On the Convex Combination of Determinantal Point Processes

Tatsuya Matsuoka TA.MATSUOKA@NEC.COM
NEC Corporation

Naoto Ohsaka OHSAKA@NEC.COM
NEC Corporation

Akihiro Yabe* A_YABE@FANFARE-KK.COM
Fanfare Inc.

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

Determinantal point processes (DPPs) are attractive probabilistic models for expressing item quality and set diversity simultaneously. Although DPPs are widely-applicable to many subset selection tasks, there exist simple small-size probability distributions that any DPP cannot express. To overcome this drawback while keeping good properties of DPPs, in this paper we investigate the expressive power of *convex combinations of DPPs*. We provide upper and lower bounds for the number of DPPs required for *exactly* expressing any probability distribution. For the *approximation* error, we give an upper bound on the Kullback–Leibler divergence $n - \lfloor \log t \rfloor + \epsilon$ for any $\epsilon > 0$ of approximate distribution from a given joint probability distribution, where t is the number of DPPs. Our numerical simulation on an online retail dataset empirically verifies that a convex combination of only two DPPs can outperform a nonsymmetric DPP in terms of the Kullback–Leibler divergence. By combining a polynomial number of DPPs, we can express probability distributions induced by bounded-degree pseudo-Boolean functions, which include weighted coverage functions of bounded occurrence.

Keywords: Determinantal point processes, Convex combination

1. Introduction

Subset selection tasks play an immense role in a wide range of situations wherein one would like to extract a small number of representative items. Often good subsets are expected to consist of high-quality and diverse items. For example, in recommender systems, a few candidates that possess both high reputation and distinct characteristics should be presented to users so that any user can examine every candidate and adopts at least one of them.

Determinantal point processes (DPPs) offer an appealing probabilistic model which achieves a balance between item quality and set diversity, and have been widely applied to many tasks in machine learning, e.g., recommender systems (Wilhelm et al., 2018), image search (Kulesza and Taskar, 2011a), and document summarization (Kulesza and Taskar, 2011b). Given a kernel matrix L, a DPP defines the probability mass for each subset A in proportion to $\det(L_A)$, the principal minor of L indexed by A. DPPs exhibit *negative*

*. The work of the third author was done while he worked at NEC Corporation.

correlations, which intuitively mean that similar items rarely appear simultaneously. Unlike other probabilistic models such as graphical models (Cooper, 1990; Dagum and Luby, 1993), DPPs make various inference tasks tractable; e.g., normalization and sampling can be performed in polynomial time (Hough et al., 2006; Kulesza and Taskar, 2012).

However, one serious drawback of DPPs is that their expressive power is limited; there exist simple probability distributions that any DPP cannot express exactly. Several classes of point processes generalizing DPPs have been proposed, e.g., signed DPPs (Brunel, 2018), nonsymmetric DPPs (Gartrell et al., 2019), exponentiated DPPs (Anari and Gharan, 2017; Mariet et al., 2018), and Π-DPPs (Ohsaka and Matsuoka, 2020). However, efficient inference is sacrificed, or the expressive power is still limited (see Section 2). Thus, we quest for a class of point processes that are more expressive than DPPs while offering efficient algorithms for inference tasks.

Our choice for the above purpose is a *convex combination* of DPPs (hereafter called a *CC-DPP*), appearing in Kulesza and Taskar (2011a). We can conduct some inference tasks for CC-DPPs by utilizing efficient algorithms for inference tasks on positive semidefinite kernel DPPs. For example, for sampling, we can adapt *any* sampling algorithms defined for positive semidefinite kernel DPPs. Let t denote the number of DPPs constructing the CC-DPP (we call it the size of a CC-DPP)[1]. If we utilize the algorithm by Dereziński et al. (2019) for sampling on a positive semidefinite kernel DPP, we can draw a random sample from a CC-DPP in $\mathrm{O}(t+\mathrm{poly}(\mathbb{E}(|S|)))$ time with $\mathrm{O}(tn\cdot\mathrm{poly}(\mathbb{E}(|S|))\mathrm{polylog}(n))$-time preprocessing, where n is the cardinality of the ground set, $\mathbb{E}$ denotes the expectation, and S is a set sampled ($\mathbb{E}(|S|)\leq\mathrm{rank}(L)$). For other fundamental inference tasks including nomalization, marginalization, and conditioning, the computational cost is roughly t times that of a positive semidefinite kernel DPP. For the space of the model, a CC-DPP takes space roughly t times that for storing a positive semidefinite kernel DPP.

In addition to time and space efficiency, the major interest in the model is its representability. For representability, however, not much is known even for DPPs. In this paper, we focus on the representability of convex combinations of DPPs from both theoretical and empirical approaches.

1.1. Our Contributions

In this paper, we conduct a study on convex combinations of DPPs systematically[2]. Given t kernel matrices $L^1,\dots,L^t$ along with nonnegative real numbers $\lambda_1,\dots,\lambda_t$ with $\sum_{i=1}^t\lambda_i=1$, we define the probability mass for each subset A as $\sum_{i=1}^t\lambda_i\cdot\det(L^i_A)/\det(L^i+I)$. We call a CC-DPP with t DPPs a *size-t CC-DPP*. We investigate the expressive power of CC-DPPs through theoretical analysis, numerical study, and introduction of a concrete class. Our contributions are detailed below.

Properties and Inference (Section 3.3). We demonstrate that some properties of DPPs and inference algorithms for DPPs can be extended to CC-DPP case with certain computational cost.

1. There exist CC-DPPs with different sizes for the common probability distribution.
2. Note that a convex combination of DPPs is quite different from a DPP defined by a convex combination of kernel matrices.

Number of Necessary DPPs for Exact Representation (Section 4). We investigate the exact representability of CC-DPPs. We devise upper and lower bounds on the number of DPPs required for *exactly* expressing *any* probability distribution.

Approximation Error (Section 5). Since the bounds appearing in Section 4 are bounds for the *exact* representation, we expect that combining a smaller number of DPPs is enough for *approximating* any distributions. Thus, we investigate the approximation error for CC-DPPs. For a given joint probability distribution $\boldsymbol{q}^*$ and a size-t CC-DPP of $n \times n$ matrices, we give an upper bound $n - \lfloor \log t \rfloor + \epsilon$ on the Kullback–Leibler divergence of approximate distribution from $\boldsymbol{q}^*$ for an arbitrary positive ϵ.

Numerical Study (Section 6). Since the upper bound for the approximation error presented in Section 5 is for the *worst case*, we investigate the empirical performance of the model for real data. Our simulation results indicate the superiority of a convex combination of DPPs over a DPP on the Kullback–Leibler divergence. Using a real dataset of online retail, we confirm that the obtained size-2 CC-DPP is superior to the obtained nonsymmetric DPP, and one kernel seems to roughly represent importance of each element and the other kernel seems to roughly represent negative correlations between each pair of elements.

Representable Distribution Class (Section 7). By a convex combination of a polynomial number of DPPs, we can exactly express certain point processes. We can represent some point processes appearing in Iyer and Bilmes (2015) as subclasses of the submodular point process, by polynomial-size CC-DPPs. Concretely, we represent point processes induced by bounded-degree pseudo-Boolean functions, which include weighted coverage functions of bounded occurrence.

1.2. Scope of This Paper

We remark that, if not explicitly written, we mostly deal with *real-valued positive semidefinite* kernel DPPs and their convex combinations in this paper since we can conduct some inference tasks for real-valued positive semidefinite DPPs faster than for other types of DPPs, see, e.g., Kulesza and Taskar (2012). In this paper, we often write simply as a "DPP" to mean a real-valued positive semidefinite kernel DPP. Nonsymmetric DPPs have stronger expressive power (Gartrell et al., 2019), and hence we adopt a nonsymmetric DPP for comparison in the numerical simulation (Section 6).

2. Related Work

The determinantal point process is initially proposed by Macchi (1975) as a model for fermions in statistical physics. In the machine learning community, DPPs have received significant attention from researchers by virtue of a number of desirable features (see, e.g., Kulesza and Taskar (2012)). For example, DPPs are closed under complement and restriction, and inference tasks including (constrained) sampling can be conducted in polynomial time (Kulesza and Taskar, 2011a; Celis et al., 2017).

Several studies have been conducted on devising a rich class of point processes by generalizing DPPs. *Signed DPPs* considered by Brunel (2018) ensure that $K_{ji} = \pm K_{ij}$ for any $i \neq j$ for the marginal kernel K, which is no longer symmetric. Signed DPPs can encode

both repulsion and attraction to a certain degree, but the precise expressive power is not well-understood. Gartrell et al. (2019) investigate *nonsymmetric DPPs*, which admit a general nonsymmetric real matrix as a kernel matrix. This class is obviously more general than signed DPPs as well as DPPs but still cannot represent the example introduced in Section 7. Exponentiated DPPs (Anari and Gharan, 2017; Mariet et al., 2018) define the probability of taking a subset A as in proportion to $\det(L_A)$ to the p-th power, where p is a fixed number. Although exponentiated DPPs have stronger expressive power and the algorithms on approximation have been investigated, performing exact normalizing is hard (Gurvits, 2005; Ohsaka and Matsuoka, 2020). Probability distributions defined by the product of DPPs (Π-DPPs) are studied (Ohsaka and Matsuoka, 2020), but although the paper gives some FPT results, hardness results are also given for computing a nomarlizing constant (Ohsaka and Matsuoka, 2020). *Log-submodular point processes (Log-SPPs)* by Djolonga and Krause (2014) and Gotovos et al. (2015) are a generalization that defines the probability mass for any subset A in proportion to $\exp(\beta f(A))$, where f is a submodular set function, and $\beta > 0$ is a scaling parameter. A related class is *submodular point processes (SPPs)* proposed by Iyer and Bilmes (2015), probability mass for any subset A of which is given in proportion to $f(A)$ for a submodular set function f. Since computing the normalizing constant for Log-SPPs and SPPs requires exponential time (Iyer and Bilmes, 2015), we have to resort to approximate inference by sampling. Unlike the above models, the convex combination of DPPs expresses practical classes of point processes (Section 7) and allows us to perform inference tasks exactly.

The most closely related work is that of Kulesza and Taskar (2011a), who propose to take a convex combination of DPPs. The original motivation of Kulesza and Taskar (2011a) is to learn k-DPPs by optimizing weights over fixed DPPs. Our study reveals that taking convex combinations over multiple DPPs results in a rich class of point processes while maintaining computational tractability.

3. Preliminaries

This section gives notations used in this paper, reviews the definition and basic properties of DPPs, and introduces the definition of CC-DPPs. Table 1 lists definitions and notations frequently used in this paper.

Notations. Let $V = \{1, \dots, n\}$ denote the ground set of a finite number of items. A *point process* $\mathcal{P}$ on V is defined as a probability measure on the power set 2^V. Consider a random subset $\boldsymbol{Y}$ drawn from a point process on V, which can be the empty set $\emptyset$, the ground set V, or anything between them. Then, for a set $A \subseteq V$, we use $\mathcal{P}(A \subseteq \boldsymbol{Y})$ to denote the probability that $\boldsymbol{Y}$ includes A, which will be referred to as the *marginal probability*, and we use $\mathcal{P}(\boldsymbol{Y} = A)$ to denote the probability that $\boldsymbol{Y}$ is exactly equal to A, which will be referred to as the *joint probability*. We have a simple relation between them that is $\mathcal{P}(A \subseteq \boldsymbol{Y}) = \sum_{A \subseteq B \subseteq V} \mathcal{P}(\boldsymbol{Y} = B)$. Since a point process can be thought of as a probability vector over the possible subsets of V, we use $\boldsymbol{p} = (p_A)_{A \subseteq V} \in \mathbb{R}^{2^V}$ to denote the vector corresponding to the marginal probabilities, where $p_A = \mathcal{P}(A \subseteq \boldsymbol{Y})$ for some point process, and we use $\boldsymbol{q} = (q_A)_{A \subseteq V} \in \mathbb{R}^{2^V}$ to denote the *joint probability vector*, where $q_A = \mathcal{P}(\boldsymbol{Y} = A)$ for some point process. Note that we align elements in such vectors in lexicographical order.

Table 1: Notations used in this paper.

notation	description
$V = \{1, \dots, n\}$	the ground set of n items
$\mathcal{P}$	a point process defined on V
$\boldsymbol{p} = (p_A)_{A \subseteq V}$	a vector in $\mathbb{R}^{2^V}$ for a point process over V with $p_A = \mathcal{P}(A \subseteq \boldsymbol{Y})$ (representing marginal probabilities)
$\boldsymbol{q} = (q_A)_{A \subseteq V}$	a joint probability vector in $\mathbb{R}^{2^V}$ for a point process over V with $q_A = \mathcal{P}(\boldsymbol{Y} = A)$
K	a marginal kernel in $\mathbb{R}^{V \times V}$ that defines a DPP
L	a joint probability kernel in $\mathbb{R}^{V \times V}$ that defines an L-ensemble
$M_A = (M_{ij})_{i,j \in A}$	the restriction of a matrix M to the elements indexed by $A \subseteq V$
t	the number of kernels over which a convex combination is taken
$\boldsymbol{\lambda} = (\lambda_i)_{i \in [t]}$	a nonnegative real vector with $\sum_{i=1}^t \lambda_i = 1$ (probability vector)
$(K^1, \dots, K^t; \boldsymbol{\lambda})$	a tuple of t marginal kernels and a nonnegative real vector that defines a CC-DPP
$(L^1, \dots, L^t; \boldsymbol{\lambda})$	a tuple of t joint probability kernels and a nonnegative real vector that defines a CC-DPP

In this paper, we adopt 2 as a base of logarithm. For probability distribution $\boldsymbol{q}$ and $\boldsymbol{q}'$, $D(\boldsymbol{q} || \boldsymbol{q}') = \sum_{S \subseteq V} q_S \log(q_S / q'_S)$ is called *Kullback–Leibler divergence* (*KL-divergence*) of $\boldsymbol{q}'$ from $\boldsymbol{q}$ (define the summand as 0 for S with $q'_S = 0$). Note that KL-divergence is 0 if and only if $\boldsymbol{q} = \boldsymbol{q}'$ holds. KL-divergence is nonsymmetric but widely used because of good properties (cf. e.g., Cover and Thomas (2012)).

Determinantal Point Process. A point process $\mathcal{P}$ on V is called a *determinantal point process (DPP)* (Macchi, 1975) if there exists a real positive semidefinite matrix $K \in \mathbb{R}^{V \times V}$ such that $\mathcal{P}(A \subseteq \boldsymbol{Y}) = \det(K_A)$ holds for all $A \subseteq V$, where $K_A = (K_{ij})_{i,j \in A} \in \mathbb{R}^{A \times A}$ is the restriction of K to the elements indexed by A, and we define $\det(K_\emptyset) = 1$. We call K a *marginal kernel*. Every real positive semidefinite DPP satisfies the *log-submodular* property; i.e., $\det(K_A) \det(K_B) \geq \det(K_{A \cup B}) \det(K_{A \cap B})$ for all $A, B \subseteq V$.

We then introduce a slightly restricted class of DPPs called L-ensembles. Given a real positive semidefinite matrix $L \in \mathbb{R}^{V \times V}$, an *L-ensemble* (Borodin and Rains, 2005) defines a DPP whose joint probability satisfies $\mathcal{P}(\boldsymbol{Y} = A) \propto \det(L_A)$ for all $A \subseteq V$. We call L a *joint probability kernel*. The normalizing constant has a simple closed form $\sum_{A \subseteq V} \det(L_A) = \det(L + I)$ (Kulesza and Taskar, 2012), and we thus have that $\mathcal{P}(\boldsymbol{Y} = A) = \det(L_A) / \det(L + I)$. We use marginal kernels K and joint probability kernels L interchangeably.

3.1. Definition of CC-DPPs

A convex combination of DPPs is simply a point process whose probability distribution is given by a convex combination of multiple DPPs. Formally, let $K^1, \dots, K^t$ be t marginal kernels on V, and let $\boldsymbol{\lambda} = (\lambda_i)_{i \in [t]}$ be a nonnegative real vector satisfying $\sum_{i=1}^t \lambda_i = 1$. Then, a *convex combination of DPPs (CC-DPP)* is defined as a point process whose marginal probability for subset $A \subseteq V$ is given by $\mathcal{P}(A \subseteq \boldsymbol{Y}) = \sum_{i=1}^t \lambda_i \cdot \det(K^i_A)$. Similarly, for given t joint probability kernels $L^1, \dots, L^t$ and $\boldsymbol{\lambda} = (\lambda_i)_{i \in [t]}$, we define a CC-DPP as a point

process whose joint probability for subset $A \subseteq V$ is given by $\mathcal{P}(\boldsymbol{Y} = A) = \sum_{i=1}^{t} \lambda_i \cdot \frac{\det(L^i_A)}{\det(L^i+I)}$. We call the number t of marginal (or joint) kernels the *size* of a CC-DPP. We denote CC-DPPs by tuples $(K^1, \dots, K^t; \boldsymbol{\lambda})$ or $(L^1, \dots, L^t; \boldsymbol{\lambda})$.

3.2. Example

Here, we demonstrate that the class of CC-DPPs is wider than that of DPPs. Consider a joint probability vector $\boldsymbol{q} = (\frac{1}{3}, \frac{5}{12}, \frac{1}{12}, \frac{1}{6})$ on $V = \{1, 2\}$. The corresponding marginal probability vector is then $\boldsymbol{p} = (1, \frac{7}{12}, \frac{1}{4}, \frac{1}{6})$. Since $\boldsymbol{p}$ *violates* the log-submodularity; i.e., $p_1 p_2 - p_\emptyset p_{12} = -1/48 < 0$, it cannot be represented by any real positive semidefinite DPP. On the other hand, we can express $\boldsymbol{p}$ by a size-2 CC-DPP: $\frac{1}{2}\mathcal{P}_1 + \frac{1}{2}\mathcal{P}_2$, where marginal kernels for $\mathcal{P}_1$ and $\mathcal{P}_2$ are given by

$$K^1 = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix} \text{ and } K^2 = \begin{pmatrix} \frac{2}{3} & 0 \\ 0 & \frac{1}{2} \end{pmatrix},$$

respectively. Note that, as this example, it may be impossible to represent a size-2 CC-DPP by a single real positive semidefinite DPP even if both kernels are diagonal matrices.

3.3. Properties and Inference

We extend known useful properties and existing efficient inference algorithms for DPPs to CC-DPPs. As a consequence, we can perform a variety of inference tasks for the class of point processes introduced in Section 7 efficiently. For properties and inference tasks, we refer Section 2.2 of Kulesza and Taskar (2012).

3.3.1. Useful Properties

We first demonstrate that the class of CC-DPPs is closed under the following operations. Here, let $\boldsymbol{Y}$ be a random set of a CC-DPP $(K^1, \dots, K^t; \boldsymbol{\lambda})$. (We use marginal kernels following Kulesza and Taskar (2012).)

- **Scaling:** For a scaling parameter $\gamma \in [0, 1]$, create a point process $\mathcal{P}'$ on V with $\mathcal{P}'(S \subseteq \boldsymbol{Y}) = \gamma^{|S|} \cdot \mathcal{P}(S \subseteq \boldsymbol{Y})$ for all $S \subseteq V$.
- **Complement:** Create a point process $\mathcal{P}'$ on V with $\mathcal{P}'(S \subseteq \boldsymbol{Y}) = \mathcal{P}(S \subseteq V \setminus \boldsymbol{Y})$.
- **Restriction:** For a subset $A \subseteq V$, create a point process $\mathcal{P}'$ on V with $\mathcal{P}'(S \subseteq \boldsymbol{Z}) = \sum_{\boldsymbol{Y}\colon \boldsymbol{Z} = \boldsymbol{Y} \cap A} \mathcal{P}(S \subseteq \boldsymbol{Y} \cap A)$.

Claim 1 *For any positive integer t, the class of CC-DPPs of size t is closed under scaling, complement, and restriction. Specifically, for a CC-DPP $(K^1, \dots, K^t; \boldsymbol{\lambda})$ of size t, CC-DPPs obtained by applying each operation have the following form and are of size t: $(\gamma K^1, \dots, \gamma K^t; \boldsymbol{\lambda})$ (scaling), $(I - K^1, \dots, I - K^t; \boldsymbol{\lambda})$ (complement), and $(K^1_A, \dots, K^t_A; \boldsymbol{\lambda})$ (restriction).*

Proof We can adapt Section 2.3 of Kulesza and Taskar (2012) to verify the correctness by simple calculations: $\sum_i \lambda_i \det(\gamma K^i_S) = \sum_i \lambda_i \gamma^{|S|} \mathcal{P}(S \subseteq \boldsymbol{Y})$ for all $S \subseteq V$ (scaling),

$\sum_i \lambda_i \det(I - K^i_S) = \sum_i \lambda_i \mathcal{P}(S \cap \boldsymbol{Y} = \emptyset) = \sum_i \lambda_i \mathcal{P}(S \subseteq V \setminus \boldsymbol{Y})$ for all $S \subseteq V$ (complement), and $\sum_i \lambda_i \det((K^i_A)_S) = \sum_i \lambda_i \det(K^i_S) = \sum_i \lambda_i \mathcal{P}(S \subseteq \boldsymbol{Y} \cap A)$ for all $S \subseteq A$ (restriction). ■

3.3.2. Efficient Inference

We can naturally adapt existing efficient inference algorithms for DPPs to CC-DPPs with roughly t times the cost of that of existing algorithms in the time complexity. Following previous studies (Kulesza and Taskar, 2011a, 2012; Celis et al., 2017, 2018), we use joint probability kernels (except for L-ensemble representation task); let $(L^1, \dots, L^t; \boldsymbol{\lambda})$ be a CC-DPP on V of size t.

- **Normalization:** For a CC-DPP $(L^1, \dots, L^t; \boldsymbol{\lambda})$, one can normalize this CC-DPP just by normalizing each L^i as an L-ensemble.

- **Marginalization:** The marginal probability of a random set including subset $A \subseteq V$ is equal to $\sum_{A \subseteq S \subseteq V} \sum_{i=1}^t \lambda_i \det(L^i_S) / \det(L^i + I)$. Simple calculation (cf. Theorem 2.2 of Kulesza and Taskar (2012)) reveals that any joint-kernel representation $(L^1, \dots, L^t; \boldsymbol{\lambda})$ is translated into an equivalent marginal-kernel representation $(K^1, \dots, K^t; \boldsymbol{\lambda})$, where $K^i = L^i(L^i + I)^{-1}$ for all $i \in [t]$. The whole computation can be done by t times of matrix multiplication in $\mathrm{O}(tn^\omega)$ time, where ω is the exponent of matrix multiplication.

- **L-ensemble Representation:** Conversely, we can translate a marginal-kernel representation $(K^1, \dots, K^t; \boldsymbol{\lambda})$ into an equivalent joint-kernel representation $(L^1, \dots, L^t; \boldsymbol{\lambda})$, where $L^i = K^i(I - K^i)^{-1}$ for all $i \in [t]$ provided that no eigenvalue of K^i's achieves 1, which is the same assumption as for DPPs.

- **Conditioning:** Consider conditioning a CC-DPP on the event that none of the elements in F appears and all of the elements in T appear in the random set for two disjoint sets $F, T \subseteq V$. The resulting point process is a CC-DPP defined by $(\bar{L}^1, \dots, \bar{L}^t; \boldsymbol{\lambda}')$, where $\bar{L}^i = ([(L^i_{V \setminus F} + I_{V \setminus T})^{-1}]_{V \setminus T})^{-1} - I$ and $\lambda'_i \propto \lambda_i \cdot \det(L^i_{V \setminus F} + I_{V \setminus T}) / \det(L^i + I)$ for all $i \in [t]$, where $I_{V \setminus T}$ is a diagonal matrix such that $(I_{V \setminus T})_{ii}$ is 1 if $i \in V \setminus T$ and 0 otherwise, which is a straightforward application of Kulesza and Taskar (2012). Hence, the class of CC-DPPs is closed under this conditioning operation, and the transformation takes $\mathrm{O}(tn^\omega)$ time.

- **Sampling:** By choosing one joint probability kernel L from $L^1, \dots, L^t$ according to the probabilities from $\boldsymbol{\lambda}$ and sampling from L, we can sample a random subset from a CC-DPP $(L^1, \dots, L^t; \boldsymbol{\lambda})$ in $\mathrm{O}(t + n^{\omega+1})$ time by applying a generic algorithm, see, e.g., Lemma 22 of Celis et al. (2017). (Note that, as we see in Section 1, there is an algorithm for sampling on a DPP in $\mathrm{O}(t + \mathrm{poly}(\mathbb{E}(|S|)))$ time with preprocessing in $\mathrm{O}(tn \cdot \mathrm{poly}(\mathbb{E}(|S|)) \mathrm{poly}\log(n))$ time by Dereziński et al. (2019), where $\mathbb{E}$ is the expectation and S is a subset of the ground set, and remark that $\mathbb{E}(|S|) \leq \mathrm{rank}(L)$ holds.)

Thanks to the simple nature of convex combinations, we can adapt existing algorithms for constrained sampling to CC-DPPs as follows.

Fixed-size Sampling. Given a point process and integer k, consider conditioning on the event that the sample is limited to one of size k; i.e., a new point process whose joint probability is proportional to the original one for size-k subsets and 0 for the others. Such constrained point processes for DPPs are called *k-DPPs* (Kulesza and Taskar, 2011a). We here present how to sample a fixed-size set from a CC-DPP $(L^1, \dots, L^t; \boldsymbol{\lambda})$. Normalization for any k is an easy task since $\sum_{S:|S|=k} \sum_{i\in[t]} \lambda_i \cdot \det(L^i_S)/\det(L^i + I)$ is a convex combination of the normalizing constant for k-DPPs defined by L^i's, each of which can be computed in time $\mathrm{O}(n^3)$. Then, observing that the class of size-t CC-DPPs with constraint such that the size of a set sampled is k (over each k) is closed under conditioning, we can use a generic algorithm for sampling on a DPP.

Partition Constrained Sampling. Subset sampling under *partition (or fairness) constraints* is also useful and known as P-DPPs (Celis et al., 2018). Let $P_1, \dots, P_p$ be a partition of V and $b_1, \dots, b_p$ be integers, a partition family is defined as $\mathcal{C} = \{S \subseteq V \mid |S \cap P_j| = b_j \ \forall j \in [p]\}$. Celis et al. (2017) show that if one can construct an oracle calculating a *generating function* $g(\boldsymbol{x}) = \sum_{S\subseteq V} q_S \prod_{i\in S} x_i$ for joint probabilities $\boldsymbol{q}$ and an arbitrary $\boldsymbol{x} \in \mathbb{R}^V$, then one can draw a subset in $\mathcal{C}$ from $\boldsymbol{q}$. Now, let $g_1, \dots, g_t$ be the generating functions for L-ensembles of $L^1, \dots, L^t$. We then have that the generating function for the CC-DPP is $g(\boldsymbol{x}) = \sum_{i=1}^t \lambda_i g_i(\boldsymbol{x})$, which can be computed in time $\mathrm{O}(tn^\omega)$ since we can compute each g_i in time $\mathrm{O}(n^\omega)$ (Fact 8 of Celis et al. (2017)). Consequently, the entire sampling completes in $tn^{p+\mathrm{O}(1)}$ time (Corollary 10 of Celis et al. (2017)).

Budget Constrained Sampling. The tractability of evaluating the generating function of CC-DPPs immediately implies efficient subset sampling under *budget constraints*. Let $\boldsymbol{c} \in \mathbb{Z}^V_{\geq 0}$ be a cost vector, and $b \in \mathbb{Z}_{\geq 0}$ be a budget, and let us denote $\boldsymbol{c}(S) = \sum_{i\in S} c_i$. Then, a linear family is defined as $\{S \subseteq V \mid \boldsymbol{c}(S) = b\}$. Celis et al. (2017)'s method works for this constraint, where the time complexity is bounded by $t\|\boldsymbol{c}\|_1 n^{\mathrm{O}(1)}$, which actually works for even the case of multiple budget constraints (see Corollary 5 of Celis et al. (2017)).

4. Bounds on the Number of Necessary Kernels for Exact Representation

In this section, we provide an analysis on the representability of CC-DPPs. Concretely, we give upper and lower bounds on the number of required kernels for representing *any* given probability distribution *exactly*. Note that it is not the purpose that we actually obtain the DPPs by proofs in this section. For smaller-size CC-DPPs for approximation, we argue in Sections 5 and 6. Note that proofs in this paper majorly use DPPs with diagonal matrix kernels.

4.1. Upper Bound

We first show an upper bound on the number of kernels.

Theorem 1 *There exists a CC-DPP which represents a given probability distribution on $\{1, \dots, n\}$, the size of which is at most 2^{n-1}.*

Proof Let $\boldsymbol{q} \in \mathbb{R}^{2^V}$ be an input joint probability vector. Fix an arbitrary element $v \in V$. Let $\mathcal{F} = \{V'(\subseteq V) \mid v \in V',\ (q_{V'\setminus\{v\}} > 0) \text{ or } (q_{V'} > 0)\}$. For each $V' \in \mathcal{F}$, let $K^{V'}$ be a

diagonal matrix such that $K^{V'}_{ij} = 0$ if $i \neq j$ or $i = j \notin V'$, $K^{V'}_{ii} = 1$ if $i \in V' \setminus \{v\}$, and that $K^{V'}_{vv} = \frac{q_{V'}}{q_{V'\setminus\{v\}}+q_{V'}}$. This $K^{V'}$ corresponds to a joint probability vector with $\frac{q_{V'}}{q_{V'\setminus\{v\}}+q_{V'}}$ for V' and $\frac{q_{V'\setminus\{v\}}}{q_{V'\setminus\{v\}}+q_{V'}}$ for $V' \setminus \{v\}$. Then, one can obtain a CC-DPP presenting the given probability distribution by taking the convex combination of DPPs for all $V' \in \mathcal{F}$ such that $q_{V'\setminus\{v\}}+q_{V'}$ is a coefficient $\lambda_{V'}$ of $K^{V'}$. Since $|\mathcal{F}| \leq 2^{n-1}$, we need at most 2^{n-1} DPPs. ■

4.2. Lower Bound

We show a linear lower bound on the size with the *constructive proof*.

Theorem 2 *There exists a probability distribution on $\{1, \ldots, n\}$ such that we cannot represent by a CC-DPP whose size is less than $n/2$.*

Proof Let $V = \{1, \ldots, n\}$. Let $\boldsymbol{q}$ be the input joint probability vector such that $q_V = q_{V\setminus\{1,2\}} = q_{V\setminus\{1,2,3,4\}} = \cdots = q_{V\setminus\{1,2,\ldots,2\lfloor\frac{n}{2}\rfloor\}} = 1/\lfloor\frac{n}{2}\rfloor$ and all the other elements are 0. Consider constructing a CC-DPP $(\{K^j\}_j; \boldsymbol{\lambda})$ which represents $\boldsymbol{q}$. Since $q_V > 0$ and $q_{V\setminus\{i\}} = 0$ for all $i \in V$, at least one kernel matrix K^j must correspond to a joint probability vector $\hat{\boldsymbol{q}}$ such that $\hat{q}_V > 0$ and that $\hat{q}_{V\setminus\{i\}} = 0$ for all $i \in V$. Without loss of generality, we assume K^1 is such a K^j. Then, since $\det(K^1_V) = \hat{q}_V$ and $\det(K^1_{V\setminus\{i\}}) = \hat{q}_{V\setminus\{i\}} + \hat{q}_V$, $\hat{q}_{V\setminus\{i\}} = 0$ requires the condition $\det(K^1_V) = \det(K^1_{V\setminus\{i\}})$. For $K^1_V = \begin{pmatrix} d & \boldsymbol{v} \\ \boldsymbol{v}^\top & K^1_{V\setminus\{i\}} \end{pmatrix}$, since $\det(K^1_V) = \det(K^1_{V\setminus\{i\}})\det(d - \boldsymbol{v}K^1_{V\setminus\{i\}}\boldsymbol{v}^\top)$, we have $d = 1$ and $\boldsymbol{v} = \boldsymbol{0}$. By doing the same discussion on all i, we have that K^1 is an identity matrix. This means that if $\hat{q}_V > 0$, then all the other elements of $\hat{\boldsymbol{q}}$ are 0. By inductive argument, only one element is positive for each vector corresponding to $q_{V\setminus\{1,2\}}, \ldots, q_{V\setminus\{1,2,\ldots,2\lfloor\frac{n}{2}\rfloor\}}$. Therefore, we need at least $1 + \lfloor\frac{n}{2}\rfloor \geq n/2$ DPPs. ■

5. Bound on Approximation Error

In the previous section, we evaluate lower and upper bounds of the number of necessary DPP kernels for exactly expressing an arbitrary distribution. However, we expect that smaller number of DPPs are enough for a CC-DPP for approximation. Thus, in this section, we investigate an upper bound of KL-divergence. We utilize the method by Montufar et al. (2014); one for evaluation of KL-divergence on a certain mixture model with disjoint supports. For KL-divergence, we obtain the following bound.

Theorem 3 *Let t be a positive integer and ϵ be a positive real number. For an arbitrary joint probability vector $\boldsymbol{q}^* \in \mathbb{R}^{2^V}_{\geq 0}$, there exists a joint probability vector $\boldsymbol{r}$ represented by a CC-DPP of size t satisfying*

$$D(\boldsymbol{q}^*||\boldsymbol{r}) \leq n - \lfloor \log t \rfloor + \epsilon,$$

where n is the cardinality of V. In other words, there exists $\boldsymbol{r}$ such that $D(\boldsymbol{q}^*||\boldsymbol{r})$ is upper bounded by a number which is larger but arbitrarily close to $n - \lfloor \log t \rfloor$.

Proof Let $k = 2^{\lfloor \log t \rfloor}$. For an arbitrary joint probability vector $\boldsymbol{q}^*$, take a joint probability vector $\boldsymbol{q}$ such that $\boldsymbol{q} = \sum_{S \subseteq [k]} \alpha_S \boldsymbol{\eta}_S$ where α_S is nonnegative and satisfies $\sum_{S \subseteq [k]} \alpha_S = 1$, and $(\eta_S)_X$ is defined as $\frac{1}{2^{n-k}}$ if $X \cap [k] = S$ and 0 otherwise for each $X \subseteq V$.

Using the above $\{\alpha_S\}_{S \subseteq [k]}$, we can define a joint probability vector $\boldsymbol{r}$ as follows: for a real number M, let $(\eta_S^M)_X$ be $\frac{M^{|X \cap S|}}{2^{n-k}(M+1)^{|S|}}$ if $X \cap [k] \subseteq S$ and 0 otherwise for each $X \subseteq V$, and $\boldsymbol{r} = \sum_{S \subseteq [k]} \alpha_S \boldsymbol{\eta}_S^M$. For each $S \subseteq [k]$, let R^S be a diagonal matrix defined as M if $v \in S$, 0 if $v \in [k] \setminus S$, and 1 otherwise. The above R^S is an L-ensemble of a DPP corresponding to $\boldsymbol{\eta}_S^M$. Note that $\lim_{M \to \infty} \boldsymbol{r} = \boldsymbol{q}$ holds. We show the following lemma.

Lemma 1 *If $M \geq \frac{1}{2^{\frac{\epsilon}{k}} - 1}$, then $2^{-\epsilon} q_X \leq r_X$ holds for each $X \subseteq V$ with $q_X > 0$.*

Proof From the definitions of $\boldsymbol{q}$ and $\boldsymbol{r}$, for each $X \subseteq V$, $\max\{q_X - r_X, 0\}$ is equal to $\frac{1}{2^{n-k}}\left(1 - \left(\frac{M}{M+1}\right)^{|S|}\right)$ if $X \cap [k] = S$ and 0 otherwise holds. Then, for each $X \subseteq V$ with $q_X > 0$, $q_X - r_X \leq \frac{1}{2^{n-k}}\left(1 - \left(\frac{M}{M+1}\right)^k\right) \leq \frac{1}{2^{n-k}}\left(1 - \left(\frac{1}{2^{\epsilon/k}}\right)^k\right) = (1 - 2^{-\epsilon}) q_X$ holds since $q_X = 1/2^{n-k}$. Thus, $2^{-\epsilon} q_X \leq r_X$ holds. ■

Since $\log \frac{q_X^*}{r_X} \leq \log \frac{q_X^*}{2^{-\epsilon} q_X} = \log \frac{q_X^*}{q_X} + \log \frac{1}{2^{-\epsilon}}$ for $q_X^* > 0$ by Lemma 1, we have $D(\boldsymbol{q}^* || \boldsymbol{r}) \leq \sum_{X \subseteq V \colon q_X^* > 0} q_X^* \log \frac{q_X^*}{q_X} + \sum_{X \subseteq V \colon q_X^* > 0} q_X^* \log \frac{1}{2^{-\epsilon}} = D(\boldsymbol{q}^* || \boldsymbol{q}) + \epsilon$. By Corollary 3.3 of Montufar et al. (2014), $\min_{\boldsymbol{q}} D(\boldsymbol{q}^* || \boldsymbol{q}) \leq \log(2^n / k) = n - \log k = n - \lfloor \log t \rfloor$ holds. Thus, by taking the minimizer $\boldsymbol{q}$ of the above KL-divergence and defining $\boldsymbol{r}$ by the same $\{\alpha_S\}$ as $\boldsymbol{q}$, we obtain $\boldsymbol{r}$ for which the inequality in Theorem 3 is satisfied. (Note that if one wants to obtain an exactly-size-t CC-DPP, then such a CC-DPP can be constructed by adding arbitrary $t - k (= t - 2^{\lfloor \log t \rfloor})$ DPPs with coefficient 0.) ■

6. Numerical Simulation

Having established an upper bound on the KL-divergence (Theorem 3), we now analyze the expressive power of CC-DPPs *empirically*. One can naturally expect that fewer kernel matrices than those envisioned from theoretical analyses are sufficient to express *real-world* distributions. We verified this by numerically optimizing CC-DPPs given a real-world retail dataset. Our experimental results indicate that (1) the empirical KL-divergence is far smaller than the theoretical bound in Theorem 3, (2) a CC-DPP with a small t (e.g., $t = 3$) can suffice to achieve a moderately small KL-divergence, and (3) even a CC-DPP with $t = 2$ can outperform nonsymmetric DPPs.

Setup. We will explain how to optimize CC-DPPs and nonsymmetric DPPs, as a reasonable competitor, which include signed DPPs as well as DPPs. Given a log of m observations $Y_1, \ldots, Y_m \subseteq V$ for a ground set V, we first construct a joint probability vector $\boldsymbol{q} \in \mathbb{R}^{2^V}$ such that q_A for $A \subseteq V$ is the fraction of observing A, i.e., $q_A = \frac{|\{i \in [m] | Y_i = A\}|}{m}$. The objective is then to minimize the KL-divergence of a CC-DPP $(L^1, \ldots, L^t; \boldsymbol{\lambda})$ on V from $\boldsymbol{q}$; in other words, we would like to solve the following optimization problem:

$$\max_{L^i, \ldots, L^t; \boldsymbol{\lambda}} \sum_{A \subseteq V} q_A \cdot \log \left(\frac{q_A}{\sum_{i \in [t]} \lambda_i \cdot \frac{\det(L_A^i)}{\det(L_A^i + I)}} \right). \tag{1}$$

Even though a CC-DPP can have positive probabilities for exponentially many subsets, to evaluate the objective function in Eq. (1), we do not need to run through every subset $A \subseteq V$, but only need to run A such that $q_A > 0$.

We now explain how to optimize a CC-DPP. Define $\boldsymbol{\lambda}' \in \mathbb{R}^t$ with $\lambda_i = \lambda_i'^2$ for all $i \in [t]$ and $\|\boldsymbol{\lambda}'\|^2 = 1$. Given that (i) L^i is positive semidefinite for all $i \in [t]$ and that (ii) $\boldsymbol{\lambda}'$ is a point on the unit sphere, the space spanned by $L^1, \ldots, L^t$ and $\boldsymbol{\lambda}'$ forms a manifold. We can thus use optimization techniques on manifolds (Absil et al., 2009) to directly optimize Eq. (1). Note that the number of trainable parameters of a CC-DPP of size t is equal to $t\frac{n(n+1)}{2} + t$, where n is the size of V.

We next explain how to optimize nonsymmetric DPPs. It is known that a nonsymmetric DPP defines a probability distribution if and only if its joint probability kernel is a P_0-matrix [3] (Gartrell et al., 2019). To apply manifold optimization techniques, we represent a P_0-matrix $L \in \mathbb{R}^{n \times n}$ by $L = S + A$, where $S \in \mathbb{R}^{n \times n}$ is positive semidefinite matrix, and $A = BC^\top - CB^\top \in \mathbb{R}^{n \times n}$ is skew-symmetric with arbitrary matrices $B, C \in \mathbb{R}^{n \times n}$, according to Gartrell et al. (2019). Note that the number of trainable parameters of nonsymmetric DPPs is n^2.

We used Pymanopt (Townsend et al., 2016), an off-the-shelf solver for manifold optimization, and executed a steepest descent algorithm with back-tracking linear-search to minimize Eq. (1) for both CC-DPPs and a nonsymmetric DPP. Optimization terminates when the gradient norm is less than 10^{-6}. For each setting, we ran Pymanopt 20 times and calculated the average and the standard deviation of the KL-divergence. We conducted experiments on a Linux server with an Intel Xeon E5-2699 2.30GHz CPU and 792GB memory.

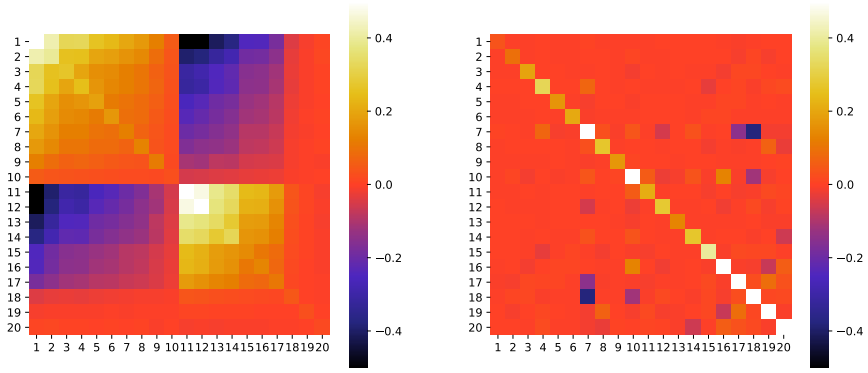
Dataset. We used the Online Retail Data Set (Chen et al., 2012)[4], which has been tested in the literature of learning DPPs (Warlop, 2018; Gartrell et al., 2019; Mariet et al., 2019; Warlop et al., 2019), to construct the joint probability vector. This is a public dataset that contains 25,900 observations of subsets from 4,070 unique stocks, consisting of transactions between December 2010 and December 2011 from an online retail based in the UK. We excluded every item but the 20 most frequently occurring items so that the resulting dataset would contain $m = 12{,}375$ observations, each of which is a subset of $n = 20$ items. Each observation was randomly assigned to either the training set or test set in the ratio of 3 to 1, resulting in the observations being split into the training set of 9,316 subsets and test set of 3,059 subsets used to construct the training and test probability vectors $\boldsymbol{q}^{\mathrm{tr}}$ and $\boldsymbol{q}^{\mathrm{te}}$, respectively. More specifically, we use $\boldsymbol{q}^{\mathrm{tr}}$ to optimize the parameters of CC-DPPs and nonsymmetric DPPs and evaluate the resulting model by calculating the KL-divergence from $\boldsymbol{q}^{\mathrm{te}}$.

Results. Table 2 shows the average and standard deviation of the KL-divergence over 20 runs for each of the DPP, the CC-DPPs of sizes $t = 2, \ldots, 5$, and the nonsymmetric DPP. The $\boldsymbol{\lambda}$'s are calculated as the average over 20 runs and are in decreasing order.[5] The number of trainable parameters for each model is also shown in the table. We first compared the empirical KL-divergence to the theoretical bound. Theorem 3 states that any probability distribution can be expressed by a size-3 CC-DPP of KL-divergence $\approx 20 - \lfloor \log 3 \rfloor = 19$.

3. A matrix $L \in \mathbb{R}^{n \times n}$ is called a P_0-matrix if its all principal minors are nonnegative.
4. Available from https://archive.ics.uci.edu/ml/datasets/online+retail
5. We confirmed that the standard deviation for each λ_i is less than 0.012.

Table 2: Experimental results for the DPP, CC-DPPs of size $t = 2, \ldots, 5$, and nonsymmetric DPP on the Online Retail Data Set. $\boldsymbol{q}^{\mathrm{tr}}$ and $\boldsymbol{q}^{\mathrm{te}}$ denote the training and test probability vectors constructed from 9,316 and 3,059 observations of subsets, respectively (see also Dataset paragraph). $D(\boldsymbol{q}^{\mathrm{tr}}||\cdot)$ and $D(\boldsymbol{q}^{\mathrm{te}}||\cdot)$ denote the KL-divergence (average $\pm$ standard deviation) of each model from $\boldsymbol{q}^{\mathrm{tr}}$ and $\boldsymbol{q}^{\mathrm{te}}$, respectively. Each $(\lambda_i)_{i\in[t]}$ is calculated as the average over 20 trials and is in decreasing order.

model	**#params.**	$D(\boldsymbol{q}^{\mathrm{tr}}\|\|\cdot)$	$D(\boldsymbol{q}^{\mathrm{te}}\|\|\cdot)$	λ_1	λ_2	λ_3	λ_4	λ_5
DPP	210	2.123 ± 0.007	2.502 ± 0.006	–				
Nonsym.	400	1.624 ± 0.020	2.017 ± 0.017	–				
$t = 2$	421	1.571 ± 0.015	1.993 ± 0.015	0.781	0.219			
$t = 3$	632	1.462 ± 0.029	1.876 ± 0.032	0.750	0.226	0.024		
$t = 4$	843	1.459 ± 0.009	$\mathbf{1.872 \pm 0.009}$	0.748	0.225	0.026	0.000	
$t = 5$	1,054	1.460 ± 0.007	$\mathbf{1.872 \pm 0.007}$	0.747	0.227	0.026	0.000	0.000



(a) Matrix L^1 (b) Matrix L^2

Figure 1: Heatmaps of two kernel matrices L^1, L^2 of the obtained size-2 CC-DPP (rows and columns are ordered manually).

On the other hand, Table 2 indicates that the empirical KL-divergence is 1.876, which is ten times smaller than the theoretical bound. This is because Theorem 3 is based on the worst-case analysis.

When comparing among DPPs, nonsymmetric DPPs and CC-DPPs, the size-4 CC-DPP achieved 7% lower KL-divergence than the nonsymmetric DPP, while the nonsymmetric DPP achieved a 19% lower KL-divergence than that the DPP achieved. Overall, we have the following trend regarding the KL-divergence on test data: (size 3) $\approx$ (size 4) $\approx$ (size 5) $<$ (size 2) $<$ (Nonsym.) $<$ (DPP). Since λ_4 and λ_5 have a weight of less than 10^{-100}, a convex combination of three DPPs seems to be sufficient to express $\boldsymbol{q}$ for this dataset.

We finally analyzed the kernel matrix of CC-DPPs through visualization. Figure 1 shows heatmaps of two kernel matrices L^1 and L^2 of the size-2 CC-DPP, the rows and columns of which are ordered manually. The two matrices play a different role in constructing

probability distributions. Clearly, L^1 has two clusters, which means that two or more items are unlikely to be chosen from the same cluster at the same time. On the other hand, L^2 has small off-diagonal and large diagonal elements, and no cluster structures are observed. Such structural properties indicate that the DPP *solely* defined by L^2 is close to independent distributions (i.e., Poisson point processes). In this way, CC-DPPs can have the ability to *decompose* a given distribution into easy-to-interpret DPPs.

7. Compactly Representable Class by a CC-DPP

We introduce a basic class of point processes; pseudo-Boolean functions, captured by convex combinations of a polynomial number of DPPs, that cannot be represented by a single DPP. This class appears in Iyer and Bilmes (2015) as the subclass of submodular point processes.

In the following, we introduce set functions $f\colon 2^V \to \mathbb{R}_{\geq 0}$ and examine a point process such that $\mathcal{P}(\boldsymbol{Y} = S) \propto f(S)$ for all $S \subseteq V$. We demonstrate the number of kernels required for expressing classes below is bounded by a polynomial in n. We can thus perform inference tasks in polynomial time. Let $V = \{1, \dots, n\}$. A function $f\colon \{0,1\}^V \to \mathbb{R}$ is called a *pseudo-Boolean function*, which has a unique polynomial forms $f(\boldsymbol{x}) = \sum_{T \subseteq V} a_T \prod_{i \in T} x_i$ for a real vector $\boldsymbol{a} = (a_T)_{T \subseteq V} \in \mathbb{R}^{2^V}$. Pseudo-Boolean functions naturally include graph cuts, and optimization on them has a diverse range of applications (see, e.g., the survey in Boros and Hammer (2002)).

Let us consider a point process on V such that $\mathcal{P}(\boldsymbol{Y} = S) \propto f(\boldsymbol{x}(S))$ for all $S \subseteq V$, where $\boldsymbol{x}(S) \in \{0,1\}^V$ is an indicator vector of S. Assume here that $\boldsymbol{a} \geq \boldsymbol{0}$ in order to ensure $\mathcal{P}$ is valid. Let $\mathrm{supp}(\boldsymbol{a}) = \{T \subseteq V \mid a_T > 0\}$. Now, we show that f can be written by a CC-DPP of size $|\mathrm{supp}(\boldsymbol{a})|$. Observe that $\mathcal{P}(S \subseteq \boldsymbol{Y})$ is $\frac{\sum_{T \subseteq V} a_T \cdot 2^{|V \setminus (S \cup T)|}}{\sum_{T \subseteq Y} a_T \cdot 2^{|V \setminus T|}} = \sum_{T \subseteq V} \lambda_T \cdot 2^{-|S \setminus T|}$, where $\lambda_T = \frac{a_T \cdot 2^{-|T|}}{\sum_{T' \subseteq V} a_{T'} \cdot 2^{-|T'|}}$ for all $T \subseteq V$. Hence, we construct for each $T \in \mathrm{supp}(\boldsymbol{a})$, a marginal kernel K^T such that $K^T_{ii} = 1$ for all $i \in T$, all the other diagonal elements are $1/2$, and all the other elements are 0. The resulting CC-DPP $(\{K^T\}_{T \in \mathrm{supp}(\boldsymbol{a})}; \boldsymbol{\lambda})$ of size $|\mathrm{supp}(\boldsymbol{a})|$ matches $\mathcal{P}$. Because $|\mathrm{supp}(\boldsymbol{a})| \leq \sum_{k=0}^{d} \binom{n}{k} = \mathrm{O}(n^d)$, where d is the degree of f (which is defined the maximum size of $T \subseteq V$ such that $a_T \neq 0$), point processes induced by pseudo-Boolean functions of *bounded degree* can be expressed by a polynomial-size CC-DPP.

Weighted Coverage Functions of Bounded Occurrence. As an application of a pseudo-Boolean function representation, we show that *weighted coverage functions* have a polynomial-size CC-DPP representation if the number of occurrences of each element is bounded. For a universal set U, let $\mathcal{F} = \{S_1, \dots, S_n\}$ be a family of subsets of U, $\boldsymbol{c} \in \mathbb{R}^U_{\geq 0}$ be a nonnegative real vector, and $V = \{1, \dots, n\}$. Consider a weighted coverage function $f : 2^V \to \mathbb{R}_{\geq 0}$ defined as $f(A) = \boldsymbol{c}(\bigcup_{i \in A} S_i)$ for subset $A \subseteq V$. Then, we denote the set of indices i for which S_i includes an element $u \in U$ by $I(u) = \{i \in V \mid u \in S_i\}$, and we define a polynomial g as $g(\boldsymbol{x}) = \sum_{u \in U} c_u \Big(1 - \prod_{i \in I(u)} (1 - x_i)\Big)$. Observe that $g(\boldsymbol{x}(A)) = f(A)$ for all $A \subseteq V$, and the degree of g is equivalent to the maximum number of *occurrences* in $\mathcal{F}$ of each element u of U, i.e., $\max_{u \in U} |I(u)|$. Hence, we can express weighted coverage functions using a polynomial number of DPPs whenever the number of occurrences is bounded by a constant. Almost the same argument holds for probabilistic coverage functions, see, e.g., Iyer and Bilmes (2015).

Undirected Cut. Given an undirected graph $G = (V, E)$, the *cut function* $c\colon 2^V \to \mathbb{Z}_{\geq 0}$ is defined as the number of edges whose one endpoint is in S and the other one is not in S for subset $S \subseteq V$, i.e., $c(S) = |\{\{u, v\} \in E \mid (u \in S) \wedge (v \notin S)\}|$. Then, the point process derived by c is $\mathcal{P}(\boldsymbol{Y} = S) \propto c(S)$, or equivalently, $\mathcal{P}(S \subseteq \boldsymbol{Y}) = \sum_{\{u,v\} \in E} 2^{-|S|} m^{-1} \cdot [(u \notin S) \vee (v \notin S)]$, where $[\psi]$ is 1 if ψ is true and 0 otherwise.

Here, we give an example showing strong representability of size-3 CC-DPP model compared with some (generalized) real DPP models. The joint probability vector $\boldsymbol{q} = (0, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0)$ can be represented by a size-3 CC-DPP with

$$K_1 = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 \end{pmatrix}, \; K_2 = \begin{pmatrix} 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \end{pmatrix}, \; K_3 = \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix},$$

and $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$.

Claim 2 *The point process whose joint probability is proportional to the cut function in the undirected 3-clique (triangle graph) cannot be expressed by DPPs, signed DPPs, or nonsymmetric DPPs.*

Proof Showing the claim for nonsymmetric DPPs, of which kernel is not necessarily symmetric, is sufficient. Let $G = (V, E)$ be the 3-clique. Assume the existence of a marginal kernel K that yields the desired point process. Recall that K is a real matrix (but not necessarily symmetric). Since the cut value of $S \subseteq V$ is 0 if S is $\emptyset$ or V and 2 otherwise, the joint probability vector is $\boldsymbol{q} = (0, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0)$, and the marginal probability vector is then $\boldsymbol{p} = (1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0)$. Hence, it must hold that $K_{11} = K_{22} = K_{33} = \frac{1}{2}$, and the equality $K_{11}K_{22} - K_{12}K_{21} = \det(K_{12}) = \frac{1}{6}$ derives that $K_{12}K_{21} = \frac{1}{12}$, and similarly we have that $K_{13}K_{31} = K_{23}K_{32} = \frac{1}{12}$. However, we eventually find that $\det(K_{123}) = K_{12}K_{23}K_{31} + K_{21}K_{32}K_{13}$, which is equal to $(K_{12}K_{23}K_{31})^{-1}((K_{12}K_{23}K_{31})^2 + \frac{1}{12^3}) \neq 0$, a contradiction to that $\det(K_{123}) = p_{123} = 0$. ■

8. Conclusion

We presented a systematic study on convex combinations of determinantal point processes. Our contributions were fivefold: (1) extend some properties and inference algorithms on DPPs to CC-DPPs, (2) provide lower and upper bounds on the number of kernels, (3) show an approximation upper bound, (4) verify the superiority of the convex combination through numerical simulations, and (5) introduce a class of distributions that can be expressed by a polynomial number of kernels.

References

P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Nima Anari and Shayan Oveis Gharan. A generalization of permanent inequalities and applications in counting and optimization. In *STOC*, pages 384–396, 2017.

Alexei Borodin and Eric M. Rains. Eynard-Mehta theorem, Schur process, and their Pfaffian analogs. *J. Stat. Phys.*, 121(3–4):291–317, 2005.

Endre Boros and Peter L. Hammer. Pseudo-Boolean optimization. *Discrete Appl. Math.*, 123(1–3):155–225, 2002.

Victor-Emmanuel Brunel. Learning signed determinantal point processes through the principal minor assignment problem. In *NeurIPS*, pages 7365–7374, 2018.

L. Elisa Celis, Amit Deshpande, Tarun Kathuria, Damian Straszak, and Nisheeth K. Vishnoi. On the complexity of constrained determinantal point processes. In *APPROX/RANDOM*, pages 36:1–36:22, 2017.

L. Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth K. Vishnoi. Fair and diverse DPP-based data summarization. In *ICML*, pages 715–724, 2018.

Daqing Chen, Sai Laing Sain, and Kun Guo. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, 19(3):197–208, 2012.

Gregory F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artif. Intell.*, 42(2-3):393–405, 1990.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley & Sons, 2012.

Paul Dagum and Michael Luby. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artif. Intell.*, 60(1):141–153, 1993.

Michal Dereziński, Daniele Calandriello, and Michal Valko. Exact sampling of determinantal point processes with sublinear time preprocessing. In *NeurIPS*, pages 11542–11554, 2019.

Josip Djolonga and Andreas Krause. From MAP to marginals: Variational inference in Bayesian submodular models. In *NIPS*, pages 244–252, 2014.

Mike Gartrell, Victor-Emmanuel Brunel, Elvis Dohmatob, and Syrine Krichene. Learning nonsymmetric determinantal point processes. In *NeurIPS*, pages 6718–6728, 2019.

Alkis Gotovos, S. Hamed Hassani, and Andreas Krause. Sampling from probabilistic submodular models. In *NIPS*, pages 1945–1953, 2015.

Leonid Gurvits. On the complexity of mixed discriminants and related problems. In *MFCS*, pages 447–458, 2005.

J. Ben Hough, Manjunath Krishnapur, Yuval Peres, and Bálint Virág. Determinantal processes and independence. *Probab. Surv.*, 3:206–229, 2006.

Rishabh Iyer and Jeffrey Bilmes. Submodular point processes with applications to machine learning. In *AISTATS*, pages 388–397, 2015.

Alex Kulesza and Ben Taskar. k-DPPs: Fixed-size determinantal point processes. In *ICML*, pages 1193–1200, 2011a.

Alex Kulesza and Ben Taskar. Learning determinantal point processes. In *UAI*, pages 419–427, 2011b.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Found. Trends Mach. Learn.*, 5(2–3):123–286, 2012.

Odile Macchi. The coincidence approach to stochastic point processes. *Adv. Appl. Probab.*, 7(1):83–122, 1975.

Zelda Mariet, Mike Gartrell, and Suvrit Sra. Learning determinantal point processes by corrective negative sampling. In *AISTATS*, pages 2251–2260, 2019.

Zelda E. Mariet, Suvrit Sra, and Stefanie Jegelka. Exponentiated strongly Rayleigh distributions. In *NeurIPS*, pages 4464–4474, 2018.

Guido Montufar, Johannes Rauh, and Nihat Ay. Expressive power and approximation errors of restricted Boltzmann machines. *arXiv preprint arXiv:1406.3140*, 2014. Preliminary version: Guido Montufar, Johannes Rauh, and Nihat Ay. Expressive power and approximation errors of restricted Boltzmann machines. In *NIPS*, pages 415–423, 2011.

Naoto Ohsaka and Tatsuya Matsuoka. On the (in)tractability of computing normalizing constants for the product of determinantal point processes. In *ICML*, pages 7414–7423, 2020.

James Townsend, Niklas Koep, and Sebastian Weichwald. Pymanopt: A Python toolbox for optimization on manifolds using automatic differentiation. *J. Mach. Learn. Res.*, 17(137): 1–5, 2016.

Romain Warlop. *Novel Learning and Exploration-Exploitation Methods for Effective Recommender Systems*. PhD thesis, 2018.

Romain Warlop, Jérémie Mary, and Mike Gartrell. Tensorized determinantal point processes for recommendation. In *KDD*, pages 1605–1615, 2019.

Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. Practical diversified recommendations on YouTube with determinantal point processes. In *CIKM*, pages 2165–2173, 2018.