Hilbert Sinkhorn Divergence for Optimal Transport

Qian Li[1*] Zhichao Wang [2*†] Gang Li[3] Jun Pang[4] Guandong Xu[1]

[1] Faculty of Engineering and Information Technology, University of Technology Sydney, Australia
[2] School of Electrical Engineering and Telecommunications, University of New South Wales, Australia
[3] Centre for Cyber Security Research and Innovation, Deakin University, Geelong, VIC 3216, Australia
[4] Faculty of Science, Technology and Medicine, University of Luxembourg

{qian.li, guandong.xu}@uts.edu.au, zchaoking@gmail.com, gang.li@deakin.edu.au, jun.pang@uni.lu

Abstract

The Sinkhorn divergence has become a very popular metric to compare probability distributions in optimal transport. However, most works resort to the Sinkhorn divergence in Euclidean space, which greatly blocks their applications in complex data with nonlinear structure. It is therefore of theoretical demand to empower the Sinkhorn divergence with the capability of capturing nonlinear structures. We propose a theoretical and computational framework to bridge this gap. In this paper, we extend the Sinkhorn divergence in Euclidean space to the reproducing kernel Hilbert space, which we term "Hilbert Sinkhorn divergence" (HSD). In particular, we can use kernel matrices to derive a closed form expression of the HSD that is proved to be a tractable convex optimization problem. We also prove several attractive statistical properties of the proposed HSD, i.e., strong consistency, asymptotic behavior and sample complexity. Empirically, our method yields state-of-the-art performances on image classification and topological data analysis.

1. Introduction

As an important tool to compare probability distributions, optimal transport theory [52] has found many successful applications in machine learning. Examples include generative modeling [56, 19], domain adaptation [17], dictionary learning [42], text mining [29], sampling [54, 55] and single-cell genomics [41]. Optimal transport aims at minimizing the cost of moving a source distribution to a target distribution. The minimal transportation cost defines a divergence between the two distributions, which is called the Wasserstein or Earth-Mover distance [51, 40]. Roughly speaking, the Wasserstein distance measures the minimal cost required to deform a distribution to another distribution. Different from other divergence, such as Kullback–Leibler divergence and the L_2 distance, the Wasserstein distance could compare probability distributions in a geometrically faithful manner. This entails a rich geometric structure on the space of probability distributions.

Related work. Existing optimal transport schemes can be mainly categorized into three classes. Methods in the first class are the regularization-based Wasserstein distance. Such numerical schemes add a regularization penalty to the original optimal transport problem. For instance, the Sinkhorn divergence [12, 13] provides a fast approximation to the Wasserstein distance by regularizing the original optimal transport with an entropy term. Greedy [2], Nystrom [1] and stochastic [18] versions of Sinkhorn algorithm with better empirical performance have also been explored. Other representative contributions towards regularization-based optimal transport include quantum regularization [35], sparse regularization [7] and Boltzmann-Shannon entropy [16].

An alternative principle for approximating the Wasserstein distance comes from Radon transform: to project high-dimensional distribution to one-dimensional distributions. One representative example is the sliced Wasserstein distance [8, 23, 14, 24], which is defined as the average Wasserstein distance obtained between random one dimensional projections. In other words, the sliced Wasserstein distance is calculated via linear slicing of the probability distribution. Its important extensions, such as [34, 31], are proposed recently to search for the k-dimensional subspace that would maximize the Wasserstein distance between two measures after projection. The sample complexity of such estimators is investigated [15, 14] between two measures and their empirical counterparts.

Methods in the third class include the Gromov-Wasserstein distance, it extends optimal transport to scenario where heterogeneous distributions are involved, i.e.,

*Equal contribution
†Corresponding author

distributions defined on different metric spaces. Such an approach was successfully applied for tasks where source and target samples do not lie in the same Euclidean space, e.g., for heterogeneous domain adaptation [57], shapes [46], word embedding [3] and generative modeling [9]. From the computational perspective, the Gromov-Wasserstein distance involves a non-convex quadratic problem, and it is hard to lift it to large scale settings [36]. Such a heavy computation burden could be remedied by the Sinkhorn divergence [37] or the sliced technique [50].

Motivations. All these works consider the optimal transport in original sample space (usually Euclidean space $\mathbb{R}^n$). However, various machine learning tasks are kernel dependent, such as computerized tomography [32], topological data analysis [25], geometric domain [45], kernel mean embedding [30] and adaptive Monte Carlo [43]. There is no straightforward way to formulate the optimal transport problem for such tasks. The performance of these tasks highly depends on comparing the distributions in reproducing kernel Hilbert spaces (RKHS) [48]. Thus defining the optimal transport in the original sample space may lead to sub-optimum performance for the above applications.

In fact, RKHS provides a platform for optimal transport in functional spaces to be applied in real-world problems. Recent relevant works can be found in [32, 58], where efforts have been devoted to the intersection of RKHS and optimal transport. But they involve a linear program so that evaluating the Wasserstein distance in RHKS is a computational bottleneck in general, and they did not deliver the convergence result such as *asymptotic behavior* or *sample complexity*. Although various techniques mentioned previously have been extensively considered Sinhorn divergence to accelerate the computation of Wasserstein distance, it is still non-trivial to extend the analysis for RKHS cases. This is mainly due to the fact that the convex relaxation of the Sinkhorn divergence is formulated through an implicit nonlinear map in infinite dimension, making it a challenge to optimize. Another important reason is that the theoretical convergence of optimal transport in RKHS has not been well studied. There are practical and theoretical demands to develop a general framework to analyze the theory between optimal transport and RKHS, which will favor various kernel-dependent tasks in machine learning.

Contributions. This paper makes two valuable contributions, including both practical formulations and theoretical convergence.

- We prove that the Sinkhorn divergence has a special structure in RKHS, which allows us to propose an equivalent and computable "Hilbert Sinkhorn divergence" (HSD) that could be fully determined by the kernel function (Thm. 1) and then could be computed via solving a convex optimization problem. Thm. 1 shows that the proposed HSD could be solved in an efficient manner using Sinkhorn iterations.
- We analyze the strong consistency and sample complexity of the proposed HSD. We give the error bound and prove the strong consistency when approximating original Wasserstein distances with our HSD (Thm. 3). We also focus on how the asymptotic behavior (Prop. 3) and sample complexity (Thm. 4) affect the convergence of the proposed HSD.

2. Preliminary

We begin with the background of optimal transportation. Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denote the cost function where $\mathcal{X} \in \mathbb{R}^n$ is the sample space. We define $\Pi(\mu, \nu)$ as the set of all probabilistic couplings π with marginals μ and ν. Formally, the Wasserstein distance is thus defined as

$$\mathcal{W}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x, y) d\pi(x, y) \tag{1}$$

Thus the Wasserstein distance aims to find a couplings π so as to minimize the cost function of moving a probability mass from μ to ν. The Wasserstein distance (1) is numerically intractable in general due to its high computational complexity. Consequently, the Sinkhorn divergence [12] is proposed to approximate (1) by regularizing the original problem with an entropy term.

Definition 1 (Sinkhorn divergence) *The Sinkhorn divergence is formally defined as:*

$$\mathcal{W}_\epsilon(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left[\int_{\mathcal{X}\times\mathcal{X}} c(x, y) d\pi(x, y) + \epsilon H(\pi) \right] \tag{2}$$

where $\epsilon > 0$ is a coefficient and the entropic regularization $H(\pi)$ is given below

$$H(\pi) = \log \left(\frac{d\pi}{d\mu\, d\nu}(x, y) \right) \tag{3}$$

The entropic regularization makes the Sinkhorn divergence (2) strictly convex to guarantee the unique minimizer. The Sinkhorn divergence allows an extremely simple iterative algorithm [12], which can be implemented using only matrix-vector products and converges quickly to a solution of (1). Note that the Sinkhorn divergence (2) is defined in the sample space. To further reformulate the Sinkhorn divergence in RKHS, we need Hilbert embedding to transform the probability measures from the sample space to RKHS.

Definition 2 (Hilbert embedding) *Let $\mathbb{P}(\mathcal{X})$ be the set of probability measures on sample set $\mathcal{X}$ and $\mathbb{P}(\mathcal{H})$ be the set of probability measures on reproducing kernel Hilbert*

space $\mathcal{H}$. Given a probability measure $\mu \in \mathbb{P}(\mathcal{X})$, the implicit feature map $\phi : \mathcal{X} \to \mathcal{H}$ will induce the Hilbert embedding of μ:

$$\phi_* : \mathbb{P}(\mathcal{X}) \to \mathbb{P}(\mathcal{H}),\ \mu \mapsto \phi_*\mu = \int_{\mathcal{X}} \phi(x) d\mu(x) \tag{4}$$

For the map $(\phi, \phi) : \mathcal{X} \times \mathcal{X} \to \mathcal{H} \times \mathcal{H}$, we similarly have

$$(\phi, \phi)_* : (\mu, \nu) \mapsto (\phi_*\mu, \phi_*\nu) \tag{5}$$

3. Hilbert Sinkhorn divergence

In this section, we introduce the nonlinear version of the Sinkhorn divergence in RKHS, which we term "Hilbert Sinkhorn divergence" (HSD). Section 3.1 provides an equivalent and computable formulation of HSD. In Section 3.2, we further discuss how HSD could be used to compare empirical probability measures.

3.1. Formulation

Let $\mathcal{X}$ be a sample space. A function $k : \mathcal{X} \times \mathcal{X}$ is called a kernel function of reproducing kernel Hilbert space $\mathcal{H}$, i.e., $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{H}$. The kernel function k satisfies the reproducing property

$$f \in \mathcal{H}, \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x) \tag{6}$$

Define the implicit feature map $\phi : \mathcal{X} \to \mathcal{H}$ as $x \mapsto \phi(x) = k(\cdot, x)$. Then we have kernel trick such that $\langle \phi(x), \phi(y)\rangle_{\mathcal{H}} = k(x, y)$.

In order to capture the nonlinearity of optimal transport, we employ a compact representation that not only preserves statistical properties of arbitrary distributions, but also permits efficient computation. That is, we adopt the Hilbert embedding [6, 44, 48] to represent the probability measure as a mean function in the RKHS.

Definition 3 (Hilbert Sinkhorn divergence, HSD) *Given measures $\mu, \nu \in \mathbb{P}(\mathcal{X})$ and elements $u, v \in \mathcal{H}$, the Hilbert Sinkhorn divergence between embedding $\phi_*\mu$ and $\phi_*\nu$ is written as*

$$\mathcal{S}_\epsilon(\phi_*\mu, \phi_*\nu) = \inf_{\pi_\phi} \int_{\mathcal{H}\times\mathcal{H}} c_\phi(u, v) d\pi_\phi(u, v) + \epsilon\Phi(\pi_\phi) \tag{7}$$

where $\pi_\phi \in \Pi(\phi_*\mu, \phi_*\nu)$ is a joint probability measure with two marginals $\phi_*\mu$ and $\phi_*\nu$, and

$$c_\phi(u, v) = \|u - v\|^2_{\mathcal{H}}$$
$$\Phi(\pi_\phi) = \log\left(\frac{d\pi_\phi}{d(\phi_*\mu)\, d(\phi_*\nu)}(u, v)\right)$$

HSD is a natural nonlinear version of the Sinkhorn divergence (2) that is extended to RKHS. Therefore, HSD can be solved by finding an optimal joint probability measure in $\mathcal{P}(\mathcal{H} \times \mathcal{H})$. However, as HSD in (7) is expressed by an unknown implicit nonlinear map ϕ, making it difficult to solve. Thus, we provide an equivalent and solvable formulation of (7), which is completely determined by the kernel function.

Theorem 1 *Given two measures $\mu, \nu \in \mathbb{P}(\mathcal{X})$, we write*

$$\mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu) = \inf_{\pi} \int_{\mathcal{X}\times\mathcal{X}} c_{\mathcal{H}}(x, y) d\pi(x, y) + \epsilon H(\pi) \tag{8}$$

where $\pi \in \Pi(\mu, \nu)$ is the joint probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals μ and ν, and

$$c_{\mathcal{H}}(x, y) = \|\phi(x) - \phi(y)\|^2_{\mathcal{H}} = k(x, x) + k(y, y) - 2k(x, y)$$
$$H(\pi) = \log\left(\frac{d\pi}{d\mu\, d\nu}(x, y)\right)$$

Then we have the following conclusions:

- $\mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu) = \mathcal{S}_\epsilon(\phi_*\mu, \phi_*\nu)$
- *If π^* is a minimizer of (8), its Hilbert embedding $(\phi, \phi)_*\pi^*$ is a minimizer of (7).*

Proof *See proof in the supplement.*

The reformulation (8) can be viewed as the Sinkhorn divergence in sample domain $\mathcal{X} \times \mathcal{X}$ with cost function $c_{\mathcal{H}}$ induced by the nonlinear kernel $k(\cdot, \cdot)$, e.g., RBF kernel.

The reformulation (8) improves (7) in two promising aspects: First, the integral domain in formulation (8) is transformed into a more tractable and computatble sample domain $\mathcal{X} \times \mathcal{X}$ rather than the Hilbert domain $\mathcal{H} \times \mathcal{H}$. This is a practically useful transformation as it effectively simplifies the problem from an infinite dimensional space to a finite dimensional space. Second, this reformulation is beneficial for establishing the strong consistency and sample complexity via leveraging standard tools from RKHS theory, as shown in Section 4.

Remark. We have a special case for Eq. (8) that enables a connection between HSD and the Wasserstein distance. That is, if $\epsilon = 0$, then $\mathcal{S}_{\mathcal{H},\epsilon}$ degenerates to the Wasserstein distance in RKHS.

$$\mathcal{W}_{\mathcal{H}}(\mu, \nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c_{\mathcal{H}}(x, y) d\pi(x, y) \tag{9}$$

3.2. Discrete Hilbert Sinkhorn divergence

In most applications, probability measures μ and ν are only discrete, i.e., empirical measures. Accordingly, we define a discrete case of our HSD for empirical measures

$$\mu_n = \sum_{i=1}^{n} \hat{\mu}_i \delta_{x_i}, \quad \nu_n = \sum_{j=1}^{n} \hat{v}_j \delta_{y_j} \tag{10}$$

where δ_{x_i} is the Dirac measure, and $\hat{\mu}_i$ is the probability of mass associated to x_i. Similarly, δ_{y_i} and $\hat{v}_j$ are defined for

y_j. Let $\mathcal{B}$ denote a set of probabilistic couplings between the two empirical measures:

$$\mathcal{B} = \left\{ \pi \in \left(\mathbb{R}^+\right)^{n \times n} \mid \pi \mathbf{1}_n = \mu, \pi^\top \mathbf{1}_n = \nu \right\} \tag{11}$$

where $\mathbf{1}_n$ is a n-dimensional vector of ones. Then the discrete version for HSD in (8) is

$$\pi^* = \underset{\pi \in \mathcal{B}}{\operatorname{argmin}} \langle \pi, K \rangle_F + \epsilon H(\pi) \tag{12}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius dot product, and K is the kernel matrix with entries defined on a kernel function k

$$K_{i,j} = k(x_i, x_i) + k(y_j, y_j) - 2k(x_i, y_j) \tag{13}$$

Apparently, the discrete HSD in (12) is a strict convex problem. The Sinkhorn algorithm [12, 36] enables this problem with a unique solution π^*, which is a very simple iterative algorithm involving only matrix-vector products.

4. Theoretical properties

In this section, we present several theoretical properties of HSD $\mathcal{S}_{\mathcal{H},\epsilon}$: *consistency*, *asymptotic behavior* and *sample complexity*. All the proofs are provided in the supplement.

4.1. Strong consistency

Proving the consistency of our HSD can be decomposed into two problems: *variational representation* and *approximation*. The first one leads to two propositions. Prop. 1 states that HSD admits the *variational representation*, which results in the lower bound in Prop. 2. We use Corollary 1 to prove the second problem in Thm. 2, which states the almost sure convergence of $\mathcal{S}_{\mathcal{H},\epsilon}(\mu_n, \nu_n)$ to $\mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu)$.

Proposition 1 (Variational representation) *The Hilbert Sinkhorn divergence (7) admits the following variational representation in the reproducing kernel Hilbert space:*

$$\mathcal{S}_\epsilon\left(\phi_*\mu, \phi_*\nu\right) = \epsilon \left(1 + \min_{\pi_\phi} \mathbb{E}_{\pi_\phi}[T] - \log\left(\mathbb{E}_{\xi_\phi}\left[e^T\right]\right)\right)$$

where coefficient $\epsilon > 0$, $\pi_\phi \in \Pi\left(\phi_*\mu, \phi_*\nu\right)$, $\xi_\phi(x, y) = e^{-\|u-v\|^2_{\mathcal{H}}/\epsilon}$, and $T = \log \frac{d\pi_\phi}{d\xi_\phi} + C$ for $C \in \mathbb{R}$.

The variational representation relies on T that is a mapping from product Hilbert space $\mathcal{H} \times \mathcal{H}$ to $\mathbb{R}$. Alternatively, we hope to restrict the result of Prop. 1 on a more solvable space $\mathcal{X} \times \mathcal{X}$. Prop. 2 will give a lower bound for our HSD between arbitrary measures in RKHS.

Proposition 2 (Lower bound) *The Hilbert Sinkhorn distance has the following lower bound:*

$$\mathcal{S}_\epsilon\left(\phi_*\mu, \phi_*\nu\right) \geq \epsilon \left(1 + \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi[k] - \log\left(\mathbb{E}_\xi\left[e^k\right]\right)\right)$$

where $\epsilon > 0$, $\phi_*\mu$ and $\phi_*\nu$ are Hilbert embedding in Eq. (4), and k is a kernel function.

More interestingly, Thm. 1 ensures that Prop. 1 and Prop. 2 are also valid for our HSD between two measures μ and ν, as summarized in the following corollary.

Corollary 1 *Given notations in Prop 1 and 2, the reformulation (8) admits the following variational representation and lower bound:*

$$\mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu) = \epsilon \left(1 + \min_{\pi_\phi} \mathbb{E}_{\pi_\phi}[T] - \log\left(\mathbb{E}_{\xi_\phi}\left[e^T\right]\right)\right)$$

$$\mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu) \geq \epsilon \left(1 + \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi[k] - \log\left(\mathbb{E}_\xi\left[e^k\right]\right)\right)$$

Results in Corollary 1 are the necessary conditions to prove that empirical Hilbert Sinkhorn divergence closes to its exact version in probability, i.e., $\mathcal{S}_{\mathcal{H},\epsilon}(\mu_n, \nu_n)$ is the strongly consistent estimator of $\mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu)$ as follows.

Theorem 2 (Strong consistency) *Given μ_n, ν_n defined in Eq. (10) and $\epsilon, \eta > 0$, there exists $N > 0$ such that*

$$\forall n \geq N,\ \mathbb{P}\left(\left|\mathcal{S}_{\mathcal{H},\epsilon}(\mu_n, \nu_n) - \mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu)\right| \leq \epsilon\eta\right) = 1$$

The strong consistency theorem states the almost sure convergence of empirical estimator $\mathcal{S}_{\mathcal{H},\epsilon}(\mu_n, \nu_n)$, but it relies on the sample number $n \geq N$. We will analyze the upper bound of sample number in Subsection 4.3.

4.2. Asymptotic behavior

In this section, we investigate the asymptotic properties of the proposed HSD (8) with respect to the Wasserstein distance (1). The following proposition states that $\mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu)$ can approximate the Wasserstein distance with the bound determined by the diameter and dimension of sample space.

Proposition 3 (Approximation error) *Define the sample space $\mathcal{X}$ as a subset of $\mathbb{R}^d$ and its diameter $|\mathcal{X}| = \sup\{\|x - y\| \mid x, y \in \mathcal{X}\}$, we have*

$$\begin{aligned} &\left|\mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu) - \mathcal{W}_\epsilon(\mu, \nu)\right| \leq \epsilon\eta \\ &\left|\mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu) - \mathcal{W}(\mu, \nu)\right| \leq \epsilon \left(\eta + 2d \log \frac{e^2 LD}{\sqrt{d}\epsilon}\right) \end{aligned} \tag{14}$$

where $\epsilon > 0$, $D \geq |\mathcal{X}|$ and L is a Lipschitz constant.

We are also interested in the asymptotic bound when approximating the Wasserstein distance $\mathcal{W}(\mu, \nu)$ with the discrete HSD. Combining Prop. 3 and Thm. 2 with the triangular inequality, we can have the following bound.

Theorem 3 (Asymptotic bound) *With the notations in Prop. 3, the discrete Hilbert Sinkhorn divergence $\mathcal{S}_{\mathcal{H},\epsilon}(\mu_n, \nu_n)$ approximates the Wasserstein distance $\mathcal{W}(\mu, \nu)$ with the following bound*

$$\forall n \geq N,\ \mathbb{P}\left(\left|\mathcal{S}_{\mathcal{H},\epsilon}(\mu_n, \nu_n) - \mathcal{W}(\mu, \nu)\right| \leq \zeta\right) = 1 \tag{15}$$

where $\zeta = 2\epsilon \left(\eta + d \log \frac{e^2 LD}{\sqrt{d}\epsilon}\right)$.

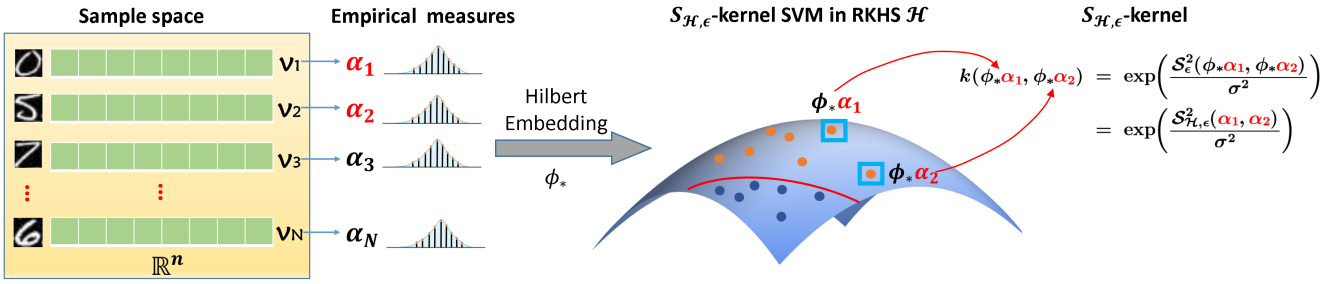

Figure 1. We map the vector v to an empirical measure α as in (18), and then use Hilbert embedding ϕ_* to transform α in reproducing kernel Hilbert space. By comparing the embedded measures $\phi_*(\alpha)$ in pairwise, we obtain a new kernel upon Hilbert Sinkhorn divergence $\mathcal{S}_{\mathcal{H},\epsilon}$, which can be used in any kernel-based learning machine, such as SVM.

4.3. Sample complexity

We will discuss the *sample complexity* of our HSD, i.e., how many samples we need for the discrete HSD at a desired accuracy of a high confidence. To answer this question, we rely on the well-known covering number for RKHS and attained the refinement Thm. 2 of Thm. 4. The lemma below shows there exists finite disks such that covering $\mathcal{H}$, and it will be useful to prove sample complexity in Thm. 4.

Lemma 1 *[59] We assume that arbitrary function $f \in \mathcal{H}$ is bounded (i.e., $\|f\|_{\mathcal{H}} \leq M$). Given the covering disk $B_\eta = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq \eta\}$, the covering number of $\mathcal{H}$ is $\mathcal{N}(\mathcal{H}, \eta) \leq \left(\frac{3M}{\eta}\right)^m$ where m is the number of basis that span the function f.*

The covering number measures the size of reproducing kernel Hilbert space with respect to the norm $\|\cdot\|_{\mathcal{H}}$. We will show how this quantity affects the sample complexity of the Hilbert Sinkhorn divergence $\mathcal{S}_{\mathcal{H},\epsilon}$.

Theorem 4 *Given the desired accuracy parameters $\eta, \epsilon > 0$ and the confidence parameter η, we have,*

$$\mathbb{P}\left(|\mathcal{S}_{\mathcal{H},\epsilon}(\mu,\nu) - \mathcal{S}_{\mathcal{H},\epsilon}(\mu_n,\nu_n)| \leq \epsilon\eta\right) \geq 1-\delta, \qquad (16)$$

whenever the number n of samples satisfies

$$n \geq \frac{2M^2(\log(2/\delta) + m\log(24M/\eta))}{\eta^2} \qquad (17)$$

where m and M are given in Lem. 1.

5. Applications

In this section, we apply our HSD on real-world tasks: image classification and topological data analysis.

5.1. Image classification

The whole process is depicted in Fig. 1. Each image is unfolded as a vector $v = (u_1, \cdots, u_n) \in \mathbb{R}^n$ that is further mapped to a empirical measure

$$\alpha = \sum_{i=1}^n a_i \delta_i \quad \text{with} \quad a_i = u_i/\|v\| \qquad (18)$$

where δ_i is the Dirac measure at position i such that the probability mass of α_i is 1. Once the probability distributions α are embedded as $\phi_*\alpha$ in RKHS, we then apply the HSD reformulation (8) to construct

$$\mathbf{S}_{\mathcal{H},\epsilon}\textbf{-kernel}: \quad k_{\mathcal{S}}(I_1, I_2) = k(\phi_*\alpha_1, \phi_*\alpha_2) \stackrel{\text{def}}{=} \exp\left(-\frac{\mathcal{S}^2_{\mathcal{H},\epsilon}(\alpha_1,\alpha_2)}{\sigma^2}\right) \qquad (19)$$

for two images I_1 and I_2 on the RKHS. We are able to apply $\mathcal{S}_{\mathcal{H},\epsilon}$-kernel in a straightforward fashion for SVM. This actually generalizes the usual kernel SVM such as RBF-SVM to $\mathcal{S}_{\mathcal{H},\epsilon}$-SVM that could deal with probability distributions in RKHS. Namely, this amounts to solving a SVM problem with $\mathcal{S}_{\mathcal{H},\epsilon}$-kernel.

For comparison, we also apply the traditional Sinkhorn divergence (2) to define $\mathcal{W}_\epsilon$ kernel

$$k_{\mathcal{W}}(I_1, I_2) = k(\alpha_1, \alpha_2) \stackrel{\text{def}}{=} \exp\left(-\frac{\mathcal{W}^2_\epsilon(\alpha_1,\alpha_2)}{\sigma^2}\right) \qquad (20)$$

5.2. Topological data analysis (TDA)

Topological data analysis extracts the topological features that are complementary to statistical quantities, which has found many applications in computer vision [4, 39, 25, 10, 26, 53, 47]. In TDA, persistence diagram (PD) is a favoured tool for describing topological feature, such as shape or graph. The space of PDs is a metric space with the *p-Wasserstein distance*, which is neither an Euclidean space nor a Hilbert space. Therefore, PDs cannot be directly used as inputs for a variety of machine learning methods, such as SVM. To analyze topological features encoded in PDs, recent efforts have been devoted to vectorizing PDs in the Hilbert space [4, 25, 39].

Motivated by these considerations, we propose a novel intuitive kernel for TDA based on our Hilbert Sinkhorn divergence $\mathcal{S}_{\mathcal{H},\epsilon}$. The whole process is depicted in Fig. 2. For

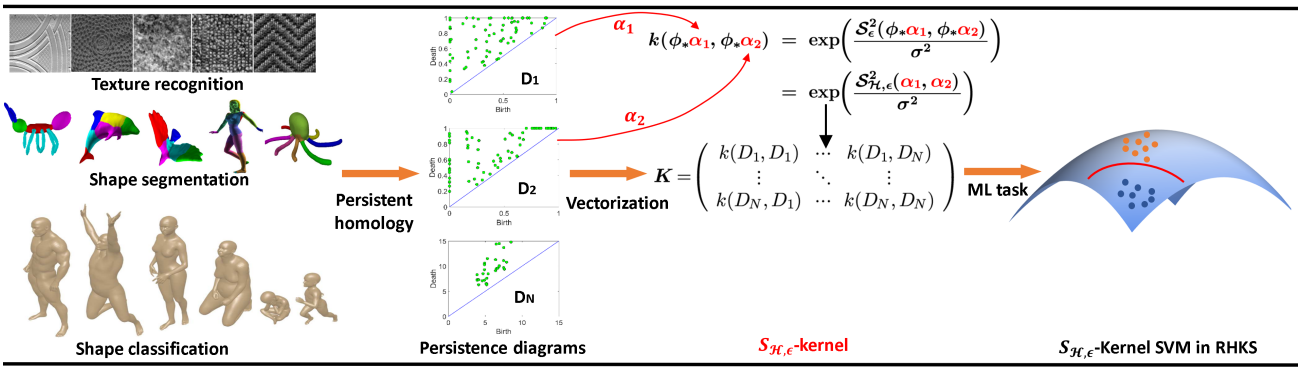

Figure 2. We represent the topological properties of visual data using the persistence diagram D_i. Next, the Hilbert embedding (22) maps persistence diagram D_i to $\phi_*\alpha_i$ in reproducing kernel Hilbert space. Finally, we define $\mathcal{S}_{\mathcal{H},\epsilon}$-kernel (23) for persistence diagrams to enable the downstream machine learning tasks.

the persistence diagram D, we introduce the measure

$$\alpha(D) := \sum_{x \in D} w(x)\delta_x \tag{21}$$

with a weight $w(x) > 0$ for each generator $x = (b, d) \in D$, where δ_x is the Dirac delta measure at x. We follow [25] to define a weight function $w(x) = \arctan(C \cdot (b-d)^p)$ with C and $p > 0$. As in Def. 2, the measure $\alpha(D)$ can be embedded in RKHS via

$$\alpha(D) \mapsto \phi_*\alpha(D) := \sum_{x \in D} w(x)k(\cdot, x) \tag{22}$$

Since $\phi_*\alpha(D)$ serves as a vector representation of the persistence diagram, we can apply $\mathcal{S}_{\mathcal{H},\epsilon}$ divergence to define a kernel over the vector representation. Similar to (19) for image classification, we consider the kernel

$$\mathbf{S}_{\mathcal{H},\epsilon}\textbf{-kernel}: \quad k_{\mathcal{S}}(D_1, D_2) = k(\phi_*\alpha_1, \phi_*\alpha_2) \overset{\text{def}}{=} \exp\left(-\frac{\mathcal{S}^2_{\mathcal{H},\epsilon}(\alpha_1, \alpha_2)}{\sigma^2}\right) \tag{23}$$

for two persistence diagrams D_1 and D_2 on the RKHS.

6. Experiments

6.1. Synthetic data

We study the behavior of the proposed HSD converging to the Wasserstein distance (1). Due to the approximation (14), we could have $\mathcal{S}_{\mathcal{H},\epsilon}(\mu,\nu) \to \mathcal{W}_\epsilon(\mu,\nu)$ if $\epsilon \to 0$. According to Def. (2), we also have $\mathcal{W}_\epsilon(\mu,\nu)$ converges to the Wasserstein distance, i.e., $\mathcal{W}_\epsilon(\mu,\nu) \to \mathcal{W}(\mu,\nu)$ if regularization $\epsilon \to 0$. So by the triangle inequality, the gap $|\mathcal{S}_{\mathcal{H},\epsilon} - \mathcal{W}|$ is expected to decrease as the parameter ϵ decreases. We use the universal kernel $k(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{\sigma^2}\right)$ with $\sigma = 0.1$ to construct the matrix K defined in (13). Fig. 4 plots $|\mathcal{S}_{\mathcal{H},\epsilon} - \mathcal{W}|/\mathcal{S}_{\mathcal{H},\epsilon}$ over 100 random pairs (x, y), where x and y are sampled from the uniform distribution $U(0, 1)$. This simulation reveals that the Hilbert Sinkhorn divergence $\mathcal{S}_{\mathcal{H},\epsilon}$ typically approximates the Wasserstein distance with a high accuracy when ϵ is less than 0.01.

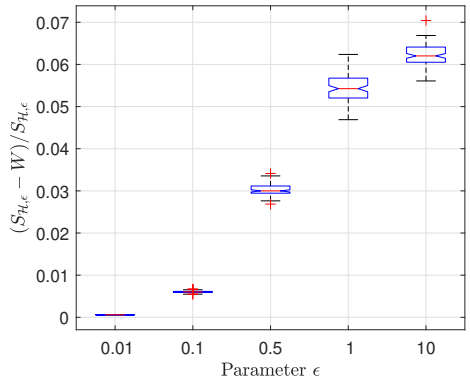

Figure 3. The gap between the HSD and the Wasserstein distance is a function of ϵ.

6.2. Image classification

We use the handwritten dataset USPS [21], which consists of 10 classes (i.e., from digital 0 to 9). The training and datasets are sampled uniformly at random. We fix the testing size as 1000 and vary the training size accordingly. Because identifying the handwritten digit is a multi-class classification problem, we use one-versus-one encoding.

Note that these two kernels (19) and (20) involve regularization parameter ϵ and width σ, we choose a small subset in the training dataset to tune ϵ in $\{10^{-2}, 0.1, 1, 10, 10^2\}$ and σ in $\{10^{-2}, 0.1, 0.2, 1, 5, 10\} \times M$, where M is the median of all the squared $\mathcal{S}_{\mathcal{H},\epsilon}$ divergence or $\mathcal{W}_\epsilon$ divergence. After computing two kernels (19) and (20) between any pair of images, we employ SVMs as the classifier for the image

classification. In our experiments, we compare three kernel methods based on RBF, Hilbert Sinkhorn divergence $\mathcal{S}_{\mathcal{H},\epsilon}$ and traditional Sinkhorn divergence $\mathcal{W}_\epsilon$. We use the universal kernel $k(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{\tau^2}\right)$ to construct the cost function $c_{\mathcal{H}}$ in (8), where τ is the median of squared Euclidean distances among all samples.

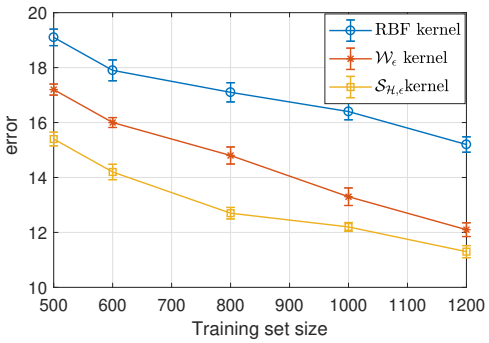

Figure 4. Average accuracy of SVM on USPS by using RBF, $\mathcal{W}_\epsilon$-kernel and $\mathcal{S}_{\mathcal{H},\epsilon}$-kernel.

Figure 4 reports the mean together with standard deviation of the classification error for all the compared methods. It is observed that $\mathcal{W}_\epsilon$ and $\mathcal{S}_{\mathcal{H},\epsilon}$ kernels outperform the RBF. In particular, $\mathcal{S}_{\mathcal{H},\epsilon}$ kernel yields the best performance when training size is varied. The out-performance attained by our method can be explained as follows. As indicated in Figure 1, our method can embed the image in sample space through $v \to \alpha \to \phi_*\alpha$ into the RKHS. Namely, rather than training on vectorial training examples, $\mathcal{S}_{\mathcal{H},\epsilon}$ kernel learns by using a collection of embedded probability distributions $\phi_*\alpha$. This can potentially incorporate higher-level statistical information that represents the discriminative features of images. Thus, $\mathcal{S}_{\mathcal{H},\epsilon}$ kernel should intuitively enhance the classification power.

6.3. Topological data analysis

We compare our **$\mathbf{S}_{\mathcal{H},\epsilon}$-kernel** (23) with several popular TDA methods on classical tasks: 3D shape analysis and texture recognition. Because the construction of persistence diagram (PD) is required for TDA methods, we construct their PDs of input data using the software `DIPHA` [5]. Parameter setting of all comparison methods will be first described.

PSS. Persistence scale space kernel [39] utilizes the multi-scale kernel function to map PD into Hilbert space. The kernel function is defined by the solutions of heat diffusion equation and has the expression: $k(D_1, D_2) = \frac{1}{8\pi t}\sum_{p\in D_1}\sum_{q\in D_2}\exp\left(-\frac{\|p-q\|^2}{8t}\right) - \exp\left(-\frac{\|p-\bar{q}\|^2}{8t}\right)$, where $\bar{q} = (d,b)$ is the symmetric of $q = (b,d)$ along the diagonal. Since there is no clear heuristic on how to tune t, we choose the kernel scale parameter t from $\{10^{-3}, 10^{-2}, 0.2, 1, 50, 10^2, 10^3\}$.

PSR. Persistence square-root representation [4] models a PD as a probability density function on the Hilbert Sphere. Then it discretizes the density function into a $K \times K$ grid. A smaller K reduces the accuracy, and a larger K improves the accuracy as well as the computational cost. Parameter K is chosen from the set $\{10, 15, 30, 50, 80\}$.

PWG. Then the persistence weighted Gaussian kernel [25] is defined as $k(D_1, D_2) = \exp\left(-\frac{\|\alpha_1-\alpha_2\|_{\mathcal{H}_\tau}}{\sigma^2}\right)$, where α is the embedding measure of PD defined by (22). We use cross-validation to chose τ and σ, each of which is taken values in 5 factors: 10^{-2}, 0.2, 1, 5 and 100. This leads to $5^2 = 25$ different possible grid values.

PWK. Persistence Wasserstein kernel uses $\mathcal{W}_\epsilon$ kernel (20) to classify the persist diagram. Measure α is defined as (21) with same parameters C and p as PWG. We choose cross-validation to tune ϵ in $\{10^{-2}, 0.1, 1, 10\}$ and σ in $\{10^{-2}, 0.1, 0.2, 1, 5, 10, 100\} \times M$ where M is the median of all the squared $\mathcal{W}_\epsilon$ distances.

PSK. Persistence Hilbert Sinkhorn kernel refers to the proposed $\mathcal{S}_{\mathcal{H},\epsilon}$-kernel in Eq. (23). We use universal kernel $k(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{\tau^2}\right)$ to construct symmetric matrix K in (13). Parameter τ is set to be the median of squared Euclidean distances among samples. Meanwhile, parameters ϵ, C, p and σ are set as same as PWK.

PHK. Persistence Hilbert Wasserstein kernel applies (9) to classify the persist diagram. As discussed in (9), PHK is just the special case of PSK under setting $\epsilon = 0$ while keeping other parameters invariant.

6.3.1 3D shape analysis

3D shape analysis uses sketch as input to retrieve 3D objects models, which mainly involves shape segmentation and shape classification.

Shape segmentation aims to design a classifier that assigns the class labels to different locations in a mesh shape. We use seven datasets of SHREC2010 [22] for both training and testing in shape segmentations, containing `ANT`, `FISH`, `BIRD`, `HUMAN`, `OCTOPUS` , `LIMB`, `BEAR`. Motivated by [11], we use the geodesic balls to construct 1-dimensional PD that characterizes the specific bumps in the shape. In particular, we construct the PD using the geodesic distance function on the shape.

Shape classification is performed on the 3D mesh benchmark dataset SHREC2014 [38] which consists of both synthetic (`SYN`) and real shapes (`REAL`). `SYN` contains 300 meshes from 15 classes of humans and `REAL` contains 400 meshes from 40 classes of humans. Shape classification aims to distinguish humans of different classes within `SYN` or `REAL` dataset. We use the popular *heat kernel signature* (HKS) [49] as the feature function for constructing 1-dimensional PDs [39, 26]. The time parameter t for HKS

Table 1. Classification performance (%) with different kernels for shape analysis.

DATA	PSS	PSR	PWG	PWK	PHK	PSK
ANT	85.3 ± 1.2	89.1 ± 0.5	88.2 ± 0.4	90.6 ± 0.3	91.3 ± 0.5	$\mathbf{92.7} \pm 0.3$
FISH	74.2 ± 1.5	75.7 ± 0.8	$\mathbf{79.3} \pm 0.6$	77.2 ± 0.7	75.6 ± 0.4	78.8 ± 0.6
BIRD	65.5 ± 2.0	67.2 ± 1.3	72.2 ± 1.4	71.8 ± 0.4	72.1 ± 0.2	$\mathbf{73.2} \pm 0.4$
HUMAN	69.8 ± 1.7	68.8 ± 0.7	69.5 ± 1.1	72.3 ± 0.6	73.2 ± 0.4	$\mathbf{74.3} \pm 0.5$
OCTOPUS	78.2 ± 1.5	77.1 ± 0.7	80.2 ± 0.6	84.4 ± 0.5	83.9 ± 0.5	$\mathbf{85.9} \pm 0.2$
LIMB	68.1 ± 1.9	66.0 ± 1.3	70.8 ± 0.5	72.8 ± 0.8	71.7 ± 0.6	$\mathbf{74.2} \pm 0.6$
BEAR	67.5 ± 1.3	67.6 ± 0.2	69.3 ± 0.7	$\mathbf{73.9} \pm 0.4$	72.8 ± 0.3	73.4 ± 0.6
SYN	97.3 ± 2.8	96.2 ± 2.2	97.2 ± 1.3	97.4 ± 1.9	97.2 ± 1.5	$\mathbf{98.1} \pm 0.7$
REAL	63.1 ± 1.7	60.1 ± 1.7	65.8 ± 1.7	66.3 ± 2.0	65.2 ± 1.9	$\mathbf{68.9} \pm 0.5$

function is set as a fixed value in $[0.005, 10]$, which controls the smoothness of the input data.

Results. We summarize the 3D shape analysis results in Tab. 1. The difference between the performance on SHREC2010 and SHREC2014 is consistent across all methods. Shape analysis on BIRD and LIMB is "hard", because there are many small prominent bumps in these shapes. Small prominent bumps having short persistences in PD may be mistaken for topological noise, which thus fools the training process resulting classification accuracy below 75% for all methods. Our persistence Hilbert Sinkhorn kernel PSK achieves the best accuracy in most cases, followed by PWG and PWK. The best shape segmentation accuracy of PSK is 92.7 ± 0.3 on ANT and the best shape classification result of PSK is 98.1 ± 0.7 on SYN. The variance of PSK is less than that of compared methods. Shape analysis results verify that $\mathcal{S}_{\mathcal{H}_\epsilon}$ metric extracts more preferable nonlinear and meaningful features from probability measures in RKHS when compared with other methods. Although PWG can preserve these features, it may lose some important statistical information when matching distributions by using the metric $\|\mu_1 - \mu_2\|_{\mathcal{H}}$ in RKHS.

6.3.2 Texture recognition

We use dataset OUTEX00000 [33] for texture recognition, including 480 texture images with 24 classes and 100 predefined training/testing splits. Following [39], texture images are downsampled to 32×32 images. We apply CLBP descriptors [20] to obtain the local region of an texture image, named as Sign (CLBP-S) and Magnitude (CLBP-M). Then we construct PDs for the component CLBP-S or CLBP-M.

Texture recognition results are reported in Tab. 2. Our kernel (PSK) outperforms all comparison methods. Although CLBP is sensitive to noise [20] and thus results in the perturbations in PDs, our PSK can remedy such perturbations via the higher-level statistical information encoded in RKHS probability measures $\phi_*\alpha$. However, the worst performance of PSR confirms that the Hilbert sphere manifold is not robust to such perturbations. Notice that $\mathcal{W}_{\mathcal{H}}$-kernel PHK achieves higher recognition rates than $\mathcal{W}_\epsilon$-kernel PWK and weighted Guassian kernel PWG but less than $\mathcal{S}_{\mathcal{H}_\epsilon}$ kernel PSK. This verifies that $\mathcal{S}_{\mathcal{H}_\epsilon}$ metric is more favorable to extract the discriminative non-linear feature representation, which can obviously improve the classification performance.

Table 2. Texture recognition (%) with different kernels.

METHODS	CLBP-S	CLBP-M
PSS	70.5 ± 2.9	56.2 ± 2.3
PSR	68.4 ± 1.6	54.3 ± 0.9
PWG	73.1 ± 1.3	59.6 ± 1.8
PWK	72.2 ± 1.2	57.3 ± 1.2
PHK	73.8 ± 1.0	60.3 ± 1.4
PSK	$\mathbf{75.3} \pm 1.0$	$\mathbf{62.3} \pm 1.4$

7. Conclusion

In this paper, we present a novel computational framework, i.e., Hilbert Sinkhorn divergence (HSD), to compare distributions in RKHS. We proved that it is theoretically robust due to strong consistency, asymptotic behavior and sample complexity. Our approach can be naturally extended to other kernel dependent machine learning tasks such as metric learning, domain adaptation and manifold learning. Moreover, it has great potential to succeed in non-vectorial data (e.g., graph or diagram) by a valid $\mathcal{S}_{\mathcal{H},\epsilon}$-kernel. While HSD can increase the accuracy of classification tasks, it is noted that the training also requires extra time. Our future work will consider the scalable Sinkhorn [1] via the Nystrom method to accelerate the computation, and investigate optimal transport in other non-Euclidean space, such as the low rank manifold [28] and Grassmannian manifold [27].

Acknowledgment: This work was supported by the Australian Research Council under Grant DP200101374 and Grant LP170100891.

References

[1] J. Altschuler, F. Bach, A. Rudi, and J. Niles-Weed. Massively scalable sinkhorn distances via the nyström method. In *Advances in Neural Information Processing Systems*, pages 4427–4437, 2019. 1, 8

[2] J. Altschuler, J. Niles-Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in neural information processing systems*, pages 1964–1974, 2017. 1

[3] D. Alvarez-Melis and T. Jaakkola. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, 2018. 2

[4] R. Anirudh, V. Venkataraman, K. Natesan Ramamurthy, and P. Turaga. A riemannian framework for statistical analysis of topological persistence diagrams. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 68–76, 2016. 5, 7

[5] U. Bauer, M. Kerber, and J. Reininghaus. Distributed computation of persistent homology. In *2014 Proceedings of the Sixteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 31–38. SIAM, 2014. 7

[6] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011. 3

[7] M. Blondel, V. Seguy, and A. Rolet. Smooth and sparse optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 880–889, 2018. 1

[8] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015. 1

[9] C. Bunne, D. Alvarez-Melis, A. Krause, and S. Jegelka. Learning generative models across incomparable spaces. *arXiv preprint arXiv:1905.05461*, 2019. 2

[10] M. Carriere, M. Cuturi, and S. Oudot. Sliced wasserstein kernel for persistence diagrams. In *International Conference on Machine Learning*, pages 664–673. PMLR, 2017. 5

[11] M. Carrière, S. Y. Oudot, and M. Ovsjanikov. Stable topological signatures for points on 3d shapes. In *Computer Graphics Forum*, volume 34, pages 1–12. Wiley Online Library, 2015. 7

[12] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013. 1, 2, 4

[13] M. Cuturi and G. Peyré. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016. 1

[14] I. Deshpande, Y.-T. Hu, R. Sun, A. Pyrros, N. Siddiqui, S. Koyejo, Z. Zhao, D. Forsyth, and A. G. Schwing. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10648–10656, 2019. 1

[15] I. Deshpande, Z. Zhang, and A. G. Schwing. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3483–3491, 2018. 1

[16] A. Dessein, N. Papadakis, and J.-L. Rouas. Regularized optimal transport and the rot mover's distance. *The Journal of Machine Learning Research*, 19(1):590–642, 2018. 1

[17] R. Flamary, N. Courty, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, 2016. 1

[18] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *Advances in neural information processing systems*, pages 3440–3448, 2016. 1

[19] A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018. 1

[20] Z. Guo, L. Zhang, and D. Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663, 2010. 8

[21] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994. 6

[22] E. Kalogerakis, A. Hertzmann, and K. Singh. Learning 3d mesh segmentation and labeling. *Acm Transactions on Graphics*, 29(4):1–12, 2010. 7

[23] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde. Generalized sliced wasserstein distances. In *Advances in Neural Information Processing Systems*, pages 261–272, 2019. 1

[24] S. Kolouri, Y. Zou, and G. K. Rohde. Sliced wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016. 1

[25] G. Kusano, Y. Hiraoka, and K. Fukumizu. Persistence weighted gaussian kernel for topological data analysis. In *International Conference on Machine Learning*, pages 2004–2013, 2016. 2, 5, 6, 7

[26] C. Li, M. Ovsjanikov, and F. Chazal. Persistence-based structural recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1995–2002, 2014. 5, 7

[27] Q. Li, W. Niu, G. Li, Y. Cao, J. Tan, and L. Guo. Lingo: linearized grassmannian optimization for nuclear norm minimization. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 801–809, 2015. 8

[28] Q. Li and Z. Wang. Riemannian submanifold tracking on low-rank algebraic variety. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 8

[29] T. Lin, Z. Hu, and X. Guo. Sparsemax and relaxed wasserstein for topic sparsity. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 141–149, 2019. 1

[30] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *arXiv preprint arXiv:1605.09522*, 2016. 2

[31] J. Niles-Weed and P. Rigollet. Estimation of wasserstein distances in the spiked transport model. *arXiv preprint arXiv:1909.07513*, 2019. 1

[32] J. H. Oh, M. Pouryahya, A. Iyer, A. P. Apte, J. O. Deasy, and A. 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*, page 103731, 2020. 2

[33] T. Ojala, T. Mäenpää, M. Pietikäinen, J. Viertola, J. Kyllönen, and S. Huovinen. Outex - new framework for empirical evaluation of texture analysis algorithms. In *International Conference on Pattern Recognition, 2002. Proceedings*, pages 701–706 vol.1, 2002. 8

[34] F.-P. Paty and M. Cuturi. Subspace robust wasserstein distances. *arXiv preprint arXiv:1901.08949*, 2019. 1

[35] G. Peyré, L. Chizat, F.-X. Vialard, and J. Solomon. Quantum entropic regularization of matrix-valued optimal transport. *European Journal of Applied Mathematics*, 30(6):1079–1102, 2019. 1

[36] G. Peyré, M. Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 2, 4

[37] G. Peyré, M. Cuturi, and J. Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672, 2016. 2

[38] D. Pickup, X. Sun, P. L. Rosin, R. R. Martin, Z. Cheng, Z. Lian, M. Aono, A. B. Hamza, A. Bronstein, and M. Bronstein. Shape retrieval of non-rigid 3d human models. In *Eurographics Workshop on 3d Object Retrieval*, pages 101–110, 2014. 7

[39] J. Reininghaus, S. Huber, U. Bauer, and R. Kwitt. A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4741–4748, 2015. 5, 7, 8

[40] F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015. 1

[41] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019. 1

[42] M. A. Schmitz, M. Heitz, N. Bonneel, F. Ngole, D. Coeurjolly, M. Cuturi, G. Peyré, and J.-L. Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018. 1

[43] D. Sejdinovic, H. Strathmann, M. L. Garcia, C. Andrieu, and A. Gretton. Kernel adaptive metropolis-hastings. In *International conference on machine learning*, pages 1665–1673, 2014. 2

[44] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007. 3

[45] J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015. 2

[46] J. Solomon, G. Peyré, V. G. Kim, and S. Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (TOG)*, 35(4):1–13, 2016. 2

[47] A. Som, K. Thopalli, K. N. Ramamurthy, V. Venkataraman, A. Shukla, and P. Turaga. Perturbation robust representations of topological persistence diagrams. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 617–635, 2018. 5

[48] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010. 2, 3

[49] J. Sun, M. Ovsjanikov, and L. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009. 7

[50] V. Titouan, R. Flamary, N. Courty, R. Tavenard, and L. Chapel. Sliced gromov-wasserstein. In *Advances in Neural Information Processing Systems*, pages 14753–14763, 2019. 2

[51] C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003. 1

[52] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008. 1

[53] Z. Wang, Q. Li, G. Li, and G. Xu. Polynomial representation for persistence diagram. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6123–6132, 2019. 5

[54] Z. Wang and V. Solo. Lie group state estimation via optimal transport. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5625–5629. IEEE, 2020. 1

[55] Z. Wang and V. Solo. Particle filtering on the stiefel manifold with optimal transport. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 4111–4116. IEEE, 2020. 1

[56] J. Wu, Z. Huang, D. Acharya, W. Li, J. Thoma, D. P. Paudel, and L. V. Gool. Sliced wasserstein generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3713–3722, 2019. 1

[57] Y. Yan, W. Li, H. Wu, H. Min, M. Tan, and Q. Wu. Semi-supervised optimal transport for heterogeneous domain adaptation. In *IJCAI*, pages 2969–2975, 2018. 2

[58] Z. Zhang, M. Wang, and A. Nehorai. Optimal transport in reproducing kernel hilbert spaces: Theory and applications. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[59] D.-X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49(7):1743–1752, 2003. 5