Optimal Transport for Kernel Gaussian Mixture Models

Jung Hun Oh OHJ@MSKCC.ORG
Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Rena Elkin ELKINR@MSKCC.ORG
Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Anish K. Simhal SIMHALA@MSKCC.ORG
Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Jiening Zhu JIENING.ZHU@STONYBROOK.EDU
Department of Applied Mathematics & Statistics, Stony Brook University, Stony Brook, NY, USA

Joseph O. Deasy DEASYJ@MSKCC.ORG
Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Allen Tannenbaum ALLEN.TANNENBAUM@STONYBROOK.EDU
Departments of Computer Science and Applied Mathematics & Statistics, Stony Brook University, Stony Brook, NY, USA

Abstract

The Wasserstein distance from optimal mass transport (OMT) is a powerful mathematical tool with numerous applications that provides a natural measure of the distance between two probability distributions. Several methods to incorporate OMT into widely used probabilistic models, such as Gaussian or Gaussian mixture, have been developed to enhance the capability of modeling complex multimodal densities of real datasets. However, very few studies have explored the OMT problems in a reproducing kernel Hilbert space (RKHS), wherein the *kernel trick* is utilized to avoid the need to explicitly map input data into a high-dimensional feature space. In the current study, we propose a Wasserstein-type metric to compute the distance between two Gaussian mixtures in a RKHS via the kernel trick, *i.e.*, kernel Gaussian mixture models.

1. Introduction

The Gaussian mixture model (GMM) is a probabilistic model defined as a weighted sum of several Gaussian distributions [6, 30]. Due to its mathematical simplicity and efficiency, GMMs are widely used to model complex multimodal densities of real datasets [7].

Optimal mass transport (OMT) is an active and ever-growing field of research, originating in the work of the French civil engineer Gaspard Monge in 1781, which was formulated as the optimal way (via the minimization of some transportation cost) to move a pile of soil from one site to another [9, 18, 37, 46]. OMT has made significant progress due to the pioneering effort of Leonid Kantorovich in 1942, who introduced a relaxed version of the original problem that is solved using simple linear programming [18]. Recently, there has been an ever increasing growth in applications of OMT in numerous fields, including medical imaging analysis, statistical physics, machine learning, and genomics [2, 3, 24, 50].

© J.H. Oh, R. Elkin, A.K. Simhal, J. Zhu, J.O. Deasy & A. Tannenbaum.

Here we briefly sketch the basic theory of OMT. Suppose that ν_0 and ν_1 are two absolutely continuous probability measures with compact support on $X = \mathbb{R}^d$. (The theory is valid on more general metric measure spaces.) A Borel map $T : X \to X$ is called a *transport plan* from ν_0 to ν_1 if it "push-forward" ν_0 to ν_1 ($T_{\#}\nu_0 = \nu_1$) which is equivalent to say that

$$\nu_1(B) = \nu_0(T^{-1}(B)), \tag{1}$$

for every Borel subset $B \subset \mathbb{R}^d$ [20, 23]. Let $c(x, y)$ be the transportation cost to move one unit of mass from x to y. The Monge version of OMT problem seeks an optimal transport map $T : X \to X$ such that the total transportation cost $\int_X c(x, T(x))\nu_0(dx)$ is minimized over the set of all transport maps T. We note that the original OMT problem is highly non-linear and may not admit a viable solution. To ease this computational difficulty, Leonid Kantorovich proposed a relaxed formulation, solved by employing a linear programming method [18, 41] which defines the W_p Wasserstein distance between ν_0 and ν_1 on $\mathbb{R}^d$ as follows:

$$W_p^p(\nu_0, \nu_1) = \inf_{\pi \in \Pi(\nu_0, \nu_1)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\pi(x, y), \tag{2}$$

where $\Pi(\nu_0, \nu_1)$ is the set of all joint probability measures π on $X \times X$ with ν_0 and ν_1 as its two marginals, and $c(x, y)$ is taken as a specific form of $c(x, y) = ||x - y||^p, p \geq 1$. In the present study, we focus on the W_2 Wasserstein distance using the squared Euclidean distance ($p = 2$) as the cost function [25].

While OMT ensures that the displacement interpolation (weighted barycenters) between two Gaussian distributions remains Gaussian, this property does not hold for Gaussian mixtures. To cope with this issue in GMMs, Chen *et al.* proposed a new Wasserstein-type distance [3]. This approach optimizes the transport map between the two probability vectors of the respective Gaussian mixtures using the discrete linear program where the cost function is computed as the closed-form formulation of the W_2 Wasserstein distance between Gaussian distributions. This ensures that the displacement interpolation between two Gaussian mixtures preserves the Gaussian mixture structure. Note that the sum of probabilities of all Gaussian components in a Gaussian mixture is 1, and therefore the total mass for two Gaussian mixtures is equal.

In machine learning, kernel methods provide a powerful framework for non-linear extensions of classical linear models by implicitly mapping the data into a high-dimensional feature space corresponding to a reproducing kernel Hilbert space (RKHS) via a non-linear mapping function [28, 31, 43]. Recently, a formulation for the W_2 Wasserstein distance metric between two Gaussian distributions in a RKHS was introduced [33, 34, 49]. Extending this concept, we propose an OMT framework to compute a Wasserstein-type distance between two Gaussian mixtures in a RKHS.

2. Methods

In this section, we first describe the underlying technical methods of the present work, and then introduce our proposed methodology.

2.1. Kernel function

Suppose that we are given a dataset of n samples, denoted by $X = [x_1, x_2, \cdots, x_n] \in \mathbb{R}^d$, each of which consists of d features. The data in the original input space can be mapped into a high-

dimensional feature space via a non-linear mapping function ϕ [1, 21, 32]:

$$\Phi : X \to \mathcal{F}, \tag{3}$$

where $\mathcal{F}$ is a Hilbert space called the *feature space* and a (Mercer) kernel function $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = \phi(x_i)^{\mathrm{T}} \phi(x_j)$ is defined as an inner dot product in $\mathcal{F}$ for a positive semi-definite kernel $k : X \times X \to \mathbb{R}$ in which the kernel function is symmetric: $k(x_i, x_j) = k(x_j, x_i)$ and $\Phi(X) := [\phi(x_1), \phi(x_2), \cdots, \phi(x_n)]$ [38, 39]. The resulting Gram matrix $K = \Phi^{\mathrm{T}}\Phi$ is positive semi-definite ($K \succeq 0$) with $K_{ij} := k(x_i, x_j), \forall i, j \in \{1, \cdots, n\}$. This process is called the *kernel trick*. Common choices of kernel functions are the Gaussian radial basis function (RBF) and polynomial kernels [19, 36, 42]. Kernels are widely used in machine learning algorithms such as support vector machines [4], linear discriminant analysis [32], and principal component analysis [40, 44]. In this study, the following RBF kernel is employed:

$$k(x_i, x_j) = \exp\big(- \gamma ||x_i - x_j||^2 \big), \tag{4}$$

where $\gamma > 0$ controls the kernel width. The mean and the covariance matrix in the feature space are given by

$$m = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) = \Phi s, \;\; \Sigma = \frac{1}{n} \sum_{i=1}^{n} (\phi(x_i) - m)(\phi(x_i) - m)^{\mathrm{T}} = \Phi J J^{\mathrm{T}} \Phi^{\mathrm{T}}, \tag{5}$$

where $s_{n\times 1} = \frac{1}{n}\mathbf{1}^{\mathrm{T}}$, $J = \frac{1}{\sqrt{n}}(I_n - s\mathbf{1})$, $\mathbf{1} = [1, 1, \cdots, 1]$, and I_n is the $n \times n$ identity matrix. Then, denoting ΦJ by S, we have

$$S = \Phi J = \frac{1}{\sqrt{n}} [(\phi(x_1) - m), \cdots, (\phi(x_n) - m)]. \tag{6}$$

A key idea to compute a Wasserstein-type distance between two Gaussian mixtures in a RKHS is to efficiently convert the mapping functions represented by m and Σ to kernel functions by putting two mapping functions next to each other in a certain way.

2.2. OMT between Gaussian distributions

Given two Gaussian distributions, ν_0 and $\nu_1 \in \mathbb{R}^d$, with mean θ_i and covariance matrix C_i for $i = 0$ and 1, the W_2 Wasserstein distance between the two distributions has the following closed formula:

$$W_2(\nu_0, \nu_1)^2 = \|\theta_0 - \theta_1\|^2 + \mathrm{tr}\left(C_0 + C_1 - 2\left(C_0^{\frac{1}{2}} C_1 C_0^{\frac{1}{2}} \right)^{\frac{1}{2}} \right), \tag{7}$$

where tr is the trace and $\mathrm{tr}\left(C_0 + C_1 - 2(C_0^{\frac{1}{2}} C_1 C_0^{\frac{1}{2}})^{\frac{1}{2}} \right) = \mathrm{tr}\left(C_0 + C_1 - 2(C_1 C_0)^{\frac{1}{2}} \right)$. In particular, when $C_0 = C_1$, we have $W_2(\nu_0, \nu_1)^2 = \|\theta_0 - \theta_1\|^2$. If C_0 and C_1 are non-degenerate, the geodesic path (displacement interpolation) $(\nu_t)_{t\in[0,1]}$ between ν_0 and ν_1 remains Gaussian and satisfies

$$\nu_t \in \mathrm{argmin}_{\rho} (1-t) W_2(\nu_0, \rho)^2 + t W_2(\nu_1, \rho)^2, \tag{8}$$

with mean $\theta_t = (1-t)\theta_0 + t\theta_1$ and covariance matrix given by:

$$C_t = ((1-t)I_d + tQ) C_0 ((1-t)I_d + tQ), \tag{9}$$

where I_d is the $d \times d$ identity matrix and $Q = C_1^{\frac{1}{2}} (C_1^{\frac{1}{2}} C_0 C_1^{\frac{1}{2}})^{-\frac{1}{2}} C_1^{\frac{1}{2}}$ [7].

2.3. OMT between Gaussian distributions in a RKHS

The data distribution in a RKHS represented via proper kernel functions is assumed to approximately follow a Gaussian distribution under suitable conditions, as justified by Huang *et al.* [13]. Let ρ_0 and $\rho_1 \in \mathbb{R}^l$ be two Gaussian distributions in a RKHS with mean m_i and covariance matrix Σ_i for $i = 0, 1$. The W_2 Wasserstein distance in a RKHS, denoted as KW_2 (*kernel Wasserstein distance*), is then defined as follows:

$$KW_2(\rho_0, \rho_1)^2 = \|m_0 - m_1\|^2 + \operatorname{tr}\left(\Sigma_0 + \Sigma_1 - 2\left(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right), \tag{10}$$

where $\Sigma_i^{\frac{1}{2}}$ is the unique semi-definite positive square root of a symmetric semi-definite positive matrix Σ_i [8, 48].

Suppose that there are two sets of data in the original input space, $X = [x_1, x_2, \cdots, x_n]$ and $Y = [y_1, y_2, \cdots, y_m] \in \mathbb{R}^d$ $(d < l)$, associated with ρ_0 and ρ_1, respectively. The first term in Eq. (10) is the squared maximum mean discrepancy (MMD) [14, 35] and may be expressed with kernel functions as follows:

$$\|m_0 - m_1\|^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k\left(x_i, x_j\right) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k\left(x_i, y_j\right) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k\left(y_i, y_j\right). \tag{11}$$

By using Eq. (5), the second term in Eq. (10) is expressed as follows:

$$\begin{aligned}
&\operatorname{tr}\left(\Sigma_0 + \Sigma_1 - 2\left(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right) = \operatorname{tr}\left(\Sigma_0 + \Sigma_1 - 2\left(\Sigma_1\Sigma_0\right)^{\frac{1}{2}}\right) \qquad (12)\\
&= \operatorname{tr}\left(J_0 J_0^{\mathrm{T}}\Phi_0^{\mathrm{T}}\Phi_0\right) + \operatorname{tr}\left(J_1 J_1^{\mathrm{T}}\Phi_1^{\mathrm{T}}\Phi_1\right) - 2\operatorname{tr}\left(\Phi_1 J_1 J_1^{\mathrm{T}} K_{10} J_0 J_0^{\mathrm{T}}\Phi_0^{\mathrm{T}}\right)^{\frac{1}{2}}\\
&= \operatorname{tr}\left(J_0 J_0^{\mathrm{T}} K_{00}\right) + \operatorname{tr}\left(J_1 J_1^{\mathrm{T}} K_{11}\right) - 2\operatorname{tr}\left(\Phi_1 Q \Phi_0^{\mathrm{T}}\right)^{\frac{1}{2}},
\end{aligned}$$

where $\operatorname{tr}\left(\Phi_1 Q \Phi_0^{\mathrm{T}}\right)^{\frac{1}{2}} = \operatorname{tr}\left(K_{01} J_1 J_1^{\mathrm{T}} K_{10} J_0 J_0^{\mathrm{T}}\right)^{\frac{1}{2}}$, $Q = J_1 J_1^{\mathrm{T}} K_{10} J_0 J_0^{\mathrm{T}}$, and $K_{ij} = \Phi_i^{\mathrm{T}}\Phi_j$. Note that Σ_0 and Σ_1 are symmetric positive semi-definite. Plugging the two terms together into Eq. (10), the KW_2 distance between ρ_0 and ρ_1 may be expressed as:

$$\begin{aligned}
KW_2(\rho_0, \rho_1)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k\left(x_i, x_j\right) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k\left(x_i, y_j\right) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k\left(y_i, y_j\right) + \qquad (13)\\
\operatorname{tr}\left(J_0 J_0^{\mathrm{T}} K_{00}\right) + \operatorname{tr}\left(J_1 J_1^{\mathrm{T}} K_{11}\right) - 2\operatorname{tr}\left(K_{01} J_1 J_1^{\mathrm{T}} K_{10} J_0 J_0^{\mathrm{T}}\right)^{\frac{1}{2}}.
\end{aligned}$$

Eq. (13) will be used as a key component for the computation of a Wasserstein-type distance between two Gaussian mixtures in a RKHS in the following section. In the special case, when $\Sigma_0 = \Sigma_1$, we have $KW_2(\rho_0, \rho_1)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k\left(x_i, x_j\right) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k\left(x_i, y_j\right) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k\left(y_i, y_j\right)$.

2.4. OMT between GMMs in a RKHS

Based on the OMT method between GMMs introduced in [3], we propose a Wasserstein-type metric to compute the distance between two Gaussian mixtures in a RKHS via the kernel trick, which

preserves the Gaussian mixture structure in the displacement interpolation. Let μ be a Gaussian mixture in a RKHS:

$$\mu = \sum_{k=1}^{N} p^k v^k, \tag{14}$$

where each v^k is a Gaussian distribution in a RKHS with $v^k = \mathcal{N}(m_k, \Sigma_k)$ and p^k is a probability of v^k with $\sum_{k=1}^{N} p^k = 1$. Let μ_0 and μ_1 denote two Gaussian mixtures in a RKHS in the following form:

$$\mu_i = p_i^1 v_i^1 + p_i^2 v_i^2 + ... + p_i^{N_i} v_i^{N_i}, \quad i = 0, 1, \tag{15}$$

where N_i is the number of Gaussian components of μ_i. The distance between μ_0 and μ_1 is then defined according to the discrete OMT formulation for discrete measures [3, 27]:

$$d(\mu_0, \mu_1)^2 = \min_{\pi \in \Pi(p_0, p_1)} \sum_{i,j} c_{ij} \pi_{ij}, \tag{16}$$

where $\Pi(p_0, p_1)$ is the set of all joint probability measures between p_0 and p_1, defined as:

$$\Pi(p_0, p_1) = \{\pi \in \mathbb{R}_+^{N_0 \times N_1} | \sum_j \pi_{kj} = p_0^k, \ \sum_k \pi_{kj} = p_1^j\}.$$

The cost c_{ij} is taken to be the square of the KW_2 distance between v_0^i and v_1^j in a RKHS:

$$c_{ij} = KW_2(v_0^i, v_1^j)^2. \tag{17}$$

Since v_0^i and v_1^j are Gaussian distributions in a RKHS, the KW_2 distance can be computed using Eq. (13). Let π^* be the optimal solution of Eq. (16). Then, the distance $d(\mu_0, \mu_1)$ between two Gaussian mixtures in a RKHS is defined as follows:

$$d(\mu_0, \mu_1) = \sqrt{\sum_{i,j} c_{ij} \pi_{ij}^*}, \tag{18}$$

where $d(\mu_0, \mu_1) \geq KW_2(\mu_0, \mu_1)$ and the following property holds [3]:

$$d(\mu_s, \mu_t) = (t - s) d(\mu_0, \mu_1), \quad 0 \leq s < t \leq 1. \tag{19}$$

The geodesic path μ_t between μ_0 and μ_1 is defined as

$$\mu_t = \sum_{i,j} \pi_{ij}^* v_t^{ij}, \tag{20}$$

where v_t^{ij} is the displacement interpolation between two Gaussian distributions, v_0^i and v_1^j.

3. Experiments

Suppose that we are given two Gaussian mixtures, $\mu_0 = 0.3\mathcal{N}(0.2, 0.002) + 0.7\mathcal{N}(0.4, 0.004)$ and $\mu_1 = 0.6\mathcal{N}(0.6, 0.005) + 0.4\mathcal{N}(0.8, 0.004)$ as a 1-dimensional example, as shown in Figure 1. Figure 2 shows the displacement interpolation μ_t for both the metric $d(\mu_0, \mu_1)$ and general Wasserstein distance between the μ_0 and μ_1 at t = 0, 0.2, 0.4, 0.6, 0.8, and 1.0. It is observed that the displacement interpolation for the metric $d(\mu_0, \mu_1)$ preserves the Gaussian mixture structure, whereas the displacement interpolation for the general Wasserstein distance does not.

As another example, three datasets were generated in the original input space, each of which has two distributions (Figure 3). Each dataset consists of 1,000 data points with 500 data points for each distribution. The $d(\cdot, \cdot)$ was then computed between each pair of datasets in a RKHS, assuming that each dataset (μ_0, μ_1, and μ_2) has two Gaussian components with v_0^1 (purple)/v_0^2 (orange), v_1^1 (red)/v_1^2 (blue), and v_2^1 (pink)/v_2^2 (brown), respectively in a RKHS (Tables 1 and 2). In the current study, $\gamma = 1$ (Table 1) and $\gamma = 10$ (Table 2) were used in the RBF kernel shown in Eq. (4).

The $d(\cdot, \cdot)$ values with $\gamma = 10$ were larger than those corresponding to $\gamma = 1$. Overall, the $d(\cdot, \cdot)$ values between dataset 2 and dataset 3 were smaller than dataset 1 vs. dataset 2 and dataset 1 vs. dataset 3. When $\gamma = 1$ and (p_0^1, p_0^2) is (0.1, 0.9) and (0.3, 0.7), the $d(\cdot, \cdot)$ values between dataset 1 and dataset 2 were larger than those between dataset 1 and dataset 3. By contrast, when (p_0^1, p_0^2) is one of (0.5, 0.5), (0.7, 0.3), and (0.9, 0.1), the $d(\cdot, \cdot)$ values between dataset 1 and dataset 2 were smaller than those between dataset 1 and dataset 3. When $\gamma = 10$, the $d(\cdot, \cdot)$ values between dataset 1 and dataset 3 were larger than those between dataset 1 and dataset 2 in more cases compared to when $\gamma = 1$.

Using dataset 1 and dataset 2, simulation tests were conducted with (0.1, 0.9), (0.5, 0.5), and (0.9, 0.1) for both (p_0^1, p_0^2) and (p_1^1, p_1^2) by randomly sampling 200, 400, 600, and 800 data points from 1,000 data points of each dataset. For each sampling experiment of the combination of (p_0^1, p_0^2) and (p_1^1, p_1^2), 100 tests were conducted and the average distance and standard deviation were computed. In Figure 4, the horizontal dot line indicates $d(\cdot, \cdot)$ when the original data with 1,000 data points for each dataset were analyzed. As can be seen, as the number of randomly selected data points increases, the standard deviation becomes narrower, converging to the horizontal dot line. Not surprisingly, the average distance after 100 repetitions of each sampling experiment was very similar to $d(\cdot, \cdot)$ computed on the original data with 1,000 data points for each dataset.

Figure 5 illustrates the elapsed time to compute $d(\cdot, \cdot)$ between dataset 1 and dataset 2 when each test was conducted only once with randomly sampled data points (200, 400, 600, 800) from each dataset, compared to the elapsed time in the original datasets (each 1,000 data points). As the number of data points increased, the computational time to compute $d(\cdot, \cdot)$ sharply increased. Therefore, there is a tradeoff between the computational time and the precision of computed distance in the sampling approach. Use of advanced sampling techniques will help compute the distance within reasonable computational time and with minimal error. All experiments in the current study were carried out using Python language in Google Colab-Pro environment.

Acknowledgments

This study was supported in part by National Institutes of Health (NIH)/National Cancer Institute Cancer Center support grant (P30 CA008748), NIH grant (R01-AG048769), AFOSR grant (FA9550-20-1-0029), Army Research Office grant (W911NF2210292), Breast Cancer Research Foundation grant (BCRF-17-193), and a grant from the Cure Alzheimer's Foundation.

References

[1] Gaston Baudat and Fatiha Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000.

[2] Yongxin Chen, Filemon Dela Cruz, Romeil Sandhu, Andrew L. Kung, Prabhjot Mundi, Joseph O. Deasy, and Allen Tannenbaum. Pediatric sarcoma data forms a unique cluster measured via the earth mover's distance. *Scientific Reports*, 7(1):7035, 2017.

[3] Yongxin Chen, Tryphon T. Georgiou, and Allen Tannenbaum. Optimal transport for gaussian mixture models. *IEEE Access*, 7:6269–6278, 2019.

[4] Zhenyu Chen, Jianping Li, and Liwei Wei. A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue. *Artificial Intelligence in Medicine*, 41(2):161–175, 2007.

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

[6] Sanjoy Dasgupta and Leonard J. Schulman. A two-round variant of em for gaussian mixtures. *The Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 152–159, 2000.

[7] Julie Delon and Agnès Desolneux. A wasserstein-type distance in the space of gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

[8] Julie Delon, Agnès Desolneux, and Antoine Salmona. Gromov-wasserstein distances between gaussian distributions. *Journal of Applied Probability*, 59(4):1–21, 2022.

[9] Lawrence C. Evans. Partial differential equations and monge-kantorovich mass transfer. *Current Developments in Mathematics*, 1997:65–126, 1999.

[10] Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Reproducing kernel hilbert space, mercer's theorem, eigenfunctions, nyström method, and use of kernels in machine learning: Tutorial and survey. *CoRR*, abs/2106.08443, 2021.

[11] João V. Graça, Kuzman Ganchev, and Ben Taskar. Expectation maximization and posterior constraints. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 569–576, 2007.

[12] Chunzhi Gu, Haoran Xie, Xuequan Lu, and Chao Zhang. Cgmvae: Coupling gmm prior and gmm estimator for unsupervised clustering and disentanglement. *IEEE Access*, 9:65140–65149, 2021.

[13] Su-Yun Huang, Chii-Ruey Hwang, and Miao-Hsiang Lin. Kernel fisher's discriminant analysis in gaussian reproducing kernel hilbert space. *Taiwan: Academia Sinica*, 2005.

[14] Arun Iyer, Saketha Nath, and Sunita Sarawagi. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. *International Conference on Machine Learning (ICML)*, 32(1):530–538, 2014.

[15] Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. Entropic optimal transport between unbalanced gaussian measures has a closed form. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10468–10479, 2020.

[16] Hongwei Jin and Xun Chen. Gromov-wasserstein discrepancy with local differential privacy for distributed structural graphs. *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2115–2121, 2022.

[17] Hongwei Jin, Zishun Yu, and Xinhua Zhang. Orthogonal gromov-wasserstein discrepancy with efficient lower bound. *The Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 917–927, 2022.

[18] Leonid Kantorovich. On the translocation of masses, dokl. akad. nauk sssr 37 (1942) 227-229, english translation:. *Journal of Mathematical Sciences*, 133:1381–1382, 2006.

[19] Soheil Kolouri, Yang Zou, and Gustavo K. Rohde. Sliced wasserstein kernels for probability distributions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5258–5267, 2016.

[20] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced wasserstein distances. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[21] Heesung Kwon and Nasser M. Nasrabadi. Kernel orthogonal subspace projection for hyperspectral signal classification. *IEEE Transactions on Geoscience and Remote Sensing*, 43(12): 2952–2962, 2005.

[22] Khang Le, Dung Q. Le, Huy Nguyen, Dat Do, Tung Pham, and Nhat Ho. Entropic gromov-Wasserstein between Gaussian distributions. *International Conference on Learning Representations (ICLR)*, 162:12164–12203, 2022.

[23] Na Lei, Kehua Su, Li Cui, Shing-Tung Yau, and Xianfeng David Gu. A geometric view of optimal transportation and generative model. *Computer Aided Geometric Design*, 68:1–21, 2019.

[24] Giulia Luise, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto. Differential properties of sinkhorn approximation for learning with wasserstein distance. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5864–5874, 2018.

[25] Anton Mallasto and Aasa Feragen. Learning from uncertain curves: The 2-wasserstein metric for gaussian processes. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5660–5670, 2017.

[26] Anton Mallasto, Augusto Gerolin, and Hà Quang Minh. Entropy-regularized 2-wasserstein distance between gaussian measures. *Information Geometry*, 5:289–323, 2022.

[27] James C. Mathews, Saad Nadeem, Maryam Pouryahya, Zehor Belkhatir, Joseph O. Deasy, Arnold J. Levine, and Allen R. Tannenbaum. Functional network analysis reveals an immune tolerance mechanism in cancer. *Proc. Natl. Acad. Sci. U. S. A.*, 117(28):16339–16345, 2020.

[28] Giacomo Meanti, Luigi Carratino, Lorenzo Rosasco, and Alessandro Rudi. Kernel methods through the roof: Handling billions of points efficiently. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 14410–14422, 2020.

[29] Hà Quang Minh. Entropic regularization of wasserstein distance between infinite-dimensional gaussian measures and gaussian processes. *Journal of Theoretical Probability*, 2022.

[30] Luminita Moraru, Simona Moldovanu, Lucian Traian Dimitrievici, Nilanjan Dey, Amira S. Ashour, Fuqian Shi, Simon James Fong, Salam Khan, and Anjan Biswas. Gaussian mixture model for texture characterization with application to brain DTI images. *Journal of Advanced Research*, 16:15–23, 2019.

[31] Jung Hun Oh and Jean Gao. A kernel-based approach for detecting outliers of high-dimensional biological data. *BMC Bioinformatics*, 10 Suppl 4(Suppl 4):S7, 2009.

[32] Jung Hun Oh and Jean Gao. Fast kernel discriminant analysis for classification of liver cancer mass spectra. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(6): 1522–1534, 2011.

[33] Jung Hun Oh, Maryam Pouryahya, Aditi Iyer, Aditya P. Apte, Joseph O. Deasy, and Allen Tannenbaum. Kernel wasserstein distance. *arXiv*, 1905.09314, 2019.

[34] Jung Hun Oh, Maryam Pouryahya, Aditi Iyer, Aditya P. Apte, Joseph O. Deasy, and Allen Tannenbaum. A novel kernel wasserstein distance on gaussian measures: An application of identifying dental artifacts in head and neck computed tomography. *Computers in Biology and Medicine*, 120:103731, 2020.

[35] Liwen Ouyang and Aaron Key. Maximum mean discrepancy for generalization in the presence of distribution and missingness shift. *CoRR*, abs/2111.10344, 2021.

[36] Karl Ezra Pilario, Mahmood Shafiee, Yi Cao, Liyun Lao, and Shuang-Hua Yang. A review of kernel methods for feature extraction in nonlinear process monitoring. *Processes*, 8(1), 2020.

[37] Maryam Pouryahya, Jung Hun Oh, Pedram Javanmard, James C. Mathews, Zehor Belkhatir, Joseph O. Deasy, and Allen R. Tannenbaum. AWCluster: A novel integrative network-based clustering of multiomics for subtype analysis of cancer data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(3):1472–1483, 2022.

[38] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1177–1184, 2007.

[39] Bernhard Schölkopf. The kernel trick for distances. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 301–307, 2000.

[40] Bernhard Schölkopf, Sebastian Mika, Christopher J. C. Burges, Phil Knirsch, Klaus-Robert Muller, Gunnar Rätsch, and Alexander J. Smola. Input space versus feature space in kernel-based methods. *IEEE Transactions on Neural Networks*, 10(5):1000–1017, 1999.

[41] Dai Shi, Junbin Gao, Xia Hong, S. T. Boris Choy, and Zhiyong Wang. Coupling matrix manifolds assisted optimization for optimal transport problems. *Machine Learning*, 110:533–558, 2021.

[42] Haochen Shi, Haipeng Xiao, Jianjiang Zhou, Ning Li, and Huiyu Zhou. Radial basis function kernel parameter optimization algorithm in support vector machine based on segmented dichotomy. *International Conference on Systems and Informatics (ICSAI)*, pages 383–388, 2018.

[43] Carl-Johann Simon-Gabriel and Bernhard Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29, 2018.

[44] Nicholas Sterge, Bharath Sriperumbudur, Lorenzo Rosasco, and Alessandro Rudi. Gain with no pain: Efficiency of kernel-pca by nyström sampling. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3642–3652, 2020.

[45] Qijun Tong and Kei Kobayashi. Entropy-regularized optimal transport on multivariate normal and q-normal distributions. *Entropy*, 23(3), 2021.

[46] Cedric Villani. *Topics in optimal transportation*. American Mathematical Soc., 2003.

[47] Jingdong Wang, Jianguo Lee, and Changshui Zhang. Kernel trick embedded gaussian mixture model. *Algorithmic Learning Theory*, pages 159–174, 2003.

[48] Gui-Song Xia, Sira Ferradans, Gabriel Peyré, and Jean-François Aujol. Synthesizing and mixing stationary gaussian texture models. *SIAM Journal on Imaging Sciences*, 7(1):476–508, 2014.

[49] Zhen Zhang, Mianzhi Wang, and Arye Nehorai. Optimal transport in reproducing kernel hilbert spaces: Theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1741–1754, 2020.

[50] Xin Zhao, Zhengyu Su, Xianfeng David Gu, Arie Kaufman, Jian Sun, Jie Gao, and Feng Luo. Area-preservation mapping using optimal mass transport. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2838–2847, 2013.

Appendix A. Related work

The Gaussian model has numerous applications in data analysis due to its mathematical tractability and simplicity. In particular, a closed-form formulation of W_2 Wasserstein distance for Gaussian densities allows for the extension of applications of OMT in connection with Gaussian processes. As a further extension, Janati $et\ al.$ proposed an entropy-regularized OMT method between two Gaussian measures, by solving the fixed-point equation underpinning the Sinkhorn algorithm for both the balanced and unbalanced cases [5, 15, 45]. Mallasto $et\ al.$ introduced an alternative approach for the entropy-regularized optimal transport, providing closed-form expressions and interpolations between Gaussian measures [26]. Le $et\ al.$ investigated the entropic Gromov-Wasserstein distance between (unbalanced) Gaussian distributions and the entropic Gromov-Wasserstein barycenter of multiple Gaussian distributions [16, 17, 22].

Kernel methods are extensively employed in machine learning, providing a powerful capability to efficiently handle data in a non-linear space by implicitly mapping data into a high-dimensional space, a method known as the *kernel trick*. Ghojogh $et\ al.$ comprehensively reviewed the background theory of kernels and their applications in machine learning [10]. As an effort to incorporate kernels into OMT, Zhang $et\ al.$ proposed a solution to compute the W_2 Wasserstein distance between Gaussian measures in a RKHS [49] and Oh $et\ al.$ proposed an alternative algorithm called the *kernel Wasserstein distance*, giving an explicit detailed proof [34]. Minh introduced a novel algorithm to compute the entropy-regularized W_2 Wasserstein distance between Gaussian measures in a RKHS via the finite kernel Gram matrices, providing explicit closed-form formulas along with the Sinkhorn barycenter equation with a unique non-trivial solution [29].

Numerous studies have been conducted on GMMs in connection with other techniques. Wang $et\ al.$ proposed a kernel trick embedded GMM method by employing the Expectation Maximization (EM) algorithm to deduce a parameter estimation method for GMMs in the feature space, and introduced a Monte Carlo sampling technique to speed up the computation in large-scale data problems [11, 12, 47]. Chen $et\ al.$ proposed a new algorithm to compute a Wasserstein-type distance between two Gaussian mixtures, employing the closed-form solution between Gaussian measures as the cost function, represented as the discrete optimization problem [3]. Following this latter work, Delon and Desolneux investigated the theory of the Wasserstein-type distance on GMMs, showing several applications including color transfer between images [7]. Mathews $et\ al.$ applied the GMM Wasserstein-type distance to functional network analysis utilizing RNA-Seq gene expression profiles from The Cancer Genome Atlas (TCGA) [27]. This approach enabled the identification of gene modules (communities) based on the local connection structure of the gene network and the collection of joint distributions of nodal neighborhoods. However, to date, no study has explored the incorporation of the kernel trick embedded GMM into OMT. To address this, we propose an OMT solution to compute the distance between Gaussian mixtures in a RKHS.

Appendix B. Entropy-regularized optimal transport between Gaussians

Given $\mu, \nu \in \mathcal{P}$ and the cost function c, the entropic optimal transport (OT) problem proposed by Mallasto *et al.* is expressed as follows [26]:

$$\mathrm{OT}^{\epsilon}(\mu,\nu) = \min_{\gamma\in\pi(\mu,\nu)}\big\{\mathbb{E}_{\gamma}[c] + \epsilon D_{\mathrm{KL}}(\gamma||\mu\otimes\nu)\big\}, \tag{21}$$

which relaxes the OT problem employing a Kullback-Leibler (KL)-divergence term, yielding a strictly convex problem for $\epsilon > 0$. Let ν_0 and $\nu_1 \in \mathbb{R}^d$ be two Gaussian distributions with mean θ_i

and covariance matrix C_i for $i = 0$ and 1. Then, the entropy-regularized W_2 Wasserstein distance can be computed using a closed-form defined as:

$$W_2^\epsilon(\nu_0, \nu_1)^2 = \|\theta_0 - \theta_1\|^2 + \mathrm{tr}(C_0) + \mathrm{tr}(C_1) - \mathcal{B}, \\ \mathcal{B} = \frac{\epsilon}{2}\left(\mathrm{tr}(M^\epsilon) - \log\det(M^\epsilon) + d\log 2 - 2d\right), \tag{22}$$

where $M_{ij}^\epsilon = I + (I + \frac{16}{\epsilon^2} C_i C_j)^{1/2}$, assuming

$$\gamma = \mathcal{N}(0, \Gamma), \;\; \Gamma = \begin{bmatrix} C_0 & G^{\mathrm{T}} \\ G & C_1 \end{bmatrix}. \tag{23}$$

The entropic displacement interpolation between ν_0 and ν_1 also follows Gaussian as $\nu_t = \mathcal{N}(\theta_t, C_t)$ for $t \in [0, 1]$ with $\theta_t = t\nu_0 + (1-t)\nu_1$ and

$$C_t = (1-t)^2 C_0 + t^2 C_1 + t(1-t)\left(\left(\frac{\epsilon^2}{16} I + C_0 C_1\right)^{1/2} + \left(\frac{\epsilon^2}{16} I + C_1 C_0\right)^{1/2}\right). \tag{24}$$

The entropic barycenter for N distributions with $\nu_0, \nu_1, \cdots, \nu_{N-1}$ is expressed as follows:

$$\theta^* = \sum_{i=1}^N \lambda_i \theta_i, \;\; C^* = \frac{\epsilon}{4} \sum_{i=1}^N \lambda_i \left(-I + \left(I + \frac{16}{\epsilon^2} C^{*1/2} C_i C^{*1/2}\right)^{1/2}\right). \tag{25}$$

Similarly, Janati *et al.* defined a closed-form for the entropy-regularized W_2 Wasserstein distance between Gaussians as follows [15]:

$$W_2^\sigma(\nu_0, \nu_1)^2 = \|\theta_0 - \theta_1\|^2 + \mathrm{tr}(C_0) + \mathrm{tr}(C_1) - \mathcal{F}, \\ \mathcal{F} = \mathrm{tr}(D_\sigma) - d\sigma^2(1 - \log(2\sigma^2)) - \sigma^2 \log\det(D_\sigma + \sigma^2 I), \tag{26}$$

where $D_\sigma = (4C_0C_1 + \sigma^4 I)^{1/2}$ and $\sigma > 0$.

Appendix C. Entropy-regularized optimal transport between Gaussians in RKHS

Here we propose two new formulas to solve the entropy-regularized W_2 Wasserstein distance between Gaussians in RKHS using kernel trick.

Proposition 1. Let $\rho_i = \mathcal{N}(m_i, \Sigma_i)$ for $i = 0, 1$ be two Gaussian distributions on $\mathbb{R}^l$ in RKHS, and let two sets of data in the input space be $X = [x_1, x_2, \cdots, x_n]$ and $Y = [y_1, y_2, \cdots, y_m] \in \mathbb{R}^d$ $(l > d)$ associated with ρ_0 and ρ_1, respectively. Then, a closed-form solution for Eq. (22) in RKHS exists, denoted as $KW_2^\epsilon(\rho_0, \rho_1)^2$:

$$KW_2^\epsilon(\rho_0, \rho_1)^2 = \|m_0 - m_1\|^2 + \mathrm{tr}(\Sigma_0) + \mathrm{tr}(\Sigma_1) - \mathcal{B}, \\ \mathcal{B} = \frac{\epsilon}{2}\left(\mathrm{tr}(M^\epsilon) - \log\det(M^\epsilon) + l\log 2 - 2l\right). \tag{27}$$

Proof. Let $A = \Sigma_0\Sigma_1$. The trace of a square matrix A is defined as the trace of the eigenvalue matrix of A, i.e., $\mathrm{tr}(A) = \mathrm{tr}(\Lambda)$ in $AP = P\Lambda$ where P and Λ are the estimated eigenvector and eigenvalue matrices, respectively. Let $\lambda_1, \lambda_2, \cdots, \lambda_k > 0$ be the distinct eigenvalues of A. Then, the eigenvalues of $(I + \frac{16}{\epsilon^2}A)^{1/2}$ are $\sqrt{1 + \frac{16}{\epsilon^2}\lambda_1}, \sqrt{1 + \frac{16}{\epsilon^2}\lambda_2}, \cdots, \sqrt{1 + \frac{16}{\epsilon^2}\lambda_k}$ and

$$\mathrm{tr}(M^\epsilon) = \mathrm{tr}\left(I + \left(I + \frac{16}{\epsilon^2}\Sigma_0\Sigma_1\right)^{1/2}\right) = l + \sum_{i=1}^{k}\sqrt{1+\frac{16}{\epsilon^2}\lambda_i}. \tag{28}$$

By $\det(I+tA) = (1+t\lambda_1)(1+t\lambda_2)\cdots(1+t\lambda_k)$, we compute $\log\det(M^\epsilon)$ as follows:

$$\begin{aligned}\log\det(M^\epsilon) &= \log\det\left(I + \left(I + \frac{16}{\epsilon^2}\Sigma_0\Sigma_1\right)^{1/2}\right) \\ &= \log\left(1+\sqrt{1+\frac{16}{\epsilon^2}\lambda_1}\right)\left(1+\sqrt{1+\frac{16}{\epsilon^2}\lambda_2}\right)\cdots\left(1+\sqrt{1+\frac{16}{\epsilon^2}\lambda_k}\right) \\ &= \sum_{i=1}^{k}\log\left(1+\sqrt{1+\frac{16}{\epsilon^2}\lambda_i}\right).\end{aligned} \tag{29}$$

Therefore, $\mathcal{B}$ is expressed as:

$$\begin{aligned}\mathcal{B} &= \frac{\epsilon}{2}\left(\mathrm{tr}(M^\epsilon) - \log\det(M^\epsilon) + l\log 2 - 2l\right) \\ &= \frac{\epsilon}{2}\left(\sum_{i=1}^{k}\sqrt{1+\frac{16}{\epsilon^2}\lambda_i} - \sum_{i=1}^{k}\log\left(1+\sqrt{1+\frac{16}{\epsilon^2}\lambda_i}\right) - l\log 5\right).\end{aligned} \tag{30}$$

Finally, we have a closed-form solution for the entropy-regularized W_2 Wasserstein distance between Gaussians in RKHS:

$$\begin{aligned}KW_2^\epsilon(\rho_0,\rho_1)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k\left(x_i,x_j\right) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}k\left(x_i,y_j\right) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k\left(y_i,y_j\right) + \\ \mathrm{tr}\left(J_0J_0^{\mathrm{T}}K_{00}\right) + \mathrm{tr}\left(J_1J_1^{\mathrm{T}}K_{11}\right) - \frac{\epsilon}{2}\left(\sum_{i=1}^{k}\sqrt{1+\frac{16}{\epsilon^2}\lambda_i} - \sum_{i=1}^{k}\log\left(1+\sqrt{1+\frac{16}{\epsilon^2}\lambda_i}\right) - l\log 5\right).\end{aligned} \tag{31}$$

Proposition 2. Similarly, we have a closed-form solution for Eq. (26) in RKHS.

Proof. Since the first two terms in Eq. (26) are the same as those in Eq. (22), we only solve $\mathcal{F} = \mathrm{tr}(D_\sigma) - l\sigma^2(1-\log(2\sigma^2)) - \sigma^2\log\det(D_\sigma+\sigma^2 I)$ in RKHS. As in Proposition 1, let $\lambda_1,\lambda_2,\cdots,\lambda_k > 0$ be the distinct eigenvalues of $\Sigma_0\Sigma_1$ that can be computed using $\mathrm{tr}(\Sigma_0\Sigma_1) = \mathrm{tr}\left(K_{10}J_0J_0^{\mathrm{T}}K_{01}J_1J_1^{\mathrm{T}}\right)$. Then, the eigenvalues of $D_\sigma = (4\Sigma_0\Sigma_1+\sigma^4 I)^{1/2} = \sigma^2(\frac{4}{\sigma^4}\Sigma_0\Sigma_1+I)^{1/2}$ are $\sigma^2\sqrt{\frac{4}{\sigma^4}\lambda_1+1}, \sigma^2\sqrt{\frac{4}{\sigma^4}\lambda_2+1},\cdots,\sigma^2\sqrt{\frac{4}{\sigma^4}\lambda_k+1}$. Therefore, $\mathrm{tr}(D_\sigma)$ is expressed as follows:

$$\mathrm{tr}(D_\sigma) = \sigma^2\sum_{i=1}^{k}\sqrt{\frac{4}{\sigma^4}\lambda_i+1}. \tag{32}$$

Here $\det(D_\sigma+\sigma^2 I)$ can be solved as follows:

$$\begin{aligned}\det(D_\sigma+\sigma^2 I) &= \sigma^{2l}\det\left(\frac{D_\sigma}{\sigma^2}+I\right) \\ &= \sigma^{2l}\left(\sqrt{\frac{4}{\sigma^4}\lambda_1+1}+1\right)\left(\sqrt{\frac{4}{\sigma^4}\lambda_2+1}+1\right)\cdots\left(\sqrt{\frac{4}{\sigma^4}\lambda_k+1}+1\right).\end{aligned} \tag{33}$$

Therefore, $\log\det(D_\sigma + \sigma^2 I)$ is

$$\log\det(D_\sigma + \sigma^2 I) = 2l\log\sigma + \sum_{i=1}^{k} \log\left(\sqrt{\frac{4}{\sigma^4}\lambda_i + 1} + 1\right). \tag{34}$$

Taken together,

$$\begin{aligned}
\mathcal{F} &= \mathrm{tr}(D_\sigma) - l\sigma^2\left(1 - \log(2\sigma^2)\right) - \sigma^2\log\det\left(D_\sigma + \sigma^2 I\right) \qquad (35)\\
&= \sigma^2\sum_{i=1}^{k}\sqrt{\frac{4}{\sigma^4}\lambda_i + 1} - l\sigma^2\left(1 - \log(2\sigma^2)\right) - \sigma^2\left(2l\log\sigma + \sum_{i=1}^{k}\log\left(\sqrt{\frac{4}{\sigma^4}\lambda_i + 1} + 1\right)\right)\\
&= \sigma^2\sum_{i=1}^{k}\sqrt{\frac{4}{\sigma^4}\lambda_i + 1} - \sigma^2\sum_{i=1}^{k}\log\left(\sqrt{\frac{4}{\sigma^4}\lambda_i + 1} + 1\right) - l\sigma^2(1 - \log 2)\\
&= \sigma^2\left(\sum_{i=1}^{k}\sqrt{\frac{4}{\sigma^4}\lambda_i + 1} - \sum_{i=1}^{k}\log\left(\sqrt{\frac{4}{\sigma^4}\lambda_i + 1} + 1\right) - l\log 5\right).
\end{aligned}$$

Interestingly, we found that the two solutions are the same with $2\sigma^2 = \epsilon$.

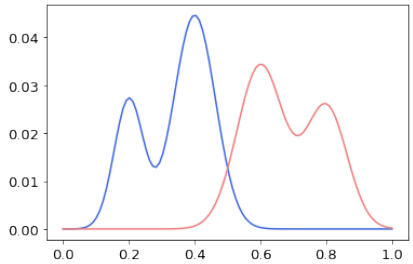

Figure 1: As an example, two Gaussian mixtures were created with $\mu_0 = 0.3\mathcal{N}(0.2, 0.002) + 0.7\mathcal{N}(0.4, 0.004)$ in blue and $\mu_1 = 0.6\mathcal{N}(0.6, 0.005) + 0.4\mathcal{N}(0.8, 0.004)$ in red, in a 1-dimensional space.

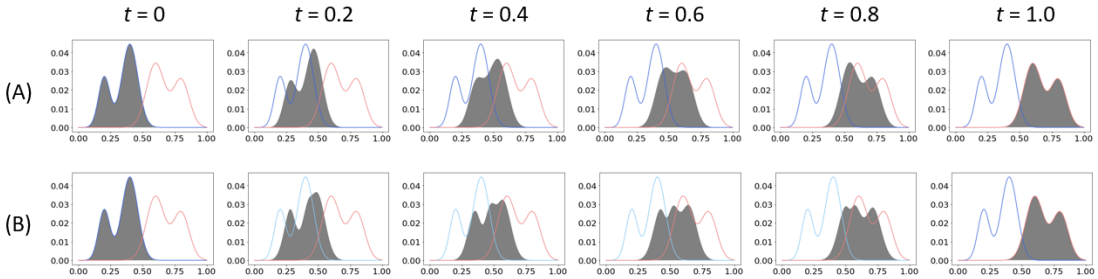

Figure 2: Displacement interpolation (gray) μ_t between two Gaussian mixtures, μ_0 (blue curve) and μ_1 (red curve). (A) Displacement interpolation for the metric $d(\mu_0, \mu_1)$ and (B) general Wasserstein distance.

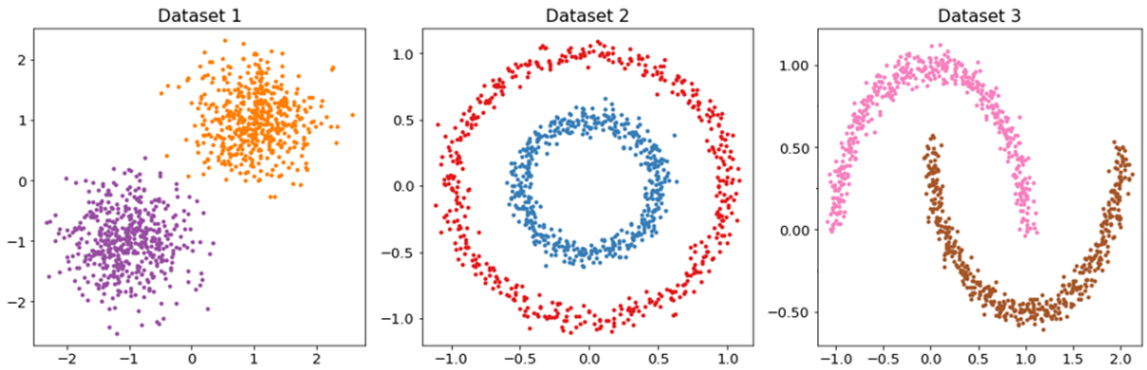

Figure 3: Three datasets, each of which has two distributions, which were used for Gaussian mixture analysis in a RKHS. We assume that each dataset follows a Gaussian mixture, *i.e.*, $\mu_0 = p_0^1 v_0^1 + p_0^2 v_0^2$, $\mu_1 = p_1^1 v_1^1 + p_1^2 v_1^2$, and $\mu_2 = p_2^1 v_2^1 + p_2^2 v_2^2$, respectively, in a RKHS.

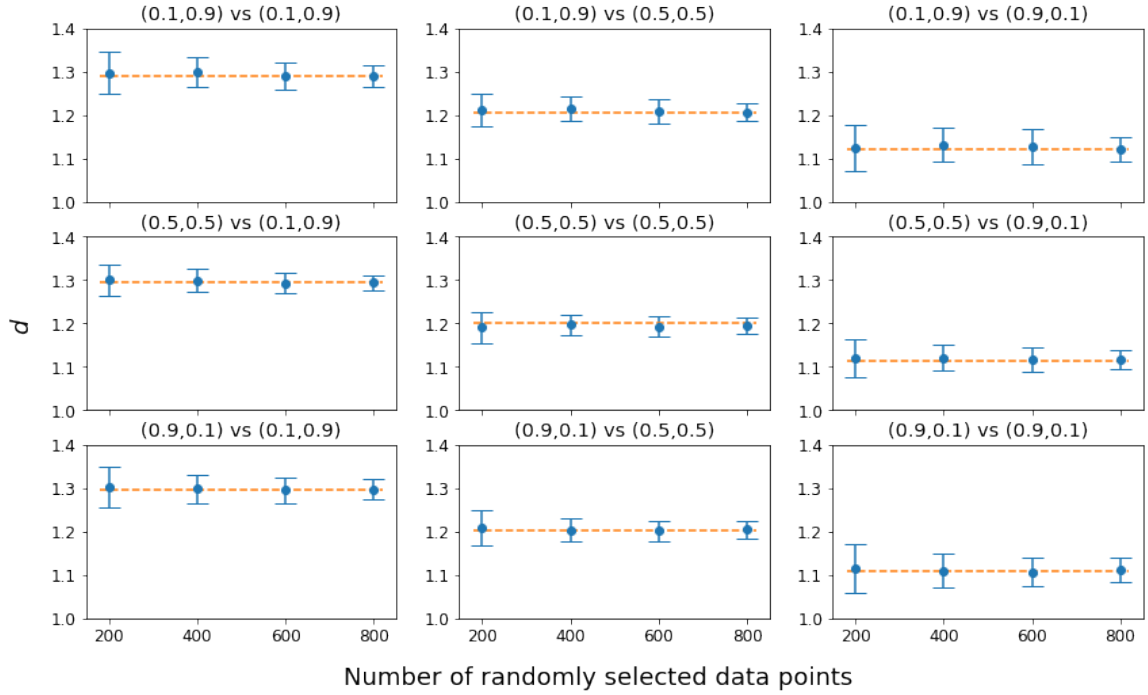

Figure 4: Using dataset 1 and dataset 2, simulation tests were conducted with (0.1, 0.9), (0.5, 0.5), and (0.9, 0.1) for both (p_0^1, p_0^2) and (p_1^1, p_1^2) by randomly sampling 200, 400, 600, and 800 data points from each dataset. The blue dot and error bar indicate the average distance and standard deviation after 100 repetitions of each sampling experiment with the combination of (p_0^1, p_0^2) and (p_1^1, p_1^2). The horizontal dot line indicates $d(\cdot, \cdot)$ when the original data with 1,000 data points for each dataset were tested.

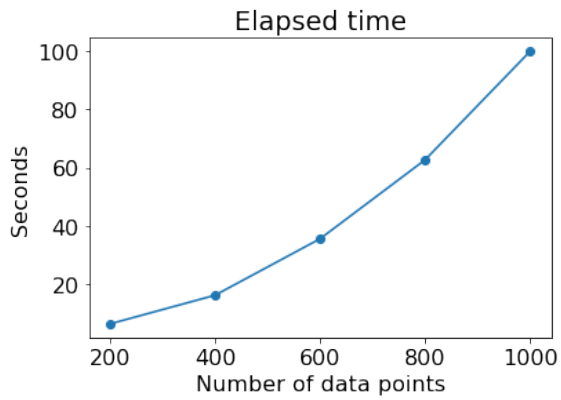

Figure 5: Elapsed time to compute $d(\cdot,\cdot)$ between dataset 1 and dataset 2 with randomly sampled data points (200, 400, 600, 800) from each dataset, compared to the elapsed time in the original datasets (each 1,000 data points).

Table 1: $d(\cdot,\cdot)$ with $\gamma = 1$ in a RKHS (A1) between dataset 1 and dataset 2, (A2) between dataset 1 and dataset 3, and (A3) between dataset 2 and dataset 3 using five different probability combinations {(0.1,0.9), (0.3,0.7), (0.5,0.5), (0.7,0.3), (0.9,0.1)} for two Gaussian components of each Gaussian mixture.

$\gamma = 1$					
(A1) Dataset 1 [row: (p_0^1, p_0^2)] vs Dataset 2 [column: (p_1^1, p_1^2)]					
$(p_0^1, p_0^2)/(p_1^1, p_1^2)$	(0.1, 0.9)	(0.3, 0.7)	(0.5, 0.5)	(0.7, 0.3)	(0.9, 0.1)
(0.1, 0.9)	1.292	1.249	1.206	1.164	1.121
(0.3, 0.7)	1.293	1.246	1.203	1.160	1.118
(0.5, 0.5)	1.294	1.247	1.200	1.157	1.114
(0.7, 0.3)	1.295	1.248	1.201	1.154	1.111
(0.9, 0.1)	1.295	1.248	1.201	1.155	1.108
(A2) Dataset 1 [row: (p_0^1, p_0^2)] vs Dataset 3 [column: (p_2^1, p_2^2)]					
$(p_0^1, p_0^2)/(p_2^1, p_2^2)$	(0.1, 0.9)	(0.3, 0.7)	(0.5, 0.5)	(0.7, 0.3)	(0.9, 0.1)
(0.1, 0.9)	1.185	1.123	1.060	0.998	0.935
(0.3, 0.7)	1.289	1.227	1.164	1.102	1.091
(0.5, 0.5)	1.394	1.331	1.268	1.258	1.247
(0.7, 0.3)	1.498	1.435	1.425	1.414	1.404
(0.9, 0.1)	1.602	1.591	1.581	1.570	1.560
(A3) Dataset 2 [row: (p_1^1, p_1^2)] vs Dataset 3 [column: (p_2^1, p_2^2)]					
$(p_1^1, p_1^2)/(p_2^1, p_2^2)$	(0.1, 0.9)	(0.3, 0.7)	(0.5, 0.5)	(0.7, 0.3)	(0.9, 0.1)
(0.1, 0.9)	0.890	0.844	0.798	0.753	0.707
(0.3, 0.7)	0.905	0.796	0.751	0.705	0.659
(0.5, 0.5)	0.921	0.812	0.703	0.657	0.612
(0.7, 0.3)	0.936	0.827	0.718	0.609	0.564
(0.9, 0.1)	0.952	0.843	0.734	0.625	0.516

Table 2: $d(\cdot,\cdot)$ with $\gamma = 10$ in a RKHS (B1) between dataset 1 and dataset 2, (B2) between dataset 1 and dataset 3, and (B3) between dataset 2 and dataset 3 using five different probability combinations {(0.1,0.9), (0.3,0.7), (0.5,0.5), (0.7,0.3), (0.9,0.1)} for two Gaussian components of each Gaussian mixture.

$\gamma = 10$					
(B1) Dataset 1 [row: (p_0^1, p_0^2)] vs Dataset 2 [column: (p_1^1, p_1^2)]					
$(p_0^1, p_0^2)/(p_1^1, p_1^2)$	(0.1, 0.9)	(0.3, 0.7)	(0.5, 0.5)	(0.7, 0.3)	(0.9, 0.1)
(0.1, 0.9)	1.661	1.618	1.575	1.532	1.489
(0.3, 0.7)	1.666	1.607	1.564	1.521	1.478
(0.5, 0.5)	1.670	1.611	1.552	1.509	1.466
(0.7, 0.3)	1.675	1.616	1.557	1.498	1.455
(0.9, 0.1)	1.680	1.621	1.562	1.503	1.443
(B2) Dataset 1 [row: (p_0^1, p_0^2)] vs Dataset 3 [column: (p_2^1, p_2^2)]					
$(p_0^1, p_0^2)/(p_2^1, p_2^2)$	(0.1, 0.9)	(0.3, 0.7)	(0.5, 0.5)	(0.7, 0.3)	(0.9, 0.1)
(0.1, 0.9)	1.692	1.610	1.528	1.446	1.364
(0.3, 0.7)	1.742	1.660	1.578	1.496	1.480
(0.5, 0.5)	1.792	1.710	1.627	1.612	1.596
(0.7, 0.3)	1.841	1.759	1.743	1.728	1.712
(0.9, 0.1)	1.891	1.875	1.859	1.843	1.828
(B3) Dataset 2 [row: (p_1^1, p_1^2)] vs Dataset 3 [column: (p_2^1, p_2^2)]					
$(p_1^1, p_1^2)/(p_2^1, p_2^2)$	(0.1, 0.9)	(0.3, 0.7)	(0.5, 0.5)	(0.7, 0.3)	(0.9, 0.1)
(0.1, 0.9)	1.282	1.371	1.461	1.550	1.639
(0.3, 0.7)	1.355	1.124	1.214	1.303	1.392
(0.5, 0.5)	1.427	1.197	0.967	1.056	1.145
(0.7, 0.3)	1.500	1.270	1.039	0.809	0.898
(0.9, 0.1)	1.573	1.342	1.112	0.882	0.651