Further Explanation on Determinantal Point Process

Richard Xu

July 16, 2022

1 What is DPP?

Most of this note is based on the original DPP paper [1]. The original paper is very detailed and well written. However, there may be some points that need further clarification, especially for students lacking linear algebra skill. Therefore, I hope to explain them in a slightly simpler way (hopefully). Please read the original paper for more details.

1.1 definition of marginal DPP distribution

Start with a **marginal** distribution:

$$\Pr(A \subseteq \mathbf{Y}) = \det(K_A) \tag{1}$$

An example: given $\Omega = \{1,2,3,4,5\}$, $A = \{1,2,3\}$ and $\mathbf{Y} \in \Omega$

$$\begin{aligned}
\Pr(A \subseteq \mathbf{Y}) &= \Pr(\{1,2,3\} \subseteq \mathbf{Y}) \\
&\equiv \Pr_K(y_1 = 1, y_2 = 1, y_3 = 1) \\
&= \sum_{t_4=0}^{1} \sum_{t_5=0}^{1} \Pr(y_1 = 1, y_2 = 1, y_3 = 1, y_4 = t_4, y_5 = t_5) \\
&= \det(K_A)
\end{aligned} \tag{2}$$

note that $\Pr(A \in \mathbf{Y})$ is analogous to $\Pr(X = x)$ for marginal DPP.

1.2 Something about marginal distribution

1. $\Pr(A \subseteq \mathbf{Y})$ is marginal, so $\Pr(A_1 \subseteq \mathbf{Y}) + \Pr(A_2 \subseteq \mathbf{Y}) + \dots$ don't need to add to 1, i.e., it may be possible that: $\Pr(A_1 \subseteq \mathbf{Y}) + \Pr(A_2 \subseteq \mathbf{Y}) > 1$
2. $\Pr(\emptyset \subseteq \mathbf{Y}) = \det(K_\emptyset) = 1$ This is obvious, as any $\mathbf{Y}$ is a superset of $\emptyset$.
3. $\Pr(i \subseteq \mathbf{Y}) = \det(K_{ii}) = K_{ii}$
4. however, its property is best determined from two elements case:

$$\begin{aligned}
\Pr(i, j \in \mathbf{Y}) &= \begin{vmatrix} K_{ii} & K_{ij} \\ K_{ji} & K_{jj} \end{vmatrix} \\
&= K_{ii}K_{jj} - K_{ij}K_{ji} \\
&= \Pr(i \subseteq \mathbf{Y}) \Pr(j \subseteq \mathbf{Y}) - K_{ij}^2
\end{aligned} \tag{3}$$

By convention, off-diagonal elements determine negative correlations between pairs.

Large absolute values of $K_{i,j}$ imply that the probability that i^{th} and j^{th} elements are both selected tend to have **low** density.

1.2.1 Example of K

Any $K, 0 \preceq K \preceq I$ defines a DPP.

If $A \preceq B$, that is, $B - A$ is positive semi-definite.

1.2.2 where K does not define DPP

example $K = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$ does **not** define DPP, we check if $K \preceq I$?

$$\begin{aligned} I - \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix} &= \begin{pmatrix} 0 & 0.5 \\ 0.5 & 0 \end{pmatrix} \\ \implies \bar{\lambda}(K) &= [-0.5, 0.5]^\top \end{aligned} \tag{4}$$

Another way to see the above is incorrect, where we let $\Omega = \{1, 2\}$:

$$\begin{aligned} \Pr\big(\{1\} \subseteq \mathbf{Y}\big) &\equiv \Pr\big((\mathbf{Y} = \{1\}) \cup (\mathbf{Y} = \{1, 2\})\big) \\ &= \det(K_1) = 1 \end{aligned} \tag{5}$$

$$\begin{aligned} \Pr\big(\{2\} \subseteq \mathbf{Y}\big) &\equiv \Pr\big((\mathbf{Y} = \{2\}) \cup (\mathbf{Y} = \{1, 2\})\big) \\ &= \det(K_2) = 1 \end{aligned} \tag{6}$$

note LHS uses $\subseteq$ and RHS uses $=$. However:

$$\begin{aligned} \Pr(\{1, 2\} \subseteq \mathbf{Y}) &\equiv \Pr\big(\mathbf{Y} = \{1, 2\}\big) \\ &= \det(K_{\{1,2\}}) = 0.75 \end{aligned} \tag{7}$$

1. The first two equation says $\{1\}$ and $\{2\}$ must be included
2. The third equation says both may NOT always be included

1.2.3 Example of K define DPP

example $K = \begin{bmatrix} 0.3 & -0.1 \\ -0.1 & 0.4 \end{bmatrix}$ **does** define DPP:

$$\begin{aligned} I - \begin{pmatrix} 0.3 & -0.1 \\ -0.1 & 0.4 \end{pmatrix} &= \begin{pmatrix} 0.7 & 0.1 \\ 0.1 & 0.6 \end{pmatrix} \\ \implies \bar{\lambda}(K) &= [0.5382, 0.7618]^\top \end{aligned} \tag{8}$$

$$\begin{aligned} \Pr\big(\{1\} \subseteq \mathbf{Y}\big) &\equiv \Pr\big((\mathbf{Y} = \{1\}) \cup (\mathbf{Y} = \{1, 2\})\big) \\ &= \det(K_1) = 0.3 \end{aligned} \tag{9}$$

$$\begin{aligned} \Pr\big(\{2\} \subseteq \mathbf{Y}\big) &\equiv \Pr\big((\mathbf{Y} = \{2\}) \cup (\mathbf{Y} = \{1, 2\})\big) \\ &= \det(K_2) = 0.4 \end{aligned} \tag{10}$$

$$\begin{aligned}\Pr(\{1,2\} \subseteq \mathbf{Y}) &\equiv \Pr\big(\mathbf{Y} = \{1,2\}\big) \\ &= \det(K_{\{1,2\}}) = 0.11\end{aligned} \tag{11}$$

the event:

$$\begin{aligned}\Pr\big(\{1,2\} \subseteq \mathbf{Y}\big) &\equiv \Pr\big(\{1,2\} = \mathbf{Y}\big) \\ &= \Pr\big(\{1\} \subseteq \mathbf{Y}\big) \cap \big(\{2\} \subseteq \mathbf{Y}\big)\end{aligned} \tag{12}$$

$$\begin{aligned}\Pr\big((\{1\} = \mathbf{Y}) \cup (\{2\} = \mathbf{Y})\big) &= \Pr\big(\{1\} \subseteq \mathbf{Y}\big) + \Pr\big(\{2\} \subseteq \mathbf{Y}\big) - \Pr(\{1,2\} \subseteq \mathbf{Y}) \\ &= 0.3 + 0.4 - 0.11 \\ &= 0.59\end{aligned} \tag{13}$$

what about the probability of selecting **exactly** the $\emptyset$?

$$\begin{aligned}\Pr(\mathbf{Y} = \emptyset) &\equiv 1 - \Pr\big((\{1\} = \mathbf{Y}) \cup (\{2\} = \mathbf{Y})\big) \\ &= 0.41\end{aligned} \tag{14}$$

2 L-Ensembles

Marginal distributions does **not** define probabilities in terms of a **particular** set directly, i.e., instead of having $\Pr(\mathbf{Y} \subseteq Y)$, we want $\Pr(\mathbf{Y} = Y)$:

$$\Pr_L(\mathbf{Y} = Y) \propto \det(L_Y) \tag{15}$$

L must be positive semi-definite.
Only a statement of proportionality, eigenvalues of L is **not** < 1

2.1 Geometry interpretation

$$\begin{aligned} X = [x_1 \quad x_2 \quad \dots \quad x_n] &\implies \\ L(x_1, \dots, x_n) = X^\top X &= \begin{pmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle & \dots & \langle x_1, x_n \rangle \\ \langle x_2, x_1 \rangle & \langle x_2, x_2 \rangle & \dots & \langle x_2, x_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x_n, x_1 \rangle & \langle x_n, x_2 \rangle & \dots & \langle x_n, x_n \rangle \end{pmatrix}\end{aligned} \tag{16}$$

Gram determinant is the square of the volume of the parallelotope formed by the vectors
vectors are linearly independent if and only if the Gram determinant is non-zero:

$$\Pr_L(Y) \propto \det(L_Y) = \text{Vol}^2\big(\{x_i\}_{i \in Y}\big) \tag{17}$$

note that the volume is span in dimensions of data $\{x_i \in \mathbb{R}^d\}$, not in the dimension of the gram matrix itself.

2.2 Proof for the Geometry interpretation

2.2.1 in 1-element case

$\text{Vol}^2(\mathbf{u}_1) = \mathbf{u}_1^\top \mathbf{u}_1$, i.e., length square of a line

2.2.2 in k-element case

$$\text{Vol}^2(\mathbf{u}_1 \dots \mathbf{u}_k, \mathbf{u}_{k+1}) = \text{Vol}^2(\mathbf{u}_1, \dots, \mathbf{u}_k)\|\tilde{\mathbf{u}}_{k+1}\|^2 \tag{18}$$

$\tilde{\mathbf{u}}_{k+1}$ is the orthogonal projection of $\mathbf{u}_{k+1}$ onto span $(\mathbf{u}_1, \dots, \mathbf{u}_k)$:
Let $(\mathbf{u}_1, \dots, \mathbf{u}_k)$ is an $n \times k$ matrix $\mathbf{Y}$:
Then there exists a vector $\mathbf{c} \in \mathbb{R}^k$ such that:

$$\begin{aligned}
\underbrace{\mathbf{u}_{k+1}}_{\text{normal}} &= \mathbf{U}\mathbf{c} + \underbrace{\tilde{\mathbf{u}}_{k+1}}_{\text{orthogonalized}} \qquad \text{split } \mathbf{u}_{k+1} \text{ into } \parallel \text{ and } \perp \text{ components regarding span } (\mathbf{u}_1, \dots, \mathbf{u}_k) \\
&= \underbrace{\begin{bmatrix} | & \vdots & | \\ \mathbf{u}_1 & \vdots & \mathbf{u}_k \\ | & \vdots & | \end{bmatrix}}_{\mathbf{U}} \begin{bmatrix} c_1 \\ \vdots \\ c_k \end{bmatrix} + \tilde{\mathbf{u}}_{k+1} \qquad \text{or } \mathbf{u}_{k+1} = c_1\mathbf{u}_1 + c_2\mathbf{u}_2 \dots c_k\mathbf{u}_k \quad + \tilde{\mathbf{u}}_{k+1}
\end{aligned} \tag{19}$$

extending $\mathbf{U} \to \mathbf{X}$ by adding one more column $\mathbf{u}_{k+1}$:

$$\begin{aligned}
\mathbf{X} &= [\mathbf{U} \quad \mathbf{u}_{k+1}] = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \dots \quad \mathbf{u}_k \quad \mathbf{u}_{k+1}] = [\mathbf{U} \quad \mathbf{U}\mathbf{c} + \tilde{\mathbf{u}}_{k+1}] \\
\implies \mathbf{X}^\top\mathbf{X} &= \begin{bmatrix} \mathbf{U}^\top\mathbf{U} & \mathbf{U}^\top\mathbf{u}_{k+1} \\ \mathbf{u}_{k+1}^\top\mathbf{U} & \mathbf{u}_{k+1}^\top\mathbf{u}_{k+1} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{U}^\top\mathbf{U} & \mathbf{U}^\top(\mathbf{U}\mathbf{c} + \tilde{\mathbf{u}}_{k+1}) \\ (\mathbf{U}\mathbf{c} + \tilde{\mathbf{u}}_{k+1})^\top\mathbf{U} & (\mathbf{U}\mathbf{c} + \tilde{\mathbf{u}}_{k+1})^\top(\mathbf{U}\mathbf{c} + \tilde{\mathbf{u}}_{k+1}) \end{bmatrix} \qquad \text{using } \mathbf{u}_{k+1} = \mathbf{U}\mathbf{c} + \tilde{\mathbf{u}}_{k+1} \\
&= \begin{bmatrix} \mathbf{U}^\top\mathbf{U} & \mathbf{U}^\top\mathbf{U}\mathbf{c} \\ \mathbf{c}^\top\mathbf{U}^\top\mathbf{U} & \mathbf{c}^\top\mathbf{U}^\top\mathbf{U}\mathbf{c} + \tilde{\mathbf{u}}_{k+1}^\top\tilde{\mathbf{u}}_{k+1} \end{bmatrix} \qquad \text{since } \mathbf{U}^\top\tilde{\mathbf{u}}_{k+1} = \mathbf{0} \\
&= \begin{bmatrix} \mathbf{U}^\top\mathbf{U} & \mathbf{U}^\top\mathbf{U}\mathbf{c} \\ \mathbf{c}^\top\mathbf{U}^\top\mathbf{U} & \mathbf{c}^\top\mathbf{U}^\top\mathbf{U}\mathbf{c} + \|\tilde{\mathbf{u}}_{k+1}\|^2 \end{bmatrix} \\
&= \left[\begin{bmatrix} \mathbf{U}^\top\mathbf{U} \\ \mathbf{c}^\top\mathbf{U}^\top\mathbf{U} \end{bmatrix} \quad \left(\begin{bmatrix} \mathbf{U}^\top\mathbf{U}\mathbf{c} \\ \mathbf{c}^\top\mathbf{U}^\top\mathbf{U}\mathbf{c} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \|\tilde{\mathbf{u}}_{k+1}\|^2 \end{bmatrix} \right) \right]
\end{aligned} \tag{20}$$

$$\begin{aligned}
\det([a_1 + b_1, a_2, \dots, a_k]) &= \det\big([a_1, a_2, \dots, a_k]\big) + \det\big([b_1, a_2, \dots, a_k]\big) \qquad \text{using } \textbf{Multi-linearity} \\
\implies \det\left(\mathbf{X}^\top\mathbf{X}\right) &= \det\left(\left[\begin{bmatrix} \mathbf{U}^\top\mathbf{U} \\ \mathbf{c}^\top\mathbf{U}^\top\mathbf{U} \end{bmatrix} \quad \left(\begin{bmatrix} \mathbf{U}^\top\mathbf{U}\mathbf{c} \\ \mathbf{c}^\top\mathbf{U}^\top\mathbf{U}\mathbf{c} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \|\tilde{\mathbf{u}}_{k+1}\|^2 \end{bmatrix} \right) \right] \right) \\
&= \det\left(\begin{bmatrix} \mathbf{U}^\top\mathbf{U} & \mathbf{U}^\top\mathbf{U}\mathbf{c} \\ \mathbf{c}^\top\mathbf{U}^\top\mathbf{U} & \mathbf{c}^\top\mathbf{U}^\top\mathbf{U}\mathbf{c} \end{bmatrix} \right) + \det\left(\begin{bmatrix} \mathbf{U}^\top\mathbf{U} & \mathbf{0} \\ \mathbf{c}^\top\mathbf{U}^\top\mathbf{U} & \|\tilde{\mathbf{u}}_{k+1}\|^2 \end{bmatrix} \right) \\
&= \mathbf{0} + \det\left(\begin{bmatrix} \mathbf{U}^\top\mathbf{U} & \mathbf{0} \\ \mathbf{c}^\top\mathbf{U}^\top\mathbf{U} & \|\tilde{\mathbf{u}}_{k+1}\|^2 \end{bmatrix} \right) \\
&= \det\left(\left[\mathbf{U}^\top\mathbf{U}\right] \right)\|\tilde{\mathbf{u}}_{k+1}\|^2 \\
&= \det\left(\left[\mathbf{U}^\top\mathbf{U}\right] \right)\text{Vol}^2(\tilde{\mathbf{u}}_{k+1})
\end{aligned} \tag{21}$$

2.3 Normalization constant in L-Ensembles

without proof, stating the **Theorem** says:

Theorem 1

$$\sum_{A \subseteq Y \subseteq \Omega} \det(L_Y) = \det(L + \mathbf{I}_{\bar{A}}) \tag{22}$$

2.3.1 2-element example

from this, it can be easily understood by multilinear rule:

$$\begin{aligned}
L &= \begin{pmatrix} 3.0 & 1.0 \\ 1.5 & 1.2 \end{pmatrix} \\
\det(L + \mathbf{I}) &= \det\left(\begin{bmatrix} 3.0 + 1 & 1.0 + 0 \\ 1.5 + 0 & 1.2 + 1 \end{bmatrix} \right) \\
&= \det\left(\begin{bmatrix} 3.0 & 1.0 + 0 \\ 1.5 & 1.2 + 1 \end{bmatrix} \right) + \det\left(\begin{bmatrix} 1 & 1.0 + 0 \\ 0 & 1.2 + 1 \end{bmatrix} \right) \\
&= \det\left(\begin{bmatrix} 3.0 & 1.0 \\ 1.5 & 1.2 \end{bmatrix} \right) + \det\left(\begin{bmatrix} 3.0 & 0 \\ 1.5 & 1 \end{bmatrix} \right) + \det\left(\begin{bmatrix} 1 & 1.0 \\ 0 & 1.2 \end{bmatrix} \right) + \det\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \\
&= \det\left(\begin{bmatrix} 3.0 & 1.0 \\ 1.5 & 1.2 \end{bmatrix} \right) + \det\big([3.0] \big) + \det\big([1.2] \big) + \det\big(\mathbf{I}\big) \\
&= \underbrace{\det\big(L\big)}_{\{1,2\}} + \underbrace{\det\big([3.0] \big)}_{\{1\}} + \underbrace{\det\big([1.2] \big)}_{\{2\}} + \underbrace{1}_{\emptyset}
\end{aligned} \tag{23}$$

Note that unless we are interested to compute $\sum_{A \subseteq Y} L_Y$ where $A = \emptyset$, we will **not** have the term $\det(\mathbf{I}) = 1$. This is not suprising, as $\sum_{A \subseteq Y} L_Y$ terms do not contain $\emptyset$. from determinant computation point of view, multilinear rule will not result to a full I, there will be some zeros.

1. those include $\{1\}$

$$\begin{aligned}
\sum_{\{1\} \subseteq Y} \det(L_Y) &= \underbrace{\det\big(L\big)}_{\{1,2\}} + \underbrace{\det\big([3.0] \big)}_{\{1\}} \quad \text{by derivation} \\
&= \det(L + \mathbf{I}_{\{\bar{1}\}}) \quad \text{by theorem 1} \\
&= \det\left(\begin{bmatrix} 3.0 + 0 & 1.0 + 0 \\ 1.5 + 0 & 1.2 + 1 \end{bmatrix} \right)
\end{aligned} \tag{24}$$

2. those include $\{2\}$

$$\begin{aligned}
\sum_{\{2\} \subseteq Y} \det(L_Y) &= \underbrace{\det\big(L\big)}_{\{1,2\}} + \underbrace{\det\big([1.2] \big)}_{\{2\}} \quad \text{by derivation} \\
&= \det(L + \mathbf{I}_{\{\bar{2}\}}) \quad \text{by theorem 1} \\
&= \det\left(\begin{bmatrix} 3.0 + 1 & 1.0 + 0 \\ 1.5 + 0 & 1.2 + 0 \end{bmatrix} \right)
\end{aligned} \tag{25}$$

2.3.2 3-element example

$$L = \begin{bmatrix} 2.8 & 4.9 & 1.8 \\ 4.9 & 2.6 & 1.1 \\ 1.8 & 1.1 & 2.0 \end{bmatrix} \tag{26}$$

1. $A = \{1, 2\} \implies \bar{A} = \{3\} \implies \mathbf{I}_{\bar{A}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

$$\begin{aligned} \sum_{\{1,2\} \subseteq Y} \det(L_Y) &= \det\left(\begin{bmatrix} 2.8 & 4.9 & 1.8 \\ 4.9 & 2.6 & 1.1 \\ 1.8 & 1.1 & 2.0 + 1.0 \end{bmatrix} \right) \\ &= \det\left(\begin{bmatrix} 2.8 & 4.9 & 1.8 \\ 4.9 & 2.6 & 1.1 \\ 1.8 & 1.1 & 2.0 \end{bmatrix} \right) + \det\left(\begin{bmatrix} 2.8 & 4.9 & 0 \\ 4.9 & 2.6 & 0 \\ 1.8 & 1.1 & 1.0 \end{bmatrix} \right) \\ &= \underbrace{\det\left(\begin{bmatrix} 2.8 & 4.9 & 1.8 \\ 4.9 & 2.6 & 1.1 \\ 1.8 & 1.1 & 2.0 \end{bmatrix} \right)}_{\{1,2,3\}} + \underbrace{\det\left(\begin{bmatrix} 2.8 & 4.9 \\ 4.9 & 2.6 \end{bmatrix} \right)}_{\{1,2\}} \end{aligned} \tag{27}$$

2. $A = \{1\} \implies \bar{A} = \{2, 3\} \implies \mathbf{I}_{\bar{A}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

$$\begin{aligned} \sum_{\{1\} \subseteq Y} \det(L_Y) &= \det\left(\begin{bmatrix} 2.8 & 4.9 & 1.8 \\ 4.9 & 2.6 + 1.0 & 1.1 \\ 1.8 & 1.1 & 2.0 + 1.0 \end{bmatrix} \right) \\ &= \det\left(\begin{bmatrix} 2.8 & 4.9 & 1.8 \\ 4.9 & 2.6 + 1.0 & 1.1 \\ 1.8 & 1.1 & 2.0 \end{bmatrix} \right) + \det\left(\begin{bmatrix} 2.8 & 4.9 & 0 \\ 4.9 & 2.6 + 1.0 & 0 \\ 1.8 & 1.1 & 1.0 \end{bmatrix} \right) \quad \text{right most column first} \\ &= \det\left(\begin{bmatrix} 2.8 & 4.9 & 1.8 \\ 4.9 & 2.6 & 1.1 \\ 1.8 & 1.1 & 2.0 \end{bmatrix} \right) + \det\left(\begin{bmatrix} 2.8 & 0 & 1.8 \\ 4.9 & 1.0 & 1.1 \\ 1.8 & 0 & 2.0 \end{bmatrix} \right) \\ &\quad \det\left(\begin{bmatrix} 2.8 & 4.9 & 0 \\ 4.9 & 2.6 & 0 \\ 1.8 & 1.1 & 1.0 \end{bmatrix} \right) + \det\left(\begin{bmatrix} 2.8 & 0 & 0 \\ 4.9 & 1.0 & 0 \\ 1.8 & 0 & 1.0 \end{bmatrix} \right) \\ &= \underbrace{\det\left(\begin{bmatrix} 2.8 & 4.9 & 1.8 \\ 4.9 & 2.6 & 1.1 \\ 1.8 & 1.1 & 2.0 \end{bmatrix} \right)}_{\{1,2,3\}} + \underbrace{\det\left(\begin{bmatrix} 2.8 & 1.8 \\ 1.8 & 2.0 \end{bmatrix} \right)}_{\{1,2\}} + \underbrace{\det\left(\begin{bmatrix} 2.8 & 4.9 \\ 4.9 & 2.6 \end{bmatrix} \right)}_{\{1,2\}} + \underbrace{\det\left(\begin{bmatrix} 2.8 \end{bmatrix} \right)}_{\{1\}} \end{aligned} \tag{28}$$

2.3.3 let $A = \emptyset$

from Therorem , normalisation constant (or partition function) is: $\bar{\emptyset} = \Omega$:

$$\begin{aligned} \sum_{\emptyset \subseteq Y \subseteq \Omega} \det(L_Y) &= \sum_{Y \subseteq \Omega} \det(L_Y) \\ &= \det(L + \mathbf{I}_{\bar{\emptyset}}) \\ &= \det(L + \mathbf{I}_{\Omega}) \\ &= \det(L + \mathbf{I}) \end{aligned} \tag{29}$$

2.4 Conversion to Marginal distribution

since both $\Pr_L(\mathbf{Y} = Y)$ and K defines DPP, therefore we must have:

$$\Pr_L(\mathbf{Y} = Y) \propto \det(L_Y) \implies \Pr_L(\mathbf{Y} = Y) = \frac{\det(L_Y)}{\det(L_Y + I)} \tag{30}$$

An L-ensemble is a DPP, and its marginal kernel is:

$$K = L(L + I)^{-1} = I - (L + I)^{-1} \tag{31}$$

an important identity:

$$L(L + I)^{-1} = I - (L + I)^{-1} \tag{32}$$

for any L where $(L + I)^{-1}$ exist, this can be easily seen to multiply R.H.S by $(L + I)(L + I)^{-1}$:

$$\begin{aligned} &(I - (L + I)^{-1})(L + I)(L + I)^{-1} \\ &= \big((L + I) - I\big)(L + I)^{-1} \\ &= L(L + I)^{-1} \end{aligned} \tag{33}$$

$$\begin{aligned} \Pr_L(A \subseteq \mathbf{Y}) &= \frac{\sum_{A \subseteq Y \subseteq \Omega} \det(L_Y)}{\sum_{Y \subseteq \Omega} \det(L_Y)} \\ &= \frac{\det(L + I_{\bar{A}})}{\det(L + I)} \\ &= \det\big((L + I_{\bar{A}})(L + I)^{-1}\big) \quad \because \det(A^{-1}) = \frac{1}{\det(A)} \quad \det(AB) = \det(A)\det(B) \end{aligned} \tag{34}$$

$$\begin{aligned} \Pr_L(A \subseteq \mathbf{Y}) &= \det\big((L + I_{\bar{A}})(L + I)^{-1}\big) \\ &= \det\big(L(L + I)^{-1} + I_{\bar{A}}(L + I)^{-1}\big) \quad \text{expand} \\ &= \det\big(I - (L + I)^{-1} + I_{\bar{A}}(L + I)^{-1}\big) \quad \because \text{of Eq. (33)} \\ &= \det\big(I - (I - I_{\bar{A}})(L + I)^{-1}\big) \quad \text{combine last two terms together} \\ &= \det\big(I - I_A(L + I)^{-1}\big) \quad \because I_A = I - I_{\bar{A}} \\ &= \det\big((I_A + I_{\bar{A}}) - I_A(L + I)^{-1}\big) \quad \text{expanding } I = I_A + I_{\bar{A}} \\ &= \det\big(I_{\bar{A}} + \underline{I_A - I_A(L + I)^{-1}}\big) \\ &= \det\big(I_{\bar{A}} + I_A\underbrace{\big(I - (L + I)^{-1}\big)}\big) \quad \because K = I - (L + I)^{-1} \\ &= \det\big(I_{\bar{A}} + I_A K\big) \end{aligned} \tag{35}$$

left multiplication by I_A **zeros out rows** of a matrix except those corresponding to A. We split the marginal kernel matrix K into K_A and $K_{\bar{A}}$:

$$\begin{aligned} K &= \begin{pmatrix} K_{\bar{A}} & K_{\bar{A}A} \\ K_{A\bar{A}} & K_A \end{pmatrix} \\ \implies I_A K &= \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{|A| \times |A|} \end{pmatrix} \begin{pmatrix} K_{\bar{A}} & K_{\bar{A}A} \\ K_{A\bar{A}} & K_A \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ K_{A\bar{A}} & K_A \end{pmatrix} \end{aligned} \tag{36}$$

Re-organise:

$$
\begin{aligned}
\Pr_L(A \subseteq \mathbf{Y}) &= \det(I_{\bar{A}} + I_A K) \\
&= \det\left(\begin{bmatrix} I_{|\bar{A}|\times|\bar{A}|} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ K_{A\bar{A}} & K_A \end{bmatrix} \right) \\
&= \det\left(\begin{bmatrix} I_{|\bar{A}|\times|\bar{A}|} & \mathbf{0} \\ K_{A\bar{A}} & K_A \end{bmatrix} \right) \\
&= \det(K_A)
\end{aligned}
\tag{37}
$$

therefore, the conversion formula is:

$$
K = L(L+I)^{-1} = I - (L+I)^{-1} \tag{38}
$$

2.4.1 Eigen-value conversion

$$
K = L(L+I)^{-1} = I - (L+I)^{-1} \tag{39}
$$

Properties

$$
\begin{aligned}
\lambda_n \in \operatorname{eig}(A) &\implies \lambda_n + 1 \in \operatorname{eig}(A+I) \\
&\implies (\lambda_n)^{-1} \in \operatorname{eig}(A^{-1})
\end{aligned}
\tag{40}
$$

Apply it to $K = I - (L+I)^{-1}$:

$$
\begin{aligned}
(\lambda_n + 1) \in \operatorname{eig}(L+I) &\implies \frac{1}{\lambda_n + 1} \in \operatorname{eig}\big((L+I)^{-1}\big) \\
&\implies 1 - \frac{1}{\lambda_n + 1} \in \operatorname{eig}\big(I - (L+I)^{-1}\big)
\end{aligned}
\tag{41}
$$

$$
1 - \frac{1}{\lambda_n + 1} = \frac{\lambda_n + 1 - 1}{\lambda_n + 1} = \frac{\lambda_n}{\lambda_n + 1} \tag{42}
$$

Therefore,

$$
L = \sum_{n=1}^{N} \lambda_n v_n v_n^\top \implies K = \sum_{n=1}^{N} \frac{\lambda_n}{\lambda_n + 1} v_n v_n^\top \tag{43}
$$

both L and K share the same eigen vectors

2.4.2 Conversions from K to L

$$
K = L(L+I)^{-1} = I - (L+I)^{-1} \tag{44}
$$

$$
\begin{aligned}
K = I - (L+I)^{-1} &\implies I - K = (L+I)^{-1} \\
&\implies (L+I)(I-K) = I \\
&\implies L + I - LK - K = I \\
&\implies L(I-K) = K \\
&\implies L = K(I-K)^{-1}
\end{aligned}
\tag{45}
$$

3 Complement

If $\mathbf{Y}$ is distributed as a DPP with marginal kernel K, then $\Omega - \mathbf{Y}$ is also distributed as a DPP, with marginal kernel $\bar{K} = I - K$:

$$\begin{aligned} \Pr\big((A \cap \mathbf{Y}) = \emptyset\big) &= \det(\bar{K}_A) \\ &= \det(I_A - K_A) \end{aligned} \tag{46}$$

For example:

$$K = \begin{pmatrix} 0.4 & 0.1 & -0.1 \\ 0.05 & 0.5 & 0.1 \\ -0.01 & 0.1 & 0.3 \end{pmatrix} \qquad \bar{A} = \{3\} \tag{47}$$

and $A = \{1, 2\}$:

$$\begin{aligned} \bar{A} &= \{3\} \\ \bar{K} = I - K &= \begin{pmatrix} 0.6 & -0.1 & 0.1 \\ -0.05 & 0.5 & -0.1 \\ 0.01 & -0.1 & 0.7 \end{pmatrix} \\ \implies \bar{K}_{A=\{1,2\}} &= \begin{pmatrix} 0.6 & -0.1 \\ -0.05 & 0.5 \end{pmatrix} \end{aligned} \tag{48}$$

It's easy to see that $\bar{K}_A = (I_A - K_A)$, basically difference of sub-matrix equal the sub-matrix of the difference.

therefore,

$$\begin{aligned} \Pr(\mathbf{Y} = \emptyset) &\equiv \Pr\big((\Omega \cap \mathbf{Y}) = \emptyset\big) \\ &= \det(\bar{K}_\Omega) \\ &= \det(I_\Omega - K_\Omega) \end{aligned} \tag{49}$$

when we look at Eq.(14), we see given matrix to be $\begin{pmatrix} 0.7 & 0.1 \\ 0.1 & 0.6 \end{pmatrix}$:

$$\begin{aligned} \bar{K}_\Omega = I - \begin{pmatrix} 0.3 & -0.1 \\ -0.1 & 0.4 \end{pmatrix} &= \begin{pmatrix} 0.7 & 0.1 \\ 0.1 & 0.6 \end{pmatrix} \\ \implies \Pr(\mathbf{Y} = \emptyset) = 0.41 \end{aligned} \tag{50}$$

3.0.1 Complement in two point cases

this is just a generalization of Eq.(14):

$$\begin{aligned} \Pr(i, j \notin \mathbf{Y}) &\equiv \Pr(i \notin \mathbf{Y} \cap j \notin \mathbf{Y}) \\ &= 1 - \Pr\big((i \in \mathbf{Y}) \cup (j \in \mathbf{Y})\big) \\ &= 1 - \big(\Pr(i \in \mathbf{Y}) + \Pr(j \in \mathbf{Y}) - \Pr(i, j \in \mathbf{Y})\big) \qquad \text{we removed } \notin \\ &= 1 - \Pr(i \in \mathbf{Y}) - \Pr(j \in \mathbf{Y}) + \Pr(i, j \in \mathbf{Y}) \\ &\leq 1 - \Pr(i \in \mathbf{Y}) - \Pr(j \in \mathbf{Y}) + \Pr(i \in \mathbf{Y}) \Pr(j \in \mathbf{Y}) \qquad \text{from DPP definition: } \Pr(i \in \mathbf{Y}) \Pr(j \in \mathbf{Y}) \geq \Pr(i, j \in \mathbf{Y}) \\ &= 1 - \Pr(i \in \mathbf{Y}) + (1 - \Pr(j \in \mathbf{Y})) - 1 + (1 - \Pr(i \notin \mathbf{Y}))(1 - \Pr(j \notin \mathbf{Y})) \qquad \in \to \notin \\ &= \Pr(i \notin \mathbf{Y}) + \Pr(j \notin \mathbf{Y}) - 1 + \underline{(1 - \Pr(i \notin \mathbf{Y}))(1 - \Pr(j \notin \mathbf{Y}))} \\ &= \Pr(i \notin \mathbf{Y}) + \Pr(j \notin \mathbf{Y}) - 1 + \underline{1 - \Pr(i \notin \mathbf{Y}) - \Pr(j \notin \mathbf{Y}) + \Pr(i \notin \mathbf{Y}) \Pr(j \notin \mathbf{Y})} \qquad \text{expand out} \\ &= \Pr(i \notin \mathbf{Y}) \Pr(j \notin \mathbf{Y}) \end{aligned} \tag{51}$$

Complement of a diversifying process also encourage diversity. (the determinant $\bar{K}_A$ also has the property).

3.0.2 Larger marginal distribution

$$K \preceq K' \implies \det\left(K_A\right) \leq \det\left(K'_A\right) \qquad \forall A \subseteq \Omega \tag{52}$$

DPP defined by K' is "larger" than the one defined by K in the sense that it assigns higher marginal probabilities to every set A.

4 Quality vs Diversity

Let $\mathbf{x}_i$ be each column of data matrix $\mathbf{X}$, and let's normalize:

$$\begin{aligned} q_i &= \|\mathbf{x}_i\|_2 \\ &= \frac{\mathbf{x}_i}{q_i} \implies \|\bar{\mathbf{x}}_i\|_2 = 1 \end{aligned} \tag{53}$$

Let:

$$\begin{aligned} Q &= \begin{bmatrix} q_1 & 0 & \dots & 0 \\ 0 & q_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \dots & \dots & \dots & q_n \end{bmatrix} \\ \implies [q_1\bar{\mathbf{x}}_1 \quad q_2\bar{\mathbf{x}}_2 \quad \dots \quad q_n\bar{\mathbf{x}}_n] &= \mathbf{X} \\ &= \bar{\mathbf{X}}Q \end{aligned} \tag{54}$$

$$\begin{aligned} L(\mathbf{x}_1, \dots, \mathbf{x}_n) &= \mathbf{X}^\top \mathbf{X} \\ &= (\bar{\mathbf{X}}Q)^\top (\bar{\mathbf{X}}Q) \\ &= Q^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}} Q \\ \implies L_{ij} &= q_i \bar{\mathbf{x}}_i^\top \bar{\mathbf{x}}_j q_j \end{aligned} \tag{55}$$

$$\begin{aligned} S_{i,j} &\equiv \bar{\mathbf{x}}_i^\top \bar{\mathbf{x}}_j \in [-1, 1] \\ \implies S_{ij} &= \frac{L_{ij}}{\sqrt{L_{ii}L_{jj}}} \end{aligned} \tag{56}$$

$\Pr_L(\mathbf{Y} = Y)$ can be viewed as the product of four determinants

$$\Pr_L(\mathbf{Y} = Y) \propto \Big(\prod_{i \in Y} q_i^2\Big) \det(S_Y) \tag{57}$$

5 Conditional

5.1 $\Pr_L(\mathbf{Y} = B \mid \mathbf{Y} \cap A = \emptyset)$

given $\mathbf{Y}$ does not contain A, what is its probability it is B?

obviously, we need to assume A and B has no overlaps, i.e., $B \cap A = \emptyset$, and $B \subseteq \Omega$:

$$
\begin{aligned}
\Pr_L(\mathbf{Y} = B \mid \mathbf{Y} \cap A = \emptyset) &= \frac{\Pr_L\big((\mathbf{Y} = B) \cap (A \cap \mathbf{Y} = \emptyset)\big)}{\Pr_L(A \cap \mathbf{Y} = \emptyset)} \\
&= \frac{\Pr_L\big(A \cap \mathbf{Y} = \emptyset | \mathbf{Y} = B\big) \Pr_L\left(\mathbf{Y} = B\right)}{\Pr_L(A \cap \mathbf{Y} = \emptyset)} \\
&= \frac{1 \times \Pr_L\left(\mathbf{Y} = B\right)}{\Pr_L(A \cap \mathbf{Y} = \emptyset)} \qquad \because B \cap A = \emptyset \implies \Pr_L\big(A \cap \mathbf{Y} = \emptyset | \mathbf{Y} = B\big) = 1 \\
&= \frac{\Pr_L\left(\mathbf{Y} = B\right)}{\Pr_L(A \cap \mathbf{Y} = \emptyset)} \\
&= \frac{\frac{\det(L_B)}{\det\left(L_\Omega + I\right)}}{\frac{\sum_{B' : B' \cap A = \emptyset} \det(L_{B'})}{\det\left(L_\Omega + I\right)}} \qquad \text{definition of L-Ensembles} \\
&= \frac{\det(L_B)}{\sum_{B' : B' \cap A = \emptyset} \det(L_{B'})} \\
&= \frac{\det(L_B)}{\sum_{\bar{A}} \det(L_{\bar{A}})} \qquad \{B' : B' \cap A = \emptyset\} = \bar{A} \\
&= \frac{\det(L_B)}{\det(L_{\bar{A}} + I_{|\bar{A}| \times |\bar{A}|})} \qquad \because \Omega \to \bar{A} \quad \text{which also include } \emptyset
\end{aligned}
\tag{58}
$$

where $L_{\bar{A}}$ is L indexed by elements in $\Omega \setminus A$.

note that by definition, $I_{|\bar{A}| \times |\bar{A}|} \neq I_{\bar{A}}$. This is because we are computing **full** set sum $\sum_{\bar{A}} \det(L_{\bar{A}})$, instead of **partial** set sum $\sum_{A \subseteq Y \subseteq \Omega} \det(L_Y) = \det(L + \mathbf{I}_{\bar{A}})$

5.2 $\Pr_L(\mathbf{Y} = A \cup B | A \subseteq \mathbf{Y})$

again, assuming $B \cap A = \emptyset$, and $B \subseteq \Omega$

$$
\begin{aligned}
\Pr_L(\mathbf{Y} = A \cup B | A \subseteq \mathbf{Y}) &= \frac{\Pr_L\left((\mathbf{Y} = A \cup B) \cap (A \subseteq \mathbf{Y})\right)}{\Pr_L(A \subseteq \mathbf{Y})} \\
&= \frac{\underbrace{\Pr_L\big(A \subseteq \mathbf{Y} | \mathbf{Y} = A \cup B\big)}_{\Pr = 1} \Pr_L\left(\mathbf{Y} = A \cup B\right)}{\Pr_L(A \subseteq \mathbf{Y})} \\
&= \frac{\Pr_L\left(\mathbf{Y} = A \cup B\right)}{\Pr_L(A \subseteq \mathbf{Y})} \\
&= \frac{\det(L_{A \cup B})}{\det(L + I_{\bar{A}})}
\end{aligned}
\tag{59}
$$

6 Sampling DPP:

6.1 express in terms of mixture of elementary DPPs

$$\Pr_L = \frac{1}{\det(L+I)} \sum_{J \subseteq \{1,2,\ldots,N\}} \mathcal{P}^{V_J} \prod_{n \in J} \lambda_n \tag{60}$$

where $\mathbf{W}_J \equiv \mathbf{W}_{V_J}$ is the associated (elementary) marginal kernel for $\mathcal{P}^{V_J}$ - we choose to use $\mathbf{W}_J$ instead of K^V, as K is reserved for generic marginal kernel.

We can easily verify that, since all eigen values of $\mathbf{W}_J$ is either zero or one, then:

$$\mathbf{0} \preceq \mathbf{W}_J \preceq \mathbf{I} \tag{61}$$

V_J is a set of **orthonormal** vectors, associated with an elementary DPP with marginal kernel $\mathbf{W}_J = \sum_{\mathbf{v} \in V} \mathbf{v}\mathbf{v}^\top$ where $\mathbf{v}_i \in V$ are eigen-vector of L.

6.1.1 advantage of elementary DPP

the most important factor (during first loop) we decides $|J| = |V|$ from by its mixture weight. Then, if we can prove to sample an elementary DPP with marginal kernel $\mathbf{W}_J$:

$$\Pr_{\mathbf{W}_J}\big(|\mathbf{Y}| = |J|\big) = 1 \tag{62}$$

we only need to sample elements of $\{Y_i\}_{i=1}^{|J|}$.

6.1.2 proof for $\Pr_{\mathbf{W}_J}\big(|\mathbf{Y}| = |J|\big) = 1$

To begin the proof, we simplify the notation by letting:

$$\mathbf{W}_{V_J} \equiv \mathbf{W}_J \tag{63}$$

Firstly, we know that $\Pr_{\mathbf{W}_J}[|\mathbf{Y}|] = 0 \quad \forall |J| < |\mathbf{Y}|$. Since matrix indexed by $\mathbf{Y}$ will have determinant being zero. However, after we prove that $\mathbb{E}_{\mathbf{W}_J}[|\mathbf{Y}|] = |J|$, so the only way for both to be true is that $|\mathbf{Y}| = |J|$ almost surely:

$$\begin{aligned}
\mathbb{E}_{\mathbf{W}_J}[|\mathbf{Y}|] &= \sum_{i=1}^{N} \mathbb{E}_{\mathbf{W}_J}[\mathbb{1}_{y_i \in \mathbf{Y}}] \quad \because \mathbb{E}[\text{sum of Bernoulli}] = \text{sum of } \mathbb{E}[\text{Bernoulli}] \\
&= \sum_{i=1}^{N} \Pr_{\mathbf{W}_J}(y_i \in \mathbf{Y}) \\
&= \sum_{i=1}^{N} \mathbf{W}_{J\,i,i} \quad \text{definition of DPP} \\
&= \mathrm{Tr}(\mathbf{W}_J) \\
&= |J| \quad \because J \text{ is sum of } |J| \text{ rank one matrix } V_i V_i^\top \text{ each with eigenvalue 1}
\end{aligned} \tag{64}$$

Of course, we also need sampling an elementary DPP with $\det(\mathbf{W}_J)$ kernel has a lot faster computation.

6.2 mixture weight $\frac{\prod_{n\in J}\lambda_n}{\det(L+I)}$

When mixture weights expressed as $\frac{\prod_{n\in J}\lambda_n}{\det(L+I)}$, for example when $J = \{1, 3, 5\}$, its corresponding mixture weights is:

$$\frac{\lambda_1\lambda_3\lambda_5}{\prod_{n=1}^{N}(\lambda_n + 1)} \tag{65}$$

note that denominator is the product of **all** eigen values. But the numerator is the product of the **selected** ones.

If we let selecting $\mathbf{v}_i$ to be $\frac{\lambda_i}{\lambda_i+1}$, and therefore, not selecting it to be $\frac{1}{\lambda_i+1}$.

Then, the probability of **only** selecting J set is:

$$\begin{aligned}&\frac{\lambda_1}{\lambda_1+1}\,\frac{1}{\lambda_2+1}\,\frac{\lambda_3}{\lambda_3+1}\,\frac{1}{\lambda_4+1}\,\frac{\lambda_5}{\lambda_5+1}\,\frac{1}{\lambda_6+1}\;\times\ldots\\ =&\frac{\lambda_1\lambda_3\lambda_5}{\prod_{n=1}^{N}(\lambda_n+1)}\end{aligned} \tag{66}$$

6.3 sampling $\mathcal{P}^V$

6.3.1 Elementary DPP:

A DPP is called **elementary** if every eigenvalue of its marginal kernel is $\in \{0, 1\}$

1. **example 1**: $V \equiv \left\{ \begin{bmatrix}0\\1\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix} \right\}$

$$\begin{aligned}
\mathbf{W}_J &= \begin{bmatrix}1 & 0 & 0\\0 & 0 & 0\\0 & 0 & 1\end{bmatrix} \\
&= 0 \begin{bmatrix}0\\1\\0\end{bmatrix} \begin{bmatrix}0 & 1 & 0\end{bmatrix} + 1 \begin{bmatrix}1\\0\\0\end{bmatrix} \begin{bmatrix}1 & 0 & 0\end{bmatrix} + 1 \begin{bmatrix}0\\0\\1\end{bmatrix} \begin{bmatrix}0 & 0 & 1\end{bmatrix}
\end{aligned} \tag{70}$$

2. **example 2**: $V \in \left\{ \begin{bmatrix}-0.5735\\0.7781\\-0.2562\end{bmatrix}, \begin{bmatrix}-0.3243\\0.0716\\0.9432\end{bmatrix}, \begin{bmatrix}0.7523\\0.6240\\0.2113\end{bmatrix} \right\}$

$$\begin{aligned}
\mathbf{W}_J = \begin{bmatrix}0.3945 & -0.0557 & -0.4856\\-0.0557 & 0.9949 & -0.0447\\-0.4856 & -0.0447 & 0.6106\end{bmatrix} &= 1 \times \begin{bmatrix}-0.5735\\0.7781\\-0.2562\end{bmatrix} \begin{bmatrix}-0.5735 & 0.7781 & -0.2562\end{bmatrix} \\
&+ 0 \times \begin{bmatrix}-0.3243\\0.0716\\0.9432\end{bmatrix} \begin{bmatrix}-0.3243 & 0.0716 & 0.9432\end{bmatrix} \\
&+ 1 \times \begin{bmatrix}0.7523\\0.6240\\0.2113\end{bmatrix} \begin{bmatrix}0.7523 & 0.6240 & 0.2113\end{bmatrix}
\end{aligned} \tag{71}$$

$\mathbf{W}_J$ is a sum of a set of rank one matrix, each constructed from an ortho-normal set. $\mathbf{W}_J$ is still a valid DPP marginal kernel, although a lot of larger sets will have zero probability.

6.3.2 Multi-Linearity

Lemma 2 *Let each $\mathbf{W}_n$ to be rank-one matrix, and sum of $\mathbf{W}_J = \sum_{n \in J} \mathbf{W}_n$: then we have:*

$$\det(\mathbf{W}_J) = \underbrace{\sum_{n_1, n_2, \ldots, n_k \in J}}_{\textit{are distinct}} \det\left([(\mathbf{W}_{n_1})_1, (\mathbf{W}_{n_2})_2, \ldots, (\mathbf{W}_{n_k})_k]\right) \tag{72}$$

RHS can be visualized as when we have a set of $|J|$ matrices $\{\mathbf{W}_n\}_{n=1}^{|J|}$, if we take a column from each of the matrices to form a new matrix $\mathbf{W}$ and to compute its determinant, and then, sum over these determinant of all combinations. Then we get the determinant of the sum of $\{\mathbf{W}_n\}_{n=1}^{|J|}$!

note also that $|J| \geq k$

6.3.3 proof of lemma

write out each column explicitly:

$$\begin{aligned}
\det(\mathbf{W}_J) &= \det\left(\Big[(\mathbf{W}_J)_1, (\mathbf{W}_J)_2, \ldots, (\mathbf{W}_J)_k\Big]\right) \\
&= \det\left(\Big[\Big(\sum_{n \in J} \mathbf{W}_n\Big)_1, (\mathbf{W}_J)_2, \ldots, (\mathbf{W}_J)_k\Big]\right) \qquad \text{expand first term}
\end{aligned} \tag{73}$$

for example:

$$\mathbf{W}_1 = \begin{bmatrix}3\\2\end{bmatrix}\begin{bmatrix}3 & 2\end{bmatrix} = \begin{bmatrix}9 & 6\\6 & 4\end{bmatrix}$$
$$\mathbf{W}_2 = \begin{bmatrix}1\\2\end{bmatrix}\begin{bmatrix}1 & 2\end{bmatrix} = \begin{bmatrix}1 & 2\\2 & 4\end{bmatrix}$$
$$\mathbf{W}_J = \mathbf{W}_1 + \mathbf{W}_2 = \begin{bmatrix}10 & 8\\8 & 8\end{bmatrix}$$
$$\left(\sum_{n\in J}\mathbf{W}_n\right)_1 = \begin{bmatrix}10\\8\end{bmatrix}$$

because **Multi-linearity** states:

$$\det\big([\mathbf{a}_1+\mathbf{b}_1,\mathbf{a}_2,\dots,\mathbf{a}_k]\big) = \det\big([\mathbf{a}_1,\mathbf{a}_2,\dots,\mathbf{a}_k]\big) + \det\big([\mathbf{b}_1,\mathbf{a}_2,\dots,\mathbf{a}_k]\big) \tag{74}$$

Therefore,

$$\begin{aligned}\det(\mathbf{W}_J) &= \det\left(\left[\left(\sum_{n\in J}\mathbf{W}_n\right)_1,(\mathbf{W}_J)_2,\dots,(\mathbf{W}_J)_k\right]\right)\\ &= \sum_{n\in J}\det\big([(\mathbf{W}_n)_1,(\mathbf{W}_J)_2,\dots,(\mathbf{W}_J)_k]\big)\end{aligned} \tag{75}$$

Now, we repeat the same thing for the second term and all subsequent terms, But we can't use the same index n for different columns. Therefore, we give a different index $n_i \in J \quad \forall i$:

$$\det(\mathbf{W}_J) = \sum_{n_1\in J}\sum_{n_2\in J}\cdots\sum_{n_k\in J}\det\big(\underbrace{[(\mathbf{W}_{n_1})_1,(\mathbf{W}_{n_2})_2,\dots,(\mathbf{W}_{n_k})_k]}_{\mathbf{W}}\big) \tag{76}$$

6.3.4 loop index $n_1,\dots n_k$ need to be distinct

when we look at:

$$\det(\mathbf{W}_J) = \sum_{n_1\in J}\sum_{n_2\in J}\cdots\sum_{n_k\in J}\det\big([(\mathbf{W}_{n_1})_1,(\mathbf{W}_{n_2})_2,\dots,(\mathbf{W}_{n_k})_k]\big) \tag{77}$$

not every term is non-zero.

Since $\mathbf{W}_n$ is rank one matrix, $(\mathbf{W}_n)_i$ and $(\mathbf{W}_n)_j$ are linearly dependant. Therefore, the determinant of any matrix containing two or more columns of the **same** $\mathbf{W}_n$ is zero, for example:

$$\det(\mathbf{W}_J) = \det\big([(\mathbf{W}_{n_1})_1,(\mathbf{W}_{n_1})_2,\dots,(\mathbf{W}_{n_k})_k]\big) = 0 \tag{78}$$

Thus the terms in the sum vanish unless $n_1,n_2,\dots n_k$ are distinct.

$$\begin{aligned}\det(\mathbf{W}_J) &= \sum_{n_1\in J}\sum_{n_2\in J}\cdots\sum_{n_k\in J}\det\big([(\mathbf{W}_{n_1})_1,(\mathbf{W}_{n_2})_2,\dots,(\mathbf{W}_{n_k})_k]\big)\\ &= \underbrace{\sum_{n_1\in J}\sum_{n_2\in J}\cdots\sum_{n_k\in J}}_{n_1,n_2,\dots,n_k \text{ are distinct}}\det\big([(\mathbf{W}_{n_1})_1,(\mathbf{W}_{n_2})_2,\dots,(\mathbf{W}_{n_k})_k]\big)\\ &= \underbrace{\sum_{n_1,n_2,\dots,n_k\in J}}_{\text{distinct}}\det\big([(\mathbf{W}_{n_1})_1,(\mathbf{W}_{n_2})_2,\dots,(\mathbf{W}_{n_k})_k]\big)\end{aligned} \tag{79}$$

6.4 Why mixture of elementary DPPs works

Most importantly, we need to show a DPP with L-ensemble kernel $L = \sum_{n=1}^{N} \lambda_n v_n v_n^\top$ is a mixture of elementary DPPs:

$$\frac{1}{\det(L+I)} \sum_{J \subseteq 1,2,\ldots,N} \mathcal{P}^{V_J} \prod_{n \in J} \lambda_n \tag{80}$$

where each $\mathcal{P}^{V_J}$ associate with its own kernel $\mathbf{W}_J$.

6.4.1 show $\Pr(A \in \mathbf{Y})$ from mixture model also equal $\det(K_A)$

for a particular index set A, we have $k = |A|$ and the associated $\mathbf{W}_n^A = [\mathbf{v}_n \mathbf{v}_n^\top]_A$. This means each of the rank-one matrix of $\mathbf{v}_n \mathbf{v}_n^\top$ gets "chop-off" by the index set A to become $\mathbf{W}_n^A$. Therefore, we need to show that summation of J (from all the mixture weights) of $\det\left(\mathbf{W}_J^A\right)$ gives the right marginal probability $\Pr(A \in \mathbf{Y}) = \det(K_A)$

Start from from mixture of elementary DPPs definition:

$$\begin{aligned}
\Pr(A \in \mathbf{Y}) &= \frac{1}{\det(L+I)} \sum_{J \subseteq \{1,2,\ldots,N\}} \det\left(W_J^A\right) \prod_{n \in J} \lambda_n \\
&= \frac{1}{\det(L+I)} \sum_{J \subseteq \{1,2,\ldots,N\}} \det\left(\sum_{n \in J} W_n^A\right) \prod_{n \in J} \lambda_n \qquad \text{let } W_J^A \equiv \mathbf{W}^J \\
&= \frac{1}{\det(L+I)} \sum_{J \subseteq \{1,2,\ldots,N\}} \underbrace{\sum_{n_1,n_2,\ldots,n_k \in J}}_{\text{distinct}} \det\left(\left[(W_{n_1}^A)_1, (W_{n_2}^A)_2, \ldots, (W_{n_k}^A)_k\right]\right) \prod_{n \in J} \lambda_n \qquad \text{from lemma (2)}
\end{aligned} \tag{81}$$

For the outer loop, $\sum_{J \subseteq \{1,2,\ldots,N\}}$ when $|J| < k$, then, the inner loop becomes zero. Since it's impossible for $|J| < k$ points to be distinct. Therefore, we need only a subset of $\{1, \ldots N\}$:

$$J \supseteq \{n_1, n_2, \ldots n_k\} \tag{82}$$

Remove the combinations of sums resulting zero determinant and then swapping the inner and outer loops, we have:

$$\begin{aligned}
&= \frac{1}{\det(L+I)} \sum_{J \supseteq \{n_1,n_2,\ldots,n_k\}} \underbrace{\sum_{n_1,n_2,\ldots,n_k \in J}}_{\text{distinct}} \det\left(\left[(W_{n_1}^A)_1, (W_{n_2}^A)_2, \ldots, (W_{n_k}^A)_k\right]\right) \prod_{n \in J} \lambda_n \\
&= \frac{1}{\det(L+I)} \underbrace{\sum_{n_1,n_2,\ldots,n_k}}_{\text{distinct}} \det\left(\left[(W_{n_1}^A)_1, (W_{n_2}^A)_2, \ldots, (W_{n_k}^A)_k\right]\right) \sum_{J \supseteq \{n_1,n_2,\ldots,n_k\}} \prod_{n \in J} \lambda_n
\end{aligned} \tag{83}$$

For example, let $J \subseteq \{1, 2, 3, 4, 5\}$, and let $\{n_1, n_2, \ldots n_k\} = \{1, 2, 3\}$. Then, $J \supseteq \{n_1, n_2, \ldots, n_k\} = \{\{1, 2, 3\}, \{1, 2, 3, 4\}, \{1, 2, 3, 5\}, \{1, 2, 3, 4, 5\}\}$:

$$
\begin{aligned}
\sum_{J \supseteq \{n_1, n_2, \ldots, n_k\}} \prod_{n \in J} \lambda_n &= \lambda_1\lambda_2\lambda_3 + \lambda_1\lambda_2\lambda_3\lambda_4 + \lambda_1\lambda_2\lambda_3\lambda_5 + \lambda_1\lambda_2\lambda_3\lambda_4\lambda_5 \qquad \text{using the example} \\
&= \lambda_1\lambda_2\lambda_3(1 + \lambda_4 + \lambda_5 + \lambda_4\lambda_5) \\
&= \lambda_1\lambda_2\lambda_3(1 + \lambda_4)(1 + \lambda_5) \qquad \text{this step is the key} \\
&= \lambda_1\lambda_2\lambda_3(1 + \lambda_4)(1 + \lambda_5)\frac{(\lambda_1 + 1)(\lambda_2 + 1)(\lambda_3 + 1)(\lambda_4 + 1)(\lambda_5 + 1)}{(\lambda_1 + 1)(\lambda_2 + 1)(\lambda_3 + 1)(\lambda_4 + 1)(\lambda_5 + 1)} \\
&= \frac{\lambda_1}{\lambda_1 + 1}\frac{\lambda_2}{\lambda_2 + 1}\frac{\lambda_3}{\lambda_3 + 1}(\lambda_1 + 1)(\lambda_2 + 1)(\lambda_3 + 1)(\lambda_4 + 1)(\lambda_5 + 1) \\
&= \frac{\lambda_{n_1}}{\lambda_{n_1} + 1} \cdots \frac{\lambda_{n_k}}{\lambda_{n_k} + 1} \prod_{n=1}^{N}(\lambda_n + 1) \qquad \text{we generalise it to } N \text{ terms}
\end{aligned} \tag{84}
$$

substituting the expression for $\sum_{J \supseteq \{n_1, n_2, \ldots, n_k\}} \prod_{n \in J} \lambda_n$:

$$
\begin{aligned}
\mathrm{Pr}_L &= \frac{1}{\det(L + I)} \underbrace{\sum_{n_1, n_2, \ldots, n_k}^{N}}_{\text{distinct}} \det\left(\left[(\mathbf{W}_{n_1}^A)_1, (\mathbf{W}_{n_2}^A)_2, \ldots, (\mathbf{W}_{n_k}^A)_k\right]\right) \sum_{J \supseteq \{n_1, n_2, \ldots, n_k\}} \prod_{n \in J} \lambda_n \\
&= \frac{1}{\prod_{n=1}^{N}(\lambda_n + 1)} \underbrace{\sum_{n_1, n_2, \ldots, n_k}^{N}}_{\text{distinct}} \det\left(\left[(\mathbf{W}_{n_1}^A)_1, (\mathbf{W}_{n_2}^A)_2, \ldots, (\mathbf{W}_{n_k}^A)_k\right]\right) \frac{\lambda_{n_1}}{\lambda_{n_1} + 1} \cdots \frac{\lambda_{n_k}}{\lambda_{n_k} + 1} \prod_{n=1}^{N}(\lambda_n + 1) \\
&= \underbrace{\sum_{n_1, n_2, \ldots, n_k}^{N}}_{\text{distinct}} \det\left(\left[(\mathbf{W}_{n_1}^A)_1, (\mathbf{W}_{n_2}^A)_2, \ldots, (\mathbf{W}_{n_k}^A)_k\right]\right) \frac{\lambda_{n_1}}{\lambda_{n_1} + 1} \cdots \frac{\lambda_{n_k}}{\lambda_{n_k} + 1} \\
&= \underbrace{\sum_{n_1, n_2, \ldots, n_k}^{N}}_{\text{distinct}} \det\left(\left[(\mathbf{W}_{n_1}^A)_1 \frac{\lambda_{n_1}}{\lambda_{n_1} + 1}, (\mathbf{W}_{n_2}^A)_2 \frac{\lambda_{n_2}}{\lambda_{n_2} + 1}, \ldots, (\mathbf{W}_{n_k}^A)_k \frac{\lambda_{n_k}}{\lambda_{n_k} + 1}\right]\right) \\
&\qquad \because \alpha\beta \det([\mathbf{a}_1 \quad \mathbf{a}_2]) = \det([\alpha\mathbf{a}_1 \quad \beta\mathbf{a}_2]) \\
&= \det\left(\sum_{n=1}^{N} \frac{\lambda_n}{\lambda_n + 1} \mathbf{W}_n^A\right) \quad \text{apply lemma (2) again, with } J \equiv \{1, \ldots, N\} \\
&= \det(K_A) \qquad \text{using Eq.(43) by noting } \sum_{n=1}^{N} \frac{\lambda_n}{\lambda_n + 1} \mathbf{W}_n = K
\end{aligned} \tag{85}
$$

6.5 Sampling algorithm

6.5.1 first step: determine the $\{\mathbf{v}_n\}_{n \in J}$, where $\mathbf{v}_n \in \mathbb{R}^N$

this is described in Eq.(66).

6.5.2 second step: sample a single DPP $\mathcal{P}^{V_J}$

According to Eq.(62), which states that $\mathrm{Pr}_{\mathbf{W}_J}(|\mathbf{Y}| = |J|) = 1$, therefore, we just need to sample $i \in \{1, \ldots, N\}$ $|J|$ times.

Although $\mathcal{P}^{V_J}$ (parametrized by $\mathbf{W}_J$) is an (elementary) **marginal DPP**, but its parameter $\mathbf{W}_J = \sum_{\mathbf{v}_n \in J} \mathbf{v}_n \mathbf{v}_n^\top$ itself in fact is a form of Gram-matrix!

Although it may look bizarre at first, please note that $\mathbf{W}_J$ is in fact a Gram matrix, where the i^{th} "data" of this gram matrix is formed by taking the i^{th} dimension of each of $\mathbf{v} \in J$.

For example: $J = \{\mathbf{v}_1, \mathbf{v}_2\}$ where:

$$\mathbf{v}_1 = [3 \quad 6 \quad 6 \quad 7] \qquad \mathbf{v}_2 = [3 \quad 5 \quad 1 \quad 2] \tag{86}$$

be the un-normalized vectors (generalized version of orthornormal sets). Then $\mathbf{W}_J = \mathbf{v}_1\mathbf{v}_1^\top + \mathbf{v}_2\mathbf{v}_2^\top$ can be equivalently viewed as as a Gram-matrix formed by "data":

$$\left\{ [3 \quad 3] \,, [6 \quad 5] \,, [6 \quad 1] \,, [7 \quad 2] \right\} \tag{87}$$

Both views will result to the same $\mathbf{W}_J$!

Then sampling can just follow the geometric property of Gram matrix, i.e., in section 2.2. Except this time we do know $|\mathbf{Y}| = |J|$.

This can be done by repetitively:

1. choosing the **base** of the remaining parallel-pip:

$$\begin{aligned} \Pr(i) &\propto \text{square of the volume align with } \mathbf{e}_i \\ &\propto \frac{1}{|V|} \sum_{\mathbf{v}_n \in V} \mathbf{v}_n^\top \mathbf{e}_i \end{aligned} \tag{88}$$

2. and then "tilt" the parallel-pip such that the remaining dimension is orthogonal to the base just removed:

$$\begin{aligned} &Y \leftarrow Y \cup i \\ &V \leftarrow V_\perp \quad \text{an orthonormal basis for the subspace of } V \text{ orthogonal to } \mathbf{e}_i \end{aligned} \tag{89}$$

References

[1] Alex Kulesza, Ben Taskar, et al., "Determinantal point processes for machine learning," *Foundations and Trends® in Machine Learning*, vol. 5, no. 2–3, pp. 123–286, 2012.