

A Unified Framework for Domain Adaptation using Metric Learning on Manifolds

Sridhar Mahadevan¹ , Bamdev Mishra², and Shalini Ghosh³

¹ University of Massachusetts, Amherst, MA 01003

mahadeva@cs.umass.edu

² Microsoft, Hyderabad 500032, Telangana, India

bamdevm@microsoft.com

³ Samsung Research America, Mountain View, CA 94043

shalini.ghosh@samsung.com

Abstract. We present a novel framework for domain adaptation, whereby both *geometric* and *statistical* differences between a labeled source domain and unlabeled target domain can be reconciled using a unified mathematical framework that exploits the curved Riemannian geometry of statistical manifolds. We exploit a simple but important observation that as the space of covariance matrices is both a Riemannian space as well as a homogeneous space, the shortest path geodesic between two covariances on the manifold can be computed analytically. Statistics on the SPD matrix manifold, such as the geometric mean of two SPD matrices can be reduced to solving the well-known Riccati equation. We show how the Riccati-based solution can be constrained to not only reduce the statistical differences between the source and target domains, such as aligning second order covariances and minimizing the maximum mean discrepancy, but also the underlying geometry of the source and target domains using diffusions on the underlying source and target manifolds. Our solution also emerges as a consequence of optimal transport theory, which shows that the optimal transport mapping between source and target distributions that are multivariate Gaussians is a function of the geometric mean of the source and target covariances, a quantity that also minimizes the Wasserstein distance. A key strength of our proposed approach is that it enables integrating multiple sources of variation between source and target in a unified way, by reducing the combined objective function to a nested set of Riccati equations where the solution can be represented by a cascaded series of geometric mean computations. In addition to showing the theoretical optimality of our solution, we present detailed experiments using standard transfer learning testbeds from computer vision comparing our proposed algorithms to past work in domain adaptation, showing improved results over a large variety of previous methods.

1 Introduction

When we apply machine learning [19] to real-world problems, e.g., in image recognition [18] or speech recognition [13], a significant challenge is the need for having large amounts of (labeled) training data, which may not always be available. Consequently, there has been longstanding interest in developing machine learning techniques that

can transfer knowledge across domains, thereby alleviating to some extent the need for training data as well as the time required to train the machine learning system. A detailed survey of transfer learning is given in [22].

Traditional machine learning assumes that the distribution of test examples follows that of the training examples [8], whereas in transfer learning, this assumption is usually violated. *Domain adaptation* (DA) is a well-studied formulation of transfer learning that is based on developing methods that deal with the change of distribution in test instances as compared with training instances [5, 11]. In this paper, we propose a new framework for domain adaptation, based on formulating transfer from source to target as a problem of geometric mean metric learning on manifolds. Our proposed approach enables integrating multiple sources of variation between source and target in a unified framework with a theoretically optimal solution. We also present detailed experiments using standard transfer learning testbeds from computer vision, showing how our proposed algorithms give improved results compared to existing methods.

A general way to model domain adaptation is using the framework of optimal transport (OT) [9, 23]. The OT problem, originally studied by Monge in 1781, is defined as the effort needed to move a given quantity of dirt (formally modeled by some source distribution) to a new location with no loss in volume (modeled formally by a target distribution), minimizing some cost of transportation per unit volume of dirt. A revised formulation by Kantorovich in 1942 allowed the transport map to be (in the discrete case) a bistochastic matrix, whose rows sum to the source distribution, and whose columns sum to the target distribution. When the cost of transportation is given by a distance metric, the optimal transport distance is defined as the *Wasserstein distance*. The solution we propose in this paper, based on the geometric mean of the source and target covariances, can be derived from optimal transport theory, under the assumption that the source and target distributions correspond to multivariate Gaussian distributions. Our approach goes beyond the simple solution proposed by optimal transport theory in that we take into account not only the cost of transportation, but also other factors, such as the geometry of source and target domains. In addition, we use the *weighted* geometric mean, giving us additional flexibility in tuning the solution.

Background: One standard approach of domain adaptation is based on modeling the *covariate shift* [1]. Unlike traditional machine learning, in DA, the training and test examples are assumed to have different distributions. It is usual in DA to categorize the problem into different types: (i) semi-supervised domain adaptation (ii) unsupervised domain adaptation (iii) multi-source domain adaptation (iv) heterogeneous domain adaptation.

Another popular approach to domain adaptation is based on aligning the distributions between source and target domains. A common strategy is based on the maximum mean discrepancy (MMD) metric [7], which is a nonparametric technique for measuring the dissimilarity of empirical distributions between source and target domains. Domain-invariant projection is one method that seeks to minimize the MMD measure using optimization on the Grassmannian manifold of fixed-dimensional subspaces of n -dimensional Euclidean space [2].

Linear approaches to domain adaptation involve the use of alignment of lower-dimensional subspaces or covariances from a data source domain $D_s = \{x_i^s\}$ with

labels $L_s = \{y_i\}$ to a target data domain $D_t = \{x_i^t\}$. We assume both x_i^s and x_i^t are n -dimensional Euclidean vectors, representing the values of n features of each training example. One popular approach to domain adaptation relies on first projecting the data from the source and target domains onto a low-dimensional subspace, and then finding correspondences between the source and target subspaces. Of these approaches, the most widely used one is Canonical Correlation Analysis (CCA) [17], a standard statistical technique used in many applications of machine learning and bioinformatics. Several nonlinear versions [15] and deep learning variants [27] of CCA have been proposed. These methods often require explicit correspondences between the source and target domains to learn a common subspace. Because CCA finds a linear subspace, a family of manifold alignment methods have been developed that extend CCA [25, 10] to exploit the nonlinear structure present in many datasets.

In contrast to using a single shared subspace across source and target domains, *subspace alignment* finds a linear mapping that transforms the source data subspace into the target data subspace [14]. To explain the basic algorithm, let $P_S, P_T \in \mathbb{R}^{n \times d}$ denote the two sets of basis vectors that span the subspaces for the “source” and “target” domains. Subspace alignment attempts to find a linear mapping M that minimizes

$$F(M) = \|P_S M - P_T\|_F^2.$$

It can be shown that the solution to the above optimization problem is simply the dot product between P_S and P_T , i.e.:

$$M^* = \operatorname{argmin}_M F(M) = P_S^T P_T.$$

Another approach exploits the property that the set of k -dimensional subspaces in n -dimensional Euclidean space forms a curved manifold called the *Grassmannian* [12], a type of matrix manifold. The domain adaptation method called geodesic flow kernels (GFK) [16] is based on constructing a distance function between source and target subspaces that is based on the geodesic or shortest path between these two elements on the Grassmannian.

Rather than aligning subspaces, a popular technique called CORAL [24] aligns correlations between source and target domains. Let μ_s, μ_t and A_s, A_t represent the mean and covariance of the source and target domains, respectively. CORAL finds a linear transformation A that minimizes the distance between the second-order statistics of the source and target features (which can be assumed as normalized with zero means). Using the Frobenius (Euclidean) norm as the matrix distance metric, CORAL is based on solving the following optimization problem:

$$\min_A \|A^T A_s A - A_t\|_F^2. \quad (1)$$

where A_s, A_t are of size $n \times n$. Using the singular value decomposition of A_s and A_t , CORAL [24] computes a particular closed-form solution⁴ to find the desired linear transformation A .

⁴ The solution characterization in [24] is non unique. [24, Theorem 1] shows that the optimal A , for full-rank A_s and A_t , is characterized as $A = U_s \Sigma_s^{-1} U_s^T U_t \Sigma_t^{1/2} U_t^T$, where $U_s \Sigma_s U_s^T$

Novelty of our Approach: Our proposed solution differs from the above previous approaches in several fundamental ways: one, we explicitly model the space of covariance matrices as a curved Riemannian manifold of symmetric positive definite (SPD) matrices. Note the difference of two SPD matrices is not an SPD matrix, and hence they do not form a vector space. Second, our approach can be shown to be both unique and globally optimal, unlike some of the above approaches. Uniqueness and optimality derive from the fact that we reduce all domain adaptation computations to nested equations involving solving the well-known *Riccati* equation [6].

The organization of the paper is as follows. In Section 2, we show the connection between the domain adaptation problem to the metric learning problem. In particular, we discuss the Riccati point of view for the domain adaptation problem. Section 3 discusses briefly the Riemannian geometry of the space of SPD matrices. Sections 4 and 5 discuss additional domain adaptation formulations. Our proposed algorithms are presented in Sections 6 and 7. Finally, in Section 8 we show the experimental results on the standard Office and the extended Office-Caltech10 datasets, where our algorithms show clear improvements over CORAL.

2 Domain Adaptation using Metric Learning

In this section, we will describe the central idea of this paper: modeling the problem of domain adaptation as a geometric mean metric learning problem. Before explaining the specific approach, it will be useful to introduce some background. The metric learning problem [4] involves taking input data in \mathbb{R}^n and constructing a (non)linear mapping $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^m$, so that the distance between two points x and y in \mathbb{R}^n can be measured using the distance $\|\Phi(x) - \Phi(y)\|$. A simple approach is to learn a squared *Mahalanobis* distance: $\delta_A^2(x, y) = (x - y)^T A (x - y)$, where $x, y \in \mathbb{R}^n$ and A is an $n \times n$ symmetric positive definite (SPD) matrix. If we represent $A = W^T W$, for some linear transformation matrix W , then it is easy to see that $\delta_A^2(x, y) = \|Wx - Wy\|_F^2$, thereby showing that the Mahalanobis distance is tantamount to projecting the data into a potentially lower-dimensional space, and measuring distances using Euclidean (Frobenius) norm. Typically, the matrix A is learned using some weak supervision, given two sets of training examples of the form:

$$\mathcal{S} = \{(x_i, x_j) | x_i \text{ and } x_j \text{ are in the same class}\},$$

$$\mathcal{D} = \{(x_i, x_j) | x_i \text{ and } x_j \text{ are in different classes}\}.$$

A large variety of metric learning methods can be designed based on formulating different optimization objectives based on functions over the \mathcal{S} and \mathcal{D} sets to extract information about the distance matrix A .

and $U_t \Sigma_t U_t^T$ are the eigenvalue decompositions of A_s and A_t , respectively. However, it can be readily checked that there exists a continuous set of optimal solutions characterized as $A = U_s \Sigma_s^{-1/2} U_s^T O U_t \Sigma_t^{1/2} U_t^T$, where O is any orthogonal matrix, i.e., $O O^T = O^T O = I$ of size $n \times n$. A similar construction for non-uniqueness of the CORAL solution also holds for rank deficient A_s and A_t .

For our purposes, the method that will provide the closest inspiration to our goal of designing a domain adaptation method based on metric learning is the recently proposed *geometric mean metric learning* (GMML) algorithm [28]. GMML models the distance between points in the S set by the Mahalanobis distance $\delta_A(x_i, x_j), x_i, x_j \in S$ by exploiting the geometry of the SPD matrices, and crucially, also models the distance between points in the disagreement set D by the *inverse* metric $\delta_{A^{-1}}(x_i, x_j), x_i, x_j \in D$. GMML is based on solving the objective function over all SPD matrices A :

$$\min_{A \succ 0} \sum_{(x_i, x_j) \in S} \delta_A^2(x_i, x_j) + \sum_{(x_i, x_j) \in D} \delta_{A^{-1}}^2(x_i, x_j),$$

where $A \succ 0$ refers to the set of all SPD matrices.

Several researchers have previously explored the connection between domain adaptation and metric learning. One recent approach is based on constructing a transformation matrix A that both minimizes the difference between the source and target distributions based on the previously noted MMD metric, but also captures the manifold geometry of source and target domains, and attempts to preserve the discriminative power of the label information in the source domain [26]. Our approach builds on these ideas, with some significant differences. One, we use an objective function that is based on finding a solution that lies on the geodesic between source and target (estimated) covariance matrices (which are modeled as symmetric positive definite matrices). Second, we use a cascaded series of geometric mean computations to balance multiple factors. We describe these ideas in more detail in this and the next section.

We now describe how the problem of domain adaptation can be considered as a type of metric learning problem, called geometric mean metric learning (GMML) [28]. Recall that in domain adaptation, we are given a source dataset D_s (usually with a set of training labels) and a target dataset D_t (unlabeled). The aim of domain adaptation, as reviewed above, is to construct an intermediate representation that combines some of the features of both the source and target domains, with the rationale being that the distribution of target features differs from that of the source. Relying purely on either the source or the target features is therefore suboptimal, and the challenge is to determine what intermediate representation will provide optimal transfer between the domains.

To connect metric learning to domain adaptation, note that we can define the two sets S and D in the metric learning problem as associated with the source and target domains respectively, whereby ⁵

$$\begin{aligned} S &\subseteq D_s \times D_s = \{(x_i, x_j) | x_i \in D_s, x_j \in D_s\} \\ D &\subseteq D_t \times D_t = \{(x_i, x_j) | x_i \in D_t, x_j \in D_t\}. \end{aligned}$$

Our approach seeks to exploit the nonlinear geometry of covariance matrices to find a Mahalanobis distance matrix A , such that we can represent distances in the source domain using A , but crucially we measure distances in the target domain using the inverse A^{-1} .

$$\min_{A \succ 0} \sum_{(x_i, x_j) \in S} (x_i - x_j)^T A (x_i - x_j) + \sum_{(x_i, x_j) \in D} (x_i - x_j)^T A^{-1} (x_i - x_j).$$

⁵ We note that while there are alternative ways to define the S and D sets, the essence of our approach remains similar.

To provide some intuition here, we observe that as we vary A to reduce the distance $\sum_{(x_i, x_j) \in S} (x_i - x_j)^T A (x_i - x_j)$ in the source domain, we simultaneously increase the distance in the target domain by minimizing $\sum_{(x_i, x_j) \in D} (x_i - x_j)^T A^{-1} (x_i - x_j)$, and vice versa. Consequently, by appropriately choosing A , we can seek to minimize the above sum. We can now use the matrix trace to reformulate the Mahalanobis distances:

$$\min_{A \succ 0} \sum_{(x_i, x_j) \in S} \text{tr}(A(x_i - x_j)(x_i - x_j)^T) + \sum_{(x_i, x_j) \in D} \text{tr}(A^{-1}(x_i - x_j)(x_i - x_j)^T).$$

Denoting the source and target covariance matrices A_s and A_t as:

$$A_s := \sum_{(x_i, x_j) \in S} (x_i - x_j)(x_i - x_j)^T \quad (2)$$

$$A_t := \sum_{(x_i, x_j) \in D} (x_i - x_j)(x_i - x_j)^T, \quad (3)$$

we can finally write a new formulation of the domain adaptation problem as minimizing the following objective function to find the SPD matrix A such that:

$$\min_{A \succ 0} \omega(A) := \text{tr}(AA_s) + \text{tr}(A^{-1}A_t). \quad (4)$$

3 Riemannian Geometry of SPD Matrices

In this section, we outline some other formulations of domain adaptation that will be useful to discuss for presenting our overall approach.

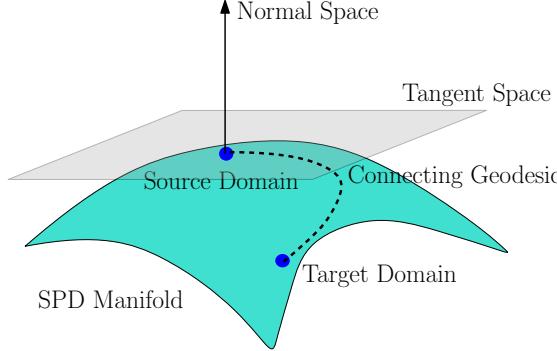


Fig. 1: The space of symmetric positive definite matrices forms a Riemannian manifold, as illustrated here. The methods we propose are based on computing geodesics (the shortest distance), shown in the dotted line, between source domain information and target domain information.

As Figure 1 shows, our proposed approach to domain adaptation builds on the non-linear geometry of the space of SPD (or covariance) matrices, we review some of this

material first [6]. Taking a simple example of a 2×2 SPD matrix M , where:

$$M = \begin{bmatrix} a & b \\ b & c \end{bmatrix},$$

where $a > 0$, and the SPD requirement implies the positivity of the determinant $ac - b^2 > 0$. Thus, the set of all SPD matrices of size 2×2 forms the interior of a cone in \mathbb{R}^3 . More generally, the space of all $n \times n$ SPD matrices forms a manifold of non-positive curvature in \mathbb{R}^{n^2} [6]. In the CORAL objective function in Equation (1), the goal is to find a transformation A that makes the source covariance resemble the target as closely as possible. Our approach simplifies Equation (1) by restricting the transformation matrix A to be a SPD matrix, i.e., $A \succ 0$, and furthermore, we solve the resulting nonlinear equation *exactly* on the manifold of SPD matrices. More formally, we solve the *Riccati equation* [6]:

$$AA_s A = A_t, \text{ for } A \succ 0, \quad (5)$$

where A_s and A_t are source and target covariances SPD matrices, respectively. Note that in comparison with the CORAL approach in Equation (1), the A matrix is symmetric (and positive definite), so A and A^T are the same. The solution to the above Riccati equation is the well-known *geometric mean* or *sharp mean*, of the two SPD matrices, A_s^{-1} and A_t .

$$A = A_s^{-1} \sharp_{\frac{1}{2}} A_t = A_s^{-\frac{1}{2}} (A_s^{\frac{1}{2}} A_t A_s^{\frac{1}{2}})^{\frac{1}{2}} A_s^{-\frac{1}{2}},$$

where $\sharp_{\frac{1}{2}}$ denotes the geometric mean of SPD matrices [28]. The sharp mean has an intuitive geometric interpretation: it is the midpoint of the geodesic connecting the source domain A_s^{-1} and target domain A_t matrices, where length is measured on the Riemannian manifold of SPD matrices. As mentioned earlier, the geometric mean is also the solution to the domain adaptation problem from optimal transport theory when the source and target distributions are given by multivariate Gaussians [23].

In a manifold, the shortest path between two elements, if it exists, can be represented by a geodesic. For the SPD manifold, it can be shown that the geodesic $\gamma(t)$ for a scalar $0 \leq t \leq 1$ between A_s^{-1} and A_t , is given by [6]:

$$\gamma(t) = A_s^{-1/2} (A_s^{1/2} A_t A_s^{1/2})^t A_s^{-1/2}.$$

It is common to denote $\gamma(t)$ as the so-called “weighted” sharp mean $A_s^{-1} \sharp_t A_t$. It is easy to see that for $t = 0$, $\gamma(0) = A_s^{-1}$, and for $t = 1$, we have $\gamma(1) = A_t$. For the distinguished value of $t = 1/2$, it turns out that $\gamma(1/2)$ is the *geometric mean* of A_s^{-1} and A_t , respectively, and satisfies all the properties of a geometric mean [6].

$$\gamma(1/2) = A_s^{-1/2} \sharp_{1/2} A_t = A_s^{-1/2} (A_s^{1/2} A_t A_s^{1/2})^{1/2} A_s^{-1/2}.$$

The following theorem summarizes some of the properties of the objective function given by Equation (4).

Theorem 1. [28, Theorem 3] *The cost function $\omega(A)$ in Equation (4) is both strictly convex as well as strictly geodesically convex on the SPD manifold.*

Theorem 1 ensures the uniqueness of the solution to Equation (4).

4 Statistical Alignment Across Domains

A key strength of our approach is that it can exploit both geometric and statistical information, and multiple sources of alignment are integrated by solving nested sets of Riccati equations. To illustrate this point, in this section we explicitly introduce a secondary criterion of aligning the source and target domains so that the underlying (marginal) distributions are similar. As our results show later, we obtain a significant improvement over CORAL on a standard computer vision dataset (Office/Caltech/Amazon problem). The reason our approach outperforms CORAL is that not only are we able to solve the Riccati equation uniquely, whereas the CORAL solution proposed is [24] is only a particular solution due to non-uniqueness⁴, we can exploit multiple sources of information.

A common way to incorporate the statistical alignment constraint is based on minimizing the maximum mean discrepancy metric (MMD) [7], a nonparametric measure of the difference between two distributions.

$$\left\| \frac{1}{n} \sum_{i=1}^n Wx_i^s - \frac{1}{m} \sum_{i=1}^m Wx_i^t \right\|^2 = \text{tr}(AXLX^T), \quad (6)$$

where $X = \{x_1^s, \dots, x_n^s, x_1^t, \dots, x_t^t\} \in \mathbb{R}^{(n+m)}$ and $L \in \mathbb{R}^{(n+m)(n+m)}$, where $L(i, j) = \frac{1}{n^2}$ if $x_i, x_j \in X_s$, $L(i, j) = \frac{1}{m^2}$ if $x_i, x_j \in X_t$, and $L(i, j) = -\frac{1}{mn}$ otherwise. It is straightforward to show that $L \succeq 0$, a symmetric positive-semidefinite matrix [26]. We can now combine the MMD objective in Equation (6) with the previous geometric mean objective in Equation (4) to give rise to the following modified objective function:

$$\min_{A \succ 0} \xi(A) := \text{tr}(AA_s) + \text{tr}(A^{-1}A_t) + \text{tr}(AXLX^T). \quad (7)$$

We can once again find a closed-form solution to the modified objective in Equation (7) by taking gradients:

$$\nabla \xi(A) = A_s - A^{-1}A_tA^{-1} + XLX^T = 0,$$

whose solution is now given $A = A_m^{-1} \sharp_{\frac{1}{2}} A_t$, where $A_m = A_s + XLX^T$.

5 Geometrical Diffusion on Manifolds

So far we have shown how the solution to the domain adaptation problem can be shown to involve finding the geometric mean of two terms, one involving the source covariance information and the Maximum Mean Discrepancy (MMD) of source and target training instances, and the second involving the target covariance matrix. In this section, we impose additional geometrical constraints on the solution that involve modeling the nonlinear manifold geometry of the source and target domains.

The usual approach is to model the source and target domains as a nonlinear manifold and set up a diffusion on a discrete graph approximation of the continuous manifold [20], using a random walk on a nearest neighbor graph connecting nearby points.

Standard results have been established showing asymptotic convergence of the graph Laplacian to the underlying manifold Laplacian [3]. We can use the above algorithm to find two graph kernels K_s and K_t that are based on the eigenvectors of the random walk on the source and target domain manifold, respectively.

$$K_s = \sum_{i=1}^m e^{-\frac{\sigma_s^2}{2\lambda_i^s}} v_i^s (v_i^s)^T$$

$$K_t = \sum_{i=1}^n e^{-\frac{\sigma_t^2}{2\lambda_i^t}} v_i^t (v_i^t)^T.$$

Here, v_i^s and v_i^t refer to the eigenvectors of the random walk diffusion matrix on the source and target manifolds, respectively, and λ_i^s and λ_i^t refer to the corresponding eigenvalues.

We can now introduce a new objective function that incorporates the source and target domain manifold geometry:

$$\min_{A \succ 0} \eta(A) := \text{tr}(AX(K + \mu L)X^T) + \text{tr}(AA_s) + \text{tr}(A^{-1}A_t), \quad (8)$$

where $K \succ 0$ and $K = \begin{pmatrix} K_s^{-1} & 0 \\ 0 & K_t^{-1} \end{pmatrix}$, and μ is a weighting term that combines the geometric and statistical constraints over A .

Once again, we can exploit the SPD nature of the matrices involved, the closed-form solution to Equation (8) is $A = A_{gs} \sharp_{\frac{1}{2}} A_t$, where $A_{gs} = A_s + X(K + \mu L)X^T$.

6 Cascaded Weighted Geometric Mean

One additional refinement that we use is the notion of a *weighted* geometric mean. To explain this idea, we introduce the following Riemannian distance metric on the nonlinear manifold of SPD matrices:

$$\delta_R^2(X, Y) \equiv \left\| \log(Y^{-\frac{1}{2}}XY^{-\frac{1}{2}}) \right\|_F^2$$

for two SPD matrices X and Y . Using this metric, we now introduce a *weighted* version of the previous objective functions in Equations (4), (7), and (8).

For the first objective function in Equation (4), we get:

$$\min_{A \succ 0} \omega_t(A) := (1 - t) \delta_R^2(A, A_s^{-1}) + t \delta_R^2(A, A_t), \quad (9)$$

where $0 \leq t \leq 1$ is the weight parameter. The unique solution to (9) is given by the *weighted geometric mean* $A = A_s^{-1} \sharp_t A_t$ [28]. Note that the weighted metric mean is no longer strictly convex (in the Euclidean sense), but remains geodesically strictly convex [28, 6, Chapter 6].

Similarly, we introduce the weighted variant of the objective function given by Equation (7):

$$\min_{A \succ 0} \xi_t(A) := (1 - t) \delta_R^2(A, (A_s + XUX^T)^{-1}) + t \delta_R^2(A, A_t), \quad (10)$$

whose unique solution is given by $A = A_m^{-1} \sharp_t A_t$, where $A_m = A_s + X L X^T$ as before. A cascaded variant is obtained when we further exploit the SPD structure of A_s and $X L X^T$, i.e., $A_m = A_s \sharp_\gamma (X L X^T)$ (weighted geometric mean of A_s and $(X L X^T)$ instead of $A_m = A_s + X L X^T$ (which is akin to the Euclidean mean of A_s and $(X L X^T)$). Here, $0 \leq \gamma \leq 1$ is the weight parameter.

Finally, we obtain the weighted variant of the third objective function in Equation (8):

$$\min_{A \succ 0} \eta_t(A) := (1-t) \delta_R^2(A, (A_s + X(K + \mu I)X^T)^{-1} + t \delta_R^2(A, A_t)), \quad (11)$$

whose unique solution is given by $A = A_{gs}^{-1} \sharp_t A_t$, where $A_{gs} = A_s + X(K + \mu L)X^T$ as previously noted. Additionally, the cascaded variant is obtained when $A_s \sharp_\gamma (X(K + \mu L)X^T)$ instead of $A_{gs} = A_s + X(K + \mu L)X^T$.

7 Domain Adaptation Algorithms

We now describe the proposed domain adaptation algorithms, based on the above development of approaches reflecting geometric and statistical constraints on the inferred solution. All the proposed algorithms are summarized in Algorithm 1. The algorithms are based on finding a Mahalanobis distance matrix A interpolating source and target covariances (GCA1), incorporating an additional MMD metric (GCA2) and finally, incorporating the source and target manifold geometry (GCA3). It is noteworthy that all the variants rely on computation of the sharp mean, a unifying motif that ties together the various proposed methods. Modeling the Riemannian manifold underlying SDP matrices ensures the optimality and uniqueness of our proposed methods.

8 Experimental Results

We present experimental results using the standard computer vision testbed used in prior work: the Office [16] and extended Office-Caltech10 [14] benchmark datasets. The Office-Caltech10 dataset contains 10 object categories from an office environment (e.g., keyboard, laptop, and so on) in four image domains: Amazon (**A**), Caltech256 (**C**), DSLR (**D**), and Webcam (**W**). The Office dataset has 31 categories (the previous 10 categories and 21 additional ones).

An exhaustive comparison of the three proposed methods with a variety of previous methods is summarized by the table in Table 1. The previous methods compared in the table refer to the unsupervised domain adaptation approach where a support vector machine (SVM) classifier is used. The experiments follow the standard protocol established by previous works in domain adaptation using this dataset. The features used (SURF) are encoded with 800-bin bag-of-words histograms and normalized to have zero mean and unit standard deviation in each dimension. As there are four domains, there are 12 ensuing transfer learning problems, denoted in Table 1 below as **A** \rightarrow **D** (for Amazon to DSLR, etc.). For each of the 12 transfer learning tasks, the best performing method is indicated in boldface. We used 30 randomized trials for each experiment, and randomly sample the same number of labeled images in the source

Algorithm 1 Algorithms for Domain Adaptation using Metric Learning

Given: A source dataset of labeled points $x_i^s \in D_s$ with labels $L_s = \{y_i\}$, and an unlabeled target dataset $x_i^t \in D_t$, with hyperparameters t, μ , and γ .

1. Define $X = \{x_1^s, \dots, x_n^s, x_1^t, \dots, x_m^t\}$.
2. Compute the source and target matrices A_s and A_t using Equations (2) and (3).
3. **Algorithm GCA1:** Compute the weighted geometric mean $A = A_s^{-1} \sharp_t A_t$ (see Equation (9)).
4. **Algorithm GCA2:** Compute the weighted geometric mean taking additionally into account the MMD metric $A = A_m^{-1} \sharp_t A_t$ (see Equation (10)), where $A_m = A_s + X L X^T$.
5. **Algorithm Cascaded-GCA2:** Compute the *cascaded* weighted geometric mean taking additionally into account the MMD metric $A = A_m^{-1} \sharp_t A_t$ (see Equation (10)), $A_m = A_s \sharp_\gamma X L X^T$.
6. **Algorithm GCA3:** Compute the weighted geometric mean taking additionally into account the source and target manifold geometry $A = A_{gs}^{-1} \sharp_t A_t$ (see Equation (11)), where $A_{gs} = A_s + X(K + \mu L)X^T$.
7. **Algorithm Cascaded-GCA3:** Compute the *cascaded* weighted geometric mean taking additionally into account the source and target manifold geometry $A = A_{gs}^{-1} \sharp_t A_t$ (see Equation (11)), where $A_{gs} = A_s \sharp_\gamma (X(K + \mu L)X^T)$.
8. Use the learned A matrix to adapt source features to the target domain, and perform classification (e.g., using support vector machines).

domain as training set, and use all the unlabeled data in the target domain as the test set. All experiments used a support vector machine (SVM) method to measure classifier accuracy, using a standard `libsvm` package. The methods compared against in Table 1 include the following alternatives:

- **Baseline-S:** This approach uses the projection defined by using PCA in the source domain to project both source and target data.
- **Baseline-T:** Here, PCA is used in the target domain to extract a low-dimensional subspace.
- **NA:** No adaptation is used and the original subspace is used for both source and target domains.
- **GFK:** This approach refers to the *geodesic flow kernel* [16], which computes the geodesic on the Grassmannian between the PCA-derived source and target subspaces computed from the source and target domains.
- **TCA:** This approach refers to the transfer component analysis method [21].
- **SA:** This approach refers to the subspace alignment method [14].
- **CORAL:** This approach refers to the correlational alignment method [24].
- **GCA1:** This is a new proposed method, based on finding the weighted geometric mean of the inverse of the source matrix A_s and the target matrix A_t .
- **GCA2:** This is a new proposed method, based on finding the (non-cascaded) weighted geometric mean of the inverse of the source matrix A_m and the target matrix A_t .
- **GCA3:** This is a new proposed method, based on finding the (non-cascaded) weighted geometric mean of the inverse of the source matrix A_{gs} and the target matrix A_t .
- **Cascaded-GCA2:** This is a new proposed method, based on finding the cascaded geometric mean of the inverse revised source matrix A_m and target matrix A_t

Method	C → A	D → A	W → A	A → C	D → C	W → C
NA	44.0	34.6	30.7	35.7	30.6	23.4
B-S	44.3	36.8	32.9	36.8	29.6	24.9
B-T	44.5	38.6	34.2	37.3	31.6	28.4
GFK	44.8	37.9	37.1	38.3	31.4	29.1
TCA	47.2	38.8	34.8	40.8	33.8	30.9
SA	46.1	42.0	39.3	39.9	35.0	31.8
CORAL	47.1	38.0	37.7	40.5	33.9	34.4
GCA1	48.4	40.1	38.6	41.4	36.0	35.0
GCA2	48.9	40.0	38.6	41.0	35.9	35.0
GCA3	48.4	40.9	37.6	41.1	36	33.6
Cascaded-GCA2	49.5	41.1	38.6	41.1	36	35.1
Cascaded-GCA3	49.4	41.0	38.5	41.0	35.9	35

Method	A → D	C → D	W → D	A → W	C → W	D → W
NA	34.5	36.0	67.4	26.1	29.1	70.9
B-S	36.1	38.9	73.6	42.5	34.6	75.4
B-T	32.5	35.3	73.6	37.3	34.2	80.5
GFK	32.5	36.1	74.6	39.8	34.9	79.1
TCA	36.4	39.2	72.1	38.1	36.5	80.3
SA	44.5	38.6	34.2	37.3	31.6	28.4
CORAL	38.1	39.2	84.4	38.2	39.7	85.4
GCA1	39.2	40.9	85.1	40.9	41.1	87.2
GCA2	39.9	41.4	85.1	40.1	41.1	87.2
GCA3	38.9	41.6	84.9	39.9	41.4	86.8
Cascaded-GCA2	40.4	40.7	85.5	41.3	40.3	87.2
Cascaded-GCA3	38.8	42.4	85.3	40.1	40.9	86.9

Table 1: Recognition accuracy with unsupervised domain adaptation using SVM classifier (Office dataset + Caltech10). Our proposed methods (labeled GCA_{XX} and Cascaded-GCA_{XX}) perform better than all other methods in a majority of the 12 transfer learning tasks.

– **Cascaded-GCA3:** This is a new proposed method, based on finding the cascaded geometric mean of the inverse revised source matrix A_{gs} and target matrix A_t

One question that arises in proposed algorithms is how to choose the value of t in computing the weighted sharp mean. Figure 2 illustrates the variation in performance of the cascaded GCA3 method over CORAL over the range $t, \gamma \in (0.1, 0.9)$, and μ fixed for simplicity. Repeating such experiments over all 12 transfer learning tasks, Figure 3 shows the percentage improvement of the cascaded GCA3 method over correlational alignment (CORAL), using the best discovered value of all three hyperparameters using cross-validation. Figure 4 compares the performance of the proposed GCA1, GCA2, and GCA3 methods where just the t hyperparameter was varied between 0.1 and 0.9 for the Amazon to DSLR domain adaptation task. Note the variation in performance with t occurs at different points for the three points, and while their performance is

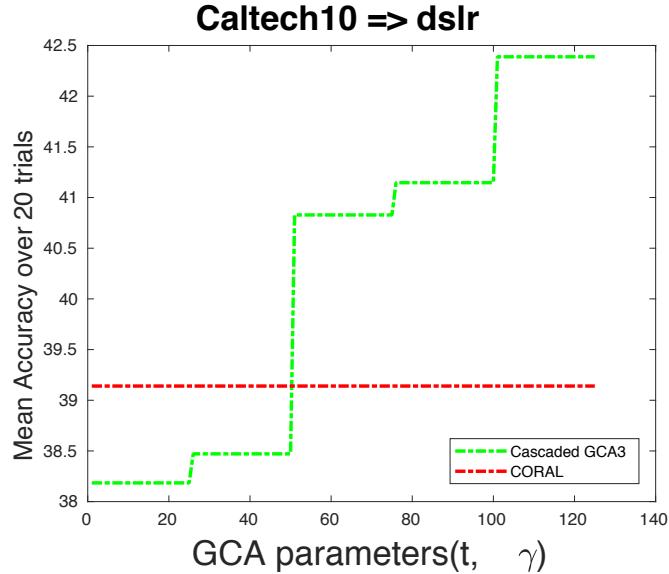


Fig. 2: Comparison of the cascaded GCA3 method with correlational alignment (CORAL). The parameters t and γ were varied between 0.1 and 0.9.

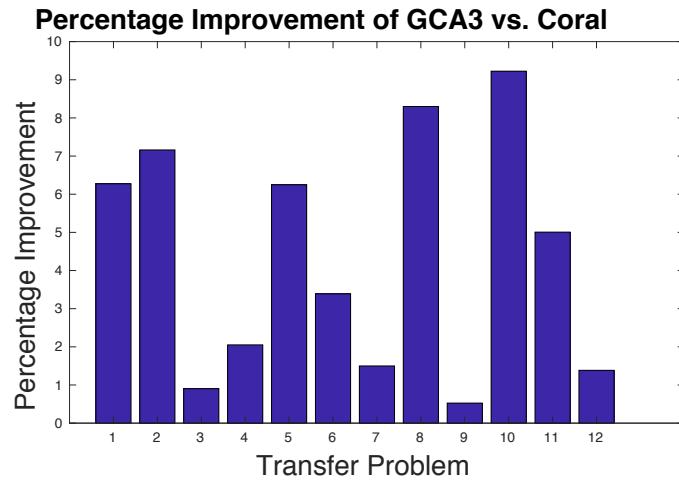


Fig. 3: Comparison of the cascaded GCA3 method with correlational alignment (CORAL) in terms of percentage improvement.

superior overall to CORAL, their relative performances at the maximum values are not very different from each other. Figure 5 once again repeats the same comparison for the Caltech10 to the Webcam domain adaptation task. As these plots clearly reveal, the values of the hyperparameters has a crucial influence on the performance of all the proposed GCAXX methods. The plot compares the performance of GCA1 to the fixed performance of the CORAL method.

9 Summary and Future Work

In this paper, we introduced a novel formulation of the classic domain adaptation problem in machine learning, based on computing the cascaded geometric mean of second order statistics from source and target domains to align them. Our approach builds on the nonlinear Riemannian geometry of the open cone of symmetric positive definite matrices (SPDs), using which the geometric mean lies along the shortest path geodesic that connects source and target covariances. Our approach has three key advantages over previous work: (a) Simplicity: The Riccati equation is a mathematically elegant solution to the domain adaptation problem, enabling integrating geometric and statistical information. (b) Theory: Our approach exploits the Riemannian geometry of SPD matrices. (c) Extensibility: As our algorithm development indicates, it is possible to easily extend our approach to capture more types of constraints, from geometrical to statistical.

There are many directions for extending our work. We briefly alluded to optimal transport theory as providing an additional theoretical justification for our solution of using the geometric mean of the source and target covariances, a link that deserves further exploration in a subsequent paper. Also, while we did not explore nonlinear variants of our approach, it is possible to extend our approach to develop a deep learning version where the gradient of the three objective functions is used to tune the weights of a multi-layer neural network. As in the case of correlational alignment (CORAL), we anticipate that the deep learning variants may perform better due to the construction of improved features of the training data. The experimental results show that the performance improvement tends to be more significant in some cases than in others.

10 Acknowledgments

Portions of this research were completed when the first and third authors were at SRI International, Menlo Park, CA and when the second author was at Amazon.com, Bangalore, India.

References

1. Adel, T., Zhao, H., Wong, A.: Unsupervised domain adaptation with a relaxed covariate shift assumption. In: AAAI. pp. 1691–1697 (2017)
2. Baktashmotagh, M., Harandi, M.T., Lovell, B.C., Salzmann, M.: Unsupervised domain adaptation by domain invariant projection. In: ICCV. pp. 769–776 (2013)

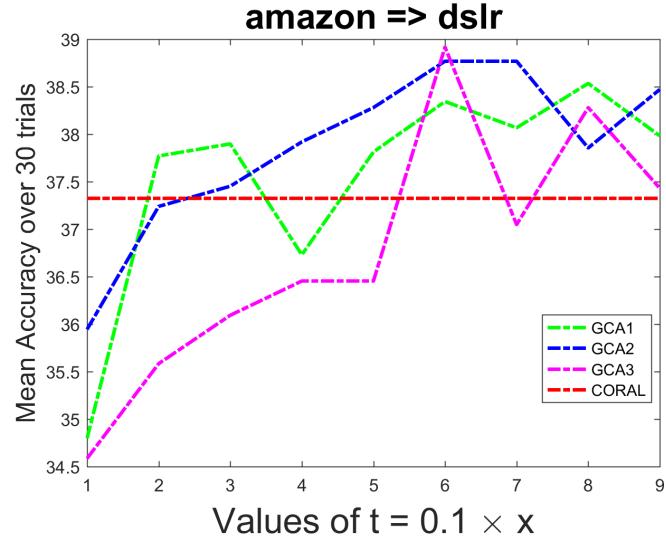


Fig. 4: Comparison of the three proposed GCA methods (GCA1, GCA2, and GCA3) with correlational alignment (CORAL).

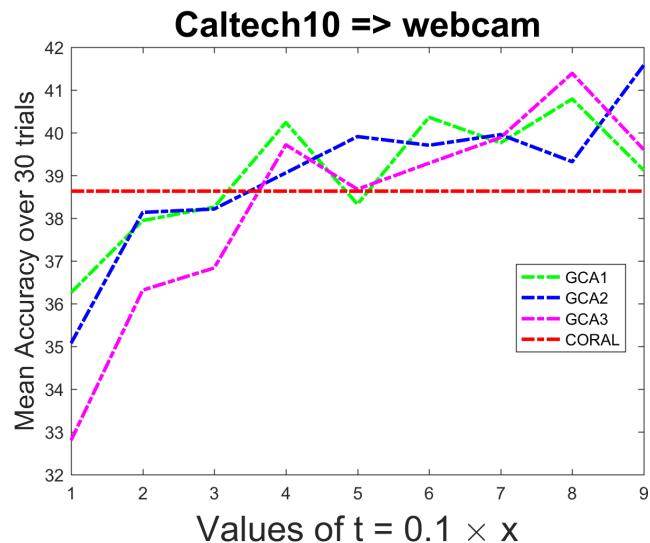


Fig. 5: Comparison of the three proposed GCA methods (GCA1, GCA2, and GCA3) with correlational alignment (CORAL).

3. Belkin, M., Niyogi, P.: Convergence of laplacian eigenmaps. In: NIPS. pp. 129–136 (2006)
4. Bellet, A., Habrard, A., Sebban, M.: Metric Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan and Claypool Publishers (2015)
5. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. In: NIPS (2006)
6. Bhatia, R.: Positive Definite Matrices. Princeton Series in Applied Mathematics, Princeton University Press, Princeton, NJ, USA (2007)
7. Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schölkopf, B., Smola, A.J.: Integrating structured biological data by kernel maximum mean discrepancy. In: ISMB (Supplement of Bioinformatics). pp. 49–57 (2006)
8. Cesa-Bianchi, N.: Learning the distribution in the extended pac model. In: ALT. pp. 236–246 (1990)
9. Courty, N., Flamary, R., Tuia, D., Rakotomamonjy, A.: Optimal transport for domain adaptation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(9), 1853–1865 (2017)
10. Cui, Z., Chang, H., Shan, S., Chen, X.: Generalized unsupervised manifold alignment. In: NIPS. pp. 2429–2437 (2014)
11. Daume, H.: Frustratingly easy domain adaptation. In: ACL (2007)
12. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20**(2), 303–353 (1998)
13. Fayek, H.M., Lech, M., Cavedon, L.: Evaluating deep learning architectures for speech emotion recognition. Neural Networks **92**, 60–68 (2017)
14. Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Subspace alignment for domain adaptation. Tech. rep., arXiv preprint arXiv:1409.5241 (2014)
15. Fukumizu, K., Bach, F.R., Gretton, A.: Statistical convergence of kernel cca. In: NIPS. pp. 387–394 (2005)
16. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: CVPR. pp. 2066–2073 (2012)
17. Hotelling, H.: Relations between two sets of variates. Biometrika **28**, 321–377 (1936)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
19. Murphy, K.P.: Machine learning : a probabilistic perspective. MIT Press, Cambridge, Mass. [u.a.] (2013)
20. Nadler, B., Lafon, S., Coifman, R.R., Kevrekidis, I.G.: Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In: NIPS. pp. 955–962 (2005)
21. Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. IEEE Trans Neural Netw **22**(2), 199–210 (Feb 2011). <https://doi.org/10.1109/TNN.2010.2091281>
22. Pan, S., Yang, Q.: A Survey on Transfer Learning. IEEE Trans Knowl Data Eng **22**(10), 1345–1359 (2010)
23. Peyré, G., Cuturi, M.: Computational Optimal Transport. ArXiv e-prints (Mar 2018)
24. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. In: AAAI. pp. 2058–2065 (2016)
25. Wang, C., Mahadevan, S.: Manifold alignment without correspondence. In: IJCAI. pp. 1273–1278 (2009)
26. Wang, H., Wang, W., Zhang, C., Xu, F.: Cross-domain metric learning based on information theory. In: AAAI. pp. 2099–2105 (2014)
27. Wang, W., Arora, R., Livescu, K., Srebro, N.: Stochastic optimization for deep cca via non-linear orthogonal iterations. In: ALLERTON. pp. 688–695 (2015)
28. Zadeh, P., Hosseini, R., Sra, S.: Geometric mean metric learning. In: ICML. pp. 2464–2471 (2016)