The *unreasonable effectiveness* of determinantal processes

Subhro Ghosh
National University of Singapore

Unreasonable Effectiveness

- E.P. Wigner, *The Unreasonable Effectiveness of Mathematics in the Natural Sciences*, Communications on Pure and Applied Mathematics, 1960; 13:001-14.

Unreasonable Effectiveness

- E.P. Wigner, *The Unreasonable Effectiveness of Mathematics in the Natural Sciences*, Communications on Pure and Applied Mathematics, 1960; 13:001-14.
- Contends that mathematical concepts have applicability that is often far beyond the context in which they were originally developed.

Unreasonable Effectiveness

- "The miracle of .. the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve. We should be grateful for it and hope that ... it will extend, .. to our pleasure, .. to wide branches of learning."

Constrained systems : I.I.D. and beyond

- The most popular model of randomness in science is perhaps that of Independent and Identically Distributed (I.I.D.) random variables.

Constrained systems : I.I.D. and beyond

- The most popular model of randomness in science is perhaps that of Independent and Identically Distributed (I.I.D.) random variables.
- The I.I.D. paradigm has led to ground breaking progress and a vast body of literature, including many ideas and methodologies that have become second nature for applications.

Constrained systems : I.I.D. and beyond

- The most popular model of randomness in science is perhaps that of Independent and Identically Distributed (I.I.D.) random variables.
- The I.I.D. paradigm has led to ground breaking progress and a vast body of literature, including many ideas and methodologies that have become second nature for applications.
- E.g.s include, but are not limited to, the fundamental theories behind Principal Component Analysis (P.C.A.) and other dimension reduction techniques, Maximum Likelihood based methods (M.L.E.), a wide array of information-theoretic approaches, and so on.

Constrained systems : I.I.D. and beyond

- Though there is scientific interest in exploring beyond the I.I.D. paradigm, progress mostly involves specific structures with limited or 'localised' dependence - e.g., various kinds of Markov processes. Also, some results in the direction of independent but not identically distributed random fields.

Constrained systems : I.I.D. and beyond

- Though there is scientific interest in exploring beyond the I.I.D. paradigm, progress mostly involves specific structures with limited or 'localised' dependence - e.g., various kinds of Markov processes. Also, some results in the direction of independent but not identically distributed random fields.
- In general, lack of independence is largely believed to be an obstacle or a hindrance to overcome, and many approaches involve trying to 'locate' independence or approximate independence in the overall dependency structure.

Constrained systems : I.I.D. and beyond

- Though there is scientific interest in exploring beyond the I.I.D. paradigm, progress mostly involves specific structures with limited or 'localised' dependence - e.g., various kinds of Markov processes. Also, some results in the direction of independent but not identically distributed random fields.
- In general, lack of independence is largely believed to be an obstacle or a hindrance to overcome, and many approaches involve trying to 'locate' independence or approximate independence in the overall dependency structure.
- In this talk, we will take a different point of view - namely, try to exploit dependence structures in stochastic systems in order to make substantive progress in fundamental learning problems.

Constrained Stochastic systems: some natural models

- In this talk, we will focus on a significant class of natural strongly dependent random systems, known as *Determinantal Processes* or DPPs.

Constrained Stochastic systems: some natural models

- In this talk, we will focus on a significant class of natural strongly dependent random systems, known as *Determinantal Processes* or DPPs.
- A DPP is a random set of points that all interact with each other, and where the interaction is encoded by a kernel.

Constrained Stochastic systems: some natural models

- In this talk, we will focus on a significant class of natural strongly dependent random systems, known as *Determinantal Processes* or DPPs.
- 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 random point sets.

Constrained Stochastic systems: some natural models

- DPPs are well-motivated by their origins in quantum and statistical physics.

Constrained Stochastic systems: some natural models

- DPPs are well-motivated by their origins in quantum and statistical physics.
- DPP strtucture arises natrually, e.g. as *Slater determinants* in wave-functions for Fermions (following earlier work by Heisenberg and Dirac)

Constrained Stochastic systems: some natural models

- DPPs are well-motivated by their origins in quantum and statistical physics.
- DPP strtucture arises natrually, e.g. 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, interacting particle systems ...

Determinantal processes

- Any model of a random point set is characterised by its 'correlation functions', which are essentially the joint probabilities of having points at specified locations

Determinantal processes

- Any model of 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.

Determinantal processes

- Any model of 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., if the model of random point set is to pick points independently and uniformly at random from a domain, then $\rho_m(\alpha_1, \ldots, \alpha_m) = \rho^m$ where ρ is the mean density of points per unit area.

Determinantal processes

- Determinantal processes are models of random point sets that are parameterised by a kernel function K.
- The m-point correlation functions are given by determinants :

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

- Clearly, if α_i and α_j are the close to each other (in some feature space) for different i and j, then under mild continuity assumption on the kernel K, the probability density ρ_m is very close to 0.

Determinantal processes

- Determinantal processes are models of random point sets that are parameterised by a kernel function K.
- The m-point correlation functions are given by determinants :

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

- Clearly, if α_i and α_j are the close to each other (in some feature space) for different i and j, then under mild continuity assumption on the kernel K, the probability density ρ_m is very close to 0.
- Thus, a DPP penalizes points for getting too close to each other, and therefore naturally embodies repulsive interaction between the points, albeit in a highly non-linear and complex manner.

Determinantal Processes

- DPPs are, therefore, effective in modelling situations where the sample points need to be very different from each other.
- E.g., in diversity sampling, the population may be represented by points in some (high dimensional) feature space, and the kernel K incorporates the proximity between these points in the feature space, which in turn encodes the 'similarity' between different points that we want to sample from.

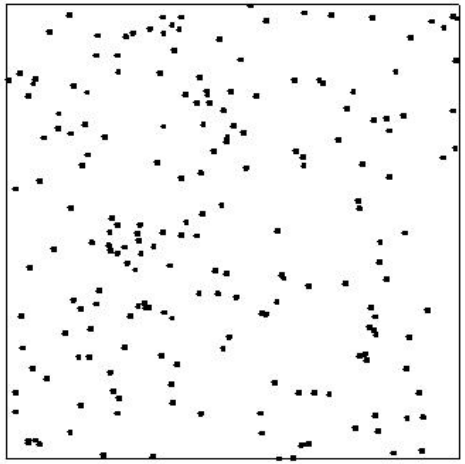
Determinantal Processes

- DPPs are, therefore, effective in modelling situations where the sample points need to be very different from each other.
- E.g., in diversity sampling, the population may be represented by points in some (high dimensional) feature space, and the kernel K incorporates the proximity between these points in the feature space, which in turn encodes the 'similarity' between different points that we want to sample from.
- Recently, DPPs have emerged as a fundamental component of a rapidly developing learning toolbox based on negative dependence

Determinantal Processes

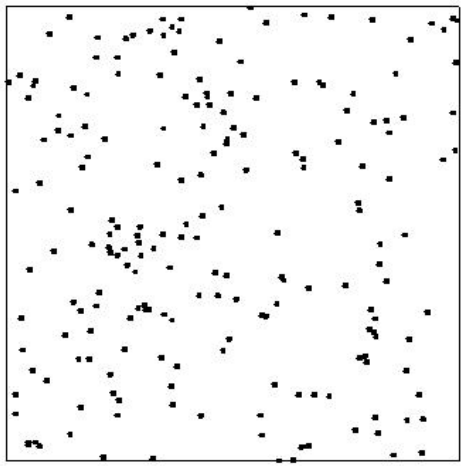
- DPPs are, therefore, effective in modelling situations where the sample points need to be very different from each other.
- E.g., in diversity sampling, the population may be represented by points in some (high dimensional) feature space, and the kernel K incorporates the proximity between these points in the feature space, which in turn encodes the 'similarity' between different points that we want to sample from.
- Recently, DPPs have emerged as a fundamental component of a rapidly developing learning toolbox based on negative dependence that, in many applications, shines over state-of-the-art methods based on statistical independence.

Simulations

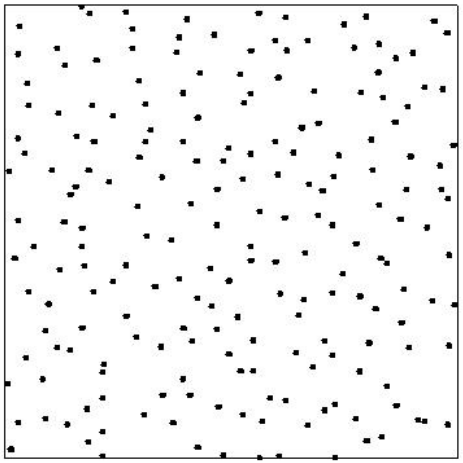


I.I.D.

Simulations



I.I.D.



Determinantal

Gaussian Determinantal Processes

- In the analysis of DPPs, 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.

Gaussian Determinantal Processes

- In the analysis of DPPs, 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.
- To this end, we propose the model of **Gaussian Determinantal Process** (GDP), that will be indexed by the space of positive definite matrices of a given dimension, which we will call the *scattering matrix*.

Gaussian Determinantal Processes

- In the analysis of DPPs, 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.
- To this end, we propose the model of **Gaussian Determinantal Process** (GDP), that will be indexed by the space of positive definite matrices of a given dimension, which we will call the *scattering matrix*.
- This would in turn be a 'testing ground' to understand the response of the spatial behaviour of the point process to parameter modulations in the space of scattering matrices.

Gaussian Determinantal Processes

- In the analysis of DPPs, 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.
- To this end, we propose the model of **Gaussian Determinantal Process** (GDP), that will be indexed by the space of positive definite matrices of a given dimension, which we will call the *scattering matrix*.
- This would in turn be a 'testing ground' to understand the response of the spatial behaviour of the point process to parameter modulations in the space of scattering matrices.
- Connection to *Spiked Models* of random matrices and *Spiked* PCA.

Gaussian Determinantal Processes

- In the analysis of DPPs, 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.
- To this end, we propose the model of **Gaussian Determinantal Process** (GDP), that will be indexed by the space of positive definite matrices of a given dimension, which we will call the *scattering matrix*.
- This would in turn be a 'testing ground' to understand the response of the spatial behaviour of the point process to parameter modulations in the space of scattering matrices.
- Connection to *Spiked Models* of random matrices and *Spiked* PCA.
- Based on joint work with P. Rigollet.

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

- Let $B(t)$ denotes the Euclidean ball of radius t in $\mathbb{R}^d$, and $|B(t)|$ be its volume. Let $\{X_1, \ldots, X_n\}$ be the observed data points. Let $r > 0$ be a threshold, to be detailed later.

Parameter estimation in GDP

- Let $B(t)$ denotes the Euclidean ball of radius t in $\mathbb{R}^d$, and $|B(t)|$ be its volume. Let $\{X_1, \ldots, X_n\}$ be the observed data points. Let $r > 0$ be a threshold, to be detailed later.
- Then

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

is a consistent estimator of Σ.

Parameter estimation in GDP

- Let $B(t)$ denotes the Euclidean ball of radius t in $\mathbb{R}^d$, and $|B(t)|$ be its volume. Let $\{X_1, \ldots, X_n\}$ be the observed data points. Let $r > 0$ be a threshold, to be detailed later.
- Then

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

  is a consistent estimator of Σ.
- Bias variance tradeoff leads to optimal choice of $r = \Theta(\sqrt{d \log n})$, as $n \to \infty$.

Testing for directionality in GDP

- We want to test for directional bias in the dependency structure of the observed data points, against a null hypothesis of isotropic dependence.
- In terms of the spiked model, this is equivalent to testing for the presence of a spike.

Testing for directionality in GDP

- We want to test for directional bias in the dependency structure of the observed data points, against a null hypothesis of isotropic dependence.
- In terms of the spiked model, this is equivalent to testing for the presence of a spike.

Theorem (G.-Rigollet)

Based on the test statistic $\|\hat{\Sigma}\|_{\mathrm{op}}$, *we can detect the spike with high probability if the spike size* λ *is above the threshold*

$$\lambda \gtrsim d^2 \log R \left(\frac{c\sqrt{d} \log R}{R} \right)^{d/2}.$$

Testing for directionality in GDP

- We want to test for directional bias in the dependency structure of the observed data points, against a null hypothesis of isotropic dependence.
- In terms of the spiked model, this is equivalent to testing for the presence of a spike.

Theorem (G.-Rigollet)

Based on the test statistic $\|\hat{\Sigma}\|_{\mathrm{op}}$, we can detect the spike with high probability if the spike size λ is above the threshold

$$\lambda \gtrsim d^2 \log R \left(\frac{c\sqrt{d} \log R}{R} \right)^{d/2}.$$

The leading eigenvector of $\hat{\Sigma}$ is a consistent estimator of the direction of the spike.

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.
- It has led to significant methodological progress,

Dimension Reduction and Directionality in Data

- The problem of dimension reduction is one of the central problems in the applied mathematics.
- It has led to significant methodological progress, e.g. Principal Component Analysis and its derivatives, the entire suite of methods involving the Johnson-Lindenstrauss Lemma and related random projection based techniques, and so on.

Dimension Reduction and Directionality in Data

- The problem of dimension reduction is one of the central problems in the applied mathematics.
- It has led to significant methodological progress, e.g. Principal Component Analysis and its derivatives, the entire suite of methods involving the Johnson-Lindenstrauss Lemma and related random projection based techniques, and so on.
- 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.). This is going to be one of the directions of focus in our talk.

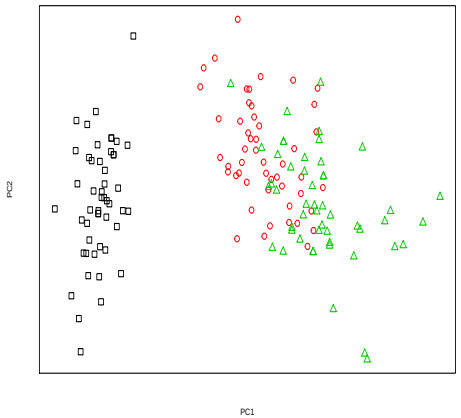
Dimension Reduction

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

Dimension Reduction

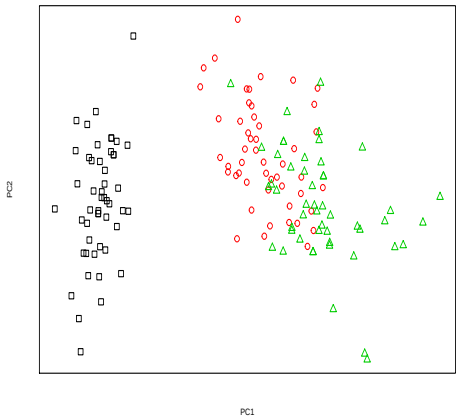
- We use the GDP model as an *ansatz* for proposing a dimension reduction methodology.
- We may compute the quantity $\hat{\Sigma}$ for any observed data set in $\mathbb{R}^d$. We then perform SVD on $\hat{\Sigma}$ and project the data points on to the principal eigen-directions of $\hat{\Sigma}$ in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

Dimension Reduction : Fisher's Iris

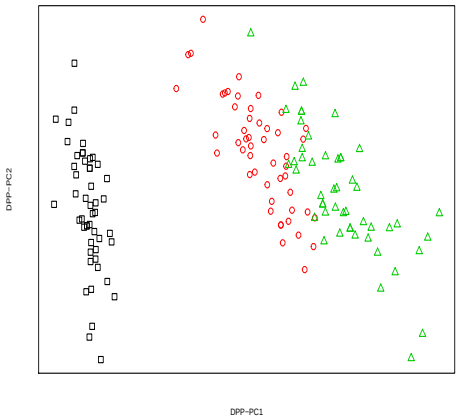


PCA

Dimension Reduction : Fisher's Iris

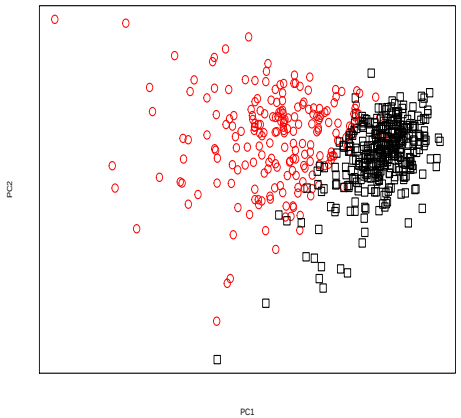


PCA



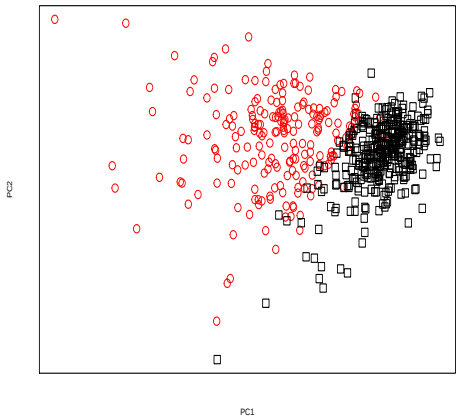
GDP

Dimension Reduction : Wisconsin Breast Cancer

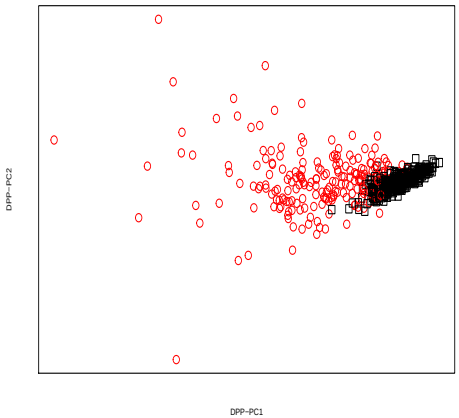


PCA

Dimension Reduction : Wisconsin Breast Cancer



PCA



GDP

Stochastic Gradient Descent and DPP

- Stochastic gradient descent (SGD) is a cornerstone of modern machine learning.

Stochastic Gradient Descent and DPP

- Stochastic gradient descent (SGD) is a cornerstone of modern machine learning.
- In large datasets, SGD relies on constructing an unbiased estimator of the gradient using a small subset of the original dataset, called a minibatch.

Stochastic Gradient Descent and DPP

- Stochastic gradient descent (SGD) is a cornerstone of modern machine learning.
- In large datasets, SGD relies on constructing an unbiased estimator of the gradient using a small subset of the original dataset, called a minibatch.
- Default minibatch construction involves uniformly sampling a subset of the desired size.

- We contribute an *orthogonal polynomial*-based DPP paradigm for minibatch sampling in SGD, and substantiate it with a robust theoretical foundation.

Stoch. Grad. Descent, Orthogonal Polynomials and DPP

- We contribute an *orthogonal polynomial*-based DPP paradigm for minibatch sampling in SGD, and substantiate it with a robust theoretical foundation.
- Our approach *leverages the specific data distribution* at hand, which endows it with greater sensitivity and power over existing data-agnostic methods.

Stoch. Grad. Descent, Orthogonal Polynomials and DPP

- We contribute an *orthogonal polynomial*-based DPP paradigm for minibatch sampling in SGD, and substantiate it with a robust theoretical foundation.
- Our approach *leverages the specific data distribution* at hand, which endows it with greater sensitivity and power over existing data-agnostic methods.
- Joint work with R. Bardenet and M. Lin.

- We obtain a detailed theoretical analysis of its convergence properties, *interweaving between the discrete data set and the underlying continuous domain*.
- We propose *gradient estimators* whose *variance decays provably faster with the batchsize* than under uniform sampling.
- For a large enough batchsize and a fixed budget, DPP minibatches lead to a *smaller bound on the mean square approximation error* than uniform minibatches.

The SGD minibatch problem

- Stochastic Gradient Descent (SGD) is the workhorse of modern machine learning
- Useful in a wide array of optimization scenario, ranging from maximum likelihood problems of parametric statistics to back propagation in training deep neural networks, and beyond ....

The SGD minibatch problem

- The fundamental step in a gradient descent approach 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

The SGD minibatch problem

- The fundamental step in a gradient descent approach 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 $\frac{1}{N} \cdot \sum_{i=1}^{N} \nabla_\theta \mathcal{L}(z_i, \theta)$
- $\widehat{\mathcal{L}(A, \theta_t)}$ is *based on a relatively small subsample $A \subset \mathcal{D}$ of the full data set $\mathcal{D}$.*

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 $\frac{1}{N} \cdot \sum_{i=1}^{N} \nabla_\theta \mathcal{L}(z_i, \theta)$
- $\widehat{\mathcal{L}(A, \theta_t)}$ is *based on a relatively small subsample $A \subset \mathcal{D}$ of the full data set $\mathcal{D}$.*
- Such a subsample A is called a minibatch.

The SGD minibatch problem

- A minibatch is a (random) subset $A \subset \mathcal{D}$ 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 $\Xi_N = \frac{1}{N} \sum_{z_i \in \mathcal{D}} \nabla_\theta \mathcal{L}(z_i, \theta)$.

The SGD minibatch problem

- A minibatch is a (random) subset $A \subset \mathcal{D}$ 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 $\Xi_N = \frac{1}{N} \sum_{z_i \in \mathcal{D}} \nabla_\theta \mathcal{L}(z_i, \theta)$.
- Natural problem : How to select the *random* subset $A \subset \mathcal{D}$? 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 p.

The default choice : select A uniformly at random

- The default choice in minibatch selection is to sample A to be uniformly at random, and take the usual empirical average on A to construct Ξ_A.

The default choice : select A uniformly at random

- The default choice in minibatch selection is to sample A to be uniformly at random, and take the usual empirical average on A to construct Ξ_A.
- Sampling p data points with / without replacement, Poissonian sampling (select each data point to be in A independently with probability p/N) ...

The default choice : select A uniformly at random

- The default choice in minibatch selection is to sample A to be uniformly at random, and take the usual empirical average on A to construct Ξ_A.
- Sampling p data points with / without replacement, Poissonian sampling (select each data point to be in A independently with probability p/N) ...
- Unbiased estimate of the variance

$$\sigma^2(\Xi_{\text{Unif}}) = O_P(p^{-1}).$$

Enter DPPs

- Our approach : select A according to a DPP to effect variance reduction.

Enter DPPs

- Our approach : select A according to a DPP to effect variance reduction.
- Our DPP sampler will be tailored to the data distribution, implemented via Orthogonal Polynomials.

Multivariate Orthogonal Polynomials

- For definiteness, we consider our domain on which the data points live to be $[-1, 1]^d$.

Multivariate Orthogonal Polynomials

- For definiteness, we consider our domain on which the data points live to be $[-1,1]^d$.
- Multivariate Orthogonal polynomials : consider a reference measure $q(x)dx$ on $[-1,1]^d$.
- Consider the monomial functions $(x_1,\ldots,x_d) \mapsto x_1^{\alpha_1}\cdots x_d^{\alpha_d}$ in the graded lexical order.

Multivariate Orthogonal Polynomials

- For definiteness, we consider our domain on which the data points live to be $[-1,1]^d$.
- Multivariate Orthogonal polynomials : consider a reference measure $q(x)dx$ on $[-1,1]^d$.
- 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 $L^2(q(x)\,\mathrm{d}x)$ to these ordered monomials.

Multivariate Orthogonal Polynomials

- For definiteness, we consider our domain on which the data points live to be $[-1,1]^d$.
- Multivariate Orthogonal polynomials : consider a reference measure $q(x)dx$ on $[-1,1]^d$.
- 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 $L^2(q(x)\,\mathrm{d}x)$ 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. q.

DPPs based on Multivariate Orthogonal Polynomials

- Construct a DPP with the kernel given by the projection

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

with respect to the background measure $q(x)dx$.

DPPs based on Multivariate Orthogonal Polynomials

- Construct a DPP with the kernel given by the projection

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

  with respect to the background measure $q(x)dx$.
- We obtain a projection DPP with kernel denoted as $K_q^{(p)}$, referred to as the *Multivariate OPE* (i.e., Multivariate Orthogonal Polynomial Ensemble) associated with the measure $q(x)dx$.

DPPs based on Multivariate Orthogonal Polynomials

- Construct a DPP with the kernel given by the projection

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

  with respect to the background measure $q(x)dx$.
- We obtain a projection DPP with kernel denoted as $K_q^{(p)}$, referred to as the *Multivariate OPE* (i.e., Multivariate Orthogonal Polynomial Ensemble) associated with the measure $q(x)dx$.
- # of sampled points = rank of the projection = p (always !)

An OP based minibatch selector for SGD

- Let $\tilde{\gamma}(z) = \frac{1}{Nh^d} \sum_{i=1}^{N} k\left(\frac{z-z_i}{h}\right)$ be a kernel density estimator of the pdf of the data-generating distribution γ with window size $h > 0$.
- Let $q(x) = q_1(x_1) \dots q_d(x_d)$ be a separable pdf on $[-1, 1]^d$, where each q_i is positive on $[-1, 1]$.

An OP based minibatch selector for SGD

- Let $\tilde{\gamma}(z) = \frac{1}{Nh^d} \sum_{i=1}^{N} k\left(\frac{z-z_i}{h}\right)$ be a kernel density estimator of the pdf of the data-generating distribution γ with window size $h > 0$.
- Let $q(x) = q_1(x_1) \dots q_d(x_d)$ be a separable pdf on $[-1, 1]^d$, where each q_i is positive on $[-1, 1]$.
- Let $K_q^{(p)}$ be the Multivariate OPE kernel with respect to the measure q.

A DPP based minibatch selector for SGD

- Define a new kernel that factors in $\tilde{\gamma}$:

$$K^{(p)}_{q,\tilde{\gamma}}(x,y) := \sqrt{\frac{q(x)}{\tilde{\gamma}(x)}} K^{(p)}_{q}(x,y) \sqrt{\frac{q(y)}{\tilde{\gamma}(y)}}$$

A DPP based minibatch selector for SGD

- Define a new kernel that factors in $\tilde{\gamma}$:

$$K^{(p)}_{q,\tilde{\gamma}}(x,y) := \sqrt{\frac{q(x)}{\tilde{\gamma}(x)}} K^{(p)}_q(x,y) \sqrt{\frac{q(y)}{\tilde{\gamma}(y)}}$$

- Both $K^{(p)}_q, K^{(p)}_{q,\tilde{\gamma}}$ are projection DPPs on the space $[-1,1]^d$.

A DPP based minibatch selector for SGD

- Define a new kernel that factors in $\tilde{\gamma}$:

$$K^{(p)}_{q,\tilde{\gamma}}(x,y) := \sqrt{\frac{q(x)}{\tilde{\gamma}(x)}} K^{(p)}_q(x,y) \sqrt{\frac{q(y)}{\tilde{\gamma}(y)}}$$

- Both $K^{(p)}_q, K^{(p)}_{q,\tilde{\gamma}}$ are projection DPPs on the space $[-1,1]^d$. Need : a kernel on $\mathcal{D} \subset [-1,1]^d$.

A DPP based minibatch selector for SGD

- Define a new kernel that factors in $\tilde{\gamma}$:

$$K^{(p)}_{q,\tilde{\gamma}}(x,y) := \sqrt{\frac{q(x)}{\tilde{\gamma}(x)}} K^{(p)}_q(x,y) \sqrt{\frac{q(y)}{\tilde{\gamma}(y)}}$$

- Both $K^{(p)}_q, K^{(p)}_{q,\tilde{\gamma}}$ are projection DPPs on the space $[-1,1]^d$. Need : a kernel on $\mathcal{D} \subset [-1,1]^d$.
- Idea : Simply restrict $K^{(p)}_{q,\tilde{\gamma}}$ to $\mathcal{D}$!

A DPP based minibatch selector for SGD

- Define a new kernel that factors in $\tilde{\gamma}$:

$$K^{(p)}_{q,\tilde{\gamma}}(x,y) := \sqrt{\frac{q(x)}{\tilde{\gamma}(x)}} K^{(p)}_q(x,y) \sqrt{\frac{q(y)}{\tilde{\gamma}(y)}}$$

- Both $K^{(p)}_q, K^{(p)}_{q,\tilde{\gamma}}$ are projection DPPs on the space $[-1,1]^d$. Need : a kernel on $\mathcal{D} \subset [-1,1]^d$.
- Idea : Simply restrict $K^{(p)}_{q,\tilde{\gamma}}$ to $\mathcal{D}$! Problem : No longer a projection (important for variance reduction purposes)

A DPP based minibatch selector for SGD

- Define a new kernel that factors in $\tilde{\gamma}$:

$$K^{(p)}_{q,\tilde{\gamma}}(x,y) := \sqrt{\frac{q(x)}{\tilde{\gamma}(x)}} K^{(p)}_q(x,y) \sqrt{\frac{q(y)}{\tilde{\gamma}(y)}}$$

- Both $K^{(p)}_q, K^{(p)}_{q,\tilde{\gamma}}$ are projection DPPs on the space $[-1,1]^d$. Need : a kernel on $\mathcal{D} \subset [-1,1]^d$.
- Idea : Simply restrict $K^{(p)}_{q,\tilde{\gamma}}$ to $\mathcal{D}$! Problem : No longer a projection (important for variance reduction purposes)
- However : $K^{(p)}_{q,\tilde{\gamma}}$ is *approximately* a rank-p projection with respect to the uniform distribution on $\mathcal{D}$

A DPP based minibatch selector for SGD

- Define a new kernel that factors in $\tilde{\gamma}$:

$$K^{(p)}_{q,\tilde{\gamma}}(x,y) := \sqrt{\frac{q(x)}{\tilde{\gamma}(x)}} K^{(p)}_q(x,y) \sqrt{\frac{q(y)}{\tilde{\gamma}(y)}}$$

- Both $K^{(p)}_q, K^{(p)}_{q,\tilde{\gamma}}$ are projection DPPs on the space $[-1,1]^d$. Need : a kernel on $\mathcal{D} \subset [-1,1]^d$.
- Idea : Simply restrict $K^{(p)}_{q,\tilde{\gamma}}$ to $\mathcal{D}$! Problem : No longer a projection (important for variance reduction purposes)
- However : $K^{(p)}_{q,\tilde{\gamma}}$ is *approximately* a rank-p projection with respect to the uniform distribution on $\mathcal{D}$
- Solution : Spectrally round-off $K^{(p)}_{q,\tilde{\gamma}}$ to a rank-p projection $\widetilde{\mathrm{K}}$.

A DPP based gradient estimator for SGD

- We consider a minibatch $A \sim \mathrm{DPP}(\widetilde{\mathrm{K}}, \hat{\gamma}_N)$, where $\widetilde{\mathrm{K}}$ is the projection as obtained above and the background measure $\hat{\gamma}_N$ is the uniform distribution on $\mathcal{D}$.

A DPP based gradient estimator for SGD

- We consider a minibatch $A \sim \mathrm{DPP}(\widetilde{\mathrm{K}}, \hat{\gamma}_N)$, where $\widetilde{\mathrm{K}}$ is the projection as obtained above and the background measure $\hat{\gamma}_N$ is the uniform distribution on $\mathcal{D}$.
- Since $(\widetilde{\mathrm{K}}, \hat{\gamma}_N)$ is a projection kernel, $|A| = p$ almost surely.

A DPP based gradient estimator for SGD

- We consider a minibatch $A \sim \mathrm{DPP}(\widetilde{\mathsf{K}}, \hat{\gamma}_N)$, where $\widetilde{\mathsf{K}}$ is the projection as obtained above and the background measure $\hat{\gamma}_N$ is the uniform distribution on $\mathcal{D}$.
- Since $(\widetilde{\mathsf{K}}, \hat{\gamma}_N)$ is a projection kernel, $|A| = p$ almost surely.
- Define the gradient estimator

$$\Xi_{A,\mathrm{DPP}} := \sum_{z_i \in A} \frac{\nabla_\theta \mathcal{L}(z_i, \theta)}{\widetilde{\mathsf{K}}(z_i, z_i)}.$$

A DPP based gradient estimator for SGD

- We consider a minibatch $A \sim \mathrm{DPP}(\widetilde{\mathsf{K}}, \hat{\gamma}_N)$, where $\widetilde{\mathsf{K}}$ is the projection as obtained above and the background measure $\hat{\gamma}_N$ is the uniform distribution on $\mathcal{D}$.
- Since $(\widetilde{\mathsf{K}}, \hat{\gamma}_N)$ is a projection kernel, $|A| = p$ almost surely.
- Define the gradient estimator

$$\Xi_{A,\mathrm{DPP}} := \sum_{z_i \in A} \frac{\nabla_\theta \mathcal{L}(z_i, \theta)}{\widetilde{\mathsf{K}}(z_i, z_i)}.$$

- $\mathbb{E}[\Xi_{A,\mathrm{DPP}}] = \sum_{i=1}^N \left(\frac{\nabla_\theta \mathcal{L}(z_i,\theta)}{\widetilde{\mathsf{K}}(z_i,z_i)} \right) \cdot \widetilde{\mathsf{K}}(z_i, z_i) \cdot \frac{1}{N} = \frac{1}{N} \sum_{i=1}^N \nabla_\theta \mathcal{L}(z_i, \theta)$
- So, $\Xi_{A,\mathrm{DPP}}$ is unbiased, as desired.

Why does $\Xi_{A,\mathrm{DPP}}$ have reduced fluctuations ?

$$\sigma^2(\Xi_{A,\mathrm{DPP}})$$

$$= \frac{1}{N^2} \sum_{i,j} \left\| \frac{\nabla_\theta \mathcal{L}(z_i, \theta)}{\widetilde{K}(z_i, z_i)} - \frac{\nabla_\theta \mathcal{L}(z_j, \theta)}{\widetilde{K}(z_j, z_j)} \right\|_2^2 |\widetilde{K}(z_i, z_j)|^2 \quad \text{[Projection Kernel]}$$

$$\lesssim \mathcal{M}(\theta) \cdot \frac{1}{p^2} \int \int \|z - w\|_2^2 \, |K_q^{(p)}(z, w)|^2 dq(z) dq(w) \quad \text{[under regularity]}$$

Why does $\Xi_{A,\mathrm{DPP}}$ have reduced fluctuations ?

$$\sigma^2(\Xi_{A,\mathrm{DPP}})$$

$$= \frac{1}{N^2} \sum_{i,j} \left\| \frac{\nabla_\theta \mathcal{L}(z_i, \theta)}{\widetilde{K}(z_i, z_i)} - \frac{\nabla_\theta \mathcal{L}(z_j, \theta)}{\widetilde{K}(z_j, z_j)} \right\|_2^2 |\widetilde{K}(z_i, z_j)|^2 \quad \text{[Projection Kernel]}$$

$$\lesssim \mathcal{M}(\theta) \cdot \frac{1}{p^2} \int \int \|z - w\|_2^2 \, |K_q^{(p)}(z, w)|^2 dq(z) dq(w) \quad \text{[under regularity]}$$

- If $\|z - w\|_2^2$ was not present, then $\int\int |K_q^{(p)}(z, w)|^2 dq(z) dq(w) = p$ implies $\mathrm{Var} \lesssim 1/p$, which is the same as uniform random sampling of A.

Why does $\Xi_{A,\mathrm{DPP}}$ have reduced fluctuations ?

$$\sigma^2(\Xi_{A,\mathrm{DPP}})$$

$$= \frac{1}{N^2} \sum_{i,j} \left\| \frac{\nabla_\theta \mathcal{L}(z_i, \theta)}{\widetilde{K}(z_i, z_i)} - \frac{\nabla_\theta \mathcal{L}(z_j, \theta)}{\widetilde{K}(z_j, z_j)} \right\|_2^2 |\widetilde{K}(z_i, z_j)|^2 \quad \text{[Projection Kernel]}$$

$$\lesssim \mathcal{M}(\theta) \cdot \frac{1}{p^2} \int \int \|z - w\|_2^2 \, |K_q^{(p)}(z, w)|^2 dq(z) dq(w) \quad \text{[under regularity]}$$

- If $\|z - w\|_2^2$ was not present, then $\int\int |K_q^{(p)}(z, w)|^2 dq(z) dq(w) = p$ implies $\mathrm{Var} \lesssim 1/p$, which is the same as uniform random sampling of A.
- However, main contribution to $\int\int |K_q^{(p)}(z, w)|^2 dq(z) dq(w)$ comes from near the diagonal $z = w$, which is precisely suppressed by the term $\|z - w\|_2^2$.

Why does $\Xi_{A,\mathrm{DPP}}$ have reduced fluctuations ?

$$\sigma^2(\Xi_{A,\mathrm{DPP}})$$

$$= \frac{1}{N^2} \sum_{i,j} \left\| \frac{\nabla_\theta \mathcal{L}(z_i, \theta)}{\widetilde{K}(z_i, z_i)} - \frac{\nabla_\theta \mathcal{L}(z_j, \theta)}{\widetilde{K}(z_j, z_j)} \right\|_2^2 |\widetilde{K}(z_i, z_j)|^2 \quad \text{[Projection Kernel]}$$

$$\lesssim \mathcal{M}(\theta) \cdot \frac{1}{p^2} \int \int \|z - w\|_2^2 \, |K_q^{(p)}(z, w)|^2 dq(z) dq(w) \quad \text{[under regularity]}$$

- If $\|z - w\|_2^2$ was not present, then $\int\int |K_q^{(p)}(z, w)|^2 dq(z) dq(w) = p$ implies $\mathrm{Var} \lesssim 1/p$, which is the same as uniform random sampling of A.
- However, main contribution to $\int\int |K_q^{(p)}(z, w)|^2 dq(z) dq(w)$ comes from near the diagonal $z = w$, which is precisely suppressed by the term $\|z - w\|_2^2$.
- Use *Christoffel-Darboux formula* to make this control precise

Why does $\Xi_{A,\mathrm{DPP}}$ have reduced fluctuations ?

- Obtain

Theorem (Bardenet-G.-Lin)

$\mathrm{Var}[\Xi_{A,DPP}|\mathcal{D}] = O_P(p^{-(1+1/d)})$.

- Improvement in exponent of p compared to uniform random sampling of A !

Why does $\Xi_{A,\mathrm{DPP}}$ have reduced fluctuations ?

- Obtain

Theorem (Bardenet-G.-Lin)

$$\mathrm{Var}[\Xi_{A,DPP}|\mathcal{D}] = O_P(p^{-(1+1/d)}).$$

- Improvement in exponent of p compared to uniform random sampling of A !
- Near-completion (with R. Bardenet and M. Lin) : a technique for sampling the gradient estimator *directly*,

Why does $\Xi_{A,\mathrm{DPP}}$ have reduced fluctuations ?

- Obtain

Theorem (Bardenet-G.-Lin)

$\mathrm{Var}[\Xi_{A,DPP}|\mathcal{D}] = O_P(p^{-(1+1/d)})$.

- Improvement in exponent of p compared to uniform random sampling of A !
- Near-completion (with R. Bardenet and M. Lin) : a technique for sampling the gradient estimator *directly*, *without having to sample the DPP* minibatch.

Why does $\Xi_{A,\mathrm{DPP}}$ have reduced fluctuations ?

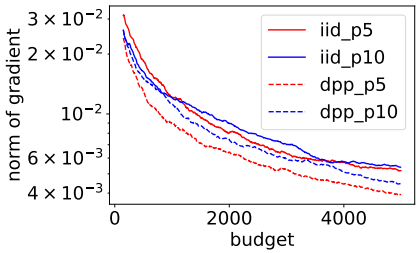
- Obtain

Theorem (Bardenet-G.-Lin)

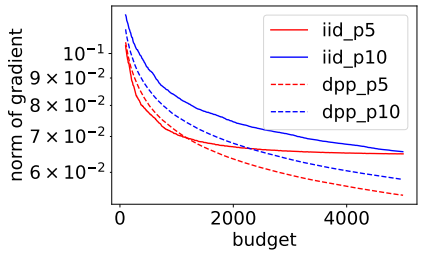
$\mathrm{Var}[\Xi_{A,DPP}|\mathcal{D}] = O_P(p^{-(1+1/d)})$.

- Improvement in exponent of p compared to uniform random sampling of A !
- Near-completion (with R. Bardenet and M. Lin) : a technique for sampling the gradient estimator *directly*, *without having to sample the DPP* minibatch. Applications to a wide array of DPP based approaches in machine learning (such as coresets), spatial statistics ....

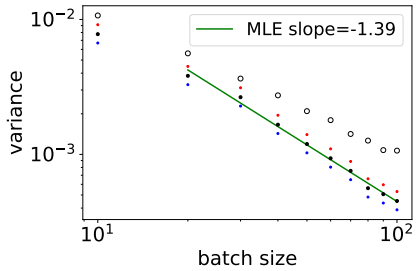
Performance in experiments



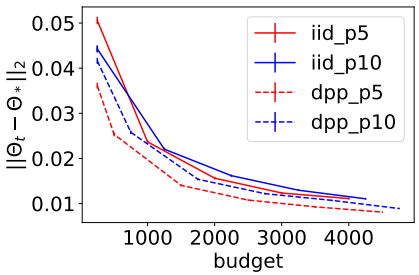
(a) d=3, Lin. Reg.



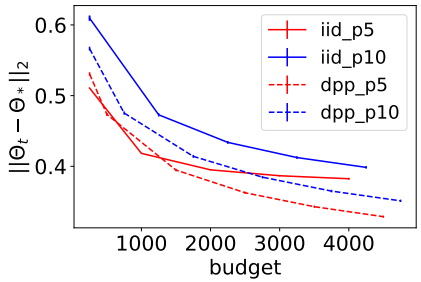
(b) d=11, Log. Reg.



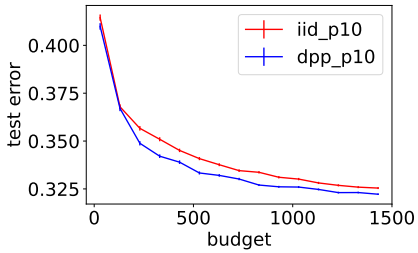
(c) d=3, Var vs log p



(d) d=3, Lin. Reg.



(e) d=11, Log. Reg.



(f) Letter. binary

Figure: Summary of the performance of two sampling strategies in SGD.

References

- Gaussian determinantal processes: A new model for directionality in data, with P. Rigollet,
  Proceedings of the National Academy of Sciences, vol. 117, no. 24 (2020), pp. 13207–13213
  (PNAS Direct Submission)
- Determinantal point processes based on orthogonal polynomials for sampling minibatches in SGD, with R. Bardenet and M. Lin
  Advances in Neural Information Processing Systems 34 (2021)
  (Spotlight at NeurIPS 2021)