Determinantal Point Processes For Machine Learning

Alex Kulesza, Ben Taskar

Wittawat Jitkrittum

Gatsby Machine Learning Journal Club
31 Oct 2016

What is a determinantal point process (DPP)?

- A distribution over finite subsets of a fixed ground set $\mathcal{Y}$.
  - In general, $\mathcal{Y}$ can be uncountable.
  - Here, we assume $\mathcal{Y} = \{1, 2, \ldots, N\}$.
- If $\mathcal{Y} = \{a, b, c, d\}$ and $\mathbf{Y} \sim$ DPP, then we can ask $\mathcal{P}(\{a, d\} \subseteq \mathbf{Y})$.
- Parametrized by a similarity matrix $K \in \mathbb{R}^{N \times N}$. Similar items are less likely to co-occur. Encourage diversity.

DPP

Independent

- **Applications:**
  - Text summarization. Choose a diverse subset of sentences.
  - Diverse search results for a search engine (conditional DPPs).
  - A set of spike times.

What is a determinantal point process (DPP)?

- A distribution over finite subsets of a fixed ground set $\mathcal{Y}$.
  - In general, $\mathcal{Y}$ can be uncountable.
  - Here, we assume $\mathcal{Y} = \{1, 2, \ldots, N\}$.
- If $\mathcal{Y} = \{a, b, c, d\}$ and $\mathbf{Y} \sim \mathrm{DPP}$, then we can ask $\mathcal{P}(\{a, d\} \subseteq \mathbf{Y})$.
- Parametrized by a similarity matrix $K \in \mathbb{R}^{N \times N}$. Similar items are less likely to co-occur. Encourage diversity.

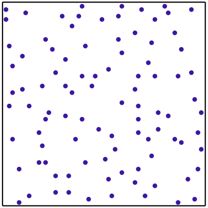
DPP

Independent

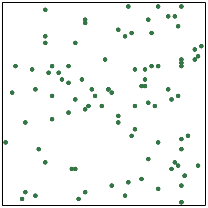
- **Applications:**
  - Text summarization. Choose a diverse subset of sentences.
  - Diverse search results for a search engine (conditional DPPs).
  - A set of spike times.

What is a determinantal point process (DPP)?

- A distribution over finite subsets of a fixed ground set $\mathcal{Y}$.
  - In general, $\mathcal{Y}$ can be uncountable.
  - Here, we assume $\mathcal{Y} = \{1, 2, \ldots, N\}$.
- If $\mathcal{Y} = \{a, b, c, d\}$ and $\mathbf{Y} \sim \mathrm{DPP}$, then we can ask $\mathcal{P}(\{a, d\} \subseteq \mathbf{Y})$.
- Parametrized by a similarity matrix $K \in \mathbb{R}^{N \times N}$. Similar items are less likely to co-occur. Encourage diversity.



DPP

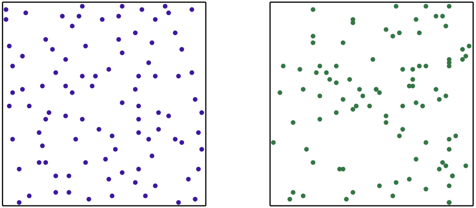


Independent

- **Applications:**
  - Text summarization. Choose a diverse subset of sentences.
  - Diverse search results for a search engine (conditional DPPs).
  - A set of spike times.

What is a determinantal point process (DPP)?

- A distribution over finite subsets of a fixed ground set $\mathcal{Y}$.
  - In general, $\mathcal{Y}$ can be uncountable.
  - Here, we assume $\mathcal{Y} = \{1, 2, \ldots, N\}$.
- If $\mathcal{Y} = \{a, b, c, d\}$ and $\mathbf{Y} \sim \text{DPP}$, then we can ask $\mathcal{P}(\{a, d\} \subseteq \mathbf{Y})$.
- Parametrized by a similarity matrix $K \in \mathbb{R}^{N \times N}$. Similar items are less likely to co-occur. Encourage diversity.



DPP Independent

- **Applications:**
  - Text summarization. Choose a diverse subset of sentences.
  - Diverse search results for a search engine (conditional DPPs).
  - A set of spike times.

Outline

- Basics of DPPs
- Properties: conditioning, restriction, complementation, etc.
- L-ensembles: normalization, marginalization.
- Sampling from a DPP
- Dual representation.

- Mainly from

Determinantal point processes for machine learning Alex Kulesza, Ben Taskar https://arxiv.org/abs/1207.6083

- Took some slides from Lobato & Ge, 2014.
  https://jmhldotorg.files.wordpress.com/2014/02/slidesrcc-dpps.pdf

Formal definition

- Let $\mathcal{Y} = \{1, \ldots, N\}$ be a fixed ground set.
- A point process $\mathcal{P}$ on $\mathcal{Y}$ is a probability distribution on $2^{\mathcal{Y}}$.
- Let $K \in \mathbb{R}^{N \times N}$ be a symmetric similarity matrix such that $\mathbf{0} \preceq K \preceq I$.
  - Eigenvalues are in $[0, 1]$.
- $\mathcal{P}$ is a DPP if, when $\mathbf{Y} \sim \mathcal{P}$, then for every $A \subseteq \mathcal{Y}$,

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

where $K_A := [K_{ij}]_{i,j \in A}$. Define $\det(K_\emptyset) = 1$.

- K is called the marginal kernel.

Formal definition

- Let $\mathcal{Y} = \{1, \ldots, N\}$ be a fixed ground set.
- A point process $\mathcal{P}$ on $\mathcal{Y}$ is a probability distribution on $2^{\mathcal{Y}}$.
- Let $K \in \mathbb{R}^{N \times N}$ be a symmetric similarity matrix such that $\mathbf{0} \preceq K \preceq I$.
  - Eigenvalues are in $[0, 1]$.
- $\mathcal{P}$ is a DPP if, when $\mathbf{Y} \sim \mathcal{P}$, then for every $A \subseteq \mathcal{Y}$,

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

where $K_A := [K_{ij}]_{i,j \in A}$. Define $\det(K_\emptyset) = 1$.

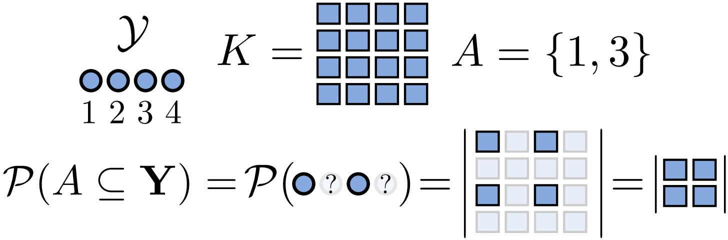
- K is called the marginal kernel.

Formal definition

- Let $\mathcal{Y} = \{1, \dots, N\}$ be a fixed ground set.
- A point process $\mathcal{P}$ on $\mathcal{Y}$ is a probability distribution on $2^{\mathcal{Y}}$.
- Let $K \in \mathbb{R}^{N \times N}$ be a symmetric similarity matrix such that $\mathbf{0} \preceq K \preceq I$.
  - Eigenvalues are in $[0, 1]$.
- $\mathcal{P}$ is a DPP if, when $\mathbf{Y} \sim \mathcal{P}$, then for every $A \subseteq \mathcal{Y}$,

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

where $K_A := [K_{ij}]_{i,j \in A}$. Define $\det(K_\emptyset) = 1$.



- K is called the marginal kernel.

Negative correlations in DPPs

If $A = \{i, j\}$, then

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

- Off-diagonal entries determine the negative correlations.
- If K is diagonal, items in $\mathbf{Y}$ are independent (Poisson point process).
- If $K_{ij} = \sqrt{K_{ii}K_{jj}}$, then i and j never appear together in $\mathbf{Y}$.
- Correlations are always negative.

"DPPs cannot represent distributions where elements are more likely to co-occur than if they were independent."

Negative correlations in DPPs

If $A = \{i, j\}$, then

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

- Off-diagonal entries determine the negative correlations.
- If K is diagonal, items in $\mathbf{Y}$ are independent (Poisson point process).
- If $K_{ij} = \sqrt{K_{ii}K_{jj}}$, then i and j never appear together in $\mathbf{Y}$.
- Correlations are always negative.

"DPPs cannot represent distributions where elements are more likely to co-occur than if they were independent."

Negative correlations in DPPs

If $A = \{i, j\}$, then

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

- Off-diagonal entries determine the negative correlations.
- If K is diagonal, items in $\mathbf{Y}$ are independent (Poisson point process).
- If $K_{ij} = \sqrt{K_{ii}K_{jj}}$, then i and j never appear together in $\mathbf{Y}$.
- Correlations are always negative.

"DPPs cannot represent distributions where elements are more likely to co-occur than if they were independent."

Negative correlations in DPPs

If $A = \{i, j\}$, then

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

- Off-diagonal entries determine the negative correlations.
- If K is diagonal, items in $\mathbf{Y}$ are independent (Poisson point process).
- If $K_{ij} = \sqrt{K_{ii}K_{jj}}$, then i and j never appear together in $\mathbf{Y}$.
- Correlations are always negative.

"DPPs cannot represent distributions where elements are more likely to co-occur than if they were independent."

Property 1: conditioning in DPPs

(slide from Lobato & Ge, 2014)

- DPPs are closed under conditioning.

$$
\begin{aligned}
\mathcal{P}(B \subseteq \mathbf{Y} | A \subseteq \mathbf{Y}) &= \frac{\mathcal{P}(A \cup B \subseteq \mathbf{Y})}{\mathcal{P}(A \subseteq \mathbf{Y})} \\
&= \frac{\det(K_{A\cup B})}{\det(K_A)} = \det(K_B - K_{BA} K_A^{-1} K_{AB}) \\
&= \det([K - K_{\star A} K_A^{-1} K_{A\star}]_B)\,.
\end{aligned}
$$

$$
K_{A\cup B} = \begin{bmatrix} K_A & K_{AB} \\ K_{BA} & K_B \end{bmatrix}
$$

Schur Complement of K_A.

$\det(K_{A\cup B}) =$
$\det(K_A)\det(K_B - K_{BA} K_A^{-1} K_{AB})\,.$

Restriction, and complement

- Assume $\mathbf{Z} \sim \mathrm{DPP}(K)$ with ground set $\mathcal{Y}$.

Property 2: restriction

- Let $A \subseteq \mathcal{Y}$.
- Define $\mathbf{Y} := \mathbf{Z} \cap A$.
- Then, $\mathbf{Y} \sim \mathrm{DPP}(K_A)$.

Same as changing the ground set $\mathcal{Y}$ to A.

Property 3: complement

- Define $\mathbf{Y} := \mathcal{Y} \backslash \mathbf{Z}$.
- Then, $\mathbf{Y} \sim \mathrm{DPP}(I - K)$.

That is, $\mathcal{P}(A \cap \mathbf{Y} = \emptyset) = \det(\overline{K}_A) = \det(I - K_A)$.

Restriction, and complement

- Assume $\mathbf{Z} \sim \mathrm{DPP}(K)$ with ground set $\mathcal{Y}$.

Property 2: restriction

- Let $A \subseteq \mathcal{Y}$.
- Define $\mathbf{Y} := \mathbf{Z} \cap A$.
- Then, $\mathbf{Y} \sim \mathrm{DPP}(K_A)$.

Same as changing the ground set $\mathcal{Y}$ to A.

Property 3: complement

- Define $\mathbf{Y} := \mathcal{Y} \backslash \mathbf{Z}$.
- Then, $\mathbf{Y} \sim \mathrm{DPP}(I - K)$.

That is, $\mathcal{P}(A \cap \mathbf{Y} = \emptyset) = \det(\overline{K}_A) = \det(I - K_A)$.

Domination, and scaling

- Assume $\mathbf{Z} \sim \mathrm{DPP}(K)$ with ground set $\mathcal{Y}$.

Property 4: domination

If $K \preceq K'$ (i.e., $K' - K$ is positive semidefinite), then for any $A \subseteq \mathcal{Y}$,

$$\det(K_A) \leq \det(K'_A).$$

- $\mathrm{DPP}(A')$ assigns higher marginal probabilities to every set A.

Property 5: scaling

If $K = \gamma K'$ for some $0 \leq \gamma < 1$, then for $A \subseteq \mathcal{Y}$,

$$\det(K_A) = \gamma^{|A|} \det(K'_A).$$

- $\mathrm{DPP}(K)$ generates sample from $\mathrm{DPP}(K')$. Then, delete each item with probability $1 - \gamma$.

Domination, and scaling

- Assume $\mathbf{Z} \sim \text{DPP}(K)$ with ground set $\mathcal{Y}$.

Property 4: domination

If $K \preceq K'$ (i.e., $K' - K$ is positive semidefinite), then for any $A \subseteq \mathcal{Y}$,

$$\det(K_A) \leq \det(K'_A).$$

- $\text{DPP}(A')$ assigns higher marginal probabilities to every set A.

Property 5: scaling

If $K = \gamma K'$ for some $0 \leq \gamma < 1$, then for $A \subseteq \mathcal{Y}$,

$$\det(K_A) = \gamma^{|A|} \det(K'_A).$$

- $\text{DPP}(K)$ generates sample from $\text{DPP}(K')$. Then, delete each item with probability $1 - \gamma$.

L-ensembles

- An L-ensemble defines a DPP through a real, symmetric matrix $L \succeq \mathbf{0}$:

$$\mathcal{P}_L(\mathbf{Y} = Y) \propto \det(L_Y).$$

- No need $L \preceq I$.
- For modeling data, this parameterization is more convenient.

To get the normalizer,

Theorem (2.1)

For any $A \subseteq \mathcal{Y}$,

$$\sum_{A \subseteq Y \subseteq \mathcal{Y}} \det(L_Y) = \det(L + I_{\overline{A}}),$$

where $\overline{A} = \mathcal{Y} \backslash A$.

- The normalizer is $\sum_{Y \subseteq \mathcal{Y}} \det(L_Y) = \det(L + I)$. Complexity: $O(N^3)$.

L-ensembles

- An L-ensemble defines a DPP through a real, symmetric matrix $L \succeq \mathbf{0}$:

$$\mathcal{P}_L(\mathbf{Y} = Y) \propto \det(L_Y).$$

- No need $L \preceq I$.
- For modeling data, this parameterization is more convenient.

To get the normalizer,

Theorem (2.1)

For any $A \subseteq \mathcal{Y}$,

$$\sum_{A \subseteq Y \subseteq \mathcal{Y}} \det(L_Y) = \det(L + I_{\overline{A}}),$$

where $\overline{A} = \mathcal{Y} \backslash A$.

- The normalizer is $\sum_{Y \subseteq \mathcal{Y}} \det(L_Y) = \det(L + I)$. Complexity: $O(N^3)$.

L-ensembles

- An L-ensemble defines a DPP through a real, symmetric matrix $L \succeq \mathbf{0}$:

$$\mathcal{P}_L(\mathbf{Y} = Y) \propto \det(L_Y).$$

- No need $L \preceq I$.
- For modeling data, this parameterization is more convenient.

To get the normalizer,

Theorem (2.1)

For any $A \subseteq \mathcal{Y}$,

$$\sum_{A \subseteq Y \subseteq \mathcal{Y}} \det(L_Y) = \det(L + I_{\overline{A}}),$$

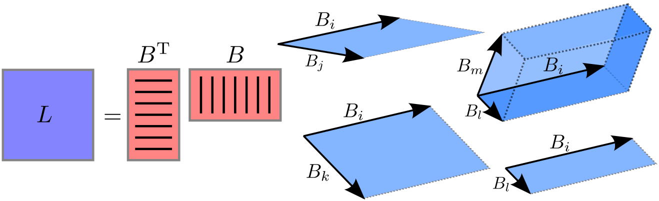
where $\overline{A} = \mathcal{Y} \backslash A$.

- The normalizer is $\sum_{Y \subseteq \mathcal{Y}} \det(L_Y) = \det(L + I)$. Complexity: $O(N^3)$.

Geometric interpretation

- Let $L = B^\top B$ for some $B = (B_1|\cdots|B_N) \in \mathbb{R}^{D\times N}$.

$$\mathcal{P}_L(\mathbf{Y} = Y) \propto \det(L_Y) = \mathrm{vol}^2(\{B_i\}_{i\in Y})$$

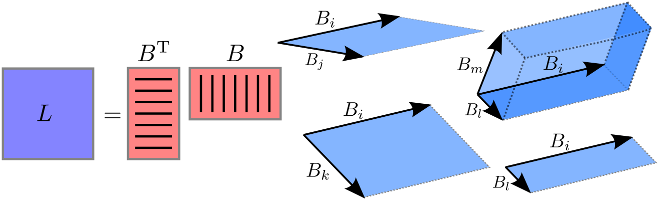


- Probability determined by the volume spanned by $\{B_i\}_{i\in Y}$.
- Diverse set $\implies$ $\approx$ orthogonal vectors $\implies$ span large volumes.
- Items with large-magnitude feature vectors (B_i) are more likely to appear.

Geometric interpretation

- Let $L = B^\top B$ for some $B = (B_1|\cdots|B_N) \in \mathbb{R}^{D\times N}$.

$$\mathcal{P}_L(\mathbf{Y} = Y) \propto \det(L_Y) = \mathrm{vol}^2(\{B_i\}_{i\in Y})$$



- Probability determined by the volume spanned by $\{B_i\}_{i\in Y}$.
- Diverse set $\implies$ $\approx$ orthogonal vectors $\implies$ span large volumes.
- Items with large-magnitude feature vectors (B_i) are more likely to appear.

Inference: normalization & marginalization

- Assume $L = \sum_{n=1}^{N} \boldsymbol{\lambda}_n \mathbf{v}_n \mathbf{v}_n^\top = VDV^\top$.

Normalizer: $\det(L+I)$.

- Then,

$$\det(L+I) = \det(VDV^\top + VV^\top)$$
$$= \det(V)\det(D+I)\det(V^\top) = \prod_{n=1}^{N}(\lambda_n + 1).$$

Marginalization: (get $\mathcal{P}(A \subseteq \mathbf{Y})$ from $\mathcal{P}_L(\mathbf{Y} = Y)$)

Theorem (2.2)

An L-ensemble is a DPP with marginal kernel

$$K = I - (L+I)^{-1} = L(L+I)^{-1}$$
$$= VDV^\top\left[V(D+I)^{-1}V^\top\right] = \sum_{n=1}^{N} \frac{\lambda_n}{\lambda_n + 1}\mathbf{v}_n\mathbf{v}_n^\top.$$

Inference: normalization & marginalization

- Assume $L = \sum_{n=1}^{N} \boldsymbol{\lambda}_n \mathbf{v}_n \mathbf{v}_n^\top = VDV^\top$.

Normalizer: $\det(L + I)$.

- Then,

$$\det(L+I) = \det(VDV^\top + VV^\top)$$
$$= \det(V)\det(D+I)\det(V^\top) = \prod_{n=1}^{N}(\lambda_n + 1).$$

Marginalization: (get $\mathcal{P}(A \subseteq \mathbf{Y})$ from $\mathcal{P}_L(\mathbf{Y} = Y)$)

Theorem (2.2)

An L-ensemble is a DPP with marginal kernel

$$K = I - (L+I)^{-1} = L(L+I)^{-1}$$
$$= VDV^\top \left[V(D+I)^{-1}V^\top\right] = \sum_{n=1}^{N} \frac{\lambda_n}{\lambda_n + 1}\mathbf{v}_n\mathbf{v}_n^\top.$$

Sampling from a DPP

- Let $L = \sum_{n=1}^{N} \boldsymbol{\lambda}_n \mathbf{v}_n \mathbf{v}_n^\top$. Let $I = (\mathbf{e}_1 | \dots | \mathbf{e}_N)$.

1 $J \leftarrow \emptyset$

2 for $n = 1, \dots, N$:

- $J \leftarrow J \cup \{n\}$ with probability $\frac{\lambda_n}{\lambda_n + 1}$

3 $V \leftarrow (\mathbf{v}_n)_{n \in J} \in \mathbb{R}^{N \times |J|}$ % $|J|$ is the number of items to sample

4 while $|V| > 0$:

  1 $\mathbf{p} := \left[\frac{1}{|V|} \sum_{\mathbf{v} \in V} (\mathbf{v}^\top \mathbf{e}_i)^2 \right]_{i=1}^{N} = \frac{1}{|V|} \left(\| V(1,:) \|^2, \dots, \| V(N,:) \|^2 \right)$

  2 Draw $i \sim \text{Discrete}(\mathbf{p})$

  3 $Y \leftarrow Y \cup \{i\}$

  4 $V \leftarrow V_\perp$, an orthonormal basis for the subspace of V orthogonal to $\mathbf{e}_i$. (run Gram-Schmidt)

- 4.4: The dimension of V is reduced by 1.
- Runs in time $O(Nk^3)$ where $k = |J|$. Gram-Schmidt costs $O(Nk^2)$.
- Eigen-decomposition of L: $O(N^3)$ (only once).
  - Can be approximated by computing only top k eigenvectors.

Sampling from a DPP

- Let $L = \sum_{n=1}^{N} \lambda_n \mathbf{v}_n \mathbf{v}_n^\top$. Let $I = (\mathbf{e}_1 | \ldots | \mathbf{e}_N)$.

1. $J \leftarrow \emptyset$
2. for $n = 1, \ldots, N$:
   - $J \leftarrow J \cup \{n\}$ with probability $\frac{\lambda_n}{\lambda_n + 1}$
3. $V \leftarrow (\mathbf{v}_n)_{n \in J} \in \mathbb{R}^{N \times |J|}$ % $|J|$ is the number of items to sample
4. while $|V| > 0$:
   1. $\mathbf{p} := \left[\frac{1}{|V|} \sum_{\mathbf{v} \in V} (\mathbf{v}^\top \mathbf{e}_i)^2\right]_{i=1}^{N} = \frac{1}{|V|} \left(\|V(1,:)\|^2, \ldots, \|V(N,:)\|^2\right)$
   2. Draw $i \sim \text{Discrete}(\mathbf{p})$
   3. $Y \leftarrow Y \cup \{i\}$
   4. $V \leftarrow V_\perp$, an orthonormal basis for the subspace of V orthogonal to $\mathbf{e}_i$. (run Gram-Schmidt)

- 4.4: The dimension of V is reduced by 1.
- Runs in time $O(Nk^3)$ where $k = |J|$. Gram-Schmidt costs $O(Nk^2)$.
- Eigen-decomposition of L: $O(N^3)$ (only once).
  - Can be approximated by computing only top k eigenvectors.

Sampling from a DPP

- Let $L = \sum_{n=1}^{N} \lambda_n \mathbf{v}_n \mathbf{v}_n^\top$. Let $I = (\mathbf{e}_1 | \ldots | \mathbf{e}_N)$.

1 $J \leftarrow \emptyset$

2 for $n = 1, \ldots, N$:

- $J \leftarrow J \cup \{n\}$ with probability $\frac{\lambda_n}{\lambda_n + 1}$

3 $V \leftarrow (\mathbf{v}_n)_{n \in J} \in \mathbb{R}^{N \times |J|}$ % $|J|$ is the number of items to sample

4 while $|V| > 0$:

  1 $\mathbf{p} := \left[\frac{1}{|V|} \sum_{\mathbf{v} \in V} (\mathbf{v}^\top \mathbf{e}_i)^2\right]_{i=1}^{N} = \frac{1}{|V|} \left(\|V(1,:)\|^2, \ldots, \|V(N,:)\|^2\right)$

  2 Draw $i \sim \text{Discrete}(\mathbf{p})$

  3 $Y \leftarrow Y \cup \{i\}$

  4 $V \leftarrow V_\perp$, an orthonormal basis for the subspace of V orthogonal to $\mathbf{e}_i$. (run Gram-Schmidt)

- 4.4: The dimension of V is reduced by 1.
- Runs in time $O(Nk^3)$ where $k = |J|$. Gram-Schmidt costs $O(Nk^2)$.
- Eigen-decomposition of L: $O(N^3)$ (only once).
  - Can be approximated by computing only top k eigenvectors.

Sampling from a DPP

- Let $L = \sum_{n=1}^{N} \lambda_n \mathbf{v}_n \mathbf{v}_n^\top$. Let $I = (\mathbf{e}_1 | \dots | \mathbf{e}_N)$.

1 $J \leftarrow \emptyset$
2 for $n = 1, \dots, N$:
  - $J \leftarrow J \cup \{n\}$ with probability $\frac{\lambda_n}{\lambda_n + 1}$
3 $V \leftarrow (\mathbf{v}_n)_{n \in J} \in \mathbb{R}^{N \times |J|}$ % $|J|$ is the number of items to sample
4 while $|V| > 0$:
  1 $\mathbf{p} := \left[\frac{1}{|V|} \sum_{\mathbf{v} \in V} (\mathbf{v}^\top \mathbf{e}_i)^2 \right]_{i=1}^{N} = \frac{1}{|V|} \left(\| V(1,:) \|^2, \dots, \| V(N,:) \|^2 \right)$
  2 Draw $i \sim \text{Discrete}(\mathbf{p})$
  3 $Y \leftarrow Y \cup \{i\}$
  4 $V \leftarrow V_\perp$, an orthonormal basis for the subspace of V orthogonal to $\mathbf{e}_i$. (run Gram-Schmidt)

- 4.4: The dimension of V is reduced by 1.
- Runs in time $O(Nk^3)$ where $k = |J|$. Gram-Schmidt costs $O(Nk^2)$.
- Eigen-decomposition of L: $O(N^3)$ (only once).
  - Can be approximated by computing only top k eigenvectors.

Sampling from a DPP

- Let $L = \sum_{n=1}^{N} \lambda_n \mathbf{v}_n \mathbf{v}_n^\top$. Let $I = (\mathbf{e}_1 | \ldots | \mathbf{e}_N)$.

1 $J \leftarrow \emptyset$

2 for $n = 1, \ldots, N$:

- $J \leftarrow J \cup \{n\}$ with probability $\frac{\lambda_n}{\lambda_n + 1}$

3 $V \leftarrow (\mathbf{v}_n)_{n \in J} \in \mathbb{R}^{N \times |J|}$ % $|J|$ is the number of items to sample

4 while $|V| > 0$:

   1 $\mathbf{p} := \left[\frac{1}{|V|} \sum_{\mathbf{v} \in V} (\mathbf{v}^\top \mathbf{e}_i)^2\right]_{i=1}^{N} = \frac{1}{|V|} \left(\| V(1,:)\|^2, \ldots, \| V(N,:)\|^2\right)$

   2 Draw $i \sim \text{Discrete}(\mathbf{p})$

   3 $Y \leftarrow Y \cup \{i\}$

   4 $V \leftarrow V_\perp$, an orthonormal basis for the subspace of V orthogonal to $\mathbf{e}_i$. (run Gram-Schmidt)

- 4.4: The dimension of V is reduced by 1.
- Runs in time $O(Nk^3)$ where $k = |J|$. Gram-Schmidt costs $O(Nk^2)$.
- Eigen-decomposition of L: $O(N^3)$ (only once).
  - Can be approximated by computing only top k eigenvectors.

Sampling from a DPP

- Let $L = \sum_{n=1}^{N} \lambda_n \mathbf{v}_n \mathbf{v}_n^\top$. Let $I = (\mathbf{e}_1 | \dots | \mathbf{e}_N)$.

1 $J \leftarrow \emptyset$

2 for $n = 1, \dots, N$:

- $J \leftarrow J \cup \{n\}$ with probability $\frac{\lambda_n}{\lambda_n + 1}$

3 $V \leftarrow (\mathbf{v}_n)_{n \in J} \in \mathbb{R}^{N \times |J|}$ % $|J|$ is the number of items to sample

4 while $|V| > 0$:

    1 $\mathbf{p} := \left[\frac{1}{|V|} \sum_{\mathbf{v} \in V} (\mathbf{v}^\top \mathbf{e}_i)^2 \right]_{i=1}^{N} = \frac{1}{|V|} \left(\| V(1,:) \|^2, \dots, \| V(N,:) \|^2 \right)$

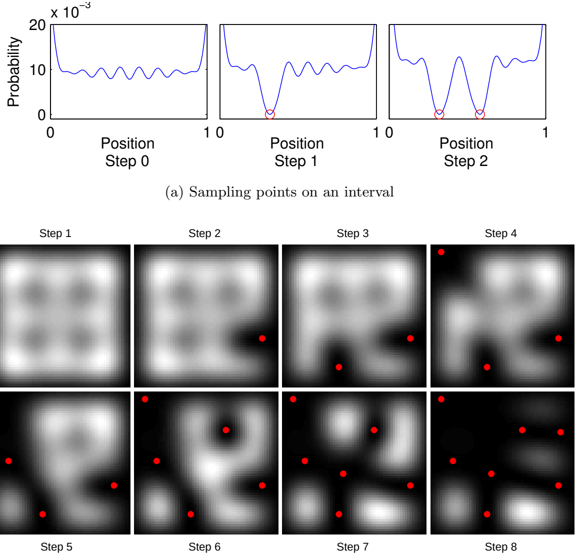
    2 Draw $i \sim \mathrm{Discrete}(\mathbf{p})$

    3 $Y \leftarrow Y \cup \{i\}$

    4 $V \leftarrow V_\perp$, an orthonormal basis for the subspace of V orthogonal to $\mathbf{e}_i$. (run Gram-Schmidt)

- 4.4: The dimension of V is reduced by 1.
- Runs in time $O(Nk^3)$ where $k = |J|$. Gram-Schmidt costs $O(Nk^2)$.
- Eigen-decomposition of L: $O(N^3)$ (only once).
  - Can be approximated by computing only top k eigenvectors.

Visualization of the sampling process



(a) Sampling points on an interval

(b) Sampling points in the plane

- Discrete $[0, 1]$ (2D plane). Show $\mathbf{p}$ in step 4.1.
- Diversifying.

Property 6: cardinality

- Recall $L = \sum_{n=1}^{N} \boldsymbol{\lambda}_n \mathbf{v}_n \mathbf{v}_n^\top$ and $K = \sum_{n=1}^{N} \frac{\lambda_n}{\lambda_n+1} \mathbf{v}_n \mathbf{v}_n^\top$.
- Let $h_n \sim \text{Bernoulli}\left(\frac{\lambda_n}{\lambda_n+1}\right)$. $h_n \in \{0, 1\}$.
- Then, $|\mathbf{Y}| = \sum_{n=1}^{N} h_n$.
  - Follows from step 2 of the sampling procedure.

Consequences

1. $|\mathbf{Y}| \leq \text{rank}(L)$ because $\text{rank}(L) = \#\text{nonzero } \lambda_n$.
2. $\mathbb{E}[|\mathbf{Y}|] = \sum_{n=1}^{N} \frac{\lambda_n}{\lambda_{n1}+1} = \text{tr}(K)$.
3. $\mathbb{V}[|\mathbf{Y}|] = \sum_{n=1}^{N} \frac{\lambda_n}{\lambda_{n1}+1}\left(1 - \frac{\lambda_n}{\lambda_{n1}+1}\right) = \sum_{n=1}^{N} \frac{\lambda_n}{(\lambda_n+1)^2}$

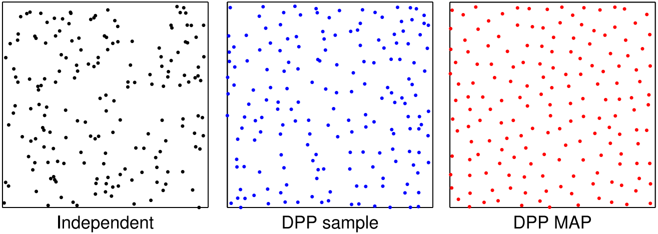
Property 6: cardinality

- Recall $L = \sum_{n=1}^{N} \boldsymbol{\lambda}_n \mathbf{v}_n \mathbf{v}_n^\top$ and $K = \sum_{n=1}^{N} \frac{\lambda_n}{\lambda_n+1} \mathbf{v}_n \mathbf{v}_n^\top$.
- Let $h_n \sim \text{Bernoulli}\left(\frac{\lambda_n}{\lambda_n+1}\right)$. $h_n \in \{0, 1\}$.
- Then, $|\mathbf{Y}| = \sum_{n=1}^{N} h_n$.
  - Follows from step 2 of the sampling procedure.

Consequences

1. $|\mathbf{Y}| \leq \text{rank}(L)$ because $\text{rank}(L) = \#\text{nonzero } \boldsymbol{\lambda}_n$.
2. $\mathbb{E}[|\mathbf{Y}|] = \sum_{n=1}^{N} \frac{\lambda_n}{\lambda_{n1}+1} = \text{tr}(K)$.
3. $\mathbb{V}[|\mathbf{Y}|] = \sum_{n=1}^{N} \frac{\lambda_n}{\lambda_{n1}+1} \left(1 - \frac{\lambda_n}{\lambda_{n1}+1}\right) = \sum_{n=1}^{N} \frac{\lambda_n}{(\lambda_n+1)^2}$

Finding the mode



Finding the set $Y \subseteq \mathcal{Y}$ that maximizes $\mathcal{P}_L(Y)$ is NP-hard.

Submodularity: $\mathcal{P}_L$ is log-submodular, that is,

$$\log \mathcal{P}_L(Y \cup \{i\}) - \log \mathcal{P}_L(Y) \geq \log \mathcal{P}_L(Y' \cup \{i\}) - \log \mathcal{P}_L(Y'),$$

whenever $Y \subseteq Y' \subseteq \mathcal{Y} - \{i\}$.

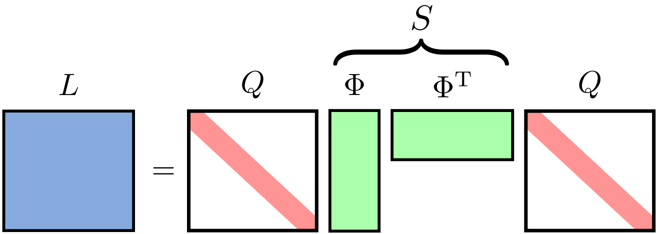
Many results exists for approximately maximizing monotone submodular functions. However, $\mathcal{P}_L$ is highly non-monotone! In practice, this is not a problem [Kulesza et al., 2012].

DPP decomposition: quality vs diversity

We can take the notation $L = B^{\mathrm{T}}B$ one step further.

Each column B_i satisfies $B_i = q_i \phi_i$, where

- $q_i \in \mathbb{R}^+$ is a quality term.
- $\phi_i \in \mathbb{R}^D$, $||\phi_i|| = 1$ is a vector of diversity features.



We now have $\mathcal{P}_L(Y) \propto \left[\prod_{i\in Y} q_i^2\right] \det(S_Y)$.

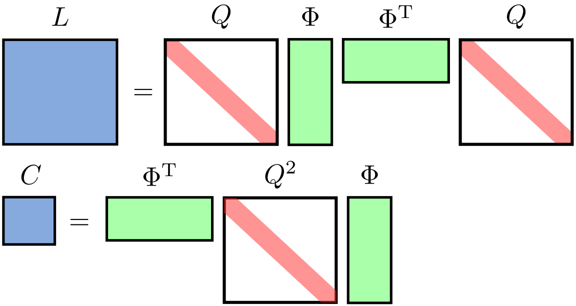
The first factor increases with the quality of the items in Y.

The second factor increases with the diversity of the items in Y.

Dual representation I

Most algorithms require manipulating L through inversion, eigendecomposition, etc...

When N is very large, directly working with the $N \times N$ matrix L is not efficient.



Let B be the $D \times N$ matrix with $B_i = q_i \phi_i$ so that $L = B^{\mathrm{T}} B$. Instead, we work with the $D \times D$ matrix $C = BB^{\mathrm{T}}$.

Dual representation II

- C and L have the same (non-zero) eigenvalues.
- Their eigenvectors are linearly related.
- Working with C scales as a function of $D \ll N$.

Proposition:

$$C = BB^{\mathrm{T}} = \sum_{n=1}^{D} \lambda_n \hat{\mathbf{v}}_n \hat{\mathbf{v}}_n^{\mathrm{T}}$$

is an eigendecomposition of C if and only if

$$L = B^{\mathrm{T}}B = \sum_{n=1}^{D} \lambda_n \left[\frac{1}{\sqrt{\lambda_n}} B^{\mathrm{T}} \hat{\mathbf{v}}_n\right] \left[\frac{1}{\sqrt{\lambda_n}} B^{\mathrm{T}} \hat{\mathbf{v}}_n\right]^{\mathrm{T}}$$

is an eigendecomposition of L.

- C is sufficient to perform nearly all forms of DPP inference efficiently.

Inference in the dual form

- Assume $L = B^\top B \in \mathbb{R}^{N\times N}$ and $C = BB^\top = \sum_{n=1}^{D} \boldsymbol{\lambda} \hat{\mathbf{v}}_n \hat{\mathbf{v}}_n^\top \in \mathbb{R}^{D\times D}$

Normalization:

$$\det(L+I) = \prod_{n=1}^{D} (\lambda_n + 1) = \det(C+I).$$

Marginalization:

- Recall $L = \sum_{n=1}^{D} \lambda_n \left[\frac{1}{\lambda_n} B^\top \hat{\mathbf{v}}_n\right] \left[\frac{1}{\lambda_n} B^\top \hat{\mathbf{v}}_n\right]^\top$.
- So,

$$\begin{aligned} K_{ij} &= \sum_{n=1}^{D} \frac{\lambda_n}{\lambda_n+1} \left[\frac{1}{\lambda_n} B_i^\top \hat{\mathbf{v}}_n\right] \left[\frac{1}{\lambda_n} B_j^\top \hat{\mathbf{v}}_n\right]^\top \\ &= q_i q_j \sum_{n=1}^{D} \frac{1}{\lambda_n+1} (\phi_i^\top \hat{\mathbf{v}}_n)(\phi_j^\top \hat{\mathbf{v}}_n), \end{aligned}$$

  which can be computed in $O(D^2)$ time.
- Say $A \in \mathcal{Y}$ such that $|A| = k$. Then, $\mathcal{P}(A \subseteq \mathbf{Y}) = \det(K_A)$ can be computed in $O(D^2 k^2 + k^3)$.

Other interesting things we did not discuss

- Proofs of all the results.
- Related processes: Poisson point processes, Matern repulsive, random sequential adsorption.
- Decomposing DPPs into elementary DPPs.
- Random projections approximately preserve volumes. Can be used to reduce D. Faster.
- Supervised learning with conditional DPPs $\mathcal{P}(\mathbf{Y} = Y|X) \propto \det(L_Y(X))$.
- k-DPPs: a distribution over all subsets $Y \subseteq \mathbf{Y}$ with cardinality k.
- Learning θ for K_θ?

Questions?

Thank you

References I

- Some materials from https://jmhldotorg.files.wordpress.com/2014/02/slidesrcc-dpps.pdf.