Wasserstein Exponential Kernels

Henri De Plaen, Michaël Fanuel and Johan A. K. Suykens

Department of Electrical Engineering, ESAT-STADIUS, KU Leuven
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
Email: {henri.deplaen, michael.fanuel, johan.suykens}@esat.kuleuven.be

***Abstract*—In the context of kernel methods, the similarity between data points is encoded by the kernel function which is often defined thanks to the Euclidean distance; the squared exponential kernel is a common example. Recently, other distances relying on optimal transport theory – such as the Wasserstein distance between probability distributions – have shown their practical relevance for different machine learning techniques. In this paper, we study the use of exponential kernels defined thanks to the regularized Wasserstein distance and discuss their positive definiteness. More specifically, we define Wasserstein feature maps and illustrate their interest for supervised learning problems involving shapes and images. Empirically, Wasserstein squared exponential kernels are shown to yield smaller classification errors on small training sets of shapes, compared to analogous classifiers using Euclidean distances.**

***Index Terms*—Machine Learning, Kernel Methods, Optimal Transport.**

I. Introduction

Contemporary machine learning methods frequently rely on neural networks and shape recognition relies more specifically on convolutional neural networks. The big advantage of the latter is its ability to take into account the underlying structure of the data by treating neighboring pixels together. If these methods are often impressive by their performance, they are also known for their drawbacks such as a weak robustness and a difficult explainability. On the other side, although not always as accurate as neural networks, kernel methods are praised for their easy explainability and robustness. Another advantage of kernel methods is their versatility as they can easily be used in supervised and unsupervised methods, as well as for generation [1]. In this paper, we emphasize the interest of choosing a particular kernel based on Wasserstein distance for classifying small datasets consisting of shapes.

In the context of kernel methods, squared exponential kernel functions are widely used, mainly because of their universal approximation properties and their empirical success. These Gaussians consist of the exponential of the negative Euclidean distance squared. However, the Euclidean distance might not always be appropriate to compare data points when data has some particular structure. Indeed, it measures the correspondence of each feature independently of the other features. For example, let's consider the case of two identical 2D-shapes. When the two shapes overlap, their Euclidean distance is zero. However, if they do not overlap, their relative Euclidean distance becomes large although the shapes are identical. In other words, the Euclidean distance only compares each pixel at the same place on the grid and does not take the neighbouring pixels into account. The general structure of the features is not taken into account, only their strict correspondence. Another distance – the Wasserstein distance – gained popularity in recent years since it can incorporate the structure of the data if the dataset can be processed in such a manner that the datapoints can be considered as probability distributions.

Contributions

The contributions of this paper are the following. Empirically, we demonstrate that squared exponential kernels (1) based on a regularized Wassertein distance are performant on small scale classification problems involving shape datasets, compared for instance to the popular Gaussian RBF kernel [2]. Also, an approximation technique is proposed, with the so-called Wasserstein feature map, so that a positive semi-definite (psd) kernel can be defined from the Wasserstein squared exponential kernel which is not necessarily psd.

Notations and conventions

In the sequel, we denote vectors by bold lower case letters. Let $\mathbf{1}$ be the all ones column vector. Also, we define δ_y to be the Dirac measure at point y. A kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called positive semi-definite if all kernel matrices $K = [k(\boldsymbol{x_i}, \boldsymbol{x_j})]_{i,j=1}^n$ are positive semi-definite.

Wasserstein distances

The Wasserstein distance is a central notion in optimal transport theory. Also known as the *earth mover's distance*, it corresponds to the optimal transportation cost between two measures [3], [4]. Let $p > 0$. We then define two normalized empirical measures $\boldsymbol{\alpha} = \sum_{i=1}^m a_i \delta_{\boldsymbol{y_i}}$ and $\boldsymbol{\beta} = \sum_{j=1}^n b_j \delta_{\boldsymbol{z_j}}$ such that $\boldsymbol{a}^\top \mathbf{1} = 1$ and $\boldsymbol{b}^\top \mathbf{1} = 1$, and where $\{\boldsymbol{y_i} \in \mathbb{R}^d\}_{i=1}^m$, $\{\boldsymbol{z_j} \in \mathbb{R}^d\}_{j=1}^n$ are support points. Also, we define a distance matrix $d_{ij} = d(\boldsymbol{y_i}, \boldsymbol{z_j})$, *e.g.* the Euclidean distance $\|\boldsymbol{y_i} - \boldsymbol{z_j}\|_2$. Then, the p-Wasserstein distance is given by

$$\mathcal{W}_p(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \left(\min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\alpha}, \boldsymbol{\beta})} \sum_{i,j} \pi_{ij} d_{ij}^p \right)^{1/p},$$

with $\Pi(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \left\{ \Pi \in \mathbb{R}^{m \times n} | \Pi \mathbf{1} = \boldsymbol{a} \text{ and } \Pi^\top \mathbf{1} = \boldsymbol{b} \right\}$, the set of joint distributions π with specified marginals given by α and β. Intuitively, the optimal probability distribution $\pi^\star$ represents the optimal mass transportation scheme from

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

α to β. A particular result occurs in the one-dimensional case ($d = 1$) assuming the support points are ordered, *i.e.*, $y_1 \leq \ldots \leq y_m$ and $z_1 \leq \ldots \leq z_n$ with $n = m$, where the Wasserstein distance reduces to an ℓ^p-norm: $\mathcal{W}_p^p\left(\frac{1}{n}\sum_{i=1}^n \delta_{y_i}, \frac{1}{n}\sum_{j=1}^n \delta_{z_j}\right) = \frac{1}{n}||\boldsymbol{y} - \boldsymbol{z}||_p^p$ [4]. This connection between ℓ^p-norms and Wasserstein distances is only clear in one dimension, illustrating here again the fact that ℓ^p-norms do not take into account the underlying structure. To do so, we need to consider the case $d > 1$. In this way we can define the following kernel function

$$k_W(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \exp\left(-\frac{W_2^2(\boldsymbol{\alpha}, \boldsymbol{\beta})}{2\sigma^2}\right), \tag{1}$$

where $\sigma > 0$ is a bandwidth parameter.

This has however some undesirable consequences concerning positive definiteness. A kernel $k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-tf(\boldsymbol{x}, \boldsymbol{y}))$ is positive semi-definite for all $t > 0$ if and only if $f(\boldsymbol{x}, \boldsymbol{y})$ is Hermitian and *conditionally* negative semi-definite [5]. Recall that a kernel is *conditionally* negative semi-definite if any Gram matrix $F = [f(\boldsymbol{x_i}, \boldsymbol{x_j})]_{i,j=1}^n$ (with $n \geq 2$) built from a discrete sample satisfies $\boldsymbol{c}^\top F \boldsymbol{c} \leq 0$ for all $\boldsymbol{c}$ such that $\mathbf{1}^\top \boldsymbol{c} = 0$. However, the Wasserstein distance for $d > 1$ is not necessarily *conditionally* negative definite [4]. By consequence, we cannot guarantee that any resulting squared exponential kernel matrix built with the 2-Wasserstein distance is positive definite. This property is fundamental in kernel theory and more specifically for defining *reproducing kernel Hilbert spaces* (RKHS; see [6] for more details).

II. Dealing with indefinite exponential kernels

This restriction has lead authors to consider only some specific cases of Wasserstein distances which are known to be positive definite. The one-dimensional generic case is proven to be positive definite and has lead to the introduction of sliced Wasserstein distances [7], [8]. Another notable case is the Wasserstein distance between two multivariate normal distributions in more than one dimension, which can even be written in closed form [4].

Some kernel methods have been used with indefinite kernels, such as LS-SVMs [9], [10]. This leads however to a slightly different interpretation of the global problem, using Kreĭn spaces for which a weaker version of the representer theorem holds [11]. In this paper, we propose an alternative which allows to continue working with a positive definite kernel approximating the squared exponential kernel. If the Wasserstein exponential kernel can not be used, we can always find a parameter $\sigma > 0$ and a finite dimensional feature map resulting in a positive definite kernel.

A. Positive definite squared exponential kernels and bandwidth choice

In this section, we show that for a given dataset, the corresponding Gram matrix of k_W is positive definite if the bandwidth parameter $\sigma > 0$ is small enough.

Definition II.1. *Let $d : \mathcal{D} \times \mathcal{D} \to \mathbb{R}_{\geq 0}$ be a symmetric function such that $d(\boldsymbol{x}, \boldsymbol{x}) = 0$ and let $\{\boldsymbol{x}_i \in \mathcal{D}\}_{i=1}^N$ be a dataset. A squared exponential kernel matrix is defined as*

$$\boldsymbol{K}_{d,\sigma} = \left[\exp\left(\frac{-d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)}{2\sigma^2}\right)\right]_{i,j=1}^N.$$

By construction, this squared exponential kernel matrix will be symmetric and have a diagonal consisting only of ones. Its eigenvalues are real. To investigate its (semi)-definiteness, we have to investigate the sign of the minimum eigenvalue. The minimum eigenvalue $\lambda_{\min}(\sigma)$ of $\boldsymbol{K}_{d,\sigma}$ is the function $\lambda_{\min} : \mathbb{R}_{>0} \to \mathbb{R}, \sigma \mapsto \min\{\lambda_1, \ldots, \lambda_N\}$ where $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of $\boldsymbol{K}_{d,\sigma}$. We can now prove the following results.

Lemma II.1. *The eigenvalues of the squared exponential kernel matrix $\boldsymbol{K}_{d,\sigma}$ are continuous functions of σ. In particular, $\lambda_{\min}(\sigma)$ is continuous.*

Proof: This is a direct consequence of the continuity of the roots of a polynomial with continuous coefficients. Therefore, we have to prove that the coefficients of the characteristic polynomial of the squared exponential kernel matrix $\boldsymbol{K}_{d,\sigma}$ is continuous as a function of σ. The characteristic polynomial is given by $\det(\boldsymbol{K}_{d,\sigma} - \lambda \boldsymbol{I})$ and by the formula of Leibniz, we ultimately have that the characteristic polynomial is a sum of products of elements of $\boldsymbol{K}_{d,\sigma} - \lambda \boldsymbol{I}$, which are continuous in function of σ. Hence, the coefficients are continuous and so are the eigenvalues. ∎

Lemma II.2. $\lim_{\sigma \to 0} \boldsymbol{K}_{d,\sigma} = \mathrm{id}$ *and thus* $\lambda_{\min}(0) = 1$.

Proof: From Definition II.1, we know that $[\boldsymbol{K}_{d,\sigma}]_{i,j} = \exp\left(\frac{-d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)}{2\sigma^2}\right)$ with $d^2(\boldsymbol{x}_i, \boldsymbol{x}_i) = 0$ and $d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) > 0$ for $i \neq j$. Denote $C_{i,j} = d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for simplicity. We have $\lim_{\sigma \to 0} \exp\left(\frac{0}{2\sigma^2}\right) = 1$ and $\lim_{\sigma \to 0} \exp\left(-\frac{C_{i,j}}{2\sigma^2}\right) = 0$ with $C_{i,j} > 0$ for $i \neq j$, thus the identity matrix. By consequence, all the eigenvalues are equal to 1. ∎

Lemma II.3. *We have* $\lim_{\sigma \to \infty} \boldsymbol{K}_{d,\sigma} = \mathbf{1}\mathbf{1}^T$ *and thus* $\lim_{\sigma \to \infty} \lambda_{\min}(\sigma) = 0$.

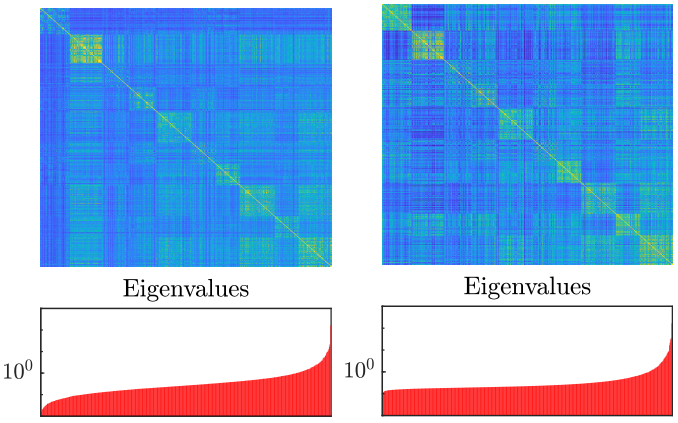
Proof: Similarly, we have $\lim_{\sigma \to +\infty} [\boldsymbol{K}_{d,\sigma}]_{i,j} = 1$ everywhere. By consequence, we have $\lambda_{\max} = N$ and all others equal to zero, hence $\lambda_{\min} = 0$. ∎

Proposition II.4. *There exists a $\sigma_{\mathrm{PSD}} \in \mathbb{R}_+$ such that $\boldsymbol{K}_{d,\sigma}$ is positive semi-definite for all $\sigma \leq \sigma_{\mathrm{PSD}}$.*

Proof: Let us proceed *ad absurdum* and suppose this is not the case. We consider the sequence $(\sigma_n)_n$ converging to 0 with $\sigma_0 = \sigma_{\mathrm{PSD}}$. There must exist some subsequence $(\sigma_{n_j})_j$ such that $(\lambda_{\min}(\sigma_{n_j}))_j < 0$. If this sequence is finite, then it is sufficient to consider a new sequence with $\sigma_{\mathrm{PSD}} = \sigma_{n_{j_{\max}+1}}$. If this subsequence is infinite, then $(\lambda_{\min}(\sigma_n))_n$ cannot converge to 1. This is impossible because of the continuity of $\lambda_{\min}(\sigma)$ (Lemma II.1) and its convergence to 1 (Lemma II.2). Hence, there exist some $\sigma_{\mathrm{PSD}} > 0$ such that

$\lambda_{\min}(\sigma) \geq 0$ for all $\sigma \leq \sigma_{\text{PSD}}$. This proves our proposition. ■

We can empirically see the result of Proposition II.4 in Fig. 1, where all eigenvalues are positive. Intuitively, decreasing the bandwidth σ tends to make the smallest distances more predominant, pushing the smallest eigenvalue progressively to the positive side. In this sense, an indefinite kernel matrix with σ close to σ_{PSD} will lead to proportionally very small negative eigenvalues in magnitude. In this case, a finite positive definite approximation can be justified.



(a) ℓ^2 (b) $\mathcal{W}_2^2$

Fig. 1: Comparison of the classical squared exponential kernel matrix (based on a ℓ^2-distance) and the introduced Wasserstein exponential kernel matrix on 500 normalized digits of the MNIST dataset [12]. The digits are ordered by class in ascending order. For the color legend, please refer to Fig. 2.

B. Wasserstein features

We can consider a finite dimensional feature map $\boldsymbol{\phi}(\boldsymbol{x})$ such that the positive semi-definite kernel $\boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{y})$ approximates $k_W(\boldsymbol{x}, \boldsymbol{y})$ given in (1). This finite approximation is based on a training dataset $\{\boldsymbol{x}_i\}_{i=1}^N$ for constructing an original kernel matrix $\boldsymbol{K} = [k_W(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^N \in \mathbb{R}^{N \times N}$. It suffices to truncate the spectral decomposition of the kernel matrix $\boldsymbol{K} = \sum_{l=1}^N \lambda_l \boldsymbol{v}_l \boldsymbol{v}_l^\top$ to the ℓ largest strictly positive eigenvalues. This will result in a new positive semi-definite kernel matrix $\boldsymbol{K}^{(\ell)} \overset{\text{def}}{=} \sum_{l=1}^{\ell} \lambda_l \boldsymbol{v}_l \boldsymbol{v}_l^\top \succeq 0$ with $\lambda_1 \geq \cdots \geq \lambda_N$. We can now reconstruct the different components of an approximate feature map

$$\phi_l(\boldsymbol{x}) \overset{\text{def}}{=} \frac{1}{\sqrt{\lambda_l}} \boldsymbol{k}_{\boldsymbol{x}}^\top \boldsymbol{v}_l, \qquad \text{for } i = 1, \ldots, \ell, \tag{2}$$

with $\boldsymbol{k}_{\boldsymbol{x}} \overset{\text{def}}{=} [k_W(\boldsymbol{x}, \boldsymbol{x}_1) \cdots k_W(\boldsymbol{x}, \boldsymbol{x}_N)]^\top$. We refer to these different components as the *Wasserstein features* as they compose the approximate feature map $\boldsymbol{\phi}(\boldsymbol{x}) \overset{\text{def}}{=} [\phi_1(\boldsymbol{x}) \cdots \phi_\ell(\boldsymbol{x})]^\top$ of the Wasserstein exponential kernel. This approximate feature map is constructed by using a training dataset, but can afterwards be evaluated at any out-of-sample point. By construction, we can verify that the Wassertein features evaluated on the training dataset result in the truncated kernel matrix.

Proposition II.5. *We have* $\left[\boldsymbol{\phi}(\boldsymbol{x}_i)^\top \boldsymbol{\phi}(\boldsymbol{x}_j)\right]_{i,j=1}^N = \boldsymbol{K}^{(\ell)}$.

Proof: It suffices to observe that $\boldsymbol{k}_{\boldsymbol{x}_i} = \sum_{l=1}^N \lambda_l \boldsymbol{v}_l [\boldsymbol{v}_l]_i$. By consequence, we have $\phi_l(\boldsymbol{x}_i) = \sqrt{\lambda_l} [\boldsymbol{v}_l]_i$. ■

Proposition II.4 suggests that even if no suitable σ can be found such that the kernel matrix is psd, the negative eigenvalues will remain very small in magnitude. By consequence, we can suppress them without much information loss. A truncated kernel is thus very close to the original one in spectral norm. This justifies the Wasserstein features in this sense that they are very close to the Wasserstein exponential kernel as well as being positive definite by construction. This fact can be visualized on Fig. 2.

Clearly, the *Wasserstein features* yield a positive semi-definite kernel. Moreover, it is also advantageous to work with finite dimensional feature maps to reduce the training time. Indeed, the computation of the Wasserstein distance (or an approximation with *e.g.* Sinkhorn's algorithm [13]) is still relatively expensive compared to ℓ^2 distance.

III. Experiments

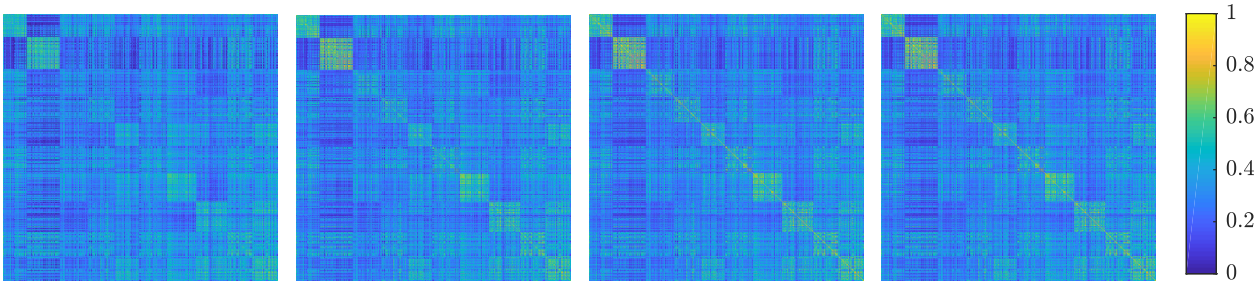
A. Setup for 2D shape classification

Let $\boldsymbol{u}$ be a greyscale image that we unfold as a vector of length m and so that $u_i > 0$ is the "grey" value at the pixel $\boldsymbol{y_i}$ of a pixel grid. It is mapped to a probability $\boldsymbol{\alpha} = \sum_{i=1}^m a_i \delta_{\boldsymbol{y_i}}$ by defining $a_i = u_i / \|\boldsymbol{u}\|_1$, so that the mass of $\boldsymbol{\alpha}$ is one. In practice, the $p = 2$ Wassertein distance is computed in this paper with the help of the well-known entropic regularization

$$\mathcal{W}_{2,\epsilon}^2(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\alpha}, \boldsymbol{\beta})} \sum_{i,j} \left(\pi_{ij} d_{ij}^2 + \epsilon \pi_{ij} \log \pi_{ij}\right),$$

where $\epsilon > 0$ is a small regularization term and d_{ij}^2 is the Euclidean distance between pixels located at $\boldsymbol{y_i}$ and $\boldsymbol{y_j}$ in a pixel grid. The advantage of this regularized problem is that its solution can be efficiently obtained thanks to the Sinkhorn algorithm, which can be parallelized. For more details, we refer to [4]. All the simulations used $\epsilon = 2.5$ and the diagonal of the distance matrix set to zero. This choice of value was motivated by trial-and-error in order to make the regularization parameter large enough to be close to the exact Wasserstein distance, without being too large, which increases drastically the computation time and eventually leads to unstable iterations of Sinkhorn's algorithm.

The full Wasserstein kernel matrix has N^2 elements, with N the number of datapoints. Computing the full kernel matrix requires thus to compute $N^2/2$ pairwise distances as they are symmetric. By using the Wasserstein features, the number of pairwise distances to compute can be reduced to $N_1^2/2 + N_1 N_2$, where N_1 is the size of the training dataset and N_2 of the out-of-sample dataset. The computation of each pairwise Wasserstein distance has a complexity of $\widetilde{O}\left(n^2/\epsilon^3\right)$ where n is the dimension of each datapoint and ϵ the regularization parameter [14], compared to $O(n)$ for the Euclidean distance. However, Sinkhorn's algorithm can be implemented on GPUs,



(a) $\ell = 5$ (b) $\ell = 15$ (c) $\ell = 500$ (d) Exact

Fig. 2: Kernel matrices constructed as the inner products of a different number Wasserstein features of a test set. These matrices are compared with the exact Wasserstein squared exponential kernel matrix of the test set. Both the training set and the test set are of size $N = 500$.

TABLE I: Percentage of classification error on the test set of three datasets. The standard deviation is given in parenthesis. The number of repeated simulations is 7 for MNIST, 8 for Quickdraw and 6 for USPS.

Dataset	**MNIST**		**Quickdraw**		**USPS**	
Method	**Avg.**	**Best**	**Avg.**	**Best**	**Avg.**	**Best**
Wass. LS-SVM (Core+OOS)	3.95 ($\pm$ 0.18)	3.74	11.45 ($\pm$ 0.39)	10.97	6.77 ($\pm$ 0.52)	6.20
Wass. LS-SVM (Core)	3.81 ($\pm$ 0.34)	3.28	10.80 ($\pm$ 0.19)	10.52	7.93 ($\pm$ 1.45)	6.35
Wass. LS-SVM (Indef.)	**3.40** ($\pm$ 0.11)	**3.23**	**10.75** ($\pm$ 0.27)	10.35	6.15 ($\pm$ 0.67)	5.45
R. Wass. LS-SVM (Core+OOS)	3.91 ($\pm$ 0.27)	3.45	11.79 ($\pm$ 0.48)	10.95	6.68 ($\pm$ 0.80)	5.70
R. Wass. LS-SVM (Core)	3.71 ($\pm$ 0.15)	3.46	10.99 ($\pm$ 0.44)	**10.07**	6.35 ($\pm$ 0.11)	6.20
R. Wass. LS-SVM (Indef.)	3.48 ($\pm$ 0.13)	3.29	12.43 ($\pm$ 0.43)	11.95	**5.70** ($\pm$ 0.29)	**5.40**
Wass. kNN	6.31 ($\pm$ 0.33)	5.81	12.26 ($\pm$ 0.33)	11.91	6.60 ($\pm$ 0.44)	6.00
RBF LS-SVM	4.26 ($\pm$ 0.10)	4.07	11.46 ($\pm$ 0.20)	11.23	6.75 ($\pm$ 0.04)	6.70
ℓ^2 kNN	7.20 ($\pm$ 0.15)	6.95	15.32 ($\pm$ 0.40)	14.68	7.52 ($\pm$ 0.38)	7.20
Set size	**Core + OOS**	**Others**	**Core + OOS**	**Others**	**Core + OOS**	**Others**
Training	1500 + 2500	4000	500 + 750	1250	1000 + 1500	2500
Validation	5000	5000	5000	5000	2000	2000
Test	10 000	10 000	10 000	10 000	2000	2000

benefiting from their massive parallelization capabilities, allowing to compute pairwise distances from an arbitrary group of datapoints to a common datapoint together.

B. Shape recognition

We illustrate the use of the Wasserstein based kernels in the context of shape classifications. Namely, we train a Least Squares Support Vector Machine [15] classifier on subsets of the MNIST [12], Quickdraw[1] and USPS [16] datasets, which are sampled uniformly at random. These three datasets contain handwritten digits and shapes. The multiclass problem is solved by a one-versus-one encoding. One instance of these binary classifiers $f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^{\star\top}\boldsymbol{\phi}(\boldsymbol{x}) + b^\star)$ is obtained by solving

$$\min_{\substack{\boldsymbol{w}\in\mathbb{R}^\ell; b\in\mathbb{R} \\ e_i\in\mathbb{R}}} \boldsymbol{w}^\top\boldsymbol{w} + \frac{\gamma}{N}\sum_{i=1}^{N} e_i^2 \text{ s.t. } e_i = y_i - \boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_i) - b, \quad (3)$$

[1]https://quickdraw.withgoogle.com/data

where $y_i \in \{-1, 1\}$ and $\boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^\ell$ is a feature map obtained for instance thanks to (2). The solution is obtained by solving

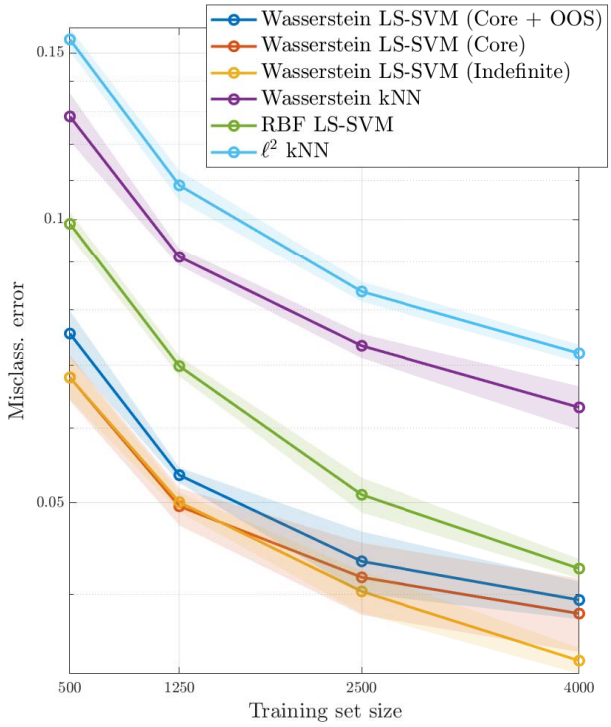
$$\begin{bmatrix} \sum_i \boldsymbol{\phi}(\boldsymbol{x_i})\boldsymbol{\phi}(\boldsymbol{x_i})^\top + \frac{N}{\gamma}\mathbb{I} & \sum_i \boldsymbol{\phi}(\boldsymbol{x_i}) \\ \sum_i \boldsymbol{\phi}(\boldsymbol{x_i})^\top & N \end{bmatrix} \begin{bmatrix} \boldsymbol{w} \\ b \end{bmatrix} = \begin{bmatrix} \sum_i y_i\boldsymbol{\phi}(\boldsymbol{x_i}) \\ \sum_i y_i \end{bmatrix},$$

which is a $(\ell+1)\times(\ell+1)$ linear system. A classifier can also be obtained by solving the dual problem of (3). The optimality conditions of this dual problem yield the following $(N+1)\times(N+1)$ linear system

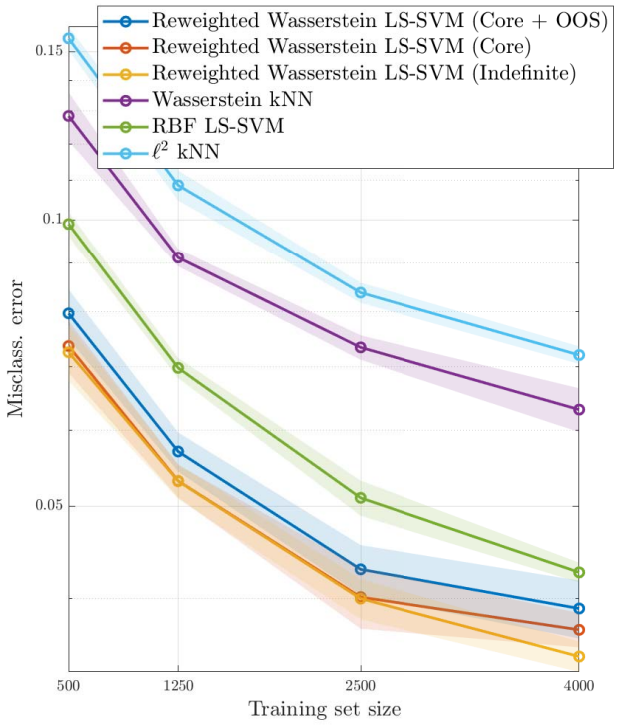
$$\begin{bmatrix} K + \frac{N}{\gamma}\mathbb{I} & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix}. \quad (4)$$

The resulting classifier has then the expression $f(\boldsymbol{x}) = \text{sign}(\sum_{i=1}^N \alpha_i^\star k(\boldsymbol{x}, \boldsymbol{x_i}) + b^\star)$. The optimal hyperparameters $\sigma > 0$ and $\gamma > 0$ are estimated using grid search with validation on a hold-out set. The final classification is done by minimizing the Hamming distance on the one-versus-one outputs [17]. In order to account for the amount of ink in the grey images $\boldsymbol{u}$ and $\boldsymbol{v}$, we also introduce a reweighted kernel that is defined as

$$k_{RW}(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u}\|_1\|\boldsymbol{v}\|_1 k_W\left(\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_1}, \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_1}\right). \quad (5)$$



(a) Comparison of Wasserstein exponential kernels with other similar methods.



(b) Comparison of *reweighted* Wasserstein exponential kernels with other similar methods.

Fig. 3: Mean misclassification rates for various subset sizes of the MNIST dataset, computed on 7 simulations. The standard deviation is given by the errors bars. For the specific case of "Core + OOS", the out-of-sample subset represents 300 datapoints on 500, 750 on 1250, 1500 on 2500 and 2500 on 4000. The size of validation set is always 5000 and of the test set always 10 000.

Notice that a similar kernel has been defined with the Euclidean distance in [18], [19].

In our experiments, we compare several methods based on k_W and k_{RW}, Wasserstein and Euclidean distances.

1) Core Wasserstein kernel: The "Core" method consists of solving (3) thanks to the feature map (2) associated to $K^{(\ell)}$. The parameter ℓ is selected such that all the selected eigenvalues are larger than 10^{-6} to avoid numerical instabilities. The optimal $\boldsymbol{w}^\star$ and $b^\star$ are then obtained by solving a linear system.

2) Core Wasserstein kernel with out-of-sample: Our second method named "Core + OOS" uses almost the same methodology as "Core". However, a subset of the training set is used to construct the truncated Wasserstein kernel of Proposition II.5. Then the out-of-sample (OOS) formula (2) is used to construct an approximation of the kernel matrix on the full training dataset. The advantage of this approximation is that it can avoid the full eigendecomposition of the kernel matrix which is necessary for the "Core" method.

3) Indefinite Wasserstein kernel: For this third method, we simply use the indefinite Gram matrix associated to (5) for the kernel matrix and solve the system (4) associated to the dual formulation of LS-SVM. While the associated optimization problem is not necessarily bounded in that case, the linear system (4) still has often a solution in practice. We name this method "Indefinite Wasserstein" in Fig. 3.

4) Gaussian RBF: The previous methods are compared with a classical LS-SVM classifier with kernel

$$k(\boldsymbol{u}, \boldsymbol{v}) = \exp\left(\frac{-\|\boldsymbol{u} - \boldsymbol{v}\|_2^2}{2\sigma^2}\right).$$

The parameters σ and γ are obtained by validation in the same way as above.

5) KNN: The same task is also performed for a kNN classifiers defined both with Euclidean and Wasserstein distances [20]. Those two methods are considered as benchmarks to assess the accuracy of the kernel methods hereabove. Notice that the number of nearest neighbours k is selected by validation.

C. Description of the simulations

The simulations are repeated several times and the mean classification error rate is given as well as the standard

deviation. We emphasize that the classes are balanced in each of the datasets. The code is provided on GitHub[2].

D. Discussion

The results obtained by classifiers defined with Wasserstein exponential kernel k_W outperform the Euclidean and Wasserstein kNN classifiers, as well as LS-SVM with a Gaussian RBF kernel (see Fig. 3 and Table I). The latter is especially outperformed when the number of training data points is limited to a few thousands. We observed empirically that the advantage of k_W is indeed reduced as the size of the training set further increases. The reweighted version of the kernel k_{RW} also proves to be competitive probably because it incorporates the information about the amount of ink, which was suppressed in the normal k_W due to the normalization. However, the amount of ink seems to be a better class indicator in the Quickdraw dataset than in the two datasets consisting of digits. This counter-intuitive result may point towards the need for an alternative way of reincorporating the suppressed information due to the normalization. Surprisingly, the classifier obtained for the indefinite k_W kernel yields the best performance when the training set is larger. This observation certainly deserves further research. For moderate size training sets, LS-SVM classifiers can be competitive with respect to other methods that do not rely on convolutional neural networks. The latter are known to be performant for relatively large training datasets. While an advantage of Wasserstein based methods is an increased accuracy in the classification tasks of this paper, a main disadvantage is the increased training time.

IV. Conclusion

In this paper, we proposed the use of Wasserstein squared exponential kernels for classifying shapes given relatively small training datasets. Although the computation of Wasserstein distances is expensive, it can be made possible thanks to the entropic regularization and the Sinkhorn algorithm, as it is well known. The so-called Wasserstein features are also proposed to serve as an approximation of the Wasserstein squared exponential kernel which is not necessarily positive semidefinite. In particular, this construction is possible if the bandwidth parameter is small enough as it is explained by elementary theoretical results. These theoretical results also open a door to more general exponential kernels based on any measure of similarity.

Acknowledgment

EU: The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Advanced Grant E-DUALITY (787960). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. Research Council KUL: Optimization frameworks for deep kernel machines C14/18/068. Flemish Government: FWO: projects: GOA4917N (Deep Restricted Kernel Machines: Methods and Foundations), PhD/Postdoc grant. This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. Ford KU Leuven Research Alliance Project KUL0076 (Stability analysis and performance improvement of deep reinforcement learning algorithms). The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government – department EWI.

[2]https://github.com/hdeplaen/Wasserstein_Exponential_Kernels

References

[1] A. Pandey, J. Schreurs, and J. A. K. Suykens, "Generative restricted kernel machines." arXiv:1906.08144.

[2] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[3] C. Villani, *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, 2008.

[4] G. Peyré and M. Cuturi, *Computational Optimal Transport: With Applications to Data Science*. Now Foundations and Trends, 2019.

[5] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic Analysis on Semigroups*, vol. 100 of *Graduate Texts in Mathematics*. New York, NY: Springer New York, 1984.

[6] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[7] M. Carrière, M. Cuturi, and S. Oudot, "Sliced wasserstein kernel for persistence diagrams," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, p. 664–673, JMLR.org, 2017.

[8] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde, "Generalized sliced wasserstein distances," in *Advances in Neural Information Processing Systems 32*, pp. 261–272, Curran Associates, Inc., 2019.

[9] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.

[10] X. Huang, A. Maier, J. Hornegger, and J. A. K. Suykens, "Indefinite kernels in least squares support vector machines and principal component analysis," *Applied and Computational Harmonic Analysis*, vol. 43, no. 1, pp. 162–172, 2017.

[11] C. S. Ong, X. Mary, S. Canu, and A. J. Smola, "Learning with non-positive kernels," in *Twenty-first international conference on Machine learning - ICML '04*, (New York, New York, USA), p. 81, ACM Press, 2004.

[12] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.

[13] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems 26*, pp. 2292–2300, Curran Associates, Inc., 2013.

[14] 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 30*, pp. 1964–1974, Curran Associates, Inc., 2017.

[15] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, p. 293–300, June 1999.

[16] J. J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 550–554, May 1994.

[17] J. A. K. Suykens and J. Vandewalle, "Multiclass least squares support vector machines," in *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No.99CH36339)*, vol. 2, pp. 900–903 vol.2, July 1999.

[18] J. Mairal, "End-to-end kernel learning with supervised convolutional kernel networks," in *Advances in Neural Information Processing Systems 29*, pp. 1399–1407, Curran Associates, Inc., 2016.

[19] D. Chen, L. Jacob, and J. Mairal, "Biological sequence modeling with convolutional kernel networks," *Bioinformatics*, vol. 35, pp. 3294–3302, 02 2019.

[20] M. Snow and J. Van Lent, "Monge's optimal transport distance for image classification." arXiv:1612.00181.