Maximum likelihood estimation of determinantal point processes

Victor-Emmanuel Brunel, Ankur Moitra[†],
Philippe Rigollet[*] and John Urschel

Massachusetts Institute of Technology

Abstract. Determinantal point processes (DPPs) have wide-ranging applications in machine learning, where they are used to enforce the notion of diversity in subset selection problems. Many estimators have been proposed, but surprisingly the basic properties of the maximum likelihood estimator (MLE) have received little attention. The difficulty is that it is a non-concave maximization problem, and such functions are notoriously difficult to understand in high dimensions, despite their importance in modern machine learning. Here we study both the local and global geometry of the expected log-likelihood function. We prove several rates of convergence for the MLE and give a complete characterization of the case where these are parametric. We also exhibit a potential curse of dimensionality where the asymptotic variance of the MLE scales exponentially with the dimension of the problem. Moreover, we exhibit an exponential number of saddle points, and give evidence that these may be the only critical points.

AMS 2000 subject classifications: Primary 62F10; secondary 60G55.
Key words and phrases: Determinantal point processes, statistical estimation, maximum likelihood, L-ensembles.

1. INTRODUCTION

Determinantal point processes (DPPs) describe a family of repulsive point processes; they induce probability distributions that favor configurations of points that are far away from each other. DPPs are often split into two categories: discrete and continuous. In the former case, realizations of the DPP are vectors from the Boolean hypercube $\{0,1\}^N$, while in the latter, they occupy a continuous space such as $\mathbb{R}^d$. In both settings, the notion of distance can be understood in the sense of the natural metric with which the space is endowed. Such processes were formally introduced in the context of quantum mechanics to model systems of fermions [Mac75] that were known to have a repulsive behavior, though DPPs have appeared implicitly in earlier work on random matrix theory, e.g. [Dys62]. Since then, they have played a central role in various corners of probability,

[*] This work was supported in part by NSF CAREER DMS-1541099, NSF DMS-1541100, DARPA BAA-16-46 and a grant from the MIT NEC Corporation.

[†] This work was supported in part by NSF CAREER Award CCF-1453261, NSF Large CCF-1565235, a David and Lucile Packard Fellowship, an Alfred P. Sloan Fellowship, an Edmund F. Kelley Research Award, a Google Research Award and a grant from the MIT NEC Corporation.

algebra and combinatorics [BO00, BS03, Bor11, Oko01, OR03], for example, by allowing exact computations for integrable systems.

Following the seminal work of Kulesza and Taskar [KT12], both discrete and continuous DPPs have recently gained attention in the machine learning literature where the repulsive character of DPPs has been used to enforce the notion of diversity in subset selection problems. Such problems are pervasive to a variety of applications such as document or timeline summarization [LB12, YFZ$^+$16], image search [KT11, AFAT14], audio signal processing [XO16], image segmentation [LCYO16], bioinformatics [BQK$^+$14], neuroscience [SZA13] and wireless or cellular networks modelization [MS14, TL14, LBDA15, DZH15]. DPPs have also been employed as methodological tools in Bayesian and spatial statistics [KK16, BC16], survey sampling [LM15, CJM16] and Monte Carlo methods [BH16].

Even though most of the aforementioned applications necessitate estimation of the parameters of a DPP, statistical inference for DPPs has received little attention. In this context, maximum likelihood estimation is a natural method, but generally leads to a non-convex optimization problem. This problem has been addressed by various heuristics, including Expectation-Maximization [GKFT14], MCMC [AFAT14], and fixed point algorithms [MS15]. None of these methods come with global guarantees, however. Another route used to overcome the computational issues associated with maximizing the likelihood of DPPs consists in imposing additional modeling constraints, initially in [KT12, AFAT14, BTRA15], and, more recently, [DB16, GPK16a, GPK16b, MS16], in which assuming a specific low rank structure for the problem enabled the development of sublinear time algorithms.

The statistical properties of the maximum likelihood estimator for such problems have received attention only in the continuous case and under strong parametric assumptions [LMR15, BL16] or smoothness assumptions in a nonparametric context [Bar13]. However, despite their acute relevance to machine learning and several algorithmic advances (see [MS15] and references therein), the statistical properties of general discrete DPPs have not been established. Qualitative and quantitative characterizations of the likelihood function would shed light on the convergence rate of the maximum likelihood estimator, as well as aid in the design of new estimators.

In this paper, we take an information geometric approach to understand the asymptotic properties of the maximum likelihood estimator. First, we study the curvature of the expected log-likelihood around its maximum. Our main result is an exact characterization of when the maximum likelihood estimator converges at a parametric rate (Theorem 8). Moreover, we give quantitative bounds on the strong convexity constant (Proposition 9) that translate into lower bounds on the asymptotic variance and shed light on what combinatorial parameters of a DPP control said variance. Second, we study the global geometry of the expected log-likelihood function. We exhibit an exponential number of saddle points that correspond to partial decouplings of the DPP (Theorem 11). We conjecture that these are the only critical points, which would be a key step in showing that the maximum likelihood estimator can be computed efficiently after all, in spite of the fact that it is attempting to maximize a non-concave function.

The remainder of the paper is as follows. In Section 2, we provide an introduction to DPPs together with notions and properties that are useful for our purposes. In Section 3, we study the information landscape of DPPs and specifically, the local behavior of the expected log-likelihood around its critical points. Finally, we translate these results into rates of convergence for maximum likelihood estimation in Section 4. All proofs are gathered in Section 6 in order to facilitate the narrative.

Notation. Fix a positive integer N and define $[N] = \{1, 2, \dots, N\}$. Throughout the paper, $\mathcal{X}$ denotes a subset of $[N]$. We denote by $\wp(\mathcal{X})$ the power set of $\mathcal{X}$.

We implicitly identify the set of $|\mathcal{X}| \times |\mathcal{X}|$ matrices to the the set of mappings from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. As a result, we denote by $I_{\mathcal{X}}$ the identity matrix in $\mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ and we omit the subscript whenever $\mathcal{X} = [N]$. For a matrix $A \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ and $J \subset \mathcal{X}$, denote by A_J the restriction of A to $J \times J$. When defined over $\mathcal{X} \times \mathcal{X}$, A_J maps elements outside of $J \times J$ to zero.

Let $\mathcal{S}_{\mathcal{X}}$ denote the set of symmetric matrices in $\mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ matrices and denote by $\mathcal{S}_{\mathcal{X}}^{\Lambda}$ the subset of matrices in $\mathcal{S}_{\mathcal{X}}$ that have eigenvalues in $\Lambda \subset \mathbb{R}$. Of particular interest are $\mathcal{S}_{\mathcal{X}}^{+} = \mathcal{S}_{\mathcal{X}}^{[0,\infty)}$, $\mathcal{S}_{\mathcal{X}}^{++} = \mathcal{S}_{\mathcal{X}}^{(0,\infty)}$, the subsets of positive semidefinite and positive definite matrices respectively.

For a matrix $A \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$, we denote by $\|A\|_F$, $\det(A)$ and $\mathrm{Tr}(A)$ its Frobenius norm, determinant and trace respectively. We set $\det A_\varnothing = 1$ and $\mathrm{Tr}\, A_\varnothing = 0$. Moreover, we denote by $\mathrm{diag}(A)$ the vector of size $|\mathcal{X}|$ with entries given by the diagonal elements of A. If $x \in \mathbb{R}^N$, we denote by $\mathrm{Diag}(x)$ the $N \times N$ diagonal matrix with diagonal given by x.

For $\mathcal{A} \in \mathcal{S}_{\mathcal{X}}$, $k \geqslant 1$ and a smooth function $f : \mathcal{A} \to \mathbb{R}$, we denote by $\mathrm{d}^k f(A)$ the k-th derivative of f evaluated at $A \in \mathcal{A}$. This is a k-linear map defined on $\mathcal{A}$; for $k = 1$, $\mathrm{d}f(A)$ is the gradient of f, $\mathrm{d}^2 f(A)$ the Hessian, etc.

Throughout this paper, we say that a matrix $A \in \mathcal{S}_{\mathcal{X}}$ is block diagonal if there exists a partition $\{J_1, \dots, J_k\}$, $k \geqslant 1$, of $\mathcal{X}$ such that $A_{ij} = 0$ if $i \in J_a, j \in J_b$ and $a \neq b$. The largest number k such that such a representation exists is called the *number of blocks* of A and in this case $J_1, \dots, J_k$ are called *blocks* of L.

2. DETERMINANTAL POINT PROCESSES AND L-ENSEMBLES

In this section we gather definitions and useful properties, old and new, about determinantal point processes.

2.1 Definitions

A (discrete) *determinantal point process* (DPP) on $\mathcal{X}$ is a random variable $Z \in \wp(\mathcal{X})$ with distribution

$$\mathbb{P}[J \subset Z] = \det(K_J), \quad \forall\, J \subset \mathcal{X}, \tag{2.1}$$

where $K \in \mathcal{S}_{\mathcal{X}}^{[0,1]}$, is called the *correlation kernel* of Z.

If it holds further that $K \in \mathcal{S}_{\mathcal{X}}^{(0,1)}$, then Z is called L-*ensemble* and there exists a matrix $L = K(I - K)^{-1} \in \mathcal{S}_{\mathcal{X}}^{++}$ such that

$$\mathbb{P}[Z = J] = \frac{\det(L_J)}{\det(I + L)}, \quad \forall\, J \subset \mathcal{X}, \tag{2.2}$$

Using the multilinearity of the determinant, it is easy to see that (2.2) defines a probability distribution (see Lemma 17). We call L the *kernel* of the L-ensemble Z.

Using the inclusion-exclusion principle, it follows from (2.1) that $\mathbb{P}(Z = \varnothing) = \det(I - K)$. Hence, a DPP Z with correlation kernel K is an L-ensemble if and only if Z can be empty with positive probability.

In this work, we only consider DPPs that are L-ensembles. In that setup, we can identify L-ensembles and DPPs, and the kernel L and correlation kernel K are related by the identities

$$L = K(I - K)^{-1}\,, \qquad K = L(I + L)^{-1}. \tag{2.3}$$

Note that we only consider kernels L that are positive definite. In general L-ensembles may also be defined for $L \in \mathcal{S}_{\mathcal{X}}^{+}$, when $K \in \mathcal{S}_{\mathcal{X}}^{[0,1)}$. We denote by $\mathsf{DPP}_{\mathcal{X}}(L)$ the probability distribution associated with the DPP with kernel L and refer to L as the *parameter* of the DPP in the context of statistical estimation. If $\mathcal{X} = [N]$, we drop the subscript and only write $\mathsf{DPP}(L)$ for a DPP with kernel L on $[N]$.

2.2 Negative association

Perhaps one of the most distinctive feature of DPPs is their repellent nature. It can be characterized by the notion of *negative association*, which has been extensively covered in the mathematics literature [BBL09]. To define this notion, we recall that a function $f : \{0,1\}^N \to \mathbb{R}$ is *non decreasing* if for all $x = (x_1, \dots, x_N)$, $y = (y_1, \dots, y_N) \in \{0,1\}^N$ such that $x_i \leqslant y_i$, $\forall\, i \in [N]$, it holds that $f(x) \leqslant f(y)$.

Let Z be a DPP on $[N]$ with kernel $L \in \mathcal{S}_{[N]}^{++}$ and correlation kernel $K = L(I + L)^{-1} \in \mathcal{S}_{[N]}^{(0,1)}$. Denote by $\chi(Z) \in \{0,1\}^N$ the (random) characteristic vector of Z. Note that $\mathbb{E}[\chi(Z)] = \mathrm{diag}(K)$, moreover, the entries of $\chi(Z)$ are *conditionally negatively associated*.

Definition 1. *Let Z be a random subset of $[N]$ with characteristic vector $X = \chi(Z) \in \{0,1\}^N$. The coordinates $X_1, \dots, X_N \in \{0,1\}$ of X are said to be negatively associated Bernoulli random variables if for all $J, J' \subset [N]$ such that $J \cap J' = \varnothing$ and all non decreasing functions f and g on $\{0,1\}^N$, it holds*

$$\mathbb{E}\big[f(\chi(Z \cap J))g(\chi(Z \cap J'))\big] \leqslant \mathbb{E}\big[f(\chi(Z \cap J))\big]\mathbb{E}\big[g(\chi(Z \cap J'))\big]\,.$$

Moreover, $X_1, \dots, X_N$ are conditionally negatively associated if it also holds that for all $S \subset [N]$,

$$\begin{aligned}\mathbb{E}\big[f(\chi(Z \cap J))g(\chi(Z \cap J'))\big|Z \cap S\big]&\\ \leqslant \mathbb{E}\big[f(\chi(Z \cap J))\big|Z \cap S\big]&\mathbb{E}\big[g(\chi(Z \cap J'))\big|Z \cap S\big]\end{aligned}$$

almost surely.

Negative association is much stronger than pairwise non positive correlations. Conditional negative association is even stronger, and this property will be essential for the proof of Theorem 11. The following lemma is a direct consequence of Theorem 3.4 of [BBL09].

Lemma 2. *Let $Z \sim \mathsf{DPP}(L)$ for some $L \in \mathcal{S}_{[N]}^{++}$ and denote by $\chi(Z) = (X_1, \dots, X_N) \in \{0,1\}^N$ its characteristic vector. Then, the Bernoulli random variables $X_1, \dots, X_N$ are conditionally negatively associated.*

Now we introduce the notion of a *partial decoupling* of a DPP. This notion will be relevant in the study of the likelihood geometry of DPPs.

DEFINITION 3. *Let $\mathcal{P}$ be a partition of $[N]$. A* partial decoupling *Z' of a DPP Z on $[N]$ according to partition $\mathcal{P}$ is a random subset of $[N]$ such that $\{\chi(Z' \cap J), J \in \mathcal{P}\}$ are mutually independent and $\chi(Z' \cap J)$ has the same distribution as $\chi(Z \cap J)$ for all $J \in \mathcal{P}$. We say that the partial decoupling is strict if and only if Z' does not have the same distribution as Z.*

It is not hard to see that a partial decoupling Z' associated to a partition $\mathcal{P}$ of a DPP Z is also a DPP with correlation kernel K' given by

$$K'_{i,j} = \begin{cases} K_{i,j} & \text{if } i,j \in J \text{ for some } J \in \mathcal{P}\,, \\ 0 & \text{otherwise.} \end{cases}$$

In particular, note that if Y' is a strict partial decoupling of a DPP Y, then its correlation kernel K and thus its kernel L are both block diagonal with at least two blocks.

2.3 Identifiability

The probability mass function (2.2) of $\mathsf{DPP}(L)$ depends only on the principal minors of L and on $\det(I + L)$. In particular, L is not fully identified by $\mathsf{DPP}(L)$ and the lack of identifiability of L has been characterized exactly [Kul12, Theorem 4.1]. Denote by $\mathcal{D}$ the collection of $N \times N$ diagonal matrices with ± 1 diagonal entries. Then, for $L_1, L_2 \in \mathcal{S}^{++}_{[N]}$,

$$\mathsf{DPP}(L_1) = \mathsf{DPP}(L_2) \iff \exists D \in \mathcal{D}, L_2 = DL_1D. \tag{2.4}$$

We define the degree of identifiability of a kernel L as follows.

DEFINITION 4. *Let $L \in \mathcal{S}^{++}_{[N]}$. The degree $\mathsf{Deg}(L)$ of identifiability of L is the cardinality of the family $\{DLD : D \in \mathcal{D}\}$. We say that L is* irreducible *whenever $\mathsf{Deg}(L) = 2^{N-1}$ and* reducible *otherwise. If $Z \sim \mathsf{DPP}(L)$ for some $L \in \mathcal{S}^{++}_{[N]}$, we also say that Z is irreducible if L is irreducible, and that Z is reducible otherwise.*

For instance, the degree of identifiability of a diagonal kernel is 1. It is easy to check that diagonal kernels are the only ones with degree of identifiability equal to 1. These kernels are perfectly identified. Intuitively, the higher the degree, the less the kernel is identified. It is clear that for all $L \in \mathcal{S}^{++}_{[N]}, 1 \leqslant \mathsf{Deg}(L) \leqslant 2^{N-1}$.

As we will see in Proposition 6, the degree of identifiability of a kernel L is completely determined by its block structure. The latter can in turn be characterized by the connectivity of certain graphs that we call *determinantal graphs*.

DEFINITION 5. *Fix $\mathcal{X} \subset [N]$. The* determinantal graph *$\mathcal{G}_L = (\mathcal{X}, E_L)$ of a DPP with kernel $L \in \mathcal{S}^{++}_{\mathcal{X}}$ is the undirected graph with vertices $\mathcal{X}$ and edge set $E_L = \big\{\{i,j\} : L_{i,j} \neq 0\big\}$. If $i,j \in \mathcal{X}$, write $i \sim_L j$ if there exists a path in $\mathcal{G}_L$ that connects i and j.*

It is not hard to see that a DPP with kernel L is irreducible if and only if its determinantal graph $\mathcal{G}_L$ is connected. The blocks of L correspond to the connected components of $\mathcal{G}_L$. Moreover, it follows directly from (2.2) that if $Z \sim \mathsf{DPP}(L)$ and L has blocks $J_1, \ldots, J_k$, then $Z \cap J_1, \ldots, Z \cap J_k$ are mutually independent DPPs with correlation kernels $K_{J_1}, \ldots, K_{J_k}$ respectively, where $K = L(I+L)^{-1}$ is the correlation kernel of Z.

The main properties regarding identifiability of DPPs are gathered in the following straightforward proposition.

Proposition 6. *Let $L \in \mathcal{S}^{++}_{[N]}$ and $Z \sim \mathsf{DPP}(L)$. Let $1 \leqslant k \leqslant N$ and $\{J_1, \ldots, J_k\}$ be a partition of $[N]$. The following statements are equivalent:*

1. *L is block diagonal with k blocks $J_1, \ldots, J_k$,*
2. *K is block diagonal with k blocks $J_1, \ldots, J_k$,*
3. *$Z \cap J_1, \ldots, Z \cap J_k$ are mutually independent irreducible DPPs,*
4. *$\mathcal{G}_L$ has k connected components given by $J_1, \ldots, J_k$,*
5. *$L = D_j L D_j$, for $D_j = \mathrm{Diag}(2\chi(J_j) - 1) \in \mathcal{D}$, for all $j \in [k]$.*

In particular, Proposition 6 shows that the degree of identifiability of $L \in \mathcal{S}^{++}_{[N]}$ is $\mathsf{Deg}(L) = 2^{N-k}$, where k is the number of blocks of L.

Now that we have reviewed useful properties of DPPs, we are in a position to study the information landscape for the statistical problem of estimating the kernel of a DPP from independent observations.

3. GEOMETRY OF THE LIKELIHOOD FUNCTIONS

3.1 Definitions

Our goal is to estimate an unknown kernel $L^* \in \mathcal{S}^{++}_{[N]}$ from n independent copies of $Z \sim \mathsf{DPP}(L^*)$. In this paper, we study the statistical properties of what is arguably the most natural estimation technique: maximum likelihood estimation.

Let $Z_1, \ldots, Z_n$ be n independent copies of $Z \sim \mathsf{DPP}(L^*)$ for some unknown $L^* \in \mathcal{S}^{++}_{[N]}$. The (scaled) log-likelihood associated to this model is given for any $L \in \mathcal{S}^{++}_{[N]}$,

$$\hat{\Phi}(L) = \frac{1}{n}\sum_{i=1}^{n} \log p_{Z_i}(L) = \sum_{J \subset [N]} \hat{p}_J \log\det(L_J) - \log\det(I+L)\,, \tag{3.1}$$

where $p_J(L) = \mathbb{P}[Z = J]$ is defined in (2.2) and $\hat{p}_J$ is its empirical counterpart defined by

$$\hat{p}_J = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Z_i = J)\,.$$

Here $\mathbb{1}(\cdot)$ denotes the indicator function.

Using the identity (2.3), it is also possible to write $p_J(L)$ as

$$p_J(L) = |\det(K - I_{\bar{J}})|,$$

where $\bar{J}$ is the complement of J. Hence, the log-likelihood function can be defined with respect to $K \in \mathcal{S}^{(0,1)}_{[N]}$ as

$$\hat{\Psi}(K) = \sum_{J \subset [N]} \hat{p}_J \log|\det(K - I_{\bar{J}})|\,. \tag{3.2}$$

We denote by Φ_{L*} (resp. Ψ_{L*}) the expected log-likelihood as a function of L (resp. K):

$$\Phi_{L*}(L) = \sum_{J\subset[N]} p_J(L^*) \log\det(L_J) - \log\det(I+L)\,. \tag{3.3}$$

and

$$\Psi_{L*}(K) = \sum_{J\subset[N]} p_J(L^*) \log|\det(K - I_{\bar{J}})|\,. \tag{3.4}$$

For the ease of notation, we assume in the sequel that L^* is fixed, and write simply $\Phi = \Phi_{L*}$, $\Psi = \Psi_{L*}$ and $p^*_J = p_J(L^*)$, for $J \subset [N]$.

We now proceed to studying the function Φ. Namely, we study its critical points and their type: local/global maxima, minima and saddle points. We also give a necessary and sufficient condition on L^* so that Φ is locally strongly concave around $L = L^*$, i.e., the Hessian of Φ evaluated at $L = L^*$ is definite negative. All our results can also be rephrased in terms of Ψ.

3.2 Global maxima

Note that $\Phi(L)$ is, up to an additive constant that does not depend on L, the Kullback-Leibler (KL) divergence between $\mathsf{DPP}(L)$ and $\mathsf{DPP}(L^*)$:

$$\Phi(L) = \Phi(L^*) - \mathsf{KL}\left(\mathsf{DPP}(L^*), \mathsf{DPP}(L)\right), \forall L \in \mathcal{S}^{++}_{[N]}\,,$$

where KL stands for the Kullback-Leibler divergence between probability measures. In particular, by the properties of this divergence, $\Phi(L) \leqslant \Phi(L^*)$ for all $L \in \mathcal{S}^{++}_{[N]}$, and

$$\Phi(L) = \Phi(L^*) \iff \mathsf{DPP}(L) = \mathsf{DPP}(L^*) \iff L = DL^*D, \quad \text{for some } D \in \mathcal{D}.$$

As a consequence, the global maxima of Φ are exactly the matrices DL^*D, for D ranging in $\mathcal{D}$. The following theorem gives a more precise description of Φ around L^* (and, equivalently, around each DL^*D for $D \in \mathcal{D}$).

THEOREM 7. *Let $L^* \in \mathcal{S}^{++}_{[N]}$, $Z \sim \mathsf{DPP}(L^*)$ and $\Phi = \Phi_{L*}$, as defined in (3.3). Then, L^* is a critical point of Φ. Moreover, for any $H \in \mathcal{S}_{[N]}$,*

$$\mathrm{d}^2\Phi(L^*)(H,H) = -\operatorname{Var}[\operatorname{Tr}((L^*_Z)^{-1}H_Z)].$$

In particular, the Hessian $\mathrm{d}^2\Phi(L^*)$ is negative semidefinite.

The first part of this theorem is a consequence of the facts that L^* is a global maximum of a smooth Φ over the open parameter space $\mathcal{S}^{++}_{[N]}$. The second part of this theorem follows from the usual fact that the Fisher information matrix has two expressions: the opposite of the Hessian of the expected log-likelihood and the variance of the score (derivative of the expected log-likelihood). We also provide a purely algebraic proof of 7 in the appendix.

Our next result characterizes the null space of $d^2\Phi(L^*)$ in terms of the determinantal graph $\mathcal{G}_{L*}$.

THEOREM 8. *Under the same assumptions of Theorem 7, the null space of the quadratic Hessian map $H \in \mathcal{S}_{[N]} \mapsto \mathrm{d}^2\Phi(L^*)(H,H)$ is given by*

$$\mathcal{N}(L^*) = \left\{H \in \mathcal{S}_{[N]} \,:\, H_{i,j} = 0 \text{ for all } i,j \in [N] \text{ such that } \; i \sim_{L^*} j\right\}\,. \tag{3.5}$$

In particular, $\mathrm{d}^2\Phi(L^*)$ is definite negative if and only if L^* is irreducible.

The set $\mathcal{N}(L^*)$ has an interesting interpretation using perturbation analysis when L^* is reducible. On the one hand, since L^* is reducible, there exits $D_0 \in \mathcal{D}\backslash\{-I, I\}$ such that $L^* = D_0 L^* D_0$ is a global maximum for Φ_{L^*}. On the other hand, take any $H \in \mathcal{S}_{[N]}$ such that $L^* + H \in \mathcal{S}_{[N]}^{++}$ and observe that $D(L^* + H)D$ are all global maxima for Φ_{L^*+H} and in particular, $D_0(L^* + H)D_0$ is a global maximum for Φ_{L^*+H}. The Frobenius distance between L^* and $D_0(L^* + H)D_0$ is $\|H - D_0 H D_0\|_F$, which is maximized over H with fixed norm if and only if $D_0 H D_0 = -H$. Such matrices span precisely the null space $\mathcal{N}(L^*)$ (see Lemma 19). This leads to the following interpretation of $\mathcal{N}(L^*)$: The directions along which Φ_{L^*} has vanishing second derivative $L = L^*$ are spanned by the matrices H that push away any two merged modes of Φ_L^* as much as possible.

It follows from Theorem 8 that Φ_{L^*} is locally strongly concave around L^* if and only if L^* is irreducible since, in that case, the smallest eigenvalue of $-\mathrm{d}^2\Phi(L^*)$ is positive. Nevertheless, this positive eigenvalue may be exponentially small in N, leading to a small curvature around the maximum of Φ_{L^*}. This phenomenon is illustrated by the following example.

Consider the tridiagonal matrix L^* given by:

$$L^*_{i,j} = \begin{cases} a \text{ if } i = j, \\ b \text{ if } |i-j| = 1, \\ 0 \text{ otherwise,} \end{cases}$$

where a and b are real numbers.

PROPOSITION 9. *Assume that $a > 0$ and $a^2 > 2b^2$. Then, $L^* \in \mathcal{S}_{[N]}^{++}$ and there exist two positive numbers c_1 and c_2 that depend only on a and b such that*

$$0 < \inf_{H \in \mathcal{S}_{[N]} : \|H\|_F = 1} -\mathrm{d}^2\Phi(L^*)(H,H) \leqslant c_1 e^{-c_2 N}.$$

While the Hessian cancels in some directions $H \in \mathcal{N}(L^*)$ for any reducible $L^* \in \mathcal{S}_{[N]}^{++}$, the next theorem shows that the fourth derivative is negative in *any* nonzero direction $H \in \mathcal{N}(L^*)$ so that Φ is actually curved around L^* in any direction.

THEOREM 10. *Let $H \in \mathcal{N}(L^*)$. Then,*

(i) $\mathrm{d}^3\Phi(L^*)(H,H,H) = 0$;

(ii) $\mathrm{d}^4\Phi(L^*)(H,H,H,H) = -\dfrac{2}{3}\operatorname{Var}\left[\operatorname{Tr}\left(((L_Z^*)^{-1}H_Z)^2\right)\right] \leqslant 0$;

(iii) $\mathrm{d}^4\Phi(L^*)(H,H,H,H) = 0 \iff H = 0$.

The first part of Theorem 10 is obvious, since L^* is a global maximum of Φ. However, we give an algebraic proof of this fact, which is instructive for the proof of the two remaining parts of the theorem.

3.3 Other critical points

The function Φ_{L^*} is not concave and so finding its global maximum is fraught with difficulty. A standard approach is to work with a concave relaxation [CT04, CR09, ABH16], which has proven to be successful in applications such as compressed sensing, matrix completion and community detection. More recently, algorithms that attempt to directly optimize a non-concave objective have received growing attention, primarily driven by a good empirical performance and simple implementation (see [AGMM15, CLS15, BWY17] for example).

In fact, there are *two* issues that confound such approaches. The first is spurious local maxima where gradient ascent can get trapped. In some instances such as matrix completion [GLM16] it can be shown that the non-concave objective has no spurious local maxima, while in others such as Gaussian mixture models [JZB$^+$16], it does. The second issue is the presence of a large and often exponential number of saddle points. Empirically, it has been postulated [DPG$^+$14] that escaping saddle points is the main difficulty in optimizing large non-concave objectives. However if certain conditions on the saddle points are met then it is known that one can efficiently find a local maximum [NP06, GHJY15].

Here we show that the function Φ_{L^*} has exponentially many saddle points that correspond to all possible partial decouplings of the DPP.

THEOREM 11. *Let $L^* \in \mathcal{S}_{[N]}^{++}$ and $K^* = L^*(I + L^*)^{-1}$. Let $Z \sim \mathsf{DPP}(L^*)$. Then, the kernel L of any partial decoupling of Z is a critical point of Φ_{L^*}. Moreover, it is always a saddle point when the partial decoupling is strict.*

We conjecture that these are the only saddle points, which would be a major step in showing that despite the fact that Φ_{L^*} is non-concave, one can find its maximum via first and second order methods. This would give a compelling new example of a problem arising from big data where non-concave optimization problems can be tamed.

CONJECTURE 12. *Let $L^* \in \mathcal{S}_{[N]}^{++}$ and $Z \sim \mathsf{DPP}(L^*)$. The kernels of the partial decouplings of Z are the only critical points of Φ_{L^*}.*

The following proposition provides some evidence, by verifying a consequence of the conjecture:

PROPOSITION 13. *Let $L^* \in \mathcal{S}_{[N]}^{++}$ and let L be a critical point of Φ_{L^*}. Let $K^* = L^*(I + L^*)^{-1}$ and $K = L(I + L)^{-1}$. Then, K^* and K have the same diagonal.*

4. MAXIMUM LIKELIHOOD ESTIMATION

Let $Z_1, \ldots, Z_n$ be n independent copies of $Z \sim \mathsf{DPP}(L^*)$ with unknown kernel $L^* \in \mathcal{S}_{[N]}^{++}$. The maximum likelihood estimator (*MLE*) $\hat{L}$ of L^* is defined as a maximizer of the likelihood $\hat{\Phi}$ defined in (3.1). Since for all $L \in \mathcal{S}_{[N]}^{++}$ and all $D \in \mathcal{D}$, $\hat{\Phi}(L) = \hat{\Phi}(DLD)$, there is more than one kernel $\hat{L}$ that maximizes $\hat{\Phi}$ in general. We will abuse notation and refer to any such maximizer as *"the" MLE.* Since there is a bijection (2.3) between kernels L and correlation kernels K, the random correlation kernel $\hat{K} = \hat{L}(I + \hat{L})^{-1}$ maximizes the function $\hat{\Psi}$ defined in (3.2) and therefore, is the maximum likelihood estimator of the unknown correlation kernel $K^* = L^*(I + L^*)^{-1}$.

We measure the performance of the MLE using the *loss* ℓ defined by

$$\ell(\hat{L}, L^*) = \min_{D \in \mathcal{D}} \|\hat{L} - DL^*D\|_F$$

where we recall that $\|\cdot\|_F$ denotes the Frobenius norm.

The loss $\ell(\hat{L}, L^*)$ being a random quantity, we also define its associated *risk* $\mathcal{R}_n$ by

$$\mathcal{R}_n(\hat{L}, L^*) = \mathbb{E}\big[\ell(\hat{L}, L^*)\big],$$

where the expectation is taken with respect to the joint distribution of the iid observation $Z_1, \ldots, Z_n \sim \mathsf{DPP}(L^*)$.

Our first statistical result establishes that the MLE is a consistent estimator.

THEOREM 14.

$$\ell(\hat{L}, L^*) \xrightarrow[n\to\infty]{} 0\,, \qquad \textit{in probability.}$$

Theorem 14 shows that consistency of the MLE holds for all $L^* \in \mathcal{S}_{[N]}^{++}$. However, the MLE can be $\sqrt{n}$-consistent only when L^* is irreducible. Indeed, this is the only case when the Fisher information is invertible, by Theorem 8.

Let $M \in \mathcal{S}_{[N]}$ and Σ be a symmetric, positive definite bilinear form on $\mathcal{S}_{[N]}$. We write $A \sim \mathcal{N}_{\mathcal{S}_{[N]}}(M, \Sigma)$ to denote a Wigner random matrix $A \in \mathcal{S}_{[N]}$, such that for all $H \in \mathcal{S}_{[N]}$, $\mathrm{Tr}(AH)$ is a Gaussian random variable, with mean $\mathrm{Tr}(MH)$ and variance $\Sigma(H, H)$.

Assume that L^* is irreducible and let $\hat{L}$ be the MLE. Let $\hat{D} \in \mathcal{D}$ be such that

$$\|\hat{D}\hat{L}\hat{D} - L^*\|_F = \min_{D \in \mathcal{D}} \|D\hat{L}D - L^*\|_F$$

and set $\tilde{L} = \hat{D}\hat{L}\hat{D}$. Recall that by Theorem 8, the bilinear operator $\mathrm{d}^2\Phi(L^*)$ is invertible and let $V(L^*)$ be denote its inverse. Then, by Theorem 5.41 in [vdV98],

$$\sqrt{n}(\tilde{L} - L^*) = -V(L^*)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big((L^*_{Z_i})^{-1} - (I + L^*)^{-1}\big) + \rho_n, \tag{4.1}$$

where $\|\rho_n\|_F \xrightarrow[n\to\infty]{} 0$. Hence, we get the following theorem.

THEOREM 15. *Let L^* be irreducible. Then, $\tilde{L}$ is asymptotically normal, with asymptotic covariance operator $V(L^*)$:*

$$\sqrt{n}(\tilde{L} - L^*) \xrightarrow[n\to\infty]{} \mathcal{N}_{\mathcal{S}_{[N]}}\left(0, V(L^*)\right)\,,$$

where the above convergence holds in distribution.

Recall that we exhibited in Proposition 9 an irreducible kernel $L^* \in \mathcal{S}_{[N]}^{++}$ that is non-degenerate—its entries and eigenvalues are either zero or bounded away from zero—such that $V(L^*)[H, H] \geqslant c^N$ for some positive constant c and unit norm $H \in \mathcal{S}_{[N]}$. Together with Theorem 15, it implies that while the MLE $\tilde{L}$ converges at the parametric rate $n^{1/2}$, $\sqrt{n}\,\mathrm{Tr}[(\tilde{L}-L^*)^\top H]$ has asymptotic variance of order at least c^N for some positive constant c. It implies that the MLE suffers from a *curse of dimensionality.*

In the sequel, we say that an estimator $\hat{\theta}$ of an unknown quantity θ is n^α-consistent (for a given $\alpha > 0$) if the sequence $n^\alpha(\hat{\theta}-\theta)$ is bounded in probability. In particular, if the sequence $n^\alpha(\hat{\theta}-\theta)$ converges in distribution, then $\hat{\theta}$ is n^α-consistent.

When L^* is not irreducible, the MLE is no longer a $\sqrt{n}$-consistent estimator of L^*; it is only $n^{1/4}$-consistent. Nevertheless, in this case, the blocks of L^* may still be estimated at the parametric rate, as indicated by the following theorem.

If $A \in \mathbb{R}^{N\times N}$ and $J, J' \subset [N]$, we denote by $A_{J,J'}$ the $N \times N$ matrix whose entry (i,j) is $A_{i,j}$ if $(i,j) \in J \times J'$ and 0 otherwise. We have the following theorem.

THEOREM 16. *Let $L^* \in \mathcal{S}_{[N]}^{++}$ be block diagonal with blocks $\mathcal{P}$. Then, for $J, J' \in \mathcal{P}$, $J \neq J'$,*

$$\min_{D\in\mathcal{D}} \|\hat{L}_{J,J'} - DL^*_{J,J'}D\|_F = O_{\mathbb{P}}(n^{-1/4}) \tag{4.2}$$

and

$$\min_{D\in\mathcal{D}} \|\hat{L}_J - DL^*_J D\|_F = O_{\mathbb{P}}(n^{-1/2}). \tag{4.3}$$

Theorem 16 may also be stated in terms of K^* and its MLE $\hat{K} = \hat{L}(I+\hat{L})^{-1}$. In particular, the MLE $\hat{K}$ estimates the diagonal entries of K^* at the speed $n^{-1/2}$, no matter whether L^* (or, equivalently, K^*) is irreducible. Actually, it is possible to compute $\hat{K}_{j,j}$, for all $j \in [N]$: It is equal to the estimator of $K^*_{j,j}$ obtained by the method of moments. Indeed, recall that $\hat{L}$ satisfies the first order condition

$$\sum_{J\subset[N]} \hat{p}_J \hat{L}_J^{-1} = (I+\hat{L})^{-1}.$$

Post-multiplying by $\hat{L}$ both sides of this equality and identifying the diagonal entries yields

$$\hat{K}_{j,j} = \sum_{J\subset[N]:J\ni j} \hat{p}_J = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{j\in Z_i},$$

for all $j = 1,\ldots,N$. This is the estimator of $K^*_{j,j}$ obtained by the method of moments and it is $\sqrt{n}$-consistent by the central limit theorem.

5. CONCLUSION AND OPEN PROBLEMS

In this paper, we studied the local and global geometry of the log-likelihood function. We gave a nuanced treatment of the rates achievable by the maximum likelihood estimator and we establish when it can achieve parametric rates, and even when it cannot, which sets of parameters are the bottleneck. The main open question is to resolve Conjecture 12, which would complete our geometric picture of the log-likelihood function.

In a companion paper [BMRU17], using an approach based on the method of moments, we devise an efficiently computable estimator that converges at a parametric rate for any kernel whose underlying graph has bounded cycle sparsity. Moreover, we show an almost matching information-theoretic lower bound on the sample complexity as a function of the cycle sparsity. This work also indicates via minimax lower bounds that in the case of a cycle , not only the MLE suffers from asymptotically large asymptotic variance (Cf. Proposition 9) but actually *all* estimators suffer from this limitation.

6. PROOFS

6.1 A key determinantal identity and its consequences

We start this section by giving a key yet simple identity for determinants.

Lemma 17. *For all square matrices $L \in \mathbb{R}^{N\times N}$,*

$$\det(I+L) = \sum_{J\subset[N]} \det(L_J). \tag{6.1}$$

This identity is a direct consequence of the multilinearity of the determinant. Note that it gives the value of the normalizing constant in (2.2). Successive differentiations of (6.1) with respect to L lead to further useful identities. To that end, recall that if $f(L) = \log\det(L), L \in \mathcal{S}_{[N]}^{++}$, then for all $H \in \mathcal{S}_{[N]}$,

$$\mathrm{d}f(L)(H) = \mathrm{Tr}(L^{-1}H).$$

Differentiating (6.1) once over $L \in \mathcal{S}_{[N]}^{++}$ yields

$$\sum_{J\subset[N]} \det(L_J)\,\mathrm{Tr}(L_J^{-1}H_J) = \det(I+L)\,\mathrm{Tr}((I+L)^{-1}H), \quad \forall H \in \mathcal{S}_{[N]}. \tag{6.2}$$

In particular, after dividing by $\det(I+L)$,

$$\sum_{J\subset[N]} p_J(L)\,\mathrm{Tr}(L_J^{-1}H_J) = \mathrm{Tr}((I+L)^{-1}H), \quad \forall H \in \mathcal{S}_{[N]}. \tag{6.3}$$

In matrix form, (6.3) becomes

$$\sum_{J\subset[N]} p_J(L)L_J^{-1} = (I+L)^{-1}. \tag{6.4}$$

Here we use a slight abuse of notation. For $J \subset [N]$, L_J^{-1} (the inverse of L_J) has size $|J|$, but we still denote by L_J^{-1} the $N \times N$ matrix whose restriction to J is L_J^{-1} and which has zeros everywhere else.

Let us introduce some extra notation, for the sake of presentation. For any positive integer k and $J \subset [N]$, define

$$a_{J,k} = \mathrm{Tr}\left((L_J^{-1}H_J)^k\right) \quad \text{and} \quad a_k = \mathrm{Tr}\left(((I+L)^{-1}H)^k\right),$$

where we omit the dependency in $H \in \mathcal{S}_{[N]}$. Then, differentiating again (6.2) and rearranging terms yields

$$\sum_{J\subset[N]} p_J(L)a_{J,2} - a_2 = \sum_{J\subset[N]} p_J(L)a_{J,1}^2 - a_1^2, \tag{6.5}$$

for all $H \in \mathcal{S}_{[N]}$. In the same fashion, further differentiations yield

$$\begin{aligned}\sum_{J\subset[N]} p_J(L)a_{J,3} - a_3 = -\frac{1}{3}\Big(\sum_{J\subset[N]} p_J(L)a_{J,1}^3 - a_1^3\Big) + \frac{2}{3}\Big(\sum_{J\subset[N]} p_J(L)a_{J,2} - a_2\Big)\\ + \frac{1}{3}\Big(\sum_{J\subset[N]} p_J(L)a_{J,1}a_{J,2} - a_1a_2\Big)\end{aligned} \tag{6.6}$$

and

$$\begin{aligned}&\sum_{J\subset[N]} p_J(L)a_{J,4} - a_4\\ &\quad= \frac{1}{9}\Big(\sum_{J\subset[N]} p_J(L)a_{J,1}^4 - a_1^4\Big) - \frac{4}{9}\Big(\sum_{J\subset[N]} p_J(L)a_{J,1}^2a_{J,2} - a_1^2a_2\Big)\\ &\quad- \frac{2}{9}\Big(\sum_{J\subset[N]} p_J(L)a_{J,1}a_{J,2} - a_1a_2\Big) + \frac{5}{9}\Big(\sum_{J\subset[N]} p_J(L)a_{J,1}a_{J,3} - a_1a_3\Big)\\ &\quad+ \frac{1}{9}\Big(\sum_{J\subset[N]} p_J(L)a_{J,2}^2 - a_2^2\Big) + \frac{4}{9}\Big(\sum_{J\subset[N]} p_J(L)a_{J,3} - a_3\Big),\end{aligned} \tag{6.7}$$

for all $H \in \mathcal{S}_{[N]}$.

6.2 The derivatives of Φ

Let $L^* \in \mathcal{S}_{[N]}^{++}$ and $\Phi = \Phi_{L^*}$. In this section, we give the general formula for the derivatives of Φ.

LEMMA 18. *For all positive integers k and all $H \in \mathcal{S}_{[N]}$,*

$$\begin{aligned}&\mathrm{d}^k\Phi(L^*)(H,\ldots,H)\\ &\quad= (-1)^{k-1}(k-1)!\left(\sum_{J\subset[N]} p_J^* \operatorname{Tr}\left(((L_J^*)^{-1}H_J)^k\right) - \operatorname{Tr}\left(((I+L^*)^{-1}H)^k\right)\right).\end{aligned}$$

Proof

This lemma can be proven by induction, using the two following facts. If $f(M) = \log\det(M)$ and $g(M) = M^{-1}$ for $M \in \mathcal{S}_{[N]}^{++}$, then for all $M \in \mathcal{S}_{[N]}^{++}$ and $H \in \mathcal{S}_{[N]}$,

$$\mathrm{d}f(M)(H) = \operatorname{Tr}(M^{-1}H)$$

and

$$\mathrm{d}g(M)(H) = -M^{-1}HM^{-1}.$$

□

6.3 Auxiliary lemma

LEMMA 19. *Let $L^* \in \mathcal{S}_{[N]}^{++}$ and $\mathcal{N}(L^*)$ be defined as in (3.5). Let $H \in \mathcal{N}(L^*)$. Then, H can be decomposed as $H = H^{(1)} + \ldots + H^{(k)}$ where for each $j = 1,\ldots,k$, $H^{(j)} \in \mathcal{S}_{[N]}$ is such that $D^{(j)}H^{(j)}D^{(j)} = -H^{(j)}$, for some $D^{(j)} \in \mathcal{D}$ satisfying $D^{(j)}L^*D^{(j)} = L^*$.*

Proof. Let $H \in \mathcal{N}(L^*)$. Denote by $J_1, \ldots, J_M$ the blocks of L^* ($M = 1$ and $J_1 = [N]$ whenever L^* is irreducible). For $i = 1, \ldots, M$, let $D^{(i)} = \mathrm{Diag}(2\chi(J_i) - 1) \in \mathcal{D}$. Hence, $D^{(i)} L^* D^{(i)} = L^*$, for all $i = 1, \ldots, k$.

For $i, j \in [k]$ with $i < j$, define

$$H^{(i,j)} = \mathrm{Diag}(\chi(J_i)) H \, \mathrm{Diag}(\chi(J_j)) + \mathrm{Diag}(\chi(J_j)) H \, \mathrm{Diag}(\chi(J_i)) \, .$$

Then, it is clear that

$$H = \sum_{1 \leqslant i < j \leqslant M} H^{(i,j)} \qquad \text{and} \qquad D^{(i)} H^{(i,j)} D^{(i)} = -H^{(i,j)}, \ \forall \, i < j \, .$$

The lemma follows by renumbering the matrices $H^{(i,j)}$. □

6.4 Proof of Theorem 7

Theorem 7 is a direct consequence of Lemma 18 and identities (6.3) and (6.5). □

6.5 Proof of Theorem 8

Let $H \in \mathcal{S}_{[N]}$ be in the null space of $\mathrm{d}^2\Phi(L^*)$, i.e., satisfy $\mathrm{d}^2\Phi(L^*)(H, H) = 0$. We need to prove that $H_{i,j} = 0$ for all pairs $i, j \in [N]$ such that $i \sim_{L^*} j$. To that end, we proceed by (strong) induction on the distance between i and j in $\mathcal{G}_{L^*}$, i.e., the length of the shortest path from i to j (equal to ∞ if there is no such path). Denote this distance by $d(i, j)$.

First, by Theorem 7, $\mathrm{Var}[\mathrm{Tr}((L_Z^*)^{-1} H_Z)] = 0$ so the random variable $\mathrm{Tr}((L_Z^*)^{-1} H_Z)$ takes only one value with probability one. Therefore since $p_J^* > 0$ for all $J \subset [N]$ and $\mathrm{Tr}((L_\varnothing^*)^{-1} H_\varnothing) = 0$, we also have

$$\mathrm{Tr}([L_J^*]^{-1} H_J) = 0, \quad \forall J \subset [N]. \tag{6.8}$$

We now proceed to the induction.

If $d(i, j) = 0$, then $i = j$ and since L^* is definite positive, $L_{i,i}^* \neq 0$. Thus, using (6.8) with $J = \{i\}$, we get $H_{i,i} = 0$.

If $d(i, j) = 1$, then $L_{i,j}^* \neq 0$, yielding $H_{i,j} = 0$, using again (6.8), with $J = \{i, j\}$ and the fact that $H_{i,i} = H_{j,j} = 0$, established above.

Let now $m \geqslant 2$ be an integer and assume that for all pairs $(i, j) \in [N]^2$ satisfying $d(i, j) \leqslant m$, $H_{i,j} = 0$. Let $i, j \in [N]$ be a pair satisfying $d(i, j) = m + 1$. Let $(i, k_1, \ldots, k_m, j)$ be a shortest path from i to j in $\mathcal{G}_{L^*}$ and let $J = \{k_0, k_1, \ldots, k_m, k_{m+1}\}$, where $k_0 = i$ and $k_{m+1} = j$. Note that the graph $\mathcal{G}_{L_J^*}$ induced by L_J^* is a path graph and that for all $s, t = 0, \ldots, m + 1$ satisfying $|s - t| \leqslant m$, $d(k_s, k_t) = |s - t| \leqslant m$, yielding $H_{k_s, k_t} = 0$ by induction. Hence,

$$\mathrm{Tr}\left((L_J^*)^{-1} H_J\right) = 2 \left((L_J^*)^{-1}\right)_{i,j} H_{i,j} = 0, \tag{6.9}$$

by (6.8) with $J = \{i, j\}$. Let us show that $\left((L_J^*)^{-1}\right)_{i,j} \neq 0$, which will imply that $H_{i,j} = 0$. By writing $(L_J^*)^{-1}$ as the ratio between the adjugate of L_J^* and its determinant, we have

$$\left((L_J^*)^{-1}\right)_{i,j} = \frac{\det L_{J \setminus \{i\}, J \setminus \{j\}}}{\det L_J}, \tag{6.10}$$

where $L_{J\backslash\{i\},J\backslash\{j\}}$ is the submatrix of L_J obtained by deleting the i-th line and j-th column. The determinant of this matrix can be expanded as

$$\det L_{J\backslash\{i\},J\backslash\{j\}} = \sum_{\sigma\in\mathcal{M}_{i,j}} \varepsilon(\sigma) L^*_{i,\sigma(i)} L^*_{k_1,\sigma(k_1)} \dots L^*_{k_m,\sigma(k_m)}\,, \tag{6.11}$$

where $\mathcal{M}_{i,j}$ stands for the collection of all one-to-one maps from $J\backslash\{j\}$ to $J\backslash\{i\}$ and, for any such map σ, $\varepsilon(\sigma) \in \{-1,1\}$. There is only one term in (6.11) that is nonzero: Let $\sigma \in \mathcal{M}_{i,j}$ for which the product in (6.11) is nonzero. Recall that the graph induced by L^*_J is a path graph. Since $\sigma(i) \in J\backslash\{i\}$, $L^*_{i,\sigma(i)} = 0$ unless $\sigma(i) = k_1$. Then, $L^*_{k_1,\sigma(k_1)}$ is nonzero unless $\sigma(k_1) = k_1$ or k_2. Since we already have $\sigma(i) = k_1$ and σ is one-to-one, $\sigma(k_1) = k_2$. By induction, we show that $\sigma(k_s) = k_{s+1}$, for $s = 1,\dots,m-1$ and $\sigma(k_m) = j$. As a consequence, $\det L^*_{J\backslash\{i\},J\backslash\{j\}} \neq 0$ and, by (6.9) and (6.10), $H_{i,j} = 0$, which we wanted to prove.

Hence, by induction, we have shown that if $\mathrm{d}^2\Phi(L^*)(H,H) = 0$, then for any pair $i,j \in [N]$ such that $d(i,j)$ is finite, i.e., with $i \sim_{L^*} j$, $H_{i,j} = 0$.

Let us now prove the converse statement: Let $H \in \mathcal{S}_{[N]}$ satisfy $H_{i,j} = 0$, for all i,j with $i \sim_{L^*} j$. First, using Lemma 19 with its notation, for any $J \subset [N]$ and $j = 1,\dots,k$,

$$D_J^{(j)}(L_J^*)^{-1}D_J^{(j)} = \left(D_J^{(j)}L_J^*D_J^{(j)}\right)^{-1} = (L_J^*)^{-1}$$

and

$$D_J^{(j)}H_J^{(j)}D_J^{(j)} = -H_J^{(j)}\,.$$

Hence,

$$\operatorname{Tr}\left((L_J^*)^{-1}H_J^{(j)}\right) = \operatorname{Tr}\left(D^{(j)}(L_J^*)^{-1}D^{(j)}H_J^{(j)}\right) = -\operatorname{Tr}\left((L_J^*)^{-1}H_J^{(j)}\right) = 0\,.$$

Summing over $j = 1,\dots,k$ yields

$$\operatorname{Tr}\left((L_J^*)^{-1}H_J\right) = 0. \tag{6.12}$$

In a similar fashion,

$$\operatorname{Tr}\left((I+L)^{-1}H\right) = 0. \tag{6.13}$$

Hence, using (6.5),

$$\mathrm{d}^2\Phi(L^*)(H,H) = -\sum_{J\subset[N]} p_J^* \operatorname{Tr}^2\left((L_J^*)^{-1}H_J\right) + \operatorname{Tr}^2\left((I+L^*)^{-1}H\right) = 0,$$

which ends the proof of the theorem. □

6.6 Proof of Proposition 9

Consider the matrix $H \in \mathcal{S}_{[N]}$ with zeros everywhere but in positions $(1,N)$ and $(N,1)$, where its entries are 1. Note that $\operatorname{Tr}\left((L_J^*)^{-1}H_J\right)$ is zero for all $J \subset [N]$ such that $J \neq [N]$. This is trivial if J does not contain both 1 and N, since H_J will be the zero matrix. If J contains both 1 and N but does not contain the whole path that connects them in $\mathcal{G}_{L^*}$, i.e., if J does not contain the whole space $[N]$, then the subgraph $\mathcal{G}_{L^*_J}$ has at least two connected components, one containing 1 and another containing N. Hence, L^*_J is block diagonal, with 1 and N being in different blocks. Therefore, so is $(L^*_J)^{-1}$ and $\operatorname{Tr}\left((L^*_J)^{-1}H_J\right) = 2\left((L^*_J)^{-1}\right)_{1,N} = 0$.

Now, let $J = [N]$. Then,

$$\begin{aligned}\operatorname{Tr}\left((L^*_J)^{-1}H_J\right) &= 2\left((L^*)^{-1}\right)_{1,N} \\ &= 2(-1)^{N+1}\frac{\det(L^*_{[N]\setminus\{1\},[N]\setminus\{N\}})}{\det L^*} \\ &= 2(-1)^{N+1}\frac{b^{N-1}}{\det L^*}. \qquad (6.14)\end{aligned}$$

Write $\det L^* = u_N$ and observe that

$$u_k = au_{k-1} + b^2 u_{k-2}, \quad \forall k \geqslant 2$$

and $u_1 = a, u_2 = a^2 - b^2$. Since $a^2 > 4b^2$, there exists $\mu > 0$ such that

$$u_k \geqslant \mu\left(\frac{a+\sqrt{a^2-4b^2}}{2}\right)^k, \quad \forall k \geqslant 1. \qquad (6.15)$$

Hence, (6.14) yields

$$\left|\operatorname{Tr}\left((L^*_J)^{-1}H_J\right)\right| \leqslant \frac{2}{\mu|b|}\left(\frac{2|b|}{a+\sqrt{a^2-4b^2}}\right)^N,$$

which proves the second part of Proposition 9, since $a + \sqrt{a^2 - 4b^2} > a > 2|b|$.

Finally note that (6.15) implies that all the principal minors of L^* are positive so that $L \in \mathcal{S}^{++}_{[N]}$. □

6.7 Proof of Theorem 10

Let $H \in \mathcal{N}(L^*)$. By Lemma 18, the third derivative of Φ at L^* is given by

$$\mathrm{d}^3\Phi(L^*)(H,H,H) = 2\sum_{J\subset[N]} p^*_J \operatorname{Tr}\left(((L^*_J)^{-1}H_J)^3\right) - 2\operatorname{Tr}\left(((I+L^*)^{-1}H)^3\right).$$

Together with (6.6), it yields

$$\begin{aligned}\mathrm{d}^3\Phi(L^*)(H,H,H) = -\frac{2}{3}\Big(\sum_{J\subset[N]} p_J(L)a^3_{J,1} - a^3_1\Big) + \frac{4}{3}\Big(\sum_{J\subset[N]} p_J(L)a_{J,2} - a_2\Big) \\ + \frac{2}{3}\Big(\sum_{J\subset[N]} p_J(L)a_{J,1}a_{J,2} - a_1 a_2\Big).\end{aligned}$$

Each of the three terms on the right hand side of the above display vanish because of (6.12), $H \in \mathcal{N}(L^*)$ and (6.13) respectively. This concludes the proof of *(i)*.

Next, the fourth derivative of Φ at L^* is given by

$$\mathrm{d}^4\Phi(L^*)(H,H,H,H) = -6\sum_{J\subset[N]} p^*_J \operatorname{Tr}\left(((L^*_J)^{-1}H_J)^4\right) + 6\operatorname{Tr}\left(((I+L^*)^{-1}H)^4\right).$$

Using (6.7) together with (6.12), (6.13) and $\mathrm{d}^3\Phi(L^*)(H,H,H)=0$, it yields

$$\mathrm{d}^4\Phi(L^*)(H,H,H,H) = -\frac{2}{3}\Big(\sum_{J\subset[N]} p_J^* \operatorname{Tr}^2\left((L_J^*)^{-1}H_J)^2\right) - \operatorname{Tr}^2\left(((I+L^*)^{-1}H)^2\right)\Big).$$

Since $H \in \mathcal{N}(L^*)$, meaning $\mathrm{d}^2\Phi(L^*)(H,H)=0$, we also have

$$\operatorname{Tr}\left(((I+L^*)^{-1}H)^2\right) = \sum_{J\subset[N]} p_J^* \operatorname{Tr}\left((L_J^*)^{-1}H_J)^2\right).$$

Hence, we can rewrite $\mathrm{d}^4\Phi(L^*)(H,H,H,H)$ as

$$\mathrm{d}^4\Phi(L^*)(H,H,H,H) = -\frac{2}{3}\left(\mathbb{E}\left[\operatorname{Tr}^2\left((L_Z^*)^{-1}H_Z)^2\right)\right] - \mathbb{E}\left[\operatorname{Tr}\left((L_Z^*)^{-1}H_Z)^2\right)\right]^2\right).$$

This concludes the proof of *(ii)*.

To prove *(iii)*, note first that if $H=0$ then trivially $\mathrm{d}^4\Phi(L^*)(H,H,H,H)=0$. Assume now that $\mathrm{d}^4\Phi(L^*)(H,H,H,H)=0$, which, in view of *(ii)* is equivalent to $\operatorname{Var}[\operatorname{Tr}(((L_Z^*)^{-1}H_Z)^2)]=0$. Since $\operatorname{Tr}(((L_\varnothing^*)^{-1}H_\varnothing)^2)=0$, and $p_J^*>0$ for all $J\subset[N]$, it yields

$$\operatorname{Tr}(((L_J^*)^{-1}H_J)^2) = 0 \quad \forall\, J\subset[N]. \tag{6.16}$$

Fix $i,j\in[N]$. If i and j are in one and the same block of L^*, we know by Theorem 8 that $H_{i,j}=0$. On the other hand, suppose that i and j are in different blocks of L^* and let $J=\{i,j\}$. Denote by $h=H_{i,j}=H_{j,i}$. Since L_J^* is a 2×2 diagonal matrix with nonzero diagonal entries and $H_{i,i}=H_{j,j}=0$, (6.16) readily yields $h=0$. Hence, $H=0$, which completes the proof of *(iii)*. □

6.8 Proof of Theorem 11

Denote by $\Phi=\Phi_{L^*}$ and $K^*=L^*(I+L^*)^{-1}$. Let L be the kernel of a partial decoupling of Z according to a partition $\mathcal{P}$ of $[N]$. By definition, the correlation kernel $K=L(I+L)^{-1}$ is block diagonal, with blocks $K_J=D_JK_J^*D_J$, $J\in\mathcal{P}$, for some matrix $D\in\mathcal{D}$. Without loss of generality, assume that $D=I$. Since $L=K(I-K)^{-1}$, L is also block diagonal, with blocks $L_J=K_J^*(I_J-K_J^*)^{-1}$, $J\in\mathcal{P}$. To see that L is a critical point of Φ, note that the first derivative of Φ can be written in matrix form as

$$\mathrm{d}\Phi(L) = \sum_{J'\subset[N]} p_{J'}^* L_{J'}^{-1} - (I+L)^{-1}, \tag{6.17}$$

where $L_{J'}^{-1}$ stands for the $N\times N$ matrix with the inverse of $L_{J'}$ on block J' and zeros everywhere else. Note that since L is block diagonal, so are each of the terms of the right-hand side of (6.17), with the same blocks. Hence, it is enough to prove that for all $J\in\mathcal{P}$, the block J of $\mathrm{d}\Phi(L)$ (i.e., $(\mathrm{d}\Phi(L))_J$) is zero. Using elementary block matrix operations, for all $J\subset[N]$, the block J of $L_{J'}^{-1}$ is given by $L_{J\cap J'}^{-1}$, using the same abuse of notation as before. Hence, the block J of $\mathrm{d}\Phi(L)$ is given by

$$(\mathrm{d}\Phi(L))_J = \sum_{J'\subset[N]} p_{J'}^* L_{J'\cap J}^{-1} - (I_J+L_J)^{-1},$$

which can also be written as

$$(\mathrm{d}\Phi(L))_J = \sum_{J' \subset J} \tilde{p}^*_{J'} L_{J'}^{-1} - (I_J + L_J)^{-1}, \tag{6.18}$$

where

$$\tilde{p}^*_{J'} = \sum_{J'' \subset \bar{J}} p^*_{J' \cup J''} = \sum_{J'' \subset \bar{J}} \mathbb{P}\left[Z = J' \cup J''\right] = \mathbb{P}\left[Z \cap J = J'\right]. \tag{6.19}$$

Recall that $Z \cap J$ is a DPP on J with correlation kernel K^*_J. Hence, its kernel is L_J and (6.19) yields

$$\tilde{p}^*_{J'} = p_{J'}(L_J).$$

Together with (6.18), it yields

$$(\mathrm{d}\Phi(L))_J = \mathrm{d}\Phi_{L_J}(L_J),$$

which is zero by Theorem 7. This proves that L is a critical point of Φ.

Next, we prove that if L is the kernel of a strict partial decoupling of Z, then it is a saddle point of Φ. To that end, we exhibit two matrices $H, H' \in \mathcal{S}_{[N]}$ such that $\mathrm{d}^2\Phi(L)(H,H) > 0$ and a $\mathrm{d}^2\Phi(L)(H',H') < 0$.

Consider a strict partial decoupling of Z according to a partition $\mathcal{P}$. Let L and K be its kernel and correlation kernel, respectively. In particular, there exists $J \in \mathcal{P}$, $i \in J$ and $j \in \bar{J}$ such that $K^*_{i,j} \neq 0$. Consider the matrix H with zeros everywhere but in positions (i,j) and (j,i), where its entries are 1. By simple matrix algebra,

$$\begin{aligned} &\mathrm{d}^2\Phi(L)(H,H) \\ &\quad = -\sum_{J' \subset [N]} p^*_{J'} \operatorname{Tr}\left((L_{J'}^{-1} H_{J'})^2\right) + \operatorname{Tr}\left(((I+L)^{-1}H)^2\right) \\ &\quad = -2 \sum_{J' \subset [N]} p^*_{J'} \left(L_{J' \cap J}^{-1}\right)_{i,i} \left(L_{J' \cap \bar{J}}^{-1}\right)_{j,j} + 2\left((I+L)^{-1}\right)_{i,i} \left((I+L)^{-1}\right)_{j,j}, \end{aligned} \tag{6.20}$$

where we recall that for all $J' \subset [N]$ and $k \in [N]$, $(L_{J'}^{-1})_{k,k}$ is set to zero if $k \notin J'$.

Denote by $Y_i = (L_{Z \cap J}^{-1})_{i,i}$ and $Y_j = (L_{Z \cap \bar{J}}^{-1})_{j,j}$. Note that $\mathbb{E}[Y_i] = \left((I+L)^{-1}\right)_{i,i}$. Indeed,

$$\begin{aligned} \mathbb{E}[Y_i] &= \sum_{J' \subset [N]} p^*_{J'} \left(L_{J' \cap J}^{-1}\right)_{i,i} = \sum_{J' \subset J} \sum_{J'' \subset \bar{J}} p^*_{J' \cup J''} \left(L_{J'}^{-1}\right)_{i,i} \\ &= \sum_{J' \subset J} \mathbb{P}[Z \cap J = J'](L_{J'}^{-1})_{i,i} = \sum_{J' \subset J} p_{J'}(L_J) \left(L_{J'}^{-1}\right)_{i,i} \\ &= (I_J + L_J)^{-1}_{i,i} = (I+L)^{-1}_{i,i}. \end{aligned}$$

Here, the third equality follows from the fact that L_J is the kernel of the DPP $Z \cap J$, the fourth equality follows from (6.4) and the last equality comes from the block diagonal structure of L. It can be checked using the same argument that $\mathbb{E}[Y_j] = \left((I+L)^{-1}\right)_{j,j}$. Together with (6.20), it yields

$$\mathrm{d}^2\Phi(L)(H,H) = -2\mathbb{E}[Y_i Y_j] + 2\mathbb{E}[Y_i]\mathbb{E}[Y_j]. \tag{6.21}$$

Next, recall that $(X_1, \dots, X_N) = \chi(Z)$ denotes the characteristic vector of Z and observe that $Y_iY_j = 0$ whenever $X_j = 0$ or $X_j = 0$ so that $Y_iY_j = Y_iY_jX_iX_j$. Hence,

$$\mathbb{E}[Y_iY_j] = \mathbb{E}[Y_iY_j|X_i = 1, X_j = 1]\mathbb{P}[X_i = 1, X_j = 1] .$$

Since $L \in \mathcal{S}^{++}_{[N]}$, we have $\mathbb{E}[Y_iY_j] > 0$, yielding $\mathbb{E}[Y_iY_j|X_i = 1, X_j = 1] > 0$ by the previous equality. Moreover,

$$\mathbb{P}[X_i = 1, X_j = 1] = K^*_{i,i}K^*_{j,j} - (K^*_{i,j})^2 < K^*_{i,i}K^*_{j,j} = \mathbb{P}[X_i = 1]\mathbb{P}[X_j = 1] ,$$

where the inequality follows from the assumption $K^*_{i,j} \neq 0$. Hence,

$$\mathbb{E}[Y_iY_j] < \mathbb{E}[Y_iY_j|X_i = 1, X_j = 1]\mathbb{P}[X_i = 1]\mathbb{P}[X_j = 1] . \tag{6.22}$$

We now use conditional negative association. To that end, we check that $Y_i = f_i(\chi(Z \cap J))$ and $Y_j = f_j(\chi(Z \cap \bar{J}))$, for some non decreasing functions f_i and f_j. For any $J' \subset J$, define $f_i(J') = (L^{-1}_{J'})_{i,i}$. It is sufficient to check that

$$(L^{-1}_{J'})_{i,i} \leqslant (L^{-1}_{J' \cup \{k\}})_{i,i} , \quad \forall\, k \in J \backslash J' \tag{6.23}$$

First, note that (6.23) is true if $i \notin J'$, since in this case, $(L^{-1}_{J'})_{i,i} = 0$ and $(L^{-1}_{J' \cup \{k\}})_{i,i} \geqslant 0$. Assume now that $i \in J'$ and consider the matrix $L_{J' \cup \{k\}}$, of which $L_{J'}$ is a submatrix. Using the Schur complement, we get that

$$\big(L^{-1}_{J' \cup \{k\}}\big)_{J'} = \big(L_{J'} - \frac{1}{L_{k,k}} AA^\top\big)^{-1} , \tag{6.24}$$

where $A = L_{J',\{k\}}$. Since $L_{k,k} > 0$ and $AA^\top$ is positive semidefinite, then

$$L_{J'} - \frac{1}{L_{k,k}} AA^\top \preceq L_{J'} ,$$

where $\preceq$ denotes the Löwner order on $\mathcal{S}^+_{[N]}$. Moreover, it follows from the Löwner-Heinz theorem that if $A \preceq B$, then $B^{-1} \preceq A^{-1}$ for any nonsingular $A, B \in \mathcal{S}_{[N]}$. Therefore,

$$L^{-1}_{J'} \preceq \big(L_{J'} - \frac{1}{L_{k,k}} AA^\top\big)^{-1} .$$

In particular, the above display yields, together with (6.24),

$$\big(L^{-1}_{J'}\big)_{i,i} \preceq \big(\big(L_{J'} - \frac{1}{L_{k,k}} AA^\top\big)^{-1}\big)_{i,i} = \big(L^{-1}_{J' \cup \{k\}}\big)_{(i,i)} .$$

This completes the proof of (6.23) and monotonicity of f_j follows from the same arguments.

We are now in a position to use the conditional negative association property from Lemma 2. Together with (6.22), it yields

$$\mathbb{E}[Y_iY_j] < \mathbb{E}[Y_i|X_i = 1, X_j = 1]\mathbb{E}[Y_2|X_i = 1, X_j = 1]\mathbb{P}[X_i = 1]\mathbb{P}[X_j = 1] . \tag{6.25}$$

Next, note that

$$\mathbb{E}[Y_i|X_i = 1, X_j = 1] \leqslant \mathbb{E}[Y_i|X_i = 1] ,$$

and

$$\mathbb{E}[Y_j|X_i = 1, X_j = 1] \leqslant \mathbb{E}[Y_j|X_j = 1]\,.$$

These inequalities are also a consequence of the conditional negative association property. Indeed, using Bayes formula and the fact that $j \notin J$ respectively, we get

$$\begin{aligned}\mathbb{E}[Y_i|X_i = 1, X_j = 1] &= \frac{\mathbb{E}[Y_i X_j|X_i = 1]}{\mathbb{E}[X_j|X_i = 1]} \\ &\leqslant \frac{\mathbb{E}[Y_i|X_i = 1]\mathbb{E}[X_j|X_i = 1]}{\mathbb{E}[X_j|X_i = 1]} = \mathbb{E}[Y_i|X_i = 1]\,.\end{aligned}$$

The second inequality follows from the same argument and the fact that $i \notin \bar{J}$. Finally, (6.25) becomes

$$\mathbb{E}[Y_i Y_j] < \mathbb{E}[Y_i]\mathbb{E}[Y_j]$$

and hence, (6.21) yields that $\mathrm{d}^2\Phi(L)(H, H) > 0$.

We now exhibit H' such that $\mathrm{d}^2\Phi(L)(H, H) < 0$. To that end, let H' be the matrix with zeros everywhere but in position $(1, 1)$, where $H'_{1,1} = 1$. Let J be the element of $\mathcal{P}$ that contains 1. By simple matrix algebra,

$$\begin{aligned}\mathrm{d}^2\Phi(L)(H', H') &= -\sum_{J' \subset [N]} p^*_{J'} \big(L_{J'}^{-1}\big)^2_{1,1} + \big((I + L)^{-1}\big)^2_{i,i} \\ &= -\sum_{J' \subset J} \sum_{J'' \subset \bar{J}} p^*_{J' \cup J''} \big(L_{J'}^{-1}\big)^2_{1,1} + \big((I + L)^{-1}\big)^2_{i,i} \\ &= -\sum_{J' \subset J} \Big(\sum_{J'' \subset \bar{J}} p^*_{J' \cup J''}\Big) \big(L_{J'}^{-1}\big)^2_{1,1} + \big((I_J + L_J)^{-1}\big)^2_{i,i} \\ &= -\sum_{J'' \subset J} p_{J'}(L_J) \big(L_{J'}^{-1}\big)^2_{1,1} + \big((I + L)^{-1}\big)^2_{i,i} \\ &= \mathrm{d}^2\Phi_{L_J}(H'_J, H'_J). \qquad (6.26)\end{aligned}$$

By Theorem 7, $\mathrm{d}^2\Phi_{L_J}(H'_J, H'_J) \leqslant 0$. In addition, by Theorem 8, $\mathrm{d}^2\Phi_{L_J}(H'_J, H'_J) \neq 0$ since H'_J has at least one nonzero diagonal entry. Hence, $\mathrm{d}^2\Phi_{L_J}(H'_J, H'_J) < 0$ and it follows from (6.26) that $\mathrm{d}^2\Phi(L)(H', H') < 0$, which completes the proof of Theorem 11. □

6.9 Proof of Proposition 13

Let L be a critical point of Φ and $K = L(I + L)^{-1}$. Then, for all $N \times N$ matrices H,

$$\mathrm{d}\Phi(L)(H) = \sum_{J \subset [N]} p^*_J \operatorname{Tr}\big(L_J^{-1} H_J\big) - \operatorname{Tr}\big((I + L)^{-1} H\big) = 0.$$

Fix $t_1, \ldots, t_N \in \mathbb{R}$ and define $T = \mathrm{Diag}(t_1, \ldots, t_N)$, $H = LT$. Then, since T is diagonal, $H_J = L_J T_J$, for all $J \subset [N]$. Using the above equation and the fact that L and $(I + L)^{-1}$ commute, we have

$$\sum_{J \subset [N]} p^*_J \sum_{j \in J} t_j = \operatorname{Tr}(KT) = \sum_{j=1}^{N} K_{j,j} t_j. \qquad (6.27)$$

Since (6.27) holds for any $t_1, \dots, t_N \in \mathbb{R}$, we conclude that

$$K_{j,j} = \sum_{J \subset [N]: J \ni j} p_J^* = K_{j,j}^*,$$

for all $j \in [N]$, which ends the proof. □

6.10 Proof of Theorem 14

Our proof is based on Theorem 5.14 in [vdV98]. We need to prove that there exists a compact subset E of $\mathcal{S}_{[N]}^{++}$ such that $\hat{L} \in E$ eventually almost surely. Fix $\alpha, \beta \in (0,1)$ to be chosen later such that $\alpha < \beta$ and define the compact set of $\mathcal{S}_{[N]}^{++}$ as

$$E_{\alpha,\beta} = \left\{ L \in \mathcal{S}_{[N]}^{++} \, : \, K = L(I+L)^{-1} \in \mathcal{S}_{[N]}^{[\alpha,\beta]} \right\}.$$

Let $\delta = \min_{J \subset [N]} p_J^*$. Since L^* is definite positive, $\delta > 0$. Define the event $\mathcal{A}$ by

$$\mathcal{A} = \bigcap_{J \subset [N]} \left\{ p_J^* \leqslant 2\hat{p}_J \leqslant 3p_J^* \right\}.$$

and observe that on $\mathcal{A}$, we have $3\Phi(L) \leqslant 2\hat{\Phi}(L) \leqslant \Phi(L)$ simultaneously for all $L \in \mathcal{S}_{[N]}^{++}$. In particular,

$$\Phi(\hat{L}) \geqslant 2\hat{\Phi}(\hat{L}) \geqslant 2\hat{\Phi}(L^*) \geqslant 3\Phi(L^*), \tag{6.28}$$

where the second inequality follows from the definition of the MLE.

Using Hoeffding's inequality together with a union bound, we get

$$\mathbb{P}[\mathcal{A}] \geqslant 1 - 2^{N+1} e^{-\delta^2 n/2}. \tag{6.29}$$

Observe that $\Phi(L^*) < 0$, so we can define $\alpha < \exp(3\Phi(L^*)/\delta)$ and $\beta > 1 - \exp(3\Phi(L^*)/\delta)$ such that $0 < \alpha < \beta < 1$. Let $L \in \mathcal{S}_{[N]}^{++} \backslash E_{\alpha,\beta}$ and $K = L(I+L)^{-1}$. Then, either *(i)* K has an eigenvalue that is less than α, or *(ii)* K has an eigenvalue that is larger than β. Since all the eigenvalues of K lie in $(0,1)$, we have that $\det(K) \leqslant \alpha$ in case *(i)* and $\det(I-K) \leqslant 1 - beta$ in case *(ii)*. Recall that

$$\Phi(L) = \sum_{J \subset [N]} p_J^* \log|\det(K - I_{\bar{J}})|, ,$$

and observe that each term in this sum is negative. Hence, by definition of α and β,

$$\Phi(L) \leqslant \begin{cases} p_{[N]}^* \log\alpha \leqslant \delta\log\alpha < 3\Phi(L^*) \leqslant \Phi(\hat{L}) & \text{in case } \textit{(i)} \\ p_{\varnothing}^* \log(1-\beta) \leqslant \delta\log(1-\beta) < 3\Phi(L^*) \leqslant \Phi(\hat{L}) & \text{in case } \textit{(ii)} \end{cases}$$

using (6.28). Thus, on $\mathcal{A}$, $\Phi(L) < \Phi(\hat{L})$ for all $L \in \mathcal{S}_{[N]}^{++} \backslash E_{\alpha,\beta}$. It yields that on this event, $\hat{L} \in E_{\alpha,\beta}$.

Now, let $\varepsilon > 0$. For all $J \subset [N]$, $p_J(\cdot)$ is a continuous function; hence, we can apply Theorem 5.14 in [vdV98], with the compact set $E_{\alpha,\beta}$. This yields

$$\begin{aligned} \mathbb{P}[\ell(\hat{L}, L^*) > \varepsilon] &\leqslant \mathbb{P}[\ell(\hat{L}, L^*) > \varepsilon, \hat{L} \in E_{\alpha,\beta}] + \mathbb{P}[\hat{L} \notin E_{\alpha,\beta}] \\ &\leqslant \mathbb{P}[\ell(\hat{L}, L^*) > \varepsilon, \hat{L} \in E_{\alpha,\beta}] + (1 - \mathbb{P}[\mathcal{A}]). \end{aligned}$$

Using Theorem 5.14 in [vdV98], the first term goes to zero, and the second term goes to zero by (6.29). This ends the proof of Theorem 14. □

6.11 Proof of Theorem 16

The first statement of Theorem 16 follows from Theorem 5.52 in [vdV98], with $\alpha = 4$ and $\beta = 2$. For the second statement, note that since the DPPs $Z \cap J, J \in \mathcal{P}$ are independent, each $\hat{L}_J, J \in \mathcal{P}$ is the maximum likelihood estimator of L^*_J. Since L^*_J is irreducible, the $n^{1/2}$-consistency of $\hat{L}_J$ follows from Theorem 15. □

REFERENCES

[ABH16] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, Jan 2016.

[AFAT14] R. H. Affandi, E. B. Fox, R. P. Adams, and B. Taskar. 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*, pages 1224–1232, 2014.

[AGMM15] S. Arora, R. Ge, T. Ma, and A. Moitra. Simple, efficient, and neural algorithms for sparse coding. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, pages 113–149, 2015.

[Bar13] Y. Baraud. Estimation of the density of a determinantal process. *Confluentes Math.*, 5(1):3–21, 2013.

[BBL09] J. Borcea, P. Brändén, and T. M. Liggett. Negative dependence and the geometry of polynomials. *J. Amer. Math. Soc.*, 22(2):521–567, 2009.

[BC16] C. A. N. Biscio and J.-F. Coeurjolly. Standard and robust intensity parameter estimation for stationary determinantal point processes. *Spat. Stat.*, 18(part A):24–39, 2016.

[BH16] R. Bardenet and A. Hardy. Monte Carlo with Determinantal Point Processes. working paper or preprint, May 2016.

[BL16] C. A. N. Biscio and F. Lavancier. Contrast estimation for parametric stationary determinantal point processes. *Scandinavian Journal of Statistics*, 2016.

[BMRU17] V.-E. Brunel, A. Moitra, P. Rigollet, and J. Urschel. Learning determinantal point processes with moments and cycles. ArXiv:1703.00539, 2017.

[BO00] A. Borodin and G. Olshanski. Distributions on partitions, point processes, and the hypergeometric kernel. *Comm. Math. Phys.*, 211(2):335–358, 2000.

[Bor11] A. Borodin. Determinantal point processes. In *The Oxford handbook of random matrix theory*, pages 231–249. Oxford Univ. Press, Oxford, 2011.

[BQK$^+$14] N. K. Batmanghelich, G. Quon, A. Kulesza, M. Kellis, P. Golland, and L. Bornn. Diversifying sparsity using variational determinantal point processes. *CoRR*, abs/1411.6307, 2014.

[BS03] A. Borodin and A. Soshnikov. Janossy densities. i. determinantal ensembles. *Journal of Statistical Physics*, 113(3):595–610, 2003.

[BTRA15] R. Bardenet and M. Titsias RC AUEB. Inference for determinantal point processes without spectral knowledge. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Ad-*

vances in Neural Information Processing Systems 28, pages 3393–3401. Curran Associates, Inc., 2015.

[BWY17] S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Ann. Statist. (to appear)*, 2017.

[CJM16] A. Chambaz, E. Joly, and X. Mary. Survey sampling targeted inference. Preprint HAL, August 2016.

[CLS15] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Trans. Information Theory*, 61(4):1985–2007, 2015.

[CR09] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[CT04] E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Information Theory*, 51:4203 – 4215, 2004.

[DB16] C. Dupuy and F. Bach. Learning determinantal point processes in sublinear time. arXiv:1610.05925, 2016.

[DPG$^+$14] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 2933–2941, Cambridge, MA, USA, 2014. MIT Press.

[Dys62] F. J. Dyson. Statistical theory of the energy levels of complex systems. III. *J. Mathematical Phys.*, 3:166–175, 1962.

[DZH15] N. Deng, W. Zhou, and M. Haenggi. The ginibre point process as a model for wireless networks with repulsion. *IEEE Trans. Wireless Communications*, 14(1):107–121, 2015.

[GHJY15] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, pages 797–842, 2015.

[GKFT14] J. Gillenwater, A. Kulesza, E. Fox, and B. Taskar. Expectation-maximization for learning determinantal point processes. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 3149–3157, Cambridge, MA, USA, 2014. MIT Press.

[GLM16] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2973–2981. Curran Associates, Inc., 2016.

[GPK16a] M. Gartrell, U. Paquet, and N. Koenigstein. Bayesian low-rank determinantal point processes. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 349–356, New York, NY, USA, 2016. ACM.

[GPK16b] M. Gartrell, U. Paquet, and N. Koenigstein. Low-rank factorization of determinantal point processes for recommendation. arXiv:1602.05436, 2016.

[JZB$^+$16] C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4116–4124. Curran Associates, Inc., 2016.

[KK16] M. Kojima and F. Komaki. Determinantal point process priors for Bayesian variable selection in linear regression. *Statist. Sinica*, 26(1):97–117, 2016.

[KT11] A. Kulesza and B. Taskar. 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*, pages 1193–1200, 2011.

[KT12] A. Kulesza and B. Taskar. *Determinantal Point Processes for Machine Learning.* Now Publishers Inc., Hanover, MA, USA, 2012.

[Kul12] A. Kulesza. *Learning with determinantal point processes.* PhD thesis, University of Pennsylvania, 2012.

[LB12] H. Lin and J. A. Bilmes. 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*, pages 479–490, 2012.

[LBDA15] Y. Li, F. Baccelli, H. S. Dhillon, and J. G. Andrews. Statistical modeling and probabilistic analysis of cellular networks with determinantal point processes. *IEEE Trans. Communications*, 63(9):3405–3422, 2015.

[LCYO16] D. Lee, G. Cha, M. Yang, and S. Oh. 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*, pages 330–346, 2016.

[LM15] V. Loonis and X. Mary. Determinantal Sampling Designs. ArXiv:1510.06618, 2015.

[LMR15] F. Lavancier, J. Møller, and E. Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):853–877, 2015.

[Mac75] O. Macchi. The coincidence approach to stochastic point processes. *Advances in Appl. Probability*, 7:83–122, 1975.

[MS14] N. Miyoshi and T. Shirai. *Cellular Networks with α-Ginibre Configurated Base Stations*, pages 211–226. Springer Japan, Tokyo, 2014.

[MS15] Z. Mariet and S. Sra. Fixed-point algorithms for learning determinantal point processes. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2389–2397, 2015.

[MS16] Z. E. Mariet and S. Sra. Kronecker determinantal point processes. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2694–2702. Curran Associates, Inc., 2016.

[NP06] Y. Nesterov and B. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[Oko01] A. Okounkov. Infinite wedge and random partitions. *Selecta Math.*

(N.S.), 7(1):57–81, 2001.

[OR03] A. Okounkov and N. Reshetikhin. Correlation function of Schur process with application to local geometry of a random 3-dimensional Young diagram. *J. Amer. Math. Soc.*, 16(3):581–603 (electronic), 2003.

[SZA13] J. Snoek, R. S. Zemel, and R. P. Adams. 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.*, pages 1932–1940, 2013.

[TL14] G. L. Torrisi and E. Leonardi. Large deviations of the interference in the ginibre network model. *Stoch. Syst.*, 4(1):173–205, 2014.

[vdV98] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[XO16] H. Xu and H. Ou. 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.

[YFZ$^+$16] J. Yao, F. Fan, W. X. Zhao, X. Wan, E. Y. Chang, and J. Xiao. Tweet timeline generation with determinantal point processes. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 3080–3086, 2016.

Victor-Emmanuel Brunel
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Cambridge, MA 02139-4307, USA
(vebrunel@math.mit.edu)

Philippe Rigollet
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Cambridge, MA 02139-4307, USA
(rigollet@math.mit.edu)

Ankur Moitra
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Cambridge, MA 02139-4307, USA
(moitra@mit.edu)

John Urschel
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Cambridge, MA 02139-4307, USA
(urschel@mit.edu)