riemannian Manifolds. University of Clermont Auvergne. 2023. hal-04281544

HAL Id: hal-04281544
https://hal.science/hal-04281544

Submitted on 13 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Transfer Learning on the Riemannian Manifold of Probability Measures

T.T. Tran, I. Adouani, C. Samir

[a] *University of Clermont Auvergne, France, France*

Abstract

Keywords: Transfer learning, Probability measure, Information geometry, Fisher-Rao metric, Levi-civita parallel transport.

1. Introduction

Machine learning methods have achieved significant success in solving problems across various domains, including computer vision, medical analysis, image and speech recognition. Despite these achievements, the limited availability of sufficiently large training datasets remains a bottleneck, restricting the advancement of machine learning methods. Most machine learning algorithms typically assume that both training and testing data originate from the same feature space and share the same distribution. Any disparities in data distribution or feature spaces can significantly degrade the model's performance. Moreover, collecting a substantial number of new samples for retraining a new model can be both challenging and expensive. Therefore, there is a growing need to find ways to reuse existing learning models. To address this issue, recent research has introduced the concept of Transfer Learning [1, 2].

In this work, we introduce a novel geometric transfer learning approach of learning models on the space of probability measures, denoted as $\mathcal{P}_+$. The statistical analysis of probability measures is gaining increasing importance in both applications and theory. For many applications in signal processing, text mining, data analysis, and machine learning, the natural way to model objects is as a probability distribution. Our objective is to systematically explore the geometry of $\mathcal{P}_+$ and develop a powerful transfer learning algorithm that enhances the performance of statistical models on target data.

2. Geometry of the manifold of probability measures

In this section, we introduce the problem formulation and the Riemannian geometric structures of the space of probability measures $\mathcal{P}_+$ equipped with a Riemannian metric.

2.1. Problem Formulation

In this paper, we address the problem of transfer learning on the Riemannian manifold of probability measures. Specifically, we are given two datasets: $P_{N_1} = \{\mu_i\}_{i=1}^{N_1}$ and $P_{N_2} = \{\mu_i\}_{i=1}^{N_2}$ from two distinct domains. These datasets consist of N_1 and N_2 probability measures, with $N_2 \ll N_1$. Our objective is to transfer a model that has been developed for a learning

task in the source domain, P_{N_1}, which could include PCA, linear regression, and logistic regression models, to construct an improved model for the target data, $P_{N_2} = \{\mu_i\}_{i=1}^{N_2}$. To achieve this, we establish the Riemannian structure of the probability simplex embedded with the Fisher-Rao metric and introduce a method for transfer learning using Levi-Civita parallel transport.

2.2. Riemannian calculus on $\mathcal{P}_+$

In this section, we develop Riemannian calculus on $\mathcal{P}_+$ with the Fisher-Rao metric, deriving various geometric concepts, such as geodesics, exponential maps, logarithm maps, and the Levi-Civita parallel transport.

2.2.1. Manifold structure

Let $I = \{1, \ldots, n, n+1\}, n \in \mathbb{N}$, be a finite sample space. Let $\mathcal{F}(I) = \{f : I \to \mathbb{R}\}$ be the algebra of real functions on I. Its unity function $\mathbb{1}_I$ or simply $\mathbb{1}$ is given by $\mathbb{1}(i) = 1$, for $i = 1, \ldots, n, n+1$. A canonical basis of $\mathcal{F}(I)$ is defined by

$$e_i(j) = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

and hence, every $f \in \mathcal{F}(I)$ has the representation

$$f = \sum_{i \in I} f^i e_i, \tag{2}$$

where $f^i = f(i)$. We will denote by $\mathcal{S}(I)$ the dual space of $\mathcal{F}(I)$, the space of $\mathbb{R}$-valued linear forms on $\mathcal{F}(I)$. With the Riesz representation theorem, this vector space is interpreted as the vector space of signed measures on I, namely

$$\mathcal{S}(I) = \{\mu : \mathcal{F}(I) \to \mathbb{R} \mid \ \mu = \sum_{i \in I} \mu_i \delta^i\}, \tag{3}$$

where $\mu_i = \mu(e_i)$ and δ^i is considered as the Dirac measure supported at $i \in I$. It is also shown that $\mathcal{S}(I)$ is a smooth manifold. Besides we have a vector space isomorphism between the space $\mathcal{F}(I)$ and $\mathcal{S}(I)$, given by

$$\begin{aligned} \mathcal{F}(I) &\longrightarrow \mathcal{S}(I) \\ f &\longmapsto f\mu := \sum_{i \in I} f^i \mu_i \delta^i. \end{aligned} \tag{4}$$

The inverse is the Radon-Nikodym derivative with respect to μ, denoted as ϕ_μ,

$$\begin{aligned} \phi_\mu : \mathcal{S}(I) &\longrightarrow \mathcal{F}(I) \\ \nu = \sum_{i \in I} \nu_i \delta^i &\longmapsto \frac{d\nu}{d\mu} := \sum_{i \in I} \frac{\nu_i}{\mu_i} e_i. \end{aligned} \tag{5}$$

In particular, the tangent space at the point $\mu \in \mathcal{S}(I)$ is given by

$$T_\mu \mathcal{S}(I) = \{\mu\} \times \mathcal{S}(I). \tag{6}$$

Let us consider the following submanifolds of $\mathcal{S}(I)$:

$$\mathcal{S}_\epsilon(I) = \left\{ \mu = \sum_{i\in I} \mu_i \delta^i \mid \sum_{i\in I} \mu_i = \epsilon, \quad \epsilon \in \mathbb{R} \right\},$$

and

$$\mathcal{M}_+(I) = \{\mu \in \mathcal{S}(I) \mid \ \mu_i > 0, \quad \forall \quad i \in I\},$$

the space of finite strictly positive measures on I.

Definition 1. A probability measure on a finite sample space I is a map $\mu : I \to \mathbb{R}$ defined for any $A \subset I$ by $\mu(A) = \sum_{i\in A} \mu_i$ and which satisfies:

1. For all $i \in I, \mu_i \geq 0$ and $\mu(\emptyset) = 0$.
2. $\sum_{i\in I} \mu_i = 1$.
3. $\mu(\{i\}) = \mu_i$.

We denote by $\mathcal{P}_+(I)$ the space of strictly positive probability measures on I,

$$\mathcal{P}_+(I) = \left\{ \mu = \sum_{i\in I} \mu_i \delta^i \mid \mu_i > 0, \quad \forall i \in I, \text{ and } \sum_{i\in I} \mu_i = 1 \right\}.$$

We check at once that $\mathcal{P}_+(I) \subset \mathcal{M}_+(I) \subset \mathcal{S}(I)$. Therefore, as an open submanifold of $\mathcal{S}(I)$, $\mathcal{M}_+(I)$ has the same tangent space at the point $\mu \in \mathcal{M}_+(I)$. $\mathcal{P}_+(I)$ is a submanifold of $\mathcal{S}(I)$, and clearly, for $\mu \in \mathcal{P}_+(I)$, we have:

$$\begin{aligned} T_\mu \mathcal{P}_+(I) &= \{\mu\} \times \mathcal{S}_0(I) \\ &= \{(\mu, v) \mid \mu \in \mathcal{P}_+(I) \text{ and } v \in \mathcal{S}_0(I)\}. \end{aligned}$$

We want to endow $\mathcal{P}_+(I)$ with a Riemannian metric. To this end, we define a local coordinate map on $\mathcal{P}_+(I)$. Let U be an open set of $\mathbb{R}^n$ given by

$$U = \left\{ x = (x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_i > 0, \forall i \in I, \text{ and } \sum_{i=1}^n x_i < 1 \right\}.$$

We define a map φ as

$$\begin{aligned} \varphi : \mathcal{P}_+(I) &\longrightarrow U, \\ \mu = \sum_{i\in I} \mu_i \delta^i &\longmapsto (\varphi^1(\mu), ..., \varphi^n(\mu)) = (x^1(\mu), ..., x^n(\mu)), \end{aligned}$$

such that $(\varphi^1(\mu), ..., \varphi^n(\mu)) = (\mu_1, \ldots, \mu_n)$. Clearly, φ is an homomorphism and its inverse is given by

$$\begin{aligned} \varphi^{-1} : U &\longrightarrow \mathcal{P}_+(I), \\ (x_1, \ldots, x_n) &\longmapsto \mu = \sum_{i=1}^n x_i \delta^i + \left(1 - \sum_{i=1}^n x_i\right) \delta^{n+1}. \end{aligned}$$

Given a point $\mu \in \mathcal{P}_+(I)$, let $\frac{\partial}{\partial x^i}\Big|_\mu$ be the tangent vector at μ given by

$$\frac{\partial}{\partial x^i}\Big|_\mu = \frac{\partial}{\partial x_i}\Big|_{\varphi(\mu)} \varphi^{-1} = \left(\delta^i - \delta^{n+1}\right), \quad \text{for } i = 1, \ldots, n.$$

Thus, $\left\{ \frac{\partial}{\partial x^i}\Big|_\mu, i = 1, ..., n \right\}$ define a local frame field of $T_\mu \mathcal{P}_+(I)$ at a point $\mu \in \mathcal{P}_+(I)$. Similarly we can define the dual basis of $\frac{\partial}{\partial x^i}\Big|_\mu$, the basis of the cotangent bundle $T^*_\mu \mathcal{P}_+(I) = \{\mu\} \times (\mathcal{F}(I)/\mathbb{R})$ by $dx^i = e_i + \mathbb{R}, \quad i = 1, \ldots, n$.

Remark 1. Let $\mu \in \mathcal{P}_+(I)$ and $v = \sum_{i \in I} v_i \delta^i \in T_\mu \mathcal{P}_+(I)$. It can be easily seen that, for $v \in \mathcal{S}_0(I)$

$$\begin{aligned} v &= \sum_{i=1}^{n+1} v_i \delta^i = \sum_{i=1}^{n} v_i \delta^i - \sum_{i=1}^{n} v_i \delta^{n+1} \\ &= \sum_{i=1}^{n} v_i (\delta^i - \delta^{n+1}) = \sum_{i=1}^{n} v_i \frac{\partial}{\partial x^i}. \end{aligned}$$

$S(I)$ is a finite-dimensional linear space, and therefore, it can be naturally equipped with a metric. For $v, w \in T_\mu S(I)$, we define the inner product as

$$< v, w >_\mu = \mu \left(\frac{dv}{d\mu} . \frac{dw}{d\mu} \right) = \sum_i \frac{v_i w_i}{\mu_i}, \tag{7}$$

where $\frac{dv}{d\mu} = \sum_{i \in I} \frac{v_i}{\mu_i} e_i \in \mathcal{F}(I)$, represents a simple version of the Radon–Nikodym derivative with respect to μ. This metric induces a metric on $\mathcal{M}_+(I)$. The probability manifold $\mathcal{P}_+(I)$ as a submanifold of $\mathcal{M}_+(I)$, is endowed with the Fisher-Rao metric. Hence, following the geometry structures in $\mathcal{M}_+(I)$ equipped with Fisher information metric, we derive the corresponding ones in $\mathcal{P}_+(I)$.

Definition 2. Let μ be a probability measure in $\mathcal{P}_+(I)$. Given two tangents vectors v and w in $T_\mu \mathcal{P}_+(I)$, the Fisher-Rao metric $\mathfrak{g}_\mu : T_\mu \mathcal{P}_+(I) \times T_\mu \mathcal{P}_+(I) \to \mathbb{R}$ is defined by

$$\mathfrak{g}_\mu(v, w) = \sum_{i \in I} \frac{v_i w_i}{\mu_i}, \tag{8}$$

and $||v||_\mu = \sqrt{\mathfrak{g}_\mu(v, v)}$. With respect to the coordinate map $(\mathcal{P}_+(I), \varphi)$, the Fisher-Rao metric is expressed as

$$g_{ij}(\mu) = \begin{cases} \frac{1}{\mu_i} + \frac{1}{\mu_{i+1}}, & \text{if } i = j, \\ \frac{1}{\mu_{n+1}}, & \text{otherwise,} \end{cases}$$

for $i, j = 1, \cdots, n$. And the components of the inverse matrix are given by

$$g^{ij}(\mu) = \begin{cases} \mu_i(1 - \mu_i), & \text{if } i = j, \\ -\mu_i \mu_j, & \text{otherwise.} \end{cases}$$

Our goal to make $\mathcal{P}_+(I)$ as a Riemannian manifold is fully satisfied. Our next goal is to compute explicit expressions of geometric structures on $\mathcal{P}_+(I)$, especially, the Levi-civita parallel transport which will be essential to make our transfer learning approach of statistical models on $\mathcal{P}_+(I)$.

2.2.2. Fisher-Rao metric on $\mathcal{P}_+$

Let $\mathcal{X}(\mathcal{P}_+(I))$ denote the set of smooth vector fields on $\mathcal{P}_+(I)$. Essentially, at each point $\mu \in \mathcal{P}_+(I)$, the Levi-Civita connection associated with the Fisher-Rao metric $\nabla : \mathcal{X}(\mathcal{P}_+(I)) \times \mathcal{X}(\mathcal{P}_+(I)) \to \mathcal{X}(\mathcal{P}_+(I))$ gives a new vector field, notated $\nabla_X Y$, telling us how the vector field Y is changing in the direction X and satisfying for all $X, Y, Z \in \mathcal{X}(\mathcal{P}_+(I))$,

$$\begin{cases} X\mathfrak{g}(Y,Z) = \mathfrak{g}(\nabla_X Y, Z) + \mathfrak{g}(Y, \nabla_X Z), \\ \nabla_X Y - \nabla_Y X = [X,Y]. \end{cases} \tag{9}$$

In the local coordinate map $(\mathcal{P}_+(I), \varphi)$, the Levi-Civita connection is defined by the Christoffel symbols $\Gamma_{ij}^k : \mathcal{P}_+(I) \to \mathbb{R}$ such that $\nabla_{\partial x_i} \partial x_j = \Gamma_{ij}^k \partial x_k$.

Proposition 1. With respect to the local coordinate map $(\mathcal{P}_+(I), \varphi)$, the Christoffel symbols associated with the Fisher-Rao metric are given by

$$\Gamma_{ij}^k = \begin{cases} \frac{1}{2} \frac{x_k}{1 - \sum_{h=1}^n x_h}, & i \neq j, \\ \frac{1}{2} \frac{x_k}{1 - \sum_{h=1}^n x_h} + \frac{1}{2} \frac{x_k}{x_i}, & i = j \neq k, \\ \frac{1}{2} \frac{x_k}{1 - \sum_{h=1}^n x_h} - \frac{1}{2} \frac{1 - x_k}{x_k}, & i = j = k, \end{cases} \tag{10}$$

Proof. The smooth functions Γ_{ij}^k are easily computed through the characterization of the Levi-Civita connection by the Koszul formula obtained from (9) computed for all the circular permutations of $X, Y, Z \in \mathcal{X}(\mathcal{P}_+(I))$,

$$\begin{aligned} \mathfrak{g}(\nabla_X Y, Z) =& \frac{1}{2} \{ X\mathfrak{g}(Y,Z) + Y\mathfrak{g}(Z,X) - Z\mathfrak{g}(X,Y) \\ &+ \mathfrak{g}([X,Y],Z) - \mathfrak{g}([Y,Z],X) - \mathfrak{g}([X,Z],Y) \}. \end{aligned} \tag{11}$$

Now, in the Koszul formula we set $X = \partial x_i, Y = \partial x_j$ and $Z = \partial x_l$. We get

$$\Gamma_{ij}^k = \frac{1}{2} \sum_{l=1}^n g^{kl} \left(g_{il,j} + g_{jl,i} - g_{ij,l} \right), \tag{12}$$

for $i, j, k \in \{1, ..., n\}$, where $g_{il,j} = \frac{\partial g_{il}}{\partial x_j}$, $g_{jl,i} = \frac{\partial g_{jl}}{\partial x_i}$, and $g_{ij,l} = \frac{\partial g_{ij}}{\partial x_l}$. In the local coordinate system, the Fisher-Rao metric and its inverse are given by

$$g_{ij} = \begin{cases} \frac{1}{x_i} + \frac{1}{1 - \sum_{h=1}^n x_h}, & \text{if } i = j, \\ \frac{1}{1 - \sum_{h=1}^n x_h}, & \text{if } i \neq j, \end{cases} \tag{13}$$

$$g^{ij} = \begin{cases} x_i(1 - x_i), & \text{if } i = j, \\ -x_i x_j, & \text{if } i \neq j, \end{cases} \tag{14}$$

for $i, j = 1, \ldots, n$. Now if we take the derivative of (13) by x_l, we get

$$g_{ij,l} = \begin{cases} -\dfrac{1}{(x_i)^2} + \dfrac{1}{(1-\sum_{h=1}^n x_h)^2}, & \text{if } i = j = l, \\ \dfrac{1}{(1-\sum_{h=1}^n x_h)^2}, & \text{otherwise.} \end{cases} \tag{15}$$

Replace (15) in (12), the formula follows. □

Definition 3. Let $X \in \mathcal{X}(\mathcal{P}_+(I))$ be a vector field on $\mathcal{P}_+(I)$. Then in the local coordinate $(\mathcal{P}_+(I), \varphi)$, we have the representation $X = \sum_{i=1}^n X_i \partial x_i$. X is called a constant vector field on $\mathcal{P}_+(I)$ if all X_i are independent of μ.

Theorem 1. *Given two constant vector fields X, Y on $\mathcal{P}_+(I)$, the Levi-Civita connection at $\mu \in \mathcal{P}_+(I)$ is given by*

$$\nabla_X Y(\mu) = -\frac{1}{2}\left(\frac{dX}{d\mu}\frac{dY}{d\mu} - \mathfrak{g}_\mu(X, Y)\right)\mu. \tag{16}$$

Proof. Let $X = \sum_{i\in I} X_i\delta^i, Y = \sum_{i\in I} Y_i\delta^i$ and $Z = \sum_{i\in I} Z_i\delta^i$ be constant vector fields on $\mathcal{P}_+(I)$. Thus, we get $[X, Y] = [Y, Z] = [X, Z] = 0$ and consequently (11) gives

$$\mathfrak{g}(\nabla_X Y, Z) = \frac{1}{2}\left\{X\mathfrak{g}(Y, Z) + Y.\mathfrak{g}(X, Z) - Z.\mathfrak{g}(X, Y)\right\}. \tag{17}$$

Set $\mu = \sum_{i\in I} \mu_i\delta^i \in \mathcal{P}_+(I)$, and $\alpha(t) = \mu + vt$, a curve on $\mathcal{P}_+(I)$ such that $\mu(0) = \mu$ and $\dot{\mu}(0) = v = X(\mu)$. We have

$$\begin{aligned} X\mathfrak{g}_\mu(Y, Z) &= \frac{d}{dt}\Big|_{t=0} \mathfrak{g}_{\mu(t)}(Y, Z) \\ &= \frac{d}{dt}\Big|_{t=0} \sum_{i\in I} \frac{Y_i Z_i}{\mu_i + tv_i} \\ &= -\sum_{i\in I} \frac{v_i Y_i Z_i}{\mu_i^2} = -\sum_{i\in I} \frac{X_i Y_i Z_i}{\mu_i^2}. \end{aligned}$$

Similarly, one obtains formulae for $Y\mathfrak{g}(X, Z)$ and $Z\mathfrak{g}(X, Y)$. Now replacing the above results in (17), we get

$$\begin{aligned} &\mathfrak{g}_\mu(\nabla_X Y, Z) \\ &= \frac{1}{2}\left\{-\sum_{i\in I} \frac{X_i Y_i Z_i}{\mu_i^2} - \sum_{i\in I} \frac{X_i Y_i Z_i}{\mu_i^2} + \sum_{i\in I} \frac{X_i Y_i Z_i}{\mu_i^2}\right\} \\ &= -\frac{1}{2}\sum_{i\in I} \frac{X_i Y_i Z_i}{\mu_i^2}. \end{aligned} \tag{18}$$

On the other hand, we have

$$\sum_{i\in I} \mathfrak{g}_\mu(X, Y)Z_i = \mathfrak{g}_\mu(X, Y)\sum_{i\in I} Z_i = 0, \tag{19}$$

since Z is a constant vector field on $\mathcal{P}_+(I)$. Then (18) can be written as

$$\begin{aligned}
&\mathfrak{g}_\mu(\nabla_X Y, Z) \\
&= -\frac{1}{2}\sum_{i\in I}\left(\frac{X_i Y_i}{\mu_i^2} - \mathfrak{g}_\mu(X,Y)\right)\mu_i\frac{Z_i}{\mu_i} \\
&= \mathfrak{g}_\mu\left(-\frac{1}{2}\left(\frac{dX}{d\mu}\frac{dY}{d\mu} - \mathfrak{g}_\mu(X,Y)\right)\mu, Z\right).
\end{aligned}$$

which completes the proof. □

2.2.3. Geodesics on $\mathcal{P}_+$

Theorem 2. *Let $\mu = \sum_{i\in I}\mu_i\delta^i$ be a probability measure in $\mathcal{P}_+(I)$ and $v \in T_\mu\mathcal{P}_+(I)$ a unit tangent vector, i.e., $||v||_\mu = 1$. Then the geodesic α that satisfies $\alpha(0) = \mu$ and $\dot\alpha(0) = v$ is given by $\alpha(t) = \sum_{i\in I}\alpha_i(t)\delta^i$ with*

$$\alpha_i(t) = \left(\cos\frac{t}{2} + \frac{\dot\alpha_i(0)}{\alpha_i(0)}\sin\frac{t}{2}\right)^2\alpha_i(0), \tag{20}$$

where $\alpha_i(0) = \mu_i$ and $\dot\alpha_i(0) = v_i$, $\forall i \in I$.

Proof. Let $\alpha(t) = \sum_{i\in I}\alpha_i(t)\delta^i$ and $\dot\alpha(t) = \sum_{i\in I}\dot\alpha_i(t)\delta^i$. Then for each t, we have

$$\begin{cases} \sum_{i\in I}\alpha_i(t) &= 1, \quad \text{and } \alpha_i(t) > 0, \forall i \in I, \\ \sum_{i\in I}\dot\alpha_i(t) &= 0. \end{cases} \tag{21}$$

Set X a constant vector field in $\mathcal{P}_+(I)$. From the condition (9) of Levi–Civita connection, we have

$$\mathfrak{g}_{\alpha(t)}(\nabla_{\dot\alpha(t)}\dot\alpha(t), X) = \dot\alpha(t)\left(\mathfrak{g}_{\alpha(t)}(\dot\alpha(t), X)\right) - \mathfrak{g}_{\alpha(t)}(\dot\alpha(t), \nabla_{\dot\alpha(t)}X). \tag{22}$$

With the properties of Levi-Civita connection, to compute $\nabla_{\dot\alpha(t)}X$, the tangent vector $\dot\alpha(t)$ can be considered as a constant vector field on $\mathcal{P}_+(I)$ when t is fixed. Therefore, applying (16) for $\dot\alpha(t)$ and X we get,

$$\begin{aligned}
\nabla_{\dot\alpha(t)}X &= -\frac{1}{2}\left(\frac{d\dot\alpha(t)}{d\alpha(t)}\frac{dX}{d\alpha(t)} - \mathfrak{g}_{\alpha(t)}(\dot\alpha(t), X)\right)\alpha(t) \\
&= -\frac{1}{2}\sum_{i\in I}\left(\frac{\dot\alpha_i}{\alpha_i}\frac{X_i}{\alpha_i} - \sum_{j\in I}\frac{\dot\alpha_j X_j}{\alpha_j}\right)\alpha_i\delta^i.
\end{aligned} \tag{23}$$

Taking into account of (21), the last term in (22) becomes

$$\begin{aligned}
&\mathfrak{g}(\dot\alpha(t), \nabla_{\dot\alpha(t)}X) \\
&= \left\langle\frac{d\dot\alpha}{d\alpha}, \frac{d\nabla_{\dot\alpha(t)}X}{d\alpha}\right\rangle_{\alpha(t)} \\
&= -\frac{1}{2}\sum_{i\in I}\frac{\dot\alpha_i}{\alpha_i}\left(\frac{\dot\alpha_i}{\alpha_i}\frac{X_i}{\alpha_i} - \sum_{j\in I}\frac{\dot\alpha_j X_j}{\alpha_j}\right)\alpha_i \\
&= -\frac{1}{2}\sum_{i\in I}\frac{\dot\alpha_i^2 X_i}{\alpha_i^2}.
\end{aligned} \tag{24}$$

Now, we compute the second term in (22). We have

$$\dot{\alpha}(t)\left(\mathfrak{g}_{\alpha(t)}(\dot{\alpha}(t),X)\right)=\frac{d}{dt}\mathfrak{g}_{\alpha(t)}(\dot{\alpha}(t),X)=\sum_{i\in I}\frac{d}{dt}\left(\frac{\dot{\alpha}_i}{\alpha_i}\right)X_i. \tag{25}$$

Combining (24) and (25) in (22), we get

$$\mathfrak{g}_{\alpha(t)}(\nabla_{\dot{\alpha}(t)}\dot{\alpha}(t),X)=\sum_{i\in I}\left(\frac{d}{dt}\left(\frac{\dot{\alpha}_i}{\alpha_i}\right)+\frac{1}{2}\frac{\dot{\alpha}_i^2}{\alpha_i^2}\right)X_i. \tag{26}$$

Let's define the function $F(t)$ as

$$\begin{aligned}F(t)&=-\sum_{i\in I}\left(\frac{d}{dt}\left(\frac{\dot{\alpha}_i}{\alpha_i}\right)+\frac{1}{2}\frac{\dot{\alpha}_i^2}{\alpha_i^2}\right)\alpha_i(t)\\&=-\sum_{i\in I}\frac{d}{dt}\left(\frac{\dot{\alpha}_i}{\alpha_i}\right)\alpha_i(t)-\frac{1}{2}\mathfrak{g}_{\alpha(t)}(\dot{\alpha}(t),\dot{\alpha}(t)).\end{aligned} \tag{27}$$

Hence, the measure

$$\nu(t)=\sum_{i\in I}\left(\frac{d}{dt}\left(\frac{\dot{\alpha}_i}{\alpha_i}\right)+\frac{1}{2}\frac{\dot{\alpha}_i^2}{\alpha_i^2}+F(t)\right)\alpha_i\delta^i \tag{28}$$

belongs to $T_{\alpha(t)}\mathcal{P}_+$. In this way, (26) can be written as $\mathfrak{g}_\alpha(\nabla_{\dot{\alpha}}\dot{\alpha},X)=\mathfrak{g}_\alpha(\nu,X)$. Since X is an arbitrary constant vector field, we get

$$\nabla_{\dot{\alpha}}\dot{\alpha}=\nu=\sum_{i\in I}\left(\frac{d}{dt}\left(\frac{\dot{\alpha}_i}{\alpha_i}\right)+\frac{1}{2}\frac{\dot{\alpha}_i^2}{\alpha_i^2}+F(t)\right)\alpha_i\delta^i. \tag{29}$$

Therefore, $\alpha(t)=\sum_{i\in I}\alpha_i(t)\delta^i$ is a geodesic if and only if

$$\begin{cases}\dfrac{d}{dt}\left(\frac{\dot{\alpha}_i}{\alpha_i}\right)+\dfrac{1}{2}\left(\dfrac{\dot{\alpha}_i}{\alpha_i}\right)^2+F(t)=0, \quad \forall i\in I,\\ \sum_{i\in I}\dot{\alpha}_i(t)=0, \quad \forall t.\end{cases} \tag{30}$$

Our next goal is to solve (30). We may remark that if α is a geodesic then $\mathfrak{g}_{\alpha(t)}(\dot{\alpha}(t),\dot{\alpha}(t))$ is constant along $\alpha(t)$. Consequently, taking into account of the assumption that $||\dot{\gamma}(0)||_\mu=1$, we can assert that

$$\mathfrak{g}_{\alpha(t)}(\dot{\alpha}(t),\dot{\alpha}(t))=\sum_{i\in I}\frac{\dot{\alpha}_i^2}{\alpha_i}\equiv 1. \tag{31}$$

Thus

$$\sum_{i\in I}\frac{d}{dt}\left(\frac{\dot{\alpha}_i}{\alpha_i}\right)\alpha_i=\frac{d}{dt}\sum_{i\in I}\left(\frac{\dot{\alpha}_i}{\alpha_i}\alpha_i\right)-\sum_{i\in I}\frac{\dot{\alpha}_i^2}{\alpha_i}=-1. \tag{32}$$

Which gives that $F(t) = \frac{1}{2}$. Substituting this result in (30), we obtain

$$\frac{d}{dt}\left(\frac{\dot{\alpha}_i}{\alpha_i}\right) + \frac{1}{2}\left(\frac{\dot{\alpha}_i}{\alpha_i}\right)^2 + \frac{1}{2} = 0, \quad \forall i \in I. \tag{33}$$

Set $\omega_i(t) = \dfrac{\dot{\alpha}_i(t)}{\alpha_i(t)}$. Equation (33) is written as

$$\frac{d}{dt}\omega_i + \frac{1}{2}\omega_i^2 + \frac{1}{2} = 0, \quad \forall i \in I,$$

The solution of this differential equation is given by $\omega_i = \tan\left(-\frac{t}{2} + \Theta^i\right)$, where Θ^i is constant, $i \in I$. Hence, we have

$$\frac{\dot{\alpha}_i}{\alpha_i} = \tan\left(-\frac{1}{2}t + \Theta^i\right), \forall i \in I$$

and $\alpha_i(t) = \Omega^i \cos^2\left(-\frac{t}{2} + \Theta^i\right)$, where Ω^i is constant, and $i \in I$. Taking into account initial conditions, we find that

$$\Theta^i = \arctan\left(\frac{\dot{\alpha}_i(0)}{\alpha_i(0)}\right), \tag{34}$$

$$\Omega^i = \frac{\alpha_i^2(0) + \dot{\alpha}_i^2(0)}{\alpha_i(0)}. \tag{35}$$

which proves the theorem. □

Corollary 1. The geodesic $\alpha(t)$ with $\alpha(0) = \mu$ and $\dot{\alpha}(0) = v$, where v is a nontrivial tangent vector (not necessary unit), is given by

$$\alpha(t) = \sum_{i \in I}\left(\cos\frac{t\|v\|_\mu}{2} + \frac{v_i}{\mu_i\|v\|_\mu}\sin\frac{t\|v\|_\mu}{2}\right)^2 \mu_i\delta^i. \tag{36}$$

Proposition 2. The Fisher Rao distance $d^{FR} : \mathcal{P}_+(I) \times \mathcal{P}_+(I) \to [0, \pi)$ between two measures $\mu, \nu \in \mathcal{P}_+(I)$ under the Fisher-Rao metric is given by

$$d^{FR}(\mu, \nu) = 2\arccos\left(\sum_{i \in I}\sqrt{\mu_i\nu_i}\right). \tag{37}$$

To prove proposition (2), we will show the following lemma given in [39].

Lemma 3. *Let*

$$\mathbb{S}^+_{(0,2)}(I) = \left\{ f \in \mathcal{F}(I) \mid f^i > 0, \forall i \in I \text{ and } \sum_{i \in I}(f^i)^2 = 4 \right\}$$

be the positive sector of the sphere centered at 0 *with radius* 2*. As a submanifold of* $\mathcal{F}(I)$ *it carries the induced standard metric of* $\mathcal{F}(I)$*. That is for a given point* $f \in \mathbb{S}^+_{(0,2)}(I)$ *and two tangents vectors* $p, q \in T_f\mathbb{S}^+_{(0,2)}(I)$*, we have*

$$\langle p, q\rangle_f = \sum_{i\in I} p^i q^i. \tag{38}$$

Then the map $\Phi : \mathcal{P}_+(I) \longrightarrow \mathbb{S}^+_{(0,2)}(I)$ *defined by*

$$\mu = \sum_{i\in I} \mu_i \delta^i \longmapsto 2\sum_{i\in I} \sqrt{\mu_i} e_i$$

is an isometry.

Proof of the lemma. It is clear that Φ is bijective. Now, let v, w be in $T_\mu\mathcal{P}_+(I)$. We have

$$\begin{aligned}
&\left\langle \frac{\partial\Phi}{\partial v}(\mu), \frac{\partial\Phi}{\partial w}(\mu)\right\rangle \\
&= \left\langle \frac{d}{dt}\Phi(\mu + vt)\Big|_{t=0}, \frac{d}{dt}\Phi(\mu + wt)\Big|_{t=0}\right\rangle \\
&= \left\langle \sum_{i\in I} \frac{v_i}{\sqrt{\mu_i}} e_i, \sum_{i\in I} \frac{w_i}{\sqrt{\mu_i}} e_i\right\rangle \\
&= \sum_{i\in I} \frac{v_i w_i}{\mu_i} = \mathfrak{g}_\mu(v, w).
\end{aligned}$$

□

Proof of the Proposition. By virtue of Lemma 3, we get

$$d^{FR}(\mu, \nu) = d(\Phi(\mu), \Phi(\nu)) = 2\arccos\left(\sum_{i\in I} \sqrt{\mu_i\nu_i}\right).$$

□

Theorem 4. *Let* μ, ν *be two different probability measures in* $\mathcal{P}_+(I)$*. Then there exists a unique geodesic* $\alpha : [0, l] \to \mathcal{P}_+(I)$*,* $t \to \alpha(t)$*, joining two points* μ *and* ν*, with* $\alpha(0) = \mu$*,* $\alpha(l) = \nu$ *and* $l = d^{FR}(\mu, \nu)$*, given by*

$$\alpha(t) = \sum_{i\in I} \left(\cos\frac{t}{2} + \frac{d\tau}{d\mu}(i)\sin\frac{t}{2}\right)^2 \mu_i\delta^i, \tag{39}$$

where τ *is the tangent vector in* $T_\mu\mathcal{P}_+(I)$ *defined by*

$$\tau = \frac{1}{\sin\frac{l}{2}} \sum_{i\in I} \left(\sqrt{\frac{d\nu}{d\mu}} - \sum_{j\in I} \sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)\mu_i\delta^i. \tag{40}$$

Proof. The proof falls naturally into three parts.
Step 1 First, let us check that τ is a tangent vector in $T_\mu\mathcal{P}_+(I)$. Indeed,

$$\begin{aligned}
&\frac{1}{\sin\frac{l}{2}}\sum_{i\in I}\left(\sqrt{\frac{d\nu}{d\mu}}(i)-\sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)\mu_i\\
&=\frac{1}{\sin\frac{l}{2}}\left(\sum_{i\in I}\sqrt{\frac{d\nu}{d\mu}}(i)\mu_i-\sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)\\
&=0. \qquad (41)
\end{aligned}$$

Then, since

$$\left(\sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)^2=\left(\sum_{j\in I}\sqrt{\mu_j\nu_j}\right)^2=\cos^2\frac{l}{2}. \qquad (42)$$

it follows that

$$\begin{aligned}
\langle\tau,\tau\rangle_\mu&=\frac{1}{\sin^2\frac{l}{2}}\sum_{i\in I}\left(\sqrt{\frac{d\nu}{d\mu}}(i)-\sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)^2\mu_i\\
&=\frac{1}{\sin^2\frac{l}{2}}\left(\sum_{i\in I}\nu(i)-\left(\sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)^2\right)\\
&=\frac{1}{\sin^2\frac{l}{2}}\left(1-\cos^2\frac{l}{2}\right)=1. \qquad (43)
\end{aligned}$$

hence τ is a unit tangent vector.
Step 2 Now let us examine that the curve $\alpha(t)$ defined in (39) satisfies $\alpha(0)=\mu$ and $\alpha(1)=\nu$. It is easily seen that for $t=0$, $\alpha(0)=\mu$. Now for $t=l$, we have

$$\alpha(l)=\sum_{i\in I}\left(\cos\frac{l}{2}+\frac{d\tau}{d\mu}(i)\sin\frac{l}{2}\right)^2\mu_i\delta^i, \qquad (44)$$

By (40) we get

$$\begin{aligned}
\frac{d\tau}{d\mu}\sin\frac{l}{2}&=\sum_{i\in I}\left(\sqrt{\frac{d\nu}{d\mu}}(i)-\sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)e_i\\
&=\sum_{i\in I}\left(\sqrt{\frac{d\nu}{d\mu}}(i)-\cos\frac{l}{2}\right)e_i. \qquad (45)
\end{aligned}$$

Hence,

$$\begin{aligned}
\alpha(l)&=\sum_{i\in I}\left(\cos\frac{l}{2}+\sqrt{\frac{d\nu}{d\mu}}(i)-\cos\frac{l}{2}\right)^2\mu_i\delta^i\\
&=\sum_{i\in I}\nu_i\delta^i=\nu. \qquad (46)
\end{aligned}$$

Step 3 Now we go to prove the uniqueness of the curve. Let $\mu(t) = \exp_\mu \tau t$ and $\tilde{\mu}(t) = \exp_\mu \tilde{\tau} t$ be unit speed geodesics corresponding with τ and $\tilde{\tau}$, and satisfying $\mu(0) = \tilde{\mu}(0) = \mu$ and $\mu(l) = \tilde{\mu}(l) = \nu$. By means of Theorem 2, we have

$$\mu(t) = \sum_{i \in I} \left(\cos \frac{t}{2} + \frac{d\tau}{d\mu} \sin \frac{t}{2} \right)^2 \mu_i \delta^i, \tag{47}$$

$$\tilde{\mu}(t) = \sum_{i \in I} \left(\cos \frac{t}{2} + \frac{d\tilde{\tau}}{d\mu} \sin \frac{t}{2} \right)^2 \mu_i \delta^i. \tag{48}$$

From later condition, we have

$$\left(\cos \frac{l}{2} + \frac{d\tau}{d\mu}(i) \sin \frac{l}{2} \right)^2 = \left(\cos \frac{l}{2} + \frac{d\tilde{\tau}}{d\mu}(i) \sin \frac{l}{2} \right)^2, \forall i \in I \tag{49}$$

$$\Rightarrow \cos \frac{l}{2} + \frac{d\tau}{d\mu}(i) \sin \frac{l}{2} = \pm \left(\cos \frac{l}{2} + \frac{d\tilde{\tau}}{d\mu}(i) \sin \frac{l}{2} \right), \forall i \in I. \tag{50}$$

Define

$$\begin{aligned} I_\pm = \Bigg\{ i \in I \Big| & \cos \frac{l}{2} + \frac{d\tau}{d\mu}(i) \sin \frac{l}{2} \\ & = \pm \left(\cos \frac{l}{2} + \frac{d\tilde{\tau}}{d\mu}(i) \sin \frac{l}{2} \right) \Bigg\} \end{aligned} \tag{51}$$

Then we have $I_- \cup I_+ = I$. Moreover $I_- \cap I_+ = \emptyset$. Indeed, if there exists $i \in I_- \cap I_+$ then

$$\nu_i = \left(\cos \frac{t}{2} + \frac{d\tau}{d\mu} \sin \frac{t}{2} \right)^2 \mu_i = 0, \tag{52}$$

contradict to $\nu \in \mathcal{P}_+$. Sine $0 < l < \pi$, we have

$$I_+ = \{ i \in I | \tau_i = \tilde{\tau}_i \}, \tag{53}$$

$$I_- = \left\{ i \in I | \tau_i + \tilde{\tau}_i = -2\mu_i \cot \frac{l}{2} \right\}. \tag{54}$$

Suppose $I_- \neq \emptyset$, since τ and $\tilde{\tau}$ are unit tangent vectors at μ, we have

$$\sum_{i \in I_+} \tau_i + \sum_{i \in I_-} \tau_i = \sum_{i \in I_+} \tilde{\tau}_i + \sum_{i \in I_-} \tilde{\tau}_i = 0 \tag{55}$$

$$\Rightarrow \sum_{i \in I_-} \left(\tilde{\tau}_i + 2\mu_i \cot \frac{l}{2} \right) + \sum_{i \in I_-} \tilde{\tau}_i = 0. \tag{56}$$

Since (56) we see that if $I_- = I$, then $\cot \frac{l}{2} = 0$ contradicts to $0 < l < \pi$. So $I_- \neq I$. We have the claim below.

Claim 1. For all $\mu \in \mathcal{P}_+$ and $0 < l < \pi$. If $\tau, \tilde{\tau} \in T_\mu \mathcal{P}_+$. Let

$$I_+ = \{i \in I | \tau_i = \tilde{\tau}_i\}, \tag{57}$$

$$I_- = \left\{ i \in I | \tau_i + \tilde{\tau}_i = -2\mu_i \cot \frac{l}{2} \right\}. \tag{58}$$

then $I_- = \emptyset$.

By means of the Claim, we prove the uniqueness of the geodesic (39) defined with the unit tangent vector (40). □

The proof of Claim 1 will be given in Appendix 1.

Corollary 2. Let

$$\varepsilon = \{(\mu, v) \, | \alpha(t, \mu, v) \text{ is defined on an interval containing } [0, l] \, \}$$

The exponential map $\exp_\mu : \varepsilon \longrightarrow \mathcal{P}_+(I)$ is defined as

$$\exp_\mu(v) = \sum_{i \in I} \left(\cos \frac{\|v\|_\mu}{2} + \frac{v_i}{\mu_i \|v\|_\mu} \sin \frac{\|v\|_\mu}{2} \right)^2 \mu_i \delta^i. \tag{59}$$

Similarly, given two points μ and ν on $\mathcal{P}_+(I)$, the inverse exponential map (also known as the logarithmic map) at μ, $\log_\mu : \mathcal{P}_+(I) \longrightarrow \varepsilon$ is defined for any $\nu \in \mathcal{P}_+(I)$ by

$$\log_\mu(\nu) = \frac{l}{\sin \frac{l}{2}} \sum_{i \in I} \left(\sqrt{\frac{d\nu}{d\mu}}(i) - \sum_{j \in I} \sqrt{\frac{d\nu}{d\mu}}(j) \mu(j) \right) \mu_i \delta^i. \tag{60}$$

2.2.4. Levi-civita parallel transport on $\mathcal{P}_+$

Let us consider two points $\mu, \nu \in \mathcal{P}_+(I)$, a tangent vector $v \in T_\mu \mathcal{P}_+(I)$ and a geodesic curve $\alpha : [0, l] \to \mathcal{P}_+(I)$ on $\mathcal{P}_+(I)$ such that $\alpha(0) = \mu$ and $\alpha(l) = \nu$. We would like to map v from $T_\mu \mathcal{P}_+(I) = T_{\alpha(0)} \mathcal{P}_+(I)$ to $T_\nu \mathcal{P}_+(I) = T_{\alpha(l)} \mathcal{P}_+(I)$. We introduce X, a vector field defined along the geodesic α, such that $X(\mu) = v$ and $\nabla_{\dot{\alpha}(t)} X(\alpha(t)) = 0$. We say that the tangent vector v is constant along the geodesic curve α with respect to ∇.

Definition 4. A metric parallel transport on $\mathcal{P}_+(I)$ is the map

$$\Gamma_{\alpha(0) \rightarrowtail \alpha(t)} : T_{\alpha(0)} \mathcal{P}_+(I) \to T_{\alpha(t)} \mathcal{P}_+(I) \tag{61}$$

such that for any $v, w \in T_\mu \mathcal{P}_+(I)$, and for $t \in [0, l]$ we have

$$\mathfrak{g}_{\alpha(0)}(v, w) = \mathfrak{g}_{\alpha(t)} \left(\Gamma_{\alpha(0) \to \alpha(t)}(v), \Gamma_{\alpha(0) \to \alpha(t)}(w) \right). \tag{62}$$

Γ is the Levi-Civita parallel transport along the geodesic curve α on $\mathcal{P}_+(I)$ with respect to the Fisher-Rao metric.

Rewriting equation $\nabla_{\dot{\alpha}(t)}X(\alpha(t)) = 0$, we conclude that computing $X(t) = X(\alpha(t))$ requires solving a linear first order differential equations on $\mathcal{P}_+(I)$ given by

$$\frac{dX_k}{dt} + \sum_{i,j} \alpha_{ij}^k \frac{d\alpha_i}{dt} X_j = 0, \quad \text{for } k = 1, \ldots, n. \tag{63}$$

We check at once that it is difficult to solve Eq.(63) directly. Hence we will use Eq.(16).

Theorem 5. *Let μ be a probability measure in $\mathcal{P}_+(I)$ and $v \in T_\mu\mathcal{P}_+(I)$ a unit tangent vector, i.e., $||v||_\mu = 1$. Let $\alpha : [0, l] \to \mathcal{P}_+(I)$ be a geodesic curve such that $\alpha(0) = \mu$ and $\dot{\alpha}(0) = v$. The Levi-civita parallel transport of a vector $w \in T_\mu\mathcal{P}_+(I)$ to $T_{\alpha(t)}\mathcal{P}_+(I)$, is given by*

$$\begin{aligned}\Gamma_{\alpha(0)\rightarrowtail\alpha(t)}(w) = &\sum_{i\in I} \sqrt{\alpha_i(t)} \left(-F(0)\sqrt{\mu_i} \left(2\sin\frac{t}{2} - 2\frac{v_i}{\mu_i}\cos\frac{t}{2} \right) \right. \\ &\left. + \frac{w_i}{\sqrt{\mu_i}} - 2F(0)\frac{v_i}{\sqrt{\mu_i}} \right) \delta^i,\end{aligned} \tag{64}$$

where $F(0) = \frac{1}{2}\mathfrak{g}_\mu(v, w)$.

Proof. We can proceed analogously to the proof of Theorem 2. Thus, let $\alpha(t) = \sum_{i\in I} \alpha_i(t)\delta^i$ be a geodesic curve, and define $\dot{\alpha}(t) = \sum_{i\in I} \dot{\alpha}_i(t)\delta^i$. Consider the vector field X on α defined by $X(\alpha(t)) = \sum_{i\in I} X_i(\alpha(t))\delta^i$, for $t \in [0, l]$, as the parallel transport of vector w along α. Then

$$\begin{cases} \nabla_{\dot{\alpha}(t)}X(t) = 0 \\ X(0) = w \end{cases}, \tag{65}$$

where we write $X(\alpha(t))$ simply $X(t)$ when no confusion can arise. Let Y be a constant vector field (in the sense of Definition 3) on $\mathcal{P}_+(I)$, we have

$$\begin{aligned}&\mathfrak{g}_{\alpha(t)}\left(\nabla_{\dot{\alpha}(t)}X(t), Y\right) \\ &= \dot{\alpha}(t)\left(\mathfrak{g}_{\alpha(t)}(X(t), Y)\right) - \mathfrak{g}_{\alpha(t)}\left(X(t), \nabla_{\dot{\alpha}(t)}Y\right).\end{aligned} \tag{66}$$

Applying Theorem 1, we get

$$\nabla_{\dot{\alpha}}Y = -\frac{1}{2}\sum_{i\in I}\left(\frac{\dot{\alpha}_i}{\alpha_i}\frac{Y_i}{\gamma_i} - \sum_{j\in I}\frac{\dot{\alpha}_j Y_j}{\alpha_j}\right)\alpha_i\delta^i. \tag{67}$$

Hence the last term in (66) becomes

$$\begin{aligned}\mathfrak{g}_\alpha(X, \nabla_{\dot{\alpha}}Y) &= -\frac{1}{2}\sum_{i\in I}\frac{X_i}{\alpha_i}\left(\frac{\dot{\alpha}_i}{\alpha_i}\frac{Y_i}{\alpha_i} - \sum_{j\in I}\frac{\dot{\alpha}_j Y_j}{\alpha_j}\right)\alpha_i \\ &= -\frac{1}{2}\sum_{i\in I}\frac{X_i Y_i \dot{\alpha}_i}{\alpha_i^2}.\end{aligned} \tag{68}$$

Let us now compute the second term in (66). We obtain

$$\begin{aligned}\dot{\alpha}(t)\left(\mathfrak{g}_{\alpha(t)}(X,Y)\right) &= \frac{d}{dt}\mathfrak{g}_{\alpha(t)}(X(t),Y)\\ &= \sum_{i\in I}\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right)Y_i. \end{aligned} \tag{69}$$

Consequently, Equation (66) becomes

$$\mathfrak{g}_{\alpha}(\nabla_{\dot{\alpha}}X,Y) = \sum_{i\in I}\left(\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right)+\frac{1}{2}\frac{X_i\dot{\alpha}_i}{\alpha_i^2}\right)Y_i. \tag{70}$$

Define the function $F(t)$ by

$$\begin{aligned}F(t) &= -\sum_{i\in I}\left(\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right)+\frac{1}{2}\frac{X_i\dot{\alpha}_i}{\alpha_i^2}\right)\alpha_i(t)\\ &= -\sum_{i\in I}\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right)\alpha_i(t) - \frac{1}{2}\mathfrak{g}_{\alpha(t)}(X(t),\dot{\alpha}(t)). \end{aligned} \tag{71}$$

Then, $\forall t\in[0,l]$, the probability measure

$$\nu(t) = \sum_{i\in I}\left(\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right)+\frac{1}{2}\frac{X_i\dot{\alpha}_i}{\alpha_i^2}+F(t)\right)\alpha_i\delta^i$$

belongs to $T_{\alpha(t)}\mathcal{P}_+(I)$. Thus, Equation (70) can be written as

$$\mathfrak{g}_{\alpha}(\nabla_{\dot{\alpha}}X,Y) = \mathfrak{g}_{\alpha}(\nu,Y). \tag{72}$$

Since Y is an arbitrary constant vector field, we get

$$\nabla_{\dot{\alpha}}X = \nu = \sum_{i\in I}\left(\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right)+\frac{1}{2}\frac{X_i\dot{\alpha}_i}{\alpha_i^2}+F(t)\right)\alpha_i\delta^i. \tag{73}$$

Therefore, $X(t)$ is the parallel transport of the vector w along the geodesic curve $\alpha(t)$ if and only if

$$\begin{cases}\dfrac{d}{dt}\left(\dfrac{X_i}{\alpha_i}\right)+\dfrac{1}{2}\dfrac{X_i\dot{\alpha}_i}{\alpha_i^2}+F(t)=0, \quad \forall i\in I,\\ X(0)=w.\end{cases} \tag{74}$$

Our next concern will be to solve Eq.(74). We remind that

$$\mathfrak{g}_{\alpha(t)}(X(t),\dot{\alpha}(t)) = \mathfrak{g}_{\alpha(0)}(X(0),\dot{\alpha}(0)). \tag{75}$$

Moreover

$$\begin{aligned}\sum_{i\in I}\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right)\alpha_i &= \frac{d}{dt}\sum_{i\in I}\left(\frac{X_i}{\alpha_i}\alpha_i\right) - \sum_{i\in I}\left(\frac{X_i\dot{\alpha}_i}{\alpha_i}\right)\\ &= -\mathfrak{g}_{\alpha(0)}(X(0),\dot{\alpha}(0)). \end{aligned} \tag{76}$$

Which gives that $F(t)$ is a constant function and $F(t) = F(0) = \frac{1}{2}\mathfrak{g}_{\alpha(0)}(X(0), \dot{\alpha}(0))$. Hence, substituting this result in Eq.(74) we get

$$\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right) + \frac{1}{2}\frac{X_i\dot{\alpha}_i}{\alpha_i^2} + F(0) = 0, \quad \forall i \in I. \tag{77}$$

Set $\omega_i = \dfrac{X_i}{\alpha_i}$. Equation (77) can be written as

$$\frac{d}{dt}\omega_i + \frac{1}{2}\frac{\dot{\alpha}_i}{\alpha_i}\omega_i + F(0) = 0, \quad \forall i \in I. \tag{78}$$

Solution of the first order differential equation (78) is given by

$$\begin{aligned}\omega_i(t) = \frac{1}{\sqrt{\alpha_i(t)}}\Bigg(-F(0)\sqrt{\alpha_i(0)}\left(2\sin\frac{t}{2} - 2\frac{\dot{\alpha}_i(0)}{\alpha_i(0)}\cos\frac{t}{2}\right) \\ + \Theta_i) \quad \text{for } \Omega_i \text{ constant, } i \in I.\end{aligned} \tag{79}$$

Therefore,

$$X_i = \sqrt{\alpha_i(t)}\left(-F(0)\sqrt{\alpha_i(0)}\left(2\sin\frac{t}{2} - 2\frac{\dot{\alpha}_i(0)}{\alpha_i(0)}\cos\frac{t}{2}\right) + \Theta\right), \tag{80}$$

for Θ_i constant, $i \in I$. According to the initial conditions, it follows that

$$\Theta = \frac{w_i}{\sqrt{\mu_i}} - 2F(0)\frac{v_i}{\sqrt{\mu_i}}. \tag{81}$$

We conclude that

$$\begin{aligned}X_i(t) =& \sqrt{\alpha_i(t)}\left(-F(0)\sqrt{\mu_i}\left(2\sin\frac{t}{2} - 2\frac{v_i}{\mu_i}\cos\frac{t}{2}\right)\right. \\ &\left. + \frac{w_i}{\sqrt{\mu_i}} - 2F(0)\frac{v_i}{\sqrt{\mu_i}}\right), \quad i \in I.\end{aligned} \tag{82}$$

and it is easy to check that, $\forall t \in [0, l]$, $X(t) = \sum_{i\in I} X_i(t)\delta^i \in T_{\gamma(t)}P_+(I)$ and it is the Levi-civita parallel transport of the vector w along the geodesic curve $\alpha(t)$. □

Theorem 6. *Given two distinct probability measures μ and ν in $\mathcal{P}_+(I)$, a nontrivial tangent vector $w \in T_\mu\mathcal{P}_+(I)$ and a geodesic curve $\alpha : [0, l] \to \mathcal{P}_+(I)$ such that $\alpha(0) = \mu$ and $\alpha(l) = \nu$. The Levi-Civita parallel transport, $\Gamma_{\mu\rightarrowtail\nu} : T_\mu\mathcal{P}_+(I) \to T_\nu\mathcal{P}_+(I)$, that transports a vector w from $T_\mu\mathcal{P}_+(I) = T_{\alpha(0)}\mathcal{P}_+(I)$ to $T_\nu\mathcal{P}_+(I) = T_{\alpha(l)}\mathcal{P}_+(I)$ given by*

$$\begin{aligned}\Gamma_{\mu\rightarrowtail\nu}(w) = \sum_{i\in I}\sqrt{\nu_i}\Bigg(-F(0)\sqrt{\mu_i}\left(2\sin\frac{l}{2} - 2\frac{\tau_i}{\mu_i}\cos\frac{l}{2}\right) \\ + \frac{w_i}{\sqrt{\mu_i}} - 2F(0)\frac{\tau_i}{\sqrt{\mu_i}}\Bigg)\delta^i,\end{aligned} \tag{83}$$

where $l = 2\arccos\sum_{i\in I}\sqrt{\mu_i\nu_i}$, $F(0) = \frac{1}{2}\mathfrak{g}_\mu(w, \tau)$, and τ is the unit tangent vector

$$\tau = \frac{1}{\sin\frac{l}{2}}\sum_{i\in I}\left(\sqrt{\frac{d\nu}{d\mu}}(i) - \sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)\mu_i\delta^i. \tag{84}$$

Proof. It suffices to use the equation of the geodesic curve $\alpha(t)$ joining two points μ and ν given by Theorem 4 together with taking $t = l$ in theorem 5, the proof follows. □

2.2.5. Riemannian mean on $\mathcal{P}_+$

Using Riemannian geodesic distance (37), the Riemannian mean of a set of probability measures $\{\mu_i\}_{i=1}^N$ on $\mathcal{P}_+(I)$ is given by

$$\mu^* = \text{argmin}_\mu \sum_{i=1}^{N} d^{FR}(\mu, \mu_i)^2 \tag{85}$$

In the literature, local optima of the optimization problem (85) are known as Karcher means while a global optimum is called the Fréchet mean [44, 42, 43]. In general Riemannian manifold, the Riemannian mean for a set of points is not unique. Given a set of probability measure on $\mathcal{P}_+(I)$, existence and unicity of the Karcher mean can be proved [40, 42] and the solution can be obtained using gradient descent algorithm.

3. Transfer Learning With Parallel Transport

Motivated by the theoretical insight from the previous section, this section studies the transfer learning problem on the space of probability measures $\mathcal{P}_+(I)$. Specifically, we explore the benefits of using the Riemannian geometry of $\mathcal{P}_+(I)$ and the integration of geodesic tools into transfer learning. To tackle this problem, we consider two subsets, $P_{N_1} = \{\mu_i\}_{i=1}^{N_1}$ and $P_{N_2} = \{\mu_i\}_{i=1}^{N_2}$ from two different domains consisting of N_1 and N_2 probability measures, respectively, with $N_2 \ll N_1$. We denote their respective Karcher means as μ_1^* and μ_2^*. Let $\alpha : [0, l] \to \mathcal{P}_+(I)$ be the geodesic curve given in (39), joining μ_1^* and μ_2^*, with $\alpha(0) = \mu_1^*$ and $\alpha(l) = \mu_2^*$. Finally, let $\Gamma_{\mu_1^* \rightarrowtail \mu_2^*} : T_{\mu_1^*}\mathcal{P}_+(I) \to T_{\mu_1^*}\mathcal{P}_+(I)$, be the Levi-Civita parallel transport along the geodesic curve $\alpha(t)$ given by (83). The goal of transfer learning is to leverage useful information from the source data P_{N_1}, to enhance the model for the target data P_{N_2}. However, direct data transport from the source to the target can be computationally expensive, especially with large source datasets. To address this challenge, we introduce a transfer learning algorithm for statistical models, aiming to leverage the model developed for the source domain, P_{N_1}, as a foundation to create a robust learning model for the target domain, P_{N_2}. The proposed algorithm consists of four main steps:

- **Step** 1: Project the set of probability measure P_{N_1} to the tangent space $T_{\mu_1^*}\mathcal{P}_+(I)$ by $a_i = \log_{\mu_1^*}(\mu_i), i = 1, ..., N_1$, using the logarithm map (60). Similarly, lift the set of probability measure P_{N_2} to the tangent space $T_{\mu_2^*}\mathcal{P}_+(I)$ by $b_i = \log_{\mu_2^*}(\mu_i), i = 1, ..., N_2$.
- **Step** 2: Learn a statistical model S_1 on $T_{\mu_1^*}\mathcal{P}_+(I)$ (respectively a statistical model S_2 on $T_{\mu_2^*}\mathcal{P}_+(I)$).
- **Step** 3: Parallel transport S_1 to $T_{\mu_2^*}\mathcal{P}_+(I)$ along the geodesic curve α by computing $S_T = \Gamma_{\mu_1^* \rightarrowtail \mu_2^*}(S_1)$.
- **Step** 4: Compute the fused model on $T_{\mu_2^*}\mathcal{P}_+(I)$ using shrinkage estimation [41]: $S_\lambda = \lambda S_T + (1 - \lambda)S_2,\ 0 \leq \lambda \leq 1$.

We conclude that the transfer step relies on parallel transport to move a statistical model from $T_{\mu_1^*}\mathcal{P}_+(I)$ to $T_{\mu_2^*}\mathcal{P}_+(I)$ along the unique geodesic α joining μ_1^* and μ_2^*. We also

mention that the preceding steps primarily utilized the existence of a unique geodesic curve connecting two points on $\mathcal{P}+_(I)$. In the sequel, we will provide a detailed account of how parallel transport is applied to transfer some statistical models of interest, including Linear Regression, Logistic Regression, and Principal Component Analysis (PCA).

3.1. Linear regression transport

Let $\mathfrak{g}_{\mu_1^*} : T_{\mu_1^*}\mathcal{P}_+(I) \times T_{\mu_1^*}\mathcal{P}_+(I) \to \mathbb{R}, (v, w) \to \mathfrak{g}_{\mu_1^*}(v, w)$ be the inner product on $T_{\mu_1^*}\mathcal{P}_+(I)$ given by (8). It can equivalently be expressed as $\mathfrak{g}_{\mu_1^*}(v, w) = v^T G_{\mu_1^*} w$, where $G_{\mu_1^*}$ denotes the Fisher-Rao matrix. Given a data set $\mathcal{D} = \{(a_i, t_i), i = 1, ..., N_1\}$ where $\{a_i\}_{i=1}^{N_1}$ are tangents vectors on $T_{\mu_1^*}\mathcal{P}_+(I)$ defined by $a_i = \log_{\mu_1^*}(\mu_i), i = 1, ..., N_1$ and $\{t_i\}_{i=1}^{N_1} \in \mathbb{R}$ denote their corresponding label, a linear regression model $T_{\mu_1^*}\mathcal{P}_+(I) \to \mathbb{R}$ is defined as

$$y_i = a_i^T \beta + \beta_0 = \mathfrak{g}_{\mu_1^*}(a_i, G_{\mu_1^*}^{-1}\beta) + \beta_0, \tag{86}$$

where $\beta_0 \in \mathbb{R}$ and $\beta \in T_{\mu_1^*}\mathcal{P}_+(I)$. Such parameters can be estimated using the squared error loss function $l_i : \mathbb{R} \to \mathbb{R}_+$, $l_i(y_i) = (y_i - t_i)^2 = (a_i^T\beta + \beta_0 - t_i)^2$. More precisely, the estimates $(\widehat{\beta}_0, \widehat{\beta})$ are solutions of the minimization problem

$$(\widehat{\beta}_0, \widehat{\beta}) = \underset{\beta \in T_{\mu_1^*}\mathcal{P}_+(I), \beta_0 \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{N_1} l_i(a_i^T\beta + \beta_0). \tag{87}$$

Proposition 3 (Linear regression transport). The tangent vector $\widehat{\delta} \in T_{\mu_2^*}\mathcal{P}_+(I)$ defined by $\widehat{\delta} = G_{\mu_2^*}\Gamma_{\mu_1^* \rightarrowtail \mu_2^*}(G_{\mu_1^*}^{-1}\widehat{\beta})$ is the estimate for the weight parameter $\delta \in T_{\mu_2^*}\mathcal{P}_+(I)$ on the linear regression model $T_{\mu_2^*}\mathcal{P}_+(I) \to \mathbb{R}$ given by $\tilde{y}_i = \left(\Gamma_{\mu_1^* \rightarrowtail \mu_2^*}(a_i)\right)^T \delta + \beta_0$. Specifically, we have

$$\widehat{\delta} = \underset{\delta \in T_{\mu_2^*}\mathcal{P}_+(I)}{\operatorname{argmin}} \sum_{i=1}^{N_2} l_i(\left(\Gamma_{\mu_1^* \rightarrowtail \mu_2^*}(a_i)\right)^T \delta + \beta_0) \tag{88}$$

Proof. The proof is immediate since the Levi-civita parallel transport (83) conserve inner product. More accurately, we have $\mathfrak{g}_{\mu_1^*}(a_i, G_{\mu_1^*}^{-1}\beta) = \mathfrak{g}_{\mu_2^*}(\Gamma_{\mu_1^* \rightarrowtail \mu_2^*}(a_i), \Gamma_{\mu_1^* \rightarrowtail \mu_2^*}(G_{\mu_1^*}^{-1}\beta))$. □

As shown in Proposition (3), the Levi-civita parallel transport (83) enables us to efficiently transport linear regression model defined on $T_{\mu_1^*}\mathcal{P}_+(I)$ to the tangent space $T_{\mu_2^*}\mathcal{P}_+(I)$. Actually, parallel transport (83) allows a vector a_i in the tangent space $T_{\mu_1^*}\mathcal{P}_+(I)$ to be transported to the tangent space $T_{\mu_2^*}\mathcal{P}_+(I)$, by ensuring that the inner product between a_i and the direction of the geodesic α, joining μ_1^* and μ_2^*, is conserved. Therefore, the estimated parameter $\widehat{\delta}$ is exactly the solution of the linear regression model on $T_{\mu_2^*}\mathcal{P}_+(I)$ defined by the transported tangent vectors $\Gamma_{\mu_1^* \rightarrowtail \mu_2^*}(a_i)$ and is easily computed as the parallel transport of the solution (87) of the linear regression model (86) on $T_{\mu_1^*}\mathcal{P}_+(I)$. Furthermore, let $\overline{y}_i = b_i^T\eta + \eta_0$ be the linear regression model on $T_{\mu_2^*}\mathcal{P}_+(I)$ defined by tangents vectors $b_i = \log_{\mu_2^*}(\mu_i), i = 1, ..., N_2$. Since the two linear regression models $\tilde{y}_i$ and $\overline{y}_i$ live in the same tangent space $T_{\mu_2^*}\mathcal{P}_+(I)$, we can use shrinkage estimation to compute a solution for the fused model as $\eta_\lambda = \lambda\widehat{\delta} + (1 - \lambda)\widehat{\eta},\ 0 \leq \lambda \leq 1$, where $\widehat{\eta}$ represent the least square estimate for the slope parameter η of the linear regression model $\overline{y}_i$. It is clear that the larger λ is, the better is the influence of the linear regression model $\tilde{y}_i$. We summarize the different steps of our approach for linear regression transport in Algorithm **??**.

3.2. Logistic regression transport

Let $\mathcal{D} = \{(a_i, t_i)\}_{i=1}^{N_1}$ be a data set with respect to class labels, where $a_i \in T_{\mu_1^*}\mathcal{P}_+(I)$ defined by $a_i = \log_{\mu_1^*}(\mu_i)$ and $t_i \in \{0, 1\}$, $i = 1, ..., N_1$. The probability of t_i being in class 1 can be represented by a logistic regression function defined by

$$p(a_i) = \frac{1}{1 + e^{-\left(a_i^T\omega + \omega_0\right)}} = \frac{1}{1 + e^{-\left(\mathfrak{g}_{\mu_1^*}(a_i, G_{\mu_1^*}^{-1}\omega) + \omega_0\right)}} \tag{89}$$

where $\omega_0 \in \mathbb{R}$ and $\omega \in T_{\mu_1^*}\mathcal{P}_+(I)$. The probability of t_i being in class 0 is given by: $P(t_i = 0|a_i) = 1 - p(a_i)$. Instead of least-squares, we make use of the maximum likelihood (MLE) to find the estimate prameters ω_0 and ω in the logistic regression function. Nevertheless, it is easily seen that the logistic regression function is defined by means of the inner product $\mathfrak{g}_{\mu_1^*}$ on $T_{\mu_1^*}\mathcal{P}_+(I)$. Hence, with the property that Levi-civita parallel transport preserves the inner product between tangent vectors, we can parallel transport logistic regression model to $T_{\mu_2^*}\mathcal{P}_+(I)$ to obtain a better classification model in this last tangent space. More precisely, if $(\widehat{\omega}_0, \widehat{\omega})$ denote the maximum likelihood estimators (MLE) of (ω_0, ω), then the tangent vector $\widehat{\varpi} \in T_{\mu_2^*}\mathcal{P}_+(I)$ defined by $\widehat{\varpi} = G_{\mu_2^*}\Gamma_{\mu_1^* \rightarrowtail \mu_2^*}(G_{\mu_1^*}^{-1}\widehat{\omega})$ is the estimate for the weight parameter ϖ on the logistic regression model given by

$$p(\Gamma_{\mu_1^* \rightarrowtail \mu_2^*}(a_i)) = \frac{1}{1 + e^{-\left(\Gamma_{\mu_1^* \rightarrowtail \mu_2^*}(a_i)^T\varpi + \omega_0\right)}} \tag{90}$$

Similar to the linear regression model, we make use of shrinkage estimation to compute a solution for the fused logistic regression model on $T_{\mu_2^*}\mathcal{P}_+(I)$ represented by the transported data $\Gamma_{\mu_1^* \rightarrowtail \mu_2^*}(a_i), i = 1, ..., N_1$ and tangents vectors $b_i = \log_{\mu_2^*}(\mu_i), i = 1, ..., N_2$. An equivalent version of Algorithm **??** holds for the transport of logistic regression model.

3.3. Principle component analysis transport

Given the two populations $P_{N_1} = \{\mu_i\}_{i=1}^{N_1}$ and $P_{N_2} = \{\mu_i\}_{i=1}^{N_2}$, the commonly used covariance matrix estimator is the sample covariance matrix defined as

$$C_{N_1} = \frac{1}{N_1 - 1}\sum_{i=1}^{N_1} \log_{\mu_1^*}(\mu_i)\log_{\mu_1^*}(\mu_i)^T = \frac{1}{N_1 - 1}\sum_{i=1}^{N_1} a_i a_i^T = \frac{1}{N_1 - 1}AA^T$$

and

$$C_{N_2} = \frac{1}{N_2 - 1}\sum_{i=1}^{N_2} \log_{\mu_2^*}(\mu_i)\log_{\mu_2^*}(\mu_i)^T = \frac{1}{N_2 - 1}\sum_{i=1}^{N_2} b_i b_i^T = \frac{1}{N_2 - 1}BB^T$$

where $A = [a_1, ..., a_{N_1}] \in \mathbb{R}^{n \times N_1}$ and $B = [b_1, ..., b_{N_2}] \in \mathbb{R}^{n \times N_2}$. Since P_{N_2} is of small size, C_{N_2} may be a poor estimate of the true covariance matrix of P_{N_2}. Hence, our goal is to enhance the covariance estimation C_{N_2} by exploiting C_{N_1}. As a consequence, a well-performed PCA model is constructed on the tangent space $T_{\mu_2^*}\mathcal{P}_+(I)$.

Proposition 4 (PCA transport). Let VDU^T be the SVD of $A = [a_1, ..., a_{N_1}] \in T_{\mu_1^*}\mathcal{P}_+(I)$ with the diagonal entries of D is sorted in non-increasing order and let VD^2V^T be the eigen-decomposition of AA^T. Then

a) $\tilde{V}DU^T$ is the SVD of $\tilde{A} = [\tilde{a}_1, \ldots, \tilde{a}_{N_1}] = \Gamma_{\mu_1^* \to \mu_2^*}(\{a_i\}_{i=1}^{N_1})$ in $T_{\mu_1^*}\mathcal{P}_+(I)$ and $\tilde{V}D^2\tilde{V}^T$ is the eigen-decomposition of $\tilde{A}\tilde{A}^T$.

b) If $k_1 < n_1$, then the k_1-dimensional PCA model of $\tilde{A}$ is given by $\{\tilde{v}_i\}_{i=1}^{k_1}$ and $\{D_{i,i}/\sqrt{N_1 - 1}\}_{i=1}^{k_1}$.

Proof. a) Let VDU^T be the SVD of $A = [a_1, ..., a_{N_1}] \in T_{\mu_1^*}\mathcal{P}_+(I)$, thus $V = [v_1, ..., v_n] \in \mathcal{R}^{n \times n}$ is an orthogonal matrix which contains eigenvectors of AA^T and $U = [u_1, ..., u_{N_1}] \in \mathcal{R}^{N_1 \times N_1}$ is an orthogonal matrix which contains the eigenvectors of A^TA. Let's define the matrix $Z = [z_1, ..., z_{N_1}] \in \mathcal{R}^{n \times N_1}$ by $Z = V^TA = SU^T$. Thus $A = VZ$ and its vector column a_i is consequently defined by $a_i = Vz_i = \sum_{j=1}^{n} v_j Z_{j,i} \cong \mathcal{R}^{n_1}$. By linearity of $\Gamma_{\mu_1^* \rightarrowtail \mu_2^*}$, we get

$$\tilde{a}_i = \Gamma_{\mu_1^* \rightarrowtail \mu_2^*}(a_i) = \Gamma_{\mu_1^* \rightarrowtail \mu_2^*}\left(\sum_{j=1}^{N_1} v_j Z_{j,i}\right) = \sum_{j=1}^{N_1} \tilde{v}_{j\intercal} Z_{j,i}, \tag{91}$$

which means

$$\tilde{A} = [\tilde{v}_1, ..., \tilde{v}_{n_1}]Z = [\tilde{v}_1, ..., \tilde{v}_{n_1}]SU^T = \tilde{V}SU^T. \tag{92}$$

Since $\Gamma_{\mu_1^* \rightarrowtail \mu_2^*}$ is a metric parallel transport, $\tilde{V}$ is an orthogonal matrix. Furthermore, we assert that $\tilde{V}SU^T$ is the SVD of $\tilde{A}$. Moreover,

$$\tilde{A}\tilde{A}^T = \tilde{V}DU^TUD\tilde{V}^T = \tilde{V}D^2\tilde{V}^T. \tag{93}$$

Hence, $\tilde{V}D^2\tilde{V}^T$ is the eigen-decomposition of the covariance matrix $\tilde{A}\tilde{A}^T$.

b) Follows directly from the proof of part a). □

Again, we can apply shrinkage estimation to combine the two covariance matrices C_{N_2} and $\tilde{C}_{N_1} = \frac{1}{N_1 - 1}\tilde{A}\tilde{A}^T$ as follow,

$$C_\lambda = \lambda\tilde{C}_{N_1} + (1 - \lambda)C_{N_2}, \quad 0 \leq \lambda \leq 1. \tag{94}$$

It is easily seen that in the case $\lambda = 1$, C_λ performs extremely well in modeling unseen examples of the small-sample population P_{N_2}. Moreover, for PCA, let $\overline{V} = [\overline{v}_1, ..., \overline{v}_n] \in \mathcal{R}^{n \times n}$ an orthogonal matrix which contains eigenvectors of BB^T. We can consider $\{\tilde{v}_i\}_{i=1}^{k_1}$ and $\{D_{i,i}/\sqrt{N_1 - 1}\}_{i=1}^{k_1}$ as a PCA model on $T_{\mu_2^*}\mathcal{P}_+(I)$, as we can also construct a fused model represented by an orthonormalization of the subspace $\{\tilde{v}, \overline{v}\}$ formed by k_1 eigenvectors of $\tilde{V}$ and k_2 eigenvectors of $\overline{V}$ and their corresponding (k_1, k_2) eigen-values. We summarize different steps of transfer learning covariance matrix and PCA model in Algorithm **??**.

4. Conclusion

In this paper, we proposed a new transfer learning approach of learning models on the space of probability measures $\mathcal{P}_+$ using the analytic expression of parallel transport under the Levi-Civita connection. We provided the mathematical foundation of the proposed transfer learning method. Specifically, we established new results in the geometry of the probability manifold. Finally, an experimental study was conducted to demonstrate the effectiveness of our framework.

Appendix 1. Proof of Claim 1

Proof. We proof the Claim by induction on the degree of I. If $|I|$ is one or two the Claim is true since I_+ is not empty. Suppose the Claim is trus for $|I| = n$. We go to prove the Claim for $|I| = n+1$. Let μ ,τ, $\tilde{\tau}$ and l like in the Claim. Suppose $I_- \neq \emptyset$ then $|I_-| \geq 2$. Let g, h be two distinct index in I_-, this means $\tau_g + \tilde{\tau}_g = -2\mu_g \cot \frac{l}{2}$ and $\tau_h + \tilde{\tau}_h = -2\mu_h \cot \frac{l}{2}$. Now let $k \in I_+$ and define three measures $\tau', \tilde{\tau}', \mu'$ on $I \setminus \{k\}$ as follow

$$\tau' = \sum_{i \in I, i \neq k,h,g} \tau_i \delta^i + \tau_g \delta^g + (\tau_h + \tau_k)\delta^h, \tag{.1}$$

$$\tilde{\tau}' = \sum_{i \in I, i \neq k,h,g} \tilde{\tau}_i \delta^i + (\tilde{\tau}_g + 2\tilde{\tau}_k)\delta^g + (\tilde{\tau}_h - \tilde{\tau}_k)\delta^h, \tag{.2}$$

$$\mu' = \sum_{i \in I, i \neq k,h,g} \mu_i \delta^i + (\mu_g + \mu_k)\delta^g + \mu_h \delta^h. \tag{.3}$$

We have $\tau', \tilde{\tau}' \in T_{\mu'}\mathcal{P}_+(I \setminus \{k\})$, and $h \in I_- \neq \emptyset$. This contradicts to the hypothesis. This shows the Claim for $|I| = n+1$. □

References

[1] S. Thrun, and , and L. Pratt, Learning to Learn: Introduction and Overview, 3-17, Springer, 1998.

[2] S.J. Pan, and , and Q. Yang, A Survey on Transfer Learning, IEE Transactions on Knowledge and Data Engineering, 22 (10), 1345 - 1359, 2010.

[3] X. Li, and Y. Grandvalet, and F. Davoine, and J. Cheng, and Y. Cui, and H. Zhang, and S. Belongie, and Y. H. Tsai, and M. H. Yang, Transfer learning in computer vision tasks: Remember where you come from, Image and Vision Computing ,93, 2020.

[4] A. Brodzicki, and M. Piekarski, and D. Kucharski, and J. J Korjakowska, and M. Gorgon, Transfer Learning Methods as a New Approach in Computer Vision Tasks with Small Datasets, Foundations of Computing and Decision Sciences, 45, 2020.

[5] M. Oquab, and L. Bottou, and I. Laptev, and J. Sivic, Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks, Proceedings of the IEEE conference on computer vision and pattern recognition, 2014.

[6] S. Ruder, and M.E. Peters, and S. Swayamdipta, and T. Wolf, Transfer Learning in Natural Language Processing, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, 93, 2019.

[7] J. Tao and X. Fang, Toward multi-label sentiment analysis: a transfer learning based approach, Journal of Big Data, 7, 2020.

[8] P. Mignone, and G. Pio, and D.D'Elia, and M. Ceci, Exploiting transfer learning for the reconstruction of the human gene regulatory network, Bioinformatics, 36(5), 1553–61, 2020.

[9] M.O. Arowolo, and M.O.Adebiyi, and A.A.Adebiyi, and O. Olugbara, Optimized hybrid investigative based dimensionality reduction methods for malaria vector using knn classifier, Journal of Big Data, 8 (1), 1-14, 2021.

[10] M. Byra, and M. Wu, and X. Zhang, and H. Jang, and Y.J Ma, and E.Y Chang, and S.Shah, and J.Du, Knee menisci segmentation and relaxometry of 3D ultrashort echo time cones MR imaging using attention U-net with transfer learning, Magnetic Resonance in Medicine, 83, 2020.

[11] P. Zhao, and S.C.H. Hoi, OTL: A framework of online transfer learning, Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel.

[12] B.Tan, and Y. Song, and E. Zhong, and Q. Yang, Transitive transfer learning, In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1155-1164, 2015.

[13] H.B. Ammar, and E. Eaton, and J.M. Luna, and P. Ruvolo, Autonomous cross-domain knowledge transfer in lifelong policy gradient reinforcement learning, Proc. 24th International Joint Conference on Artificial Intelligence, 3345–3351, 2015.

[14] M.E. Taylor, and P. Stone, Transfer learning for reinforcement learning domains: A survey, Journal of Machine Learning Research, 10, 1633-1685, 2009.

[15] B. Tan, and E. Zhong, and E.W. Xiang, and Q. Yang, Multi-transfer: transfer learning with multiple views and multiple sources, Statistical Analysis and Data Mining, 4, 282–293, 2014.

[16] Y. He, and Y. Tian, and D. Liu, Multi-view transfer learning with privileged learning framework, Neurocomputing, 335, 131–142, 2019.

[17] H. Liang, and W. Fu, and F. Yi, A Survey of Recent Advances in Transfer Learning, IEEE 19th International Conference on Communication Technology, 362–373, 2019.

[18] F. Zhuang, and Z. Qi, K. Duan, and D. Xi, and Y. Zhu, and H. Zhu, and H. Xiong, and Q. He, A Comprehensive Survey on Transfer Learning, Proceedings of the IEEE, 48, 43-76, 2021.

[19] S. Amari, and H. Nagaoka, Methods of Information Geometry, Translations of Mathematical Monographs, 191, 2000.

[20] S. T. Rachev, and L. B. Klebanov, and S. Stoyanov, and F. J. Fabozzi, The Methods of Distances in the Theory of Probability and Statistics, Springer New York, 2013.

[21] L. Devroye, and L. Györfi, Nonparametric Density Estimation: The L_1 View, John Wiley Sons, 1985.

[22] C. Villani, Optimal Transport: Old and New, Springer Science & Business Media, 2009.

[23] N. Ay, and J. Jost, and H. Le, and L. Schwachhofer, Information geometry, Springer, 2017.

[24] A. Karimi, and L. Ripani, and T.T Georgiou, Statistical Learning in Wasserstein Space, IEEE Control Systems Letters, 05, 2021.

[25] R. Kyng, and J. M. Phillips, and S. Venkatasubramanian, Dimensionality Reduction on the Simplex, Computer Science, Mathematics, 2011.

[26] H. Gzyl, and F. Nielsen, Geometry of the probability simplex and its connection to the maximum entropy method, ournal of Applied Mathematics Statistics and Informatics, 16(1), 25-35, 2020.

[27] J. C. Nascimento, and M. Barão, and J. S. Marques, and J.M. Lemos, An information geometric framework for the optimization on a discrete probability spaces: Application to human trajectory classification, Neurocomputing, 150, 155-162, 2015.

[28] F. Nielsen, and K. Sun, Clustering in Hilbert simplex geometry, preprint arXiv:1704.00454, 2017.

[29] S. Hauberg, and F. Lauze, and K. Pedersen, Unscented Kalman filtering on Riemannian manifolds, Journal of Mathematical Imaging and Vision, 46, 1-18, 2013.

[30] Q. Xie, and S. Kurtek, and H. Le, and A. Srivastava, Transitive transfer learning, IEEE International Conference on Computer Vision, 2013.

[31] P. Zanini, and M. Congedo, and C. Jutten, and S. Said, and Y. Berthoumieu, Transfer learning: A Riemannian geometry framework with applications to braincomputer interfaces, IEEE Trans. Biomed. Eng, 65 (5), 1107–1116, 2018.

[32] O. Yair, and M. Ben-Chen, and R. Talmon, Parallel Transport on the Cone Manifold of SPD Matrices for Domain Adaptation, IEEE Transactions on Signal Processing, 67 (7), 1797 - 1811, 2019.

[33] N. Guigui, and X. Pennec, Parallel transport, a central tool in geometric statistics for computational anatomy: Application to cardiac motion modeling, Handbook of Statistics, 46, 285-326, 2022.

[34] L. Younes, and A. Qiu, and R.L. Winslow, and M.I. Miller, Transport of Relational Structures in Groups of Diffeomorphisms, Journal of mathematical imaging and vision, 32(01), 41-56, 2008.

[35] M. Lorenzi, and N. Ayache, and X. Pennec, Schild's Ladder for the Parallel Transport of Deformations in Time Series of Images, Information Processing in Medical Imaging, 463–474, 2011.

[36] O. Freifeld, and S. Hauberg and M. J. Black, Model Transport: Towards Scalable Transfer Learning on Manifolds, CVPR, 1378-1385, 2014.

[37] S. Amari, and H. Nagaoka, Methods of Information Geometry, American Mathematical Society: Providence, RL, 2007.

[38] C. R. Radhakrishna, Information and accuracy attainable in the estimation of statistical parameters, 37 (3), 81–91, 1945.

[39] N. Ay, and J. Jost, and H. Le, and L. Schwachhofer, *Information geometry.* Springer, 2017.

[40] S. Lang, Fundamentals of Differential Geometry, Springer New York, 1999.

[41] J. Schäfer, and K. Strimmer, A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, SAGMB, 4(1), 2005.

[42] X. Pennec, Intrinsic statistics on Riemannian manifolds: basic tools for geometric measurements, 25 (1), 127–154, 2006.

[43] P.T. Fletcher, Statistics on manifolds, Riemannian Geometric Statistics in Medical Image Analysis, 39-74, 2020;

[44] H. Karcher, Riemannian Center of Mass and Mollifier Smoothing, Comm. Pure and Applied Math., 30, 509-541, 1977.