From Sampling to Optimization on Discrete Domains with Applications to Determinant Maximization

Nima Anari ANARI@STANFORD.EDU and **Thuy-Duong Vuong** TDVUONG@STANFORD.EDU
Stanford University, Computer Science Department

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We establish a connection between sampling and optimization on discrete domains. For a family of distributions μ defined on size k subsets of a ground set of elements, that is closed under external fields, we show that rapid mixing of natural local random walks implies the existence of simple approximation algorithms to find $\max \mu(\cdot)$. More precisely, we show that if t-step down-up random walks have spectral gap at least inverse polynomially large, then t-step local search finds $\max \mu(\cdot)$ within a factor of $k^{O(k)}$. As the main application of our result, we show that 2-step local search achieves a nearly-optimal $k^{O(k)}$-factor approximation for MAP inference on nonsymmetric k-DPPs. This is the first nontrivial multiplicative approximation algorithm for this problem.

In our main technical result, we show that an exchange inequality, a concept rooted in discrete convex analysis, can be derived from fast mixing of local random walks. We further advance the state of the art on the mixing of random walks for nonsymmetric DPPs and more generally sector-stable distributions, by obtaining the tightest possible bound on the step size needed for polynomial-time mixing of random walks. We bring the step size down by a factor of 2 compared to prior works, and consequently get a quadratic improvement on the runtime of local search steps; this improvement is potentially of independent interest in sampling applications.

Keywords: Markov chains, approximate optimization, Poincare inequality, exchange inequality, determinant maximization

Contents

1. Introduction

Sampling and optimization are fundamental computational tasks. In continuous settings, sampling and optimization are known to be intimately connected; convex sets, and more generally log-concave distributions, are the natural domains where either task is algorithmically tractable. For a survey of sampling and optimization in continuous settings see (Lovász and Vempala, 2006; Lee et al., 2018).

On discrete/combinatorial domains, the relationship between sampling and optimization is less clear. For example, optimization over independent sets of a bipartite graph is easy, but it is not known how to sample them (see, e.g., Jenssen et al., 2022). The opposite holds for determinantal point processes, which are easy to sample from (see, e.g., Anari et al., 2016) and hard to optimize (Çivril and Magdon-Ismail, 2010).

In this work, we establish a new connection between sampling and optimization in discrete settings. Informally, we show that rapid mixing of natural local random walks that sample from a family of distributions μ, implies that local search algorithms are able to approximately find $\max \mu(\cdot)$, within a nearly-optimal approximation factor.

In our setup, we have a weight function μ on k-sized subsets of a ground set $[n]$, that we denote by $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$.[1] Our goal is to approximately find

$$\max \left\{ \mu(S) \;\middle|\; S \in \binom{[n]}{k} \right\}.$$

We study a family of natural local search algorithms (algorithm 2) for this optimization problem. Formally, each local search algorithm is defined with a parameter $r \geq 0$ which specifies the "radius of the local neighborhood" the algorithm searches over in each iteration. The r-neighborhood of $S \in \binom{[n]}{k}$ are all the sets that can be reached by swapping at most r elements:

$$\mathcal{N}_r(S) := \left\{ T \in \binom{[n]}{k} \;\middle|\; |S - T| \leq r \right\}.$$

Each iteration of local search goes from a set S to the $S' \in \mathcal{N}_r(S)$ which maximizes $\mu(S')$. If we reach a local optimum, i.e., $S = S'$, then we stop and return S as our approximate solution. Despite the simplicity and practical popularity of local search algorithms, it is often challenging to obtain theoretical approximation guarantees for them.

We show that rapid mixing of natural local random walks designed to sample from a family of distributions μ, implies that *local maxima* of μ are *approximate global maxima*.

Definition 1 (Down-Up Random Walk) *For a density $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$, and an integer $\ell \leq k$, the $k \leftrightarrow \ell$ down-up random walk is the sequence of random sets $S_0, S_1, \ldots$ generated as follows:*

1. The restriction of the domain to size k subsets of a ground set should be thought of as a "canonical form"; other discrete domains, including product spaces can be naturally transformed into this domain (Anari et al., 2020).

for $t = 0, 1, \ldots$ **do**
  Select T_t *uniformly at random from subsets of size* ℓ *of* S_t.
  Select S_{t+1} *with probability* $\propto \mu(S_{t+1})$ *from supersets of size* k *of* T_t.

Down-up random walks are time-reversible, always have μ as their stationary measure, and have been widely studied as part of the emerging area of analyzing Markov chains via high-dimensional-expanders (see, e.g., Lubetzky et al., 2017; Kaufman and Oppenheim, 2018; Dinur and Kaufman, 2017; Kaufman and Mass, 2016; Alev and Lau, 2020; Anari et al., 2020). Down-up walks generalize other well-known local random walks like the Glauber dynamics (Anari et al., 2020). Note that the down-up random walk is *local* in the sense that $S_{t+1} \in \mathcal{N}_{k-\ell}(S_t)$. Naturally, we tie mixing of these random walks to local search with neighborhoods of radius $r = k - \ell$.

To state our main result, we need to define the notion of an external field. For a distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ and $\lambda = (\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n_{>0}$, the λ-*external field* applied to μ is another distribution on $\binom{[n]}{k}$, denoted by $\lambda * \mu$, defined as follows:

$$\mathbb{P}_{\lambda * \mu}[S] \propto \mu(S) \cdot \prod_{i \in S} \lambda_i.$$

In this work we consider families of distribution that are closed under external fields, and tie rapid mixing of local random walks on such a family to approximation guarantees of local search. External fields are well-studied operations and often preserve algorithmic tractability of sampling and/or mixing of random walks (see, e.g., Alimohammadi et al., 2021; Eldan and Shamir, 2016; Anari et al., 2021a).

Theorem 2 *Consider a distribution* $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$. *Suppose that for some* $r = O(1)$, *the* $k \leftrightarrow (k - r)$ *down-up random walk on* $\lambda * \mu$ *has spectral gap at least* $k^{-O(1)}$ *for all external fields* $\lambda \in \mathbb{R}^n_{\geq 0}$.[2] *Then any approximate local maximum, that is a set* $S \in \binom{[n]}{k}$ *such that*

$$\mu(S) \geq \Omega(1) \cdot \max \{\mu(T) \mid T \in \mathcal{N}_r(S)\}$$

is a $k^{O(k)}$*-approximate global maximum, i.e.,*

$$\mu(S) \geq k^{-O(k)} \cdot \max \left\{ \mu(T) \;\middle|\; T \in \binom{[n]}{k} \right\}.$$

Moreover, such an approximate local maximum can be found efficiently given oracle access to μ *and a starting point in the support of* μ.

1.1. Applications to Determinant Maximization

The main application of theorem 2 that we highlight in this work is to the problem of MAP inference on nonsymmetric determinantal point proceesses. Determinantal point processes (DPPs) have found many applications in machine learning, such as data summarization (Gong et al., 2014; Lin and Bilmes, 2012), recommender systems (Gartrell et al., 2016; Wilhelm et al., 2018), neural network

2. We remark that the spectral gap of $k^{-O(1)}$ can be replaced with $n^{-O(1)}$ as long as we take n to mean the number of elements in the ground set with a nonzero external field. However, we do not emphasize this relaxed assumption, as in our applications we always have inverse $\mathrm{poly}(k)$ spectral gap.

compression (Mariet and Sra, 2015), kernel approximation (Li et al., 2016), and multi-modal output generation (Elfeki et al., 2019).

A DPP on a set of items $[n] = \{1, \ldots, n\}$ is a probability distribution over subsets $S \subseteq [n]$ parameterized by a matrix $L \in \mathbb{R}^{n \times n}$ where S is chosen with probability proportional to the determinant of the principal submatrix L_S whose columns and rows are indexed by S:

$$\mathbb{P}[Y] \propto \det(L_S).$$

A related and more widely used model, is a k-DPP, where the size of S is constrained to be exactly k. In applications, usually k is much smaller than n. We study k-DPPs in this work (see Kulesza and Taskar, 2012, for a survey on fixed-size DPPs and their applications).

The fundamental optimization problem associated with probabilistic models, including DPPs, is to find the maximum a posteriori (MAP) configuration (see, e.g., Gillenwater et al., 2012), that is $\max \{\mathbb{P}[S]\}$. The main application of our results is to the problem of MAP inference on DPPs.

Most of the literature on DPPs assumes the matrix L is symmetric, but such symmetric matrices can only encode repulsive (negatively correlated) interactions between items (Borcea et al., 2009). This limits their modeling power. To remedy this, more recent works have considered the more general class of nonsymmetric DPPs (NDPPs) which can additionally capture positive correlations (Brunel, 2018; Gartrell et al., 2019). Gartrell et al. (2019) observed that for nonsymmetric positive semidefinite (nPSD) matrices L, the principal minors are still nonnegative and therefore the distribution is well-defined.

Definition 3 *A matrix $L \in \mathbb{R}^{n \times n}$ is nonsymmetric positive semidefinite (nPSD) if $L + L^\intercal \succeq 0$.*

Throughout, we will consider only NDPPs with nPSD kernels and refer to them simply as NDPPs or nonsymmetric DPPs. Alimohammadi et al. (2021) showed how to efficiently sample from fixed-size NDPPs using down-up random walks.

Theorem 4 ((Alimohammadi et al., 2021)) *The $k \leftrightarrow (k-4)$ down-up random walk on a nonsymmetric k-DPP has spectral gap $\Omega(k^{-4})$.*

k-NDPPs are closed under external fields, because applying an external field L is equivalent to multiplication of the kernel matrix on the left and the right with a diagonal matrix (see, e.g., Alimohammadi et al., 2021). This immediately gives a corollary of theorem 2: 4-step local search can approximately find the maximum of a k-NDPP within a factor of $k^{O(k)}$. This corollary would already give the first nontrivial unconditional multiplicative approximation for MAP inference on NDPPs. However, we go even beyond the results of Alimohammadi et al. (2021), and show that the $k \leftrightarrow (k-4)$ down-up walk can be replaced by the $k \leftrightarrow (k-2)$ down-up walk.

Theorem 5 *The $k \leftrightarrow (k-2)$ down-up random walk on a k-NDPP has spectral gap $\Omega(k^{-O(1)})$.*

Remark 6 *Note that we only improve the step size, actually at the expense of a worse (but still inverse-polynomially large) spectral gap. Nevertheless, we conjecture the $k \leftrightarrow (k-2)$ down-up walk should have spectral gap $\Omega(k^{-2})$ and leave this as an open question. The importance of this question is mostly in sampling applications; in the optimization problems we study here, the improvement in step size directly results in much faster local search algorithms, while still maintaining a $k^{O(k)}$ approximation factor.*

More generally, we obtain the tightest possible step size for rapid mixing of down-up random walks for all sector-stable (see section 2.5 for definition) distributions. The proof of theorem 5 is deferred to section A. As a corollary, we obtain the following result for optimization.

Corollary 7 *Any (approximate) 2-local optimum of a k-NDPP is a global optimum within an approximation factor of $k^{O(k)}$.*

To appreciate the importance of the improvement from 4 to 2, note that in 2-local search, in each step we have to search over $\sim (kn)^2$ many possible sets, which is significantly smaller than $\sim (kn)^4$. Theorem 5 is potentially of independent interest in sampling applications. We state concrete runtimes below, but we emphasize that we did not attempt to optimize the runtime of the algorithm; it is likely that by employing linear-algebraic tricks one can further shave polynomial factors.

Theorem 8 *There is a polynomial time algorithm that on input $L \in \mathbb{R}^{n\times n}$ that is nPSD, outputs a set of indices $S \in \binom{[n]}{k}$ guaranteeing*

$$\det(L_S) \geq k^{-O(k)} \cdot \max\left\{ \det(L_S) \;\middle|\; S \in \binom{[n]}{k} \right\}.$$

Moreover, the algorithm runs in $O(n^4k + n^2k^5 \log n)$ time given the entries of L, and $O(n^2d^2k + n^2d^2k^3 \log n)$ time given a rank-d decomposition of L, i.e., $L = BCB^\intercal$ with $B \in \mathbb{R}^{n\times d}, C \in \mathbb{R}^{d\times d}$.

Remark 9 *We remark that the approximation factor of $k^{O(k)}$ is nearly optimal amongst efficient algorithms. The special case of symmetric determinantal point processes was shown to be hard to approximate within a factor of c^k for some constant $c > 1$ (Çivril and Magdon-Ismail, 2010). Further, the factor of $k^{O(k)}$ is tight for local search, even in the special case of symmetric determinantal point processes (Anari and Vuong, 2020). Additionally, the step size of 2 for local search is also optimal, as we demonstrate in section C.*

1.2. Related Work

For continuous distributions/functions defined on the Euclidean space $\mathbb{R}^n$, there is a rich history connecting sampling and optimization. Bertsimas and Vempala (2004); Lovász and Vempala (2006) show how to solve convex programs by performing random walks. Besides general black-box reductions, popular algorithms for sampling such as Langevin dynamics are known to be intimately connected to popular algorithms for optimization such as gradient descent (Dalalyan, 2017).

The relationship between optimization and sampling is not as well understood in discrete settings. For example, the problem of sampling independent sets from bipartite graphs, equivalent to the so-called #BIS problem is believed to be hard and is often used as a hardness assumption (Galanis et al., 2016), whereas maximum independent sets in bipartite graphs are easy to find due to connections with bipartite maximum matching. On the other hand, for DPPs, sampling is known to be easy (Kulesza and Taskar, 2012), but optimization is NP-hard within subexponential factors (Çivril and Magdon-Ismail, 2010). We are not aware of prior works that study the optimization algorithms derived from sampling in general discrete settings, but in many specific cases, optimization and sampling are studied hand-in-hand. For some recent examples on the Ising model and the Sherrington-Kirkpatrick spin glass, which correspond to the optimization problem of Max-Cut, see (Montanari, 2021; Koehler et al., 2022).

Our main application is the optimization problem of MAP inference for NDPPs. Earlier, the work of Gartrell et al. (2020) studied the greedy algorithm for MAP inference on NDPPs, as greedy was previously used successfully for symmetric DPPs. The greedy algorithm (algorithm 1) runs for k iterations, in each iteration adding the item that most increases the DPP score. Unfortunately, the greedy algorithm does not achieve any finite approximation guarantees for NDPPs. As such, Gartrell et al. (2020), put extra, potentially limiting, assumptions on the matrix L, in order to be able to analyze the performance of greedy.

Algorithm 1: Standard greedy for DPPs

Initialize $S \leftarrow \emptyset$.
while $|S| < k$ **do**
  Pick $i \notin S$ that maximizes $\det(L_{S \cup \{i\}})$, and update $S \leftarrow S \cup \{i\}$.

Although greedy obtains a $k^{O(k)}$-approximation for symmetric DPPs (Çivril and Magdon-Ismail, 2010), it cannot achieve even a finite approximation factor for NDPPs. For example, on a skew-symmetric matrix L, i.e., $L = -L^\intercal$, since all odd-sized principle minors of X are zero, algorithm 1 would necessarily resort to picking an arbitrary/random item at every other iteration, which can result in an arbitrarily bad final answer. Many variants of the greedy algorithm, including 1-step local search, fail as well. For more details see section C.

Our work gives the first nontrivial *multiplicative* approximation algorithm for MAP inference on NDPPs, and we do not put any assumptions on L. Prior related work (Gartrell et al., 2020) obtained multiplicative approximations for $\log\det(L_S)$ under further assumptions on L; this is often a weaker approximation guarantee when OPT is sufficiently large – roughly super-exponentially large in k. The assumptions behind prior works implicitly imply that OPT is at least exponentially large in k, making the unconditional factor $k^{O(k)}$-approximation guarantee of this work attractive.

The $k^{O(k)}$ approximation factor matches that of the standard greedy heuristic on *symmetric* DPPs, as well as the guarantee of other simple heuristics proposed for *symmetric* DPPs (Çivril and Magdon-Ismail, 2010; Kathuria and Deshpande, 2016). As mentioned earlier, Çivril and Magdon-Ismail (2010)'s greedy and Kathuria and Deshpande (2016)'s local search algorithm do not achieve any finite approximation factor for NDPPs. Our result is incomparable to Gartrell et al. (2020) as (i) multiplicative approximations for maximizing $\log\det(L_S)$ do not imply similar results for $\det(L_S)$, and (ii) we place no additional assumption on L. As demonstrated in section C, our approximation guarantees hold for matrices L where algorithm 1 fails to achieve even a finite approximation factor.

Our local search algorithm for NDPPs searches over 2-neighborhoods, unlike most prior related works which typically use 1 neighborhoods; using 2-neighborhoods is necessary, and is compatible with intuition from prior work of Anari and Vuong (2020) who first studied 2-neighborhood local search for the related problem of finding the maximum $k \times k$ subdeterminant of a rectangular matrix. Unlike (Anari and Vuong, 2020), our analysis of local search is not based on algebraic identities, which we believe do not have a counterpart in the world of NDPPs, but rather we resort to mixing properties of random walks.

1.3. Techniques

Our main tool for proving theorem 2 is a form of (approximate) exchange inequality. Exchange inequalities have been traditionally been studied in discrete convex analysis (Murota et al., 2016), but

have recently been extended and used in sampling (Anari et al., 2021b) and optimization (Anari and Vuong, 2020) problems beyond the reach of traditional discrete convex analysis. Unlike prior works, here we go in the opposite direction and show that efficient sampling implies a form of exchange inquality. To prove theorem 2, we set the external field λ appropriately to focus the distribution on the two sets found by local search and the global optimum, and use the lower bound on the spectral gap of the down-up walk on $\lambda * \mu$ to derive our approximate exchange property lemma 30.

We then show that approximate exchange implies the desired approximation factor for local search (proposition 31). Since nonsymmetric DPPs are $1/4$-fractionally log-concave (Alimohammadi et al., 2021), theorem 2 already implies an efficient algorithm (algorithm 2 with $r = 4$) to get $k^{O(k)}$-approximation factor for the MAP inference problem on nonsymmetric DPPs. We can further improve the the local search radius r to 2, and get a faster algorithm that matches the runtime stated in theorem 8 by showing a stronger approximate exchange property (**??**).

As a tangential implication of our techniques, in the appendix, section D, we demonstrate how to construct core-sets for optimization for log-concave polynomials, a slight generalization of earlier techniques of Mahabadi et al. (2019).

2. Preliminaries

We use $[n]$ to denote the set $\{1, \ldots, n\}$ and $\binom{[n]}{k}$ to denote the family of size k subsets of $[n]$. We use $\mathbb{1}$ to denote the all 1 vector. When n is clear from context, we use $\mathbb{1}_S \in \mathbb{R}^n$ to denote the indicator vector of the set $S \subseteq [n]$, having a coordinate of 0 everywhere except for elements of S, where the coordinate is 1. For sets S, T of the same size we define their *distance* to be $d(S,T) := |S \Delta T|/2 = |S \setminus T| = |T \setminus S|$. With this notion of distance, we can define neighborhoods:

Definition 10 *For $r \geq 0$ the r-neighborhood of $S \in \binom{[n]}{k}$ is*

$$\mathcal{N}_r(S) := \left\{ T \in \binom{[n]}{k} \;\middle|\; d(S,T) \leq r \right\}.$$

2.1. Determinantal Point Processes (DPPs)

A DPP on a set of n items is a probability distribution over subsets $Y \subseteq [n]$. It is parameterized by a matrix $L \in \mathbb{R}^{n \times n}$: $\mathbb{P}_L[Y] \propto \det(L_Y)$, where L_Y denote the principle submatrix whose columns and rows are indexed by Y. We call L the kernel matrix.

For $Y \subseteq [n]$, if we condition the distribution $\mathbb{P}_L$ on the event that items in Y are included in the sample, we still get a DPP; the new kernel is given by the Schur complement $L^Y = L_{\tilde{Y}} - L_{\tilde{Y},Y} L_{Y,Y}^{-1} L_{Y,\tilde{Y}}$ where $\tilde{Y} = [n] \setminus Y$.

Given a cardinality constraint k, the k-DPP paremeterized by L is a distribution over subsets of size k of Y defined by $\mathbb{P}_L^k[Y] = \frac{\det(L_Y)}{\sum_{|Y'|=k} \det(L_{Y'})}$.

To ensure that $\mathbb{P}_L$ defines a probability distribution, all principal minors of L must be non-negative: $\det(L_S) \geq 0$. Matrices that satisfy this property are called P_0-matrices (Fang, 1989, Definition 1). Any nonsymmetric (or symmetric) PSD matrix is automatically P_0 (Gartrell et al., 2019, Lemma 1).

We say a NDPP kernel $L \in \mathbb{R}^{n \times n}$ has a low-rank decomposition (Gartrell et al., 2019, 2020) if L can be written as $L = BCB^\intercal$ for some $d \leq n$, $B \in \mathbb{R}^{n \times d}, C \in \mathbb{R}^{d \times d}$. Clearly, $\mathrm{rank}(L) = d$, and we say $L = BCB^\intercal$ is a rank-d decomposition of L. We will need the following identity, which is

derived from Schur complements; it has previously appeared in (Gartrell et al., 2020). For $S \subseteq [n]$, let B_S denote the sub-matrix of B consisting of rows in S; then $L_S = B_S C B_S^\intercal$ and

$$\begin{aligned} \det(L_{Y \cup D}) &= \det(L_Y) \det(L_D - L_{D,Y} L_Y^{-1} L_{Y,D}) \\ &= \det(L_Y) \det(L_D - B_D C (B_Y^\intercal L_Y^{-1} B_Y) C B_D^T). \end{aligned} \tag{1}$$

Given $\det(L_Y)$ and L_Y^{-1}, we can compute $\det(L_{Y \cup D})$ in $O(|D|d^2 + |D|^2 d + |D|^3)$ time.

2.2. MAP Inference

Given a density $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$, optimization with respect to μ or MAP inference on μ is to find

$$S^* := \arg\max \left\{ \mu(S) \;\middle|\; S \in \binom{[n]}{k} \right\}.$$

Throughout the paper, we let $\mathrm{OPT} := \max \{\mu(S)\}$. We say that an algorithm gives a factor c-approximation for MAP inference on μ if it outputs $\hat{S} \in \binom{[n]}{k}$ such that $c \cdot \mu(\hat{S}) \geq \mathrm{OPT}$. When μ is defined by a DPP, i.e., $\mu(S) = \det(L_{S,S})$ for a $n \times n$ matrix L, MAP inference on μ is also called the determinant maximization problem (e.g., see Mahabadi et al., 2019).

2.3. Markov Chains and Spectral Gap

A Markov chain on a state space Ω is defined by a row-stochastic matrix $P \in \mathbb{R}^{\Omega \times \Omega}$. We view distributions μ on Ω as row vectors, and as such μP would be the distribution after one transition according to P, if we started from a sample of μ. A distribution μ is stationary if $\mu P = \mu$. Under mild assumptions on P (ergodicity), stationary distributions are unique and νP^t converges to this stationary distribution as $t \to \infty$ for any starting ν (Levin and Peres, 2017).

We will only consider reversible Markov chain. A Markov chain P is (time-)reversible if

$$\mu(x) P(x, y) = \mu(y) P(y, x) \quad \forall x, y \in \Omega.$$

Reversible Markov chains have only real eigenvalues, and we call the difference between the top two eigenvalues $1 - \lambda_2(P)$ the spectral gap. The spectral gap is intimately connected to the mixing time and convergence properties of a Markov chain.

The conductance[3] of a subset S of states in a Markov chain is

$$\Phi(S) = \frac{Q(S, \Omega \setminus S)}{\mu(S)}$$

where $Q(S, \Omega \setminus S) = \sum_{x \in S, y \in \Omega \setminus S} \mu(x) P(x, y)$ is the ergodic flow between S and $\Omega \setminus S$, and $\mu(S) = \sum_{x \in S} \mu(x)$. The conductance of a Markov chain is defined as the minimum conductance over all subsets S with $\mu(S) \leq 1/2$, i.e.,

$$\Phi = \min \{\Phi(S) \mid S \subseteq \Omega, \mu(S) \leq 1/2\}.$$

Theorem 11 (Cheeger's Inequality (see, e.g., Levin and Peres, 2017, Thm. 13.10)) *Let λ_2 be the second largest eigenvalue of the transition matrix P. Then*

$$\Phi^2/2 \leq 1 - \lambda_2 \leq 2\Phi.$$

3. also known as bottleneck ratio in (Levin and Peres, 2017)

2.4. The Down-Up Random Walk

Consider a distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$. The down-up walk is given by the composition of two row-stochastic operators, known as the down and up operators.

Definition 12 (Down Operator) *For a ground set Ω, and $|\Omega| \geq k \geq \ell$, define the down operator $D_{k\to\ell} \in \mathbb{R}^{\binom{\Omega}{k}\times\binom{\Omega}{\ell}}$ as*

$$D_{k\to\ell}(S,T) = \begin{cases} \frac{1}{\binom{k}{\ell}} & \text{if } T \subseteq S, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 13 (Up Operator) *For a ground set Ω, $|\Omega| \geq k \geq \ell$, and density $\mu : \binom{\Omega}{k} \to \mathbb{R}_{\geq 0}$, define the up operator $U_{\ell\to k} \in \mathbb{R}^{\binom{\Omega}{\ell}\times\binom{\Omega}{k}}$ as*

$$U_{\ell\to k}(T,S) = \begin{cases} \frac{\mu(S)}{\sum_{S'\supseteq T}\mu(S')} & \text{if } T \subseteq S, \\ 0 & \text{otherwise.} \end{cases}$$

If we define $\mu_k = \mu$ and more generally μ_ℓ as $\mu_k D_{k\to\ell}$, then the down and up operators satisfy

$$\mu_k(S)D_{k\to\ell}(S,T) = \mu_\ell(T)U_{\ell\to k}(T,S).$$

This property ensures that the composition of the down and up operators have the appropriate μ as a stationary distribution, are reversible, and have nonnegative real eigenvalues (see, e.g., Kaufman and Oppenheim, 2018; Alev and Lau, 2020).

Definition 14 (Down-Up Walk) *For a ground set Ω, $|\Omega| \geq k \geq \ell$, and density $\mu : \binom{\Omega}{k} \to \mathbb{R}_{\geq 0}$, the $k \leftrightarrow \ell$ down-up walk is defined by the row-stochastic matrix $D_{k\to\ell}U_{\ell\to k}$.*

2.5. Geometry of Polynomials

For a density $\mu : 2^{[n]} \to \mathbb{R}_{\geq 0}$, the *generating polynomial* of μ is defined as

$$g_\mu(z_1,\ldots,z_n) = \sum_{S\in 2^{[n]}} \mu(S)\prod_{i\in S} z_i.$$

We use $\mathbb{F}[z_1,\ldots,z_n]$ to denote n-variate polynomials with coefficients from $\mathbb{F}$, where we usually take $\mathbb{F}$ to be $\mathbb{R}$ or $\mathbb{C}$. We denote the degree of a polynomial g by $\deg(g)$. We call a polynomial homogeneous of degree k if all nonzero terms in it are of degree k.

Definition 15 (Stability) *For an open subset $U \subseteq \mathbb{C}^n$, we call a polynomial $g \in \mathbb{C}[z_1,\ldots,z_n]$ U-stable if $(z_1,\ldots,z_n) \in U \implies g(z_1,\ldots,z_n) \neq 0$.*

For convenience, we also call the identically 0 polynomial U-stable. This ensures that limits of U-stable polynomials are U-stable. For convenience, when n is clear from context, we abbreviate stability w.r.t. regions of the form $U \times U \times \cdots \times U$ where $U \subseteq \mathbb{C}$ simply as U-stability.

Our choice of the region U in this work is the product of open sectors in the complex plane.

Definition 16 (Sectors) *The open sector of aperture $\alpha\pi$ centered around the positive real axis is:*

$$\Gamma_\alpha := \{\exp(x+iy) \mid x \in \mathbb{R}, y \in (-\alpha\pi/2, \alpha\pi/2)\}.$$

Note that Γ_1 is the right-half-plane, and Γ_1-stability is the same as the more well-studied notion of Hurwitz-stability (see, e.g., Brändén, 2007). Another closely related notion is that of real-stability where the region U is the upper-half-plane $\{z \mid \mathrm{Im}(z) > 0\}$ (see, e.g., Borcea et al., 2009). For *homogeneous* polynomials, stability w.r.t. U is the same as stability w.r.t. any rotation/scaling of U; so Hurwitz-stability and real-stability are the same for *homogeneous* polynomials. Real-stability of the generating polynomial of a distriution is also known as being strongly Rayleigh, and we use the two notions interchangeably.

We use α-sector-stable as a shorthand for Γ_α-stable. Naturally, we call a distribution α-sector-stable if its generating polynomial is α-sector-stable.

Proposition 17 ((Alimohammadi et al., 2021)) *These operations preserve Γ_α-sector-stability on homogeneous multi-affine polynomials:*

1. *Specialization: $g \mapsto g(a, z_2, \ldots, z_n)$, for $a \in \bar{\Gamma}_\alpha$, where $\bar{\Gamma}_\alpha$ is the closure of Γ_α in $\mathbb{C}$.*
2. *Derivative: $g \mapsto \frac{\partial}{\partial z_1} g(z_1, \cdots, z_n)$.*
3. *Scaling: $g \mapsto g(\lambda_1 z_1, \ldots, \lambda_n z_n)$, for $\lambda \in \mathbb{R}^n_{\geq 0}$.*

We state some examples of sector stable distributions.

Lemma 18 ((Alimohammadi et al., 2021)) *Consider $L \in \mathbb{R}^{n \times n}$ that is nPSD, i.e., $L + L^T \succcurlyeq 0$, then $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ defined by $\mu(S) = \det(L_S)$ is $1/2$-sector-stable.*

Lemma 19 ((Alimohammadi et al., 2021)) *Given a density $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ and a partition $T_1 \cup T_2 \cup \cdots \cup T_s = [n]$, and numbers $c_1, \ldots, c_s \in \mathbb{Z}_{\geq 0}$, let the partition-constrained density $\mu_{T,c}$ be μ restricted to sets $S \in \binom{[n]}{k}$ where $|S \cap T_i| = c_i$. When μ is strongly Rayleigh, $\mu_{T,c}$ is $1/2^s$-sector-stable.*

We next define *log-concavity* for distributions over size-k subsets of n elements, and its direct generalization, *fractional log-concavity*.

Definition 20 *We say a probability distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ is* log-concave *if its generating polynomial is a log-concave function over $\mathbb{R}_{\geq 0}$, i.e., if $\log g_\mu(z_1, \ldots, z_n)$ is concave over $\mathbb{R}_{\geq 0}$.*

The notion of fractional log-concavity (Alimohammadi et al., 2021) generalizes the above.

Definition 21 ((Alimohammadi et al., 2021)) *A probability distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ is* α-fractionally-log-concave *if $g_\mu(z_1^\alpha, \ldots, z_n^\alpha)$ is log-concave for $z_1, \ldots, z_n \in \mathbb{R}^n_{\geq 0}$.*

? showed that for homogeneous multiaffine polynomials, real-stability implies log-concavity. A similar relationship holds for sector stability and fractional log-concavity.

Lemma 22 (Lemma 67 from (Alimohammadi et al., 2021)) *If a polynomial g is α-sector-stable, then it is $\frac{\alpha}{2}$-fractionally-log-concave.*

We note that scaling preserves α-log-concavity of homogeneous distributions (Alimohammadi et al., 2021), i.e., if μ is α-fractionally-log-concave, then so is $\lambda * \mu$ for all $\lambda \in \mathbb{R}^n_{\geq 0}$.

Theorem 23 ((Alimohammadi et al., 2021; Anari et al., 2021a)) *Suppose $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ is α-fractionally log-concave. The $k \leftrightarrow (k - \lceil 1/\alpha \rceil)$-down-up-walk on μ has spectral gap $\geq \Omega(k^{-1/\alpha})$.*

3. Local Search Algorithms

In this section, we show how to efficiently find a local optima[4] of a given distribution μ. We run a two stage algorithm:

1. First, we find some "good" initial subset $S_0 \in \binom{[n]}{k}$, i.e., one such that the ratio $\mathrm{OPT}/\mu(S_0)$ is bounded by $2^{\mathrm{poly}(n,k)}$ (see lemma 26).
2. Then, for a suitably chosen radius r, we run a simple local search (algorithm 2) that starts with $S \leftarrow S_0$, and finds better and better solutions by swapping *at most* r elements in S for elements outside of S until no more improvement in terms of $\mu(S)$ can be found.

To ensure that our algorithm terminates within polynomial time, we will only take improvements that increase the value by at least a factor of $1/\zeta$, which can we set to a constant like 2.

Algorithm 2: Local-Search-r (LS_r)

input : improvement factor $\zeta \in [0,1]$ and starting point $S_0 \in \binom{[n]}{k}$ with $\mu(S_0) > 0$.
Initialize $S \leftarrow S_0$.
while $\mu(S) < \zeta \cdot \mu(T)$ *for some* $T \in \mathcal{N}_r(S)$ **do**
  Update $S \leftarrow \arg\max \{\mu(T) \mid T \in \mathcal{N}_r(S)\}$.

We prove the algorithmic part of theorem 2, that with a suitable choice for S_0, algorithm 2 runs in polynomial time.

Proposition 24 *The number of steps taken by algorithm 2 with $r = O(1)$ is at most*

$$\log_{1/\zeta}\left(\mathrm{OPT}/\mu(S_0)\right).$$

Each step can be implemented using $O((nk)^r)$ oracle queries to μ.

Proof Each iteration improves $\mu(S)$ by a factor of at least $1/\alpha$. On the other hand, this value can never exceed OPT, and it starts as $\mu(S_0) > 0$.

Clearly, to perform local search in the r-neighborhood of a set S, we only need to query $\mu((S \setminus U_1) \cup U_2)$ for $U_1 \in \binom{S}{\leq r}$ and $U_2 \in \binom{[n]}{\leq r}$. The total number of such queries is $O((nk)^r)$. ■

Definition 25 *For $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ and $\zeta > 0$, we say $S \in \binom{[n]}{k}$ is a (r,ζ)-local maximum w.r.t. μ if $\mu(S) \geq \zeta \cdot \mu(T)$ for all $T \in \mathcal{N}_r(S)$.*

Clearly, when algorithm 2 terminates, the output is a (r,ζ)-local maximum.

Next, we show how to obtain a "good" initialization S_0 by a simple greedy algorithm, which we call Induced-Greedy, that is based on maximizing the *marginal gain* defined by the distribution on size $\leq k$ subsets. This gain is *induced* by the distribution μ, as defined below.

For this algorithm, we extend the domain of our distribution μ from subsets of size k to all subsets of size *at most* k as follows: For subset T of $[n]$ of size $\leq k$, let $\mu(T) = \sum_{S \in \binom{[n]}{k} : S \supseteq T} \mu(S)$.

4. More precisely, we show how to find an approximate local optima, which is sufficient for our purpose.

Algorithm 3: Induced-Greedy

Initialize $S \leftarrow \emptyset$.
while $|S| < k$ **do**
  Pick $i \notin S$ that maximizes $\mu(S \cup \{i\})$ and update $S \leftarrow S \cup \{i\}$.

Lemma 26 *Algorithm 3 returns S with*

$$O(n^k) \cdot \mu(S) \geq \mathrm{OPT}.$$

Proof For $j \in [k]$, let i_j be the element added to S at the j-th iteration of the while loop. Let $S_0 = \emptyset$, $S_j = S_{j-1} \cup \{i_j\}$. Observe that $|S_j| = j$ and for each $j \geq 0$

$$\mu(S_j) = \frac{1}{k - |S_j|} \sum_{i \notin S_j} \mu(S_j \cup \{i\}) \leq \frac{n-j}{k-j} \mu(S_{j+1}).$$

Thus $\binom{n}{k}\mu(S_k) \geq \mu(S_0) = \mu(\emptyset) = \sum_{S' \in \binom{[n]}{k}} \mu(S') \geq \mathrm{OPT}$. ■

Remark 27 *In algorithm 3, it is enough to find i that approximately maximizes $\mu(S \cup i)$, i.e., for some constant $\zeta \in (0,1)$, $\mu(S \cup i) \geq \zeta \mu(S \cup j)$ for all $j \notin S$. In that case, lemma 26 still holds, and algorithm 3 can be efficiently implemented given access to efficient algorithms that approximately sample from $\lambda * \mu$ for $\lambda \in \mathbb{R}^n_{\geq 0}$. Indeed, note that $\mu(S \cup i)/\mu(S)$ is the marginal of $\lambda * \mu$ where*

$$\lambda_i = \begin{cases} \infty & \text{for } i \in S, \\ 1 & \text{else.} \end{cases}$$

Thus, Induced-Greedy can be implemented by a sampling algorithm for our family of distributions.

4. From Sampling to Optimization

In this section, we prove theorem 2.

Definition 28 (r-exchange) *For $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$, $r \in \mathbb{N}$ and $S, T \in \binom{[n]}{k}$, we let*

$$\mathcal{E}^r(S,T) := \{U \subseteq S \Delta T \mid |U \cap S| = |U \cap T| = r\}$$

be the set of all r-exchanges between S and T.

Definition 29 (Weak (r,β)-approximate exchange) *We say a distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ satisfies* weak (r,β)-approximate exchange *if for any $S, T \in \binom{[n]}{k}$, there exists $s \in \{1, \cdots, r\}$ and $U \in \mathcal{E}^s(S,T)$ such that*

$$\mu(S) \leq \beta \cdot \mu(S \Delta U) \left(\frac{\mu(S)}{\mu(T)}\right)^{s/d(S,T)}$$

Lemma 30 *Consider $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ such that for all external field $\lambda \in \mathbb{R}^n_{\geq 0}$, the conductance of the $k \leftrightarrow (k-r)$-down-up walk on $\lambda * \mu$ is at least $\Omega(k^{-c})$. Then μ satisfies weak $(r, O(k^{r+c}))$-approximate exchange.*

Proposition 31 *If $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ satisfies weak (r, β)-approximate exchange then any (r, ζ)-local max with respect to μ is also an $O((\beta/\zeta)^k)$-approximate global max.*

In particular, when μ is α-fractionally log-concave, lemma 30 and proposition 31 hold with $r = \lceil 1/\alpha \rceil$ and $c = 1/\alpha$ and $\beta = O(k^{r+c})$. The local search guarantee in theorem 2 follows from theorems 11 and 23, lemma 30, and proposition 31, and the runtime bound follows from remark 27 and proposition 24.

Proof [of lemma 30] If $d(S,T) \leq r$ then the lemma holds trivially by setting $U = S \Delta T$. In what follows, we assume $d(S,T) \geq r$. W.l.o.g. we can assume that $S = \{1, \ldots, t\} \cup C$ and $T = \{t+1, \ldots, 2t\} \cup C$ with $C = \{2t+1, \ldots, t+k\}$ and $t = d(S,T)$.

Consider distribution $\mu' = \lambda * \mu$ with $\lambda_i = \begin{cases} 1 & \text{if } 1 \leq i \leq t \\ (\mu(S)/\mu(T))^t & \text{if } t+1 \leq i \leq 2t \\ \infty & \text{if } 2t+1 \leq i \leq t+k \\ 0 & \text{else} \end{cases}$.

Note that μ' is supported on $W \in \binom{[n]}{k}$ where $(S \cap T) = C \subseteq W \subseteq (S \cup T)$. Let Φ be the conductance of the $k \leftrightarrow (k-r)$-down-up walk on μ', then $\Phi \geq \Omega(k^{-c})$. On the other hand, since $\mu'(S) = \mu'(T) = \mu(S) \leq \frac{\sum_W \mu'(W)}{2}$, we have that by definition of Φ,

$$\Phi = \min_{\mu'(\mathcal{S}) \leq \mu'(\Omega)/2} \frac{Q(\mathcal{S}, \Omega \setminus \mathcal{S})}{\mu'(\mathcal{S})} \leq \frac{Q(\{S\}, \Omega \setminus \{S\})}{\mu'(S)}$$

where we can rewrite $Q(\{S\}, \Omega \setminus \{S\})$ as

$$Q(\{S\}, \Omega \setminus \{S\}) = \mu'(S) \frac{1}{\binom{k}{r}} \sum_{U_1 \in \binom{S}{r}} \sum_{\substack{W \supseteq S \setminus U_1 \\ W \in \mathrm{supp}(\mu') \setminus \{S\}}} \frac{\mu'(W)}{\mu'(S \setminus U_1)}$$

where $\mu'(S \setminus U_1) = \sum_{W \in \binom{[n]}{k}, W \supseteq S \setminus U_1} \mu'(W)$.

Note that

$$\left\{ W \in \mathrm{supp}(\mu') \setminus \{S\} \;\middle|\; W \supseteq S \setminus U_1 \right\} \subseteq \left\{ (S \setminus U_1) \cup U_2 \;\middle|\; U_2 \in \binom{T \cup U_1}{r} \setminus \{U_1\} \right\}$$

thus

$$\bigcup_{U_1 \in \binom{S}{r}} \left\{ W \in \mathrm{supp}(\mu') \setminus \{S\} \;\middle|\; W \supseteq S \setminus U_1 \right\} \subseteq \left\{ S \Delta U \;\middle|\; U \in \bigcup_{s=1}^{r} \mathcal{E}^s(S,T) \right\}.$$

Moreover, $|\binom{S}{r}| = \binom{k}{r}$ and for each $U_1 \in \binom{S}{r}$, the cardinality of $\{W \in \mathrm{supp}(\mu') \setminus \{S\} \mid W \supseteq S \setminus U_1\}$ is at most $\leq \binom{k+r}{r} - 1 \leq k^r$. Hence, there must exist $s \in [r]$ and $U \in \mathcal{E}^s(S,T)$ such that

$$\frac{\mu'(S \Delta U)}{\mu'(S \setminus U)} \geq \frac{1}{k^r} \frac{Q(\{S\}, \Omega \setminus \{S\})}{\mu'(S)} \geq \Omega(k^{-(r+c)}).$$

Thus

$$\mu(S) = \mu'(S) \leq \mu'(S \setminus U) \leq O(k^{r+c})\mu'(S\Delta U) = O(k^{r+c})\mu(S\Delta U)(\frac{\mu(S)}{\mu(T)})^{s/t}.$$

■

Proof [of proposition 31] Apply lemma 30 for S being a (r, ζ)-local max and $T := \arg\max \mu(W)$. Let $t = d(S, T)$. For some $s \in [r]$ and $U \in \mathcal{E}^s(S, T)$

$$\mu(S) \leq O(k^{r+c})\mu(S\Delta U)(\frac{\mu(S)}{\mu(T)})^{s/t} \leq O(k^{r+c}/\zeta)\mu(S)(\frac{\mu(S)}{\mu(T)})^{s/t}$$

where the inequality follows from definition of (r, ζ)-local max. Divide both sides by $\mu(S) > 0$, we get

$$\mu(T) \leq O(k^{r+c}/\zeta)^{t/s}\mu(S) \leq O(k^{r+c}/\zeta)^{k}\mu(S).$$

where we use the fact that $t/s \leq k$. ■

Acknowledgments

Nima Anari and Thuy-Duong Vuong are supported by NSF CAREER award CCF-2045354, a Sloan Research Fellowship, and a Google Faculty Research Award.

References

Vedat Levi Alev and Lap Chi Lau. Improved analysis of higher order random walks and applications. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1198–1211, 2020.

Yeganeh Alimohammadi, Nima Anari, Kirankumar Shiragur, and Thuy-Duong Vuong. Fractionally log-concave and sector-stable polynomials: Counting planar matchings and more. *arXiv preprint arXiv:2102.02708*, 2021.

Nima Anari and Thuy-Duong Vuong. An extension of plücker relations with applications to subdeterminant maximization. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

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.

Nima Anari, Shayan Oveis Gharan, and Cynthia Vinzant. Log-concave polynomials, entropy, and a deterministic approximation algorithm for counting bases of matroids. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 35–46. IEEE, 2018.

Nima Anari, Kuikui Liu, and Shayan Oveis Gharan. Spectral independence in high-dimensional expanders and applications to the hardcore model. In *Proceedings of the 61st IEEE Annual Symposium on Foundations of Computer Science*. IEEE Computer Society, 2020.

Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic independence in high-dimensional expanders: Modified log-sobolev inequalities for fractionally log-concave polynomials and the ising model. *CoRR*, abs/2106.04105, 2021a. URL https://arxiv.org/abs/2106.04105.

Nima Anari, Kuikui Liu, Shayan Oveis Gharan, Cynthia Vinzant, and Thuy-Duong Vuong. Log-concave polynomials iv: approximate exchange, tight mixing times, and near-optimal sampling of forests. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 408–420, 2021b.

Dimitris Bertsimas and Santosh Vempala. Solving convex programs by random walks. *Journal of the ACM (JACM)*, 51(4):540–556, 2004.

Julius Borcea, Petter Brändén, and Thomas Liggett. Negative dependence and the geometry of polynomials. *Journal of the American Mathematical Society*, 22(2):521–567, 2009.

Petter Brändén. Polynomials with the half-plane property and matroid theory. *Advances in Mathematics*, 216(1):302–320, 2007.

Victor-Emmanuel Brunel. Learning signed determinantal point processes through the principal minor assignment problem. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 7365–7374. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/e1228be46de6a0234ac22ded31417bc7-Paper.pdf.

Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689. PMLR, 2017.

Irit Dinur and Tali Kaufman. High dimensional expanders imply agreement expanders. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 974–985. IEEE, 2017.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940. PMLR, 2016.

Mohamed Elfeki, Camille Couprie, Morgane Riviere, and Mohamed Elhoseiny. Gdpp: Learning diverse generations using determinantal point processes. In *International Conference on Machine Learning*, pages 1774–1783. PMLR, 2019.

Li Fang. On the spectra of p- and p0-matrices. *Linear Algebra and its Applications*, 119:1 – 25, 1989. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(89)90065-7. URL http://www.sciencedirect.com/science/article/pii/0024379589900657.

Andreas Galanis, Daniel Stefankovic, Eric Vigoda, and Linji Yang. Ferromagnetic potts model: Refined# bis-hardness and related results. *SIAM Journal on Computing*, 45(6):2004–2065, 2016.

Mike Gartrell, Ulrich Paquet, and Noam Koenigstein. Bayesian low-rank determinantal point processes. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 349–356, New York, NY, USA, 2016. Association for Computing Machinery. ISBN

9781450340359. doi: 10.1145/2959100.2959178. URL https://doi.org/10.1145/2959100.2959178.

Mike Gartrell, Victor-Emmanuel Brunel, Elvis Dohmatob, and Syrine Krichene. Learning nonsymmetric determinantal point processes. *ArXiv*, abs/1905.12962, 2019.

Mike Gartrell, Insu Han, Elvis Dohmatob, Jennifer Gillenwater, and Victor-Emmanuel Brunel. Scalable learning and map inference for nonsymmetric determinantal point processes, 2020.

J. Gillenwater, A. Kulesza, and B. Taskar. Near-optimal map inference for determinantal point processes. *Advances in Neural Information Processing Systems*, 4:2735–2743, 01 2012.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, USA, 1996. ISBN 0801854148.

Boqing Gong, Wei lun Chao, Kristen Grauman, and Fei Sha. Large-margin determinantal point processes, 2014.

Piotr Indyk, Sepideh Mahabadi, Mohammad Mahdian, and Vahab S Mirrokni. Composable core-sets for diversity and coverage maximization. In *Proceedings of the 33rd ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 100–108, 2014.

Matthew Jenssen, Aditya Potukuchi, and Will Perkins. Approximately counting independent sets in bipartite graphs via graph containers. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 499–516. SIAM, 2022.

Tarun Kathuria and Amit Deshpande. On sampling and greedy map inference of constrained determinantal point processes. *arXiv preprint arXiv:1607.01551*, 2016.

Tali Kaufman and David Mass. High dimensional random walks and colorful expansion. *arXiv preprint arXiv:1604.02947*, 2016.

Tali Kaufman and Izhar Oppenheim. High order random walks: Beyond spectral gap. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2018.

Frederic Koehler, Holden Lee, and Andrej Risteski. Sampling approximately low-rank ising models: Mcmc meets variational methods. *arXiv preprint arXiv:2202.08907*, 2022.

Alex Kulesza and Ben Taskar. k-dpps: Fixed-size determinantal point processes, 2012.

Yin Tat Lee, Aaron Sidford, and Santosh S Vempala. Efficient convex optimization with membership oracles. In *Conference On Learning Theory*, pages 1292–1294. PMLR, 2018.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Fast DPP sampling for nyström with application to kernel methods. *CoRR*, abs/1603.06052, 2016. URL http://arxiv.org/abs/1603.06052.

Hui Lin and Jeff Bilmes. Learning mixtures of submodular shells with application to document summarization. *Uncertainty in Artificial Intelligence - Proceedings of the 28th Conference, UAI 2012*, 10 2012.

László Lovász and Santosh Vempala. Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 57–68. IEEE, 2006.

Eyal Lubetzky, Alex Lubotzky, and Ori Parzanchevski. Random walks on ramanujan complexes and digraphs. *arXiv preprint arXiv:1702.05452*, 2017.

S. Mahabadi, P. Indyk, S. Oveis Gharan, and A. Rezaei. Composable core-sets for determinant maximization: A simple near-optimal algorithm. *ArXiv*, abs/1907.03197, 2019.

Zelda Mariet and Suvrit Sra. Fixed-point algorithms for determinantal point processes. *CoRR, abs/1508.00792*, 2015.

Andrea Montanari. Optimization of the sherrington–kirkpatrick hamiltonian. *SIAM Journal on Computing*, (0):FOCS19–1, 2021.

Kazuo Murota et al. Discrete convex analysis: A tool for economics and game theory. *Journal of Mechanism and Institution Design*, 1(1):151–273, 2016.

Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. Practical diversified recommendations on youtube with determinantal point processes. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 2165–2173, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3272018. URL https://doi.org/10.1145/3269206.3272018.

Ali Çivril and Malik Magdon-Ismail. Exponential inapproximability of selecting a maximum volume sub-matrix. *Algorithmica*, 65, 06 2010. doi: 10.1007/s00453-011-9582-6.

Appendix A. Improved Step Size for Down-Up Random Walks

In this section we prove theorem 5. More generally we prove that for any distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ whose generating polynomial is sector-stable, the step size of the down-up random walks can be improved by a factor of 2 compared to lemma 22 and theorem 23, while preserving an inverse-polynomially large spectral gap.

Theorem 32 *Suppose that $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ has a generating polynomial g_μ that is Γ_α-stable for some $\alpha = \Omega(1)$. Then the $k \leftrightarrow (k - \lceil 1/\alpha \rceil)$ down-up random walk on μ has spectral gap $1/\mathrm{poly}(k)$.*

Directly plugging $\alpha = 1/2$ in the above, combined with lemma 18 yields a proof of theorem 5.

Proof [of theorem 32] Note that w.l.o.g., we can assume that $1/\alpha$ is an integer (by perhaps decreasing α slightly). We use the same strategy as in all prior works on high-dimensional expanders to establish a spectral gap for the down-up random walks, namely we use the local-to-global theorem of Alev and Lau (2020).

Consider a family of distributions μ that are closed under conditioning and removal of elements (a.k.a. links (Alev and Lau, 2020)): That is if we condition μ on sets S that contain some element i, and then remove i from all the sets, the new distribution obtained on $\binom{[n]-\{i\}}{k-1}$ is part of the family. Sector-stable distributions are closed under this operation (Alimohammadi et al., 2021).

For each distribution μ in this family on $\binom{[n]}{k}$ we can consider a local weighted graph on n nodes defined by the edge weights:

$$w_{ij} = \mathbb{P}_{S\sim\mu}[i, j \in S].$$

Suppose that the simple random walk on the graph defined by w has spectral gap δ_k. Then Alev and Lau (2020) showed that the spectral gap of the $k \leftrightarrow \ell$ is at least $\Omega(\frac{1}{k}\prod_{i=k-\ell+1}^{k}\delta_i)$. Alimohammadi et al. (2021) managed to bound all such spectral gaps δ_i by roughly $\max(0, 1 - 2/\alpha \cdot i)$, which is only nonzero for $i > 2/\alpha$. As a result ℓ could be taken to be only as large as $k - 2/\alpha$.

Our strategy is to prove a crude bound for δ_i. Already by the analysis of Alimohammadi et al. (2021), we know that the product of δ_i for all $i > 2\alpha$ is inverse-polynomially large. If we show a $1/\mathrm{poly}(k)$ bound for each δ_i for $i \in \{1/\alpha+1, \ldots, 2/\alpha\}$, we can combine the product of these crude bounds to still get a $1/\mathrm{poly}(k)$ lower bound on the spectral gap of the $k \leftrightarrow (k - 1/\alpha)$ random walk.

Now consider a Γ_α-stable distribution μ on $\binom{[n]}{k}$ with $k > 1/\alpha$, and consider the n-node graph defined by the weights $w_{ij} = \mathbb{P}_{S\sim\mu}[i, j \in S]$. We will show a $1/\mathrm{poly}(k)$ lower bound on the Cheeger constant of this graph. Together with the Cheeger inequality (theorem 11), this implies a $1/\mathrm{poly}(k)$ lower bound on the spectral gap.

Consider a partition of the graph defined by some set $S \subseteq [n]$ and $[n] - S$ which achieves the Cheeger constant. Let z be a variable and define

$$z_i = \begin{cases} z & i \in S, \\ z^{-1} & i \notin S. \end{cases}$$

Define the univariate function $f(z) = g(z_1, \ldots, z_n)$. Note that this function can be written as

$$c_k z^k + c_{k-2} z^{k-2} + \cdots + c_{-k} z^{-k}.$$

We claim that to lower bound the Cheeger constant, it is enough to show that there is at least one coefficient $i \in (-k, k)$ such that $c_i > \min\{c_{-k}, c_k\}/\mathrm{poly}(k)$. This is because for any set $T \in \binom{[n]}{k}$ in the support of μ, if T is neither a subset of S nor a subset of $[n] - S$, then at least some of the edges w_{ij} that T contributes to cross the cut $(S, [n] - S)$. So, up to $\mathrm{poly}(k)$ factors, the weight of such sets T lower bounds $w(S, [n] - S)$. The weight of all such sets T is exactly the sum of c_i for $i \neq k, -k$. On the other hand, sets that are completely inside T or $[n] - T$ contribute to the degrees of the two sides of the cut $(S, [n] - S)$ without contributing anything to the crossing edges. The weight of such sets on the S side is c_k and the weight on the $[n] - S$ side is c_{-k}. So it is enough to lower bound one of the middle c_is by $\min\{c_k, c_{-k}\}$.

Now consider the polynomial $h(z) = c_k z^k + c_{k-2} z^{k-1} + \cdots + c_{-k} z^0$. Since the value of f is never 0 on the sector Γ_α, it is easy to see that the value of h is never zero on $\Gamma_{2\alpha}$. We write h as $h_1 + h_2$ where $h_1(z) = c_k z^k + c_{-k} z^0$, and h_2 is defined as the sum of the remaining terms. Note that h_1 has a complex root ω whose argument is π/k, which means that this root completely lies inside Γ_α. Now by elementary results in complex analysis, we know that if on the boundary of a region U that contains this root ω, we always have $|h_2| < |h_1|$, then h_1 and $h = h_1 + h_2$ would have the same number of roots inside U. In other words, if we find a region $U \subseteq \Gamma_{2\alpha}$, where this

condition happens, we get a contradiction. We let this region U be the sector whose sides bisect the angle between the roots of h_1 with the smallest arguments. For any z on the rays defining the boundaries of this region we have $|h_1(z)| \simeq c_k |z|^k + c_{-k} \geq \min\{c_k, c_{-k}\} \max\{1, |z|^k\}$. On the other hand by the triangle inequality, we have that $h_2(z) \leq (\sum_{i \neq k,-k} c_i) \max\{1, |z|^k\}$. This means one of the middle c_i has to be large enough and this completes the proof. ■

Appendix B. Implementation of Local Search for Determinant Maximization

Now, we are ready to prove theorem 8.

Proof [of theorem 8] We let $\mu(S) = \det(L_S)$ and run the two stage algorithm in section 3 with $r = 2$. The approximation guarantee is a direct consequence of lemmas 18 and 30 and theorem 32.

Suppose we are given access to the entries of L. Each iteration of algorithm 2 clearly runs in $O(n^2k^5)$ time, since $\mathcal{N}_2(S)$ has at most $O(k^2n^2)$ elements and computing the determinant of $k \times k$ matrices costs $O(k^3)$ time. The cost of LS_2 can be reduced to $O(n^2k^4)$ time using Schur complements to compute all $\det(L_{Y \cup D})$ for each fixed Y and all D of size $\leq r$ in $O(k^3 + n^2k^2)$ time (see eq. (1) or Gartrell et al., 2020, for example). If we are only given B, C, then each of these submatrices and their determinant can be computed in $O(d^2)$ time, so that each iteration takes $O(n^2d^2k^2)$ time.

Now, we bound the runtime of algorithm 3. To implement each iteration of algorithm 3, we need to compute $\mu(Y) = \sum_{S \in \binom{[n]}{k} : S \supseteq Y} \det(L_Y)$, which is the coefficient of λ^{n-k} in $g(\lambda) = \det(L + \lambda \cdot \mathrm{diag}(\mathbb{1}_{\tilde{Y}}))$ where $\tilde{Y} = [n] \setminus Y$.

There are several ways to compute $\mu(Y)$. To compute the coefficients of polynomial $g(\lambda)$ of degree $\leq n$, we can evaluate g at $n+1$ distinct points λ and use polynomial interpolation, i.e., solve a linear system of equations involving the the Vandermonde matrix. A more efficient way, which costs $O(n^3)$ per computation of $\mu(T)$, for a total runtime of $O(n^4k)$, is as follow:

1. Let $D = \mathrm{diag}(\mathbb{1}_{\tilde{Y}})$. We use the QZ decomposition algorithm (Golub and Van Loan, 1996, Section 7.7, p. 313) to compute unitary matrices Q, Z such that

   $$L = Q\tilde{A}Z^*, D = Q\tilde{D}Z^*$$

   where $\tilde{A}, \tilde{D}$ are both upper triangular. Note that $\deg(g) \leq n - |T|$.

   Compute the roots of $g(\lambda) = \det(L + \lambda D)$, which are exactly the generalized eigenvalues $\lambda_1, \ldots, \lambda_{\deg(g)}$ defined by $\lambda_i = \frac{\tilde{A}_{i,i}}{\tilde{B}_{i,i}}$ where we may assume w.l.o.g. that $\tilde{D}_{i,i} \neq 0$ for $i = 1, \ldots, \deg(g)$, and is zero otherwise. Let $c := \prod_{i \in [n] : B_{i,i} \neq 0} \tilde{D}_{i,i} \prod_{i \in [n] : D_{i,i} = 0} \tilde{A}_{i,i}$. Then

   $$g(\lambda) = c \prod_{i \in [\deg(g)]} (\lambda - \lambda_i)$$

2. We then compute the $(k - n + n')^{th}$-symmetric polynomial of $\lambda_1, \ldots, \lambda_{n-|T|}$ where $e_t = \sum_{W \in \binom{[k-|T|]}{t}} \prod_{j \in W} \lambda_j$ using the recursion (Kulesza and Taskar, 2012)

   $$te_t = e_{t-1}p_1 - e_{t-2}p_2 + e_{t-3}p_3 - \cdots \pm p_k$$

   with $p_t = \sum \lambda_j^t$, and output $\mu(Y) = |ce_{k-(n-\deg(g))}|$.

Given the low-rank decomposition $L = BCB^\intercal$, we can further optimize by reducing the cost of step (i) to $O(nd^2)$. Then the total runtime will be $O(n^2kd^2)$.

Let L^Y be the kernel of $\mathbb{P}_L$ conditioned on the inclusion of items in Y. The eigenvalues of L^Y are exactly the roots of $g(\lambda)$. By eq. (1), L^Y can be rewritten as product of two matrices of rank $\leq d$, thus the nonzero eigenvalues of L^Y can be computed in $O(d^3)$ time. Indeed, let $D_Y := B_Y^\intercal(B_YCB_Y^\intercal)^{-1}B_Y$ then $L^Y = B_{\tilde{Y}}(C - CD_YC)B_{\tilde{Y}}^\intercal$ and $\mathrm{rank}(D_Y) \leq k$ and D_Y can be computed in $O(kd^2)$ time (see eq. (1)).

The matrix $F_Y := \left((C - CD_YC)B_{\tilde{Y}}^\intercal\right) B_{\tilde{Y}}$ has the same characteristic polynomial and nonzero eigenvalues as L^Y. Clearly, $\mathrm{rank}(F_Y) \leq \mathrm{rank}(B) \leq d$, so F_Y and its eigenvalues can be computed in $O(nd^2)$ time. ■

Appendix C. Lower Bound on Step Size for NDPPs

In this section we provide an example of an NDPP, which helps explain why greedy and even 1-step local search fail to achieve a meaningful approximation factor. This example also shows that local search greedy (Kathuria and Deshpande, 2016)[5], another candidate MAP inference algorithm with theoretical performance guarantees for symmetric DPPs, also used by Gartrell et al. (2020) as a baseline to compare their greedy method, fails to achieve a meaningful approximation factor.

Example 1 *Consider L composed of 2×2 blocks $D_i = \begin{bmatrix} c_i & x_i \\ -x_i & c_i \end{bmatrix}$ where $c_i > 1$:*

$$L := \begin{bmatrix} D_1 & 0 & \dots & 0 \\ 0 & D_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & D_n \end{bmatrix},$$

We further assume that $c_1 > c_2 > \cdots > c_n$, $x_1 < x_2 < \cdots < x_n$ and $x_i \gg c_j \ \forall i, j$. It is easy to check that algorithm 1 (greedy) on input $k = 2t$ will select $S = \{1, 2, \dots, 2t-1, 2t\}$. Indeed, algorithm 1 first picks item 1 since $\det(L_{\{1\}}) = c_1$ is maximum among all $L_{\{i\}}$, then picks item 2 since $\det(L_{\{1,2\}}) = c_1^2 + x_1^2 \gg \det(L_{\{1,i\}}) = c_1c_i \forall i \neq 1$, and so on. On the other hand, the optimal subset is $\{n - 2t + 1, \dots, n\}$ by our choice of x, and this could be arbitrarily better than algorithm 1's solution. We may think of items $2i - 1$ and $2i$ as complementary items, say, e.g., toothpaste and toothbrush proposed in a recommender system. The conditions on c_i's and x_i's mean that the degree of complementarity between these pairs increases with i. So $2n - 1$ and $2n$ are the most likely pair to appear together, but each one of $2n - 1$ and $2n$ is most unlikely to appear as a singleton, and the opposite holds for item 1 and 2; for example, think of $2n - 1$ and $2n$ as a tea cup and tea cup lid, which are almost always bought together, but 1 and 2 as toothpaste and toothbrush, which are sometimes purchased separately.

Furthermore, switching out any item $2i-1$ or $2i$ in S for an item $2j-1$ or $2j$ outside of S reduces the determinant by $(c_i^2 + x_i^2)/c_ic_j > c_i > 1$, so S is also maximum among its 1-neighborhood. Thus local search greedy, or equivalently, local search initialized at S, will simply output S itself.

5. This algorithm starts with the output S of algorithm 1, then continuously swaps out an element in S with one outside S to increase the DPP score, until either a local maximum is reached or $k^2 \log k$ swaps have been performed.

We remark that it is easy to construct an example where algorithm 1 produces a subset with zero determinant, whereas the optimal subset can have arbitrarily large determinant. E.g., in example 1, we can make all diagonal entries except for $L_{1,1}$ zero; then, algorithm 1 with even k will necessarily produces a zero determinant.

Appendix D. Composable Core-Sets via Local Search

As further application of our methods, we extend prior work of Mahabadi et al. (2019) on the construction of composable core-sets for maximizing symmetric DPPs to the more general class of distributions that satisfy the strongly Rayleigh property (Borcea et al., 2009) or have a log-concave generating polynomial (Anari et al., 2018).

Definition 33 ((Mahabadi et al., 2019, Definition 2.2)) *A function $c(P)$ that maps the input set $P \subseteq \mathbb{R}^d$ to one of its subsets is called an α-composable core-set for a function $f : 2^{\mathbb{R}^d} \to \mathbb{R}$ if, for any collection of sets $P_1, \ldots, P_n \subseteq \mathbb{R}^d$ we have $f(C) \geq f(P)/\alpha$ where $P = \bigcup_{i \leq n} P_i$ and $C = \bigcup_{i \leq n} c(P_i)$.*

Composable core-sets are a tool (Indyk et al., 2014) to handle computational problems involving large amounts of data. Roughly speaking, a core-set is a summary of a dataset that is enough to solve the computational problem at hand; a *composable* core-set has the additional property that the union of summaries for multiple datasets is itself a good summary for union of all datasets. More precisely, in the context of the optimization problem on $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$, a function c that maps any set $P \subseteq [n]$ to one of its subsets is called an α-composable core-set ((Mahabadi et al., 2019)) if it satisfies the following condition: given any integer m and any collection of sets $P_1, \cdots, P_m \subseteq [n]$

$$\alpha \cdot \max \left\{ \mu(S) \;\middle|\; S \subseteq \bigcup_{i=1}^{m} c(P_i) \right\} \geq \max \left\{ \mu(S) \;\middle|\; S \subseteq \bigcup_{i=1}^{m} P_i \right\}.$$

We also say c is a core-set of size t if $|c(P)| \leq t$ for all sets P. Composable core-sets are very versatile; when a composable core-set is designed for a task, they automatically imply efficient streaming and distributed algorithms for the same task.

One strategy for constructing composable core-sets is local search. Mahabadi et al. (2019) showed that for k-DPP parameterized by symmetric PSD matrix L, (1-step)-local search (algorithm 2 with $r = 1$) gives a $k^{O(k)}$-composable core-sets of size k. The approximation factor of $k^{O(k)}$ is nearly optimal.

Recall that k-DPP parameterized by symmetric PSD matrix L belongs to the family of homogeneous *strongly Rayleigh* distributions, i.e., distributions μ whose generating polynomial g_μ is nonvanishing on the upper half plane (Borcea et al., 2009). An even more general family of distributions is the family of log-concave distributions (Anari et al., 2018). We extend (Mahabadi et al., 2019)'s result to any distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ that is strongly Rayleigh or has a log-concave generating polynomial.

Theorem 34 *Given a distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$, let c be a map that takes $P \subseteq [n]$ to some $c(P) \in \binom{P}{k}$ that is an ζ-approximate local maximum in the 1-neighborhood with respect to μ, for some fixed constant $\zeta \in (0, 1)$:*

$$\mu(c(P)) \geq \zeta \cdot \max \{\mu(S) \mid S \in \mathcal{N}_1(c(P))\}.$$

Then c is an α-composable core-set of size k for the MAP-inference problem on μ with $\alpha = k^{O(k)}$ for strongly Rayleigh μ, and $\alpha = 2^{O(k^2)}$ when μ has a log-concave generating polynomial.

Here we prove that (1-step)-local search yields composable core-sets for distributions that satisfy a strong form of exchange.

Definition 35 (β-strong approximate basis exchange) *For $\beta \geq 1$, we say $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ satisfy β-strong approximate basis exchange if, for $S, T \in \binom{[n]}{k}$ and $j \in T \setminus S$, there exists $i \in S \setminus T$ such that:*

$$\mu(S)\mu(T) \leq \beta\mu(S - i + j)\mu(T + i - j) \tag{2}$$

Lemma 36 *Suppose $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ satisfies β-strong approximate basis exchange, then the Local Search algorithm[6] achieves an $O(\beta)^k$-composable core-set of size k for the MAP-inference problem for μ.*

Proof Consider a partition $P_1 \cup \cdots \cup P_r$ of $[n]$, and let $C_i \in \binom{P_i}{k}$ be a $(1, \zeta)$-local optimum in P_i with $\mu(C_i) > 0$. We want to show

$$\left(\frac{\beta}{\zeta}\right)^k \mathrm{OPT}(C) \geq \mathrm{OPT}(\bigcup_{i=1}^{r} P_i)$$

where $C := \bigcup_{i=1}^{r} C_i$.

Let S^* be such that $\mu(S^*) = \mathrm{OPT}(\bigcup_{i=1}^{r} P_i)$. We need the following fact.

Proposition 37 *For any $W \in \binom{[n]}{k}$ with non-empty $(W \setminus C_i)$, there exists $W' \in \binom{[n]}{k}$ s.t. $|W' \setminus C_i| = |W \setminus C_i| - 1$ and $\frac{\beta}{\zeta}\mu(W') \geq \mu(W)$.*

Proof [of proposition 37] Take an arbitrary $j \in (W \cap P_i) \setminus C_i$. There exists $e \in C_i \setminus W$ s.t.

$$\mu(C_i)\mu(W) \leq \beta\mu(C_i - e + j)\mu(W + e - j) \leq \frac{\beta}{\zeta}\mu(C_i)\mu(W + e - j)$$

Setting $W' = W + e - j$ and dividing both sides by $\mu(C_i) > 0$ gives the desired inequality, since $|W' \setminus C_i| = |W \setminus C_i| - 1$. ■

We can iteratively apply proposition 37 for up to k times to obtain the desired inequality. Indeed, let $W_0 := S^*$, and for $j \geq 1$ let $i_j \in [r]$ and $W_j \in \binom{[n]}{k}$ be such that $\mu(W_{j-1}) \leq \frac{\beta}{\zeta} \cdot \mu(W_j)$ and $|W_j \setminus C_{i_j}| = |W_{j-1} \setminus C_{i_j}| - 1$. proposition 37 guarantees the existence of such i_j and W_j, as long as $W_{j-1} \not\subseteq C$. Let s be the minimum index such that $W_s \subseteq C$. Note that $s \leq k$ and $\mu(W_s) \leq \mathrm{OPT}(C)$. We have

$$\begin{aligned}\mathrm{OPT}([n]) = \mu(W_0) \leq \frac{\beta}{\zeta} \cdot \mu(W_1) \leq \left(\frac{\beta}{\zeta}\right)^2 \cdot \mu(W_2) \leq \cdots \\ \leq \left(\frac{\beta}{\zeta}\right)^s \cdot \mu(W_s) \leq \left(\frac{\beta}{\zeta}\right)^k \cdot \mathrm{OPT}(C)\end{aligned}$$

6. algorithm 2 with $r = 1$

■

Theorem 34 is a direct consequence of lemma 36 and the fact that strongly Rayleigh (respectively log concave) distributions satisfy k^2-strong approximate basis exchange ($2^{O(k)}$-strong approximate basis exchange respectively) (Anari et al., 2021b).