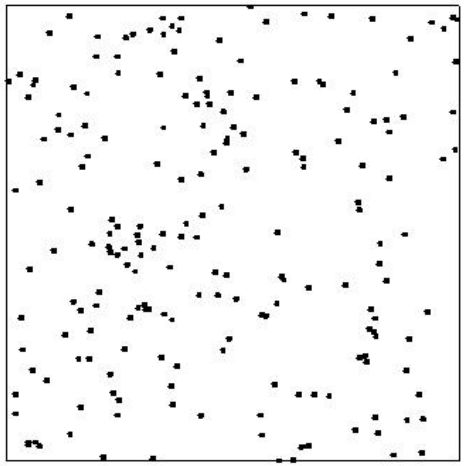
Strongly correlated particle systems:
a toolbox for machine intelligence

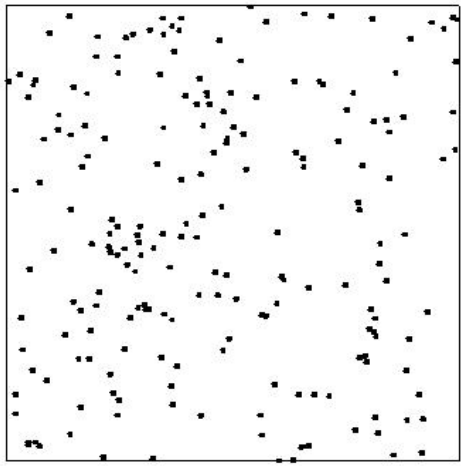
Subhro Ghosh
National University of Singapore

Particle systems : ideal vs. real

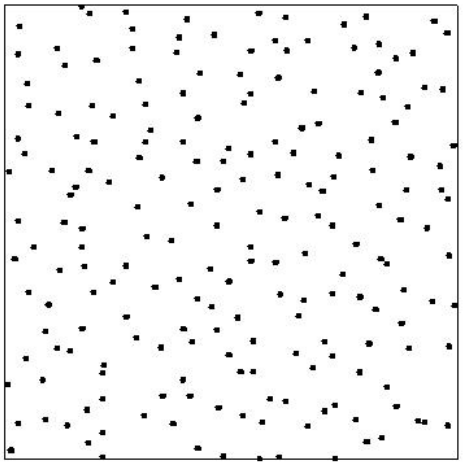


Ideal Gas

Particle systems : ideal vs. real



Ideal Gas



Coulomb Gas

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

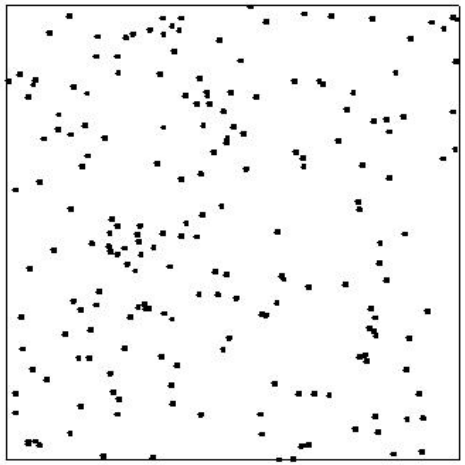
The IID paradigm of randomness

- 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

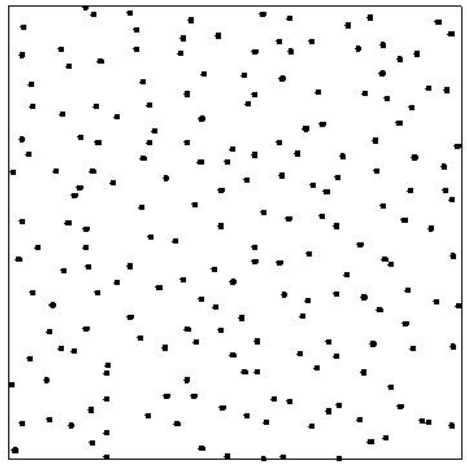
Questions

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

Particle systems : ideal vs. real



Ideal Gas



Coulomb Gas

IID samples may be less representative or less stable

IID vs strongly correlated samples

Sampling for quadrature

$$\Lambda_m(f) = \frac{1}{m}\sum_{i=1}^{m} f(X_i) \longleftrightarrow \int f(x)\mathrm{d}\mu(x)$$

IID vs strongly correlated samples

Sampling for quadrature

$$\Lambda_m(f) = \frac{1}{m}\sum_{i=1}^{m} f(X_i) \longleftrightarrow \int f(x)\mathrm{d}\mu(x)$$

IID samples

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

IID vs strongly correlated samples

Sampling for quadrature

$$\Lambda_m(f) = \frac{1}{m}\sum_{i=1}^{m} f(X_i) \longleftrightarrow \int f(x)\mathrm{d}\mu(x)$$

IID samples

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

Strongly correlated samples

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

- Importance sampling typically leads to an improvement in the constant on O
- Aim for improvement in the exponent on m via having points "see each other"

Gradient Descent

- Stochastic gradient descent (SGD) is the workhorse of modern machine learning and AI
- Routinely deployed for training models of enormous size, such as large scale neural networks

Gradient Descent

- Stochastic gradient descent (SGD) is the workhorse of modern machine learning and AI
- Routinely deployed for training models of enormous size, such as large scale neural networks
- Goal: Minimize loss $\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(z_i,\theta)$
- The fundamental step in gradient descent can be coded as

$$\theta_{t+1} \leftarrow \theta_t - \eta_t \cdot \left[\frac{1}{N}\cdot\sum_{i=1}^{N}\nabla_\theta\mathcal{L}(z_i,\theta)\right],$$

where η_t is the step-size

Gradient Descent

- Stochastic gradient descent (SGD) is the workhorse of modern machine learning and AI
- Routinely deployed for training models of enormous size, such as large scale neural networks
- Goal: Minimize loss $\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(z_i,\theta)$
- The fundamental step in gradient descent can be coded as

$$\theta_{t+1} \leftarrow \theta_t - \eta_t \cdot \left[\frac{1}{N}\cdot\sum_{i=1}^{N}\nabla_\theta\mathcal{L}(z_i,\theta)\right],$$

  where η_t is the step-size
- However, for large data sets, computing the empirical average at each step can be prohibitively expensive.

The SGD minibatch problem

- The fundamental step in stochastic gradient descent can be coded as
$$\theta_{t+1} \leftarrow \theta_t - \gamma_t \widehat{\mathcal{L}(A, \theta_t)}$$
- $\widehat{\mathcal{L}(A, \theta_t)}$ is an estimate of the full data set gradient $\frac{1}{N} \cdot \sum_{i=1}^{N} \nabla_\theta \mathcal{L}(\mathbf{z}_i, \theta)$
- $\widehat{\mathcal{L}(A, \theta_t)}$ is *based on a relatively small subsample $A \subset D$ of the full data set D.*

The SGD minibatch problem

- The fundamental step in stochastic gradient descent can be coded as
$$\theta_{t+1} \leftarrow \theta_t - \gamma_t \widehat{\mathcal{L}(A, \theta_t)}$$
- $\widehat{\mathcal{L}(A, \theta_t)}$ is an estimate of the full data set gradient $\frac{1}{N} \cdot \sum_{i=1}^{N} \nabla_\theta \mathcal{L}(\mathbf{z}_i, \theta)$
- $\widehat{\mathcal{L}(A, \theta_t)}$ is *based on a relatively small subsample $A \subset D$ of the full data set D.*
- Such a subsample A is called a minibatch.

The SGD minibatch problem

- A minibatch is a (random) subset $A \subset D$ of size $|A| = p \ll N$ such that the random variable

$$\Xi_A = \Xi_A(\theta) := \sum_{z_i \in A} w_i \nabla_\theta \mathcal{L}(z_i, \theta), \tag{1}$$

for suitable weights (w_i), provides a good approximation for $\Xi_N = \frac{1}{N} \sum_{\mathbf{z}_i \in \mathcal{D}} \nabla_\theta \mathcal{L}(\mathbf{z}_i, \theta)$.

The SGD minibatch problem

- A minibatch is a (random) subset $A \subset D$ of size $|A| = p \ll N$ such that the random variable

$$\Xi_A = \Xi_A(\theta) := \sum_{z_i \in A} w_i \nabla_\theta \mathcal{L}(z_i, \theta), \tag{1}$$

  for suitable weights (w_i), provides a good approximation for $\Xi_N = \frac{1}{N} \sum_{\mathbf{z}_i \in \mathcal{D}} \nabla_\theta \mathcal{L}(\mathbf{z}_i, \theta)$.
- Natural problem : How to select the *random* subset $A \subset D$? What is the *nature of randomness* that leads to improved performance of the SGD algorithm ?

The impact of randomness

- The impact of the randomness of A (equivalently, that of Ξ_A) is captured by the following theorem.

Theorem (Moulines-Bach '11)

For smooth and strongly convex expected loss function, compact parameter space, an unbiased estimator Ξ_A for the gradient and step size $\gamma_t \sim t^{-\alpha}$ for some $0 < \alpha < 1$, we have

$$\mathbb{E}\|\theta_t - \theta_\star\|^2 \leq C \cdot \frac{\sigma^2}{t^\alpha} + \mathcal{O}(e^{-t^\alpha}),$$

where $\sigma^2 = \mathbb{E}[\|\Xi_A(\theta_\star)\|^2|\mathcal{D}]$ is the trace of the covariance matrix of the gradient estimator, evaluated at the true optimizer $\theta_\star$.

The impact of randomness

- The impact of the randomness of A (equivalently, that of Ξ_A) is captured by the following theorem.

Theorem (Moulines-Bach '11)

For smooth and strongly convex expected loss function, compact parameter space, an unbiased estimator Ξ_A for the gradient and step size $\gamma_t \sim t^{-\alpha}$ for some $0 < \alpha < 1$, we have

$$\mathbb{E}\|\theta_t - \theta_\star\|^2 \leq C \cdot \frac{\sigma^2}{t^\alpha} + \mathcal{O}(e^{-t^\alpha}),$$

where $\sigma^2 = \mathbb{E}[\|\Xi_A(\theta_\star)\|^2|\mathcal{D}]$ is the trace of the covariance matrix of the gradient estimator, evaluated at the true optimizer $\theta_\star$.

- Therefore, the *goal is to make σ^2 small*, as a function of the batch size m.

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

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

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

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

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

Strongly correlated particle systems: some natural models

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

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

DPPs originate canonically in quantum and statistical physics

Strongly correlated particle systems: some natural models

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

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

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

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

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

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

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

DPP correlation functions

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.

DPP correlation functions

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

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

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

DPP models

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

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

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

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 $(x_1, \ldots, x_d) \mapsto x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ in the graded lexical order.
- Then apply the Gram-Schmidt algorithm in $L^2(\gamma)$ to these ordered monomials.
- 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

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

Orthogonal polynomials and DPPs

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*

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*

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

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 (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*

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

$$\sup_{f\in\mathcal{F}} |\Lambda_m(f) - \mathbb{E}\left[\Lambda_m(f)\right]| = \tilde{O}_P\left(m^{-(1/2+1/(2d))}\right)$$

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*

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

$$\sup_{f\in\mathcal{F}} |\Lambda_m(f) - \mathbb{E}\left[\Lambda_m(f)\right]| = \tilde{O}_P\left(m^{-(1/2+1/(2d))}\right)$$

- $\mathcal{F}$ is a function class of interest, eg the gradient of the loss $f_\theta(\mathbf{z}) = \nabla_\theta \mathcal{L}(\mathbf{z}, \theta)$ as the parameter θ varies along the gradient descent path

Sampling from DPPs

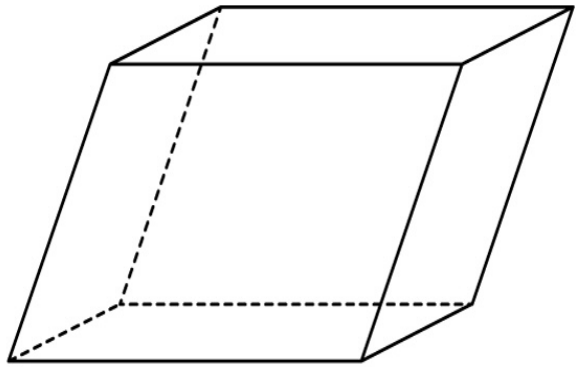
Hough-Krishnapur-Peres-Virag '06

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

Sampling from DPPs

Hough-Krishnapur-Peres-Virag '06

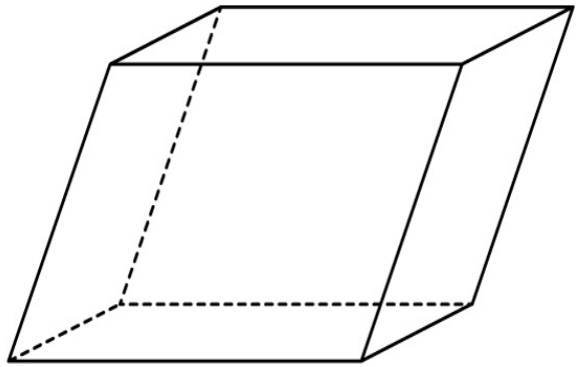
Spectral sampling algorithm available, leveraging Hilbert space geometry of the kernel mapping



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

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 geometric grids may not even be defined

Some heuristics

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

Some heuristics

- 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

- 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

- 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

Probabilistic theory

- Classic work by Diaconis and Evans ('80s): Orthogonal Polynomials on the unit circle (related to random matrix theory)

Probabilistic theory

- Classic work by Diaconis and Evans ('80s): Orthogonal Polynomials on the unit circle (related to random matrix theory)
- Obtained CLT for $\left(\sum_{i=1}^{m} f(x_i) - \int f(x)\mathrm{d}x\right)$ without any scaling

Probabilistic theory

- Classic work by Diaconis and Evans ('80s): Orthogonal Polynomials on the unit circle (related to random matrix theory)
- Obtained CLT for $\left(\sum_{i=1}^{m} f(x_i) - \int f(x)\mathrm{d}x\right)$ without any scaling
  Divide throughout by m to obtain the m^{-1} rate for $d = 1$

Probabilistic theory

- Classic work by Diaconis and Evans ('80s): Orthogonal Polynomials on the unit circle (related to random matrix theory)
- Obtained CLT for $\left(\sum_{i=1}^{m} f(x_i) - \int f(x)\mathrm{d}x\right)$ without any scaling
  Divide throughout by m to obtain the m^{-1} rate for $d = 1$
- Bardenet and Hardy ('20): Multivariate Orthogonal Polynomials ($d > 1$)

Probabilistic theory

- Classic work by Diaconis and Evans ('80s): Orthogonal Polynomials on the unit circle (related to random matrix theory)
- Obtained CLT for $\left(\sum_{i=1}^{m} f(x_i) - \int f(x)\mathrm{d}x\right)$ without any scaling Divide throughout by m to obtain the m^{-1} rate for $d = 1$
- Bardenet and Hardy ('20): Multivariate Orthogonal Polynomials ($d > 1$) Obtained analogous CLT (with more general sampling marginals μ) with the scaling $m^{\frac{1}{2} - \frac{1}{2d}}$

Probabilistic theory

- Classic work by Diaconis and Evans ('80s): Orthogonal Polynomials on the unit circle (related to random matrix theory)
- Obtained CLT for $\left(\sum_{i=1}^{m} f(x_i) - \int f(x)\mathrm{d}x\right)$ without any scaling Divide throughout by m to obtain the m^{-1} rate for $d = 1$
- Bardenet and Hardy ('20): Multivariate Orthogonal Polynomials ($d > 1$) Obtained analogous CLT (with more general sampling marginals μ) with the scaling $m^{\frac{1}{2} - \frac{1}{2d}}$
- CLTs are significant, since they provide confidence interval guarantees for $\frac{1}{m}\sum_{i=1}^{m} f(x_i)$ as an estimator of $\int f(x)\mathrm{d}\mu(x)$

Probabilistic theory

- Classic work by Diaconis and Evans ('80s): Orthogonal Polynomials on the unit circle (related to random matrix theory)
- Obtained CLT for $\left(\sum_{i=1}^{m} f(x_i) - \int f(x)\mathrm{d}x\right)$ without any scaling Divide throughout by m to obtain the m^{-1} rate for $d = 1$
- Bardenet and Hardy ('20): Multivariate Orthogonal Polynomials ($d > 1$) Obtained analogous CLT (with more general sampling marginals μ) with the scaling $m^{\frac{1}{2}-\frac{1}{2d}}$
- CLTs are significant, since they provide confidence interval guarantees for $\frac{1}{m}\sum_{i=1}^{m} f(x_i)$ as an estimator of $\int f(x)\mathrm{d}\mu(x)$
- CLTs are known to hold for linear statistics of DPPs under very general conditions, at the scale of its standard deviation (which is kernel and model dependent)

Interweaving between discrete and continuous spaces

- Rate guarantees easier to estimate in continuous spaces – analytical tools

Interweaving between discrete and continuous spaces

- Rate guarantees easier to estimate in continuous spaces – analytical tools
- Many sampling tasks are natural in discrete spaces – eg, SGD minibatches or coresets

Interweaving between discrete and continuous spaces

- Rate guarantees easier to estimate in continuous spaces – analytical tools
- Many sampling tasks are natural in discrete spaces – eg, SGD minibatches or coresets
- Need to transition between discrete and continuous spaces in terms of kernel design and rate analysis – important technical ingredient in DPP samplers

Interweaving between discrete and continuous spaces

- Rate guarantees easier to estimate in continuous spaces – analytical tools
- Many sampling tasks are natural in discrete spaces – eg, SGD minibatches or coresets
- Need to transition between discrete and continuous spaces in terms of kernel design and rate analysis – important technical ingredient in DPP samplers
- Spectral approximation (Bardenet-G.-Lin '21)

Interweaving between discrete and continuous spaces

- Rate guarantees easier to estimate in continuous spaces – analytical tools
- Many sampling tasks are natural in discrete spaces – eg, SGD minibatches or coresets
- Need to transition between discrete and continuous spaces in terms of kernel design and rate analysis – important technical ingredient in DPP samplers
- Spectral approximation (Bardenet-G.-Lin '21)
- General theory (Jaquard-Keriven '26)
    - Discrete Orthogonal Polynomial ensembles

Interweaving between discrete and continuous spaces

- Rate guarantees easier to estimate in continuous spaces – analytical tools
- Many sampling tasks are natural in discrete spaces – eg, SGD minibatches or coresets
- Need to transition between discrete and continuous spaces in terms of kernel design and rate analysis – important technical ingredient in DPP samplers
- Spectral approximation (Bardenet-G.-Lin '21)
- General theory (Jaquard-Keriven '26)
  - Discrete Orthogonal Polynomial ensembles
- General theory (Tran-Gupta-Bardenet-G. '26)
  - Novel exact identities for DPPs

What rates are possible?

- d in the exponent $m^{-\frac{1}{2}-\frac{1}{2d}}$ can in practice be taken to be the intrinsic dimension of the data (as opposed to ambient dimension), after a dimension reduction pre-processing step

What rates are possible?

- d in the exponent $m^{-\frac{1}{2}-\frac{1}{2d}}$ can in practice be taken to be the intrinsic dimension of the data (as opposed to ambient dimension), after a dimension reduction pre-processing step
- Intrinsic dimension can be much smaller than the ambient dimension – eg, in clustering problems, intrinsic dimension roughly equals the number of clusters

What rates are possible?

- d in the exponent $m^{-\frac{1}{2}-\frac{1}{2d}}$ can in practice be taken to be the intrinsic dimension of the data (as opposed to ambient dimension), after a dimension reduction pre-processing step
- Intrinsic dimension can be much smaller than the ambient dimension – eg, in clustering problems, intrinsic dimension roughly equals the number of clusters or in PCA, intrinsic dimension roughly equals the number of significant principal components

What rates are possible?

- d in the exponent $m^{-\frac{1}{2}-\frac{1}{2d}}$ can in practice be taken to be the intrinsic dimension of the data (as opposed to ambient dimension), after a dimension reduction pre-processing step
- Intrinsic dimension can be much smaller than the ambient dimension – eg, in clustering problems, intrinsic dimension roughly equals the number of clusters or in PCA, intrinsic dimension roughly equals the number of significant principal components

Theorem (Wavelet kernels and smoothness; Tran-Gupta-Bardenet-G. '26)

With wavelet kernels and $f \in \mathcal{H}^s, C^s; \quad s > 0$
Approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{s\wedge 1}{d}})$

What rates are possible?

Theorem (Wavelet kernels and smoothness; Tran-Gupta-Bardenet-G. '26)

With wavelet kernels and $f \in \mathcal{H}^s, C^s; \quad s > 0$

Approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{s\wedge 1}{d}})$

What rates are possible?

Theorem (Wavelet kernels and smoothness; Tran-Gupta-Bardenet-G. '26)

With wavelet kernels and $f \in \mathcal{H}^s, C^s; \quad s > 0$

Approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{s\wedge 1}{d}})$

- Improved exponent beyond orthogonal polynomials from $1/2d$ to $1/d$ (for $s \geq 1$)

What rates are possible?

Theorem (Wavelet kernels and smoothness; Tran-Gupta-Bardenet-G. '26)

With wavelet kernels and $f \in \mathcal{H}^s, C^s; \quad s > 0$

Approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{s\wedge 1}{d}})$

- Improved exponent beyond orthogonal polynomials from $1/2d$ to $1/d$ (for $s \geq 1$)
- On the unit circle, with a C^1 function f, we get an approximation guarantee of $m^{-3/2}$, compared to m^{-1} in the spirit of Diaconis-Evans

What rates are possible?

Theorem (Wavelet kernels and smoothness; Tran-Gupta-Bardenet-G. '26)

With wavelet kernels and $f \in \mathcal{H}^s, C^s; \quad s > 0$

Approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{s\wedge 1}{d}})$

- Improved exponent beyond orthogonal polynomials from $1/2d$ to $1/d$ (for $s \geq 1$)
- On the unit circle, with a C^1 function f, we get an approximation guarantee of $m^{-3/2}$, compared to m^{-1} in the spirit of Diaconis-Evans
- Works for objective functions f with very low regularity – any $s > 0$! Applicable to very rough data, eg from finance

What rates are possible?

Theorem (Wavelet kernels and smoothness; Tran-Gupta-Bardenet-G. '26)

With wavelet kernels and $f \in \mathcal{H}^s, C^s; \quad s > 0$

Approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{s\wedge 1}{d}})$

- Improved exponent beyond orthogonal polynomials from $1/2d$ to $1/d$ (for $s \geq 1$)
- On the unit circle, with a C^1 function f, we get an approximation guarantee of $m^{-3/2}$, compared to m^{-1} in the spirit of Diaconis-Evans
- Works for objective functions f with very low regularity – any $s > 0$! Applicable to very rough data, eg from finance
- Identifies *stratified sampling* as a DPP sampler with Haar wavelet kernel!

Quasi Monte Carlo and DPP samplers

- Quasi Monte Carlo methods to generate samples (designs) known to beat iid rates of $m^{-1/2}$

Quasi Monte Carlo and DPP samplers

- Quasi Monte Carlo methods to generate samples (designs) known to beat iid rates of $m^{-1/2}$
- Approximation rates known to scale as $(\log m)^d/m$

Quasi Monte Carlo and DPP samplers

- Quasi Monte Carlo methods to generate samples (designs) known to beat iid rates of $m^{-1/2}$
- Approximation rates known to scale as $(\log m)^d/m$ – difficult to scale with dimension d

Quasi Monte Carlo and DPP samplers

- Quasi Monte Carlo methods to generate samples (designs) known to beat iid rates of $m^{-1/2}$
- Approximation rates known to scale as $(\log m)^d/m$ – difficult to scale with dimension d
- Quasi Monte Carlo generally known to work in very strongly structured spaces, principally Euclidean spaces

Quasi Monte Carlo and DPP samplers

- Quasi Monte Carlo methods to generate samples (designs) known to beat iid rates of $m^{-1/2}$
- Approximation rates known to scale as $(\log m)^d/m$ – difficult to scale with dimension d
- Quasi Monte Carlo generally known to work in very strongly structured spaces, principally Euclidean spaces
- DPP samplers can be used on very general unstructured spaces, due to its highly abstract nature

Quasi Monte Carlo and DPP samplers

- Quasi Monte Carlo methods to generate samples (designs) known to beat iid rates of $m^{-1/2}$
- Approximation rates known to scale as $(\log m)^d/m$ – difficult to scale with dimension d
- Quasi Monte Carlo generally known to work in very strongly structured spaces, principally Euclidean spaces
- DPP samplers can be used on very general unstructured spaces, due to its highly abstract nature eg general geometries (such as manifolds), networks etc

Quasi Monte Carlo and DPP samplers

- Quasi Monte Carlo methods to generate samples (designs) known to beat iid rates of $m^{-1/2}$
- Approximation rates known to scale as $(\log m)^d/m$ – difficult to scale with dimension d
- Quasi Monte Carlo generally known to work in very strongly structured spaces, principally Euclidean spaces
- DPP samplers can be used on very general unstructured spaces, due to its highly abstract nature eg general geometries (such as manifolds), networks etc where natural grids might not even exist

DPP samplers on general spaces

Theorem (Tran-Gupta-G.'26)

On a manifold, DPP kernels based on eigenfunctions of the manifold Laplacian provide approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{1}{2d_{\mathrm{man}}}})$

- Depends on the dimension d_{man} of the manifold on (as opposed to the much larger ambient dimension), in a manner similar to OP kernels on Euclidean spaces

DPP samplers on general spaces

Theorem (Tran-Gupta-G.'26)

On a manifold, DPP kernels based on eigenfunctions of the manifold Laplacian provide approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{1}{2d_{\mathrm{man}}}})$

- Depends on the dimension d_{man} of the manifold on (as opposed to the much larger ambient dimension), in a manner similar to OP kernels on Euclidean spaces

Theorem (Tran-Gupta-G.'26)

On k-NN graphs and weighted geometric graphs, DPP kernels based on eigenfunctions of the graph Laplacian provides approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{1}{2d_{\mathrm{int}}}})$

DPP samplers on general spaces

Theorem (Tran-Gupta-G.'26)

On a manifold, DPP kernels based on eigenfunctions of the manifold Laplacian provide approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{1}{2d_{\mathrm{man}}}})$

- Depends on the dimension d_{man} of the manifold on (as opposed to the much larger ambient dimension), in a manner similar to OP kernels on Euclidean spaces

Theorem (Tran-Gupta-G.'26)

On k-NN graphs and weighted geometric graphs, DPP kernels based on eigenfunctions of the graph Laplacian provides approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{1}{2d_{\mathrm{int}}}})$

- Automatically adapts to the intrinsic dimensionality d_{int} of the manifold on which the data distribution is supported, without having to a-priori know anything about the underlying manifold (or that even such a structure exists)

DPP samplers on general spaces

Theorem (Tran-Gupta-G.'26)

On a manifold, DPP kernels based on eigenfunctions of the manifold Laplacian provide approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{1}{2d_{\mathrm{man}}}})$

- Depends on the dimension d_{man} of the manifold on (as opposed to the much larger ambient dimension), in a manner similar to OP kernels on Euclidean spaces

Theorem (Tran-Gupta-G.'26)

On k-NN graphs and weighted geometric graphs, DPP kernels based on eigenfunctions of the graph Laplacian provides approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{1}{2d_{\mathrm{int}}}})$

- Automatically adapts to the intrinsic dimensionality d_{int} of the manifold on which the data distribution is supported, without having to a-priori know anything about the underlying manifold (or that even such a structure exists)
- Connects to the so-called *Manifold Hypothesis*

DPP samplers on general spaces

Theorem (Tran-Gupta-G.'26)

On a manifold, DPP kernels based on eigenfunctions of the manifold Laplacian provide approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{1}{2d_{\mathrm{man}}}})$

Theorem (Tran-Gupta-G.'26)

On k-NN graphs and weighted geometric graphs, DPP kernels based on eigenfunctions of the graph Laplacian provides approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{1}{2d_{\mathrm{int}}}})$

DPP samplers on general spaces

Theorem (Tran-Gupta-G.'26)

On a manifold, DPP kernels based on eigenfunctions of the manifold Laplacian provide approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{1}{2d_{\mathrm{man}}}})$

Theorem (Tran-Gupta-G.'26)

On k-NN graphs and weighted geometric graphs, DPP kernels based on eigenfunctions of the graph Laplacian provides approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{1}{2d_{\mathrm{int}}}})$

- Analytical techniques leverage
  - Dirichlet Forms and theory of Markov diffusions

DPP samplers on general spaces

Theorem (Tran-Gupta-G.'26)

On a manifold, DPP kernels based on eigenfunctions of the manifold Laplacian provide approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{1}{2d_{\mathrm{man}}}})$

Theorem (Tran-Gupta-G.'26)

On k-NN graphs and weighted geometric graphs, DPP kernels based on eigenfunctions of the graph Laplacian provides approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{1}{2d_{\mathrm{int}}}})$

- Analytical techniques leverage
  - Dirichlet Forms and theory of Markov diffusions
  - Weyl's Law for manifold spectrum

DPP samplers on general spaces

Theorem (Tran-Gupta-G.'26)

On a manifold, DPP kernels based on eigenfunctions of the manifold Laplacian provide approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{1}{2d_{\mathrm{man}}}})$

Theorem (Tran-Gupta-G.'26)

On k-NN graphs and weighted geometric graphs, DPP kernels based on eigenfunctions of the graph Laplacian provides approximation guarantees $O_P(m^{-\frac{1}{2}-\frac{1}{2d_{\mathrm{int}}}})$

- Analytical techniques leverage
  - Dirichlet Forms and theory of Markov diffusions
  - Weyl's Law for manifold spectrum
  - Ingredients from pseudodifferential operator theory

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

Young Collaborators

Hoang Son Tran

Pranav Gupta

References

• "Small coresets via negative dependence: DPPs, linear statistics, and concentration"
R. Bardenet, S.Ghosh, H. Simon-Onfroy, H.S. Tran (*alphabetically ordered*)
NeurIPS 2024 (Spotlight)

• "D.P.P.s based on orthogonal polynomials for sampling minibatches in SGD"
R. Bardenet, S.Ghosh, M. Lin (*alphabetically ordered*)
NeurIPS 2021 (Spotlight)

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

• "State-of-art minibatches via novel DPP kernels: discretization, wavelets, and rough objectives"
H.S. Tran, P. Gupta, R. Bardenet, S.Ghosh
Arxiv preprint

• "Fast determinantal sampling on general spaces and diffusion geometry"
H.S. Tran, P. Gupta, S. Ghosh
Arxiv preprint