Kernel approximations using determinantal point processes

Ayoub Belhadji
ENS de Lyon

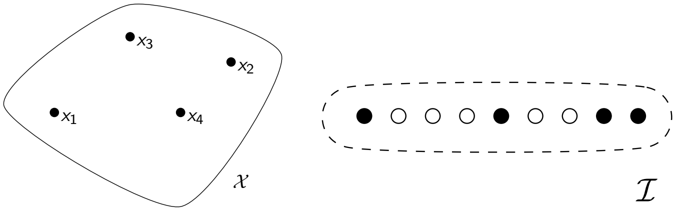
Joint work with Pierre Chainais and Rémi Bardenet

Centrale Lille, CRIStAL, Université de Lille, CNRS

JDS'22
14 juin 2022

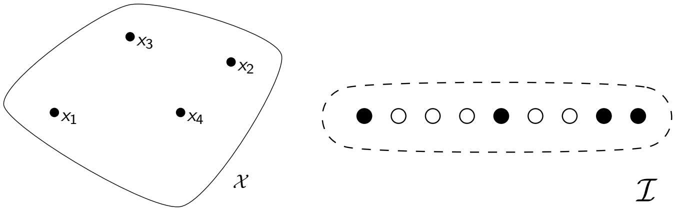
Outline

A **determinantal point process** (DPP) is a **distribution over subsets** of some set $\mathcal{X}, \mathcal{I}, \ldots$



Introduction

A **determinantal point process** (DPP) is a **distribution over subsets** of some set $\mathcal{X}, \mathcal{I}, \ldots$

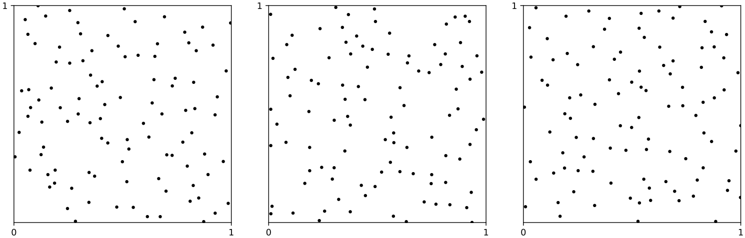


...with the **negative correlation** property:

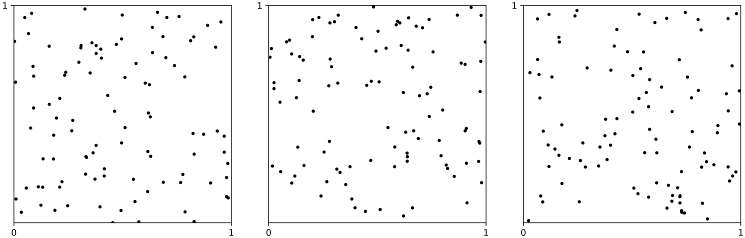
$$\forall B, B' \subset \mathcal{X}, \;\; B \cap B' = \emptyset \implies \mathbb{C}\mathrm{ov}(n_{\boldsymbol{x}}(B), n_{\boldsymbol{x}}(B')) \leq 0,$$

$$\text{where } n_{\boldsymbol{x}}(B) := |B \cap \boldsymbol{x}|$$

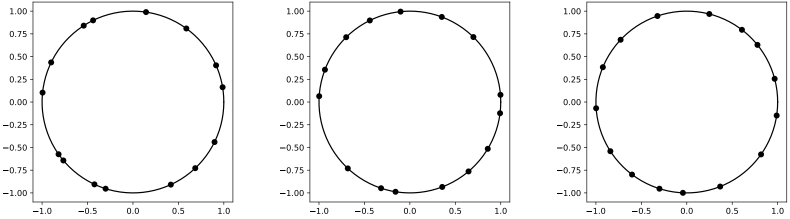
A DPP on $[0,1]^2$



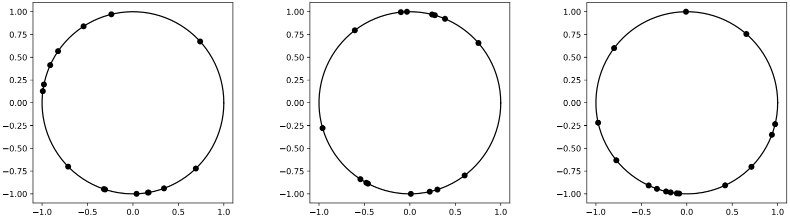
i.i.d. particles on $[0,1]^2$



A DPP on the unit circle



i.i.d. particles on the unit circle



Introduction

DPPs were used as **tools of modelisation**

- models for (fermions in particle physics) [Macchi (1975)]
- eigenvalues of random matrices [Weyl (1946), Dyson (1962), Ginibre (1965)]
- statistical models (spatial statistics) [Lavancier et al. (2012)]

DPPs were used as **tools of modelisation**

- models for (fermions in particle physics) [Macchi (1975)]
- eigenvalues of random matrices [Weyl (1946), Dyson (1962), Ginibre (1965)]
- statistical models (spatial statistics) [Lavancier et al. (2012)]

they were also used as **tools of simulation**

- subset selection (feature selection, subsampling of nodes in graphs...)[Belhadji et al. (2018), Tremblay et al. (2017) ...]
- numerical integration [Bardenet and Hardy (2016)]

Kernel approximations

We study the quality of the kernel approximations

$$\mu \approx \sum_{i=1}^{N} w_i k(x_i, .),$$

where the $\boldsymbol{x} = \{x_1, \dots, x_N\}$ follows the distribution of a DPP and

- k is a kernel over a domain $\mathcal{X}$
- $\mu \in \mathcal{F}$ the RKHS associated to the kernel k

Assumption: the Mercer decomposition of the kernel k

$$k(x, y) = \sum_{m \in \mathbb{N}^*} \sigma_m e_m(x) e_m(y)$$

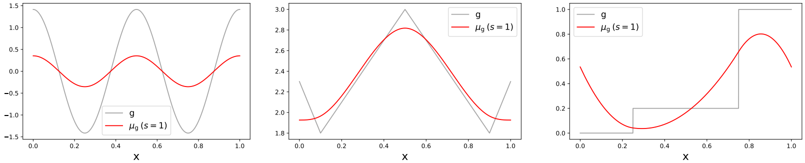
with $\sigma_1 \geq \sigma_2 \geq ... > 0$

Kernel approximations

Definition: an embedding of an element of $\mathbb{L}_2(\omega)$

Let $g \in \mathbb{L}_2(\omega)$, the *embedding* of g is defined by

$$\mu_g := \int_{\mathcal{X}} k(x,.)g(x)\mathrm{d}\omega(x).$$



In particular, we have

$$\forall f \in \mathcal{F}, \ \ \langle f, \mu_g \rangle_{\mathcal{F}} = \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x).$$

Kernel approximations for the study of quadrature rules

For a given $f \in \mathcal{F}$ and $g \in \mathbb{L}_2(\omega)$, we have

$$\begin{aligned}\Big| \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x) - \sum_{i \in [N]} w_i f(x_i) \Big| &= \big| \langle f, \mu_g - \sum_{i \in [N]} w_i k(x_i, .) \rangle_{\mathcal{F}} \big|, \\ &\leq \|f\|_{\mathcal{F}} \|\mu_g - \sum_{i \in [N]} w_i k(x_i, .)\|_{\mathcal{F}}.\end{aligned}$$

Kernel approximations for the study of quadrature rules

For a given $f \in \mathcal{F}$ and $g \in \mathbb{L}_2(\omega)$, we have

$$\Big| \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x) - \sum_{i \in [N]} w_i f(x_i) \Big| = \big| \langle f, \mu_g - \sum_{i \in [N]} w_i k(x_i, .) \rangle_{\mathcal{F}} \big|,$$

$$\leq \|f\|_{\mathcal{F}} \|\mu_g - \sum_{i \in [N]} w_i k(x_i, .)\|_{\mathcal{F}}.$$

Definition: the worst integration error on the unit ball

$$\Big\| \mu_g - \sum_{i \in [N]} w_i k(x_i, .) \Big\|_{\mathcal{F}} = \sup_{\|f\|_{\mathcal{F}} \leq 1} \Big| \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\omega(x) - \sum_{i \in [N]} w_i f(x_i) \Big|$$

The optimal kernel quadrature

$$x_1, \dots, x_N = \text{i.i.d.} \sim \omega \implies \mathbb{E}\|\mu_g - \sum_{i \in [N]} \frac{1}{N} k(x_i, .)\|_{\mathcal{F}}^2 = \mathcal{O}(1/N).$$

Can we improve on the rate $\mathcal{O}(1/N)$?

The optimal kernel quadrature

$$x_1, \dots, x_N = \text{i.i.d.} \sim \omega \implies \mathbb{E}\|\mu_g - \sum_{i \in [N]} \frac{1}{N} k(x_i, .)\|_{\mathcal{F}}^2 = \mathcal{O}(1/N).$$

Can we improve on the rate $\mathcal{O}(1/N)$?

Definition

Given a set of nodes $\boldsymbol{x} = \{x_1, \dots, x_N\}$ s.t. $\boldsymbol{K}(\boldsymbol{x})$ is non-singular, the **optimal kernel quadrature** is the couple $(\boldsymbol{x}, \hat{\boldsymbol{w}})$ such that

$$\left\| \mu_g - \sum_{i \in [N]} \hat{w}_i k(x_i, .) \right\|_{\mathcal{F}} = \min_{\boldsymbol{w} \in \mathbb{R}^N} \left\| \mu_g - \sum_{i \in [N]} w_i k(x_i, .) \right\|_{\mathcal{F}}$$

The optimal kernel quadrature

$$x_1, \dots, x_N = \text{i.i.d.} \sim \omega \implies \mathbb{E}\|\mu_g - \sum_{i \in [N]} \frac{1}{N} k(x_i, .)\|_{\mathcal{F}}^2 = \mathcal{O}(1/N).$$

Can we improve on the rate $\mathcal{O}(1/N)$?

Definition

Given a set of nodes $\boldsymbol{x} = \{x_1, \dots, x_N\}$ s.t. $\boldsymbol{K}(\boldsymbol{x})$ is non-singular, the **optimal kernel quadrature** is the couple $(\boldsymbol{x}, \hat{\boldsymbol{w}})$ such that

$$\left\| \mu_g - \sum_{i \in [N]} \hat{w}_i k(x_i, .) \right\|_{\mathcal{F}} = \min_{\boldsymbol{w} \in \mathbb{R}^N} \left\| \mu_g - \sum_{i \in [N]} w_i k(x_i, .) \right\|_{\mathcal{F}}$$

In particular

$$\left\| \mu_g - \sum_{i \in [N]} \hat{w}_i k(x_i, .) \right\|_{\mathcal{F}} = \left\| \mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})} \mu_g \right\|_{\mathcal{F}},$$

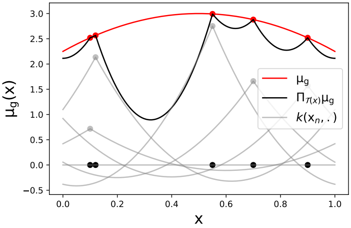
$\mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}$: the orthogonal projection onto $\mathcal{T}(\boldsymbol{x}) = \mathrm{Span}(k(x_i, .))_{i \in [N]}$.

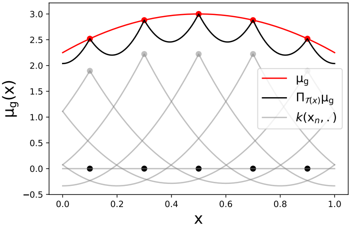
The optimal kernel quadrature

Kernel interpolation

The optimal mixture $\hat{\mu}_g := \sum_{i\in[N]} \hat{w}_i k(x_i, .)$ satisfies

$$\forall i \in [N],\ \hat{\mu}_g(x_i) = \mu_g(x_i).$$





How to choose the nodes in general?

The determinantal distributions

Definition

Let κ be a kernel s.t. $\int_{\mathcal{X}} \kappa(x,x)\mathrm{d}\omega(x) < +\infty$. The function

$$f_\kappa(x_1, \ldots, x_N) \propto \operatorname{Det} \boldsymbol{\kappa}(\boldsymbol{x})$$

is a p.d.f. on $\mathcal{X}^N$.

The determinantal distributions

Definition

Let κ be a kernel s.t. $\int_{\mathcal{X}} \kappa(x,x)\mathrm{d}\omega(x) < +\infty$. The function

$$f_{\kappa}(x_1, \dots, x_N) \propto \operatorname{Det} \boldsymbol{\kappa}(\boldsymbol{x})$$

is a p.d.f. on $\mathcal{X}^N$.

We study two cases:

- Projection DPP:

$$\kappa(x,y) := \mathfrak{K}(x,y) = \sum_{n \in [N]} e_n(x) e_n(y)$$

- Continuous volume sampling (CVS):

$$\kappa(x,y) = k(x,y) = \sum_{m \in \mathbb{N}^*} \sigma_m e_m(x) e_m(y)$$

Main results

Theorem (Belhadji et al. (2019); Belhadji (2021))

Under the distribution of the Projection DPP, we have

$$\forall g \in \mathbb{L}_2(\omega), \ \mathbb{E}_{\mathrm{DPP}} \left\| \mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})} \mu_g \right\|_{\mathcal{F}}^2 = \mathcal{O}(r_{N+1}),$$

where $r_{N+1} := \sum_{m \geq N+1} \sigma_m$.

Main results

Theorem (Belhadji et al. (2019); Belhadji (2021))

Under the distribution of the Projection DPP, we have

$$\forall g \in \mathbb{L}_2(\omega),\ \ \mathbb{E}_{\mathrm{DPP}} \left\|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\right\|_{\mathcal{F}}^2 = \mathcal{O}(r_{N+1}),$$

where $r_{N+1} := \sum_{m \geq N+1} \sigma_m$.

Theorem (Belhadji et al. (2020))

Under the distribution of CVS, we have

$$\forall g \in \mathbb{L}_2(\omega),\ \ \mathbb{E}_{\mathrm{DPP}} \left\|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\right\|_{\mathcal{F}}^2 = \mathcal{O}(\sigma_{N+1}).$$

Theorem: a lower bound (Pinkus (1985); Belhadji et al. (2020))

For any configuration $\boldsymbol{x} \in \mathcal{X}^N$ such that $\dim \mathcal{T}(\boldsymbol{x}) = N$,

$$\sup_{\|g\|_\omega \leq 1} \|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 \geq \sigma_{N+1}$$

$\mathcal{X}$	$\mathcal{F}$ or k	σ_{N+1}	(e_m)
$[0,1]$	Sobolev	$\mathcal{O}(N^{-2s})$	Fourier
$[0,1]^d$	Korobov	$\mathcal{O}(\log(N)^{2s(d-1)}N^{-2s})$	$\otimes$ of Fourier
$[0,1]^d$	Sobolev	$\mathcal{O}(N^{-2s/d})$	"Fourier"
$\mathbb{S}^d$	Dot product	"-"	Spherical Harmonics
$\mathbb{R}$	Gaussian	$\mathcal{O}(e^{-\alpha N})$	Hermite Polys.
$\mathbb{R}^d$	Gaussian	$\mathcal{O}(e^{-\alpha d N^{1/d}})$	$\otimes$ of Hermite Polys.
...	...	...	...

Numerical simulations: DPP vs uniform grid

$\mathcal{F}$ = Korobov space of order $s = 1$, $\mathcal{X} = [0,1]^2$
We report $\epsilon_m(N) = \mathbb{E}_\kappa \|\mu_{e_m} - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_{e_m}\|_{\mathcal{F}}^2$
under DPP ($\kappa = \mathfrak{K}$).

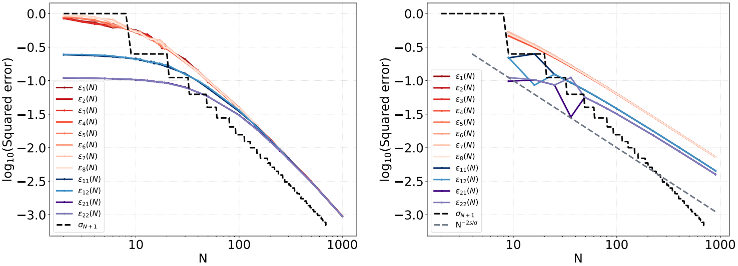


Figure: OKQ using DPPs (left) vs OKQ using the uniform grid (right)

Numerical simulations: the Gaussian space

$\mathcal{F}$ = the RKHS associated to the Gaussian kernel
We report $\epsilon_m(N) = \mathbb{E}_\kappa \|\mu_{e_m} - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_{e_m}\|^2_{\mathcal{F}}$ for $m \in \{1, 15\}$
under DPP ($\kappa = \mathfrak{K}$).

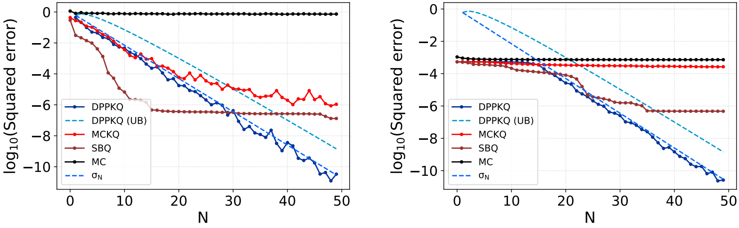
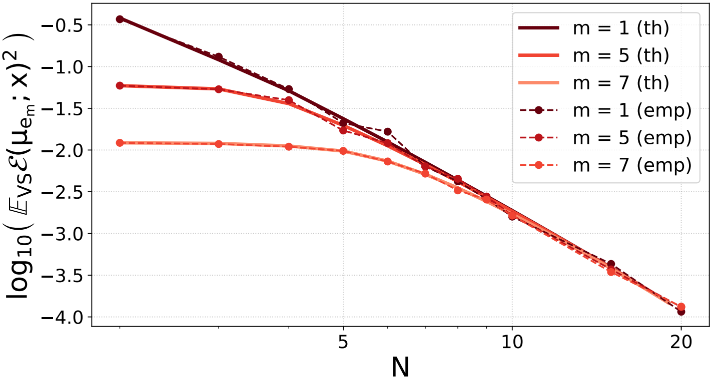


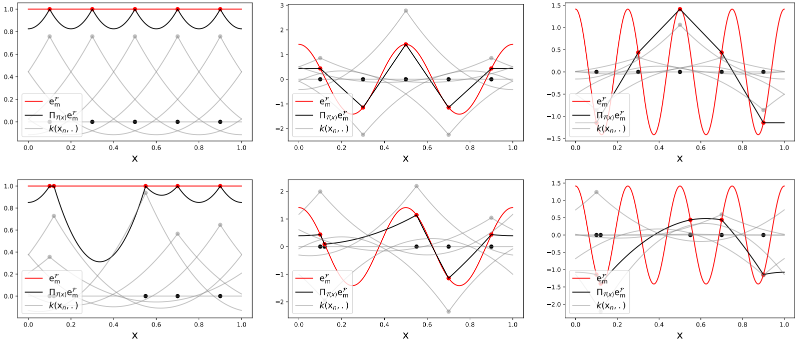
Figure: The squared interpolation error for e_1 (Left), vs e_{15} (Right).

$\mathcal{F} =$ the periodic Sobolev space of order $s = 2$, $\mathcal{X} = [0, 1]$
We report $\epsilon_m(N) = \mathbb{E}_\kappa \|\mu_{e_m} - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_{e_m}\|^2_{\mathcal{F}}$ for $m \in \{1, 5, 7\}$
under CVS ($\kappa = k$).



An intuition

The reconstruction on the RKHS is governed by the reconstruction of the eigenfunctions e_m



A summary

Take Home Messages

- A general **theoretical analysis of kernel interpolation** for nodes sampled according to **the projection DPP or CVS**
- New **geometric intuitions** behind the use of **DPPs** for this task
- Empirical validation on different RKHSs: **the optimal rate of convergence $\mathcal{O}(\sigma_{N+1})$ is achieved**

Distribution	Theoretical rate	Empirical rate
P. DPP	$N^2\mathcal{O}(r_{N+1})$ [Belhadji et al. (2019)] $\mathcal{O}(r_{N+1})$ [Belhadji (2021)]	$\mathcal{O}(\sigma_{N+1})$ [Belhadji et al. (2019)]
CVS	$\mathcal{O}(\sigma_{N+1})$ [Belhadji et al. (2020)]	$\mathcal{O}(\sigma_{N+1})$ [Belhadji et al. (2020)]

Thank you for your attention!