Isotropy and Log-Concave Polynomials: Accelerated Sampling and High-Precision Counting of Matroid Bases

Nima Anari
Computer Science Department
Stanford University
Stanford, USA
anari@cs.stanford.edu

Michał Dereziński
Department of Statistics
University of California, Berkeley
Berkeley, USA
mderezin@berkeley.edu

***Abstract*—We define a notion of isotropy for discrete set distributions. If μ is a distribution over subsets S of a ground set $[n]$, we say that μ is in isotropic position if $\mathbb{P}_{S\sim\mu}[e \in S]$ is the same for all $e \in [n]$. We design a new approximate sampling algorithm that leverages isotropy for the class of distributions μ that have a log-concave generating polynomial; this class includes determinantal point processes, strongly Rayleigh distributions, and uniform distributions over matroid bases. We show that when μ is in approximately isotropic position, the running time of our algorithm depends polynomially on the size of the set S, and only logarithmically on n. When n is much larger than the size of S, this is significantly faster than prior algorithms, and can even be sublinear in n. We then show how to transform a non-isotropic μ into an equivalent approximately isotropic form with a polynomial-time preprocessing step, accelerating subsequent sampling times. The main new ingredient enabling our algorithms is a class of negative dependence inequalities that may be of independent interest.**

As an application of our results, we show how to approximately count bases of a matroid of rank k over a ground set of n elements to within a factor of $1+\epsilon$ in time $O((n+1/\epsilon^2)\cdot \mathrm{poly}(k,\log n))$. This is the first algorithm that runs in nearly linear time for fixed rank k, and achieves an inverse polynomially low approximation error.

The full version of this paper is available at: https://arxiv.org/abs/2004.09079

***Keywords*-isotropy; log-concave polynomials; matroids; Markov chain Monte Carlo;**

I. Introduction

Designing efficient algorithms for sampling from continuous distributions of convex type, i.e., those with a log-concave density, has a long history [see 39, 40, 31, for relevant surveys]. A challenging part of many algorithms for this problem has been transforming the distribution via a linear map to an equivalent standard form called the *isotropic position*. To date, the running times for the fastest algorithms for sampling from convex polytopes and log-concave distributions are dominated by finding the correct scaling linear transform [33, 34]. As our main contribution, we introduce an analogous notion of isotropic position for discrete distributions, and design sampling algorithms that can take advantage of the isotropic position for distributions with a log-concave generating polynomial; a class that should in many ways be thought of as the discrete analog of log-concave distributions.

We study the problem of sampling from a distribution on size k subsets of $[n] = \{1,\dots,n\}$ given by a density function $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$.[1] The task of approximate sampling is to query the function μ repeatedly in order to produce a random set $S \in \binom{[n]}{k}$, such that approximately $\mathbb{P}[S] \propto \mu(S)$. While there is no hope for efficient algorithms that run in nontrivial time ($\ll n^k$) for general μ, recent works have identified a tractable class of distributions with many nice structural properties [5, 3, 8, 12, 4], namely, the class of distributions with a log-concave generating polynomial. This class consists of all μ where the following polynomial is log-concave as a function over $\mathbb{R}^n_{\geq 0}$:

$$g_\mu(z_1,\dots,z_n) := \sum_{S\in\binom{[n]}{k}} \mu(S)\prod_{i\in S} z_i.$$

For any fixed n and k, this class has nonempty interior in the space of all distributions on $\binom{[n]}{k}$. But more importantly, it includes well-studied distributions such as the uniform distribution over bases and/or independent sets of a matroid [3, 8], strongly Rayleigh measures and specifically determinantal point processes [7]. For more examples refer to [3, 8]. Approximate sampling and counting in matroids has connections to many natural combinatorial problems, which motivates the need for finding fast algorithms for these problems. For example, estimating network reliability can be cast as counting independent sets in a cographic matroid [10] and estimating reliability of a linear code against erasures can be cast as counting independent sets

[1] We note that the choice of $\binom{[n]}{k}$ as the domain is simply a standard form and trivial transformations can be applied to many high-dimensional discrete distributions to obtain this form [see, e.g., 2].

2575-8454/20/$31.00 ©2020 IEEE
DOI 10.1109/FOCS46700.2020.00126

in a linear matroid [9]. For an exposition on similar reliability quantities of rigidity matroids see [20].

Anari, Liu, Oveis Gharan, and Vinzant [3] used natural random walks studied in the context of high-dimensional expanders [28, 17, 29] to show that distributions μ with a log-concave generating polynomial can be approximately sampled in polynomial time. Specifically, they showed that the following "down-up" random walk can be used to sample from μ:

- Starting from a set $S_0 \in \binom{[n]}{k}$, repeat for $i = 0, \ldots, t-1$:
  - Sample $e \in S_i$ uniformly at random and let $T_i = S_i - \{e\}$.
  - From all $f \in [n] - T_i$ pick one with probability $\propto \mu(T_i \cup \{f\})$ and let $S_{i+1} = T_i \cup \{f\}$.
- Output S_t.

Our understanding of the mixing time of this random walk has gradually improved to a tight bound in a series of works [3, 12, 4]. We now know that after $t = O(k \log(k/\epsilon))$ steps the distribution of S_t becomes ϵ-close to μ in total variation distance [4]. However, each step of the random walk requires $O(n)$ evaluations of the density μ, which brings the total complexity to $O(nk \log(k/\epsilon))$, much worse than nearly-linear in k. This problem is exacerbated when sampling is used to solve counting [27], that is approximating the partition function $\sum_S \mu(S)$. Known reductions from approximate counting to approximate sampling [27] tack on at least an additional multiplicative factor of $1/\epsilon^2$, where ϵ is the desired relative error, yielding a running time that grows at least as quickly as $\frac{n}{\epsilon^2}\,\mathrm{poly}(k, \log n)$. This is a barrier against using these sampling and counting algorithms.

The dependence of the running time on n can be prohibitive in some natural applications where k is of moderate size but n is very large or even potentially ∞ in the case of continuous determinantal point processes [1]. Other variants of this random walk have been studied for subclasses of log-concave polynomials [24], but they too have a similar total running time.

The starting point of this work are the following questions:

> When can the dependence on n be avoided?
> Can we sample from μ in time $\mathrm{poly}(k)$?

A natural barrier to a positive answer is the existence of *important elements* in the ground set $[n]$. Consider a distribution defined by $\mu(S) \propto \prod_{i \in S} \lambda_i$, where $\lambda_1, \ldots, \lambda_n \in \mathbb{R}_{\geq 0}$. If one of the λ_is is much larger than the others, we should return a set that with high probability contains i. However no algorithm can find this *important* i with fewer than n/k queries to μ. Our main result shows that in a sense, identifying the important elements is the only barrier.

To state our main result, it is convenient to assume that we have an oracle $\mathcal{O}$ that can produce i.i.d. samples of a fixed distribution on $[n]$. One should think of the oracle as trying to sample approximately proportional to the marginals of the target distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$. Note that the sum of marginals $\sum_i \mathbb{P}_{S \sim \mu}[i \in S]$ is always k for a distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$.

Theorem 1. *Given oracle access to a density function $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ with a log-concave generating polynomial, and access to a distribution $p : [n] \to \mathbb{R}_{\geq 0}$ with an i.i.d. sampling oracle $\mathcal{O}$ such that for all i,*

$$p(i) \geq \frac{\mathbb{P}_{S \sim \mu}[i \in S]}{k + O(1)},$$

there is an algorithm that generates approximate samples from μ that are ϵ-close in total variation distance, running in time $\mathrm{poly}(k, \log(1/\epsilon))$.

A natural question is, where does the oracle $\mathcal{O}$ come from? When μ is in approximately isotropic position, that is when $\mathbb{P}_{S \sim \mu}[i \in S]$ is nearly the same for all i, then $\mathcal{O}$ can simply return a uniformly random element in $[n]$. See [36] for some examples of isotropic distributions with log-concave generating polynomials (where the isotropy stems from symmetries in the ground set).

Any non-isotropic distribution μ can be put in near-isotropic position by an operation we call subdivision, akin to linear transformation of continuous log-concave densities. In this operation a larger ground set $[m]$ together with a projection map $\pi : [m] \to [n]$ is used to define a new $\mu' : \binom{[m]}{k} \to \mathbb{R}_{\geq 0}$: A sample $S \sim \mu'$ is obtained by first sampling $\{e_1, \ldots, e_k\} \sim \mu$ and then replacing each e_i with a *uniformly random* element of $\pi^{-1}(e_i)$. The marginals of the new elements are of the form $\mathbb{P}_{S \sim \mu}[i \in S]/|\pi^{-1}(i)|$. By choosing $|\pi^{-1}(i)|$ to be approximately proportional to $\mathbb{P}_{S \sim \mu}[i \in S]$, near-isotropic position is achieved.

To keep the exposition clean, instead of "transforming" distributions μ into near-isotropic position, we instead change our algorithms to be aware of the marginals through the distribution p and the oracle $\mathcal{O}$. This is similar in the continuous sampling literature where, instead of changing the distribution, we change, say the ball-walk algorithm, to an "ellipsoid"-walk. Our algorithms perform exactly the same way as if we had performed a subdivision with infinitely large m, and we invite the reader to verify the equivalent subdivided forms of the algorithms.

For any given probability distribution $p : [n] \to \mathbb{R}_{\geq 0}$, an oracle $\mathcal{O}$ can be constructed with preprocessing time $O(n)$ that generates i.i.d. samples in time $O(\log n)$. So to leverage Theorem 1 it is enough to "approximate" the marginals of μ, namely $\mathbb{P}_{S \sim \mu}[i \in S]$ sufficiently well.

Approximating the marginals can be done by approximate counting which reduces back to the approximate sampling task [27]. Naively one could use the "down-up" random walk to once-and-for-all approximate the marginals of μ. Subsequently there is no need to recompute the marginals, and each subsequent sample can be generated in $\mathrm{poly}(k, \log n, \log(1/\epsilon))$ time. This approach is not satisfactory as the preprocessing step requires at least $\Omega(n)$ samples from μ to even cover all elements; with each sample taking time linear in n, the resulting running time will be quadratic in n. Instead we show how to use a careful cooling schedule, combined with Theorem 1, to both approximate the marginals well enough, and to approximate the partition function.

Theorem 2. *Given oracle access to a distribution μ with a log-concave generating polynomial, there is a randomized algorithm that constructs a sampling oracle $\mathcal{O}$ satisfying the assumptions of Theorem 1 in time $n\ \mathrm{poly}(k, \log n, \log(1/\delta))$ with probability $1 - \delta$.*

We show additionally that the partition function $\sum_S \mu(S)$ can be approximated using the same cooling schedule in time that improves significantly over $(n/\epsilon^2)\,\mathrm{poly}(k, \log n)$.

Theorem 3. *Given oracle access to a density μ with a log-concave generating polynomial, there is a randomized algorithm that computes an ϵ-relative error approximation of $\sum_S \mu(S)$ which runs in time $(n + 1/\epsilon^2)\,\mathrm{poly}(k, \log n, \log(1/\delta))$ and succeeds with probability $1 - \delta$.*

As a corollary we obtain algorithms that can count bases of small-rank matroids with high precision in nearly linear time.

Corollary 4. *Given an oracle that answers independence queries for a matroid of rank $O(\mathrm{poly}\log n)$ over a ground set of n elements, there is an algorithm that approximately counts bases within a multiplicative factor of $1 + 1/\sqrt{n}$ with high probability in nearly linear time $\tilde{O}(n)$.*

By trivial reductions, this result also automatically allows us to estimate the number of small independent sets in nearly linear time, regardless of the rank of the matroid. For example, we can use this result to count forests of size $k = O(\mathrm{poly}\log n)$ in graphs in nearly linear time; for some motivating applications of counting forests, see [19].

A. Related Work

Our results are related to recent work on accelerated sampling for determinantal point processes (DPP), which form a subset of distributions with log-concave polynomials and have a variety of applications in machine learning [30], statistics [6], and graph theory [21] (e.g., uniform sampling of spanning trees). DPPs enjoy a number of properties not satisfied by the general class of distributions with log-concave polynomials, such as closed form expressions for the partition function and the marginals, which make accelerated sampling easier. In particular, recent results by [16, 15, 13, 14] take advantage of these properties to obtain $\mathrm{poly}(k)$ time sampling algorithms based on importance sampling proportional to the marginals. Crucially, these algorithms take advantage of the special structure of DPPs (such as the closed form of the partition function), which is why they cannot be directly extended to general distributions with log-concave polynomials. Even slight variants of DPPs, such as exponentiated DPPs [35] do not enjoy the closed form expressions of DPPs and need our new framework. The differences in our approach, which make accelerated sampling possible for this broader class, include a hierarchical Markov chain procedure to avoid computing an exact partition function, and a cooling schedule for efficiently approximating the marginals.

Negative dependence inequalities related to those we prove in our analysis have been studied for the class of strongly Rayleigh (SR) distributions [7]. This class includes all DPPs, but not all distributions with log-concave polynomials. For example, a uniform distribution over the bases of a matroid is not SR if the matroid is not balanced. Surprisingly, despite extensive literature on the negative correlation properties of SR distributions [37], the negative dependence inequality we prove for all distributions with log-concave polynomials appears to be new even for SR distributions. Some special forms of approximate negative correlation have been obtained for matroids and log-concave polynomials [25]; as an additional corollary of our proof techniques we rederive these correlation bounds and significantly generalize them.

The question of approximately sampling or counting bases of a matroid has been studied for a long time [see, e.g., 18, 26, 22], but most of the attention has been focused on proving just polynomial time efficiency, with some exceptions. Sampling random spanning trees can now be done in nearly linear time [38, 4], but for the seemingly related problem of sampling random forests [19], the jury is still out. When it comes to counting algorithms, the situation is much worse since most results are only based on powerful but generic counting to sampling reductions [27]; these reductions often blow up the running time and make the algorithm impractical. This is despite the fact that the class of reliability problems for graphs, code, truss systems, etc. [10, 9], are all natural counting questions.

B. Techniques

Unlike most prior work on sampling from matroids, our algorithm does not just perform a walk on the basis exchange graph. Rather we combine walks on the basis exchange graph with macro steps that choose a small important subset of the ground set, and only permit walks on that small part. This significantly speeds up the sampling algorithm, by not allowing the basis exchange walks to focus on unimportant elements.

One of the key ingredients in our proof is a set of negative dependence inequalities, which yield interesting facts about matroids. If S is a random basis of a matroid, think of a random forest of size k in a graph, and T is a fixed set, our inequalities upper bound probabilities of the following types of events: $S = T$, $S \supseteq T$, $S \subseteq T$, based on marginals of S. We apply recently derived Modified Log-Sobolev Ineqaulities for matroids and log-concave polynomials [12] in novel ways to derive these inequalities.

II. Preliminaries

We use $[n]$ to denote the set $\{1, \ldots, n\}$, and $\binom{[n]}{k}$ to denote the family of size k subsets of $[n]$. All logs are taken in base e.

We denote a sequence of length t by $\sigma = \langle \sigma_1, \ldots, \sigma_t \rangle$ and for an index set $I \subseteq [t]$ we let

$$\sigma_I := \{\sigma_i \mid i \in I\}.$$

We use one of the common forms of the Chernoff bound.

Lemma 5 (Chernoff Bound)**.** *Let $X_1, ..., X_m$ be independent Bernoulli variables and let $S = \sum_i X_i$. Then, for any $\epsilon \in (0,1)$, we have:*

$$\mathbb{P}[|S - \mathbb{E}[S]| \geq \epsilon \cdot \mathbb{E}[S]] \leq 2e^{-\epsilon^2 \mathbb{E}[S]/3}.$$

We denote the directional derivative operator in direction $v \in \mathbb{R}^n$ with ∂_v:

$$\partial_v = v_1 \partial_1 + \cdots + v_n \partial_n.$$

We use $\mathbb{R}[z_1, \ldots, z_n]$ to denote the set of polynomials with real coefficients in variables $z_1, \ldots, z_n$. We call a polynomial homogeneous (of degree k) if all of its terms have the same degree (equal to k). We call a polynomial multiaffine if no variable in it appears with degree more than 1.

For a distribution or density function $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$, we define the generating polynomial g_μ to be

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

Note that by definition g_μ is both multiaffine and homogeneous.

We use $\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu)$ to denote the Kullback-Leibler divergence between distributions ν and μ defined as follows:

$$\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu) := \mathbb{E}_{S \sim \mu}\left[\frac{\nu(S)}{\mu(S)} \log \frac{\nu(S)}{\mu(S)}\right] = \mathbb{E}_{S \sim \nu}\left[\log \frac{\nu(S)}{\mu(S)}\right].$$

We use $\|\nu - \mu\|_{\mathrm{TV}}$ to denote the total variation distance between distributions ν and μ:

$$\|\nu - \mu\|_{\mathrm{TV}} := \frac{1}{2} \sum_S |\nu(S) - \mu(S)|.$$

A. Markov Chains

For a Markov chain P with stationary distribution μ, we define the mixing time $t_{\mathrm{mix}}(P, \epsilon)$ to be the minimum time t such that for all starting states $S \in \mathrm{supp}(\mu)$

$$\|P^t(S, \cdot) - \mu\|_{\mathrm{TV}} \leq \epsilon.$$

Theorem 6 ([see, e.g., 32])**.** *If an irreducible aperiodic Markov chain with stationary distribution μ and transition matrix P satisfies $\|P^t(S, \cdot) - \mu\|_{\mathrm{TV}} \leq 1/4$ for all $S \in \mathrm{supp}(\mu)$ and some $t \geq 1$, then for any $\epsilon \in (0, 1/4]$*

$$t_{\mathrm{mix}}(P, \epsilon) \leq t \log(1/\epsilon).$$

B. Log-Concave Polynomials

We call a polynomial $g(z_1, \ldots, z_n)$ with nonnegative coefficients log-concave if it is log-concave as a function over $\mathbb{R}^n_{\geq 0}$, i.e., for any $x, y \in \mathbb{R}^n_{\geq 0}$ and $\lambda \in (0, 1)$ we have

$$g(\lambda x + (1 - \lambda) y) \geq g(x)^\lambda g(y)^{1-\lambda}.$$

One of the key operations preserving log-concavity of a polynomial is composition with a linear operator $T : \mathbb{R}^m \to \mathbb{R}^n$ for which $T(\mathbb{R}^m_{\geq 0}) \subseteq \mathbb{R}^n_{\geq 0}$ [5].

Log-concavity of a polynomial is in general not preserved under differentiation. Prior work has considered two classes of polynomials, called strongly log-concave [23], and completely log-concave or Lorentzian [5, 8], to deal with this issue. However, not being closed under derivatives is an artifact of high or low powers of variables dominating others and masking non-log-concavity at middle scales. In particular, for homogeneous multiaffine polynomials (which includes all of the polynomials considered here), this masking does not happen and all three notions coincide.

Lemma 7 ([see, e.g., 4])**.** *Let $g \in \mathbb{R}[z_1, \ldots, z_n]$ be a multiaffine homogeneous polynomial with nonnegative coefficients. If g is log-concave, then it is completely log-concave as well, which means that for any $k \in \mathbb{Z}_{\geq 0}$ and directions $v_1, \ldots, v_k \in \mathbb{R}^n_{\geq 0}$, the following polynomial is log-concave:*

$$\partial_{v_1} \cdots \partial_{v_k} g.$$

C. Down and Up Operators

We define two operators borrowed from the literature on high-dimensional expanders [28, 17, 29]. The down operator acts on the distribution of a random set S of size k and extracts a uniformly random subset of size l.

Definition 8. For nonnegative integers $l \le k \le n$, let $D_{k\to l} \in \mathbb{R}^{\binom{[n]}{k}\times\binom{[n]}{l}}$ be the k to l down operator defined as

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

Note that if $\mu : \binom{[n]}{k} \to \mathbb{R}_{\ge 0}$ is a distribution on sets of size k, then $\mu D_{k\to l}$ is a distribution on sets of size l. Further note that the down operators compose in the expected way:

$$D_{k\to l} D_{l\to m} = D_{k\to m}.$$

When there is a background distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\ge 0}$ we can define an up operator, which is the time-reversal of the down operator for stationary distribution μ. We do not use time-reversal and time-reversibility of Markov chains here, but for definitions of these terms, see [32].

Definition 9. For nonnegative integers $l \le k \le n$ and background distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\ge 0}$, let $U_{l\to k} \in \mathbb{R}^{\binom{[n]}{l}\times\binom{[n]}{k}}$ be the up operator defined as

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

It is easy to see that $D_{k\to l}U_{l\to k}$ defines a time-reversible Markov chain on the state space $\binom{[n]}{k}$ with stationary distribution μ. This can be seen by considering a bipartite graph between $\binom{[n]}{k}$ and $\binom{[n]}{l}$ with an edge of weight $\mu(S)$ between S and T whenever $T \subseteq S$. Then $D_{k\to l}$ is the operation of randomly walking from the top side to the bottom side, and $U_{l\to k}$ is randomly walking from the bottom to the top. Further the weighted degree of every node on the top is $\binom{k}{l}\mu(S) \propto \mu(S)$, which makes $\mu(S)$ the stationary distribution of $D_{k\to l}U_{l\to k}$.

These Markov chains are efficiently implementable, with oracle access to μ, when $k - l = O(1)$. The difficult part is implementing $U_{l\to k}$ which naively takes time n^{k-l}. A sequence of works [3, 12, 4] have identified tight mixing times for distributions μ with a log-concave generating polynomial, when $l = k - 1$:

Theorem 10 ([4]). *Let $\mu : \binom{[n]}{k} \to \mathbb{R}_{\ge 0}$ be a distribution with a log-concave generating polynomial. Then $P := D_{k\to k-1}U_{k-1\to k}$ is a Markov chain with stationary distribution μ and*

$$t_{\mathrm{mix}}(P,\epsilon) \le O(k\log(k/\epsilon)).$$

One of the important steps yielding the above result is a discrete-time variant of the Modified Log-Sobolev Inequality (MLSI) proved by Cryan, Guo, and Mousa [12]. We will use the main ingredient from the proof of the MLSI inequality, that the down operator shrinks KL-divergence:

Theorem 11 ([12]). *If $\mu : \binom{[n]}{k} \to \mathbb{R}_{\ge 0}$ is a distribution with a log-concave generating polynomial, and $\nu : \binom{[n]}{k} \to \mathbb{R}_{\ge 0}$ is an arbitrary distribution, then*

$$\mathcal{D}_{\mathrm{KL}}(\nu D_{k\to k-1} \,\|\, \mu D_{k\to k-1}) \le \frac{k-1}{k}\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu).$$

III. Instant Mixing via Isotropy

Let $\mu : \binom{[n]}{k} \to \mathbb{R}_{\ge 0}$ be a density function on sets of size k. Here we describe a general random-walk-based algorithm for approximately sampling a set S with $\mathbb{P}[S] \propto \mu(S)$, and show that isotropy of μ guarantees almost-instantaneous mixing. However, implementing each step of this algorithm is nontrivial, and later in Section V we show how to implement the steps using an *inner random walk*, proving Theorem 1.

Without loss of generality, assume that every element of $[n]$ is drawn by μ with positive probability, i.e., that $\mathbb{P}_{S\sim\mu}[i \in S] > 0$ for all $i \in [n]$. Also, by a slight abuse of notation, let $\mu(S)$ be equal to zero for all sets S of size other than k. Moreover, let $p : [n] \to \mathbb{R}_{\ge 0}$ be a probability distribution over $[n]$ such that $p(i) > 0$ for all i. We assume there is an oracle $\mathcal{O}$ that can produce i.i.d. samples from p. Consider the following Markov chain $\mathcal{M}^t_{\mu,p}$ defined for any positive integer t, with the state space $\mathrm{supp}(\mu)$. Starting from $S_0 \in \mathrm{supp}(\mu)$, one step of the chain is given by:

1) Draw a sequence $\rho = \langle \rho_1, ..., \rho_t \rangle$ of i.i.d. samples from p by calling $\mathcal{O}$.
2) Arrange S_0 together with ρ, permuting the elements uniformly at random to obtain a sequence $\sigma = \langle \sigma_1, ..., \sigma_{t+k} \rangle$.
3) Return $S_1 = \sigma_S$, where S is drawn from a distribution $\mu_{\sigma,p} : \binom{[t+k]}{k} \to \mathbb{R}_{\ge 0}$ defined by:

$$\mu_{\sigma,p}(S) = \frac{\mu(\sigma_S)\prod_{i\in S} p(\sigma_i)^{-1}}{Z_{\mu,p}(\sigma)} \text{ where } \sigma_S = \{\sigma_i \mid i \in S\}.$$

Here, $Z_{\mu,p}(\sigma)$ denotes the normalization constant of $\mu_{\sigma,p}$. We next establish some basic properties of $\mathcal{M}^t_{\mu,p}$. In particular, we show that, regardless of the choice of the sampling distribution p, the stationary distribution of $\mathcal{M}^t_{\mu,p}$ is μ.

Lemma 12. *If μ has a log-concave generating polynomial then for any $t \geq 1$ the chain $\mathcal{M}^t_{\mu,p}$ is irreducible and aperiodic with stationary distribution μ.*

Proof: Let P denote the transition probability matrix of $\mathcal{M}^t_{\mu,p}$. The aperiodicity follows because for any $S \in \text{supp}(\mu)$, we have $P(S,S) > 0$. To establish irreducibility, note that since $t \geq 1$, for any $S, S' \in \text{supp}(\mu)$ that differ only by swapping a pair of elements (i.e., $|S \cap S'| = k-1$), we have $P(S,S') > 0$. Since μ has a log-concave generating polynomial, $\text{supp}(\mu)$ is a matroid [8, 3] so there is a sequence of such swaps within $\text{supp}(\mu)$ that converts any $S \in \text{supp}(\mu)$ into any $S' \in \text{supp}(\mu)$.

It remains to find the stationary distribution. To that end, suppose that we perform one step of the chain starting from $S_0 \sim \mu$. We first derive the distribution of the intermediate sequence σ, using $r = t+k$ and $p(\tau) = \prod_{i=1}^r p(\tau_i)$ for $\tau \in [n]^r$ as shorthands:

$$\begin{aligned}\mathbb{P}[\sigma = \tau] &= \frac{1}{k!\binom{r}{k}} \sum_{S \in \binom{[r]}{k}} \mu(\tau_S) \prod_{i \in [r] \setminus S} p(\tau_i) \\ &= \frac{p(\tau)}{k!\binom{r}{k}} \sum_{S \in \binom{[r]}{k}} \mu(\tau_S) \prod_{i \in S} p(\tau_i)^{-1} \\ &= \frac{p(\tau) Z_{\mu,p}(\tau)}{k!\binom{r}{k}}, \end{aligned} \tag{1}$$

where the summation enumerates different placements of S_0 in the sequence σ resulting from uniformly permuting the sequence. Finally, we derive the distribution of S_1 as follows:

$$\begin{aligned}\mathbb{P}[S_1] &= \sum_{\tau \in [n]^r} \mathbb{P}[\sigma_S = S_1 \mid \sigma = \tau] \mathbb{P}[\sigma = \tau] \\ &= \sum_{\tau \in [n]^r} \sum_{S \in \binom{[r]}{k}} \mathbb{1}_{[\tau_S = S_1]} \mu_{\tau,p}(S) \mathbb{P}[\sigma = \tau] \\ &= \sum_{S \in \binom{[r]}{k}} \sum_{\tau \in [n]^r} \mathbb{1}_{[\tau_S = S_1]} \frac{\mu(\tau_S) \prod_{i \in S} p(\sigma_i)^{-1}}{Z_{\mu,p}(\tau)} \frac{Z_{\mu,p}(\tau) p(\tau)}{k!\binom{r}{k}} \\ &= k! \binom{r}{k} \sum_{\tilde{\tau} \in [n]^t} \mu(S_1) \frac{p(\tilde{\tau})}{k!\binom{r}{k}} = \mu(S_1) \sum_{\tilde{\tau} \in [n]^t} p(\tilde{\tau}) = \mu(S_1), \end{aligned}$$

which means that μ is the stationary distribution of $\mathcal{M}^t_{\mu,p}$, concluding the proof. ∎

Since $\mathcal{M}^t_{\mu,p}$ is irreducible and aperiodic, it converges to its stationary distribution from any starting state. We next establish the rate of this convergence, under some additional assumptions.

A. Negative Dependence Property

To further analyze the Markov chain $\mathcal{M}^t_{\mu,p}$, we establish a new property of distributions with a log-concave generating polynomial. This property, which may be of independent interest, allows us to bound the normalization constant $Z_{\mu,p}(\sigma)$, as shown in the following lemma (the proof is deferred to Section IV).

Lemma 13. *Let $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ be a distribution with a log-concave generating polynomial. Then, for any function $p : [n] \to \mathbb{R}_{>0}$, integer $t \geq k$ and sequence $\tau \in [n]^t$, we have*

$$\sum_{S \subseteq \binom{[t]}{k}} \mu(\tau_S) \prod_{i \in S} p(\tau_i)^{-1} \leq \left(\sum_{i=1}^t \frac{\mathbb{P}_{S \sim \mu}[\tau_i \in S]}{k \cdot p(\tau_i)} \right)^k.$$

Note that when p is the sampling distribution from $\mathcal{M}^t_{\mu,p}$, then the left-hand side of the inequality is exactly the normalization constant $Z_{\mu,p}(\tau)$. To give some intuition behind this bound, suppose that $p(i) = 1$ for all $i \in [n]$, and let τ be a sequence of t unique elements from some set T. Then, the left-hand side can be concisely written as $\mathbb{P}_{S \sim \mu}[S \subseteq T]$. Furthermore, let $s_1, ..., s_k$ be a random permutation of the elements of S (i.e., s_i are identically distributed according to the marginal distribution of μ, but they may not be independent). Then, the inequality can be stated as a new negative correlation property that may be of independent interest (for all T of size at least k):

$$\mathbb{P}\left[\bigwedge_{i=1}^k s_i \in T \right] \leq \left(\frac{1}{k} \sum_{i \in T} \mathbb{P}[i \in S] \right)^k = \prod_{i=1}^k \mathbb{P}[s_i \in T]. \tag{2}$$

When the size of T is k, then Eq. (2) is implied by an even stronger property of log-concave distributions:

$$\mathbb{P}[S = T] \overset{(a)}{\leq} \prod_{i \in T} \mathbb{P}[i \in S] \overset{(b)}{\leq} \left(\frac{1}{k} \sum_{i \in T} \mathbb{P}[i \in S] \right)^k,$$

where (a) will be proved in Section IV and (b) is the arithmetic-geometric mean inequality. For $t > k$, Eq. (2) no longer follows from (a). For example, suppose that $\mathbb{P}[i \in S] = \frac{k}{n}$ for all i. Then, for a set T of size $t > k$, inequality (a) implies that $\mathbb{P}[S \subseteq T] \leq \binom{t}{k}\left(\frac{k}{n}\right)^k$, whereas Eq. (2) states that $\mathbb{P}[S \subseteq T] \leq \left(\frac{t}{n}\right)^k$, which is tighter by a factor that can be as large as 2^k.

B. Coupling Argument

To establish the convergence rate of $\mathcal{M}^t_{\mu,p}$, we couple a single step of this Markov chain with a single step of the chain $\mathcal{M}^{t-k}_{\mu,p}$. Crucially, we allow the former to start from a fixed arbitrary state $S_0 \in \text{supp}(\mu)$, but we assume that the latter is already at the stationary distribution, with starting state $S'_0 \sim \mu$. The overall coupling procedure is illustrated in Fig. 1.

We next bound the probability that the output states of the two chains are different, i.e., $\mathbb{P}[S_1 \neq S'_1]$, which implies a bound on the total variation distance of $\mathcal{M}^t_{\mu,p}$ from μ after one step. To achieve this, in addition to μ having a log-concave polynomial, we must also

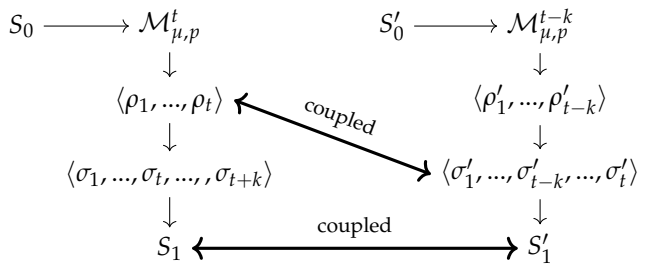

Figure 1. The coupling between $\mathcal{M}^t_{\mu,p}$ started at fixed $S_0 \in \text{supp}(\mu)$ and $\mathcal{M}^{t-k}_{\mu,p}$ started at $S'_0 \sim \mu$.

assume that the importance sampling distribution p is a sufficiently good approximation of the marginal distribution of μ, i.e., that $p(i) \approx \frac{1}{k}\mathbb{P}_{S\sim\mu}[i \in S]$. This ensures we obtain the tightest upperbound from the inequality in Lemma 13. It also makes intuitive sense, since if an element i has high marginal probability, then we need to ensure that the chain is likely to find it in its i.i.d. sampling phase. The precise formulation of the result is given in the following lemma.

Lemma 14. *Let $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ be a distribution with a log-concave generating polynomial. For any $c > 1$, if $t \geq ck^2$ and $k \cdot p(i) \geq (1 + \frac{1}{ck})^{-1}\mathbb{P}_{S\sim\mu}[i \in S]$ for all $i \in [n]$, then the transition matrix P of $\mathcal{M}^t_{\mu,p}$ satisfies for any starting state S_0:*

$$\|P(S_0, \cdot) - \mu\|_{\text{TV}} \leq 5/c.$$

Proof: As discussed above, it suffices to show that the chain $\mathcal{M}^t_{\mu,p}$ started at S_0 can be coupled with chain $\mathcal{M}^{t-k}_{\mu,p}$ started at $S'_0 \sim \mu$, so that $\mathbb{P}[S_1 = S'_1] \geq 1 - \frac{5}{c}$. First, letting $\epsilon = \frac{1}{ck}$, observe that by combining Lemma 13 with the approximation guarantee $k \cdot p(i) \geq (1+\epsilon)^{-1}\mathbb{P}_{S\sim\mu}[i \in S]$, we can upper bound the normalization constant of the distribution $\mu_{\tau,p}$ for any $\tau \in [n]^t$ as follows:

$$Z_{\mu,p}(\tau) \leq \left(\sum_{i=1}^{t} \frac{\mathbb{P}_{S\sim\mu}[\tau_i \in S]}{k \cdot p(\tau_i)}\right)^k \leq (t + \epsilon t)^k.$$

We use this bound first to couple the intermediate sequences generated by the two chains, namely $\rho = \langle\rho_1, ..., \rho_t\rangle$ and $\sigma' = \langle\sigma'_1, ..., \sigma'_t\rangle$ (see Fig. 1). Note that the lengths of the sequences match because the former does not have the starting state S_0 permuted into it. Using the derivation in Eq. (1), we bound the probability distribution of σ' in terms of the distribution of ρ:

$$\begin{aligned}\mathbb{P}[\sigma' = \tau] &= \frac{p(\tau)Z_{\mu,p}(\tau)}{k!\binom{t}{k}} \leq p(\tau)\frac{(t+\epsilon t)^k}{k!\binom{t}{k}} \\ &= \mathbb{P}[\rho = \tau]\prod_{i=0}^{k-1}\frac{t+\epsilon t}{t-i} \leq \mathbb{P}[\rho = \tau]\left(\frac{t+\epsilon t}{t-k}\right)^k,\end{aligned}$$

so using the assumptions that $t \geq ck^2$ and $\epsilon = \frac{1}{ck}$ we conclude that $\mathbb{P}[\rho = \tau] \geq (1 - \frac{2}{c})\mathbb{P}[\sigma' = \tau]$ for all $\tau \in [n]^t$, which means that by coupling we can ensure that $\mathbb{P}[\rho = \sigma'] \geq 1 - \frac{2}{c}$.

We next couple the output sets S_1 and S'_1 by first conditioning on the event that $\rho = \sigma'$, and then observing that distributions $S \sim \mu_{\sigma,p}$ and $S' \sim \mu_{\sigma',p}$ are identical as long as S does not select any indices corresponding to the input set S_0. Without loss of generality, fix the permutation of σ so that $\sigma = \langle\rho_1, ..., \rho_t, s_1, ..., s_k\rangle$, where $S_0 = \{s_1, ..., s_k\}$. Letting $r = t + k$, we now lower bound the probability that S does not select any of the last k indices:

$$\begin{aligned}\mathbb{P}\left[S \subseteq [t]\right] &= \sum_{\tau\in[n]^t}\mathbb{P}[\rho = \tau]\sum_{S\subseteq[t]}\frac{\mu(\tau_S)\prod_{i\in S}p(\tau_i)^{-1}}{Z_{\mu,p}(\langle\tau, s_1, ..., s_k\rangle)} \\ &\geq \sum_{\tau\in[n]^t}\frac{p(\tau)Z_{\mu,p}(\tau)}{(r+\epsilon r)^k} = \frac{k!\binom{t}{k}}{(r+\epsilon r)^k} \geq 1 - \frac{3}{c}.\end{aligned}$$

Observe that, conditioned on the events that $\rho = \sigma'$ and that $S \subseteq [t]$, the distributions of S_1 and S'_1 are identical. A union bound shows that the two events occur together with probability at least $1 - \frac{5}{c}$, completing the proof. ■

Setting $c = 20$ in Lemma 14, we obtain that if $t \geq 20k^2$ and $k \cdot p(i) \geq \left(1 + \frac{1}{20k}\right)^{-1}\mathbb{P}_{S\sim\mu}[i \in S]$, then the transition matrix P of $\mathcal{M}^t_{\mu,p}$ for any state $S_0 \in \text{supp}(\mu)$ satisfies $\|P(S_0, \cdot) - \mu\|_{\text{TV}} \leq 1/4$. This, combined with Lemma 12 and Theorem 6, implies that:

$$\sup_{S_0\in\text{supp}(\mu)}\left\|P^s(S_0, \cdot) - \mu\right\|_{\text{TV}} \leq \epsilon \quad \text{for} \quad s \geq \log 1/\epsilon.$$

IV. Negative Dependence Inequalities

In this section we prove Lemma 13. As a corollary of our techniques, we will additionally prove an approximate negative correlation inequality which generalizes results of Huh, Schröter, and Wang [25] and may be of independent interest. First we strengthen Theorem 11 to show how KL-divergences contract under the $D_{k\to l}$ operator.

Lemma 15. *If $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ is a distribution with a log-concave generating polynomial, and $\nu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ is an arbitrary distribution, then*

$$\mathcal{D}_{\text{KL}}(\nu D_{k\to l} \,\|\, \mu D_{k\to l}) \leq \frac{l}{k}\mathcal{D}_{\text{KL}}(\nu \,\|\, \mu).$$

Proof: We prove this by induction on $k - l$. When $k - l = 1$, this is already the contents of Theorem 11. Note that the distribution $\mu D_{k\to l}$ always has a log-concave generating polynomial. This is because

$$g_{\mu D_{k\to l}} \propto (\partial_1 + \cdots + \partial_n)^{k-l} g_\mu$$

and by Lemma 7, the above polynomial is log-concave. So assuming the statement is correct for k and l, we

can prove it for k and $l-1$ by applying Theorem 11 to the distributions $\nu D_{k\to l}$ and $\mu D_{k\to l}$; we obtain

$$\mathcal{D}_{\mathrm{KL}}(\nu D_{k\to l} D_{l\to l-1} \,\|\, \mu D_{k\to l} D_{l\to l-1}) \le \frac{l-1}{l}\mathcal{D}_{\mathrm{KL}}(\nu D_{k\to l} \,\|\, \mu D_{k\to l}).$$

But note that $D_{k\to l} D_{l\to l-1} = D_{k\to l-1}$. Combining this with the shrinkage of KL-divergence for $D_{k\to l}$ we get

$$\mathcal{D}_{\mathrm{KL}}(\nu D_{k\to l-1} \,\|\, \mu D_{k\to l-1}) \le \frac{l-1}{l}\cdot\frac{l}{k}\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu) = \frac{l-1}{k}\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu),$$

which completes the proof. ∎

In the rest of this section we will apply Lemma 15 for $l=1$. Note that $\mu D_{k\to 1}$ is a distribution on $\binom{[n]}{1} \cong [n]$, which assigns a weight of $\frac{1}{k}\mathbb{P}_{S\sim\mu}[i\in S]$ to every singleton $\{i\}$. As a warmup, we first prove a form of negative correlation for distributions with a log-concave generating polynomial.

Proposition 16. *If $\mu : \binom{[n]}{k} \to \mathbb{R}_{\ge 0}$ is a distribution with a log-concave generating polynomial, then for any set $T \in \binom{[n]}{k}$,*

$$\mu(T) = \mathbb{P}_{S\sim\mu}[S = T] \le \prod_{i\in T}\mathbb{P}_{S\sim\mu}[i\in S].$$

Proof: Let $\nu : \binom{[n]}{k} \to \mathbb{R}_{\ge 0}$ be defined to be 0 everywhere, except at T, where $\nu(T) = 1$. Then

$$\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu) := \mathbb{E}_{S\sim\nu}\left[\log\frac{\nu(S)}{\mu(S)}\right] = \log\frac{1}{\mu(T)}.$$

Note that $\nu D_{k\to 1}$ is the uniform distribution over elements of T. Therefore

$$\mathcal{D}_{\mathrm{KL}}(\nu D_{k\to 1} \,\|\, \mu D_{k\to 1}) = \mathbb{E}_{\{i\}\sim\nu D_{k\to 1}}\left[\log\frac{\nu D_{k\to 1}(\{i\})}{\mu D_{k\to 1}(\{i\})}\right] = \frac{1}{k}\sum_{i\in T}\log\frac{1/k}{\mathbb{P}_{S\sim\mu}[i\in S]/k}.$$

Using Lemma 15 for $l=1$, we obtain that

$$\sum_{i\in T}\log\frac{1}{\mathbb{P}_{S\sim\mu}[i\in S]} \le \log\frac{1}{\mu(T)}.$$

Exponentiating and rearranging the above yields the desired inequality. ∎

Next we prove a negative correlation property that generalizes some results of Huh, Schröter, and Wang [25]. This result can be interpreted as negative correlation up to a factor of 2, and it may be of independent interest.

Lemma 17. *If $\mu : \binom{[n]}{k} \to \mathbb{R}_{\ge 0}$ is a distribution with a log-concave generating polynomial, then for any set T of size $l \le k$, we have*

$$\mathbb{P}_{S\sim\mu}[T\subseteq S] \le \alpha(k,l)\cdot\prod_{i\in T}\mathbb{P}_{S\sim\mu}[i\in S],$$

where $\alpha(k,l) = \binom{k}{l}\left(\frac{l}{k}\right)^l \le l^l/l!$. In particular for $l=2$, $\alpha(k,l)\le 2$, and more generally $\alpha(k,l)\le e^l$.

Proof: We apply Proposition 16 to the distribution $\mu D_{k\to l}$. Note that the marginals of $\mu D_{k\to l}$ are simply l/k times those of μ, because in $D_{k\to l}$ we only keep l out of k elements uniformly at random:

$$\mathbb{P}_{S\sim\mu D_{k\to l}}[i\in S] = \frac{l}{k}\mathbb{P}_{S\sim\mu}[i\in S].$$

Further $\mu D_{k\to l}(T)$ is simply the chance that a sample $S\sim\mu$ contains T times the chance that in dropping from k to l elements, we exactly choose elements of T, which is $1/\binom{k}{l}$:

$$\mu D_{k\to l}(T) = \frac{1}{\binom{k}{l}}\mathbb{P}_{S\sim\mu}[T\subseteq T].$$

Putting these together we get

$$\frac{1}{\binom{k}{l}}\mathbb{P}_{S\sim\mu}[T\subseteq S] \le \prod_{i\in T}\left(\frac{l}{k}\mathbb{P}_{S\sim\mu}[i\in S]\right).$$

∎

So far we have been bounding the probability of $\mathbb{P}_{S\sim\mu}[T\subseteq S]$. Next, we prove a bound on the event that $S\subseteq T$. This is the key ingredient for proving Lemma 13.

Lemma 18. *If $\mu : \binom{[n]}{k} \to \mathbb{R}_{\ge 0}$ is a distribution with a log-concave generating polynomial, then for any set T, we have*

$$\mathbb{P}_{S\sim\mu}[S\subseteq T] \le \left(\frac{\sum_{i\in T}\mathbb{P}_{S\sim\mu}[i\in S]}{k}\right)^k.$$

Proof: Consider the distribution $\nu : \binom{n}{k} \to \mathbb{R}_{\ge 0}$ defined as μ conditioned on the sampled set being $\subseteq T$:

$$\nu(S) \propto \begin{cases}\mu(S) & S\subseteq T,\\ 0 & \text{otherwise.}\end{cases}$$

Notice that the normalizing constant in the above definition is $\mathbb{P}_{S\sim\mu}[S\subseteq T]$. It is easy to see that

$$\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu) = \log\left(\frac{1}{\mathbb{P}_{S\sim\mu}[S\subseteq T]}\right).$$

Let $D_{k\to 1} \in \mathbb{R}^{\binom{n}{k}\times n}$ be the operator that maps a distribution on sets of size k to a distribution on singletons, by first sampling a set from the original distribution, and then sampling a uniformly random element inside of it.

$$D_{k\to 1}(S,i) = \begin{cases}\frac{1}{k} & i\in S,\\ 0 & i\notin S.\end{cases}$$

Applying $D_{k\to 1}$ shrinks the KL-divergence by a factor of k by Lemma 15. In other words

$$\mathcal{D}_{\mathrm{KL}}(\nu D_{k\to 1} \,\|\, \mu D_{k\to 1}) \le \frac{1}{k}\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu).$$

It is easy to see that $\mu D_{k\to 1}(i) = \mathbb{P}_{S\sim\mu}[i \in S]/k$. On the other hand, $\nu D_{k\to 1}$ is not easily expressible in terms of the marginals. But notice that $\nu D_{k\to 1}$ is a distribution on $[n]$ whose support is inside of T. We can ask, among all distributions supported on T, which one minimizes the KL-divergence w.r.t. $\mu D_{k\to 1}$? It is exactly the one obtained from $\mu D_{k\to 1}$ by conditioning on being inside T [see, e.g., 11]. In other words, if we define

$$\omega(i) \propto \begin{cases} \mu D_{k\to 1}(i) & i \in T \\ 0 & i \notin T, \end{cases}$$

then

$$\begin{aligned} \mathcal{D}_{\mathrm{KL}}(\omega \,\|\, \mu D_{k\to 1}) &\le \mathcal{D}_{\mathrm{KL}}(\nu D_{k\to 1} \,\|\, \mu D_{k\to 1}) \\ &\le \frac{1}{k}\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu). \end{aligned}$$

With this definition of ω, it is easy to compute $\mathcal{D}_{\mathrm{KL}}(\omega \,\|\, \mu D_{k\to 1})$:

$$\begin{aligned} \mathcal{D}_{\mathrm{KL}}(\omega \,\|\, \mu D_{k\to 1}) &= \log\left(\frac{1}{\sum_{i\in T} \mu D_{k\to 1}(i)}\right) \\ &= \log\left(\frac{k}{\sum_{i\in T} \mathbb{P}_{S\sim\mu}[i \in S]}\right). \end{aligned}$$

Putting this altogether we get that

$$\log\left(\frac{k}{\sum_{i\in T} \mathbb{P}_{S\sim\mu}[i \in S]}\right) \le \frac{1}{k}\log\left(\frac{1}{\mathbb{P}_{S\sim\mu}[S \subseteq T]}\right).$$

Exponentiating and rearranging gives us the result. ∎

We are now ready to prove Lemma 13.

Proof of Lemma 13: First, note that if there are any repetitions in τ, we can reduce the size of τ by merging all duplicates of an element e together, and replacing $p(e)$ by $p(e)/m$, where m is the number of duplicates. This does not change the r.h.s. of the desired inequality; previously there were m terms of the form $\frac{\mathbb{P}_{S\sim\mu}[e\in S]}{kp(e)}$, and now there is one term equal to $\frac{\mathbb{P}_{S\sim\mu}[e\in S]}{kp(e)/m}$. Similarly, the l.h.s. does not change because every set S containing e previously appeared m times, once for each copy of e, and now it appears once, but with m times the value.

Now that there are no duplicates in τ, we can think of τ as a set T. The inequality we need to prove is

$$\sum_{S\in\binom{T}{k}} \left(\frac{\mu(S)}{\prod_{i\in S} p(i)}\right) \le \left(\sum_{i\in T} \frac{\mathbb{P}_{S\sim\mu}[i \in S]}{kp(i)}\right)^k.$$

Note that if all $p(i)$ were equal to 1, this would be exactly Lemma 18.

Our strategy is to first prove this inequality when p takes integral values, by reducing to $p = 1$, and then generalize to rational p values. Note that both sides of the inequality are k-homogeneous in the p values. So by scaling the inequality for integral p implies the same inequality for rational p. Finally by continuity, we get the inequality for general $p : [n] \to \mathbb{R}_{\ge 0}$. So from now on assume that $p : [n] \to \mathbb{Z}_{\ge 0}$ takes only integer values.

We will apply Lemma 18 to a subdivision of μ. Consider a set of m elements, where $m = p(1) + \cdots + p(n)$. Let $\pi : [m] \to [n]$ be a projection defined by mapping $p(i)$ many distinct elements in $[m]$ to i. Now consider the distribution $\mu' : \binom{[m]}{k} \to \mathbb{R}_{\ge 0}$ defined as follows: First we sample $\{e_1, \ldots, e_k\} \sim \mu$, and then we replace each e_i with a uniformly random element of $\pi^{-1}(e_i)$ to obtain a set $S' \in \binom{[m]}{k}$. This operation preserves log-concavity of the generating polynomial, because $g_{\mu'}$ can be obtained from g_μ by composing with a linear map:

$$g_{\mu'}(z_1, \ldots, z_m) = g_\mu\left(\frac{\sum_{i\in\pi^{-1}(1)} z_i}{p(1)}, \ldots, \frac{\sum_{i\in\pi^{-1}(n)} z_i}{p(n)}\right).$$

Now let T' be a set of the same size as T such that $\pi(T') = T$; that is, for each element $i \in T$ we choose an arbitrary representative of $\pi^{-1}(i)$ and call the collection of these representatives T'. What is the chance that $S' \sim \mu'$ is contained in T'? To produce S', we can first sample $S \sim \mu$. It must be that $S \in \binom{T}{k}$, or else S' will not be contained in T'; but this is not enough. For each such set S, the chance that the random replacements of all elements end up coinciding with our representative choices for T' is exactly $\prod_{i\in S} p(i)^{-1}$. So we get that

$$\mathbb{P}_{S'\sim\mu'}[S' \subseteq T'] = \sum_{S\in\binom{T}{k}} \left(\mu(S)\prod_{i\in S} p(i)^{-1}\right).$$

On the other hand, the marginals of μ' are just $1/p(i)$ fraction of those of μ. That is

$$\mathbb{P}_{S'\sim\mu'}[i \in S'] = \frac{\mathbb{P}_{S\sim\mu}[\pi(i) \in S]}{p(\pi(i))}.$$

So

$$\sum_{i\in T'} \frac{\mathbb{P}_{S'\sim\mu'}[i \in S']}{k} = \sum_{i\in T} \frac{\mathbb{P}_{S\sim\mu}[i \in S]}{kp(i)}.$$

The desired inequality follows from applying Lemma 18 to μ'. ∎

V. Hierarchical Sampling

In this section we complete the proof of Theorem 1 by presenting an efficient implementation of the Markov chain $\mathcal{M}^t_{\mu,p}$ defined in Section III. In the next section we show how to compute the sampling distribution p so that it approximates the marginals of μ sufficiently well, as required by Lemma 14. We measure the complexity by the number of queries of $\mu(S)$, since they represent the dominant cost of all the algorithms. The main tool we use to accomplish both of these tasks is the

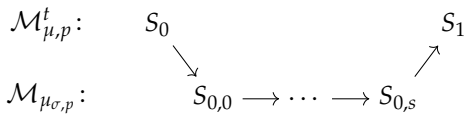

Figure 2. Hierarchical sampling which uses $\mathcal{M}_{\mu_{\sigma,p}}$ to approximate each step of $\mathcal{M}^t_{\mu,p}$.

"down-up" Markov chain, denoted here as $\mathcal{M}_\mu$, which for any μ with a log-concave generating polynomial converges to within ϵ total variation distance of μ after $O(k\log(k/\epsilon))$ steps, and each step requires n queries (Theorem 10).

Our implementation of $\mathcal{M}^t_{\mu,p}$ can be viewed as a hierarchical sampler with two levels: the top level is the chain $\mathcal{M}^t_{\mu,p}$ itself, whereas the bottom level is a separate MCMC sampler which is used to approximate each step of the top level. The expensive part of each step of $\mathcal{M}^t_{\mu,p}$ involves sampling from the distribution $\mu_{\sigma,p}$ which is defined over $\binom{[t+k]}{k}$. Since $\mu_{\sigma,p}$ has a log-concave generating polynomial (see discussion in Section IV), we can use $\mathcal{M}_{\mu_{\sigma,p}}$ to approximately sample from it. The overall implementation of one step of $\mathcal{M}^t_{\mu,p}$ given state S_i, illustrated in Figure 2, proceeds as follows:

1) Construct random sequence σ by i.i.d. sampling $\rho_1, ..., \rho_t$ from p and permuting S_i into it.
2) Return $S_{i+1} = S_{i,s}$ obtained from s steps of $\mathcal{M}_{\mu_{\sigma,p}}$ started at $S_{i,0} = S_i$.

Lemma 19. *Let $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ be a distribution with a log-concave generating polynomial and take any $0 < \epsilon \leq 1/2$. If $s = \lceil 3k\log(k/\epsilon)\rceil$, $t = 20k^2$ and $k \cdot p(i) \geq \left(1 + \frac{1}{20k}\right)^{-1}\mathbb{P}_{S\sim\mu}[i \in S]$ for all $i \in [n]$, then starting from any $S_0 \in \mathrm{supp}(\mu)$, the above procedure returns S_l within ϵ total variation distance of μ for $l = \lceil \log(2/\epsilon)\rceil$ and requires $O(k^3\log^2(k/\epsilon))$ queries.*

Proof: The proof is again based on a coupling argument. Let S'_i denote the sequence of random states produced by the chain $\mathcal{M}^t_{\mu,p}$ started at S_0. From Lemma 14 it follows that S'_l is within $\epsilon/2$ total variation distance of μ. To couple this chain with our sampling procedure note that since $s \geq 3k\log(k/\epsilon) \geq k\log(k/\delta)$ for $\delta = \frac{\epsilon}{2\log(2/\epsilon)}$, log-concavity of the generating polynomial of $\mu_{\sigma,p}$ together with Theorem 10 imply that any given step of our procedure is only δ total variation distance away from an exact step of the chain $\mathcal{M}^t_{\mu,p}$. So, a union bound implies that with probability at least $1 - l\delta \geq 1 - \frac{\epsilon}{2}$ the entire procedure is identical to the procedure that generates S'_l. We conclude that S_l must be within ϵ total variation distance from μ. Since the support of each distribution $\mu_{\sigma,p}$ is contained in $2^{[20k^2+k]}$, the number of queries per step of $\mathcal{M}_{\mu_{\sigma,p}}$ is bounded by $21k^2$. Further, the total number of those steps is $ls \leq 3k\log(k/\epsilon)\log(2/\epsilon)$, so the number of queries is $O(k^3\log^2(k/\epsilon))$. ∎

VI. Marginal Estimation and High-Precision Counting

In this section we prove Theorems 2 and 3. We use the standard trick of introducing a cooling schedule. Our strategy is to introduce a sequence of distributions $\mu_0, \ldots, \mu_t = \mu$ that together define a cooling schedule. All μ_i will have a log-concave generating polynomial. We will guarantee that μ_i and μ_{i+1} do not differ by more than a factor of $1 + O(1/k)$ pointwise. We will make sure that μ_0's marginals are easy to estimate, and $t = \mathrm{poly}(\log n, k)$ is not too large. Then we use this cooling schedule to successively estimate the marginals of each μ_i. This will prove Theorem 2. Afterwards we use standard unbiased estimators for the ratios of partition functions of successive μ_i and combine the estimates to prove Theorem 3.

Our cooling schedule construction will be based on one initial set $U \in \binom{[n]}{k}$. We call U admissible if $\mathbb{P}_\mu[U] \geq \frac{1}{2\binom{n}{k}}$. Note that a random $U \sim \mu$ is guaranteed to be admissible with probability at least $1/2$:

$$\sum_{U \text{ not admissible}} \mathbb{P}_\mu[U] \leq \sum_{U\in\binom{[n]}{k}} \frac{1}{2\binom{n}{k}} \leq \frac{1}{2}.$$

Proposition 20. *There is a randomized algorithm that outputs an admissible set with probability $1 - \delta$ in time $O(nk\log(k)\log(1/\delta))$.*

Proof: We can use the naive down-up random walk of Theorem 10 to approximately sample $\log(1/\delta)$ many points from μ. We then simply return the sample U with the highest $\mu(U)$. Each sample takes time $O(nk\log(k))$ to produce, and the chance that none of them are admissible is $2^{-\log(1/\delta)} \leq \delta$. ∎

Once we have found the admissible U, we let μ_i be defined as follows:

$$\mu_i(S) = \lambda_i^{|S\cap U|}\mu(S),$$

where $\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_t = 1$ define the cooling schedule. Note that each μ_i has a log-concave generating polynomial, because g_{μ_i} can be obtained from g_μ by scaling variables z_e for $e \in U$ by λ_i, a linear transformation.

We will guarantee that $\lambda_{i-1} \leq (1 + O(1/k^2)) \cdot \lambda_i$. This implies that the distributions μ_i and μ_{i-1} do not differ by more than a factor of $1 + O(1/k)$ pointwise, because

$$\frac{\mu_{i-1}(S)}{\mu_i(S)} \leq \left(\frac{\lambda_{i-1}}{\lambda_i}\right)^k = (1 + O(1/k^2))^k = 1 + O(1/k).$$

Our goal is for $\lambda_t = 1$, so that $\mu_t = \mu$. On the other hand, we would like μ_0 to be a distribution whose marginals are easy to estimate, and for this, we would like μ_0 to be very much close to a point mass distribution on the set U. This can be achieved by setting λ_0 to a very large value, but the tradeoff is that we need $t = k^2 O(\log(\lambda_0))$ many steps in our cooling process. A happy middle ground is $\lambda_0 = n^{O(k)}$. Starting with this λ_0, we only need $t = k^2 \log(n^{O(k)}) = k^3 \log n = \mathrm{poly}(k, \log n)$ many cooling steps. This choice also leads to easy bounds on the marginals of μ_0:

Proposition 21. *Assuming that U is admissible, the total variation distance between μ_0 and the point mass distribution $\mathbb{1}_U$ is $\leq 2\binom{n}{k}/\lambda_0$.*

Proof: It is easy to see that

$$\begin{aligned}\|\mathbb{1}_U - \mu_0\|_{\mathrm{TV}} &= 1 - \mathbb{P}_{\mu_0}[U] \\ &\leq \frac{\sum_{S \neq U} \mu(S) \lambda_0^{|S \cap U|}}{\mu(U)\lambda_0^k + \sum_{S \neq U} \mu(S) \lambda_0^{|S \cap U|}} \\ &\leq \frac{\lambda_0^{k-1} \sum_S \mu(S)}{\lambda_0^k \mu(U)} \leq \frac{1}{\lambda_0 \mathbb{P}_\mu[U]}.\end{aligned}$$

and the desired inequality follows from admissibility of U. ∎

So whilst keeping $\lambda_0 = n^{O(k)}$, we can make sure μ_0 is inverse-polynomially close to $\mathbb{1}_U$. This allows us to easily construct a sampling distribution $p : [n] \to \mathbb{R}_{\geq 0}$ that satisfies the assumptions of Theorem 1 for μ_0.

Proposition 22. *If $\|\mathbb{1}_U - \mu_0\|_{\mathrm{TV}} \leq 1/n$, then the following distribution $p : [n] \to \mathbb{R}_{\geq 0}$ satisfies the assumptions of Theorem 1 for μ_0:*

$$p(i) := \begin{cases} \frac{1}{k+1} & i \notin U, \\ \frac{1}{(k+1)(n-k)} & i \in U. \end{cases}$$

Proof: For $i \in U$, we trivially have

$$p(i) = \frac{1}{k+1} = \frac{1}{k + O(1)} \geq \frac{\mathbb{P}_{S \sim \mu_0}[i \in S]}{k + O(1)}.$$

For $i \notin U$, by the bound on total variation distance, we have $\mathbb{P}_{S \sim \mu_0}[i \in S] \leq 1/n$. This implies

$$p(i) = \frac{1}{(k+1)(n-k)} \geq \frac{1}{n(k+1)} \geq \frac{\mathbb{P}_{S \sim \mu_0}[i \in S]}{k + O(1)}.$$

∎

We are now ready to finish the proof of Theorem 2.

Proof of Theorem 2: Having constructed a sampling distribution p for μ_0, we can apply Theorem 1 and use samples to get an even more precise estimate of the marginals. We then use the updated marginal estimates for μ_1. Since μ_0 and μ_1 differ by at most a factor of $1 + O(1/k)$, our marginal estimates are valid for Theorem 1 to be applied to μ_1. However we cannot keep going forward, or else we accumulate error. Instead we use Theorem 1 repeatedly to sample and from these samples extract a fresh good quality estimate of the marginals of μ_1. We then continue the same procedure for μ_2, μ_3, and so on.

All that we need to show is how to leverage samples produced by Theorem 1 to produce a new valid estimate p of marginals whose quality is independent of what is fed to Theorem 1. This is the contents of Lemma 23. Setting $\epsilon = O(1/k)$ in Lemma 23, we use $n\,\mathrm{poly}(k, \log n)$ many samples, each obtained in $\mathrm{poly}(\log n, k)$ time, to produce a fresh distribution p satisfying the assumptions of Theorem 1. ∎

Lemma 23. *Let $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ be a distribution with a log-concave generating polynomial. For any $0 < \epsilon < 1$, it takes $O(nk^{-1}\epsilon^{-3}\log(n/\delta))$ samples from μ to generate a distribution $p : [n] \to \mathbb{R}_{>0}$ such that with probability $1 - \delta$ we have $k \cdot p(i) \geq (1 - \epsilon)\mathbb{P}_{\mathcal{S} \sim \mu}[i \in \mathcal{S}]$ for all $i \in [n]$.*

Proof: Throughout the proof, let q_i denote the marginal $\mathbb{P}_{S \sim \mu}[i \in S]$. Since we only require a one-sided approximation of the marginals, it is acceptable that we significantly over-estimate those that are sufficiently small. To that end, we define $T = \{i \in [n] \mid q_i \geq \frac{\epsilon k}{4n}\}$ as the set of marginals which are "large". Let $S_1, \ldots, S_s$ be sampled i.i.d. from μ with $s = \lceil 3 \cdot 4^3 nk^{-1}\epsilon^{-3}\ln(2n/\delta) \rceil$. Let $\hat{q}_i = \frac{1}{s}\sum_{j=1}^{s} \mathbb{1}_{[i \in S_j]}$ and define $\hat{T} = \{i \in [n] \mid \hat{q}_i \geq \frac{\epsilon k}{3n}\}$. We construct our distribution p as follows:

1) For every $i \in \hat{T}$, we let $p(i) = \frac{1}{k}(1 - \frac{3}{4}\epsilon)\hat{q}_i$.
2) For all remaining i, we let $p(i) = (1 - \sum_{i' \in \hat{T}} p(i'))\frac{1}{|[n] \setminus \hat{T}|}$.

Note that $\sum_i p(i) = 1$. Furthermore, it follows from Lemma 5 that for any $i \in T$ we have:

$$\mathbb{P}\left[|\hat{q}_i - q_i| \geq \frac{\epsilon}{4} \cdot q_i\right] \leq 2e^{-\log(\frac{2n}{\delta})\frac{4n}{\epsilon k} q_i} \leq \frac{\delta}{n}.$$

Thus, a union bound implies that with probability $1 - \delta$ we have $|\hat{q}_i - q_i| \leq \frac{\epsilon}{4} q_i$ for all $i \in T$. From now on, condition on this event. First, it implies that $\hat{T} \subseteq T$, since if $i \in \hat{T}$ then $q_i \geq (1 + \frac{1}{4})^{-1}\hat{q}_i \geq \frac{4}{5} \cdot \frac{\epsilon k}{3n} \geq \frac{\epsilon k}{4n}$. This ensures that for all $i \in \hat{T}$:

$$k \cdot p(i) = (1 - \tfrac{3}{4}\epsilon)\hat{q}_i \geq (1 - \tfrac{3}{4}\epsilon)(1 - \tfrac{1}{4}\epsilon)q_i \geq (1 - \epsilon)q_i.$$

To lower-bound the remaining $p(i)$, we upper bound the total probability mass of p in the set $\hat{T}$:

$$\sum_{i \in \hat{T}} p(i) \leq (1 - \tfrac{3}{4}\epsilon)(1 + \tfrac{1}{4}\epsilon)\frac{1}{k}\sum_{i \in \hat{T}} \mu_i \leq 1 - \frac{\epsilon}{2}.$$

Therefore, if $i \notin \hat{T}$ then $k \cdot p(i) \geq \frac{\epsilon k}{2n} \geq \frac{3}{2}\hat{q}_i$. Furthemore, either $i \in T$, in which case $q_i \leq \frac{4}{3}\hat{q}_i \leq k \cdot p(i)$, or $i \notin T$, and then $q_i \leq \frac{\epsilon k}{4n} \leq k \cdot p(i)$. ∎

Next we prove Theorem 3.

Proof of Theorem 3: By using Theorem 2, we can prepare marginal estimates sufficient for Theorem 1 for each one of the distributions μ_i. The total running time is $n\,\mathrm{poly}(\log n, k, \log(1/\delta))$ for a success probability of $1-\delta/\mathrm{poly}(n)$. So from now on we assume that we can produce a sample from each μ_i in time $\mathrm{poly}(k, \log n)$.

Let $Z_i := \sum_S \mu_i(S)$ be the partition function for μ_i. We use the standard trick of writing a telescoping product

$$Z_t = \frac{Z_t}{Z_{t-1}} \cdot \frac{Z_{t-1}}{Z_{t-2}} \cdots \frac{Z_1}{Z_0} \cdot Z_0,$$

and estimating each fraction individually. In order for the total multiplicative error to be $1+O(\epsilon)$, we need to make sure each factor gets approximated within a factor of $1+O(\epsilon/t)$. We will show that this part takes only $1/\epsilon^2 \cdot \mathrm{poly}(\log n, k, \log(1/\delta))$ time. To approximate Z_{i+1}/Z_i, we can use empirical averages of an unbiased estimator. If $S \sim \mu_i$, then $\mu_{i+1}(S)/\mu_i(S)$ becomes an unbiased estimator for Z_{i+1}/Z_i. Because we made sure μ_i and μ_{i+1} are not too different pointwise, the range of this unbiased estimator is $1 \pm O(1/k)$. So taking an empirical mean of $(t/k\epsilon)^2 \log(1/\delta)$ many such samples yields a $1+O(\epsilon/t)$-approximation of Z_{i+1}/Z_i with probability $1-\delta/\mathrm{poly}(n)$. By Theorem 1, each sample can be produced in time $\mathrm{poly}(k, \log n, \log(1/\delta))$ (where we push the total variation distance into the failure probability δ, to assume our samples yield unbiased estimators).

It remains to estimate Z_0. Note that $\mu_0(U)$ is already a very good estimate of Z_0, because we can make sure λ_0 is set such that $\mathbb{P}_{\mu_0}[U] \geq 1-1/\mathrm{poly}(n)$. However in rare cases where the desired accuracy ϵ is smaller than this $1/\mathrm{poly}(n)$, we can do the following procedure: We continue introducing distributions behind μ_0, namely μ_{-1}, μ_{-2}, and so on, each with a larger and larger λ_i. We only need to go back far enough that $2\binom{[n]}{k}/\lambda_{-i} \ll \epsilon/t$. This happens at $i = \mathrm{poly}(\log n, k)$. We then use $\mu_{-i}(U)$ as our estimate for Z_{-i} and as before estimate the ratios $Z_{-(i-1)}/Z_{-i}$ and so on using empirical means of unbiased estimators. Note that to sample from $\mu_{-1}, \mu_{-2}, \ldots$, we do not need to use new sampling distributions p applicable to Theorem 1. The one we have precomputed for μ_0 works for all of them. ∎

Acknowledgment

The first author thanks Jan Vondrák for stimulating discussions related to negative dependence inequalities. The first author was partially supported by a Google Faculty Research Award. The second author thanks the NSF for funding via the NSF TRIPODS program.

References

[1] Raja Hafiz Affandi, Emily Fox, and Ben Taskar. "Approximate Inference in Continuous Determinantal Processes". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Curran Associates, Inc., 2013, pp. 1430–1438.

[2] Nima Anari, Kuikui Liu, and Shayan Oveis Gharan. "Spectral Independence in High-Dimensional Expanders and Applications to the Hardcore Model". In: *arXiv preprint arXiv:2001.00303* (2020).

[3] Nima Anari, Kuikui Liu, Shayan Oveis Gharan, and Cynthia Vinzant. "Log-Concave Polynomials II: High-Dimensional Walks and an FPRAS for Counting Bases of a Matroid". In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. ACM, June 2019.

[4] Nima Anari, Kuikui Liu, Shayan Oveis Gharan, and Cynthia Vinzant. "Log-Concave Polynomials IV: Approximate Exchange, Tight Mixing Times, and Faster Sampling of Spanning Trees". In: *arXiv preprint arXiv:2004.07220* (2020).

[5] Nima Anari, Shayan Oveis Gharan, and Cynthia Vinzant. "Log-Concave Polynomials I: Entropy and a Deterministic Approximation Algorithm for Counting Bases of Matroids". In: *Proceedings of the 59th IEEE Annual Symposium on Foundations of Computer Science*. IEEE Computer Society, Oct. 2018. DOI: 10.1109/focs.2018.00013.

[6] Bardenet, Rémi, Lavancier, Frédéric, Mary, Xavier, and Vasseur, Aurélien. "On a few statistical applications of determinantal point processes". In: *ESAIM: Procs* 60 (2017), pp. 180–202. DOI: 10.1051/proc/201760180.

[7] Julius Borcea, Petter Brändén, and Thomas Liggett. "Negative dependence and the geometry of polynomials". In: *Journal of the American Mathematical Society* 22.2 (2009), pp. 521–567.

[8] Petter Brändén and June Huh. "Lorentzian polynomials". In: *arXiv preprint arXiv:1902.03719* (2019).

[9] Peter J Cameron. *Notes on matroids and codes*. 1998.

[10] Manoj Chari and Charles J Colbourn. "Reliability polynomials: a survey". In: *J. Combin. Inform. System Sci* 22.3-4 (1997), pp. 177–193.

[11] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[12] Mary Cryan, Heng Guo, and Giorgos Mousa. "Modified log-Sobolev inequalities for strongly log-concave distributions". In: *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2019, pp. 1358–1370.

[13] Michał Dereziński. "Fast determinantal point processes via distortion-free intermediate sampling". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by Alina Beygelzimer and Daniel Hsu. Vol. 99. Proceedings of Machine Learning Research. Phoenix, USA, 25–28 Jun 2019, pp. 1029–1049.

[14] 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*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 11542–11554.

[15] Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. "Correcting the bias in least squares regression with volume-rescaled sampling". In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 16–18 Apr 2019, pp. 944–953.

[16] Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. "Leveraged volume sampling for linear regression". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., 2018, pp. 2510–2519.

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

[18] Tomás Feder and Milena Mihail. "Balanced matroids". In: *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*. 1992, pp. 26–38.

[19] Ashish Goel, Sanjeev Khanna, Sharath Raghvendra, and Hongyang Zhang. "Connectivity in random forests and credit networks". In: *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2014, pp. 2037–2048.

[20] Jack E Graver. "Rigidity matroids". In: *SIAM Journal on Discrete Mathematics* 4.3 (1991), pp. 355–368.

[21] Alain Guenoche. "Random spanning tree". In: *Journal of Algorithms* 4.3 (1983), pp. 214–220.

[22] Heng Guo and Mark Jerrum. "Approximately counting bases of bicircular matroids". In: *arXiv preprint arXiv:1808.09548* (2018).

[23] Leonid Gurvits. "On multivariate Newton-like inequalities". In: *Advances in combinatorial mathematics*. Springer, 2009, pp. 61–78.

[24] Jonathan Hermon and Justin Salez. "Modified log-Sobolev inequalities for strong-Rayleigh measures". In: *arXiv preprint arXiv:1902.02775* (2019).

[25] June Huh, Benjamin Schröter, and Botong Wang. "Correlation bounds for fields and matroids". In: *arXiv preprint arXiv:1806.02675* (2018).

[26] Mark Jerrum and Jung-Bae Son. "Spectral gap and log-Sobolev constant for balanced matroids". In: *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.* IEEE. 2002, pp. 721–729.

[27] Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. "Random generation of combinatorial structures from a uniform distribution". In: *Theoretical Computer Science* 43 (1986), pp. 169–188.

[28] Tali Kaufman and David Mass. "High dimensional random walks and colorful expansion". In: *arXiv preprint arXiv:1604.02947* (2016).

[29] Tali Kaufman and Izhar Oppenheim. "High order random walks: Beyond spectral gap". In: *Combinatorica* (2020), pp. 1–37.

[30] Alex Kulesza and Ben Taskar. *Determinantal Point Processes for Machine Learning*. Hanover, MA, USA: Now Publishers Inc., 2012.

[31] Yin Tat Lee and Santosh S Vempala. "The Kannan-Lovász-Simonovits Conjecture". In: *arXiv preprint arXiv:1807.03465* (2018).

[32] David A Levin and Yuval Peres. *Markov chains and mixing times*. Vol. 107. American Mathematical Soc., 2017.

[33] László Lovász and Santosh Vempala. "Simulated annealing in convex bodies and an O*(n4) volume algorithm". In: *Journal of Computer and System Sciences* 72.2 (2006), pp. 392–417.

[34] Oren Mangoubi and Nisheeth K Vishnoi. "Faster Polytope Rounding, Sampling, and Volume Computation via a Sub-Linear Ball Walk". In: *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2019, pp. 1338–1357.

[35] Zelda E Mariet, Suvrit Sra, and Stefanie Jegelka. "Exponentiated strongly rayleigh distributions". In: *Advances in Neural Information Processing Systems*. 2018, pp. 4459–4469.

[36] Shayan Oveis Gharan and Alireza Rezaei. "A Polynomial Time MCMC Method for Sampling from Continuous DPPs". In: *arXiv preprint arXiv:1810.08867* (2018).

[37] Robin Pemantle and Yuval Peres. "Concentration of Lipschitz functionals of determinantal and other strong Rayleigh measures". In: *Combinatorics, Probability and Computing* 23.1 (2014), pp. 140–160.

[38] Aaron Schild. "An almost-linear time algorithm for uniform random spanning tree generation". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM. 2018, pp. 214–227.

[39] Santosh Vempala. "Geometric random walks: a survey". In: *Combinatorial and computational geometry* 52.573-612 (2005), p. 2.

[40] Santosh S Vempala. "Recent progress and open problems in algorithmic convex geometry". In: *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2010)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2010.