Electronic Colloquium on Computational Complexity, Report No. 83 (2022)

Hardness of Maximum Likelihood Learning of DPPs

Elena Grigorescu* Brendan Juba† Karl Wimmer‡ Ning Xie§

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 (2011) that the problem is NP-complete. The lack of a formal proof prompted Brunel, Moitra, Rigollet and Urschel (COLT 2017) 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 $\left(1 - O(\frac{1}{\log^9 N})\right)$-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 (FOCS 2002), and is further enhanced by the strong expanders of Alon and Capalbo (FOCS 2007) to serve our purposes.

*Purdue University, elena-g@purdue.edu
†Washington University in St. Louis, bjuba@wustl.edu
‡Duquesne University, wimmerk@duq.edu
§Florida International University, nxie@cis.fiu.edu

ISSN 1433-8092

Contents

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 [Mac75], but they also arise naturally in diverse fields such as random matrix theory, probability theory, number theory, random spanning trees and non-intersecting paths [Dys62, BP93, RS96, Sos00]. After the seminal work of [KT12], 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 [HKPV06, KT10, RK15, LJS16a, LJS16b, AGR16, DCV19, LGD20]. Furthermore, DPPs have since found a vast variety of applications throughout machine learning and data analysis, including text and image search, segmentation and summarization [LB12, KT12, ZA12, GKT12b, GKT12a, YFZ+16, KT11a, AFAT14, LCYO16, AKFT13, CGGS15, AFT13], signal processing [XO16, KSG08, GKS05], clustering [ZA12, Kan13, SG13], recommendation systems [ZKL+10], revenue maximization [DRS09], multi-agent reinforcement learning [OR19, YWW+20], modeling neural spikes [SZA13], sketching for linear regression [DW18, DLM20], low-rank approximation [GS12], 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 [KT12], 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. [BMRU17b] 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 [GKFT14], fixed point algorithms [MS15], and MCMC [AFAT14]. In the continuous case, the problem has been studied under strong parametric assumptions [LMR15], or smoothness assumptions [Bar13].

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 [KT12, AFAT14, GPK16, MS16, GPK17, UBMR17, DB18], while the other invokes heuristics to maximize the likelihood in an unconstrained search over the parameter

space [KT11b, GKFT14, AFAT14, MS15]. 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 [KT11b, Kul12] 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 sample set.*

Remark 1. *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 [PV88], 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.

Theorem 2 (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 [UBMR17], 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

[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.

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 [Kul12] 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 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).

[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.

[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 1 below.

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 [BMRU17a], 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 [KLS00, GK04, DMR09, BG16] 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 [GGR98].

We circumvent this obstacle by adapting a *sparse graph* construction of Bogdanov, Obata and Trevisan [BOT02] (referred to as BOT graph henceforth). Based on the hardness of approximating Max-3SAT, BOT graph was used in [BOT02] to prove query lower bound for testing 3-Colorability in the bounded-degree model. We fix some minor mistakes in the construction and analysis of [BOT02], further enhance the robustness of BOT graph with the strong expander construction of Alon and Capalbo [AC07]. 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

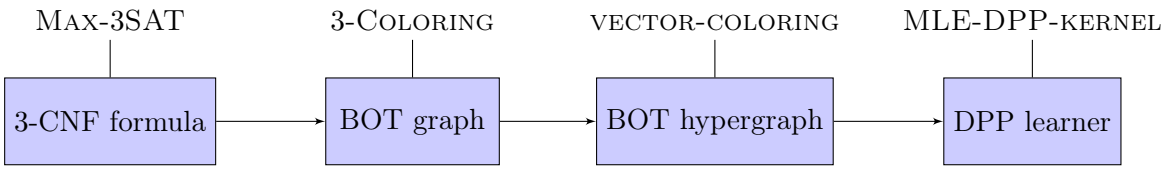


Figure 1: High level overview of our reductions.

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, [UBMR17] 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, [BMRU17a] 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 [KT11b], various empirical learning algorithms have been proposed for learning discrete DPPs, such as Bayesian methods [AFAT14], expectation-maximization (EM) algorithms [GKFT14], fixed-point iteration [MS15], learning non-symmetric DPPs [GBDK19], learning with negative sampling [MGS19], and minimizing Wasserstein distance [AGR$^+$20]. However, none of these algorithms has theoretical guarantees. Efficient learning algorithms have also been designed for restricted classes of DPPs [MS16, GPK17, DB18, ORG$^+$18]. A related problem, namely testing DPPs, recently has been investigated by [GAJ20].

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 [KLQ95, CMI09]. The inapproximability for this problem was recently strengthened substantially by [Ohs21].

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 [Lov79], 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. [Lov79, LSS89, PP89, KMS98, LV99, LSS00, HPS$^+$10] and the recent textbook [Lov19].

Matrix completion problem. Geometric representations of graphs are intimated connected to another class of problems, *matrix completion problems*. For instance, the celebrated result of [Pee96], 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. [Lau09] for an overview of the important results in this area. Interestingly, a recent work of [HMRW14], 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]

1.4 Organization of the paper

The rest of the paper is organized as follows. In Section 2 we summarize notation, definitions and theorems to be used later, define formally the problem of maximum likelihood learning of DPPs and state our main theorem on hardness of MLE learning DPPs. Then we outline the proof of the main theorem, while deferring some technical proofs to later sections. Specifically, in Section 3, we provide a detailed description of our slightly modified BOT graph construction and prove some useful properties of these graphs. In Section 4, we show the following: given any training set, at least one of its optimal DPP kernels satisfy that their diagonals are equal to the empirical frequencies of elements in the ground set. In Section 5, we explicitly construct a rank-3 optimal kernel for 3-colorable BOT graphs. For general BOT graphs, we further prove in Section 6 that one may restrict to optimizing over rank-3 DPP kernels only, without sacrificing too much in likelihood. Then in Section 7 we put these pieces together and prove the soundness theorem, which completes the proof of our main NP-hardness of DPP learning theorem. In Section 8 we present and analyze our simple approximation algorithm. Finally, we conclude in Section 9 with discussions and open problems.

[4]Some authors, for example [Lov79], 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 *Maximum Likelihood Learning of DPPs* problem as follows. A learning algorithm receives a training data sample $\{X'_t\}_{t=1}^{T'}$ (viewed as a multiset) 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 sample so that to minimize[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 sample $\{X'_t\}_{t=1}^{T'}$ 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 3 (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 sample $\{X_t\}_{t=1}^{N/2}$ 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.*

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

[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 1.

[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.

clauses of ϕ that can be satisfied by any truth assignment of the variables. A classical hardness result is the Håstad's 3-bit PCP theorem [Hås01], which states that it is NP-hard to $(7/8+\epsilon)$-approximate MAX-3SAT for any constant $\epsilon > 0$. However, for our purpose, we need the formula ϕ to have bounded occurrence of any variable. Let MAX-3SAT(k) to denote a subclass of MAX-3SAT, in which the instances satisfy that every variable occurs in at most k clauses. [Hås00] 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 1 ([Hås00]). *There are 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 [BOT02], 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 2 ([BOT02]). *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 [BOT02] 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 our 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.

Definition 1 (Very strong expanders [AC07]). *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 [AC07].

[9] We may also use the NP-hardness results of [BKS03] for 3-SAT instance in which every variable appears exactly 4 times, or assuming **RP** $\neq$ **NP** and use the hardness result of [Tre01] with better parameters.

Theorem 4 ([AC07]). *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 [LPS88, Mar88] yield families of d-regular strong expanders on n vertices for infinitely many n's.

3-uniform hypergraph. To obtain the training data sample, 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 sample $\{X_t\}_{t=1}^m$, 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 5. *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*

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

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.

[10] In order to not overload the statement in Theorem 3 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 (see Section 8 for details), 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.

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 5 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 6 (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 7. *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 1 (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 7, 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 sample, 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 8 (Soundness theorem)**.** *Let ℓ^* be the optimal log likelihood as in Theorem 6. Then there exists a constant $C > 0$ which depends only on k and ϵ' as those defined in Theorem 2, such that the following holds. If there is a DPP marginal kernel K of rank 3, which satisfies $\ell(K) \leq \ell^* + \frac{C}{\log^2 n}$ 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 1.*

To see why Theorem 8 implies our main theorem, Theorem 3, 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 6. What is the log likelihood estimator of ϕ_N? Well, by Theorem 7, if the marginal kernel of G_{ϕ_N} has log likelihood $\ell(K) \leq \ell^* + \delta$ for some small enough $\delta > 0$, then there exists a marginal kernel K' of dimension 3 such that $\ell(K') \leq \ell^* + C_k \delta^{1/4}$. By Theorem 8, 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 sample from H_{ϕ_Y} and H_{ϕ_N}, respectively.

3 Robustness of very strong expanders – Proof of Lemma 2

3.1 BOT graphs

We now provide a detailed description and analysis of the construction in [BOT02], as our hardness proof relies crucially on the properties of BOT graphs.

Recall that in the classic reduction from 3-SAT to 3-Coloring, there are three designated nodes True, False and Dummy which form a triangle so that in any 3-coloring of the graph they must be assigned to different colors. The reduction then connects the nodes in literal gadgets and clause gadgets to these three nodes so that every literal is connected to the Dummy node, and any 3-coloring of the graph can be decoded into a satisfying assignment simply by reading the color of every literal node. On the other hand, any satisfying assignment of the variables can also be transformed into a valid 3-coloring of the graph by assigning "True" literal nodes the same color as the True node and assigning "False" literal nodes the same color as the False node.

The basic idea of the construction in [BOT02] is to make k "local" copies for each literal node as well as for the three designated True, False and Dummy nodes, so that each local copy of these nodes used only *once* in the clause gadgets. The benefit of doing so is that deleting any single node or any single edge will affect (or "destroy") a single clause, hence makes the reduction from 3-SAT to 3-Coloring (up to a constant factor) distance-preserving. However, this introduces a new difficulty: we need to make sure that all k copies of literal nodes (True, False and Dummy nodes) should be

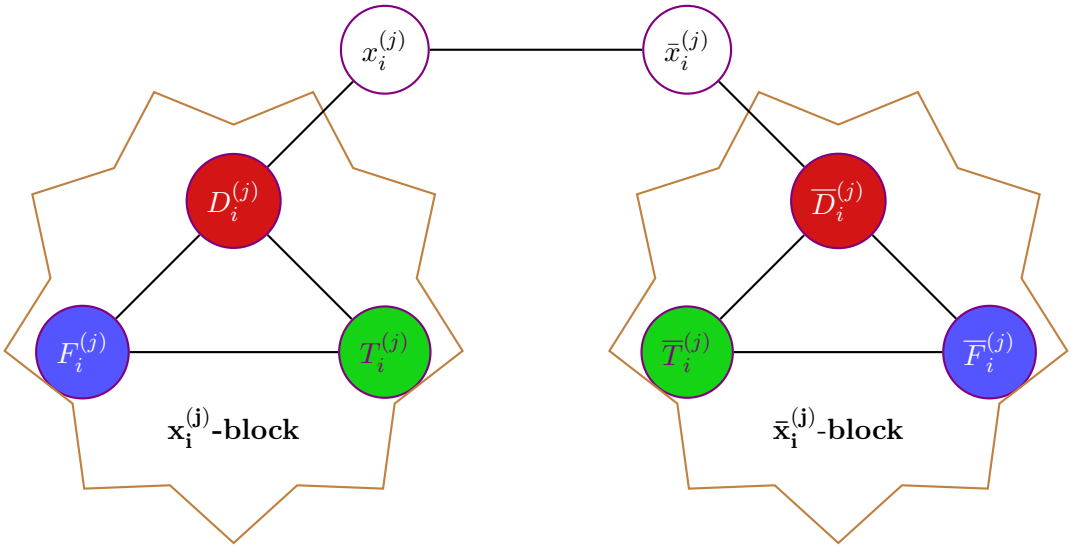
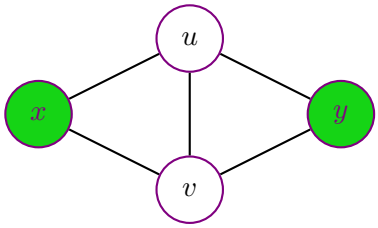
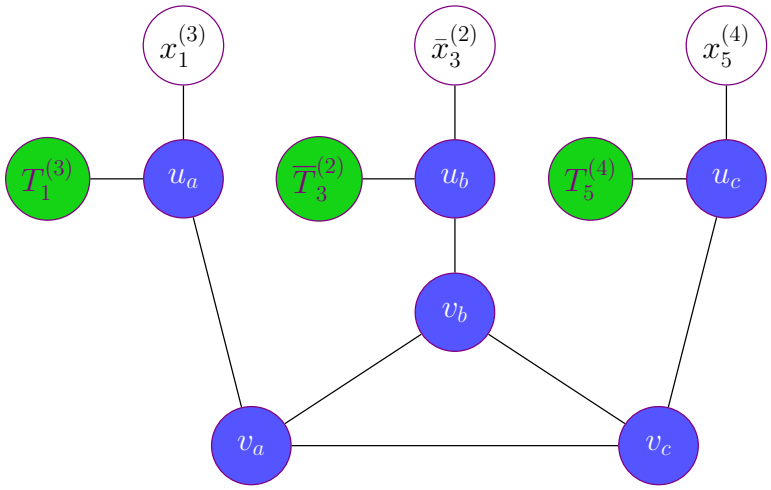


Figure 2: Literal blocks. This gadget enforces that the j^{th} copy of literal x_i and $\bar{x}_i$ will always be assigned opposite truth values, as long as the assignments of TRUE, FALSE and DUMMY nodes in $x_i^{(j)}$-block and $\bar{x}_i^{(j)}$-block are consistent with their corresponding nodes in the rest of the graph.

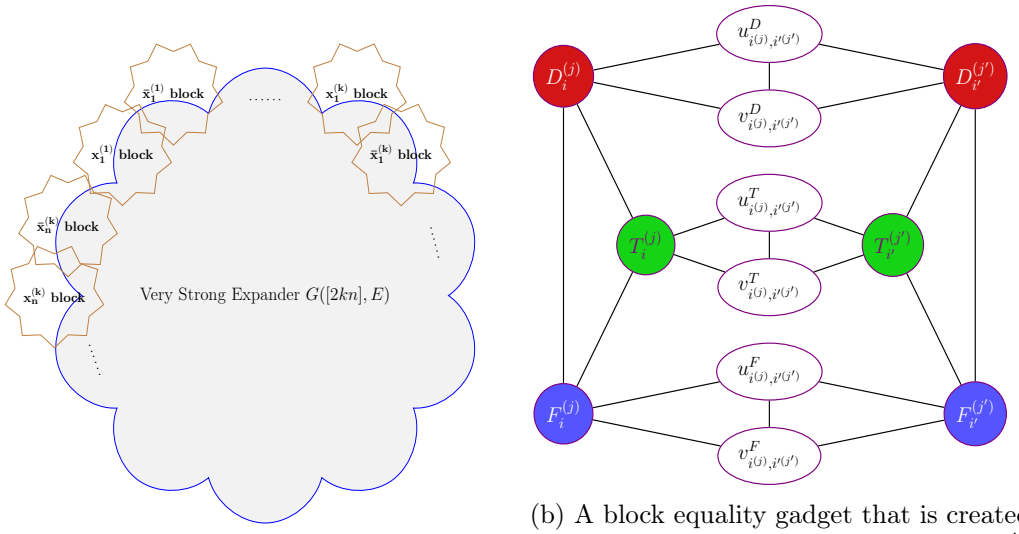


(a) The equality gadget for $x = y$.



(b) The clause gadget for a clause $(\ell_a \vee \ell_b \vee \ell_c)$. For concreteness, we illustrate the construction with $\ell_a = x_1$ (using the 3rd copy), $\ell_b = \bar{x}_3$ (using the 2nd copy) and $\ell_c = x_5$ (using the 4th copy).

Figure 3: Two basic gadgets of BOT graphs.



(a) The underlying expander with literal blocks as vertices.

(b) A block equality gadget that is created in step 6 if $x_i^{(j)}$-block is connected with $x_{i'}^{(j')}$-block in the underlying expander.

Figure 4: The expander part of BOT graph.

assigned with the same color. Such consistency is ensured by embedding these local copies on an expander so that the graph is robust against deletions of a small number of edges or nodes, and at the same time the blowup in the maximum degree is bounded.

Specifically, given a Boolean formula ϕ in 3-CNF form, in which ϕ has n variables, m clauses and each literal appears in at most k clauses, construct an undirected graph $G_\phi(V, E)$ as follows:

1. Create $2kn$ literal nodes: for each variable x_i, $1 \le i \le n$, create k copies of it $x_i^{(j)}$, and another k copies of its negation $\bar{x}_i^{(j)}$, where $j = 1, \ldots, k$;

2. For each of the $2kn$ literal nodes $x_i^{(j)}$ (resp. $\bar{x}_i^{(j)}$), create a block — called $x_i^{(j)}$-*block* (resp. $\bar{x}_i^{(j)}$-*block*) — of 3 color nodes $T_i^{(j)}, F_i^{(j)}, D_i^{(j)}$ (resp. $\bar{T}_i^{(j)}, \bar{F}_i^{(j)}, \bar{D}_i^{(j)}$), which serve as local copies of TRUE, FALSE and DUMMY nodes. Add a triangle on $(T_i^{(j)}, F_i^{(j)}, D_i^{(j)})$ (resp. $(\bar{T}_i^{(j)}, \bar{F}_i^{(j)}, \bar{D}_i^{(j)})$) for each literal block;

3. For each literal, add an edge between the literal node and its negation node, so that literals and their negations will always be assigned to different colors. These edges, as well as the $x_i^{(j)}$-block and the $\bar{x}_i^{(j)}$-block, are illustrated in Fig. 2;

4. For each literal x_i (resp. $\bar{x}_i$), use the *equality gadget* as illustrated in Fig. 3a to connect between each pair of its k copies. That is, create a "k-clique" among $\{x_i^{(j)}\}_{j\in[k]}$ (resp. $\{\bar{x}_i^{(j)}\}_{j\in[k]}$) such that each "edge" in the "clique" is actually an independent copy of the equality gadget;

5. For each clause, add a *clause gadget* to connect 3 literals in that clause together with their corresponding TRUE nodes (in the blocks that are created in step 1) with 6 auxiliary nodes as shown in Fig 3b. Note that all literal nodes and TRUE nodes in the clause gadgets are used only once; and

6. Build a constant-degree expander graph of size $2kn$, with $x_i^{(j)}$-blocks and $\bar{x}_i^{(j)}$-blocks, $1 \le i \le n$ and $1 \le j \le k$, as its vertices. Then, for any two blocks that are connected by an edge in the expander, connect their corresponding TRUE nodes (resp. FALSE nodes and DUMMY nodes) with an independent copy of the equality gadget. See Fig. 4 for an illustration.

Note that for a bounded occurrence instance ϕ of 3-CNF formula, the number of clauses is at most $m \le 2kn/3$.

3.2 Properties of BOT graph and proof of Lemma 2

We first record the following correctness result of G_ϕ from [BOT02], which follows a similar proof of the classic reduction from 3-SAT to 3-COLORING.

Proposition 1 ([BOT02])**.** *The 3-CNF formula ϕ is satisfiable if and only if graph G_ϕ is 3-colorable.*

Next we do some simple calculations on the parameters of G_ϕ.

Proposition 2. *Suppose the underlying expander is d-regular, then G_ϕ satisfies the following.*

- *The maximum degree is* $\max(2d+3, 2k+1) \le 2\max(k,d)+3$*;*
- $|V(G_\phi)| = O(nk) \cdot \max(k,d)$*; and*
- $|E(G_\phi)| = O(nk) \cdot \max(k,d)$.

Proof. The maximum degree follows by noticing that (for simplicity, we only consider nodes for the un-negated literals) $\deg(D_i^{(j)}) = 2d+3$, $\deg(F_i^{(j)}) = 2d+2$, $\deg(T_i^{(j)}) \le 2d+3$, and $\deg(x_i^{(j)}) \le 2k+1$.

Note that every equality gadget introduces 2 auxiliary nodes and 5 additional edges. Similarly, every clause gadget introduces 6 auxiliary nodes and 12 additional edges. The nodes in G_ϕ consists of $2nk$ literals nodes, $6nk$ literal block nodes and auxiliary nodes introduced by gadgets. That is

$$\begin{aligned} |V(G_\phi)| &= 8nk + 2n \cdot \frac{k(k-1)}{2} \cdot 2 + 3 \cdot \frac{2nk \cdot d}{2} \cdot 2 + 6m \\ &\le 8nk + 2n(k^2-k) + 6nkd + 4nk = O(nk) \cdot \max(k,d), \end{aligned}$$

where in the second last step we use the bound $m \le 2kn/3$. Finally,

$$\begin{aligned} |E(G_\phi)| &= nk(3+6) + 2n \cdot \frac{k(k-1)}{2} \cdot 5 + 3 \cdot \frac{2nk \cdot d}{2} \cdot 5 + 12m \\ &\le 9nk + 5n(k^2-k) + 15nkd + 8nk = O(nk) \cdot \max(k,d), \end{aligned}$$

where the first term counts the edges in Fig. 2 for all k copies of each *variable*, and the next two terms count the total number of edges in the equality constraints and the last term counts the total number of edges in the clause gadgets. □

We say an equality gadget is *broken* if any of the 5 edges in Fig. 3a is deleted, and say a literal node ($x_i^{(j)}$ (resp. $\bar{x}_i^{(j)}$)) is *isolated* if either all its $k-1$ equality connections with other copies of x_i (resp. $\bar{x}_i$) are all broken, or all its d equality connections with other literal nodes in the expander graph are all broken. We say a literal pair $(x_i^{(j)}, \bar{x}_i^{(j)})$ is *damaged* if any of the edges within the two literal blocks in Fig. 2 is deleted, or either $x_i^{(j)}$ node or $\bar{x}_i^{(j)}$ node becomes isolated.

Let $E' \subset E(G_\phi)$ be a subset of the edges of graph G_ϕ. What happens if we delete all the edges in E'? Define the *survived subgraph induced by* E' of G_ϕ, denoted $G'_{\phi,E'}$ as follows. Delete all literal pairs that are damaged from deleting edges in E', and delete all the edges connected with damage pairs; delete all clause gadgets that contain one or more damaged literals. We make the following two simple observations on a survived subgraph $G'_{\phi,E'}$.

Claim 1. *Let ϕ' be the 3-CNF formula by keeping only the survived clauses in $G'_{\phi,E'}$. Then ϕ' is satisfiable if and only if $G'_{\phi,E'}$ is 3-colorable.*

Proof. This follows directly by noticing that $G'_{\phi,E'}$ can be regarded as the BOT graph constructed from formula ϕ'. □

Claim 2. *If we delete a subset of edges E', then the number of survived clauses in ϕ' is at least $m - 2|E'|$.*

Proof. This is because deleting an edge causes at most one literal pair damaged, and thus affects at most two clauses. □

Our main goal in this Section is the prove the following lemma on the gap-preserving reduction from Max-3SAT(k) to 3-Coloring of bounded-degree graphs of Bogdanov, Obata and Trevisan [BOT02].

Lemma 2 (restatement)**.** *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.*

Proof. Recall that, by Lemma 1, there are constant integer k and constant $\epsilon(k) > 0$ such that there are two families of instances ϕ_Y and ϕ_N in Max-3SAT(k) of size n each, ϕ_Y is satisfiable and any truth assignment can satisfy at most $1-\epsilon(k)$ fraction of the clauses in ϕ_N, and it is NP-hard to distinguish between these two formulas. Now, use the very strong expander given in Theorem 4 to construct the corresponding BOT graphs $G_{\phi,Y}$ and $G_{\phi,N}$ for ϕ_Y and ϕ_N, respectively (we just need to make sure that the number of vertices of the very strong expander is at least $2nk$; if the number of vertices is larger, then we can simply make some of the literals more than k copies). Let n be the number of variables and m be the number of clauses of the Boolean formula, where $m \geq n$. Define $\epsilon' = \frac{\epsilon \cdot n}{2|E(G_{\phi,N})|}$. Since both k and d are constants, then by Proposition 2, ϵ' is an absolute constant depending on ϵ, k and the degree d of the underlying expander graph. Clearly $G_{\phi,Y}$ is 3-colorable. It is also not hard to prove that $G_{\phi,N}$ is not $(1-\epsilon')$ 3-colorable. Indeed, suppose that $G_{\phi,N}$ is $(1-\epsilon')$ 3-colorable. Then we can delete at most $\epsilon'|E(G_{\phi,N})|$ edges from $G_{\phi,N}$, obtain a survived subgraph with at least $m - \epsilon \cdot n \geq (1-\epsilon)m$ clauses which is 3-colorable. Now by Claim 1, the 3-coloring of the survived subgraph can be decoded into a satisfying assignment for the survived, at least $(1-\epsilon)m$ clauses of ϕ_N, a contradiction. Finally the constant degree bound of the constructed BOT graphs follows from the first item of Proposition 2. □

4 Structure of max-likelihood solutions – Proof of Theorem 5

For convenience of the reader, we repeat Theorem 5 here.

Theorem 5 (restatement)**.** *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*

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

The proof follows from the following slightly stronger lemma, which implies that we can assume that the diagonal of a max likelihood kernel K (indexed by vertices and edges of G_ϕ) satisfies that K_{ii} is equal to the normalized frequency of element i in the training set.

Lemma 3. *Let $[N]$ be the ground set, $\mathcal{S} = \{X_j\}_{j=1}^m$ be a set of m samples with each $X_j \subseteq [N]$. Let D be the empirical distribution induced by the sample set $\mathcal{S}$: namely, for every $X \subseteq [N]$, $D(X) = |\{X \in \mathcal{S}\}|/m$ if X is in the sample set, and $D(X)$ is zero otherwise. Then there is a maximum likelihood marginal kernel K which satisfies that the diagonal entry of K at each element in $[N]$ is equal to the element's empirical frequency in the sample set. That is, $K_{ii} = \sum_{X \in \mathcal{S}: X \ni i} D(X)$ for every $i \in [N]$.*

The rest of this section is devoted to a proof of Lemma 3.

4.1 The L-ensemble case

We start by assuming that all of K's eigenvalues are in $(0,1)$, and follow a similar strategy as of the proof of Proposition 13 in [BMRU17a]. Note that in this case the DPP is an L-ensemble, where $L = K(I-K)^{-1}$. If K is a maximum likelihood DPP, then the corresponding L is a critical point of the likelihood function, and accordingly the directional derivative of the likelihood function in every direction is zero. Thus, in "direction" of any matrix $H \in \mathbb{R}^{N \times N}$, we get

$$\sum_{X \subseteq [N]} D(X) \mathrm{Tr}(L_X^{-1} H_X) - \mathrm{Tr}((I+L)^{-1} H) = 0$$

Fix $t_1, t_2, \ldots, t_n \in \mathbb{R}$, and let T be the diagonal matrix with diagonal $t_1, t_2, \ldots, t_n$. Setting $H = \frac{1}{2}LT + \frac{1}{2}TL$ (this ensures that we consider symmetric kernels), we get $H_X = \frac{1}{2}L_X T_X + \frac{1}{2}T_X L_X$. Since $(I+L)^{-1}$ and L commute, and the trace is invariant under cyclic permutations, we get

$$\sum_{X \subseteq [N]} D(X) \sum_{j \in X} t_j = \mathrm{Tr}(KT) = \sum_{j=1}^N K_{jj} t_j$$

Fix $i \in [N]$, and consider the setting where $t_i = 1$ and $t_j = 0$ for all $j \neq i$. The above equation becomes

$$\sum_{X \subseteq [N]: X \ni i} D(X) = K_{ii}, \tag{1}$$

completing the proof of the lemma for the L-ensemble case.

4.2 Perturbing the training example distribution

To deal with the non L-ensemble case, we apply a continuity argument: specifically, we employ a Lipschitz property of the log likelihood function with respect to small perturbations of the training example distribution. In the rest of the proof we think N (hence both n and m) as a fixed constant and let the small quantities such as ϵ tend to zero independent of N.

Let D be the original empirical distribution induced by the sample set $\mathcal{S}$ of size m. Without loss of generality, we assume[12] both $D(\emptyset) = 0$ and $D([N]) = 0$. We perturb D by adding very tiny probabilities at $\emptyset$ and $[N]$ so that the corresponding correlation kernel is an L-ensemble.[13] Formally, define D_ϵ to be a distribution on $2^{[N]}$ such that for any $X \subseteq [N]$

$$D_\epsilon(X) = \begin{cases} (1-\epsilon)D(X) & \text{if } X \in \mathcal{S}; \\ \epsilon/2 & \text{if } X = \emptyset; \\ \epsilon/2 & \text{if } X = [N]; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

We stress that the small quantity ϵ is independent of the DPP learning instance under consideration.

In the following, we will use K^* to denote a (non L-ensemble) optimal marginal kernel for the original training distribution D, and use K_ϵ to denote an optimal marginal kernel for the perturbed training distribution D_ϵ. Since $D_\epsilon(\emptyset) > 0$ and $D_\epsilon([N]) > 0$, K_ϵ is necessarily an L-ensemble and hence by our previous argument, the diagonals of K_ϵ satisfy (1). In particular, for every $i \in [N]$

$$(K_\epsilon)_{ii} = \frac{\epsilon}{2} + \sum_{X \in \mathcal{S}: X \ni i} (1-\epsilon)D(X), \tag{3}$$

which approaches $\sum_{X \in \mathcal{S}: X \ni i} D(X)$ when ϵ tends to zero.

To make the dependency of the maximum log likelihood estimator on the empirical distribution explicit, we write $\ell(K, D)$ for the log likelihood function of DPP kernel K on training distribution D. Therefore,

$$\ell(K^*, D) = \sum_{X \in \mathcal{S}} D(X) \log \left(\frac{1}{\Pr_{\boldsymbol{Y} \sim \mathcal{P}_{K^*}}[\boldsymbol{Y} = X]} \right), \tag{4}$$

and

$$\begin{aligned} \ell(K_\epsilon, D_\epsilon) =& \frac{\epsilon}{2} \log \left(\frac{1}{\Pr_{\boldsymbol{Y} \sim \mathcal{P}_{K_\epsilon}}[\boldsymbol{Y} = \emptyset]} \right) + \sum_{X \in \mathcal{S}} (1-\epsilon) D(X) \log \left(\frac{1}{\Pr_{\boldsymbol{Y} \sim \mathcal{P}_{K_\epsilon}}[\boldsymbol{Y} = X]} \right) \\ &+ \frac{\epsilon}{2} \log \left(\frac{1}{\Pr_{\boldsymbol{Y} \sim \mathcal{P}_{K_\epsilon}}[\boldsymbol{Y} = [N]]} \right). \end{aligned} \tag{5}$$

The main goal in this subsection is to prove the following inequality, which says that the L-ensemble marginal kernel K_ϵ, which is defined to be optimal for the perturbed training distribution D_ϵ, is also close to being optimal for the original training distribution D.

[12] For if $D(\emptyset) \neq 0$, any optimal DPP kernel for D is necessarily an L-ensemble. We further assume $D([N]) = 0$ to simplify exposition. It is easy to see that the there are two parts in the definition of K_ϵ^* below: the rescaling part is to deal with DPP kernel singularity at $\emptyset$, and the spectrum shifting part is to deal with DPP kernel singularity at $[N]$. One can therefore keep only the rescaling part in K_ϵ^* when $D([N])$ is bounded away from zero.

[13] Making the empirical distribution to be non-zero at $\emptyset$ ensures that the kernel is an L-ensemble; however, in order to apply the matrix derivative argument in [BMRU17a], we need to make the distribution to be non-zero at $[N]$ as well.

Proposition 3. *Let D, K^*, D_ϵ and K_ϵ be defined as before. Then for every $\epsilon > 0$ that is small enough*

$$\ell(K_\epsilon, D) \leq \ell(K^*, D) + O(\epsilon \log \frac{1}{\epsilon}), \tag{6}$$

where the hidden constant in the $O(\epsilon \log \frac{1}{\epsilon})$ term depends only on N.

Proof. Let I_N denote the $N \times N$ identity matrix. Define a new marginal kernel K^*_ϵ as

$$K^*_\epsilon = (1-\epsilon)K^* + \frac{\epsilon}{2} I_N.$$

Clearly, since K^* is PSD, $\lambda_N(K^*_\epsilon) \geq \epsilon/2$. It is also not hard to verify that $\lambda_1(K^*_\epsilon) = (1-\epsilon)\lambda_1(K^*) + \epsilon/2 \leq 1 - \epsilon/2$. It follows that K^*_ϵ is indeed a DPP marginal kernel.

We will prove the following sequence of inequalities

$$\ell(K^*_\epsilon, D) \leq \ell(K^*, D) + O(\epsilon). \tag{7}$$

$$\ell(K^*_\epsilon, D_\epsilon) \leq \ell(K^*_\epsilon, D) + O(\epsilon \log \frac{1}{\epsilon}). \tag{8}$$

$$\ell(K_\epsilon, D_\epsilon) \leq \ell(K^*_\epsilon, D_\epsilon). \tag{9}$$

$$\ell(K_\epsilon, D) \leq \ell(K_\epsilon, D_\epsilon) + O(\epsilon). \tag{10}$$

First of all, let's state and prove two useful bounds.

Fact 1. *Let D' be an arbitrary empirical distribution over the ground set $[N]$, and let $\ell^*(D')$ denote the optimal value of the DPP maximum log likelihood estimator for D'. Then $\ell^*(D') \leq N \log 2$.*

Proof. This is because the trivial diagonal kernel $K_I := \frac{1}{2} I_N$, which corresponds to the uniform distribution over all $2^{[N]}$ subsets, is a legal DPP marginal kernel for any distribution D' over $[N]$, and its log likelihood estimator is easily seen to be $N \log 2$. □

Fact 2. *Let $\mathcal{S} = \{X_j\}_{j=1}^m$ be a set of m samples with each $X_j \subseteq [N]$, the ground set. Let K be any maximum likelihood marginal kernel of the empirical distribution induced by the sample set $\mathcal{S}$. Then, for any $X_j \in \mathcal{S}$,*

$$\Pr_{\boldsymbol{Y} \sim \mathcal{P}_K}[\boldsymbol{Y} = X_j] \geq e^{-\log 2 \cdot mN} > e^{-\log 2 \cdot N^2}.$$

Proof. This follows by combining Fact 1 with the fact that in (4), $D(X_j) \geq 1/m$. □

4.2.1 Proof of inequality (7)

For any DPP marginal kernel K and $X \subseteq [N]$, it is well-known that using inclusion-exclusion principle, we can express $\Pr_{\boldsymbol{Y} \sim \mathcal{P}_K}[\boldsymbol{Y} = X]$ in terms of K as

$$\begin{aligned}\Pr_{\boldsymbol{Y} \sim \mathcal{P}_K}[\boldsymbol{Y} = X] &= \sum_{T \supseteq X} (-1)^{|T \setminus X|} \det(K_T) = (-1)^{|X|} \sum_{T \supseteq X} (-1)^{|T|} \det(K_T) \\ &= (-1)^{|X|} \det(I_{\bar{X}} - K) = |\det(K - I_{\bar{X}})|,\end{aligned}$$

where $I_{\bar{X}}$ stands for the diagonal matrix whose (i,i)-entry is 1 if $i \in \bar{X}$ and is 0 otherwise.

Now fix any $X \in \mathcal{S}$. Then $\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K^*}}[\boldsymbol{Y} = X] = |\det(K^* - I_{\bar{X}})| = |\det(A)|$, where $A := K^* - I_{\bar{X}}$. By Fact 2, $|\det(A)| \geq e^{-mN\log 2}$. Similarly,

$$\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K^*_\epsilon}}[\boldsymbol{Y} = X] = |\det(K^*_\epsilon - I_{\bar{X}})| = |\det(A + \epsilon E)|,$$

where $E := \frac{1}{2}I_N - K^*$.

We use the following result to bound the difference between $\det(A + \epsilon E)$ and $\det(A)$.

Theorem 9 ([IR08])**.** *Let A and E be two $N \times N$ complex matrices. Let $\sigma_1 \geq \cdots \geq \sigma_N$ be the singular values of A, and suppose that the spectral norm of E satisfies $\|E\|_2 < 1$. Then*

$$|\det(A + E) - \det(A)| \leq s_{N-1}\|E\|_2 + O(\|E\|_2^2),$$

where s_{N-1} is the $(N-1)^{st}$ elementary symmetric function in the singular values of A and is upper bounded by $N\sigma_1 \cdots \sigma_{N-1}$.

Since $A = K^* - I_{\bar{X}}$ is symmetric, $\{\sigma_1(A), \ldots, \sigma_N(A)\} = \{|\lambda_1(A)|, \ldots, |\lambda_N(A)|\}$. To bound the singular values of A, we recall Weyl's theorem on the changes to eigenvalues of a Hermitian matrix that is perturbed. Specifically, the theorem states that if A, B and C are Hermitian matrices of size $n \times n$, $C = A + B$, then

$$\begin{gathered}\lambda_1(A) + \lambda_n(B) \leq \lambda_1(C) \leq \lambda_1(A) + \lambda_1(B)\\ \cdots\cdots\cdots\\ \lambda_n(A) + \lambda_n(B) \leq \lambda_n(C) \leq \lambda_n(A) + \lambda_1(B)\end{gathered}$$

Now all the eigenvalues of K^* are in $[0, 1]$, and all the eigenvalues of $-I_{\bar{X}}$ are in $\{-1, 0\}$, therefore, all the eigenvalues of $A = K^* - I_{\bar{X}}$ are in $[-1, 1]$. It follows that $\sigma_1, \ldots, \sigma_N \in [0, 1]$ and $s_{N-1} \leq N$.

Since $E = \frac{1}{2}I_N - K^*$, its eigenvalues are just the corresponding eigenvalues of $-K^*$ but shifted by $\frac{1}{2}$, hence all eigenvalues of E are in $[-\frac{1}{2}, \frac{1}{2}]$ and $\|E\|_2 \leq \frac{1}{2}$.

Now we can plug in all these bounds into Theorem 9 to obtain

$$\begin{aligned}\left|\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K^*}}[\boldsymbol{Y} = X] - \Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K^*_\epsilon}}[\boldsymbol{Y} = X]\right| &\leq \left|\det(K^* - I_{\bar{X}}) - \det\left(K^* - I_{\bar{X}} + \epsilon(\frac{I_N}{2} - K^*)\right)\right|\\ &\leq \frac{\epsilon N}{2} + O(\frac{\epsilon^2}{4}) < \epsilon N,\end{aligned}$$

for all small enough ϵ.

It follows that, for all small enough ϵ,

$$\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K^*_\epsilon}}[\boldsymbol{Y} = X] \geq \Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K^*}}[\boldsymbol{Y} = X] - \epsilon N \geq \Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K^*}}[\boldsymbol{Y} = X](1 - Ne^{\log 2\cdot N^2}\epsilon) \geq \Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K^*}}[\boldsymbol{Y} = X]e^{-O(\epsilon)},$$

where the last step follows from the inequality that $1 - x \geq e^{-\frac{3}{2}x}$ for all $0 \leq x \leq 1/2$.

Since this lower bound on DPP probability of kernel K^*_ϵ holds for every $X \in \mathcal{S}$, plugging it into the log likelihood estimator (4) for empirical distribution D proves inequality (7).

4.2.2 Proof of inequality (8)

Since $D(\emptyset) = D([N]) = 0$, we have

$$\begin{aligned}\ell(K_\epsilon^*, D_\epsilon) &= (1-\epsilon)\ell(K_\epsilon^*, D) + \frac{\epsilon}{2}\log\left(\frac{1}{\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K_\epsilon^*}}[\boldsymbol{Y}=\emptyset]}\right) + \frac{\epsilon}{2}\log\left(\frac{1}{\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K_\epsilon^*}}[\boldsymbol{Y}=[N]]}\right)\\ &\le \ell(K_\epsilon^*, D) + \frac{\epsilon}{2}\log\left(\frac{1}{\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K_\epsilon^*}}[\boldsymbol{Y}=\emptyset]}\right) + \frac{\epsilon}{2}\log\left(\frac{1}{\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K_\epsilon^*}}[\boldsymbol{Y}=[N]]}\right).\end{aligned}$$

We can lower bound $\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K_\epsilon^*}}[\boldsymbol{Y}=\emptyset]$ and $\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K_\epsilon^*}}[\boldsymbol{Y}=[N]]$, using the bounds on $\lambda_1(K_\epsilon^*)$ and $\lambda_N(K_\epsilon^*)$ obtained right after the definition of K_ϵ^*, as follows.

$$\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K_\epsilon^*}}[\boldsymbol{Y}=\emptyset] = \det(I-K_\epsilon^*) = \prod_{i=1}^{N}(1-\lambda_i(K_\epsilon^*)) \ge (\frac{\epsilon}{2})^N.$$

Similarly,

$$\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K_\epsilon^*}}[\boldsymbol{Y}=[N]] = \det(K_\epsilon^*) = \prod_{i=1}^{N}\lambda_i(K_\epsilon^*) \ge (\frac{\epsilon}{2})^N.$$

Therefore,

$$\ell(K_\epsilon^*, D_\epsilon) \le \ell(K_\epsilon^*, D) + N\epsilon\log\frac{2}{\epsilon} = \ell(K_\epsilon^*, D) + O(\epsilon\log\frac{1}{\epsilon}).$$

4.2.3 Proof of inequality (9)

This follows directly from the fact that K_ϵ is an optimal kernel for the sample distribution D_ϵ.

4.2.4 Proof of inequality (10)

Recall that

$$\begin{aligned}\ell(K_\epsilon, D_\epsilon) =& \frac{\epsilon}{2}\log\left(\frac{1}{\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K_\epsilon}}[\boldsymbol{Y}=\emptyset]}\right) + \sum_{X\in\mathcal{S}}(1-\epsilon)D(X)\log\left(\frac{1}{\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K_\epsilon}}[\boldsymbol{Y}=X]}\right)\\ &+ \frac{\epsilon}{2}\log\left(\frac{1}{\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K_\epsilon}}[\boldsymbol{Y}=[N]]}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\ell(K_\epsilon, D) &= \sum_{X\in\mathcal{S}} D(X)\log\left(\frac{1}{\Pr_{\boldsymbol{Y}\sim\mathcal{P}_{K_\epsilon}}[\boldsymbol{Y}=X]}\right)\\ &\le \frac{1}{1-\epsilon}\ell(K_\epsilon, D_\epsilon)\\ &\le (1+2\epsilon)\ell(K_\epsilon, D_\epsilon) \qquad \text{(as long as } \epsilon \le 1/2)\\ &\le \ell(K_\epsilon, D_\epsilon) + (2\log 2)N\epsilon\\ &= \ell(K_\epsilon, D_\epsilon) + O(\epsilon),\end{aligned}$$

where in the second last step we use the fact that K_ϵ is an optimal DPP kernel for D_ϵ and the bound in Fact 1.

4.2.5 Putting everything together

Adding inequalities (7), (8), (9) and (10) together yields $\ell(K_\epsilon, D) \leq \ell(K^*, D) + O(\epsilon \log \frac{1}{\epsilon})$, which completes the proof of Proposition 3. □

4.3 An optimal kernel as a limiting matrix

Let ε be a small enough positive number so that Proposition 3 holds. Define an infinite decreasing sequence $\{\epsilon_1, \ldots, \epsilon_k, \ldots\}$, where $\epsilon_k = \varepsilon/k$. Define an infinite sequence of distributions $\{D_{\epsilon_1}, \ldots, D_{\epsilon_k}, \ldots\}$ as in (2) for each ϵ_k. Correspondingly, let $\{K_{\epsilon_1}, \ldots, K_{\epsilon_k} \ldots\}$ be a sequence of optimal DPP kernels for distributions $\{D_{\epsilon_1}, \ldots, D_{\epsilon_k}, \ldots\}$. Note that, since there can be more than one optimal DPP kernel for each distribution, such an infinite sequence of kernels is in general not unique. We just fix one such sequence.

If we view[14] the set of all (symmetric) $N \times N$ positive semidefinite matrices whose maximum eigenvalues are bounded by 1 as a subset $\mathcal{P}$ of $\mathbb{R}^{N^2}$, then for any $M \in \mathcal{P}$,

$$\sum_{i,j=1}^{N} |M_{i,j}|^2 = \sum_{k=1}^{N} \sigma_k^2(M) = \sum_{k=1}^{N} \lambda_k^2(M) \leq N,$$

where $\sigma_k(M)$ denotes the k^{th} singular value of matrix M. It follows that $\mathcal{P}$ is a bounded set.

Consider $\{K_{\epsilon_1}, \ldots, K_{\epsilon_k}, \ldots\}$, which is an infinite sequence in $\mathcal{P}$. Recall the Bolzano-Weierstrass theorem, which states that every bounded infinite sequence in a finite-dimensional Euclidean space $\mathbb{R}^{N^2}$ has a convergent subsequence. Therefore there exists an infinite sequence of indices $i_1, \ldots, i_k, \ldots$ of $\mathbb{N}$ such that the infinite subsequence of matrices $\{K_{\epsilon_{i_1}}, \ldots, K_{\epsilon_{i_k}}, \ldots\}$ converge in $\mathcal{P}$.

Let K_∞ be the limiting matrix of the converging sequence $\{K_{\epsilon_{i_1}}, \ldots, K_{\epsilon_{i_k}}, \ldots\}$. That is,

$$K_\infty := \lim_{k \to \infty} K_{\epsilon_{i_k}}.$$

Since each of the marginal kernel in $\{K_{\epsilon_{i_1}}, \ldots, K_{\epsilon_{i_k}}, \ldots\}$ satisfies the diagonal entry condition in (3), so does K_∞. That is, for every $j \in [N]$

$$(K_\infty)_{jj} = \lim_{k \to \infty} (K_{\epsilon_{i_k}})_{jj} = \lim_{\epsilon \to 0} \left(\frac{\epsilon}{2} + \sum_{X \in \mathcal{S}: X \ni j} (1 - \epsilon) D(X) \right) = \sum_{X \in \mathcal{S}: X \ni j} D(X).$$

Moreover, by Proposition 3,

$$\ell(K_\infty, D) = \lim_{k \to \infty} \ell(K_{\epsilon_{i_k}}, D) \leq \ell(K^*, D) + \lim_{k \to \infty} O(\epsilon_{i_k} \log \frac{1}{\epsilon_{i_k}}) = \ell(K^*, D),$$

which shows that K_∞ is an optimal DPP kernel for the original sample distribution D. This completes the proof of Lemma 3.

[14] Correspondingly, the underlying matrix norm is the Frobenius norm.

5 An optimal kernel for 3-colorable graphs – Proof of Theorem 6

We now prove Theorem 6, which states that if formula ϕ is satisfiable (hence the corresponding BOT graph G_ϕ is 3-colorable), then there exists a rank-3 optimal DPP marginal kernel K, and the log likelihood function is $\ell(K) = 3\log m - \frac{1}{m}\sum_{(u,v)\in E(G_\phi)}\Big(\log(\deg_{G_\phi}(u)) + \log(\deg_{G_\phi}(v))\Big)$.

Proof. From Theorem 5, we can assume that the diagonal of K (indexed by vertices and edges of G_ϕ) satisfies

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

We first prove a lower bound on the log likelihood function of any marginal kernel K, by means of the Hadamard inequality.

Lemma 4 (Hadamard's inequality, see e.g. Theorem 7.8.1 in [HJ12])**.** *For every positive semidefinite matrix B, $\det(B) \le \prod_i B_{ii}$, with equality if and only if B is diagonal.*

Therefore, for every example $T = (a_u, a_v, a_{(u,v)}) \in E'(H_\phi)$, we can upper bound the probability that marginal kernel K outputs T as

$$\Pr_{\boldsymbol{Y}\sim\mathcal{P}_K}[\boldsymbol{Y} = T] \le \Pr_{\boldsymbol{Y}\sim\mathcal{P}_K}[T \subseteq \boldsymbol{Y}] = \det(K_T) \le \frac{\deg_{G_\phi}(u)}{m}\frac{\deg_{G_\phi}(v)}{m}\frac{1}{m} = \frac{\deg_{G_\phi}(u)\deg_{G_\phi}(v)}{m^3},$$

The average log likelihood of the DPP associated with any marginal kernel K thus satisfies

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

Next, for any 3-colorable BOT graph G_ϕ, we construct a rank-3 marginal kernel K with matching log likelihood function, hence proving the optimality of the kernel. The kernel K is constructed by the natural 3-dimensional embedding induced by the vertex coloring of G_ϕ.

Let $\chi : V(G_\phi) \to \{1,2,3\}$ be a proper 3-coloring of G_ϕ. We extend the coloring function χ to include the edge set $E(G_\phi)$ in the natural way: for every $(u,v) \in E(G_\phi)$, let $\chi((u,v)) = \{1,2,3\} \setminus \{\chi(u), \chi(v)\}$. Since χ is a proper coloring of G_ϕ, such an extended definition of χ is unambiguous.

Let $\mathbf{e}_1 = \begin{pmatrix}1\\0\\0\end{pmatrix}$, $\mathbf{e}_2 = \begin{pmatrix}0\\1\\0\end{pmatrix}$, and $\mathbf{e}_3 = \begin{pmatrix}0\\0\\1\end{pmatrix}$. Let $N = |V(G_\phi)| + |E(G_\phi)|$. Let Q, the embedding matrix of coloring χ, be a $3 \times N$ matrix whose columns are indexed by the vertices and edges of G_ϕ:

$$Q = \begin{pmatrix} | & | & & | \\ q_1 & q_2 & \cdots & q_N \\ | & | & & | \end{pmatrix},$$

Each q_i is a 3-dimensional column vector set as follows:

- If $i = u \in V(G_\phi)$, then set $q_i = \sqrt{\frac{\deg_{G_\phi}(u)}{m}}\mathbf{e}_{\chi(u)}$;

- if $i = (u, v) \in E(G_\phi)$, then set $q_i = \sqrt{\frac{1}{m}}\mathbf{e}_{\chi((u,v))}$.

Finally, define the marginal kernel as $K = Q^\top Q$.

Clearly, $\mathrm{rank}(K) = 3$. Since K is a Gram matrix, it is positive semidefinite and the diagonals of K satisfy the conditions specified in Theorem 5. Next we prove that $\lambda_1(K)$ — the largest eigenvalue of K — is equal to 1, which would imply that K is indeed a DPP marginal kernel.

By Courant-Fischer's variational formulation of the eigenvalues,

$$\lambda_1(K) = \sup_{x \in \mathbb{R}^N,\ \|x\|_2 = 1} x^\top K x = \sup_{x \in \mathbb{R}^N,\ \|x\|_2 = 1} \|Qx\|^2.$$

For notational convenience, for each column vector in Q, let $q_i = \sqrt{p_i}\mathbf{e}_j$ where $j \in \{1, 2, 3\}$ is the color of node i in hypergraph H_ϕ. That is, $p_i = \frac{\deg_{G_\phi}(i)}{m}$ if node i corresponds to a vertex in G_ϕ and $p_i = \frac{1}{m}$ if node i corresponds to an edge in G_ϕ. Let $W_1 = \sum_{i \in [N]:\chi(\text{node } i)=1} x_i\sqrt{p_i}$, and similarly define W_2 and W_3. Also define P_1 (respectively P_2 and P_3) to be $W_1 = \sum_{i \in [N]:\chi(\text{node } i)=1} p_i$. An important observation is that, since χ is a proper 3-coloring of G_ϕ, $P_1 = P_2 = P_3 = 1$.

Clearly

$$Qx = \begin{pmatrix} W_1 \\ W_2 \\ W_3 \end{pmatrix},$$

and by the Cauchy-Schwarz inequality

$$W_1^2 \le \left(\sum_{i \in [N]:\chi(\text{node } i)=1} p_i \right) \left(\sum_{i \in [N]:\chi(\text{node } i)=1} x_i^2 \right) = \sum_{i \in [N]:\chi(\text{node } i)=1} x_i^2.$$

Similar inequalities hold for W_2 and W_3. Therefore,

$$\begin{aligned} \lambda_1(K) &= \sup_{x \in \mathbb{R}^N,\ \|x\|_2 = 1} \|Qx\|_2^2 = W_1^2 + W_2^2 + W_3^2 \\ &\le \sum_{i \in [N]:\chi(\text{node } i)=1} x_i^2 + \sum_{i \in [N]:\chi(\text{node } i)=2} x_i^2 + \sum_{i \in [N]:\chi(\text{node } i)=3} x_i^2 \\ &= \sum_{i \in [N]} x_i^2 = 1. \end{aligned}$$

Furthermore, it is easy to see that we can choose q_i's such that q_i is proportional to $\sqrt{p_i}$ for every $i \in [N]$, thus making the Cauchy-Schwarz inequalities to be equalities. This shows that the spectral norm of K is indeed one.[15]

It is well-known that, if $q_1, \ldots, q_k$ are k vectors in an inner product space and form the $k \times k$ Gram matrix A whose (i, j) entry is the inner product between q_i and q_j, then $\det(A)$ is equal to the square of the volume of the k-dimensional parallelepiped spanned by $q_1, \ldots, q_k$. Now by our

[15] We remark that, for G_ϕ that is not 3-colorable, a marginal kernel constructed from a 3-coloring may fail to be an optimal DPP kernel for two reasons. Firstly, such a coloring necessarily make some edges in G_ϕ monochromatic, and hence the corresponding DPP probabilities vanish. Secondly, it may incurs some discrepancy among three colors. Since $P_1 + P_2 + P_3 = 3$ always holds, the discrepancy then implies $\lambda_1(K) = \max\{P_1, P_2, P_3\} > 1$, and consequently we need to scale down the matrix K by a factor larger than one in order to make it a DPP kernel. This in turn implies that kernel such constructed can not be an optimal one, as indicated by Theorem 5.

construction, for every example $T = (a_u, a_v, a_{(u,v)}) \in E'(H_\phi)$, the u^{th}, v^{th}, and $(u,v)^{\text{th}}$ column vectors of Q are pairwise orthogonal. It follows that, since $\text{rank}(K) = 3$,

$$\Pr_{\boldsymbol{Y} \sim \mathcal{P}_K}[\boldsymbol{Y} = T] = \Pr_{\boldsymbol{Y} \sim \mathcal{P}_K}[T \subseteq \boldsymbol{Y}] = \det(K_T) = \frac{\deg_{G_\phi}(u)}{m} \frac{\deg_{G_\phi}(v)}{m} \frac{1}{m}.$$

Consequently,

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

which completes the proof of the theorem. □

6 Good rank-3 kernels exist – Proof of Theorem 7

In this section we prove Theorem 7, which shows that, if the log likelihood function is already close to optimal, then there exists correspondingly a *rank*-3 kernels whose log likelihood function is also close (with worse proximity parameter) to optimal.

Theorem 7 (restatement)**.** *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 $\ell(K') \le \ell^* + C_k \delta^{1/4}$.*

6.1 Proof overview

Without loss of generality, we can assume that the diagonal of K is consistent with Theorem 5, namely it satisfies that the diagonal is

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

By our construction of G_ϕ, we may assume that every element $i \in [m+n]$ occurs in at most k training examples in $\mathcal{T}$ for some absolute constant k.

Since K is PSD, there exists a matrix Q such that $K = Q^\top Q$. Let $q_1, q_2, \ldots, q_N$ denote the columns of Q, and for a set $T \subseteq [N]$, define $Q_T = \text{span}(\{q_i : i \in T\})$. Recall that we need to compare

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

which corresponds to an embedding with perfect vector coloring, with

$$\ell(K) \ge 3 \log m - \frac{1}{m} \sum_{(u,v) \in E(G_\phi)} \left(\log(\deg_{G_\phi}(u)) + \log(\deg_{G_\phi}(u)) + \log(\sin^2 \theta_{(u,v)}) \right).$$

Note that $\ell(K)$ can be greater than the RHS above because, for some edge $(u,v) \in E(G_\phi)$, the edge vector $q_{(u,v)}$ may not be perpendicular to the plane spanned by q_u and q_v. Since "vertex" (u,v) in the

hypergraph H_ϕ is connected only with vertices u and v, and our goal is to maximize the likelihood of K (hence minimize the log likelihood $\ell(K)$), from now on we assume that $q_{(u,v)}$ is always taken to be orthogonal to both q_u and q_v, for every $(u,v) \in E(G_\phi)$.

It is more convenient to work with likelihood functions, which are $L(K) = \exp(\ell(K))$ and $L^*(K) = \exp(\ell^*(K))$. After simple manipulations, we have

$$\frac{L^*(K)}{L(K)} = \frac{\prod_{(u,v)\in E(G_\phi)} \Pr_{\boldsymbol{Y}\sim\mathcal{P}_K}[\boldsymbol{Y} = \{u, v, (u,v)\}]}{\prod_{(u,v)\in E(G_\phi)} \|q_u\|_2^2 \cdot \|q_v\|_2^2 \cdot \|q_{(u,v)}\|_2^2} = \prod_{(u,v)\in E(G_\phi)} \sin^2\theta_{(u,v)} \geq \exp(-\delta m), \tag{11}$$

by our assumption that $\ell(K) \leq \ell^* + \delta$.

That is, the ratio between the likelihood of a marginal kernel K and the likelihood of a perfect coloring kernel is just the product of $\sin^2\theta_{(u,v)}$, over all edges $(u,v) \in E(G_\phi)$.

To construct a good dimension-3 kernel K' from kernel K, our basic idea is to project each column vector of Q onto a subspace of dimension 3 (so that the dimension of the new kernel is at most 3), and show that the likelihood does not decrease too much. However, there are several issues we need to cope with. First, how to find such a subspace? Second, there can be pairs of column vectors q_u and q_v of Q such that the angle $\theta_{(u,v)}$ between them is small, and this angle may become even smaller or even zero upon projection. Last, there can be column vectors whose projections onto the subspace have extremely small or even zero lengths.

Since the log likelihood function of K is close to optimal, there exists an edge $(u,v) \in E(G_\phi)$ such that q_u and q_v are nearly orthogonal (recall that we always take $q_{(u,v)}$ to be orthogonal to the plane spanned by q_u and q_v) by a simple probabilistic argument. The subspace spanned by $\{q_u, q_v, q_{(u,v)}\}$ will be the dimension-3 subspace onto which we project each column vector of Q. To tackle the other two issues mentioned above, we employ a simple counting method and the negative association property of conditional DPPs to bound the numbers of column vectors involved in these two bad situations, and apply a greedy algorithm to assign new directions for these vectors while keep their norms unchanged. This allows us to lower bound the probabilities of seeing these "bad edges" under the new DPP kernel. Finally, we lower bound the likelihood function of the new kernel on the training set by providing an upper bound on the scaling factor needed to ensure that the new kernel's eigenvalues are bounded by 1.

6.2 Finding a good dimension-3 subspace

Let $K = Q^\top Q$ be a marginal kernel such that $L(K) \geq \exp(-\delta m)L^*(K)$. If $\operatorname{rank}(Q) \leq 3$, we are done. So we assume that $\operatorname{rank}(Q) > 3$. By geometric averaging, there exists a size-3 subset $S = \{u, v, (u,v)\}$ such that

$$\Pr_{\boldsymbol{Y}\sim\mathcal{P}_K}[\boldsymbol{Y} = S] \geq \exp(-\delta) \prod_{i\in S} \|q_i\|_2^2.$$

In the following, we use $\bar{S} := [N] \setminus S$ to denote the complement of S.

Denote the dimension-3 subspace spanned by the embedding vectors of these three elements by V, namely

$$V = \operatorname{span}\{q_u, q_v, q_{(u,v)}\}.$$

Then V will the subspace onto which we project all column vectors of Q. More importantly, we will show that the average length orthogonal to V of the column vectors of Q is small.

Lemma 5. *Let V be a dimension-3 subspace as constructed above, then*

$$\sum_{i\in\bar{S}} \|\mathrm{proj}_{V^\perp} q_i\|_2^2 \le \delta,$$

where $V^\perp$ denotes the dimension $\mathrm{rank}(Q)-3$ *subspace orthogonal to V.*

Proof. Let $\boldsymbol{Y}$ be a random set distributed according to the DPP kernel K, and let $\boldsymbol{Y}_S$ be a random set which is distributed according kernel K, conditioned on S occurring in $\boldsymbol{Y}_S$. It is well-known that conditioning a DPP on the event that all of the elements in a fixed set are observed gives rise to another DPP distribution (over the ground set $\bar{S}$) (see, e.g. [Kul12]). Therefore if we define $\boldsymbol{Z} := \boldsymbol{Y}_S \setminus S$ as the random set containing elements outside S which appear in $\boldsymbol{Y}_S$, then $\boldsymbol{Z}$ is distributed as a DPP.[16]

By our choice of subset S,

$$\begin{aligned}
&\exp(-\delta)\prod_{i\in S}\|q_i\|_2^2 \le \Pr_{\boldsymbol{Y}\sim\mathcal{P}_K}[\boldsymbol{Y}=S] \\
&= \Pr_{\boldsymbol{Y}\sim\mathcal{P}_K}[S\subseteq\boldsymbol{Y}]\cdot\Pr[\boldsymbol{Y}_S=S] = \Pr_{\boldsymbol{Y}\sim\mathcal{P}_K}[S\subseteq\boldsymbol{Y}]\cdot\Pr[\boldsymbol{Z}=\emptyset] \\
&= \det(K_S)\cdot\Pr[\boldsymbol{Z}=\emptyset].
\end{aligned}$$

Since $\det(K_S)\le\prod_{i\in S}\|q_i\|_2^2$, it follows that $\Pr[\boldsymbol{Z}=\emptyset]\ge\exp(-\delta)$.

Recall that a set of random variables $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_N$ is said to be *negatively associated* (NA) if for any two disjoint index subsets $S,T\subset[N]$ and any two functions f,g that depend only on variables in subsets S and T respectively, and are either both monotone increasing or both monotone decreasing, we have

$$\mathrm{E}\left[f(\boldsymbol{X}_i : i\in S)\cdot g(\boldsymbol{X}_j : j\in T)\right] \le \mathrm{E}\left[f(\boldsymbol{X}_i : i\in S)\right]\cdot\mathrm{E}\left[g(\boldsymbol{X}_j : j\in T)\right].$$

Lyons [Lyo03] proved that the indicator random variables of elements in the set are negatively associated if the probability distribution on the set is a DPP.

Claim 3. *If $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_N$ are negatively associated $\{0,1\}$-valued random variables such that $\Pr[\bigwedge_{i=1}^N(\boldsymbol{X}_i=0)]\ge\exp(-\alpha)$, then $\mathrm{E}[\sum_{i=1}^N\boldsymbol{X}_i]\le\alpha$.*

Proof. First of all, define $f_i(\boldsymbol{X}_i) := \mathbb{1}(\boldsymbol{X}_i=0)$ for every $i\in[N]$. Note that for disjoint subsets S and T, $f:=\prod_{i\in S}f_i$ and $g:=\prod_{i\in T}f_i$ depend on disjoint subsets of variables and are both monotone decreasing. Therefore, applying inductively the negatively associated property of $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_N$, we have $\mathrm{E}\left[\prod_{i\in[N]}f_i(\boldsymbol{X}_i)\right]\le\prod_{i\in[N]}(\mathrm{E}[f_i(\boldsymbol{X}_i)])$, or equivalently

$$\Pr[\bigwedge_{i=1}^N(\boldsymbol{X}_i=0)]\le\prod_{i=1}^N\Pr[\boldsymbol{X}_i=0].$$

[16] In fact, $\boldsymbol{Z}$ is distributed as a DPP with marginal kernel $K-K_{\bar{S}S}K_S^{-1}K_{S\bar{S}}$ [BR05], where K_{AB} denotes the submatrix of K consisting of the rows in A and columns in B. However, the actual form of the DPP kernel is unimportant for our argument here.

Therefore,

$$\begin{aligned}\exp(-\alpha) &\le \Pr[\bigwedge_{i=1}^{N}(\boldsymbol{X}_i = 0)] \le \prod_{i=1}^{N}\Pr[\boldsymbol{X}_i = 0] \\ &= \prod_{i=1}^{N}(1 - \Pr[\boldsymbol{X}_i = 1]) \le \prod_{i=1}^{N}\exp(-\Pr[\boldsymbol{X}_i = 1]) \\ &= \exp(-\sum_{i=1}^{N}\Pr[\boldsymbol{X}_i = 1]) = \exp(-\mathrm{E}[\sum_{i=1}^{N}\boldsymbol{X}_i]).\end{aligned}$$

Hence the claim follows. □

Claim 4. *Let $\boldsymbol{Z}$ be the random set containing elements outside S which appear in $\boldsymbol{Y}_S$ as before, then $\mathrm{E}[|\boldsymbol{Z}|] \le \delta$.*

Proof. Define $\boldsymbol{X}_i$ to be the $\{0,1\}$-indicator random variable corresponding to the event $i \in \boldsymbol{Z}$ for each $i \in \bar{S}$. Then applying Claim 3 to these variables, and observing that $|\boldsymbol{Z}| = \sum_{i\in\bar{S}}\boldsymbol{X}_i$ proves the claim. □

Now we have

$$\mathrm{E}[|\boldsymbol{Z}|] = \sum_{i\in\bar{S}}\Pr[i \in \boldsymbol{Z}] = \sum_{i\in\bar{S}}\frac{\Pr_{\boldsymbol{Y}\sim\mathcal{P}_K}[S\cup\{i\}\subseteq\boldsymbol{Y}]}{\Pr_{\boldsymbol{Y}\sim\mathcal{P}_K}[S\subseteq\boldsymbol{Y}]} = \sum_{i\in\bar{S}}\frac{\det(K_{S\cup\{i\}})}{\det(K_S)},$$

and interpreting the determinants as squared volumes of parallelepipeds, we have

$$\sum_{i\in\bar{S}}\frac{\det(K_{S\cup\{i\}})}{\det(K_S)} = \sum_{i\in\bar{S}}\|\mathrm{proj}_{V^\perp}q_i\|_2^2$$

and applying Claim 4 completes the proof. □

6.3 Bounding the number of "bad" vertices

Let $\epsilon_0 < \frac{1}{8}$ be a fixed constant. Define

$$B_e = \{(u,v) \in E(G_\phi) : \sin^2\theta_{(u,v)} < \epsilon_0\}.$$

Lemma 6. *The size of B_e can be upper bounded as $|B_e| \le \delta m$.*

Proof. If we rewrite (11) as

$$\exp(-\delta m)L^*(K) \le L(K) \le L^*(K)\prod_{(u,v)\in E(G_\phi)}\sin^2\theta_{(u,v)},$$

then

$$\exp(-\delta m) \le \prod_{(u,v)\in E(G_\phi)}\sin^2\theta_{(u,v)} < {\epsilon_0}^{|B_e|}.$$

Taking logarithms on both sides yields $-\delta m \le |B_e|(\log\epsilon_0) \le |B_e|\log(1/8)$, so $|B_e| \le \dfrac{\delta m}{\log 8} \le \delta m/2$. □

Define B_1 as the set of "bad" vertices of the first kind; that is, the set of vertices whose embedding vectors that are too close to some of their neighboring (in graph G_ϕ)) embedding vectors:

$$B_1 = \{u \in V(G_\phi) : \exists v \text{ such that } \sin^2 \theta_{(u,v)} < \epsilon_0\}.$$

Then clearly,

$$|B_1| \leq 2|B_e| \leq \delta m. \tag{12}$$

6.4 Assigning projections for "bad vectors"

Define B_2 to be the set of "bad" vertices of the second kind; these are vertices whose embedding vectors' (relative) norms along $V^\perp$ are not small (consequently, the norms of the projections onto V of such vectors are not large enough):

$$B_2 = \{i \in [N] : \|\text{proj}_{V^\perp} q_i\|_2^2 \geq \sqrt{\delta}\|q_i\|_2^2\}.$$

Let $B := B_1 \cup B_2$ be the set of all "bad" vertices (correspondingly, "bad" embedding vectors). We will call a vertex (and its corresponding embedding vector) "good" if it is not "bad".

First of all, since $\|q_i\|_2^2 \geq \frac{1}{m}$ for each $i \in [N]$, by Lemma 5, the size of B_2 can be bounded by

$$|B_2| \leq \sqrt{\delta} m. \tag{13}$$

Combined with (12), this gives

$$|B| \leq 2\sqrt{\delta} m. \tag{14}$$

For every "good" vertex $u \notin B$, let

Assigning embedding vectors for "good" vertices
$q'_u \leftarrow \text{proj}_V q_u$

What about "bad" vertices? For each $v \in B$, we will use the greedy algorithm outlined in Fig. 5 to find a good direction $\mathbf{z}$ (as the direction of its new embedding vector q'_v) so that the angle between $\mathbf{z}$ and the new embedding vector of any neighbor (in graph G_ϕ) of v is not too small. Note that in the description of the algorithm, we assume that the the projection subspace $V = \mathbb{R}^3$ and use $\theta(x, y)$ to denote the angle between any two vectors x and y in $\mathbb{R}^3$.

Claim 5. *There is an absolute constant τ_k depending on k only such that, for all $(u, v) \in E(G_\phi)$ with at least one of u and v is "bad", the angle between their new embedding vectors satisfies that $\sin^2 \theta(q'_u, q'_v) \geq \tau_k$.*

Proof. Recall that every element $u \in V(G_\phi)$ has at most k neighbors in $V(G_\phi)$. Then we can take τ_k to be the solution to the following optimization problem

$$\min_{x_1, x_2, \ldots, x_k \in S^2} \max_{y \in S^2} \min_i \sin^2 \theta(x_i, y),$$

Observe that τ_k is an absolute constant between 0 and 1. Indeed, one can obtain a simple lower bound on τ_k as follows: let r_k be the maximum radius of a spherical cap such that $k + 1$ identical such caps can be placed disjointly on the surface of a unit sphere. Now for any configuration of $x_1, x_2, \ldots, x_k \in S^2$, at least one cap will contain no x_i. Letting y be the center of such a cap is a certificate that $\tau_k \geq r_k^2$. □

Assigning embedding vectors for "bad" vertices

1. order all n vertices arbitrarily
2. for each $v \in V(G_\phi)$
   If v is "good"
   continue (that is, $q'_v \leftarrow \text{proj}_V q_v$ as we did before)
   Else (v is "bad")
   let $\mathcal{N}$ be the set of neighboring vertices of v that appear before v
   let $\mathbf{z} = \operatorname{argmax}_{y \in S^2} \max_{u \in \mathcal{N}} \min \sin^2 \theta(u, y)$
   (such maximal direction in general is not unique: any one works)
   assign $q'_v \leftarrow \|q_v\|_2 \mathbf{z}$

Figure 5: Algorithm for computing the new embedding vector for "bad" vertices.

Once we set new embedding vector for every vertex $v \in V(G_\phi)$, let

$$Q' = \begin{pmatrix} | & | & & | \\ q'_1 & q'_2 & \cdots & q'_N \\ | & | & & | \end{pmatrix},$$

and set

$$K' = \beta (Q')^\top Q',$$

where β is a parameter to be determined later to make the eigenvalues of K' at most 1.

6.5 Bounding the likelihood function of the new kernel

To lower bound the likelihood function of new kernel K' on the training set, we compare the probabilities of kernels K and K' seeing an arbitrary subset $T := \{u, v, (u, v)\}$ in the training set. We distinguish between two cases.

Both u and v are "good" vertices. Since u is "good", the angle θ'_u between q_u and its projection q'_u is small. Indeed, since $\|q'_u\|_2^2 > (1 - \sqrt{\delta})\|q_u\|_2^2$, we have $\sin \theta'_u < \delta^{1/4}$, or $\theta'_u < \sin^{-1} \delta^{1/4} \leq \frac{2}{\pi}\delta^{1/4} < \delta^{1/4}$. It follows that, if both u and v are "good" vertices, and denote the angle between their projected vectors by $\theta'_{(u,v)}$, then we have

$$\theta'_{(u,v)} > \epsilon_0 - 2\delta^{1/4}, \tag{15}$$

by triangle inequality.

Therefore, we have

$$\frac{\Pr_{\boldsymbol{Y}' \sim \mathcal{P}_{K'}}[\boldsymbol{Y}' = \{u, v, (u, v)\}]}{\Pr_{\boldsymbol{Y} \sim \mathcal{P}_K}[\boldsymbol{Y} = \{u, v, (u, v)\}]} \geq \frac{\Pr_{\boldsymbol{Y}' \sim \mathcal{P}_{K'}}[\boldsymbol{Y}' = \{u, v, (u, v)\}]}{\Pr_{\boldsymbol{Y} \sim \mathcal{P}_K}[\{u, v, (u, v)\} \subseteq \boldsymbol{Y}]}$$
$$= \frac{\det(K'_T)}{\det(K_T)} \qquad \text{(since } K' \text{ is a dimension-3 DPP marginal kernel)}$$

$$
\begin{aligned}
&= \beta^3 \frac{\|q'_u\|_2^2 \|q'_v\|_2^2 \sin^2\theta'_{(u,v)}}{\|q_u\|_2^2 \|q_v\|_2^2 \sin^2\theta_{(u,v)}} \\
&\geq \beta^3 (1-\sqrt{\delta})^2 \frac{\sin^2(\epsilon_0 - 2\delta^{1/4})}{\sin^2\epsilon_0} \qquad (\|q'_i\|_2^2/\|q_i\|_2^2 > 1-\sqrt{\delta} \text{ for "good" embedding vectors and using (15))} \\
&\geq \beta^3 (1-\sqrt{\delta})^2 \left(1 - \frac{2\delta^{1/4}}{\sin^2\epsilon_0}\right) \qquad (\text{using the fact that } |\frac{\mathrm{d}}{\mathrm{d}x}(\sin^2 x)| \leq 1 \text{ for all } x) \\
&\geq \beta^3 (1-\sqrt{\delta}) \left(1 - \frac{3\delta^{1/4}}{\sin^2\epsilon_0}\right),
\end{aligned}
$$

where the factor β^3 comes from the scaling factor β in our definition of K', and the fact that these probabilities are determinants of 3×3 principal minors of $\beta(Q')^\top Q'$.

At least one of u and v is "bad". Using Claim 5, we have for such edges

$$
\begin{aligned}
&\frac{\Pr_{\boldsymbol{Y}' \sim \mathcal{P}_{K'}}[\boldsymbol{Y}' = \{u,v,(u,v)\}]}{\Pr_{\boldsymbol{Y} \sim \mathcal{P}_K}[\boldsymbol{Y} = \{u,v,(u,v)\}]} \geq \frac{\Pr_{\boldsymbol{Y}' \sim \mathcal{P}_{K'}}[\boldsymbol{Y}' = \{u,v,(u,v)\}]}{\Pr_{\boldsymbol{Y} \sim \mathcal{P}_K}[\{u,v,(u,v)\} \subseteq \boldsymbol{Y}]} = \frac{\det(K'_T)}{\det(K_T)} \\
&= \beta^3 \frac{\|q'_u\|_2^2 \|q'_v\|_2^2 \sin^2\theta'_{(u,v)}}{\|q_u\|_2^2 \|q_v\|_2^2 \sin^2\theta_{(u,v)}} \\
&\geq \beta^3 (1-\sqrt{\delta}) \tau_k,
\end{aligned}
$$

since at least one of u and v is a "bad" vertex, and $\|q'_i\|_2^2/\|q_i\|_2^2 = 1$ for such vertices.

Putting the two cases together.

$$
\begin{aligned}
&\frac{L(K')}{L(K)} = \frac{\prod_{(u,v)\in E(G_\phi)} \Pr_{\boldsymbol{Y}' \sim \mathcal{P}_{K'}}[\boldsymbol{Y}' = \{u,v,(u,v)\}]}{\prod_{(u,v)\in E(G_\phi)} \Pr_{\boldsymbol{Y} \sim \mathcal{P}_K}[\boldsymbol{Y} = \{u,v,(u,v)\}]} \\
&\geq \prod_{\substack{(u,v)\in E(G_\phi) \\ u\notin B \text{ and } v\notin B}} \beta^3(1-\sqrt{\delta}) \left(1 - \frac{3\delta^{1/4}}{\sin^2\epsilon_0}\right) \prod_{\substack{(u,v)\in E(G_\phi) \\ u\in B \text{ or } v\in B}} \beta^3(1-\sqrt{\delta})\tau_k \\
&\geq \beta^{3m}(1-\sqrt{\delta})^m \left(1 - \frac{3\delta^{1/4}}{\sin^2\epsilon_0}\right)^{m-k|B|} \tau_k^{k|B|} \\
&\geq \beta^{3m}(1-\sqrt{\delta})^m \left(1 - \frac{3\delta^{1/4}}{\sin^2\epsilon_0}\right)^{(1-2k\sqrt{\delta})m} \tau_k^{2k\sqrt{\delta}m} \\
&\geq \beta^{3m} \exp\left[-\left(\frac{3\delta}{2} + \frac{3\delta^{1/4}}{\sin^2\epsilon_0}(1-2k\sqrt{\delta}) + 2k\log(\frac{1}{\tau_k})\sqrt{\delta}\right) m\right] \\
&\geq \beta^{3m} \exp\left(-C''_k \delta^{1/4} m\right),
\end{aligned}
$$

where C''_k is some constant depending only on k and in the second last step we use the inequality $1 - x \geq \exp(-3x/2)$ for all $0 \leq 1/2 \leq x$. It follows that

$$
\frac{L(K')}{L^*(K)} \geq \beta^{3m} \exp\left(-C''_k \delta^{1/4} m\right) \cdot \exp(-\delta m) \geq \beta^{3m} \exp\left(-C'_k \delta^{1/4} m\right), \tag{16}
$$

for some absolute constant C'_k depending on k only.

6.6 Bounding the scaling factor β

We finish the proof by providing a lower bound on β. First, we consider $K'' = (Q'')^\top Q''$, where the i^{th} column of Q'' is $\text{proj}_V q_i$ for *every* $i \in [N]$. Recall that the spectral norm of a matrix A, $\|A\|_2$, is the largest singular value of A, we have $\|Q''\|_2 \leq \|Q\|_2 \leq 1$, as Q'' is a projection onto a subspace spanned by a subset of its columns. By our bound on the number of "bad" embedding vectors (14), at most $2\sqrt{\delta}m$ of the columns of Q' are not the same projection as in Q''. Each column is replaced by a vector of the same length, which is at most $\sqrt{k/m}$. We use the Frobenius norm of $Q'' - Q'$ to upper bound its spectral norm as follows. For every "bad" vertex i,

$$\|q_i - q'_i\|_2^2 = \|q_i\|_2^2 + \|q'_i\|_2^2 - 2\langle q_i, q'_i\rangle \leq 4\|q_i\|_2^2 \leq 4k/m.$$

Therefore,

$$\|Q'' - Q'\|_2^2 \leq \|Q'' - Q'\|_F^2 = \sum_{i=1}^{N} \|q_i - q'_i\|_2^2 = \sum_{i \in B} \|q_i - q'_i\|_2^2 \leq 2\sqrt{\delta}m \cdot 4k/m = 8k\sqrt{\delta}.$$

Now, by the triangle inequality, we have

$$\|Q'\|_2 \leq \|Q''\|_2 + \|Q'' - Q'\|_2 \leq 1 + \sqrt{8k\sqrt{\delta}}.$$

In order for $K' = \beta Q^\top Q$ to be the marginal kernel of a DPP, we need to ensure that $\|K'\|_2 \leq 1$. As $\|K'\|_2 = |\beta|\|Q^\top Q\|_2 = |\beta|\|Q\|_2^2$, we can take $\beta = \frac{1}{(1+\sqrt{8k\sqrt{\delta}})^2} \geq 1 - 2\sqrt{8k\sqrt{\delta}} = 1 - \sqrt{32k\sqrt{\delta}} \geq \exp\left(-\sqrt{72k}\delta^{1/4}\right)$.

Finally, plugging the bound $\beta \geq \exp\left(-\sqrt{72k}\delta^{1/4}\right)$ into (16) to get that

$$\frac{L(K')}{L^*(K)} \geq \beta^{3m} \exp\left(-C'_k\delta^{1/4}m\right) \geq \exp\left(-\sqrt{648k}\delta^{1/4}m\right) \cdot \exp\left(-C'_k\delta^{1/4}m\right) = \exp\left(-C_k\delta^{1/4}m\right),$$

where $C_k = C'_k + \sqrt{648k}$. Or equivalently, $\ell(K') \leq \ell^* + C_k\delta^{1/4}$ for some constant C_k which depends only on k. This completes the proof of Theorem 7.

7 Putting it all together – Proof of Theorem 8

Now it is time to assemble the parts we have built so far and prove the following soundness theorem, which basically says that there must be a gap between the log likelihood functions of the training set derived from the YES instance BOT graph and that from the NO instance BOT graph.

Theorem 8 (Soundness theorem, restatement)**.** *Let ℓ^* be the optimal log likelihood as in Theorem 6. Then there exists a constant $C > 0$ which depends only on k and ϵ' as those defined in Theorem 2, such that the following holds. If there is a DPP marginal kernel K of rank 3, which satisfies $\ell(K) \leq \ell^* + \frac{C}{\log^2 n}$ 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 1.*

In conventional graph coloring, a coloring $\chi : V \to \{c_1, \ldots, c_k\}$ *satisfies* an edge (u, v) if the edge is non-monochromatic, i.e. $\chi(u) \neq \chi(v)$. In other words, an edge connecting u and v forces the colors of the two vertices to be different. In vector coloring, the notion of edge (u, v) being non-monochromatic is generalized to vectors assigned to u and v being (almost) orthogonal. From Lemma 2 we know that an "almost perfect" 3-coloring of the vertices in G_ϕ can be used to decode a truth assignment that satisfies "almost all" the clauses in ϕ. However, in our maximum likelihood DPP learning reduction, we represent vertices by their embedding vectors and work with vector coloring accordingly, therefore it is natural to ask: Given an *almost* perfect vector 3-coloring of the vertices in G_ϕ, can we still decode a truth assignment that satisfies "almost all" the clauses in ϕ, like what we did from a 3-coloring of G_ϕ? Our key observation is that the equality gadgets and clause gadgets in [BOT02] are "robust" even for a vector 3-coloring, with the help of several simple facts about spherical geometry.

From now on, for any vector $v \in \mathbb{R}^3$, we use (v_1, v_2, v_3) to denote its Cartesian coordinates.

Claim 6. *Let $a, b, c, d \in S^2$ be unit vectors in $\mathbb{R}^3$ and let $0 \leq t \leq 1/5$. If $|\langle a, b\rangle| \leq t$, $|\langle b, c\rangle| \leq t$, $|\langle c, a\rangle| \leq t$, $|\langle d, b\rangle| \leq t$, then $|\langle a, d\rangle| \geq 1 - 5t^2$.*

Intuitively, if vectors a, b and c in S^2 are close to forming an orthonormal basis, and a fourth vector $d \in S^2$ is almost orthogonal to both b and c, then d is necessarily close to a.

Proof. Without loss of generality, assume that $b = (0, 0, 1)$. Then $0 \leq |a_3|, |c_3|, |d_3| \leq t$. Let a', c' and d' be the projection to the X-Y plane of a, c and d, respectively. Note that $\|a'\|, \|c'\|, \|d'\| \geq \sqrt{1 - t^2}$, and since $|\langle a, c\rangle| \geq \langle a', c'\rangle - |a_3||c_3|$, we have $|\langle a', c'\rangle| \leq t + t^2$. Similarly, $|\langle d', c'\rangle| \leq t + t^2$.

Without loss of generality, assume that $c' = (0, \|c'\|)$. Then, because $|\langle a', c'\rangle| \leq t + t^2$ and $\|c'\| \geq \sqrt{1 - t^2}$, we have $|a_2| \leq \frac{t+t^2}{\sqrt{1-t^2}}$, and consequently $|a_1^2| \geq 1 - t^2 - \left(\frac{t+t^2}{\sqrt{1-t^2}}\right)^2$. Similar bounds hold for d. Now

$$\begin{aligned}
|\langle a, d\rangle| &\geq |\langle a, d\rangle| - |a_3||d_3| \\
&\geq |a_1||d_1| - |a_2||d_2| - t^2 \\
&\geq 1 - t^2 - \left(\frac{t + t^2}{\sqrt{1 - t^2}}\right)^2 - \frac{t^2(1 + t)^2}{1 - t^2} - t^2 \\
&= 1 - 2t^2 - 2t^2\frac{1 + t}{1 - t} \\
&\geq 1 - 5t^2,
\end{aligned}$$

where the last step we use $\frac{1+t}{1-t} \leq 3/2$ for $0 \leq t \leq 1/5$. □

Corollary 1. *Let a, b, c, d be unit vectors in S^2 satisfy that $\frac{\pi}{2} - \theta \leq \theta_{(a,b)}, \theta_{(a,c)}, \theta_{(b,c)}, \theta_{(b,d)}, \theta_{(c,d)} \leq \pi/2$ with $0 \leq \theta \leq \theta_0 = \sin^{-1}(1/5)$. Then $\theta_{(a,d)} \leq 3\theta$.*

Proof. Let $(x, y) \in \{(a, b), (a, c), (b, c), (b, d), (c, d)\}$ and $t = \sin\theta$. Then $|\langle x, y\rangle| = \cos\theta_{(x,y)} \leq \cos(\pi/2 - \theta) = \sin\theta = t$, so by Claim 6, $|\langle a, d\rangle| \geq 1 - 5\sin^2\theta$. It follows that $\theta_{(a,d)} \leq \cos^{-1}(1 - 5\sin^2\theta) \leq 3\theta$, because $\sin^2(3\alpha) \geq 5\sin^2\alpha$ for all $0 \leq \alpha \leq \pi/7$ and $\theta_0 < \pi/7$. □

The following Claim facilitates transforming between the angle between a pair of vectors in S^2 and their inner product, especially for the cases when the two vectors are almost orthogonal or very close to each other.

Claim 7. *We have the following inequalities between inner product of unit vectors and the angle between them.*

1. *Suppose* $0 \le \theta \le \pi/2$ *and* $\sin^2\theta = 1-\epsilon$ *for some* $0 \le \epsilon < 1$. *Let* $\alpha := \pi/2 - \theta$ *be the angular distance between* θ *and* $\pi/2$. *Then* $\sqrt{\epsilon} \le \alpha \le \frac{\pi}{2}\sqrt{\epsilon}$.
2. *If* $u, v \in S^2$ *and* $|\langle u, v\rangle| = \epsilon$ *(i.e.* u *and* v *are close to orthogonal), and let* $\alpha := \pi/2 - \theta_{(u,v)}$ *be the angular distance between* $\theta_{(u,v)}$ *and* $\pi/2$, *then* $\epsilon \le \alpha \le \frac{\pi}{2}\epsilon$.
3. *If* $u, v \in S^2$ *and* $|\langle u, v\rangle| = 1-\epsilon$ *(i.e.* u *and* v *are* ϵ*-close), then* $\sqrt{2\epsilon} \le \theta_{(u,v)} \le \frac{\pi}{\sqrt{2}}\sqrt{\epsilon}$.

Proof. Item 1 follows from the equality $\sin^2\alpha = \epsilon$ and the fact that, for $\alpha \in [0, \pi/2]$, $\frac{2}{\pi}\alpha \le \sin\alpha \le \alpha$. Item 2 follows directly from the above inequality. For Item 3, since $\sin\theta_{(u,v)} = 1-\epsilon$, $\sin^2\frac{\theta_{(u,v)}}{2} = \epsilon^2/2$. Now apply Item 1 to $\frac{\theta_{(u,v)}}{2}$. □

Remark 2. *Note that our vector* 3*-coloring only specifies the inner-product between any pair of vertex vectors* $\chi^\top(u)\chi(v)$, *and for the purpose of proving hardness of DPP learning, all we care is* $\sin^2\theta_{(u,v)} = 1 - |\chi^\top(u)\chi(v)|^2$. *This information does not uniquely determine all the vertex vectors on* S^2 *though, even if we fix certain vertex vector to be* $(1,0,0)$, *say. This is because, for a fixed vector* $\chi(v)$, $\chi(u)$ *and* $-\chi(u)$ *give rise to the same value of* $\sin^2\theta_{(u,v)}$. *Therefore, from now on, we will not distinguish between* $\chi(u)$ *and* $-\chi(u)$, *and always assume* $\theta_{(u,v)}$ *to be in* $[0, \pi/2]$.

Now we can show that the equality gadget in Fig. 3a is robust against small noise.

Lemma 7 (The equality gadget is robust)**.** *Let* $\chi : V(G_\phi) \to S^2$ *be a vector* 3*-coloring which satisfies that* $\theta_{(u,v)} \ge \pi/2 - \delta$ *for every* $(u,v) \in E(G_\phi)$. *Then, for the the equality gadget in Fig. 3a, the angle between node* x *and node* y *is at most* 3δ.

Proof. This follows directly by plugging $a = \chi(x)$, $b = \chi(y)$, $c = \chi(u)$ and $d = \chi(v)$ into Corollary 1. □

The case of clause gadget requires a bit more work. Note that we make no attempt to optimize the robustness parameters in the following statements, but focus on arguing qualitatively that the clause gadgets are robust against small noise.

Abusing notation, in the following, we use the names of nodes in the gadget to also denote the coloring vectors of the nodes.

Lemma 8 (The clause gadget is robust)**.** *Let* $\chi : V(G_\phi) \to S^2$ *be a vector* 3*-coloring which satisfies that* $\theta_{(u,v)} \ge \pi/2 - \delta_e$ *for every* $(u,v) \in E(G_\phi)$, *where* $\delta_e = o(1)$ *for our BOT hypergraph setting. Further, assume that for any pair of two* True *nodes in the clause gadgets as shown in Fig. 3b or for any pair of two* Dummy *nodes in the literal blocks as shown in Fig. 2, the angle between the two color vectors* u *and* v *satisfy that* $\theta_{(u,v)} \le \delta_p$, *where* $\delta_p = \pi/300$ *is some small absolute constant. Consider the clause gadget as shown in Fig. 3b. Then at least one of the three literals must be "assigned" value* True *in the sense that its color vector is close to that of its corresponding* True *node: there exists at least one* $i \in \{a, b, c\}$ *such that* $|\langle \ell_i, T_i\rangle| > t_0$, *where* $t_0 = 0.98$ *(this corresponds to* $\theta_{(\ell_i, T_i)} < \pi/15$*).*

We first need to argue about the coloring vectors of a simplified clause gadget as shown in Fig. 6.

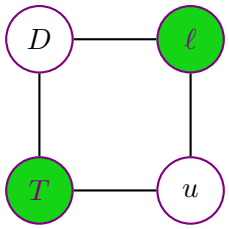


Figure 6: The simplified clause gadget.

Claim 8. *Let $D, T, u, \ell \in S^2$ be unit vectors in $\mathbb{R}^3$ and $t_0 = 0.98$ as in Lemma 8. Suppose that $|\langle D, T\rangle| \le \epsilon$, $|\langle D, \ell\rangle| \le \epsilon$, $|\langle u, T\rangle| \le \epsilon$, $|\langle u, \ell\rangle| \le \epsilon$. If $|\langle T, \ell\rangle| \le t_0$, then $|\langle D, u\rangle| \ge 1 - 202\epsilon^2$.*

Basically the Claim asserts that if the vector 3-coloring satisfies all four edges in the gadget with closeness parameter $1 - \epsilon$, then if vectors T and ℓ are not close to each other, this would force vectors D and u to be aligned along almost the same direction — or equivalently, be assigned to the same "color" (note that if T and ℓ are very close, then u would have the freedom to choose between the color of D or the color that is "orthogonal" to both T and D; namely, the color of FALSE nodes).

Proof. For convenience, we take the plane spanned by T and ℓ to be the X-Y plane and in particular set $T = (1, 0, 0)$ and $t := |\langle T, \ell\rangle| \le t_0$. Then without loss of generality, let $\ell = (t, \sqrt{1-t^2}, 0)$. Let $u = (u_1, u_2, u_3)$. From $|\langle u, T\rangle| \le \epsilon$ and $|\langle u, \ell\rangle| \le \epsilon$, we deduce that $|u_1| \le \epsilon$ and $|u_2| \le \frac{1+t}{\sqrt{1-t^2}}\epsilon = \sqrt{\frac{1+t}{1-t}}\epsilon \le 10\epsilon$. It follows that $u_3^2 \ge 1 - 101\epsilon^2$. By symmetry, the same bounds hold for the three coordinates of vector D. Therefore

$$\begin{aligned}|\langle D, u\rangle| &= |D_1u_1 + D_2u_2 + D_3u_3| \ge |D_3||u_3| - |D_1||u_1| - |D_2||u_2| \\ &\ge 1 - 101\epsilon^2 - \epsilon^2 - (10\epsilon)^2 \\ &= 1 - 202\epsilon^2.\end{aligned}$$

□

Proof of Lemma 8. Pick any fixed pair of TRUE node and DUMMY node in a literal block (see Fig. 2) and denote them by T and D respectively.

Now suppose that none of the literal color vectors ℓ_a, ℓ_b and ℓ_c in Fig. 3b is close to their corresponding TRUE node, i.e. $|\langle T_i, \ell\rangle_i| \le t_0$, for every $i \in \{a, b, c\}$.

Let us consider nodes ℓ_a and u_a. Since ℓ_a is connected to some D_a node in its literal block (so the angle between ℓ_a and D_a is $\pi/2 - o(1)$), and the angle between D_a and the fixed DUMMY node D is at most δ_p, it follows that $\theta_{(\ell_a, D)} \ge \pi/2 - \delta_p - o(1)$. On the other hand, since D and T is connected by an edge, they are very close to orthogonal; moreover, the angle between T_a and T is at most δ_p. Therefore, $\theta_{(T_a, D)} \ge \pi/2 - \delta_p - o(1)$. Since u_a is connected to both T_a and ℓ_a, we have $\theta_{(T_a, u_a)} \ge \pi/2 - o(1)$ and $\theta_{(\ell_a, u_a)} \ge \pi/2 - o(1)$. Therefore, the conditions of Claim 8 are met for vectors D, T_a, u_a and ℓ_a for $\epsilon = \sin(\delta_p + o(1)) \le \delta_p + o(1) \le 1.01\delta_p$ for all sufficiently large BOT graph size n. Hence, by Claim 8, $|\langle D, u_a\rangle| \ge 1 - 202\epsilon^2 \ge 1 - 207\delta_p^2$. By the same reasoning, we have $|\langle D, u_b\rangle| \ge 1 - 207\delta_p^2$ and $|\langle D, u_c\rangle| \ge 1 - 207\delta_p^2$. By Claim 7, item 3, the angle between D and each of the u_a, u_b and u_c is at most $32\delta_p$.

To finish the proof, note that, since each u_i is connected with v_i for every $i \in \{a, b, c\}$, it follows that the angle between D and each v_i is at least $\pi/2 - 33\delta_p$ (whenever n is large enough). Applying

Corollary 1 and noting that v_a, v_b and v_c form an almost orthonormal basis, and D is close to orthogonal to both v_a and v_b with closeness parameter $\theta = 33\delta_p$, we conclude that $\theta_{(D,v_c)} \le 99\delta_p$. But we also know that the angle between D and v_c is at least $\pi/2 - 33\delta_p$, therefore we reach a contradiction as long as $\delta_p < \pi/264$. □

Proof of Theorem 8. We begin by writing the log maximum likelihood estimator of a DPP kernel K as

$$\ell(K) = 3\log m - \frac{1}{m}\sum_{(u,v)\in E(G_\phi)} \left(\log(\deg_{G_\phi}(u)) + \log(\deg_{G_\phi}(v)) + \log(\sin^2\theta_{(u,v)})\right).$$

Let K be a DPP kernel rank 3 which satisfies that $\ell(K) \le \ell^* + \frac{C}{\log^2 n}$, where C is some very small absolute constant to be determined later. Therefore, $\mathrm{E}_{(u,v)\in E(G_\phi)}[\log(\sin^2\theta_{(u,v)})] \ge -\frac{C}{\log^2 n}$. By Markov inequality, except for an ϵ'' fraction of the edges in the BOT graph G_ϕ, all edges (u,v) satisfy that

$$\log(\sin^2\theta_{(u,v)}) \ge -\frac{C/\epsilon''}{\log^2 n},$$

where ϵ'' is an absolute constant to be specified later. Call an edge in the BOT graph that satisfies this condition *good*, and *bad* otherwise. Consequently,

$$\sin^2\theta_{(u,v)} \ge e^{-\frac{C/\epsilon'}{\log^2 n}} \ge 1 - \frac{C}{\epsilon'\log^2 n},$$

By Claim 7, the angle between all good edges (u,v) satisfies that $\theta_{(u,v)} \ge \frac{\pi}{2} - \sqrt{\frac{C\pi}{2\epsilon''}}\frac{1}{\log n}$.

Next we record some facts about the "robustness" of a very strong expander against deletion of a small fraction of its edges. First, following Alon and Capalbo, we note the following notion of expanders which is slightly weaker than *very strong* expanders.

Definition 2 (Strong expanders [AC07])**.** *A graph $G = (V,E)$ is called a* d-strong expander *on n vertices if its minimum vertex degree is at least d, the average degree in any subgraph of G on at most $n/10$ vertices is at most $2d/9$, and the average degree in any subgraph of G on at most $n/2$ vertices is at most $8d/9$.*

Lemma 9 ([AC07])**.** *Let $G = (V,E)$ be a very strong d-regular expander on n vertices. Let $E' \subset E$ be an arbitrary subset of edges satisfying $|E'| \le \frac{nd}{150}$. Consider the process that we first delete E' from G, and then repeatedly delete from G the set of vertices with degree smaller than $3d/4 + 2$ (and all the edges incident to such vertices) as long as such vertices exist. Let* $\mathrm{Dense}(G,E')$ *be the resulting subgraph of G of this vertex-trimming process, and let S denote the set of vertices deleted from G during this process. Then $|S| \le \frac{15|E'|}{d}$.*

It is easy to check (by lower bounding the number of edges between any subset of vertices X of size at most $n/10$ and $n/2$ respectively, with the rest of the vertices in G; namely $|E(X, V-X)|$) that the following lemma holds.

Lemma 10 ([AC07])**.** *Any subgraph of minimum degree at least $\frac{3d}{4}$ of a very strong d-regular expander is a $\frac{3d}{4}$-strong expander.*

It follows that $\mathrm{Dense}(G,E')$ is a $\frac{3d}{4}$-strong expander.

Lastly, the following lemma shows that the diameter of $\mathrm{Dense}(G,E')$ is $O(\log n)$.

Lemma 11 ([AC07]). *For any d and n, a d-regular strong expander on n vertices has diameter at most $\frac{2}{3}\log n + 14$ (which is at most $\log n$ for all large enough n).*

Let n be the number of variables in the 3-CNF instance ϕ. Then the underlying very strong expander has $|V(G_{\text{exp}})| = 2nk$ vertices and $|E(G_{\text{exp}})| = nkd$ edges, where d is the degree of the very strong expander. Also recall that the BOT graph has $|V(G_\phi)| = O(nk) \cdot \max(k, d)$ vertices and $|E(G_\phi)| = O(nk) \cdot \max(k, d)$ edges.

Now, if we set the constant ϵ'' such that $\epsilon''|E(G_\phi)| \cdot \frac{15}{d} = \epsilon n$, or $\epsilon'' = \frac{nd}{15|E(G_\phi)|}\epsilon$, where ϵ is the constant in Lemma 1, then removing any ϵ''-fraction of the edges in G_ϕ makes at least $(1-\epsilon)$-fraction of the clauses intact and thus can be satisfied. Note that, since both d and k are constants, $\epsilon'' = \theta(\epsilon)$, another (small) constant.

Let G'_ϕ be the subgraph of G_ϕ after removing all the bad edges. Note that this can be translated into deleting edges of the underlying very strong expander G_{exp}; namely, if any of the equality gadget edges between two literal blocks is deleted, delete the edge between the corresponding two vertices in G_{exp}. Now we perform the vertex trimming process as described in Lemma 9 to G_{exp}. That is, if an edge between two vertices in G_{exp} is deleted in the process, then we delete the gadget between the two corresponding literal blocks in G'_ϕ. Let G''_ϕ be the resulting graph. Now by our argument in the previous paragraph and Lemma 11, there is a truth assignment that satisfies at least $(1-\epsilon)$ fraction of the clauses in G''_ϕ. Moreover, every edge in G''_ϕ is good and the distance between any two vertices in G''_ϕ is $O(\log n)$.

Now we can unambiguously decode all the TRUE nodes, FALSE nodes and DUMMY nodes in G''_ϕ.

Lemma 12. *All the* TRUE *nodes (resp.* FALSE *nodes and* DUMMY *nodes) in G''_ϕ can be decoded into the same color.*

Proof. The decoding algorithm works as follows. Pick any survived literal block in G''_ϕ. Set the vector of the TRUE node in the literal block to be $(0, 0, 1)$. Set the direction in S^2 that is closest to the FALSE vector and orthogonal to direction $(0, 0, 1)$ to be $(1, 0, 0)$. Since the TRUE, FALSE and DUMMY nodes in the literal block are interconnected by good edges, the angle between FALSE vector and $(1, 0, 0)$ together with the angle between DUMMY vector and $(0, 1, 0)$ are both $O(\frac{1}{\log n})$.

Now, by the facts that the distance between any pair of TRUE nodes in G''_ϕ is $O(\log n)$, equality gadgets are robust (Lemma 7) and angle on the points in S^2 is a metric space (in particular, angle metric satisfies the triangle inequality), if the constant C in Theorem 8 is small enough, then the angle between any pair of TRUE vectors is at most $\delta_p := \pi/300$. Denote[17]

$$S_{\text{True}} = \{v \in S^2 : \text{angle}(v, (0, 0, 1)) \le \delta_p\}$$

Then all TRUE vectors in G''_ϕ fall into region S_{True}. Similarly define $S_{\text{False}} := \{v \in S^2 : \text{angle}(v, (1, 0, 0)) \le \delta_p\}$ and $S_{\text{Dummy}} := \{v \in S^2 : \text{angle}(v, (0, 1, 0)) \le \delta_p\}$. Then similar argument shows that all FALSE (resp. DUMMY) vectors in G''_ϕ fall into region S_{False} (resp. S_{Dummy}). Finally, since S_{True}, S_{False} and S_{Dummy} are clearly disjoint, we therefore unambiguously decode all the TRUE nodes, FALSE nodes and DUMMY nodes in G''_ϕ. □

[17] Recall that, by our convention, S_{True} is actually the union of two disconnected regions: $\{v \in S^2 : \text{angle}(v, (0, 0, 1)) \le \delta_p\} \cup \{v \in S^2 : \text{angle}(v, (0, 0, -1)) \le \delta_p\}$; that is, S_{True} is a small cap centered around the "north pole" plus a small cap centered around the "south pole". Similar conventions hold for S_{False} and S_{Dummy} as well.

Note that the proof of Lemma 12 makes it possible to decode a truth assignment for all n variables that satisfies at least $(1-\epsilon)$ fraction of the clauses in ϕ, as we now describe. Since the angle between every pair of TRUE node vectors is at most $\delta_p := \pi/300$, the assumption of Lemma 8 is satisfied. Therefore, by Lemma 8, at least one of the three literals in each clause is $\pi/15 + \pi/300 = 21\pi/300$ close to direction $(0,0,1)$. Moreover, since all the survived copies of the same literal are inter-connected by an equality gadget, it follows that each of these at most k copies of the literal is $21\pi/300 + o(1) < \pi/12$ close to $(0,0,1)$. On the other hand, since each literal is connected with its negation by a good edge, its negation is at least $\pi/2 - \pi/12 - o(1) > \pi/3$ far from $(0,0,1)$. Therefore, we can decode unambiguously the truth assignment of each variable as follows: if all the survived copies of x_i are $\pi/12$ close to $(0,0,1)$, assign x_i to TRUE; else, if all the survived copies of $\bar{x}_i$ are $\pi/12$ close to $(1,0,0)$, assign x_i to FALSE; otherwise, x_i is a free variable and can be assigned truth value arbitrarily. This completes the proof of Theorem 8. □

8 The approximation algorithm

Let `DiagonalKernel` be the algorithm that, given a collection of subsets $X_1, X_2, \ldots, X_m \subseteq [N]$, returns the $N \times N$ diagonal matrix K such that $K_{ii} = \frac{1}{m} \cdot |\{j : X_j \ni i\}|$ for all $i \in [N]$. Clearly, `DiagonalKernel` runs in at most $O(mN)$ time.

Theorem 10. *Let ℓ be the log likelihood of the DPP defined by the marginal kernel output by `DiagonalKernel` on examples $X_1, X_2, \ldots, X_m$, and let ℓ^* be the optimal log likelihood achieved by a DPP kernel on the same training set. Then $\ell \le \left(1 + \frac{1}{\log(\frac{m}{a_{max}})}\right)\ell^*$, where $a_{max} := \max_{i\in[N]} |\{j : X_j \ni i\}|$ is the maximum element frequency in examples. It follows immediately that $\ell \le (1+\frac{1+o(1)}{\log N})\ell^*$, when $a_{max} = O(m)/N$. As an unconditional weaker bound, we always have $\ell \le (1 + (1 + \frac{1}{m-1})\log m)\ell^*$.*

Proof. Let ℓ and ℓ^* be the log likelihoods of `DiagonalKernel` kernel and optimal kernel, respectively. In the following we upper bound the ratio $\frac{\ell}{\ell^*}$.

For each $i \in [N]$, define $a_i := |\{j : X_j \ni i\}|$ to be the frequency of element i in the examples. Then K is a diagonal matrix with $K_{ii} = \frac{a_i}{m}$ for each $i \in [N]$. If $\boldsymbol{X}$ is a random variable distributed according to this diagonal DPP, then $\boldsymbol{X}$ is distributed as a product distribution of N independent random variables, with $\Pr[i \in \boldsymbol{X}] = K_{ii}$ for every $i \in [N]$. It follows that, for every subset example $X_j \subseteq [N]$,

$$\Pr_{\boldsymbol{X}\sim\mathcal{P}_K}[\boldsymbol{X} = X_j] = \prod_{i\in X_j} K_{ii} \prod_{i\notin X_j}(1-K_{ii}) = \prod_{i\in X_j}\frac{a_i}{m}\prod_{i\notin X_j}\left(1-\frac{a_i}{m}\right).$$

Since there are exactly a_i sets that contain element i and $m - a_i$ that do not contain element i, the log likelihood of the marginal DPP kernel K is

$$\ell = -\log\left[\prod_{j=1}^{m}\left(\prod_{i\in X_j}\frac{a_i}{m}\prod_{i\notin X_j}(1-\frac{a_i}{m})\right)\right] = -\log\left[\prod_{i=1}^{N}\left(\frac{a_i}{m}\right)^{a_i}\left(1-\frac{a_i}{m}\right)^{m-a_i}\right].$$

By Theorem 5, we can assume that the diagonal entries of a kernel K^* achieving log likelihood ℓ^* match those of the diagonal entries of K. Let $\boldsymbol{X}^*$ be a random variable distributed according to such a marginal kernel, then by Lemma 4 (Hadamard's Inequality) we have

$$\Pr_{\boldsymbol{X}^* \sim \mathcal{P}_{K^*}}[\boldsymbol{X}^* = X_j] \leq \Pr_{\boldsymbol{X}^* \sim \mathcal{P}_{K^*}}[X_j \subseteq \boldsymbol{X}^*] = \det(K^*_{X_j})$$
$$\leq \prod_{i \in X_j} K^*_{ii} = \prod_{i \in X_j} K_{ii} = \prod_{i \in X_j} \frac{a_i}{m}.$$

It follows that

$$\ell^* \geq -\log\left(\prod_{i=1}^{N} \left(\frac{a_i}{m}\right)^{a_i}\right).$$

We now observe that

$$\frac{\ell}{\ell^*} \leq \frac{\log\left(\prod_{i=1}^N \left(\frac{a_i}{m}\right)^{a_i}\left(1-\frac{a_i}{m}\right)^{m-a_i}\right)}{\log\left(\prod_{i=1}^N \left(\frac{a_i}{m}\right)^{a_i}\right)} = 1 + \frac{\log(\prod_{i=1}^N \left(1-\frac{a_i}{m}\right)^{m-a_i})}{\log\left(\prod_{i=1}^N \left(\frac{a_i}{m}\right)^{a_i}\right)}$$
$$= 1 + \frac{\log\left(\prod_{i=1}^N (1-\frac{a_i}{m})^{1-\frac{a_i}{m}}\right)}{\log\left(\prod_{i=1}^N \left(\frac{a_i}{m}\right)^{\frac{a_i}{m}}\right)}$$
$$= 1 + \frac{\sum_{i=1}^N \log\left((1-\frac{a_i}{m})^{1-\frac{a_i}{m}}\right)}{\sum_{i=1}^N \log\left(\left(\frac{a_i}{m}\right)^{\frac{a_i}{m}}\right)}.$$

For $x \in (0,1)$, let $g(x) := -\log\left((1-x)^{1-x}\right)$ and $h(x) := -\log\left(x^x\right)$, and define

$$f(x) := \frac{g(x)}{h(x)} = \frac{\log\left((1-x)^{1-x}\right)}{\log\left(x^x\right)} = \frac{(1-x)\log(1-x)}{x\log x}.$$

Then it is easy to check that, for every $0 < x < 1$, both $g(x)$ and $h(x)$ and hence $f(x)$ are positive; moreover, $f(x)$ is an increasing function in x . Therefore,

$$\frac{\ell}{\ell^*} \leq 1 + \frac{\sum_{i=1}^N g(\frac{a_i}{m})}{\sum_{i=1}^N h(\frac{a_i}{m})} \leq 1 + \max_{i\in[N]} \frac{g(\frac{a_i}{m})}{h(\frac{a_i}{m})}$$
$$= 1 + \max_{i\in[N]} \frac{\log\left((1-\frac{a_i}{m})^{1-\frac{a_i}{m}}\right)}{\log\left(\left(\frac{a_i}{m}\right)^{\frac{a_i}{m}}\right)}$$
$$= 1 + f(\frac{a_{max}}{m}). \tag{17}$$

Using the inequality that $(1+x)\log(1+x) \geq x$ for all $x > -1$, we get that $f(x) \leq -\frac{1}{\log x}$ for all $0 < x < 1$. Thus, when $\frac{a_{max}}{m} \leq \frac{C}{N}$, the kernel output by `DiagonalKernel` satisfies $\ell \leq (1 - \frac{1}{\log(C/N)})\ell^* = (1 + \frac{1}{\log(N/C)})\ell^*$.

When the condition $\frac{a_{max}}{m} \leq \frac{C}{N}$ is not satisfied, in order to obtain an unconditional upper bound on the log likelihood of `DiagonalKernel` , observe that without loss of generality, we may consider only the cases when $a_{max} \leq m-1$. This is because elements occur in all the training subsets will be assigned probability 1 in a maximum likelihood DPP, hence have 1 at the diagonal entry and

0's at all other off-diagonal entries. This means that such elements can be "factored out" from the marginal kernel. Thus, plugging $\frac{a_{max}}{m} = \frac{m-1}{m}$ into (17), we obtain

$$\frac{\ell}{\ell^*} \leq 1 + \frac{\frac{1}{m}\log\frac{1}{m}}{(1-\frac{1}{m})\log(1-\frac{1}{m})} = 1 + \frac{\log m}{(m-1)\log(1+\frac{1}{m-1})} \leq 1 + (1+\frac{1}{m-1})\log m,$$

where in the last step we use again the inequality $(1+x)\log(1+x) \geq x$ for all $x > -1$ (and setting $x = \frac{1}{m-1}$). □

Thus this simple algorithm does surprisingly well, unless there are some high frequent elements that appear in a non-trivial fraction of the training data, which is often not the case in practice. Further, it's worth noting that the training set constructed in our hardness of learning proof also satisfies that $\frac{a_{max}}{m} \leq \frac{C}{N}$ for some constant C.

9 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 7. 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 [BMRU17a] 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.

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

[AC07] 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.

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

[AFT13] 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, 2013.

[AGR16] 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.

[AGR+20] Lucas Anquetil, Mike Gartrell, Alain Rakotomamonjy, Ugo Tanielian, and Clément Calauzènes. Wasserstein learning of determinantal point processes. *arXiv preprint arXiv:2011.09712*, 2020.

[AKFT13] 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, 2013.

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

[BG16] Joshua Brakensiek and Venkatesan Guruswami. New hardness results for graph and hypergraph colorings. In *31st Conference on Computational Complexity (CCC 2016)*, 2016.

[BKS03] 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.

[BMRU17a] Victor-Emmanuel Brunel, Ankur Moitra, Philippe Rigollet, and John Urschel. Maximum likelihood estimation of determinantal point processes. *arXiv preprint arXiv:1701.06501*, 2017.

[BMRU17b] 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, 2017.

[BOT02] 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.

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

[BR05] Alexei Borodin and Eric M. Rains. Eynard-Mehta theorem, Schur process, and their Pfaffian analogs. *Journal of statistical physics*, 121(3-4):291–317, 2005.

[CGGS15] 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.

[CMI09] 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.

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

[DCV19] 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.

[DLM20] 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.

[DMR09] Irit Dinur, Elchanan Mossel, and Oded Regev. Conditional hardness for approximate coloring. *SIAM Journal on Computing*, 39(3):843–873, 2009.

[DRS09] Shaddin Dughmi, Tim Roughgarden, and Mukund Sundararajan. Revenue submodularity. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 243–252, 2009.

[DW18] Michał Dereziński and Manfred K. Warmuth. Reverse iterative volume sampling for linear regression. *The Journal of Machine Learning Research*, 19(1):853–891, 2018.

[Dys62] Freeman J. Dyson. Statistical theory of the energy levels of complex systems. III. *Journal of Mathematical Physics*, 3(1):166–175, 1962.

[GAJ20] Khashayar Gatmiry, Maryam Aliakbarpour, and Stefanie Jegelka. Testing determinantal point processes. *arXiv preprint arXiv:2008.03650*, 2020.

[GBDK19] Mike Gartrell, Victor-Emmanuel Brunel, Elvis Dohmatob, and Syrine Krichene. Learning nonsymmetric determinantal point processes. *arXiv preprint arXiv:1905.12962*, 2019.

[GGR98] 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.

[GK04] 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.

[GKFT14] Jennifer A Gillenwater, Alex Kulesza, Emily Fox, and Ben Taskar. Expectation-maximization for learning determinantal point processes. *Advances in Neural Information Processing Systems*, 27:3149–3157, 2014.

[GKS05] 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.

[GKT12a] 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, 2012.

[GKT12b] 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, 2012.

[GPK16] 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.

[GPK17] Mike Gartrell, Ulrich Paquet, and Noam Koenigstein. Low-rank factorization of determinantal point processes. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[GS12] 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.

[Hås00] Johan Håstad. On bounded occurrence constraint satisfaction. *Information Processing Letters*, 74(1-2):1–6, 2000.

[Hås01] Johan Håstad. Some optimal inapproximability results. *Journal of the ACM*, 48(4):798–859, 2001.

[HJ12] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, second edition, 2012.

[HKPV06] J. Ben Hough, Manjunath Krishnapur, Yuval Peres, and Bálint Virág. Determinantal processes and independence. *Probability surveys*, 3:206–229, 2006.

[HMRW14] Moritz Hardt, Raghu Meka, Prasad Raghavendra, and Benjamin Weitz. Computational limits for matrix completion. In *Conference on Learning Theory*, pages 703–725. PMLR, 2014.

[HPS+10] Gerald Haynes, Catherine Park, Amanda Schaeffer, Jordan Webster, and Lon H Mitchell. Orthogonal vector coloring. *The Electronic Journal of Combinatorics*, 17, 2010.

[IR08] Ilse C. F. Ipsen and Rizwana Rehman. Perturbation bounds for determinants and characteristic polynomials. *SIAM Journal on Matrix Analysis and Applications*, 30(2):762–776, 2008.

[Kan13] Byungkon Kang. Fast determinantal point process sampling with application to clustering. In *Advances in Neural Information Processing Systems 25*, pages 2319–2327, 2013.

[KLQ95] Chun-Wa Ko, Jon Lee, and Maurice Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691, 1995.

[KLS00] Sanjeev Khanna, Nathan Linial, and Shmuel Safra. On the hardness of approximating the chromatic number. *Combinatorica*, 20(3):393–415, 2000.

[KMS98] David R. Karger, Rajeev Motwani, and Madhu Sudan. Approximate graph coloring by semidefinite programming. *J. ACM*, 45(2):246–265, 1998.

[KSG08] 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.

[KT10] Alex Kulesza and Ben Taskar. Structured determinantal point processes. *Advances in neural information processing systems 23*, pages 1171–1179, 2010.

[KT11a] 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, 2011.

[KT11b] Alex Kulesza and Ben Taskar. Learning determinantal point processes. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 419–427, 2011.

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

[Kul12] John A. Kulesza. *Learning with Determinantal Point Processes*. PhD thesis, University of Pennsylvania, 2012.

[Lau09] Monique Laurent. Matrix completion problems. *Encyclopedia of Optimization*, 3:221–229, 2009.

[LB12] 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.

[LCYO16] 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.

[LGD20] Claire Launay, Bruno Galerne, and Agnès Desolneux. Exact sampling of determinantal point processes without eigendecomposition. *Journal of Applied Probability*, 57(4):1198–1221, 2020.

[LJS16a] 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, 2016.

[LJS16b] 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, 2016.

[LMR15] 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.

[Lov79] László Lovász. On the Shannon capacity of a graph. *IEEE Trans. Inf. Theory*, 25(1):1–7, 1979.

[Lov19] László Lovász. *Graphs and geometry*, volume 65. American Mathematical Soc., 2019.

[LPS88] Alexander Lubotzky, Ralph Phillips, and Peter Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988.

[LSS89] L. Lovász, M. Saks, and A. Schrijver. Orthogonal representations and connectivity of graphs. *Linear Algebra and its Applications*, 114-115:439–454, 1989.

[LSS00] 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.

[LV99] László Lovász and Katalin Vesztergombi. Geometric representations of graphs. *Paul Erdos and his Mathematics*, 2, 1999.

[Lyo03] Russell Lyons. Determinantal probability measures. *Publications Mathématiques de l'IHÉS*, 98:167–212, 2003.

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

[Mar88] 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.

[MGS19] 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.

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

[MS16] Zelda E Mariet and Suvrit Sra. Kronecker determinantal point processes. *Advances in Neural Information Processing Systems*, 29:2694–2702, 2016.

[Ohs21] 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.

[OR19] Takayuki Osogami and Rudy Raymond. Determinantal reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33(1), pages 4659–4666, 2019.

[ORG+18] 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.

[Pee96] René Peeters. Orthogonal representations over finite fields and the chromatic number of graphs. *Combinatorica*, 16(3):417–431, 1996.

[PP89] T.D. Parsons and Tomaž Pisanski. Vector representations of graphs. *Discrete Mathematics*, 78(1):143–154, 1989.

[PV88] Leonard Pitt and Leslie G. Valiant. Computational limitations on learning from examples. *Journal of the ACM (JACM)*, 35(4):965–984, 1988.

[RK15] Patrick Rebeschini and Amin Karbasi. Fast mixing for discrete point processes. In *Conference on Learning Theory*, pages 1480–1500. PMLR, 2015.

[RS96] Zeév Rudnick and Peter Sarnak. Zeros of principal L-functions and random matrix theory. *Duke Mathematical Journal*, 81(2):269–322, 1996.

[SG13] 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.

[Sos00] Alexander Soshnikov. Determinantal random point fields. *Russian Mathematical Surveys*, 55(5):923, 2000.

[SZA13] 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.

[Tre01] 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.

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

[XO16] 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.

[YFZ+16] 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.

[YWW+20] 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.

[ZA12] 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.

[ZKL+10] 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.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il