Determinantal Point Processes for Machine Learning

A. Kulesza, B. Taskar, Foundation and Trends in ML, 2012

Maryan Morel

ML Journal Club

Table of contents

Introduction

Context

Structured output space: difficult inference.

Graphical models:

- Tractable when dependence graph is a tree
- or when interactions are positive in loopy graphs

Determinantal Point Processes (DPP)

- Probabilistic model of global negative correlations
- Efficient algorithms for sampling, marginalization, conditioning, etc.

Characterization

Definitions

Discrete Point Process (discrete PP)

A discrete Point Process $\mathcal{P}$ on a ground set $\mathcal{Y} = \{1, 2, \ldots, N\}$ is a probability measure over point patterns $Y \in 2^{\mathcal{Y}}$ where $2^{\mathcal{Y}}$ is set of all finite subsets of $\mathcal{Y}$.

Determinantal PP (DPP)

$\mathcal{P}$ is a DPP of marginal kernel K if the random subset $\mathbf{Y} \in 2^{\mathcal{Y}}$ drawn according to $\mathcal{P}$ verifies

$$\forall A \subseteq \mathcal{Y}, \mathcal{P}(A \subseteq \mathbf{Y}) = \det(K_A)$$

where $0 \preceq K \preceq I$ is a $N \times N$ matrix indexed by the elements of $\mathcal{Y}$, and K_A is its restriction to the entries indexed by the elements of A. This equation gives the marginal probabilities of inclusion.

Behaviour

DPP favours diversity

$$\mathcal{P}(\{i\} \subseteq \mathbf{Y}) = K_{ii}$$

$$\begin{aligned}\mathcal{P}(\{i,j\} \subseteq \mathbf{Y}) &= K_{ii}K_{jj} - K_{ij}K_{ji} \\ &= \mathcal{P}(\{i\} \subseteq \mathbf{Y})\mathcal{P}(\{j\} \subseteq \mathbf{Y}) - K_{ij}^2\end{aligned}$$

- Off-diagonal elements determine the negative correlations between pairs of elements: i and j have smaller probability to co-occur for large K_{ij}.
- if $K_{ij} = \sqrt{K_{ii}K_{jj}}$, i and j are perfectly similar
- if K is diagonal, there are no correlations between i and j
- *DPPs cannot represent distributions where elements are more likely to co-occur than if they were independent*

Behaviour

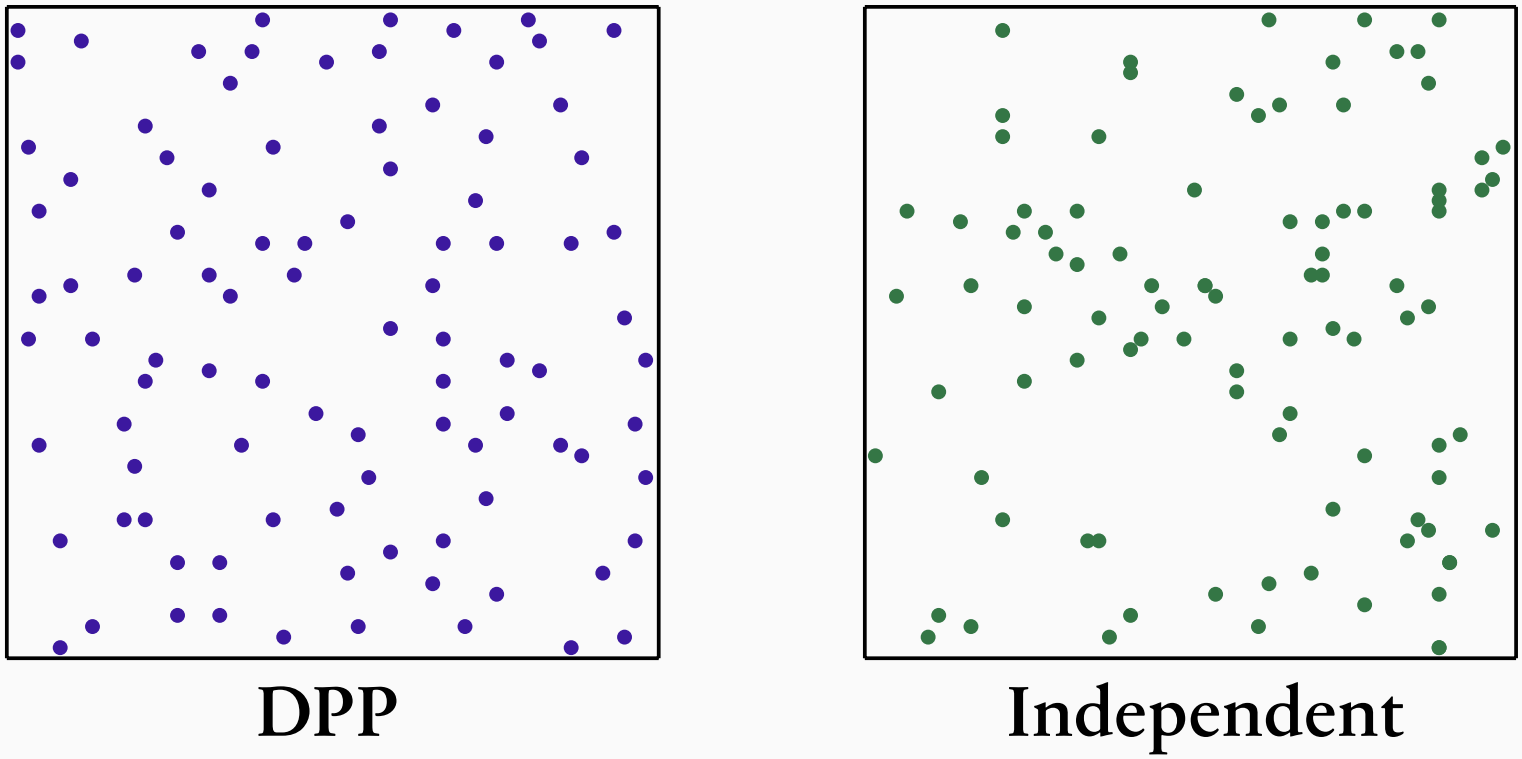

Figure 1: A set of points in the plane drawn from a DPP (left), and the same number of points sampled independently using a Poisson point process (right). (*source:* [1])

L-ensembles

Alternative representation of DPPs using a real symmetric positive semidefinite matrix L, indexed by the elements of $\mathcal{Y}$:

$$\mathcal{P}_L(Y) = \mathcal{P}_L(\mathbf{Y} = Y) = \frac{\det(L_Y)}{\det(L_Y + I)}$$

This equation specifies the atomic probabilities for every possible instantiation of $\mathbf{Y}$. As before:

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

L-ensembles vs Marginal Kernels

Alternative representations:

$$K = L(L + I)^{-1} = I - (L + I)^{-1}$$

- K gives marginal probabilities, while L-ensembles model the atomic probabilities
- L need only to be positive semidefinie, while K has also requires bounded above eigenvalues.

Geometry

Let B be a $D \times N$ matrix s.t. $L = B^\top B$ (can always be found) and B_i denote the columns of B.

- B_i can be thought as feature vectors describing elements of $\mathcal{Y}$
- L measures similarity of such elements using feature vectors dot products
- The probability assigned by a DPP to a set Y is related to the volume spanned by its associated feature vectors:

$$\mathcal{P}_L(Y) \propto \det(L_Y) = \mathrm{Vol}(\{B_i\}_{i \in Y})$$

Geometry

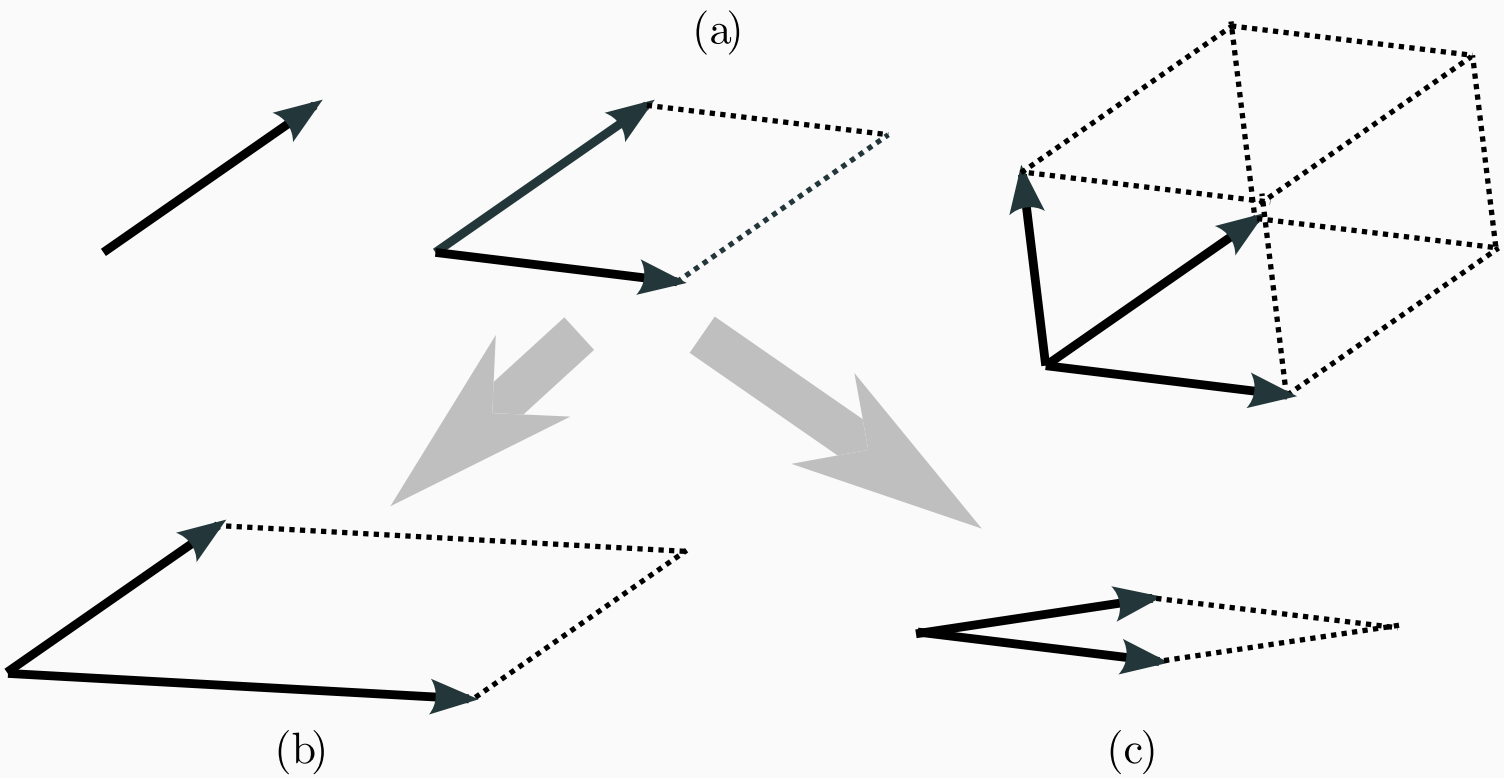

Fig. 2.3 A geometric view of DPPs: each vector corresponds to an element of $\mathcal{Y}$. (a) The probability of a subset Y is the square of the volume spanned by its associated feature vectors. (b) As the magnitude of an item's feature vector increases, so do the probabilities of sets containing that item. (c) As the similarity between two items increases, the probabilities of sets containing both of them decrease.

Figure 2: (*source:* [1])

Properties & Inference

Properties

Many useful properties

	restriction of a DPP: if $\mathbf{Y}$ is distributed as a $\mathbf{Y} \bigcap A, A \subseteq \mathcal{Y}$ is distributed as a DPP
Complement	$\mathcal{Y} - \mathbf{Y}$ is distributed as a DPP
Domination	$K \preceq K' \Rightarrow \det(K_A) \leq \det(K'_A)$
Scaling	$0 \leq \gamma < 1, K = \gamma K' \Rightarrow \det(K_A) = \gamma^{\lvert A \rvert} \det(K'_A)$
Cardinality	$\mathbb{E}(\lvert \mathbf{Y} \rvert) = \operatorname{tr}(K)$

Inference

Computation time order of magnitudes: N_1 is the upper bound for N to compute at interactive speed (i.e. 2-3 seconds of computation), N_2 is the upper bound for a 5-10 minutes computation.

Normalization of $\mathcal{P}_L(Y)$: $O(N^{2.376})$, $N_1 \approx 5000$, $N_2 \approx 40000$

Marginalization i.e. computing K from L: $O(N^3)$, $N_1 \approx 2000$, $N_2 \approx 20000$

Conditioning ex. computing $\mathcal{P}_L(\mathbf{Y} = B | A \bigcap \mathbf{Y} = \emptyset)$: requires a $N \times N$ matrix inversion ☹

Simulation $N_1 \approx 1000$, $N_2 \approx 10000$

MAP inference i.e. find $Y \subseteq \mathcal{Y}$ that maximizes $\mathcal{P}_L(Y)$. Finding the mode or approximating it is NP-hard ☹☹☹

Learning

Quality vs diversity formulation

We want diversity of items selected by a DPP to be balanced by some underlying preferences for items in $\mathcal{Y}$

Quality term $q_i \in \mathbb{R}^+$ underlying preference for items in $\mathcal{Y}$

Diversity features $\phi_i \in \mathbb{R}^D$, $||\phi_i|| = 1$, measures similarity between items

$$L_{ij} = q_i \phi_i^\top \phi_j q_j$$

Denoting similarity $S_{ij} \equiv \phi_i^\top \phi_j = \frac{L_{ij}}{\sqrt{L_{ii} L_{jj}}}$, we have

$$\mathcal{P}_L(Y) \propto \left(\prod_{i \in Y} q_i^2 \right) \det(S_Y)$$

Back to geometry

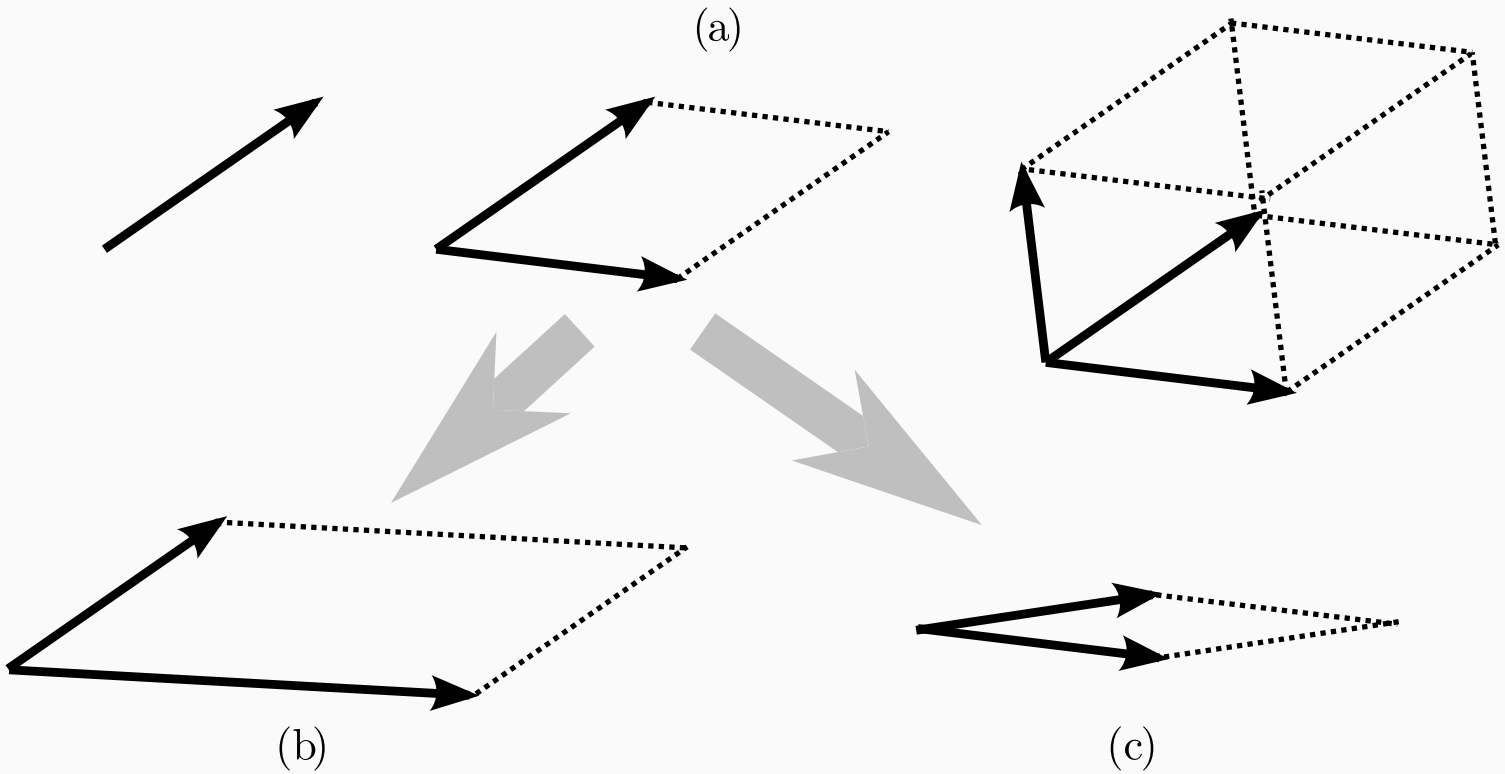

Fig. 3.1 Revisiting DPP geometry: (a) The probability of a subset Y is the square of the volume spanned by $q_i\phi_i$ for $i \in Y$. (b) As item i's quality q_i increases, so do the probabilities of sets containing item i. (c) As two items i and j become more similar, $\phi_i^\top \phi_j$ increases and the probabilities of sets containing both i and j decrease.

Figure 3: (*source:* [1])

Learning DPPs

- How do we use data ? Problem: fixed base set $\mathcal{Y}$
- How to make $\mathcal{Y}$ depend on the context ?
- Solution: Make the DPP depend on input data $X \in \mathcal{X}$

Example

In the case of article summarization, let $\mathcal{Y}$ be the set of all sentences of the corpus, and $\mathcal{X}$ be the space of news articles. Let $\mathcal{Y}(X)$ be the set of all sentences in article X.

Conditional DPPs

Definition

Using the L-ensemble formulation, we define the conditional probabilistic model $\mathcal{P}(\mathbf{Y} = Y|X)$, for all $Y \subseteq \mathcal{Y}(X)$ as

$$\mathcal{P}(\mathbf{Y} = Y|X) \propto \det(L_Y(X))$$

where $L(X)$ is a positive semidefinite $|\mathcal{Y}(X)| \times |\mathcal{Y}(X)|$ kernel matrix depending on the input

Supervised learning

Dataset

T *i.i.d.* samples $\{(X^{(t)}Y^{(t)})\}_{t=1}^{T}$ drawn from a distribution D over pairs $(X, Y) \in \mathcal{X} \times 2^{\mathcal{Y}(X)}$

Likelihood

Parameterizing the kernel $L(X;\theta)$ in terms of a generic θ, we obtain

$$\mathcal{P}_\theta(Y|X) = \frac{\det(L_Y(X;\theta))}{\det(L_Y(X;\theta)+I)}$$

$$\mathcal{L}(\theta) = \log \prod_{t=1}^{T} \mathcal{P}_\theta(Y|X)$$

$$= \sum_{t=1}^{T} [\log\det(L_Y(X;\theta)) - \log\det(L_Y(X;\theta)+I)]$$

Kernel parameterization

Quality/diversity formulation

$$L_{ij}(X) = q_i(X)\phi_i(X)^\top \phi_j(X)q_j(X)$$

with $q_i(X) \in (R)^+$ and $\phi_i(X) \in \mathbb{R}^D, ||\phi_i(X)|| = 1$

Log-linear quality scores

$$q_i(X;\theta) = \exp\left(\frac{1}{2}\theta^\top \mathbf{f}_i(X)\right)$$

where $\mathbf{f}_i(X) \in \mathbb{R}^m$ is a feature vector for item i and $\theta \in \mathbb{R}^m$.

Fixed kernel $\phi_i(X)$ (analoguous to SVM)

Optimization

- Resulting likelihood concave in θ
- Gradient can be computed in $O(N^3)$
- or in $O(ND^3)$ using a dual formulation, effective if $D \ll N$
- What if D is big ? (e.g. NLP, high-res image processing) Use random projections!

Random projections

High dimensional points can be projected onto a logarithmic number of dimensions while approximately preserving distances between them (+ array of theoretical guarantees)

Application: document summarization

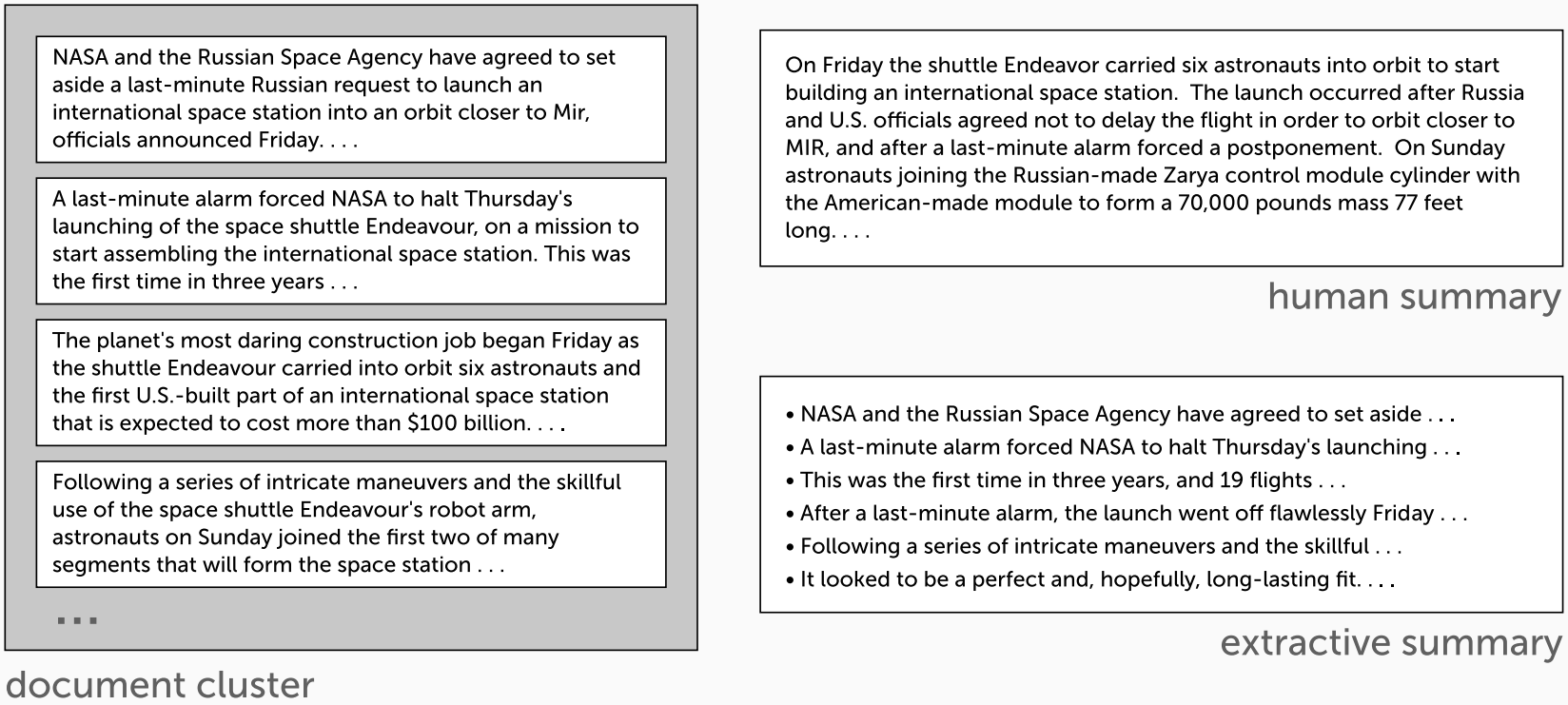

Fig. 4.1 A sample cluster from the DUC 2004 test set, with one of the four human reference summaries and an (artificial) extractive summary.

Figure 4: (source: [1])

Going further

$k-$DPP

- DPP assign probability to all sets in $2^{\mathcal{Y}}$
- For some applications (e.g. image search), we might want to assign probability only on sets of cardinality k
- Solution: condition the DPP on the cardinality of the random set $\mathbf{Y}$
- Simple in theory...
- ...but requires new algorithms for inference, sampling and learning to keep solving these problems in polynomial time

Structured DPP

- Polynomial-time inference and learning w.r.t. N with classis DPPs
- Can be too slow for some applications, e.g. modeling position of objects trajectories in space: $N = M^R$. Leveraging *structure* in data can help
- View $\mathcal{Y}$ as a set of structures $\mathbf{y}$ given by a sequence of R parts $(y_1, y_2, \ldots, y_R)$, with $y_r \in 1, 2, \ldots, M$.
- Solution: use *second order message passing* in the dual
- If D is too big, use random projections to approximate the DPP

Conclusion

Takeaway

Probabilistic semantics to model repulsive phenomena on many variables

Rich theory and efficient algorithms: practical applications

Only tractable negative correlation model ?

Questions?

References i

A. Kulesza and B. Taskar.
Determinantal point processes for machine learning.
Foundations and Trends® in Machine Learning, 5(2–3):123–286, 2012.