# Determinantal point processes and their parameter estimations

by

Haiyi Shi

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Mathematics

Department of Mathematical and Statistical Sciences
University of Alberta

# Abstract

Determinantal point processes (DPPs) arise as important tools in various aspects of mathematics, such as stochastic processes, random matrices, and combinatorics. Over the last decade, DPPs have also been widely used in machine learning community; they are especially popular in subset selection problems, for they favour subsets of high quality and diversity. These applications motivate studies in parameter estimations, of which a common method is maximum likelihood estimation. In 2017, Brunel et al first studied this non-convex optimization problem using an information geometric approach. Inspired by their work, we introduce and extend some of their results: we exhibit the strong consistency and the rates of convergence of the maximum likelihood estimator to the normality, i.e. the Berry-Essen type theorem. Moreover, in two dimensional case, we obtain the explicit form of the estimator and establish the strong consistency and central limit theorem. We also give some remarks on higher dimensional DPPs.

# Acknowledgements

First of all, I am deeply indebted to my supervisor, Prof. Yaozhong Hu for his professional guidance and constant support. Prof. Hu not only helps me finish my thesis but also has been taking care of me throughout the last two years. His seemingly endless energy, patience, and kindness keep inspiring me.

I greatly thank the rest of my committee members, Prof. Bin Han, Prof. Bei Jiang, Prof. Adam Kashlak, and Prof. Linglong Kong for their valuable guidance and suggestions about my thesis. I greatly thank Prof. Jochen Kuttler for chairing the defense.

I would like to express my sincere gratitude to Prof. Feng Dai for his insightful teaching, kind encouragement, valuable advice, and unreserved recommendation. I also greatly thank Prof. Sean Graves and Prof. Peter Minev for their help in my Ph.D. application.

Furthermore, I would like to thank all of my friends and colleagues. I am lucky to have their warm help and the happy time we spent together is memorable.

Last but not the least, I would like to devote my deepest love to my family. Their love is always something I can have to fall back on at every stage of my growth.

# Table of Contents

**1 Introduction 1**

1.1 Literature Review . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.2 Outline and notation . . . . . . . . . . . . . . . . . . . . . . . . 2

**2 Determinantal point processes 4**

2.1 Definitions . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2.2 L-ensembles . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

2.3 Properties . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.3.1 Restriction . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.3.2 Complement . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.3.3 Domination . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.3.4 Scaling . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

2.3.5 Cardinality . . . . . . . . . . . . . . . . . . . . . . . . . 11

2.3.6 Identifiability . . . . . . . . . . . . . . . . . . . . . . . . 12

2.4 Examples . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

2.4.1 Descents in random sequences . . . . . . . . . . . . . . . 13

2.4.2 Loop-free Markov chain . . . . . . . . . . . . . . . . . . 14

2.4.3 Eigenvalues of random matrices . . . . . . . . . . . . . . 14

2.4.4 Edges in random spanning trees . . . . . . . . . . . . . . 15

2.5 Geometric interpretation . . . . . . . . . . . . . . . . . . . . . 15

**3 Maximum likelihood estimation of determinantal point processes 18**

3.1 Definitions . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18

3.2 Consistency . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21

3.3 Berry-Essen theorem . . . . . . . . . . . . . . . . . . . . . . . . 26
3.4 Two-by-two block kernel . . . . . . . . . . . . . . . . . . . . . . . 32

**4 Conclusion** **39**

**Bibliography** **40**

# List of Figures

2.1 A sample from DPP with Gaussian kernel. . . . . . . . . . . . . 8

2.2 A sample drawn independently from the plane . . . . . . . . . 8

2.3 If Y has cardinality 2, the probability of Y is proportional to the area spanned by its corresponding column vectors $B_1$ and $B_2$ as shown in (a). (b) shows that when $B_2$ increases its magnitude, the area increases and hence the probability increases. (c) shows that when $B_1$ and $B_2$ get closer, the area and hence probability decreases. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

# Chapter 1

# Introduction

## 1.1 Literature Review

Determinantal point processes (DPPs) arise from random matrix theory and are first introduced to give the probability distribution of fermions system in thermal equilibrium in quantum physics [Mac75]. The Pauli exclusion principle states that no two fermions can occupy the same quantum state, which leads to the so-called "anti-bunching" effect of fermions system. Since then, DPPs have been found in various aspects of probability, algebra and combinatorics, which we will briefly mention in the second chapter.

Determinantal point processes are often defined on $\mathbb{R}^d$ or $\mathbb{Z}^d$ through joint density functions. However, in this thesis, we mainly focus on determinantal point processes defined on a finite set, for which the joint density functions simplify to the determinant of matrices.

In the seminal work of [KT12], Kulesza and Taskar show that DPPs are unique among various probabilistic models in the sense that they capture the global repulsive behavior between items, give efficient algorithms for statistical inference, and have geometrical intuition. Since then, DPPs have been widely applied to machine learning community where the repulsive character of DPPs has been used to enforce the notion of diversity and quality in subset selection problems, which includes documentary summarization, image search, pose capture [KT12] and imagine processing[LDG21], etc. These real world applications necessitate the estimation of parameters of determinantal point processes. In this context, maximum likelihood estimation is a natural choice, which in general leads to a non-convex optimization problem. Along this direction, some work focuses on partial parameters estimation. Kulesza and Taskar split DPPs model into diversity part and quality part and only learn the quality part while the first part is fixed. They conjectured that the problem of learning the whole parameters of DPPs is NP-hard. After a decade, this conjecture is proven by [GJWX22]. [DB18] proposes a low-rank factorization of the determinant point processes and the parameter learning algorithm runs in sublinear time. [PL21] considers stationary determinantal point processes approximated by Fourier series. The other work addresses this problem without any restrictions, including Expectation-Maximization algorithm [GKFT14], Markov Chain Monte Carlo method [AFAT14] and fixed point algorithms [MS15]. None of the above work gives the global guarantee of estimation error, whereas [UBMR17] learns the parameters using moments and cycles, and gives the theoretical error bound.

[BMRU17] studies the local geometry of expected maximum likelihood estimation of DPPs, that is, the curvature of likelihood function around its maximum. Then they prove the maximum likelihood estimator converges to true values in probability and establish the corresponding central limit theorem.

## 1.2 Outline and notation

The remainder of the thesis is as follows. In the second chapter we give a brief introduction to determinantal point processes. We discuss about DPPs' definitions, important properties and examples in various aspects of mathematics. In the third chapter we study the maximum likelihood estimation of determinantal point processes. Our first work is to prove that the convergence of the estimator to true values holds almost surely. The second result is the Berry-Essen type theorem of the likelihood estimator, that is, the quantitative error bound for the central limit theorem. Lastly, we present some special cases where all the parameters can be estimated analytically. In the fourth chapter, we conclude our main result.

*Notation.* Fix a positive integer $N$ and define $[N] = \{1, 2, ..., N\}$. For a matrix $A \in \mathbb{R}^{N\times N}$ and $J \subseteq [N]$, denote by $A_J$ the restriction of A to $J \times J$. Sometimes we use a slight abuse of notation of $A_J$. We refer it to an $N \times N$ matrix whose restriction to $J$ is $A_J$ and has zeros everywhere else.

Let $\mathcal{S}_{[N]}$, $\mathcal{S}^{+}_{[N]}$, $\mathcal{S}^{++}_{[N]}$ and $\mathcal{S}^{(0,1)}_{[N]}$ be the set of all symmetric matrices in $\mathbb{R}^{N\times N}$, the set of all positive semi-definite matrices, the set of all positive definite matrices, and the set of all symmetric matrices whose eigenvalues belong to interval $(0, 1)$ respectively.

Let $A$ and $B$ be matrices in $\mathcal{S}_{[N]}$. We say that $B \preceq A$ if $A - B$ is positive semidefinite. Similarly, we say that $B \prec A$ if $A - B$ is positive definite. We say that $B \leq A$ if $A_{i,j} - B_{i,j} \geq 0$ for all $i$ and $j$.

For a matrix $A \in \mathbb{R}^{N\times N}$, let $\|A\|_F$, $\det(A)$, and $\mathrm{Tr}(A)$ denote its Frobenius norm, determinant and trace respectively. If $A$ is vectorized as an $N \times N$ column vector then the Frobenius norm of $A$ is $\mathcal{L}^2$ norm $\|A\|_2$.

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

A matrix $A \in \mathcal{S}_{[N]}$ is called block diagonal if there exists a partition $\{J_1, J_2, ..., J_k\}$, $k \geq 1$, such that $A_{ij} = 0$ when $i$ and $j$ belong to different $J_a$ and $J_b$. The largest $k$ such that the partition exists is called the number of blocks of $A$ and consequently $J_1, ..., J_k$ are called blocks of $A$.

For a subset $A \subseteq \mathcal{Y}$, let $\bar{A}$ denote the complement of $A$, that is, set $\mathcal{Y} \backslash A$.

# Chapter 2

# Determinantal point processes

In this section we give definitions and properties of discrete determinantal point processes. Most of the content is from [KT12], [Kul12], and [BMRU17]. We give a proof of the sufficient and necessary condition of defining DPPs, as it is not very clear in literature.

## 2.1 Definitions

A point process $\mathcal{P}$ on a ground set $\mathcal{Y}$ is a probability measure over the subsets of the ground set $\mathcal{Y}$. This kind of process is pretty common in real life. For example, the seats taken in one classroom at each class can be described by a point process, where the ground set is all seats in the classroom. Some students like to sit together while others are used to leaving some space in between. Some students like to sit at the front while others sit at the back. Sometimes the classroom is filled to capacity and sometimes the classroom has few students. Point processes capture these seats distributions.

For the remainder of this thesis, we will focus on discrete, finite point processes, where we assume without loss of generality that the ground set $\mathcal{Y} = \{1, 2, \cdots, N\}$ endowed with some metrics. In this case, a point process is simply a probability measure on $2^{\mathcal{Y}}$, the set of all subsets of $\mathcal{Y}$. A sample from $\mathcal{P}$ might be the empty set, the entirety of $\mathcal{Y}$, or anything in between.

**Definition 1.** *$\mathcal{P}$ is called a determinantal point process if, when $\mathbf{Y}$ is a random* *subset drawn according to $\mathcal{P}$, we have, for every fixed set $A \subseteq \mathcal{Y}$,*

$$\mathbb{P}(A \subseteq \mathbf{Y}) = \det(K_A) \tag{2.1}$$

*where $K_A$ is the restriction of an $N \times N$ symmetric matrix*[1] *$K$ to entries indexed by the elements of the subset $A$, that is, $K_A := [K_{i,j}]_{i,j \in A}$.*

If we think of each of item in the ground set $\mathcal{Y}$ as the Boolean variable, the left side of (2.1) is the marginal probability and hence $K$ is called marginal kernel. The normalization is unnecessary since the marginal probability need not sum to one. However, we have the following necessary conditions:

- Since the marginal probability of empty set is the total probability space, $\mathbb{P}(\Omega) = \mathbb{P}(\emptyset \subseteq \mathbf{Y}) = 1$. We denote $\det(K_\emptyset) = 1$.

- Since $\mathcal{P}$ is a probability measure, all principal minors of $K$, i.e. $\det(K_A)$ must be nonnegative, and thus K itself must be positive semidefinite, that is, $K \succeq 0$.

- $\mathbb{P}(\emptyset = \mathbf{Y}) + \mathbb{P}(\bigcup_{i=1}^{N}\{i \in \mathbf{Y}\}) = 1$. Using inclusion–exclusion principle we get

$$\begin{aligned}
\mathbb{P}(\bigcup_{i=1}^{N}\{i \in \mathbf{Y}\}) &= \sum_{i\in[N]} \mathbb{P}(i \in \mathbf{Y}) - \sum_{\{i,j\}\subset[N]} \mathbb{P}(\{i,j\} \subseteq \mathbf{Y}) + \ldots \\
\ldots & +(-1)^{N-1}\mathbb{P}([N] \subseteq \mathbf{Y}) \\
&= \sum_{|A|=1} \det(K_A) - \sum_{|A|=2} \det(K_A) + \ldots \\
\ldots & +(-1)^{N-1}\det(K) \\
&= 1 - \det(I - K) \qquad (2.2)
\end{aligned}$$

the last equality follows from the characteristic polynomial. This means

$$\mathbb{P}(\emptyset = \mathbf{Y}) = \det(I - K) \geq 0. \tag{2.3}$$

[1] In general, K need not be symmetric. We assume this for simplicity.

Similarly, we are able to show that $\mathbb{P}(\emptyset = \mathbf{Y} \cap A) = \det(I - K_A) \geq 0$ for any subset $A \subseteq [N]$. Therefore $K \preceq I$.

So the necessary condition is $0 \preceq K \preceq I$. In particular, all the diagonal elements of the marginal kernel $K_{i,i}$ should be in the interval $[0, 1]$. We can assume $K_{i,i}$ is always greater than 0, otherwise the element $i$ can be excluded from the model. This condition turns out to be sufficient: any $0 \preceq K \preceq I$ defines a DPP. To prove this, it's sufficient to show that for every $A \subseteq [N]$, the atomic probability is well-defined, that is, $0 \leq \mathbb{P}(A = \mathbf{Y}) \leq 1$. The probability being less or equal to 1 holds since $K \preceq I$. For the other inequality, we assume $K_A$ is invertible.[2] Then using Schur complement and characteristic polynomial, we have

$$\begin{aligned}
\mathbb{P}(A = \mathbf{Y}) &= \mathbb{P}(A \subseteq \mathbf{Y}) - \mathbb{P}(\bigcup_{i \in \bar{A}} \{A \cup \{i\} \subseteq \mathbf{Y}\}) \\
&= \det(K_A) - \sum_{i \in \bar{A}} \det(K_{A \cup \{i\}}) + \sum_{\{i,j\} \subseteq \bar{A}} \det(K_{A \cup \{i,j\}}) + \\
\ldots & +(-1)^{|\bar{A}|} \det(K) \\
&= \det(K_A) - \sum_{i \in \bar{A}} \det(K_A) \det(K_{ii} - K_{\{i\},A} K_A^{-1} K_{A,\{i\}}) \\
+ & \sum_{\{i,j\} \subseteq \bar{A}} \det(K_A) \det(K_{\{i,j\}} - K_{\{i,j\},A} K_A^{-1} K_{A,\{i,j\}}) + \\
\ldots & +(-1)^{|\bar{A}|} \det(K_A) \det(K_{\bar{A}} - K_{\bar{A},A} K_A^{-1} K_{A,\bar{A}}) \\
&= (-1)^{|\bar{A}|} \det(K_A) \det((K_{\bar{A}} - K_{\bar{A},A} K_A^{-1} K_{A,\bar{A}}) - I_{\bar{A}}) \\
&= (-1)^{|\bar{A}|} \det(K - I_{\bar{A}}), \qquad (2.4)
\end{aligned}$$

where $K_{A,B}$ denotes the matrix obtained by only keeping the entries whose rows belong to $A$ and columns belong to $B$ (if $A = B$ we simply denote it $K_A$.), $|A|$ denotes the cardinality of subset $A$, and $\bar{A}$ the complement of set $A$. Here we use a slight abuse of notation of $I_{\bar{A}}$ . We refer it to an N × N matrix whose restriction to $\bar{A}$ is $I_{\bar{A}}$ and has zeros everywhere else. Since $0 \preceq K \preceq I$, $\mathbb{P}(A = \mathbf{Y}) = |\det(K - I_{\bar{A}})| \geq 0$

[2] if $K_A$ is not invertible, we immediately get $\mathbb{P}(A = \mathbf{Y}) = 0$.

From Equation 2.1, if $A = \{i\} \subseteq \mathcal{Y}$ is a singleton, then we have

$$\mathbb{P}(i \in \mathbf{Y}) = K_{ii} \tag{2.5}$$

so the diagonal of marginal kernel gives the probability of inclusion for individual elements. if $A = \{i, j\} \subseteq \mathcal{Y}$, then the probability is given by the two by two principal minor $\begin{pmatrix} K_{ii} & K_{ij} \\ K_{ji} & K_{jj} \end{pmatrix}$

$$\begin{aligned} \mathbb{P}(\{i, j\} \subseteq \mathbf{Y}) &= K_{ii}K_{jj} - K_{ij}^2 \\ &\leq K_{ii}K_{jj} \\ &= \mathbb{P}(i \in Y)\mathbb{P}(j \in Y). \end{aligned} \tag{2.6}$$

Inequality 2.6 implies that element $i$ and $j$ tend not to co-occur, especially when $K_{ij}^2$ is close to $K_{ii}K_{jj}$. This phenomenon is called repulsive behavior of determinantal point processes and the off-diagonal elements characterize the degree of repulsion. Because of this major property, points tend to repel each other and hence induce point configurations that usually spread out evenly on the space. For example, let our ground set $\mathcal{Y}$ be a 2-dimensional grid: set $\{(i, j) \in \mathbb{Z}^2 : 1 \leq i, j \leq 60\}$, and then the kernel should a 3600 by 3600 matrix. Let the matrix be a Gaussian kernel[3], where each entry is given by $L_{ij,kl} = \exp\{-\frac{1}{0.1^2}((i-k)^2 + (j-l)^2)\}$. Using the sampling algorithm proposed by Hough et al [HKPV06], we draw samples from the DPP. See Figures 2.1 and 2.2.

Machine learning community often regards the off-diagonal elements $K_{i,j}$ as a measurement of similarity between pairs of elements in $\mathcal{Y}$. For example, if $K_{i,j} = 0$ it means that element $i$ and $j$ has no similarity whereas if $|K_{ij}| = \sqrt{K_{ii}K_{jj}}$ they are identical. This idea combined with DPPs' geometric intuition gives DPPs great abilities of modeling subset selection problems.

[3] the Gaussian kernel defines an L-ensemble instead of marginal kernel.

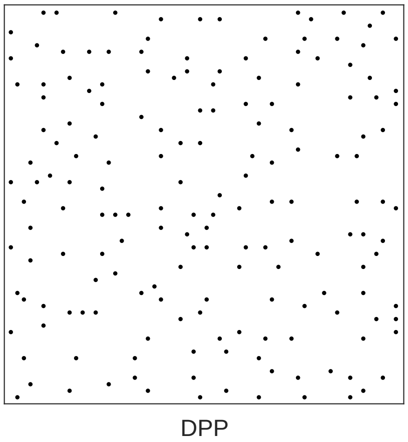

**Figure 2.1:** A sample from DPP with Gaussian kernel.

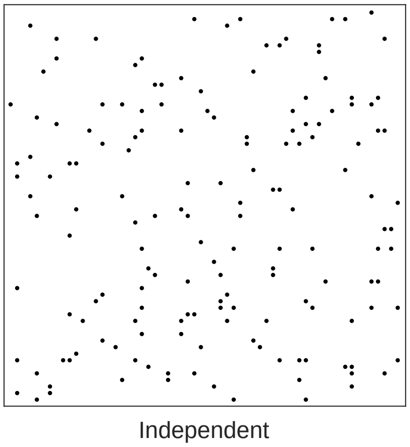

**Figure 2.2:** A sample drawn independently from the plane

## 2.2 L-ensembles

Sometimes it is quite inconvenient to work with marginal kernels since their eigenvalues should be bounded by 0 and 1, and the marginal probability is not very appropriate to describe real world data. Here we introduce a slightly smaller class of DPPs called L-ensembles.

**Definition 2.** *A point process is called an L-ensemble if it is defined through a real, symmetric matrix $L$:*

$$\mathbb{P}_L(A = \mathbf{Y}) \propto \det(L_A), \tag{2.7}$$

*where $A \subseteq \mathcal{Y}$ is a fixed subset .*

By the normalization, the proportion coefficient is equal to

$$\frac{1}{\sum_{A \subseteq \mathcal{Y}} \det(L_A)}. \tag{2.8}$$

Though this seems very cumbersome, the following theorem gives us the closed form of 2.8

**Theorem 3.** *For any $A \subseteq \mathcal{Y}$,*

$$\sum_{A \subseteq Y \subseteq \mathcal{Y}} \det(L_Y) = \det(L + I_{\bar{A}}). \tag{2.9}$$

*In particular, when $A = \emptyset$, we have $\sum_{A \subseteq \mathcal{Y}} \det(L_A) = \det(L + I)$.*

*Proof.* This can be proven by using the same argument as 2.4. □

Thus we have

$$\mathbb{P}_L(A = \mathbf{Y}) = \frac{\det(L_A)}{\det(L + I)}. \tag{2.10}$$

Moreover, we show that L-ensembles are DPPs. The following theorem is proven by [Mac75].

**Theorem 4.** *An L-ensemble is a DPP, and its marginal kernel is*

$$K = L(L + I)^{-1} = I - (L + I)^{-1}. \tag{2.11}$$

*Proof.* Using the last theorem, the marginal probability of a set A is

$$\begin{aligned} \mathbb{P}_L(A \subseteq \mathcal{Y}) &= \frac{\sum_{A \subseteq Y \subseteq \mathcal{Y}} \det(L_Y)}{\sum_{Y \subseteq \mathcal{Y}} \det(L_Y)} \\ &= \frac{\det(L + I_{\bar{A}})}{\det(L + I)} \\ &= \det((L + I_{\bar{A}})(L + I)^{-1}) \end{aligned} \tag{2.12}$$

We use the identity $L(L + I)^{-1} = I - (L + I)^{-1}$ to simplify

$$\begin{aligned} \mathbb{P}_L(A \subseteq \mathbf{Y}) &= \det(I_{\bar{A}}(L + I)^{-1} + I - (L + I)^{-1}) \\ &= \det(I - I_A(L + I)^{-1}) \\ &= \det(I_{\bar{A}} + I_A K), \end{aligned} \tag{2.13}$$

where we let $K = I - (L + I)^{-1}$. (2.13) is equal to the block matrix: $\begin{pmatrix} I_{\bar{A}} & 0 \\ K_{A,\bar{A}} & K_A \end{pmatrix}$ and hence $\mathbb{P}_L(A \subseteq \mathbf{Y}) = \det(K_A)$.

□

However, not all DPPs are L-ensembles. By inverting the 2.11, we have

$$L = K(I - K)^{-1}. \tag{2.14}$$

We see that the equality fails when the eigenvalues of K achieve the upper bound 1. Also from (2.3) we observe that the existence of L-ensembles is equivalent to the point processes giving non-zero probability to the empty set.

## 2.3 Properties

In this section we gather some basic properties of DPPs.

### 2.3.1 Restriction

If $\mathbf{Y}$ is distributed as a DPP with marginal kernel $K$, then $\mathbf{Y} \cap A$, where $A \subseteq \mathcal{Y}$, is also distributed as a DPP, with marginal kernel $K_A$.

*Proof.* For a subset $B \subseteq A$, $\mathbb{P}(B \subseteq \mathbf{Y} \cap A) = \mathbb{P}(B \subseteq \mathbf{Y}) = \det(K_B) = \det((K_A)_B)$. And for any $B$ not belonging to $A$, $\mathbb{P}(B \subseteq \mathbf{Y} \cap A) = 0$. □

### 2.3.2 Complement

If $\mathbf{Y}$ is distributed as a DPP with marginal kernel $K$, then $\mathcal{Y} - \mathbf{Y}$ is also a DPP with marginal kernel $I - K$, where $I$ denotes the identity matrix of appropriate size.

*Proof.* Since $\mathbb{P}(A \subseteq \mathcal{Y} - \mathbf{Y}) = \mathbb{P}(A \cap \mathbf{Y} = \emptyset)$, we can apply inclusion-exclusion principle in the same way as (2.2) to get the result. □

### 2.3.3 Domination

If $K \preceq K^{'}$, then for all $A \subseteq \mathcal{Y}$ we have

$$\det(K_A) \leq \det(K^{'}_A). \tag{2.15}$$

*Proof.* It is obvious if $K$ and $K'$ are positive semi-definite. For symmetric, positive definite matrices $K$, $K^{'}$ and $K^{'} - K$, using the Minkowski determinant theorem we have

$$\begin{aligned} \det(K') &= \det(K' - K + K) \geq \left(\det(K' - K)^{\frac{1}{N}} + \det(K)^{\frac{1}{N}}\right)^N \\ &\geq \det(K), \end{aligned} \tag{2.16}$$

and this also holds for all principal minors. □

### 2.3.4 Scaling

If $K = \gamma K'$ for some $0 \leq \gamma < 1$, then for all $A \subseteq \mathcal{Y}$ we have

$$\det(K_A) = \gamma^{|A|} \det(K'_A). \tag{2.17}$$

(2.17) has an interesting interpretation: the distribution of $K$ is obtained by taking a random subset distributed according to the DPP with marginal $K'$, and then independently delete each of its element by probability $1 - \gamma$.

### 2.3.5 Cardinality

Let $\lambda_1, \lambda_2, \ldots, \lambda_N$ be the eigenvalues of $L$. Then $|\mathbf{Y}|$ is distributed as the number of successes in N Bernoulli trials with n-th successful rate $\frac{\lambda_n}{1+\lambda_n}$. This fact follows from Theorem 2.3 in [KT12]. The expectation and variance of the n-th Bernoulli variable is $\frac{\lambda_n}{1+\lambda_n}$ and $\frac{\lambda_n}{(1+\lambda_n)^2}$, so we have

$$\mathbb{E}(|\mathbf{Y}|) = \sum_{i=1}^{N} \frac{\lambda_n}{1 + \lambda_n} \tag{2.18}$$

and

$$\mathrm{Var}(|\mathbf{Y}|) = \sum_{i=1}^{N} \frac{\lambda_n}{(1 + \lambda_n)^2}. \tag{2.19}$$

### 2.3.6 Identifiability

The distribution of DPPs is not identifiable, that is, multiple kernels can give rise to the same DPP. Let $\mathrm{DPP}(L^*)$ denote the L-ensemble determined by the matrix $L^*$. The identifiability problem is precisely described by Theorem 4.1 in [Kul12].

**Theorem 5.** *Denote $\mathcal{D}$ the collection of all diagonal matrices whose entry is either 1 or -1. Then for $L_1$ and $L_2 \in \mathcal{S}^{++}_{[N]}$, $\mathrm{DPP}(L_1) = \mathrm{DPP}(L_2)$ if and only if there exists a $D \in \mathcal{D}$ such that $L_2 = DL_1D$.*

[BMRU17] defines the degree of identifiability of a kernel $L$ and gives the following proposition.

**Definition 6.** *Let $L \in \mathcal{S}^{++}_{[N]}$. The degree $\mathrm{Deg}(L)$ of identifiablity of $L$ is the cardinality of the family $\{DLD : D \in \mathcal{D}\}$. We say that $L$ is irreducible if the cardinality is $2^{N-1}$ and reducible otherwise. If $\mathbf{Z} \sim \mathrm{DPP}(L)$, we also call $\mathbf{Z}$ is irreducible if $L$ is irreducible and reducible otherwise.*

If the i-th element of $D$ is -1, $DLD$ flips the sign of i-th row and column of $L$, and hence the diagonal element of D always remain the same. In fact, it is easy to check that $\mathrm{Deg}(L) = 1$ if and only if $L$ is a diagonal matrix. Moreover, $\mathrm{Deg}(L)$ is at most $2^{N-1}$, so for any $L \in \mathcal{S}^{++}_{[N]}$, $1 \leq \mathrm{Deg}(L) \leq 2^{N-1}$. The next proposition shows that the degree of identifiability is completely described by the block structure of the matrix. And the block structure is in turn characterized by the connectivity of certain graphs called determinantal graph.

**Definition 7.** *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 $j$.*

**Proposition 8.** *Let $L \in \mathcal{S}^{++}_{[N]}$, $Z \sim \mathrm{DPP}(L)$, and $K$ be the corresponding marginal kernel. Let $1 \leq k \leq N$ and $\{J_1, J_2, ..., J_k\}$ be a partition of $[N]$. The following statements are equivalent:*

*1. $L$ is block diagonal with $k$ blocks $J_1, J_2, ..., J_k$,*

*2. $K$ is block diagonal with $k$ blocks $J_1, J_2, ..., J_k$,*

*3. $Z \cap J_1, ..., Z \cap J_k$ are mutually independent irreducible DPPs,*

*4. $\mathcal{G}_L$ has $k$ connected components given by $J_1, ..., J_k$,*

*5. $L = D_j L D_j$ for all $j \in [k]$, where $D_j \in \mathcal{D}$ whose diagonal element is 1 on $J_j$ and -1 otherwise.*

From the above proposition we know that $L$ has $k$ blocks if and only if the degree of identifiability of $L$ is $2^{N-k}$. In particular, $L$ is irreducible if and only if it only has one block.

## 2.4 Examples

Because of these nice properties, determinantal point processes are also found prevalent in many areas of mathematics, such as stochastic processes, random matrix theory, and random graph theory. We introduce some examples that have been thoroughly studied.

### 2.4.1 Descents in random sequences

If a sequence of random numbers is drawn uniformly and independently from a given set, say, the set$\{0, 1, 2, \ldots, 9\}$, then the locations in the sequence where the current number is less than the previous numbers form a subset of $\{2, 3, \ldots, 9\}$, which we take as our ground set. This subset is distributed as a determinantal point process. Intuitively speaking, if in the random sequence, the k-th number is less than the previous one, it means the k-th number is probably not too large, so the next k+1-th number independently drawn from $\{0, 1, 2 \ldots, 9\}$ is less likely to be less than the k-th number. In consequence, adjacent numbers repel each other and this repulsion is precisely described by a determinantal point process. See more in [BDF10].

### 2.4.2 Loop-free Markov chain

A discrete time Markov chain on a discrete space $\mathcal{X}$ with initial distribution $\pi$ and transitional matrix $[P_{xy}]_{x,y\in\mathcal{X}}$ is called loop-free if its trajectory of the Markov chain doesn't pass through the same point twice almost surely. In other words, we assume that:

$$P_{xx}^{k} = 0 \qquad \text{for any k} > 0 \, and \, x \in \mathcal{X},$$

where $P_{xy}^{k}$ is the probability that the chain starts from x and ends at y after k steps. This condition guarantees the finiteness of the matrix elements of the matrix :

$$Q = \sum_{i=1}^{\infty} P^{i} \leq 1.$$

We consider the Markov chain as a probability measure on trajectories viewed as subsets of $\mathcal{X}$. Then this measure on $2^{\mathcal{X}}$ is a determinantal point process on $\mathcal{X}$ with marginal kernel

$$K_{xy} = \pi_x + (\pi Q)_x - Q_{yx}.$$

Markov chain is probably the most fundamental thing in stochastic processes. This theorem actually shows that DPPs are actually quite common. In fact, we can construct a Loop-free Markov chain easily. For any discrete time Markov chain $M(t_n)$ on discrete space, its graph $(t_n, M(t_n))$ then is a loop-free Markov chain since time can never go back. Full details are given in [Bor08].

### 2.4.3 Eigenvalues of random matrices

A complex Ginibre ensemble is a random matrix whose entries are i.i.d standard complex normal random variables. The eigenvalues on the complex plane of Ginibre ensemble then are distributed as a determinantal point process. The details are in [Gin65].

### 2.4.4 Edges in random spanning trees

Let $G$ be an arbitrary finite graph with N edges, which we take as our ground set. We draw spanning trees uniformly from the set of all the spanning trees in $G$. Then the edges of random spanning trees form a random subset of the ground set and is distributed as a DPP. Since the cardinality of the edges of every spanning tree is always equal to $k =$ numbers of vertices $-1$ , this special DPP only assigns probability to subsets whose cardinality is fixed k. Full details are in [BP93].

## 2.5 Geometric interpretation

Using matrix decomposition, [KT12] gives the geometric interpretation of DPPs. For every symmetric positive semi-definite kernel $L$, there exists a $D \times N$ matrix such that $L = B^T B$. (D should be greater than or equal to the rank of $L$) Denote the columns of $B$ by $B_i$, for $i = 1, 2, \ldots, N$. Then using the geometry interpretation of determinant we have for any arbitrary subset $Y \subseteq \mathcal{Y}$:

$$\mathbb{P}(Y = \mathbf{Y}) \propto \det(L_Y) = \mathrm{Vol}^2(\{B_i\}_{i \in Y}). \tag{2.20}$$

The above equation implies that the probability that $Y$ occurs is proportional to the volume of the parallelepiped spanned by the column vectors $B_i$ for which $i \in Y$. The volume of the parallelepiped depends on the magnitude of vectors and angle between vectors. In fact, we can normalize the column vector

$$B_i = ||B_i|| \cdot \frac{B_i}{||B_i||} := q_i \cdot \phi_i,$$

where $q_i$ is the magnitude and $\phi_i$ is the unit direction vector; moreover, we divide the probability model into two parts:

$$\det(L_Y) = \det(B^T B) = \left( \prod_{i \in Y} {q_i}^2 \right) \det(S_Y), \tag{2.21}$$

where $\det(S_Y) = \det([\langle \phi_i, \phi_j \rangle]_{ij})$. The first part is called the quality part and the second similarity kernel. This geometric interpretation leads to many

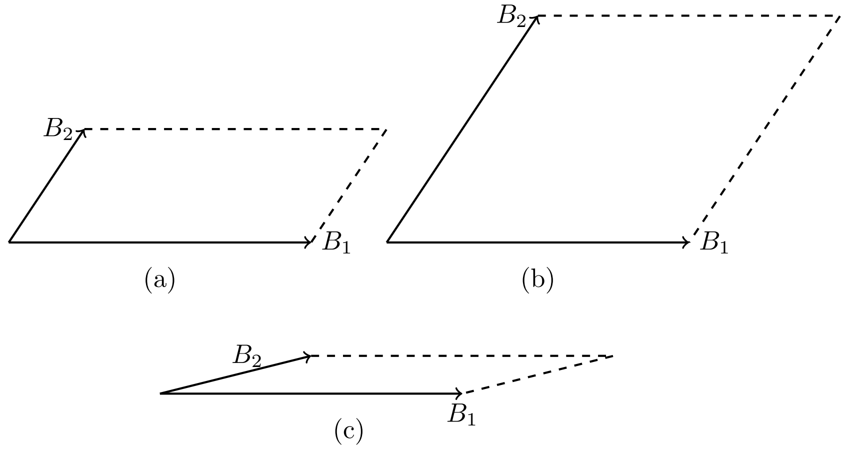

**Figure 2.3:** If Y has cardinality 2, the probability of Y is proportional to the area spanned by its corresponding column vectors $B_1$ and $B_2$ as shown in (a). (b) shows that when $B_2$ increases its magnitude, the area increases and hence the probability increases. (c) shows that when $B_1$ and $B_2$ get closer, the area and hence probability decreases.

real world applications. Usually we regard each of items in the ground set as its column vector of $B$: the magnitude of the vector represents the intrinsic goodness of the item and the angle between a pair of vectors stands for the similarity of the pairs of items. Take the document summarization in [KT12] as an example. To put it simply, the goal of the task is to obtain a summary from an article. We let the ground set consist of all the sentences from the article and a summary be a subset of the ground set. We expect a good summary should cover the most important but also diverse information from the article; It makes no sense if all the sentences convey the same information. The information of a sentence may be relevant to its position in the article (normally the first a few sentences cover the main idea of the article, so they convey more information), its length (the longer the sentence is, the more information it has.), etc; these features can be used as the quality of the sentence. As for the diversity of sentence, we can measure the similarity between a pair of sentences by counting words they both have. If two sentences have many words in common, they are very likely to be similar. Using the quality of sentences and similarity between sentences we are able to construct a DPP from which we sample a good summary.

# Chapter 3

# Maximum likelihood estimation of determinantal point processes

One of the most important questions is that how do we estimate the parameters of the kernel of a DPP based on a set of samples from it? A natural choice is maximum likelihood estimation.

Given a set of observed data, maximum likelihood estimation is a method of estimating the unknown parameters of a known probabilistic distribution. The estimation is obtained by choosing the parameters for which the likelihood function achieves its maximum. This method is a dominant means in statistical inference because it is very flexible and intuitive. The maximum likelihood estimation of determinantal point processes has been well studied by [BMRU17]. In this chapter, we introduce and extend some of their results.

## 3.1 Definitions

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

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

where

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

$\hat{p}(J)$ stands for the empirical probability that subset $J$ occurs. It is also useful to define the expected log-maximum likelihood function given the real kernel $L^\star$

$$\Phi_{L^\star}(L) = \sum_{J\subseteq[N]} p_{L^\star}(J)\log\det(L_J) - \log\det(I+L) \tag{3.2}$$

where

$$p_{L^\star}(J) = \mathrm{E}(\hat{p}(J)) = \frac{\det\left(L^\star_J\right)}{\det\left(I+L^\star\right)}.$$

Basically, we take the expectation of $\hat{p}(J)$ with respect to the true probability measure $\mathrm{DPP}(L^\star)$ and then get the expected maximum likelihood function. Let in the sequel $L^\star$ be fixed, $\hat{p}_J$ denote $\hat{p}(J)$, $p_J$ denote $p_{L^*}(J)$ and $\Phi$ denote $\Phi_{L^\star}$.

Let $\mathrm{KL}\big(\mathrm{DPP}(L^\star), \mathrm{DPP}(L)\big)$ be the Kullback-Leibler divergence, which measures the difference between distributions of $\mathrm{DPP}(L^\star)$ and of $\mathrm{DPP}(L)$. Since Kullback-Leibler divergence is always non-negative, we have

$$\mathrm{KL}\big(\mathrm{DPP}(L^\star), \mathrm{DPP}(L)\big) = \Phi(L^\star) - \Phi(L) \geq 0,\ \forall L \in \mathcal{S}^{++}_{[N]}.$$

As a consequnce $L^\star$ is the global maxima of the expected maximum function $\Phi(L)$. Due to non-identifiability of DPPs introduced in Theorem 6, $\Phi(L)$ achieves the maximum whenever $L = DL^\star D$ for some $D \in \mathcal{D}$ and hence the global maxima is the set $\{DL^\star D : D \in \mathcal{D}\}$. We introduce a useful lemma.

**Lemma 9.** *The gradient of log-likelihood function* $\hat{\Phi}(L)$ *defined in* (3.1) *is given by*

$$\mathrm{d}\hat{\Phi}(L) = \sum_{J\subseteq[N]} \hat{p}_J L_J^{-1} - (I+L)^{-1}. \tag{3.3}$$

*Proof.* For all square matrices $L \in \mathcal{S}^{++}_{[N]}$, from Theorem 3,

$$\det(L+I) = \sum_{J\subseteq[N]} \det(L_J). \tag{3.4}$$

Now thinking of determinant as a multivariate function, the directional derivative of $\det(L+I)$ along direction $H$ is given by

$$\begin{aligned}
\mathrm{d}\det(L+I)(H) =& \lim_{t\to 0} \frac{\det(L+I+tH)-\det(L+I)}{t} \\
=& \lim_{t\to 0} \det(L+I)\Big[\frac{\det(I+t(L+I)^{-1}H)-1}{t}\Big] \\
=& \lim_{t\to 0} \det(L+I)\Big[\frac{1+t\,\mathrm{Tr}((L+I)^{-1}H)+o(t^2)-1}{t}\Big] \\
=& \det(L+I)\,\mathrm{Tr}((L+I)^{-1}H), \qquad (3.5)
\end{aligned}$$

where the third equality follows from the power series representation of $\det(I+A)$. Then differentiating (3.4) once over $L \in \mathcal{S}_{[N]}^{++}$ along any $H \in \mathcal{S}_{[N]}$ yields

$$\sum_{J\subseteq[N]} \det(L_J)\,\mathrm{Tr}(L_J^{-1}H_J) = \det(I+L)\,\mathrm{Tr}((I+L)^{-1}H). \qquad (3.6)$$

By dividing both sides by $\det(I+L)$,

$$\sum_{J\subseteq[N]} p_L(J)\,\mathrm{Tr}(L_J^{-1}H_J) = \mathrm{Tr}((I+L)^{-1}H). \qquad (3.7)$$

In matrix form, the above equation becomes

$$\sum_{J\subseteq[N]} p_L(J)L_J^{-1} = (I+L)^{-1}. \qquad (3.8)$$

Using (3.5) we can obtain the gradient of log-likelihood function $\hat{\Phi}(L)$

$$\mathrm{d}\hat{\Phi}(L) = \sum_{J\subseteq[N]} \hat{p}_J L_J^{-1} - (I+L)^{-1}. \qquad (3.9)$$

□

Moreover, the following theorems by [BMRU17] characterize the curvature of the expected maximum likelihood function at its maximum.

**Theorem 10.** *Let $L^* \in \mathcal{S}_{[N]}^{++}$, $Z \sim \mathrm{DPP}(L^\star)$ and $\Phi = \Phi_{L^\star}$. Then, $L^\star$ is a* *critical point of* $\Phi$*. Moreover, for any* $H \in \mathcal{S}_{[N]}$*,*

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

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

**Theorem 11.** *Under the same assumptions of Theorem 10, the null space of the quadratic Hessian map* $H \in \mathcal{S}_{[N]} \mapsto \mathrm{d}^2\Phi(L^*)(H,H)$ *is given by*

$$\mathcal{N}(L^*) = \left\{H \in \mathcal{S}_{[N]} \,:\, H_{i,j} = 0 \text{ for all } i,j \in [N] \text{ such that } i \sim_{L^*} j\right\} . \quad (3.10)$$

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

## 3.2 Consistency

One main property of maximum likelihood estimation is the consistency. Since the distributions of determinantal point processes are not identifiable we measure the performance of maximum likelihood estimation by the distance between the likelihood maximizer $\hat{L}_n$ and the set of true values :

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

[BMRU17] proves the distance converges to zero in probability. We prove that the consistency also holds almost surely. The proof is based on theorem 14 in [BMRU17] and Wald's consistency theorem [Wal49]. Even though the latter theorem originally requires the distribution to be identifiable, this is not a problem for this setting where we consider distance between $\hat{L}_n$ and the set of true values instead of one value.

We first show that $\ell(\hat{L}_n, L^\star)$ converges to zero almost surely when parameters of matrices are restricted on a compact set. For $0 < \alpha < \beta < 1$, define a set $E_{\alpha,\beta}$

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

Choose appropriate $\alpha, \beta$ such that $L^\star \in E_{\alpha,\beta}$. $E_{\alpha,\beta}$ is compact since it's bounded and closed in $\mathbb{R}^{N\times N}$.

**Lemma 12.** *Let $Z_1, ..., Z_n$ be $n$ independent copies of $Z \sim DPP(L^\star)$ for some unknown $L^\star \in \mathcal{S}_{[N]}^{++}$. Let $\hat{L}_n$ be the maximum likelihood estimator of $\hat{\Phi}(L)$ defined on $E_{\alpha,\beta}$, then $\ell(\hat{L}_n, L^\star)$ converges to zero almost surely.*

*Proof.* Let

$$\Delta\hat{\Phi}(L) = \hat{\Phi}(L) - \hat{\Phi}(L^\star) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{P_L(Z_i)}{P_{L^\star}(Z_i)}$$

and

$$\Delta\Phi(L) = \Phi(L) - \Phi(L^\star) = \mathrm{E}_{L^\star}\big(\log \frac{P_L(Z)}{P_{L^\star}(Z)}\big).$$

$\Delta\Phi(L)$ is the Kullback-Leibler Divergence between DPP$(L^\star)$ and DPP$(L)$. By Jensen's inequality, $\Delta\Phi(L) \leq 0$ for all $L$ and by the condition for equality in Jensen's inequality, $\Phi(L) = \Phi(L^\star)$ if and only if $P_L(Z) = P_{L^\star}(Z)$ for all $Z \in [N]$, which means $L = DL^\star D$ for some $D \in \mathcal{D}$.

For each $L \in E_{\alpha,\beta}$, the strong law of large numbers implies

$$\Delta\hat{\Phi}(L) \xrightarrow{a.s.} \Delta\Phi(L).$$

However, the above convergence doesn't imply the convergence of Maximum likelihood estimator to the true values. Thus we need some kinds of uniform integrablity, which is the Wald's integrability condition: for every $L \in E_{\alpha,\beta}$, there exists $\epsilon > 0$ such that,

$$\mathrm{E} \sup_{\substack{N \in E_{\alpha,\beta} \\ \ell(L,N)<\epsilon}} \log \frac{P_N(Z)}{P_{L^\star}(Z)} < \infty. \tag{3.11}$$

Since $L \mapsto \log \frac{P_L(Z)}{P_{L^\star}(Z)}$ is continuous (the determinant function is continuous), for any arbitrary $\delta_L > 0$ there exists $\ell(L, N) < \epsilon$,

$$(1-\delta_L)\frac{P_L(Z)}{P_{L^\star}(Z)} < \frac{P_N(Z)}{P_{L^\star}(Z)} < (1+\delta_L)\frac{P_L(Z)}{P_{L^\star}(Z)},$$

the Wald's integrability condition is satisfied. Now for every sequence $\{L_n\}$

converging to L, we show that $\Delta\Phi(L_n)$ is upper semicontinuous:

$$\begin{aligned}
\limsup_{n\to\infty} \Delta\Phi(L_n) &= \limsup_{n\to\infty} \mathrm{E} \log \frac{P_{L_n}(Z)}{P_{L^\star}(Z)} \\
&\le \mathrm{E} \limsup_{n\to\infty} \log \frac{P_{L_n}(Z)}{P_{L^\star}(Z)} \\
&= \mathrm{E} \frac{P_L(Z)}{P_{L^\star}(Z)} \\
&= \Delta\Phi(L).
\end{aligned}$$

The second inequality follows from the Fatou's lemma and the third identity is the consequence of continuity of the function $\log \frac{P_{L_n}(Z)}{P_{L^\star}(Z)}$. For every $\eta > 0$ we define the set $K_\eta$:

$$\begin{aligned}
K_\eta &= \big\{L \in E_{\alpha,\beta} : \ell(L, L^\star) \ge \eta\big\} \\
&= \bigcap_{D\in\mathcal{D}} \big\{L \in E_{\alpha,\beta} : \|L - DL^\star D\|_F \ge \eta\big\}. \qquad (3.12)
\end{aligned}$$

Set $K_\eta$ is a closed set and hence a compact set.

Since $\Delta\Phi(L)$ is an upper semicontinuous function, it achieves maximum over the compact set $K_\eta$. We denote the maximum by $m(\eta)$. And we cannot have $m(\eta) = 0$ because that would imply there is a $L \in K_\eta$ such that $L = DL^\star D$ for some $D \in \mathcal{D}$. The strong law of large numbers implies

$$\begin{aligned}
\sup_{\substack{N\in E_{\alpha,\beta} \\ \ell(L,N)<\epsilon}} \Delta\hat{\Phi}(N) &\le \frac{1}{n}\sum_{i=1}^{n} \sup_{\substack{N\in E_{\alpha,\beta} \\ \ell(L,N)<\epsilon}} \log \frac{P_N(Z_i))}{P_{L^\star}(Z_i)} \\
&\xrightarrow{a.s.} \mathrm{E} \sup_{\substack{N\in E_{\alpha,\beta} \\ \ell(L,N)<\epsilon}} \log \frac{P_N(Z)}{P_{L^\star}(Z)}. \qquad (3.13)
\end{aligned}$$

By continuity,

$$\lim_{\epsilon\to 0} \sup_{\substack{N\in E_{\alpha,\beta} \\ \ell(L,N)<\epsilon}} \log \frac{P_N(Z))}{P_{L^\star}(Z)} = \log \frac{P_L(Z)}{P_{L^\star}(Z)}$$

and $\sup_\epsilon \log \frac{P_N}{P_{L^\star}}$ is a decreasing function with respect to $\epsilon$ because supremum over a smaller subset is smaller than over a bigger subset. And by (3.11) it is integrable for all small enough $\epsilon$ . Hence by dominated convergence theorem,

$$\lim_{\epsilon\to 0} \mathrm{E} \sup_{\substack{N\in E_{\alpha,\beta}\\ \ell(L,N)<\epsilon}} \log \frac{P_N(Z))}{P_{L^\star}(Z)} = \mathrm{E}\log\frac{P_L(Z)}{P_{L^\star}(Z)} = \Delta\Phi(L).$$

Thus for any $L \in K_\eta$ and any $\gamma > 0$ there exists a $\epsilon_L$ such that

$$\mathrm{E} \sup_{\substack{N\in E_{\alpha,\beta}\\ \ell(L,N)<\epsilon_L}} \log \frac{P_N(Z)}{P_{L^\star}(Z)} < m(\eta)+\gamma. \tag{3.14}$$

For each $L \in K_\eta$, we define the open set:

$$V_L = \{N \in E_{\alpha,\beta} : \ell(N,L) < \epsilon_L\}$$

and then the family $\{V_L : L \in K_\eta\}$ is an open cover of $K_\eta$ and hence has a finite subcover: $V_{L_1}, V_{L_2}, ...., V_{L_d}$. On every $V_{L_i}$ we use strong law of large numbers one more time:

$$\begin{aligned} \limsup_{n\to\infty} \sup_{N\in V_{L_i}} \Delta\hat{\Phi}(N) &\leq \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \sup_{N\in V_{L_i}} \log\frac{P_N(Z_i)}{P_{L^\star}(Z_i)} \\ &= \mathrm{E} \sup_{N\in V_{L_i}} \log\frac{P_N(Z)}{P_{L^\star}(Z)}. \end{aligned} \tag{3.15}$$

So we get

$$\limsup_{n\to\infty} \sup_{N\in V_{L_i}} \Delta\hat{\Phi}(N) < m(\eta)+\gamma \qquad i = 1,2,...,d.$$

Since $\{V_{L_i} : i = 1, 2..., d\}$ cover $K_\eta$ we have

$$\limsup_{n\to\infty} \sup_{N\in K_\eta} \Delta\hat{\Phi}(N) < m(\eta)+\gamma$$

which, since $\gamma$ is arbitrary, implies

$$\limsup_{n\to\infty} \sup_{L\in K_\eta} \Delta\hat{\Phi}(L) < \sup_{L\in K_\eta} \Delta\Phi(L) = m(\eta). \tag{3.16}$$

Notice that $m(\eta) < 0$. From (3.16) there exists a constant $N_1$ such that

$$\sup_{L \in K_\eta} \Delta\hat{\Phi}(L) < \frac{m(\eta)}{2}, \qquad n > N_1.$$

But

$$\Delta\hat{\Phi}(\hat{L}_n) = \sup_{L \in E_{\alpha,\beta}} \Delta\hat{\Phi}(L) \geq \Delta\hat{\Phi}(L^\star) \xrightarrow{a.s.} \Delta\Phi(L^\star) = 0,$$

so there exists a constant $N_2$ such that

$$\Delta\hat{\Phi}(\hat{L}_n) \geq \frac{m(\eta)}{2}, \qquad n > N_2$$

which implies that $\hat{L}_n \notin K_\eta$, that is, $\ell(\hat{L}_n, L) < \epsilon$.

□

The second step is to show that the event $\{\hat{L}_n \in E_{\alpha,\beta}\}$ holds almost sure. We adopt the proof from [BMRU17]. Let $\delta = \min_{J \subset [N]} P_{L^\star}(J)$. For simplicity, we denote $P_{L^\star}(J)$ by $p^\star_J$. Since $L^\star$ is positive definite, $\delta > 0$. Define the event $\mathcal{A}$ by

$$\mathcal{A} = \bigcap_{J \subset [N]} \{p^\star_J \leq 2\hat{p}_J \leq 3p^\star_J\}.$$

Observe that $\Phi(L^\star) < 0$, so we can define $\alpha < \exp(3\Phi(L^\star)/\delta)$ and $\beta > 1 - \exp(3\Phi(L^\star)/\delta)$ such that $0 < \alpha < \beta < 1$. Using the conclusion in Theorem 14 from [BMRU17] we know that on the event $\mathcal{A}$, $\hat{L} \in E_{\alpha,\beta}$, that is,

$$P(\hat{L} \in E_{\alpha,\beta}) \geq P(\mathcal{A}).$$

Because

$$\hat{p}_J = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(Z_i = J) \xrightarrow{a.s.} P_{L^\star}(Z = J) = p^\star_J$$

, the event $\mathcal{A}$ holds almost surely when $n$ goes to infinity and hence $\{\hat{L}_n \in E_{\alpha,\beta}\}$ holds almost surely.

**Theorem 13.** *$\ell(\hat{L}_n, L^\star)$ converges to zero almost surely.*

*Proof.* Let $\mathbb{I}_{E_n}$ denote the characteristic function of the event $\{\hat{L}_n \in E_{\alpha,\beta}\}$, then

$$\begin{aligned}
\mathbb{P}\big(\lim_{n\to\infty} \ell(\hat{L}_n, L^\star) = 0\big) &= \mathbb{P}\big(\lim_{n\to\infty} \ell(\hat{L}_n, L^\star) = 0, \lim_{n\to\infty} \mathbb{I}_{E_n} = 1\big) \\
&\quad + \mathbb{P}\big(\lim_{n\to\infty} \ell(\hat{L}_n, L^\star) = 0, \lim_{n\to\infty} \mathbb{I}_{E_n} \neq 1\big) \\
&= \mathbb{P}\big(\lim_{n\to\infty} \ell(\hat{L}_n, L^\star) = 0, \lim_{n\to\infty} \mathbb{I}_{E_n} = 1\big) \\
&= \mathbb{P}\big(\lim_{n\to\infty} \ell(\hat{L}_n, L^\star) = 0 \big| \lim_{n\to\infty} \mathbb{I}_{E_n} = 1\big)\mathbb{P}\big(\lim_{n\to\infty} \mathbb{I}_{E_n} = 1\big) \\
&= \mathbb{P}\big(\lim_{n\to\infty} \ell(\hat{L}_n, L^\star) = 0 \big| \lim_{n\to\infty} \mathbb{I}_{E_n} = 1\big) \\
&= 1.
\end{aligned}$$

The last equality follows from the fact that $\hat{L}_n \in E_{\alpha,\beta}$ almost surely and from lemma 12. □

## 3.3 Berry-Essen theorem

We observe that an $N$ by $N$ matrix $[A_{ij}]_{N\times N}$ can also be viewed as an $N \times N$ dimensional column vector: $(A_{11}, A_{12}, ..., A_{1N}, A_{21}, ..., A_{N1}, ...A_{NN})^T$. And then the Frobenius norm of the matrix is just the $\mathcal{L}^2$ norm for its corresponding column vector. In the following section we abuse the notation: without causing any confusion, sometimes we regard the matrix as the corresponding column vector.

Assume that $L^\star$ is irreducible and let $\hat{L}$ be the maximal likelihood estimator. Let $\hat{D} \in \mathcal{D}$ be such that

$$\|\hat{D}\hat{L}\hat{D} - L^\star\|_F = \min_{D\in\mathcal{D}} \|D\hat{L}D - L^\star\|_F \tag{3.17}$$

and set $\tilde{L} = \hat{D}\hat{L}\hat{D}$. Then the strong consistency of $\tilde{L}$ immediately follows from the theorem 13.

According to theorem 11, $\mathrm{d}^2\Phi(L^\star)$ is negative definite and hence invertible. Let $V(L^\star)$ denote its inverse. Here if we vectorize $L$ then $\mathrm{d}^2\Phi(L^\star)$ is an $(N \times$

$N) \times (N \times N)$ Hessian matrix. By Theorem 5.41 in [VdV00],

$$\begin{aligned}\sqrt{n}(\tilde{L}-L^\star) &= -(\mathrm{E}(\mathrm{d}^2 \log P_{L^\star}(Z)))^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{d}(\log P_{L^\star}(Z_i)) + o_P(1) \\ &= -V(L^\star)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}((L^\star_{Z_i})^{-1} - (I+L^\star)^{-1}) + o_P(1). \qquad (3.18)\end{aligned}$$

In particular, Theorem 5.41 states that the sequence $\sqrt{n}(\tilde{L}-L^\star)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $-V(L^\star)$. Hence we get the following theorem from [BMRU17].

**Theorem 14.** *Let $L^\star$ be irreducible. Then, $\tilde{L}$ is asymptotically normal:*

$$\sqrt{n}(\tilde{L}-L^\star) \xrightarrow[n\longrightarrow\infty]{} \mathcal{N}(\mathbf{0}, -V(L^\star)), \qquad (3.19)$$

*where the above convergence holds in distribution.*

Now let us take one step further. We want to find an upper error bound on the rate of convergence of the distribution of $(-V(L^\star))^{-\frac{1}{2}}\sqrt{n}(\tilde{L}-L^\star)$ to standard normal distribution $Z \sim \mathcal{N}(\mathbf{0}, I)$. We argue that when $\tilde{L} \in E_{\alpha\beta}$, the bound of the maximal error is of order $n^{-\frac{1}{4}}$. The condition is not of too much restriction. Indeed, Since $\alpha$ and $\beta$ can be arbitrarily close to 0 and 1 respectively, $E_{\alpha,\beta}$ converges to $\mathcal{S}^{++}_{[N]}$. What'more, since $\hat{L} \in E_{\alpha,\beta}$ almost surely, $\hat{D}\hat{L}\hat{D} = \tilde{L} \in E_{\alpha,\beta}$ almost surely.

**Theorem 15.** *Let $\tilde{L}$ be as defined as above and also belong to $E_{\alpha,\beta}$ and $Z$ be an $N \times N$ standard Gaussian matrix. Then for every $x \in \mathbb{R}^{N\times N}$,*

$$|\mathbb{P}((-V(L^\star))^{-\frac{1}{2}}\sqrt{n}(\tilde{L}-L^\star) < x) - \mathbb{P}(Z < x)| \leq C\frac{1}{\sqrt[4]{n}},$$

*where $C$ is a sufficient large constant, which is irrelevant to $x$, subject to $\alpha, \beta$ and proportional to $N^2$.*

According to (3.18), $(-V(L^\star))^{-\frac{1}{2}}\sqrt{n}(\tilde{L}-L^\star)$ can be decomposed into a sum

$$X_n = \sum_{i=1}^{n} \xi_i = (-V(L^\star))^{\frac{1}{2}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} ((L^\star_{Z_i})^{-1} - (I + L^\star)^{-1}) \tag{3.20}$$

and a term $\rho_n = (-V(L^\star))^{-\frac{1}{2}} o_P(1)$ whose Frobenius norm converges to zero in probability.

$$\begin{aligned}
&|\mathbb{P}(X_n + \rho_n < x) - \mathbb{P}(Z < x)| \\
&= |\mathbb{P}(X_n + \rho_n < x, \|\rho_n\|_F \geq k_n) + \mathbb{P}(X_n + \rho_n < x, \|\rho_n\|_F < k_n) - \mathbb{P}(Z < x)| \\
&\leq \mathbb{P}(\|\rho_n\|_F \geq k_n) + |\mathbb{P}(X_n + \rho_n < x, \|\rho_n\|_F < k_n) - \mathbb{P}(Z < x)| \\
&\leq \mathbb{P}(\|\rho_n\|_F \geq k_n) && \text{(I1)} \\
&\quad + |\mathbb{P}(X_n + k_n \mathbb{1} < x, \|\rho_n\|_F < k_n) - \mathbb{P}(Z < x)| && \text{(I2)} \\
&\quad + |\mathbb{P}(X_n - k_n \mathbb{1} < x, \|\rho_n\|_F < k_n) - \mathbb{P}(Z < x)|, && \text{(I3)}
\end{aligned}$$

where $\{k_n\}$ is an arbitrary sequence of positive real number and $\mathbb{1}$ is the $N \times N$ matrix whose entries are all 1. The following lemma estimates I1.

**Lemma 16.** $\mathbb{P}(\|\rho_n\| \geq k_n) \leq \frac{C_4}{\sqrt[4]{n}}$, *where* $k_n = n^{-\frac{1}{4}}$ *and* $C_4$ *is a constant.*

*Proof.* From the proof of Theorem 5.41 of [VdV00] $\rho_n$ has the following expression

$$\begin{aligned}
\rho_n = \sqrt{n}(-V(L^\star))^{\frac{1}{2}} \bigg( &\mathbf{d}^2 \hat{\Phi}_n(L^\star) - \mathrm{E}(\mathbf{d}^2 \hat{\Phi}_n(L^\star)) \\
&+ \frac{1}{2}(\tilde{L} - L^\star)^T \mathbf{d}^3 \hat{\Phi}_n(L_n) \bigg) (\tilde{L} - L^\star),
\end{aligned} \tag{3.21}$$

where $L_n$ is a point on the line segment between $\tilde{L}$ and $L^\star$. To simplify notation, let $\theta$ denote

$$\bigg( \mathbf{d}^2 \hat{\Phi}_n(L^\star) - \mathrm{E}(\mathbf{d}^2 \hat{\Phi}_n(L^\star)) + \frac{1}{2}(\tilde{L} - L^\star)^T \mathbf{d}^3 \hat{\Phi}_n(L_n) \bigg) (\tilde{L} - L^\star).$$

Then

$$\begin{aligned}
\mathrm{E}\|\rho_n\|_F &= \mathrm{E}\|\sqrt{n}(-V(L^\star))^{\frac{1}{2}}\theta\|_F \\
&= \sqrt{n}\mathrm{E}\|(-V(L^\star))^{\frac{1}{2}}\theta\|_F \\
&\leq \sqrt{n}\mathrm{E}\|(-V(L^\star))^{\frac{1}{2}}\|_{op}\|\theta\|_2 \\
&= \sqrt{n\cdot\Lambda_{max}(-V)}\cdot\mathrm{E}\|\theta\|_2. \qquad (3.22)
\end{aligned}$$

$\|\cdot\|_{op}$ denotes the operator norm induced by $\mathcal{L}^2$ norm and $\Lambda_{max}$ denotes the largest eigenvalue. For the first inequality, we regard $\theta$ as an $N\times N$ column vector and $(-V(L^\star))^{\frac{1}{2}}$ is an $(N\times N)\times(N\times N)$ matrix.

$$\begin{aligned}
\mathrm{E}\|\phi\|_2 =&\mathrm{E}\big\|\big(\mathbf{d}^2\,\hat{\Phi}_n(L^\star)-\mathrm{E}(\mathbf{d}^2\,\hat{\Phi}_n(L^\star))+\frac{1}{2}(\tilde{L}-L^\star)^T\,\mathbf{d}^3\,\hat{\Phi}_n(L_n)\big)(\tilde{L}-L^\star)\big\|_2 \\
\leq&\mathrm{E}\big\|\big(\mathbf{d}^2\,\hat{\Phi}_n(L^\star)-\mathrm{E}(\mathbf{d}^2\,\hat{\Phi}_n(L^\star))(\tilde{L}-L^\star)\big\|_2 \qquad \text{(I1-1)} \\
&+\mathrm{E}\big\|\frac{1}{2}(\tilde{L}-L^\star)^T\,\mathbf{d}^3\,\hat{\Phi}_n(L_n)\big)(\tilde{L}-L^\star)\big\|_2 \qquad \text{(I1-2)}
\end{aligned}$$

Using Cauchy-Schwartz inequality to estimate I1-1:

$$\begin{aligned}
\text{I1-1} &\leq \mathrm{E}^{\frac{1}{2}}\big\|\big(\mathbf{d}^2\,\hat{\Phi}_n(L^\star)-\mathrm{E}(\mathbf{d}^2\,\hat{\Phi}_n(L^\star))\big\|_{op}^2\mathrm{E}^{\frac{1}{2}}\|\tilde{L}-L^\star\|_2^2 \\
&\leq \frac{N^2}{\sqrt{n}}\max_{i,j}(L^{\star-1})_{ij}^2\mathrm{E}^{\frac{1}{2}}\|\tilde{L}-L^\star\|_2^2. \qquad (3.23)
\end{aligned}$$

Let $h(x)$ be a multivariate function:

$$\begin{aligned}
h:\quad &\mathbb{R}^{N\times N}\longrightarrow\mathbb{R} \\
(x_1,x_2,...,x_{NN})&\longmapsto x_1^2+x_2^2+\cdots+x_{NN}^2
\end{aligned}$$

Then $h$ is a continuous function. What's more almost surely $\tilde{L}\in E_{\alpha,\beta}$, which is a compact and convex set. Using Theorem 14 and portmanteau lemma we have

$$\mathrm{E}\big(h(\sqrt{n}(\tilde{L}-L^\star))\big)=n\mathrm{E}\|\tilde{L}-L^\star\|_F^2\longrightarrow\mathrm{E}\|\tilde{Z}\|_F^2, \qquad (3.24)$$

where $\tilde{Z}\sim\mathcal{N}(\mathbf{0},-V(L^\star))$. $\mathrm{E}\|\tilde{Z}\|_F^2$ is equal to $\mathrm{E}(\tilde{Z}_{11}^2+\cdots+\tilde{Z}_{1n}^2+\tilde{Z}_{21}^2+\cdots+$

$\tilde{Z}^2_{nn}) = \mathrm{Tr}(-V(L^\star))$. Then there exists a constant $C_1$ subject to $\alpha, \beta$ such that

$$\mathrm{E}^{\frac{1}{2}} \|\tilde{L} - L^\star\|_2^2 \leq C_1 \frac{1}{\sqrt{n}}. \tag{3.25}$$

As a result,

$$\text{I1-1} \leq C_2 N^2 \frac{1}{n} \tag{3.26}$$

where $C_2$ is a suitable constant.

Next, we estimate the second part, that is I1-2:

$$\mathrm{E}\|\frac{1}{2}(\tilde{L} - L^\star)^T \,\mathbf{d}^3\, \hat{\Phi}_n(L_n)\big)(\tilde{L} - L^\star)\big\|_2.$$

Here $\mathbf{d}^3\, \hat{\Phi}_n(L_n)$ is an $N \times N$ dimensional column vector whose entries are $N \times N$ matrices. Since $\hat{\Phi}(L)$ is infinitely many differentiable, $L_n$ is on the line segement between $\tilde{L}$ and $L^\star$, and $E_{\alpha,\beta}$ is a convex and compact set, we conclude that every entry of $\mathbf{d}^3\, \hat{\Phi}_n(L_n)$ is bounded. Hence there exists a constant $C_3 \geq 0$ such that

$$\begin{aligned} \mathrm{E}\|\frac{1}{2}(\tilde{L} - L^\star)^T \,\mathbf{d}^3\, \hat{\Phi}_n(L_n)\big)(\tilde{L} - L^\star)\big\|_2 \leq & C_3 \mathrm{E}\|\tilde{L} - L^\star\|_2^2 \\ \leq & \frac{C_1^2 C_3}{n}. \end{aligned} \tag{3.27}$$

Now let $k_n = n^{-\frac{1}{4}}$. Using Chebyshev's inequality we get:

$$\mathbb{P}(\|\rho_n\| \geq k_n) \leq \frac{\mathrm{E}\|\rho_n\|}{k_n} = \frac{C_4}{\sqrt[4]{n}} \tag{3.28}$$

for a suitable constant $C_4$. □

Our next lemma estimates I2 as follows.

**Lemma 17.** *Let $k_n$ be $\frac{1}{\sqrt[4]{n}}$. Then I2 $\leq \frac{C_7}{\sqrt[4]{n}}$ for some constant $C_7$*

*Proof.* Because

$$\begin{aligned}
&\mathbb{P}(X_n + k_n\mathbb{1} < x) - \mathbb{P}(Z < x) \\
&\geq \mathbb{P}(X_n + k_n\mathbb{1} < x, \|\rho_n\|_F < k_n) - \mathbb{P}(Z < x) \\
&= \big(\mathbb{P}(X_n + k_n\mathbb{1} < x) - \mathbb{P}(X_n + k_n\mathbb{1} < x, \|\rho_n\|_F \geq k_n)\big) - \mathbb{P}(Z < x) \\
&\geq \mathbb{P}(X_n + k_n\mathbb{1} < x) - \mathbb{P}(\|\rho_n\|_F > k_n) - \mathbb{P}(Z < x),
\end{aligned}$$

we have

$$\begin{aligned}
I2 &\leq |\mathbb{P}(X_n + k_n\mathbb{1} < x) - \mathbb{P}(Z < x)| \\
&\quad + |\mathbb{P}(X_n + k_n\mathbb{1} < x) - \mathbb{P}(|\rho_n\|_F \geq k_n) - \mathbb{P}(Z < x)| \\
&\leq 2|\mathbb{P}(X_n + k_n\mathbb{1} < x) - \mathbb{P}(Z < x)| + \mathbb{P}(\|\rho_n\|_F \geq k_n) \\
&= 2|\mathbb{P}(X_n + k_n\mathbb{1} < x) - \mathbb{P}(Z + k_n\mathbb{1} < x) \\
&\quad + \mathbb{P}(Z + k_n\mathbb{1} < x) - \mathbb{P}(Z < x)| + \mathbb{P}(\|\rho_n\|_F \geq k_n) \\
&\leq 2|\mathbb{P}(X_n + k_n\mathbb{1} < x) - P(Z + k_n\mathbb{1} < x)| \qquad \text{(I2-1)} \\
&\quad + 2|\mathbb{P}(Z + k_n\mathbb{1} < x) - P(Z < x)| \qquad \text{(I2-2)} \\
&\quad + \mathbb{P}(\|\rho_n\|_F \geq k_n). \qquad \text{(I2-3)}
\end{aligned}$$

By multidimensional Berry-Essen theorem in [Ben05],

$$\text{I2-1} \leq C_5 \cdot \sqrt{N} \cdot n \cdot \mathrm{E}\|\xi_1\|^3 \tag{3.29}$$

where $C_5$ is a constant and $\xi_1$ is defined in (3.20):

$$\begin{aligned}
\mathrm{E}\|\xi_1\|^3 &= \mathrm{E}\|\frac{1}{\sqrt{n}}(-V(L^\star))^{-\frac{1}{2}}(L^\star_{Z_i})^{-1} - (I + L^\star)^{-1}\|^3 \\
&\leq (\frac{1}{\sqrt{n}})^3\mathrm{E}\|(-V(L^\star))^{-\frac{1}{2}}\big((L^\star_{Z_i})^{-1} - (I + L^\star)^{-1}\big)\|^3. \qquad (3.30)
\end{aligned}$$

Since $\mathrm{E}\|(-V(L^\star))^{-\frac{1}{2}}\big((L^\star_{Z_i})^{-1} - (I + L^\star)^{-1}\big)\|^3$ is a constant we get

$$\text{I2-1} \leq C_6\sqrt{\frac{N}{n}} \tag{3.31}$$

For (I2-2), since Z can be viewed as a standard Guassian random vector, we have:

$$\begin{aligned} \text{I2-2} &= 2|\mathbb{P}(x - k_n I < Z_n < x)| \\ &\leq 2\sum_{i,j=1}^{N} \mathbb{P}(x_{ij} - k_n \leq (Z_n)_{ij} \leq x_{ij}) \\ &= \frac{2N^2}{\sqrt{2\pi}} k_n \end{aligned} \tag{3.32}$$

Combining 3.31, 3.32 with lemma 16, where we take $k_n = n^{-\frac{1}{4}}$ we conclude that:

$$I2 \leq \frac{C_7}{\sqrt[4]{n}},$$

where $C_5$ is a constant. □

As for I3 we can use the same argument as above and get that I3 is less than $C_8 \cdot n^{-\frac{1}{4}}$ for some constant $C_8$.

*Proof.* [Theorem 15]
The result follows from the last two lemmas. □

## 3.4 Two-by-two block kernel

In this section we show that if the kernels of determinantal point processes are two-by-two symmetric positive semi-definite matrices, the maximum likelihood estimators can be solved analytically. This result immediately extends to any two by two block matrices. However, the first method used in two by two kernel is difficult to apply to higher dimensional kernel.

Let $Z \sim \mathrm{DPP}(L^\star)$, where $L^\star = \begin{pmatrix} a^* & b^* \\ b^* & c^* \end{pmatrix}$, and the ground set be $\mathcal{Y} = [2]$. For our purpose, we assume

$$a^*, c^* > 0$$

and

$$a^*c^* - b^{*2} \geq 0.$$

For ease of notation, let $\hat{p}_0, \hat{p}_1, \hat{p}_2, \hat{p}_3$ denote the empirical probability of the subset $\{\emptyset\}, \{1\}, \{2\}, \{1,2\}$ respectively and let $p_0, p_1, p_2, p_3$ denote the theoretical probability respectively . Then its likelihood function defined in 3.1 is

$$\begin{aligned}\hat{\Phi}(L) &= \sum_{J \in [2]} \hat{p}_J \log(L_J) - \log\det(L+I) \\ &= \hat{p}_1 \log a + \hat{p}_2 \log c + \hat{p}_3 \log(ac - b^2) - \log[(a+1)(c+1) - b^2] \qquad (3.33)\end{aligned}$$

To find the critical point we first let the partial derivative of $\hat{\Phi}(L)$ with respect to $b$ equal zero and get

$$\frac{\partial \hat{\Phi}(L)}{\partial b} = \frac{2\hat{p}_3 b}{ac - b^2} + \frac{2b}{(a+1)(c+1) - b^2} = 0. \qquad (3.34)$$

Then we have $b$ is either equal to 0 or

$$b^2 = \frac{ac - (a+1)(c+1)\hat{p}_3}{1 - \hat{p}_3}. \qquad (3.35)$$

We can always assume $b$ is non-negative since by identifiability of DPPs, $\begin{pmatrix} a & b \\ b & c \end{pmatrix}$ and $\begin{pmatrix} a & -b \\ -b & c \end{pmatrix}$ give the same distribution. If $b = 0$, then by setting the partial derivative with respect to $a$ and $c$ to zero and notice that $\hat{p}_0 + \hat{p}_1 + \hat{p}_2 + \hat{p}_3 = 1$ we get the first critical point:

$$(\hat{a}, \hat{b}, \hat{c}) = \left( \frac{\hat{p}_1 + \hat{p}_3}{\hat{p}_0 + \hat{p}_2}, 0, \frac{\hat{p}_2 + \hat{p}_3}{\hat{p}_0 + \hat{p}_1} \right). \qquad (3.36)$$

This critical point exists only if $\hat{p}_0 + \hat{p}_2$ and $\hat{p}_0 + \hat{p}_1$ is nonzero. Since empirical probability converges to its corresponding theoretical probability almost surely and $p_0 > 0$, the strong law of large numbers implies the critical point exists almost surely.

If it is the other case, plugging (3.35) into $\hat{\Phi}(L)$ yields

$$\hat{\Phi}(L) = \hat{p}_1 \log a + \hat{p}_2 \log c + (\hat{p}_3 - 1)\log(a+c+1) - (\hat{p}_3 - 1)\log \frac{\hat{p}_3}{1-\hat{p}_3} + \log \hat{p}_3. \tag{3.37}$$

Let $\frac{\partial \hat{\Phi}(L)}{\partial a}$ and $\frac{\partial \hat{\Phi}(L)}{\partial c}$ equal zero we find

$$\frac{\partial \hat{\Phi}(L)}{\partial a} = \frac{\hat{p}_1}{a} + \frac{\hat{p}_3 - 1}{a+c+1} = 0 \tag{3.38}$$

$$\frac{\partial \hat{\Phi}(L)}{\partial c} = \frac{\hat{p}_2}{c} + \frac{\hat{p}_3 - 1}{a+c+1} = 0. \tag{3.39}$$

The above equations and (3.35) yield

$$(\hat{a}, \hat{b}, \hat{c}) = \Big(\frac{\hat{p}_1}{\hat{p}_0}, \frac{\sqrt{\hat{p}_1\hat{p}_2 - \hat{p}_0\hat{p}_3}}{\hat{p}_0}, \frac{\hat{p}_2}{\hat{p}_0}\Big), \tag{3.40}$$

from which we have this critical point exists only if $\hat{p}_0 > 0$ and $\hat{p}_1\hat{p}_2 - \hat{p}_0\hat{p}_3 \geq 0$. Again by strong laws of large numbers, the second critical point also exists and converges to the true value almost surely. In fact almost surely,

$$\frac{\hat{p}_1}{\hat{p}_0} \to \frac{p_1}{p_0} = a^*, \quad \frac{\sqrt{\hat{p}_1\hat{p}_2 - \hat{p}_0\hat{p}_3}}{\hat{p}_0} \to \frac{\sqrt{p_1 p_2 - p_0 p_3}}{p_0} = b^*, \quad \frac{\hat{p}_2}{\hat{p}_0} \to c^*.$$

Furthermore, we establish the central limit theorem for the estimator 3.40, which corresponds to the result in Theorem 14.

**Theorem 18.** *Assume $b > 0$, then the estimator $(\hat{a}, \hat{b}, \hat{c})$ in 3.40 is asymptotically normal,*

$$\sqrt{n}((\hat{a}, \hat{b}, \hat{c}) - (a^*, b^*, c^*)) \xrightarrow[n \longrightarrow \infty]{} \mathcal{N}(\mathbf{0}, -V(a^*, b^*, c^*)), \tag{3.41}$$

*where the convergence holds in distribution and $V(a^*, b^*, c^*)$ is the inverse of the Hessian matrix of the expected maximum likelihood function $\Phi(a, b, c) = p_1 \log a + p_2 \log c + p_3 \log(ac - b^2) - \log[(a+1)(c+1) - b^2]$.*

*Proof.* Let $Z_1, ..., Z_n$ be n independent copies of $Z \sim \mathrm{DPP}(L^*)$, where $L^* = \begin{pmatrix} a^* & b^* \\ b^* & c^* \end{pmatrix}$. Let $X_i$ be the random vector $(\mathbb{I}_{\{Z_i = \emptyset\}}, \mathbb{I}_{\{Z_i = \{1\}\}}, \mathbb{I}_{\{Z_i = \{2\}\}}, \mathbb{I}_{\{Z_i = \{1,2\}\}})^T$, where $\mathbb{I}_{\{\cdot\}}$ stands for the indicator random variable. Then $X_i$ has mean $\boldsymbol{\mu} = (p_0, p_1, p_2, p_3)^T$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} p_0 - p_0^2 & -p_0p_1 & -p_0p_2 & -p_0p_3 \\ -p_0p_1 & p_1 - p_1^2 & -p_1p_2 & -p_1p_3 \\ -p_0p_2 & -p_1p_2 & p_2 - p_2^2 & -p_2p_3 \\ -p_0p_3 & -p_1p_3 & -p_2p_3 & p_3 - p_3^2 \end{pmatrix}$$

By central limit theorem, $\sqrt{n}(\overline{X}_n - \boldsymbol{\mu})$ converges to a multivariate distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$. Let a function $g : \mathbb{R}^4 \to \mathbb{R}^3$ be such that

$$g(x_1, x_2, x_3, x_4) = (\frac{x_2}{x_1}, \frac{\sqrt{x_2x_3 - x_1x_4}}{x_1}, \frac{x_3}{x_1}).$$

Its Jacobi matrix $\dot{g}(\boldsymbol{x}) = \left[\frac{\partial g_i}{\partial x_j}\right]_{3\times 4}$ is given by

$$\begin{pmatrix} -\frac{x_2}{x_1^2} & \frac{1}{x_1} & 0 & 0 \\ -\frac{x_4}{2x_1\sqrt{x_2x_3-x_1x_4}} - \frac{\sqrt{x_2x_3-x_1x_4}}{x_1^2} & \frac{x_3}{2x_1\sqrt{x_2x_3-x_1x_4}} & \frac{x_2}{2x_1\sqrt{x_2x_3-x_1x_4}} & -\frac{1}{2\sqrt{x_2x_3-x_1x_4}} \\ -\frac{x_3}{x_1^2} & 0 & \frac{1}{x_1} & 0 \end{pmatrix}.$$

Now we are in the position to apply Delta method [VH12]; we have

$$\sqrt{n}\big((\hat{a}, \hat{b}, \hat{c}) - (a^*, b^*, c^*)\big) = \sqrt{n}\big(g(\overline{X}_n) - g(\boldsymbol{\mu})\big) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \dot{g}(\boldsymbol{\mu})\boldsymbol{\Sigma}\dot{g}(\boldsymbol{\mu})').$$

After tedious matrix computations, $\dot{g}(\boldsymbol{\mu})\boldsymbol{\Sigma}\dot{g}(\boldsymbol{\mu})'$ is found to be

$$D \begin{pmatrix} (a^* + a^{*2}) & (\frac{a^*c^*}{2b^*} + a^*b^* + \frac{a^*}{2b^*}(a^*c^* - b^{*2})) & a^*c^* \\ (\frac{a^*c^*}{2b^*} + a^*b^* + \frac{a^*}{2b^*}(a^*c^* - b^{*2})) & \frac{\frac{a^*c^*}{b^{*2}} - 1}{4}D + \frac{a^* + c^* + 4a^*c^*}{4} & \frac{a^*c^*}{2b^*} + b^*c^* + \frac{c^*}{2b^*}(a^*c^* - b^{*2}) \\ a^*c^* & \frac{a^*c^*}{2b^*} + b^*c^* + \frac{c^*}{2b^*}(a^*c^* - b^{*2}) & c^* + c^{*2} \end{pmatrix},$$

where $D = (a^* + 1)(c^* + 1) - b^{*2}$. It is straightforward to verify the above matrix is the inverse of the Hessian matrix of the expected maximum likelihood function $\Phi(L)$, that is, $-V(a^*, b^*, c^*)$, which in turn verifies theorem 14. However, in this two-by-two case, our maximum likelihood estimator is unique with no need of the definition 3.17. □

Now if $L^\star$ is a matrix with k two-by-two blocks $J_1, ..., J_k$

$$\begin{pmatrix} a_1 & b_1 & & & & & \\ b_1 & c_1 & & & & & \\ & & a_2 & b_2 & & & \\ & & b_2 & c_2 & & & \\ & & & & \ddots & & \\ & & & & & a_k & b_k \\ & & & & & b_k & c_k \end{pmatrix}, \tag{3.42}$$

where for each $1 \leq i \leq k$, $a_i$, $b_i$, $c_i > 0$ and $a_i c_i - b_i^2 \geq 0$. Let ground set $\mathcal{Y}$ of this DPP be $\{J_1^1, J_1^2, J_2^1, J_2^2, ..., J_k^1, J_k^2\}$ and for each $1 \leq i \leq k$,

$$\hat{p}_{J_i}^0 = \frac{1}{n} \sum_{m=1}^{n} \mathbb{I}\{J_i^1 \notin Z_m, J_i^2 \notin Z_m\} \tag{3.43}$$

$$\hat{p}_{J_i}^1 = \frac{1}{n} \sum_{m=1}^{n} \mathbb{I}\{J_i^1 \in Z_m, J_i^2 \notin Z_m\} \tag{3.44}$$

$$\hat{p}_{J_i}^2 = \frac{1}{n} \sum_{m=1}^{n} \mathbb{I}\{J_i^1 \notin Z_m, J_i^2 \in Z_m\} \tag{3.45}$$

$$\hat{p}_{J_i}^3 = \frac{1}{n} \sum_{m=1}^{n} \mathbb{I}\{J_i^1 \in Z_m, J_i^2 \in Z_m\}, \tag{3.46}$$

where $Z_1, ..., Z_n$ are n independent copies drawn from DPP($L^\star$). By Proposition 8, $Z \cap J_1, ..., Z \cap J_k$ are mutually independent, then the result of critical point for two by two matrix can be applied:

$$(\hat{a}_i, \hat{b}_i, \hat{c}_i) = \left( \frac{\hat{p}_{J_i}^1}{\hat{p}_{J_i}^0}, \frac{\sqrt{\hat{p}_{J_i}^1 \hat{p}_{J_i}^2 - \hat{p}_{J_i}^0 \hat{p}_{J_i}^3}}{\hat{p}_{J_i}^0}, \frac{\hat{p}_{J_i}^2}{\hat{p}_{J_i}^0} \right), \tag{3.47}$$

for every $1 \leq i \leq k$.

However the first order method is fraught with difficulties when the kernel has dimension higher than 2. For example, if the kernel is a $3 \times 3$ matrix

$$\begin{pmatrix} a & d & e \\ d & b & f \\ e & f & c \end{pmatrix},$$

we let the gradient of likelihood estimation function $\hat{\Phi}(L)$ equal zero:

$$\mathrm{d}\hat{\Phi}(L) = \sum_{J \subseteq [3]} \hat{p}_J L_J^{-1} - (L+I)^{-1} = 0.$$

Computing $L^{-1}$ and $(L+I)^{-1}$ can be troublesome. For example, $L^{-1}$ is:

$$\frac{1}{a(bc-f^2) - d(cd-ef) + e(df-be)} \begin{pmatrix} bc-f^2 & -cd+ef & -be+df \\ -cd+ef & ac-e^2 & de-af \\ -be+df & de-af & ab-d^2 \end{pmatrix}$$

which is difficult to use to solve for the true values.

Let the ground set be $[N]$. Now we bypass the difficulty by only focusing on all the two by two principal minors of the kernel. For all J such that $|J| \leq 1$, we let

$$\frac{\det(L_J)}{\det(L+I)} = \hat{p}_J, \tag{3.48}$$

where the left side is the theoretical probability of $J$ and the right side the empirical probability of $J$. We solve for all the diagonal elements of $L$

$$L_{ii} = \frac{\hat{p}_i}{\hat{p}_0}. \tag{3.49}$$

Then using equations (3.48) for $|J| = 2$ again we are able to determine the off-diagonal elements up to the sign

$$L_{ij}^2 = \frac{\hat{p}_i \hat{p}_j - \hat{p}_\emptyset \hat{p}_{\{i,j\}}}{\hat{p}_\emptyset^2}, \tag{3.50}$$

where $i \neq j$. Notice that this is the maximum likelihood estimator when $L$ is two dimensional. The recovery of signs of the off-diagonal elements has been solved by [UBMR17] using graph theory.

It is worth noting all the diagonal elements of the corresponding marginal kernel $K$ can be obtained from maximum likelihood estimation. Let the gradient of $\hat{\Phi}(L)$ equal zero:

$$\mathrm{d}\hat{\Phi}(L) = \sum_{J\subseteq[N]} \hat{p}_J L_J^{-1} - (L+I)^{-1} = 0.$$

Then moving $(L+I)^{-1}$ to the right side and multiplying both sides by $L$ yield:

$$\sum_{J\subseteq[N]} \hat{p}_J L L_J^{-1} = L(L+I)^{-1} = K, \tag{3.51}$$

from which we get

$$\hat{K}_{ii} = \sum_{\{i\}\subseteq J\subseteq[N]} \hat{p}_J,$$

for all $i = 1, 2, ..., N$. The above identity means that the maximum likelihood estimation of $K_{ii}$, the probability of inclusion of the item $i$ , is equal to the empirical marginal probability of the inclusion of item $i$.

# Chapter 4

# Conclusion

In this thesis, we first give a brief introduction to determinantal point processes (DPPs): definitions of marginal kernels and L ensembles, properties, examples in mathematics, and the rationale behind their applications in machine learning. Next we study their maximum likelihood estimation. Brunel et al show that the expected likelihood function $\Phi(L)$ is locally strongly concave around true value $L^\star$ if and only if $L^\star$ is irreducible, since the Hessian matrix of $\Phi(L)$ at $L^\star$ is negative definite. Then they prove the maximum likelihood estimator (MLE) is consistent in probability and when $L^\star$ is irreducible the MLE converges in distribution to a Gaussian random matrix. Based on their theorems, we show the MLE is also consistent almost surely; moreover, we find the $n^{-\frac{1}{4}}$ order bound on the rate of convergence of the MLE to normality. Last, we obtain the explicit form of the MLE where $L^\star$ is a two by two block matrix. The strong consistency and central limit theorem follows from the explicit form, which demonstrates the general strong consistency and central limit theorem proved earlier. It would be interesting to find the explicit form of higher dimensional DPPs. However, as the maximum likelihood learning of DPPs is proven to be NP-hard, the explicit form, even if was found, would be very difficult to compute.

# Bibliography

[AFAT14] Raja Hafiz Affandi, Emily Fox, Ryan Adams, and Ben Taskar. Learning the parameters of determinantal point process kernels. In *International Conference on Machine Learning*, pages 1224–1232. PMLR, 2014.

[BDF10] Alexei Borodin, Persi Diaconis, and Jason Fulman. On adding a list of numbers (and other one-dependent determinantal processes). *Bulletin of the American Mathematical Society*, 47(4):639–670, 2010.

[Ben05] Vidmantas Bentkus. A lyapunov-type bound in $R^d$. *Theory of Probability & Its Applications*, 49(2):311–323, 2005.

[BMRU17] Victor-Emmanuel Brunel, Ankur Moitra, Philippe Rigollet, and John Urschel. Maximum likelihood estimation of determinantal point processes. *arXiv preprint arXiv:1701.06501*, 2017.

[Bor08] Alexei Borodin. Loop-free Markov chains as determinantal point processes. In *Annales de l'IHP Probabilités et statistiques*, volume 44, pages 19–28, 2008.

[BP93] Robert Burton and Robin Pemantle. Local characteristics, entropy and limit theorems for spanning trees and domino tilings via transfer-impedances. *The Annals of Probability*, pages 1329–1371, 1993.

[DB18] Christophe Dupuy and Francis Bach. Learning determinantal point processes in sublinear time. In *International Conference on Artificial Intelligence and Statistics*, pages 244–257. PMLR, 2018.

[Gin65] Jean Ginibre. Statistical ensembles of complex, quaternion, and real matrices. *Journal of Mathematical Physics*, 6(3):440–449, 1965.

[GJWX22] Elena Grigorescu, Brendan Juba, Karl Wimmer, and Ning Xie. Hardness of maximum likelihood learning of DPPs. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3800–3819. PMLR, 02–05 Jul 2022.

[GKFT14] Jennifer A Gillenwater, Alex Kulesza, Emily Fox, and Ben Taskar. Expectation-maximization for learning determinantal point processes. *Advances in Neural Information Processing Systems*, 27, 2014.

[HKPV06] J Ben Hough, Manjunath Krishnapur, Yuval Peres, and Bálint Virág. Determinantal processes and independence. 2006.

[KT12] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012.

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

[LDG21] Claire Launay, Agnès Desolneux, and Bruno Galerne. Determinantal point processes for image processing. *SIAM Journal on Imaging Sciences*, 14(1):304–348, 2021.

[Mac75] O. Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122, 1975.

[MS15] Zelda Mariet and Suvrit Sra. Fixed-point algorithms for learning determinantal point processes. In *International Conference on Machine Learning*, pages 2389–2397. PMLR, 2015.

[PL21] Arnaud Poinas and Frédéric Lavancier. Asymptotic approximation of the likelihood of stationary determinantal point processes. *Scandinavian Journal of Statistics*, 2021.

[UBMR17] John Urschel, Victor-Emmanuel Brunel, Ankur Moitra, and Philippe Rigollet. Learning determinantal point processes with moments and cycles. In *International Conference on Machine Learning*, pages 3511–3520. PMLR, 2017.

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[VH12] Jay M Ver Hoef. Who invented the delta method? *The American Statistician*, 66(2):124–127, 2012.

[Wal49] Abraham Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949.