An introduction to Determinantal Point Processes

Pierre-Olivier Amblard

with many interactions with (and inputs from) Simon Barthelmé
and Nicolas Tremblay
and collab. with Jeff Coeurjolly & Adrien Mazoyer

CNRS, GIPSA-lab, U. Grenoble-Alpes, France

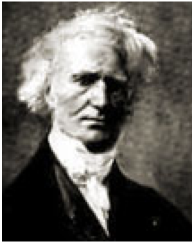
W'p "Kernel and sampling methods for design and quantization",
Gdr MascoNum 2021

A tribute to . . .

O. Macchi

$$(4.33)\quad h_k(t_1,\cdots,t_k) = \begin{vmatrix} C(t_1,t_1) & C(t_1,t_2) & \cdots & C(t_1,t_k) \\ C(t_2,t_1) & C(t_2,t_2) & \cdots & C(t_2,t_k) \\ \vdots & & & \\ C(t_k,t_1) & C(t_k,t_2) & \cdots & C(t_k,t_k) \end{vmatrix}$$

L.-A. Cauchy and J.-Ph.-M. Binet

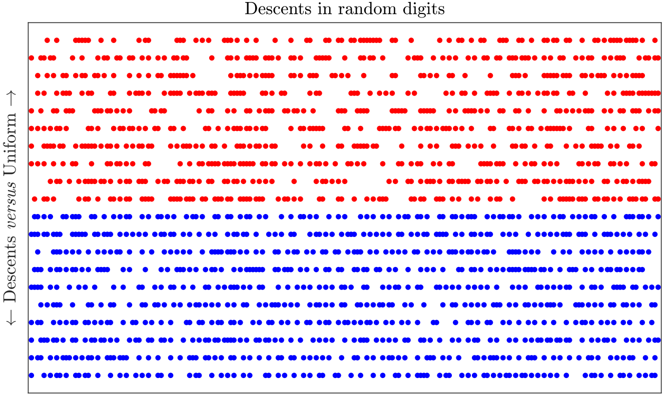


$$\text{Det }(\boldsymbol{AB})_{\mathcal{X}\mathcal{X}'} = \sum_{\mathcal{Y} \backslash |\mathcal{Y}| = |\mathcal{X}|} \text{Det } \boldsymbol{A}_{\mathcal{X}\mathcal{Y}} \text{Det } \boldsymbol{B}_{\mathcal{Y}\mathcal{X}'}$$

Determinantal Point Processes are ubiquitous...

Descents in series of i.i.d. digits : t s.t. $X_t > X_{t+1}$

4 0 6 6 7 8 0 9 1 7 7 5 2 9 5

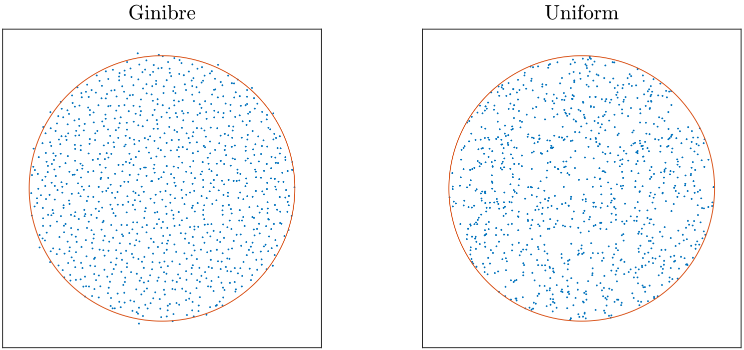


$P(i \in D) = k(0) = 0.45; \quad P((i,j) \in D) = k(0)^2 - k(i-j)k(j-i)$

where $\sum_{m \in \mathbb{Z}} k(m)t^m = (1-(1-t)^{10})^{-1}$ Borodin, Diaconis, Fulman 2010

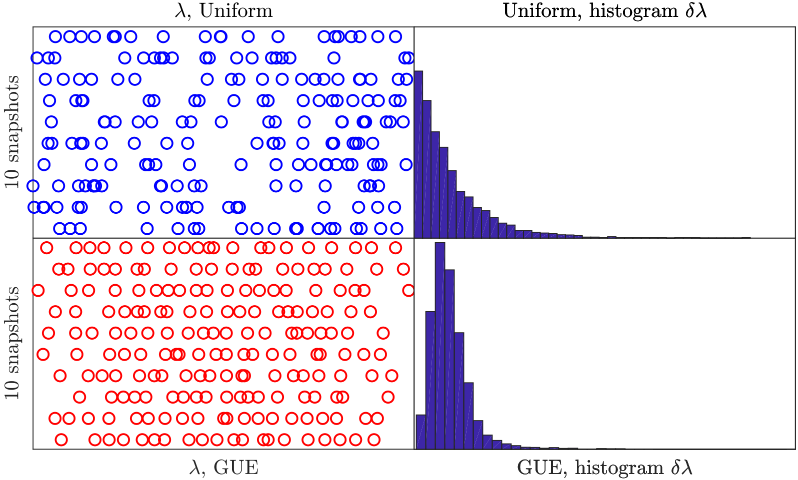
$\mathrm{Cov}(D_i, D_{i+1}) = -0.0825$

Eigenvalues of complex matrices with i.i.d. entries



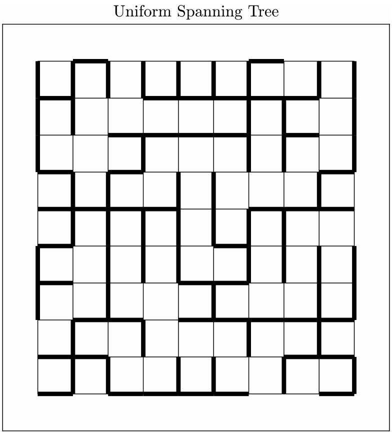
$$P(\boldsymbol{\lambda}) \propto \exp\left(-\boldsymbol{\lambda}^\dagger \boldsymbol{\lambda}\right) \underbrace{\prod_{i<j} |\lambda_i - \lambda_j|^2}_{\text{Vandermonde determinant}^2}$$

Eigenvalues of Hermitian matrices. Gaus. Unit. Ens.



$$P(\boldsymbol{\lambda}) \propto \exp\left(-\frac{\|\boldsymbol{\lambda}\|^2}{2}\right) \underbrace{\prod_{i<j}(\lambda_i - \lambda_j)^2}_{\text{Vandermonde determinant}^2}$$

Uniform Spanning Trees



Let $\mathcal{S}$ a subset of edges

$\mathcal{S}$ forms a UST

$$\Longleftrightarrow$$

$$P(\mathcal{S}) = \text{Det}\,[\boldsymbol{B}^\top \boldsymbol{L}^\dagger \boldsymbol{B}]_{\mathcal{S}\mathcal{S}} \delta_{(|\mathcal{S}|=n-1)}$$

$\boldsymbol{B}$ is the incidence matrix ;
$\boldsymbol{L}$ the Laplacian

What's interesting in DPPs

As illustrated : DPPs are repulsive !

↪ useful for diversity in random sampling / for space filling properties in DoE

Not illustrated : DPPs are theoretically tractable negatively correlated PP

↪ likelihoods analytically known, including Zs

↪ correlation functions (inclusion prob.) known at any order

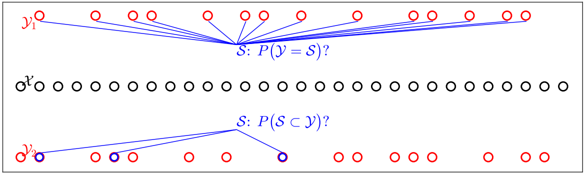
↪ have exact simulation techniques

They are to negatively correlated PP what Poisson are for independent PP, or what Gaussian are to SP

Today :

- Elements on Determinantal Point Processes (discrete case)
- Sampling DPPs
- Some applications (especially as coresets)

Point Process on a discrete space $\mathcal{X} \longleftrightarrow \{1, \ldots, |\mathcal{X}|\}$



PP on $\mathcal{X}$ = probability over $2^{\mathcal{X}}$, set of subsets of $\mathcal{X}$

Let $\mathcal{Y}$ be this process :

1. likelihoods (or probabilities of the sets)

$$\forall \mathcal{S} \subset \mathcal{X}, \Pr\left(\mathcal{Y} = \mathcal{S}\right)$$

2. inclusion probabilities (marginals)

$$\forall \mathcal{S} \subset \mathcal{X}, \Pr\left(\mathcal{S} \subset \mathcal{Y}\right) = \sum_{\mathcal{S}' \supset \mathcal{S}} \Pr\left(\mathcal{Y} = \mathcal{S}'\right)$$

Alternately : Stochastic process ε_t indexed by $\mathcal{X}$ with values in $\{0, 1\}$

Completely determined by the knowledge of $\Pr\left(\varepsilon_{i_1}, \ldots, \varepsilon_{i_n}\right), \forall \boldsymbol{i}, \forall n$

Poisson PP on a discrete space $\mathcal{X}$

Two disjoint subsets belongs to the process independently :

$$\Pr\left(\mathcal{A} \cup \mathcal{B} \subset \mathcal{Y}\right) = \Pr\left(\mathcal{A} \subset \mathcal{Y}\right) \Pr\left(\mathcal{B} \subset \mathcal{Y}\right)$$

In the alternative description, second-order inclusion

$$\Pr\left(\varepsilon_i = 1, \varepsilon_j = 1\right) = \Pr\left(\varepsilon_i = 1\right) \Pr\left(\varepsilon_j = 1\right) \qquad \forall i \neq j$$

Also called Bernoulli process since ε_t is i.i.d. $Ber(P(\varepsilon_t = 1))$

Determinantal PP on a discrete $\mathcal{X}$

Let $\boldsymbol{I} \geq \boldsymbol{K} \geq 0$ be a $|\mathcal{X}| \times |\mathcal{X}|$ matrix (often symmetric)

A DPP is a PP which samples $\mathcal{Y}$ out of $\mathcal{X}$ such that

$$\forall \mathcal{S} \subset \mathcal{X}, \mathrm{Pr}\left(\mathcal{S} \subset \mathcal{Y}\right) = \mathrm{Det}\ \boldsymbol{K}_{\mathcal{S}\mathcal{S}}$$

First and second order inclusion :

$$\begin{aligned} \mathrm{Pr}\left(\{i\} \subset \mathcal{Y}\right) &= \boldsymbol{K}_{ii} \\ \mathrm{Pr}\left(\{i,j\} \subset \mathcal{Y}\right) = \boldsymbol{K}_{ii}\boldsymbol{K}_{jj} - \boldsymbol{K}_{ij}^2 &< \boldsymbol{K}_{ii}\boldsymbol{K}_{jj} = \mathrm{Pr}\left(\{i\} \subset \mathcal{Y}\right)\mathrm{Pr}\left(\{j\} \subset \mathcal{Y}\right) \end{aligned}$$

Equivalent description : ε_i indexed by $1, \ldots, |\mathcal{X}|$, defined by

$$\mathrm{Pr}\left(\varepsilon_i = 1\right) = \boldsymbol{K}_{ii}; \mathrm{Pr}\left(\varepsilon_i = 1, \varepsilon_j = 1\right) = \boldsymbol{K}_{ii}\boldsymbol{K}_{jj} - \boldsymbol{K}_{ij}^2; \ldots$$

Likelihoods or L-ensemble

Different approach : Let $\boldsymbol{L} \geq 0$ be a $|\mathcal{X}| \times |\mathcal{X}|$ matrix (often symm.)

A L-ensemble is a PP which samples $\mathcal{Y}$ out of $\mathcal{X}$ such that

$$\begin{aligned} \forall \mathcal{S} \subset \mathcal{X}, \Pr(\mathcal{Y} = \mathcal{S}) &= \frac{\text{Det } \boldsymbol{L}_{\mathcal{S}\mathcal{S}}}{\sum_{\mathcal{S} \subset \mathcal{X}} \text{Det}(\boldsymbol{L}_{\mathcal{S}\mathcal{S}})} \\ &= \frac{\text{Det } \boldsymbol{L}_{\mathcal{S}\mathcal{S}}}{\text{Det}(\boldsymbol{L} + \boldsymbol{I})} \end{aligned}$$

DPP $\Longleftrightarrow$ L-ensemble ? Recall $\Pr(\mathcal{S} \subset \mathcal{Y}) = \sum_{\mathcal{S}' \supset \mathcal{S}} \Pr(\mathcal{Y} = \mathcal{S}')$

L-ensemble are DPP : $\boldsymbol{K} = \boldsymbol{L}(\boldsymbol{L} + \boldsymbol{I})^{-1}$

DPP are L-ensemble only if $\boldsymbol{K} < \boldsymbol{I}$: $\boldsymbol{L} = \boldsymbol{K}(\boldsymbol{K} - \boldsymbol{I})^{-1}$

$\hookrightarrow$ DPP with $\boldsymbol{K}$ a projection are not L-ensemble (*cf* UST)

(NB : Extended L-ensemble reduces the symmetry breaking ! see our arxiv sub's)

Repulsion, or negative correlation

Looking at $P(A \subset \mathcal{Y} | B \subset \mathcal{Y})$ for A, B disjoint :

- Poisson : $P(A \subset \mathcal{Y} | B \subset \mathcal{Y}) = \dfrac{P(A \subset \mathcal{Y},\ B \subset \mathcal{Y})}{P(B \subset \mathcal{Y})} = P(A \subset \mathcal{Y})$

  Thus, $\text{Cov}[\varepsilon_i, \varepsilon_j] = 0$

- DPP : $P(A \subset \mathcal{Y} | B \subset \mathcal{Y}) = \dfrac{P(A \subset \mathcal{Y},\ B \subset \mathcal{Y})}{P(B \subset \mathcal{Y})} < P(A \subset \mathcal{Y})$

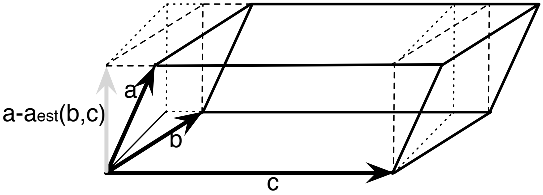
  Thus, $\text{Cov}[\varepsilon_i, \varepsilon_j] < 0$

"Strengh" of repulsion depends on $\boldsymbol{K}(or \boldsymbol{L})$: usually constructed from a kernel $k(x, y)$

Allow to quantify similarities between elements . . .

An Interpretation

Elementary geometry : $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n) \Rightarrow \text{vol}\,(P(\boldsymbol{b}_i))^2 = \text{Det}\,\boldsymbol{B}^\top \boldsymbol{B}$



Let $\boldsymbol{B} = (\boldsymbol{a}\,\boldsymbol{b}\,\boldsymbol{c}) = (\boldsymbol{a}\,\boldsymbol{B}')$ then

$$\begin{aligned}\text{vol}\,(P(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}))^2 &= \|\boldsymbol{a} - \widehat{\boldsymbol{a}}(\boldsymbol{b}, \boldsymbol{c})\|^2\, \text{vol}\,(P(\boldsymbol{b}, \boldsymbol{c}))^2 \\ \text{Det}\,\boldsymbol{B}^\top \boldsymbol{B} &= (\boldsymbol{a}^\top \boldsymbol{a} - \boldsymbol{a}^\top \boldsymbol{B}'(\boldsymbol{B}'^\top \boldsymbol{B}')^{-1} \boldsymbol{B}'^\top \boldsymbol{a})\text{Det}\,\boldsymbol{B}'^\top \boldsymbol{B}'\end{aligned}$$

For $\boldsymbol{K} = \boldsymbol{B}^\top \boldsymbol{B}$: matrix of similarities between features $\boldsymbol{b}_i$

Subsampling features by a DPP = seeking features that create a big volume ! and are thus most diverse !

k DPP, just a flavor

For a DPP, $|\mathcal{Y}|$ is random. How to get exactly k subsamples ?

1. Use a rank k projection matrix, see that soon, or
2. condition a DPP to $|\mathcal{Y}| = k$.

$$\Pr\left(\mathcal{Y} = \mathcal{S} \middle| |\mathcal{Y}| = k\right) = \frac{\text{Det } \boldsymbol{L}_{\mathcal{S}\mathcal{S}}}{\sum_{\mathcal{S}/|\mathcal{S}|=k} \text{Det } (\boldsymbol{L}_{\mathcal{S}\mathcal{S}})} \mathbf{1}(|\mathcal{S}| = k)$$

Partition function is an elementary symmetric polynomial :

$$\sum_{\mathcal{S}/|\mathcal{S}|=k} \text{Det } (\boldsymbol{L}_{\mathcal{S}}) = e_k(\boldsymbol{\lambda}) = \sum_{\mathcal{J}\subset\{1,\ldots,N\}/|\mathcal{J}|=k} \prod_{j\in\mathcal{J}} \lambda_j$$

- More complicated analytically
- Sampling, same strategy (see later) but a little bit more difficult (because of e.s.p.)
  We have designed efficient approximations to $e_k(\boldsymbol{\lambda})$...

Usefull properties–1

Let $\mathcal{Y}$ be a DPP with kernel $\boldsymbol{K}$ (or a L-ensemble $\boldsymbol{L}$).

- The size $|\mathcal{Y}|$ of the sample is random and :

  $$E[|\mathcal{Y}|] = \text{Tr}[\boldsymbol{K}] \text{ and } \text{Var}[\ |\mathcal{Y}|] = \text{Tr}[\boldsymbol{K}(\boldsymbol{I} - \boldsymbol{K})]$$

  $E[|\mathcal{Y}|] = E \sum \varepsilon_i = \sum K_{ii}$ et $\text{Var}[|\mathcal{Y}|] = \sum \text{Cov}[\varepsilon_i, \varepsilon_j] \ldots$

- If $\boldsymbol{K}$ is a projection operator on a $k < |\mathcal{X}|$ dimensional subspace, then $|\mathcal{Y}| = k$ a.s.

  $\text{Tr}\boldsymbol{K} = \sum \lambda_i$ et $\text{Var}[|\mathcal{Y}|] = \text{Tr}\boldsymbol{K}(\boldsymbol{I} - \boldsymbol{K}) = \sum_i \lambda_i(1 - \lambda_i)$

  Such a DPP is called a projection DPP

Usefull properties–2

Let $\mathcal{Y}$ be a DPP with kernel $\boldsymbol{K}$ (or $\boldsymbol{L}$).

- A DPP is a mixture of projection DPPs

Recall Cauchy-Binet formula $\text{Det}\,(\boldsymbol{AB})_{\mathcal{X}\mathcal{X}'} = \sum_{\mathcal{Y}\ \backslash |\mathcal{Y}\ |=|\mathcal{X}|} \text{Det}\,\boldsymbol{A}_{\mathcal{X}\mathcal{Y}} \text{Det}\,\boldsymbol{B}_{\mathcal{Y}\mathcal{X}'}$

If $\boldsymbol{L} = \sum_n \lambda_n \boldsymbol{v}_n \boldsymbol{v}_n^\top = \boldsymbol{V\Lambda V}^T$ then

$$\begin{aligned}
\text{Det}\,\boldsymbol{L}_{\mathcal{S}} &= \sum_{\mathcal{Y}\ \backslash |\mathcal{Y}\ |=|\mathcal{S}|} \text{Det}\,\boldsymbol{V}_{\mathcal{S}\mathcal{Y}} \text{Det}\,(\boldsymbol{\Lambda V}^\top)_{\mathcal{Y}\mathcal{S}} \\
&= \sum_{\mathcal{Y}\ \backslash |\mathcal{Y}\ |=|\mathcal{S}|} \text{Det}\,\boldsymbol{V}_{\mathcal{S}\mathcal{Y}} \text{Det}\,\boldsymbol{V}^\top_{\mathcal{S}\mathcal{Y}} \text{Det}\,\boldsymbol{\Lambda}_{\mathcal{Y}} \\
&= \sum_{\mathcal{Y}\ \backslash |\mathcal{Y}\ |=|\mathcal{S}|} \text{Det}\,\Big(\sum_{n\in\mathcal{Y}} \boldsymbol{v}_n \boldsymbol{v}_n^\top\Big)_{\mathcal{S}} \prod_{n\in\mathcal{Y}} \lambda_n \\
&= \sum_{\mathcal{Y}\ \backslash |\mathcal{Y}\ |=|\mathcal{S}|} \text{Det}\,\boldsymbol{K}^{V_{\mathcal{Y}}}_{\mathcal{S}} \prod_{n\in\mathcal{Y}} \lambda_n
\end{aligned}$$

Sampling

If $\boldsymbol{L} = \sum_n \lambda_n \boldsymbol{v}_n \boldsymbol{v}_n^\top = \boldsymbol{V\Lambda V}^T$ and $\boldsymbol{K}_{\mathcal{S}}^{V_{\mathcal{Y}}} = \left(\sum_{n \in \mathcal{Y}} \boldsymbol{v}_n \boldsymbol{v}_n^\top\right)_{\mathcal{S}}$ then

$$\Pr\left(\mathcal{S}\right) = \frac{\text{Det } \boldsymbol{L}_S}{\text{Det } (\boldsymbol{L} + \boldsymbol{I})} = \sum_{\mathcal{Y} \setminus |\mathcal{Y}| = |\mathcal{S}|} \underbrace{\text{Det } \boldsymbol{K}_{\mathcal{Y}}^{V_{\mathcal{Y}}}}_{step2} \underbrace{\prod_{n \in \mathcal{Y}} \frac{\lambda_n}{1 + \lambda_n}}_{step1}$$

1. Keep $\boldsymbol{v}_n$ if $Ber\left(\lambda_n/(1 + \lambda_n)\right) = 1$
2. Construct a projection kernel with the eigenvectors kept.

   Sample a projection DPP.

Sampling a projection DPP, idea

A fact : if $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{X}')$, then

$$\begin{aligned}\text{Det } \boldsymbol{X}^T\boldsymbol{X} &= (\boldsymbol{x}_1^\top \boldsymbol{x}_1 - \boldsymbol{x}_1^\top \boldsymbol{X}'(\boldsymbol{X}'^\top \boldsymbol{X}')^{-1}\boldsymbol{X}'^\top \boldsymbol{x}_1)\text{Det } \boldsymbol{X}'^\top \boldsymbol{X}' \\ &= \left\|\boldsymbol{x}_1 - \text{Proj}_\perp(x_1 \middle| \text{span } (\boldsymbol{X}')\right\|^2 \text{Det } \boldsymbol{X}'^\top \boldsymbol{X}'\end{aligned}$$

If $\boldsymbol{K} = \boldsymbol{V}\boldsymbol{V}^\top$ is a projection kernel, then

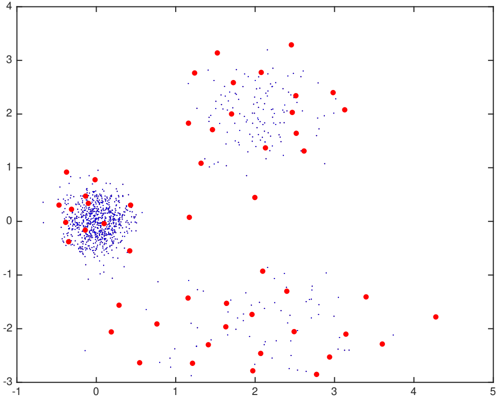
$$\text{Det } \boldsymbol{K} = \left\|\boldsymbol{K}_{1,.} - \text{Proj}_\perp(\boldsymbol{K}_{1,.} \middle| \text{span } (\boldsymbol{K}_{2,.}, \ldots, \boldsymbol{K}_{n,.})\right\|^2 \text{Det } \boldsymbol{K}'$$

- If n vectors have already been sampled, choose the $(n+1)$-th with a probability prop. to the MSE of its prediction from the first n!

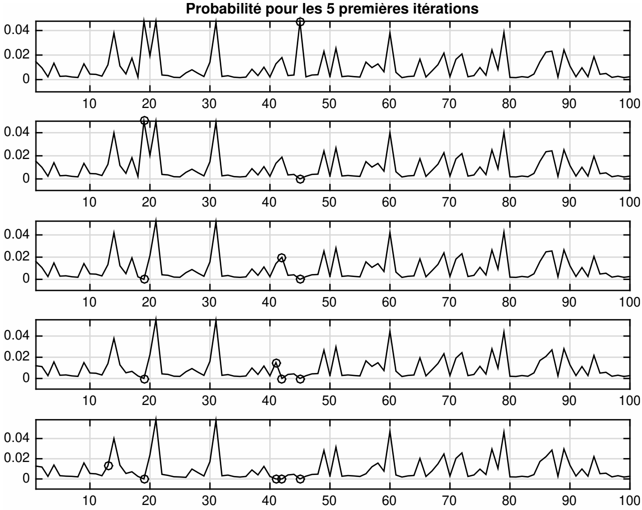
$\hookrightarrow$ Implemented in the physical space, complexity $O(Nk^3)$ for a rank k kernel in $\mathbb{R}^N$,

$\hookrightarrow$ Implemented in the associated RKHS, $O(Nk^2)$.

Illustration



Illustration



If N is large ?

Exact sampling requires spectral components of $\boldsymbol{L}$ or $\boldsymbol{K}$: $O(|\mathcal{X}|^3)$!

Ideas : be patient, or

- Taylored algorithms for special DPPs (*e.g.* Wilson's for USTs)
- Approximate sampling, *e.g.* Gibbs
- Approximate kernel :

1. If $\boldsymbol{L} = \boldsymbol{B}^\top \boldsymbol{B}$, $\boldsymbol{B} : D \times N$, eigen elements of $\boldsymbol{L}$ obtained from $\boldsymbol{B}\boldsymbol{B}^\top$.
2. If $k(x, y) = \varphi(x - y)$, then necessarily $\exists P / \varphi \propto TF^{-1}(P)$ Bochner

Random Fourier Features :

$$\widehat{\varphi(x)} = \sum_{k=1}^{K} \exp(2i\pi\nu_k x) \text{ where } \nu_k \sim P \overset{K\to+\infty}{\longrightarrow} \varphi(x) = \int e^{2i\pi\nu x} dP(\nu)$$

If $\boldsymbol{\varphi} = \left(\exp(2i\pi\boldsymbol{\nu}x_1), \ldots, \exp(2i\pi\boldsymbol{\nu}x_N) \right)$ of dim. $K \times N$ then
$\boldsymbol{\varphi}^\dagger \boldsymbol{\varphi} = \varphi\widehat{(x - y)} \overset{K\to+\infty}{\longrightarrow} \boldsymbol{L}$ and use the preceding trick

Applications of DPPs

Everywhere subsampling is interesting . . .

- Subsampling Graphs, patches in images for reconstruction (Launay *et. al.*)
- DoE (Fanuel, see Rémi later on). For example :

  Let S be a projection dpp with kernel $XX^\top, X \in \mathbb{R}^{n\times d}$. Then

  $$E[X^\dagger_{(S,:)}] = X^\dagger$$

  Consequence : $E[X_{S,:}^{-1} y_S] = X^\dagger y = \arg\min \| Xw - y\|^2$.

- Monte-Carlo integration (see Jean-François) / statistical estimation
- Coresets/Sketching

Application in estimation

Estimate statistics $\mathrm{E}[h(x)]$ where $h : \mathbb{R}^d \longrightarrow \mathbb{R}^p$

Let C_N be the empirical mean, and consider

$$C_\pi = \frac{1}{N} \sum_{i=1}^{N} \frac{h(x_i)\varepsilon_i}{\pi_i}$$

ε is a doubly stochastic PP, *i.e.* $\Pr\left(\varepsilon_i = 1 \middle| x\right) = \pi_i(x_i)$

<u>No Bias :</u> $\mathrm{E}\left[C_\pi\right] = \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}\left[\frac{h(x_i)}{\pi_i} \mathrm{E}\left[\varepsilon_i \middle| x_i\right]\right] = \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}\left[\frac{h(x_i)}{\pi_i} \pi_i\right] = \mathrm{E}[h(x)]$

<u>Variance :</u> $\mathrm{Var}[C_\pi] = \frac{1}{N^2} \sum_{i,j} \mathrm{Cov}\left[\frac{h(x_i)\varepsilon_i}{\pi_i}, \frac{h(x_j)\varepsilon_j}{\pi_j}\right]$

$$= \mathrm{Var}[C_N] + \frac{1}{N^2} \sum_{i,j} \mathrm{E}\left[h(x_i)h(x_j)\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\right]$$

Variance, possible gain from negative correlations

If $\pi_{ij} = \pi_i \pi_j$ Poisson process

$$\mathrm{Var}[C_P] = \mathrm{Var}[C_N] + \frac{1}{N^2} \sum_i \mathrm{E}\left[h(x_i)^2 \frac{1 - \pi_i}{\pi_i} \right] \tag{1}$$

Then, for C_π

$$\mathrm{Var}[C_\pi] = \mathrm{Var}[C_P] + \frac{1}{N^2} \sum_{i \neq j} \mathrm{E}\left[h(x_i) h(x_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right] \tag{2}$$

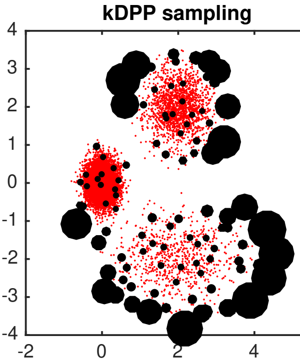
h takes value in $\mathbb{R}^+$ (resp. $\mathbb{R}^-$), a sampling process with negative correlation (resp. positive) is better than the Poisson subsampling

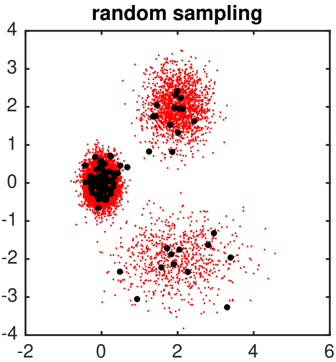
Example of the correlation matrix

Let $\boldsymbol{x}_i$ a series of 10000 i.i.d. r.v.

Subsamples 100 points, $\boldsymbol{L}_{ij} = \exp(-\| \boldsymbol{x}_i - \boldsymbol{x}_j \|^2 / 2\sigma^2)$

Mixture of 3 Gaussian : probabilities [3/4 1/6 1/12]



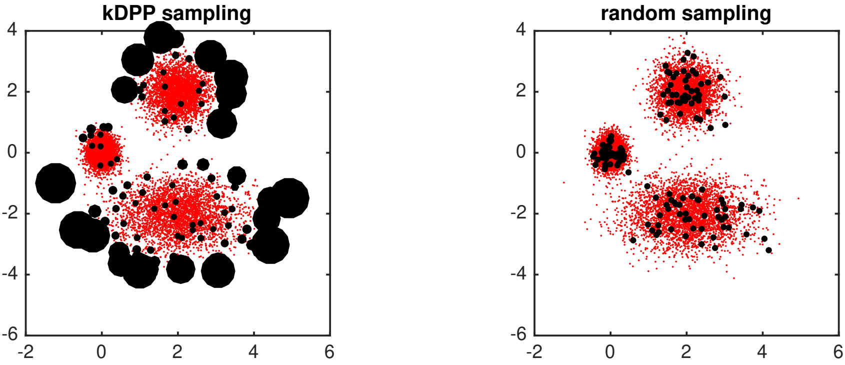


Example of the correlation matrix

Let $\boldsymbol{x}_i$ a series of 10000 i.i.d. r.v.

Subsamples 100 points, $\boldsymbol{L}_{ij} = \exp(-\| \boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2\sigma^2)$

Mixture of 3 Gaussian : probabilities [1/3 1/3 1/3]



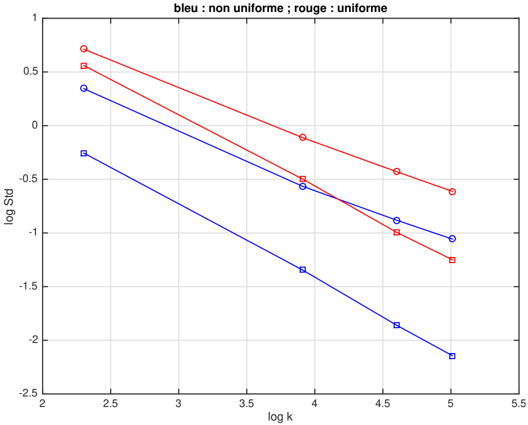
Example of the correlation matrix

Estimate $XX^\top/10000$ with $10000^{-1}\sum_i \varepsilon_i \boldsymbol{x}_i \boldsymbol{x}_i^\top/\pi_i$

$\pi_i = M/N$ for random sampling ($\circ$)

π_i approximated first order inclusion probability for k DPP ($\square$).

Plot the norm of matrices of st.d. ; $k = 10, 50, 100, 150$; 300 rff.



Coresets

Suppose observing $\mathcal{X}$ a set of N points in $\mathbb{R}^d$ bearing information over a parameter θ from a compact set $\Theta \subset \mathbb{R}^k$.

Consider estimating θ *via* minimization of a risk

$$L(\mathcal{X}, \theta) = \sum_{x \in \mathcal{X}} f(x, \theta)$$

where $f : \mathbb{R}^d \times \Theta \to \mathbb{R}^+$ a well behaved (Lipschitz in θ) cost

Suppose N too large and/or f very complicated and costly to evaluate

$\Longrightarrow$ use of a subset of $\mathcal{X}$ a possibility

An ε coreset is a $\mathcal{Y} \subset \mathcal{X}$ such that

$$\forall \theta \in \Theta, |\, L(\mathcal{X}, \theta) - L(\mathcal{Y}, \theta)| \leq \varepsilon L(\mathcal{X}, \theta)$$

Random Coresets

Let p_i a probability on $1, \ldots, N$.

Let $\mathcal{Y}$ be composed of M elements of $\mathcal{X}$ taken independently (with replacement) with probability p_i.

Recall $L(\mathcal{X}, \theta) = \sum_{i=1}^{N} f(x_i, \theta)$. Construct $\widehat{L}(\mathcal{Y}, \theta) = \sum_{i=1}^{M} \frac{f(y_i, \theta)}{M p_i}$.

Choose small ε, δ. Then, if $M(\varepsilon, \delta)$ is large enough

$$\Pr\left(\forall\theta, |\, L(\mathcal{X}, \theta) - \widehat{L}(\mathcal{Y}, \theta)| \leq \varepsilon L(\mathcal{X}, \theta)\right) \geq 1 - \delta$$

A good probability law in this problem : $p_i \propto \max_\theta \frac{f(x_i, \theta)}{L}$
Once again, some samples may be over represented and diversity may be beneficial !

DDP for coresets

To include negative correlation or diversity, sample a DPP or a k-DPP, and form

$$\widehat{L}(\mathcal{Y}, \theta) = \sum_{i \in \mathcal{Y}} \frac{f(y_i, \theta)}{\pi_i}$$

We proved that for DPP (and k DPP) :

- If $|\mathcal{Y}|$ (or k) is large enough

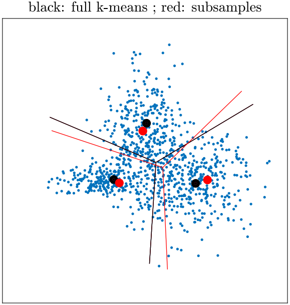
$$\Pr\left(\forall \theta, |\, L(\mathcal{X}, \theta) - \widehat{L}(\mathcal{Y}, \theta)| \leq \varepsilon L(\mathcal{X}, \theta)\right) \geq 1 - \delta$$

  We guarantee that DPP cannot be worse than i.i.d. sampling

- variance reduction due to negative correlation implies that DDP are better

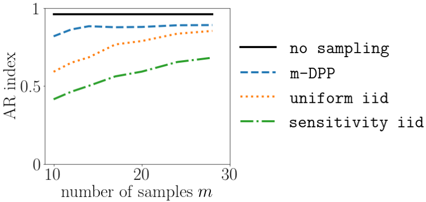
Application for clustering

$$L(\mathcal{X}, \theta) = \sum_{x \in \mathcal{X}} f(x, \theta) \quad \text{with} \quad f(x, \theta) = \min_{c \in \theta} \|x - c\|^2.$$



Use the classical MNIST data base (7.10^4 handwritten digits)

AR similarity index between ground truth and subsampling strategies (the closer to one the better)



To conclude

Continuous framework : more to follow.

- Usefulness of DPP (and negatively associated processes) for subsampling while conserving as much information as possible
- Promising applications in DoE, see works by *e.g.* Fanuel, Rémi
- A lot of work to do on these processes, notably their behavior in high dimensions ; their sampling, their use in sensor networks/distributed algorithms design . . .

More in

- Surveys Stat. (Bardenet *et. al.*, Lavancier *et. al.*), ML (Kuleza &Taskar), Prob (Lyons&Peres, Hough *et.al*, Bacchelli) . . .
- `arXiv:1803.01576` : Asymptotic Equivalence of Fixed-size and Varying-size Determinantal Point Processes, Bernoulli 2019
- `arXiv:1803.08700` : Determinantal Point Processes for Coresets, JMLR 2020
- Est/MC integration : Bardenet&Hardy, Coeurjolly, Mazoyer, POA, Stat. Spat & EJS to appear.
- Recent arXiv papers Barthelmé, Tremblay, Usevitch and POA.
- and some conf. papers (e.g. EUSIPCO 2017, IEEE SSP 2018)