Determinantal Point Process and its Time-varying model

A/Prof Richard Yi Da Xu
Yida.Xu@uts.edu.au
Wechat: aubedata
https://github.com/roboticcam/machine-learning-notes

University of Technology Sydney (UTS)

February 18, 2018

What is DPP?

Start with a **marginal** distribution:

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

An example: given $\mathcal{Y} = \{1, 2, 3, 4, 5\}$, $A = \{1, 2, 3\}$

$$\begin{aligned}\Pr(A \subseteq \mathbf{Y}) &\equiv \Pr(A \subseteq \mathbf{Y} \subseteq \mathcal{Y}) \equiv \Pr\left(\mathbf{Y} \in \{\{1,2,3\},\{1,2,3,4\},\{1,2,3,5\},\{1,2,3,4,5\}\}\right) \\ &= \det(K_A)\end{aligned}$$

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

Something about marginal distribution

- $\Pr(A \subseteq \mathbf{Y})$ is marginal, they don't need to add to 1
- it may be possible that, $\Pr(A_1 \subseteq \mathbf{Y}) + \Pr(A_2 \subseteq \mathbf{Y}) > 1$
- $\Pr(\emptyset \subseteq \mathbf{Y}) = \det(K_\emptyset) = 1$ This is obvious, as any $\mathbf{Y}$ is a superset of $\emptyset$.
- $\Pr(i \subseteq \mathbf{Y}) = \det(K_{ii}) = K_{ii}$
- Look at the two element case:

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

- Off-diagonal elements determine negative correlations between pairs.
- Large values of K_{ij} imply i and j tend **not** co-occur

Example of K does NOT define DPP

- Any $K, 0 \preceq K \preceq I$ defines a DPP.
- If $K \preceq K'$, that is, $K' - K$ is positive semidefinite.
  Therefore, $\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$ can NOT define DPP, as

$$\left| I - \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix} \right| = \begin{pmatrix} 0 & 0.5 \\ 0.5 & 0 \end{pmatrix} \implies \bar{\lambda} = [-0.5, 0.5]^\top$$

- Another way to see the above is incorrect:
  $\mathcal{Y} = \{1, 2\}$

$$\begin{aligned} \Pr(A = \{1\} \subseteq \mathbf{Y}) &\equiv \Pr(\mathbf{Y} \in \{\{1\}, \{1, 2\}\}) \\ &= \det(K_1) = 1 \\ \Pr(A = \{2\} \subseteq \mathbf{Y}) &\equiv \Pr(\mathbf{Y} \in \{\{2\}, \{1, 2\}\}) \\ &= \det(K_2) = 1 \end{aligned}$$

  However,

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

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

Example of K define DPP

- Any $K, 0 \preceq K \preceq I$ defines a DPP.
- If $K \preceq K'$, that is, $K' - K$ is positive semidefinite.
  $\begin{bmatrix} 0.3 & -0.1 \\ -0.1 & 0.4 \end{bmatrix}$ **can** define DPP:

$$\left| I - \begin{pmatrix} 0.3 & -0.1 \\ -0.1 & 0.4 \end{pmatrix} \right| = \left| \begin{pmatrix} 0.7 & 0.1 \\ 0.1 & 0.6 \end{pmatrix} \right| \implies \bar{\lambda} = [0.5382, 0.7618]^\top$$

- $\mathcal{Y} = \{1, 2\}$

$$\begin{aligned} \Pr(A = \{1\} \subseteq \mathbf{Y}) &\equiv \Pr\left(Y \in \{\{1\}, \{1,2\}\}\right) \\ &= \det(K_1) = 0.3 \\ \Pr(A = \{2\} \subseteq \mathbf{Y}) &\equiv \Pr\left(Y \in \{\{2\}, \{1,2\}\}\right) \\ &= \det(K_2) = 0.4 \\ \Pr(A = \{1,2\} \subseteq \mathbf{Y}) &\equiv \Pr\left(Y \in \{\{1,2\}\}\right) \\ &= \det(K_{\{1,2\}}) = 0.11 \end{aligned}$$

- So where do rest of probabilities go?

$$\begin{aligned} \Pr(A = \emptyset \subseteq \mathbf{Y}) &\equiv \Pr\left(Y \in \{\emptyset, \{1\}, \{2\}, \{1,2\}\}\right) \\ &= \det(K_\emptyset) = 1 \end{aligned}$$

- Some probabilities mass is assigned to $\emptyset$.

L-Ensembles

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

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

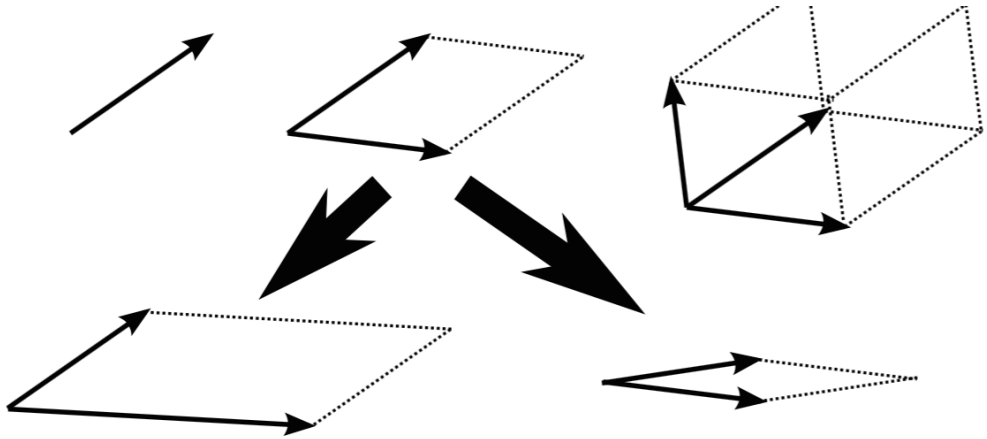
- L must be positive semidefinite.
- Only a statement of proportionality, eigenvalues of L need **not** < 1

Geometry interpretation

$$X = [x_1 \quad x_2 \quad \dots \quad x_n] \implies$$

$$L(x_1, \dots, x_n) = X^\top X = \begin{pmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle & \dots & \langle x_1, x_n \rangle \\ \langle x_2, x_1 \rangle & \langle x_2, x_2 \rangle & \dots & \langle x_2, x_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x_n, x_1 \rangle & \langle x_n, x_2 \rangle & \dots & \langle x_n, x_n \rangle \end{pmatrix}$$

- 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 nonzero
- $\Pr_L(Y) \propto \det(L_Y) = \mathrm{Vol}^2\left(\{x_i\}_{i \in Y}\right)$



Proof for the Geometry interpretation (1)

- in **1-element** case: $\text{Vol}(u) = \sqrt{u^\top u}$, i.e., length of a line
- in **k-element** case: $\text{Vol}(u_1 \dots u_k, u_{k+1}) = \text{Vol}(u_1, \dots, u_k)\|\tilde{u}_{k+1}\|$
- $\tilde{u}_{k+1}$ is the orthogonal projection of u_{k+1} onto span $(u_1, \dots, u_k)$: imagine in in the **2-element** or **3-element** case.
- Let $(u_1, \dots, u_k)$ is an $n \times k$ matrix Y:

$$Y = (u_1 \quad u_2 \quad \dots u_k)$$

- Then there exists a vector $c \in R_k$ such that $u_{k+1} = Yc + \tilde{u}_{k+1}$:

$$u_{k+1} = Yc + \tilde{u}_{k+1} = c_1 u_1 + c_2 u_2 \dots c_k u_k + \tilde{u}_{k+1}$$

- extend Y to X:

$$\begin{aligned} X &= [u_1 \quad u_2 \quad \dots \quad u_k \quad u_{k+1}] \\ &= [Y \quad Yc + \tilde{u}_{k+1}] \\ \implies X^\top X &= \begin{bmatrix} Y^\top Y & Y^\top (Yc + \tilde{u}_{k+1}) \\ (Yc + \tilde{u}_{k+1})^\top Y & (Yc + \tilde{u}_{k+1})^\top (Yc + \tilde{u}_{k+1}) \end{bmatrix} \\ &= \begin{bmatrix} Y^\top Y & Y^\top Yc \\ c^\top Y^\top Y & c^\top Y^\top Yc + \tilde{u}_{k+1}^\top \tilde{u}_{k+1} \end{bmatrix} \\ &= \begin{bmatrix} Y^\top Y & Y^\top Yc \\ c^\top Y^\top Y & c^\top Y^\top Yc + \|\tilde{u}_{k+1}\| \end{bmatrix} \end{aligned}$$

Proof for the Geometry interpretation (2)

$$
\begin{aligned}
X^\top X &= \begin{bmatrix} Y^\top Y & Y^\top Yc \\ c^\top Y^\top Y & c^\top Y^\top Yc + \|\tilde{u}_{k+1}\| \end{bmatrix} \\
&= \left[\begin{bmatrix} Y^\top Y \\ c^\top Y^\top Y \end{bmatrix} \quad \begin{bmatrix} Y^\top Yc \\ c^\top Y^\top Yc \end{bmatrix} + \begin{bmatrix} 0 \\ \|\tilde{u}_{k+1}\| \end{bmatrix} \right] \\
\implies |X^\top X| &= \begin{vmatrix} Y^\top Y & Y^\top Yc \\ c^\top Y^\top Y & c^\top Y^\top Yc \end{vmatrix} + \begin{vmatrix} Y^\top Y & 0 \\ c^\top Y^\top Y & \|\tilde{u}_{k+1}\| \end{vmatrix} \\
&= 0 + \begin{vmatrix} Y^\top Y & 0 \\ c^\top Y^\top Y & \|\tilde{u}_{k+1}\| \end{vmatrix} \\
&= |Y^\top Y| \underbrace{\|\tilde{u}_{k+1}\|}_{\text{Vol}^2(\tilde{u}_{k+1})}
\end{aligned}
$$

Proof for the Geometry interpretation (3)

$$B = \begin{bmatrix} V_{1,1} & V_{1,2} & \dots & V_{1,n} \\ V_{2,1} & V_{2,2} & \dots & V_{2,n} \\ \dots & \dots & \dots & \dots \\ V_{k,1} & V_{k,2} & \dots & V_{k,n} \end{bmatrix} = [B_1 \quad B_2 \quad \dots \quad B_n]$$

$$K^V = B^\top B = \begin{bmatrix} \langle V_1, V_1 \rangle & \langle V_1, V_2 \rangle & \dots & \langle V_1, V_n \rangle \\ \langle V_2, V_1 \rangle & \langle V_2, V_2 \rangle & \dots & \langle V_2, V_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle V_n, V_1 \rangle & \langle V_n, V_2 \rangle & \dots & \langle V_n, V_n \rangle \end{bmatrix}$$

Normalization constant

Theorem says,

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

For example,

$$L = \begin{pmatrix} 2.8599 & -0.4936 & -1.8458 \\ -0.4936 & 2.6264 & -1.1437 \\ -1.8458 & -1.1437 & 2.0522 \end{pmatrix}$$

$$A = \{1, 2\} \implies \bar{A} = \{3\} \implies I_{\bar{A}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Therefore, normalisation constant (or partition function) is:

$$\sum_{\emptyset \subseteq Y \subseteq \mathcal{Y}} \det(L_Y) = \sum_{Y \subseteq \mathcal{Y}} \det(L_Y) = \det(L + I_{\bar{\emptyset}}) = \det(L + I)$$

Conversion to Marginal distribution (1)

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

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

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

$L(L + I)^{-1} = I - (L + I)^{-1}$ is **true** for any L where $(L + I)^{-1}$ exist, think scalar case:

$$1 - \frac{1}{x + 1} = \frac{x + 1 - 1}{x + 1} = \frac{x}{x + 1}$$

Then,

$$\begin{aligned}
\Pr_L(A \subseteq \mathbf{Y}) &= \frac{\sum_{A \subseteq Y \subseteq \mathcal{Y}} \det(L_Y)}{\sum_{Y \subseteq \mathcal{Y}} \det(L_Y)} \\
&= \frac{\det(L + I_{\bar{A}})}{\det(L + I)} \\
&= \det\left((L + I_{\bar{A}})(L + I)^{-1}\right)
\end{aligned}$$

Since, $\det(A^{-1}) = \frac{1}{\det(A)}$ $\quad$ $\det(AB) = \det(A)\det(B)$

Conversion to Marginal distribution (2)

$$
\begin{aligned}
\Pr_L(A \subseteq \mathbf{Y}) &= \det\left((L + I_{\bar{A}})(L + I)^{-1}\right) \\
&= \det\left(\underbrace{L(L + I)^{-1}}_{I-(L+I)^{-1}} + I_{\bar{A}}(L + I)^{-1}\right) \\
&= \det\left(I - (L + I)^{-1} + I_{\bar{A}}(L + I)^{-1}\right) \\
&= \det\left(I - (I - I_{\bar{A}})(L + I)^{-1}\right) \\
&= \det\left(I - I_A(L + I)^{-1}\right) \\
&= \det\left(\underbrace{I_A + I_{\bar{A}}}_{I} - I_A(L + I)^{-1}\right) \\
&= \det\left(I_{\bar{A}} + \underline{I_A - I_A(L + I)^{-1}}\right) \\
&= \det\left(I_{\bar{A}} + I_A\left(\underbrace{I - (L + I)^{-1}}_{K}\right)\right)
\end{aligned}
$$

Conversion to Marginal distribution (3)

$$\Pr_L(A \subseteq \mathbf{Y}) = \det\left(I_{\bar{A}} + I_A\left(\underbrace{I - (L+I)^{-1}}_{K}\right)\right)$$

- left multiplication by I_A **zeros out rows** of a matrix except those corresponding to A, $\Longrightarrow$

$$K = \begin{pmatrix} K_{\bar{A}} & K_{\bar{A}A} \\ K_{A\bar{A}} & K_A \end{pmatrix} \Longrightarrow I_A(K) = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{|A|\times|A|} \end{pmatrix} \begin{pmatrix} K_{\bar{A}} & K_{\bar{A}A} \\ K_{A\bar{A}} & K_A \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ K_{A\bar{A}} & K_A \end{pmatrix}$$

- Re-organise:

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

$K = L(L+I)^{-1} = I - (L+I)^{-1}$ is the conversion formula!

Eigen-value conversion

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

- **Properties**

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

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

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

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

- Therefore,

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

Conversions from K to L

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

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

Complement

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

$$\Pr(A \cap \mathbf{Y} = \emptyset) = \det(\bar{K}_A) = \det(I - K_A)$$

For example:

$$K = \begin{pmatrix} 0.4 & 0.1 & -0.1 \\ 0.05 & 0.5 & 0.1 \\ -0.01 & 0.1 & 0.3 \end{pmatrix}, A = \{1, 2\}, \bar{A} = \{3\}$$

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

It's easy to see that $\bar{K}_A = (I - K_A)$

Complement

$$
\begin{aligned}
\Pr(i, j \notin \mathbf{Y}) &= 1 - \Pr(i, j \in \mathbf{Y}) \\
&= 1 - \big(\Pr(i \in \mathbf{Y}) + \Pr(j \in \mathbf{Y}) - \Pr(i, j \in \mathbf{Y}) \big) \\
&= 1 - \Pr(i \in \mathbf{Y}) - \Pr(j \in \mathbf{Y}) + \Pr(i, j \in \mathbf{Y}) \\
&\leq 1 - \Pr(i \in \mathbf{Y}) - \Pr(j \in \mathbf{Y}) + \Pr(i \in \mathbf{Y}) \Pr(j \in \mathbf{Y}) \\
&= 1 - \Pr(i \in \mathbf{Y}) + (1 - \Pr(j \in \mathbf{Y})) - 1 + (1 - \Pr(i \notin \mathbf{Y}))(1 - \Pr(j \notin \mathbf{Y})) \\
&= \Pr(i \notin \mathbf{Y}) + \Pr(j \notin \mathbf{Y}) - 1 + \underline{(1 - \Pr(i \notin \mathbf{Y}))(1 - \Pr(j \notin \mathbf{Y}))} \\
&= \Pr(i \notin \mathbf{Y}) + \Pr(j \notin \mathbf{Y}) - 1 + \underline{1 - \Pr(i \notin \mathbf{Y}) - \Pr(j \notin \mathbf{Y}) + \Pr(i \notin \mathbf{Y}) \Pr(j \notin \mathbf{Y})} \\
&= \Pr(i \notin \mathbf{Y}) \Pr(j \notin \mathbf{Y})
\end{aligned}
$$

Complement of a diversifying process also encourage diversity.

Larger marginal distribution

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

- 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.

Quality vs Diversity

- Think of a Gram matrix, let each column matrix x_i:

$$q_i = \|x_i\|_{L_2} \qquad \phi_i = \frac{x_i}{q_i} \implies \|\phi_i\| = 1$$

- Let $Q = \begin{bmatrix} q_i & 0 & \dots & 0 \\ 0 & q_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & q_n \end{bmatrix} \implies [q_1\phi_1 \quad q_2\phi_2 \quad \dots \quad q_n\phi_n] = \Phi Q$

$$L(x_1, \dots, x_n) = X^\top X = (\Phi Q)^\top (\Phi Q) = Q^\top \Phi^\top \Phi Q$$
$$\implies L_{ij} = q_i \phi_i^\top \phi_j q_j$$

- $S_{ij} \equiv \phi_i^\top \phi_j \in [-1, 1] \implies S_{ij} = \frac{L_{ij}}{\sqrt{L_{ii} L_{jj}}}$
- $\Pr_L(\mathbf{Y} = Y)$ can be viewed as the product of four determinants

$$\Pr_L(\mathbf{Y} = Y) \propto \left(\prod_{i \in Y} q_i^2 \right) \det(S_Y)$$

Conditional (1)

- for $(B \subseteq \mathcal{Y}) \cap A = \emptyset$:

$$\begin{aligned}\Pr_L(\mathbf{Y} = B | A \cap \mathbf{Y} = \emptyset) &= \frac{\Pr_L\left((\mathbf{Y} = B) \cap (A \cap \mathbf{Y} = \emptyset)\right)}{\Pr_L(A \cap \mathbf{Y} = \emptyset)} \\ &= \frac{\underbrace{\Pr_L\left(A \cap \mathbf{Y} = \emptyset | \mathbf{Y} = B\right)}_{\Pr=1 \text{ when } B \cap A = \emptyset} \Pr_L\left(\mathbf{Y} = B\right)}{\Pr_L(A \cap \mathbf{Y} = \emptyset)} \\ &= \frac{\Pr_L\left(\mathbf{Y} = B\right)}{\Pr_L(A \cap \mathbf{Y} = \emptyset)}\end{aligned}$$

- all combination of $\mathbf{Y} \subseteq \bar{A}$
- **note** the usual form is defined for $\Pr_L(\bar{A} \subseteq \mathbf{Y}) = \det(L + I_{\bar{A}})$
- normally, $\Pr_L$ is defined in terms of equality sign. However, it is now defined in terms of a set. Therefore, we need the summation of all the equal terms:

$$= \frac{\frac{\det(L_B)}{\det\left(L_{\mathcal{Y}} + I\right)}}{\frac{\sum_{B': B' \cap A = \phi} \det(L_{B'})}{\det\left(L_{\mathcal{Y}} + I\right)}} = \frac{\det(L_B)}{\underbrace{\sum_{B': B' \cap A = \phi} \det(L_{B'})}_{\text{this is the normalisation constant from } \mathcal{Y} \to \bar{A}}}$$

$$= \frac{\det(L_B)}{\det(L_{\bar{A}} + I)}$$

Conditional (2)

For $(B \subseteq \mathcal{Y}) \cap A = \emptyset$:

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

Sampling DPP:

- Can you sample directly from $\mathcal{Y} \sim L$? Computationally impossible, as there are 2^N combinations.
- You can NOT sample from K either. Since K is defined as marginal, you can NOT add up all cases.
- The solution, express DPP in terms of mixture of **elementary** DPPs:

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

- Mixing weights of using V_J to construct **elementary** DPP is: $\frac{\prod_{n \in J} \lambda_n}{\det(L+I)} = \frac{\prod_{n \in J} \lambda_n}{\prod_{n=1}^{N}(\lambda_n+1)}$
- For example, let $J = \{1, 3, 5\}$, it's mixing weights are $\frac{\lambda_1 \lambda_3 \lambda_5}{\prod_{n=1}^{N}(\lambda_n+1)}$
- Now, we can decide the probability of inclusion of a single V_j to construct **elementary** DPP! For example, let $N = 3$, and we need to decide the inclusion of the element 1:

$$\frac{\lambda_1 + \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_1\lambda_2\lambda_3}{(\lambda_1+1)(\lambda_2+1)(\lambda_3+1)} = \frac{\lambda_1(1 + \lambda_2 + \lambda_3 + \lambda_2\lambda_3)}{(\lambda_1+1)(\lambda_2+1)(\lambda_3+1)}$$
$$= \frac{\lambda_1(1+\lambda_2)(1+\lambda_3)}{(\lambda_1+1)(\lambda_2+1)(\lambda_3+1)} = \frac{\lambda_1}{(\lambda_1+1)}$$

Elementary DPP:

- A DPP, is called **elementary** if every eigenvalue of its marginal kernel is $\in \{0, 1\}$, i.e., $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
- We write $\mathcal{P}^V$, where V is a set of **orthonormal** vectors, to denote an elementary DPP with marginal kernel $K^V = \sum_{v \in V} vv^T$. In the above example, $v \in \left\{ \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = 0 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} + 1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} + 1 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$$

Multi-Linearity

- We let $W_J = \sum_{n \in J} W_n$, where W_n is rank-one matrix:

$$\det(W_J) = \det\left(\sum_{n \in J} W_n\right) = \det\left([(W_J)_1, (W_J)_2, \ldots, (W_J)_k]\right)$$
$$= \det\left(\left[\left(\sum_{n \in J} W_n\right)_1, (W_J)_2, \ldots, (W_J)_k\right]\right) \qquad \text{expand first term}$$

- because **Multi-linearity** states:

$$\det([a_1 + b_1, a_2, \ldots, a_k]) = \det([a_1, a_2, \ldots, a_k]) + \det([b_1, a_2, \ldots, a_k])$$

- Therefore,

$$\det(W_J) = \sum_{n \in J} \det\left(\left[(W_n)_1, (W_J)_2, \ldots, (W_J)_k\right]\right)$$

- Now, we repeat the same thing for the second term and all subseqnet times,
- But we can't use n again, we have to give a different index $n_i \in J \quad \forall i$:

$$\det(W_J) = \sum_{n_1 \in J} \sum_{n_2 \in J} \cdots \sum_{n_k \in J} \det\left(\left[(W_{n_1})_1, (W_{n_2})_2, \ldots, (W_{n_k})_k\right]\right)$$

Elementary DPP:

$$\det(W_J) = \sum_{n_1 \in J} \sum_{n_2 \in J} \cdots \sum_{n_k \in J} \det\left(\left[(W_{n_1})_1, (W_{n_2})_2, \ldots, (W_{n_k})_k\right]\right)$$

- Not every term is non-zero.
- Since W_n is rank one matrix, $(W_n)_i$ and $(W_n)_j$ are linearly dependant. Therefore, the determinant of any matrix containing two or more columns of W_n is zero, for example,

$$\det(W_J) = \det\left(\left[(W_{n_1})_1, (W_{n_1})_2, \ldots, (W_{n_k})_k\right]\right) = 0$$

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

$$\begin{aligned}
\det(W_J) &= \sum_{n_1 \in J} \sum_{n_2 \in J} \cdots \sum_{n_k \in J} \det\left(\left[(W_{n_1})_1, (W_{n_2})_2, \ldots, (W_{n_k})_k\right]\right) \\
&= \underbrace{\sum_{n_1 \in J} \sum_{n_2 \in J} \cdots \sum_{n_k \in J}}_{n_1, n_2, \ldots, n_k \text{ are distinct}} \det\left(\left[(W_{n_1})_1, (W_{n_2})_2, \ldots, (W_{n_k})_k\right]\right) \\
&= \underbrace{\sum_{n_1, n_2, \ldots, n_k \in J}}_{\text{are distinct}} \det\left(\left[(W_{n_1})_1, (W_{n_2})_2, \ldots, (W_{n_k})_k\right]\right)
\end{aligned}$$

Mixture of elementary DPPs

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

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

Start from from mixture of elementary DPPs:

$$\begin{aligned}
&\frac{1}{\det(L+I)} \sum_{J \subseteq \{1,2,\ldots,N\}} \mathcal{P}^{V_J} \prod_{n \in J} \lambda_n \\
&= \frac{1}{\det(L+I)} \sum_{J \subseteq \{1,2,\ldots,N\}} \det\left(K^{V_J}\right) \prod_{n \in J} \lambda_n \\
&= \frac{1}{\det(L+I)} \sum_{J \subseteq \{1,2,\ldots,N\}} \det\left(\sum_{n \in J} W_n\right) \prod_{n \in J} \lambda_n \\
&= \frac{1}{\det(L+I)} \sum_{J \subseteq \{1,2,\ldots,N\}} \underbrace{\sum_{n_1,n_2,\ldots,n_k \in J}}_{\text{are distinct}} \det\left(\left[(W_{n_1})_1, (W_{n_2})_2, \ldots, (W_{n_k})_k\right]\right) \prod_{n \in J} \lambda_n
\end{aligned}$$

Mixture of elementary DPPs

$$\frac{1}{\det(L+I)} = \frac{1}{\det(L+I)} \sum_{J\subseteq\{1,2,\ldots,N\}} \underbrace{\sum_{n_1,n_2,\ldots,n_k\in J}}_{\text{are distinct}} \det\left(\left[(W_{n_1})_1, (W_{n_2})_2, \ldots, (W_{n_k})_k\right]\right) \prod_{n\in J} \lambda_n$$

- For the outter loop, $\sum_{J\subseteq\{1,2,\ldots,N\}}$ when $|J| < k$, then, the inner loop is zero.
- Therefore, we need $J \subseteq \{1, 2, \ldots, N\} \rightarrow J \supseteq \{n_1, n_2, \ldots n_k\}$
- We can also remove $\in J$ from $n_1, n_2, \ldots n_k \in J$
- By swapping the inner and outter loops, we have:

$$\frac{1}{\det(L+I)} = \frac{1}{\det(L+I)} \sum_{J\subseteq\{1,2,\ldots,N\}} \underbrace{\sum_{n_1,n_2,\ldots,n_k\in J}}_{\text{are distinct}} \det\left(\left[(W_{n_1})_1, (W_{n_2})_2, \ldots, (W_{n_k})_k\right]\right) \prod_{n\in J} \lambda_n$$

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

Mixture of elementary DPPs

$$\Pr_L = \frac{1}{\det(L+I)} \underbrace{\sum_{n_1,n_2,\ldots,n_k}}_{\text{are distinct}} \det\left([(W_{n_1})_1, (W_{n_2})_2, \ldots, (W_{n_k})_k]\right) \sum_{J \supseteq \{n_1,n_2,\ldots,n_k\}} \prod_{n \in J} \lambda_n$$

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

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

Mixture of elementary DPPs

$$\Pr_L = \frac{1}{\det(L+I)} \underbrace{\sum_{n_1,n_2,\ldots,n_k}}_{\text{are distinct}} \det\left(\left[(W_{n_1})_1, (W_{n_2})_2, \ldots, (W_{n_k})_k\right]\right) \sum_{J \supseteq \{n_1,n_2,\ldots,n_k\}} \prod_{n \in J} \lambda_n$$

$$= \frac{1}{\prod_{n=1}^{N}(\lambda_n+1)} \underbrace{\sum_{n_1,n_2,\ldots,n_k}}_{\text{are distinct}} \det\left(\left[(W_{n_1})_1, (W_{n_2})_2, \ldots, (W_{n_k})_k\right]\right) \frac{\lambda_{n_1}}{\lambda_{n_1}+1} \cdots \frac{\lambda_{n_k}}{\lambda_{n_k}+1} \prod_{n=1}^{N}(\lambda_n+1)$$

$$= \underbrace{\sum_{n_1,n_2,\ldots,n_k}}_{\text{are distinct}} \det\left(\left[(W_{n_1})_1, (W_{n_2})_2, \ldots, (W_{n_k})_k\right]\right) \frac{\lambda_{n_1}}{\lambda_{n_1}+1} \cdots \frac{\lambda_{n_k}}{\lambda_{n_k}+1}$$

$$= \underbrace{\sum_{n_1,n_2,\ldots,n_k}}_{\text{are distinct}} \det\left(\left[(W_{n_1})_1 \frac{\lambda_{n_1}}{\lambda_{n_1}+1}, (W_{n_2})_2 \frac{\lambda_{n_2}}{\lambda_{n_2}+1}, \ldots, (W_{n_k})_k \frac{\lambda_{n_k}}{\lambda_{n_k}+1}\right]\right)$$

$$= \det\left(\sum_{n=1}^{N} \frac{\lambda_n}{\lambda_n+1} W_n\right) = \det(K_A)$$

Sequential DPP

- Data can be stream in, for example, news sourced from various places, but they do **not** change significantly between consecutive times.
- Our aim is to select the most diverse subset of news to display at each time interval - to avoid show similar pieces. This is done through DPP
- Rather than applying separate DPP sampling at each time stamp, we use Sequential Monte Carlo.
- We have a set of $\{\{y_k^{(i)}, W_k^{(i)}\}_{i=1}^N\}_{k=1}^t$ representing a distribution of selected news articles.

Vanilla Importance Sampling

- We do **not** know how to sample γ_n:

$$\pi_n(x) = \frac{\gamma_n(x)}{Z_n}$$

- but instead, we sample N particles $\{X_n^{(i)}\}$ from η_n, and have:

$$\eta_n^N(\mathrm{d}x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_n^{(i)}}(\mathrm{d}x)$$

- and compute their weights:

$$w_n(x) = \frac{\gamma_n(x)}{\eta_n(x)}$$

- These "weighted" particles can be used to approximate:
  - Expectation of a function $\Psi(x)$: $\mathbb{E}_{\pi_n}(\Psi) = Z_n^{-1} \int \Psi(x) w_n(x) \eta_n(x) \mathrm{d}x$
  - Partition function: $Z_n = \int w_n(x)\eta_n(x)\mathrm{d}x$

Vanilla Sequential Importance Sampling

- At time $n-1$, we have N particles $\{X_{n-1}^{(i)}\}$ distributed according to η_{n-1}
- We propose to move these particles using a Markov kernel K_n with associated density $K_n(x, x')$
- Particles are are marginally distributed according to:

$$\eta_n(x') = \int_E \eta_n(x) K_n(x, x') \mathrm{d}x$$

- If η_n can be computed point-wise, then it is possible to use the standard IS estimates of π_n and Z_n
- The **limitation** is that, in most cases, it's impossible to compute importance distribution $\eta_n(x_n)$:

$$\eta_n(x_n) = \int \eta_1 \prod_{k=1}^{n} K_k(x_{k-1}, x_k) \mathrm{d}x_{1:n-1}$$

Sequential Monte Carlo (1)

- Sequential Monte Carlo (SMC) performs Importance Sampling between:
  - artificial joint sequential distribution

$$\widetilde{\pi}_t(X_{1:t}) = \frac{\widetilde{\gamma}(X_{1:t})}{Z_t} = \frac{\gamma_t(x_t)\prod_{k=1}^{t-1} L_k(x_{k+1}, x_k)}{Z_t}$$

  - **and** joint importance distribution:

$$\eta_t(X_{1:t}) = \eta_1(x_1)\prod_{k=2}^{t} K_k(x_{k-1}, x_k)$$

- $\widetilde{\pi}_n(x_{1:n})$ admits $\pi_n(x_n)$ as a marginal by construction.
- $L_k(x_{k+1}, x_k)$ are backward Markov kernels.
- $K_k(x_{k-1}, x_k)$ are forward Markov kernels.

Sequential Monte Carlo (2)

- The unnormalized importance weights:

$$w_t(X_{1:t}) = \frac{\widetilde{\gamma}_t(X_{1:t})}{\eta_t(X_{1:t})} = w_{t-1}(x_{1:t-1})\widetilde{w}_t(x_{t-1}, x_t)$$

- (unnormalized) incremental weight:

$$\widetilde{w}_t(x_{t-1}, x_t) = \frac{\gamma_t(x_t)L_{t-1}(x_t, x_{t-1})}{\gamma_{t-1}(x_{t-1})K_t(x_{t-1}, x_t)}$$

- $\pi_t(x_t)$ is **marginal distribution** of $\widetilde{\pi}_t(X_{1:t})$, it can be approximated by sampled N particle-weight pairs $\{X_{1:t}^{(i)}, W_{1:t}^{(i)}\}_{i=1}^N$,

$$\pi_t^N = \sum_{i=1}^N W_t^{(i)}\delta(X_t^{(i)})$$

Sub-optimal backward kernel

- When $\pi_t \approx \pi_{t-1}$ a good approximation for optimal L_{t-1}:

$$L_{t-1}(x_t, x_{t-1}) = \frac{\pi_t(x_{t-1})K_t(x_{t-1}, x_t)}{\pi_t(x_t)}$$

- Substituting it into $\widetilde{w}_t(x_{t-1}, x_t)$, the unnormalized incremental weight is reformulated as:

$$\begin{aligned}\widetilde{w}_t(x_{t-1}, x_t) &= \frac{\gamma_t(x_t)L_{t-1}(x_t, x_{t-1})}{\gamma_{t-1}(x_{t-1})K_t(x_{t-1}, x_t)} \\ &= \frac{\gamma_t(x_t)}{\gamma_{t-1}(x_{t-1})K_t(x_{t-1}, x_t)} \times \frac{\pi_t(x_{t-1})K_t(x_{t-1}, x_t)}{\pi_t(x_t)} \\ &= \frac{\gamma_t(x_{t-1})}{\gamma_{t-1}(x_{t-1})}\end{aligned}$$

- Require the following to make it work:

γ_{t-1} or π_{t-1}

γ_t or π_t

$\{x_{t-1}^{(i)}\}_{i=1}^N$ at time stamp $t-1$

Bring DPP into SMC framework: Sequential DPP

$$\pi_t(\mathbf{Y} = y_t) = \frac{\det(M_{t,y_t})}{\det(M_t + I)}$$

- The matrix M_t is constructed from data $X_t = (x_{t1}, \ldots, x_{tN_t})^T$ with each $x_{ti} \in R^D$ using $M = X_t X_t^T$
- assume at t, dataset X_t differs from X_{t+1} by a only a few elements $\implies \pi_{t-1} \approx \pi_t$
- For simplicity, we assume that $|X_t| = |X_{t-1}|$.
- at time $t = 1$, we samples $\left\{ \mathbf{Y}_1^{(i)} \sim \det\left(L_{1,y_1^{(i)}} \right) \right\}_{i=1}^{N}$ using a **fast DPP sampling** (Kang 2013)
- at each time $t > 1$, we update these samples from $\{y_{t-1}^{(i)}\}_{i=1}^{N}$ using SMC scheme.

Incremental weight for sequential DPPs

$$\widetilde{w}_t(y_{t-1}, y_t) = \frac{\det(M_{t,y_{t-1}})/\det(M_t + I)}{\det(M_{t-1,y_{t-1}})/\det(M_{t-1} + I)}$$
$$\propto \det(M_{t,y_{t-1}})/\det(M_{t-1,y_{t-1}})$$

- difference between $\det(M_{t,x_{t-1}})$ and $\det(M_{t-1,y_{t-1}})$ is small
- Let $\mathbb{M}^{\mathbf{c},\mathbf{c}}$ denote the shared sub-matrix between $M_{t,y_{t-1}}$ and $M_{t-1,y_{t-1}}$
- $M_{t,y_{t-1}} = \begin{bmatrix} \mathbb{M}^{\mathbf{c},\mathbf{c}} & M^{c,t} \\ M^{t,c} & M^{t,t} \end{bmatrix}$
- $M_{t-1,y_{t-1}} = \begin{bmatrix} \mathbb{M}^{\mathbf{c},\mathbf{c}} & M^{c,t-1} \\ M^{t-1,c} & M^{t-1,t-1} \end{bmatrix}$
- The **trick** is determinant ratio can be computed **efficiently** by applying determinant formula of partitioned block matrices.
- **no need** to compute nominator nor denominator explicitly.
- efficiently compute incremental weights:

$$\widetilde{w}_t(y_{t-1}, y_t) \propto \frac{\det(M^{c,t} - M^{tc}(\mathbb{M}^{\mathbf{c},\mathbf{c}})^{-1}M^{c,t})}{\det(M^{c,t-1} - M^{t-1,c}(\mathbb{M}^{\mathbf{c},\mathbf{c}})^{-1}M^{c,t-1})}$$

Computation at the Inverse

$$\widetilde{w}_t(y_{t-1}, y_t) \propto \frac{\det(M^{c,t} - M^{tc}(\mathbb{M}^{\mathbf{c},\mathbf{c}})^{-1}M^{c,t})}{\det(M^{c,t-1} - M^{t-1,c}(\mathbb{M}^{\mathbf{c},\mathbf{c}})^{-1}M^{c,t-1})}$$

- $(\mathbb{M}^{\mathbf{c},\mathbf{c}})^{-1}$ may still be computational expensive.
- However it can be achieved by repeatably applying block-inversion formula (cache in the memory):

$$M^{-1} = \begin{pmatrix} L & b \\ b^\top & c \end{pmatrix}^{-1} = \begin{pmatrix} L^{-1} + L^{-1}bb^\top L^{-1}d^{-1} & -L^{-1}bd^{-1} \\ -b^\top L^{-1}d^{-1} & d \end{pmatrix}^{-1}$$

Algorithm 1 Fast sampling for sequential DPPs

Require: $\{M_i\}_{i=1}^t$

1: $\{\mathbf{Y}_1^{(i)} \sim \det(M_{1,y_1^{(i)}})\}_{i=1}^N$ by (Kang 2013)
2: $\{W_1^{(i)} = 1/N\}_{i=1}^N$
3: **for** $k = 2, ..., t$ **do**
4: $\{\widetilde{w}_k(y_{k-1}^{(i)}, y_k^{(i)}) = \det(L_{k,y_{k-1}^{(i)}})/\det(L_{k-1,y_{k-1}^{(i)}})\}_{i=1}^N$
5: $\{W_k^{(i)} = W_{k-1}^{(i)}\widetilde{w}_k^{(i)}/\sum_{i=1}^N(W_{k-1}^{(i)}\widetilde{w}_k^{(i)})\}_{i=1}^N$ [a]
6: $\{y_k^{(i)} \sim K(y_{k-1}^{(i)}, y_k^{(i)})\}_{i=1}^N$
7: $N_{k,ESS} = \{\sum_{i=1}^N(W_k^{(i)})^2\}^{-1}$
8: **if** $N_{k,ESS} < \alpha \cdot N$ **then**
9: $\{y_k^{(i)} \sim \text{Mult}(W_k^{(1)}, ..., W_k^{(N)})\}_{i=1}^N$
10: $\{W_k^{(i)} = 1/N\}_{i=1}^N$
11: move $\{y_k^{(i)}\}_{i=1}^N$ by π_k invariant MCMC kernel $K_{\pi_k}(y_k^{(i)}, \cdot)$
12: **end if**
13: **end for**
14: **return** $\{\{y_k^{(i)}, W_k^{(i)}\}_{i=1}^N\}_{k=1}^t$

[a] We use $W_k^{(i)}$ and $\widetilde{w}_k^{(i)}$ to replace $W_k(y_{1:k}^{(i)})$ and $\widetilde{w}_k(y_{k-1}^{(i)}, y_k^{(i)})$ for simplicity