Sliced Wasserstein Kernels for Probability Distributions

Soheil Kolouri
Carnegie Mellon University
skolouri@andrew.cmu.edu

Yang Zou
Carnegie Mellon University
yzou2@andrew.cmu.edu

Gustavo K. Rohde
Carnegie Mellon University
gustavor@cmu.edu

Abstract

Optimal transport distances, otherwise known as Wasserstein distances, have recently drawn ample attention in computer vision and machine learning as powerful discrepancy measures for probability distributions. The recent developments on alternative formulations of the optimal transport have allowed for faster solutions to the problem and have revamped their practical applications in machine learning. In this paper, we exploit the widely used kernel methods and provide a family of provably positive definite kernels based on the Sliced Wasserstein distance and demonstrate the benefits of these kernels in a variety of learning tasks. Our work provides a new perspective on the application of optimal transport flavored distances through kernel methods in machine learning tasks.

1. Introduction

Many computer vision algorithms rely on characterizing images or image features as probability distributions which are often high-dimensional. This is for instance the case for histogram-based methods like the Bag-of-Words (BoW) [21], feature matching [17], co-occurence matrices in texture analysis [15], action recognition [43], and many more. Having an adequate measure of similarity (or equivalently discrepancy) between distributions becomes crucial in these applications. The classic distances or divergences for probability densities include Kullback-Leibler divergence, Kolmogorov distance, Bhattacharyya distance (also known as the Hellinger distance), etc. More recently, however, the optimal transportation framework and the Wasserstein distance [41] also known as the Earth Mover Distance (EMD) [34] have attracted ample interest in the computer vision [39], machine learning [11], and biomedical image analysis [3] communities. The Wasserstein distance computes the optimal warping to map a given input probability measure μ to a second one ν. The optimality corresponds to a cost function which measures the expected value of the displacement in a warping. Informally, thinking about μ and ν as piles of dirt (or sand), the Wasserstein distance measures

a notion of displacement of each sand particle in μ times its mass to warp μ into ν.

The Wasserstein distance has been shown to provide a useful tool to quantify the geometric discrepancy between two distributions. Specifically, they've been used as distances in content-based image image retrieval [34], modeling and visualization of modes of variation of image intensity values [3, 45, 37, 6], estimate the mean of a family of probability measures (i.e. barycenters of distributions) [2, 32], cancer detection [30, 40], super-resolution [23], amongst other applications. Recent advances utilizing variational minimization [16, 9], particle approximation [44], multi-scale schemes [27, 29], and entropy regularization [11, 39, 4], have enabled transport metrics to be efficiently applied to machine learning and computer vision problems. In addition, Wang et al. [45] described a method for computing a transport distance (denoted as linear optimal transport) between all image pairs of a dataset of N images that requires only N transport minimization problems. Rabin et al. [32] and Bonneel et al. [7] proposed to leverage the fact that these problems are easy to solve for one-dimensional distributions, and introduced an alternative distance called the Sliced Wasserstein distance. Finally, recent work [31, 22, 24] has shown that the transport framework can be used as an invertible signal/image transformation framework that can render signal/image classes more linearly separable, thus facilitating a variety of pattern recognition tasks.

Due to the benefits of using the transport distances outlined above, and given the flexibility and power of kernel-based methods [18], several methods using transport-related distances in constructing kernel matrices have been described with applications in computer vision, and EEG data classification [47, 12]. Since positive definite RBF kernels require the metric space induced by the distance to be 'flat' (zero curvature) [13], and because the majority of the transport-related distances mentioned above, in particular distances for distributions of dimension higher than one utilizing the L_2 cost, do not satisfy this requirement, few options for provably positive definite transport-based kernels have emerged. Cuturi [10] for example, suggested utilizing

the permanent of the transport polytope, thus guaranteeing the positivity of the derived kernels. More recently, Gardner et al. [14] have shown that certain types of earth mover's distances (e.g. with the 0-1 cost function) can yield kernels which are positive definite.

Here we expand upon these sets of ideas and show that the Sliced Wasserstein distance satisfies the basic requirements for being used as a positive definite kernel[18] in a variety of regression-based pattern recognition tasks, and have concrete theoretical and practical advantages. Based on the recent works on kernel methods [19, 20, 13], we exploit the connection between the optimal transport framework and the kernel methods and introduce a family of provably positive definite kernels which we denote as Sliced Wasserstein kernels. We derive mathematical results enabling the application of the Sliced Wasserstein metric in the kernel framework and, in contrast to other work, we describe the explicit form for the embedding of the kernel, which is analytically invertible. Finally, we demonstrate experimentally advantages of the Sliced Wasserstein kernels over the commonly used kernels such as the radial basis function (RBF) and the polynomial kernels for a variety of regression.

Paper organization. We first review the preliminaries and formally present the p-Wasserstein distance, the Sliced Wasserstein distance, and review some of the theorems in the literature on positive definiteness of kernels in Section 2. The main theorems of the paper on the Sliced Wasserstein kernels are stated in Section 3. In Section 4 we review some of the kernel-based algorithms including the kernel k-means clustering, the kernel PCA, and the kernel SVM. Section 5 demonstrates the benefits of the Sliced Wasserstein kernel over the commonly used kernels in a variety of pattern recognition tasks. Finally we conclude our work in Section 6 and lay out future directions for research in the area.

2. Background

2.1. The p-Wasserstein distance

Let σ and μ be two probability measures on measurable spaces X and Y and their corresponding probability density functions I_0 and I_1, $d\sigma(x) = I_0(x)dx$ and $d\mu(y) = I_1(y)dy$. In computer vision and image processing applications one often deals with compact d-dimensional Euclidean spaces, hence $X = Y = [0,1]^d$.

Definition 1. The p-Wasserstein distance for $p \in [1, \infty)$ is defined as,

$$W_p(\sigma, \mu) := (\inf_{\pi \in \Pi(\sigma, \mu)} \int_{X \times Y} (x - y)^p d\pi(x, y))^{\frac{1}{p}} \quad (1)$$

where $\Pi(\sigma, \mu)$ is the set of all transportation plans, and $\pi \in$ $\Pi(\sigma, \mu)$, that satisfies the following,

$$\begin{aligned} \pi(A \times Y) = \sigma(A) & \quad \text{for any Borel subset } A \subseteq X \\ \pi(X \times B) = \mu(B) & \quad \text{for any Borel subset } B \subseteq Y \end{aligned} \quad (2)$$

Due to Brenier's theorem [8], for absolutely continuous probability measures σ and μ (with respect to Lebesgue measure) the p-Wasserstein distance can be equivalently obtained from,

$$W_p(\sigma, \mu) = (\inf_{f \in MP(\sigma, \mu)} \int_X (f(x) - x)^p d\sigma(x))^{\frac{1}{p}} \quad (3)$$

where, $MP(\sigma, \mu) = \{f : X \to Y \mid f_\# \sigma = \mu\}$ and $f_\# \sigma$ represents the pushforward of measure σ and is characterized as,

$$\int_{f^{-1}(A)} d\sigma = \int_A d\mu \text{ for any Borel subset } A \subseteq Y$$

In the one-dimensional case, the 2-Wasserstein distance has a closed form solution as the mass preserving (MP) transport map, $f \in MP$, is unique. We will show this in the following theorem.

Theorem 1. *Let σ and μ be absolutely continuous probability measures on $\mathbb{R}$ with corresponding positive density functions I_0 and I_1, and corresponding cumulative distribution functions $F_\sigma(x) := \sigma((-\infty, x])$ and $F_\mu(x) := \mu((-\infty, x])$. Then, there only exists one monotonically increasing transport map $f : \mathbb{R} \to \mathbb{R}$ such that $f \in MP(\sigma, \mu)$ and it is defined as,*

$$f(x) := min\{t \in \mathbb{R} : F_\mu(t) \geq F_\sigma(x)\} \quad (4)$$

or equivalently $f(x) = F_\mu^{-1}(F_\sigma(x))$. The 2-Wasserstein distance is then obtained from,

$$W_2(\sigma, \mu) = (\int_X (f(x) - x)^2 I_0(x)dx)^{\frac{1}{2}}. \quad (5)$$

Note that throughout the paper we use $W_2(\sigma, \mu)$ and $W_2(I_0, I_1)$ interchangeably.

Proof. Assume there exist more than one transport maps, say f and g, such that $f, g \in MP(\sigma, \mu)$, then we can write,

$$\int_{-\infty}^{f(x)} I_1(\tau)d\tau = \int_{-\infty}^{x} I_0(\tau)d\tau = \int_{-\infty}^{g(x)} I_1(\tau)d\tau$$

Above is equivalent to $F_\mu(f(x)) = F_\mu(g(x))$, but I_1 is positive everywhere, hence the CDF is monotonically increasing, therefore $F_\mu(f(x)) = F_\mu(g(x))$ implies that $f(x) = g(x)$. Therefore the transport map in one dimension is unique. □

The closed-form solution of the Wasserstein distance in one dimension is an attractive property, as it alleviates the need for often computationally intensive optimizations. Recently there has been some work on utilizing this property of the Wasserstein distance to higher dimensional problems [32, 7, 22] (i.e. images). We review such distances in the following section.

2.2. The Sliced Wasserstein distance

The idea behind the Sliced Wasserstein distance is to first obtain a family of one-dimensional representations for a higher-dimensional probability distribution through projections/slicing, and then calculate the distance between two input distributions as a functional on the Wasserstein distance of their one-dimensional representations. In this sense, the distance is obtained by solving several one-dimensional optimal transport problems, which have closed-form solutions.

Definition 2. The d-dimensional Radon transform $\mathscr{R}$ maps a function $I \in L^1(\mathbb{R}^d)$ where $L^1(\mathbb{R}^d) := \{I : \mathbb{R}^d \to \mathbb{R} | \int_{\mathbb{R}^d} |I(x)| dx \leq \infty\}$ into the set of its integrals over the hyperplanes of $\mathbb{R}^n$ and is defined as,

$$\mathscr{R}I(t, \theta) := \int_{\mathbb{R}} I(t\theta + \gamma\theta^\perp) d\gamma \qquad (6)$$

here $\theta^\perp$ is the subspace or unit vector orthogonal to θ. Note that $\mathscr{R} : L^1(\mathbb{R}^d) \to L^1(\mathbb{R} \times \mathbb{S}^{d-1})$, where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$.

We note that the Radon transform is an invertible, linear transform and we denote its inverse as $\mathscr{R}^{-1}$. For brevity we do not define the inverse Radon transform here, but the details can be found in [28]. Next, following [32, 7, 22] we define the Sliced Wasserstein distance.

Definition 3. Let μ and σ be two continuous probability measures on $\mathbb{R}^d$ with corresponding positive probability density functions I_1 and I_0. The Sliced Wasserstein distance between μ and σ is defined as,

$$
\begin{aligned}
SW(\mu, \sigma) &:= \left(\int_{\mathbb{S}^{d-1}} W_2^2(\mathscr{R}I_1(., \theta), \mathscr{R}I_0(., \theta)) d\theta\right)^{\frac{1}{2}} \\
&= \left(\int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} (f_\theta(t) - t)^2 \mathscr{R}I_0(t, \theta) dt d\theta\right)^{\frac{1}{2}}
\end{aligned}
$$
$$(7)$$

where f_θ is the MP map between $\mathscr{R}I_0(., \theta)$ and $\mathscr{R}I_1(., \theta)$ such that,

$$\int_{-\infty}^{f_\theta(t)} \mathscr{R}I_1(\tau, \theta) d\tau = \int_{-\infty}^{t} \mathscr{R}I_0(\tau, \theta) d\tau, \ \forall \theta \in \mathbb{S}^{d-1} \quad (8)$$

or equivalently in the differential form,

$$\frac{\partial f_\theta(t)}{\partial t} \mathscr{R}I_1(f_\theta(t), \theta) = \mathscr{R}I_0(t, \theta), \ \forall \theta \in \mathbb{S}^{d-1}. \qquad (9)$$

The Sliced Wasserstein distance as defined above is symmetric, and it satisfies subadditivity and coincidence axioms, and hence it is a true metric (See [22] for proof).

2.3. The Gaussian RBF Kernel on Metric Spaces

The kernel methods and specifically the Gaussian RBF kernel have shown to be very powerful tools in a plethora of applications. The Gaussian RBF kernel was initially designed for Euclidean spaces, however, recently there has been several work extending the Gaussian RBF kernel to other metric spaces. Jayasumana et al. [20], for instance, developed an approach to exploit the Gaussian RBF kernel method on Riemannian manifolds. In another interesting work, Feragen et al. [13] considered the Gaussian RBF kernel on general geodesic metric spaces and showed that the geodesic Gaussian kernel is only positive definite when the underlying Riemannian manifold is flat or in other words it is isometric to a Euclidean space. Here we review some definitions and theorems that will be used in the rest of the paper.

Definition 4. A positive definite (PD) (resp. conditional negative definite) kernel on a set M is a symmetric function $k : M \times M \to \mathbb{R}$, $k(I_i, I_j) = k(I_j, I_i)$ for all $I_i, I_j \in M$, such that for any $n \in N$, any elements $I_1, ..., I_n \in X$, and numbers $c_1, ..., c_n \in \mathbb{R}$, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(I_i, I_j) \geq 0 \quad (\text{resp.} \leq 0) \qquad (10)$$

with the additional constraint of $\sum_{i=1}^{n} c_i = 0$ for the conditionally negative definiteness.

Above definition is used in the following important theorem due to Isaac J. Schoenberg [35],

Theorem 2. *Let M be a nonempty set and $f : (M \times M) \to \mathbb{R}$ be a function. Then kernel $k(I_i, I_j) = exp(-\gamma f(I_i, I_j))$ for all $I_i, I_j \in M$ is positive definite for all $\gamma > 0$ if and only if $f(., .)$ is conditionally negative definite.*

The detailed proof of above theorem can be found in Chapter 3, Theorem 2.2 of [5].

Following the work by Jayasumana et al. [20], here we state the theorem (Theorem 6.1 in [20]) which provides the necessary and sufficient conditions for obtaining a positive definite Gaussian kernel from a given distance function defined on a generic metric space.

Theorem 3. *Let (M, d) be a metric space and define $k : M \times M \to \mathbb{R}$ by $k(I_i, I_j) := exp(-\gamma d^2(I_i, I_j))$ for all $I_i, I_j \in M$. Then $k(., .)$ is a positive definite kernel for all $\gamma > 0$ if and only if there exists an inner product space $\mathcal{V}$ and a function $\psi : M \to \mathcal{V}$ such that $d(I_i, I_j) = \|\psi(I_i) - \psi(I_j)\|_{\mathcal{V}}$.*

Proof. The detailed proof is presented in [20]. The gist of the proof, however, follows from Theorem 2 which states that positive definiteness of $k(.,.)$ for all $\gamma > 0$ and conditionally negative definiteness of $d^2(.,.)$ are equivalent conditions. Hence, for $d(I_i, I_j) = \|\psi(I_i) - \psi(I_j)\|_{\mathcal{V}}$ it is straightforward to show that $d^2(.,.)$ is conditionally negative definite and therefore $k(.,.)$ is positive definite. On the other hand, if $k(.,.)$ is positive definite then $d^2(.,.)$ is conditionally negative definite, and since $d(I_i, I_i) = 0$ for all $I_i \in M$ a vector space $\mathcal{V}$ exists for which $d(I_i, I_j) = \|\psi(I_i) - \psi(I_j)\|_{\mathcal{V}}$ for some $\psi : M \to \mathcal{V}$ [20, 5]. $\square$

3. Sliced Wasserstein Kernels

In this section we present our main theorems. We first demonstrate that the Sliced Wasserstein Gaussian kernel of probability measures is a positive definite kernel. We proceed our argument by showing that there is an explicit formulation for the nonlinear mapping to the kernel space (aka feature space) and define a family of kernels based on this mapping.

We start by proving that for one-dimensional probability density functions the 2-Wasserstein Gaussian kernel is a positive definite kernel.

Theorem 4. *Let M be the set of absolutely continuous one-dimensional positive probability density functions and define $k : M \times M \to \mathbb{R}$ to be $k(I_i, I_j) := exp(-\gamma W_2^2(I_i, I_j))$ then $k(.,.)$ is a positive definite kernel for all $\gamma > 0$.*

Proof. In order to be able to show this, we first show that for absolutely continuous one-dimensional positive probability density functions there exists an inner product space $\mathcal{V}$ and a function $\psi : M \to \mathcal{V}$ such that $W_2(I_i, I_j) = \|\psi(I_i) - \psi(I_j)\|_{\mathcal{V}}$.

Let σ, μ, and ν be probability measures on $\mathbb{R}$ with corresponding absolutely continuous positive density functions I_0, I_1, and I_2. Let $f, g, h : \mathbb{R} \to \mathbb{R}$ be transport maps such that $f_\# \sigma = \mu$, $g_\# \sigma = \nu$, and $h_\# \mu = \nu$. In the differential form this is equivalent to $f' I_1(f) = g' I_2(g) = I_0$ and $h' I_2(h) = I_1$ where $I_1(f)$ represents $I_1 \circ f$. Then we have,

$$W_2^2(I_1, I_0) = \int_{\mathbb{R}} (f(x) - x)^2 I_0(x) dx,$$
$$W_2^2(I_2, I_0) = \int_{\mathbb{R}} (g(x) - x)^2 I_0(x) dx,$$
$$W_2^2(I_2, I_1) = \int_{\mathbb{R}} (h(x) - x)^2 I_1(x) dx.$$

We follow the work of Wang et al. [45] and Park et al. [31] and define a nonlinear map with respect to a fixed probability measure, σ with corresponding density I_0, that maps an input probability density to a linear functional on the corresponding transport map. More precisely, $\psi_\sigma(I_1(.)) := (f(.) - id(.))\sqrt{I_0(.)}$ where $id(.)$ is the identity map and

$f' I_1(f) = I_0$. Notice that such ψ_σ maps the fixed probability density I_0 to zero, $\psi_\sigma(I_0(.)) = (id(.) - id(.))\sqrt{I_0(.)} = 0$ and it satisfies,

$$W_2(I_1, I_0) = \|\psi_\sigma(I_1)\|_2$$
$$W_2(I_2, I_0) = \|\psi_\sigma(I_2)\|_2.$$

More importantly, we demonstrate that $W_2(I_2, I_1) = \|\psi_\sigma(I_1) - \psi_\sigma(I_2)\|_2$. To show this, we can write,

$$
\begin{aligned}
W_2^2(I_2, I_1) &= \int_{\mathbb{R}} (h(x) - x)^2 I_1(x) dx \\
&= \int_{\mathbb{R}} (h(f(\tau)) - f(\tau))^2 f'(\tau) I_1(f(\tau)) d\tau \\
&= \int_{\mathbb{R}} (g(\tau) - f(\tau))^2 I_0(\tau) d\tau \\
&= \int_{\mathbb{R}} ((g(\tau) - \tau) - (f(\tau) - \tau))^2 I_0(\tau) d\tau \\
&= \|\psi_\sigma(I_1) - \psi_\sigma(I_2)\|_2^2
\end{aligned}
$$

where in the second line we used the change of variable $f(\tau) = x$. In the third line, we used the fact that composition of transport maps is also a transport map, in other words, since $f_\# \sigma = \mu$ and $h_\# \mu = \nu$ then $(h \circ f)_\# \sigma = \nu$. Finally, from Theorem 1 we have that the one-dimensional transport maps are unique, therefore if $(h \circ f)_\# \sigma = \nu$ and $g_\# \sigma = \nu$ then $h \circ f = g$.

We showed that there exists a nonlinear map $\psi_\sigma : M \to \mathcal{V}$ for which $W_2(I_i, I_j) = \|\psi_\sigma(I_i) - \psi_\sigma(I_j)\|_2$ and therefore according to Theorem 3, $k(I_i, I_j) := exp(-\gamma W_2^2(I_i, I_j))$ is a positive definite kernel. $\square$

Combining the results in Theorems 4 and 2 will lead to the following corollary.

Corollary 1. *The squared 2-Wasserstein distance for continuous one-dimensional positive probability density functions, $W_2^2(.,.)$, is a conditionally negative definite function.*

Moreover, following the work of Feragen et al. [13], Theorem 4 also states that the Wasserstein space in one dimension (the space of one-dimensional absolutely continuous positive probability densities endowed with the 2-Wasserstein metric) is a flat space (it has zero curvature), in the sense that it is isometric to the Euclidean space.

3.1. The Sliced Wasserstein Gaussian kernel

Now we are ready to show that the Sliced Wasserstein Gaussian kernel is a positive definite kernel.

Theorem 5. *Let M be the set of absolutely continuous positive probability density functions and define $k : M \times M \to \mathbb{R}$ to be $k(I_i, I_j) := exp(-\gamma SW^2(I_i, I_j))$ then $k(.,.)$ is a positive definite kernel for all $\gamma > 0$.*

Proof. First note that for an absolutely continuous positive probability density function, $I \in M$, each hyperplane integral, $\mathscr{R}I(.,\theta), \forall\theta \in \mathbb{S}^{d-1}$ is a one dimensional absolutely continuous positive probability density function. Therefore, following Corollary 1 for $I_1, ..., I_N \in M$ we have,

$$\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j W_2^2(\mathscr{R}I_i(.,\theta), \mathscr{R}I_j(.,\theta)) \le 0, \ \forall\theta \in \mathbb{S}^{d-1} \quad (11)$$

where $\sum_{i=1}^{N} c_i = 0$. Integrating the left hand side of above inequality over θ leads to,

$$\int_{\mathbb{S}^{d-1}}(\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j W_2^2(\mathscr{R}I_i(.,\theta), \mathscr{R}I_j(.,\theta)))d\theta \le 0 \ \Rightarrow$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j(\int_{\mathbb{S}^{d-1}} W_2^2(\mathscr{R}I_i(.,\theta), \mathscr{R}I_j(.,\theta)))d\theta \le 0 \ \Rightarrow$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j SW^2(I_i, I_j) \le 0 \quad (12)$$

Therefore $SW^2(.,.)$ is conditionally negative definite, and hence from Theorem 2 we have that $k(I_i, I_j) := exp(-\gamma SW^2(I_i, I_j))$ is a positive definite kernel for $\gamma > 0$. $\qquad\square$

The following corollary follows from Theorems 3 and 5.

Corollary 2. *Let M be the set of absolutely continuous positive probability density functions and let $SW(.,.)$ be the sliced Wasserstein distance, then there exists an inner product space $\mathcal{V}$ and a function $\phi : M \to \mathcal{V}$ such that $SW(I_i, I_j) = \|\phi(I_i) - \phi(I_j)\|_{\mathcal{V}}$, for $\forall I_i, I_j \in M$.*

In fact, using a similar argument as the one we provided in the proof of Theorem 4 it can be seen that for a fixed absolutely continuous measure, σ, with positive probability density function I_0, we can define,

$$\phi_\sigma(I_i) := (f_i(t,\theta) - t)\sqrt{\mathscr{R}I_0(t,\theta)} \quad (13)$$

where f_i satisfies $\frac{\partial f_i(t,\theta)}{\partial t}\mathscr{R}I_i(f_i(t,\theta),\theta) = \mathscr{R}I_0(t,\theta)$. It is straightforward to show that such ϕ_σ also satisfies the following,

$$SW(I_i, I_0) = \|\phi_\sigma(I_i)\|_2 \quad (14)$$
$$SW(I_i, I_j) = \|\phi_\sigma(I_i) - \phi_\sigma(I_j)\|_2 \quad (15)$$

for a detailed derivation and proof of the equations above please refer to [22]. More importantly, such nonlinear transformation $\phi_\sigma : M \to \mathcal{V}$ is invertible.

3.2. The Sliced Wasserstein polynomial kernel

In this section, using Corollary 2 we define a polynomial Kernel based on the Sliced Wasserstein distance and show that this kernel is positive definite.

Theorem 6. *Let M be the set of absolutely continuous positive probability density functions and let σ be a template probability measure with corresponding probability density function $I_0 \in M$. Let $\phi_\sigma : M \to \mathcal{V}$ be defined as in Equation (13). Define a kernel function $k : M \times M \to \mathbb{R}$ to be $k(I_i, I_j) := (\langle\phi_\sigma(I_i), \phi_\sigma(I_j)\rangle)^d$ for $d \in \{1, 2, ...\}$ then $k(.,.)$ is a positive definite kernel.*

Proof. Given $I_1, ..., I_N \in M$ and for $d = 1$ we have,

$$\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j \langle\phi_\sigma(I_i), \phi_\sigma(I_j)\rangle =$$

$$\langle\sum_{i=1}^{N} c_i\phi_\sigma(I_i), \sum_{j=1}^{N} c_j\phi_\sigma(I_j)\rangle = \|\sum_{i=1}^{N} c_i\phi_\sigma(I_i)\|_2^2 \ge 0$$

$$(16)$$

and since $k(I_i, I_j) = \langle\phi_\sigma(I_i), \phi_\sigma(I_j)\rangle$ is a positive definite kernel and from Mercer's kernel properties it follows that $k(I_i, I_j) = (\langle\phi_\sigma(I_i), \phi_\sigma(I_j)\rangle)^d$ and $k(I_i, I_j) = (\langle\phi_\sigma(I_i), \phi_\sigma(I_j)\rangle + 1)^d$ are also positive definite kernels. $\qquad\square$

4. The Sliced Wasserstein Kernel-based algorithms

4.1. Sliced Wasserstein Kernel k-means

Considering clustering problems for data with the form of probability distributions, we propose the Sliced Wasserstein k-means. For a set of input data $I_1, ..., I_N \in M$, the Sliced Wasserstein k-means with kernel $k(I_i, I_j) = \langle\phi_\sigma(I_i), \phi_\sigma(I_j)\rangle$ transforms the input data to the kernel space via $\phi_\sigma : M \to \mathcal{V}$ and perform k-means in this space. Note that since $\|\phi_\sigma(I_i) - \phi_\sigma(I_j)\|_2 = SW(I_i, I_j)$, performing k-means in $\mathcal{V}$ is equivalent to performing k-means with Sliced Wasserstein distance in M. The kernel k-means with the Sliced Wasserstein distance, essentially provides k barycenters for the input distributions. In addition, the Gaussian and polynomial Sliced Wasserstein kernel k-means are obtained by performing Gaussian and polynomial kernel k-means in $\mathcal{V}$.

4.2. Sliced Wasserstein Kernel PCA

Now we consider the key concepts of the kernel PCA. The Kernel-PCA [36] is a non-linear dimensionality reduction method that can be interpreted as applying the PCA in the kernel-space (or feature-space), $\mathcal{V}$. Performing standard PCA on $\phi_\sigma(I_1), ..., \phi_\sigma(I_N) \in \mathcal{V}$ provides

the Sliced Wasserstein kernel PCA. In addition, applying Gaussian and polynomial Sliced Wasserstein kernel PCA to $I_1, ..., I_N \in M$ is also equivalent to applying Gaussian and polynomial kernel PCA on $\phi_\sigma(I_1), ..., \phi_\sigma(I_N) \in \mathcal{V}$. We utilize the cumulative percent variance (CPV) as a quality measure for how well the principal components are capturing the variation of the dataset in M and similarly in $\mathcal{V}$.

4.3. Sliced Wasserstein Kernel SVM

For a binary classifier, given a set of training examples $\{I_i, y_i\}_{i=1}^N$ where $I_i \in M$ and $y_i \in \{-1, 1\}$, support vector machine (SVM) searches for a hyperplane in M that separates training classes while maximizing the separation margin where separation is measured with the Euclidean distance. A kernel-SVM , on the other hand, searches for a hyperplane in $\mathcal{V}$ which provides maximum margin separation between $\phi_\sigma(I_i)$s which is equivalent to finding a nonlinear classifier in M that maximizes the separation margin according to the Sliced Wasserstein distance. Note that the kernel SVM with the Sliced Wasserstein Gaussian and polynomial kernels are essentially equivalent to applying kernel SVM, with the same kernels in the transformed Sliced Wasserstein space $\mathcal{V}$. It is worthwhile to mention that, since ϕ_σ is invertible, the Sliced Wasserstein kernel SVM learned from kernel $k(I_i, I_j) = \langle \phi_\sigma(I_i), \phi_\sigma(I_j) \rangle$ can be sampled along side the orthogonal direction to the discriminating hyperplane in $\mathcal{V}$ and then inverted through ϕ_σ^{-1} to directly get the discriminating features in the space of the probability densities, M. Finally, for multiclass classification problems, the problem can be turned into several binary classification tasks using pairwise coupling as suggested by Wu et al. [46] or a one versus all approach.

5. Experimental Results

For our experiments we utilized two image datasets, namely the University of Illinois Urbana Champaign (UIUC) texture dataset [25] and the LHI animal face dataset [38]. The texture dataset contains 25 classes of texture images with 40 images per class, which includes a wide range of variations. The animal face dataset contains 21 classes of animal faces with the average number of images per class being 114. Figures 1 (a) and 2 (a) show sample images from image classes for the texture and the LHI dataset, respectively. For the texture dataset we extracted the gray-level co-occurence matrices for texture images and normalized them to obtain empirical joint probability density functions of co-occuring gray levels (See Figure 1 (b)). On the other hand, for the animal face dataset, we used the normalized HOGgles images [42] as probability distribution representations of RGB animal face images (See Figure 2 (b)). We acknowledge that the HOGgles are not designed for feature extraction but rather for visualization of the extracted HoG features [42], but our goal here is to show that for any ex-

tracted probability density features from images the Sliced Wasserstein kernels often outperform commonly used kernels.

Regarding the implementation, the fixed density I_0 for each dataset is chosen to be the average distribution over the entire dataset. The Radon transform was used to slice the probability distributions at discrete angles $\theta \in \{0, 1, ..., 179\}$. For each slice, $\mathscr{R}I_i(., \theta_k)$, the transport map f_i that satisfies Equation (9) is calculated using Equation (4). Finally, the kernel representations $\phi_\sigma(I_i)$ for the extracted probability distributions were calculated using Equation (13). The calculated representations, $\phi_\sigma(I_i)$, are shown in Figures 1 (c) and 2 (c). The Matlab code [26] and detailed description of our implementation can be found in [1]. The computational complexity of calculating the transport map for a pair of one-dimensional probability density functions (PDFs), presented as vectors of length N, is $\mathcal{O}(NlogN)$. On the other hand, the computational complexity of calculating M slices (projections) of a d-dimensional PDF, which is presented as a $N \times N \times ... \times N$ tensor, is $\mathcal{O}(MN^d)$. Therefore, the overall computational complexity of calculating the explicit mapping $\phi_\sigma(.)$ is dominated by computational complexity of calculating the M projections of the distribution and is equal to $\mathcal{O}(MN^d)$. It should be mentioned that, for higher dimensional PDFs, more projections (larger M) is required to capture the underlying variations of the distributions.

First, the PCA of the input data, $I_1, ..., I_N \in M$ (i.e. the co-occurence matrices for the texture images and the HOG-gles images for the LHI dataset) as well as the kernel-PCA of the data with the Sliced Wasserstein kernel, RBF kernel, and the Sliced Wasserstein Gaussian kernel were calculated for both datasets. Figure 3 shows the cumulative percent variance (CPV) of the dataset captured by different kernels for both datasets. It can be seen that the variations in the datasets are captured more efficiently in the Sliced Wasserstein kernel space. We note that the CPV generally depends on the choice of the kernel parameters (e.g. radius of the RBF), however, for the Sliced Wasserstein kernel, $k(I_i, I_j) = \langle \phi_\sigma(I_i), \phi_\sigma(I_j) \rangle$, there are no parameters and the comparison between PCA and Sliced Wasserstein kernel PCA is parameter free.

Next, we performed classification tasks on the texture and animal face datasets. Linear SVM, RBF kernel SVM, Sliced Wasserstein Gaussian kernel SVM, and the Sliced Wasserstein polynomial kernel were utilized for classification accuracy comparison. A five fold cross validation scheme was used, in which 20% of each class was held out for testing and the rest was used for training and parameter estimation. The hyperparameters of the kernels are calculated through grid search. The classification experiments were repeated 100 times and the means and standard deviations of the classification accuracies for both datasets are
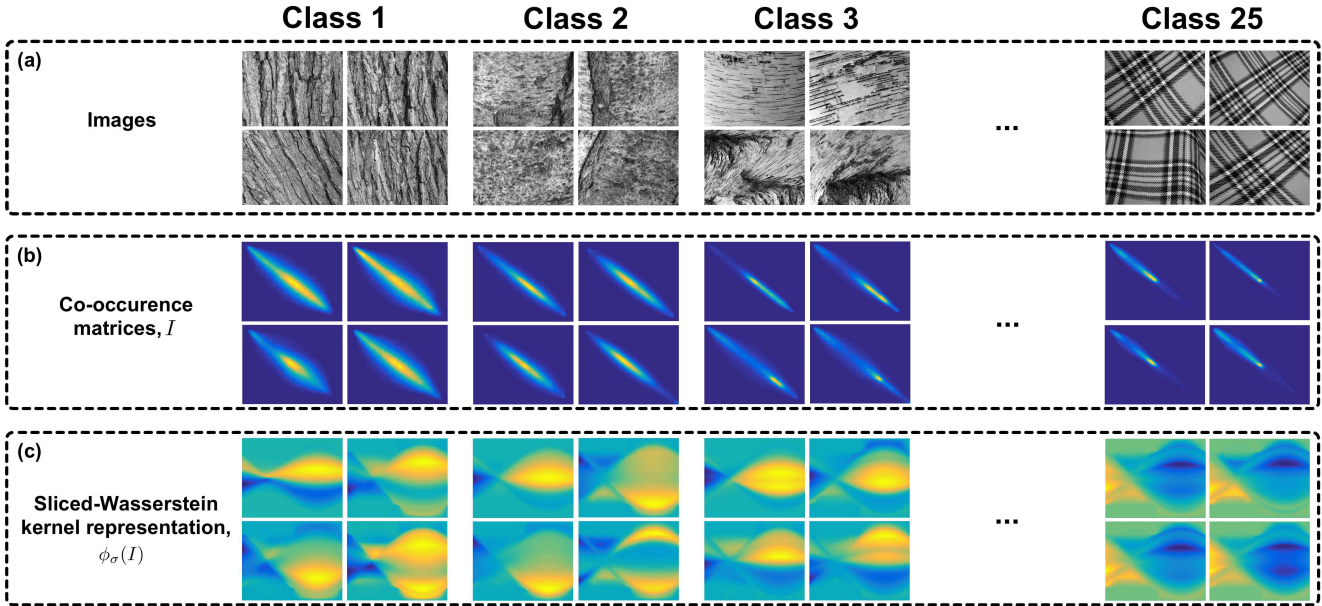
Figure 1. The UIUC texture dataset with 25 classes (a), the corresponding calculated co-occurence matrices (b), and the kernel representation (i.e. ϕ_σ) of the co-occurence matrices (c).
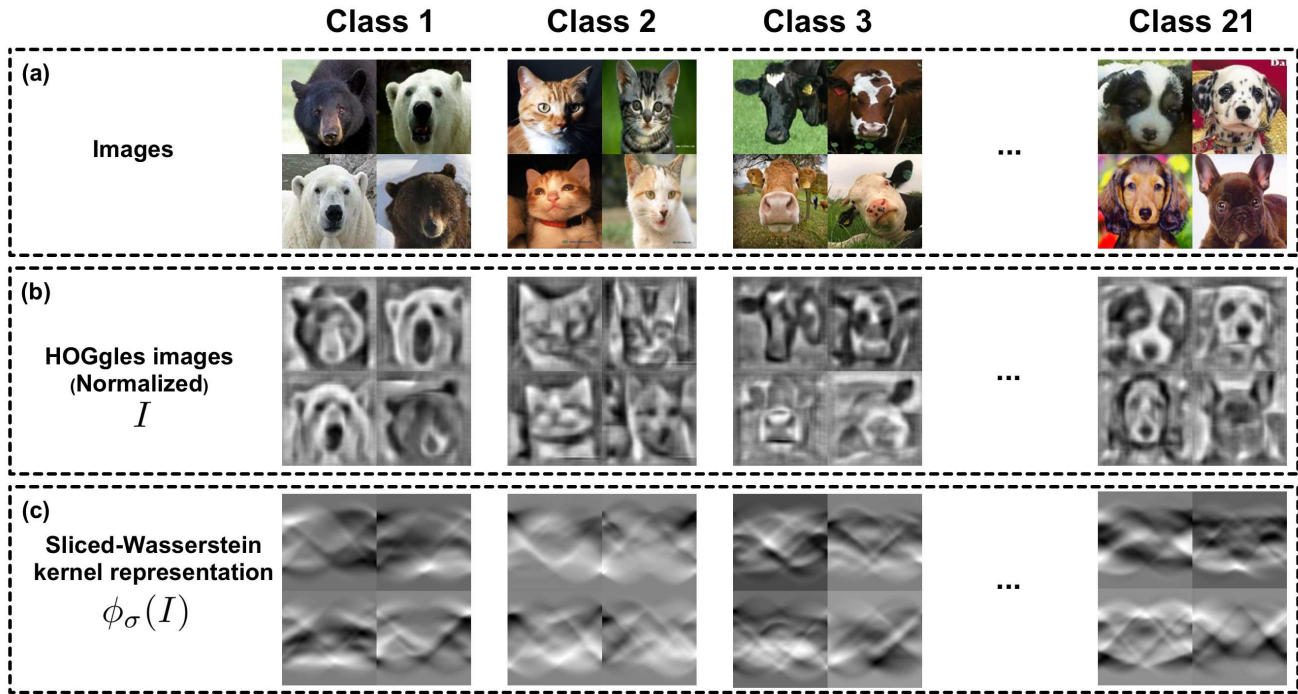

Figure 2. The LHI animal face dataset with 21 classes (a), the corresponding calculated HOGgles representation (b), and the kernel representation (i.e. ϕ_σ) of the HOGgles images (c).

reported in Figure 4.

Finally, we performed clustering on the UIUC texture and the LHI animal face dataset. We utilized the k-means algorithm and the spectral clustering method on the normalized co-occurence matrices and the HOGgles images,

$I \in M$, and their corresponding representations in the kernel space, $\phi_\sigma(I) \in \mathcal{V}$ (i.e. kernel k-means). We note that, since the mapping $\phi_\sigma()$ is known, any clustering algorithm can be performed in the Sliced Wasserstein kernel space. We repeated the k-means experiment 100 times and mea-
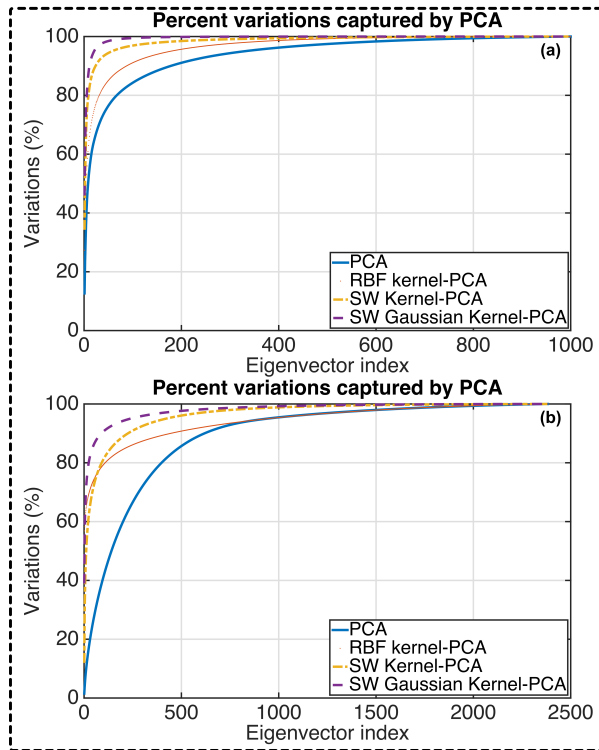
Figure 3. Percentage variations captured by eigenvectors calculated from PCA, RBF kernel-PCA, Sliced Wasserstein kernel PCA, and Sliced Wasserstein Gaussian kernel (SW-RBF) PCA for the texture dataset (a) and for the animal face dataset (b).
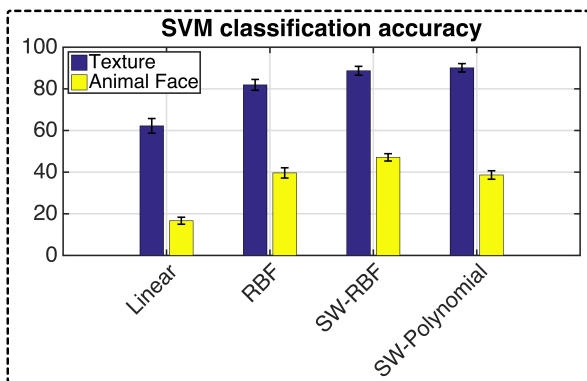

Figure 4. Kernel SVM classification accuracy with linear kernel, radial basis function kernel (RBF), Sliced Wasserstein Gaussian Kernel (SW-RBF), and Sliced Wasserstein Polynomial Kernel (SW-Polynomial).

sured the V-measure [33], which is a conditional entropy-based external cluster evaluation measure, at each iteration for both datasets. In addition, we performed spectral clustering in the kernel space $\phi_\sigma(I) \in \mathcal{V}$ and measured the V-measure. Figure 5 shows the mean and standard deviation of the V-measure for k-means, Sliced Wasserstein kernel k-means, spectral clustering, and Sliced Wasserstein spectral
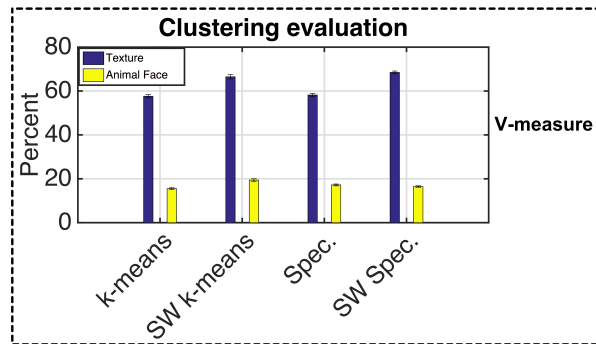

Figure 5. Cluster evaluation for k-means, Sliced Wasserstein kernel k-means, spectral clustering, and spectral clustering k-means using the V-measure.

clustering. It can be seen that for both methods, the Sliced Wasserstein kernel provides better clusters which match the texture and animal face classes better, as it leads to higher V-measure values.

6. Discussion

In this paper, we have introduced a family of provably positive definite kernels for probability distributions based on the mathematics of the optimal transport and more precisely the Sliced Wasserstein distance. We denoted our proposed family of kernels as the Sliced Wasserstein kernels. Following the work of [31, 22], we provided an explicit nonlinear and invertible formulation for mapping probability distributions to the kernel space (aka feature space). Our experiments demonstrated the benefits of the Sliced Wasserstein kernels over the commonly used RBF and Polynomial kernels in a variety of pattern recognition tasks on probability densities.

More specifically, we showed that utilizing a dimensionality reduction scheme like PCA with the Sliced Wasserstein kernel leads to capturing more of the variations of the datasets with fewer parameters. Similarly, clustering methods can benefit from the Sliced Wasserstein kernel as the clusters have higher values of V-measure and lower value of inertia. Finally, we demonstrated that the classification accuracy for a kernel classifier like the kernel SVM can also benefit from the Sliced Wasserstein kernels.

Finally, the experiments in this paper were focused on two-dimensional distributions. However, the proposed framework can be extended to higher-dimensional probability densities. We therefore intend to investigate the application of the Sliced Wasserstein kernel to higher-dimensional probability densities such as volumetric MRI/CT brain data.

7. Acknowledgement

References

[1] Sliced Wasserstein kernels. https://www.andrew.cmu.edu/user/gustavor/transport.html. Accessed: 04-10-2016. 6

[2] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011. 1

[3] S. Basu, S. Kolouri, and G. K. Rohde. Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry. *Proceedings of the National Academy of Sciences*, 111(9):3448–3453, 2014. 1

[4] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015. 1

[5] C. Berg, J. P. R. Christensen, and P. Ressel. Harmonic analysis on semigroups. 1984. 3, 4

[6] J. Bigot, R. Gouet, T. Klein, and A. López. Geodesic PCA in the Wasserstein space by convex PCA. 2015. 1

[7] 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, 3

[8] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991. 2

[9] R. Chartrand, B. Wohlberg, K. Vixie, and E. Bollt. A gradient descent solution to the Monge-Kantorovich problem. *Applied Mathematical Sciences*, 3(22):1071–1080, 2009. 1

[10] M. Cuturi. Permanents, transportation polytopes and positive definite kernels on histograms. In *International Joint Conference on Artificial Intelligence, IJCAI*, 2007. 1

[11] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013. 1

[12] M. R. Daliri. Kernel Earth Mover's distance for EEG classification. *Clinical EEG and neuroscience*, 44(3):182–187, 2013. 1

[13] A. Feragen, F. Lauze, and S. Hauberg. Geodesic Exponential Kernels: When curvature and linearity conflict. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2, 3, 4

[14] A. Gardner, C. A. Duncan, J. Kanno, and R. R. Selmic. Earth Mover's distance yields positive definite kernels for certain ground distances. *arXiv preprint arXiv:1510.02833*, 2015. 2

[15] W. Gómez, W. Pereira, and A. F. C. Infantosi. Analysis of co-occurrence texture statistics as a function of gray-level quantization for classifying breast ultrasound. *Medical Imaging, IEEE Transactions on*, 31(10):1889–1899, 2012. 1

[16] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent. Optimal mass transport for registration and warping. *International Journal of computer vision*, 60(3):225–240, 2004. 1

[17] D. C. Hauagge and N. Snavely. Image matching using local symmetry features. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 206–213. IEEE, 2012. 1

[18] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008. 1, 2

[19] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi. Kernel methods on the riemannian manifold of symmetric positive definite matrices. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 73–80. IEEE, 2013. 2

[20] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi. Kernel methods on Riemannian manifolds with Gaussian RBF kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015. 2, 3, 4

[21] H. Jégou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, 87(3):316–336, 2010. 1

[22] S. Kolouri, S. R. Park, and G. Rohde. The Radon Cumulative Distribution Transform and its application to image classification. *IEEE transactions on image processing*, 25(2):920–934, 2016. 1, 3, 5, 8

[23] S. Kolouri and G. K. Rohde. Transport-based single frame super resolution of very low resolution face images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4876–4884, 2015. 1

[24] S. Kolouri, A. B. Tosun, J. A. Ozolek, and G. K. Rohde. A continuous linear optimal transport approach for pattern analysis in image datasets. *Pattern Recognition*, 51:453–462, 2016. 1

[25] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1265–1278, 2005. 6

[26] MATLAB. *version 8.4.0 (R2014b)*. The MathWorks Inc., Natick, Massachusetts, 2014. 6

[27] Q. Mérigot. A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pages 1583–1592. Wiley Online Library, 2011. 1

[28] F. Natterer. *The mathematics of computerized tomography*, volume 32. Siam, 1986. 3

[29] A. M. Oberman and Y. Ruan. An efficient linear programming method for optimal transportation. *arXiv preprint arXiv:1509.03668*, 2015. 1

[30] J. A. Ozolek, A. B. Tosun, W. Wang, C. Chen, S. Kolouri, S. Basu, H. Huang, and G. K. Rohde. Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning. *Medical image analysis*, 18(5):772–780, 2014. 1

[31] S. R. Park, S. Kolouri, S. Kundu, and G. Rohde. The Cumulative Distribution Transform and linear pattern classification. *arXiv preprint arXiv:1507.05936*, 2015. 1, 4, 8

[32] J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2012. 1, 3

[33] A. Rosenberg and J. Hirschberg. V-Measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, volume 7, pages 410–420, 2007. 8

[34] Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000. 1

[35] I. J. Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, pages 811–841, 1938. 3

[36] B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In *Artificial Neural Networks-ICANN'97*, pages 583–588. Springer, 1997. 5

[37] V. Seguy and M. Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. 1

[38] Z. Si and S.-C. Zhu. Learning hybrid image templates (HIT) by information projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(7):1354–1367, 2012. 6

[39] J. Solomon, F. de Goes, P. A. Studios, 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 (Proc. SIGGRAPH 2015), to appear*, 2015. 1

[40] A. B. Tosun, O. Yergiyev, S. Kolouri, J. F. Silverman, and G. K. Rohde. Detection of malignant mesothelioma using nuclear structure of mesothelial cells in effusion cytology specimens. *Cytometry Part A*, 87(4):326–333, 2015. 1

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

[42] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: Visualizing object detection features. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1–8. IEEE, 2013. 6

[43] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79, 2013. 1

[44] W. Wang, J. Ozolek, D. Slepcev, A. B. Lee, C. Chen, G. K. Rohde, et al. An optimal transportation approach for nuclear structure-based pathology. *Medical Imaging, IEEE Transactions on*, 30(3):621–631, 2011. 1

[45] W. Wang, D. Slepcev, S. Basu, J. A. Ozolek, and G. K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101(2):254–269, 2013. 1, 4

[46] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 5:975–1005, 2004. 6

[47] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, 73(2):213–238, 2007. 1