1-1-2013

Advances in the Theory of Determinantal Point Processes

Justin Kyle Rising
University of Pennsylvania, jkrising@gmail.com

Advances in the Theory of Determinantal Point Processes

Abstract
The theory of determinantal point processes has its roots in work in mathematical physics in the 1960s, but it is only in recent years that it has been developed beyond several specific examples. While there is a rich probabilistic theory, there are still many open questions in this area, and its applications to statistics and machine learning are still largely unexplored.

Our contributions are threefold. First, we develop the theory of determinantal point processes on a finite set. While there is a small body of literature on this topic, we offer a new perspective that allows us to unify and extend previous results.

Second, we investigate several new kernels. We describe these processes explicitly, and investigate the new discrete distribution which arises from our computations.

Finally, we show how the parameters of a determinantal point process over a finite ground set with a symmetric kernel may be computed if infinite samples are available. This algorithm is a vital step towards the use of determinantal point processes as a general statistical model.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Statistics

First Advisor
Lawrence D. Brown

Keywords
Determinantal Point Processes

Subject Categories
Mathematics | Statistics and Probability

ADVANCES IN THE THEORY OF DETERMINANTAL POINT PROCESSES

Justin K. Rising

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

_____

Lawrence D. Brown, Miers Busch Professor of Statistics

Graduate Group Chairperson

_____

Eric Bradlow, Professor of Marketing, Statistics and Education

Dissertation Committee
Lawrence D. Brown, Professor of Statistics
Dean P. Foster, Professor of Statistics
Dylan Small, Associate Professor of Statistics

ADVANCES IN THE THEORY OF DETERMINANTAL POINT PROCESSES

In memory of Larry Shepp.

ABSTRACT

ADVANCES IN THE THEORY OF DETERMINANTAL POINT PROCESSES

Justin K. Rising

Lawrence D. Brown

The theory of determinantal point processes has its roots in work in mathematical physics in the 1960s, but it is only in recent years that it has been developed beyond several specific examples. While there is a rich probabilistic theory, there are still many open questions in this area, and its applications to statistics and machine learning are still largely unexplored.

Our contributions are threefold. First, we develop the theory of determinantal point processes on a finite set. While there is a small body of literature on this topic, we offer a new perspective that allows us to unify and extend previous results.

Second, we investigate several new kernels. We describe these processes explicitly, and investigate the new discrete distribution which arises from our computations.

Finally, we show how the parameters of a determinantal point process over a finite ground set with a symmetric kernel may be computed if infinite samples are available. This algorithm is a vital step towards the use of determinantal point processes as a general statistical model.

Contents

List of Tables

List of Algorithms

Chapter 1

Introduction

In his biography of Paul Lévy, the renowned probabilist Michel Loève observes that

> "Martingales, Markov dependence and stationarity are the only three dependence concepts so far isolated which are sufficiently general and sufficiently amenable to investigation yet with a great number of deep properties." (Loève [1973])

Any probabilist giving the same inventory of sufficiently general interesting processes today would find that the list of candidates has not grown by much. In recent years it has become clear that determinantal point processes deserve a place on this list. In this dissertation, we provide an introduction to the theory of determinantal point processes and derive several new results that we hope will encourage others to begin work in this area.

The earliest examples of determinantal point processes appeared roughly fifty years ago (Dyson [1962a,b,c,d,e], Ginibre [1965], Karlin and McGregor [1959], Mehta and Gaudin [1960]). However, it was not until Macchi [1975] that they were identified as a class. Furthermore, it is only the past two decades that they have become objects of general interest.

Fortunately, what we have lost in time, we have not lost in activity. The recent mathematical literature on determinantal point processes is vast, and there are enough good surveys of this body of literature already that to attempt to improve on them would be folly. We refer the reader to any of Borodin [2011], Hough et al. [2006], Johansson [2005b], König [2005], Lyons [2003], Shirai and Takahashi [2003a,b], Soshnikov [2000, 2006], Tao [2009] for surveys that we have found helpful in preparing this document.

In the past few years, determinantal point processes have escaped from the clutches of mathematical physics and probability to become an important tool in the machine learning community (Affandi et al. [2012, 2013], Gillenwater et al. [2012a,b], Kulesza and Taskar [2011a,b,c, 2012]). This body of literature is small but growing, and many questions concerning determinantal point processes in general are sure to arise from the work done here.

In this dissertation, we will both address a problem arising from the work of machine learners and investigate some interesting examples of determinantal point processes. In Chapter 2, we will give a survey of the general theory of determinantal point processes, with emphasis on the finite case. In Chapter 3, we will consider the aforementioned interesting examples. Finally, in Chapter 4, we will consider the problem of estimating the parameters of a determinantal point process from data.

Chapter 2

A Brief Survey of Determinantal Point Processes

In this chapter, we give a definition of determinantal point processes, survey their basic properties, and give some standard examples. Although the general theory allows us to define determinantal point processes on a very large class of spaces, we will concentrate primarily on processes defined on finite sets and the real line so that we may avoid most of the measure-theoretic technicalities that the general theory demands.

2.1 Notation

We will be interested in the cardinality of finite sets, but not their specific elements. Therefore, we will use the set $[n] = \{1, 2, \ldots, n\}$ as our prototype of a finite set. We will also have occasion to use $\binom{[n]}{k}$ to denote the set of k-element subsets of $[n]$.

As may be expected from the name, a determinantal point process is somehow related to the determinant of a matrix. Although the general theory allows for the matrix of interest to be nonsymmetric, we will restrict our attention to the class

of processes which are determined by a symmetric matrix. Following **?**, we will use S^n to denote the set of $n \times n$ symmetric matrices, and S^n_{++} to denote the set of strictly positive definite symmetric matrices. Given a matrix $\boldsymbol{L}$ and a nonempty $\alpha \subseteq [n]$, we will write $\boldsymbol{L}_\alpha$ to denote the principal submatrix of $\boldsymbol{L}$ whose rows and columns are indexed by α. The corresponding principal minor is then given by $\det(\boldsymbol{L}_\alpha)$. We will not define $\boldsymbol{L}_\emptyset$ as a matrix, but we will observe the convention that $\det(\boldsymbol{L}_\emptyset) = 1$.

Finally, we will use Y to denote the realization of a point process.

2.2 General Point Processes

How can we rigorously define the notion of a random subset of a set $\mathcal{X}$? This is exactly the question that is addressed by the theory of point processes[1]. If the ground set $\mathcal{X}$ is finite, we can simply assign a probability to each subset of $\mathcal{X}$ and be done. If $\mathcal{X}$ is not finite, then we must introduce some measure-theoretic ideas to define exactly what we mean by a point process.

In the remainder of this brief section we will give the definition of a general point process. The reader who is only interested in the case of a finite ground set may skip to Section 2.3 with no loss of insight. We emphasize this only a cursory introduction to the theory, and refer the reader to Daley and Vere-Jones [2002, 2007] for an exhaustive treatment of the topic.

To define a general point process, we require that $\mathcal{X}$ to be a completely separable metric space[2]. Then the Borel sets $\mathcal{B}(\mathcal{X})$ are defined, and we can consider measures defined on these sets. We say that a measure μ defined on $\mathcal{B}(\mathcal{X})$ is boundedly finite if $\mu(A) < \infty$ for every bounded $A \in \mathcal{B}(\mathcal{X})$, and we define $\mathcal{N}^{\#}_{\mathcal{X}}$

[1] In an an unfortunate collsion of terminology, a point process is not generally a stochastic process. However, if $\mathcal{X} = \mathbb{R}$, then any point process is isomorphic to a binary-valued stochastic process. Point processes on a different set may therefore be viewed as the generalization of a binary-valued stochastic process with an appropriately chosen index set.

[2] We note that point processes can be defined on more general spaces, but we will not need this generality. We refer the reader to Soshnikov [2000] for the details.

to be the set of boundedly finite integer-valued measures on $\mathcal{B}(\mathcal{X})$. We will have need of the class of measures $\mu \in \mathcal{N}_{\mathcal{X}}^{\#}$ with the property that $\mu(\{x\}) \in \{0,1\}$ for every $x \in \mathcal{X}$, and we will refer to this class as $\mathcal{N}_{\mathcal{X}}^{\#*}$.

With some dexterity, it can be shown that $\mathcal{N}_{\mathcal{X}}^{\#}$ is also a completely separable metric space, and so we can define its Borel sets $\mathcal{B}\left(\mathcal{N}_{\mathcal{X}}^{\#}\right)$. We can now define the general point process N on $\mathcal{X}$ as a measurable map from some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $\left(\mathcal{N}_{\mathcal{X}}^{\#}, \mathcal{B}\left(\mathcal{N}_{\mathcal{X}}^{\#}\right)\right)$. We say that a point process is simple if $\mathbb{P}\left(N \in \mathcal{N}_{\mathcal{X}}^{\#*}\right) = 1$.

Intuitively, we have defined a set-valued random variable Y taking values in $2^{\mathcal{X}}$ such that $Y \cap B$ is almost surely finite for every bounded $B \subseteq \mathcal{X}$. In particular, when $\mathcal{X} = \mathbb{R}^d$, the number of points in any compact set is almost surely finite.

2.3 Determinantal Point Processes on a Finite Set

In this section, we will lay out the theory of determinantal point processes when the ground set is $[n]$. These processes were introduced in Exercises 5.4.7 and 5.4.8 of Daley and Vere-Jones [2002], but they have largely not been studied. We refer the reader to Borodin and Rains [2005], Kulesza and Taskar [2012], Lyons [2003], Lyons and Steif [2003] for an overview of the existing literature.

2.3.1 Definition and Elementary Properties

We will begin by defining an important subclass of determinantal point processes. We will give some elementary results on these processes, and then in Section 2.3.2 we will turn to the problem of defining general determinantal point processes.

In order to define a point process on a finite set, we must assign a probability to each subset. In this case, we will do so by fixing some $\boldsymbol{L} \in S_{++}^n$, and taking

$\mathbb{P}(Y = A) \propto \det(\boldsymbol{L}_A)$ for all $A \subseteq [n]$. This gives us our first examples of determinantal point processes. In the spirit of Macchi [1975], we will refer to $\boldsymbol{L}$ as the exclusion kernel of the determinantal point process.

In order to completely specify this distribution, we must determine its normalizing constant, which is done in the following theorem:

Theorem 2.3.1 (Kulesza and Taskar [2012])**.** *Let $\boldsymbol{L}$ be any matrix, and fix a set $A \subseteq [n]$. Given $J \subseteq [n]$, define $\boldsymbol{I}_J = \operatorname{diag}(\mathbf{1}\{i \in J\})$. Then*

$$\sum_{A \subseteq S \subseteq [n]} \det(\boldsymbol{L}_S) = \det(\boldsymbol{I}_{A^c} + \boldsymbol{L})$$

In particular, when $A = \emptyset$, we have that

$$\sum_{S \subseteq [n]} \det(\boldsymbol{L}_S) = \det(\boldsymbol{I} + \boldsymbol{L})$$

Theorem 2.3.1 gives us our normalizing constant, and we may now compute the probabilities assigned to any set by a determinantal point process. Throughout this chapter, we will take the determinant point process associated with

$$\boldsymbol{L} = \begin{bmatrix} \frac{17}{28} & \frac{3}{7} & \frac{3}{28} \\ \frac{3}{7} & \frac{5}{7} & \frac{3}{7} \\ \frac{3}{28} & \frac{3}{7} & \frac{17}{28} \end{bmatrix}$$

as our example. Table 2.1 gives the probability distribution associated with this process.

We will also need to be able to compute $\mathbb{P}(A \subseteq Y)$, the point process analogue of the survival function for real random variables. By Theorem 2.3.1, we have that

$$\mathbb{P}(A \subseteq Y) = \frac{\det(\boldsymbol{I}_{A^c} + \boldsymbol{L})}{\det(\boldsymbol{I} + \boldsymbol{L})}$$

S	$\mathbb{P}(Y = S)$
$\emptyset$	$\frac{7}{27}$
$\{1\}$	$\frac{17}{108}$
$\{2\}$	$\frac{5}{27}$
$\{3\}$	$\frac{17}{108}$
$\{1,2\}$	$\frac{7}{108}$
$\{1,3\}$	$\frac{5}{54}$
$\{2,3\}$	$\frac{7}{108}$
$\{1,2,3\}$	$\frac{1}{54}$

Table 2.1: Equality probabilities of the example determinantal point process.

With a bit of algebra, we can show the following theorem:

Theorem 2.3.2. *Borodin and Rains [2005] Let Y be distributed according to a determinantal point process with exclusion kernel $\boldsymbol{L}$. If we define $\boldsymbol{K} = \boldsymbol{L}(\boldsymbol{I} + \boldsymbol{L})^{-1}$, then $\mathbb{P}(A \subseteq Y) = \det(\boldsymbol{K}_A)$.*

We will refer to $\boldsymbol{K}$ as the kernel of the determinantal point process. The kernel of our example process defined above is given by

$$\boldsymbol{K} = \begin{bmatrix} \frac{1}{3} & \frac{1}{6} & 0 \\ \frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\ 0 & \frac{1}{6} & \frac{1}{3} \end{bmatrix}$$

and Table 2.2 shows the inclusion probabilities determined by this kernel. We note that $k_{13} = 0$ and $\mathbb{P}(\{1,3\} \subseteq Y) = \mathbb{P}(\{1\} \subseteq Y)\,\mathbb{P}(\{3\} \subseteq Y)$, which is not a coincidence. We will return to this phenomenon in the next section once we have a few more theorems under our belt.

If Y is a realization of a determinantal point process on $[n]$, we define the random vector X by $X_i = \mathbf{1}\{i \in Y\}$, which we will refer to as the indicator vector of Y. X_i takes values in $\{0,1\}$ and $\mathbb{P}(X_i = 1) = k_{ii}$, so we see that $X_i \sim \text{Bern}(k_{ii})$. $\mathbb{P}(X_i = 1, X_j = 1) = k_{ii}k_{jj} - k_{ij}^2$, so with a bit of calculation we can see that

S	$\mathbb{P}(S \subseteq Y)$
$\emptyset$	1
$\{1\}$	$\frac{1}{3}$
$\{2\}$	$\frac{1}{3}$
$\{3\}$	$\frac{1}{3}$
$\{1,2\}$	$\frac{1}{12}$
$\{1,3\}$	$\frac{1}{9}$
$\{2,3\}$	$\frac{1}{12}$
$\{1,2,3\}$	$\frac{1}{54}$

Table 2.2: Inclusion probabilities of the example determinantal point process.

$\mathrm{Cov}(X_i, X_j) = -k_{ij}^2$. The joint distribution of more than two of these variables is complicated, but we can make one interesting claim about it:

Theorem 2.3.3. *Let X be the indicator vector of a determinantal point process. The random variables $\{X_i\}_{i=1}^n$ are jointly independent if and only if* $\mathrm{Cov}(X_i, X_j) = 0$ *for every i, $j \in [n]$.*

Proof. If $\{X_i\}_{i=1}^n$ are jointly independent, then we immediately know that $\mathrm{Cov}(X_i, X_j) = 0$ for every i, $j \in [n]$. $\mathrm{Cov}(X_i, X_j) = -k_{ij}^2$, so if this is equal to zero for every i, $j \in [n]$, the kernel of the determinantal point process must be diagonal. In this case, all of its principal minors are the product of the appropriate set of diagonal entries, and we see immediately that

$$\mathbb{P}\left(\bigcap_{i \in I} X_i = 1\right) = \prod_{i \in I} \mathbb{P}(X_i = 1)$$

for all $I \subseteq [n]$. Therefore the components of X are jointly independent. □

Theorem 2.3.3 illustrates a general principle regarding determinantal point processes: because a matrix is function defined on pairs of elements of $[n]$, all the properties of a determinantal point process are determined by the pairwise interactions of the elements of $[n]$. This general principle evokes the multivariate Gaussian distribution, and we note that the Gaussian is the only standard dis-

tribution with the property given in Theorem 2.3.3. In the next section, we will see further similarities between multivariate Gaussians and determinantal point processes.

We close this section by computing the characteristic function of X. This is given by

$$\mathbb{E}e^{i\vec{t}^T X} = \sum_{S \subseteq [n]} \left(\prod_{s \in S} e^{it_s} \right) \det\left(\boldsymbol{L}_S\right)$$

With a simple application of Theorem 2.3.1, we can show that this is equal to $\det\left(\boldsymbol{I} + e^{i\,\mathrm{diag}\left(\vec{t}\right)}\boldsymbol{L}\right)$. A similar computation may be used to find the moment generating function.

2.3.2 L-ensembles and General Determinantal Point Processes

Let Y be distributed according to a determinantal process with exclusion kernel $\boldsymbol{L}$, and let $\boldsymbol{K}$ denote its kernel. We note that $\boldsymbol{K}$ and $\boldsymbol{L}$ have the same set of eigenvectors. Furthermore, if $\{\lambda_i\}_{i=1}^n$ are the eigenvalues of $\boldsymbol{L}$, then the eigenvalues of $\boldsymbol{K}$ are given by $\left\{\frac{\lambda_i}{1+\lambda_i}\right\}_{i=1}^n$. We have required that the eigenvalues of $\boldsymbol{L}$ be strictly positive, which implies that the eigenvalues of $\boldsymbol{K}$ fall in the open interval $(0, 1)$. However, we can define the inclusion probabilities given any $\boldsymbol{K}$ with all of its eigenvalues fall in the closed interval $[0, 1]$, and we will take this to be the definition of a general determinantal point process.

If we allow some of the eigenvalues of $\boldsymbol{L}$ to be zero, then the corresponding eigenvalues of $\boldsymbol{K}$ are zero as well. On the other hand, we can rewrite the definition of $\boldsymbol{K}$ to show that $\boldsymbol{L} = \boldsymbol{K}\left(\boldsymbol{I} - \boldsymbol{K}\right)^{-1}$. From here we see that if a determinantal point process has a kernel with eigenvalues equal to one, then it does not possess an exclusion kernel. The reason for this simple: any determinantal point process with an exclusion kernel must assign positive probability to the empty set, and not every determinantal point process has this property.

We choose to assume that $\boldsymbol{L}$ exists and is strictly positive definite because this will considerably simplify the results in the next section. We note that most authors define a determinantal point process by its inclusion probabilities, and that a determinantal point process which possesses an exclusion kernel is generally known as an $\boldsymbol{L}$-ensemble. There is no standard term for a determinantal point process which possesses a positive definite exclusion kernel, so we will refer to these as positive $\boldsymbol{L}$-ensembles.

This leaves us with the problem of generalizing our results on positive $\boldsymbol{L}$-ensembles to the full class of determinantal point processes. Our solution is to observe that the map from a matrix to its vector of principal minors is a continuous map between finite dimensional spaces. As a result, we can define the measures associated with a general determinantal point process as the limit of positive $\boldsymbol{L}$-ensembles. Most of the properties we are interested in will be preserved by this limit operation.

2.3.3 Closure Properties

In this section, we consider operations that produce new determinantal point processes from old. We will see that the class of determinantal point processes is closed under most things that we could want to do to a probability distribution.

We introduce one new piece of notation to make the statements of our theorems clearer. If Y is distributed according a determinantal point process with exclusion kernel $\boldsymbol{L}$, we will write $Y \sim \mathrm{LDPP}(\boldsymbol{L})$. Likewise, if Y is distributed according to a determinantal point process with kernel $\boldsymbol{K}$, we will write $Y \sim \mathrm{DPP}(\boldsymbol{K})$. In a minor abuse of notation, we will also write $X \sim \mathrm{LDPP}(\boldsymbol{L})$ and $X \sim \mathrm{DPP}(\boldsymbol{K})$ to specify the joint distribution of a set of Bernoullis.

We begin with the simplest result. Our first theorem shows that the marginal distributions of a determinantal point process are themselves determinantal point processes:

Theorem 2.3.4. *Let $X \sim \mathrm{DPP}(\boldsymbol{K})$, and fix $I \subseteq [n]$. Then $\{X_i\}_{i \in I} \sim \mathrm{DPP}(\boldsymbol{K}_I)$.*

Proof. If $A \subseteq I$, then $\mathbb{P}(A \subseteq Y \cap I) = \det(\boldsymbol{K}_A)$. Since this holds for any $A \subseteq I$, the theorem follows. □

Our second theorem shows that if the distribution of a random set is a determinantal point process, then the distribution of its complement is a determinantal point process as well.

Theorem 2.3.5. *Let $X \sim \mathrm{DPP}(\boldsymbol{K})$. Then $\vec{1} - X \sim \mathrm{DPP}(\boldsymbol{I} - \boldsymbol{K})$.*

Proof. We assume X follows a positive $\boldsymbol{L}$-ensemble and take limits for the general case. If $X \sim \mathrm{DPP}(\boldsymbol{K})$, then $X \sim \mathrm{LDPP}(\boldsymbol{L})$. We then compute:

$$\begin{aligned}
\mathbb{P}(Y = A^c) &= \frac{\det(\boldsymbol{L}_{A^c})}{\det(\boldsymbol{I} + \boldsymbol{L})} \\
&= \frac{\det(\boldsymbol{L}_{A^c})}{\det(\boldsymbol{L})\det(\boldsymbol{I} + \boldsymbol{L}^{-1})} \\
&= \frac{\det([\boldsymbol{L}^{-1}]_A)}{\det(\boldsymbol{I} + \boldsymbol{L}^{-1})}
\end{aligned}$$

Here we have used the fact that $\det([\boldsymbol{L}^{-1}]_A) = \dfrac{\det(\boldsymbol{L}_{A^c})}{\det(\boldsymbol{L})}$ (Borodin and Rains [2005]). We have shown that $\vec{1} - X \sim \mathrm{LDPP}(\boldsymbol{L}^{-1})$, and from here it is easy to see that $\vec{1} - X \sim \mathrm{DPP}(\boldsymbol{I} - \boldsymbol{K})$. □

Corollary 2.3.6. *Let $Y \sim \mathrm{DPP}(\boldsymbol{K})$. Then $\mathbb{P}(Y \subseteq A) = \det((\boldsymbol{I} - \boldsymbol{K})_{A^c})$.*

Finally, we consider the operation of conditioning on the presence or absence of some set of elements. We can show directly that conditioning on non-inclusion produces a new determinantal point process:

Theorem 2.3.7. *Kulesza and Taskar [2012] Let $X \sim \mathrm{LDPP}(\boldsymbol{L})$ and fix $I \subseteq [n]$. Then $X | Y \subseteq I^c \sim \mathrm{LDPP}(\boldsymbol{L}_{I^c})$.*

As a consequence of Theorems 2.3.5 and 2.3.7, we also see that $X | I \subseteq Y$ follows a determinantal point processes. The general form of this exclusion kernel was first

given in Borodin and Rains [2005] and we refer the reader to that paper for the specific expression.

We note that a similar conditioning property holds for determinantal point processes which are not $\boldsymbol{L}$-ensembles. The form of the kernel is complicated, and as we are merely interested in its existence, we do not state it.

The closure properties of determinantal point processes with respect to conditioning and marginalization are directly analogous to properties of the multivariate Gaussian distribution. We interpret the closure under complementation as an analogue of the fact that an affine transformation of any set of Gaussian random variables is also Gaussian.

We now return to our example determinantal point process and consider applying the above operations to it. We first observe that

$$\boldsymbol{I} - \boldsymbol{K} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{6} & 0 \\ -\frac{1}{6} & \frac{2}{3} & -\frac{1}{6} \\ 0 & -\frac{1}{6} & \frac{2}{3} \end{bmatrix}$$

and that

$$\boldsymbol{L}^{-1} = \begin{bmatrix} \frac{7}{2} & -3 & \frac{3}{2} \\ -3 & 5 & -3 \\ \frac{3}{2} & -3 & \frac{7}{2} \end{bmatrix}$$

Table 2.3 gives the equality and inclusion probabilities for this procces.

The marginal distributions are given by the principal submatrices of $\boldsymbol{K}$. We note that $X_i \sim \text{Bern}\left(\frac{1}{3}\right)$, that X_1 and X_3 are marginally independent, and that $\text{Cov}\left(X_1, X_2\right)$ and $\text{Cov}\left(X_2, X_3\right)$ are both equal to $-\frac{1}{36}$.

We now consider the joint distribution of X_1 and X_3 given the value of X_2. If

S	$\mathbb{P}(Y = S)$	$\mathbb{P}(S \subseteq Y)$
$\emptyset$	$\frac{1}{54}$	1
$\{1\}$	$\frac{7}{108}$	$\frac{2}{3}$
$\{2\}$	$\frac{5}{54}$	$\frac{2}{3}$
$\{3\}$	$\frac{7}{108}$	$\frac{2}{3}$
$\{1,2\}$	$\frac{17}{108}$	$\frac{5}{12}$
$\{1,3\}$	$\frac{5}{27}$	$\frac{4}{9}$
$\{2,3\}$	$\frac{17}{108}$	$\frac{5}{12}$
$\{1,2,3\}$	$\frac{7}{27}$	$\frac{7}{27}$

Table 2.3: Equality and inclusion probabilities of the complement of the example determinantal point process.

we condition on the event $X_2 = 0$, the new exclusion kernel is given by

$$\boldsymbol{L} = \begin{bmatrix} \frac{17}{28} & \frac{3}{28} \\ \frac{3}{28} & \frac{17}{28} \end{bmatrix}$$

and the corresponding kernel is given by

$$\boldsymbol{K} = \begin{bmatrix} \frac{3}{8} & \frac{1}{24} \\ \frac{1}{24} & \frac{3}{8} \end{bmatrix}$$

We see that conditioning on $X_2 = 0$ introduces a dependence between X_1 and X_3.

We now condition on the event $X_2 = 1$. In this case, the exclusion kernel is given by

$$\boldsymbol{L} = \begin{bmatrix} \frac{7}{20} & -\frac{3}{20} \\ -\frac{3}{20} & \frac{7}{20} \end{bmatrix}$$

and the corresponding kernel is given by

$$\boldsymbol{K} = \begin{bmatrix} \frac{1}{4} & -\frac{1}{12} \\ -\frac{1}{12} & \frac{1}{4} \end{bmatrix}$$

We now see that conditioning on either value of X_2 introduces a dependence between X_1 and X_3, but the exact nature of the induced dependence can vary with the value of X_2.

We can take advantage of this fact to construct determinantal point processes which contain conditional independencies. However, the conditional independencies in a determinantal point process can depend on the values of the variables being conditioned on. This is an important difference between determinantal point processes and graphical models, and deserves some attention.

2.3.4 The Random Variable $|Y|$

In the previous section, we have seen similarities between determinantal point processes and multivariate Gaussians. However, there is an important difference: a random set drawn according to a determinantal point process has a cardinality, and there is no corresponding concept for Gaussians. We now consider the random variable $|Y|$.

We begin by observing that $|Y| = \vec{1}^T X$. We can apply some straightforward calculations to this quantity to derive the following theorem:

Theorem 2.3.8. *Let* $Y \sim \mathrm{DPP}(\boldsymbol{K})$. *Then* $\mathbb{E}|Y| = \mathrm{tr}(\boldsymbol{K})$ *and* $\mathrm{Var}(|Y|) = \mathrm{tr}(\boldsymbol{K}) - ||\boldsymbol{K}||_F^2$.

However, we can make the following far stronger claim:

Theorem 2.3.9. *Hough et al. [2006] Let* $Y \sim \mathrm{DPP}(\boldsymbol{K})$, *and let* $\{\lambda_i\}_{i=1}^n$ *be the eigenvalues of* $\boldsymbol{K}$. *Then* $|Y| \stackrel{D}{=} \sum_{i=1}^{n} Z_i$, *where* $\{Z_i\}_{i=1}^n$ *are independent Bernoullis with* $\mathbb{E}Z_i = \lambda_i$.

We have two simple corollaries to Theorem 2.3.9. The first gives us bounds on $|Y|$:

Corollary 2.3.10. *If* $Y \sim \mathrm{DPP}(\boldsymbol{K})$, *then* $n - \mathrm{rank}(\boldsymbol{I} - \boldsymbol{K}) \leq |Y| \leq \mathrm{rank}(\boldsymbol{K})$.

The second is the observation that if $\boldsymbol{K} = \lambda \boldsymbol{I}$, then $|Y| \sim \mathrm{Binom}(n, \lambda)$. This suggests that determinantal point processes can represent interesting generalizations of the binomial distribution, and we will return to this issue in Chapter 3.

2.3.5 Mixture Representations

In this section, we consider the general determinantal point process as a mixture of other distributions. In both cases these distributions are concentrated on sets of a fixed size.

Our first mixture representation requires the idea of a k-DPP (Kulesza and Taskar [2011a, 2012]), which is the distribution achieved by sampling from a determinantal point process conditional on the event $|Y| = k$. Given this definition, it is trivial to see that a general determinantal point process can be written as a mixture of k-DPPs. While this may seem too simple to be interesting, there are two reasons to give it some consideration. First, calculating $\mathbb{P}(|Y| = k)$ is not a trivial problem. Second, the class of k-DPPs contains distributions which are not determinantal point processes.

This second point is worth exploring. If $Y \sim \mathrm{LDPP}(\boldsymbol{I})$, then Y is uniformly distributed over all subsets of $[n]$. A k-DPP generated from this distribution is then a uniform distribution over sets of size $\boldsymbol{K}$. As observed in Kulesza and Taskar [2012], this is not a determinantal point process unless $k \in \{0, 1, n-1, n\}$.

This leaves us with the problem of computing $\mathbb{P}(|Y| = k)$. While there is no closed form expression, the numerical value can be computed efficiently and accurately. To describe this computation, we must introduce the family of elementary symmetric polynomials. The kth elementary symmetric polynomial on n variables

is defined by

$$e_k(x_1, x_2, \ldots, x_n) = \sum_{S \in \binom{[n]}{k}} \prod_{s \in S} x_s$$

The elementary symmetric polynomials allow us to directly compute the probabilities we want:

Theorem 2.3.11. *Kulesza and Taskar [2011a] Let* $Y \sim \mathrm{LDPP}(\boldsymbol{L})$*, and let* $\{\lambda_i\}_{i=1}^n$ *be the eigenvalues of* $\boldsymbol{L}$*. Then*

$$\mathbb{P}(|Y| = k) = \frac{e_k(\lambda_1, \lambda_2, \ldots, \lambda_n)}{\det(\boldsymbol{I} + \boldsymbol{L})}$$

The elementary symmetric polynomials can in principle be computed by observing that $e_0(x_1, x_2, \ldots, x_n) = 1$ and

$$e_k(x_1, x_2, \ldots, x_n) = e_k(x_1, x_2, \ldots, x_{n-1}) + x_n e_{k-1}(x_1, x_2, \ldots, x_{n-1})$$

for $k > 0$. However, there are numerically superior algorithms, which are detailed in Baker and Harwell [1996], and we refer the reader to this survey for details.

Before we turn our attention away from k-DPPs, we note that they have similar closure properties to those discussed for general determinantal point processes in Section 2.3.3. As the details are slightly more complicated, we refer the reader to Kulesza and Taskar [2012] for details.

We now turn our attention to our second mixture representation. This requires the notion of a determinantal projection process, which is simply a determinantal point process whose kernel is an orthogonal projection matrix. As a result, all the eigenvalues of the kernel are in $\{0, 1\}$, and these processes are not in general $\boldsymbol{L}$-ensembles. However, every positive $\boldsymbol{L}$-ensemble can be written as a mixture of determinantal projection processes.

Theorem 2.3.12. *Kulesza and Taskar [2012] Let* $\mu_{\boldsymbol{\Lambda}}$ *be the measure associated with a determinantal point process with exclusion kernel* $\boldsymbol{\Lambda}$*. Let* $\boldsymbol{L}$ *be a symmetric*

Algorithm 1 Sample from a determinantal point process with kernel $\boldsymbol{K}$

Let $\{(\lambda_i, \vec{v}_i)\}_{i=1}^n$ be the eigendecomposition of $\boldsymbol{K}$
$S \leftarrow \emptyset$
for $i \in [n]$ **do**
  $S \leftarrow S \cup \{i\}$ with probability λ_i
end for
$V \leftarrow \{\vec{v}_s\}_{s \in S}$
$Y \leftarrow \emptyset$
while $|V| > 0$ **do**
  Select i from $[n]$ with probability $\dfrac{1}{|V|} \sum_{\vec{v} \in V} v_i^2$
  $Y \leftarrow Y \cup \{i\}$
  Replace V with an orthonormal basis of the subspace of V orthogonal to $\vec{e}_i$
end while
return Y

positive definite matrix, let $\{\lambda_i\}_{i=1}^n$ *be the eigenvalues of* $\boldsymbol{L}$*, and let* $\{\vec{v}_i\}_{i=1}^n$ *be the corresponding eigenvectors. Then*

$$\mu_{\boldsymbol{L}} = \frac{1}{\det(\boldsymbol{I} + \boldsymbol{L})} \sum_{S \subseteq [n]} \left(\prod_{s \in S} \lambda_s \right) \mu_{\boldsymbol{V}_S}$$

where $\boldsymbol{V}_S = \sum_{s \in S} \vec{v}_s \vec{v}_s^T$.

In light of Theorem 2.3.12, we see that the problem of sampling from a determinal point process reduces to the problem of selecting a determinantal projection process and sampling from it. Algorithm 1 performs this procedure, and we refer the reader to Kulesza and Taskar [2012] for the proof.

There is one interesting fact regarding Algorithm 1. We can regard the indicator vector of the set S as a collection of Bernoulli random variables, and it is natural to ask if we can infer these Bernoullis from the draws of the process. Surprisingly, the answer is no. These variables are in fact not measurable with respect to the observed draws. We refer the reader to Hough et al. [2006] for a simple example.

2.3.6 Interpretations

We close our study of determinantal point processes on finite sets by examining some interpretations of the entries of the kernel $\boldsymbol{K}$ and the probabilities assigned by the exclusion kernel $\boldsymbol{L}$. In order to do so, we will need the idea of a Gramian matrix, which is an $n \times n$ matrix $\boldsymbol{M}$ such that $m_{ij} = \vec{v}_i \vec{v}_j$ for some set of vectors $\{\vec{v}_i\}_{i=1}^n$. It is immediately obvious that every Gramian matrix is symmetric and positive semidefinite. It is less immediately obvious that every symmetric positive semidefinite matrix is a Gramian (Lanckriet et al. [2004]).

Following Kulesza and Taskar [2012], we begin by examining the entries of the kernel matrix $\boldsymbol{K}$ in light of the Gramian interpretation. $\boldsymbol{K}$ is symmetric and positive semidefinite, so there is some set of vectors $\{\vec{v}_i\}_{i=1}^n$ such that $k_{ij} = \vec{v}_i^T \vec{v}_j$. We decompose $\vec{v}_i$ as $q_i \vec{\phi}_i$ with $||\vec{\phi}_i||_2 = 1$. We will refer to q_i as the quality of i, and $\vec{\phi}_i$ as the feature vector of i. We can then write $k_{ij} = q_i q_j \vec{\phi}_i^T \vec{\phi}_j$.

We can assign interpretations to these quantities by examining the moments of $X \sim \text{DPP}(\boldsymbol{K})$. By the definition of $\boldsymbol{K}$, we have $\mathbb{P}(i \in Y) = k_{ii}$ and $\mathbb{P}(\{i, j\} \subseteq Y) = k_{ii}k_{jj} - k_{ij}^2$. We observe that $k_{ii} = q_i^2$, which implies that higher quality elements of our ground set are more likely to be selected than lower quality items.

In order to interpret k_{ij}, we note that $\vec{\phi}_i^T \vec{\phi}_j = \cos(\theta_{ij})$ with θ_{ij} the angle between the vectors $\vec{\phi}_i$ and $\vec{\phi}_j$. We can then simply calculate to see that $k_{ii}k_{jj} - k_{ij}^2 = q_i^2 q_j^2 \sin^2(\theta_{ij})$. This tells us that the angles $\{\theta_{ij}\}_{i,j=1}^n$ give us a measure of similarity between the items of our ground set. If the feature vectors corresponding to two items are parallel, they will not be selected together, and if the feature vectors are orthogonal, their indicators are independent. Furthermore, the transition from mutual exclusivity to independence is quadratic in the sine of the angle between the feature vectors.

We now consider $\boldsymbol{L}$ as a Gramian matrix. There are two relevant facts here. The first is that any principal submatrix $\boldsymbol{L}_A$ is the Gramian of the set of vectors $\{\vec{v}_i\}_{i \in A}$. The second is that the determinant of a Gramian matrix is equal to the

squared volume of a parallelotope whose sides are given by the vectors that make up the Gramian (Barth [1999]).

By examining $\boldsymbol{K}$ as a Gramian matrix, we were able to interpret the probabilities assigned by a determinantal point process in terms of pairs of items. However, this interpretation does not extend naturally to larger collections until we bring in the insight gained from examining $\boldsymbol{L}$ in the same way. We now see that a determinantal point process will favor sets of high quality items whose feature vectors are mutually far apart. Following Kulesza and Taskar [2012], we refer to this quality as diversity.

The ideas of quality and diversity have led machine learning researchers to experiment with using determinantal point processes to randomly select subsets which are somehow representative of the ground set, and the results so far are encouraging (Affandi et al. [2012, 2013], Gillenwater et al. [2012a], Kulesza and Taskar [2011a]). There is one further observation that suggests that this notion of representativeness is somehow naturally connected to determinantal point processes. Recall that if a random variable $W \sim \mathcal{N}(\vec{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \in S_{++}^n$, the entropy of W is given by $H(W) = \frac{1}{2}\log(\det(2\pi e\boldsymbol{\Sigma}))$. With some simple algebra, we can rearrange this to show that

$$\det(\boldsymbol{\Sigma}) = e^{2H(W) - n(\log(2\pi)+1)}$$

If $W \sim \mathcal{N}(0, \boldsymbol{L})$, then we have that

$$\mathbb{P}(Y = A) \propto e^{2H(W_A) - |A|(\log(2\pi)+1)}$$

Campbell [1966] gives an interpretation of exponential entropy as a measure of the spread of a distribution, which fits well with the volume interpretation given above. However, to the best of our knowledge, there is no standard interpretation of the ratio of exponential entry penalized by the cardinality in this manner.

We close this section by observing that the two interpretations above give us a relationship between the entropy of a Gaussian distribution, its dimension, and the log volume of a parallelotope whose sides are a set of vectors whose Gramian is the covariance matrix of the Gaussian distribution. It appears that there is more to this puzzle than has been uncovered to date.

2.4 Determinantal Point Processes on the Real Line

We now abandon our consideration of determinantal point processes on finite sets and consider how the same class of processes can be defined on more general spaces. There is a general theory which allows us to define α-determinantal processes on second countable locally compact Hausdorff spaces endowed with a nonnegative nonatomic Radon measure (Shirai and Takahashi [2003a]), but the price that must be paid in terms of attention to detail is high. Fortunately, the real line endowed with Lebesgue measure is such a space, and so we can study some interesting aspects of the general theory by considering determinantal point processes on this space. We note that the treatment here is elementary, and we refer the reader to Hough et al. [2006], Shirai and Takahashi [2003a,b] for more detailed expositions.

We remind the reader that, following the discussion in Section 2.2, a sample from a general point process on $\mathbb{R}$ has a finite intersection with any bounded set. Any point process with this property is referred to as finite. As it happens, the theory of finite point processes is technically simpler than the theory of locally finite point processes. As a result, we will be able to avoid dealing with certain details by defining determinantal point processes on compact subsets of $\mathbb{R}$ and then extending them to larger sets by the use of an appropriate limit theorem[3].

[3]We could directly define determinantal point processes on the real line by appeal to the

We will have need of some general concepts from the theory of finite point processes[4]. Let Y be a sample from such a process. The n-point correlation[5] functions of a point process, which are denoted $\rho^{(n)}(x_1, x_2, \ldots, x_n)$, give the probabilities of the events $\{\{x_i\}_{i=1}^n \subseteq Y\}$. The related concept of the n-point Janossy densities, which are denoted $j_n(x_1, x_2, \ldots, x_n)\ \mathrm{d}x_1\ \mathrm{d}x_2 \ldots\ \mathrm{d}x_n$, may be interpreted as the probability that there are exactly n points, and that they fall in the infinitesimal intervals $\{[x_i, x_i + \mathrm{d}x_i)\}$. As recorded in the following theorem, both of these quantities are sufficient to define a finite point process.

Theorem 2.4.1. *Daley and Vere-Jones [2002], Macchi [1975] A finite point process is completely determined by either its Janossy densities or its correlation functions.*

We now have sufficient machinery to define determinantal point processes on compact subsets of $\mathbb{R}$. We say that a process defined on a compact $C \subseteq \mathbb{R}$ is determinantal if there is some kernel K such that $\rho^{(n)}(x_1, x_2, \ldots, x_n) = \det\left([K(x_i, x_j)]_{i,j=1}^n\right)$. Our first result concerns the existence and uniqueness of the probability measure associated with a determinantal point process.

Theorem 2.4.2. *Shirai and Takahashi [2003a] Let $C \subseteq \mathbb{R}$ be compact, let μ denote Lebesgue measure, and let K be an integral operator on $L^2(C, \mu)$ which is bounded, locally trace class, and has no eigenvalues outside the interval $[0, 1]$. Then there is a unique determinantal measure whose correlation functions are given by the n-point principal minors of K.*

We will refer to kernels which satisfy the hypotheses of Theorem 2.4.2 as nice. We

point process version of Kolmogorov's consistency theorem. However, the theorems that we will use to build up the general case from determinantal point processes are of independent interest, and so we prefer this approach. The full details of the direct approach may be found in Daley and Vere-Jones [2007].

[4]In this section, we are assuming that the point processes we consider possess certain densities with respect to Lebesgue measure. We refer the reader to Daley and Vere-Jones [2002] for the general theory.

[5]In another unfortunate clash of terminology, the sense of the word "correlation" as used here is unrelated to its ordinary statistical meaning.

note that the converse of Theorem 2.4.2 is true as well, and we refer the reader to Soshnikov [2000] for its statement.

We now have the continuous analogue of the kernel $\boldsymbol{K}$, and we will now describe the corresponding analogue of the exclusion kernel $\boldsymbol{L}$. We begin by observing that in the finite case, the matrices $\boldsymbol{K}$ and $\boldsymbol{L}$ are related by the equation $\boldsymbol{L} = \boldsymbol{K} + \boldsymbol{L}\boldsymbol{K}$. Following Macchi [1975], we note that the obvious generalization of this matrix equation to the continuous case is the integral equation

$$L(s,t) = K(s,t) + \int_C L(s,u)K(u,t) \ \mathrm{d}u \tag{2.1}$$

which is generally known as the resolvent equation (Smithies [1958]). We are in luck, as this integral equation does define a kernel L whose principal minors are proportional to the Janossy densities of the determinantal point process with kernel K (Macchi [1975]). Furthermore, the relationship between the eigenvalues of K and L is exactly the same as in the finite case:

Theorem 2.4.3. *Macchi [1975] Let K be a nice kernel, and let L be the corresponding kernel defined by Equation 2.1. If $\{\lambda_i\}_{i=1}^{\infty}$ are the eigenvalues of K, then the corresponding eigenvalues of L are given by $\left\{ \dfrac{\lambda_i}{1-\lambda_i} \right\}_{i=1}^{\infty}$.*

However, there is one further similarity with the finite case. Equation 2.1 only possesses a solution if all of the eigenvalues of K are strictly less than one. Fortunately, there is a convergence theorem which allows us to consider limits of sequences of kernels whose eigenvalues have this property, and we will state it as Theorem 2.4.5.

In order to use the Janossy densities associated with a determinantal point processes, we must be able to compute the normalizing constant associated with them. By analogy with the finite ground set theory, we would expect to find something like $\det(\boldsymbol{I} + \boldsymbol{L})$. As it happens, the Fredholm determinant, which is the operator-theoretic analogue of this determinant, is exactly the quantity we

need.

While there is an elegant theory of Fredholm determinants as described in Reffgen [2003], we will have need of one important property. If T is a trace class operator with eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$, then the Fredholm determinant is equal to

$$\prod_{i=1}^{\infty}(1+\lambda_i)$$

Given that K is a trace class operator, we are guaranteed that L is a trace class operator as well by the combination of Theorem 2.4.3 and the following lemma:

Lemma 2.4.4. *Let $\{\lambda_i\}_{i=1}^{\infty}$ be such that each $\lambda_i \in [0,1]$ and $\sum_{i=1}^{\infty}\lambda_i < \infty$. Then $\sum_{i=1}^{\infty}\frac{\lambda_i}{1-\lambda_i} < \infty$ as well.*

Proof. $\sum_{i=1}^{\infty}\lambda_i < \infty$ implies that $\lambda_i \to 0$. As a result, there is some N such that for any $i > N$, we have that $\frac{1}{1-\lambda_i} < \frac{3}{2}$. This implies that

$$\sum_{i=1}^{\infty}\frac{\lambda_i}{1-\lambda_i} < \sum_{i=1}^{N}\frac{\lambda_i}{1-\lambda_i} + \frac{3}{2}\sum_{i=N+1}^{\infty}\lambda_i$$

Our conclusion follows immediately. □

We now consider the problem of defining determinantal point processes on noncompact subsets of $\mathbb{R}$. We can easily define them as limits of determinantal point processes defined on compact sets with the following theorem:

Theorem 2.4.5. *Shirai and Takahashi [2003a] Let $\{K_i\}_{i=1}^{\infty}$ be a sequence of nice kernels that converges uniformly to a kernel K on every compact set. Then K is nice, there is a unique determinantal measure μ associated with K, and the sequence of measures associated with each K_i converges weakly to μ.*

Our general strategy for applying this theorem will be to take some sequence of compact sets $\{C_i\}_{i=1}^{\infty}$ such that $C_i \subseteq C_{i+1}$ and $\lim_{i\to\infty} C_i = \mathbb{R}$. We can then take

a kernel K defined on $\mathbb{R}$ and define a sequence of kernels $\{K_i\}$ by restricting K to C_i. These kernels are nice and they trivially converge uniformly to K on every compact set, so the hypotheses of Theorem 2.4.5 are satisfied.

We now present some results from the general theory of determinantal point processes on the real line. Our first result generalizes Theorem 2.3.9 to continuous determinantal point processes:

Theorem 2.4.6. *Hough et al. [2006] Let Y be distributed according to a determinantal point process with kernel K, let C be a compact subset of $\mathbb{R}$, and let $\{\lambda_i\}_{i=1}^{\infty}$ be the eigenvalues of K restricted to C. Then $|Y| \stackrel{D}{=} \sum_{i=1}^{\infty} Z_i$, where $\{Z_i\}_{i=1}^{\infty}$ are independent Bernoullis with $\mathbb{E}Z_i = \lambda_i$.*

In particular, because K is locally trace class, the number of points in a compact set is almost surely finite by the Borel-Cantelli lemma. If K is in fact trace class on $\mathbb{R}$, then the total number of points in any realization is almost surely finite.

We have one further result regarding the distribution of the number of points in a draw from a determinantal point process. A simple application of the Lindberg-Feller central limit theorem combined with Theorem 2.4.6 gives us a central limit theorem for determinantal point processes:

Theorem 2.4.7. *Hough et al. [2006] Let $\{Y_i\}_{i=1}^{\infty}$ be a sequence of independent draws from a sequence of determinantal point processes with kernels in the sequence $\{K_i\}_{i=1}^{\infty}$. Let $\{S_i\}_{i=1}^{\infty}$ be a sequence of measurable subsets of $\mathbb{R}$ such that $\mathrm{Var}\left(|Y_i \cap S_i|\right) \to \infty$. Then*

$$\frac{|Y_i \cap S_i| - \mathbb{E}|Y_i \cap S_i|}{\sqrt{\mathrm{Var}\left(|Y_i \cap S_i|\right)}} \xrightarrow{d} \mathcal{N}(0,1)$$

There are other limit theorems for determinantal point processes, but their statements are complex. These theorems may be found in Shirai and Takahashi [2003a].

We note that we have not discussed the joint distribution of the number of points occuring in disjoint sets. There are some results in this direction–in partic-

ular, these counts will generally be dependent–but the theory is not simple. We refer readers to Hough et al. [2006] for the details.

From the discussion in Section 2.3, we have the intuition that determinantal point processes on finite sets capture negative associations between points of the ground set. Our next result gives us a precise statement of this idea for general determinantal point processes:

Theorem 2.4.8. *Shirai and Takahashi [2003a] Let Y be distributed according to a determinantal point process with kernel K, and let S_1, $S_2 \subseteq \mathbb{R}$. Then $\mathbb{P}(S_1 \cup S_2 \subseteq Y) \leq \mathbb{P}(S_1 \subseteq Y)\,\mathbb{P}(S_2 \subseteq Y)$.*

Our final result in the theory of determinantal point processes on the real line requires the notion of a renewal process. We give only the basic definition, and refer the reader to any of the standard textbooks (e.g., Durrett [2010], Feller [1968, 1971], Karlin and Taylor [1975], Ross [1996]) for details. If $\{X_i\}_{i=1}^{\infty}$ are independent and identically distributed nonnegative random variables, then we can define the sequence $\{S_i\}_{i=0}^{\infty}$ by $S_0 = 0$ and $S_{n+1} = S_n + X_{n+1}$. We then define the renewal process $N(t)$ by

$$N(t) = \sup_i \{S_i \leq t\}$$

We observe that the values $\{S_i\}_{i=0}^{\infty}$ form a point process on the real line, and that this is the basis of a rich theory that we will ignore. Instead, we give the following theorem, which addresses the question of when a determinantal point process on the right half-line is a renewal process:

Theorem 2.4.9. *Soshnikov [2000] A determinantal point process on $\mathbb{R}_+$ with kernel K is a renewal process if and only if the following two conditions are satisfied almost everywhere:*

1. *$x_1 \leq x_2 \leq x_3$ implies $K(x_1, x_2)K(x_2, x_3) = K(x_1, x_3)K(x_2, x_2)$.*

2. $x_1 \leq x_2$ *implies* $K(x_2, x_2) - \dfrac{K(x_1, x_2)K(x_2, x_1)}{K(x_1, x_1)} = u(x_2 - x_1)K(x_1, x_1)$ *for some function* u.

If a determinantal point process is a renewal process, then we can completely specify the joint distribution of its points by specifying the distribution of the length of the interval between zero and the first point.

We close this section by examining the relationship between determinantal point processes on finite or discrete sets and general determinantal point processes. We first show how finite set determinantal processes can be considered as a special case of determinantal point processes. We simply define the kernel of a determinantal point process to be zero everywhere except at a finite number of points, and we assign those values so that they evaluate to the entries of the appropriate matrix. In this way, we see that all of the results of this section apply to determinantal point processes on finite sets.

We now discuss an open problem in the theory of determinantal point processes. It is a standard result in probability theory that appropriately rescaled Markov chains converge to diffusions in a sense made precise in Kushner [1974], Turner [2002]. Is there a similar sense in which determinantal point processes or other processes on finite sets can be made to converge to determinantal point processes on continuous sets? There are specific examples of this phenomenon discussed in Borodin and Gorin [2009], Borodin and Olshanski [2007], Gorin [2008], Johansson [2005a], Olshanski [2008], but there is no general theory as of this writing.

2.5 Examples

We close our survey chapter with a collection of several standard examples of determinantal point processes. To do so, we must drop our assumption that the kernels of the processes we consider are symmetric, as this is not generally the

case. We note that our treatment in this section is not mean to be exhaustive, and we refer the reader who is interested in a more encyclopedic listing to Borodin [2011].

Perhaps the most famous example of a determinantal point processes is the distribution of the eigenvalues of a matrix drawn according to the Gaussian unitary ensemble. A discussion of the physics and random matrix theory involved in this phenomenon is beyond the scope of this dissertation, so we refer the reader to Johansson [2005b], Soshnikov [2000].

The first example of a determinantal point process was given in Karlin and McGregor [1959]. Assume that n independent copies of a continuous time birth-death process are started in distinct states, and condition on the event that no pair of the sample paths intersect. Then the joint distribution of the states at any time t follows a determinantal point process. In particular, random walks on the integers have this property. Generalizations to Brownian motions with an appropriate set of starting values have been given in Johansson [2004], Katori and Tanemura [2007].

Another standard example is due to Burton and Pemantle [1993]. Let G be a graph, and let T be a spanning tree of G chosen uniformly from the set of all spanning trees of G. Then the set of edges of T is distributed according to a determinantal point process.

The next example, given in **?**, requires the notion of a loop-free Markov chain. A Markov chain is said to be loop-free if the probability of visiting any state more than once is zero. We have the perhaps surprising result that finite subsets of the sample paths of a loop-free Markov chain on a discrete space are distributed according to a determinantal point process. As any process with almost surely positive independent and identically distributed increments is a loop-free Markov chain, we have a determinantal representation for this large class of processes.

Let $\{A_i\}_{i=0}^{\infty}$ be independent random variables distributed according to the stan-

dard complex Gaussian distribution, and define the function f on the unit disk by

$$f(z) = \sum_{n=0}^{\infty} A_n z^n$$

Then the zeros of f are distributed according to a determinantal point process.

Our final example is that of a binary one-dependent process. A binary-valued stochastic process $\{X_i\}_{i=0}^{\infty}$ is one-dependent if the processes $\{X_i\}_{i=0}^{N-1}$ and $\{X_i\}_{i=N+1}^{\infty}$ are independent for all N. As shown in Borodin et al. [2010], the set $\{i : X_i = 0\}$ is distributed according to a determinantal point process. If $\{Y_i\}_{i=0}^{\infty}$ are independent uniform draws from any set, then the sequence of random variables $X_i = \mathbf{1}\{Y_i \leq Y_{i+1}\}$ form a one-dependent process. As this appears to be tied to the longest increasing subsequence problem (Logan and Shepp [1977], **?**), this hints that the theory of determinantal point processes has yet to reach its full potential in general probability.

Chapter 3

New Kernels for Determinantal Point Processes

In this chapter, we examine two new kernels which give rise to interesting determinantal point processes. In both cases we will be able to characterize the process exactly and obtain some insight into a probabilistic problem that would be difficult to solve using other methods.

3.1 The Anticardinality and Cardinality Processes

Our first example is defined on $[n]$ with $\mathbb{P}(Y = A) \propto n+1-|A|$. Although it is not immediately obvious that this is a determinantal point process, we will be able to write down a kernel and give an explicit calculation to show that the determinants of this matrix give the desired probabilities.

Before we begin studying the process itself, we need a few results on exchangeable matrices. A matrix $\boldsymbol{M}$ is said to be exchangeable if $\boldsymbol{\Sigma}\boldsymbol{M}\boldsymbol{\Sigma}^T = \boldsymbol{M}$ for every permutation matrix $\boldsymbol{\Sigma}$. If we define $\boldsymbol{J} = \vec{1}\vec{1}^T$, then we can write every exchangeable matrix in the form $\alpha\boldsymbol{I} + \beta\boldsymbol{J}$. The eigenvectors of such a matrix are $\vec{1}$ and the basis of the orthogonal complement of $\vec{1}$, and their respective eigenvalues are

This chapter is joint work with Larry Shepp.

$\alpha + n\beta$ and α. We have immediately proved our first result:

Theorem 3.1.1.

$$\det(\alpha \boldsymbol{I} + \beta \boldsymbol{J}) = \alpha^{n-1}(\alpha + n\beta)$$

We must now compute the product of a pair of exchangeable matrices and the inverse of either of them. As the set of exchangeable matrices is closed under both operations, we note that it is a group under matrix multiplication. The proof of the following theorem is simple algebra and is omitted.

Theorem 3.1.2. *Let α, β, γ, δ be arbitrary real numbers.*

1. $(\alpha \boldsymbol{I} + \beta \boldsymbol{J})(\gamma \boldsymbol{I} + \delta \boldsymbol{J}) = \alpha\gamma \boldsymbol{I} + (\alpha\delta + \beta\gamma + \beta\delta)\boldsymbol{J}$
2. *If $\alpha \neq 0$ and $\beta \neq -\frac{\alpha}{n}$, then $(\alpha \boldsymbol{I} + \beta \boldsymbol{J})^{-1}$ exists and is equal to*

$$\frac{1}{\alpha}\boldsymbol{I} - \frac{\beta}{\alpha^2 + n\alpha\beta}\boldsymbol{J}$$

The following corollary is an easy consequence of Theorem 3.1.2 and shows that determinantal point processes with exchangeable kernels cannot have arbitrary restrictions on the size of the sets they draw:

Corollary 3.1.3. *For any α, $\beta \in \mathbb{R}$, we have that* $\operatorname{rank}(\alpha \boldsymbol{I} + \beta \boldsymbol{J}) \in \{0, 1, n-1, n\}$

With these results on exchangeable matrices, we can begin studying the anticardinality process. We must first compute the normalizing constant for this process. Our next result captures this computation:

Lemma 3.1.4.

$$\sum_{k=0}^{n} \binom{n}{k}(n + 1 - k) = (n + 2)2^{n-1}$$

Proof. Let $W \sim \text{Binom}\left(n, \frac{1}{2}\right)$. We compute as follows:

$$\begin{aligned}\sum_{k=0}^{n} \binom{n}{k}(n+1-k) &= \sum_{k=0}^{n} \binom{n}{k}(k+1) \\ &= 2^n \sum_{k=0}^{n} \binom{n}{k}(k+1)2^{-n} \\ &= 2^n(\mathbb{E}W + 1) \\ &= (n+2)2^{n-1}\end{aligned}$$

□

Therefore the anticardinality process is a distribution over subsets of $[n]$ defined by

$$\mathbb{P}(Y = A) = \frac{n+1-|A|}{(n+2)2^{n-1}}$$

for all $A \subseteq [n]$. We will now show that this is a determinantal point process:

Theorem 3.1.5. *The anticardinality process is a positive $\boldsymbol{L}$-ensemble parameterized by the matrix*

$$\boldsymbol{L} = \boldsymbol{I} - \frac{1}{n+1}\boldsymbol{J}$$

The corresponding kernel is given by

$$\boldsymbol{K} = \frac{1}{2}\boldsymbol{I} - \frac{1}{2n+4}\boldsymbol{J}$$

Proof. The proof is a simple application of Theorems 3.1.1 and 3.1.2, followed by some calculation. □

From here, we can compute that $\mathbb{E}|Y| = \frac{n}{2}$ and $\text{Var}(|Y|) = \frac{n(n^2+3n+5)}{4(n+2)^2}$. While we could have done this directly, it would be considerably harder to show that $\text{Cov}(X_i, X_j) = -\frac{1}{(2n+4)^2}$. And this points out some interesting asymptotic behavior: when n is large, the components of X are very nearly independent.

k	$\mathbb{P}(\lvert Y\rvert = k)$
0	$\frac{1}{256}$
1	$\frac{3}{64}$
2	$\frac{45}{256}$
3	$\frac{5}{16}$
4	$\frac{75}{256}$
5	$\frac{9}{64}$
6	$\frac{7}{256}$

Table 3.1: The probability of drawing a set of a given size following the cardinality process for $n = 6$.

However, we can show something more interesting. We consider the distribution of Y^c, which is given by $\mathbb{P}(Y^c = A) = \mathbb{P}(Y = A^c)$. It follows that

$$\mathbb{P}(Y = A) = \frac{|A| + 1}{(n+2)2^{n-1}}$$

However, by Theorem 2.3.5, we know that Y^c is also distributed according to a determinantal process. While we could work out the kernel of this cardinality process following the computations above, there is no need to do so, as we can compute it trivially from the kernel of the anticardinality process. We record these matrices as Theorem 3.1.6.

Theorem 3.1.6. *The cardinality process is a positive $\boldsymbol{L}$-ensemble parameterized by the matrix*

$$\boldsymbol{L} = \boldsymbol{I} + \boldsymbol{J}$$

The corresponding kernel is given by

$$\boldsymbol{K} = \frac{1}{2}\boldsymbol{I} + \frac{1}{2n+4}\boldsymbol{J}$$

The cardinality process is an important counterexample to the idea that determinantal point processes favor small sets. In fact, for $n = 6$, the cardinality of a set drawn according to the cardinality process is more than twice as likely to be

above its mean than below. Table 3.1 shows the probabilities of drawing sets of various sizes for this process.

Before we move on, we will spend a short time investigating determinantal point processes over a finite set whose kernel is an exchangeable matrix. Because all of the individual elements of the ground set are not distinguished by the kernel, we must have that $\mathbb{P}(Y = A) = f(|A|)$ for some function f. It follows that $|A|$ is a minimally sufficient statistic. If a distribution over independent and identically distributed random variables admits a minimally sufficient statistic whose dimension does not grow with the dimension of the sample, then the Pitman-Koopman Darmois theorem (see Brown [1986] for details) guarantees that the distribution belongs to an exponential family. It remains an open problem to determine whether the distribution of the cardinality of a draw from an exchangeable determinantal point process is an exponential family.

3.2 The Brownian Kernel Determinantal Point Process

We now turn our the determinantal point process on the right half-line whose kernel is defined by $L(s,t) = s \wedge t$ for s, $t \geq 0$, which the reader will recognize as the covariance function of a Brownian motion. The process given by this kernel is analytically tractable, and we can explicitly write down a formula for any quantity of interest.

The kernel $s \wedge t$ is not trace class on the right half-line, so we cannot simply analyze it as is. Instead we will consider its restriction to intervals of the form $[0,T]$ and take the weak limit as $T \to \infty$.

While in principle the nature of this process is determined by $L(s,t)$, we can also derive some insight from looking at the corresponding kernel $K(s,t)$. In this case the computation is easy, and so we perform it without hesitation.

Theorem 3.2.1. *Let $\mathbb{P}$ be the determinantal measure defined by $L(s,t) = s \min t$. Then the corresponding kernel is given by $K(s,t) = \sinh(s \wedge t)e^{-(s \vee t)}$.*

Proof. We begin with the integral equation

$$s \wedge t = K(s,t) + \int_0^\infty (s \wedge u)K(u,t) \ \ \mathrm{d}u$$

If we consider the case $s \leq t$, we can write this as

$$s = K(s,t) + \int_0^s uK(u,t) \ \ \mathrm{d}u + \int_s^\infty sK(u,t) \ \ \mathrm{d}u$$

From here, we observe that $K(0,t) = 0$. We differentiate the equation with respect to s to find that

$$1 = K_s(s,t) + \int_s^\infty K(u,t) \ \ \mathrm{d}u$$

. We then repeat the differentiation to find that

$$0 = K_{ss}(s,t) - K(s,t)$$

We now have a second-order differential equation for $K(s,t)$ with the initial condition $K(0,t) = 0$, and so we can conclude that $K(s,t) = \sinh(s)f(t)$ for some function f.

To compute f, we must observe that for $u > t$, $K(u,t)$ is equal to $\sinh(t)f(u)$, as our calculation above was performed under the assumption $s \leq t$. In light of this observation, we can write our integral equation as

$$s = \sinh(s)f(0) + \int_0^s u \sinh(u)f(t) \ \ \mathrm{d}u + \int_s^t s \sinh(u)f(t) \ \ \mathrm{d}u + \int_t^\infty s \sinh(t)f(u) \ \ \mathrm{d}u$$

After a straightforward but tedious calculation, we can compute $f(t) = \cosh(t) + c\sinh(t)$ for some constant c. If $c \neq -1$, then $f(t) \to \pm\infty$ as $t \to \infty$, and $K(s,t)$ is not bounded. We therefore conclude that $f(t) = \cosh(t) - \sinh(t)$, which simplifies

to e^{-t}. Therefore $K(s,t) = \sinh(s \wedge t)e^{-(s \vee t)}$ as claimed. $\square$

From the proof, we see that K is bounded even when defined on the right half line. Second, we can easily check that $K(s,t)$ satisfies the hypotheses of Theorem 2.4.9 and our process is a renewal process. In order to verify that K is nice, we will need to compute its eigenvalues. However, the eigenvalues of K are difficult to compute, and we will need to compute the eigenvalues of L to find its Janossy densities. Once we have done this, we will be able to find the eigenvalues of K by Theorem 2.4.3.

We begin the process of finding the Janossy densities associated with L. In order to do so, we will need the Fredholm determinant of L. We can compute this easily once we have found the eigenvalues of L, and so we do so promptly:

Theorem 3.2.2. *The eigenvalues of $L(s,t) = s \wedge t$ restricted to $[0,T]$ are given by*

$$\lambda_n = \frac{T^2}{\left(n + \frac{1}{2}\right)^2 \pi^2}$$

for $n = 0, 1, 2, \ldots$. The Fredholm determinant of L restricted to $[0,T]$ is equal to $\cosh(T)$.

Proof. The eigenvalues and eigenfunctions of L are the solutions to the integral equation

$$\lambda\phi(t) = \int_0^t u\phi(u) \ \mathrm{d}u + \int_t^T t\phi(u) \ \mathrm{d}u$$

As before, we will transform this into a differential equation, so we will need the initial condition $\phi(0) = 0$ to specify the solution. We differentiate once with respect to t to derive

$$\lambda\phi'(t) = \int_t^T \phi(u) \ \mathrm{d}u$$

We observe that $\phi'(T) = 0$, and differentiate again with respect to t to derive

$$\lambda\phi''(t) = -\phi(t)$$

This allows us to conclude that $\phi(t) = c\sin(\omega t)$ with $\lambda = \omega^{-2}$. The observation $\phi'(T) = 0$ implies that $\omega T = \left(n + \frac{1}{2}\right)\pi$ for some nonnegative integer n, and so we have that

$$\lambda = \frac{T^2}{\left(n + \frac{1}{2}\right)^2 \pi^2}$$

for $n = 0, 1, 2, \ldots$ as claimed. The Fredholm determinant of L is simply the product of the eigenvalues of $I + L$, and so we simply compute

$$\prod_{n=0}^{\infty} \left(1 + \frac{T^2}{\left(n + \frac{1}{2}\right)^2 \pi^2}\right) = \cosh(T)$$

□

Corollary 3.2.3. *The eigenvalues of K restricted to $[0, T]$ are given by*

$$\lambda_n = \frac{T^2}{\left(n + \frac{1}{2}\right)\pi^2 + T^2}$$

for $n = 0, 1, 2, \ldots$. Furthermore, K is locally trace class.

In order to fully specify the Janossy densities, we must compute the determinant of a matrix $\boldsymbol{L}\left(\{t_i\}_{i=1}^n\right)$ whose (i, j)-th entry is given by $\ell_{ij} = t_i \wedge t_j$. We do so in the following theorem:

Theorem 3.2.4. *Let $\{t_i\}_{i=1}^n$ be a set of nonnegative real numbers, and let L be the matrix whose (i, j)-th entry is given by $\ell_{ij} = t_i \wedge t_j$. Then* $\det(L) = t_1(t_2 - t_1)\ldots(t_n - t_{n-1})$.

Proof. As noted in Shirai and Takahashi [2003a], whenever $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$ are complex numbers, the determinant of the matrix $\boldsymbol{M}$ defined by $m_{ij} = a_{i\wedge j} b_{i\vee j}$ is given by the product

$$a_1 \cdot \begin{vmatrix} b_1 & b_2 \\ a_1 & a_2 \end{vmatrix} \cdot \begin{vmatrix} b_2 & b_3 \\ a_2 & a_3 \end{vmatrix} \cdot \cdots \cdot \begin{vmatrix} b_{n-1} & b_n \\ a_{n-1} & a_n \end{vmatrix} \cdot b_n$$

Our determinant is computed by applying this result to $a_i = t_i$ and $b_i = 1$. $\square$

We have now shown that the n-point Janossy density associated with L is given by

$$\frac{t_1(t_2 - t_1)\dots(t_n - t_{n-1}) \;\; \mathrm{d}t_1 \;\; \mathrm{d}t_2 \dots \;\; \mathrm{d}t_n}{\cosh(T)}$$

where we have $0 \leq t_1 < t_2 < \cdots < t_n \leq T$. In order to compute the probability that there are exactly n points in the interval $[0, T]$, we must integrate this density. We can do so directly, and we find that

$$\int_0^T \int_{t_1}^T \cdots \int_{t_{n-1}}^T \frac{t_1(t_2 - t_1)\dots(t_n - t_{n-1}) \;\; \mathrm{d}t_1 \;\; \mathrm{d}t_2 \dots \;\; \mathrm{d}t_n}{\cosh(T)} = \frac{T^{2n}}{(2n)!\cosh(T)}$$

This is a valid probability mass function as $\dfrac{T^{2n}}{(2n)!}$ is the nth term in the Taylor series for $\cosh(T)$. We will refer to this as the hyperbolic cosine series distribution, and study it in Section 3.2.1.

We now want to take the weak limit as T $\to \infty$. Because the process we are studying is a renewal process, we need only find the distribution of the location of the first point. To do so, we must integrate over the remainder of the points in the interval $[0, T]$. The quantity we must compute is given by

$$\frac{t_1 \;\; \mathrm{d}t_1}{\cosh(T)} \sum_{k=0}^{\infty} \int_{t_1}^T \int_{x_1}^T \cdots \int_{x_{k-1}}^T (x_1 - t_1)(x_2 - x_1)\dots(x_k - x_{k-1}) \;\; \mathrm{d}x_1 \;\; \mathrm{d}x_2 \dots \;\; \mathrm{d}x_k$$

After some computation, we find that this is equal to $\dfrac{t_1 \cosh(T - t_1) \;\; \mathrm{d}t_1}{\cosh(T)}$. We can then let $T \to \infty$ to find that the density of the first point is given by $t_1 e^{-t_1} \;\; \mathrm{d}t_1$. We observe that this is the density of the sum of two independent exponentials with $\lambda = 1$, and we have now characterized the process.

Theorem 3.2.5. *The Brownian kernel determinantal point process is equal in distribution to the set of even-indexed jump times of a rate one Poisson process.*

Support	$0, 1, 2, \ldots$
Parameters	$T > 0$
Mean	$\frac{1}{2}T\tanh(T)$
Variance	$\frac{1}{4}T\left(\tanh(T) + T\operatorname{sech}^2(T)\right)$
Mode	$(2\log(T))^{-1}$

Table 3.2: Quantities of interest for the hyperbolic cosine series distribution with parameter T.

3.2.1 The Hyperbolic Cosine Series Distribution

In the previous section we encountered a discrete distribution with support on the nonnegative integers whoseT probability mass function is given by

$$\mathbb{P}(X = n) = \frac{T^{2n}}{(2n)!\cosh(T)}$$

for some $T > 0$. Table 3.2 lists the summary statistics of this distribution, and we now show how they may be computed as well as some other properties.

We first observe that this is an exponential family distribution. We can see this by writing the probability mass function as

$$\mathbb{P}(X = n) = \frac{1}{(2n)!}e^{2n\log(T) - \log(\cosh(T))}$$

As such, we are guaranteed that the moment generating function exists. However, we will compute the characteristic function instead, as this is of more general interest and we can derive the moment generating function easily from it. This is a straightforward computation, and so we perform it immediately.

Theorem 3.2.6. *Let W be distributed according to a hyperbolic cosine series distribution with parameter T. Then* $\mathbb{E}e^{itW} = \dfrac{\cosh\left(Te^{\frac{it}{2}}\right)}{\cosh(T)}$.

Proof.

$$\begin{aligned}
\mathbb{E}e^{itW} &= \sum_{n=0}^{\infty} \frac{e^{itn}T^{2n}}{(2n)!\cosh(T)} \\
&= \sum_{n=0}^{\infty} \frac{(Te^{\frac{it}{2}})^{2n}}{(2n)!\cosh(T)} \\
&= \frac{\cosh\left(Te^{\frac{it}{2}}\right)}{\cosh(T)} \sum_{n=0}^{\infty} \frac{(Te^{\frac{it}{2}})^{2n}}{(2n)!\cosh\left(Te^{\frac{it}{2}}\right)} \\
&= \frac{\cosh\left(Te^{\frac{it}{2}}\right)}{\cosh(T)}
\end{aligned}$$

□

By the same reasoning, we can show that the probability generating function is given by $\mathbb{E}z^X = \dfrac{\cosh\left(T\sqrt{z}\right)}{\cosh(T)}$.

In principle, we now have sufficient machinery to calculate any summary statistics for the hyperbolic cosine series distribution. However, the moments of this distribution have an elegant representation in terms of the generalized hypergeometric function, and we would be remiss in not recording it.

Theorem 3.2.7. *Let W be distributed according to a hyperbolic cosine series distribution with parameter T. Then for any positive integer j, we have that*

$$\mathbb{E}X^j = \frac{T^2}{2\cosh(T)}\ {}_{j-1}F_j\left(2,\ldots,2;1,\ldots,1,\frac{3}{2};\frac{T^2}{4}\right)$$

Proof. The proof is a straightforward computation that relies on the definition of the generalized hypergeometric function and two standard properties of the gamma function:

1. $\Gamma(z+1) = z\Gamma(z)$
2. $\Gamma(z)\Gamma\left(z+\frac{1}{2}\right) = 2^{1-2z}\sqrt{\pi}\Gamma(2z)$

We now compute with the above properties in mind:

$$\begin{aligned}
\frac{T^2}{2}\,{}_{j-1}F_j\left(2,\ldots,2;1,\ldots,1,\frac{3}{2};\frac{T^2}{4}\right) &= \frac{1}{2}\sum_{n=0}^{\infty}\frac{\Gamma(n+2)^{j-1}\Gamma\left(\frac{3}{2}\right)T^{2(n+1)}}{4^n\Gamma(n+1)^{j-1}\Gamma(n+1)\Gamma\left(n+\frac{3}{2}\right)} \\
&= \frac{1}{2}\sum_{n=0}^{\infty}\frac{(n+1)^{j-1}\Gamma\left(\frac{3}{2}\right)T^{2(n+1)}}{4^n\Gamma(n+1)\Gamma\left(n+\frac{3}{2}\right)} \\
&= \sum_{n=0}^{\infty}\frac{(n+1)^{j-1}T^{2(n+1)}}{2\Gamma(2(n+1))} \\
&= \sum_{n=1}^{\infty}\frac{n^{j-1}T^{2n}}{2\Gamma(2n)} \\
&= \sum_{n=1}^{\infty}\frac{n^{j}T^{2n}}{2n\Gamma(2n)} \\
&= \sum_{n=1}^{\infty}\frac{n^{j}T^{2n}}{\Gamma(2n+1)} \\
&= \sum_{n=0}^{\infty}\frac{n^{j}T^{2n}}{(2n)!}
\end{aligned}$$

□

We will conclude our study of the hyperbolic cosine series by observing that the probability it assigns to large integers is smaller than that assigned by the Poisson distribution with the same parameter. In this sense, we are capturing the repulsive nature of a determinantal point process.

Theorem 3.2.8.

$$\lim_{n\to\infty}\frac{T^{2n}}{(2n)!\cosh(T)}\cdot\frac{n!}{e^{-T}T^n}=0$$

Proof. This ratio of probabilities is equal to $\dfrac{T^n e^T}{(n+1)(n+2)\ldots(2n)\cosh(T)}$. The denominator is greater than n^n, and the desired limit follows immediately from this observation. □

Chapter 4

Estimation of the Kernel Matrix

In this chapter, we consider the problem of estimating the kernel matrix of a determinantal point process from data. While Kulesza and Taskar [2011b] give an efficient algorithm for estimating the parameters of an $\boldsymbol{L}$-ensemble conditional on some set of features, the general problem of estimating the kernel from samples is still open.

This is not a trivial problem. We have two issues: if we allow arbitrary kernel matrices, we have a non-identifiable parameterization, as $\det(\boldsymbol{K}_\alpha) = \det([\boldsymbol{D}\boldsymbol{K}\boldsymbol{D}^{-1}]_\alpha)$ for any diagonal matrix $\boldsymbol{D}$ and $\alpha \subseteq [n]$. Furthermore, even if we have infinite data and can compute the relevant probabilities exactly, we are left with the task of computing a matrix from its principal minors. While this "principal minor assignment problem" was solved for a restricted class of matrices in Griffin and Tsatsomeros [2006b], the general problem remains open.

We have three major results in this chapter. First, we will give a simple characterization of the set of matrices with equal corresponding principal minors to a given matrix. Second, we will show how to pick a canonical representative from this set. Finally, we will give an algorithm that reconstructs an arbitrary symmetric matrix from its principal minors. In this way we solve the infinite data

This chapter is joint work with Ben Taskar and Alex Kulesza.

problem exactly, and lay the groundwork for performing estimation with finite data.

We note that there is one significant caveat to our solution the reconstruction problem. We can guarantee that if a list of numbers corresponds to the principal minors of some matrix, then that matrix is the output of our algorithm. However, we cannot guarantee that the list corresponds to any matrix without checking every single principal minor, which is an $O\left(2^n\right)$ operation (Griffin and Tsatsomeros [2006a]). However, in any situation where we are guaranteed to be given the principal minors of some matrix, the algorithm produces the correct output.

Example: 3×3 Matrices

Before we begin with the general theory, we consider the problem of reconstructing a 3×3 matrix from its principal minors to illuminate the issues we will face. Suppose that we are given an oracle for the principal minors of

$$\boldsymbol{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{12} & h_{22} & h_{23} \\ h_{13} & h_{23} & h_{33} \end{bmatrix}$$

In this case, we can enumerate the principal minors of $\boldsymbol{H}$ and write out an analytic solution. We have $\det\left(\boldsymbol{H}_{\{i\}}\right) = h_{ii}$ for each i, so we can reconstruct the diagonal of $\boldsymbol{H}$ exactly. We also have $\det\left(\boldsymbol{H}_{\{i,j\}}\right) = h_{11}h_{22} - h_{12}^2$ for each $i \neq j$, so we can also find the magnitude of each of the off-diagonal elements. However, we get no information about the signs of these elements from the principal minors of size 2, and we must examine the determinant of the matrix to find them.

We compute $\det\left(\boldsymbol{H}\right) = h_{11}h_{22}h_{33} + 2h_{12}h_{13}h_{23} - h_{11}h_{23}^2 - h_{13}^2h_{22} - h_{12}^2h_{33}$. From this, we can infer the value of the product $h_{11}h_{22}h_{33}$. If this value is not zero, we can tell whether an even or odd number of the off-diagonal elements are negative.

By inspection, we find that any assignment of signs that preserves this parity gives the same set of determinants, so it is not possible to infer the actual signs from the entire set of principal minors. On the other hand, if the product of the off-diagonal elements is zero, then any combination of signs is consistent with the determinant.

For larger matrices, we can follow the procedure above to find the diagonal elements and the magnitude of the off-diagonal elements. However, any ad-hoc approach to discovering the signs will quickly become unmanageable. As such, we need to find a generalization of the three-element product that we can use in a systematic manner to assign signs to the off-diagonal elements in a way that is consistent with the given principal minors.

We will base our work on a certain graphical representation of the matrix $\boldsymbol{H}$ to be defined in Section 4.1.2 in which the natural generalization of the three-element product is a simple chordless cycle. Our algorithm will start from a spanning tree of this graph in which the entries of the matrix corresponding to the edges are marked positive. From there, we infer the signs of the remaining edges based on a sufficiently systematic exploration of cycles and their chords, and we create a matrix $\boldsymbol{H}'$ which has equal corresponding principal minors to $\boldsymbol{H}$. This matrix is unique in some sense to be described in Section 4.2.4, so the problem is solved.

4.1 Notation and Terminology

4.1.1 Matrices

We say that $\boldsymbol{H}$ and $\boldsymbol{K}$ are *determinantally compatible* if $h_{ii} = k_{ii}$ for all i and $|h_{ij}| = |k_{ij}|$ for all $i \neq j$. We further say that $\boldsymbol{H}$ and $\boldsymbol{K}$ are *determinantally equivalent*, and write $\boldsymbol{H} \overset{\text{det}}{\equiv} \boldsymbol{K}$, if $\det(\boldsymbol{H}_\alpha) = \det(\boldsymbol{K}_\alpha)$ for all $\alpha \subseteq [n]$. Any pair of determinantally equivalent matrices are determinantally compatible, but the converse is not true.

We also say that two matrices $\boldsymbol{H}$ and $\boldsymbol{K}$ are $\boldsymbol{D}$-*similar*, and write $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$, if there is some diagonal matrix $\boldsymbol{D}$ with nonzero entries in $\{-1, 1\}$ such that $\boldsymbol{H} = \boldsymbol{D}\boldsymbol{K}\boldsymbol{D}^{-1}$. The set of all such $n \times n$ matrices will be denoted as $D_{\pm}^{n}$. We observe that $D_{\pm}^{n}$ is an Abelian group under matrix multiplication, so its actions on S^n correspond to symmetries: $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$ if and only if $\boldsymbol{H}$ is in the orbit of $\boldsymbol{K}$ under the action of $D_{\pm}^{n}$ on S^n defined by $\boldsymbol{D} \cdot \boldsymbol{K} = \boldsymbol{D}\boldsymbol{K}\boldsymbol{D}^{-1}$. This symmetry is easy to describe: it is simply invariance with respect to the direction of the coordinate axes. As above, any pair of $\boldsymbol{D}-$*similar* matrices are determinantally compatible, but the converse is not true.

4.1.2 Graphs

While we can describe our algorithms and results purely in terms of matrices, there is an equivalent and much clearer description in graph-theoretic terms. In general, we will use n to denote the number of vertices in a graph G, m to denote the number of edges, and ℓ to denote the length of the longest cycle.

For any $\boldsymbol{H} \in S^n$, we define the graph $G(\boldsymbol{H})$ to have vertex set $[n]$ and edge set $\{(i, j) : i \neq j \text{ and } h_{ij} \neq 0\}$. As observed above, the signs of $\boldsymbol{H}$ contain essential information about its determinants, so we give $G(\boldsymbol{H})$ edge weights $w_{ij} = \operatorname{sgn}(h_{ij})$. The subgraph of $G(\boldsymbol{H})$ induced by $\alpha \subseteq [n]$ is simply $G(\boldsymbol{K}_{\alpha})$. We will use $n(\boldsymbol{H})$ to denote the number of connected components of $G(\boldsymbol{H})$.

For determinantally compatible matrices $\boldsymbol{H}$ and $\boldsymbol{K}$, the only possible differences are the signs of the off-diagonal elements. We define the graph $G(\boldsymbol{H}, \boldsymbol{K})$ to have the shared vertex and edge sets of $G(\boldsymbol{H})$ and $G(\boldsymbol{K})$, and edge weights $w_{ij} = \operatorname{sgn}(h_{ij}k_{ij})$. As above, we will use $n(\boldsymbol{H}, \boldsymbol{K})$ to denote the number of connected components of $G(\boldsymbol{H}, \boldsymbol{K})$.

Every graph we will consider in the remainder of this paper will be either $G(\boldsymbol{H})$ or $G(\boldsymbol{H}, \boldsymbol{K})$ for some $\boldsymbol{H}$ and $\boldsymbol{K}$. The discussion below does not apply to general graphs, but specifically to graphs of this form.

A rooted graph is a graph G in which some vertex r has been designated as the root. While this does not imply any special properties of G, many graphical algorithms, including ours, are most simply expressed with some node designated as the root.

A path P in a graph G is a sequence of vertices $\left\{v_{i_j}\right\}_{j=1}^{|P|}$ such that G contains the edges $\left(v_{i_j}, v_{i_{j+1}}\right)$ for $j = 1, \ldots, |C|$. A graph is connected if there is a path between any two vertices, and is otherwise partitioned into a set of connected subgraphs referred to as connected components.

A tree is a graph T such that there is a unique path between any two vertices of T. If G is a connected graph, a spanning tree T is a subgraph of G such that T is a tree, each vertex of G is present in T, and each edge of T is present in G. If G is not collected, we can define a spanning forest, which is the union of a collection of spanning trees for each connected component of G.

If G contains a path C and also contains the edge $\left(v_{i_{|C|}}, v_{i_1}\right)$, then we say that G contains a cycle C. If the vertices of C are all distinct, C is referred to as a simple cycle. The set of vertices $\left\{v_{i_j}\right\}_{j=1}^{|C|}$ is referred to as the support of C, and denoted by $\operatorname{supp}(C)$.

If there is an edge between two vertices of a cycle which is not contained in the cycle itself, this edge is referred to as a chord. A cycle with no chords is referred to as chordless. If C does contain a chord c, then there are two cycles C_1 and C_2 such that $\operatorname{supp}(C_1) \cup \operatorname{supp}(C_2) = \operatorname{supp}(C)$ and $\operatorname{supp}(C_1) \cap \operatorname{supp}(C_2) = \operatorname{supp}(p)$. In this case, we say that c separates C into the subcycles C_1 and C_2.

We will have occasion to consider the product of the edge weights of a cycle C, which we will denote as $p(C)$ and define as

$$p(C) = w_{|C|,1} \prod_{j=1}^{|C|-1} w_{j,j+1}$$

Note that $p(C)$ does not depend on the weights of any k-chords in C.

A graph coloring c is a mapping from the vertices of graph G to some set of colors which obeys some set of constraints induced by the edge weights. In traditional graph coloring problems, the vertices i and j must be assigned different colors if the edge (i, j) is present. In our case, we will be interested in a coloring c of $G(\boldsymbol{H}, \boldsymbol{K})$ taking values in $\{-1, 1\}$ such that $c(j) = w_{ij}c(i)$. If such a mapping exists, it will be referred to as a valid coloring of $G(\boldsymbol{H}, \boldsymbol{K})$.

We will be interested in determining whether the graphs we consider have unique colorings, but by the above definition, this is never true: if c is a valid coloring for G, then $-c$ is as well. We define a rooted coloring for G to be a valid coloring c_r with the property that $c(r) = 1$ for some specified vertex r, which will be referred to as the root. Any graph with a valid coloring possesses a unique rooted coloring. If we assign a root to a tree T, then the depth of each vertex v is defined to be the length of the path from the root to v.

Finally, we define the set $G_{\pm}$ of graphs with edge weights in the set $\{-1, 1\}$. We define the action of $\boldsymbol{D} \in D_{\pm}^n$ on a graph $G \in G_{\pm}$ by taking the new edge weights $w_{ij}^{'} = d_{ii}d_{jj}w_{ij}$. We can furthermore define the relation of $\boldsymbol{D}$-similarity on $G_{\pm}$ by $G_1 \overset{D}{\sim} G_2$ if and only if $G_1 = \boldsymbol{D} \cdot G_2$ for some $\boldsymbol{D} \in D_{\pm}$. By construction, we have that $G(\boldsymbol{H}) \overset{D}{\sim} G(\boldsymbol{K})$ if and only if $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$.

4.2 The Theory of D-Similarity

In this section, we will characterize the relationship between determinantal equivalence and $\boldsymbol{D}$-similarity. While these two relations are not in general equivalent, they are sufficiently close that a canonical representative can be described with ease.

4.2.1 Preliminary Results

We open with a set of observations which will be used throughout the remainder of our paper. Lemmas 4.2.1 and 4.2.2 allow us to assume without loss of generality that our graphs are connected and that our vertices are labeled in any convenient order. Theorem 4.2.3, which is an interesting result in its own right, allows us to make arbitrary choices in our algorithms without worrying about how they affect the correctness of the result.

Lemma 4.2.1. *Let σ be a permutation of $[n]$, and let $\mathbf{\Sigma}$ be the corresponding permutation matrix.*

1. $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$ *if and only if* $\mathbf{\Sigma}\boldsymbol{H}\mathbf{\Sigma}^T \overset{D}{\sim} \mathbf{\Sigma}\boldsymbol{K}\mathbf{\Sigma}^T$.

2. $\boldsymbol{H} \overset{\det}{\equiv} \boldsymbol{K}$ *if and only if* $\mathbf{\Sigma}\boldsymbol{H}\mathbf{\Sigma}^T \overset{\det}{\equiv} \mathbf{\Sigma}\boldsymbol{K}\mathbf{\Sigma}^T$.

Proof. We prove each claim in turn:

1. If $\boldsymbol{H} = \boldsymbol{D}\boldsymbol{K}\boldsymbol{D}^{-1}$, then $\mathbf{\Sigma}\boldsymbol{H}\mathbf{\Sigma}^T = (\mathbf{\Sigma}\boldsymbol{D}\mathbf{\Sigma}^T)(\mathbf{\Sigma}\boldsymbol{K}\mathbf{\Sigma}^T)(\mathbf{\Sigma}\boldsymbol{D}^{-1}\mathbf{\Sigma}^T)$. This follows from the observations that $\mathbf{\Sigma}^{-1} = \mathbf{\Sigma}^T$ and that $\mathbf{\Sigma}\boldsymbol{D}\mathbf{\Sigma}^T = \mathbf{\Sigma}^T\boldsymbol{D}\mathbf{\Sigma}$.

2. Let $\boldsymbol{I}_\alpha$ to be the diagonal matrix with $[\boldsymbol{I}_\alpha]_{ii} = \mathbf{1}\{i \in \alpha\}$. If $\boldsymbol{H}$ is any matrix and α any subset of $[n]$, we have that $\det(\boldsymbol{H}_\alpha) = \det\left(\boldsymbol{I}_\alpha H \boldsymbol{I}_\alpha + I_{[n]-\alpha}\right)$. With this observation, we can simply calculate to show that $\det(\boldsymbol{H}_\alpha) = \det\left([\mathbf{\Sigma}\boldsymbol{H}\mathbf{\Sigma}^T]_{\sigma(\alpha)}\right)$:

$$\begin{aligned}
\det\left([\mathbf{\Sigma}\boldsymbol{H}\mathbf{\Sigma}^T]_{\sigma(\alpha)}\right) &= \det\left(\boldsymbol{I}_{\sigma(\alpha)}\mathbf{\Sigma}\boldsymbol{H}\mathbf{\Sigma}^T\boldsymbol{I}_{\sigma(\alpha)} + \boldsymbol{I}_{[n]-\sigma(\alpha)}\right) \\
&= \det\left(\mathbf{\Sigma}\boldsymbol{I}_\alpha\mathbf{\Sigma}^T\mathbf{\Sigma}\boldsymbol{H}\mathbf{\Sigma}^T\mathbf{\Sigma}\boldsymbol{I}_\alpha\mathbf{\Sigma}^T + \mathbf{\Sigma}\boldsymbol{I}_{[n]-\alpha}\mathbf{\Sigma}^T\right) \\
&= \det\left(\mathbf{\Sigma}(\boldsymbol{I}_\alpha\boldsymbol{H}\boldsymbol{I}_\alpha + \boldsymbol{I}_{[n]-\alpha})\mathbf{\Sigma}^T\right) \\
&= \det\left(\boldsymbol{I}_\alpha\boldsymbol{H}\boldsymbol{I}_\alpha + \boldsymbol{I}_{[n]-\alpha}\right) \\
&= \det(\boldsymbol{H}_\alpha)
\end{aligned}$$

□

Lemma 4.2.2. *We have the following equivalences:*

1. $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$ *if and only if* $\boldsymbol{H}_\alpha \overset{D}{\sim} \boldsymbol{K}_\alpha$ *for every connected component* α *of* $G(\boldsymbol{H}, \boldsymbol{K})$.

2. $\boldsymbol{H} \overset{\det}{\equiv} \boldsymbol{K}$ *if and only if* $\boldsymbol{H}_\alpha \overset{\det}{\equiv} \boldsymbol{K}_\alpha$ *for every connected component* α *of* $G(\boldsymbol{H}, \boldsymbol{K})$.

Proof. By Lemma 4.2.1, we can assume without loss of generality that $\boldsymbol{H} = \operatorname{diag}\left(\{\boldsymbol{H}_{\alpha_i}\}_{i=1}^{n(\boldsymbol{H},\boldsymbol{K})}\right)$ and $\boldsymbol{K} = \operatorname{diag}\left(\{\boldsymbol{K}_{\alpha_i}\}_{i=1}^{n(\boldsymbol{H},\boldsymbol{K})}\right)$.

1. We write $\boldsymbol{D} = \operatorname{diag}\left(\{\boldsymbol{D}_{\alpha_i}\}_{i=1}^{n(\boldsymbol{H},\boldsymbol{K})}\right)$, and simply observe that $\boldsymbol{H} = \boldsymbol{D} \cdot \boldsymbol{K}$ if and only if $\boldsymbol{H}_{\alpha_i} = \boldsymbol{D}_{\alpha_i} \cdot \boldsymbol{K}_{\alpha_i}$ for all i, since all entries outside of the blocks corresponding to the connected components are zero.

2. This follows immediately from the fact that $\boldsymbol{H}_\beta = \operatorname{diag}\left(\{\boldsymbol{H}_{\alpha_i \cap \beta}\}_{i=1}^{n(\boldsymbol{H},\boldsymbol{K})}\right)$.

□

Theorem 4.2.3. *Let* $\boldsymbol{D}_1, \boldsymbol{D}_2 \in D_{\pm}^n$. $\boldsymbol{D}_1 \cdot \boldsymbol{H} = \boldsymbol{D}_2 \cdot \boldsymbol{H}$ *if and only if* $[\boldsymbol{D}_1]_\alpha = \pm[\boldsymbol{D}_2]_\alpha$ *for every connected component* α *of* $G(\boldsymbol{K})$.

Proof. We first assume that $[\boldsymbol{D}_1]_\alpha = \pm[\boldsymbol{D}_2]_\alpha$ for every connected component α of $G(\boldsymbol{K})$. Here we can directly apply Lemma 4.2.2 to conclude that $\boldsymbol{D}_1 \cdot \boldsymbol{K} = \boldsymbol{D}_2 \cdot \boldsymbol{K}$.

We now assume that $\boldsymbol{D}_1 \cdot \boldsymbol{H} = \boldsymbol{D}_2 \cdot \boldsymbol{H}$ and that $G(\boldsymbol{H})$ has a single connected component. Let $\theta = [D_1]_{11}[D_2]_{11}$, and assume that $D_1 \neq \theta D_2$. Then there is some least index $b > 1$ such that $[D_1]_{bb} \neq \theta[D_2]_{bb}$. This implies that $[\boldsymbol{D}_1 \cdot \boldsymbol{H}]_{bj} \neq [\boldsymbol{D}_2 \cdot \boldsymbol{H}]_{bj}$ for any $j < b$, and so we have that $\boldsymbol{D}_1 \cdot \boldsymbol{H} \neq \boldsymbol{D}_2 \cdot \boldsymbol{H}$. This contradicts our hypothesis, and so we can conclude that $\boldsymbol{D}_1 = \theta \boldsymbol{D}_2$. By Lemma 4.2.2, the argument above applies to each connected component of $G(\boldsymbol{H})$, and we have the desired result. □

4.2.2 Algorithms

In this section, we give an algorithm that correctly decides whether $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$. We then show how to extend it to be completely constructive: if $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$, we will construct a $\boldsymbol{D}$ such that $\boldsymbol{H} = \boldsymbol{D} \cdot \boldsymbol{K}$; and if $\boldsymbol{H} \overset{D}{\not\sim} \boldsymbol{K}$, we will construct a minimal α such that $\boldsymbol{H}_\alpha \overset{D}{\not\sim} \boldsymbol{K}_\alpha$. Throughout this section we assume that any pair $\boldsymbol{H}$ and $\boldsymbol{K}$ are determinantally compatible and that $G(\boldsymbol{H}, \boldsymbol{K})$ is connected.

Recall that c is a valid coloring of $G(\boldsymbol{H}, \boldsymbol{K})$ if and only if $c(j) = w_{ij}c(i)$ for all i and j, and a valid rooted coloring if $c(r) = 1$ for some $r \in [n]$. Therefore, $G(\boldsymbol{H}, \boldsymbol{K})$ possesses a valid rooted coloring if and only if this system of linear equations possesses a solution. We refer to this system as the coloring equations.

Lemma 4.2.4. *$\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$ if and only if $G(\boldsymbol{H}, \boldsymbol{K})$ possesses a valid coloring.*

Proof. $\boldsymbol{H} = \boldsymbol{D}\boldsymbol{K}\boldsymbol{D}^{-1}$ if and only if $h_{ij} = d_{ii}d_{jj}k_{ij}$. By the construction of $G(\boldsymbol{H}, \boldsymbol{K})$, we have $w_{ij} = \operatorname{sgn}(h_{ij}k_{ij})$. Therefore $\boldsymbol{H} = \boldsymbol{D}\boldsymbol{K}\boldsymbol{D}^{-1}$ if and only if $w_{ij} = d_{ii}d_{jj}$. If we define $c(i) = d_{ii}$, we have constructed a valid coloring c. If we are given a valid coloring c of $G(\boldsymbol{H}, \boldsymbol{K})$, we can use the same logic to construct a $D \in D_{\pm}^n$ such that $\boldsymbol{H} = \boldsymbol{D}\boldsymbol{K}\boldsymbol{D}^{-1}$. □

In light of Lemma 4.2.4, we can decide whether $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$ by determining whether the coloring equations for $G(\boldsymbol{H}, \boldsymbol{K})$ have a solution. Any algorithm that produces a solution to a system of linear equations may be used to find valid colorings of $G(\boldsymbol{H}, \boldsymbol{K})$, but the coloring equations are so sparse that a specialized algorithm can find the solution much more quickly. We begin our discussion of such an algorithm with a result on trees and colorability.

Lemma 4.2.5. *If $G(\boldsymbol{H}, \boldsymbol{K})$ is a tree, there is a unique valid rooted coloring of $G(\boldsymbol{H}, \boldsymbol{K})$ for any set of edge weights. If $G(\boldsymbol{H}, \boldsymbol{K})$ is not a tree, there is some set of edge weights for which no valid coloring is possible.*

Proof. We first assume that $G(\boldsymbol{H}, \boldsymbol{K})$ is a tree. Then we can permute the indices of $G(\boldsymbol{H}, \boldsymbol{K})$ so that the matrix corresponding to the coloring equations is $n \times n$

Algorithm 2 Given determinantally compatible $\boldsymbol{H}$ and $\boldsymbol{K}$, either produce a $\boldsymbol{D}$ such that $\boldsymbol{H} = \boldsymbol{D}\boldsymbol{K}\boldsymbol{D}^{-1}$ or determine that none exists

```
Let T be a spanning tree of G(H, K) with root 1
Produce a coloring c of T with c(1) = 1 and construct the corresponding D
for each edge (i, j) not in T do
    Verify that w_ij = d_ii d_jj
end for
```

upper triangular with nonzero diagonal entries. In this case, we are guaranteed that a solution exists.

We now assume that $G(\boldsymbol{H}, \boldsymbol{K})$ is not a tree. In this case, there is a pair of vertices u and v such that there is a path p_1 from u to v, and a disjoint path p_2 from v to u. If we write w_p for the sum of the edge weights along the path p, we have that $c(u) = c(u)w_{p_1}w_{p_2}$. We can always choose edge weights so that this condition is not satisfied, and so we have the desired result. □

The proof of Lemma 4.2.5 gives us a relationship between the colorability of a cycle and the sum of its edge weights. We record this result as Corollary 4.2.6.

Corollary 4.2.6. *A cycle C of $G(\boldsymbol{H}, \boldsymbol{K})$ possesses a valid coloring if and only if $p(C) = 1$.*

By Lemma 4.2.5, the coloring equations for any tree have a solution. Furthermore, they are sufficiently sparse that any of the standard graph search algorithms can be used to solve them with slight modifications. Therefore, our algorithm to find a valid coloring of an arbitrary graph is simple: we will color some spanning tree, and verify that the coloring produced is valid for the entire graph. This idea is captured in Algorithm 2 and proved correct in Theorem 4.2.7.

Theorem 4.2.7. *Algorithm 2 produces a $\boldsymbol{D} \in D_{\pm}^{n}$ such that $\boldsymbol{H} = \boldsymbol{D}\boldsymbol{K}\boldsymbol{D}^{-1}$ if any exists, and otherwise determines that no such $\boldsymbol{D}$ exists.*

Proof. By Lemma 4.2.4, if $\boldsymbol{H} = \boldsymbol{D}\boldsymbol{K}\boldsymbol{D}^{-1}$ for some $\boldsymbol{D} \in D_{\pm}^{n}$, then this $\boldsymbol{D}$ corresponds to a valid coloring of $G(\boldsymbol{H}, \boldsymbol{K})$. As discussed above, Algorithm 2 correctly

finds a valid rooted coloring of the spanning tree T and verifies that it holds for the entire graph. By Theorem 4.2.3 and the constraint $c(1) = 1$, the $\boldsymbol{D}$ produced is irrespective of the choice of spanning tree. If $\boldsymbol{H} \overset{D}{\not\sim} \boldsymbol{K}$, then the coloring equations have no solution, and Algorithm 2 will correctly verify this. □

We now consider the problem of verifying that $\boldsymbol{H} \overset{D}{\not\sim} \boldsymbol{K}$. While we can show this by failing to produce a $\boldsymbol{D}$ such that $\boldsymbol{H} = \boldsymbol{D}\boldsymbol{K}\boldsymbol{D}^{-1}$, we would like to be able to produce a $\beta \subseteq [n]$ such that $\boldsymbol{H}_\beta \overset{D}{\not\sim} \boldsymbol{K}_\beta$, but $\boldsymbol{H}_\gamma \overset{D}{\not\sim} \boldsymbol{K}_\gamma$ for every γ strictly contained in β. We will refer to such β as a *minimal counterexample*, and to any $\alpha \supseteq \beta$ as a *counterexample*. We can easily find a minimal counterexample by enumerating every $\beta \subseteq [n]$, but we would like to find a more efficient algorithm. To do so, we must examine the structure of any minimal counterexample.

Lemma 4.2.8. *If $H \overset{D}{\not\sim} K$, any minimal counterexample must be the support of some simple chordless cycle C.*

Proof. By Lemma 4.2.5 and Corollary 4.2.6, $G(\boldsymbol{H}, \boldsymbol{K})$ must contain a cycle C such that $p(C) = -1$. Assume that C has some n-chord (u, v) that separates C into C_1 and C_2. Then we must have that $p(C) = p(C_1)p(C_2)$, and either $p(C_1) = -1$ or $p(C_2) = -1$. It follows that there must be some chordless cycle C_* such that $p(C_*) \neq 1$. Any proper subgraph of C_* is a tree and does have a valid coloring, so C_* is a minimal counterexample as claimed. □

The proof of Lemma 4.2.8 suggests an algorithm for finding a minimal counterexample when $\boldsymbol{H} \overset{D}{\not\sim} \boldsymbol{K}$. The procedure is outlined in Algorithm 3, and its correctness is recorded in Theorem 4.2.9.

Theorem 4.2.9. *If $\boldsymbol{H} \overset{D}{\not\sim} \boldsymbol{K}$, Algorithm 3 correctly discovers a minimal counterexample.*

Algorithm 3 Given determinantally compatible $\boldsymbol{H}$ and $\boldsymbol{K}$ with $\boldsymbol{H}$ not $\boldsymbol{D}$-similar to $\boldsymbol{K}$, find a minimal counterexample α

Run Algorithm 2 until a contradiction is discovered along edge e
Let C be the cycle consisting of the edge e and the path from i to j in the spanning tree of $G(\boldsymbol{H}, \boldsymbol{K})$
while C contains an n-chord (u, v) **do**
    Let C_1 and C_2 be the subcycles of C separated by c
    if $p(C_1) = -1$ **then**
        $C \leftarrow C_1$
    else
        $C \leftarrow C_2$
    end if
end while

4.2.3 D-Similarity and Determinantal Equivalence

In this section, we will show $\boldsymbol{H} \overset{\det}{\equiv} \boldsymbol{K}$ if and only if $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$. We will show that simple chordless cycles play a prominent role in the theory of $\boldsymbol{D}$-similarity, and then use this insight to prove the equivalence of our two relations. We begin with two lemmas on $\boldsymbol{D}$-similarity of graphs

Lemma 4.2.10. *For any $G \in G_{\pm}$, there is some $\boldsymbol{H} \in S^n$ such that $G = G(\boldsymbol{H})$.*

Proof. Define $\boldsymbol{H}$ by $h_{ii} = 1$ for all i, $h_{ij} = 0$ if (i, j) is not an edge of G, and $h_{ij} = w_{ij}$ if (i, j) is an edge of G. Then $G = G(\boldsymbol{H})$ by construction. □

Lemma 4.2.11. *Let C_1 and C_2 be simple chordless cycles of length n with weights in $\{-1, 1\}$. $C_1 \overset{D}{\sim} C_2$ if and only if $p(C_1) = p(C_2)$.*

Proof. By Lemma 4.2.10, we can choose $\boldsymbol{H}$ and $\boldsymbol{K}$ such that $C_1 = G(\boldsymbol{H})$ and $C_2 = G(\boldsymbol{K})$. Then $C_1 \overset{D}{\sim} C_2$ if and only if $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$. By Lemma 4.2.4, $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$ if and only if $G(\boldsymbol{H}, \boldsymbol{K})$ possesses a valid coloring. By Corollary 4.2.6, $G(\boldsymbol{H}, \boldsymbol{K})$ possesses a valid coloring if and only if $p(G(\boldsymbol{H}, \boldsymbol{K})) = 1$. The edge weights of $G(\boldsymbol{H}, \boldsymbol{K})$ are the product of the corresponding edge weights in $G(\boldsymbol{H})$ and $G(\boldsymbol{K})$, so it follows that $p(G(\boldsymbol{H}, \boldsymbol{K})) = 1$ if and only if $p(C_1) = p(C_2)$. □

Given a determinantally compatible pair of matrices $(\boldsymbol{H}, \boldsymbol{K})$ and a cycle C contained in $G(\boldsymbol{H}, \boldsymbol{K})$, we write $C_{\boldsymbol{H}}$ to denote C with the edge weights inherited

from $G(\boldsymbol{H})$, $C_{\boldsymbol{K}}$ to denote C with the edge weights inherited from $G(\boldsymbol{K})$, and $C_{\boldsymbol{H},\boldsymbol{K}}$ to denote C with edge weights inherited from $G(\boldsymbol{H},\boldsymbol{K})$. With this notation, we can easily state and prove a necessary and sufficient condition for $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$ in terms of the simple chordless cycles of $G(\boldsymbol{H},\boldsymbol{K})$.

Theorem 4.2.12. *Let $\boldsymbol{H}$, $\boldsymbol{K}$ be determinantally compatible. Then $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$ if and only if $p\left(C_{\boldsymbol{H}}\right) = p\left(C_{\boldsymbol{K}}\right)$ for every simple chordless cycle C of $G(\boldsymbol{H},\boldsymbol{K})$.*

Proof. We first assume that $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$. If C is any simple chordless cycle of $G(\boldsymbol{H},\boldsymbol{K})$, we have $C_{\boldsymbol{H}} \overset{D}{\sim} C_{\boldsymbol{K}}$. Therefore $p(C_{\boldsymbol{H}}) = p(C_{\boldsymbol{K}})$ by Lemma 4.2.11.

We now assume that $\boldsymbol{H} \overset{D}{\not\sim} \boldsymbol{K}$. Then there is some minimal counterexample α. By Lemma 4.2.8, $G\left(\boldsymbol{H}_\alpha, \boldsymbol{K}_\alpha\right)$ must be a simple chordless cycle C. Lemma 4.2.11 allows us to conclude that $p(C_H) \neq p(C_K)$. □

The matrices whose graphs are simple chordless cycles are known as cyclic tridiagonal matrices (Engeln-Müllges and Uhlig [1996]). The following lemma gives an explicit expression for the determinant of a cyclic tridiagonal matrix $\boldsymbol{T}$.

Lemma 4.2.13. *Let $\boldsymbol{T}$ be a cyclic tridiagonal matrix. Then*

$$\det\left(\boldsymbol{T}\right) = t_{nn}\det\left(\boldsymbol{T}_{[n-1]}\right) - t_{n-1,n}^2\det\left(\boldsymbol{T}_{[n-2]}\right) - t_{1,n}^2\det\left(\boldsymbol{T}_{[n-1]\setminus[1]}\right) + (-1)^{n+1}t_{1n}\prod_{i=1}^{n-1}t_{i,i+1}$$

Proof. The expression is obtained by the Laplace expansion of the determinant of $\boldsymbol{T}$. The details are omitted. □

We now have sufficient machinery to to prove our main result, and we do so without hesitation.

Theorem 4.2.14. *$\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$ if and only if $\boldsymbol{H} \overset{\det}{\equiv} \boldsymbol{K}$.*

Proof. If $\boldsymbol{H}$ and $\boldsymbol{K}$ are not determinantally compatible, then $\boldsymbol{H} \overset{D}{\not\sim} \boldsymbol{K}$ and $\boldsymbol{H} \overset{\det}{\not\equiv} \boldsymbol{K}$. We therefore assume that $\boldsymbol{H}$ and $\boldsymbol{K}$ are determinantally compatible. We first assume that $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$. Then $\boldsymbol{H}$ is diagonally similar to $\boldsymbol{K}$, and as diagonally

similar matrices have equal corresponding principal minors, we can conclude that $\boldsymbol{H} \overset{\text{det}}{\equiv} \boldsymbol{K}$.

We now assume that $\boldsymbol{H} \overset{\text{det}}{\equiv} \boldsymbol{K}$. If $G(\boldsymbol{H}, \boldsymbol{K})$ contains no cycles, then $G(\boldsymbol{H}, \boldsymbol{K})$ is a tree, and we have that $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$ by Lemma 4.2.5. We therefore assume that $G(\boldsymbol{H}, \boldsymbol{K})$ contains at least one cycle, which implies that $G(\boldsymbol{H}, \boldsymbol{K})$ must contain some simple chordless cycle C. We apply Lemma 4.2.13 to the matrices $\boldsymbol{H}_{\text{supp}(C)}$ and $\boldsymbol{K}_{\text{supp}(C)}$ and cancel equal terms to see that

$$h_{|C|1} \prod_{i=1}^{|C|-1} h_{i,i+1} = k_{|C|1} \prod_{i=1}^{|C|-1} k_{i,i+1}$$

We know that $|h_{ij}| = |k_{ij}|$ for all i and j, so it must be the case that $p\left(C_{\boldsymbol{H}}\right) = p\left(C_{\boldsymbol{K}}\right)$. Since this holds for every simple chordles cycle in $G(\boldsymbol{H}, \boldsymbol{K})$, Theorem 4.2.12 allows us to conclude that $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$. □

Two comments are in order. First, as in Engel and Schneider [1980], the hypotheses of Theorem 4.2.14 can be weakened considerably. Our argument would still go through if we could merely assume that $\boldsymbol{H}$ and $\boldsymbol{K}$ are determinantally compatible and the principal minors corresponding to the cycles of $G(\boldsymbol{H}, \boldsymbol{K})$ are equal. In particular, for any determinantally compatible $\boldsymbol{H}$ and $\boldsymbol{K}$ with no zeros off the diagonal, we have that $\boldsymbol{H} \overset{\text{det}}{\equiv} \boldsymbol{K}$ if and only if all the corresponding 3×3 principal minors are equal.

Second, we can reinterpret Algorithm 3 in light of Theorem 4.2.14. A minimal counterexample α for $\boldsymbol{H}$ and $\boldsymbol{K}$ has the property that $\det\left(\boldsymbol{H}_\alpha\right) \neq \det\left(\boldsymbol{K}_\alpha\right)$, but $\det\left(\boldsymbol{H}_\beta\right) = \det\left(\boldsymbol{K}_\beta\right)$ for any β strictly contained in α. Algorithm 3 therefore finds minimal unequal corresponding principal minors.

We close this section by counting the number of matrices with equal corresponding principal minors to some fixed $\boldsymbol{H}$. The following result is an easy corollary of Theorems 4.2.3 and 4.2.14.

Corollary 4.2.15. *For any fixed* $\boldsymbol{H}$, $\left|\left\{\boldsymbol{K} : \boldsymbol{H} \overset{\text{det}}{\equiv} \boldsymbol{K}\right\}\right| = 2^{n-n(\boldsymbol{H})}$.

4.2.4 Canonicalization

We have one last major theorem to develop before we have a complete theory of $\boldsymbol{D}$-similarity and determinantal equivalence. In short, given that there infinitely many matrices which are determinantally equal to a fixed $\boldsymbol{H}$, how can we pick a canonical representative from this set? As might be expected, our algorithm is graphical in nature, but it does not give us a single canonical element. Instead, the canonicalization is relative to a deterministic spanning tree algorithm $\mathcal{A}$. We begin with two lemmas regarding real positive spanning trees and $\boldsymbol{D}$-similarity.

Lemma 4.2.16. *Let $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$, and assume that there is a spanning tree T of $G(\boldsymbol{H}, \boldsymbol{K})$ with weights identically equal to one. Then $\boldsymbol{H} = \boldsymbol{K}$.*

Proof. By Theorem 4.2.12, we must have $p(C) = 1$ for every chordless cycle C of $G(\boldsymbol{H}, \boldsymbol{K})$. Every edge in T has weight one, and so it follows that every edge of $G(\boldsymbol{H}, \boldsymbol{K})$ has weight one. Therefore $\boldsymbol{H} = \boldsymbol{K}$. □

Lemma 4.2.17. *Assume that $G(\boldsymbol{H})$ is a tree, and let $\boldsymbol{D}$ be the matrix corresponding to the solution of the coloring equations for $G(\boldsymbol{H})$. If $\boldsymbol{K} = \boldsymbol{D}^{-1}\boldsymbol{H}\boldsymbol{D}$, then $\operatorname{sgn}(k_{ij}) = 1$ whenever $i \neq j$ and $k_{ij} \neq 0$.*

Proof. Choose an arbitrary node r as the root of T, and consider the subgraph consisting of the path from r to any leaf l. Without loss of generality we assume every edge is of the form $(u, u+1)$. Then $d_{11} = 1$ and $d_{ii} = \prod_{j=1}^{i-1} w_{j,j+1}$ for $i > 1$. We have $h_{ij} = d_{ii}d_{jj}k_{ij}$ for all i and j. If $j \neq i+1$, $h_{ij} = 0$, so $k_{ij} = 0$ as well. If $j = i+1$, $d_{ii}d_{jj} = \operatorname{sgn}(h_{ij})$, which implies that $\operatorname{sgn}(k_{ij}) = 1$ as claimed. Since l was chosen arbitrarily, this holds for every path from the root to a leaf, and so it holds for the entire tree. □

The canonicalization procedure is now clear. Given $\boldsymbol{H}$, we take T to be a spanning tree of $G(\boldsymbol{H})$ generated by a spanning tree algorithm $\mathcal{A}$. We let $\boldsymbol{D}$ be

Algorithm 4 Compute the canonicalization of a symmetric matrix $\boldsymbol{H}$ with respect to a spanning tree algorithm $\mathcal{A}$

function CANONICALIZE($\boldsymbol{H}$, $\mathcal{A}$)
  Run $\mathcal{A}$ on $G(\boldsymbol{H})$ to produce a spanning tree T with weights drawn from $G(\boldsymbol{H})$
  Mark 1 as the root of T
  Produce a $\boldsymbol{D}$ corresponding to a rooted coloring for T
  return $\boldsymbol{D}^{-1}\boldsymbol{H}\boldsymbol{D}$
end function

the matrix corresponding to a valid coloring of T, and we take the canonical representation of $\boldsymbol{H}$ to be $\boldsymbol{D}^{-1}\boldsymbol{H}\boldsymbol{D}$. Algorithm 4 is a restatement of this procedure, and is shown to be correct in Theorem 4.2.18.

Theorem 4.2.18. *Let $\boldsymbol{H}_{\mathcal{A}}$ denote the output of Algorithm 4 when given $\boldsymbol{H}$ and $\mathcal{A}$ as input. $\boldsymbol{H} \overset{\text{det}}{\equiv} \boldsymbol{K}$ if and only if $\boldsymbol{H}_{\mathcal{A}} = \boldsymbol{K}_{\mathcal{A}}$.*

Proof. We first assume that $\boldsymbol{H}_{\mathcal{A}} = \boldsymbol{K}_{\mathcal{A}}$. $\boldsymbol{H}_{\mathcal{A}} \overset{D}{\sim} \boldsymbol{H}$ and $\boldsymbol{K}_{\mathcal{A}} \overset{D}{\sim} \boldsymbol{K}$, so it follows that $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$. By Theorem 4.2.14, we have that $\boldsymbol{H} \overset{\text{det}}{\equiv} \boldsymbol{K}$.

We now assume that $\boldsymbol{H} \overset{\text{det}}{\equiv} \boldsymbol{K}$. By Theorem 4.2.14, $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$. This implies that $\boldsymbol{H}_{\mathcal{A}} \overset{D}{\sim} \boldsymbol{K}_{\mathcal{A}}$, and the weights of every edge of $T_{\boldsymbol{H}_{\mathcal{A}}}$ and $T_{\boldsymbol{K}_{\mathcal{A}}}$ are identically one by Lemma 4.2.17. We can therefore apply Lemma 4.2.16 to conclude that $\boldsymbol{H}_{\mathcal{A}} = \boldsymbol{K}_{\mathcal{A}}$. □

We note that a more general version of Lemma 4.2.16 follows as a corollary to Theorem 4.2.18:

Corollary 4.2.19. *Let $\boldsymbol{H} \overset{D}{\sim} \boldsymbol{K}$, and assume that there is a spanning tree T of $G(\boldsymbol{H}, \boldsymbol{K})$ such that the weights of T inherited from $G(\boldsymbol{H})$ are equal to the correspondings weights of T inherited from $G(\boldsymbol{K})$. Then $\boldsymbol{H} = \boldsymbol{K}$.*

Proof. If we run Algorithm 4 on $\boldsymbol{H}$ and $\boldsymbol{K}$ with an algorithm $\mathcal{A}$ that outputs T, we will compute a $\boldsymbol{D}$ such that $\boldsymbol{D}\boldsymbol{H}\boldsymbol{D}^{-1} = \boldsymbol{D}\boldsymbol{K}\boldsymbol{D}^{-1}$. The desired equality follows immediately. □

4.3 Solving the Symmetric Principal Minor Assignment Problem

In Section 4.2, we have characterized the set $\{\boldsymbol{K} : \boldsymbol{H} \overset{\det}{\equiv} \boldsymbol{K}\}$ for any symmetric matrix $\boldsymbol{H}$ and described how to pick a canonical representative from this set. In this section, we show how to reconstruct this canonical representative given a constant-time oracle for the principal minors of a symmetric matrix $\boldsymbol{H}$.

As before, chordless cycles will play a prominent role in our analysis. We begin by showing that two cyclic tridiagonal matrices with equal determinants are determinantally equivalent. This allows us to then show that knowing the determinant of a cyclic tridiagonal matrix, its diagonal entries, and all but one of the off-diagonal entries allows us to infer the final entry.

Lemma 4.3.1. *Let $\boldsymbol{H}$ and $\boldsymbol{K}$ be determinantally compatible and cyclic tridiagonal such that $\det(\boldsymbol{H}) = \det(\boldsymbol{K})$. Furthermore, let $h_{i,i+1} = k_{i,i+1}$ for all i between 1 and $n-1$. Then $h_{1n} = k_{1n}$.*

Proof. This follows immediately from Lemma 4.2.13. □

We consider the problem of reconstructing a cyclic tridiagonal matrix $\boldsymbol{T}$ from its principal minors. As always, we can infer the diagonal entries from the one-element principal minors, and the magnitude of the off-diagonal entries from the two-element principal minors. If we construct a spanning tree of $G(\boldsymbol{T})$, then there is exactly one edge whose sign is unknown. This satisfies the hypotheses of Lemma 4.3.1, and so we may infer the unknown sign. We refer to this process of inferring the sign of an edge as marking, and we consider every edge not in T to start unmarked.

We move on from this simple case to reconstructing a matrix $\boldsymbol{H}$ such that $G(\boldsymbol{H})$ consists of a simple cycle C with at least one chord (u, v). In this case we can use Algorithm 5 to reconstruct $\boldsymbol{H}$. We will prove that this procedure is

Algorithm 5 Given a cycle C with exactly one unmarked edge e, infer the sign of e

```
procedure MarkCycle(C, e)
    if C contains a chord e' then
        Let C_1 be the subcycle of C not containing e, and C_2 the other subcycle
        if e' is unmarked then
            MarkCycle(C_1, e')
        end if
        MarkCycle(C_2, e)
    else
        Mark e with its sign as determined from Lemma 4.2.13
    end if
end procedure
```

correct in Theorem 4.3.2.

Theorem 4.3.2. *Let $\boldsymbol{H}$ be such that $G(\boldsymbol{H})$ consists of a simple cycle with any number of chords. Given the diagonal entries of $\boldsymbol{H}$, the magnitude of the off-diagonal entries, and a spanning tree T with each edge marked positive, Algorithm 5 will correctly infer the signs of the entries corresponding to the unmarked edges.*

Proof. The proof is by induction on the number of edges in $\{e \in G(\boldsymbol{H}) : e \notin T\}$, which will be referred to as problematic edges. If there is exactly one problematic edge e, then the sign may be inferred as argued above. We now assume that the algorithm works for any cycle with up to k problematic edges, and we consider its operation on a cycle with $k+1$ problematic edges. Because there are multiple problematic edges, $G(\boldsymbol{H})$ must contain at least one chord e'. C_1 contains at most k problematic edges, so by induction it is marked correctly. Once e' is marked, C_2 contains at most k problematic edges, so it is also marked correctly by induction. At this point the entire cycle is marked correctly, so the algorithm works on a cycle with up to $k+1$ problematic edges. We have therefore completed the induction, and the algorithm is shown to be correct. □

We finally consider the general symmetric principal minor assignment problem. In this case, the structure of $G(\boldsymbol{H})$ is arbitrary. Here we may use Algorithm 6 to

Algorithm 6 Given a principal minor oracle for $\boldsymbol{H}$ and a deterministic spanning tree algorithm $\mathcal{A}$, output the canonicalization $\boldsymbol{H}_{\mathcal{A}}$

Use the first and second order principal minors of $\boldsymbol{H}$ to infer the diagonal elements and the magnitude of the off-diagonal elements
Let T be a spanning tree of $G(\boldsymbol{H})$ generated by $\mathcal{A}$ with every edge marked as positive
while $G(\boldsymbol{H})$ contains an unmarked edge (i, j) **do**
    Let C be the cycle consisting of (i, j) and the path from i to j in T
    MarkCycle(C, (i, j))
end while

perform the reconstruction. We will prove in Theorem 4.3.3 that this procedure is correct.

Theorem 4.3.3. *Given a principal minor oracle for $\boldsymbol{H}$ and a deterministic spanning tree algorithm $\mathcal{A}$, Algorithm 6 will correctly reconstruct the canonicalization $\boldsymbol{H}_{\mathcal{A}}$.*

Proof. Let $\tilde{\boldsymbol{H}}_{\mathcal{A}}$ denote the output of Algorithm 6 when given the oracle for $\boldsymbol{H}$ and $\mathcal{A}$ as input. By construction, $p\left(C_{\tilde{\boldsymbol{H}}_{\mathcal{A}}}\right) = p\left(C_{\boldsymbol{H}_{\mathcal{A}}}\right)$ for every simple chordless cycle C of $G\left(\tilde{\boldsymbol{H}}_{\mathcal{A}}, \boldsymbol{H}_{\mathcal{A}}\right)$. Therefore $\tilde{\boldsymbol{H}}_{\mathcal{A}} \overset{D}{\sim} \boldsymbol{H}_{\mathcal{A}}$ by Theorem 4.2.12, and $\tilde{\boldsymbol{H}}_{\mathcal{A}} \overset{\det}{\equiv} \boldsymbol{H}_{\mathcal{A}}$ by Theorem 4.2.14. $\tilde{\boldsymbol{H}}_{\mathcal{A}}$ and $\boldsymbol{H}_{\mathcal{A}}$ agree on the spanning tree T, so by Lemma 4.2.16, we can conclude that $\tilde{\boldsymbol{H}}_{\mathcal{A}} = \boldsymbol{H}_{\mathcal{A}}$. □

We close with two comments. The first regards the performance of Algorithm 6. While the algorithm will succeed given any spanning tree of $G(\boldsymbol{H})$ and any sequence of unmarked edges, we may end up considering the same cycle multiple times if the sequence is chosen poorly. We can avoid this by taking a breadth-first search spanning tree of $G(\boldsymbol{H})$ and marking the edges in decreasing order of the depths of their endpoints.

The second regards the problem of computing the determinant of a matrix $\boldsymbol{H}$ given its diagonal entries, the magnitude of its off-diagonal entries, and the value of $p(C)$ for every simple chordless cycle C of $G(\boldsymbol{H})$. In this case we can modify Algorithm 6 to assign signs to the entries of $\boldsymbol{H}$ consistent with the given sign

products. From there, we may compute the determinant directly.

Chapter 5

Open Problems and Future Work

In lieu of some grand and overarching conclusion, we close by compiling a list of the open problems and interesting future research directions given in this dissertation. We hope that this will be useful to future researchers.

Chapter 2

- We have interpretations for the diagonal elements of the kernel of a determinantal process, and for the magnitudes of its off-diagonal elements. How shall we interpret the signs of the off-diagonal elements?
- Develop a theory of conditional independencies for determinantal point processes.
- Flesh out the interpretation of probabilities assigned by a determinantal point process in terms of exponential entropy.
- Develop a theory of convergence of discrete kernels to continuous.
- We have seen that determinantal point processes on a finite set behave like multivariate Gaussian distributions. To what extent do general determinantal point processes behave like Gaussian processes?

- Is the theory of determinantal point processes useful in the study of longest monotone subsequences?

Chapter 3

- Find new examples of interesting phenomena which are determinantal point processes.
- Continue investigating the class of exchangeable determinantal point processes. Is the cardinality of a draw an exponential family of distributions?
- Find new kernels which are analytically tractable and can be analyzed in the manner of the Brownian kernel.
- Find applications of the hyperbolic cosine series distribution.

Chapter 4

- Solve the principal minor assignment problem for nonsymmetric matrices. In particular, solve it for Hermitian matrices.
- Find a way to estimate a matrix given noisy measurements of its principal minors.
- Study the geometry of the quotient space of symmetric matrices modulo $\boldsymbol{D}$-similarity.

General

- Find a simple and efficiently checkable necessary and sufficient condition for a point process to be determinantal.
- Suppose $\boldsymbol{H} \succeq 0$ is determinantally compatible with $\alpha\boldsymbol{I} + \beta\boldsymbol{J}$ but $\boldsymbol{H} \overset{D}{\not\sim} \alpha\boldsymbol{I} + \beta\boldsymbol{J}$. What can we say about the distribution of $Y \sim \mathrm{LDPP}(\boldsymbol{H})$?

Bibliography

R. H. Affandi, A. Kulesza, and E. B. Fox. Markov Determinantal Point Processes. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, 2012.

R. H. Affandi, A. Kulesza, E. Fox, and B. Taskar. Nyström Approximation for Large-Scale Determinantal Processes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2013.

A. Arlotto, R. W. Chen, L. A. Shepp, and J. M. Steele. On-Line Selection of Alternating Subsequences from a Random Sample. *Journal of Applied Probability*, 48:1114–1132, 2011.

F. B. Baker and M. R. Harwell. Computing Elementary Symmetric Functions and Their Derivatives: A Didactic. *Applied Psychological Measurement*, 20:169–192, 1996.

N. Barth. The Gramian and K-Volume in N-Space: Some Classical Results in Linear Algebra. *Journal of Young Investigators*, 2, 1999.

A. Borodin. Determinantal Point Processes. In G. Akemann, J. Baik, and P. D. Francesco, editors, *The Oxford Handbook of Random Matrix Theory*, chapter 11, pages 231–249. Oxford University Press, 2011. ISBN 9780199574001. arXiv:0911.1153.

A. Borodin and V. Gorin. Shuffling Algorithm for Boxed Plane Partitions. *Advances in Mathematics*, 220(6):1739–1770, 2009. arXiv:0804.3071.

A. Borodin and G. Olshanski. Asymptotics of Plancherel–Type Random Partitions. *Journal of Algebra*, 313(1):40–60, 2007. arXiv:math/0610240.

A. Borodin and E. M. Rains. Eynard-Mehta theorem, Schur Process, and Their Pfaffian Analogs. *Journal of Statistical Physics*, 121:291–317, 2005. arXiv:math-ph/0409059.

A. Borodin, P. Diaconis, and J. Fulman. On Adding a List of Numbers (and Other One-Dependent Determinantal Processes). *Bulletin of the American Mathematical Society*, 47(4):639–670, 2010. arXiv:0904.3740.

L. D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, 1986. ISBN 0940600102.

R. Burton and R. Pemantle. Local Characteristics, Entropy and Limit Theorems for Spanning Trees and Domino Tilings Via Transfer-Impedances. *Annals of Probability*, 21(3):1329–1371, 1993.

L. L. Campbell. Exponential Entropy as a Measure of Extent of a Distribution. *Probability Theory and Related Fields*, 5(3):217–225, 1966.

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Elementary Theory and Methods*, volume 1. Springer, 2002. ISBN 0387955410.

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: General Theory and Structure*, volume 2. Springer, 2007. ISBN 0387213376.

R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 4th edition, 2010. ISBN 0521765390.

F. J. Dyson. A Brownian-Motion Model for the Eigenvalues of a Random Matrix. *Journal of Mathematical Physics*, 3:1191–1198, 1962a.

F. J. Dyson. Statistical Theory of the Energy Levels of Complex Systems. I. *Journal of Mathematical Physics*, 3:140–156, 1962b.

F. J. Dyson. Statistical Theory of the Energy Levels of Complex Systems. II. *Journal of Mathematical Physics*, 3:157–165, 1962c.

F. J. Dyson. Statistical Theory of the Energy Levels of Complex Systems. III. *Journal of Mathematical Physics*, 3:166–175, 1962d.

F. J. Dyson. The Threefold Way. Algebraic Structure of Symmetry Groups and Ensembles in Quantum Mechanics. *Journal of Mathematical Physics*, 3(6):1199–1215, 1962e.

G. M. Engel and H. Schneider. Matrices Diagonally Similar to a Symmetric Matrix. *Linear Algebra and its Applications*, 29:131–138, 1980.

G. Engeln-Müllges and F. Uhlig. *Numerical Algorithms with C*. Springer, 1996. ISBN 3540605304.

W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, 3rd edition, 1968. ISBN 0471257087.

W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, 1971. ISBN 0471257095.

J. Gillenwater, A. Kulesza, and B. Taskar. Discovering Diverse and Salient Threads in Document Collections. In *Proceedings of the 2012 Conference on Empirical Methods in Machine Learning*, 2012a.

J. Gillenwater, A. Kulesza, and B. Taskar. Near-Optimal MAP Inference for Determinantal Point Processes. In *Advances in Neural Information Processing Systems 25*, 2012b.

J. Ginibre. Statistical Ensembles of Complex, Quaternion, and Real Matrices. *Journal of Mathematical Physics*, 6:440–449, 1965.

V. Gorin. Non–Intersecting Paths and Hahn Orthogonal Polynomial Ensemble. *Functional Analysis and Its Applications*, 42(3):180–197, 2008. arXiv:0708.2349.

K. Griffin and M. J. Tsatsomeros. Principal Minors, Part I: A Method for Computing All the Principal Minors of a Matrix. *Linear Algebra and its Applications*, 419:107–124, 2006a.

K. Griffin and M. J. Tsatsomeros. Principal Minors, Part II: The Principal Minor Assignment Problem. *Linear Algebra and its Applications*, 419:125–171, 2006b.

J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. Determinantal Processes and Independence. *Probability Surveys*, 3:206–229, 2006.

J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. *Zeros of Gaussian Analytic Functions and Determinantal Point Processes*. American Mathematical Society, 2009. ISBN 9780821843734.

K. Johansson. Determinantal Processes with Number Variance Saturation. *Communications in Mathematical Physics*, 252(1):111–?148, 2004. arXiv:math/0404133.

K. Johansson. The Arctic Circle Boundary and the Airy Process. *The Annals of Probability*, 33(1):1–?30, 2005a. arXiv:math/0306216.

K. Johansson. Random Matrices and Determinantal Processes. arXiv:math-ph/0510038, 2005b. Lectures given at the summer school on Mathematical statistical mechanics in July 05 at Ecole de Physique, Les Houches.

S. Karlin and J. McGregor. Coincidence Probabilities. *Pacific Journal of Mathematics*, 9(4):1141–1164, 1959.

S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes*. Academic Press, 2nd edition, 1975. ISBN 0123985528.

M. Katori and H. Tanemura. Noncolliding Brownian Motion and Determinantal Processes. *Journal of Statistical Physics*, 129:1233–1277, 2007. arXiv:0705.2460.

W. König. Orthogonal Polynomial Ensembles in Probability Theory. *Probability Surveys*, 2:385–447, 2005. arXiv:math/0403090.

A. Kulesza. *Learning with Determinantal Point Processes*. PhD thesis, University of Pennsylvania, 2012.

A. Kulesza and B. Taskar. k-DPPs: Fixed-Size Determinantal Point Processes. In *Proceedings of the 28th International Conference on Machine Learning*, 2011a.

A. Kulesza and B. Taskar. Learning Determinantal Point Processes. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, 2011b.

A. Kulesza and B. Taskar. Structured Determinantal Point Processes. In *Advances in Neural Information Processing Systems 23*, 2011c.

A. Kulesza and B. Taskar. Determinantal Point Processes for Machine Learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012. arXiv:1207.6083.

H. J. Kushner. On the Weak Convergence of Interpolated Markov Chains to a Diffusion. *The Annals of Probability*, 2(1):40–50, 1974.

G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the Kernel Matrix with Semidefinite Programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

M. Loève. Paul Lévy. *Annals of Probability*, 1(1):1–8, 1973.

B. F. Logan and L. A. Shepp. A Variational Problem for Random Young Tableaux. *Advances in Mathematics*, 26:206–222, 1977.

R. Lyons. Determinantal Probability Measures. *Publications Mathématiques de l'IHÉS*, 98(1):167–212, 2003.

R. Lyons and J. E. Steif. Stationary Determinantal Processes: Phase Multiplicity, Bernoullicity, Entropy, and Domination. *Duke Mathematical Journal*, 120(3): 515–575, 2003.

O. Macchi. The Coincidence Approach to Stochastic Point Processes. *Advances in Applied Probability*, 7(1):83–122, 1975.

M. L. Mehta and M. Gaudin. On the Density of Eigenvalues of a Random Matrix. *Nuclear Physics*, 18:420–427, 1960.

G. Olshanski. Difference Operators and Determinantal Point Processes. *Functional Analysis and its Applications*, 42(4):317–329, 2008. arXiv:0810.3751.

Y. Peres and B. Virág. Zeros of the i.i.d. Gaussian Power Series: a Conformally Invariant Determinantal Process. *Acta Mathematica*, 194(1):1–35, 2005. arXiv:math/0310297.

A. Reffgen. The Determinant in Finite– and Infinite–Dimensional Vector Spaces. Master's thesis, Lund University, 2003. URL `http://www.nek.lu.se/NEKARE/cv/reffgen2003.pdf`.

S. M. Ross. *Stochastic Processes*. Wiley, 2nd edition, 1996. ISBN 0471120626.

T. Shirai and Y. Takahashi. Random Point Fields Associated with Certain Fredholm Determinants I: Fermion, Poisson and Boson Point Processes. *Journal of Functional Analysis*, 205(2):414–463, 2003a.

T. Shirai and Y. Takahashi. Random Point Fields Associated with Certain Fredholm Determinants II: Fermion Shifts and Their Ergodic and Gibbs Properties. *The Annals of Probability*, 31(3):1533–1564, 2003b.

F. Smithies. *Integral Equations*. Cambridge University Press, 1958. ISBN 052106502X.

A. Soshnikov. Determinantal Random Point Fields. *Russian Mathematical Surveys*, 55(5):923–975, 2000.

A. Soshnikov. Determinantal Random Fields. In *Encyclopedia of Mathematical Physics*, pages 47–53. Elsevier, 2006.

T. Tao. Determinantal Processes, August 2009. URL `http://terrytao.wordpress.com/2009/08/23/determinantal-processes/`.

A. Turner. Convergence of Markov Processes. Master's thesis, University of Cambridge, 2002. URL `http://www.maths.lancs.ac.uk/~turnera/essay.pdf`.