*Bernoulli* **25**(4B), 2019, 3555–3589
https://doi.org/10.3150/18-BEJ1102

# Asymptotic equivalence of fixed-size and varying-size determinantal point processes

SIMON BARTHELMÉ, PIERRE-OLIVIER AMBLARD and
NICOLAS TREMBLAY

*CNRS, Gipsa-lab, Grenoble INP and Université Grenoble Alpes, 11 rue des Mathématiques, Grenoble Campus, BP46, F-38402 Saint Martin d'Heres Cedex, France. E-mail: simon.barthelme@gipsa-lab.fr; pierre-olivier.amblard@gipsa-lab.fr; nicolas.tremblay@gipsa-lab.fr*

Determinantal Point Processes (DPPs) are popular models for point processes with repulsion. They appear in numerous contexts, from physics to graph theory, and display appealing theoretical properties. On the more practical side of things, since DPPs tend to select sets of points that are some distance apart (repulsion), they have been advocated as a way of producing random subsets with high diversity. DPPs come in two variants: fixed-size and varying-size. A sample from a varying-size DPP is a subset of random cardinality, while in fixed-size "$k$-DPPs" the cardinality is fixed. The latter makes more sense in many applications, but unfortunately their computational properties are less attractive, since, among other things, inclusion probabilities are harder to compute. In this work, we show that as the size of the ground set grows, $k$-DPPs and DPPs become equivalent, in the sense that fixed-order inclusion probabilities converge. As a by-product, we obtain saddlepoint formulas for inclusion probabilities in $k$-DPPs. These turn out to be extremely accurate, and suffer less from numerical difficulties than exact methods do. Our results also suggest that $k$-DPPs and DPPs also have equivalent maximum likelihood estimators. Finally, we obtain results on asymptotic approximations of elementary symmetric polynomials which may be of independent interest.

*Keywords:* determinantal point processes; point processes; saddlepoint approximation

Determinantal Point Processes originally arose in quantum physics (Macchi [13]) and random matrix theory (Soshnikov [16]), but they are such natural objects that they have also been rediscovered within computer science (Deshpande *et al.* [7], Deshpande and Rademacher [6]) and that special cases have appeared in the statistics literature as well (Chen, Dempster and Liu [2]). Within Machine Learning, their current popularity owes much to Kulesza and Taskar [11], whose overall approach we will mostly follow here. Like them, we focus on discrete DPPs.

Kulesza and Taskar [11] advocate DPPs as tractable probabilistic models for diverse subsets. Specifically, we assume that we have a ground set of $n$ items, $\Omega = x_1 \dots x_n$, of which we wish to retain a subset $\mathcal{X} \subseteq \Omega$. Our requirement is that $\mathcal{X}$ be diverse, that is, that it should not contain items that are too much alike, or, put differently, that it be representative of the range of items found in $\Omega$. A DPP is essentially a way of picking a random $\mathcal{X}$ that has this property with high probability.

We introduce DPPs formally below, but a salient feature of classical DPPs is that the cardinal of $\mathcal{X}$ is a random variable. Since this is not always suitable, Kulesza and Taskar [10] have introduced a fixed-size variant (so-called $k$-DPPs), which are nothing more than DPPs conditioned

1350-7265 © 2019 ISI/BS

on the event that $|\mathcal{X}| = k$. $k$-DPPs share some features with DPPs but unfortunately lose some tractability.

In this work, we show that this loss of tractability only matters for very small $n$. In large sets, $k$-DPPs and DPPs converge in a sense we make precise below, but roughly means that the probability that item $x_i$ ends up in set $\mathcal{X}$ is almost the same under a $k$-DPP and a matched DPP. Moreover, this is true for bi-inclusions (i.e., the event that $x_i$ and $x_j$ are in $\mathcal{X}$) or indeed for joint inclusion probabilities of any fixed order.[1]

Practically speaking, the ability to compute inclusion probabilities is essential when $k$-DPPs are used for importance sampling. For example, in Tremblay, Barthelmé and Amblard [19], $k$-DPPs are used to estimate averages: let $L = \sum_{i=1}^{n} f(x_i)$. If $\mathcal{X}$ is sampled from a $k$-DPP, the average $L$ can be estimated from the values of $f$ in $\mathcal{X}$. Since not all items have equal probability of appearing in a $k$-DPP, we have to reweight by the inverse inclusion probability to form the unbiased estimate:

$$\hat{L}(\mathcal{X}) = \sum_{i=1}^{n} \frac{f(x_i)\mathbb{I}(i \in \mathcal{X})}{p(i \in \mathcal{X})} \tag{0.1}$$

Here we therefore need first-order inclusion probabilities. To estimate a pairwise quantity (e.g., mean distance), we would need second-order inclusion probabilities, and so on.

Our results lead to stable and accurate approximations for inclusion probabilities, as described in Section 3, and stable algorithms for sampling $k$-DPPs with relatively large $k$. They also clarify the links between $k$-DPPs and DPPs, and when the one should look like the other.

The article is structured as follows: in Section 1, we introduce notation and recall results on DPPs and $k$-DPPs. Section 2 contains our main theoretical results. The practical algorithms that follow are described in Section 3. Section 4 contains simulation results.

To prove our main result, we use saddlepoint approximations and a perturbation argument, but readers who wish to skip the technical details will find an intuitive argument in Section 2.1, where we explain that DPPs are just exponentially relaxed $k$-DPPs. Essentially, the strict constraint $|X| = k$ that appears in $k$-DPPs is relaxed to a soft constraint in DPPs, and the difference between the soft and the hard constraint becomes irrelevant in large $n$.

# 1. Background

In this section, we introduce notation and some basic results.

## 1.1. Notation

We deal with finite ground sets, so without loss of generality, we may take $\Omega = \{1, \dots, n\}$. Fixed subsets of $\Omega$ are then equivalent to multi-indices and noted $\boldsymbol{\alpha}$, with cardinality noted $|\boldsymbol{\alpha}|$. Random

[1] To be precise: $k$-DPPs and DPPs cannot be equivalent in a strong sense, since they do not have the same support (one generates a fixed size set, the other doesn't). However, for $n$ and $k$ large enough, the probability that they include a certain fixed subset converges.

subsets are noted $\mathcal{X}$ or $\mathcal{Y}$. Expectation is noted $E(\cdot)$, and $\mathbb{I}$ is the indicator function, so that e.g., $E(\mathbb{I}(i \in \mathcal{X})) = p(i \in \mathcal{X})$. There are two equivalent viewpoints when dealing with finite random subsets: one is to look at $\mathcal{X}$, a subset, as the random variable. Another is to consider binary strings of length $n$, which indicate whether item $i$ is included in $\mathcal{X}$. We note such strings $z$, and depending on context one or the other viewpoint is more convenient. Matrices are in bold capitals, for example, $\mathbf{L}$. The identity matrix is noted $\mathbf{I}$. Individual entries in a matrix are noted using capitals: $L_{ij}$ is entry $(i, j)$ in matrix $\mathbf{L}$. Sub-matrices are in bold, with indices, for example $\mathbf{L}_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ is the sub-matrix of $\mathbf{L}$ with rows indexed by $\boldsymbol{\alpha}$ and columns indexed by $\boldsymbol{\beta}$. So-called "Matlab" notation is used occasionally, so that the submatrix formed by selecting all rows in $\boldsymbol{\alpha}$ is noted $\mathbf{L}_{\boldsymbol{\alpha},:}$, and $\mathbf{L}_{:,1:k}$ is the submatrix containing the first $k$ columns. For simplicity, a single index is used if it is repeated: $\mathbf{L}_{\boldsymbol{\alpha}} = \mathbf{L}_{\boldsymbol{\alpha},\boldsymbol{\alpha}}$. Sub-matrices and sub-vectors formed by excluding elements are noted with a minus sign, for example, the index $\boldsymbol{\alpha}_{-j}$ includes all elements in $\boldsymbol{\alpha}$ except index $j$.

## 1.2. Some lemmas

We will need two well-known lemmas in the course of this work. The first one (Cauchy–Binet) is central to the theory of DPPs, the second is an easy lemma on inclusion probabilities.

The Cauchy–Binet lemma expresses the determinant of a matrix product as a sum of products of determinants.

**Lemma 1.1 (Cauchy–Binet).** *Let* $\mathbf{M} = \mathbf{AB}$, *with* $\mathbf{A}$ *a* $n \times m$ *matrix*, $\mathbf{B}$ *a* $m \times n$ *matrix. We assume* $m \geq n$. *Then*:

$$\det \mathbf{M} = \sum_{\boldsymbol{\alpha}, |\boldsymbol{\alpha}|=n} \det \mathbf{A}_{:,\boldsymbol{\alpha}} \det \mathbf{B}_{\boldsymbol{\alpha},:} \tag{1.1}$$

*where* $\boldsymbol{\alpha}$ *is a multi-index of length* $n$. *The sum is over all multi-indices* $\boldsymbol{\alpha}$, *of which there are* $\binom{m}{n}$.

The second lemma is an easy lemma on sums of inclusion probabilities. An inclusion probability is the probability that a certain item (or items) appear in a random set.

**Lemma 1.2 (Sums of inclusion probabilities).** *Let* $\Omega$ *designate a base set of items, and* $\mathcal{X}$ *a random subset of* $\Omega$. *Let* $\boldsymbol{\alpha}$ *designate a fixed subset of items of cardinality* $m$. $p(\boldsymbol{\alpha} \subseteq \mathcal{X})$ *is called an inclusion probability. We have that*: $\sum_{\boldsymbol{\alpha},|\boldsymbol{\alpha}|=m} p(\boldsymbol{\alpha} \subseteq \mathcal{X}) = E(\binom{|\mathcal{X}|}{m})$, *where the expectation is over the random set* $\mathcal{X}$. *In particular*:

1. *if* $\mathcal{X}$ *is a set of fixed size* $k$, *the sum equals* $\binom{k}{m}$.
2. *if* $m = 1$, *the sum equals* $E(|\mathcal{X}|)$

**Proof.**

$$\sum_{\boldsymbol{\alpha},|\boldsymbol{\alpha}|=m} p(\boldsymbol{\alpha} \subseteq \mathcal{X}) = \sum_{\boldsymbol{\alpha},|\boldsymbol{\alpha}|=m} E\big(\mathbb{I}(\boldsymbol{\alpha} \subseteq \mathcal{X})\big)$$

$$= E\left(\sum_{\boldsymbol{\alpha}, |\boldsymbol{\alpha}|=m} \mathbb{I}(\boldsymbol{\alpha} \subseteq \mathcal{X})\right)$$

$$= E\left(\binom{|\mathcal{X}|}{m}\right) \qquad \square$$

**Remark 1.1.** For large sets, $E\left(\binom{|\mathcal{X}|}{m}\right) = \frac{1}{m!}E(|\mathcal{X}|(|\mathcal{X}|-1)\cdots(|\mathcal{X}|-m+1)) = \frac{1}{m!}E(O(|\mathcal{X}|^m))$ As a consequence, the sum of order-$m$ inclusion probabilities for a set of fixed size $k$ is $O(\frac{k^m}{m!})$. We use this fact to properly normalise the total variation distance, see Section 2.2.

## 1.3. Elementary symmetric polynomials

The Elementary Symmetric Polynomials (ESPs) of a matrix play an important role in the theory of $k$-DPPs, and one of our core problems will be to find asymptotic formulas for them. Let $\mathbf{L}$ denote a positive definite matrix and $\lambda_1 \dots \lambda_n$ its eigenvalues. The $k$-th ESP is a sum of all the products of $k$ eigenvalues:

$$e_k(\boldsymbol{\lambda}) = \sum_{\boldsymbol{\alpha}, |\boldsymbol{\alpha}|=k} \prod_{j \in \boldsymbol{\alpha}} \lambda_j \tag{1.2}$$

For example, $e_2(\boldsymbol{\lambda}) = \sum_{i<j} \lambda_i \lambda_j$. Interesting special cases include $e_1(\boldsymbol{\lambda}) = \sum \lambda_i = \mathrm{Tr}(\mathbf{L})$, and $e_n(\boldsymbol{\lambda}) = \prod \lambda_i = \det \mathbf{L}$. There is a rich theory on ESPs, going back at least to Newton, with interesting modern developments (Mariet and Sra [14], Jozsa and Mitchison [9]). As we explain below, they occur in $k$-DPPs as normalisation constants, and ratios of ESPs appear in inclusion probabilities.

## 1.4. DPPs

DPPs are defined such as to produce random subsets that are not overly redundant, where the notion of redundancy is defined with respect to a (positive definite) similarity function.

We have a collection $\Omega$ of items ordered from 1 to $n$. We associate to each pair of items a similarity score $L_{ij}$, such that the matrix $\mathbf{L}$ with entries $L_{ij}$ is positive definite. The matrix $\mathbf{L} \in \mathbb{R}^{n \times n}$ is called the L-ensemble of the DPP.[2]

**Definition 1.** A Determinantal Point Process is a random subset $\mathcal{X}$ of $1 \dots n$ with probability mass function given by:

$$p(\mathcal{X}) = \frac{\det(\mathbf{L}_{\mathcal{X}})}{\det(\mathbf{I} + \mathbf{L})} \tag{1.3}$$

[2] We find it more natural to define DPPs via the L-ensemble, since the more common definition via the marginal kernel does not carry over to fixed-size DPPs.

The preference for diverse subsets built into DPPs comes from the fact that if a subset $\mathcal{X}$ includes items that are too similar, the matrix $\mathbf{L}_{\mathcal{X}}$ will have nearly colinear columns, and its determinant will be close to 0.

An interesting aspect of DPPs is how tractable the marginals are. The inclusion probabilities, that is, the probability that item $i$ is in $\mathcal{X}$, are given by the so-called "marginal kernel" matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$, where

$$\mathbf{K} = (\mathbf{I} + \mathbf{L})^{-1}\mathbf{L} \tag{1.4}$$

Specifically, for a DPP, $p(i \in \mathcal{X}) = K_{ii}$. More generally, inclusion probabilities are given by principal minors of the marginal kernel, e.g., if $\boldsymbol{\alpha}$ is a subset of $\Omega$:

$$p(\boldsymbol{\alpha} \subseteq \mathcal{X}) = \det(\mathbf{K}_{\boldsymbol{\alpha}}) \tag{1.5}$$

A DPP can generate random subsets of any size from 1 to $n$. The expected cardinality of $\mathcal{X}$ can also be read out from the marginal kernel, specifically:

$$E\big(|\mathcal{X}|\big) = \operatorname{Tr}(\mathbf{K}) = \sum \frac{\lambda_i}{1+\lambda_i} \tag{1.6}$$

where the $\lambda_i$'s designate the eigenvalues of the L-ensemble $\mathbf{L}$.

## 1.5. *k*-DPPs

**Definition 2.** A $k$-DPP is a DPP conditioned on the size of the sampled set $|\mathcal{X}| = k$. In other words, the probability mass function stays the same but now the sample space is the set of subsets of $1 \ldots n$ of size $k$, and

$$p\big(\mathcal{X} \big| |\mathcal{X}| = k\big) \propto \begin{cases} \det(\mathbf{L}_{\mathcal{X}}) & \text{if } |\mathcal{X}| = k \\ 0 & \text{otherwise} \end{cases} \tag{1.7}$$

**Remark 1.2.** Contrary to DPPs, $k$-DPPs are insensitive to the overall scaling of the L-ensemble. Since

$$\det(\beta \mathbf{L}_{\mathcal{X}}) = \beta^k \det(\mathbf{L}_{\mathcal{X}}),$$

the probability density (1.7) is invariant to any rescaling by a factor $\beta > 0$.

An important property of $k$-DPPs, one that unlocks many analytical simplifications, is that $k$-DPPs are a mixture distribution. The mixture involves a diagonal $k$-DPP and a projection $k$-DPP, two objects that are simpler than a generic $k$-DPP.

The mixture property is a consequence of the Cauchy–Binet formula (Lemma 1.1). Let $\mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ denote the spectral decomposition of $\mathbf{L}$, with $\mathbf{D} = \operatorname{diag}(\lambda_1 \ldots \lambda_n)$ and $\mathbf{U}$ the matrix of eigenvectors. Then

$$p\big(\mathcal{X} \big| |\mathcal{X}| = k\big) = \frac{1}{Z}\det(\mathbf{L}_{\mathcal{X}}) = \frac{1}{Z}\sum_{\mathcal{Y}, |\mathcal{Y}|=k} \det\big(\mathbf{U}_{\mathcal{X},\mathcal{Y}}\mathbf{U}_{\mathcal{X},\mathcal{Y}}^\top\big)\det(\mathbf{D}_{\mathcal{Y}}) \tag{1.8}$$

where $Z$ is an integration constant (to be defined later), $\mathcal{Y}$ is a subset of *columns* of $\mathbf{U}$, and the sum is over all such subsets of size $k$. Equation (1.8) shows that the probability mass function has the form of a mixture distribution, where we first choose a set of *eigenvalues* (with indices $\mathcal{Y}$) from a $k$-DPP with *diagonal* L-ensemble $\mathbf{D}$ and then choose a set of items $\mathcal{X}$ from a $k$-DPP with L-ensemble $\mathbf{U}_{:,\mathcal{Y}}\mathbf{U}_{:,\mathcal{Y}}^{\top}$. The latter is a specific kind of DPP, called a "projection DPP".

The same mixture interpretation holds for DPPs as well. In the case of DPPs, the rule for sampling the set $\mathcal{Y}$ of eigenvalues is simpler. Each eigenvalue is sampled independently and included with probability $\frac{\lambda_i}{1+\lambda_i}$. Once we have the eigenvalues, we proceed in exactly the same way as above: form a projection kernel, and sample the corresponding projection DPP.

### 1.5.1. *Projection DPPs*

**Definition 3.** A projection DPP is a $k$-DPP whose L-ensemble has the following form:

$$\mathbf{L} = \mathbf{V}\mathbf{V}^{\top} \tag{1.9}$$

where $\mathbf{V}_{n\times k}$ has orthonormal columns (i.e., $\mathbf{V}^{\top}\mathbf{V} = \mathbf{I}$).

Projection DPPs have a set of properties that make them especially tractable. The most salient is that the marginal kernel equals the L-ensemble, e.g., the inclusion probability of item $i$ equals $L_{ii}$, as shown in the following lemma.

**Lemma 1.3.** *In a projection DPP with L-ensemble* $\mathbf{L}$, $p(\boldsymbol{\alpha} \subseteq \mathcal{X}) = \det(\mathbf{L}_{\mathcal{X}})$.

**Proof.** See appendix. □

This result is proved rigorously in the appendix, but straightforward if one looks at projection DPPs as DPPs taken to a certain limit. Consider a DPP with the following L-matrix, indexed by parameter $\gamma > 0$:

$$\mathbf{L}(\gamma) = \mathbf{R}\mathbf{D}(\gamma)\mathbf{R}^{\top} \tag{1.10}$$

where $D(\gamma)$ is a diagonal matrix with entries on the diagonal equal to $\gamma$ repeated $k$ times, followed by $\gamma^{-1}$, repeated $n-k$ times, and $\mathbf{R}$ is a $n \times n$ orthonormal matrix. Let $\gamma \to \infty$. Following the mixture interpretation of DPPs, we see that the probability of picking one of the first $k$ eigenvalues equals $\gamma/(1+\gamma)$, which tends to 1, while the probability of picking one of the latter $n-k$ tends to 0. This means that with increasing $\gamma$ we end up always picking the same $k$ eigenvalues, and hence always sampling the same $k$-DPP, one with kernel $\mathbf{R}_{:,1:k}\mathbf{R}_{:,1:k}^{\top}$. The marginal probabilities are given by the corresponding marginal kernel: $\mathbf{R}\mathbf{D}_m(\gamma)\mathbf{R}^{\top}$ where $\mathbf{D}_m(\gamma)$ has first $k$ entries equal to $\frac{\gamma}{1+\gamma}$, and the next $n-k$ equal to $\frac{1}{\gamma+1}$. In the large-$\gamma$ limit, the marginal kernel thus equals $\mathbf{R}_{:,1:k}\mathbf{R}_{:,1:k}^{\top}$ as claimed. The limit is however improper, as some entries in the L-matrix tend to infinity.

To sum up: if the L-ensemble is a projection matrix of rank $k$, then a $k$-DPP is also a DPP. We can even extend this further to *all* L-ensembles of rank $k$.

**Result 1.** Let $\mathbf{L}$ have rank $k$, with eigendecomposition $\mathbf{L} = \mathbf{UDU}^\top$. Without loss of generality, we assume that $\mathbf{U}$ is of size $n \times k$ and $\mathbf{D}$ a diagonal matrix of size $k \times k$ with non-null diagonal elements. Then a $k$-DPP with L-ensemble $\mathbf{L}$ is also a projection DPP, with marginal kernel equal to $\mathbf{UU}^\top$.

**Proof.** $\mathbf{L}$ has rank $k$, so in the eigendecomposition $\mathbf{U}$ is $n \times k$, and $\mathbf{D}$ is a diagonal matrix of size $k \times k$. If $\mathcal{X}$ is a subset of size $k$, we have

$$\det \mathbf{L}_{\mathcal{X}} = \det \mathbf{U}_{\mathcal{X},:}\mathbf{D}\mathbf{U}_{:,\mathcal{X}}^\top$$

and since the matrices involved are square, we have:

$$\det \mathbf{L}_{\mathcal{X}} = \det D\big(\det \mathbf{U}_{\mathcal{X},:}\mathbf{U}_{:,\mathcal{X}}^\top\big)$$

Then $p(\mathcal{X}) \propto (\det \mathbf{U}_{\mathcal{X},:}\mathbf{U}_{:,\mathcal{X}}^\top)$, which is the probability mass function of a projection DPP and the result follows. □

This result hints at a close kinship between $k$-DPPs and DPPs, and convergence results bear this out.

### 1.5.2. *Inclusion probabilities in k-DPPs*

Since a $k$-DPP is a mixture of projection-DPPs (eq. (1.8)), the first order inclusion probability for item $i$ can be expressed as

$$p\big(i \in \mathcal{X} || \mathcal{X}| = k\big) = E_{\mathcal{Y}}\big((\mathbf{U}_{\mathcal{Y}}\mathbf{U}_{\mathcal{Y}}^\top)_{ii}\big) \tag{1.11}$$

$$= E_{\mathcal{Y}}\left(\sum_{j=1}^{n} U_{ij}^2 \mathrm{I}(j \in \mathcal{Y})\right) \tag{1.12}$$

$$= \sum_{j=1}^{n} U_{ij}^2 P(j \in \mathcal{Y}) \tag{1.13}$$

$$= \big(\mathbf{U}\,\mathrm{diag}(\boldsymbol{\pi})\mathbf{U}^\top\big)_{ii} \tag{1.14}$$

where $\pi_j = p(j \in \mathcal{Y})$, the probability that the $j$-th eigenvector is included in set $\mathcal{Y}$. Formulas for higher-orders (joint inclusion probabilities) are in Section A.2.

Computing the inclusion probabilities for a $k$-DPP thus boils down to computing inclusion probabilities in a *diagonal* $k$-DPP, and combining them with the eigenvectors of $\mathbf{L}$.

## 1.6. Diagonal DPPs and *k*-DPPs

In the special case of diagonal DPPs and $k$-DPPs, the L-ensemble is a diagonal matrix. A diagonal DPP turns out to be nothing more than a Bernoulli process. If conditioned to be of fixed size $k$, a diagonal $k$-DPP is obtained.

So far we have kept with the usual viewpoint on DPPs, which sees them as random sets. Alternatively, a sample from a discrete DPP can be viewed as a binary string $\boldsymbol{z}$ of size $n$, where $z_i = 1$ indicates inclusion of the $i$-th item, and $\sum_{i=1}^{n} z_i = k$. In this section, we prefer the latter viewpoint, because it lightens notation.

In this notation, the inclusion probability of item $i$ equals the marginal probability of $z_i$, $p(z_i = 1)$, and similarly for joint probabilities $p(z_i = 1, z_j = 1)$, etc. $p(\boldsymbol{z}) = p(z_1 \dots z_n)$ is the likelihood of the draw.

### 1.6.1. *Diagonal DPPs*

Consider a DPP with diagonal L-ensemble

$$\mathbf{L} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$$

Following Eq. (1.4), $\mathbf{K}$ is diagonal too, with entries $K_{ii} = \pi_i = \frac{\lambda_i}{1+\lambda_i}$. The fact that the marginal kernel is diagonal implies that $p(z_i = 1, z_j = 1) = \det(\mathbf{K}_{\{i,j\}}) = \pi_i \pi_j = p(z_i = 1) p(z_j = 1)$, with similar results for higher-order probabilities. We conclude that (viewed as a binary string) a diagonal DPP is a product of independent Bernoulli variables, where each $z_i$ is drawn with probability $\pi_i$.

### 1.6.2. *Diagonal k-DPPs*

Viewed as distributions over binary strings, diagonal DPPs are a product measure, meaning that each $z_i$ is sampled independently. Diagonal $k$-DPPs are not, due to the constraint that $\sum z_i = k$. The density of a diagonal $k$-DPP is given by:

$$p(\boldsymbol{z}) = \frac{\prod_{j=1}^{n} \lambda_j^{z_j}}{Z} \mathbb{I}\left(\sum z_i = k\right) \tag{1.15}$$

The integration constant $Z$ is given by the $k$'th elementary symmetric polynomial (ESP)

$$Z = e_k(\boldsymbol{\lambda}) = \sum_{\boldsymbol{\alpha}} \prod_{j \in \boldsymbol{\alpha}} \lambda_j \tag{1.16}$$

where $\boldsymbol{\alpha}$ is a multi-index of size $k$. At this stage, it may be hard to see what sort of probability distribution eq. (1.15) defines. Indeed, it is not obvious how to sample from such a distribution, and the algorithm given in Kulesza and Taskar [11] is not trivial. We return to the issue in Section 3.3.1.

Inclusion probabilities can be computed through direct summation.

$$p(z_i = 1) = \sum_{\boldsymbol{z}_{-i}} p(z_i = 1, \boldsymbol{z}_{-i}) = \frac{\lambda_i \sum_{|\boldsymbol{\alpha}|=k-1, \boldsymbol{\alpha} \cap \{i\} = \varnothing} \prod_{j \in \alpha} \lambda_j}{e_k(\boldsymbol{\lambda})} \tag{1.17}$$

$$= \frac{\lambda_i e_{k-1}(\boldsymbol{\lambda}_{-i})}{e_k(\boldsymbol{\lambda})} \tag{1.18}$$

Computing such quantities in practice is again not completely trivial, although Kulesza and Taskar [11] gives an algorithm. We include a fairly accurate approximation below, and due to numerical instabilities in the exact algorithm, we advocate using the approximation in most cases (Section 4).

# 2. Asymptotic equivalence of $k$-DPPs and DPPs

Before stating our main results formally, we give an intuitive argument as to why $k$-DPPs and DPPs may resemble one another.

## 2.1. Some intuition

Readers familiar with statistical physics will know of a class of results known as "equivalence of ensembles" (Touchette [18]). These results justify formally a mathematical subterfuge, whereby a probability distribution that incorporates a hard constraint (the "micro-canonical ensemble") can be replaced with a more tractable variant (the "canonical ensemble"), where the hard constraint is turned into a soft constraint. Our result is a variant of this particular scenario.

We rewrite the likelihood of a $k$-DPP as the likelihood of a DPP times a hard constraint:

$$p(\mathcal{X}) \propto (\det \mathbf{L}_{\mathcal{X}}) \mathbb{I}\big(|\mathcal{X}| = k\big)$$

Deploy now the usual trick of turning the hard constraint into a soft constraint via an exponential, defining a new distribution:

$$q(\mathcal{X}) \propto (\det \mathbf{L}_{\mathcal{X}}) \exp\big(\nu |\mathcal{X}|\big) \tag{2.1}$$

where $\nu$ should be set so that $|\mathcal{X}| = k$ on average over $q$, i.e., $E_q(|\mathcal{X}|) = k$. Before we find such a value, it helps to recognise that $q$ actually has the form of a DPP: since $\det(\beta \mathbf{L}_{\mathcal{X}}) = \beta^{|\mathcal{X}]} \det \mathbf{L}_X$, we have

$$q(\mathcal{X}) \propto \det\big(\exp(\nu) \mathbf{L}_{\mathcal{X}}\big) \tag{2.2}$$

and we identify $q$ as a DPP with L-ensemble $\exp(\nu)\mathbf{L}_{\mathcal{X}}$. Using eq. (1.6), we find that:

$$E_q\big(|\mathcal{X}|\big) = \exp(\nu) \operatorname{Tr}\Big(\big(\exp(\nu)\mathbf{L} + \mathbf{I}\big)^{-1} \mathbf{L}\Big) \tag{2.3}$$

The appropriate value for $\nu$ is determined by the implicit equation that $E_q(|\mathcal{X}|) = k$. In terms of the eigenvalues, this reads:

$$\sum_i \frac{\lambda_i e^{\nu}}{1 + \lambda_i e^{\nu}} = k \tag{2.4}$$

To sum up, this development suggests that a $k$-DPP with ensemble $\mathbf{L}$ can be approximated by a (tilted) DPP with L-ensemble $\exp(\nu)\mathbf{L}$, with $\nu$ set so that the matched DPP has $k$ elements on average. The next section gives a rigorous statement for this approximation.

## 2.2. Main result

Under certain conditions, DPPs and $k$-DPPs are equivalent in a regime where we pick a fixed ratio of items from a growing set, that is, $\frac{k}{n} = r > 0$, fixed as $n \to \infty$. By equivalence, we mean that they have the same marginals (inclusion probabilities of order 1 and above). The conditions for equivalence boil down to the *number of degrees of freedom of* $\mathbf{L}$ *being high enough*, and we make that condition more precise below. In practice the approximations, we derive give excellent results in most settings we have tried, except with very small values of $n$ (less than 10, say).

We require assumptions on the L-ensembles: let $\mathbf{L}_1 \dots \mathbf{L}_n$ denote a sequence of positive definite matrices of increasing size $n \times n$. The assumption is that $\mathrm{Tr}((\mathbf{L}_n + \mathbf{I})^{-2}\mathbf{L}_n)$ diverges. The question of which sequences of matrices verify this condition is left to Section 2.3.

We associate with each $\mathbf{L}_n$ a $k$-DPP $\mathcal{X}_n$, where $k = \lfloor rn \rfloor$, a fixed fraction of the number of items. Similarly, we have a second sequence of *matched* DPPs $\tilde{\mathcal{X}}_n$ with L-ensemble $\exp(\nu_n)\mathbf{L}_n$, where $\nu_n$ verifies Eq. (2.4). Let $\boldsymbol{\alpha}$ denote a multi-index of fixed finite size $m < k$, and $\pi_n(\boldsymbol{\alpha})$ the probability that $\boldsymbol{\alpha} \subseteq \mathcal{X}_n$, and $\tilde{\pi}_n(\boldsymbol{\alpha})$ the corresponding probability for $\tilde{\mathcal{X}}_n$. We may interpret $\pi$ and $\tilde{\pi}$ as two measures over $\boldsymbol{\alpha}$, and an appropriate means of comparing these quantities is via total variation. Because $\pi$ and $\tilde{\pi}$ have total mass that grows with $k$ (see Lemma 1.2), we normalise the total variation distance with the appropriate factor.

**Definition 4.** Let $\pi$, $\tilde{\pi}$ designate two inclusion measures of order $m \geq 1$, corresponding to inclusion probabilities in point processes with $n$ elements. We define their total variation distance as:

$$D_m(\pi, \tilde{\pi}) = \binom{k}{m}^{-1} \sum_{\boldsymbol{\alpha}, |\boldsymbol{\alpha}|=m} \left|\pi(\boldsymbol{\alpha}) - \tilde{\pi}(\boldsymbol{\alpha})\right| \tag{2.5}$$

We have the following result:

**Theorem 2.1.** *Under the assumptions above, joint inclusion probabilities under a k-DPP and its matched DPP converge*:

$$D_m(\pi_n, \tilde{\pi}_n) = O\left(n^{-1}\right) \quad \textit{as } n \to \infty \tag{2.6}$$

**Remark 2.1.** Note that in our proof we have $k = O(n)$, which is needed because of a central limit argument implicit in the saddlepoint expansion.

**Remark 2.2.** A quantity of interest in many calculations are sample averages of the form $A(\mathcal{X}) = \frac{1}{m}\sum_{i \in \mathcal{X}} f_i$. Then $E_{\mathcal{X}}(A) = \frac{1}{m}\sum_{j \in \Omega} \pi(j) f_j$. An easy corollary is that $|E_{\mathcal{X}}(A) - E_{\tilde{\mathcal{X}}}(A)| \to 0$, from well-known properties of the total variation distance (DasGupta [5]).

The overall proof path for Theorem 2.1 is as follows:

1. We reduce the equivalence of $k$-DPPs and DPPs to the equivalence of *diagonal* $k$-DPPs and DPPs (Section 2.2.1)

2. Elementary symmetric polynomials (and ratios thereof) hold the key to the next step, and we show how they can be approximated using a saddlepoint approximation (Section 2.2.2)
3. We insert the asymptotic series for ESPs into the formula for inclusion probabilities, and derive the $O(1)$ and $O(n^{-1})$ terms. The $O(1)$ term corresponds to inclusion probabilities in the matched DPP, from which Theorem 2.1 follows (Section 2.2.3).

### 2.2.1. *Reduction to diagonal DPPs*

Recall (Section 1.5) that DPPs and $k$-DPPs are both mixture distributions, where we first draw a set of eigenvectors of $\mathbf{L}$, and then draw from a projection DPP formed from these eigenvectors. That second step is the same in DPPs and $k$-DPPs, only the first step differs. In DPPs, we draw from a diagonal DPP, while in $k$-DPPs we draw from a diagonal $k$-DPP. Heuristically, because it is only the first step that differs, we can focus on our asymptotic study on the first step.

Formally if we can establish that the inclusion probabilities in *diagonal* $k$-DPPs and DPPs converge (at any finite order), then the inclusion probabilities in *general* $k$-DPPs and DPPs converge as well (up to the same order). We note $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$ the diagonal DPPs associated with $\mathcal{X}$ and $\tilde{\mathcal{X}}$. The order-$m$ inclusion measures for $\mathcal{X}$ and $\tilde{\mathcal{X}}$ are noted $\pi_m$ and $\tilde{\pi}_m$, while the corresponding measures for $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$ are noted $\rho_m$ and $\tilde{\rho}_m$ (the latter correspond to the probability that certain eigenvectors are included, as per the mixture interpretation of DPPs introduced in Section 1.5.1).

The following lemma states the result.

**Lemma 2.1.** $D_m(\pi_m, \tilde{\pi}_m) \leq D_m(\rho_m, \tilde{\rho}_m)$

Lemma 2.1 implies that if diagonal $k$-DPPs converge to matched diagonal DPPs, so do general k-DPPs. The proof is deferred to the appendix (Section A.2). Armed with this lemma, we now focus only on the diagonal case.

Our goal is now to compute inclusion probabilities in diagonal $k$-DPPs. Recall that $\boldsymbol{\alpha}$ denotes a subset of $(1, \dots, n)$ of fixed size $m$. We wish to compute $p(\boldsymbol{\alpha} \in \mathcal{Y})$, or equivalently, the probability that $p(\prod_{j\in\boldsymbol{\alpha}} z_j = 1)$. This marginal probability can be computed via direct summation:

$$p\left(\prod_{j\in\boldsymbol{\alpha}} z_j = 1\right) = \frac{(\prod_{j\in\boldsymbol{\alpha}} \lambda_i) \sum_{\boldsymbol{\beta}, |\boldsymbol{\beta}|=k-|\boldsymbol{\alpha}|, \boldsymbol{\beta}\cap\boldsymbol{\alpha}=\varnothing} \prod_{j\in\boldsymbol{\beta}} \lambda_j}{e_k(\boldsymbol{\lambda})} = \left(\prod_{i\in\boldsymbol{\alpha}} \lambda_i\right) \frac{e_{k-m}(\boldsymbol{\lambda}_{-\boldsymbol{\alpha}})}{e_k(\boldsymbol{\lambda})} \tag{2.7}$$

Thus, inclusion probabilities in diagonal DPPs can be expressed using ratios of ESPs. This leads us to our next section, where we derive an asymptotic approximation for ESPs. We will then insert the asymptotic approximation into Eq. (2.7), to get an asymptotic series for inclusion probabilities.

### 2.2.2. *Saddlepoint approximation for ESPs*

ESPs are unwieldy combinatorial objects, but fortunately they lend themselves well to asymptotic approximation. This section is crucial for the rest and so we keep the details in the main text.

ESPs have an elegant probabilistic interpretation (already noted in passing in Chen, Dempster and Liu [2]). An equivalent definition for ESPs views them as the coefficients in a power series:

$$e_k(\boldsymbol{\lambda}) = [x^k]\prod_{i=1}^{n}(1+\lambda_i x) \tag{2.8}$$

We borrow the notation $[x^k]f(x)$ from combinatorics to denote the coefficient of $x^k$ in the series $f$. To uncover the probabilistic interpretation of ESPs, we transform the series into a *probability* generating function.

$$\begin{aligned} e_k(\boldsymbol{\lambda}) &= [x^k]\prod_{i=1}^{n}(1+\lambda_i)\frac{(1+\lambda_i x)}{1+\lambda_i} && (2.9)\\ &= \prod_{i=1}^{n}(1+\lambda_i)[x^k]\prod(1-p_i+xp_i) && (2.10)\end{aligned}$$

where $p_i = \frac{\lambda_i}{1+\lambda_i} \in (0,1)$ is now to be interpreted as the parameter of a Bernoulli variable, $B_i$. Let $S_n = \sum_{i=1}^{n} B_i$ designate the sum of all such independent $B_i$'s. Then:

$$p(S_n = k) = [x^k]\prod_{i=1}^{n}(1-p_i+xp_i) = \frac{e_k(\boldsymbol{\lambda})}{\prod_{i=1}^{n}(1+\lambda_i)}$$

Since $S_n$ is the sum of $n$ independent random variables, it invites a central limit approximation to the $p(S_n = k)$. First, note that:

$$\mu = E(S_n) = \sum \frac{\lambda_i}{1+\lambda_i} \tag{2.11}$$

which tells us that $e_k$, taken as a function of $k$, is likely to peak near $\mu$. The second moment,

$$\sigma^2 = \mathrm{Var}(S_n) = \sum \frac{\lambda_i}{(1+\lambda_i)^2} \tag{2.12}$$

gives a measure of scale for the peak of $e_k$ around $\mu$. Since $\frac{\lambda_i}{(1+\lambda_i)^2} \leq \frac{\lambda_i}{1+\lambda_i}$, we have:

$$\sigma^2 \leq \mu \tag{2.13}$$

In studying the convergence of $k$-DPPs and DPPs, it is $\sigma^2$, rather than $\mu$ that captures the appropriate notion of "degrees of freedom". In our case the Lyapunov central limit theorem (Billingsley [1]) requires that $\sigma^2$ diverge asymptotically, and the condition, we assumed on the sequence of L-ensembles guarantees exactly that (see Section 2.3 for a discussion).

A much better approximation than the Gaussian CLT is the saddlepoint approximation of Daniels [4]. Unlike the CLT, it is accurate in the tails and has $O(n^{-1})$ relative error. It reads:

$$p(S_n = k) = \frac{1}{\sqrt{2\pi\psi''(\nu^\star)}}\exp\bigl(\psi(\nu^\star) - k\nu^\star\bigr)\bigl(1+O(n^{-1})\bigr) \tag{2.14}$$

where $\psi(\nu) = \log E(\exp(\nu S_n))$ is the cumulant-generating function of $S_n$, and $\nu^\star$ is the solution of the saddlepoint equation:

$$\nu^\star = \operatorname*{argmin}_{\nu} \psi(\nu) - k\nu \tag{2.15}$$

In our case, we have:

$$\begin{aligned}\psi(\nu) &= \log E\big(\exp(\nu S_n)\big)\\ &= \sum_{i=1}^{n} \log E\big(\exp(\nu B_i)\big)\\ &= \sum \log\left(\frac{1}{1+\lambda_i} + \frac{\lambda_i}{1+\lambda_i} e^{\nu}\right)\\ &= \sum \log\big(1+\lambda_i e^{\nu}\big) - \sum \log(1+\lambda_i)\end{aligned} \tag{2.16}$$

We will need the derivatives of $\psi$ as well:

$$\psi'(\nu) = \sum \frac{\lambda_i e^{\nu}}{1+\lambda_i e^{\nu}} \tag{2.17}$$

$$\psi''(\nu) = \sum \frac{\lambda_i e^{\nu}}{(1+\lambda_i e^{\nu})^2} \tag{2.18}$$

Inserting (2.16) into (2.15), we see that:

$$\sum \frac{\lambda_i e^{\nu^\star}}{1+\lambda_i e^{\nu^\star}} = k$$

recovering (2.4).

To summarise: inserting (2.9) into (2.15), we have

**Lemma 2.2.**

$$e_k(\boldsymbol{\lambda}) = \frac{1}{\sqrt{2\pi\psi''(\nu^\star)}} \exp\left(\sum_{i=1}^{n}\big(\log\big(1+\lambda_i e^{\nu^\star}\big)\big) - k\nu^\star\right)\big(1+O\big(n^{-1}\big)\big) \tag{2.19}$$

**Remark 2.3.** In large $n$ the exponential term dominates (a large deviation regime, see Touchette [18]), and we have:

$$\log e_k(\boldsymbol{\lambda}) \approx \sum_{i=1}^{n} \log\big(1+\lambda_i e^{\nu^\star}\big) - k\nu^\star \tag{2.20}$$

In random matrix theory it is customary to define the *Shannon transform* of a matrix $\mathbf{L}$ as: $T(s) = \log\det(\mathbf{I} + s\mathbf{L})$ (Couillet and Debbah [3]). Eq. (2.20) says that for large matrices, the ESPs of $\mathbf{L}$ are directly related to the Legendre transform of $T(e^{\nu})$.

At this stage, we have a tractable approximation to ESPs, and we are now ready to use it to find an approximation for inclusion probabilities.

### 2.2.3. *Inclusion probabilities, and ratios of ESPs*

To study the asymptotics of inclusion probabilities, we insert approximation (2.14) into eq. (2.7), and compute the $O(1)$ and $O(n^{-1})$ terms. The calculation is lengthy and can be found in the appendix (Section A.3). The end result is as follows:

**Lemma 2.3.** *In a diagonal k-DPP $\mathcal{Y}$ with L-ensemble* $\operatorname{diag}(\lambda_1, \dots, \lambda_n)$, *inclusion probabilities have the asymptotic form*:

$$p_k(\boldsymbol{\alpha} \in \mathcal{Y}) = \left(\prod_{i \in \boldsymbol{\alpha}} \frac{\lambda_i \exp(\nu^\star)}{1 + \lambda_i \exp(\nu^\star)}\right)\left(1 + \frac{1}{n} g\left(\nu^\star\right) + O\left(\frac{1}{n^2}\right)\right) \tag{2.21}$$

*with*

$$g\left(\nu^\star\right) = -\frac{\nu_1^2}{2}\bar{\psi}''\left(\nu^\star\right) - \frac{1}{2\bar{\psi}''(\nu^\star)}\left(\bar{\psi}^{(3)}\left(\nu^\star\right)\nu_1 - m\bar{\psi}''_{\boldsymbol{\alpha}}\left(\nu^\star\right)\right)$$

*The terms appearing in the correction $g(\nu^\star)$ are defined in Appendix A.3.*

Notice that the $O(1)$ term corresponds exactly to the inclusion probability in the matched diagonal DPP, $\tilde{\mathcal{Y}}$. We now have all the elements we need to prove Theorem 2.1. Consider a $k$-DPP with $m$-th order inclusion probability $\pi_m$. Let $\tilde{\pi}_m$ be the $m$-th order inclusion probability of the matched DPP. Let the corresponding measure for the generating diagonal $k$-DPP be $\rho_m(\boldsymbol{\alpha}) = p_k(\boldsymbol{\alpha} \in \mathcal{Y})$, whose approximation $\rho_m = \tilde{\rho}_m(1 + O(1/n))$ is given by eq. (2.21). Starting with Lemma 2.1 and using the approximation leads to

$$\begin{aligned}
D_m(\pi_m, \tilde{\pi}_m) \; &\le \; D_m(\rho_m, \tilde{\rho}_m) \\
&= \binom{k}{m}^{-1} \sum_{\boldsymbol{\alpha}, |\boldsymbol{\alpha}| = m} \left(\prod_{i \in \boldsymbol{\alpha}} \frac{\lambda_i \exp(\nu^\star)}{1 + \lambda_i \exp(\nu^\star)}\right)\left(\frac{1}{n} g\left(\nu^\star\right) + O\left(\frac{1}{n^2}\right)\right) \\
&\overset{(a)}{=} k^m \binom{k}{m}^{-1} \left(\frac{1}{n} g\left(\nu^\star\right) + O\left(\frac{1}{n^2}\right)\right) \\
&\overset{(b)}{=} O\left(\frac{1}{n}\right)
\end{aligned}$$

where equality $(a)$ is due to equation (2.4) which implicitly defines $\nu^\star$, and equality $(b)$ holds because $\binom{k}{m} = O(k^m)$. This concludes the proof of the main result. A refinement is described in Appendix A.4, where we derive a tractable correction to multivariate inclusion probabilities.

A remark on the precise nature of the convergence result is in order. Regardless of how large $n$ is, a $k$-DPP will continue to produce sets of fixed size, while a DPP will continue to produce sets of variable size. This implies that DPPs and $k$-DPPs cannot be equivalent in the very strong sense of the respective probability mass functions agreeing on every possible set, since by definition they remain different. The result is of the same nature as *equivalence of ensembles* in statistical physics: it pertains to two different distributions that agree more and more as $n$ tends to infinity, but never agree completely. Practically speaking, an interpretation is that for a given $n$, a $k$-DPP and a matched DPP will have very similar moments up to a certain order: certainly, at order $m > k$, this cannot be true, since the inclusion measure for the $k$-DPP is uniformly zero, but that is not true for the DPP. To get agreement up to higher orders, one has to increase $n$.

Besides the main result, another consequence of Lemma 2.3 is that in importance sampling estimators of the form given by eq. (0.1) can be used with approximate rather than exact probabilities. Using the $O(n^{-1})$ approximation induces order $O(n^{-1})$ bias, and similarly using the $O(n^{-2})$ correction induces order $O(n^{-2})$ bias. Our recommendation is therefore that one samples $k$-DPPs, rather than DPPs, while using the approximate inclusion probabilities in computations.

## 2.3. To which sequences of matrices does this apply?

We stated earlier that the result applies to any sequence of matrices whose degrees of freedom grow as a function of $n$, with the more precise statement being that $\mathrm{Tr}((\mathbf{L}_n + \mathbf{I})^{-2}\mathbf{L}_n)$ (see eq. (2.12)) should diverge. With the caveat that the condition is sufficient and not necessary, in what sort of scenarios can we expect it to hold?

A full discussion of the issue would require significant forays into random matrix theory and take us beyond the scope of the current work, so we only give a sufficient condition that is relatively easily checked. As mentioned in Section 1.5, in $k$-DPPs, the L-ensemble can be multiplied by an arbitrary positive constant without changing the distribution. This means that we are free to scale each $\mathbf{L}_n$ by an arbitrary constant independently for each $n$, a normalisation that lets us for instance set $\lambda_{\max}$ to 1 for all $n$. For $x \leq 1$, $\frac{x}{(1+x)^2} \geq \frac{1}{4}x$, which implies that $\sigma^2(n) \geq \frac{1}{4}\mathrm{Tr}(\mathbf{L}_n)$, and a sufficient condition for the theorem to apply is therefore that $\mathrm{Tr}(\mathbf{L}_n)$ diverges.

To pick a practical scenario, consider "in-fill" asymptotics. We suppose that the original set of data is made up of $n$ vectors in $\mathbb{R}^d$ sampled i.i.d. from a density $\rho(\mathbf{x})$. The L-ensemble used is the classical squared-exponential (Gaussian) kernel. Let $\mathbf{L}_n = \frac{\mathbf{M}_n}{\lambda_{\max}(\mathbf{M}_n)}$, where $M_{ij} = \exp(-\frac{1}{2\tau^2}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$. $\mathrm{Tr}\,\mathbf{M} = n$, and from the Gershgorin circle theorem we have a bound on $\lambda_{\max}$ that reads $\lambda_{\max} \leq \max_i \sum_j M_{ij}$. A sufficient condition for convergence is then that $\frac{n}{\max_i \sum_j M_{ij}}$ diverges, which will not be the case for fixed $\tau$. The reason is that $\sum_j M_{ij} = \sum \exp(-\frac{1}{2\tau^2}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ essentially counts the number of points in a neighbourhood of size $\tau$ around $\mathbf{x}_i$, and that quantity is $O(n)$. To make $\mathrm{Tr}(\mathbf{L}_n)$ diverge, we need to shrink $\tau$ with $n$ so that each point has $O(1)$ neighbours. Similarly, the condition holds under so-called "increasing-domain" asymptotics, in which $\tau$ is fixed but we consider points in a growing window. It is likely that one could relax these criteria, but in any case we must emphasise that (a) the approximations work really well in practice, see Section 4 and (b) actual simulations of $k$-DPPs require L-ensembles that have effective rank quite a bit larger than $k$, otherwise the numerical difficulties are overwhelming even though the process may be well defined.

## 2.4. Consequences for inference

DPPs are not only used for sampling, but also as statistical models for certain types of data that exhibit repulsion. Now in this case as well the modeler has to make a choice, and use either $k$-DPPs or DPPs. The former seems to imply that the number of observations (which is the role played here by $k$) is known in advance, while the latter does not. Interestingly, the results above imply that the choice of fixed size or varying size is of no consequence, at least if maximum likelihood is used for inference, though we suspect that Bayesian inference would be the same in that regard. To be precise, what we have in mind here is a case in which we observe a set $\mathcal{X}$ of size $k$, assumed to have been drawn from a $k$-DPP of matrix $\mathbf{L}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters. For instance, $\boldsymbol{\theta}$ may control the amount of repulsion in the point process. The log-likelihood of a $k$-DPP is given by:

$$\mathcal{C}_{k\mathrm{DPP}}(\boldsymbol{\theta}) = \log\det\big(\mathbf{L}(\boldsymbol{\theta})_{\mathcal{X}}\big) - \log e_k(\boldsymbol{\lambda}_{\boldsymbol{\theta}}) \tag{2.22}$$

The corresponding Maximum Likelihood estimator of $\boldsymbol{\theta}$ is noted:

$$\hat{\boldsymbol{\theta}}_{k\mathrm{DPP}} = \operatorname{argmin} \mathcal{C}_{k\mathrm{DPP}}(\boldsymbol{\theta}) \tag{2.23}$$

Similarly, a DPP model would assume $\mathcal{X}$ to be drawn from a DPP with L-ensemble $e^{\nu}\mathbf{L}(\theta)$, where $e^{\nu}$ controls the expected cardinality of the set. The log-likelihood reads in this case:

$$\mathcal{C}_{\mathrm{DPP}}(\boldsymbol{\theta},\nu) = \nu k + \log\det\big(\mathbf{L}(\boldsymbol{\theta})_{\mathcal{X}}\big) - \log\det\big(\mathbf{I} + e^{\nu}\mathbf{L}(\boldsymbol{\theta})\big) \tag{2.24}$$

Since $\nu$ is effectively a nuisance parameter, we may use a profile likelihood:

$$\mathcal{C}^{\star}_{\mathrm{DPP}}(\boldsymbol{\theta}) = \max_{\nu} \mathcal{C}_{\mathrm{DPP}}(\boldsymbol{\theta},\nu) \tag{2.25}$$

The ML estimator of $\boldsymbol{\theta}$ in this case solves:

$$\hat{\boldsymbol{\theta}}_{\mathrm{DPP}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{C}^{\star}_{\mathrm{DPP}}(\boldsymbol{\theta})$$

To find a closed-form for the profile likelihood (eq. (2.25)), we take the derivative of $\mathcal{C}(\boldsymbol{\theta},\nu)$ with respect to $\nu$, to find:

$$\frac{\partial}{\partial\nu}\mathcal{C}(\boldsymbol{\theta},\nu) = k - \mathrm{Tr}\Big(\big(\mathbf{I} + e^{\nu}\mathbf{L}\big)^{-1} e^{\nu}\mathbf{L}\Big)$$

where we recognise the saddlepoint equation in yet another form.

Equating the above to 0, we obtain:

$$\mathcal{C}^{\star}_{\mathrm{DPP}}(\boldsymbol{\theta}) = \log\det\big(\mathbf{L}(\boldsymbol{\theta})_{\mathcal{X}}\big) + \nu^{\star}(\boldsymbol{\theta})k - \log\det\big(\mathbf{I} + e^{\nu^{\star}}\mathbf{L}(\boldsymbol{\theta})\big) \tag{2.26}$$

From eq. (2.19) we know that:

$$\log e_k(\boldsymbol{\lambda}) = \sum_{i=1}^{n}\Big(\log\big(1+\lambda_i e^{\nu^{\star}}\big)\Big) - k\nu^{\star} - \frac{1}{2}\left(\log\left(\sum_{i=1}^{n}\frac{\lambda_i}{1+\lambda_i e^{\nu^{\star}}}\right) + \log(2\pi)\right) + O\big(n^{-1}\big) \tag{2.27}$$

which tells us that:

$$\mathcal{C}^{\star}_{\text{DPP}}(\boldsymbol{\theta}) = \mathcal{C}_{k\text{DPP}}(\boldsymbol{\theta}) + O\big(\log(n) + n^{-1}\big) + \textit{constant} \tag{2.28}$$

where the term in $O(\log(n))$ comes from the second derivative of $\psi$ in (2.19) and is expected to be small compared to $\mathcal{C}$. A full formal argument showing convergence of $\hat{\boldsymbol{\theta}}_{\text{DPP}}$ to $\hat{\boldsymbol{\theta}}_{k\text{-DPP}}$ is complicated, and amounts to showing that the $O(\log(n) + n^{-1})$ term is constant in a relevant region around $\hat{\boldsymbol{\theta}}_{k\text{-DPP}}$. Informally, however, what happens is quite clear: the two cost functions are close (up to a vertical shift), and if they are sufficiently well-behaved (as a function of $\boldsymbol{\theta}$), then we expect $\hat{\boldsymbol{\theta}}_{\text{DPP}} \approx \hat{\boldsymbol{\theta}}_{k\text{-DPP}}$. We verify this conjecture in a numerical example in Section 4.3.

# 3. Algorithms and numerical results

The results above are interesting theoretically, but can also be used in practice to develop algorithms that compute (approximate) ESPs, sample diagonal $k$-DPPs, and compute inclusion probabilities. We find empirically that although approximate, they are much better behaved numerically than their nominally exact counterpart.

## 3.1. Computing ESPs

The algorithm given in Kulesza and Taskar [11] (alg. 7, p. 60) for computing ESPs of all orders is fast[3] but prone to numerical problems when $n$ is large, which is not completely surprising given that ESPs can vary over dozens of orders of magnitude. We find that the saddlepoint approximation given in eq. (2.19) is more practical, especially since it is naturally computed on a logarithmic scale, a perk exact algorithms do not share. To compute eq. (2.19), one needs to solve the saddlepoint equation (eq. (2.4)) for $\nu$. Newton's algorithm can be used (Algorithm 1), but it needs appropriate initialisation or it may not converge (when it does converge, it does so very fast). In our implementation, an initial guess for $\nu$ is found by linearising the saddlepoint equation for small $\nu$ (small $k$), and large $\nu$ (large $k$). In small $\nu$, we have:

$$\sum_i \frac{\lambda_i e^{\nu}}{1+\lambda_i e^{\nu}} \approx e^{\nu} \sum \lambda_i$$

so that for $k$ small, we may approximate $\nu$ as:

$$\nu \approx \log k - \log \sum \lambda_i \tag{3.1}$$

In large $k$ we find:

$$\sum_i \frac{\lambda_i e^{\nu}}{1+\lambda_i e^{\nu}} \approx n - e^{-\nu} \sum \frac{1}{\lambda_i}$$

[3] Their algorithm runs in $O(n^2)$, like ours, although theirs is faster in practice.

**Algorithm 1** Solving the saddlepoint equation

```
Input: eigenvalues $\boldsymbol{\lambda}$, set size $k$, initial guess $\nu_0$, tolerance $\epsilon$
  procedure SOLVE($\boldsymbol{\lambda}$,$k$,$\nu_0$)
      $\nu \leftarrow \nu_0$
      while $|\psi'(\nu) - k| < \epsilon$ do
          $\nu \leftarrow \nu - \frac{(\psi'(\nu)-k)}{\psi''(\nu)}$ (eq. (2.17) and (2.18)).
      end while
  end procedure
Return $\nu$
```

which solving for $\nu$ results in:

$$\nu \approx \log(n-k) - \log \sum \lambda_i^{-1} \tag{3.2}$$

We use the first guess for $k \leq \frac{n}{2}$ and the second otherwise, with good results. Interestingly, (3.1) and (3.2) can be used to find the worst-case relative error of the saddlepoint approximation, which is about 10%, a figure we verify in practice for all but the smallest $n$. The saddlepoint approximation is at its worst far out in the tails, that is, for $k = 1$ and $k = n - 1$. Recall that $e_1(\boldsymbol{\lambda}) = \sum \lambda_i$. We inject (3.1) into (2.19) and linearise to find:

$$\frac{1}{\sqrt{2\pi\psi''(\nu^\star)}} \exp\left(\sum_{i=1}^{n}\left(\log\left(1+\lambda_i e^{\nu^\star}\right)\right) - k\nu^\star\right) \approx \frac{1}{\sqrt{2\pi}} \exp(1) \sum \lambda_i \tag{3.3}$$

so that the relative error is about $\frac{\exp(1)}{\sqrt{2\pi}} \approx 1.08$. A similar calculation for $k = n - 1$ yields the same figure.

To compute all ESPs, it is useful to begin at $k = 1$ and then “warm-start” the optimisation rather than always use the same initial condition. The procedure is outlined in Algorithm 2.

## 3.2. Computing inclusion probabilities

### 3.2.1. *In diagonal k-DPPs*

Computing inclusion probabilities boils down to an application of the asymptotic formula developed in Section A.3. Depending on the accuracy required, the $O(\epsilon)$ in (2.21) may or may not be

**Algorithm 2** Computing all (log-) ESPs using a saddlepoint approximation

```
Input: eigenvalues $\boldsymbol{\lambda}$ (vector of length $n$), Set initial guess to $\nu = -\log \sum \lambda_i$.
  for $k \leftarrow 1 \dots (n-1)$ do
      $\nu \leftarrow \text{solve}(\boldsymbol{\lambda}, k, \nu)$
      $\log e_k \leftarrow -\frac{1}{2}\log(2\pi\psi''(\nu)) + \sum_{i=1}^{n} \log(1+\lambda_i e^{\nu}) - k\nu$
  end for
Return $\log e_1 \dots \log e_{n-1}$
```

**Algorithm 3** First order inclusion probabilities in diagonal $k$-DPPs: basic estimate

```
Input: eigenvalues $\boldsymbol{\lambda}$, set size $k$.
  procedure DIAG-BASIC($\boldsymbol{\lambda}$,$k$)
    $\nu \leftarrow \text{solve}(\boldsymbol{\lambda}, k, \nu_0)$ (Solve for saddle point)
    for $i \in 1 \dots n$ do
      $\tilde{\pi}_i \leftarrow \frac{e^{\nu}\lambda_i}{1+e^{\nu}\lambda_i}$
    end for
  end procedure
Return $\tilde{\pi}_1 \dots \tilde{\pi}_n$
```

needed. Computing the terms in (2.21) requires solving the saddlepoint equation, and computing $\psi(\nu^{\star})$ and up to three derivatives, which is $O(n)$ work in total. $\psi'$ and $\psi''$ are by-products of the Newton iteration, and only $\psi^{(3)}$ must be computed from scratch. All first order inclusion probabilities can be computed jointly based on these quantities, in $O(n)$ time. Algorithm 3 computes the uncorrected estimate, and Algorithm 4 the corrected estimate (the latter may be hard to understand without a thorough look at Section A.3).

### 3.2.2. *In general k-DPPs*

Computing (approximate) inclusion probabilities in general $k$-DPPs is an application of eq. (A.14): first compute the inclusion probabilities for the eigenfunctions, then apply (A.14). For first-order inclusion probabilities, and under the $O(\frac{1}{n})$ approximation, this boils down to com-

**Algorithm 4** First order inclusion probabilities in diagonal $k$-DPPs: corrected estimate

```
Input: eigenvalues $\boldsymbol{\lambda}$, set size $k$.
  procedure DIAG-CORRECTED($\boldsymbol{\lambda}$,$k$)
    $\nu \leftarrow \text{solve}(\boldsymbol{\lambda}, k, \nu_0)$ (Solve for saddle point)
    $\bar{\psi}'' \leftarrow \frac{1}{n}\sum \frac{\lambda_i e^{\nu}}{(1+\lambda_i e^{\nu})^2}$
    $\bar{\psi}^{(3)} \leftarrow \frac{1}{n}\sum \frac{\lambda_i e^{\nu}(1-\lambda_i e^{\nu})}{(1+\lambda_i e^{\nu})^3}$
    for $i \in 1 \dots n$ do
      $\bar{\psi}'_i \leftarrow \frac{e^{nu}\lambda_i}{1+e^{nu}\lambda_i}$
      $\bar{\psi}''_i \leftarrow \frac{\lambda_i e^{\nu}}{(1+\lambda_i e^{\nu})^2}$
      $\nu_1 \leftarrow \frac{1-\bar{\psi}'_i}{\bar{\psi}''}$
      $g \leftarrow -\frac{\nu_1^2}{2}\bar{\psi}'' - \frac{1}{2\bar{\psi}''}(\bar{\psi}^{(3)}\nu_1 - m\bar{\psi}''_{\boldsymbol{\alpha}})$
      $\tilde{\pi}_i \leftarrow \frac{e^{\nu}\lambda_i}{1+e^{\nu}\lambda_i}(1 + \frac{g}{m})$
    end for
  end procedure
Return $\tilde{\pi}_1 \dots \tilde{\pi}_n$
```

**Algorithm 5** First order inclusion probabilities in general $k$-DPPs

```
Input: L-ensemble $\mathbf{L}$, set size $k$
  $\mathbf{U} \leftarrow$ eigenvectors($\mathbf{L}$), $\boldsymbol{\lambda} \leftarrow$ eigenvalues($\mathbf{L}$)
  $\tilde{\boldsymbol{\pi}} \leftarrow$ diag-simple($\boldsymbol{\lambda}, k$) or $\tilde{\boldsymbol{\pi}} \leftarrow$ diag-corrected($\boldsymbol{\lambda}, k$)
  for $i \in 1 \dots n$ do
      $\tilde{p}_i \leftarrow \sum_{j=1}^{n} U_{ij}^2 \pi_j$
  end for
Return $\tilde{p}_1 \dots \tilde{p}_1$
```

puting

$$\operatorname{diag}\left(e^{\nu^\star}\left(e^{\nu^\star}\mathbf{L}+\mathbf{I}\right)^{-1}\mathbf{L}\right) \tag{3.4}$$

where $\nu^\star$ solves the saddlepoint equation for the appropriate value of $k$. $\nu^\star$ depends on the eigenvalues of $\mathbf{L}$, so the most straightforward way of computing (3.4) is via an eigendecomposition of $\mathbf{L}$, after which one obtains (3.4) from:

$$\left(e^{\nu^\star}\left(e^{\nu^\star}\mathbf{L}+\mathbf{I}\right)^{-1}\mathbf{L}\right)_{ii} = \sum_{j=1}^{n} \frac{e^{\nu^\star}\lambda_j}{1+e^{\nu^\star}\lambda_j} U_{ij}^2 \tag{3.5}$$

where $U_{ij}$ is the $j$-th eigenvector of $\mathbf{L}$ evaluated at index $i$. In most realistic problems the dominant cost by far is the $O(n^3)$ eigendecomposition. If $\mathbf{L}$ is sparse, or if matrix-vector products $\mathbf{Lv}$ can be computed using fast algorithms, the cost can be significantly reduced using a variety of techniques. For example, under sparse $\mathbf{L}$ the Cholesky decomposition may still be relatively cheap, and the Takahashi equations (Rue and Held [15]) can be used to obtain diagonal elements in (3.4) and solve the saddlepoint equation. Algorithm 5 computes first-order probabilities, and Algorithm 6 higher order inclusion probabilities. The optional correction used in Algorithm 6 is explained in Section A.4.

**Algorithm 6** High order inclusion probabilities in general $k$-DPPs

```
Input: L-ensemble $\mathbf{L}$, subset $\boldsymbol{\alpha}$, set size $k$
  $\boldsymbol{\lambda} \leftarrow$ eigenvalues($\mathbf{L}$).
  $\nu \leftarrow$ solve($\boldsymbol{\lambda}, k, \nu_0$)
  $\tilde{p}_{\boldsymbol{\alpha}} \leftarrow \det((\mathbf{I}+e^{\nu}\mathbf{L})^{-1}e^{\nu}\mathbf{L})_{\boldsymbol{\alpha}}$
  (Optional: compute correction
  $m \leftarrow |\boldsymbol{\alpha}|$ (size of subset)
  $\tilde{\boldsymbol{\pi}} \leftarrow$ diag-simple($\boldsymbol{\lambda}, k$)
  $v \leftarrow e_m(\tilde{\boldsymbol{\pi}})$
  $\tilde{p}_{\boldsymbol{\alpha}} \leftarrow \tilde{p} \times \binom{k}{m} v^{-1}$
Return $\tilde{p}_{\boldsymbol{\alpha}}$
```

**Algorithm 7** Sampling from a diagonal $k$-DPP

```
Input: eigenvalues $\boldsymbol{\lambda}$ (vector of length $n$), integer $k$ (set size). Init $s = 0, t = 1$
  while $t \leq n, s < k$ do
      Compute $\pi_t$, inclusion probability in a diagonal $(k - s)$-DPP with eigenvalues $\lambda_t \dots \lambda_n$, using eq. (2.21).
      Set $z_t$ to 1 with probability $\pi_t$
      $s \leftarrow \sum_{i=1}^{t} z_i$
      $t \leftarrow t + 1$
  end while
Return $\mathbf{z}$, the inclusion vector.
```

## 3.3. Sampling

There already is a large literature on sampling from ($k$-)DPPs (see, e.g., Li, Jegelka and Sra [12], Gautier, Bardenet and Valko [8] and references therein). Our goal here is only to show how our methods can be used to modify the algorithms given in Kulesza and Taskar [11] to improve numerical stability. To sample a $k$-DPP, we follow the two-step strategy of Kulesza and Taskar [11], which derives from the mixture interpretation explained in Section 1.5. We first sample a set of eigenvectors, picking $k$ of them using a diagonal $k$-DPP, then sample from the projection DPP formed from the eigenvectors we selected.

### 3.3.1. *Sampling from a diagonal k-DPP*

The first part requires sampling from a diagonal $k$-DPP. For that task, Kulesza and Taskar [11] give an algorithm that they justify using a recursive argument, but a more intuitive explanation can be found. Thinking of the $k$-DPP as sampling a binary string $\mathbf{z} = z_1 \dots z_n$, we run through the elements one by one, sampling according to $p(z_t|z_1 \dots z_{t-1})$. It is straightforward to show that in a diagonal $k$-DPP, $z_t|z_1 \dots z_{t-1}$ has a sufficient statistic: $p(z_t|z_1 \dots z_{t-1}) = p(z_t|\sum_{i=1}^{t-1} z_i)$. This occurs because $z_t \dots z_n|z_1 \dots z_{t-1}$ is a diagonal $(k - s)$-DPP, where $s = \sum_{i=1}^{t-1} z_i$. Thus, $p(z_t = 1|z_1 \dots z_{t-1})$ is the inclusion probability for item $t$ in a diagonal $(k-s)$-DPP, and we can use our approximations to compute that probability.

We state the basic algorithm in Algorithm 7, but many refinements can be made for speed. In particular, computing the approximation requires solving the saddlepoint equation, and warm-starting should be used. Beyond that, given that the cost of sampling from a $k$-DPP is mostly dominated by the eigenvalue decomposition and by the step where a projection DPP is sampled, it is not worth spending too much time optimising the diagonal step.

## 3.4. Sampling from a projection DPP

Once we have obtained $k$ eigenvectors, we can form the projection kernel $\mathbf{U}_{:,\mathcal{Y}}\mathbf{U}_{\mathcal{Y},:}^{\top}$ and use any algorithm that samples a projection DPP. There are several options in the literature, but one that is both fast and particularly easy to implement is described in Tremblay, Barthelme and Amblard [20], and that is the one we use here in our simulations.

# 4. Empirical results

We report here the accuracy of our approximations in some tests and simulations. We examine briefly the quality of the approximation for ESPs, then inclusion probabilities in diagonal $k$-DPPs, and finally inclusion probabilities in general $k$-DPPs.

## 4.1. Approximation of elementary symmetric polynomials

The approximation to ESPs given in eq. (2.19) is nothing more than a saddlepoint approximation for sums of Bernoulli variables, so it would be surprising if it did not work as advertised. Nonetheless it is interesting to see how good the approximation is, and that the figure we give in Section 3.1 for a maximum error of 10% (for $k = 1$ and $k = n - 1$) is verified in practice. It also serves to illustrate the better numerical behaviour of the approximation compared to the (nominally exact) summation algorithm.

Figure 1 and 2 show results obtained on two deterministic sequences, $\lambda_{i,n} = i$ and $\lambda_{i,n} = e^{-i}$ (for three different values of $n$). The approximation is excellent even with the second sequence, which does not verify the sufficient condition for convergence ($\sigma^2 = O(n)$). The summation algorithm overflows in the first case and underflows in the second, while the approximation shows good behaviour.

To go beyond deterministic sequences, we consider a set of $n$ points in $\mathbb{R}^2$, drawn from a unit Gaussian distribution. The $\mathbf{L}$-matrix is from a squared-exponential kernel, $L_{ij} = \exp(\frac{-\|x_i - x_j\|^2}{2\tau^2})$. Here we set $\tau = 1$. Figure 3 shows the ratio of approximation to true value $\tilde{e}_k(\boldsymbol{\lambda})/e_k(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ are the eigenvalues of $\mathbf{L}$.

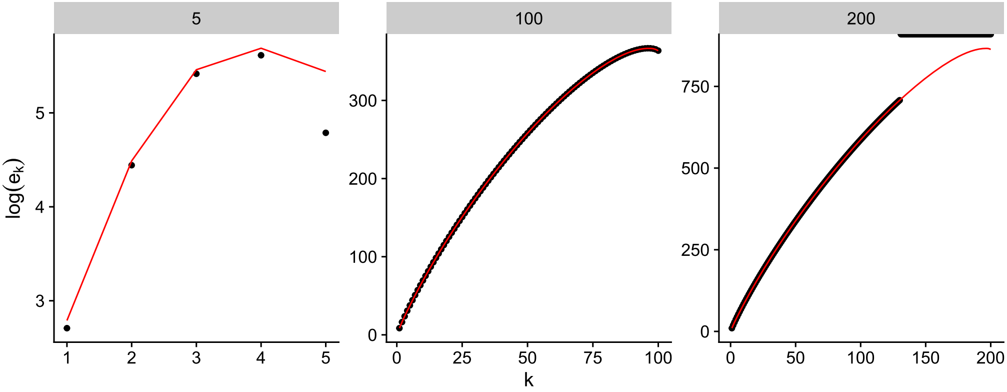

**Figure 1.** Approximation of (log) ESPs for the sequence $\lambda_i = i$, $i = 1, \dots, n$. Black dots: numerical results obtained using the "exact" summation algorithm. Red line: approximation using eq. (2.19). We show results for $n = 5$, 100 and 200. Numerical problems are already apparent at $n = 200$, with the summation algorithm overflowing at $k \approx 130$. The approximation has no such issues in this case.

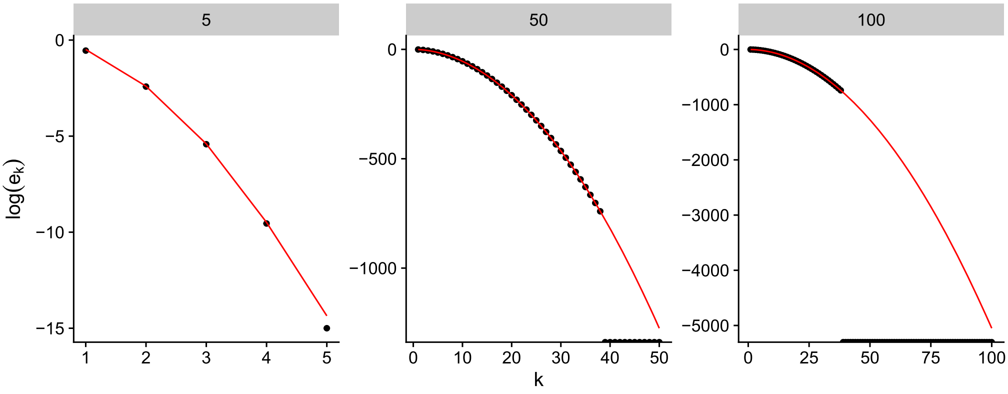

**Figure 2.** Approximation of (log) ESPs for the sequence $\lambda_i = e^{-i}$, $i = 1, \ldots, n$. Same format as Figure 1.

## 4.2. Approximation of inclusion probabilities

We begin with approximations to inclusion probabilities in diagonal $k$-DPPs. Figure 4 shows results for a diagonal $k$-DPP with diagonal values $\lambda_{i,n} = \exp(-\frac{i}{10})$, comparing true inclusion probabilities to the $O(n^{-1})$ and $O(n^{-2})$ approximations given by Lemma 2.3. The approximations are overall excellent, with even the rougher $O(n^{-1})$ approximation becoming practically exact for $n \geq 100$. Note that the conditions for convergence assumed in our theorem do not hold for the sequence in question.

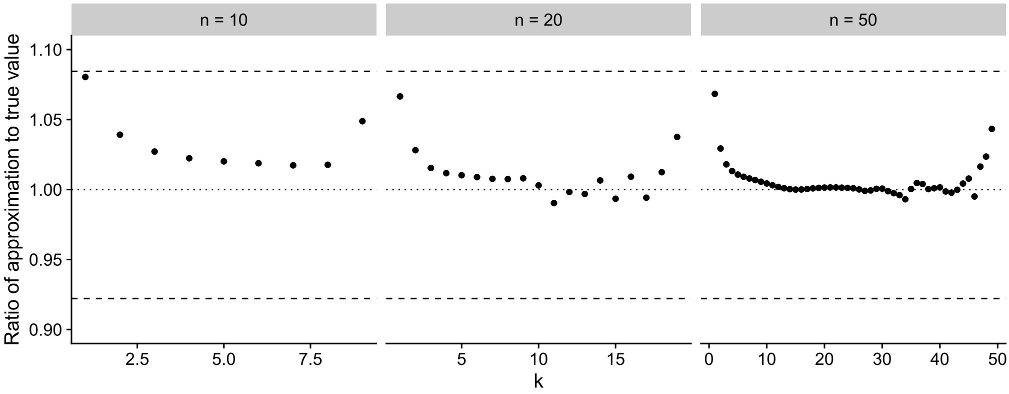

**Figure 3.** Ratio of approximated ESP to true ESP, for an example with $n$ points at random locations and a squared-exponential kernel (see text). From the arguments in Section 3.1, we expect a maximum relative error of about 1.08, shown here as upper and lower dashed lines. The central dashed lines corresponds to no error.

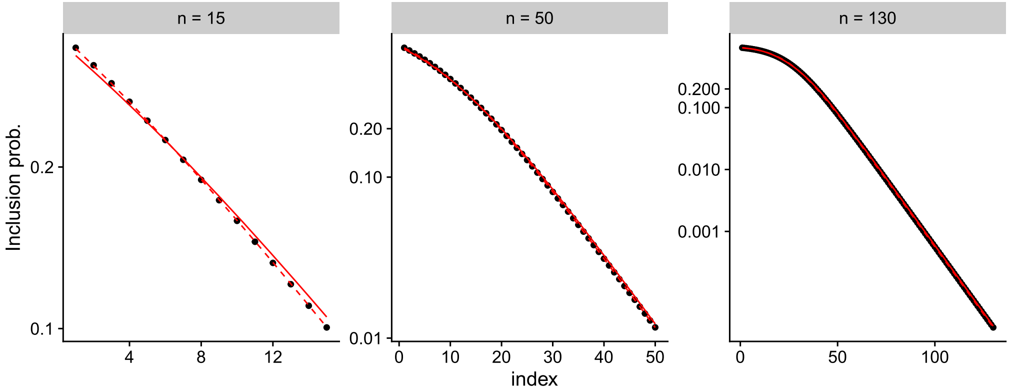

**Figure 4.** Inclusion probabilities for a diagonal $k$-DPP with diagonal entries $\exp(-\frac{i}{10})$, with $k = n/5$. Black: true values. Red, continuous: $O(n^{-1})$ approximation. Red, dashed, $O(n^{-2})$ approximation (see eq. (2.21)).

The $O(n^{-1})$ and $O(n^{-2})$ rates are asymptotic, and it is interesting to verify that they hold in practice. Figure 5 does this, in a scenario where the conditions of the theorem hold. For each $n$, the diagonal values $\lambda_{1,n}$ to $\lambda_{n,n}$ are drawn i.i.d. from the uniform distribution on the interval $(1, 10)$. We estimate convergence rates via a regression of log error on $\log n$.

Unsurprisingly given the above, approximating inclusion probabilities in general $k$-DPPs works well too. Figure 6 provides an illustration, using again $n$ points drawn i.i.d. from a Gaus-

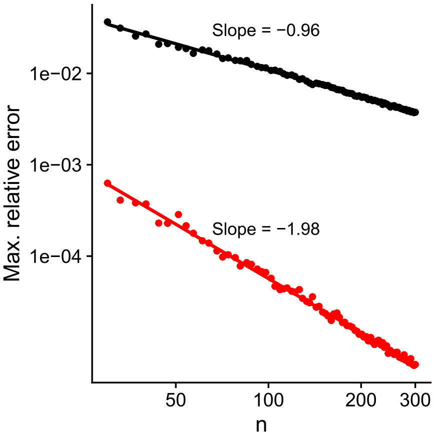

**Figure 5.** Convergence of approximations to diagonal probabilities. We verify the $O(n^{-1})$ and $O(n^{-2})$ rates empirically. See text for details.

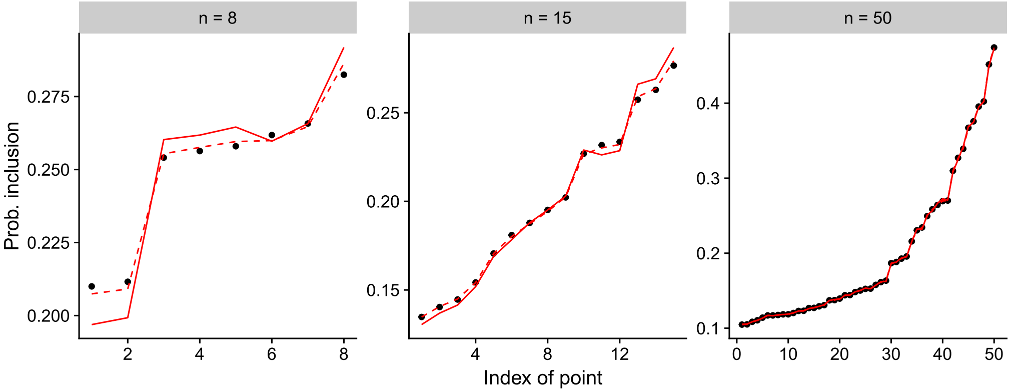

**Figure 6.** Approximation to inclusion probabilities in a full $k$-DPP. The scenario is the same as in Figure 3, namely $n$ points drawn i.i.d. and a squared exponential kernel. Each point correspond to the inclusion probabilities of point $x_i$ in a $k$-DPP. Points have been sorted according to increasing probability of inclusion.

sian, as in Figure 3. Both approximations work extremely well for realistic values of $n$, and again we stress that this is a case in which the conditions for large-$n$ convergence do not hold (because the eigenvalues decrease too fast).

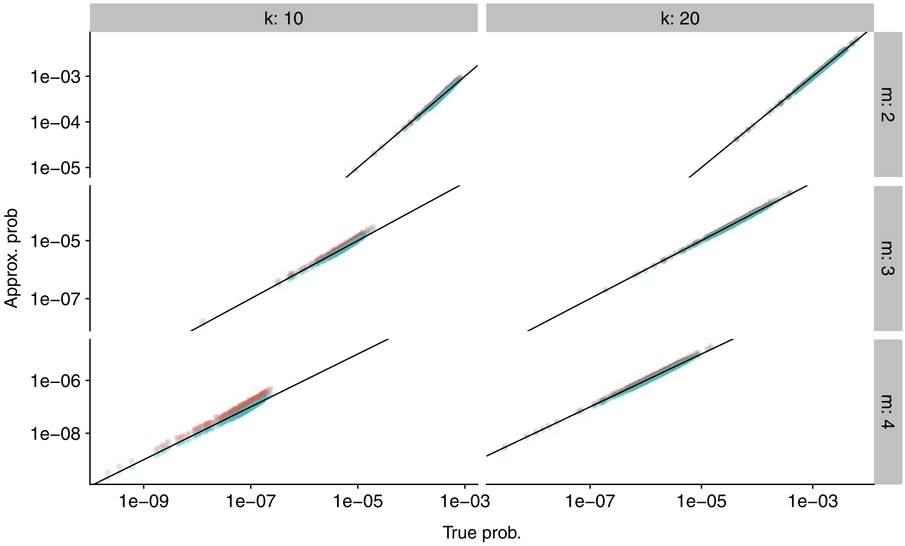

**Figure 7.** Approximation to high order inclusion probabilities in a full $k$-DPP. 400 subsets of size $m$ are drawn at random and their 2nd order inclusion probability estimated using Monte Carlo (see text). We compare the estimate to the $O(n^{-1})$ approximation (in red) and the corrected approximation (in blue), see Appendix A.4.

Theorem 2.1 implies that we can approximate inclusion probabilities for pairs, not just singletons. We show an illustration in Figure 7, where we repeated the above experiment with $n = 500$, $k = 50$ and $\tau = 0.5$. We picked 400 $m$-uples at random and estimated their true inclusion probability using Monte Carlo.[4] We compare the estimated inclusion probability to the $O(n^{-1})$ approximations, and to the corrected probabilities described in Appendix A.4. The $O(n^{-1})$ approximation shows a slight bias for high probabilities but is overall very good, and most of the bias is removed by the correction.

## 4.3. Inference

The goal of this section is to illustrate the claims of Section 2.4, namely that $k$-DPPs and DPPs have equivalent ML estimators (when used as statistical models).

We again used the same setup as in the previous section: $n$ points drawn i.i.d. from a 2D Gaussian, with a subset $\mathcal{X}$ of size $k$ drawn from a k-DPP. Contrary to the previous sections, however, the objective here is to infer something about the L-ensemble given $\mathcal{X}$. We use two statistical models:

1. That $\mathcal{X}$ is drawn from a $k$-DPP with L-ensemble $L_{ij} = \exp(-\frac{\|x_i - x_j\|^2}{2\tau^2})$ (for an unknown value of $\tau$).
2. That $\mathcal{X}$ is drawn from a DPP with L-ensemble $\tilde{L}_{ij} = e^{\nu} \exp(-\frac{\|x_i - x_j\|^2}{2\tau^2})$ (for an unknown value of $\tau$ and $\nu$).

In Figure 8, we show the log-likelihood of the $k$-DPP model as a function of $\tau$, along with the profile log-likelihood of the DPP model ($\mathcal{C}^\star$, see eq. (2.28)). The maximum likelihood estimates of $\tau$ are the argmax of these curves, and as predicted they are extremely close.

# 5. Discussion

We have shown that $k$-DPPs are for practical purposes largely equivalent to DPPs, so that one can sample from a $k$-DPP, pretend that the realisation actually came from a matched DPP, and expect no major damage. Corrections to the inclusion probabilities come at little extra cost and increase the accuracy enough so that the approximations can be used with very small $n$. The saddlepoint approximation can be used to compute ESPs as well, and if more accuracy is needed we suggest including further Edgeworth terms. The remaining hurdle is to develop appropriate algorithms that estimate the relevant functions of the L-ensemble, to remove the need for an eigenvalue decomposition. We hope to develop such methods in future work.

[4] We used an empirical version of eq. (A.11), and 1,500 samples

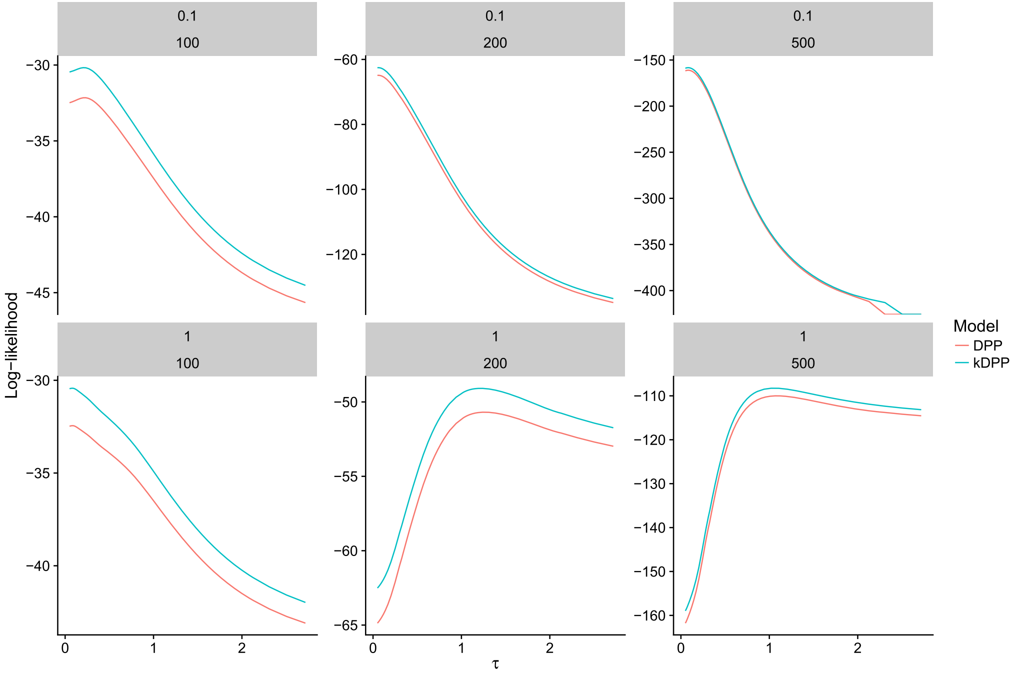

**Figure 8.** Log-likelihoods of a $k$-DPP and DPP model, for various values of $n$ (100 to 500) and true parameter $\tau$ (0.1 and 1). In blue, the $k$-DPP likelihood, red, the DPP profile log-likelihood. In all cases, $k = \frac{n}{10}$.

# Appendix

## A.1. Proof of Lemma 1.3

Here we prove Lemma 1.3, which states that the inclusion kernel of a projection DPP equals the L-ensemble. We need to compute the probability that $\boldsymbol{\alpha} \subseteq \mathcal{X}$, where $\mathcal{X}$ is a sample from a $k$-DPP with L-ensemble $\mathbf{L} = \mathbf{U}\mathbf{U}^\top$, and $\mathbf{U}$ is a $n \times k$ matrix such that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$. It is important here that we are sampling sets of size $k$ from an L-ensemble of rank $k$. As elsewhere we note $|\boldsymbol{\alpha}| = m \leq k$.

We need the following well-known result on determinants of bordered matrices:

$$\det \begin{bmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^\top & c \end{bmatrix} = (\det \mathbf{A})\left(c - \mathbf{b}^\top \mathbf{A}^{-1} \mathbf{b}\right) \tag{A.1}$$

We also make use of the following result, which lets us perform partial sums in determinants.

$$\sum_{i=1}^{n} \left(\mathbf{L}_{i,i} - \mathbf{L}_{i,\mathcal{X}} \mathbf{L}_{\mathcal{X}}^{-1} \mathbf{L}_{\mathcal{X},i}\right) = \operatorname{Tr} \mathbf{L} - \sum_{i} \operatorname{Tr}\left\{\mathbf{L}_{\mathcal{X}}^{-1} \mathbf{L}_{\mathcal{X},i} \mathbf{L}_{i,\mathcal{X}}\right\}$$

$$
\begin{aligned}
&= k - \mathrm{Tr}\left\{\mathbf{L}_{\mathcal{X}}^{-1}\sum_i \mathbf{L}_{\mathcal{X},i}\mathbf{L}_{i,\mathcal{X}}\right\} \\
&= k - \mathrm{Tr}\left\{\left(\mathbf{U}_{\mathcal{X},:}\mathbf{U}_{:,\mathcal{X}}^{\top}\right)^{-1}\sum_i \left(\mathbf{U}_{\mathcal{X},:}\mathbf{U}_{:,i}^{\top}\right)\left(\mathbf{U}_{i,:}\mathbf{U}_{:,\mathcal{X}}^{\top}\right)\right\} \\
&= k - \mathrm{Tr}\left\{\left(\mathbf{U}_{\mathcal{X},:}\mathbf{U}_{:,\mathcal{X}}^{\top}\right)^{-1}\left(\mathbf{U}_{\mathcal{X},:}\mathbf{U}_{:,\mathcal{X}}^{\top}\right)\right\} \\
&= k - |\mathcal{X}| \qquad \text{(A.2)}
\end{aligned}
$$

To simplify what follows, we change the setting a bit and look at *ordered sets*: we sample $\mathcal{X}$ from a DPP, give it a random order (one of $k!$), and thus obtain a vector $\mathbf{x}$. Instead of computing $p(\boldsymbol{\alpha} \subseteq \mathcal{X})$, we compute $p(x_1 = \alpha_1, x_2 = \alpha_2, \dots, x_m = \alpha_m)$, for one particular ordering of $\boldsymbol{\alpha}$. The two probabilities are related:

$$
p(x_1 = \alpha_1, x_2 = \alpha_2, \dots, x_m = \alpha_m) = p(\boldsymbol{\alpha} \subseteq \mathcal{X})\frac{(k-m)!}{k!}
$$

Further, note that the probability mass function for $\mathbf{x}$ is just $p(\mathbf{x}) = \frac{p(\mathcal{X})}{k!}$. Let us now compute $P_{\boldsymbol{\alpha}} = p(x_1 = \alpha_1, x_2 = \alpha_2, \dots, x_m = \alpha_m)$.

$$
\begin{aligned}
P_{\boldsymbol{\alpha}} &= \sum_{x_{m+1}\dots x_k} p\big(\mathbf{x} = [\alpha_1, \dots, \alpha_k, x_{m+1}, \dots, x_k]\big) \qquad \text{(A.3)} \\
&= \frac{1}{k!}\sum_{x_{m+1}\dots x_m} \det \mathbf{L}_{\{\boldsymbol{\alpha}, x_{m+1}, \dots, x_k\}} \qquad \text{(A.4)}
\end{aligned}
$$

Note that since the determinant equals 0 if there are repeated elements, it does not matter if we include repeated elements in the sum. Applying eq. (A.1), we obtain:

$$
P_{\boldsymbol{\alpha}} = \frac{1}{k!}\sum_{\mathbf{z}, x_k} \det \mathbf{L}_{\{\boldsymbol{\alpha},\mathbf{z}\}}\left(\mathbf{L}_{x_k} - \mathbf{L}_{x_k,\{\boldsymbol{\alpha},\mathbf{z}\}}\mathbf{L}_{\{\boldsymbol{\alpha},\mathbf{z}\}}^{-1}\mathbf{L}_{\{\boldsymbol{\alpha},\mathbf{z}\},x_k}\right) \qquad \text{(A.5)}
$$

where we have replaced $x_{m+1} \dots x_{k-1}$ with a vector $\mathbf{z}$ of length $k - m - 1$. Next, we sum over $x_k$, applying eq. (A.2):

$$
\begin{aligned}
P_{\boldsymbol{\alpha}} &= \frac{1}{k!}\sum_{\mathbf{z}} \det \mathbf{L}_{\{\boldsymbol{\alpha},\mathbf{z}\}} \sum_{x_k}\left(\mathbf{L}_{x_k} - \mathbf{L}_{x_k,\{\boldsymbol{\alpha},\mathbf{z}\}}\mathbf{L}_{\{\boldsymbol{\alpha},\mathbf{z}\}}^{-1}\mathbf{L}_{\{\boldsymbol{\alpha},\mathbf{z}\},x_k}\right) \qquad \text{(A.6)} \\
&= \frac{1}{k!}\sum_{\mathbf{z}} \det \mathbf{L}_{\{\boldsymbol{\alpha},\mathbf{z}\}}\big(k - (k-1)\big) \qquad \text{(A.7)}
\end{aligned}
$$

Doing this recursively for $x_{k-1}, x_{k-2}, \dots$ up to $x_{m+1}$, we obtain:

$$P_{\boldsymbol{\alpha}} = \frac{1}{k!}(\det \mathbf{L}_{\boldsymbol{\alpha}})(k-m)(k-m-1)\cdots 1 \tag{A.8}$$

$$= \frac{(k-m)!}{k!}\det \mathbf{L}_{\boldsymbol{\alpha}} \tag{A.9}$$

which in turns implies:

$$p(\boldsymbol{\alpha} \subseteq \mathcal{X}) = \det \mathbf{L}_{\boldsymbol{\alpha}} \tag{A.10}$$

## A.2. Reduction to diagonal DPPs

Since $k$-DPP are mixtures of diagonal $k$-DPPs we can write

$$p(\boldsymbol{\alpha} \subseteq \mathcal{X}|k) = E\big[p(\boldsymbol{\alpha} \subseteq \mathcal{X})_{\mathcal{Y}}\big] \tag{A.11}$$

where the outer expectation is over diagonal $k$-DPPs $\mathcal{Y}$, and

$$p(\boldsymbol{\alpha} \subseteq \mathcal{X})_{\mathcal{Y}} = \det\big(L(\mathcal{Y})_{\boldsymbol{\alpha}}\big) \tag{A.12}$$

with $L(\mathcal{Y}) = \mathbf{U}_{\mathcal{Y},:}\mathbf{U}_{:,\mathcal{Y}}^{\top}$.

Let $\mathbf{Y}$ be a diagonal matrix with $y_{ii} = 1$ if $i \in \mathcal{Y}$, and 0 otherwise. Then we may express the marginal probability of inclusion as:

$$\begin{aligned} p(\boldsymbol{\alpha} \subseteq \mathcal{X}) &= E\big[\det\big(\big(\mathbf{U}\mathbf{Y}\mathbf{U}^{\top}\big)_{\boldsymbol{\alpha}}\big)\big] \\ &= E\big[\det\big(\big(\mathbf{U}_{\boldsymbol{\alpha},:}\mathbf{Y}\mathbf{U}_{:,\boldsymbol{\alpha}}^{\top}\big)\big)\big] \end{aligned} \tag{A.13}$$

where the expectation is over $\mathcal{Y}$. The determinant inside the expectation can be computed using the Cauchy–Binet theorem, giving:

$$\begin{aligned} p(\boldsymbol{\alpha} \subseteq \mathcal{X}) &= E\Bigg[\sum_{\boldsymbol{\beta}/|\boldsymbol{\beta}|=|\boldsymbol{\alpha}|} \det \mathbf{U}_{\boldsymbol{\alpha}\boldsymbol{\beta}} \det\big(\mathbf{Y}\mathbf{U}^{\top}\big)_{\boldsymbol{\beta}\boldsymbol{\alpha}}\Bigg] \\ &= E\Bigg[\sum_{\boldsymbol{\beta}/|\boldsymbol{\beta}|=|\boldsymbol{\alpha}|} \det \mathbf{U}_{\boldsymbol{\alpha}\boldsymbol{\beta}}\mathbf{U}_{\boldsymbol{\alpha}\boldsymbol{\beta}}^{\top} \prod_{i\in\boldsymbol{\beta}} y_i\Bigg] \\ &= \sum_{\boldsymbol{\beta}/|\boldsymbol{\beta}|=|\boldsymbol{\alpha}|} p(\boldsymbol{\beta} \subseteq \mathcal{Y}) \det \mathbf{U}_{\boldsymbol{\alpha}\boldsymbol{\beta}}\mathbf{U}_{\boldsymbol{\alpha}\boldsymbol{\beta}}^{\top} \end{aligned} \tag{A.14}$$

In particular, for singletons $|\boldsymbol{\alpha}| = 1$ we recover the inclusion probability of order 1 given in Section 1.5.2.

Suppose now we have to measure the total variation distance between an inclusion probability of a $k$-DPP ($\pi$) and a DPP approximation of it ($\tilde{\pi}$). Recalling that each is a mixture of diagonal

DPPs with inclusion measure $\rho$ and $\tilde{\rho}$, we write

$$\begin{aligned} D_m(\pi,\widetilde{\pi}) &= \frac{1}{\binom{k}{m}}\sum_{\boldsymbol{\alpha}}|\pi(\boldsymbol{\alpha})-\widetilde{\pi}(\boldsymbol{\alpha}))| \\ &= \frac{1}{\binom{k}{m}}\sum_{\boldsymbol{\alpha}}\left|\sum_{\boldsymbol{\beta}}\det\mathbf{U}_{\boldsymbol{\alpha}\boldsymbol{\beta}}\mathbf{U}_{\boldsymbol{\alpha}\boldsymbol{\beta}}^{\top}\big(\rho(\boldsymbol{\beta})-\widetilde{\rho}(\boldsymbol{\beta})\big)\right| \\ &\leq \frac{1}{\binom{k}{m}}\sum_{\boldsymbol{\beta}}\big|\rho(\boldsymbol{\beta})-\widetilde{\rho}(\boldsymbol{\beta})\big|\sum_{\boldsymbol{\alpha}}\det\mathbf{U}_{\boldsymbol{\alpha}\boldsymbol{\beta}}\mathbf{U}_{\boldsymbol{\alpha}\boldsymbol{\beta}}^{\top} \end{aligned}$$

But $\sum_{\boldsymbol{\alpha}}\det\mathbf{U}_{\boldsymbol{\alpha}\boldsymbol{\beta}}\mathbf{U}_{\boldsymbol{\alpha}\boldsymbol{\beta}}^{\top} = \sum_{\boldsymbol{\alpha}}\det(\mathbf{U}^{\top})_{\boldsymbol{\beta}\boldsymbol{\alpha}}\mathbf{U}_{\boldsymbol{\alpha}\boldsymbol{\beta}} = \det(\mathbf{U}^{\top}\mathbf{U})_{\boldsymbol{\beta}\boldsymbol{\beta}} = 1$. Thus $D_m(\pi,\widetilde{\pi}) \leq D_m(\rho,\widetilde{\rho})$, proving Lemma 2.1.

## A.3. Computing an asymptotic expansion for diagonal inclusion probabilities

To derive the $O(1)$ and $O(n^{-1})$ terms in the inclusion probabilities, we take the exact expression (eq. 2.7) and inject the saddlepoint approximation (eq. (2.19), which yields:

$$\begin{aligned} p_k\left(\prod_{j\in\boldsymbol{\alpha}} z_j = 1\right) &= \left(\prod_{i\in\boldsymbol{\alpha}}\frac{\lambda_i}{1+\lambda_i}\right)\frac{\sqrt{\psi''(\nu^\star)}}{\sqrt{\psi''(\nu^\star_{\boldsymbol{\alpha}})-\psi''_{\boldsymbol{\alpha}}(\nu^\star_{\boldsymbol{\alpha}})}} \\ &\quad\times\exp\big(\psi\big(\nu^\star_{\boldsymbol{\alpha}}\big)-\psi\big(\nu^\star\big)-\psi_{\boldsymbol{\alpha}}\big(\nu^\star_{\boldsymbol{\alpha}}\big)+k\nu^\star-(k-m)\nu^\star_{\boldsymbol{\alpha}}\big) \qquad \text{(A.15)} \end{aligned}$$

where $\psi_{\boldsymbol{\alpha}} = \sum_{i\in\boldsymbol{\alpha}}\psi_i$, $\psi'_{\boldsymbol{\alpha}} = \sum_{i\in\boldsymbol{\alpha}}\psi'_i$ and so on. The relative error in this approximation is of order $O(n^{-2})$[5] and we neglect it from now on. To get the $O(1)$ and $O(n^{-1})$ terms, we use a perturbation approach, where we treat $\epsilon = n^{-1}$ as a (scalar) perturbation parameter. The reason we have to use a perturbation approach is for lack of an analytical expression for the saddlepoint parameter $\nu^\star$. To do so, we split eq. (A.15) into three terms:

$$\begin{aligned} A &= \left(\prod_{i\in\boldsymbol{\alpha}}\frac{\lambda_i}{1+\lambda_i}\right) \\ B &= \frac{\sqrt{\psi''(\nu^\star)}}{\sqrt{\psi''(\nu^\star_{\boldsymbol{\alpha}})-\psi''_{\boldsymbol{\alpha}}(\nu^\star_{\boldsymbol{\alpha}})}} \\ C &= \exp\big(\psi\big(\nu^\star_{\boldsymbol{\alpha}}\big)-\psi\big(\nu^\star\big)-\psi_{\boldsymbol{\alpha}}\big(\nu^\star_{\boldsymbol{\alpha}}\big)+k\nu^\star-(k-m)\nu^\star_{\boldsymbol{\alpha}}\big) \end{aligned}$$

[5] The reason the relative error is $O(n^{-2})$ is that we take a ratio of $O(n^{-1})$ errors that are actually the same up to a $O(n^{-1})$ term. Intuitively, the relative errors in the saddlepoint approximation of $e_k(\boldsymbol{\lambda})$ and $e_{k-m}(\boldsymbol{\lambda}_{-\boldsymbol{\alpha}})$ are almost the same, and thus most of the error cancels when we take the ratio.

We shall find series for $B$ and $C$ of the form $B = b_0 + \epsilon b_1 + \epsilon^2 b_2 + \cdots$, $C = \exp(c_0 + \epsilon c_1 + \epsilon^2 c_2 + \cdots)$. We will see that here $b_0 = 1$. These series can in turn be used to obtain approximations of order $\epsilon^0$ and $\epsilon^1$ to the product $ABC$, namely:

$$p_k\left(\prod_{j\in\boldsymbol{\alpha}} z_j = 1\right) = A(\exp(c_0)\big(1 + \epsilon(c_1 + b_1) + O(\epsilon^2)\big)$$

We note $\bar{\psi} = \frac{1}{n}\psi$, $\bar{\psi}_{\boldsymbol{\alpha}} = \frac{1}{m}\psi_{\boldsymbol{\alpha}}$, $r = \frac{k}{n}$. To obtain our perturbation series, we begin with the perturbed solution to the saddlepoint equation $\nu^\star_{\boldsymbol{\alpha}}$, defined by:

$$\begin{aligned}\nu^\star_{\boldsymbol{\alpha}} &= \operatorname*{argmin}_\nu \psi(\nu) - \psi_{\boldsymbol{\alpha}}(\nu) - (k-m)\nu \\ &= \operatorname*{argmin}_\nu \bar{\psi}(\nu) - r\nu + m\epsilon\big(\nu - \bar{\psi}_{\boldsymbol{\alpha}}(\nu)\big) \\ &= \operatorname*{argmin}_\nu f(\nu, \epsilon) \qquad \text{(A.16)}\end{aligned}$$

Define the ansatz $\nu^\star(\epsilon) = \operatorname{argmin} f(\nu,\epsilon) = \nu_0 + \epsilon\nu_1 + \epsilon^2\nu^2 + \cdots$. From the saddlepoint equation we obtain:

$$\bar{\psi}'(\nu^\star) - r + m\epsilon\big(1 - \bar{\psi}'_{\boldsymbol{\alpha}}(\nu^\star)\big) = 0$$

At order $\epsilon^0$, the equation implies:

$$\bar{\psi}'(\nu_0) - r = 0 \qquad \text{(A.17)}$$

so that $\nu_0$ equals the saddlepoint of the unperturbed problem. At order $\epsilon^1$, we obtain:

$$\bar{\psi}''(\nu_0)\nu_1 - m\big(1 - \bar{\psi}'_{\boldsymbol{\alpha}}(\nu_0)\big) = 0 \qquad \text{(A.18)}$$

Further orders are not needed for our purposes.

We are now ready to insert these equations back into (A.15). We begin with the exponential part.

$$\begin{aligned}C(\epsilon) &= \exp\big(n\big(f\big(\nu^\star_{\boldsymbol{\alpha}}, \epsilon\big) - \bar{\psi}\big(\nu^\star\big) - r\nu^\star\big)\big) &\text{(A.19)}\\ &= \exp\big(n\big(f\big(\nu_0 + \epsilon\nu_1 + \epsilon^2\nu_2 + \cdots, \epsilon\big) - \bar{\psi}(\nu_0) - r\nu_0\big)\big) &\text{(A.20)}\end{aligned}$$

We proceed with a similar perturbation for $f(\nu^\star(\epsilon), \epsilon)$, $f = f_0 + \epsilon f_1 + \epsilon^2 f_2 + \cdots$

$$\begin{aligned}f_0 &= \bar{\psi}(\nu_0) - r\nu_0 &\text{(A.21)}\\ f_1 &= \bar{\psi}'(\nu_0)\nu_1 - r\nu_1 + m\big(\nu_0 - \bar{\psi}_{\boldsymbol{\alpha}}(\nu_0)\big) &\text{(A.22)}\\ &= m\big(\nu_0 - \bar{\psi}_{\boldsymbol{\alpha}}(\nu_0)\big) &\text{(A.23)}\end{aligned}$$

$$f_2 = \bar{\psi}'(\nu_0)\nu_2 - r\nu_2 + \frac{1}{2}\bar{\psi}''(\nu_0)\nu_1^2 + m\big(\nu_1 - \bar{\psi}'_{\boldsymbol{\alpha}}(\nu_0)\nu_1\big) \tag{A.24}$$

$$= -\frac{\nu_1^2}{2}\bar{\psi}''(\nu_0) \tag{A.25}$$

where we have made use of eq. (A.17) and (A.18). Inserting (A.21) into (A.19), we find

$$C(\epsilon) = \exp\big(f_1 + \epsilon f_2 + O\big(\epsilon^2\big)\big) \tag{A.26}$$

We now proceed with the other factor of (A.15), $B$, involving a ratio of square roots:

$$B(\epsilon) = \frac{\sqrt{\psi''(\nu^\star)}}{\sqrt{\psi''(\nu^\star_{\boldsymbol{\alpha}}) - \psi''_{\boldsymbol{\alpha}}(\nu^\star_{\boldsymbol{\alpha}})}} = \frac{\sqrt{\bar{\psi}''(\nu_0)}}{\sqrt{\bar{\psi}''(\nu^\star_{\boldsymbol{\alpha}}) - m\epsilon\bar{\psi}''_{\boldsymbol{\alpha}}(\nu^\star_{\boldsymbol{\alpha}})}} \tag{A.27}$$

Note that

$$\sqrt{\frac{a}{a-\epsilon}} = \sqrt{\frac{1}{1-\frac{\epsilon}{a}}} = \sqrt{\left(1 + \frac{\epsilon}{a} + O\big(\epsilon^2\big)\right)} = 1 + \frac{\epsilon}{2a} + O\big(\epsilon^2\big) \tag{A.28}$$

It is immediate from the above that $B = 1 + O(\epsilon)$, and that therefore:

$$\begin{aligned} p_k\left(\prod_{j\in\boldsymbol{\alpha}} z_j = 1\right) &= \left(\prod_{i\in\boldsymbol{\alpha}} \frac{\lambda_i}{1+\lambda_i}\right)\exp\big(f_1 + O(\epsilon)\big) \\ &= \left(\prod_{i\in\boldsymbol{\alpha}} \frac{\lambda_i}{1+\lambda_i}\right)\exp\big(m\big(\nu_0 - \bar{\psi}_{\boldsymbol{\alpha}}(\nu_0)\big) + O(\epsilon)\big) \\ &= \left(\prod_{i\in\boldsymbol{\alpha}} \frac{\lambda_i \exp(\nu_0)}{1+\lambda_i\exp(\nu_0)}\right)\left(1 + O\left(\frac{1}{n}\right)\right) \end{aligned} \tag{A.29}$$

For numerical purposes it is interesting to obtain the $O(\epsilon)$ term, which requires the first-order approximation to $B$:

$$B(\epsilon) = 1 - \epsilon\frac{1}{2\bar{\psi}''(\nu_0)}\big(\bar{\psi}^{(3)}(\nu_0)\nu_1 - m\bar{\psi}''_{\boldsymbol{\alpha}}(\nu_0)\big) + O\big(\epsilon^2\big) \tag{A.30}$$

This completes the proof of Lemma 2.3.

## A.4. An easy-to-compute correction to the $O(n^{-1})$ approximation

One way to get an improved estimate of inclusion probabilities is to compute the $O(n^{-1})$ term in the saddlepoint expansion, and that is what we recommend for first-order inclusion probabilities. It is harder to use when $m > 1$, and in this section we describe a correction that is easy to compute and yields interesting insights into the approximation. From Lemma 1.2, we know what the sum of the inclusion measure for a $k$-DPP over all sets of size $m$ should equal $\binom{k}{m}$, while for a DPP with marginal kernel $\mathbf{K}$ it equals:

$$\sum_{\boldsymbol{\alpha}, |\boldsymbol{\alpha}|=m} \det \mathbf{K}_{\boldsymbol{\alpha}} = e_m(\mathbf{K}) \tag{A.31}$$

the m-th ESP of matrix $\mathbf{K}$. In the matched DPP $\mathbf{K}$ equals $(\mathbf{I} + e^{\nu}\mathbf{L})^{-1}e^{\nu}\mathbf{L}$, and the eigenvalues of $\mathbf{K}$ are $\eta_i = \frac{e^{\nu}\lambda_i}{1+e^{\nu}\lambda_i}$. What eq. (A.31) implies is that for the approximation to be exact at order $m$, we need to have $e_m(\boldsymbol{\eta}) = \binom{k}{m}$. One shows easily that this is true if and only if $\boldsymbol{\eta}$ has exactly $k$ entries that equal 1, and the rest are all zero, which happens to be just the case described in Result 1.

**Lemma A.1.** *Let* $\boldsymbol{\eta} \in [0,1]^n$, *with* $\sum \eta_i = k$, *and let* $m < k$. *Then* $\binom{k}{m} \leq e_m(\boldsymbol{\eta}) \leq \binom{n}{k}(\frac{k}{n})^m$

**Proof.** We seek the extrema of $e_m(\boldsymbol{\eta})$ under the equality constraint $\sum \eta_i = k$ and the inequality constraints $0 \leq \eta_i \leq 1$. Maximisation is easy $e_m(\boldsymbol{\eta})$ is a concave function (as a consequence of Schur concavity, Sra [17]), and we have linear constraints, so that any maximum is unique. The Lagrangian equals:

$$\mathcal{L}(\boldsymbol{\eta}, \nu, \boldsymbol{\gamma}), \boldsymbol{\delta}) = e_m(\boldsymbol{\eta}) - \nu\Big(\sum \eta_i - k\Big) - \boldsymbol{\gamma}^{\top}\boldsymbol{\eta} - \boldsymbol{\delta}^{\top}(\boldsymbol{\eta} - 1) \tag{A.32}$$

and the Karush–Kuhn–Tucker conditions imply that for all $i$:

$$\frac{\partial}{\partial \eta_i} e_m(\boldsymbol{\eta}) = e_{m-1}(\boldsymbol{\eta}_{-i}) = \nu + \gamma_i + \delta_i \tag{A.33}$$

$$\gamma_i \eta_i = 0 \tag{A.34}$$

$$\delta_i(\eta_i - 1) = 0 \tag{A.35}$$

The solution where $\eta_i = \frac{k}{n}$ for all $n$ only has inactive constraints, and is a maximum. Finding minima requires a bit more work. Let us consider a potential solution $\boldsymbol{\eta}$, and split into three consecutive parts: the zero values (active constraints under the $\gamma$ multiplier), the values contained above zero and below one (inactive constraints), and the values equal to one. The KKT conditions imply that for all $j$ such that the $j$-th constraint is inactive, $e_{m-1}(\eta_{-j}) = \nu$, meaning that removing any $0 < \eta_j < 1$ has the same effect, which implies that all these values are the same. Consequently, we can reparametrise the solution as

$$\boldsymbol{\eta} = [0, 0, \ldots, 0, a, a, \ldots, a, 1, 1, \ldots, 1]$$

where $0 < a < 1$. Since $e_m$ is invariant to permutations there is no loss of generality. Next, notice that $e_m([0\boldsymbol{\beta}]) = e_m(\boldsymbol{\beta})$ for all $m$. Further:

$$e_m\big([\boldsymbol{\beta}1]\big) = e_{m+1}(\boldsymbol{\beta}) + e_m(\boldsymbol{\beta}) \tag{A.36}$$

so that $e_m([0, 0, \dots, 0, a, a, \dots, a, 1, 1, \dots, 1])$ is just a weighted sum of elementary symmetric polynomials of the vector $[a, a, \dots, a]$, so that $a$ needs to be as small as possible under the constraints. The minima must therefore all have $k$ values equal to one, and the rest zero. Evaluating $e_m$ at the two extrema yields the bound. □

The least favorable case is therefore when $\mathbf{L}$ has a flat spectrum (which thankfully should not happen), but even then asymptotic equivalence holds: one can verify that $\binom{n}{k}(\frac{k}{n})^m \asymp \binom{k}{m}$. However, at finite orders, one can improve the approximation by making sure it sums to the right quantity: i.e., approximate inclusion probabilities via:

$$\tilde{\pi}_{\text{corrected}}(\boldsymbol{\alpha}) = \frac{\binom{k}{m}}{e_m(\boldsymbol{\eta})} \det \mathbf{K}_\alpha \tag{A.37}$$

When $m = 1$ the correction does nothing (the correction factor equals 1), but at higher orders we have found that it can sometimes reduce relative error by a factor of 10.

## Acknowledgements

This work benefited from funding from ANR GenGP (ANR-16-CE23-0008), LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01), Grenoble Data Institute (ANR-15-IDEX-02), CNRS PEPS I3A (Project RW4SPEC), and LIA CNRS/Melbourne Univ Geodesic.

## References

[1] Billingsley, P. (2008). *Probability and Measure*. John Wiley & Sons.

[2] Chen, X.-H., Dempster, A.P. and Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika* **81** 457–469. MR1311090

[3] Couillet, R. and Debbah, M. (2011). *Random Matrix Methods for Wireless Communications*. Cambridge: Cambridge Univ. Press. MR2884783

[4] Daniels, H.E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Stat.* **25** 631–650. MR0066602

[5] DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. *Springer Texts in Statistics*. New York: Springer. MR2664452

[6] Deshpande, A. and Rademacher, L. (2010). Efficient volume sampling for row/column subset selection. In 2010 *IEEE* 51*st Annual Symposium on Foundations of Computer Science—FOCS* 2010 329–338. Los Alamitos, CA: IEEE Computer Soc. MR3025206

[7] Deshpande, A., Rademacher, L., Vempala, S. and Wang, G. (2006). Matrix approximation and projective clustering via volume sampling. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms* 1117–1126. New York: ACM. MR2373839

[8] Gautier, G., Bardenet, R. and Valko, M. (2017). Zonotope hit-and-run for efficient sampling from projection DPPs. ArXiv Preprint arXiv:1705.10498.

[9] Jozsa, R. and Mitchison, G. (2015). Symmetric polynomials in information theory: Entropy and subentropy. *J. Math. Phys.* **56** 062201, 17. MR3369891

[10] Kulesza, A. and Taskar, B. (2011). k-DPPs: Fixed-size determinantal point processes. In *Proceedings of the* 28*th International Conference on Machine Learning* (*ICML*-11) 1193–1200.

[11] Kulesza, A. and Taskar, B. (2012). Determinantal point processes for machine learning. *Found. Trends Mach. Learn.* **5** 123–286.

[12] Li, C., Jegelka, S. and Sra, S. (2016). Efficient sampling for k-determinantal point processes.

[13] Macchi, O. (1975). The coincidence approach to stochastic point processes. *Adv. in Appl. Probab.* **7** 83–122. MR0380979

[14] Mariet, Z. and Sra, S. (2017). Elementary symmetric polynomials for optimal experimental design. ArXiv Eprint arXiv:1705.09677v1.

[15] Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications. Monographs on Statistics and Applied Probability* **104**. Boca Raton, FL: CRC Press/CRC. MR2130347

[16] Soshnikov, A. (2000). Determinantal random point fields. *Russian Math. Surveys* **55** 923–975.

[17] Sra, S. (2019). Logarithmic inequalities under a symmetric polynomial dominance order. *Proc. Amer. Math. Soc.* **147** 481–486. MR3894886

[18] Touchette, H. (2015). Equivalence and nonequivalence of ensembles: Thermodynamic, macrostate, and measure levels. *J. Stat. Phys.* **159** 987–1016. MR3345408

[19] Tremblay, N., Barthelmé, S. and Amblard, P.-O. (2018). Determinantal point processes for coresets. ArXiv Preprint arXiv:1803.08700.

[20] Tremblay, N., Barthelme, S. and Amblard, P.O. (2018). Optimized algorithms to sample determinantal point processes. ArXiv E-print arXiv:1802.08471.

*Received March 2018 and revised August 2018*