Learning Determinantal Point Processes with Moments and Cycles

John Urschel [1] **Victor-Emmanuel Brunel** [1] **Ankur Moitra** [1] **Philippe Rigollet** [1]

Abstract

Determinantal Point Processes (DPPs) are a family of probabilistic models that have a repulsive behavior, and lend themselves naturally to many tasks in machine learning where returning a diverse set of objects is important. While there are fast algorithms for sampling, marginalization and conditioning, much less is known about learning the parameters of a DPP. Our contribution is twofold: (i) we establish the optimal sample complexity achievable in this problem and show that it is governed by a natural parameter, which we call the *cycle sparsity*; (ii) we propose a provably fast combinatorial algorithm that implements the method of moments efficiently and achieves optimal sample complexity. Finally, we give experimental results that confirm our theoretical findings.

1. Introduction

Determinantal Point Processes (DPPs) are a family of probabilistic models that arose from the study of quantum mechanics (Macchi, 1975) and random matrix theory (Dyson, 1962). Following the seminal work of Kulesza and Taskar (Kulesza & Taskar, 2012), discrete DPPs have found numerous applications in machine learning, including in document and timeline summarization (Lin & Bilmes, 2012; Yao et al., 2016), image search (Kulesza & Taskar, 2011; Affandi et al., 2014) and segmentation (Lee et al., 2016), audio signal processing (Xu & Ou, 2016), bioinformatics (Batmanghelich et al., 2014) and neuroscience (Snoek et al., 2013). What makes such models appealing is that they exhibit repulsive behavior and lend themselves naturally to tasks where returning a diverse set of objects is important.

One way to define a DPP is through an $N \times N$ symmetric positive semidefinite matrix K, called a *kernel*, whose eigenvalues are bounded in the range $[0, 1]$. Then the DPP associated with K, which we denote by $\mathsf{DPP}(K)$, is the distribution on $Y \subseteq [N] = \{1, \dots, N\}$ that satisfies, for any $J \subseteq [N]$,

$$\mathbb{P}[J \subseteq Y] = \det(K_J),$$

where K_J is the principal submatrix of K indexed by the set J. The *graph induced by* K is the graph $G = ([N], E)$ on the vertex set $[N]$ that connects $i, j \in [N]$ if and only if $K_{i,j} \neq 0$.

There are fast algorithms for sampling (or approximately sampling) from $\mathsf{DPP}(K)$ (Deshpande & Rademacher, 2010; Rebeschini & Karbasi, 2015; Li et al., 2016b;a). Marginalizing the distribution on a subset $I \subseteq [N]$ and conditioning on the event that $J \subseteq Y$ both result in new DPPs and closed form expressions for their kernels are known (Borodin & Rains, 2005).

There has been much less work on the problem of learning the parameters of a DPP. A variety of heuristics have been proposed, including Expectation-Maximization (Gillenwater et al., 2014), MCMC (Affandi et al., 2014), and fixed point algorithms (Mariet & Sra, 2015). All of these attempt to solve a nonconvex optimization problem, and no guarantees on their statistical performance are known. Recently, Brunel *et al.* (Brunel et al., 2017) studied the rate of estimation achieved by the maximum likelihood estimator, but the question of efficient computation remains open.

Apart from positive results on sampling, marginalization and conditioning, most provable results about DPPs are actually negative. It is conjectured that the maximum likelihood estimator is NP-hard to compute (Kulesza, 2012). Actually, approximating the mode of size k of a DPP to within a c^k factor is known to be NP-hard for some $c > 1$ (Çivril & Magdon-Ismail, 2009; Summa et al., 2015). The best known algorithms currently obtain a $e^k + o(k)$ approximation factor (Nikolov, 2015; Nikolov & Singh, 2016).

In this work, we bypass the difficulties associated with maximum likelihood estimation by using the *method of moments* to achieve optimal sample complexity. We introduce a parameter ℓ, which we call the *cycle sparsity* of the graph induced by the kernel K, which governs the number of moments that need to be considered and, thus, the sample complexity. Moreover, we use a refined version of Horton's al-

[1]Department of Mathematics, MIT, USA. Correspondence to: John Urschel <urschel@mit.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

gorithm (Horton, 1987; Amaldi et al., 2010) to implement the method of moments in polynomial time.

The cycle sparsity of a graph is the smallest integer ℓ so that the cycles of length at most ℓ yield a basis for the cycle space of the graph. Even though there are in general exponentially many cycles in a graph to consider, Horton's algorithm constructs a minimum weight cycle basis and, in doing so, also reveals the parameter ℓ together with a collection of at most ℓ induced cycles spanning the cycle space.

We use such cycles in order to construct our method of moments estimator. For any fixed $\ell \geq 2$, our overall algorithm has sample complexity

$$n = O\Big((\frac{C}{\alpha})^{2\ell} + \frac{\log N}{\alpha^2 \varepsilon^2}\Big)$$

for some constant $C > 1$ and runs in time polynomial in n and N, and learns the parameters up to an additive ε with high probability. The $(C/\alpha)^{2\ell}$ term corresponds to the number of samples needed to recover the signs of the entries in K. We complement this result with a minimax lower bound (Theorem 2) to show that this sample complexity is in fact near optimal. In particular, we show that there is an infinite family of graphs with cycle sparsity ℓ (namely length ℓ cycles) on which any algorithm requires at least $(C'\alpha)^{-2\ell}$ samples to recover the signs of the entries of K for some constant $C' > 1$. Finally, we show experimental results that confirm many quantitative aspects of our theoretical predictions. Together, our upper bounds, lower bounds, and experiments present a nuanced understanding of which DPPs can be learned provably and efficiently.

2. Estimation of the Kernel

2.1. Model and definitions

Let $Y_1, \ldots, Y_n$ be n independent copies of $Y \sim \mathsf{DPP}(K)$, for some unknown kernel K such that $0 \preceq K \preceq I_N$. It is well known that K is identified by $\mathsf{DPP}(K)$ only up to flips of the signs of its rows and columns: If K' is another symmetric matrix with $0 \preceq K' \preceq I_N$, then $\mathsf{DPP}(K')$=$\mathsf{DPP}(K)$ if and only if $K' = DKD$ for some $D \in \mathcal{D}_N$, where $\mathcal{D}_N$ denotes the class of all $N \times N$ diagonal matrices with only 1 and -1 on their diagonals (Kulesza, 2012, Theorem 4.1). We call such a transform a $\mathcal{D}_N$-similarity of K.

In view of this equivalence class, we define the following pseudo-distance between kernels K and K':

$$\rho(K, K') = \inf_{D \in \mathcal{D}_N} |DKD - K'|_\infty \,,$$

where for any matrix K, $|K|_\infty = \max_{i,j \in [N]} |K_{i,j}|$ denotes the entrywise sup-norm.

For any $S \subset [N]$, we write $\Delta_S = \det(K_S)$, where K_S denotes the $|S| \times |S|$ submatrix of K obtained by keeping rows and colums with indices in S. Note that for $1 \leq i \neq j \leq N$, we have the following relations:

$$K_{i,i} = \mathbb{P}[i \in Y], \qquad \Delta_{\{i,j\}} = \mathbb{P}[\{i,j\} \subseteq Y],$$

and $|K_{i,j}| = \sqrt{K_{i,i}K_{j,j} - \Delta_{\{i,j\}}}$. Therefore, the principal minors of size one and two of K determine K up to the sign of its off-diagonal entries. In fact, for any K, there exists an ℓ depending only on the graph G_K induced by K, such that K can be recovered up to a $\mathcal{D}_N$-similarity with only the knowledge of its principal minors of size at most ℓ. We will show that this ℓ is exactly the cycle sparsity.

2.2. DPPs and graphs

In this section, we review some of the interplay between graphs and DPPs that plays a key role in the definition of our estimator.

We begin by recalling some standard graph theoretic notions. Let $G = ([N], E)$, $|E| = m$. A cycle C of G is any connected subgraph in which each vertex has even degree. Each cycle C is associated with an incidence vector $x \in GF(2)^m$ such that $x_e = 1$ if e is an edge in C and $x_e = 0$ otherwise. The *cycle space* $\mathcal{C}$ of G is the subspace of $GF(2)^m$ spanned by the incidence vectors of the cycles in G. The dimension ν_G of the cycle space is called *cyclomatic number*, and it is well known that $\nu_G := m - N + \kappa(G)$, where $\kappa(G)$ denotes the number of connected components of G.

Recall that a *simple cycle* is a graph where every vertex has either degree two or zero and the set of vertices with degree two form a connected set. A *cycle basis* is a basis of $\mathcal{C} \subset GF(2)^m$ such that every element is a simple cycle. It is well known that every cycle space has a cycle basis of induced cycles.

Definition 1. *The* cycle sparsity *of a graph G is the minimal ℓ for which G admits a cycle basis of induced cycles of length at most ℓ, with the convention that $\ell = 2$ whenever the cycle space is empty. A corresponding cycle basis is called a* shortest maximal cycle basis.

A *shortest maximal cycle basis* of the cycle space was also studied for other reasons by (Chickering et al., 1995). We defer a discussion of computing such a basis to Section 4.

For any subset $S \subseteq [N]$, denote by $G_K(S) = (S, E(S))$ the subgraph of G_K induced by S. A matching of $G_K(S)$ is a subset $M \subseteq E(S)$ such that any two distinct edges in M are not adjacent in $G(S)$. The set of vertices incident to some edge in M is denoted by $V(M)$. We denote by $\mathcal{M}(S)$ the collection of all matchings of $G_K(S)$. Then, if $G_K(S)$ is an induced cycle, we can write the principal

minor $\Delta_S = \det(K_S)$ as follows:

$$\Delta_S = \sum_{M \in \mathcal{M}(S)} (-1)^{|M|} \prod_{\{i,j\} \in M} K_{i,j}^2 \prod_{i \notin V(M)} K_{i,i} + 2 \times (-1)^{|S|+1} \prod_{\{i,j\} \in E(S)} K_{i,j}. \quad (1)$$

Others have considered the relationship between the principal minors of K and recovery of $\mathsf{DPP}(K)$. There has been work regarding the *symmetric principal minor assignment problem*, namely the problem of computing a matrix given an oracle that gives any principal minor in constant time (Rising et al., 2015).

In our setting, we can approximate the principal minors of K by empirical averages. However the accuracy of our estimator deteriorates with the size of the principal minor, and we must therefore estimate the smallest possible principal minors in order to achieve optimal sample complexity. Here, we prove a new result, namely, that the smallest ℓ such that all the principal minors of K are uniquely determined by those of size at most ℓ is exactly the cycle sparsity of the graph induced by K.

Proposition 1. *Let $K \in \mathbb{R}^{N \times N}$ be a symmetric matrix, G_K be the graph induced by K, and $\ell \geq 3$ be some integer. The kernel K is completely determined up to $\mathcal{D}_N$-similarity by its principal minors of size at most ℓ if and only if the cycle sparsity of G_K is at most ℓ.*

Proof. Note first that all the principal minors of K completely determine K up to a $\mathcal{D}_N$-similarity (Rising et al., 2015, Theorem 3.14). Moreover, recall that principal minors of degree at most 2 determine the diagonal entries of K as well as the magnitude of its off-diagonal entries. In particular, given these principal minors, one only needs to recover the signs of the off-diagonal entries of K. Let the *sign* of cycle C in K be the product of the signs of the entries of K corresponding to the edges of C.

Suppose G_K has cycle sparsity ℓ and let $(C_1, \dots, C_\nu)$ be a cycle basis of G_K where each $C_i, i \in [\nu]$ is an induced cycle of length at most ℓ. By (1), the sign of any $C_i, i \in [\nu]$ is completely determined by the principal minor Δ_S, where S is the set of vertices of C_i and is such that $|S| \leq \ell$. Moreover, for $i \in [\nu]$, let $x_i \in GF(2)^m$ denote the incidence vector of C_i. By definition, the incidence vector x of any cycle C is given by $\sum_{i \in \mathcal{I}} x_i$ for some subset $\mathcal{I} \subset [\nu]$. The sign of C is then given by the product of the signs of $C_i, i \in \mathcal{I}$ and thus by corresponding principal minors. In particular, the signs of all cycles are determined by the principal minors Δ_S with $|S| \leq \ell$. In turn, by Theorem 3.12 in (Rising et al., 2015), the signs of all cycles completely determine K, up to a $\mathcal{D}_N$-similarity.

Next, suppose the cycle sparsity of G_K is at least $\ell + 1$, and let $\mathcal{C}_\ell$ be the subspace of $GF(2)^m$ spanned by the induced cycles of length at most ℓ in G_K. Let $x_1, \dots, x_\nu$ be a basis of $\mathcal{C}_\ell$ made of the incidence column vectors of induced cycles of length at most ℓ in G_K and form the matrix $A \in GF(2)^{m \times \nu}$ by concatenating the x_i's. Since $\mathcal{C}_\ell$ does not span the cycle space of G_K, $\nu < \nu_{G_K} \leq m$. Hence, the rank of A is less than m, so the null space of $A^\top$ is non trivial. Let $\bar{x}$ be the incidence column vector of an induced cycle $\bar{C}$ that is not in $\mathcal{C}_\ell$, and let $h \in GL(2)^m$ with $A^\top h = 0$, $h \neq 0$ and $\bar{x}^\top h = 1$. These three conditions are compatible because $\bar{C} \notin \mathcal{C}_\ell$. We are now in a position to define an alternate kernel K' as follows: Let $K'_{i,i} = K_{i,i}$ and $|K'_{i,j}| = |K_{i,j}|$ for all $i, j \in [N]$. We define the signs of the off-diagonal entries of K' as follows: For all edges $e = \{i, j\}, i \neq j$, $\operatorname{sgn}(K'_e) = \operatorname{sgn}(K_e)$ if $h_e = 0$ and $\operatorname{sgn}(K'_e) = -\operatorname{sgn}(K_e)$ otherwise. We now check that K and K' have the same principal minors of size at most ℓ but differ on a principal minor of size larger than ℓ. To that end, let x be the incidence vector of a cycle C in $\mathcal{C}_\ell$ so that $x = Aw$ for some $w \in GL(2)^\nu$. Thus the sign of C in K is given by

$$\prod_{e\,:\,x_e=1} K_e = (-1)^{x^\top h} \prod_{e\,:\,x_e=1} K'_e = (-1)^{w^\top A^\top h} \prod_{e\,:\,x_e=1} K'_e = \prod_{e\,:\,x_e=1} K'_e$$

because $A^\top h = 0$. Therefore, the sign of any $C \in \mathcal{C}_\ell$ is the same in K and K'. Now, let $S \subseteq [N]$ with $|S| \leq \ell$, and let $G = G_{K_S} = G_{K'_S}$ be the graph corresponding to K_S (or, equivalently, to K'_S). For any induced cycle C in G, C is also an induced cycle in G_K and its length is at most ℓ. Hence, $C \in \mathcal{C}_\ell$ and the sign of C is the same in K and K'. By (Rising et al., 2015, Theorem 3.12), $\det(K_S) = \det(K'_S)$. Next observe that the sign of $\bar{C}$ in K is given by

$$\prod_{e\,:\,\bar{x}_e=1} K_e = (-1)^{\bar{x}^\top h} \prod_{e\,:\,\bar{x}_e=1} K'_e = -\prod_{e\,:\,x_e=1} K'_e.$$

Note also that since $\bar{C}$ is an induced cycle of $G_K = G_{K'}$, the above quantity is nonzero. Let $\bar{S}$ be the set of vertices in $\bar{C}$. By (1) and the above display, we have $\det(K_{\bar{S}}) \neq \det(K'_{\bar{S}})$. Together with (Rising et al., 2015, Theorem 3.14), it yields $K \neq DK'D$ for all $D \in \mathcal{D}_N$. □

2.3. Definition of the Estimator

Our procedure is based on the previous result and can be summarized as follows. We first estimate the diagonal entries (i.e., the principal minors of size one) of K by the method of moments. By the same method, we estimate the principal minors of size two of K, and we deduce estimates of the magnitude of the off-diagonal entries. To use these estimates to deduce an estimate $\hat{G}$ of G_K, we make the following assumption on the kernel K.

Assumption 1. *Fix $\alpha \in (0,1)$. For all $1 \leq i < j \leq N$, either $K_{i,j} = 0$, or $|K_{i,j}| \geq \alpha$.*

Finally, we find a shortest maximal cycle basis of $\hat{G}$, and we set the signs of our non-zero off-diagonal entry estimates by using estimators of the principal minors induced by the elements of the basis, again obtained by the method of moments.

For $S \subseteq [N]$, set $\hat{\Delta}_S = \frac{1}{n}\sum_{p=1}^{n} \mathbb{1}_{S \subseteq Y_p}$, and define

$$\hat{K}_{i,i} = \hat{\Delta}_{\{i\}} \quad \text{and} \quad \hat{B}_{i,j} = \hat{K}_{i,i}\hat{K}_{j,j} - \hat{\Delta}_{\{i,j\}},$$

where $\hat{K}_{i,i}$ and $\hat{B}_{i,j}$ are our estimators of $K_{i,i}$ and $K_{i,j}^2$, respectively.

Define $\hat{G} = ([N], \hat{E})$, where, for $i \neq j$, $\{i,j\} \in \hat{E}$ if and only if $\hat{B}_{i,j} \geq \frac{1}{2}\alpha^2$. The graph $\hat{G}$ is our estimator of G_K. Let $\{\hat{C}_1, ..., \hat{C}_{\nu_{\hat{G}}}\}$ be a shortest maximal cycle basis of the cycle space of $\hat{G}$. Let $\hat{S}_i \subseteq [N]$ be the subset of vertices of $\hat{C}_i$, for $1 \leq i \leq \nu_{\hat{G}}$. We define

$$\hat{H}_i = \hat{\Delta}_{\hat{S}_i} - \sum_{M \in \mathcal{M}(\hat{S}_i)} (-1)^{|M|} \prod_{\{i,j\} \in M} \hat{B}_{i,j} \prod_{i \notin V(M)} \hat{K}_{i,i},$$

for $1 \leq i \leq \nu_{\hat{G}}$. In light of (1), for large enough n, this quantity should be close to

$$H_i = 2 \times (-1)^{|\hat{S}_i|+1} \prod_{\{i,j\} \in E(\hat{S}_i)} K_{i,j}\,.$$

We note that this definition is only symbolic in nature, and computing $\hat{H}_i$ in this fashion is extremely inefficient. Instead, to compute it in practice, we will use the determinant of an auxiliary matrix, computed via a matrix factorization. Namely, let us define the matrix $\widetilde{K} \in \mathbb{R}^{N \times N}$ such that $\widetilde{K}_{i,i} = \hat{K}_{i,i}$ for $1 \leq i \leq N$, and $\widetilde{K}_{i,j} = \hat{B}_{i,j}^{1/2}$. We have

$$\begin{aligned}\det \widetilde{K}_{\hat{S}_i} = &\sum_{M \in \mathcal{M}} (-1)^{|M|} \prod_{\{i,j\} \in M} \hat{B}_{i,j} \prod_{i \notin V(M)} \hat{K}_{i,i} \\ &+ 2 \times (-1)^{|\hat{S}_i|+1} \prod_{\{i,j\} \in \hat{E}(\hat{S}_i)} \hat{B}_{i,j}^{1/2},\end{aligned}$$

so that we may equivalently write

$$\hat{H}_i = \hat{\Delta}_{\hat{S}_i} - \det(\widetilde{K}_{\hat{S}_i}) + 2 \times (-1)^{|\hat{S}_i|+1} \prod_{\{i,j\} \in \hat{E}(\hat{S}_i)} \hat{B}_{i,j}^{1/2}.$$

Finally, let $\hat{m} = |\hat{E}|$. Set the matrix $A \in GF(2)^{\nu_{\hat{G}} \times \hat{m}}$ with i-th row representing $\hat{C}_i$ in $GF(2)^m$, $1 \leq i \leq \nu_{\hat{G}}$, $b = (b_1, \ldots, b_{\nu_{\hat{G}}}) \in GF(2)^{\nu_{\hat{G}}}$ with $b_i = \frac{1}{2}[\mathrm{sgn}(\hat{H}_i) + 1]$, $1 \leq i \leq \nu_{\hat{G}}$, and let $x \in GF(2)^m$ be a solution to the linear system $Ax = b$ if a solution exists, $x = \mathbb{1}_m$ otherwise.

We define $\hat{K}_{i,j} = 0$ if $\{i,j\} \notin \hat{E}$ and $\hat{K}_{i,j} = \hat{K}_{j,i} = (2x_{\{i,j\}} - 1)\hat{B}_{i,j}^{1/2}$ for all $\{i,j\} \in \hat{E}$.

2.4. Geometry

The main result of this subsection is the following lemma which relates the quality of estimation of K in terms of ρ to the quality of estimation of the principal minors Δ_S.

Lemma 1. *Let K satisfy Assumption 1, and let ℓ be the cycle sparsity of G_K. Let $\varepsilon > 0$. If $|\hat{\Delta}_S - \Delta_S| \leq \varepsilon$ for all $S \subseteq [N]$ with $|S| \leq 2$ and if $|\hat{\Delta}_S - \Delta_S| \leq (\alpha/4)^{|S|}$ for all $S \subseteq [N]$ with $3 \leq |S| \leq \ell$, then*

$$\rho(\hat{K}, K) < 4\varepsilon/\alpha\,.$$

Proof. We can bound $|\hat{B}_{i,j} - K_{i,j}^2|$, namely,

$$\begin{aligned}\hat{B}_{i,j} &\leq (K_{i,i} + \alpha^2/16)(K_{j,j} + \alpha^2/16) - (\Delta_{\{i,j\}} - \alpha^2/16) \\ &\leq K_{i,j}^2 + \alpha^2/4\end{aligned}$$

and

$$\begin{aligned}\hat{B}_{i,j} &\geq (K_{i,i} - \alpha^2/16)(K_{j,j} - \alpha^2/16) - (\Delta_{\{i,j\}} + \alpha^2/16) \\ &\geq K_{i,j}^2 - 3\alpha^2/16,\end{aligned}$$

giving $|\hat{B}_{i,j} - K_{i,j}^2| < \alpha^2/4$. Thus, we can correctly determine whether $K_{i,j} = 0$ or $|K_{i,j}| \geq \alpha$, yielding $\hat{G} = G_K$. In particular, the cycle basis $\hat{C}_1, \ldots, \hat{C}_{\nu_{\hat{G}}}$ of $\hat{G}$ is a cycle basis of G_K. Let $1 \leq i \leq \nu_{\hat{G}}$. Denote by $t = (\alpha/4)^{|S_i|}$. We have

$$\begin{aligned}&\left|\hat{H}_i - H_i\right| \\ &\leq |\hat{\Delta}_{\hat{S}_i} - \Delta_{\hat{S}_i}| + |\mathcal{M}(\hat{S}_i)| \max_{x \in \pm 1} \left[(1+4tx)^{|\hat{S}_i|} - 1\right] \\ &\leq (\alpha/4)^{|\hat{S}_i|} + |\mathcal{M}(\hat{S}_i)| \left[(1+4t)^{|\hat{S}_i|} - 1\right] \\ &\leq (\alpha/4)^{|\hat{S}_i|} + T\left(|\hat{S}_i|, \left\lfloor \frac{|\hat{S}_i|}{2} \right\rfloor\right) 4t\; T(|\hat{S}_i|, |\hat{S}_i|) \\ &\leq (\alpha/4)^{|\hat{S}_i|} + 4t\; (2^{\frac{|\hat{S}_i|}{2}} - 1)(2^{|\hat{S}_i|} - 1) \\ &\leq (\alpha/4)^{|\hat{S}_i|} + t2^{2|\hat{S}_i|} \\ &< 2\alpha^{|\hat{S}_i|} \leq |H_i|,\end{aligned}$$

where, for positive integers $p < q$, we denote by $T(q,p) = \sum_{i=1}^{p} \binom{q}{i}$. Therefore, we can determine the sign of the product $\prod_{\{i,j\} \in E(\hat{S}_i)} K_{i,j}$ for every element in the cycle basis and recover the signs of the non-zero off-diagonal entries of $K_{i,j}$. Hence, $\rho(\hat{K}, K) = \max_{1 \leq i,j \leq N} \left||\hat{K}_{i,j}| - |K_{i,j}|\right|$. For $i = j$, $\left||\hat{K}_{i,j}| - |K_{i,j}|\right| = |\hat{K}_{i,i} - K_{i,i}| \leq \varepsilon$. For $i \neq j$ with $\{i,j\} \in \hat{E} = E$, one can easily show that $\left|\hat{B}_{i,j} - K_{i,j}^2\right| \leq 4\varepsilon$, yielding

$$|\hat{B}_{i,j}^{1/2} - |K_{i,j}|| \leq \frac{4\varepsilon}{\left|\hat{B}_{i,j}^{1/2} + |K_{i,j}|\right|} \leq \frac{4\varepsilon}{\alpha},$$

which completes the proof. □

We are now in a position to establish a sufficient sample size to estimate K within distance ε.

Theorem 1. *Let K satisfy Assumption 1, and let ℓ be the cycle sparsity of G_K. Let $\varepsilon > 0$. For any $A > 0$, there exists $C > 0$ such that*

$$n \geq C\Big(\frac{1}{\alpha^2\varepsilon^2} + \ell\big(\frac{4}{\alpha}\big)^{2\ell}\Big)\log N\,,$$

yields $\rho(\hat{K}, K) \leq \varepsilon$ with probability at least $1 - N^{-A}$.

Proof. Using the previous lemma, and applying a union bound,

$$\begin{aligned}\mathbb{P}\Big[\rho(\hat{K}, K) > \varepsilon\Big] &\leq \sum_{|S|\leq 2}\mathbb{P}\Big[|\hat{\Delta}_S - \Delta_S| > \alpha\varepsilon/4\Big] \\ &\quad + \sum_{2\leq|S|\leq\ell}\mathbb{P}\Big[|\hat{\Delta}_S - \Delta_S| > (\alpha/4)^{|S|}\Big] \\ &\leq 2N^2 e^{-n\alpha^2\varepsilon^2/8} + 2N^{\ell+1}e^{-2n(\alpha/4)^{2\ell}}, \end{aligned} \tag{2}$$

where we used Hoeffding's inequality. □

3. Information theoretic lower bound

We prove an information-theoretic lower bound that holds already if G_K is an ℓ-cycle. Let $D(K\|K')$ and $\mathbb{H}(K, K')$ denote respectively the Kullback-Leibler divergence and the Hellinger distance between $\mathsf{DPP}(K)$ and $\mathsf{DPP}(K')$.

Lemma 2. *For $\eta \in \{-, +\}$, let K^η be the $\ell\times\ell$ matrix with elements given by*

$$K_{i,j} = \begin{cases} 1/2 & \text{if } j = i \\ \alpha & \text{if } j = i \pm 1 \\ \eta\alpha & \text{if } (i,j) \in \{(1,\ell),(\ell,1)\} \\ 0 & \text{otherwise} \end{cases}.$$

Then, for any $\alpha \leq 1/8$, it holds

$$D(K\|K') \leq 4(6\alpha)^\ell, \qquad \text{and} \qquad \mathbb{H}(K, K') \leq (8\alpha^2)^\ell\,.$$

Proof. It is straightforward to see that

$$\det(K^+_J) - \det(K^-_J) = \begin{cases} 2\alpha^\ell & \text{if } J = [\ell] \\ 0 & \text{else} \end{cases}.$$

If Y is sampled from $\mathsf{DPP}(K^\eta)$, we denote by $p_\eta(S) = \mathbb{P}[Y = S]$, for $S \subseteq [\ell]$. It follows from the inclusion-exclusion principle that for all $S \subseteq [\ell]$,

$$\begin{aligned} p_+(S) - p_-(S) &= \sum_{J\subseteq[\ell]\setminus S}(-1)^{|J|}(\det K^+_{S\cup J} - \det K^-_{S\cup J}) \\ &= (-1)^{\ell-|S|}(\det K^+ - \det K^-) = \pm 2\alpha^\ell\,, \end{aligned} \tag{3}$$

where $|J|$ denotes the cardinality of J. The inclusion-exclusion principle also yields that $p_\eta(S) = |\det(K^\eta - I_{\bar{S}})|$ for all $S \subseteq [l]$, where $I_{\bar{S}}$ stands for the $\ell\times\ell$ diagonal matrix with ones on its entries (i,i) for $i \notin S$, zeros elsewhere.

Denote by $D(K^+\|K^-)$ the Kullback Leibler divergence between $\mathsf{DPP}(K^+)$ and $\mathsf{DPP}(K^-)$:

$$\begin{aligned} D(K^+\|K^-) &= \sum_{S\subseteq[\ell]} p_+(S)\log\left(\frac{p_+(S)}{p_-(S)}\right) \\ &\leq \sum_{S\subseteq[\ell]}\frac{p_+(S)}{p_-(S)}(p_+(S) - p_-(S)) \\ &\leq 2\alpha^\ell\sum_{S\subseteq[\ell]}\frac{|\det(K^+ - I_{\bar{S}})|}{|\det(K^- - I_{\bar{S}})|}, \end{aligned} \tag{4}$$

by (3). Using the fact that $0 < \alpha \leq 1/8$ and the Gershgorin circle theorem, we conclude that the absolute value of all eigenvalues of $K^\eta - I_{\bar{S}}$ are between $1/4$ and $3/4$. Thus we obtain from (4) the bound $D(K^+\|K^-) \leq 4(6\alpha)^\ell$.

Using the same arguments as above, the Hellinger distance $\mathbb{H}(K^+, K^-)$ between $\mathsf{DPP}(K^+)$ and $\mathsf{DPP}(K^-)$ satisfies

$$\begin{aligned}\mathbb{H}(K^+, K^-) &= \sum_{J\subseteq[\ell]}\left(\frac{p_+(J) - p_-(J)}{\sqrt{p_+(J)} + \sqrt{p_-(J)}}\right)^2 \\ &\leq \sum_{J\subseteq[\ell]}\frac{4\alpha^{2\ell}}{2\cdot 4^{-\ell}} = (8\alpha^2)^\ell\end{aligned}$$

which completes the proof. □

The sample complexity lower bound now follows from standard arguments.

Theorem 2. *Let $0 < \varepsilon \leq \alpha \leq 1/8$ and $3 \leq \ell \leq N$. There exists a constant $C > 0$ such that if*

$$n \leq C\Big(\frac{8^\ell}{\alpha^{2\ell}} + \frac{\log(N/\ell)}{(6\alpha)^\ell} + \frac{\log N}{\varepsilon^2}\Big),$$

then the following holds: for any estimator $\hat{K}$ based on n samples, there exists a kernel K that satisfies Assumption 1 and such that the cycle sparsity of G_K is ℓ and for which $\rho(\hat{K}, K) \geq \varepsilon$ with probability at least $1/3$.

Proof. Recall the notation of Lemma 2. First consider the $N\times N$ block diagonal matrix K (resp. K') where its first block is K^+ (resp. K^-) and its second block is $I_{N-\ell}$. By a standard argument, the Hellinger distance $\mathbb{H}_n(K, K')$ between the product measures $\mathsf{DPP}(K)^{\otimes n}$ and $\mathsf{DPP}(K')^{\otimes n}$ satisfies

$$1 - \frac{\mathbb{H}^2_n(K, K')}{2} = \big(1 - \frac{\mathbb{H}^2(K, K')}{2}\big)^n \geq \big(1 - \frac{\alpha^{2\ell}}{2\times 8^\ell}\big)^n,$$

which yields the first term in the desired lower bound.

Next, by padding with zeros, we can assume that $L = N/\ell$ is an integer. Let $K^{(0)}$ be a block diagonal matrix where each block is K^+ (using the notation of Lemma 2). For $j = 1, \ldots, L$, define the $N \times N$ block diagonal matrix $K^{(j)}$ as the matrix obtained from $K^{(0)}$ by replacing its jth block with K^- (again using the notation of Lemma 2).

Since $\mathsf{DPP}(K^{(j)})$ is the product measure of L lower dimensional DPPs that are each independent of each other, using Lemma 2 we have $D(K^{(j)} \| K^{(0)}) \leq 4(6\alpha)^\ell$. Hence, by Fano's lemma (see, e.g., Corollary 2.6 in (Tsybakov, 2009)), the sample complexity to learn the kernel of a DPP within a distance $\varepsilon \leq \alpha$ is

$$\Omega\left(\frac{\log(N/\ell)}{(6\alpha)^\ell}\right)$$

which yields the second term.

The third term follows from considering $K_0 = (1/2)I_N$ and letting K_j be obtained from K_0 by adding ε to the jth entry along the diagonal. It is easy to see that $D(K_j \| K_0) \leq 8\varepsilon^2$. Hence, a second application of Fano's lemma yields that the sample complexity to learn the kernel of a DPP within a distance ε is $\Omega(\frac{\log N}{\varepsilon^2})$. □

The third term in the lower bound is the standard parametric term and is unavoidable in order to estimate the magnitude of the coefficients of K. The other terms are more interesting. They reveal that the cycle sparsity of G_K, namely, ℓ, plays a key role in the task of recovering the sign pattern of K. Moreover the theorem shows that the sample complexity of our method of moments estimator is near optimal.

4. Algorithms

4.1. Horton's algorithm

We first give an algorithm to compute the estimator $\hat{K}$ defined in Section 2. A well-known algorithm of Horton (Horton, 1987) computes a cycle basis of minimum total length in time $O(m^3 N)$. Subsequently, the running time was improved to $O(m^2 N/\log N)$ time (Amaldi et al., 2010). Also, it is known that a cycle basis of minimum total length is a shortest maximal cycle basis (Chickering et al., 1995). Together, these results imply the following.

Lemma 3. *Let $G = ([N], E)$, $|E| = m$. There is an algorithm to compute a shortest maximal cycle basis in $O(m^2 N/\log N)$ time.*

In addition, we recall the following standard result regarding the complexity of Gaussian elimination (Golub & Van Loan, 2012).

Algorithm 1 Compute Estimator $\hat{K}$

Input: samples $Y_1, ..., Y_n$, parameter $\alpha > 0$.

Compute $\hat{\Delta}_S$ for all $|S| \leq 2$.
Set $\hat{K}_{i,i} = \hat{\Delta}_{\{i\}}$ for $1 \leq i \leq N$.
Compute $\hat{B}_{i,j}$ for $1 \leq i < j \leq N$.
Form $\widetilde{K} \in \mathbb{R}^{N \times N}$ and $\hat{G} = ([N], \hat{E})$.
Compute a shortest maximal cycle basis $\{\hat{v}_1, ..., \hat{v}_{\nu_{\hat{G}}}\}$.
Compute $\hat{\Delta}_{\hat{S}_i}$ for $1 \leq i \leq \nu_{\hat{G}}$.
Compute $\hat{C}_{\hat{S}_i}$ using $\det \widetilde{K}_{\hat{S}_i}$ for $1 \leq i \leq \nu_{\hat{G}}$.
Construct $A \in GF(2)^{\nu_{\hat{G}} \times m}$, $b \in GF(2)^{\nu_{\hat{G}}}$.
Solve $Ax = b$ using Gaussian elimination.
Set $\hat{K}_{i,j} = \hat{K}_{j,i} = (2x_{\{i,j\}} - 1)\hat{B}_{i,j}^{1/2}$, for all $\{i,j\} \in \hat{E}$.

Lemma 4. *Let $A \in GF(2)^{\nu \times m}$, $b \in GF(2)^\nu$. Then Gaussian elimination will find a vector $x \in GF(2)^m$ such that $Ax = b$ or conclude that none exists in $O(\nu^2 m)$ time.*

We give our procedure for computing the estimator $\hat{K}$ in Algorithm 1. In the following theorem, we bound the running time of Algorithm 1 and establish an upper bound on the sample complexity needed to solve the recovery problem as well as the sample complexity needed to compute an estimate $\hat{K}$ that is close to K.

Theorem 3. *Let $K \in \mathbb{R}^{N \times N}$ be a symmetric matrix satisfying $0 \preceq K \preceq I$, and satisfying Assumption 1. Let G_K be the graph induced by K and ℓ be the cycle sparsity of G_K. Let $Y_1, ..., Y_n$ be samples from DPP(K) and $\delta \in (0,1)$. If*

$$n > \frac{\log(N^{\ell+1}/\delta)}{(\alpha/4)^{2\ell}},$$

then with probability at least $1 - \delta$, Algorithm 1 computes an estimator $\hat{K}$ which recovers the signs of K up to a $\mathcal{D}_N$-similarity and satisfies

$$\rho(K, \hat{K}) < \frac{1}{\alpha}\left(\frac{8\log(4N^{\ell+1}/\delta)}{n}\right)^{1/2} \tag{5}$$

in $O(m^3 + nN^2)$ time.

Proof. (5) follows directly from (2) in the proof of Theorem 1. That same proof also shows that with probability at least $1 - \delta$, the support of G_K and the signs of K are recovered up to a $\mathcal{D}_N$-similarity. What remains is to upper bound the worst case run time of Algorithm 1. We will perform this analysis line by line. Initializing $\hat{K}$ requires $O(N^2)$ operations. Computing Δ_S for all subsets $|S| \leq 2$ requires $O(nN^2)$ operations. Setting $\hat{K}_{i,i}$ requires $O(N)$ operations. Computing $\hat{B}_{i,j}$ for $1 \leq i < j \leq N$ requires $O(N^2)$ operations. Forming $\widetilde{K}$ requires $O(N^2)$ operations. Forming G_K requires $O(N^2)$ operations. By

Lemma 3, computing a shortest maximal cycle basis requires $O(mN)$ operations. Constructing the subsets S_i, $1 \le i \le \nu_{\hat{G}}$, requires $O(mN)$ operations. Computing $\hat{\Delta}_{S_i}$ for $1 \le i \le \nu_{\hat{G}}$ requires $O(nm)$ operations. Computing $\hat{C}_{S_i}$ using $\det(\widetilde{K}[S_i])$ for $1 \le i \le \nu_{\hat{G}}$ requires $O(m\ell^3)$ operations, where a factorization of each $\widetilde{K}[S_i]$ is used to compute each determinant in $O(\ell^3)$ operations. Constructing A and b requires $O(m\ell)$ operations. By Lemma 4, finding a solution x using Gaussian elimination requires $O(m^3)$ operations. Setting $\hat{K}_{i,j}$ for all edges $\{i,j\} \in E$ requires $O(m)$ operations. Put this all together, Algorithm 1 runs in $O(m^3 + nN^2)$ time. □

4.2. Chordal Graphs

Here we show that it is possible to obtain faster algorithms by exploiting the structure of G_K. Specifically, in the case where G_K chordal, we give an $O(m)$ time algorithm to determine the signs of the off-diagonal entries of the estimator $\hat{K}$, resulting in an improved overall runtime of $O(m + nN^2)$. Recall that a graph $G = ([N], E)$ is said to be *chordal* if every induced cycle in G is of length three. Moreover, a graph $G = ([N], E)$ has a *perfect elimination ordering (PEO)* if there exists an ordering of the vertex set $\{v_1, ..., v_N\}$ such that, for all i, the graph induced by $\{v_i\} \cup \{v_j | \{i,j\} \in E, j > i\}$ is a clique. It is well known that a graph is chordal if and only if it has a PEO. A PEO of a chordal graph with m edges can be computed in $O(m)$ operations using lexicographic breadth-first search (Rose et al., 1976).

Lemma 5. *Let $G = ([N], E)$, be a chordal graph and $\{v_1, ..., v_n\}$ be a PEO. Given i, let $i^* := \min\{j | j > i, \{v_i, v_j\} \in E\}$. Then the graph $G' = ([N], E')$, where $E' = \{\{v_i, v_{i^*}\}\}_{i=1}^{N-\kappa(G)}$, is a spanning forest of G.*

Proof. We first show that there are no cycles in G'. Suppose to the contrary, that there is an induced cycle C of length k on the vertices $\{v_{j_1}, ..., v_{j_k}\}$. Let v be the vertex of smallest index. Then v is connected to two other vertices in the cycle of larger index. This is a contradiction to the construction.

What remains is to show that $|E'| = N - \kappa(G)$. It suffices to prove the case $\kappa(G) = 1$. Suppose to the contrary, that there exists a vertex v_i, $i < N$, with no neighbors of larger index. Let P be the shortest path in G from v_i to v_N. By connectivity, such a path exists. Let v_k be the vertex of smallest index in the path. However, it has two neighbors in the path of larger index, which must be adjacent to each other. Therefore, there is a shorter path. □

Now, given the chordal graph G_K induced by K and the estimates of principal minors of size at most three, we provide an algorithm to determine the signs of the edges of G_K, or, equivalently, the off-diagonal entries of K.

Algorithm 2 Compute Signs of Edges in Chordal Graph

Input: $G_K = ([N], E)$ chordal, $\hat{\Delta}_S$ for $|S| \le 3$.

Compute a PEO $\{v_1, ..., v_N\}$.
Compute the spanning forest $G' = ([N], E')$.
Set all edges in E' to have positive sign.
Compute $\hat{C}_{\{i,j,i^*\}}$ for all $\{i,j\} \in E \setminus E'$, $j < i$.
Order edges $E \setminus E' = \{e_1, ..., e_\nu\}$ such that $i > j$ if $\max e_i < \max e_j$.
Visit edges in sorted order and for $e = \{i,j\}$, $j > i$, set

$$\operatorname{sgn}(\{i,j\}) = \operatorname{sgn}(\hat{C}_{\{i,j,i^*\}}) \operatorname{sgn}(\{i,i^*\}) \operatorname{sgn}(\{j,i^*\}).$$

Theorem 4. *If G_K is chordal, Algorithm 2 correctly determines the signs of the edges of G_K in $O(m)$ time.*

Proof. We will simultaneously perform a count of the operations and a proof of the correctness of the algorithm. Computing a PEO requires $O(m)$ operations. Computing the spanning forest requires $O(m)$ operations. The edges of the spanning tree can be given arbitrary sign, because it is a cycle-free graph. This assigns a sign to two edges of each 3-cycle. Computing each $\hat{C}_{\{i,j,i^*\}}$ requires a constant number of operations because $\ell = 3$, requiring a total of $O(m - N)$ operations. Ordering the edges requires $O(m)$ operations. Setting the signs of each remaining edge requires $O(m)$ operations. □

Therefore, when G_K is chordal, the overall complexity required by our algorithm to compute $\hat{K}$ is reduced to $O(m + nN^2)$.

5. Experiments

Here we present experiments to supplement the theoretical results of the paper. We test our algorithm on two types of random matrices. First, we consider the matrix $K \in \mathbb{R}^{N \times N}$ corresponding to the cycle on N vertices,

$$K = \frac{1}{2} I + \frac{1}{4} A,$$

where A is symmetric and has non-zero entries only on the edges of the cycle, either $+1$ or -1, each with probability $1/2$. By the Gershgorin circle theorem, $0 \preceq K \preceq I$. Next, we consider the matrix $K \in \mathbb{R}^{N \times N}$ corresponding to the clique on N vertices,

$$K = \frac{1}{2} I + \frac{1}{4\sqrt{N}} A,$$

where A is symmetric and has all entries either $+1$ or -1, each with probability $1/2$. It is well known that $-2\sqrt{N} \preceq A \preceq 2\sqrt{N}$ with high probability, implying $0 \preceq K \preceq I$.

For both cases and for a range of values of matrix dimension N and samples n, we run our algorithm on 50 randomly generated instances. We record the proportion of trials where we recover the graph induced by K, and the proportion of the trials where we recover both the graph and correctly determine the signs of the entries.

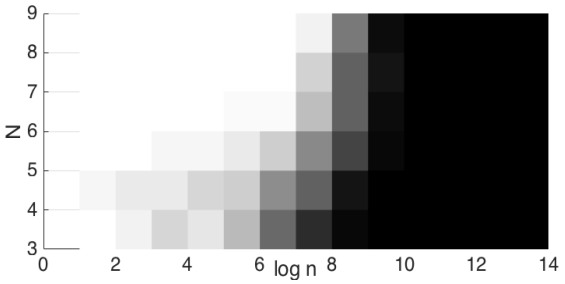
In Figure 1, the shade of each box represents the proportion of trials where recovery was successful for a given pair N, n. A completely white box corresponds to zero success rate, black to a perfect success rate.

The plots corresponding to the cycle and the clique are telling. We note that for the clique, recovering the sparsity pattern and recovering the signs of the off-diagonal entries come hand-in-hand. However, for the cycle, there is a noticeable gap between the number of samples required to recover the sparsity pattern and the number of samples required to recover the signs of the off-diagonal entries. This empirically confirms the central role that cycle sparsity plays in parameter estimation, and further corroborates our theoretical results.

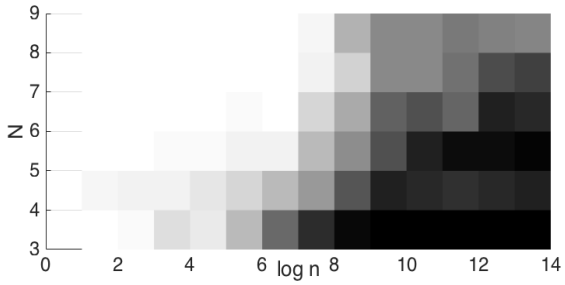
6. Conclusion and open questions

In this paper, we gave the first provable guarantees for learning the parameters of a DPP. Our upper and lower bounds reveal the key role played by the parameter ℓ, which is the cycle sparsity of graph induced by the kernel of the DPP. Our estimator does not need to know ℓ beforehand, but can adapt to the instance. Moreover, our procedure outputs an estimate of ℓ, which could potentially be used for further inference questions such as testing and confidence intervals. An interesting open question is whether on a graph by graph basis, the parameter ℓ exactly determines the optimal sample complexity. Moreover when the number of samples is too small, can we exactly characterize which signs can be learned correctly and which cannot (up to a similarity transformation by D)? Such results would lend new theoretical insights into the output of algorithms for learning DPPs, and which individual parameters in the estimate we can be confident about and which we cannot.

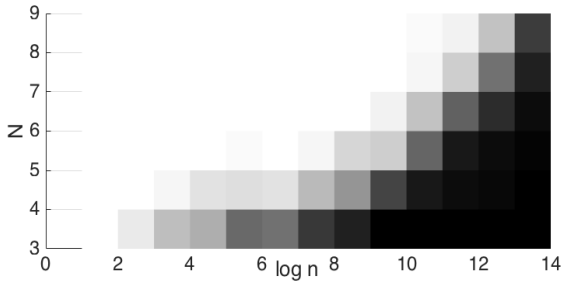
Acknowledgements. A.M. is supported in part by NSF CAREER Award CCF-1453261, NSF Large CCF-1565235, a David and Lucile Packard Fellowship and an Alfred P. Sloan Fellowship. P.R. is supported in part by NSF CAREER DMS-1541099, NSF DMS-1541100, DARPA W911NF-16-1-0551, ONR N00014-17-1-2147 and a grant from the MIT NEC Corporation.



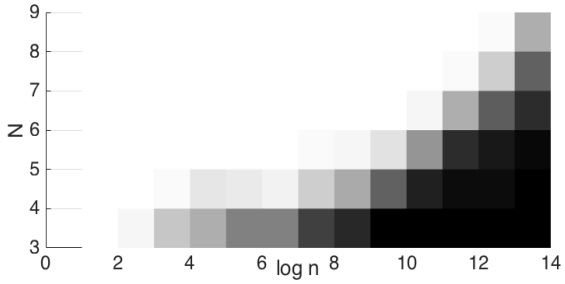
(a) graph recovery, cycle



(b) graph and sign recovery, cycle



(c) graph recovery, clique



(d) graph and sign recovery, clique

Figure 1: Plots of the proportion of successive graph recovery, and graph and sign recovery, for random matrices with cycle and clique graph structure, respectively. The darker the box, the higher the proportion of trials that were recovered successfully.

References

Affandi, Raja Hafiz, Fox, Emily B., Adams, Ryan P., and Taskar, Benjamin. Learning the parameters of determinantal point process kernels. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pp. 1224–1232, 2014.

Amaldi, Edoardo, Iuliano, Claudio, and Rizzi, Romeo. Efficient deterministic algorithms for finding a minimum cycle basis in undirected graphs. In *International Conference on Integer Programming and Combinatorial Optimization*, pp. 397–410. Springer, 2010.

Batmanghelich, Nematollah Kayhan, Quon, Gerald, Kulesza, Alex, Kellis, Manolis, Golland, Polina, and Bornn, Luke. Diversifying sparsity using variational determinantal point processes. *ArXiv: 1411.6307*, 2014.

Borodin, Alexei and Rains, Eric M. Eynard–mehta theorem, schur process, and their pfaffian analogs. *Journal of statistical physics*, 121(3):291–317, 2005.

Brunel, Victor-Emmanuel, Moitra, Ankur, Rigollet, Philippe, and Urschel, John. Maximum likelihood estimation of determinantal point processes. *arXiv:1701.06501*, 2017.

Chickering, David M., Geiger, Dan, and Heckerman, David. On finding a cycle basis with a shortest maximal cycle. *Information Processing Letters*, 54(1):55 – 58, 1995.

Çivril, Ali and Magdon-Ismail, Malik. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoretical Computer Science*, 410(47-49):4801–4811, 2009.

Deshpande, Amit and Rademacher, Luis. Efficient volume sampling for row/column subset selection. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pp. 329–338. IEEE, 2010.

Dyson, Freeman J. Statistical theory of the energy levels of complex systems. III. *J. Mathematical Phys.*, 3:166–175, 1962. ISSN 0022-2488.

Gillenwater, Jennifer A, Kulesza, Alex, Fox, Emily, and Taskar, Ben. Expectation-maximization for learning determinantal point processes. In *NIPS*, 2014.

Golub, Gene H and Van Loan, Charles F. *Matrix computations*, volume 3. JHU Press, 2012.

Horton, Joseph Douglas. A polynomial-time algorithm to find the shortest cycle basis of a graph. *SIAM Journal on Computing*, 16(2):358–366, 1987.

Kulesza, A. *Learning with determinantal point processes*. PhD thesis, University of Pennsylvania, 2012.

Kulesza, Alex and Taskar, Ben. k-DPPs: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 1193–1200, 2011.

Kulesza, Alex and Taskar, Ben. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc., Hanover, MA, USA, 2012. ISBN 1601986289, 9781601986283.

Lee, Donghoon, Cha, Geonho, Yang, Ming-Hsuan, and Oh, Songhwai. Individualness and determinantal point processes for pedestrian detection. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pp. 330–346, 2016.

Li, Chengtao, Jegelka, Stefanie, and Sra, Suvrit. Fast sampling for strongly rayleigh measures with application to determinantal point processes. *1607.03559*, 2016a.

Li, Chengtao, Jegelka, Stefanie, and Sra, Suvrit. Fast dpp sampling for nystrom with application to kernel methods. *International Conference on Machine Learning (ICML)*, 2016b.

Lin, Hui and Bilmes, Jeff A. Learning mixtures of submodular shells with application to document summarization. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, August 14-18, 2012*, pp. 479–490, 2012.

Macchi, Odile. The coincidence approach to stochastic point processes. *Advances in Appl. Probability*, 7:83–122, 1975. ISSN 0001-8678.

Mariet, Zelda and Sra, Suvrit. Fixed-point algorithms for learning determinantal point processes. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 2389–2397, 2015.

Nikolov, Aleksandar. Randomized rounding for the largest simplex problem. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pp. 861–870. ACM, 2015.

Nikolov, Aleksandar and Singh, Mohit. Maximizing determinants under partition constraints. In *STOC*, pp. 192–201, 2016.

Rebeschini, Patrick and Karbasi, Amin. Fast mixing for discrete point processes. In *COLT*, pp. 1480–1500, 2015.

Rising, Justin, Kulesza, Alex, and Taskar, Ben. An efficient algorithm for the symmetric principal minor assignment problem. *Linear Algebra and its Applications*, 473:126 – 144, 2015.

Rose, Donald J, Tarjan, R Endre, and Lueker, George S. Algorithmic aspects of vertex elimination on graphs. *SIAM Journal on computing*, 5(2):266–283, 1976.

Snoek, Jasper, Zemel, Richard S., and Adams, Ryan Prescott. A determinantal point process latent variable model for inhibition in neural spiking data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pp. 1932–1940, 2013.

Summa, Marco Di, Eisenbrand, Friedrich, Faenza, Yuri, and Moldenhauer, Carsten. On largest volume simplices and sub-determinants. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 315–323. Society for Industrial and Applied Mathematics, 2015.

Tsybakov, Alexandre B. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

Xu, Haotian and Ou, Haotian. Scalable discovery of audio fingerprint motifs in broadcast streams with determinantal point process based motif clustering. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 24(5): 978–989, 2016.

Yao, Jin-ge, Fan, Feifan, Zhao, Wayne Xin, Wan, Xiaojun, Chang, Edward Y., and Xiao, Jianguo. Tweet timeline generation with determinantal point processes. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pp. 3080–3086, 2016.