Kernel over sets of vectors.

Phd Student: Babacar SOW

**Contract duration:** From 01/11/2021 to 31/10/2024

**Project:** ANR SAMOURAI

**University:** Ecole Des Mines de Saint-Etienne

**Supervisors:** Merlin KELLER (EDF), Rodolphe LE RICHE (CNRS/LIMOS), Julien PELAMATTI (EDF), Sanaa ZANNANE (EDF)

ETICS : October 3 to 7, 2022
Table of contents

1. Context and Problem

2. Bayesian Approach and Kernels between clouds of points

3. Perspectives

4. Bibliography
Context and problem

### Functions defined over sets of vectors

- Let $\mathcal{F}$ be the family of considered functions.
- Suppose $f \in \mathcal{F}$ and $D_f$ its domain of definition.
- $u \in D_f \Rightarrow \exists n \in \mathbb{N}, d \in \mathbb{N}, u = \{x_1, ..., x_n\}, \forall i, x_i \in \mathbb{R}^d$
- $n$ belongs to a finite discrete set.
- For any permutation $\pi$ of the set $\{1, ..., n\}$ to a new one $\{\pi(1), ..., \pi(n)\}$, we denote by $u_\pi$ the following set $\{x_{\pi(1)}, ..., x_{\pi(n)}\}$.
- Note that $\forall \pi, f(u_\pi) = f(u)$: $f$ is invariant under permutation.
- The variables $u$ will be called clouds of points.
### Related works and domains

#### Learning functions defined over sets of objects with kernels

- Kernels on bags of vectors, applied to SVM Classification on images in [7].
- Same technique to define kernel on graphs by averaging over kernels between paths in [13] to measure similarity between shapes.
- Classification on text data with a set representation view in [14].
- A Kernel between sets of points is used in [5] to optimize the layout of a wind farm.

#### Focus of this presentation

- In this presentation, we discuss some general methods to construct such kernels.
- Confronting them numerically on a test function mimicking the production of a windfarm.
Bayesian Approach

A Gaussian process prior

- Gaussian process is defined by a mean function $m$ and a kernel $k$ over the spaces of inputs $\mathcal{X}$ to approximate the functions.

- Observing $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i \in \mathcal{X}$ and $y \in \mathbb{R}$ as training data, the predictive mean and covariance for a new point $x$ are given by:

$$
\mu(x; D) = m(x) + K(X, x)^T K(X, X)^{-1}(y - m(X))
$$

$$
\Sigma(x, x; D) = K(x, x) - K(X, x)^T K(X, X)^{-1} K(X, x)
$$

Necessary Conditions on $k$

- $k$ must be **symmetric** and **positive definite,** i.e, for any $M$ distinct clouds of points, for any vector $c \in \mathbb{R}^M$, the following inequality must hold: $\sum_{i=1}^M \sum_{j=1}^M c_i c_j k(X_i, X_j) \geq 0$
Bayesian approach: Kernel trick and Mapping

Comparing two clouds

![Comparison of two clouds](attachment:image.png)
k is a positive definite kernel if and only if there exists a Hilbert space $\mathcal{H}$, and a function $\phi : \mathcal{X} \rightarrow \mathcal{H}$ such that $\forall x, y, k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$.

Explicit and Implicit Mappings

- Explicit Mapping: in some cases $\phi$ and the scalar product, $\langle ., . \rangle_{\mathcal{H}}$ are known by definition or by construction
- Implicit Mapping: in some cases, we just use the compact formula of $k$
  - **Substitutions Kernels** as in Haasdonk and Bahlmann [8].
Substitution with Hilbertian Distance

Substitution with Exponential

- Firstly, we consider covariance kernels of the form: \( k(X, Y) = \sigma^2 \exp\left(-\frac{\Psi(X,Y)}{2\theta^2}\right) \).
- Semi-definite positiveness is equivalent to \( \Psi \) be Hermitian (symmetric in the real case) and **conditionally negative semi-definite** [2].
- In other words, for any \( M \) distinct points and \( c \in \mathbb{R}^M \) with \( \sum_{i=1}^{M} c_i = 0 \), the following inequality must hold: \[ \sum_{i=1}^{M} \sum_{j=1}^{M} c_i c_j \Psi(X_i, X_j) \leq 0 \]

Metric Cases

- We consider cases where \( \Psi(X, Y) = d(\tilde{X}, \tilde{Y})^2 \)
- \( d \) is the distance between \( \tilde{X} \) and \( \tilde{Y} \) the respective images of \( X \) and \( Y \) into a known metric Space.
- The above conditions are equivalent to ensuring that the metric be **Hilbertian**, as stated in Haasdonk and Bahlmann [8].
How to construct $\tilde{X}$ and $\tilde{Y}$?

With probabilities

- **Case 1:** Suppose we have two clouds $X = (x_1, \ldots, x_n)$, $Y = (y_1, \ldots, y_m)$ and $P_X = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$, $P_Y = \frac{1}{m} \sum_{j=1}^{m} \delta_{y_j}$, the respective associated empirical uniform distributions.

- **Case 2:** We associate to each cloud of point $X = (x_1, \ldots, x_n)$, $Y = (y_1, \ldots, y_m)$, its empirical Gaussian: $N_X(m_X, \Sigma_X)$ and $N_Y(m_Y, \Sigma_Y)$. item $N_X$ is defined by $m_X = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\Sigma_X = \frac{1}{n} \sum_{i=1}^{n} (x_i - m_X)(x_i - m_X)^T$

With vectors: vectorization

- $\tilde{X}$ and $\tilde{Y}$ can be two vectors of features characteristics of the clouds.
Wasserstein Distances

For two measures $\mu$ and $\nu$ defined over a space $\mathcal{X}$, the Wasserstein distance of positive cost function $\rho$ and order $p$ is defined as follows: 

$$W_p^\rho = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} \rho(x, y)^p \, d\pi(x, y)$$

Substitution with Hilbertian distance: Sliced Wasserstein Distance (see Annex)

- Let $S = \{\theta \in \mathbb{R}^2, ||\theta|| = 1\}$. Consider the projected empirical measure on the line directed by $\theta \in S$: $\theta^* P_X = \frac{1}{n} \sum_{i=1}^n \delta_{\langle x_i, \theta \rangle}$ and $\theta^* P_Y = \frac{1}{m} \sum_{i=1}^m \delta_{\langle y_i, \theta \rangle}$

- $SW_2^2(P_X, P_Y) = \int_S W_2^2(\theta^* P_X, \theta^* P_Y) \, d\theta$. Implementation using POT [6]

- The covariance kernel $k(X, Y) = \sigma^2 \exp\left(-\frac{SW_2^2(P_X, P_Y)}{2\theta^2}\right)$ is symmetric and semi-definite positive as in Carriere, Cuturi, and Oudot [4]. It will be denoted Sliced Wasserstein subs

Approximate For Gaussian Modeling (see Annex), Gaussian Wasserstein subs

$$W_2^2 \approx ||m_X - m_Y||^2 + ||\Sigma_X^{1/2} - \Sigma_Y^{1/2}||_{\text{Frobenius}}^2$$

as in Bui et al. [3] (if $\Sigma_X^{1/2} \Sigma_Y^{1/2} = \Sigma_X^{1/2} \Sigma_Y^{1/2}$)
Substitution with Hilbertian distance: MMD

- There exists a Reproducing Kernel Hilbert Space, $\mathcal{H}$ with a characteristic kernel such as $k_{\mathcal{H}}(x, .) = \exp(-\frac{||x-||^2}{2\theta^2})$.

- The characteristic nature guarantees the injectivity of the embedding map Muandet et al. [11]: $P_X \mapsto \mu_X(.) = \int P_X(x) k_{\mathcal{H}}(x, .) \, dx$.

- $\text{MMD}^2(P_X, P_Y) = \|\mu_X - \mu_Y\|^2_{\mathcal{H}}$

- For any kernel $k_{\mathcal{H}}$ of the RKHS, the uniform discrete (supported by points) laws give $\text{MMD}^2(P_X, P_Y) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k_{\mathcal{H}}(x_i, x_j) + \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} k_{\mathcal{H}}(y_i, y_j) - 2 \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k_{\mathcal{H}}(x_i, y_j)$

- The covariance kernel $k(X, Y) = \sigma^2 \exp(-\frac{||\mu_X - \mu_Y\|^2_{\mathcal{H}}}{2\theta^2})$ is symmetric and definite positive.

- We will denote the latter as MMD.
Constructing Features of a cloud

Relevant Features Map Kernel

- We consider a final kernel of the form
  \[ k(X, Y) = \sigma^2 \exp \left( - \sum_{j=1}^{n'} \frac{|w'_j(X) - w'_j(Y)|^2}{\theta_j^2} \right) \]
  with \((w'_1(X), \ldots, w'_{n'}(X))\) a vector of features.

- As features we consider:
  - The coordinates of the mean
  - The eigenvalues and eigenvectors of the empirical covariance matrix.
  - The number of points in the set
  - Greatest and shortest distances between points of the set.

- This kernel will be called Relevant Feature Kernel.
Explicit Mappings: Probability Product Kernels and Embeddings

Explicit Mappings (see Annex)

- Recall $k(x, y) = \langle \phi(x), \phi(y) \rangle$

- We consider the case where the mapping $\phi$ is known.
  - $\phi(X) = P^\rho_X$ with $\rho \in [0, 1]$ where $P_X$ is an underlying empirical distribution. $k(x, y) = \int_\Omega P(x)^\rho P'(x)^\rho dx$, Jebara and Kondor [9]. This family of kernels are called Probability Product Kernels. For two Gaussians $P_X = \mathcal{N}(\mu, \Sigma)$ and $P_Y = \mathcal{N}(\mu', \Sigma')$, one gets:

$$k(x, y) = (2\pi)^{(1-2\rho)D/2} |\Sigma^+|^{1/2} |\Sigma|^{-\rho/2} |\Sigma|^{-\rho/2} \exp \left( -\frac{\rho}{2} \mu^\top \Sigma^{-1} \mu - \frac{\rho}{2} \mu'^\top \Sigma'^{-1} \mu' + \frac{1}{2} \mu^+ \Sigma^+ \mu + \frac{1}{2} \mu'^+ \Sigma'^+ \mu' \right)$$

where $\Sigma^+ = (\rho \Sigma^{-1} + \rho' \Sigma'^{-1})^{-1}$ and $\mu^+ = \rho \Sigma^{-1} \mu + \rho' \Sigma'^{-1} \mu'$

- If $\rho = \frac{1}{2}$, it is called the **Bhattacharrya Kernel** and when $\rho = 1$ Expected Likelihood Kernel.

- $\phi(X) = \mu_X$ where $\mu_X$ is the embedding of the underlying empirical distribution into an RKHS. $k(x, y) = \langle \mu_X, \mu_Y \rangle$ it will be called **MMK**, Mean Map Kernel, for the remainder.
A test function

Mimicking wind farms

• We consider the following family of test functions mimicking wind-farms productions

\[ F(\{x_1, \ldots, x_n\}) = \sum_{i=1}^{n} \sum_{j} f_p(x_j, x_i)f_0(x_i) \]

where \( f_p(x_j, x_i) \) expresses the energy loss over \( x_i \) that is caused by \( x_j \) and \( f_0 \) is a constant.

\( x_i \in \mathbb{R}^2 \) and \( i \in \{10, 11, \ldots, 20\} \)

• The function \( x_i \mapsto f_p(x_j, x_i) \) can be parametrized differently:
  
  • It can be unidirectional with an arbitrary angle
  
  • It can be multi-directional
A test function

Mimicking wind farms: Example

In the following we represent: $x_i \mapsto f_p(x_0, x_i)$ on the left, F with a one varying point on the right. We note F with $f_p$ on left $F_0$.

Mimicking wind farms: Illustration
A test function

Mimicking wind farms: Example

In the following we represent: $x_i \mapsto f_p(x_0, x_i)$ with $\pi/4$ rotated direction, and 40 directions on the right. We note $F$ with $f_p$ on left $F_{45}$ and $F_{40d}$ for the $f_p$ on the right.

Mimicking wind farms: Illustration
Preleminary Results: 0° Interaction Function

- Modeling with Gaussians distributions is weaker than with discrete uniform ones for this function.
- Sliced Wasserstein Kernel is very competitive with MMD ;

![Graphs showing prediction results for different kernels.](image-url)
Results: 45° direction Interaction

- 45° direction does not change performance for lot of kernels but Feature Map Kernel.

![Graphs showing prediction performance for different kernels.](image)

Figure: Prediction performance on 45° direction Interaction Function
Preleminary Results: 40 directions integrated

- 40 directions integrated Function improves slightly Gaussian based kernels.
- MMD shows better results than Relevant Feature kernel and Sliced Wasserstein

Figure: Prediction performance on 40 directions integrated function
Table: Summary of the Q2 observed: Battacha refers to Bhattacharrya kernel, RFK (Relevant Feature kernel), SWS (Sliced Wasserstein subs), GWS (Gaussian Wasserstein subs)

<table>
<thead>
<tr>
<th>Function</th>
<th>Kernels</th>
<th>MMD</th>
<th>MMK</th>
<th>Battacha</th>
<th>RFK</th>
<th>SWS</th>
<th>GWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_0$</td>
<td></td>
<td>0.917</td>
<td>0.711</td>
<td>0.144</td>
<td>0.813</td>
<td>0.812</td>
<td>0.174</td>
</tr>
<tr>
<td>$F_{45}$</td>
<td></td>
<td>0.887</td>
<td>0.739</td>
<td>0.186</td>
<td>0.74</td>
<td>0.841</td>
<td>0.189</td>
</tr>
<tr>
<td>$F_{40d}$</td>
<td></td>
<td>0.88</td>
<td>0.279</td>
<td>0.314</td>
<td>0.688</td>
<td>0.798</td>
<td>0.259</td>
</tr>
</tbody>
</table>

- MMD remains the most robust kernels. MMK fails to model a lot of directions integrated.
- Modeling clouds as Gaussian seem very poor in front of discrete uniforms modelization.
- SWS and RFK are very competitive with MMD.
Perspectives

Scientific Perspectives

- Concerning Relevant Feature kernel, find automatically the most relevant features for a given function.
- For MMD and MMK, model with non uniform probabilities. Considering different weights on points could allow giving more importance to some specific points of the cloud.
- Define the directions of Sliced Wasserstein Distance by Log Likelihood.
Thanks For Your Attention!


Distance between laws: Wasserstein Distance

**Substitution with Hilbertian distance: Wasserstein Distance in 1D Case**

- Definition and properties see Carriere, Cuturi, and Oudot [4] and Kolouri, Zou, and Rohde [10]
- Let $\mu$ and $\nu$ be two nonnegative measures in $\mathbb{R}$ with $\mu(\mathbb{R}) = \nu(\mathbb{R}) = 1$. The Wasserstein distance of order 2 between $\mu$ and $\nu$ is defined as follows:

$$\mathcal{W}_2^2(\mu, \nu) = \inf_{P \in \Pi(\mu, \nu)} \int \int_{\mathbb{R} \times \mathbb{R}} |x - y|^2 P(dx, dy)$$

- Let $C_\mu(x) = \int_{-\infty}^x d\mu$, $C_\nu(x) = \int_{-\infty}^x d\nu$ their cumulative distribution function.
- Pseudo-inverse: $\forall r \in [0, 1], C_\mu^{-1}(r) = \min_x \{x \in \mathbb{R} \cup \{-\infty\} : C_\mu(r) \geq x\}$
- Then $\mathcal{W}_2^2(\mu, \nu) = \|C_\mu^{-1} - C_\nu^{-1}\|_{L^p([0,1])}^2$, see Peyré, Cuturi, et al. [12]
- $\mathcal{W}_2^2(\mu, \nu)$ is symmetric and conditionally negative definite. (Kolouri, Zou, and Rohde [10])
- If $\mu$ and $\nu$ are defined in $\mathbb{R} \times \mathbb{R}$, the above condition is no longer guaranteed.
For two measures $\mu$ and $\nu$ defined over a space $X$, the Wasserstein distance of positive cost function $\rho$ and order $p$ is defined as follows: 
\[ W_p^p = \inf_{\pi \in \Pi(\mu, \nu)} \int_{X \times X} \rho(x, y)^p d\pi(x, y) \]

We consider the case $2$

For an Euclidean cost in 2D, the Wasserstein distance of two Gaussians is given in a closed form as:
\[ W_2^2 = ||m_X - m_Y||^2 + \text{tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_X^{1/2} \Sigma_Y \Sigma_X^{1/2})^{1/2}) \]

Consider the version $W_2^2 = ||m_X - m_Y||^2 + ||\Sigma_X^{1/2} - \Sigma_Y^{1/2}||_{\text{Frobenius}}$ as in Bui et al. [3]

The above distance is conditionally negative definite and $k(X, Y) = \sigma^2 \exp(-\frac{W_2^2}{2\theta^2})$ is therefore a valid kernel.