Proceedings of Machine Learning Research vol 178:1–20, 2022 35th Annual Conference on Learning Theory

Hardness of Maximum Likelihood Learning of DPPs*

Elena Grigorescu ELENA-G@PURDUE.EDU
Purdue University

Brendan Juba BJUBA@WUSTL.EDU
Washington University in St. Louis

Karl Wimmer WIMMERK@DUQ.EDU
Duquesne University

Ning Xie NXIE@CIS.FIU.EDU
Florida International University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Determinantal Point Processes (DPPs) are a widely used probabilistic model for negatively correlated sets. DPPs have been successfully employed in Machine Learning applications to select a diverse, yet representative subset of data. In these applications, the parameters of the DPP need to be fitted to match the data; typically, we seek a set of parameters that maximize the likelihood of the data. The algorithms used for this task to date either optimize over a limited family of DPPs, or use local improvement heuristics that do not provide theoretical guarantees of optimality.

It is natural to ask if there exist efficient algorithms for finding a maximum likelihood DPP model for a given data set. In seminal work on DPPs in Machine Learning, Kulesza conjectured in his PhD Thesis (2012) that the problem is NP-complete. The lack of a formal proof prompted Brunel, Moitra, Rigollet and Urschel (2017a) to conjecture that, in opposition to Kulesza's conjecture, there exists a polynomial-time algorithm for computing a maximum-likelihood DPP. They also presented some preliminary evidence supporting their conjecture.

In this work we prove Kulesza's conjecture. In fact, we prove the following stronger hardness of approximation result: even computing a $1 - \frac{1}{\operatorname{poly}\log N}$-approximation to the maximum log-likelihood of a DPP on a ground set of N elements is NP-complete. At the same time, we also obtain the first polynomial-time algorithm that achieves a nontrivial worst-case approximation to the optimal log-likelihood: the approximation factor is $\frac{1}{(1+o(1))\log m}$ unconditionally (for data sets that consist of m subsets), and can be improved to $1-\frac{1+o(1)}{\log N}$ if all N elements appear in a $O(1/N)$-fraction of the subsets.

In terms of techniques, we reduce approximating the maximum log-likelihood of DPPs on a data set to solving a gap instance of a "vector coloring" problem on a hypergraph. Such a hypergraph is built on a bounded-degree graph construction of Bogdanov, Obata and Trevisan (2002), and is further enhanced by the strong expanders of Alon and Capalbo (2007) to serve our purposes.

* Please see the associated technical report for complete proofs: http://arxiv.org/abs/2205.12377

1. Introduction

Determinantal Point Processes (DPPs) are a family of probability distributions on sets that feature repulsion among elements in the ground set. Roughly speaking, a DPP is a distribution over all 2^N subsets of $\{1, \ldots, N\}$ defined by a positive semidefinite (PSD) $N \times N$ matrix K (called a *marginal kernel* or *correlation kernel*) whose eigenvalues all lie in $[0, 1]$, such that, for any $S \subseteq \{1, \ldots, N\}$, random subsets $\boldsymbol{X}$ drawn according to the distribution satisfy $\Pr[S \subseteq \boldsymbol{X}] = \det(K_S)$, where K_S is the principal submatrix of K indexed by S.

DPPs were first proposed in quantum statistical physics to model fermion systems in thermal equilibrium (Macchi, 1975), but they also arise naturally in diverse fields such as random matrix theory, probability theory, number theory, random spanning trees and non-intersecting paths (Dyson, 1962; Burton and Pemantle, 1993; Rudnick and Sarnak, 1996; Soshnikov, 2000). After the seminal work of Kulesza and Taskar (2012), DPPs have attracted a flurry of attention from the machine learning community due to their computational tractability and excellent capability at producing diverse but representative subsets, and subsequently fast algorithms have been developed for sampling from DPPs (Hough et al., 2006; Kulesza and Taskar, 2010; Rebeschini and Karbasi, 2015; Li et al., 2016b,a; Anari et al., 2016; Dereziński et al., 2019; Launay et al., 2020). Furthermore, DPPs have since found a vast variety of applications throughout machine learning and data analysis, including text and image search, segmentation and summarization (Lin and Bilmes, 2012; Kulesza and Taskar, 2012; Zou and Adams, 2012; Gillenwater et al., 2012b,a; Yao et al., 2016; Kulesza and Taskar, 2011b; Affandi et al., 2014; Lee et al., 2016; Affandi et al., 2013b; Chao et al., 2015; Affandi et al., 2013a), signal processing (Xu and Ou, 2016; Krause et al., 2008; Guestrin et al., 2005), clustering (Zou and Adams, 2012; Kang, 2013; Shah and Ghahramani, 2013), recommendation systems (Zhou et al., 2010), revenue maximization (Dughmi et al., 2009), multi-agent reinforcement learning (Osogami and Raymond, 2019; Yang et al., 2020), modeling neural spikes (Snoek et al., 2013), sketching for linear regression (Dereziński and Warmuth, 2018; Derezinski et al., 2020), low-rank approximation (Guruswami and Sinop, 2012), and likely many more.

Maximum likelihood estimation. Many of these applications require inferring a set of parameters for a DPP capturing a given data set. As a DPP specifies a probability distribution, hence in contrast to supervised learning problems, the quality of a DPP cannot be estimated by the "error rate" of the model's predictions. The standard approach to estimate a DPP from data is to find parameters that maximize the *likelihood* of the given data set being produced by a sample from the DPP (Kulesza and Taskar, 2012), i.e., the probability density of the observed data in the joint distribution. When the samples are identically and independently chosen from the DPP, the likelihood is the product of the probability densities the DPP assigns to the sampled subsets. The goal of the maximum likelihood estimator (MLE) method is to find a kernel matrix that maximizes the likelihood of the data set. Brunel et al. (2017b) showed that the maximum likelihood estimate indeed converges to the true kernel at a polynomial rate. In general, maximizing the likelihood of a DPP gives rise to a non-convex optimization problem, and has been approached with heuristics such as expectation maximization (Gillenwater et al., 2014), fixed point algorithms (Mariet and Sra, 2015), and MCMC (Affandi et al., 2014). In the continuous case, the problem has been studied under strong parametric assumptions (Lavancier et al., 2015), or smoothness assumptions (Baraud, 2013).

1.1. Our results

In spite of the wide applications of DPPs and the central role of the learning step, no efficient algorithms are known to find a maximum likelihood DPP. Instead, as mentioned above, two families of algorithms are known: one seeks to learn an optimal DPP within certain parameterized families of DPP structures (Kulesza and Taskar, 2012; Affandi et al., 2014; Gartrell et al., 2016; Mariet and Sra, 2016; Gartrell et al., 2017; Urschel et al., 2017; Dupuy and Bach, 2018), while the other invokes heuristics to maximize the likelihood in an unconstrained search over the parameter space (Kulesza and Taskar, 2011a; Gillenwater et al., 2014; Affandi et al., 2014; Mariet and Sra, 2015). Neither of these approaches provides any guarantees for how close the likelihood of the obtained parameters are to the maximum over all DPPs.

Indeed, Kulesza (2011a; 2012) conjectured over a decade ago that the problem of finding a set of parameters is NP-hard, but indicated that he was unable to formally establish a reduction: his thesis includes a sketch of a reduction from EXACT-3-COVER to a related problem[1] with numerical evidence suggesting its correctness but without a formal proof. The subsequent literature adopted this belief, despite the lack of a solid theoretical foundation.

In this work, we resolve this question by proving Kulesza's conjecture: computing maximum likelihood DPP kernels is indeed NP-hard. In fact, we establish a stronger, gapped hardness result: even approximating the maximum likelihood is NP-hard.

Theorem 1 (Informal Statement of the Main Theorem) *There is a ground set of size N such that it is NP-hard to $\left(1 - O(\frac{1}{\log^9 N})\right)$-approximate the maximum DPP log-likelihood value of a training set.*

Remark 2 *Some comments on our (somewhat confusing) convention of approximation factors are in place. Since log likelihood functions are always negative real numbers and it is a bit awkward to work with optimizing negative quantities, we therefore add a minus sign at the front of our definition of log likelihood functions. Consequently, we* minimize *log likelihood functions instead of maximizing them. On the other hand, as our hard learning instances are reduced from* MAX-3SAT *and* 3-COLORING*, to be consistent with hardness results in the literature on these problems, we use α-approximation (where $0 < \alpha < 1$, for maximization problems) in the statements of our hardness and algorithmic results. Note that here "α-approximation" actually means that the log likelihood function (in our definition and ought to be minimized) output by an algorithm is at most $\frac{1}{\alpha}$ time the optimal log likelihood function.*

Therefore, the difficulty of learning a DPP is not tied to any particular representation of the marginal kernel, as in fact estimating the maximum likelihood *value* itself is NP-hard. Note, however, that many problems in learning are hard merely due to the difficulty of finding a specific representation (Pitt and Valiant, 1988), which is not the case for our problem.

The NP-hardness of maximum likelihood learning naturally raises the question of whether any nontrivial approximation is possible. We show that such an approximation is possible: we present a simple, polynomial-time algorithm obtaining a $\frac{1}{(1+o(1))\log m}$-approximation for a data set with m subsets.

1. Technically, the reduction proposed by Kulesza targets a variant of the maximum-likelihood DPP learning problem in which the instance specifies a set of positive-semidefinite matrices along with the data, and the objective is to find a DPP kernel in the cone of the given matrices that maximizes the likelihood.

Theorem 3 (Informal Statement of the Approximation Algorithm) *There is a polynomial-time approximation algorithm achieving the following: on an input data set consisting of m subsets over a ground set of size N, it returns a kernel that $\frac{1}{(1+o(1))\log m}$-approximates the maximum log likelihood. Moreover, if every element in the ground set appears in at most $O(1/N)$-fraction of the subsets, the kernel achieves a $(1-\frac{1+o(1)}{\log N})$-approximation to the maximum log likelihood.*

We stress that in contrast to the prior work on learning DPP kernels with guarantees (Urschel et al., 2017), our algorithm does not rely on the data being produced by a DPP to have a "cycle basis" of short cycles or large nonzero entries. We obtain an approximation to the optimal likelihood for *any* data set. Although a $\frac{1}{(1+o(1))\log m}$-approximation is weak, when every element appears in relatively few subsets (which is common in practice), our algorithm is much better: the actual approximation factor is $1-\frac{1}{\log(m/a_{\max})}$, where $a_{\max}/m$ is the fraction of the data subsets containing the most frequent element in the ground set. Hence, if all elements appear in at most a $\sim 1/N$-fraction of the subsets, we obtain a $(1-\frac{1+o(1)}{\log N})$-approximation to the log likelihood. Although we don't expect our algorithm to obtain substantially better likelihood than various heuristics employed in practice, it may nevertheless serve as a benchmark to estimate how close to optimal these solutions are. Moreover, the family of instances constructed in our reduction indeed satisfies this condition; therefore, to improve the hardness of approximation bound beyond $1-\frac{1+o(1)}{\log N}$, the hard instance of data set must have some elements appearing in $\omega(1/N)$-fraction of the subsets.

1.2. Our approach and techniques

We show that it is NP-hard to approximate the optimal DPP likelihood function by reducing from a coloring problem, rather than from EXACT-3-COVER, which was Kulesza's (2012) initial approach.

We begin with some intuition leading to a notion of *vector coloring* that we use in the reduction. As any marginal kernel $K \in \mathbb{R}^{N\times N}$ is positive semidefinite, it can be factored as $K = Q^\top Q$, where $Q \in \mathbb{R}^{k\times N}$, $Q^\top$ stands for the transpose of M, and k is called the *dimension* of the kernel. If we denote the column vectors of Q by $q_1, \ldots, q_N$, each $q_i \in \mathbb{R}^k$, then one can think of these q_i's as providing an embedding of the elements in $\{1, \ldots, N\}$ into the space $\mathbb{R}^k$, and the embedding vectors capture similarities among elements. Specifically, the preference of DPPs for diverse subsets, roughly speaking, stems from the following fact: if a subset S includes elements that are similar, the submatrix K_S would contain columns that are nearly co-linear embedding vectors, and hence its determinant (and correspondingly, the probability that S is the random subset generated by the marginal kernel K) is close to zero.[2]

Consider, for simplicity, a training set that consists of a collection of subsets of $\{1, \ldots, N\}$, each of size r where r is a constant. What can we say about a maximum-likelihood DPP kernel for such a data set? Ideally, the embedding vectors should encode an "r-vector-coloring" of the elements in the following sense. Each of the r colors is represented by a unit vector (after normalizations) in an orthonormal basis; to maximize the likelihood function, every subset $S = \{i_1, \ldots, i_r\}$ that appears in the training set corresponds to a "rainbow coloring" of the embedding vectors $\{q_{i_1}, \ldots, q_{i_r}\}$ ("rainbow coloring" means that the r embedding vectors are all colored differently), while for the

2. Recall that, since K and hence K_S are Gram matrices, $\det(K_S)$ is equal to the square of the volume of the parallelepiped spanned by the embedding vectors of elements in S.

r-subsets that do not appear in the training set, we would like as many as possible of them to contain some repeated color.[3]

Thus, it is natural to attempt a reduction from GRAPH r-COLORING to Maximum Likelihood Learning of DPP. However, if we view each edge as a 2-subset, then we fail to get a hard problem to begin with, since graph 2-coloring is easy. We overcome this by "lifting" each edge to a size-3 subset (or equivalently, we transform a graph into a 3-uniform hypergraph, and view all its hyperedges as size-3 subsets in the data set) so that we can still work with 3-COLORABILITY. On an input graph $G = (V, E)$, our goal would be to show the following: if G is 3-colorable, then there is a DPP kernel whose likelihood is large (completeness); if G is not 3-colorable, then the likelihood of *every* DPP kernel is small (soundness).

As the column vectors of DPP kernels are in Euclidean spaces instead of discrete ones, the continuous variant of coloring that works for us is the notion of *vector coloring* of graphs. A *vector coloring* of a graph $G = (V, E)$ is a mapping from vertices in V to points (vectors) in some low-dimensional metric space $\mathcal{M}$, such that the presence or absence of edges between any pair of vertices prescribes the value of the inner product between the two corresponding vectors. (See Section 1.3.) Our problem differs from this one in two important ways: first, we do not care too much about the minimum dimension of the Euclidean space in which a vector representation exists; second, which is more subtle and will be elaborated below, we need a "gapped" reduction.

There are several technical challenges we need to address in order to make the reduction from 3-COLORABILITY work. First, we need to understand the structure of kernels that achieve maximum likelihood values. To this end, by an extension of an argument of Brunel et al. (2017a), we prove that the square of the norm of every embedding vector is equal to the empirical frequency of the element. Furthermore, via a projection argument, we show that there always exists a good *rank*-3 DPP kernel without giving up too much likelihood. This greatly simplifies the analysis of the gadgets employed in our reduction.

Secondly, instead of a "decision" hardness result on vector coloring of hypergraphs (e.g. it is NP-hard to decide if there is a 3-dimensional orthogonal representation for the elements in the set that satisfies certain orthogonality conditions), we rather require a "*gapped*" reduction, obtaining something like "starting with a YES instance, we end up with a set of embedding vectors so that the *average* volume of the 3-dimensional parallelepipeds spanned by these embedding vectors of hyperedges is large; starting with a NO instance, then every possible embedding scheme will make the *average* volume of those parallelepipeds small". Namely, in an "averaging" sense, we need the resulting hypergraph transformed from a NO instance to be "far from" 3-vector colorable. Accordingly, the NO instance of 3-COLORABILITY should have the property that, even after removing a small fraction of the edges, it is still not 3-colorable. On the other hand, the strongest known hardness results (Khanna et al., 2000; Guruswami and Khanna, 2004; Dinur et al., 2009; Brakensiek and Guruswami, 2016) on coloring 3-colorable graphs are based on dense graphs; the requirement on NO instance mentioned above, when applied to dense graphs, would make the problem fall into the regime of property testing, which is unfortunately known to be computationally easy (Goldreich et al., 1998).

We circumvent this obstacle by adapting a *sparse graph* construction of Bogdanov, Obata and Trevisan (2002) (referred to as BOT graph henceforth). Based on the hardness of approximating MAX-3SAT, BOT graph was used in Bogdanov et al. (2002) to prove query lower bound for testing

3. Implicitly, we would also like to have $k = r$ so that no non-degenerate parallelepiped of dimension higher than r exists, as the number of size-$(r+1)$ or larger subsets dominates those of size-r subsets; see Conjecture 12 below.

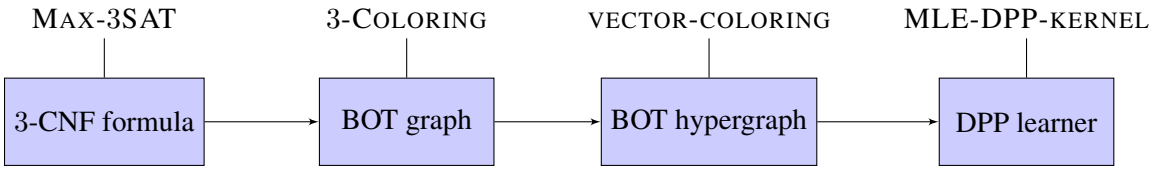

Figure 1: High level overview of our reductions.

3-COLORABILITY in the bounded-degree model. We fix some minor mistakes in the construction and analysis of Bogdanov et al. (2002), further enhance the robustness of BOT graph with the strong expander construction of Alon and Capalbo (2007). These modifications allow us to show that, for some absolute constant δ, we can decode a 3-coloring of the vertices which satisfies at least $(1-\delta)$-fraction of the edges in the original BOT graph, as long as the DPP log likelihood of a training set constructed from the edges of the BOT graph were close enough to the maximum value of a 3-colorable graph. An overview of the reduction sequence is illustrated in Figure 1.

Algorithmic results. For the upper bound, we obtain an approximation algorithm by using some of the properties required for the analysis of our reduction. The algorithm itself is very simple: given a data set $X_1, X_2, \ldots, X_m$, output the DPP marginal kernel given by the $N \times N$ diagonal matrix K such that $K_{ii} = |\{j : X_j \ni i\}|/m$ for all i in the ground set. In other words, the diagonal entries of the marginal kernel are just the empirical probabilities of elements in the data set. Hadamard's inequality gives a lower bound on the optimal likelihood that is similar to the likelihood of the diagonal kernel; if the elements all appear in at most a $a_{\max}/m$-fraction of the subsets, the ratio $\frac{\text{log likelihood output by the algorithm}}{\text{optimal log likelihood}}$ is at most $1 + \frac{\log((1-\frac{a_{\max}}{m})^{1-a_{\max}/m})}{\log((\frac{a_{\max}}{m})^{a_{\max}/m})} \approx 1 + \frac{1}{\log N}$ when $a_{\max}/m$ is $O(1/N)$. For an unconditional upper bound, observe that elements in all m sets should occur with probability 1 in the maximum likelihood DPP (and thus may be disregarded without loss of generality), hence we may plug $a_{max} = m - 1$ into the aforementioned bound and obtain a $(1 + o(1)) \log m$ upper bound on the ratio.

1.3. Related work

Learning DPPs. As mentioned earlier, Urschel et al. (2017) in particular proposed an algorithm for recovering a DPP's kernel up to similarity, which is efficient when (i) the graph defined by interpreting the kernel as a weighted adjacency matrix has a "cycle basis" of cycles of bounded length and (ii) the nonzero entries are not too small. Furthermore, they gave a lower bound on the sample complexity of estimating the DPP kernel, showing that it indeed depends similarly on these quantities. Thus, when there is enough data to permit exact recovery of the kernel, this algorithm will perform well, but otherwise there is no guarantee that the algorithm produces a kernel for a DPP with likelihood close to maximum.

In a companion paper, Brunel et al. (2017a) also studied the rate of estimation obtained by the maximum likelihood kernel. Again, they determined classes of DPPs for which it is efficient (or not). Moreover, they identified an exponential number of saddle points, and conjectured that these are the only critical points; they further suggested that a proof of this conjecture might lead to an efficient algorithm for computing a maximum likelihood kernel. But, they did not actually provide algorithms for computing the likelihood or the kernel itself.

Starting with the pioneering work of Kulesza and Taskar (2011a), various empirical learning algorithms have been proposed for learning discrete DPPs, such as Bayesian methods (Affandi et al., 2014), expectation-maximization (EM) algorithms (Gillenwater et al., 2014), fixed-point iteration (Mariet and Sra, 2015), learning non-symmetric DPPs (Gartrell et al., 2019), learning with negative sampling (Mariet et al., 2019), and minimizing Wasserstein distance (Anquetil et al., 2020). However, none of these algorithms has theoretical guarantees. Efficient learning algorithms have also been designed for restricted classes of DPPs (Mariet and Sra, 2016; Gartrell et al., 2017; Dupuy and Bach, 2018; Osogami et al., 2018). A related problem, namely testing DPPs, recently has been investigated by (Gatmiry et al., 2020).

We note that in contrast to the problem of learning the DPP kernel from a data set as considered here, the problem of computing the mode ("MAP estimate") of a DPP given by its kernel has long been known to be NP-complete (Ko et al., 1995; Civril and Magdon-Ismail, 2009). The inapproximability for this problem was recently strengthened substantially by (Ohsaka, 2021).

Vector coloring problems. The notion of *orthogonal representation* (in which there is an edge between two vertices if and only if the two corresponding representation vectors are orthogonal[4]) was introduced by Lovász (1979), and was used in the definition of the famous Lovász's ϑ function, which has been employed to bound quantities such as Shannon capacity, the clique numbers or the chromatic numbers of graphs. More generally, a *geometric representation* of a graph is a mapping from vertices in V to points in a metric space $\mathcal{M}$, such that two vertices are connected by an edge if and only if the distance between the two corresponding points satisfies certain condition. For example, orthogonal representation is a special case of the *unit-distance* graph where (in the framework of geometric representation) the underlying metric space is the unit sphere (with distance 1 replaced by sphere distance $\pi/2$). Geometric representation of graphs is a well-studied subject, revealing many surprising connections between parameters (e.g. dimension) of geometric representations and properties of the original graph, such as chromatic number, connectivity, excluded subgraphs, tree width, planarity, etc; see e.g. (Lovász, 1979; Lovász et al., 1989; Parsons and Pisanski, 1989; Karger et al., 1998; Lovász and Vesztergombi, 1999; Lovász et al., 2000; Haynes et al., 2010) and the recent textbook (Lovász, 2019).

Matrix completion problem. Geometric representations of graphs are intimated connected to another class of problems, *matrix completion problems*. For instance, the celebrated result of Peeters (1996), showing NP-hardness of deciding whether a 3-dimensional orthogonal representation over a finite field exists for a graph, was obtained through reducing 3-COLORABILITY to a low-rank matrix completion problem. Matrix completion studies under what conditions a partially specified matrix can be completed into one which belongs to a certain class of matrices, such as low-rank matrices, semidefinite matrices, Euclidean distance matrices, etc. See e.g. Laurent (2009) for an overview of the important results in this area. Interestingly, a recent work of Hardt et al. (2014), which proved the hardness of low-rank matrix completion problem under the incoherence assumption (a commonly used assumption for many matrix completion results), was also based on gapped versions of computationally hard problems on graphs such as the r-COLORING problem and the (r, ϵ)-INDEPENDENT-SET problem.[5]

4. Some authors, for example Lovász (1979), define orthogonal representation by mandating the vectors of two non-adjacent vertices to be orthogonal.
5. In this problem, one is given an undirected graph that is promised to be r-colorable and is asked to find an independent set of size ϵn, where $\epsilon < 1/r$ and n is the number of the vertices in the graph.

2. Maximum likelihood learning of DPP and our main hardness result

2.1. Preliminaries

Unless stated otherwise, all logarithms in this paper are to the base e (i.e. natural logarithms). For positive integer n, we write $[n]$ to denote the set $\{1, 2, \ldots, n\}$. For an n-dimensional real vector x, we use $\|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$ to denote the ℓ_2 or Euclidean norm of x.

Matrix analysis. Let A be an $m \times n$ matrix. The $(i,j)^{\text{th}}$ entry of A will be denoted by $A_{i,j}$. All matrices in this paper are over real numbers $\mathbb{R}$; therefore the Hermitian adjoint of A, A^H is the same as $A^\top$, the transpose of A. By the spectral theorem, the eigenvalues of a real, symmetric matrix $M \in \mathbb{R}^{n\times n}$ are all real numbers, and will be denoted $\lambda_1(M) \geq \lambda_2(M) \geq \cdots \geq \lambda_n(M)$. A real, symmetric matrix M is called *positive semidefinite* (PSD) if all its eigenvalues are non-negative (i.e., $\lambda_n(M) \geq 0$). Well-known equivalent characterizations of PSD matrices include $x^\top M x \geq 0$ for all $x \in \mathbb{R}^n$, and the existence of a matrix $Q \in \mathbb{R}^{k\times n}$ for some $k > 0$ such that $M = Q^\top Q$.

A useful variational characterization of the eigenvalues of real, symmetric matrices is the Courant-Fischer theorem, which states that for every $1 \leq k \leq n$ (when a set of vectors whose indices are outside the range $[n]$, then the set is understood to be empty) we have

$$\lambda_k(A) = \min_{x_1,\ldots,x_{k-1}\in\mathbb{R}^n} \max_{\substack{y\neq 0,\, y\in\mathbb{R}^n \\ y\perp x_1,\ldots,x_{k-1}}} \frac{y^\top A y}{y^\top y},$$

and

$$\lambda_k(A) = \max_{x_{k+1},\ldots,x_n\in\mathbb{R}^n} \min_{\substack{y\neq 0,\, y\in\mathbb{R}^n \\ y\perp x_{k+1},\ldots,x_n}} \frac{y^\top A y}{y^\top y}.$$

The *singular values* of a matrix $A \in \mathbb{R}^{m\times n}$ are defined as the (positive) square roots of the eigenvalues of $A^H A = A^\top A$ (a real, symmetric $n \times n$ matrix). Namely, $\sigma_i(A) = \sqrt{\lambda_i(A^\top A)}$, $i = 1, \ldots, n$. We also arrange the singular values of a matrix A in decreasing order, that is $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_n(A)$. The *Frobenius norm* of A, denoted $\|A\|_F$, is defined to be $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |A_{i,j}|^2}$. It is well-known that $\|A\|_F^2 = \sigma_1^2(A) + \cdots + \sigma_n^2(A)$. Finally, the *spectral norm* of a square $n \times n$ matrix A is defined as the square root of the maximum eigenvalue of $A^H A$, i.e.,

$$\|A\|_2 = \sqrt{\lambda_1(A^\top A)} = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_1(A).$$

Discrete determinantal point processes. A *discrete determinantal point process* (DPP) $\mathcal{P}$ over a finite set $\mathcal{X}$ is a probability measure over the set of all subsets of the ground set $\mathcal{X}$. The distribution of $\mathcal{P}$ is specified by a *marginal kernel* $K \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$, which is a positive semidefinite matrix with eigenvalues in $[0, 1]$, in the following manner: if $\boldsymbol{Y} \subseteq \mathcal{X}$ is a random subset drawn according to $\mathcal{P}$, then its probability mass function $\mathcal{P}_K$ is defined such that, for every $S \subseteq \mathcal{X}$,

$$\Pr_{\boldsymbol{Y}\sim\mathcal{P}_K}[S \subseteq \boldsymbol{Y}] = \det(K_S).$$

Here K_S is the principal submatrix of K indexed by $S \subseteq \mathcal{X}$.

If it is the case that all eigenvalues of K are in $[0, 1)$, then $\mathcal{P}$ is called an *L-ensemble*, whose kernel can be defined to be the positive definite[6] matrix $L = K(I - K)^{-1}$. In this case, the corresponding probability mass function, denoted $\mathcal{P}_L$, can be shown to be

$$\Pr_{\boldsymbol{Y} \sim \mathcal{P}_L}[\boldsymbol{Y} = S] = \frac{\det(L_S)}{\det(I + L)},$$

for every $S \subseteq \mathcal{X}$. Hence, $\Pr_{\boldsymbol{Y} \sim \mathcal{P}_L}[\boldsymbol{Y} = \emptyset] = \det(I - K)$, and consequently a DPP is an L-ensemble if and only if the random variable $\boldsymbol{Y} = \emptyset$ with non-zero probability.

2.2. Maximum Likelihood Learning of DPPs

We define the problem of *Maximum Likelihood Learning of DPPs* as follows. A learning algorithm receives a training data set $\{X'_t\}_{t=1}^{T'}$ (viewed as a multiset), typically drawn independently and identically from a distribution D over the subsets of a ground set $\mathcal{X}$. The goal of the learning algorithm is to find a DPP kernel K based on the training set that minimizes[7] the following *maximum log likelihood estimator*

$$\ell(K) = -\frac{1}{T'} \log \prod_{t=1}^{T} \Pr_{\boldsymbol{Y} \sim \mathcal{P}_K}[\boldsymbol{Y} = X_t] = -\frac{1}{T'} \sum_{t=1}^{T} \log \Pr_{\boldsymbol{Y} \sim \mathcal{P}_K}[\boldsymbol{Y} = X_t],$$

where $\{X_1, \ldots, X_T\}$ is the set of distinct elements in the multiset $\{X'_t\}_{t=1}^{T'}$. When the training data set $\{X'_t\}_{t=1}^{T'}$ is clear from context, we simply denote the value of the maximum log likelihood of an optimal DPP kernel by ℓ^*.

One common way to establish the hardness of maximum likelihood learning problems is to show that even computing the maximum value of the log likelihood ℓ^* is hard (if one could efficiently find an optimal DPP kernel K, then since evaluations of determinants can be performed in polynomial time, clearly the corresponding log likelihood ℓ^* would be efficiently computable as well). That is also the approach we take in this paper. In fact, the lower bound actually proved is much stronger: we show that it is NP-hard even to compute a $1 - O(\frac{1}{\log^9 N})$-approximation of ℓ^* for a ground set of size N.

Theorem 4 (Main) *There are infinitely many positive even integers N such that the following holds.*[8] *Let $\mathcal{X} = \{1, 2, \ldots, N\}$. There is a training data set $\{X_t\}_{t=1}^{N/2}$ of size $N/2$, where $X_i \subseteq \mathcal{X}$ for each i, such that it is NP-hard to $\left(1 - O(\frac{1}{\log^9 N})\right)$-approximate the maximum log likelihood value that a DPP kernel can achieve on the training set.*

6. A real, symmetric matrix is called *positive definite* (PD) if all its eigenvalues are positive.
7. We add a negative sign at the front to make the estimator to be always positive, and thus changing the maximization problem into a minimization one. To be consistent, we still call the quantity as a "maximum" likelihood estimator; see Remark 2.
8. As we will see later that $N/2$ is the number of edges in a specially constructed graph. We then construct a 3-uniform hypergraph based on this graph, and add a new node to the hypergraph for each edge in the graph. The ground set consists of these newly added nodes together with the set of vertices in the original graph, hence the cardinality of the ground set is at most N.

2.3. Proof of the Main Theorem: an outline

MAX-3SAT with bounded occurrence. Our starting point is the hardness of MAX-3SAT, in which given a Boolean formula ϕ in 3-CNF form, the goal is to output the maximum number of clauses of ϕ that can be satisfied by any truth assignment of the variables. A classical hardness result is Håstad's 3-bit PCP theorem (Håstad, 2001), which states that it is NP-hard to $(7/8+\epsilon)$-approximate MAX-3SAT for any constant $\epsilon > 0$. However, for our purposes, we need the formula ϕ to have bounded occurrences of any variable. Let MAX-3SAT(k) denote a subclass of MAX-3SAT, in which the instances satisfy that every variable occurs in at most k clauses. Håstad (2000) showed that it is NP-hard to $7/8+1/(\log k)^c$-approximate MAX-3SAT(k) where c is some absolute constant.[9] Therefore,

Lemma 5 (Håstad (2000)) *There is a constant integer k and constant $\epsilon > 0$ (depending only on k) such that it is NP-hard to $(1-\epsilon)$-approximate* MAX-3SAT(k).

That is, for infinitely many integers n, there are two families of instances ϕ_Y and ϕ_N in MAX-3SAT(k) of size n each with the following property: ϕ_Y is satisfiable; every truth assignment can satisfy at most an $1-\epsilon$ fraction of the clauses in ϕ_N; and it is NP-hard to distinguish between the two cases.

3-COLORING for bounded degree graphs. Next, we adapt a gap-preserving reduction of Bogdanov, Obata and Trevisan (2002), which was originally used to prove an $\Omega(n)$ query lower bound for testing 3-Colorability in bounded-degree graphs under the property testing model. On input an instance ϕ of MAX-3SAT(k), the reduction outputs a degree-bounded graph G_ϕ (BOT graph) which satisfies the following: if ϕ is satisfiable then G_ϕ is 3-colorable; and if every truth assignment can satisfy at most $1-\epsilon$ fraction of the clauses in ϕ then every 3-coloring of the vertices of G_ϕ can make at most $1-\epsilon'$ fraction of the edges in G_ϕ non-monochromatic. Here ϵ' is a constant depending only on ϵ and k.

Lemma 6 (Bogdanov et al. (2002)) *There are absolute constants d and $\epsilon' > 0$ such that the following holds. For infinitely many integers n, there are two degree-d bounded graphs $G_{\phi,Y}$ and $G_{\phi,N}$ of size n such that: $G_{\phi,Y}$ is 3-colorable; no $1-\epsilon'$ fraction of the edges of $G_{\phi,N}$ is 3-colorable; and yet it is NP-hard to distinguish between the two cases.*

Very strong expanders. The main idea of the reduction of Bogdanov et al. (2002) is to make k copies of TRUE, FALSE and DUMMY for each variable and its negation, and use an expander to connect these copies together to ensure truth value consistency among different copies. Any constant-degree expander with reasonable vertex-expansion suffices: on one hand, the resulting graph G_ϕ is of bounded degree; on the other hand, by the expansion property, deleting a small fraction of the edges in G_ϕ will still leave the graph with a large connected component, using which, one can — from the coloring of G_ϕ — decode a satisfying assignment that satisfies most of the clauses.

However, for the purpose of proving hardness of DPP Maximum Likelihood Learning, we need the expander to have some additional properties, which are encapsulated in the following definition.

9. We may also use the NP-hardness results of Berman et al. (2003) for 3-SAT instances in which every variable appears exactly 4 times, or assuming $\mathbf{RP} \neq \mathbf{NP}$, use the hardness result of Trevisan (2001) with better parameters.

Definition 7 (Very strong expanders (Alon and Capalbo, 2007)) *A graph $G = (V, E)$ is called a* d-regular very strong expander *on n vertices if the average degree in any subgraph of G on at most $n/10$ vertices is at most $d/6$, and the average degree in any subgraph of G on at most $n/2$ vertices is at most $2d/3$.*

The nice properties of very strong expanders that we require are summarized in the following theorem from Alon and Capalbo (2007).

Theorem 8 (Alon and Capalbo (2007)) *Let $G = (V, E)$ be a very strong d-regular expander on n vertices. If we delete an arbitrary subset of $m' \leq nd/150$ edges from G and denote the resulting graph by G', then G' contains a subgraph H on at least $n - 15m'/d$ vertices and the diameter of H is $O(\log n)$.*

The known *explicit* constructions of Ramanujan graphs (Lubotzky et al., 1988; Margulis, 1988) yield families of d-regular strong expanders on n vertices for infinitely many n's.

3-uniform hypergraph. To obtain the training data set, we transform a BOT graph $G_\phi = (V, E)$ into a 3-uniform hypergraph $H_\phi = (V', E')$ as follows. The vertex set $V'(H_\phi)$ is $V(G) \cup E(G)$, and for notational convenience, we will simply label the "graph-vertex" vertices by a_v for every $v \in V(G)$, and label the "graph-edge" vertices in $V'(H_\phi)$ by $a_{(u,v)}$ for every edge $(u, v) \in E(G)$. Then the set of hyper-edges is $E'(H_\phi) = \{(a_u, a_v, a_{(u,v)}) : (u, v) \in E(G)\}$. It follows that H_ϕ is a 3-uniform hypergraph with[10] $N = |V'(H_\phi)| = n + m$ and $|E'(H_\phi)| = m$, where n and m denote the number of vertices and edges of the BOT graph G_ϕ, respectively.

Now, what happens if we use the set of hyper-edges of H_ϕ as the training data set $\{X_t\}_{t=1}^m$, and feed it to a DPP Maximum Likelihood Learning algorithm? Our first step in understanding the optimal DPP kernel is to establish a connection between DPP kernel of learning BOT hypergraphs and a problem called "vector coloring".

Connecting DPP kernels with vector colorings. Since K is a positive semidefinite matrix, we can write K as $K = Q^\top Q$ for some matrix Q. Let $q_1, \ldots, q_N \in \mathbb{R}^k$ be the columns of Q. We can further decompose these vectors as $q_i = \|q_i\|_2 \chi_i$, where $\chi_i \in \mathbb{R}^k$ is a unit vector. The quantity $\|q_i\|_2$ is a measure of the "importance" of item i, and χ_i is a normalized vector which encodes diversity features of item i. Now the entries of the marginal kernel satisfy $K_{ij} = \|q_i\|_2 \chi_i^\top \chi_j \|q_j\|_2$, where $\chi_i^\top \chi_j \in [-1, 1]$ is a signed measure of the similarity between item i and item j. In particular, the diagonal entries satisfy that $K_{ii} = \|q_i\|_2^2$ for every $i \in [N]$.

We prove the following theorem, which allows us to somewhat decouple $\|q_i\|_2$ and χ_i for each item i, and identify (from the training set) the value of the "importance" (i.e. value $\|q_i\|_2$) for each item.[11] Our result essentially determines at least one of the optimal settings of the importance of each item.

Theorem 9 *Let K be a marginal kernel with likelihood $\ell(K)$. Then there exists a marginal kernel K' with $\ell(K') \leq \ell(K)$ such that the diagonal of K' (indexed by vertices and edges of G_ϕ) satisfies*

$$K'_{ii} = \begin{cases} \frac{\deg_{G_\phi}(u)}{m} & \text{for } i = u \in V(G_\phi); \\ \frac{1}{m} & \text{for } i = (u, v) \in E(G_\phi). \end{cases}$$

10. In order to not overload the statement in Theorem 4 with multiple parameters, we may add isolated vertices to graph G_ϕ so that $n = m$ and hence the ground set size is N and the sample size is $m = N/2$.
11. It is no coincidence that our simple algorithm, using this "first-moment" information from the training set in a similar manner, constructs its DPP that achieves nontrivial worst-case approximation to the optimal log likelihood.

Thus, it remains to determine the diversity features that maximize the likelihood. To this end, we use a variant of 3-colorability in which the "colors" are generalized to vectors in $\mathbb{R}^k$ and the "coloring constraint" for an edge (i, j) is satisfied if the vectors χ_i and χ_j assigned to the vertices i and j are orthogonal.

More formally, let S^{k-1} be the unit sphere in k-dimensional Euclidean space; that is, $S^{k-1} = \{x \in \mathbb{R}^k : \|x\|_2 = 1\}$. Given a graph $G = (V, E)$, we define a *vector k-coloring* of G to be a function $\chi : V(G) \to S^{k-1}$. We say that a vector k-coloring χ is *orthogonal* if, for every edge $(u, v) \in E(G)$, we have $\chi_u^\top \chi_v = 0$. We define the *error* of a vector k-coloring χ of G to be

$$\mathrm{err}_\chi(G) := \frac{1}{|E(G)|} \sum_{(u,v)\in E(G)} |\chi_u^\top \chi_v|^2$$

so that a vector k-coloring χ of G is orthogonal if and only if $\mathrm{err}_\chi(G) = 0$. Since χ_u and χ_v are unit vectors, $|\chi_u^\top \chi_v| = |\cos \theta_{uv}|$ for all $(u, v) \in E(G)$, where θ_{uv} is the angle between χ_u and χ_v.

Now, by combining Theorem 9 with the fact that any 3-coloring of G naturally induces a vector 3-coloring of G, it is not hard to prove the following "completeness" theorem.

Theorem 10 (Completeness theorem) *Let G_ϕ be a BOT graph, and let $n = |V(G_\phi)|$ be the number of vertices and $m = |E(G_\phi)|$ be the number of edges of G_ϕ, respectively. If ϕ is satisfiable, then there exists a rank-3 DPP marginal kernel K such that*

$$\ell(K) = \ell^* = 3 \log m - \frac{1}{m} \sum_{(u,v)\in E(G_\phi)} \Big(\log(\deg_{G_\phi}(u)) + \log(\deg_{G_\phi}(v)) \Big) .$$

Projecting DPP kernels to $\mathbb{R}^3$. Intuitively, the maximum likelihood marginal kernel has dimension 3 so that zero probability measure will be assigned for subsets of size at least 4. We were unable to prove this, but we nevertheless manage to show that the loss in making such an assumption is not too great:

Theorem 11 *Let G_ϕ be a BOT graph with maximum degree at most k. There is a constant C_k depending only on k such that the following holds. Let K be an optimal marginal kernel with likelihood $\ell(K) \le \ell^* + \delta$ for some $0 < \delta \le 1/(128k)^2$, then there exists a marginal kernel K' of dimension 3 such that $\ell(K') \le \ell^* + C_k \delta^{1/4}$.*

We conjecture that an even stronger statement is in fact true.

Conjecture 12 (Cardinality-rank conjecture) *If the cardinality of every subset in a training set is at most $k \ge 1$, then every optimal maximum likelihood marginal kernel for the training set has dimensional at most k.*

This conjecture may be of independent interest outside the realm of maximum likelihood learning of DPPs.

Decoding truth assignments from vector colorings. Because of Theorem 11, from now on we assume that the dimension of Q is 3. Therefore, for each $(a_u, a_v, a_{(u,v)}) \in E'(H_\phi)$,

$$\Pr_{\boldsymbol{Y}\sim\mathcal{P}_K}[\boldsymbol{Y} = \{a_u, a_v, a_{(u,v)}\}] = \det(K_{\{a_u,a_v,a_{(u,v)}\}}),$$

where $K = Q^\top Q$ is the marginal kernel of DPP $\mathcal{P}$. To maximize the likelihood, we want to maximize the product of determinants of the above form. Since each $a_{(u,v)}$ occurs in only one example, we can assume that $\chi_{a_{(u,v)}}$ is always taken to be orthogonal to χ_{a_u} and χ_{a_v}. Thus, the likelihood contribution from $(a_u, a_v, a_{(u,v)})$ is maximized when $\chi_{a_u}^\top \chi_{a_v} = 0$; or equivalently, when χ_{a_u} and χ_{a_v} are orthogonal. We formally prove this correspondence between the "orthogonality" of the associated vector 3-coloring of G_ϕ and the likelihood of the corresponding DPP with marginal kernel $K = Q^\top Q$, where the column of Q corresponding to vertex u is $\|q_{a_u}\|_2 \chi_{a_u}$. Moreover, we can even decode the truth-assignment of the Boolean formula ϕ if the vector 3-coloring of G_ϕ is very close to satisfying all edges of G_ϕ.

Theorem 13 (Soundness theorem) *Let ℓ^* be the optimal log likelihood as in Theorem 10. Then there exists a constant $C > 0$ which depends only on k and ϵ' as those defined in Theorem 6, such that the following holds. If there is a DPP marginal kernel K of rank 3, which satisfies $\ell(K) \le \ell^* + \frac{C}{\log^2 n}$ where $n = |V(G_\phi)|$ is the number of vertices in the BOT graph, then there is a truth assignment that satisfies at least $(1-\epsilon)$ fraction of the clauses in ϕ, where ϵ is the constant defined in Theorem 5.*

To see why Theorem 13 implies our main theorem, Theorem 4, consider that we start the reduction with two families of instances ϕ_Y and ϕ_N in MAX-3SAT(k) which are NP-hard to distinguish. Then we construct their corresponding BOT hypergraphs, H_{ϕ_Y} and H_{ϕ_N}, and use the edge sets of these two hypergraphs as training sets of size m for a DPP maximum likelihood learning algorithm. The log likelihood estimator of ϕ_Y is $\ell^* = \Theta(\log N)$ by Theorem 10. What is the log likelihood estimator of ϕ_N? Well, by Theorem 11, if the marginal kernel of G_{ϕ_N} has log likelihood $\ell(K) \le \ell^* + \delta$ for some small enough $\delta > 0$, then there exists a marginal kernel K' of dimension 3 such that $\ell(K') \le \ell^* + C_k \delta^{1/4}$. By Theorem 13, we must have $\delta = \Omega(\frac{1}{\log^2 N})$ for ϕ_N. That is, the log likelihood estimator of ϕ_N is $\ell^* + \Omega(\frac{1}{\log^8 N})$. Now if there were an polynomial-time algorithm $\mathcal{A}$ which approximates the log likelihood estimator within a factor of $1 - 1/\Omega(\log^9 N)$, then $\mathcal{A}$ would be able to tell apart ϕ_Y from ϕ_N — thus solving an NP-complete problem — simply by approximating the maximum log likelihood estimators on training data sets from H_{ϕ_Y} and H_{ϕ_N}, respectively.

3. Discussion and open problems

In this work, we establish that it is NP-hard to obtain a $1 - O(\frac{1}{\log^9 N})$-approximation to the maximum log likelihood of DPPs. We also demonstrate a simple polynomial-time algorithm that achieves $\frac{1}{(1+o(1))\log m}$-approximation. One immediate open problem is to close this large gap. A natural and plausible approach is to prove the cardinality-rank conjecture or at least to improve the bound in Theorem 11. Note that our hardness result does not rule out efficient learning with some constant factor of approximation: it is still possible that there is a polynomial-time algorithm that obtains a DPP kernel with a 99%-approximation to the maximum log likelihood. As observed earlier, we cannot preclude constant-factor approximations by a better analysis of the constructed hard instance: the approximation algorithm shows that our hardness result is tight up to a polynomial factor for the type of subset collections employed in the proof. Therefore any stronger hardness proof would require constructing a collection of subsets in which some element appears in a non-trivial fraction of the subsets.

Our investigation just takes a first stab at understanding the computational landscape of learning DPPs. In particular, our knowledge for the complexity of learning DPPs when the data set is indeed generated by an unknown DPP is still very limited: Can one design efficient algorithms for such a task? Can the DPP kernel be learned with arbitrary accuracy? And if not, what is the best approximation factor can such an algorithm achieve? Note that the data here is no longer a worst-case *data set*, but only sampled from a worst-case *DPP*. The underlying model is thus a semi-random one, and it seems challenging to extend NP-completeness hardness to such settings; some kind of *average-case hardness* is likely to be the best one can hope for. This is essentially the approach that Brunel et al. (2017a) had envisioned when examining the optimization landscape of the likelihood function for DPP kernels. The convergence of the empirical log-likelihood function to the true log-likelihood function only holds with high probability, and so in particular doesn't carry over to the kinds of worst-case data sets produced by our reduction. Thus, their conjectured property may still hold, and may be a route to an efficient algorithm in this setting. On the other hand, "realizability" is probably a too strong assumption for practical purposes; DPPs are generally used to model processes featuring negative association, and it often seems implausible that the data actually follows a DPP distribution. Therefore, finding more appropriate assumptions is yet to be explored, and an algorithm for a realizable setting would be a natural first step along this direction.

As the other side of the coin, it is entirely conceivable that such efficient algorithms may not exist at all. One may view our main result as proving the hardness of "agnostic-learning" DPPs, while here the task would be proving hardness of "PAC-learning" DPPs. Presumably this is more difficult, as PAC-learning is in general easier than agnostic learning, it is thus harder to obtain lower bounds in the former setting. In particular, the usual approach of proving PAC-learning lower bounds involves uniform distributions over some prescribed collections of subsets. Such distributions are within the scope of PAC-learning model as it allows arbitrary distributions. By contrast, DPPs are normally unable to generate the uniform distribution over an arbitrary collection of subsets. Indeed, we believe that the data set we construct would be atypical for all DPPs. This is why contrary to the usual representation-specific hardness theorems in PAC-learning, we believe that an average-case hardness assumption will be necessary here.

Indeed, this discussion also highlights that maximum likelihood estimation is a kind of representation-specific (or "proper") learning. One more potential route to learning DPPs is thus to use a more expressive representation that may possibly achieve better likelihood than the best DPP for a given data set, thwarting our reduction. Probabilistic Generating Circuits (Zhang et al., 2021), for example, can efficiently represent DPPs, and therefore may be an appropriate representation to use for this purpose.

Acknowledgments

We thank the anonymous reviewers for carefully reading the manuscript and providing useful comments and suggestions. B.J. was partially supported by NSF awards IIS-1908287, IIS-1939677, and CCF-1718380. E.G. was partially supported by NSF CCF-1910659 and NSF CCF-1910411. N.X. was partially supported by U.S. Army Research Office (ARO) under award number W911NF1910362.

References

Raja Hafiz Affandi, Emily B. Fox, and Ben Taskar. Approximate inference in continuous determinantal point processes. In *Advances in Neural Information Processing Systems 25*, pages 1430–1438, 2013a.

Raja Hafiz Affandi, Alex Kulesza, Emily Fox, and Ben Taskar. Nyström approximation for large-scale determinantal processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *PMLR*, pages 85–98, 2013b.

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

Noga Alon and Michael Capalbo. Finding disjoint paths in expanders deterministically and online. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 518–524. IEEE, 2007.

Nima Anari, Shayan Oveis Gharan, and Alireza Rezaei. Monte Carlo Markov chain algorithms for sampling strongly Rayleigh distributions and determinantal point processes. In *Conference on Learning Theory*, pages 103–115. PMLR, 2016.

Lucas Anquetil, Mike Gartrell, Alain Rakotomamonjy, Ugo Tanielian, and Clément Calauzènes. Wasserstein learning of determinantal point processes. *arXiv preprint arXiv:2011.09712*, 2020.

Yannick Baraud. Estimation of the density of a determinantal process. *Confluentes Mathematici*, 5 (1):3–21, 2013.

Piotr Berman, Marek Karpinski, and Alex D. Scott. Approximation hardness and satisfiability of bounded occurrence instances of SAT. *Electronic Colloquium in Computational Complexity*, TR03-022, 2003.

Andrej Bogdanov, Kenji Obata, and Luca Trevisan. A lower bound for testing 3-colorability in bounded-degree graphs. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pages 93–102. IEEE, 2002.

Joshua Brakensiek and Venkatesan Guruswami. New hardness results for graph and hypergraph colorings. In *31st Conference on Computational Complexity (CCC 2016)*, 2016.

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

Victor-Emmanuel Brunel, Ankur Moitra, Philippe Rigollet, and John Urschel. Rates of estimation for determinantal point processes. In *Conference on Learning Theory*, pages 343–345, 2017b.

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

Wei-Lun Chao, Boqing Gong, Kristen Grauman, and Fei Sha. Large-margin determinantal point processes. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 191–200, 2015.

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

Michał Dereziński and Manfred K. Warmuth. Reverse iterative volume sampling for linear regression. *The Journal of Machine Learning Research*, 19(1):853–891, 2018.

Michał Dereziński, Daniele Calandriello, and Michal Valko. Exact sampling of determinantal point processes with sublinear time preprocessing. In *Advances in Neural Information Processing Systems 32*, 2019.

Michal Derezinski, Feynman T. Liang, and Michael W. Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. *Advances in Neural Information Processing Systems 33*, 2020.

Irit Dinur, Elchanan Mossel, and Oded Regev. Conditional hardness for approximate coloring. *SIAM Journal on Computing*, 39(3):843–873, 2009.

Shaddin Dughmi, Tim Roughgarden, and Mukund Sundararajan. Revenue submodularity. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 243–252, 2009.

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

Freeman J. Dyson. Statistical theory of the energy levels of complex systems. III. *Journal of Mathematical Physics*, 3(1):166–175, 1962.

Mike Gartrell, Ulrich Paquet, and Noam Koenigstein. Bayesian low-rank determinantal point processes. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 349–356, 2016.

Mike Gartrell, Ulrich Paquet, and Noam Koenigstein. Low-rank factorization of determinantal point processes. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Mike Gartrell, Victor-Emmanuel Brunel, Elvis Dohmatob, and Syrine Krichene. Learning nonsymmetric determinantal point processes. *arXiv preprint arXiv:1905.12962*, 2019.

Khashayar Gatmiry, Maryam Aliakbarpour, and Stefanie Jegelka. Testing determinantal point processes. *arXiv preprint arXiv:2008.03650*, 2020.

Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. Discovering diverse and salient threads in document collections. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 710–720, 2012a.

Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. Near-optimal map inference for determinantal point processes. In *Advances in Neural Information Processing Systems 24*, pages 2744–2752, 2012b.

Jennifer A Gillenwater, Alex Kulesza, Emily Fox, and Ben Taskar. Expectation-maximization for learning determinantal point processes. *Advances in Neural Information Processing Systems*, 27: 3149–3157, 2014.

Oded Goldreich, Shari Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM (JACM)*, 45(4):653–750, 1998.

Carlos Guestrin, Andreas Krause, and Ajit Paul Singh. Near-optimal sensor placements in Gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 265–272, 2005.

Venkatesan Guruswami and Sanjeev Khanna. On the hardness of 4-coloring a 3-colorable graph. *SIAM Journal on Discrete Mathematics*, 18(1):30–40, 2004.

Venkatesan Guruswami and Ali Kemal Sinop. Optimal column-based low-rank matrix reconstruction. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1207–1214. SIAM, 2012.

Moritz Hardt, Raghu Meka, Prasad Raghavendra, and Benjamin Weitz. Computational limits for matrix completion. In *Conference on Learning Theory*, pages 703–725. PMLR, 2014.

Johan Håstad. On bounded occurrence constraint satisfaction. *Information Processing Letters*, 74 (1-2):1–6, 2000.

Johan Håstad. Some optimal inapproximability results. *Journal of the ACM*, 48(4):798–859, 2001.

Gerald Haynes, Catherine Park, Amanda Schaeffer, Jordan Webster, and Lon H Mitchell. Orthogonal vector coloring. *The Electronic Journal of Combinatorics*, 17, 2010.

J. Ben Hough, Manjunath Krishnapur, Yuval Peres, and Bálint Virág. Determinantal processes and independence. *Probability surveys*, 3:206–229, 2006.

Byungkon Kang. Fast determinantal point process sampling with application to clustering. In *Advances in Neural Information Processing Systems 25*, pages 2319–2327, 2013.

David R. Karger, Rajeev Motwani, and Madhu Sudan. Approximate graph coloring by semidefinite programming. *J. ACM*, 45(2):246–265, 1998.

Sanjeev Khanna, Nathan Linial, and Shmuel Safra. On the hardness of approximating the chromatic number. *Combinatorica*, 20(3):393–415, 2000.

Chun-Wa Ko, Jon Lee, and Maurice Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691, 1995.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(2), 2008.

Alex Kulesza and Ben Taskar. Structured determinantal point processes. *Advances in neural information processing systems 23*, pages 1171–1179, 2010.

Alex Kulesza and Ben Taskar. Learning determinantal point processes. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 419–427, 2011a.

Alex Kulesza and Ben Taskar. k-DPPs: fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1193–1200, 2011b.

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

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

Claire Launay, Bruno Galerne, and Agnès Desolneux. Exact sampling of determinantal point processes without eigendecomposition. *Journal of Applied Probability*, 57(4):1198–1221, 2020.

Monique Laurent. Matrix completion problems. *Encyclopedia of Optimization*, 3:221–229, 2009.

Frédéric Lavancier, Jesper Møller, and Ege Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77 (4):853–877, 2015.

Donghoon Lee, Geonho Cha, Ming-Hsuan Yang, and Songhwai Oh. Individualness and determinantal point processes for pedestrian detection. In *European Conference on Computer Vision*, pages 330–346. Springer, 2016.

Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Fast mixing Markov chains for strongly Rayleigh measures, DPPs, and constrained sampling. In *Advances in Neural Information Processing Systems 29*, pages 4195–4203, 2016a.

Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Fast DPP sampling for Nyström with application to kernel methods. In *International Conference on Machine Learning*, pages 2061–2070. PMLR, 2016b.

Hui Lin and Jeff Bilmes. Learning mixtures of submodular shells with application to document summarization. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 479–490, 2012.

L. Lovász, M. Saks, and A. Schrijver. Orthogonal representations and connectivity of graphs. *Linear Algebra and its Applications*, 114-115:439–454, 1989.

L. Lovász, M. Saks, and A. Schrijver. A correction: orthogonal representations and connectivity of graphs. *Linear Algebra and its Applications*, 313(1):101–105, 2000.

László Lovász. On the Shannon capacity of a graph. *IEEE Trans. Inf. Theory*, 25(1):1–7, 1979.

László Lovász. *Graphs and geometry*, volume 65. American Mathematical Soc., 2019.

László Lovász and Katalin Vesztergombi. Geometric representations of graphs. *Paul Erdos and his Mathematics*, 2, 1999.

Alexander Lubotzky, Ralph Phillips, and Peter Sarnak. Ramanujan graphs. *Combinatorica*, 8(3): 261–277, 1988.

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

Grigorii Aleksandrovich Margulis. Explicit group-theoretical constructions of combinatorial schemes and their application to the design of expanders and concentrators. *Problemy peredachi informatsii*, 24(1):51–60, 1988.

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

Zelda Mariet, Mike Gartrell, and Suvrit Sra. Learning determinantal point processes by corrective negative sampling. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2251–2260, 2019.

Zelda E Mariet and Suvrit Sra. Kronecker determinantal point processes. *Advances in Neural Information Processing Systems*, 29:2694–2702, 2016.

Naoto Ohsaka. Unconstrained map inference, exponentiated determinantal point processes, and exponential inapproximability. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 154–162, 2021.

Takayuki Osogami and Rudy Raymond. Determinantal reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33(1), pages 4659–4666, 2019.

Takayuki Osogami, Rudy Raymond, Akshay Goel, Tomoyuki Shirai, and Takanori Maehara. Dynamic determinantal point processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32(1), 2018.

T.D. Parsons and Tomaž Pisanski. Vector representations of graphs. *Discrete Mathematics*, 78(1): 143–154, 1989.

René Peeters. Orthogonal representations over finite fields and the chromatic number of graphs. *Combinatorica*, 16(3):417–431, 1996.

Leonard Pitt and Leslie G. Valiant. Computational limitations on learning from examples. *Journal of the ACM (JACM)*, 35(4):965–984, 1988.

Patrick Rebeschini and Amin Karbasi. Fast mixing for discrete point processes. In *Conference on Learning Theory*, pages 1480–1500. PMLR, 2015.

Zeév Rudnick and Peter Sarnak. Zeros of principal L-functions and random matrix theory. *Duke Mathematical Journal*, 81(2):269–322, 1996.

Amar Shah and Zoubin Ghahramani. Determinantal clustering process-a nonparametric bayesian approach to kernel based semi-supervised clustering. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 566–575, 2013.

Jasper Snoek, Richard Zemel, and Ryan Prescott Adams. A determinantal point process latent variable model for inhibition in neural spiking data. *Advances in Neural Information Processing Systems 25*, 2013.

Alexander Soshnikov. Determinantal random point fields. *Russian Mathematical Surveys*, 55(5): 923, 2000.

Luca Trevisan. Non-approximability results for optimization problems on bounded degree instances. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 453–461, 2001.

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

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

Yaodong Yang, Ying Wen, Jun Wang, Liheng Chen, Kun Shao, David Mguni, and Weinan Zhang. Multi-agent determinantal q-learning. In *International Conference on Machine Learning*, pages 10757–10766. PMLR, 2020.

Jin-ge Yao, Feifan Fan, Wayne Xin Zhao, Xiaojun Wan, Edward Chang, and Jianguo Xiao. Tweet timeline generation with determinantal point processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30(1), 2016.

Honghua Zhang, Brendan Juba, and Guy Van Den Broeck. Probabilistic generating circuits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12447–12457. PMLR, 2021.

Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.

James Y. Zou and Ryan P. Adams. Priors for diversity in generative latent variable models. In *Advances in Neural Information Processing Systems 24*, pages 2996–3004, 2012.