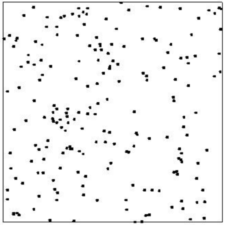
Strongly correlated particle systems:

a toolbox for machine intelligence

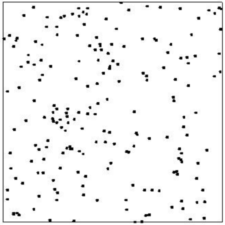
Subhro Ghosh
National University of Singapore

A gas of particles : ideal vs. real

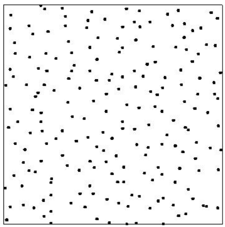


Ideal

A gas of particles : ideal vs. real



Ideal



Coulomb

The IID paradigm

The IID paradigm of randomness

- Independent and Identically Distributed (I.I.D.) is the most popular model of randomness in science

The IID paradigm

The IID paradigm of randomness

- Independent and Identically Distributed (I.I.D.) is the most popular model of randomness in science
- Tractable, interpretable, easy

The IID paradigm

The IID paradigm of randomness

- Independent and Identically Distributed (I.I.D.) is the most popular model of randomness in science
- Tractable, interpretable, easy
- Lead to many foundational ideas and methods in ML

The IID paradigm

The IID paradigm of randomness

- Independent and Identically Distributed (I.I.D.) is the most popular model of randomness in science
- Tractable, interpretable, easy
- Lead to many foundational ideas and methods in ML

Questions

- Is there a real benefit to exploring beyond IID ?

The IID paradigm

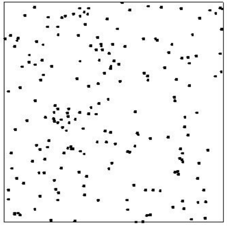
The IID paradigm of randomness

- Independent and Identically Distributed (I.I.D.) is the most popular model of randomness in science
- Tractable, interpretable, easy
- Lead to many foundational ideas and methods in ML

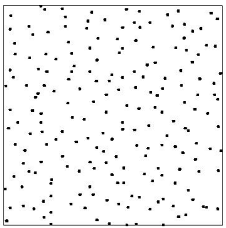
Questions

- Is there a real benefit to exploring beyond IID ?
- Even if there is, would it be realistic ?

Particle systems : ideal vs. real



Ideal



Coulomb

Problem

IID samples may be less representative or less stable

IID vs strongly correlated samples

Monte Carlo sampling

$$\Lambda(f) := \int f(x)\mathrm{d}\mu(x) \longleftrightarrow \Lambda_{\mathcal{S}}(f) := \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} f(X_i)$$

IID vs strongly correlated samples

Monte Carlo sampling

$$\Lambda(f) := \int f(x)\mathrm{d}\mu(x) \longleftrightarrow \Lambda_{\mathcal{S}}(f) := \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} f(X_i)$$

Sampling Coresets

$$L(f) = \sum_{x\in\mathcal{X}} f(x), \quad f \in \mathcal{F} \longleftrightarrow L_{\mathcal{S}}(f) := \sum_{x\in\mathcal{S}} w(x)f(x)$$

Many applications of sampling

Feature Selection

Sample a subset of columns of a low rank matrix to be representative of the entire matrix

Many applications of sampling

Feature Selection

Sample a subset of columns of a low rank matrix to be representative of the entire matrix

Neural Network Pruning

Delete redundant edges from a neural network without compromising on quality of output

....

• "Negative Dependence as a toolbox for machine learning : review and new developments"
H.S. Tran, V. Petrovich, R. Bardenet, S.Ghosh
Arxiv preprint

Many applications of sampling

IID samples

Approximation guarantee $O(m^{-1/2})$

Many applications of sampling

IID samples

Approximation guarantee $O(m^{-1/2})$

Strongly Correlated samples

Approximation guarantee $O(m^{-\gamma}), \gamma > 1/2$

IID vs strongly correlated samples

IID vs strongly correlated samples

Suitable strongly correlated samplers provide similar approximation guarantees with much smaller sample size than IID

IID vs strongly correlated samples

IID vs strongly correlated samples

Suitable strongly correlated samplers provide similar approximation guarantees with much smaller sample size than IID

Key benefits

Significant benefit in settings where function evaluation is costly

IID vs strongly correlated samples

IID vs strongly correlated samples

Suitable strongly correlated samplers provide similar approximation guarantees with much smaller sample size than IID

Key benefits

Significant benefit in settings where function evaluation is costly

- Stochastic Gradient Descent (SGD) for highly complex functions

IID vs strongly correlated samples

IID vs strongly correlated samples

Suitable strongly correlated samplers provide similar approximation guarantees with much smaller sample size than IID

Key benefits

Significant benefit in settings where function evaluation is costly

- Stochastic Gradient Descent (SGD) for highly complex functions
  - Large scale neural networks
  - Large scale conditional random field (CRF) models

Strongly correlated particle systems: some natural models

Determinantal Point Processes

A significant class of natural strongly correlated particle systems are *Determinantal Point Processes* or DPPs :

Strongly correlated particle systems: some natural models

Determinantal Point Processes

A significant class of natural strongly correlated particle systems are *Determinantal Point Processes* or DPPs : combination of tractability and feasibility

Strongly correlated particle systems: some natural models

Determinantal Point Processes

A significant class of natural strongly correlated particle systems are *Determinantal Point Processes* or DPPs : combination of tractability and feasibility

- A DPP is a random set of points that all interact with each other, and where the interaction is encoded by a kernel.
- DPPs are, in a sense, the kernel machine of particle systems.

Strongly correlated particle systems: some natural models

Origins in physics

DPPs originate canonically in quantum and statistical physics

Strongly correlated particle systems: some natural models

Origins in physics

DPPs originate canonically in quantum and statistical physics

- DPP strtucture arises as *Slater determinants* in wave-functions for Fermions (following earlier work by Heisenberg and Dirac)

Strongly correlated particle systems: some natural models

Origins in physics

DPPs originate canonically in quantum and statistical physics

- DPP strtucture arises as *Slater determinants* in wave-functions for Fermions (following earlier work by Heisenberg and Dirac)
- Connections to a wide interface of physics and mathematics, including random matrices, random polynomials, random networks, Coulomb gases ...

Correlation functions

Correlation functions

A random point set is characterised by its ‘correlation functions’, which are essentially the joint probabilities of having points at specified locations

Correlation functions

Correlation functions

A random point set is characterised by its ‘correlation functions’, which are essentially the joint probabilities of having points at specified locations

- If $\alpha_1, \ldots, \alpha_m$ are m fixed locations, then the m-point correlation function $\rho_m(\alpha_1, \ldots, \alpha_m)$ is the joint probability (density) of having points at the locations $\alpha_1, \ldots \alpha_m$ in a realization of the random point set.

Correlation functions

Correlation functions

A random point set is characterised by its ‘correlation functions’, which are essentially the joint probabilities of having points at specified locations

- If $\alpha_1, \ldots, \alpha_m$ are m fixed locations, then the m-point correlation function $\rho_m(\alpha_1, \ldots, \alpha_m)$ is the joint probability (density) of having points at the locations $\alpha_1, \ldots \alpha_m$ in a realization of the random point set.
- E.g.: for iid points with density ρ then $\rho_m(\alpha_1, \ldots, \alpha_m) = \rho^m$

DPP correlation functions

Correlation functions of a DPP

The m-point correlation functions of a DPP are given by determinants of a kernel K :

$$\rho_m(\alpha_1, \dots, \alpha_m) = \mathrm{Det} \begin{bmatrix} K(\alpha_1, \alpha_1), & \dots & \dots & K(\alpha_1, \alpha_m) \\ \dots & \dots & \dots & \dots \\ K(\alpha_m, \alpha_1), & \dots & \dots & K(\alpha_m, \alpha_m) \end{bmatrix}$$

DPP correlation functions

Correlation functions of a DPP

The m-point correlation functions of a DPP are given by determinants of a kernel K :

$$\rho_m(\alpha_1, \dots, \alpha_m) = \mathrm{Det} \begin{bmatrix} K(\alpha_1, \alpha_1), & \dots & \dots & K(\alpha_1, \alpha_m) \\ \dots & \dots & \dots & \dots \\ K(\alpha_m, \alpha_1), & \dots & \dots & K(\alpha_m, \alpha_m) \end{bmatrix}$$

- DPPs are particle systems parameterised by a kernel K.

DPP correlation functions

Correlation functions of a DPP

The m-point correlation functions of a DPP are given by determinants of a kernel K :

$$\rho_m(\alpha_1, \ldots, \alpha_m) = \mathrm{Det} \begin{bmatrix} K(\alpha_1, \alpha_1), & \ldots & \ldots & K(\alpha_1, \alpha_m) \\ \ldots & \ldots & \ldots & \ldots \\ K(\alpha_m, \alpha_1), & \ldots & \ldots & K(\alpha_m, \alpha_m) \end{bmatrix}$$

- DPPs are particle systems parameterised by a kernel K.
- Repulsion : if α_i and α_j are the close to each other for different i and j, then ρ_m is close to 0.

DPP correlation functions

Correlation functions of a DPP

The m-point correlation functions of a DPP are given by determinants of a kernel K :

$$\rho_m(\alpha_1, \ldots, \alpha_m) = \mathrm{Det} \begin{bmatrix} K(\alpha_1, \alpha_1), & \ldots & \ldots & K(\alpha_1, \alpha_m) \\ \ldots & \ldots & \ldots & \ldots \\ K(\alpha_m, \alpha_1), & \ldots & \ldots & K(\alpha_m, \alpha_m) \end{bmatrix}$$

- DPPs are particle systems parameterised by a kernel K.
- Repulsion : if α_i and α_j are the close to each other for different i and j, then ρ_m is close to 0.

DPP models

Sampling diversity

DPPs are, therefore, effective in modelling situations where the sample points are desired to be very different from each other.

DPP models

Sampling diversity

DPPs are, therefore, effective in modelling situations where the sample points are desired to be very different from each other.

- E.g., in diversity sampling,
  - Population is represented by points in some (high dimensional) feature space

DPP models

Sampling diversity

DPPs are, therefore, effective in modelling situations where the sample points are desired to be very different from each other.

- E.g., in diversity sampling,
  - Population is represented by points in some (high dimensional) feature space
  - Kernel K incorporates the proximity between these points in the feature space

DPP models

Sampling diversity

DPPs are, therefore, effective in modelling situations where the sample points are desired to be very different from each other.

- E.g., in diversity sampling,
  - Population is represented by points in some (high dimensional) feature space
  - Kernel K incorporates the proximity between these points in the feature space
  - This in turn encodes the 'similarity' between points in the ground set we want to sample from

DPP models

Sampling diversity

DPPs are, therefore, effective in modelling situations where the sample points are desired to be very different from each other.

- E.g., in diversity sampling,
  - Population is represented by points in some (high dimensional) feature space
  - Kernel K incorporates the proximity between these points in the feature space
  - This in turn encodes the 'similarity' between points in the ground set we want to sample from
- Applications include Feature Selection, Monte Carlo integration, Coreset selection, Dimensionality Reduction ....

Orthogonal Polynomials

- For a probability distribution γ on a euclidean space $\mathbb{R}^d$, consider the monomial functions $(x_1, \ldots, x_d) \mapsto x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ in the graded lexical order.

Orthogonal Polynomials

- For a probability distribution γ on a euclidean space $\mathbb{R}^d$, consider the monomial functions $(x_1, \ldots, x_d) \mapsto x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ in the graded lexical order.
- Then apply the Gram-Schmidt algorithm in $L^2(\gamma)$ to these ordered monomials.

Orthogonal Polynomials

- For a probability distribution γ on a euclidean space $\mathbb{R}^d$, consider the monomial functions $(x_1, \dots, x_d) \mapsto x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ in the graded lexical order.
- Then apply the Gram-Schmidt algorithm in $L^2(\gamma)$ to these ordered monomials.
- This yields a sequence of orthonormal polynomial functions $(\varphi_k)_{k \in \mathbb{N}}$, the multivariate orthonormal polynomials w.r.t. γ.
- Construct a DPP with the kernel given by the projection

$$K(x, y) = \sum_{k=0}^{m-1} \varphi_k(x) \varphi_k(y),$$

Orthogonal polynomials and DPPs

Orthogonal polynomial models

An effective choice of kernel for sampling applications: Projection kernel onto spaces of orthogonal polynomials (OP) $\subseteq L^2(\mu)$

Orthogonal polynomials and DPPs

Orthogonal polynomial models

An effective choice of kernel for sampling applications: Projection kernel onto spaces of orthogonal polynomials (OP) $\subseteq L^2(\mu)$

- Computing OP kernels equivalent to computing moments of μ
- Rank of kernel $=$ number of moments needed $=$ sample size
- Can be naturally extended to Reproducing Kernel Hilbert Space (RKHS) of approximating functions

How much mileage does a DPP sampler give?

Theorem (Bardenet-G.-Simon-Tran'24 ; Bardenet-G.-Lin '21)

- *OP based DPP samplers give approximation guarantees* $O_P\left(m^{-(1/2+1/(2d))}\right)$

How much mileage does a DPP sampler give?

Theorem (Bardenet-G.-Simon-Tran'24 ; Bardenet-G.-Lin '21)

- *OP based DPP samplers give approximation guarantees* $O_P\left(m^{-(1/2+1/(2d))}\right)$
- *PAC bounds of the same order (upto log factors) for several ML applications*

How much mileage does a DPP sampler give?

Theorem (Bardenet-G.-Simon-Tran'24 ; Bardenet-G.-Lin '21)

- *OP based DPP samplers give approximation guarantees* $O_P\left(m^{-(1/2+1/(2d))}\right)$
- *PAC bounds of the same order (upto log factors) for several ML applications*

How much mileage does a DPP sampler give?

Theorem (Bardenet-G.-Simon-Tran'24 ; Bardenet-G.-Lin '21)

- *OP based DPP samplers give approximation guarantees* $O_P\left(m^{-(1/2+1/(2d))}\right)$
- *PAC bounds of the same order (upto log factors) for several ML applications*

- Here d may be taken to be the reduced dimension of the data (after pre-processing via dimension reduction)
- DPP is fundamentally a Hilbert space based technique, so it works well with linear projection based dimension reduction methods

How much mileage does a DPP sampler give?

Theorem (Bardenet-G.-Simon-Tran'24 ; Bardenet-G.-Lin '21)

- *OP based DPP samplers give approximation guarantees* $O_P\left(m^{-(1/2+1/(2d))}\right)$
- *PAC bounds of the same order (upto log factors) for several ML applications*

How much mileage does a DPP sampler give?

Theorem (Bardenet-G.-Simon-Tran'24 ; Bardenet-G.-Lin '21)

- *OP based DPP samplers give approximation guarantees* $O_P\left(m^{-(1/2+1/(2d))}\right)$
- *PAC bounds of the same order (upto log factors) for several ML applications*

Regularity and rates

- Better rates possible for function classes with smoothness properties using RKHS techniques
- Interplay between smoothness of objective and rate of DPP approximation remains to be fully understood

How much mileage does a DPP sampler give?

Theorem (Concentration bounds ; Bardenet-G.-Simon-Tran'24)

Let $\Phi = (\varphi_1, \ldots, \varphi_m)^\top$ *be a vector-valued test function,*

How much mileage does a DPP sampler give?

Theorem (Concentration bounds ; Bardenet-G.-Simon-Tran'24)

Let $\Phi = (\varphi_1, \ldots, \varphi_m)^\top$ *be a vector-valued test function, and set* $\Lambda_m(\Phi) := (\Lambda_m(\varphi_i))_{i=1}^m$, $\mathbb{V}(\Phi) = (\mathrm{Var}\Lambda_m(\varphi_i)^{1/2})_{i=1}^m$

How much mileage does a DPP sampler give?

Theorem (Concentration bounds ; Bardenet-G.-Simon-Tran'24)

Let $\Phi = (\varphi_1, \ldots, \varphi_m)^\top$ *be a vector-valued test function, and set* $\Lambda_m(\Phi) := (\Lambda_m(\varphi_i))_{i=1}^m$, $\mathbb{V}(\Phi) = (\mathrm{Var}\Lambda_m(\varphi_i)^{1/2})_{i=1}^m$. *Then*

$$\mathbb{P}(\|\Lambda_m(\Phi) - \mathbb{E}[\Lambda_m(\Phi)]\| \geq \varepsilon) \leq 2m \exp\Big(- \frac{\varepsilon^2}{4A\|\mathbb{V}(\Phi)\|^2}\Big),$$

for $0 \leq \varepsilon \leq \frac{2A\|\mathbb{V}(\Phi)\|}{3} \cdot \min_{1 \leq i \leq m} \frac{\sqrt{\mathrm{Var}\Lambda_m(\varphi_i)}}{\|\varphi_i\|_\infty}$.

How much mileage does a DPP sampler give?

Theorem (Concentration bounds ; Bardenet-G.-Simon-Tran'24)

$$\mathbb{P}(\|\Lambda_m(\Phi) - \mathbb{E}[\Lambda_m(\Phi)]\| \geq \varepsilon) \leq 2m \exp\Big(- \frac{\varepsilon^2}{4A\|\mathbb{V}(\Phi)\|^2}\Big),$$

for $0 \leq \varepsilon \leq \frac{2A\|\mathbb{V}(\Phi)\|}{3} \cdot \min_{1\leq i\leq m} \frac{\sqrt{\mathrm{Var}\Lambda_m(\varphi_i)}}{\|\varphi_i\|_\infty}$.

How much mileage does a DPP sampler give?

Theorem (Concentration bounds ; Bardenet-G.-Simon-Tran'24)

$$\mathbb{P}(\|\Lambda_m(\Phi) - \mathbb{E}[\Lambda_m(\Phi)]\| \geq \varepsilon) \leq 2m \exp\Big(- \frac{\varepsilon^2}{4A\|\mathbb{V}(\Phi)\|^2}\Big),$$

for $0 \leq \varepsilon \leq \frac{2A\|\mathbb{V}(\Phi)\|}{3} \cdot \min_{1\leq i\leq m} \frac{\sqrt{\mathrm{Var}\Lambda_m(\varphi_i)}}{\|\varphi_i\|_\infty}$.

Concentration and strong dependence

- This is an example of a concentration phenomenon (with features similar to those for independent r.v.s), but in the context of strongly dependent stochastic model.
- Underscores how DPP blends strong dependence with structure and tractability, which allows for such powerful results

General structure and scope

- DPP samplers provide representative samples in very general settings

General structure and scope

- DPP samplers provide representative samples in very general settings
- This includes complicated geometries or even discrete, combinatorial or non-geometric background spaces

General structure and scope

- DPP samplers provide representative samples in very general settings
- This includes complicated geometries or even discrete, combinatorial or non-geometric background spaces
- Other methods, such as grid points or low discrepancy sequences
  - are very specific to simple geometric settings (such as Euclidean spaces)

General structure and scope

- DPP samplers provide representative samples in very general settings
- This includes complicated geometries or even discrete, combinatorial or non-geometric background spaces
- Other methods, such as grid points or low discrepancy sequences
  - are very specific to simple geometric settings (such as Euclidean spaces)
  - may scale poorly with dimension

General structure and scope

- DPP samplers provide representative samples in very general settings
- This includes complicated geometries or even discrete, combinatorial or non-geometric background spaces
- Other methods, such as grid points or low discrepancy sequences
  - are very specific to simple geometric settings (such as Euclidean spaces)
  - may scale poorly with dimension
- DPP samplers provide a notion of a ”stochastic grid pattern” when there may not even be any geometry to support the notion of a grid

General structure and scope

- DPP samplers provide representative samples in very general settings
- This includes complicated geometries or even discrete, combinatorial or non-geometric background spaces
- Other methods, such as grid points or low discrepancy sequences
  - are very specific to simple geometric settings (such as Euclidean spaces)
  - may scale poorly with dimension
- DPP samplers provide a notion of a "stochastic grid pattern" when there may not even be any geometry to support the notion of a grid
- The power of abstraction: DPPs are able to handle abstract spaces without geometry by looking at the function space on top of that space, thereby bringing to bear many tools from Euclidean spaces

Some heuristics

The power of the second-order effect

- It is crucial that the sample points "see" each other for improvement in the exponent on m :

Some heuristics

The power of the second-order effect

- It is crucial that the sample points "see" each other for improvement in the exponent on m : interactions must be at least two body

Some heuristics

The power of the second-order effect

- It is crucial that the sample points "see" each other for improvement in the exponent on m : interactions must be at least two body
- If not, we are reduced to importance sampling, which achieves improvement only in the leading constant

Some heuristics

The power of the second-order effect

- It is crucial that the sample points "see" each other for improvement in the exponent on m : interactions must be at least two body
- If not, we are reduced to importance sampling, which achieves improvement only in the leading constant
- It is crucial that the kernel K is a *projection operator* on $L^2(\mu)$, otherwise fluctuations are Poissonian in nature, and usually of similar order as independent sampling

Sampling from DPPs

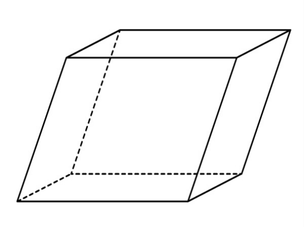
Hough-Krishnapur-Peres-Virag '06

Spectral sampling algorithm available, leveraging Hilbert space geometry of the kernel mapping

Sampling from DPPs

Hough-Krishnapur-Peres-Virag '06

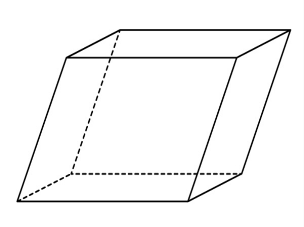
Spectral sampling algorithm available, leveraging Hilbert space geometry of the kernel mapping



Sampling from DPPs

Hough-Krishnapur-Peres-Virag '06

Spectral sampling algorithm available, leveraging Hilbert space geometry of the kernel mapping



Gillenwater et al '19 ; Tremblay et al '23 ; Anari et al '24

New tree-based algorithms built on top of the classical spectral sampler gives fast computational load of $O(m^2 \log N)$ (and similar) for each sample

Statistical Learning Theory

Key challenges

- Lack of independence of sample points renders fundamental empirical process techniques ineffective
- Eg, the basic symmetrization trick fails!
- Concentration phenomena understood only to a limited extent

Statistical Learning Theory

Key challenges

- Lack of independence of sample points renders fundamental empirical process techniques ineffective
- Eg, the basic symmetrization trick fails!
- Concentration phenomena understood only to a limited extent

Theorem (Dong-G.-Mendelson-Tran, preliminary results)

In 1D, we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{H}^2(\mathbb{T})} |\Lambda_m(f) - \mathbb{E}\left[\Lambda_m(f)\right]|\right] \lesssim \frac{\sqrt{\log m}}{m}$$

- Better than Gaussian rates (compared to independent samples)
- Compare: Discrepancy lower bound is $O(1/m)$

Towards a parametric theory of DPPs

- DPPs have emerged as a clearly interesting class of stochastic models to use as a sampling toolbox
- In statistical applications of DPPs (spatial statistics, biological data

Towards a parametric theory of DPPs

- DPPs have emerged as a clearly interesting class of stochastic models to use as a sampling toolbox
- In statistical applications of DPPs (spatial statistics, biological data c.f. Lavancier, Moller, Rubak, Taskar, Brunel, Baccelli...)
- A robust **parametric model** with naturally interpretable parameter modulation is **squarely lacking**.

Towards a parametric theory of DPPs

- DPPs have emerged as a clearly interesting class of stochastic models to use as a sampling toolbox
- In statistical applications of DPPs (spatial statistics, biological data c.f. Lavancier, Moller, Rubak, Taskar, Brunel, Baccelli...)
- A robust **parametric model** with naturally interpretable parameter modulation is **squarely lacking**.
- Compare, e.g., to the well-known exponential family models in probability, or Exponential Random Graph Models (ERGM) that are popular in the study of stochastic networks.

Towards a parametric theory of DPPs

- DPPs have emerged as a clearly interesting class of stochastic models to use as a sampling toolbox
- In statistical applications of DPPs (spatial statistics, biological data c.f. Lavancier, Moller, Rubak, Taskar, Brunel, Baccelli...)
- A robust **parametric model** with naturally interpretable parameter modulation is **squarely lacking**.
- Compare, e.g., to the well-known exponential family models in probability, or Exponential Random Graph Models (ERGM) that are popular in the study of stochastic networks.
- A parametric model will be a 'testing ground' to understand how the spatial behaviour of the points responds to parameter modulation, in turn leading to newer applications and even stronger models ...

Towards a parametric theory of DPPs

- DPPs have emerged as a clearly interesting class of stochastic models to use as a sampling toolbox
- In statistical applications of DPPs (spatial statistics, biological data c.f. Lavancier, Moller, Rubak, Taskar, Brunel, Baccelli...)
- A robust **parametric model** with naturally interpretable parameter modulation is **squarely lacking**.
- Compare, e.g., to the well-known exponential family models in probability, or Exponential Random Graph Models (ERGM) that are popular in the study of stochastic networks.
- A parametric model will be a 'testing ground' to understand how the spatial behaviour of the points responds to parameter modulation, in turn leading to newer applications and even stronger models ...

Towards a parametric theory of DPPs

- To this end, we propose the model of **Gaussian Determinantal Process** (GDP) [G. & Rigollet, PNAS (2020)]

Towards a parametric theory of DPPs

- To this end, we propose the model of **Gaussian Determinantal Process** (GDP) [G. & Rigollet, PNAS (2020)]
- The GDP is indexed by the space of positive definite matrices of a given dimension, which we will call the *scattering matrix*

Towards a parametric theory of DPPs

- To this end, we propose the model of **Gaussian Determinantal Process** (GDP) [G. & Rigollet, PNAS (2020)]
- The GDP is indexed by the space of positive definite matrices of a given dimension, which we will call the *scattering matrix*
- Connection to *Spiked Models* of random matrices and *Spiked* PCA

Towards a parametric theory of DPPs

- To this end, we propose the model of **Gaussian Determinantal Process** (GDP) [G. & Rigollet, PNAS (2020)]
- The GDP is indexed by the space of positive definite matrices of a given dimension, which we will call the *scattering matrix*
- Connection to *Spiked Models* of random matrices and *Spiked* PCA
- Applications to dimension reduction and clustering
- New kinds of random matrix phenomena based on truncated covariance matrices, leading to novel spectral asymptotics and connections with free probability ...

Gaussian Determinantal Processes

- A DPP is specified by the underlying kernel.
- The points of a GDP lives on $\mathbb{R}^d$, and the kernel is simply the d-dimensional Gaussian density with some positive definite covariance matrix Σ (which is the scattering matrix parameterizing the GDP):

$$K(x, y) = \frac{1}{(2\pi)^{d/2}\sqrt{\mathrm{Det}(\Sigma)}} \exp\left(-\frac{1}{2}(x-y)^T\Sigma^{-1}(x-y)\right).$$

Gaussian Determinantal Processes

- A DPP is specified by the underlying kernel.
- The points of a GDP lives on $\mathbb{R}^d$, and the kernel is simply the d-dimensional Gaussian density with some positive definite covariance matrix Σ (which is the scattering matrix parameterizing the GDP):

$$K(x, y) = \frac{1}{(2\pi)^{d/2}\sqrt{\mathrm{Det}(\Sigma)}} \exp\left(-\frac{1}{2}(x-y)^T\Sigma^{-1}(x-y)\right).$$

- The mean density of points in a DPP with kernel K is simply given by $K(x, x)$ - so the mean density of points in GDP is $= \frac{1}{(2\pi)^{d/2}\sqrt{\mathrm{Det}(\Sigma)}}$.

Gaussian Determinantal Processes

- A DPP is specified by the underlying kernel.
- The points of a GDP lives on $\mathbb{R}^d$, and the kernel is simply the d-dimensional Gaussian density with some positive definite covariance matrix Σ (which is the scattering matrix parameterizing the GDP):

$$K(x, y) = \frac{1}{(2\pi)^{d/2}\sqrt{\mathrm{Det}(\Sigma)}} \exp\left(-\frac{1}{2}(x-y)^T\Sigma^{-1}(x-y)\right).$$

- The mean density of points in a DPP with kernel K is simply given by $K(x, x)$ - so the mean density of points in GDP is $= \frac{1}{(2\pi)^{d/2}\sqrt{\mathrm{Det}(\Sigma)}}$.
- Our observation consists of the points in a realisation of the GDP inside a ball of large radius R.

Parametric modulation in GDP

- Our goal is to interpret the stochastic implication of varying or modulating the parameter Σ in the space $\mathcal{P}_d$ of $d \times d$ positive definite matrices.

Parametric modulation in GDP

- Our goal is to interpret the stochastic implication of varying or modulating the parameter Σ in the space $\mathcal{P}_d$ of $d \times d$ positive definite matrices.
- Note that modulating Σ such that $\mathrm{Det}(\Sigma)$ changes will lead to a change in the mean density of points, and can be detected simply by estimating this average density from the observed points.

Parametric modulation in GDP

- Our goal is to interpret the stochastic implication of varying or modulating the parameter Σ in the space $\mathcal{P}_d$ of $d \times d$ positive definite matrices.
- Note that modulating Σ such that $\mathrm{Det}(\Sigma)$ changes will lead to a change in the mean density of points, and can be detected simply by estimating this average density from the observed points.
- We will therefore focus on parametric modulation that leaves the determinant $\mathrm{Det}(\Sigma)$ invariant - similar to *shear mappings* or *shear transformations*.

Parametric modulation in GDP

- A key family of modulations that we will consider will be in the form of a **Spiked Model** in the space $\mathcal{P}_d$.

Parametric modulation in GDP

- A key family of modulations that we will consider will be in the form of a **Spiked Model** in the space $\mathcal{P}_d$.
- Formally, for a unit vector u and $\lambda > 0$, we will consider

$$\Sigma = (1+\lambda)uu^T + (1+\lambda)^{-\frac{1}{d-1}}(I_d - uu^T).$$

Parametric modulation in GDP

- A key family of modulations that we will consider will be in the form of a **Spiked Model** in the space $\mathcal{P}_d$.
- Formally, for a unit vector u and $\lambda > 0$, we will consider

$$\Sigma = (1+\lambda)uu^T + (1+\lambda)^{-\frac{1}{d-1}}(I_d - uu^T).$$

- $\lambda = 0$ makes $\Sigma = I_d$ - the 'isotropic' model with no directional bias in the dependency structure of the points.

Parametric modulation in GDP

- A key family of modulations that we will consider will be in the form of a **Spiked Model** in the space $\mathcal{P}_d$.
- Formally, for a unit vector u and $\lambda > 0$, we will consider

$$\Sigma = (1+\lambda)uu^T + (1+\lambda)^{-\frac{1}{d-1}}(I_d - uu^T).$$

- $\lambda = 0$ makes $\Sigma = I_d$ - the 'isotropic' model with no directional bias in the dependency structure of the points.
- $\lambda > 0$ corresponds to a spiked model that introduces directional bias in the strength of the dependency structure.

Parametric modulation in GDP

- The dependence (in this case, repulsion) between the points is much stronger, e.g. much more long-ranged (on the scale $1+\lambda$), in the spike direction u.

Parametric modulation in GDP

- The dependence (in this case, repulsion) between the points is much stronger, e.g. much more long-ranged (on the scale $1 + \lambda$), in the spike direction u.
- The dependence in the directions orthogonal to the spike is much weaker, and decouples to almost independent behaviour at relatively short length scales.

Parameter estimation in GDP

- $$\hat{\Sigma} = |B(1)|\frac{r^{d+2}}{d+2}I_d - \frac{1}{|B(R-r)|}\sum_{\|X_i - X_j\| < r}(X_i - X_j)(X_i - X_j)^T$$

  is a consistent estimator of Σ.

Parameter estimation in GDP

- $$\hat{\Sigma} = |B(1)|\frac{r^{d+2}}{d+2}I_d - \frac{1}{|B(R-r)|}\sum_{\|X_i - X_j\| < r}(X_i - X_j)(X_i - X_j)^T$$

  is a consistent estimator of Σ.
- Bias variance tradeoff leads to optimal choice of $r = \Theta(\sqrt{d \log n})$.
- Test statistics $\lambda_{\max}(\hat{\Sigma}), u_{\max}(\hat{\Sigma})$ allow detection of anisotropic direction with high probability if the spike size λ is above a threshold $\lambda_{n,d}$ (connections to BBP phase transition in spiked models of random matrices)
- Leads to study of truncated, and more generally, kernelized covariance matrices

Dimension Reduction and Directionality in Data

- The problem of dimension reduction is one of the central problems in the applied mathematics.

Dimension Reduction and Directionality in Data

- The problem of dimension reduction is one of the central problems in the applied mathematics.
- Roughly speaking, dimension reduction involves finding a low-dimensional subspace, or equivalently, a small number of 'significant directions', which contains most of the information about the (high dimensional) data.

Dimension Reduction and Directionality in Data

- Thus, the problem of dimensional reduction and the problem of detecting directionality in data are closely related.

Dimension Reduction and Directionality in Data

- Thus, the problem of dimensional reduction and the problem of detecting directionality in data are closely related.
- In P.C.A., we are interested in the directions of maximal variability, which are obtained by taking the principal eigen-directions of the empirical covariance matrix of the data.

Dimension Reduction and Directionality in Data

- Thus, the problem of dimensional reduction and the problem of detecting directionality in data are closely related.
- In P.C.A., we are interested in the directions of maximal variability, which are obtained by taking the principal eigen-directions of the empirical covariance matrix of the data.
- We may view the problem more generally, where dimension reduction will be performed by finding the optimal directions with respect to some other feature (as opposed to variance in the case of P.C.A.)

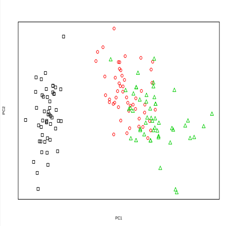
Dimension Reduction

- We use the GDP model as an *ansatz* for proposing a dimension reduction methodology.

Dimension Reduction

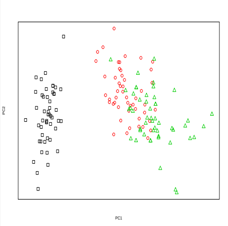
- We use the GDP model as an *ansatz* for proposing a dimension reduction methodology.
- We may compute the quantity $\hat{\Sigma}$ for any observed data set in $\mathbb{R}^d$. We then perform SVD on $\hat{\Sigma}$ and project the data points on to the principal eigen-directions of $\hat{\Sigma}$ in order to uncover low dimensional directional features in the data.

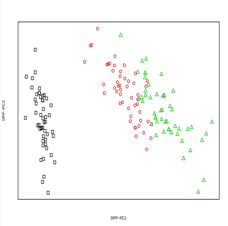
Dimension Reduction : Fisher's Iris



PCA

Dimension Reduction : Fisher's Iris

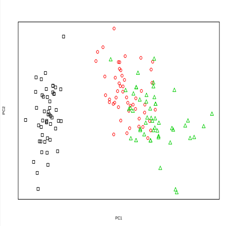




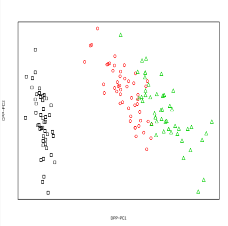
PCA

GDP

Dimension Reduction : Fisher's Iris



PCA



GDP

Theoretical analysis

In progress (with Dong, Mukherjee and Talukdar): minimax optimal guarantees for GDP-based clustering algorithms.

Kernelized sample covariance matrices

Kernelized sample covariance matrix

$$M_{\beta,n} = \frac{1}{2n^2} \sum_{1 \leq i,j \leq n} K_\beta(X_i, X_j)(X_i - X_j)(X_i - X_j)^\top$$

Kernelized sample covariance matrices

Kernelized sample covariance matrix

$$M_{\beta,n} = \frac{1}{2n^2} \sum_{1 \le i,j \le n} K_\beta(X_i, X_j)(X_i - X_j)(X_i - X_j)^\top$$

Typically:

$$K_\beta(x, y) = \varphi_\beta(D(x, y)),$$

where D is a distance and β can scale with dimension.

- The indicator kernel $\mathbb{I}(\|x - y\| \le r_{n,d})$
- The Gaussian kernel $1 - \exp\left(-\frac{\|x-y\|^2}{2\tau_{n,d}^2}\right)$

Limiting Spectral Laws of kernelized covariance matrices (with Mukherjee and Talukdar, arxiv)

- Depending on regularity properties of the kernel K_β, different asymptotic spectral laws can be established when the X_i's are i.i.d.

Limiting Spectral Laws of kernelized covariance matrices

Limiting Spectral Laws of kernelized covariance matrices (with Mukherjee and Talukdar, arxiv)

- Depending on regularity properties of the kernel K_β, different asymptotic spectral laws can be established when the X_i's are i.i.d.
- For smooth kernels, shifted and scaled family of Marcenko-Pastur type laws

Limiting Spectral Laws of kernelized covariance matrices (with Mukherjee and Talukdar, arxiv)

- Depending on regularity properties of the kernel K_β, different asymptotic spectral laws can be established when the X_i's are i.i.d.
- For smooth kernels, shifted and scaled family of Marcenko-Pastur type laws
- For non-smooth kernels, provable convergence to a more complicated spectral asymptotics that is characterised implicitly by an equation on the Stieltjes transform

Limiting Spectral Laws of kernelized covariance matrices

Limiting Spectral Laws of kernelized covariance matrices (with Mukherjee and Talukdar, arxiv)

- Depending on regularity properties of the kernel K_β, different asymptotic spectral laws can be established when the X_i's are i.i.d.
- For smooth kernels, shifted and scaled family of Marcenko-Pastur type laws
- For non-smooth kernels, provable convergence to a more complicated spectral asymptotics that is characterised implicitly by an equation on the Stieltjes transform
- Crucial issue is the dependence between the coefficient $K_\beta(X_1, X_j)$ and the rank 1 matrix $(X_i - X_j)(X_i - X_j)^\top$.

Zeros of Gaussian power series

A different strongly correlated random point field: zeros of Gaussian power series

$$\sum_{k=0}^{\infty} \xi_k \frac{z^k}{\sqrt{k!}}$$

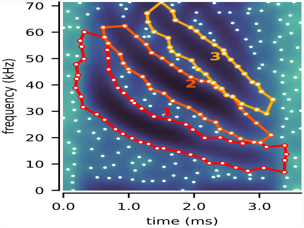
Zeros of Gaussian power series

A different strongly correlated random point field: zeros of Gaussian power series

$$\sum_{k=0}^{\infty} \xi_k \frac{z^k}{\sqrt{k!}}$$

- Arises in the study of quantum chaotic eigenstates (Bogomolny, Bohigas, Lebeouf, Nonenmacher, Voros ...)
- Captures the distribution of zeros of short time Fourier transform (STFT) of white noise

Gaussian zeros and signal processing



Bat echolocation signal

Gaussian zeros and signal processing

- Gaussian zeros form a strongly correlated point set
- Strongly rigid geometry makes deviations easy to detect
- Amenable to topological data analysis tools for signal reconstruction in a wide class of time indexed problems (incl. **gravitational waves**)

Why do DPP samples have reduced fluctuations ?

$$\mathrm{Var}\left[\Lambda_m(f)\right] = \iint \|f(\mathbf{z}) - f(\mathbf{w})\|_2^2 \, |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w})$$

$$\lesssim \mathcal{M}(f) \cdot \frac{1}{m^2} \int\int \|\mathbf{z} - \mathbf{w}\|_2^2 \, |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w})$$

- If $\|z - w\|_2^2$ was not present, then $\iint |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w}) = m$ implies $\mathrm{Var} \lesssim m^{-1}$, same as uniform random sampling
- Main contribution to $\iint |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w})$ comes from near the diagonal $\mathbf{z} = \mathbf{w}$, which is precisely suppressed by the term $\|\mathbf{z} - \mathbf{w}\|_2^2$.
- Use *Christoffel-Darboux formula* to make this control precise

Why do DPP samples have reduced fluctuations ?

$$\mathrm{Var}\left[\Lambda_m(f)\right] = \iint \|f(\mathbf{z}) - f(\mathbf{w})\|_2^2 \, |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w})$$

$$\lesssim \mathcal{M}(f) \cdot \frac{1}{m^2} \int \int \|\mathbf{z} - \mathbf{w}\|_2^2 \, |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w})$$

Why do DPP samples have reduced fluctuations ?

$$\mathrm{Var}\left[\Lambda_m(f)\right] = \iint \|f(\mathbf{z}) - f(\mathbf{w})\|_2^2 \, |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w})$$

$$\lesssim \mathcal{M}(f) \cdot \frac{1}{m^2} \int \int \|\mathbf{z} - \mathbf{w}\|_2^2 \, |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w})$$

- If $\|z - w\|_2^2$ was not present, then $\iint |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w}) = m$ implies $\mathrm{Var} \lesssim m^{-1}$, same as uniform random sampling

Why do DPP samples have reduced fluctuations ?

$$\mathrm{Var}\left[\Lambda_m(f)\right] = \iint \|f(\mathbf{z}) - f(\mathbf{w})\|_2^2 \, |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w})$$

$$\lesssim \mathcal{M}(f) \cdot \frac{1}{m^2} \int \int \|\mathbf{z} - \mathbf{w}\|_2^2 \, |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w})$$

- If $\|z - w\|_2^2$ was not present, then $\iint |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w}) = m$ implies $\mathrm{Var} \lesssim m^{-1}$, same as uniform random sampling
- Main contribution to $\iint |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w})$ comes from near the diagonal $\mathbf{z} = \mathbf{w}$, which is precisely suppressed by the term $\|\mathbf{z} - \mathbf{w}\|_2^2$.

Why do DPP samples have reduced fluctuations ?

$$\mathrm{Var}\left[\Lambda_m(f)\right] = \iint \|f(\mathbf{z}) - f(\mathbf{w})\|_2^2 \, |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w})$$

$$\lesssim \mathcal{M}(f) \cdot \frac{1}{m^2} \int \int \|\mathbf{z} - \mathbf{w}\|_2^2 \, |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w})$$

- If $\|z - w\|_2^2$ was not present, then $\iint |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w}) = m$ implies $\mathrm{Var} \lesssim m^{-1}$, same as uniform random sampling
- Main contribution to $\iint |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z})\mathrm{d}\mu(\mathbf{w})$ comes from near the diagonal $\mathbf{z} = \mathbf{w}$, which is precisely suppressed by the term $\|\mathbf{z} - \mathbf{w}\|_2^2$.
- Use *Christoffel-Darboux formula* to make this control precise

Why do DPP samples have reduced fluctuations ?

$$\mathrm{Var}\left[\Lambda_m(f)\right] = \iint \|f(\mathbf{z}) - f(\mathbf{w})\|_2^2 \, |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z}) \mathrm{d}\mu(\mathbf{w})$$

$$\lesssim \mathcal{M}(f) \cdot \frac{1}{m^2} \int \int \|\mathbf{z} - \mathbf{w}\|_2^2 \, |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z}) \mathrm{d}\mu(\mathbf{w})$$

- If $\|z - w\|_2^2$ was not present, then $\iint |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z}) \mathrm{d}\mu(\mathbf{w}) = m$ implies $\mathrm{Var} \lesssim m^{-1}$, same as uniform random sampling
- Main contribution to $\iint |K_m(\mathbf{z},\mathbf{w})|^2 \mathrm{d}\mu(\mathbf{z}) \mathrm{d}\mu(\mathbf{w})$ comes from near the diagonal $\mathbf{z} = \mathbf{w}$, which is precisely suppressed by the term $\|\mathbf{z} - \mathbf{w}\|_2^2$.
- Use *Christoffel-Darboux formula* to make this control precise

References

- “Gaussian determinantal processes: A new model for directionality in data”
S.Ghosh, P. Rigollet
Proceedings of the National Academy of Sciences, vol. 117, no. 24 (2020)

- “Small coresets via negative dependence: DPPs, linear statistics, and concentration”
R. Bardenet, S.Ghosh, H. Simon-Onfroy and H.S. Tran
NeurIPS 2024 (Spotlight)

- “D.P.P.s based on orthogonal polynomials for sampling minibatches in SGD”
R. Bardenet, S.Ghosh and M. Lin
NeurIPS 2021 (Spotlight)

- “Filtering through a topological lens : point processes and persistent homology on the time-frequency plane”
R. Bardenet, S.Ghosh, J.M. Miramont, S. Mukherjee and K.A. Tan
Arxiv preprint

- “Negative Dependence as a toolbox for machine learning : review and new developments”

H.S. Tran, V. Petrovich, R. Bardenet, S.Ghosh

Arxiv preprint