A Stable Multi-Scale Kernel for Topological Machine Learning

Supplementary material

Jan Reininghaus[1] Stefan Huber[1] Ulrich Bauer[1,2] Roland Kwitt[3]

[1]IST Austria, [2]TU München, [3]University of Salzburg

This supplementary material contains additional technical details and illustrations.

1 Indefiniteness of $d_{W,p}$

It is tempting to try to employ the Wasserstein distance for constructing a kernel on persistence diagrams. For instance, in Euclidean space, $k(x,y) = -\|x-y\|^2, x, y \in \mathbb{R}^n$ is conditionally positive definite and can be used within SVMs. Hence, the question arises if $k(x,y) = -d_{W,p}(x,y), x, y \in \mathcal{D}$ can be used as well.

In the following, we demonstrate (via counterexamples) that neither $-d_{W,p}$ nor $\exp(-\xi d_{W,p}(\cdot,\cdot))$ – for different choices of p – are (conditionally) positive definite. Thus, they cannot be employed in kernel-based learning techniques.

First, we briefly repeat some definitions to establish the terminology; this is done to avoid potential confusion, w.r.t. references [2, 1, 5]), about what is referred to as (conditionally) positive/negative definiteness in the context of kernel functions.

Definition 1. *A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is called positive definite (p.d.) if $\mathbf{c}^\top \mathbf{A}\mathbf{c} \geq 0$ for all $\mathbf{c} \in \mathbb{R}^n$. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is called negative definite (n.d.) if $\mathbf{c}^\top \mathbf{A}\mathbf{c} \leq 0$ for all $\mathbf{c} \in \mathbb{R}^n$.*

Note that in literature on linear algebra the notion of definiteness as introduced above is typically known as semidefiniteness. For the sake of brevity, in the kernel literature the prefix "semi" is typically dropped.

Definition 2. *A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is called conditionally positive definite (c.p.d.) if $\mathbf{c}^t \mathbf{A}\mathbf{c} \geq 0$ for all $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n$ s.t. $\sum_i c_i = 0$. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is called conditionally negative definite (c.n.d.) if $\mathbf{c}^\top \mathbf{A}c \leq 0$ for all $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n$ s.t. $\sum_i c_i = 0$.*

Definition 3. *Given a set $\mathcal{X}$, a function $k\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is a* positive definite kernel *if there exists a Hilbert space $\mathcal{H}$ and a map $\Phi\colon \mathcal{X} \to \mathcal{H}$ such that $k(x,y) = \langle \Phi(x), \Phi(y)\rangle_{\mathcal{H}}$.*

Typically a positive definite kernel is simply called *kernel*. Roughly speaking, the utility of p.d. kernels comes from the fact that they enable the "kernel-trick", *i.e.*, the use of algorithms that can be formulated in terms of dot products in an implicit feature space [5]. However, as shown by Schölkopf

in [4], this "kernel-trick" also works for distances, leading to the larger class of c.p.d. kernels (see Definition 4), which can be used in kernel-based algorithms that are translation-invariant (*e.g.*, SVMs or kernel PCA).

Definition 4. *A function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is (conditionally) positive (negative, resp.) definite kernel if and only if k is symmetric and for every finite subset $\{x_1, \dots, x_m\} \subseteq \mathcal{X}$ the Gram matrix $(k(x_i, x_j))_{i,j=1,1}^{m,m}$ is (conditionally) positive (negative, resp.) definite.*

To demonstrate that a function is not c.p.d. or c.n.d., resp., we can look at the eigenvalues of the corresponding Gram matrices. In fact, it is known that a matrix $\mathbf{A}$ is p.d. if and only if all its eigenvalues are nonnegative. The following lemmas from [1] give similar, but weaker results for (nonnegative) c.n.d. matrices, which will be useful to us.

Lemma 1 (see Lemma 4.1.4 of [1])**.** *If $\mathbf{A}$ is a c.n.d. matrix, then $\mathbf{A}$ has at most one positive eigenvalue.*

Corollary 1 (see Corollary 4.1.5 of [1])**.** *Let $\mathbf{A}$ be a nonnegative, nonzero matrix that is c.n.d. Then $\mathbf{A}$ has exactly one positive eigenvalue.*

The following theorem establishes a relation between c.n.d. and p.d. kernels.

Theorem 2 (see Chapter 2, §2, Theorem 2.2 of [2])**.** *Let $\mathcal{X}$ be a nonempty set and let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be symmetric. Then k is a conditionally negative definite kernel if and only if $\exp(-\xi k(\cdot, \cdot))$ is a positive definite kernel for all $\xi > 0$.*

In the MATLAB code (`test_negative_type_simple.m`) attached to the supplementary material, we generate simple examples for which the Gram matrix $\mathbf{A} = (d_{W,p}(x_i, x_j))_{i,j=1,1}^{m,m}$ – for various choices of p – has at least two positive and two negative eigenvalue. Thus, it is neither (c.)n.d. nor (c.)p.d. according to Corollary 1. Consequently, the function $\exp(-d_{W,p})$ is not p.d. either, by virtue of Theorem 2. To run the MATLAB code, simply execute:

```
load options_cvpr15.mat; % this will load a variable 'options'
test_negative_type_simple(options);
```

This will generate a short summary of the eigenvalue computations for a selection of values for p, including $p = \infty$ (bottleneck distance). The output of the MATLAB script can also be found in `test_negative_type_simple_output.pdf`.

Remark. While our simple counterexamples suggest that typical kernel constructions using $d_{W,p}$ for different p (including $p = \infty$) do not lead to (c.)p.d. kernels, a formal assessment of this question remains an open research question.

2 Plots of the feature map Φ_σ

Given a persistence diagram D, we consider the solution $u\colon \Omega \times \mathbb{R}_{\geq 0} \to \mathbb{R}, (x, t) \mapsto u(x, t)$ of the following partial differential equation

$$\begin{aligned} \Delta_x u &= \partial_t u && \text{in } \Omega \times \mathbb{R}_{>0}, \\ u &= 0 && \text{on } \partial\Omega \times \mathbb{R}_{\geq 0}, \\ u &= \sum_{p \in D} \delta_p && \text{on } \Omega \times \{0\}. \end{aligned}$$

To solve the partial differential equation, we extend the domain from Ω to $\mathbb{R}^2$ and consider for each $p \in D$ a Dirac delta δ_p and a Dirac delta $-\delta_{\overline{p}}$, as illustrated in Fig. 1 (left). By convolving $\sum_{p\in D} \delta_p - \delta_{\overline{p}}$ with a Gaussian kernel, see Fig. 1 (right), we obtain a solution $u\colon \mathbb{R}^2 \times \mathbb{R}_{\geq 0} \to \mathbb{R}, (x,t) \mapsto u(x,t)$ for the following partial differential equation:

$$\begin{aligned} \Delta_x u &= \partial_t u && \text{in } \mathbb{R}^2 \times \mathbb{R}_{>0}, \\ u &= \sum_{p\in D} \delta_p - \delta_{\overline{p}} && \text{on } \mathbb{R}^2 \times \{0\}. \end{aligned}$$

Restricting the solution u to $\Omega \times \mathbb{R}_{\geq 0}$, we then obtain the following solution $u\colon \Omega \times \mathbb{R}_{\geq 0} \to \mathbb{R}$,

$$u(x,t) = \frac{1}{4\pi t} \sum_{p\in D} e^{-\frac{\|x-p\|^2}{4t}} - e^{-\frac{\|x-\overline{p}\|^2}{4t}} \tag{1}$$

for the original partial differential equation and $t > 0$. This yields the feature map $\Phi_\sigma\colon \mathcal{D} \to L_2(\Omega)$:

$$\Phi_\sigma(D)\colon \Omega \to \mathbb{R}, \quad x \mapsto \frac{1}{4\pi\sigma} \sum_{p\in D} e^{-\frac{\|x-p\|^2}{4\sigma}} - e^{-\frac{\|x-\overline{p}\|^2}{4\sigma}}. \tag{2}$$

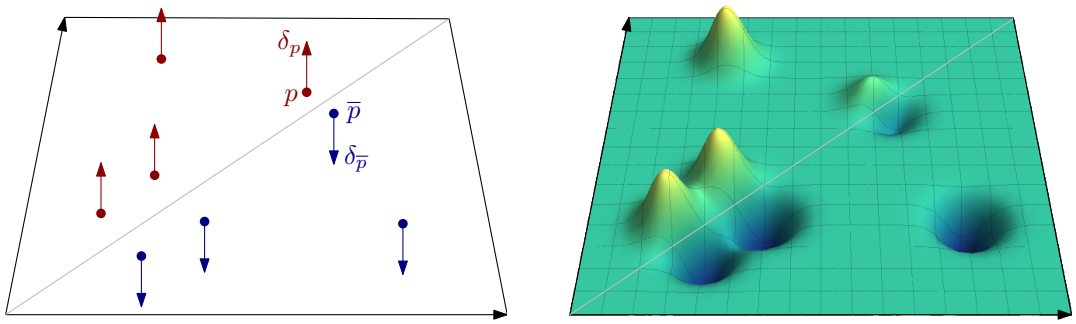

Figure 1: Solving the partial differential equation: First (left), we extend the domain from Ω to $\mathbb{R}^2$ and consider for each $p \in D$ a Dirac delta δ_p (red) and a Dirac delta $-\delta_{\overline{p}}$ (blue). Next (right), we convolve $\sum_{p\in D} \delta_p - \delta_{\overline{p}}$ with a Gaussian kernel.

In Fig. 2, we illustrate the effect of an increasing scale σ on the feature map $\Phi_\sigma(D)$. Note that in the right plot the influence of the low-persistence point close to the diagonal basically vanishes. This effect is essentially due to the Dirichlet boundary condition and is responsible for gaining stability for our persistence scale-space kernel k_σ.

3 Closed-form solution for k_σ

For two persistence diagrams F and G, the persistence scale-space kernel $k_\sigma(F,G)$ is defined as $\langle \Phi_\sigma(F), \Phi_\sigma(G) \rangle_{L_2(\Omega)}$, which is

$$k_\sigma(F,G) = \int_\Omega \Phi_\sigma(F)\, \Phi_\sigma(G)\, dx.$$

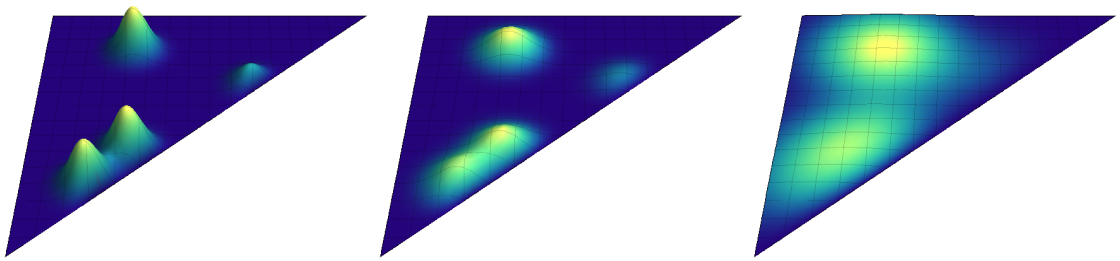

Figure 2: An illustration of the feature map $\Phi_\sigma(D)$ as a function in $L_2(\Omega)$ at growing scales σ (from left to right).

By extending its domain from Ω to $\mathbb{R}^2$, we see that $\Phi_\sigma(D)(x) = -\Phi_\sigma(D)(\overline{x})$ for all $x \in \mathbb{R}^2$. Hence, $\Phi_\sigma(F)(x) \cdot \Phi_\sigma(G)(x) = \Phi_\sigma(F)(\overline{x}) \cdot \Phi_\sigma(G)(\overline{x})$ for all $x \in \mathbb{R}^2$, and we obtain

$$
\begin{aligned}
k_\sigma(F,G) &= \frac{1}{2}\int_{\mathbb{R}^2} \Phi_\sigma(F)\,\Phi_\sigma(G)\,dx \\
&= \frac{1}{2}\frac{1}{(4\pi\sigma)^2}\int_{\mathbb{R}^2}\left(\sum_{p\in F} e^{-\frac{\|x-p\|^2}{4\sigma}} - e^{-\frac{\|x-\overline{p}\|^2}{4\sigma}}\right)\cdot\left(\sum_{q\in G} e^{-\frac{\|x-q\|^2}{4\sigma}} - e^{-\frac{\|x-\overline{q}\|^2}{4\sigma}}\right) dx \\
&= \frac{1}{2}\frac{1}{(4\pi\sigma)^2}\sum_{\substack{p\in F\\ q\in G}}\int_{\mathbb{R}^2}\left(e^{-\frac{\|x-p\|^2}{4\sigma}} - e^{-\frac{\|x-\overline{p}\|^2}{4\sigma}}\right)\cdot\left(e^{-\frac{\|x-q\|^2}{4\sigma}} - e^{-\frac{\|x-\overline{q}\|^2}{4\sigma}}\right) dx \\
&= \frac{1}{(4\pi\sigma)^2}\sum_{\substack{p\in F\\ q\in G}}\int_{\mathbb{R}^2} e^{-\frac{\|x-p\|^2+\|x-q\|^2}{4\sigma}} - e^{-\frac{\|x-p\|^2+\|x-\overline{q}\|^2}{4\sigma}}\,dx.
\end{aligned}
$$

We calculate the integrals as follows:

$$
\begin{aligned}
\int_{\mathbb{R}^2} e^{-\frac{\|x-p\|^2+\|x-q\|^2}{4\sigma}}\,dx &= \int_{\mathbb{R}^2} e^{-\frac{\|x-(p-q)\|^2+\|x\|^2}{4\sigma}}\,dx \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} e^{-\frac{(x_1-\|p-q\|)^2+x_2^2\,+\,x_1^2+x_2^2}{4\sigma}}\,dx_1\,dx_2 \\
&= \int_{\mathbb{R}} e^{-\frac{x_2^2}{2\sigma}}\,dx_2 \cdot \int_{\mathbb{R}} e^{-\frac{(x_1-\|p-q\|)^2+x_1^2}{4\sigma}}\,dx_1 \\
&= \sqrt{2\pi\sigma}\cdot\int_{\mathbb{R}} e^{-\frac{(x_1-\|p-q\|)^2+x_1^2}{4\sigma}}\,dx_1 \\
&= \sqrt{2\pi\sigma}\cdot\int_{\mathbb{R}} e^{-\frac{(2x_1-\|p-q\|)^2+\|p-q\|^2}{8\sigma}}\,dx_1 \\
&= \sqrt{2\pi\sigma}\,e^{-\frac{\|p-q\|^2}{8\sigma}}\cdot\int_{\mathbb{R}} e^{-\frac{(2x_1-\|p-q\|)^2}{8\sigma}}\,dx_1 \\
&= \sqrt{2\pi\sigma}\,e^{-\frac{\|p-q\|^2}{8\sigma}}\cdot\int_{\mathbb{R}} e^{-\frac{x_1^2}{2\sigma}}\,dx_1 \\
&= 2\pi\sigma\,e^{-\frac{\|p-q\|^2}{8\sigma}}.
\end{aligned}
$$

In the first step, we applied a coordinate transform that moves $x - q$ to x. In the second step, we performed a rotation such that $p - q$ lands on the positive x_1-axis at distance $\|p - q\|$ to the origin and we applied Fubini's theorem. We finally obtain the closed-form expression for the kernel k_σ as:

$$\begin{aligned} k_\sigma(F, G) &= \frac{1}{(4\pi\sigma)^2} \, 2\pi\sigma \sum_{\substack{p \in F \\ q \in G}} e^{-\frac{\|p-q\|^2}{8\sigma}} - e^{-\frac{\|p-\overline{q}\|^2}{8\sigma}} \\ &= \frac{1}{8\pi\sigma} \sum_{\substack{p \in F \\ q \in G}} e^{-\frac{\|p-q\|^2}{8\sigma}} - e^{-\frac{\|p-\overline{q}\|^2}{8\sigma}} . \end{aligned}$$

4 Additional retrieval results on SHREC 2014

HKS t_i	d_{k_L}	d_{k_σ}	Δ	d_{k_L}	d_{k_σ}	Δ
t_1	59.9	71.3	+11.4	26.0	21.4	−4.6
t_2	**75.1**	76.0	+0.9	23.8	22.7	−1.1
t_3	49.6	64.8	+15.2	19.1	20.7	+1.6
t_4	59.4	**77.5**	+18.1	23.5	26.1	+2.6
t_5	68.1	75.2	+7.1	22.7	27.4	+4.7
t_6	50.0	55.2	+5.2	18.9	26.2	+7.3
t_7	47.6	53.6	+6.0	27.4	31.8	+4.4
t_8	53.1	62.4	+9.3	**45.3**	**39.8**	−5.5
t_9	51.2	56.3	+5.1	24.4	30.3	+5.9
t_{10}	39.6	49.7	+10.1	2.5	21.8	+19.3
Top-3 [3]	83.2 – 76.4 – 76.0			54.1 – 47.2 – 45.1		

Table 1: T1 retrieval performance. *Left:* SHREC 2014 (synthetic); *Right:* SHREC 2014 (real).

HKS t_i	d_{k_L}	d_{k_σ}	Δ	d_{k_L}	d_{k_σ}	Δ
t_1	87.7	91.4	+3.7	41.5	34.6	−6.9
t_2	**91.1**	**95.1**	+4.0	40.8	37.1	−3.7
t_3	70.4	83.4	+13.0	36.5	36.8	+0.3
t_4	77.7	93.6	+15.9	39.8	43.4	+3.6
t_5	90.8	92.3	+1.5	35.1	41.8	+6.7
t_6	73.9	75.4	+1.5	31.6	40.2	+8.6
t_7	70.6	74.4	+3.8	38.6	47.6	+9.0
t_8	73.3	79.3	+6.0	**56.5**	**57.6**	+1.1
t_9	72.7	76.2	+3.5	31.8	42.5	+10.7
t_{10}	57.8	66.6	+8.8	4.8	31.0	+26.2
Top-3 [3]	98.7 – 97.1 – 94.9			74.2 – 65.9 – 65.7		

Table 2: T2 retrieval performance. *Left:* SHREC 2014 (synthetic); *Right:* SHREC 2014 (real).

HKS t_i	d_{k_L}	d_{k_σ}	Δ	d_{k_L}	d_{k_σ}	Δ
t_1	60.6	65.3	+4.7	25.4	22.8	−2.6
t_2	**65.0**	67.4	+2.4	25.0	23.4	−1.6
t_3	48.4	58.8	+10.4	24.0	24.0	+0.0
t_4	55.2	**67.6**	+12.4	25.3	27.4	+2.1
t_5	63.7	66.2	+2.5	21.6	25.2	+3.6
t_6	51.0	52.7	+1.7	20.7	23.7	+3.0
t_7	48.4	51.7	+3.3	22.5	27.5	+5.0
t_8	51.1	56.5	+5.4	**30.2**	**33.2**	+3.0
t_9	50.4	53.2	+2.8	15.8	25.3	+9.5
t_{10}	39.8	46.7	+6.9	3.6	19.0	+15.4
Top-3 [3]	70.6 – 69.1 – 65.9			38.7 – 35.6 – 35.4		

Table 3: EM retrieval performance. *Left:* SHREC 2014 (synthetic); *Right:* SHREC 2014 (real).

HKS t_i	d_{k_L}	d_{k_σ}	Δ	d_{k_L}	d_{k_σ}	Δ
t_1	81.3	91.5	+10.2	53.0	49.6	−3.4
t_2	**92.1**	93.4	+1.3	51.1	51.3	+0.2
t_3	80.3	89.3	+9.0	47.7	48.4	+0.7
t_4	85.0	**93.8**	+8.8	52.7	55.5	+2.8
t_5	89.0	93.2	+4.2	51.2	55.5	+4.3
t_6	78.6	82.5	+3.9	48.1	54.2	+6.1
t_7	77.2	81.6	+4.4	55.7	60.5	+4.8
t_8	80.4	86.3	+5.9	**72.8**	**68.3**	−4.5
t_9	79.7	83.9	+4.2	50.4	61.0	+10.6
t_{10}	70.8	78.9	+8.1	27.7	51.3	+23.6
Top-3 [3]	97.7 – 93.8 – 92.7			78.1 – 71.7 – 71.2		

Table 4: DCG retrieval performance. *Left:* SHREC 2014 (synthetic); *Right:* SHREC 2014 (real).

References

[1] R. Bapat and T. Raghavan. *Nonnegative Matrices and Applications*. Cambridge University Press, 1997.

[2] C. Berg, J.-P. Reus-Christensen, and P. Ressel. *Harmonic Analysis on Semi-Groups – Theory of Positive Definite and Related Functions*. Springer, 1984.

[3] Pickup, D. *et al*.. SHREC '14 track: Shape retrieval of non-rigid 3d human models. In *Proceedings of the 7th Eurographics workshop on 3D Object Retrieval*, EG 3DOR'14. Eurographics Association, 2014.

[4] B. Schölkopf. The kernel-trick for distances. In *NIPS*, 2001.

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