Kernels between clouds of points.

Phd Student: Babacar SOW since 01/11/2021

University: Ecole Des Mines Saint-Etienne
PHD Supervisors: Rodolphe LE RICHE, Julien PELAMATTI, Sanaa ZANNANE, Merlin KELLER

March 28, 2022

Table of contents

Context

Main objective in wind farm

- In a wind farm, one seeks to find the positions that minimize the losses of interaction.
- At same time, minimizing the cost of installation and maintenance.

Wind farm view

Problem Formulation

Optimization problem

- f takes its values in $\mathbb{R}$
- g may have vectorial outputs, examples of h, $\mathcal{A}$, and ϕ are given below.

$$\min_{d,X,C} f(X,C)$$

$$d \in \{n, \ldots, N\}$$

$$X \in \mathcal{X} \subset (\mathbb{R}^2)^d$$

$$C \in \{(E, V, L), E = h(X), V \in \mathcal{A}, L \in \phi(V)\}$$

$$g(X,C) \leq 0$$

Equivalence in Wind Farm

In wind farm optimization, d represents the number of turbines, X is the set of positions and C is a graph of the cable with labels as local costs and f can be cost-production.

Problem Formulation: Examples of families of h, $\mathcal{A}$, L, ϕ and V

Examples of h

$$\begin{array}{rrcl} h: & (\mathbb{R}^2)^d & \longrightarrow & (\mathbb{R}^2)^d \cup (\mathbb{R}^2)^{d+k}. \\ h: & (x_1, x_2, \ldots x_d) & \longmapsto & (x_1, x_2, \ldots x_d, x'_1, .., x'_k) \end{array}$$

Example of $\mathcal{A}$

- In this case C is composed of k connected components.
- $\mathcal{A}$ is the set of vertices achieving the minimum spanning tree.
- The labels L are the pairwise distances and $V = (V_1, \ldots, V_k)$ where V_i is the vertices achieving the minimum spanning tree of each component

Example of ϕ

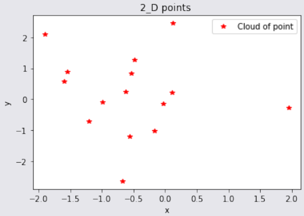
So $V = (v_1, \ldots, v_M)$ with $v_i = \big((a_i, b_i), (c_i, d_i)\big)$, $a_i, b_i, c_i, d_i \in R$ and $\phi(V) = (||(a_1, b_1) - (c_1, d_1)||_2, \ldots, ||(a_M, b_M) - (c_M, d_M)||_2)$.

Problem Formulation: Sub-optimize C

Optimize while sub-optimizing C

- Consider $f(d, X) = min_C f(d, X, C)$ where the minimization is carried out by an algorithm for minimum spanning trees.
- The global variable of optimization can be written as $(x_1, ..., x_d)$ where $\forall k \in \{1, ..d\}, x_k \in \mathbb{R}^2$.
- It is invariant under permutation and can be represented as a cloud of points.

A cloud



Problem Formulation: Example of Analytical Function

An analytical function depending only on the positions

- Consider the set of discrete values $\{15, ..20\}$
- For any cloud of 2D points $X = \{x_1, ..., x_n\}$ with $x_i = (x_{i,1}, x_{i,2}), \forall i$
- Consider a positive function $f_p(x_k, x_l)$ measuring the influence of x_k on x_l with $x_{k,1} \leq x_{l,1}$ and bounded by 1.
- A constant positive function $f_0(x_i)$
- $f(x_1, ..., x_n) = \sum_{i=1}^{n} \sum_{j, x_{j,1} \leq x_{i,1}} f_p(x_j, x_i) f_0(x_i)$
- The function is invariant under permutation.
- The variable of optimization can be represented as a cloud of point.

Problem Formulation: Example of Analytical Function

Simplistic analytical function

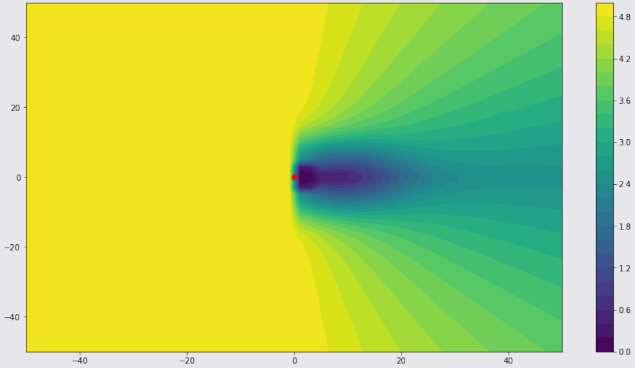


Figure: Value of $f_p((-40, 0), x)$

State of Art on Turbine Positions Optimization

Stochastic

- Genetic Algorithms as in [12] based on cross-over, mutation and selection on the positions of the turbines. The global variable is coded as a binary string.
- Particle Swarm Optimization [7]: consider layouts as swarms, elements as particles and use best local and global result for updating.
- Random Search [8], random individual move, check feasibility with stopping criterion.
- Simulated Annealing [3]: cooling temperature, metropolis hasting to avoid local convergence.

Non Stochastic

- Greedy Heuristic Algorithms [6]: adding clouds successively on best positions.
- Mixed Integer Linear Programming Approach [1] : Modeling the problem as linear including integer and real variables for simplification.

Gaussian Processes over clouds of points: Kernel Function

- We consider a function f in a compact space $\mathcal{X} \subset \bigcup_{d=n}^{N} (R^2)^d$.
- Metamodel f with a Gaussian process.
- First necessary ingredient : a **kernel function**, $k(X, Y)$ between the clouds of points X and Y.

Necessary Conditions on k

To define a Gaussian process k must satisfy the following conditions.

- Symmetry: for two cloud of points, X and Y, $k(X, Y) = k(Y, X)$.
- Positive definite: for any M distinct clouds of points, the gram matrix K defined by $K_{ij} = k(X_i, X_j)$ must be semi-definite positive. In other words, for any vector $c \in R^M$, the following inequality must hold: $\sum_{i=1}^{M} \sum_{j=1}^{M} c_i c_j k(X_i, X_j) \geq 0$

Gaussian Processes over clouds of points : Substitution Kernels

Substitution with Exponential

- Firstly, we consider correlation kernels of the form: $k(X, Y) = exp(-\frac{\Psi(X,Y)}{2\theta^2})$.
- We know that k defined above is a valid kernel (symmetric and positive semi-definite) if and only if Ψ is Hermitian (symmetric in the real case) and conditionally negative semi-definite [2].
- In other words, for any M distinct points and $c \in R^M$ 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

- If $\Psi(X, Y) = d(\phi(X), \phi(Y))^2$ where d is the distance between $\phi(X)$ and $\phi(Y)$ the respective images of X and Y in an metric space Space, the above conditions are equivalent to the fact that the metric be Hilbertian..

Gaussian Processes over clouds of points : Substitution Kernels

Hilbertian Cases: MMD

- Suppose we have two clouds $X = (x_1, ..x_n)$, $Y = (y_1, ..., y_m)$ 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.
- 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 [13]: $P_X \longmapsto \mu_X = \int P_X(x)k_{\mathcal{H}}(x, .)\mathrm{d}x$.
- $MMD^2(P_X, P_Y) = ||\mu_X - \mu_Y||^2_{\mathcal{H}}$
- For any kernel $k_{\mathcal{H}}$ of the RKHS, the uniform empirical laws give $MMD^2(P_X, P_Y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k_{\mathcal{H}}(x_i, x_j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k_{\mathcal{H}}(y_i, y_j) - 2\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}k_{\mathcal{H}}(x_i, y_j)$
- The correlation kernel $k(X, Y) = exp(-\frac{||\mu_X-\mu_Y||^2_{\mathcal{H}}}{2\theta^2})$ is symmetric and definite positive.

Gaussian Processes over clouds of points : Substitution Kernels

Wasserstein Distance in 1D Case: Definition and properties see [5] and [11]

- Let μ and ν be two nonnegative measures in $\mathbb{R}$ with $\mu(\mathbb{R}) = \nu(\mathbb{R}) = 1$. The Wasserstein distance of order 2 between μ and ν is defined as folllows:

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

- Let $\mathcal{C}_\mu(x) = \int_{-\infty}^{x} d\mu$, $\mathcal{C}_\nu(x) = \int_{-\infty}^{x} d\nu$ their cumulative distribution function.
- Pseudo-inverse : $\forall r \in [0,1], \mathcal{C}_\mu^{-1}(r) = \min_x\{x \in \mathbb{R}\cup\{-\infty\} : \mathcal{C}_\mu(r) \geq x\}$
- Then $\mathcal{W}_2^2(\mu,\nu) = ||\mathcal{C}_\mu^{-1} - \mathcal{C}_\nu^{-1}||^2_{L^p([0,1])}$, see [15]
- $\mathcal{W}_2^2(\mu,\nu)$ is symmetric and conditionally negative definite. ([11])
- If μ and ν are defined in $\mathbb{R}\times\mathbb{R}$, the above condition is no longer guaranteed.

Gaussian Processes over clouds of points : Substitution Kernels

Sliced Wasserstein Distance on empirical uniforms in 2D

- For a cloud of points $X = (x_1, ..., x_n), x_j \in \mathbb{R} \times \mathbb{R}, \forall j \in \{1, ..., n\}$, the empirical uniform measure, $P_X = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$, is defined in $\mathbb{R} \times \mathbb{R}$.
- Solution : Use Sliced Wasserstein Distance
- Consider the empirical probabilities $P_X = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$, $P_Y = \frac{1}{m} \sum_{j=1}^{m} \delta_{y_j}$ of two clouds of points
- Let $\mathcal{S} = \{\theta \in \mathbb{R}^2, ||\theta|| = 1\}$
- Consider the projected empirical measure on the line directed by $\theta \in \mathcal{S}$: $\theta^* P_X = \frac{1}{n} \sum_{i=1}^{n} \delta_{<x_i,\theta>}$ and $\theta^* P_Y = \frac{1}{m} \sum_{i=1}^{m} \delta_{<y_i,\theta>}$
- $SW_2^2(P_X, P_Y) = \int_{\mathcal{S}} \mathcal{W}_2^2(\theta^* P_X, \theta^* P_Y) \mathrm{d}\theta$,
- Implementation using POT [9]
- The correlation kernel $k(X, Y) = exp(-\frac{SW_2^2(P_X, P_Y)}{2\theta^2})$ is symmetric and semi-definite positive as in [5]

Gaussian Processes over clouds of points : Substitution Kernels

Wasserstein Distance Between Empirical Gaussians

- For two measures μ and ν defined over a space X, the Wasserstein distance of positive cost function ρ and order p is defined as follows : $W_p^p = \inf_{P \in \Pi(\mu,\nu)} \int_{X \times X} \rho(x,y) \mathrm{d}\pi(x,y)$
- We associate to each cloud of point $X = (x_1, ..x_n)$, $Y = (y_1, ..., y_m)$, its empirical Gaussian: $\mathcal{N}_{\mathcal{X}}(m_X, \Sigma_X)$ and $\mathcal{N}_{\mathcal{Y}}(m_Y, \Sigma_Y)$.
- $\mathcal{N}_{\mathcal{X}}$ is defined by $m_X = \frac{1}{n} \sum_{i=1}^n x_i$ and $\Sigma_X = \frac{1}{n} \sum_{i=1}^n (x_i - m_X)(x_i - m_X)^T$
- $\mathcal{N}_{\mathcal{Y}}$ is defined by $m_Y = \frac{1}{m} \sum_{j=1}^m y_j$ and $\Sigma_Y = \frac{1}{m} \sum_{i=1}^m (y_j - m_Y)(y_j - m_Y)^T$
- For an Euclidean cost in 2D , the Wasserstein distance of two Gaussians is given in a closed form as : $W_2^2 = ||m_X - m_Y||^2 + tr(\Sigma_X + \Sigma_Y - 2(\Sigma_X^{1/2} \Sigma_Y \Sigma_X^{1/2})^{1/2})$
- Consider the version $\mathcal{W}_2^2 = ||m_X - m_Y||^2 + ||\Sigma_X^{1/2} - \Sigma_Y^{1/2}||^2_{Frobenius}$ as in [4]
- The above distance is conditionally negative definite and $k(X,Y) = exp(-\frac{W_2^2}{2\theta^2})$ is therefore a valid kernel.

Gaussian Processes over clouds of points: Probability Product Kernels

Battachayra kernel, see [10]

- As in [10], we consider at first the following kernel between two distributions : $k(p, p') = \int_\Omega p(x)^\rho p'^\rho(x)\mathrm{d}x$
- It is symmetric semi-definite positive.
- For two clouds of points X and Y their empirical underlying laws, P_X and P_Y, consider $k(X, Y) = \int P_X^\rho(x) P_Y^\rho(x)\mathrm{d}x$
- For two Gaussians $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\mu', \Sigma')$, one gets:

$$(2\pi)^{(12)D/2}|\Sigma^+|^{1/2}|\Sigma|^{-\rho/2}|\Sigma|^{-\rho/2}\exp\big(-\frac{\rho}{2}\mu^T\Sigma^{-1}\mu - \frac{\rho}{2}\mu'^T\Sigma'^{-1}\mu' + \frac{1}{2}\mu^{+T}\Sigma^{+T}\mu^+\big)$$

  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 Battachayra Kernel and Expected Likelihood Kernel when $\rho = 1$

Gaussian Processes over clouds of points : Inner Product of Embeddings

- Consider 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 : $P_X \longmapsto \mu_X = \int P_X(x)k(x,.)\mathrm{d}x$.
- We define $k(X,Y) =< \mu_X, \mu_Y >$
- Depending on the empirical distribution and the kernel of the RKHS, we get different kernels between the clouds of points.
- For any kernel $k_{\mathcal{H}}$ of the RKHS, the uniform empirical laws give $k(X,Y) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k_{\mathcal{H}}(x_i, y_j)$
- With Gaussian and $k_{\mathcal{H}}(x,.) = exp(-\gamma||x-.||^2)$, we get a cosed form [14]:

$$K(X,Y) =< \mu_X, \mu_Y >= \exp\left(-\frac{1}{2}(\mu-\mu')^T(\Sigma+\Sigma'+\gamma^{-1}I)^{-1}(\mu-\mu')\right)/|\gamma\Sigma+\gamma\Sigma'+I|^{1/2}$$

,

Numerical Geometrical Properties of the Kernels: Rotation

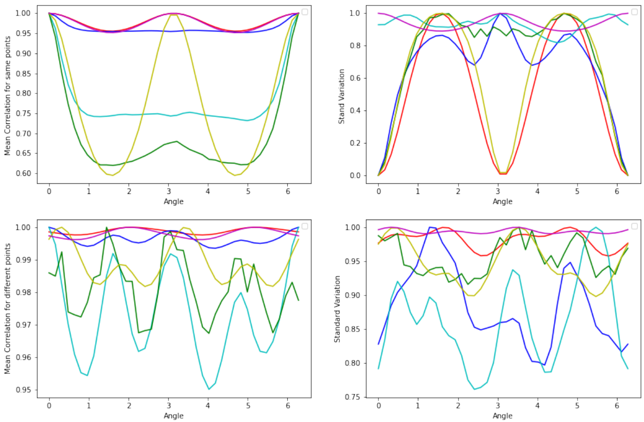


Figure: Normalized mean(100 points) correlations between clouds as measured by different kernels: Battachayra kernel(red color), sliced Wassertein (green color), Wassertein between Gaussians (yellow color), MMD substitution(blue), uniform embeddings (darkturquoise), and Gaussian embeddings (violet color). The same clouds are rotated (top right), two different clouds are rotated bottom left.

Numerical Geometrical Properties of Kernels: Translation

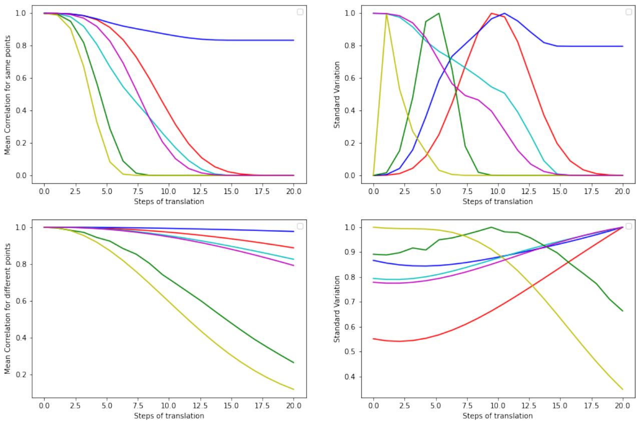


Figure: Normalized mean(100 points) correlations between clouds as measured by different kernels: Battachayra kernel(red color), sliced Wassertein (green color), Wassertein between Gaussians (yellow color), MMD substitution(blue), uniform embeddings (darkturquoise), and Gaussian embeddings (violet color). The same clouds are translated (top right), two different clouds are translated bottom left.

Numerical Geometrical Properties of Kernels: Dilatation

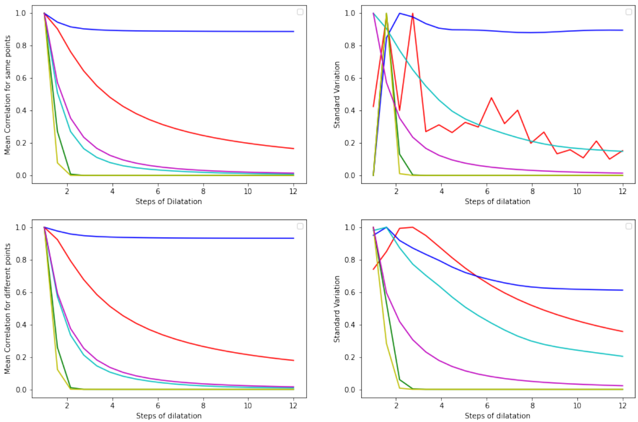


Figure: Normalized mean(100 points) correlations between clouds as measured by different kernels: Battachayra kernel(red color), sliced Wassertein (green color), Wassertein between Gaussians (yellow color), MMD substitution(blue), uniform embeddings (darkturquoise), and Gaussian embeddings (violet color). The same clouds are dilated (top right), two different clouds are dilated bottom left.

Numerical Geometrical Properties of Kernels: Size of points

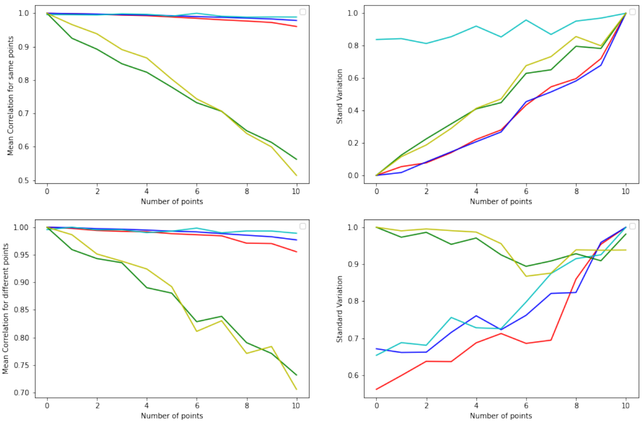


Figure: Normalized mean(100 points) correlations between clouds as measured by different kernels: Battachayra kernel(red color), sliced Wassertein (green color), Wassertein between Gaussians (yellow color), MMD substitution(blue), uniform embeddings (darkturquoise). The same clouds vary in size (top right), different clouds vary in size bottom left.

Perspectives

Perspectives

- Create a kernel with features of a cloud of points.
- Ability of Prediction of the different kernels
- Define optimization algorithms for the acquisition function.
- Think about the design of experiments.
- Enrich the base of the test functions.

Thanks For Your Attention !

Bibliography I

[1] Rosalind Archer et al. "Wind turbine interference in a wind farm layout optimization mixed integer linear programming model". In: *Wind Engineering* 35.2 (2011), pp. 165–175.

[2] Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic analysis on semigroups: theory of positive definite and related functions*. Vol. 100. Springer, 1984.

[3] Martin Bilbao and Enrique Alba. "Simulated annealing for optimization of wind farm annual profit". In: *2009 2nd International symposium on logistics and industrial informatics*. IEEE. 2009, pp. 1–5.

[4] Thi Thien Trang Bui et al. "Distribution regression model with a Reproducing Kernel Hilbert Space approach". In: *arXiv preprint arXiv:1806.10493* (2018).

Bibliography II

[5] Mathieu Carriere, Marco Cuturi, and Steve Oudot. "Sliced Wasserstein kernel for persistence diagrams". In: *International conference on machine learning*. PMLR. 2017, pp. 664–673.

[6] K Chen et al. "Wind turbine positioning optimization of wind farm using greedy algorithm". In: *Journal of Renewable and Sustainable Energy* 5.2 (2013), p. 023128.

[7] Souma Chowdhury et al. "Unrestricted wind farm layout optimization (UWFLO): Investigating key factors influencing the maximum power generation". In: *Renewable Energy* 38.1 (2012), pp. 16–30.

[8] Ju Feng and Wen Zhong Shen. "Solving the wind farm layout optimization problem using random search algorithm". In: *Renewable Energy* 78 (2015), pp. 182–192.

[9] Rémi Flamary et al. "Pot: Python optimal transport". In: *Journal of Machine Learning Research* 22.78 (2021), pp. 1–8.

Bibliography III

[10] Tony Jebara and Risi Kondor. "Bhattacharyya and expected likelihood kernels". In: *Learning theory and kernel machines*. Springer, 2003, pp. 57–71.

[11] Soheil Kolouri, Yang Zou, and Gustavo K Rohde. "Sliced Wasserstein kernels for probability distributions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 5258–5267.

[12] GPCDB Mosetti, Carlo Poloni, and Bruno Diviacco. "Optimization of wind turbine positioning in large windfarms by means of a genetic algorithm". In: *Journal of Wind Engineering and Industrial Aerodynamics* 51.1 (1994), pp. 105–116.

[13] Krikamol Muandet et al. "Kernel mean embedding of distributions: A review and beyond". In: *Foundations and Trends® in Machine Learning* 10.1-2 (2017), pp. 1–141.

[14] Krikamol Muandet et al. "Learning from distributions via support measure machines". In: *Advances in neural information processing systems* 25 (2012).

Bibliography IV

[15] Gabriel Peyré, Marco Cuturi, et al. "Computational optimal transport". In: *Center for Research in Economics and Statistics Working Papers* 2017-86 (2017).