Determinantal Point Processes

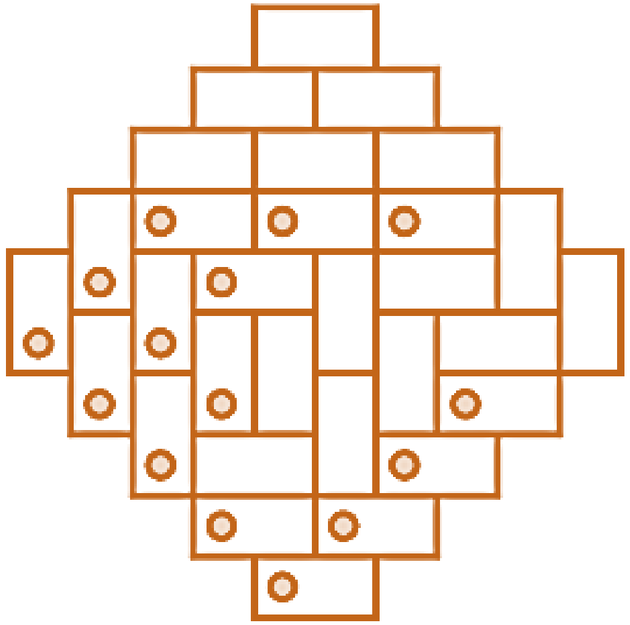
Jose Gallego-Posada

April 2021

sli.do -- #MilaDPP

Today's agenda

- Why DPPs?
- Definition and properties
- Sampling
- Applications

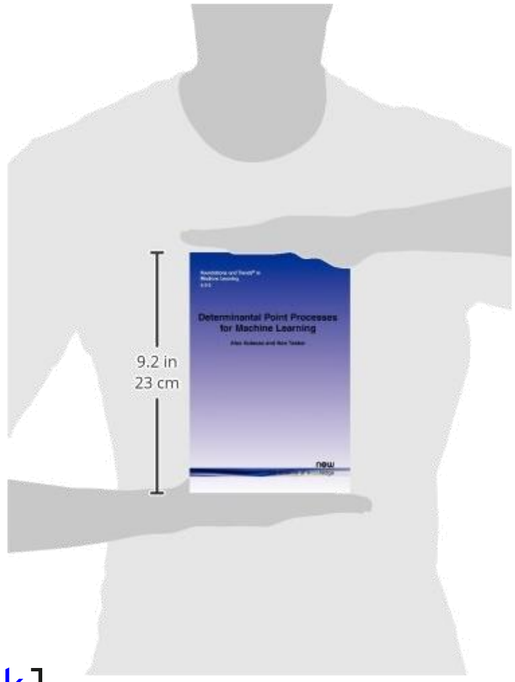


Bible for DPP in ML:

Foundations and Trends in Machine Learning

Determinantal Point Processes for Machine Learning

Alex Kulesza and Ben Taskar (2012) [link]



Presentation based on slides by :

- Simon Barthelmé, Nicolas Tremblay, EUSIPCO19 [link]
- Alex Kulesza, Ben Taskar and Jennifer Gillenwater – CVPR13 [link]

DPPy: Sampling Determinantal Point Processes with Python

"Anything that can go wrong, will go wrong" – Murphy's Law

Introduction

Determinantal point processes (DPPs) are specific probability distributions over clouds of points that have been popular as models or computational tools across physics, probability, statistics, and more recently of booming interest in machine learning. Sampling from DPPs is a nontrivial matter, and many approaches have been proposed. DPPy is a Python library that puts together all exact and approximate sampling algorithms for DPPs.

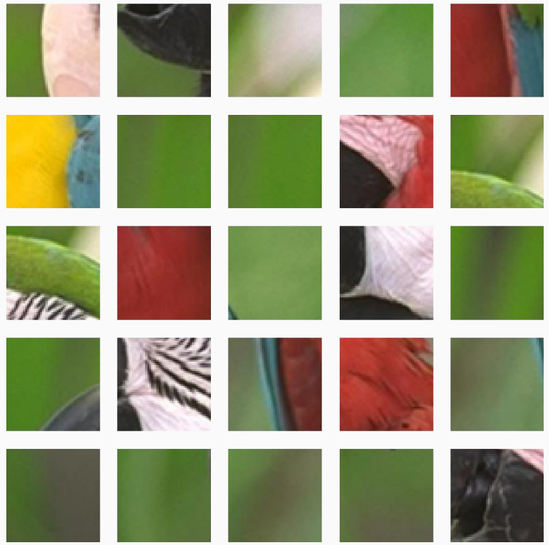


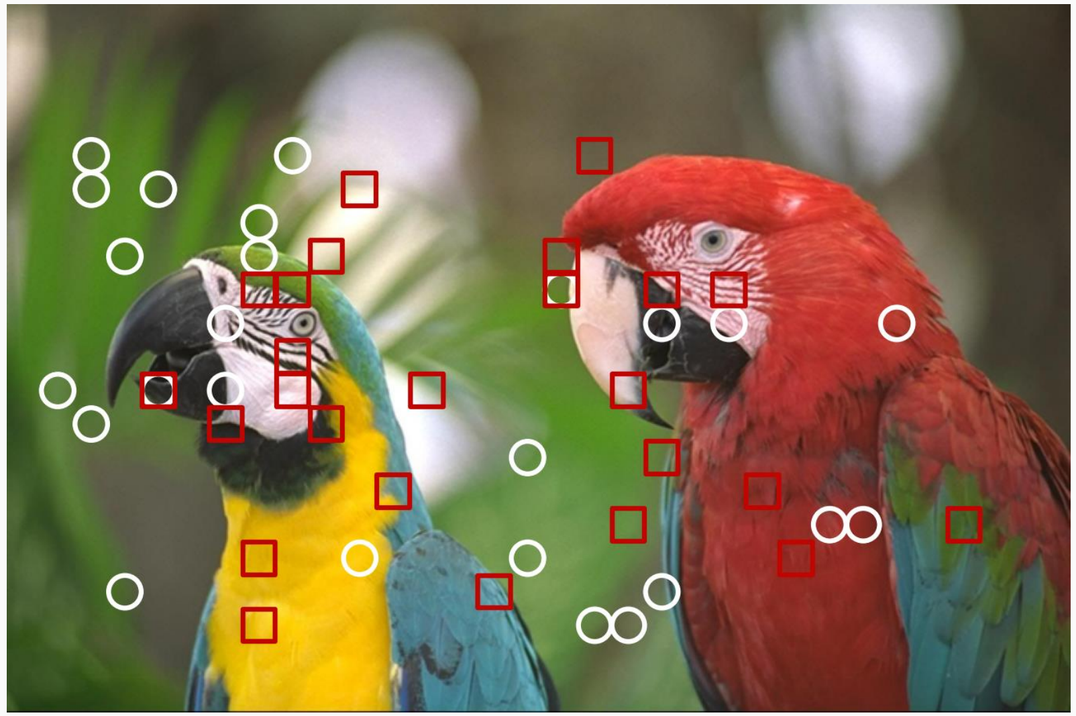


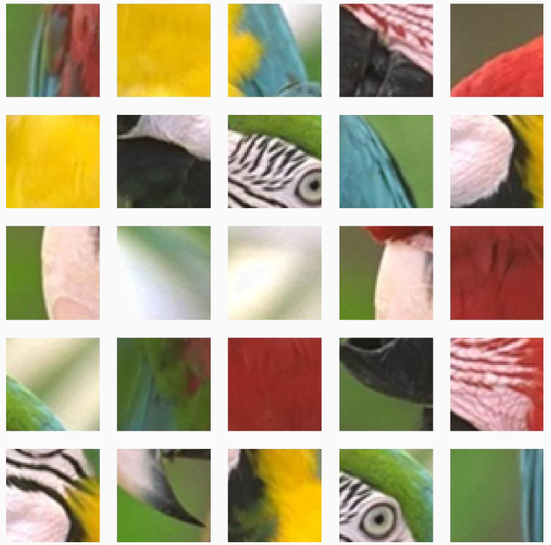
Guillaume Gautier, Rémi Bardenet, Guillermo Polito, Michal Valko

https://github.com/jgalle29/dpp_slides

BioDiversity

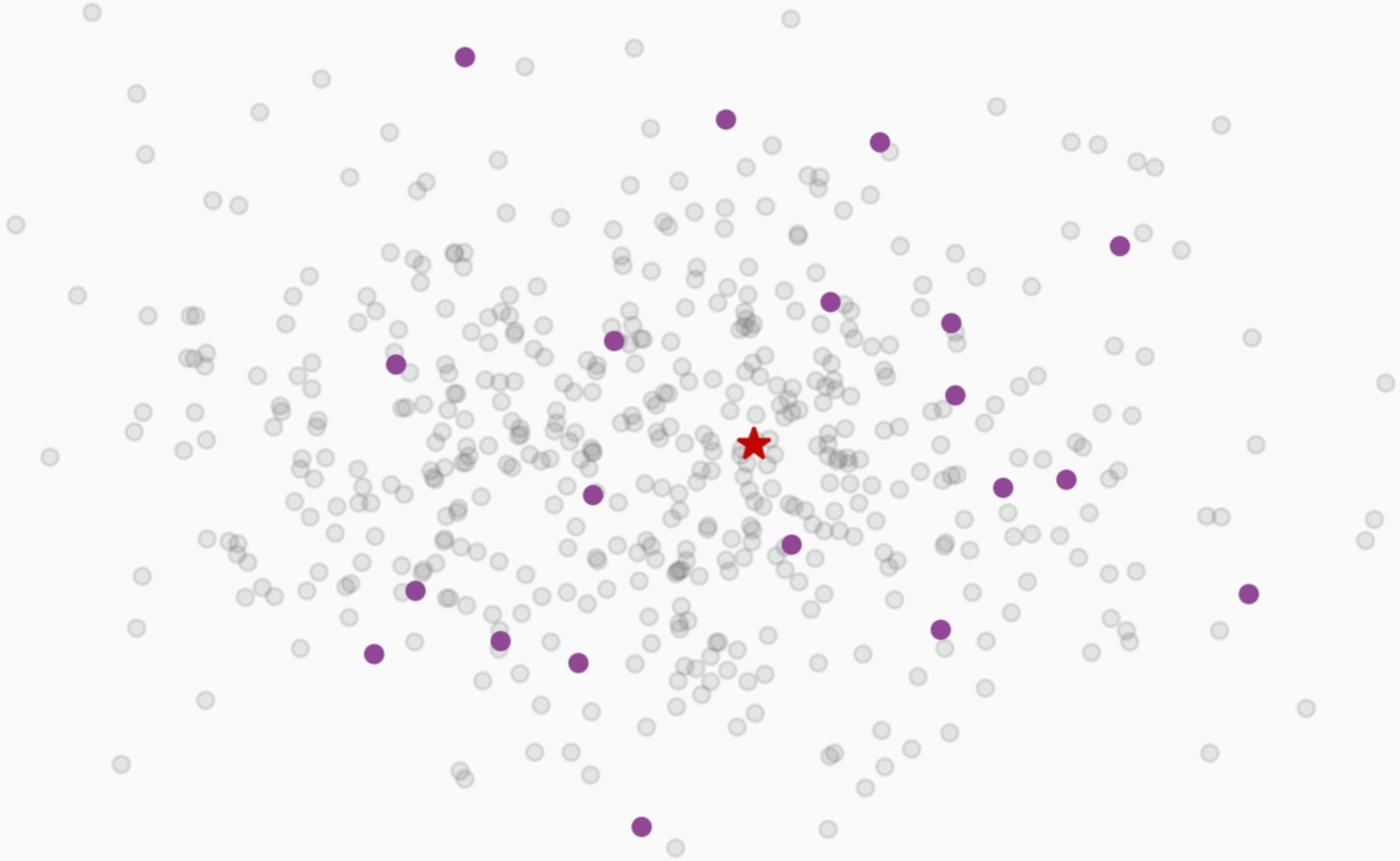






[BT19, dpp_demo]

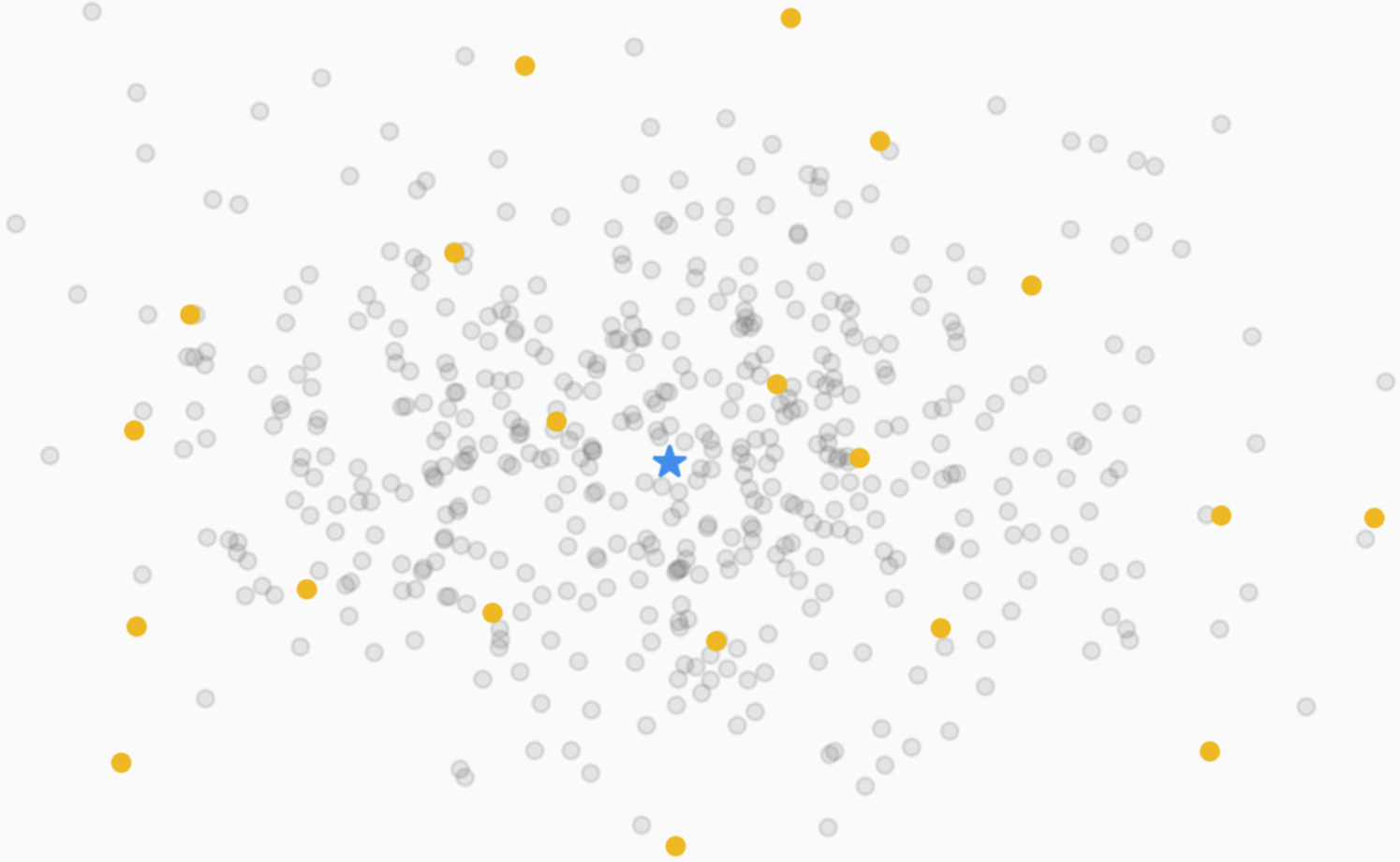
Variance reduction – Mean estimation



IID Samples

[BT19, dpp_demo]

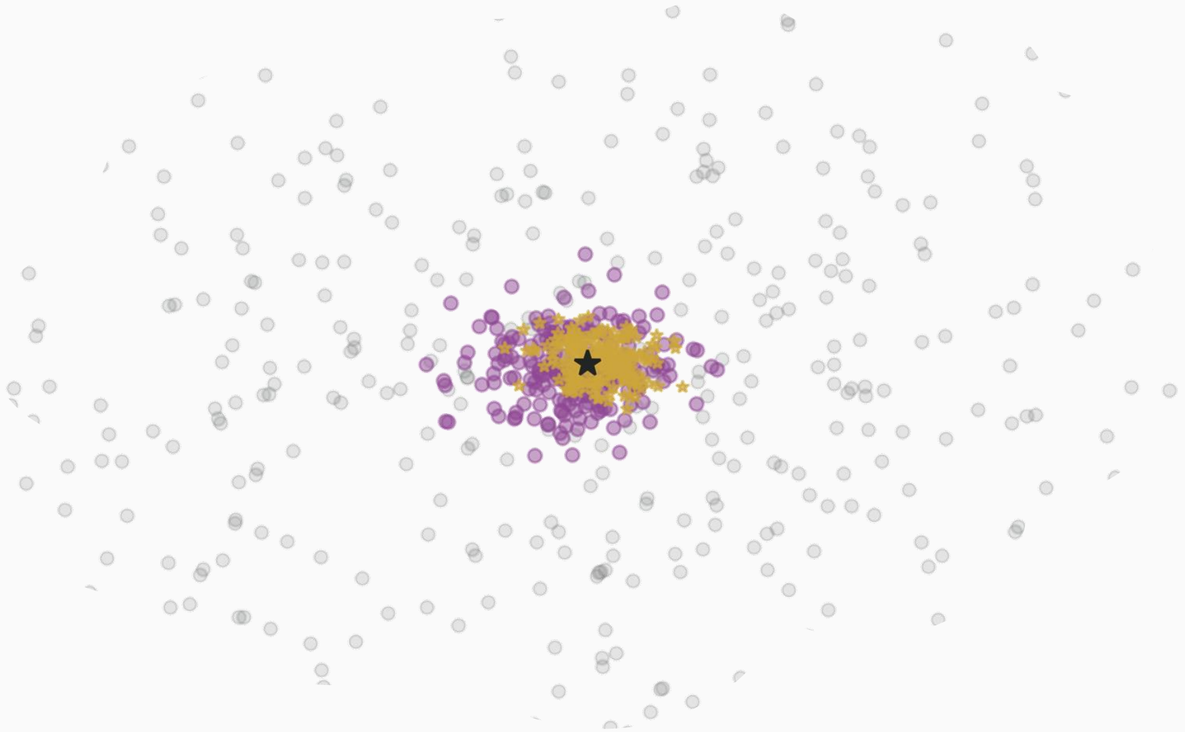
Variance reduction – Mean estimation

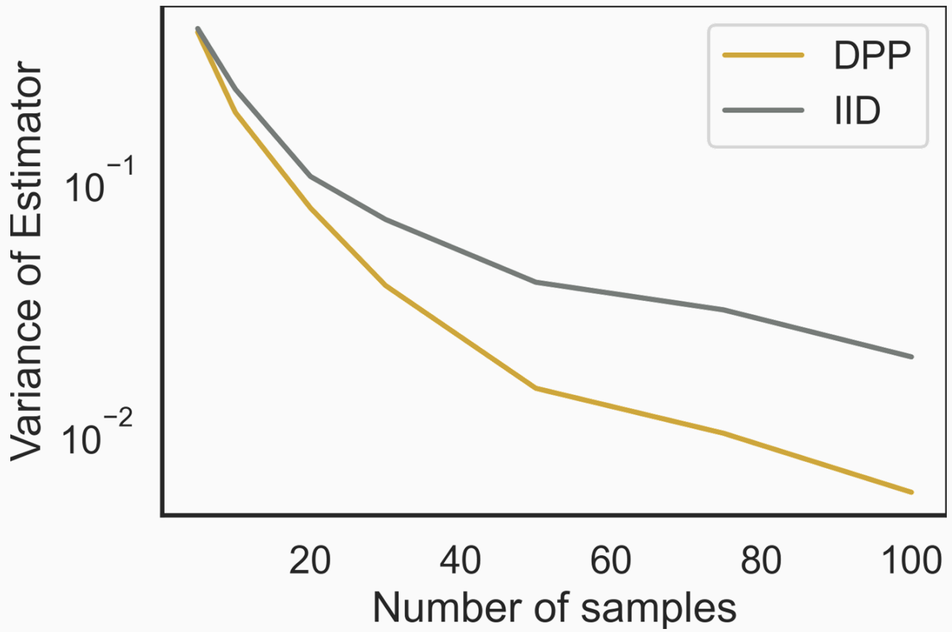


DPP Samples

[BT19, dpp_demo]

Variance reduction – Mean estimation





[BT19, dpp_demo]

$\mathcal{P}(\boldsymbol{Y} = Y)$ depends on the determinant of a

matrix selected based on the elements of Y.

Determinantal Point Process

Base set $\mathcal{Y}= \{1, \dots, n\}$ from which we sample a random subset $\boldsymbol{Y}$.

$\boldsymbol{Y}$ is distributed according to a point process $\mathcal{P}$ over $2^{\mathcal{Y}}$.

Poisson Process

- Simplest point process… too simple!
- Element memberships are parameterized by independent Bernoulli rvs.
- Special case of a DPP with marginal kernel $\mathfrak{K} = D_{\boldsymbol{p}}$.

$$\mathcal{P}(\boldsymbol{Y} = Y) = \prod_{i \in Y} p_i \prod_{i \notin Y} (1 - p_i)$$

Representing repulsion

Desiderata:

i. Density is tractable; including normalization constant

ii. Inclusion probabilities (marginals) are tractable

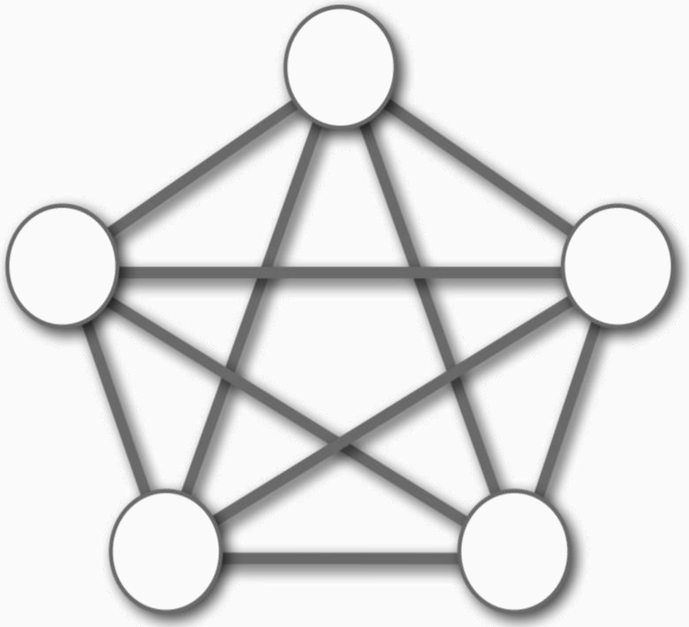
iii. Sampling is tractable

iv. Model is easy to understand

Contrary to most Gibbs processes (normalized, exponentiated potentials)**,**

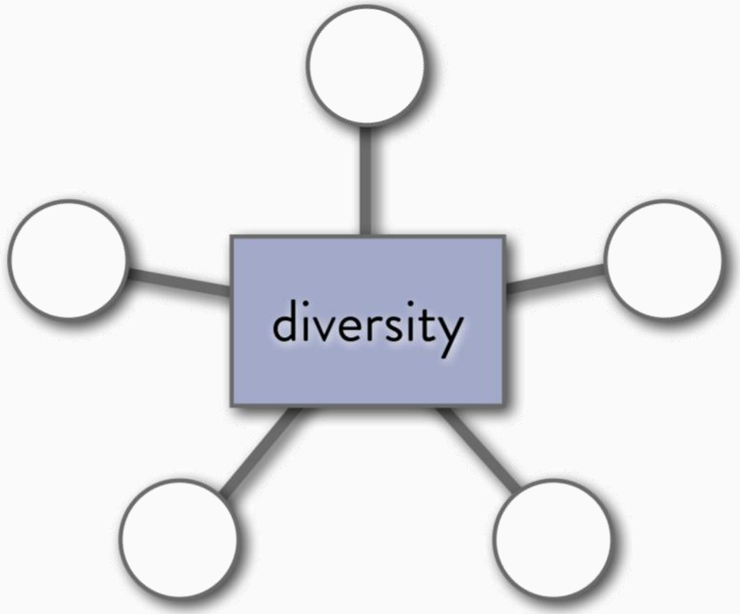
DPPs tick all the boxes

GMs vs DPPs



Loopy, negative interactions are hard

(Inference becomes intractable; worst case)



Global, negative interactions are easy

IO

[KTG13]

$\mathfrak{L}$-ensembles

- Model repulsion based on similarity between elements of $\mathcal{Y}$.
- Similarity between elements i and j is stored in $\mathfrak{L}_{ij}$.
- We assume $\mathfrak{L}$ to be positive definite.
- $\mathfrak{L}$ is known as the likelihood kernel.

We say that $\boldsymbol{Y}$ is distributed according to a DPP if:

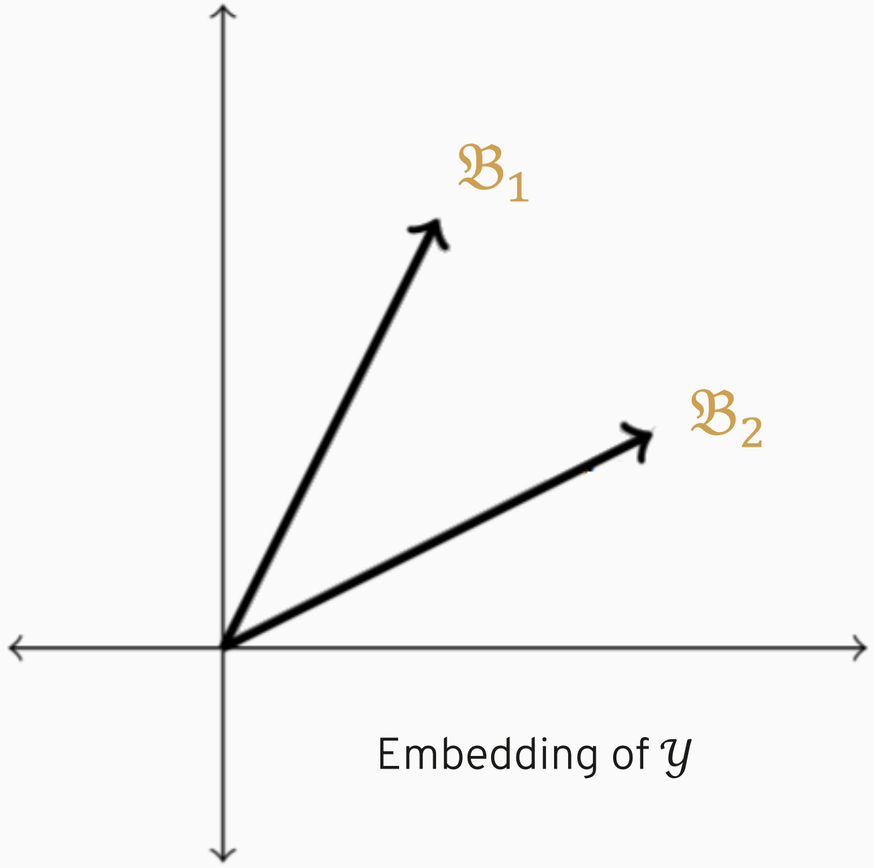
$$\mathcal{P}(\boldsymbol{Y} = Y) \propto \det \mathfrak{L}_Y$$

Where did the repulsion go?

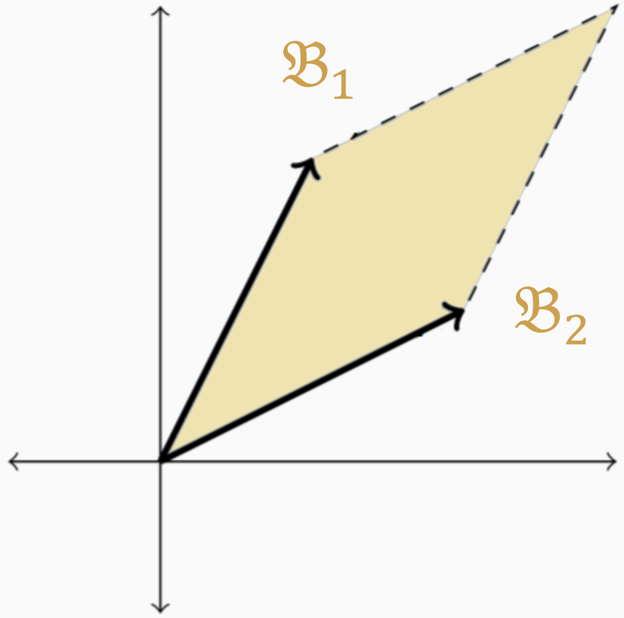
$$\mathfrak{L}_Y = [\mathfrak{B}^T \mathfrak{B}]_Y$$

$$\mathfrak{L}_{\{1,2,4\}} =$$

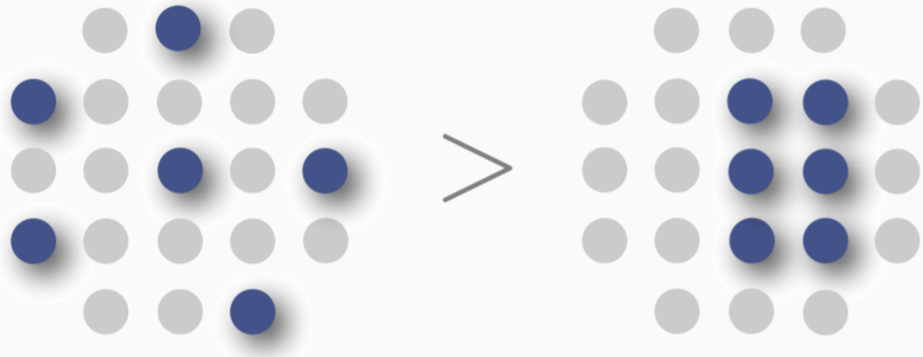
$$\mathcal{P}(\boldsymbol{Y} = Y) \propto \det \mathfrak{L}_Y = \det{}^2 \mathfrak{B}_Y$$



Where did the repulsion go?



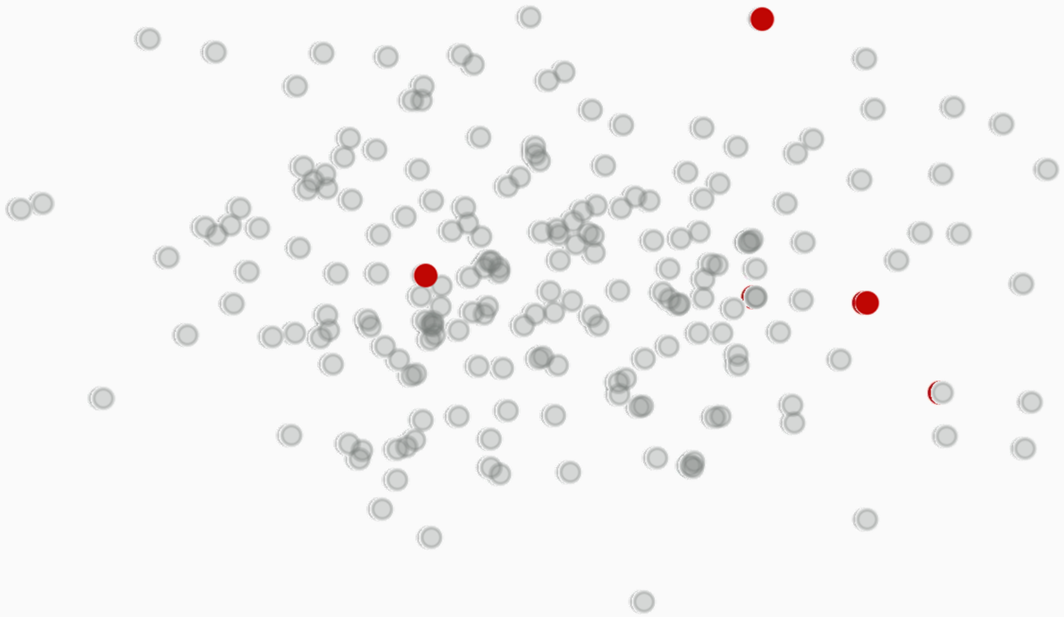
$$\mathrm{Vol}(\{\mathfrak{B}_i\}) = \det \mathfrak{B}$$

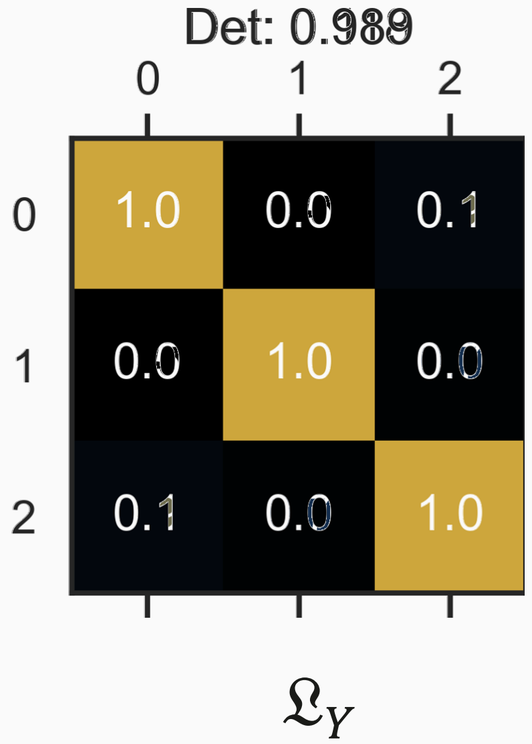


$$\mathcal{P}(\{i,j\}) \propto \mathcal{P}(\{i\})\mathcal{P}(\{j\}) - \left(\frac{\mathfrak{L}_{ij}}{\det(\mathfrak{L}+\mathbb{I})}\right)^2$$

[KTG13]

Where did the repulsion go?





Probability under a DPP grows with the spanned volume

[BT19, dpp_demo]

Normalization

Sum of squares of determinants of principal minors

Asked 9 years, 11 months ago Active 9 years, 11 months ago Viewed 2k times

5

I am interested in computing the sum of squares of determinants of principal minors. Let A be an $n \times n$ positive semidefinite matrix and A_S be a principal minor of A indexed by the set $S \subseteq \{1, \ldots, n\}$. The classical result (without squares) is:

$\sum_{S \subseteq \{1,\ldots,n\}} \det(A_S) = \det(A + I)$

4

Are there any results on computing

$\sum_{S \subseteq \{1,\ldots,n\}} \det^2(A_S)$

or any other powers?

determinants | sums-of-squares

Share Cite Improve this question Follow

asked Apr 11 '11 at 3:04

Ben

53 • 3

$$\sum_{A \subset Y \subset \mathcal{Y}} \det \mathfrak{L}_Y = \det(\mathfrak{L} + \mathbb{I}_{\bar{A}})$$

Analytic normalization constant!

Exploit linear-algebraic properties to make inference/sampling easy

(or feasible in high-dims)

Marginal kernels

- Consider a DPP with L-ensemble $\mathfrak{L}$.
- The inclusion (marginal) probability that $\boldsymbol{Y}$ contains a set S is given by:

$$\mathcal{P}(S \subset \boldsymbol{Y}) = \frac{1}{\mathfrak{S}} \sum_{S \subset Y} \det \mathfrak{L}_Y = \det \mathfrak{K}_S$$

with $\mathfrak{K} = \mathfrak{L}(\mathfrak{L} + \mathbb{I})^{-1}$.

- $\mathfrak{K}$ is known as the marginal kernel of the DPP.
- $\mathcal{P}(i \in \boldsymbol{Y}) = \mathfrak{K}_{ii}$.
- $\mathbb{E}[|\boldsymbol{Y}|] = \mathbb{E}[\sum_i \mathbb{1}_{i \in \boldsymbol{Y}}] = \sum_i \mathcal{P}(i \in \boldsymbol{Y}) = \operatorname{tr} \mathfrak{K}$.

Conditioning

$$\mathfrak{K}_{A\cup B} = \begin{bmatrix} \mathfrak{K}_A & \mathfrak{K}_{AB} \\ \mathfrak{K}_{BA} & \mathfrak{K}_B \end{bmatrix}$$

Schur complement

$$\det \mathfrak{K}_{A\cup B} = \det \mathfrak{K}_A \det(\mathfrak{K}_B - \mathfrak{K}_{BA}\mathfrak{K}_A^{-1}\mathfrak{K}_{AB})$$

DPPs are closed under conditioning!

$$\mathcal{P}(B \subset \boldsymbol{Y} \,|A \subset \boldsymbol{Y}) = \frac{\mathcal{P}(A \cup B \subset \boldsymbol{Y})}{\mathcal{P}(A \subset \boldsymbol{Y})} = \frac{\det \mathfrak{K}_{A\cup B}}{\det \mathfrak{K}_A} = \det(\mathfrak{K}_B - \mathfrak{K}_{BA}\mathfrak{K}_A^{-1}\mathfrak{K}_{AB})$$

Complexity?

- Evaluation of $\mathfrak{L}$ - $\mathcal{O}(n^2)$
- Normalization constant - $\mathcal{O}(n^3)$ [determinant]
- Marginal probabilities - $\mathcal{O}(n^3)$ [matrix inversion]
- Conditional probabilities - $\mathcal{O}(n^3)$ [Schur complement]

Questions?

Extensions

Conditional

k-

Structured

Non-symmetric

DPPs

k-DPPs

- In practical applications, often preferred to limit cardinality of the set
  - Search results
  - Minibatch selection
  - Summarization

$$\mathcal{P}(\boldsymbol{Y} = Y) \propto \det \mathfrak{L}_Y \, \mathbf{1}_{|\boldsymbol{Y}|=k}$$

- Normalization constant $\sum_{|\boldsymbol{Y}|=k} \det \mathfrak{L}_Y = e_k(\lambda_1, \ldots, \lambda_N)$ [k-th elementary sym. polynomial]

- Special case: 1-DPP

- Need not have a corresponding marginal kernel

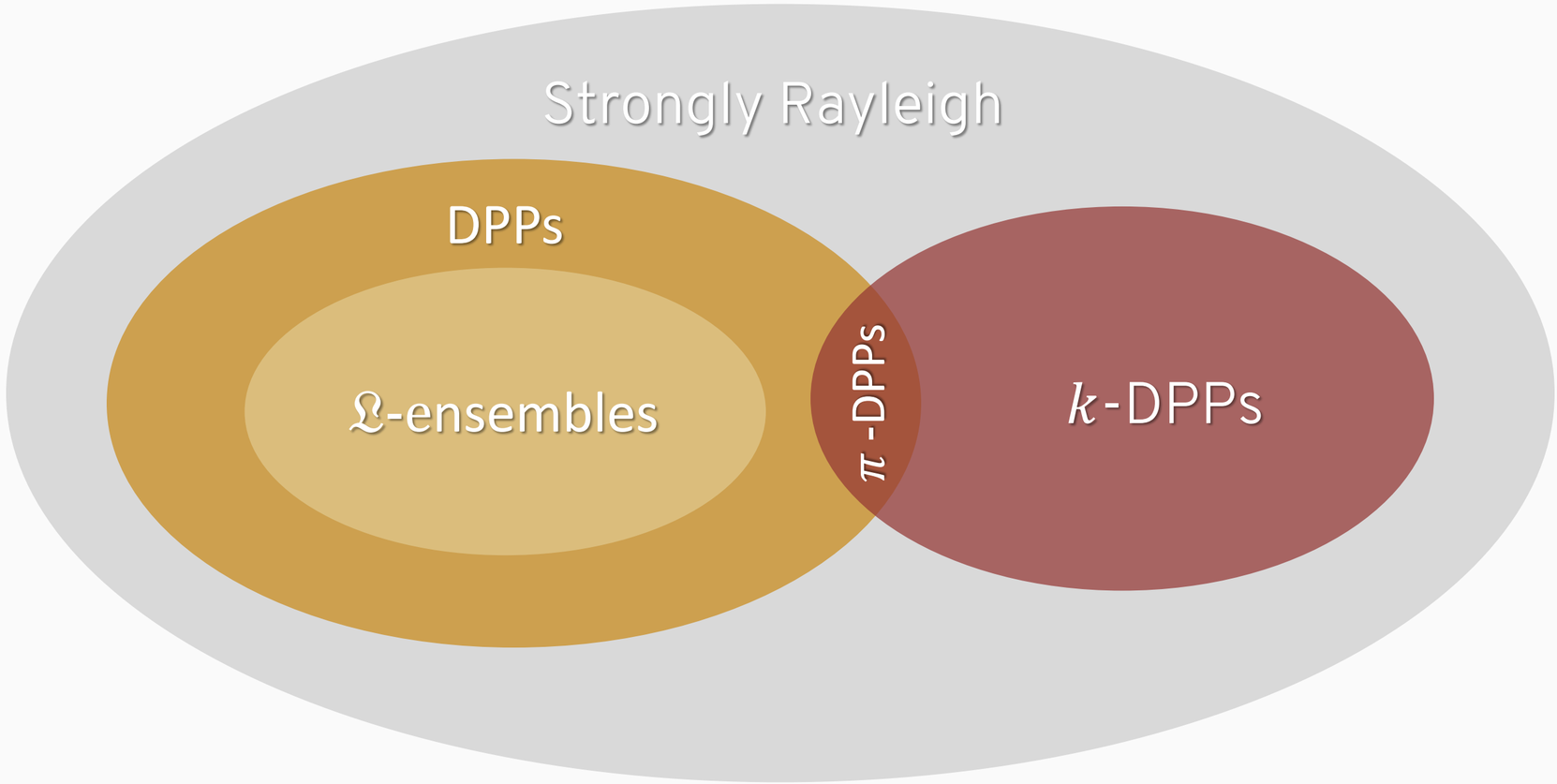
Elementary π-DPPs

- Special case: k-DPP with $k = \operatorname{rank} \mathfrak{L}$ and $\mathfrak{L} = V \Lambda V^T$, has marginal kernel $\mathfrak{K} = V V^T$
- A DPP is called **elementary** if the spectrum of its marginal kernel is $\{0, 1\}$.

$$\mathfrak{K}^V = \sum_{\boldsymbol{v} \in V} \boldsymbol{v} \boldsymbol{v}^T$$

- We denote this process as $\mathcal{P}^V$.
- If $\boldsymbol{Y} \sim \mathcal{P}^V$, then $|\boldsymbol{Y}| = |V|$ with probability one. ($|\boldsymbol{Y}|$ is a sum of Bernoulli rvs.)
- $\mathfrak{K}$ is a projection matrix – also called projection DPPs

Hierarchy of DPPs



Cauchy-Binet Lemma

- Consider matrices A of size $r \times s$ and B of size $s \times r$
- For each r-subset $Y \subset [1, \ldots, r]$, construct square matrices $A_{:Y}$ and $B_{Y:}$

$$\det AB = \sum_{|Y|=r} \det A_{:Y} \ \det B_{Y:}$$

$$F_n = \frac{\varphi^n - \psi^n}{\varphi - \psi}$$

JPM Binet

[Proof]

DPPs as mixture models

$$\mathcal{P}(\boldsymbol{Y} = Y) \propto \det \mathfrak{L}_Y = \det [V \Lambda V]_Y$$

$$= \det V_{Y:} \sqrt{\Lambda}\, \sqrt{\Lambda}\, V_{:Y}$$

$$= \sum_{|Z|=|Y|} \det V_{YZ} \sqrt{\Lambda_{ZZ}} \quad \det \sqrt{\Lambda_{ZZ}} V_{ZY}$$

$$= \sum_{|Z|=|Y|} \det V_{YZ} V_{YZ}^T \qquad \det \Lambda_{ZZ}$$

Elementary DPP

Diagonal $\mathfrak{L}$-ensemble

Sampling

- Consider a DPP with L-ensemble $\mathfrak{L} = \sum_n \lambda_n \boldsymbol{v}_n \boldsymbol{v}_n^T$.
- For each subset $J \subset \mathcal{Y}$, let V_J denote the set $\{\boldsymbol{v}_n\}_{n \in J}$ and the elementary DPP $\mathcal{P}^{V_J}$.

$$\mathcal{P} \propto \sum_{J \subset \mathcal{Y}} \mathcal{P}^{V_J} \prod_{n \in J} \lambda_n = \sum_{J \subset \mathcal{Y}} \mathcal{P}^{V_J} \det \boldsymbol{V}_J$$

Factorize the original DPP as a

mixture of *elementary* DPPs

Sampling via spectral decomposition

STAGE ONE

Choose elementary DPP $\mathcal{P}^J$

based on mixture weight

$$\Pr(J) \propto \prod_{n \in J} \lambda_n$$

STAGE TWO

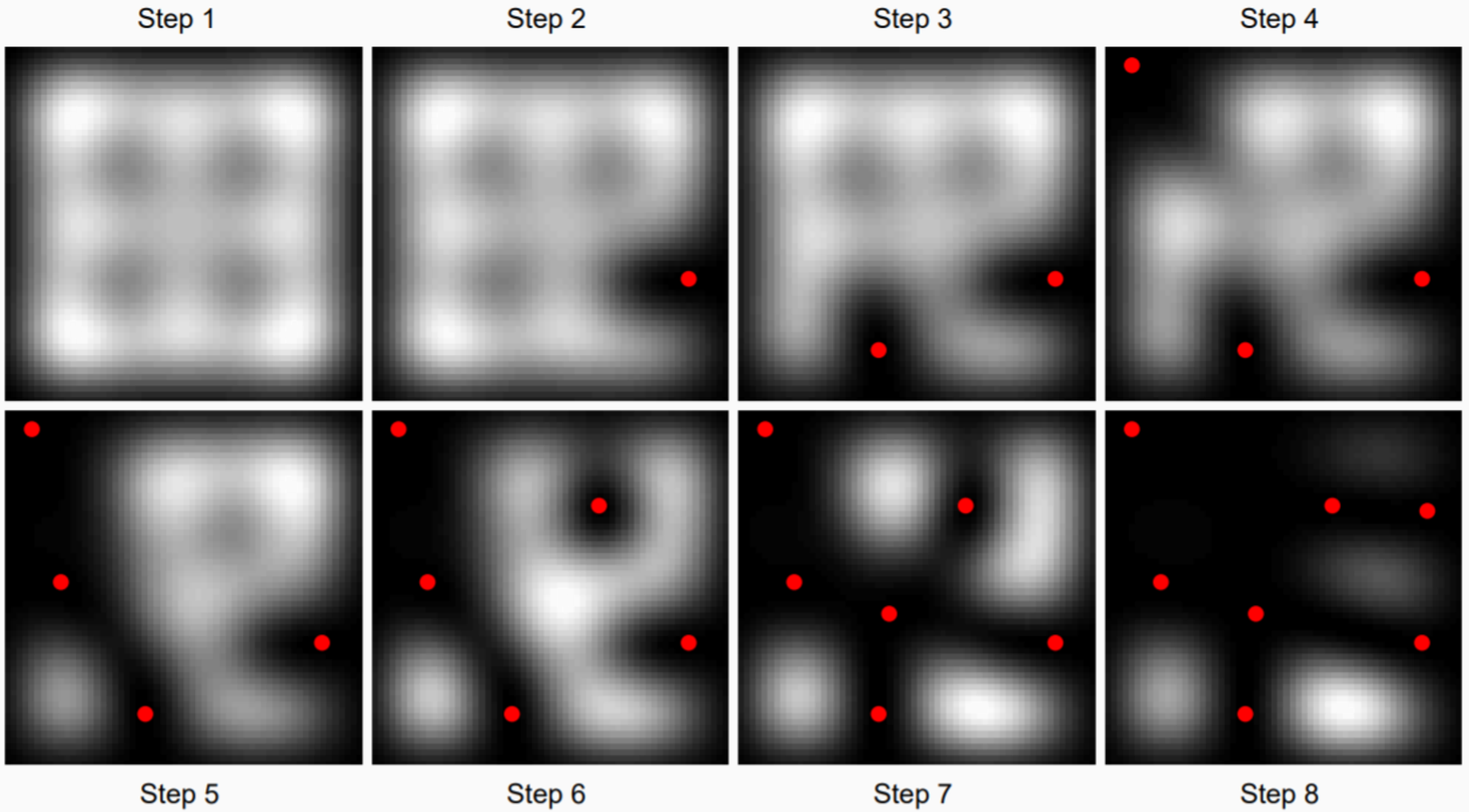
Draw a sample from $\mathcal{P}^J$

by sequential exploiting closure

of DPPs under conditioning

[KT12 – p.145]

Sampling in action



[KT12]

Advanced sampling

- Spectral method for sampling has cost $\mathcal{O}(n^2 + n^3 + nk^2)$
  - Dual sampling: instead of using $\mathfrak{L} = \mathfrak{B}^T\mathfrak{B}$ with $\mathfrak{B}$ $d \times n$ use $\mathfrak{C} = \mathfrak{B}\mathfrak{B}^T$ [KT12§3.3]
  - Random projections
  - Nyström approximations: Low rank approximation [Li, Jegelka, Sra 16a]
- MCMC sampling [LJS16b]
  - Add, remove, swap
  - Prove fast mixing for chains in terms of total variation
- Distortion-free intermediate sampling [Derezinski 18; CDV20]
  - Suitably construct an intermediate subset σ and then subsample from it

Learning DPPs

- Basic setting: Maximum Likelihood
  - Given $\{Y_t\}_{t=1}^{T}$ subsets of $\mathcal{Y}$. Parameterize $\mathfrak{L}$-ensemble as $\mathfrak{L}(\theta)$

$$\underset{\theta}{\operatorname{argmax}} \log \prod_t \mathcal{P}_\theta(Y_t) = \sum_t \log\det \mathfrak{L}_{Y_t}(\theta) - \log\det(\mathfrak{L}(\theta) + \mathbb{I})$$

- Can use gradient-based methods for optimizing θ
- Can be extended to conditioning on a covariate X: $\mathfrak{L}(\theta, X)$
  - For each X we have a DPP
  - X may be a query during search on which we want to condition the distribution over results
- See [KT12§4] for more details

Applications

Image search

{Relevance vs Diversity}

Mila Kunis - Wikipedia
en.wikipedia.org

Mila Kunis Facts | POPSUGAR Cel...
popsugar.com

Mila Kunis | POPSUGAR Celebrity
popsugar.com

Mila Kunis - Wikipedia, la ...
es.wikipedia.org

Mila Ximénez confiesa en 'Sálvame' ...
lecturas.com

Mila

Mila - Quebec Artificial Intelli...
sv.linkedin.com

Extractive summarization

document cluster

On Friday the shuttle Endeavor carried six astronauts into orbit to start building an international space station. The launch occurred after Russia and U.S. officials agreed not to delay the flight in order to orbit closer to MIR, and after a last-minute alarm forced a postponement. On Sunday astronauts joining the Russian-made Zarya control module cylinder with the American-made module to form a 70,000 pounds mass 77 feet long. . . .

human summary

- NASA and the Russian Space Agency have agreed to set aside . . .
- A last-minute alarm forced NASA to halt Thursday's launching . . .
- This was the first time in three years, and 19 flights . . .
- After a last-minute alarm, the launch went off flawlessly Friday . . .
- Following a series of intricate maneuvers and the skillful . . .
- It looked to be a perfect and, hopefully, long-lasting fit. . . .

extractive summary

[KT12]

Applications

- (Quasi) Monte-Carlo integration (Gautier et al., On two ways to use DPPs for Monte Carlo integration, 2019)

$$\int f(x)\mu(dx) \simeq \sum_{n=1}^{N} \omega_n f(x_n)$$

- Mini-batch sampling for SGD (Zhang et al., DPPs for Mini-Batch Diversification, 2017)

$$\theta \leftarrow \theta - \eta \sum_{i \in \mathcal{X}} \nabla L_i(\theta)$$

- Coresets (Tremblay et al., DPPs for Coresets, 2018)

$$\hat{L}(\mathcal{S}, \theta) = \sum_{s \in \mathcal{S}} \omega_s f(\boldsymbol{s}, \theta) \qquad \forall \theta \in \Theta \qquad (1-\epsilon)L(\mathcal{X}, \theta) \leqslant \hat{L}(\mathcal{S}, \theta) \leqslant (1+\epsilon)L(\mathcal{X}, \theta)$$

DPPs in Randomized LinAlg

- Consider a linear regression problem with a tall, full-rank matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ with $n \gg d$

$$w^* = \operatorname*{argmin}_w \|\boldsymbol{X}w - y\|^2 = \boldsymbol{X}^\dagger y$$

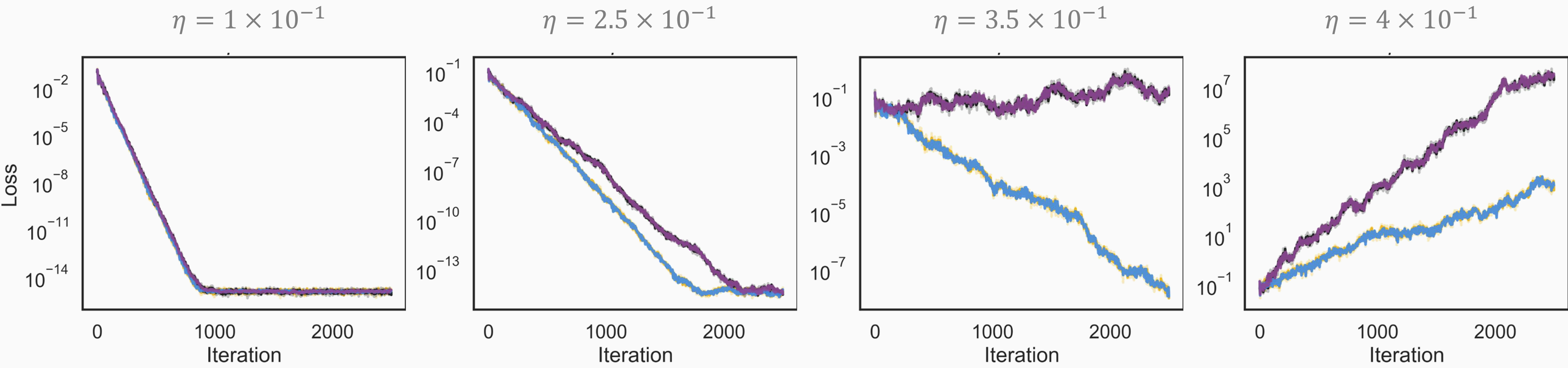
- Sketching: approximating matrix $\widetilde{\boldsymbol{X}}$ (subset of rows, low-rank)
- Usual bounds have (ε,δ)-PAC flavour
- If $S \sim d\text{-DPP}(\boldsymbol{X}\boldsymbol{X}^T)$, then $\mathbb{E}[\boldsymbol{X}_{S:}^{-1} y] = w^*$ [leverage scores]
- If $S \sim \text{DPP}\left(\frac{1}{\lambda}\boldsymbol{X}\boldsymbol{X}^T\right)$, then $\mathbb{E}[\boldsymbol{X}_{S:}^{\dagger} y] = \operatorname*{argmin}_w \|\boldsymbol{X}w - y\|^2 + \lambda\|w\|^2$ [ridge l.s.]

[DM20]

Minibatch sampling for LinReg

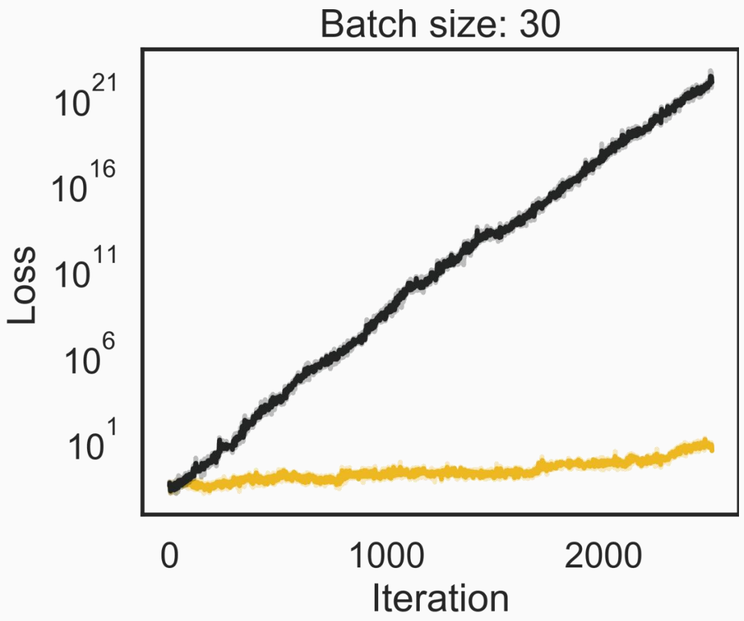
- Previously we related sampling with properties of analytic solution
- What is the influence of non-iid sampling **during stochastic optimization**?
  - Previous work by [Zhang, Kjellström, Mandt 17] for variance reduction
- Toy example: linear model
  - Gradients are ‘constant’ and correspond to points
  - Redundant points lead to redundant sampled gradients
- Sample minibatches $S \sim d\text{-DPP}(\boldsymbol{X}\boldsymbol{X}^T)$ and run SGD with momentum

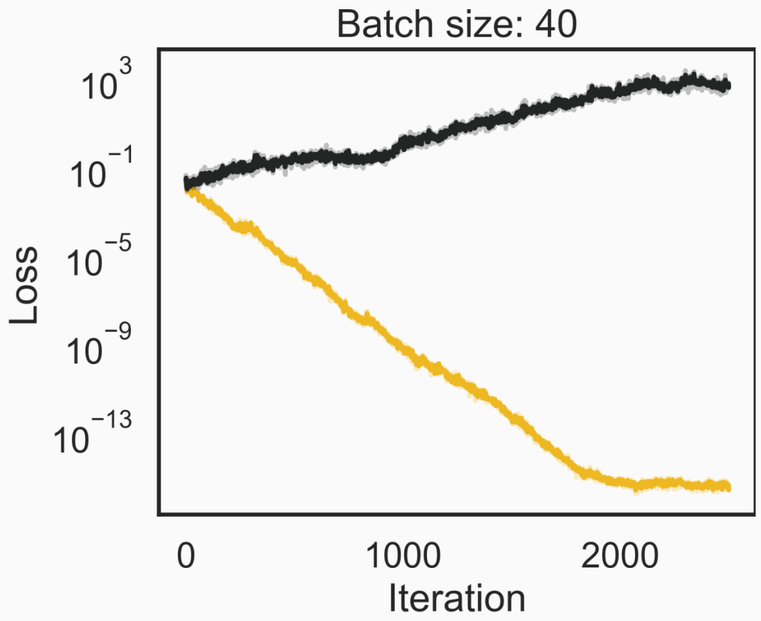
Minibatch sampling for LinReg

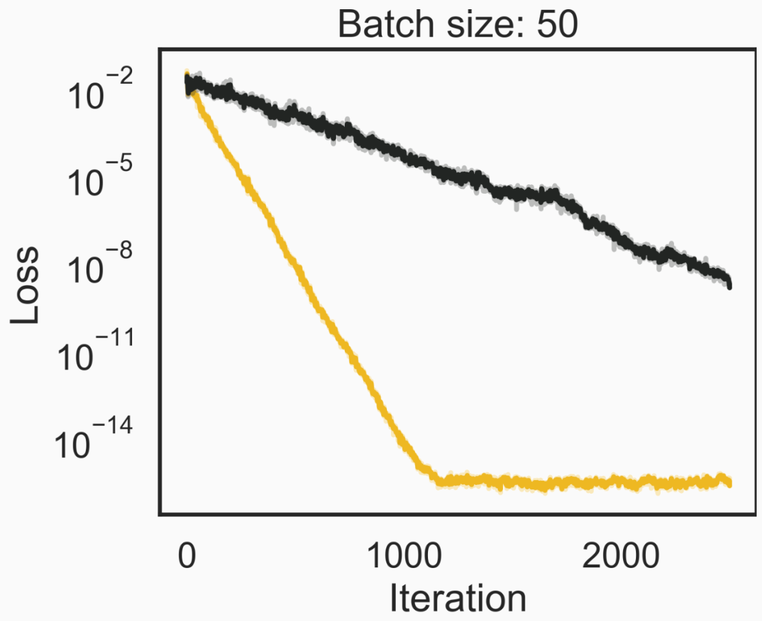


[optim_demo]

Overparameterized regime







[optim_demo]

Determinantal Point Processes

are elegant, efficient and useful

models of repulsion