Dynamic Flows on Curved Space Generated by Labeled Data

Xinru Hua[1] , **Truyen Nguyen**[2] , **Tam Le**[3] , **Jose Blanchet**[1] and **Viet Anh Nguyen**[4]
[1] Stanford University, United States
[2]The University of Akron, United States
[3]The Institute of Statistical Mathematics / RIKEN AIP, Japan
[4]Chinese University of Hong Kong, China
{huaxinru, jose.blanchet}@stanford.edu, truyennguyen3@gmail.com, tam@ism.ac.jp,
nguyen@se.cuhk.edu.hk

Abstract

The scarcity of labeled data is a long-standing challenge for many machine learning tasks. We propose our gradient flow method to leverage the existing dataset (i.e., source) to generate new samples that are close to the dataset of interest (i.e., target). We lift both datasets to the space of probability distributions on the feature-Gaussian manifold, and then develop a gradient flow method that minimizes the maximum mean discrepancy loss. To perform the gradient flow of distributions on the curved feature-Gaussian space, we unravel the Riemannian structure of the space and compute explicitly the Riemannian gradient of the loss function induced by the optimal transport metric. For practical applications, we also propose a discretized flow, and provide conditional results guaranteeing the global convergence of the flow to the optimum. We illustrate the results of our proposed gradient flow method on several real-world datasets and show our method can improve the accuracy of classification models in transfer learning settings.

1 Introduction

A major challenge in many data science applications is the scarcity of labeled data. Data augmentation methods have been studied in the literature; see for example, the noise injection methods [Moreno-Barea *et al.*, 2018], generative models [Yi *et al.*, 2019], and [Shorten and Khoshgoftaar, 2019] for a survey. We consider a setting where one domain has only a few labeled samples for each class, so we cannot train a well-performing classifier with the available data. To alleviate the data scarcity problem in this setting, we propose to enrich the target dataset by generating additional labeled samples. Using generative models is not possible in our setting because they usually require more than a few samples for each class to learn and generate high-quality new samples [Gao *et al.*, 2018]. In our work, we choose a source dataset with extensive labeled data and then flow the labeled data to the target dataset. Precisely, we introduce a novel data augmentation methodology based on a gradient flow approach that minimizes the maximum mean discrepancy (MMD) distance between the target and the augmented data. Therefore, when minimizing the MMD distance, we are able to obtain an efficient scheme which generates additional labeled data from the target distribution. Our scheme is model-independent and can be applied to any datasets regardless of the number of classes or dimensionality[1].

Mathematically, we consider a feature space $\mathcal{X} = \mathbb{R}^m$ and a *categorical* label space $\mathcal{Y}$. We have a source domain dataset consisting of N samples $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \ldots, N$, and a target domain dataset of M samples $(\bar{x}_j, \bar{y}_j) \in \mathcal{X} \times \mathcal{Y}$ for $j = 1, \ldots, M (M \ll N)$. The ultimate goal of this paper is to generate new samples in the target domain, and we aim to generate new samples whose distribution is as close as possible to the distribution that governs the target domain.

We here introduce a gradient flow method [Arbel *et al.*, 2019; Mroueh *et al.*, 2019] to synthesize new, unseen data samples. Gradient flow is a continuous flow along the path where a considered loss function decreases its value. Because we have extensive source domain samples, it is possible to flow each source sample towards the target data while minimizing the loss function. The terminal product of the flow will be new samples that can sufficiently approximate the distribution of the target domain. Thus, gradient flow is an approach to synthesize new target domain samples, and is a complement to data augmentation methods, like adding random noise.

Unfortunately, formulating a gradient flow algorithm for labeled data with categorical set $\mathcal{Y}$ is problematic. Indeed, there is no clear metric structure on $\mathcal{Y}$ in order to define the topological neighborhood, this in turn leads to the difficulty of forming the gradients with respect to the categorical component. To overcome this difficulty, we lift each individual label to a richer structure. For example, a label such as "0" is replaced by a mean vector and a covariance matrix based on the whole distribution of the information associated to this particular label. Then it will be much more natural to apply gradient flow algorithms in the space of the lifted representation. A gradient flow on the dataset space with this idea was recently proposed in [Alvarez-Melis and Fusi, 2021] by leveraging a new notion of distance between datasets in [Alvarez-Melis and Fusi, 2020; Courty *et al.*, 2017; Damodaran *et al.*, 2018]. The main idea behind this approach is to reparametrize the categorical space

[1] Our code and supplementary are available at https://github.com/LucyXH/Dynamic_Flows_Curved_Space/

$\mathcal{Y}$ using the conditional distribution of the features, which is assumed to be Gaussian, and then construct a gradient flow on the feature-Gaussian space. Nevertheless, the theoretical analysis in [Alvarez-Melis and Fusi, 2021] focuses solely on the gradients with respect to the feature with no treatment of the flow with respect to the Gaussian component. In fact, the space of Gaussian distributions is not a (flat) vector space, and extracting gradient information depends on the choice of the metric imposed on this Gaussian space. On the other hand, our method computes the full gradient with respect to the Gaussian component (the mean and covariance matrix component that correspond to the label component).

Our gradient flows minimize the MMD loss function, thus it belongs to the family of MMD gradient flows that was pioneered in [Mroueh *et al.*, 2019; Arbel *et al.*, 2019], and further extended in [Mroueh and Nguyen, 2021]. The MMD function compares two distributions via their kernel mean embeddings on a *flat* reproducing kernel Hilbert space (RKHS). In contrast to the Kullback-Leibler divergence flow, the MMD flow can employ a sample approximation for the target distribution [Liu, 2017]. Further, the squared MMD possesses unbiased sample gradients [Bińkowski *et al.*, 2018; Bellemare *et al.*, 2017]. While existing literature on MMD flows focus on distributions on flat Euclidean spaces, the flow developed in our paper is for distributions on a *curved* Riemannian feature-Gaussian space. Moreover, our approach is distinctive from the flow in [Alvarez-Melis and Fusi, 2021] because we impose a specific metric on the Gaussian component, and we compute explicitly the Riemannian gradient of the MMD loss function with respect to this metric to formulate our flow. Table 1 compares our work with two recent papers on gradient flow in theory and numerical experiments.

Recently, generative models [Rezende *et al.*, 2016; Wang *et al.*, 2021] are successful in generating image samples from given distributions. The most important difference with our method is that generative models learn a prior distribution from massive data that are similar to the target data and generate new target samples conditioning on the prior distribution [Wang *et al.*, 2020; Gao *et al.*, 2018]. Comparatively, our algorithm can transfer between two non-similar and non-related distributions, for example, from random Gaussian noise to MNIST in Supplementary B.8. Another benefit of our method is that we provide conditions for global convergence of our algorithms in Section 4, whereas generative models or more specific, generative adversarial networks (GANs), currently do not guarantee global convergence [Wiatrak *et al.*, 2019].

The application of our gradient flow is few-shot transfer learning, where we want to train classifiers with limited labeled data in the target domain. The numerical experiments in Section 5 demonstrate that our gradient flows can effectively augment the target data, and thus can significantly boost the accuracy in the classification task in the few-shot learning setting. Moreover, we run experiments on Tiny ImageNet datasets to highlight that our algorithm is scalable to higher-dimensional image data, that is higher than recent gradient flow works [Alvarez-Melis and Fusi, 2021; Fan and Alvarez-Melis, 2022]. We also compare our method with [Alvarez-Melis and Fusi, 2021], mixup method [Zhang *et al.*, 2017], and traditional data augmentation methods in Supplementary B.7, which show that our method improves the accuracy in transfer learning more than these methods.

Some works study nonparametric gradient flows using the 2-Wasserstein distance between distributions [Ambrosio *et al.*, 2008; Jordan *et al.*, 1998; Otto, 2001; Villani, 2008; Santambrogio, 2015; Santambrogio, 2017; Frogner and Poggio, 2020], but only for distributions on Euclidean spaces and different metrics. Nonparametric gradient flows with other metrics include Sliced-Wasserstein Descent [Liutkus *et al.*, 2019], Stein Descent [Liu, 2017; Liu and Wang, 2016], and Sobolev Descent [Mroueh *et al.*, 2019], but only for distributions on Euclidean spaces. In particular, [Liu, 2017] introduce Riemannian structures for the Stein geometry on flat spaces, while ours is on a curved space. Parametric flows for training GANs are studied in [Chizat and Bach, 2018; Arbel *et al.*, 2020; Mroueh and Nguyen, 2021].

Contributions. We study a gradient flow approach to synthesize new labeled samples related to the target domain. To construct this flow, we consider the space of probability distributions on the feature-Gaussian manifold, and we are metrizing this space with an optimal transport distance. We summarize the contributions of this paper as follows.

- We study in details the Riemannian structure of the feature-Gaussian manifold in Section 3, as well as the Riemannian structure of the space of probability measures supported on this manifold in Supplementary A.1.
- We consider a gradient flow that minimizes the squared MMD loss function to the target distribution. We describe explicitly the (Riemannian) gradient of the squared MMD in Lemma 5, and we provide a partial differential equation describing the evolution of the gradient flow that follows the (Riemannian) steepest descent direction.
- We propose two discretized schemes to approximate the continuous gradient flow equation in Section 4.1 and 4.2. We provide conditions guaranteeing the global convergence of our gradient flows to the optimum in both schemes.
- In Section 5, we demonstrate numerical results with our method on real-world image datasets. We show that our method can generate high-fidelity images and improve the classification accuracy in transfer learning settings.

Notations. We use $\mathbb{S}^n$ to denote the set of $n \times n$ real and symmetric matrices, and $\mathbb{S}^n_{++} \subset \mathbb{S}^n$ consists of all positive definite matrices. For $A \in \mathbb{S}^n$, $\mathrm{tr}(A) := \sum_i A_{ii}$. We use $\langle \cdot, \cdot \rangle$ and $\| \cdot \|_2$ to denote the standard inner product and norm on Euclidean spaces. Let $\mathcal{P}(X)$ be the collection of all probability distributions with finite second moment on metric space X. If $\varphi : X \to Y$ is a Borel map and $\nu \in \mathcal{P}(X)$, then the push-forward $\varphi_{\#}\nu$ is the distribution on Y given by $\varphi_{\#}\nu(E) = \nu(\varphi^{-1}(E))$ for all Borel sets $E \subset Y$. For a function f of the continuous time variable t, f_t denotes the value of f at t while $\partial_t f$ denotes the standard derivative of f w.r.t. t. Also, δ_z denotes the Dirac delta measure at z.

All proofs are provided in the Supplementary material.

Paper	Dataset	On curved Riemannian space	Gradient has mean and covariance component
[Alvarez-Melis and Fusi, 2021]	synthetic, *NIST, and CIFAR10	✗	✗
[Arbel *et al.*, 2019]	synthetic	✗	✗
Ours	synthetic, *NIST, and TinyImageNet	✓	✓

Table 1: To the best of our knowledge, we provide the *first* results on the full gradient of the features and lifted labels on a curved Riemannian space. We also conduct numerical experiments on the highest-dimension real-world datasets.

2 Labeled Data Synthesis via Gradient Flows of Lifted Distributions

In this section, we describe our approach to synthesize target domain samples using gradient flows. A holistic view of our method is presented in Fig. 1.

In the first step, we would need to lift the feature-label space $\mathcal{X} \times \mathcal{Y}$ to a higher dimensional space where a metric can be defined. Consider momentarily the source data samples $(x_i, y_i)_{i=1}^N$. Notice that this data can be represented as an empirical distribution ν on $\mathcal{X} \times \mathcal{Y}$. More precisely, we have $\nu = N^{-1}\sum_{i=1}^N \delta_{(x_i,y_i)}$. As $\mathcal{Y}$ is discrete, the law of conditional probabilities allows us to dis-integrate ν into the conditional distributions ν_y of $X|Y = y$ satisfying $\nu(E \times F) = \int_F \nu_y(E)\nu^2(\mathrm{d}y)$ for every $E \subset \mathcal{X}$ and $F \subset \mathcal{Y}$, where $\nu^2 := N^{-1}\sum_{i=1}^N \delta_{y_i}$ is the second marginal of ν [Ambrosio *et al.*, 2008, Theorem 5.3.1]. The lifting procedure is obtained by employing a pre-determined mapping $\phi : \mathcal{X} \to \mathbb{R}^n$, and any categorical value $y \in \mathcal{Y}$ can now be represented as an n-dimensional distribution $\phi_\# \nu_y$. Using this lifting, any source sample $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ is lifted to a point $(x_i, \phi_\# \nu_{y_i}) \in \mathcal{X} \times \mathcal{P}(\mathbb{R}^n)$ and the source dataset is representable as an empirical distribution of the form $N^{-1}\sum_{i=1}^N \delta_{(x_i, \phi_\# \nu_{y_i})}$.

The lifted representation of a categorical value $y \in \mathcal{Y}$ as an n-dimensional distribution $\phi_\#\nu_y \in \mathcal{P}(\mathbb{R}^n)$ is advantageous because $\mathcal{P}(\mathbb{R}^n)$ is metrizable, for example, using the 2-Wasserstein distance. The downside is that $\mathcal{P}(\mathbb{R}^n)$ is infinite dimensional, and encoding the datasets in this lifted representation is not efficient. To resolve this issue, we assume that $\phi_\#\nu_y$ is Gaussian for all $y \in \mathcal{Y}$, and thus any distribution $\phi_\#\nu_y$ can be characterized by the mean vector $\mu_y \in \mathbb{R}^n$ and covariance matrix $\Sigma_y \in \mathbb{S}_{++}^n$ defined as $\mu_y = \int_\mathcal{X} \phi(x)\nu_y(\mathrm{d}x)$ and $\Sigma_y = \int_\mathcal{X} \big[\phi(x) - \mu_y\big]\big[\phi(x) - \mu_y\big]^\top \nu_y(\mathrm{d}x)$ for all $y \in \mathcal{Y}$, where $^\top$ denotes the transposition of a vector. In real-world settings, the conditional moments of $\phi(X)|Y$ are sufficiently different for $y \neq y'$, and thus the representations using (μ_y, Σ_y) will unlikely lead to any loss of label information. With this lifting, the source data thus can be represented as an empirical distribution ρ^0 on $\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{S}_{++}^n$ via $\rho^0 = N^{-1}\sum_{i=1}^N \delta_{(x_i, \mu_{y_i}, \Sigma_{y_i})}$. By an analogous construction to compute $\bar{\mu}_y$ and $\bar{\Sigma}_y$ using the target data, the target domain data $(\bar{x}_j, \bar{y}_j)_{j=1}^M$ can be represented as another empirical distribution $\varrho = M^{-1}\sum_{j=1}^M \delta_{(\bar{x}_j, \bar{\mu}_{\bar{y}_j}, \bar{\Sigma}_{\bar{y}_j})}$. Let us denote the shorthand $\mathcal{Z} = \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{S}_{++}^n$, then ρ^0 and ϱ are both probability measures on $\mathcal{Z}$. We refer to ρ^0 and ϱ as the feature-Gaussian representations of the source and target datasets.

We now consider the gradient flow associated with the optimization problem

$$\min_{\rho \in \mathcal{P}(\mathcal{Z})} \left\{ \mathcal{F}(\rho) := \frac{1}{2}\mathrm{MMD}(\rho, \varrho)^2 \right\}$$

under the initialization $\rho = \rho^0$. The objective function $\mathcal{F}(\rho)$ quantifies how far an incumbent solution ρ is from the target distribution ϱ, measured using the MMD distance. In Sections 3 and 4, we will provide the necessary ingredients to construct this flow.

Suppose that after T iterations of the discretized gradient flow algorithm, we obtain a distribution $\rho^T \in \mathcal{P}(\mathcal{Z})$ that is sufficiently close to ϱ, i.e., $\mathcal{F}(\rho^T)$ is close to zero. Then we can recover new target labels by projecting the samples of the distribution ρ^T to the locations on $\mathcal{X} \times \mathcal{Y}$. This projection can be computed efficiently by solving a linear optimization problem, as discussed in Supplementary B.3.

Remark 1 (Reduction of dimensions)**.** *If $m = n$ and ϕ is the identity map, then our lifting procedure coincides with that proposed in [Alvarez-Melis and Fusi, 2020]. However, a large n is redundant, especially when the cardinality of $\mathcal{Y}$ is low. If $n \ll m$, then ϕ offers significant reduction in the number of dimensions, and will speed up the gradient flow algorithms.*

Remark 2 (Generalization to elliptical distributions)**.** *Our framework can be extended to the symmetric elliptical distributions because the Bures distance for elliptical distributions admits the same closed-form as for the Gaussian distributions [Gelbrich, 1990]. In this paper, we use ϕ as the t-SNE embedding. According to [van der Maaten and Hinton, 2008], t-SNE's low-dimensional embedded space forms a Student-t distribution, which is an elliptical distribution.*

3 Riemannian Geometry of $\mathcal{Z}$ and $\mathcal{P}(\mathcal{Z})$

If we opt to measure the distance between two Gaussian distributions using the 2-Wasserstein metric, then this choice would induce a natural distance d on the space $\mathcal{Z} = \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{S}_{++}^n$ prescribed as

$$\begin{aligned} &d\big((x_1, \mu_1, \Sigma_1), (x_2, \mu_2, \Sigma_2)\big) \\ &\quad := \big[\|x_1 - x_2\|_2^2 + \|\mu_1 - \mu_2\|_2^2 + \mathbb{B}(\Sigma_1, \Sigma_2)^2\big]^{\frac{1}{2}}, \quad (3.1) \end{aligned}$$

where $\mathbb{B}$ is the Bures metric on $\mathbb{S}_{++}^n$ given by $\mathbb{B}(\Sigma_1, \Sigma_2) := \big[\mathrm{tr}(\Sigma_1 + \Sigma_2 - 2[\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}]^{\frac{1}{2}})\big]^{\frac{1}{2}}$.

As $\mathbb{B}$ is a metric on $\mathbb{S}_+^n$ [Bhatia *et al.*, 2019, p.167], d is hence a product metric on $\mathcal{Z}$. In this section, first, we study the non-Euclidean geometry of $\mathcal{Z}$ under the ground

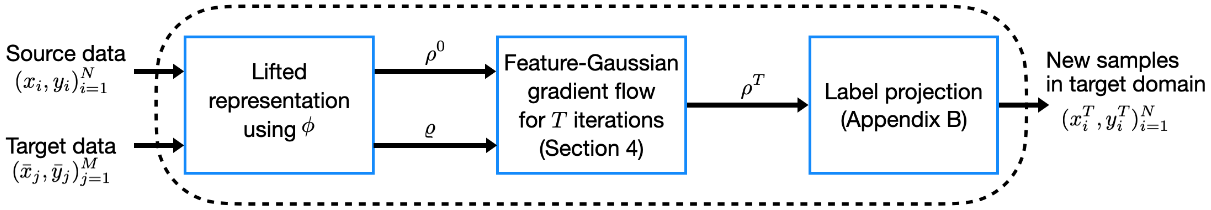
Source data
$(x_i, y_i)_{i=1}^N$
Target data
$(\bar{x}_j, \bar{y}_j)_{j=1}^M$
Lifted representation using ϕ
ρ^0
ϱ
Feature-Gaussian gradient flow for T iterations (Section 4)
ρ^T
Label projection (Appendix B)
New samples in target domain
$(x_i^T, y_i^T)_{i=1}^N$

Figure 1: Schematic view of our approach: The source and target datasets are first lifted to distributions ρ^0 and ϱ on the feature-Gaussian space (left box). We then run a gradient flow for T iterations to get a terminal distribution ρ^T (middle). Atoms of ρ^T are projected to get labeled target samples (right).

metric d. Second, we investigate the Riemannian structure on $\mathcal{P}(\mathcal{Z})$, the space of all distributions supported on $\mathcal{Z}$ and with finite second moment, that is induced by the optimal transport distance. These Riemannian structures are required to define the Riemannian gradients of any loss functionals on $\mathcal{P}(\mathcal{Z})$, and will remain important in our development of the gradient flow for the squared MMD.

The space $\mathcal{Z}$ is not a linear vector space. In this section, we reveal the Riemannian structure on $\mathcal{Z}$ associated to the ground metric d. As we shall see, $\mathcal{Z}$ is a curved space as its geodesics are not straight lines and involve solutions to the Lyapunov equation. For any positive definite matrix $\Sigma \in \mathbb{S}^n_{++}$ and any symmetric matrix $V \in \mathbb{S}^n$, the Lyapunov equation

$$H\Sigma + \Sigma H = V \tag{3.2}$$

has a unique solution $H \in \mathbb{S}^n$ [Bhatia, 1997, Theorem VII.2.1]. Let $\mathrm{L}_\Sigma[V]$ denote this unique solution H.

The space $\mathbb{S}^n_{++}$ is a Riemannian manifold with the Bures metric $\mathbb{B}$ as the associated distance function, see [Takatsu, 2011, Proposition A]. Since $\mathcal{Z}$ is the product of two Euclidean spaces and $\mathbb{S}^n_{++}$, this gives rise to the following geometric structure for $\mathcal{Z}$.

Proposition 3 (Geometry of $\mathcal{Z}$). *The space $\mathcal{Z}$ is a Riemannian manifold: at each point $z = (x, \mu, \Sigma) \in \mathcal{Z}$, the tangent space is $\mathrm{T}_z\mathcal{Z} = \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{S}^n$ and the Riemannian metric is*

$$\begin{aligned}\big\langle (w_1, v_1, V_1), (w_2, v_2, V_2) \big\rangle_z& \\ := \langle w_1, w_2\rangle + \langle v_1, v_2\rangle &+ \langle V_1, V_2\rangle_\Sigma\end{aligned} \tag{3.3}$$

for two tangent vectors (w_1, v_1, V_1) and (w_2, v_2, V_2) in $\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{S}^n$, where $\langle V_1, V_2\rangle_\Sigma := tr\Big(\mathrm{L}_\Sigma[V_1]\,\Sigma\,\mathrm{L}_\Sigma[V_2]\Big)$. Moreover, the distance function corresponding to this Riemannian metric coincides with the distance d given by (3.1).

As $\mathcal{Z}$ is a product Riemannian manifold, any geodesic in $\mathcal{Z}$ is of the form (θ, γ, Γ) with θ, γ being the Euclidean geodesics (straight lines) and Γ being a geodesic in the Riemannian manifold $\mathbb{S}^n_{++}$. More precisely, for each $\Sigma \in \mathbb{S}^n_{++}$ and each tangent vector $V \in \mathbb{S}^n$, the geodesic in the manifold $\mathbb{S}^n_{++}$ emanating from Σ with direction V is given by

$$\Gamma(t) = (I + t\mathrm{L}_\Sigma[V])\Sigma(I + t\mathrm{L}_\Sigma[V]) \quad \text{for } t \in J^*, \tag{3.4}$$

where J^* is the open interval about the origin given by $J^* = \{t \in \mathbb{R} : I + t\mathrm{L}_\Sigma[V] \in \mathbb{S}^n_{++}\}$ [Malagò *et al.*, 2018]. As a consequence, for each point $(x, \mu, \Sigma) \in \mathcal{Z}$ and each tangent vector $(w, v, V) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{S}^n$, the Riemannian exponential map in $\mathcal{Z}$ for $t \in J^*$ is given by

$$\exp_{(x,\mu,\Sigma)}(t(w, v, V)) := (\theta(t), \gamma(t), \Gamma(t)). \tag{3.5}$$

where $\theta(t) := x + tw$, $\gamma(t) := \mu + tv$, and $\Gamma(t)$ is defined by (3.4). By definition, $t \mapsto \exp_{(x,\mu,\Sigma)}(t(w, v, V))$ is the geodesic emanating from (x, μ, Σ) with direction (w, v, V).

Given the Riemannian metric (3.3), one can define the corresponding notion of gradient and divergence [Lee, 2003]. For a differentiable function $\varphi : \mathcal{Z} \to \mathbb{R}$, its gradient $\widetilde{\nabla}_d\varphi(z)$ w.r.t. the metric d defined by (3.1) is the unique element in the tangent space $\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{S}^n$ satisfying

$$\big\langle \widetilde{\nabla}_d\varphi(z), (w, v, V)\big\rangle_z = D\varphi_z(w, v, V)$$

for all $(w, v, V) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{S}^n$ with $D\varphi_z(w, v, V)$ denoting the standard directional derivative of φ at z in the direction (w, v, V). By exploiting the special form of $\langle \cdot, \cdot\rangle_z$ in (3.3), we can compute $\widetilde{\nabla}_d\varphi(z)$ explicitly:

Lemma 4 (Gradients). *For a differentiable function $\varphi : \mathcal{Z} \to \mathbb{R}$, we have for $z = (x, \mu, \Sigma)$ that*

$$\widetilde{\nabla}_d\varphi(z) = \big(\nabla_x\varphi(z),\ \nabla_\mu\varphi(z),\ 2[\nabla_\Sigma\varphi(z)]\Sigma + 2\Sigma[\nabla_\Sigma\varphi(z)]\big), \tag{3.6}$$

where $(\nabla_x, \nabla_\mu, \nabla_\Sigma)$ are the standard (Euclidean) gradients of the respective components.

The last component in formula (3.6) for $\widetilde{\nabla}_d\varphi$ reflects the curved geometry of $\mathcal{Z}$, and can be interpreted as the Riemannian gradient of the function $\Sigma \mapsto \varphi(x, \mu, \Sigma)$ w.r.t. the Bures distance $\mathbb{B}$.

For a continuous vector field $\Phi : \mathcal{Z} \to \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{S}^n$ and a distribution $\rho \in \mathcal{P}(\mathcal{Z})$, the divergence $\mathrm{div}_d(\rho\Phi)$ is the signed measure on $\mathcal{Z}$ satisfying the integration by parts formula

$$\int_{\mathcal{Z}} \varphi(z)\,\mathrm{div}_d(\rho\Phi)(\mathrm{d}z) = -\int_{\mathcal{Z}} \langle \Phi(z), \widetilde{\nabla}_d\varphi(z)\rangle_z\ \rho(\mathrm{d}z)$$

for every differentiable function $\varphi : \mathcal{Z} \to \mathbb{R}$ with compact support. In case ρ has a density w.r.t. the Riemannian volume form on $\mathcal{Z}$, then this definition coincides with the standard divergence operator induced by Riemannian metric (3.3). The optimal transport distance and its induced Riemannian metric on the space $\mathcal{P}(\mathcal{Z})$ are relegated to Supplementary A.1.

4 Gradient Flow for Maximum Mean Discrepancy

As $\mathcal{P}(\mathcal{Z})$ is an infinite dimensional curved space, many machine learning methods based on finite dimensional or linear structure cannot be directly applied to this manifold. To circumvent this problem, we use a positive definite kernel to map $\mathcal{P}(\mathcal{Z})$ to a RKHS and then perform our analysis on it. Let k be a positive definite kernel on $\mathcal{Z}$, and let $\mathcal{H}$ be the RKHS generated by k. The inner product on $\mathcal{H}$ is denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, and the kernel mean embedding $\rho \in \mathcal{P}(\mathcal{Z}) \longmapsto \mathbf{m}_\rho(\cdot) \in \mathcal{H}$ is given by $\mathbf{m}_\rho(z) := \int_{\mathcal{Z}} k(z, w)\, \rho(\mathrm{d}w)$ for z in $\mathcal{Z}$. The MMD [Gretton *et al.*, 2012] between $\rho \in \mathcal{P}(\mathcal{Z})$ and the target ϱ is defined as the maximum of the mean difference between the two distributions over all test functions in the unit ball of $\mathcal{H}$ (see Supplementary A.3). Moreover, it can be expressed by $\mathrm{MMD}(\rho, \varrho) = \|\mathbf{m}_\rho - \mathbf{m}_\varrho\|_{\mathcal{H}}$. When k is characteristic, the kernel mean embedding $\rho \mapsto \mathbf{m}_\rho$ is injective and therefore, $\mathrm{MMD}(\rho, \varrho) = 0$ if and only if $\rho = \varrho$.

Consider the loss function $\mathcal{F}[\rho] := \frac{1}{2}\mathrm{MMD}(\rho, \varrho)^2 = \frac{1}{2}\|\mathbf{m}_\rho - \mathbf{m}_\varrho\|_{\mathcal{H}}^2$. As explained in the introduction, there are three advantages of MMD over Kullback-Leibler divergence: its associated gradient flow can employ a sample approximation for the target distribution, the input distribution ρ does not have to be absolutely continuous w.r.t. the target distribution ϱ, and the squared MMD possesses unbiased sample gradients. For each ρ, the Riemannian gradient $\mathrm{grad}\,\mathcal{F}[\rho]$ is defined as the unique element in $\mathrm{T}_\rho\mathcal{P}(\mathcal{Z})$ satisfying $g_\rho(\mathrm{grad}\,\mathcal{F}[\rho], \zeta) = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}\mathcal{F}[\rho_t]$ for every differentiable curve $t \mapsto \rho_t \in \mathcal{P}(\mathcal{Z})$ passing through ρ at $t = 0$ with tangent vector $\partial_t \rho_t|_{t=0} = \zeta$. By using the Riemannian metric tensor (eq. A.3), we can compute explicitly this gradient.

Lemma 5 (Gradient formula)**.** *The Riemannian gradient of* $\mathcal{F}$ *satisfies* $\mathrm{grad}\,\mathcal{F}[\rho] = -\mathrm{div}_d\left(\rho\widetilde{\nabla}_d[\mathbf{m}_\rho - \mathbf{m}_\varrho]\right)$.

The Riemannian gradient $\mathrm{grad}\,\mathcal{F}$ on $\mathcal{P}(\mathcal{Z})$ depends not only on the gradient operator $\widetilde{\nabla}_d$ but also on the divergence operator. Using Lemma 5, we can rewrite the gradient flow equation $\partial_t \rho_t = -\mathrm{grad}\,\mathcal{F}[\rho_t]$ explicitly as

$$\partial_t \rho_t = \mathrm{div}_d\big(\rho_t \widetilde{\nabla}_d[\mathbf{m}_{\rho_t} - \mathbf{m}_\varrho]\big) \quad \text{for} \quad t \geq 0. \tag{4.1}$$

The next result exhibits the rate at which $\mathcal{F}$ decreases its value along the flow.

Proposition 6 (Rate of decrease)**.** *Along the gradient flow* $t \mapsto \rho_t \in \mathcal{P}(\mathcal{Z})$ *given by* (4.1)*, we have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}[\rho_t] = -\int_{\mathcal{Z}} \left\|\widetilde{\nabla}_d[\mathbf{m}_{\rho_t} - \mathbf{m}_\varrho]\right\|_z^2 \rho_t(\mathrm{d}z) \quad \textit{for} \quad t \geq 0.$$

Proposition 6 implies that $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}[\rho_t] = 0$ if and only if $\widetilde{\nabla}_d[\mathbf{m}_{\rho_t} - \mathbf{m}_\varrho](z) = 0$ for every z in the support of the distribution ρ_t. Thus, the objective function will decrease whenever the gradient $\widetilde{\nabla}_d[\mathbf{m}_{\rho_t} - \mathbf{m}_\varrho]$ is not identically zero.

4.1 Riemannian Forward Euler Scheme

We propose the Riemannian version of the forward Euler scheme to discretize continuous flow (4.1):

$$\boxed{\begin{aligned} &\rho^{\tau+1} = \exp(s_\tau \Phi^\tau)_\# \rho^\tau \\ &\text{with } \Phi^\tau := -\widetilde{\nabla}_d[\mathbf{m}_{\rho^\tau} - \mathbf{m}_\varrho], \end{aligned} \qquad (4.2)}$$

where $s_\tau > 0$ is the step size. Here, for a vector field $\Phi = (\Phi_1, \Phi_2, \Phi_3) : \mathcal{Z} \to \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{S}^n$ and for $\varepsilon \geq 0$, $\exp(\varepsilon\Phi) : \mathcal{Z} \to \mathcal{Z}$ is the Riemannian exponential map induced by (3.5), i.e., for $z = (x, \mu, \Sigma) \in \mathcal{Z}$:

$$\exp_z(\varepsilon\Phi(z)) = \begin{pmatrix} x + \varepsilon\Phi_1(z) \\ \mu + \varepsilon\Phi_2(z) \\ (I + \varepsilon\mathrm{L}_\Sigma[\Phi_3(z)])\Sigma(I + \varepsilon\mathrm{L}_\Sigma[\Phi_3(z)]) \end{pmatrix}.$$

Notice in the above equation that the input z affects simultaneously the bases of the exponential map $\exp_z$ as well as the direction $\Phi(z)$. This map is the ε-perturbation of the identity map along geodesics with directions Φ. When $\rho^\tau = N^{-1}\sum_{i=1}^N \delta_{z_i^\tau}$ is an empirical distribution, scheme (4.2) flows each particle z_i^τ to the new position $z_i^{\tau+1} = \exp_{z_i^\tau}(s_\tau \Phi(z_i^\tau))$. The next lemma shows that Φ^τ is the steepest descent direction for $\mathcal{F}$ w.r.t. the exponential map among all directions in the space $\mathbb{L}^2(\rho^\tau)$, which is the collection of all vector fields Φ on $\mathcal{Z}$ satisfying $\|\Phi\|_{\mathbb{L}^2(\rho^\tau)}^2 := \int_{\mathcal{Z}} \|\Phi(z)\|_z^2 \rho^\tau(\mathrm{d}z) < \infty$.

Lemma 7 (Steepest descent direction)**.** *Fix a distribution* $\rho^\tau \in \mathcal{P}(\mathcal{Z})$*. For any vector field* $\Phi : \mathcal{Z} \to \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{S}^n$*, we have*

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0}\mathcal{F}[\exp(\varepsilon\Phi)_\#\rho^\tau] = \int_{\mathcal{Z}} \langle \widetilde{\nabla}_d[\mathbf{m}_{\rho^\tau} - \mathbf{m}_\varrho](z), \Phi(z)\rangle_z\, \rho^\tau(\mathrm{d}z).$$

If $\hat{\Phi}^\tau$ *is the unit vector field (w.r.t. the* $\|\cdot\|_{\mathbb{L}^2(\rho^\tau)}$ *norm) in the direction of* Φ^τ *given in* (4.2)*, then*

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\big|_{\varepsilon=0}\mathcal{F}[\exp(\varepsilon\hat{\Phi}^\tau)_\#\rho^\tau] = -\|\widetilde{\nabla}_d[\mathbf{m}_{\rho^\tau} - \mathbf{m}_\varrho]\|_{\mathbb{L}^2(\rho^\tau)}$$

and this is the fastest decay rate among all unit directions Φ *in* $\mathbb{L}^2(\rho^\tau)$.

It follows from Lemma 7 that the discrete scheme (4.2) satisfies the Riemannian gradient descent property: if $\widetilde{\nabla}_d[\mathbf{m}_{\rho^\tau} - \mathbf{m}_\varrho]$ is nonzero and if $s_\tau > 0$ is chosen sufficiently small, then $\mathcal{F}[\rho^{\tau+1}] < \mathcal{F}[\rho^\tau]$. In Proposition 14 in the Supplementary, we quantify the amount of decrease of $\mathcal{F}$ at each iteration. Algorithm 1 implements the flow (4.2) iteratively. Each iteration in Algorithm 1 has complexity $O(N(Nm + n^3))$, where m is the feature's dimension, n is the reduced dimension ($n \ll m$), N is the number of particles.

Convergence. We now study the (weak) convergence of the solution ρ_t of the continuous gradient flow (4.1), as well as the discretized counterpart ρ^τ of flow (4.2), to the target distribution ϱ. When the kernel k is characteristic, this convergence is equivalent to $\lim_{t\to\infty}\mathrm{MMD}(\rho_t, \varrho) = 0$. Because the objective function $\mathcal{F}$ is not displacement convex [Arbel *et al.*, 2019, Section 3.1], the convergent theory for gradient flows in [Ambrosio *et al.*, 2008] does not apply even in the case of Euclidean spaces. In general, there is a possibility that $\mathrm{MMD}(\rho_t, \varrho)$ does not decrease to zero as $t \to \infty$. In view of Proposition 6, this happens if the solutions ρ_t are trapped inside the set $\big\{\rho : \int_{\mathcal{Z}} \big\|\widetilde{\nabla}_d[\mathbf{m}_\rho - \mathbf{m}_\varrho]\big\|_z^2 \rho(\mathrm{d}z) = 0\big\}$. For each distribution ρ on $\mathcal{Z}$, we define in Supplementary A.3 a symmetric linear and positive operator $\mathbb{K}_\rho : \mathcal{H} \to \mathcal{H}$ with the property that $\langle \mathbb{K}_\rho[\mathbf{m}_\rho - \mathbf{m}_\varrho], \mathbf{m}_\rho - \mathbf{m}_\varrho\rangle_{\mathcal{H}} = \int_{\mathcal{Z}} \big\|\widetilde{\nabla}_d[\mathbf{m}_\rho -$

Algorithm 1 Discretized Gradient Flow Algorithm for Scheme (4.2)

```
Input: a source distribution $\rho^0 = N^{-1}\sum_{i=1}^N \delta_{z_i^0}$, a target distribution $\varrho = M^{-1}\sum_{j=1}^M \delta_{\bar{z}_j}$, a number of iterations $T$, a sequence of step sizes $s_\tau > 0$ with $\tau = 0, 1, ..., T$ and a kernel $k$
Initialization: Compute $\bar{\Psi}(z) = M^{-1}\sum_{j=1}^M \widetilde{\nabla}_d^1 k(z, \bar{z}_j)$ with $\widetilde{\nabla}_d^1 k(z, \bar{z}_j)$ is $\widetilde{\nabla}_d$ of $z \mapsto k(z, \bar{z}_j)$
repeat for each $\tau = 0, \ldots, T-1$:
    Compute $\Psi^\tau(z) = N^{-1}\sum_{i=1}^N \widetilde{\nabla}_d^1 k(z, z_i^\tau)$
    for $i = 1, \ldots, N$
        do $z_i^{\tau+1} \leftarrow \exp_{z_i^\tau}\big(s_\tau(\bar{\Psi} - \Psi^\tau)(z_i^\tau)\big)$
    end for
Output: $\rho^T = N^{-1}\sum_{i=1}^N \delta_{z_i^T}$
```

$\mathbf{m}_\varrho]\|_z^2\, \rho(\mathrm{d}z)$. We further show in Proposition 16 that ρ_t globally converges in MMD if the minimum eigenvalue λ_t of the operator $\mathbb{K}_{\rho_t}$ satisfies an integrability condition.

4.2 Noisy Riemannian Forward Euler Scheme

The analysis in Section 4.1 reveals that the gradient flows suffer from convergence issues if the residual $\mathbf{m}_{\rho_t} - \mathbf{m}_\varrho$ belongs to the null space of the operator $\mathbb{K}_{\rho_t}$. To resolve this, we employ graduated optimization [Arbel *et al.*, 2019; Gulcehre *et al.*, 2016; Gulcehre *et al.*, 2017; Hazan *et al.*, 2016] used for non-convex optimization in Euclidean spaces. Specifically, we modify algorithm (4.2) by injecting Gaussian noise into the exponential map at each iteration τ to obtain

$$\boxed{\begin{aligned}&\rho^{\tau+1} = \exp(s_\tau \Phi^\tau)_{\#}\rho^{\tau,\beta_\tau} \qquad (4.3)\\ &\text{with } f^{\beta_\tau}\!:\!(z,u) \mapsto \exp_z(\beta_\tau u),\ \rho^{\tau,\beta_\tau}\!:=\!f^{\beta_\tau}{}_{\#}(\rho^\tau \otimes g).\end{aligned}}$$

Here g is a Gaussian measure with distribution $\mathcal{N}_{\mathbb{R}^m}(0,1) \otimes \mathcal{N}_{\mathbb{R}^n}(0,1) \otimes \mathcal{N}_{\mathbb{S}^n}(0,1)$ on the tangent space and $\mathcal{N}_{\mathbb{S}^n}(0,1)$ denotes an n-by-n symmetric matrix whose upper triangular elements are i.i.d. standard Gaussian random variables. When $\rho^\tau = N^{-1}\sum_{i=1}^N \delta_{z_i^\tau}$, scheme (4.3) flows each particle z_i^τ first to $z_i^{\tau,\beta_\tau} := \exp_{z_i^\tau}(\beta_\tau U)$ with noise $U \sim g$ and then to $z_i^{\tau+1} = \exp_{z_i^{\tau,\beta_\tau}}(s_\tau \Phi(z_i^{\tau,\beta_\tau}))$. Our next result extends Proposition 8 in [Arbel *et al.*, 2019] for the standard quadratic cost on the Euclidean space to the nonstandard cost function d^2 on the *curved* Riemannian manifold $\mathcal{Z}_{++}$. It demonstrates that scheme (4.3) achieves the global minimum of $\mathcal{F}$ provided that k is a Lipschitz-gradient kernel and both the noise level β_τ and the step size s_τ are well controlled. The proof of Proposition 8 is given in Supplementary A.3 and relies on arguments that are different from that of [Arbel *et al.*, 2019].

Proposition 8 (Objective value decay for noisy scheme)**.** *Suppose that k is a Lipschitz-gradient kernel*[2] *with constant L, and the noise level β_τ satisfies*

$$\lambda\beta_\tau^2\mathcal{F}[\rho^\tau] \le \int_{\mathcal{Z}} \|\Phi^\tau(z)\|_z^2\, \rho^{\tau,\beta_\tau}(\mathrm{d}z) \qquad (4.4)$$

for some constant $\lambda > 0$. Then for $\rho^{\tau+1}$ obtained from scheme (4.3)*, we have*

$$\mathcal{F}[\rho^{\tau+1}] \le \mathcal{F}[\rho^0]\, \exp\Big(-\lambda\sum\nolimits_{i=0}^{\tau}\big[s_i\big(1-2Ls_i\big)\beta_i^2\big]\Big).$$

In particular, $\mathcal{F}[\rho^\tau]$ tends to zero if the sequence $\sum_{i=0}^\tau s_i\big(1-2Ls_i\big)\beta_i^2$ goes to positive infinity. For an adaptive step size $s_\tau \le 1/4L$, this condition is met if, for example, β_τ is chosen of the form $(\tau s_\tau)^{-\frac{1}{2}}$ while still satisfying (4.4). The noise perturbs the direction of descent, whereas the step size determines how far to move along this perturbed direction. The noise level needs to be adjusted so that the gradient is not too blurred, but it does not necessarily decrease at each iteration. When the incumbent distribution ρ^τ is close to a local optimum, it is helpful to increase the noise level to escape the local optimum. We demonstrate in Lemma 13 in the Supplementary that any positive definite kernel k with bounded Hessian w.r.t. distance d is a Lipschitz-gradient kernel. Algorithm 2 in the Supplementary describes (4.3) in details.

5 Numerical Experiments

We evaluate the proposed gradient flow on real-world datasets and then illustrate its applications in transfer learning. We augment samples for the target dataset, where only a few samples in the dataset are available. We consider three datasets: the MNIST (M) [LeCun and Cortes, 2010], Fashion-MNIST (F) [Xiao *et al.*, 2017], Kuzushiji-MNIST (K) [Clanuwat *et al.*, 2018]. To satisfy the Gaussianity assumption of the conditional distributions, we cluster all the images from each class of the datasets and keep the largest cluster for each class. To demonstrate the scalability of our algorithm to higher-dimensional images, we run experiments on Tiny ImageNet (TIN) [Russakovsky *et al.*, 2015] and upscaled SVHN [Netzer *et al.*, 2011] datasets, where images are of $3 \times 64 \times 64$ size.

Our mapping ϕ is from $\mathbb{R}^m$ to $\mathbb{R}^2$ in the lifting procedure. To compute the MMD distance using kernel embeddings, we use a tensor kernel k on $\mathcal{Z}$ composed from three standard Gaussian kernels corresponding for each component of the feature space $\mathbb{R}^m$, the mean space $\mathbb{R}^2$ and the covariance matrix space $\mathbb{S}_{++}^2$. As a consequence, k is a characteristic kernel by [Szabó and Sriperumbudur, 2018, Theorem 4].

Experiment: Gradient Flow between Datasets. We visualize the path travelled by each sample from the source domain to the target domain, as depicted in Fig. 2. We draw randomly $N = 200$ images equally for 10 classes of the source domain, and $M = 50$ images equally for 10 classes of the target domain ($M = 10$ for the TIN and SVHN datasets). In each subfigure, each column represents a snapshot of a certain timestep and the samples flow from the source (left) to the target (right) as the number of steps increases. The first column in Fig. 2 are the images from the source domain, where the gradient flows start. Empirically, the algorithm converges after step 140 for *NIST datasets and step 6000 for TIN and SVHN. The experiments are run on a C5.4xlarge AWS instance (a CPU instance) and all finish in about one hour.

5.1 Application in Transfer Learning

Our gradient flow can alleviate the problem of insufficient labeled data by synthesizing new samples to augment the target

[2] See Definition A.3 for the technical definition

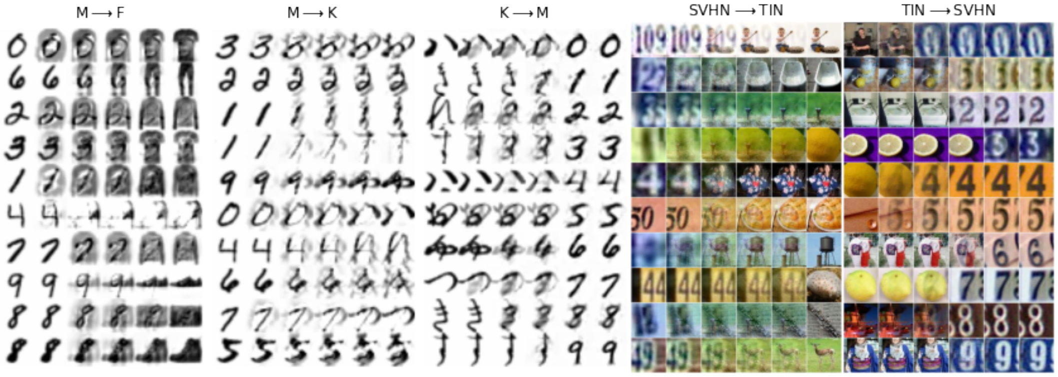

Figure 2: Sample path visualizations for five pairs of source-target domain. The original image and additional results are in the supplementary.

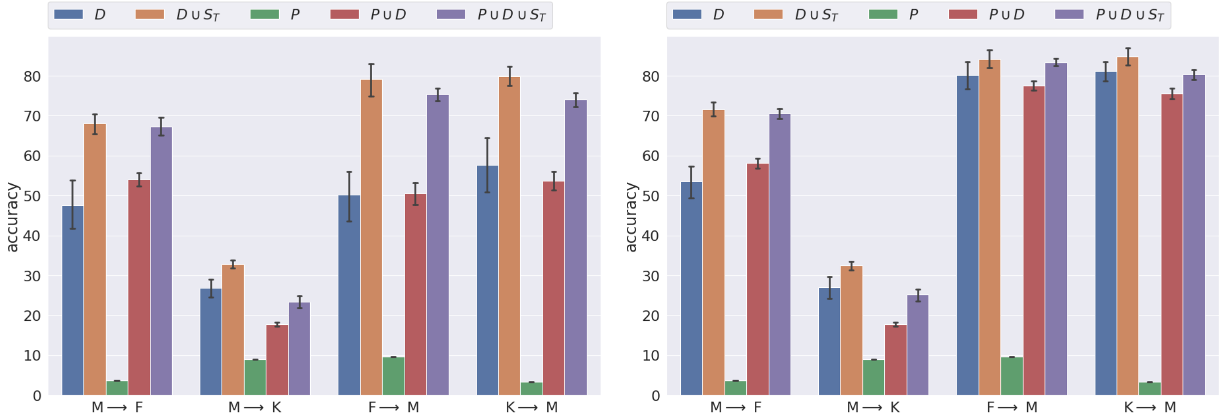

Figure 3: Average target domain accuracy on the test split for transfer learning with one-shot (left) and five-shot (right). Results are taken over 10 independent replications, and the range of accuracy is displayed by the error bars.

dataset. In this section, we demonstrate that the generated target domain samples can improve the accuracy in one-shot and five-shot transfer learning tasks.

First, we fix a source domain and pretrain a classifier P on this domain. We draw randomly N samples from the source domain to form the source dataset $(x_i, y_i)_{i=1}^N$. Next, we pick a target domain and draw randomly a few samples from this target domain: for example, in 1-shot learning, only 1 image per class from the target domain is selected to form the target dataset $D = (\bar{x}_j, \bar{y}_j)_{j=1}^M$. We then perform a noisy gradient flow scheme (4.3) from the source dataset to the target dataset to get N new samples $S_T = (x_i^T, y_i^T)_{i=1}^N$. With the target dataset D and new samples S_T, we can retrain the classifier P. Similarly, we can also train new classifiers from scratch using datasets D and $D \cup S_T$. Finally, we test the classifiers on the test set of the target domain.

Fig. 3 presents the accuracy of five transfer learning strategies on four pairs of source and target domain. For the labels above the plot, labels without P mean training a new classifier from scratch, whereas labels with P mean transferring the pretrained classifier. D and S_T represent the samples in the target domain and our flowed samples. We observe a common trend that the addition of the flowed samples S_T always improves the accuracy of the classifiers, as we compare $D \cup S_T$ with D and compare $P \cup D \cup S_T$ with $P \cup D$. Moreover, the data augmentation with S_T leads to a higher increase of accuracy for the 1-shot learning, where the data scarcity problem is more severe. The transfer learning results for SVHN and TIN datasets are provided in the Supplementary B.6. Although few-shot learning is more challenging due to the high complexity of the datasets, the addition of S_T always improves the accuracy. We also compare with baseline[3], mixup method and image augmentation methods in Supplementary B.7.

Conclusions. This paper focuses on a gradient flow approach to generate new labeled data samples in the target domain. To overcome the discrete nature of the labels, we represent datasets as distributions on the feature-Gaussian space, and the flow is formulated to minimize the MMD loss function under an optimal transport metric. Contrary to existing gradient flows on linear structure, our flows are developed on the *curved* Riemannian manifold of Gaussian distributions. We provide explicit formula for the Riemannian gradient of MMD, and analyze in details the flow equations and the convergence properties of both continuous and discretized forms. The numerical experiments demonstrate that our method can efficiently generate high-fidelity labeled training data for real-world datasets, and improve the classification accuracy in few-shot learning. The main limitation exists in the assumption that the data of one label forms an elliptical distribution.

[3] The only gradient flow work that has experiments on *NIST datasets, but it does not run experiments on TIN and SVHN.

Ethical Statement

Our work has positive societal impacts. Our work can enlarge the number of labelled data on any domains, so it alleviates the data scarcity problem in supervised learning settings. As a result, it can help reduce repetitive data collection and labeling work. It does not have possible negative societal impacts in the current stage.

Acknowledgements

We thank the anonymous reviewers for their comments. XH is supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0397. Additional support from NSF 1915967, NSF 2118199, and the Ford-Stanford alliance is gratefully acknowledged. The research of TN is supported in part by a grant from the Simons Foundation (#318995). TL gratefully acknowledges the support of JSPS KAKENHI Grants 20K19873 and 23K11243. VAN acknowledges the generous support from the CUHK's Improvement on Competitiveness in Hiring New Faculties Funding Scheme.

References

[Alvarez-Melis and Fusi, 2020] David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. In *Advances in Neural Information Processing Systems*, volume 33, pages 21428–21439, 2020.

[Alvarez-Melis and Fusi, 2021] David Alvarez-Melis and Nicolò Fusi. Dataset dynamics via gradient flows in probability space. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 219–230, 2021.

[Ambrosio *et al.*, 2008] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser Verlag, 2008.

[Arbel *et al.*, 2019] Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*, volume 32, pages 6481–6491, 2019.

[Arbel *et al.*, 2020] M Arbel, A Gretton, W Li, and G Montufar. Kernelized Wasserstein natural gradient. In *International Conference on Learning Representations*, 2020.

[Bellemare *et al.*, 2017] Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Remi Munos. The Cramer distance as a solution to biased Wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.

[Bhatia *et al.*, 2019] R. Bhatia, T. Jain, and Y. Lim. On the Bures-Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

[Bhatia, 1997] Rajendra Bhatia. *Matrix Analysis*. Springer, 1997.

[Bińkowski *et al.*, 2018] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

[Chizat and Bach, 2018] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[Clanuwat *et al.*, 2018] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical Japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

[Courty *et al.*, 2017] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[Damodaran *et al.*, 2018] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. DeepJDOT: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2018.

[Fan and Alvarez-Melis, 2022] Jiaojiao Fan and David Alvarez-Melis. Generating synthetic datasets by interpolating along generalized geodesics. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022.

[Frogner and Poggio, 2020] Charlie Frogner and Tomaso Poggio. Approximate inference with Wasserstein gradient flows. In *International Conference on Artificial Intelligence and Statistics*, pages 2581–2590. PMLR, 2020.

[Gao *et al.*, 2018] Hang Gao, Zheng Shou, Alireza Zareian, Hanwang Zhang, and Shih-Fu Chang. Low-shot learning via covariance-preserving adversarial augmentation networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[Gelbrich, 1990] M. Gelbrich. On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

[Gretton *et al.*, 2012] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[Gulcehre *et al.*, 2016] Caglar Gulcehre, Marcin Moczulski, Misha Denil, and Yoshua Bengio. Noisy activation functions. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 3059–3068, 2016.

[Gulcehre *et al.*, 2017] Caglar Gulcehre, Marcin Moczulski, Francesco Visin, and Yoshua Bengio. Mollifying networks. In *5th International Conference on Learning Representations*, 2017.

[Hazan *et al.*, 2016] Elad Hazan, Kfir Yehuda Levy, and Shai Shalev-Shwartz. On graduated optimization for stochastic non-convex problems. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1833–1841, 2016.

[Jordan *et al.*, 1998] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29:1–17, 1998.

[LeCun and Cortes, 2010] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[Lee, 2003] John Lee. *Introduction to Smooth Manifolds*. Springer-Verlag, 2003.

[Liu and Wang, 2016] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[Liu, 2017] Qiang Liu. Stein variational gradient descent as gradient flow. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[Liutkus *et al.*, 2019] Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter. Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[Malagò *et al.*, 2018] L. Malagò, L. Montrucchio, and G. Pistone. Wasserstein Riemannian geometry of Gaussian densities. *Information Geometry*, 1(2):137–179, 2018.

[Moreno-Barea *et al.*, 2018] Francisco J. Moreno-Barea, Fiammetta Strazzera, José M. Jerez, Daniel Urda, and Leonardo Franco. Forward noise adjustment scheme for data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 728–734, 2018.

[Mroueh and Nguyen, 2021] Youssef Mroueh and Truyen Nguyen. On the convergence of gradient descent in GANs: MMD GAN as a gradient flow. In *International Conference on Artificial Intelligence and Statistics*, 2021.

[Mroueh *et al.*, 2019] Youssef Mroueh, Tom Sercu, and Anant Raj. Sobolev descent. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 2976–2985, 2019.

[Netzer *et al.*, 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

[Otto, 2001] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26:101–174, 2001.

[Rezende *et al.*, 2016] Danilo Rezende, Ivo Danihelka, Karol Gregor, Daan Wierstra, et al. One-shot generalization in deep generative models. In *International conference on machine learning*, pages 1521–1529. PMLR, 2016.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[Santambrogio, 2015] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs and Modeling*. Birkhäuser, 2015.

[Santambrogio, 2017] Filippo Santambrogio. Euclidean, metric, and Wasserstein gradient flows: An overview. *Bullentin of Mathematical Sciences*, 7:87–154, 2017.

[Shorten and Khoshgoftaar, 2019] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[Szabó and Sriperumbudur, 2018] Zoltán Szabó and Bharath K. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.

[Takatsu, 2011] Asuka Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011.

[van der Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[Villani, 2008] C. Villani. *Optimal Transport: Old and New*. Springer Science & Business Media, 2008.

[Wang *et al.*, 2020] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[Wang *et al.*, 2021] Gefei Wang, Yuling Jiao, Qian Xu, Yang Wang, and Can Yang. Deep generative learning via Schrödinger bridge. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10794–10804. PMLR, 2021.

[Wiatrak *et al.*, 2019] Maciej Wiatrak, Stefano V Albrecht, and Andrew Nystrom. Stabilizing generative adversarial networks: A survey. *arXiv preprint arXiv:1910.00927*, 2019.

[Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[Yi *et al.*, 2019] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, 2019.

[Zhang *et al.*, 2017] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.