Optimal Sublinear Sampling of Spanning Trees and Determinantal Point Processes via Average-Case Entropic Independence

Nima Anari[1], Yang P. Liu[1], and Thuy-Duong Vuong[1]

[1]Stanford University, {anari,yangpliu,tdvuong}@stanford.edu

Abstract

We design fast algorithms for repeatedly sampling from strongly Rayleigh distributions, which include as special cases random spanning tree distributions and determinantal point processes. For a graph $G = (V, E)$, we show how to approximately sample uniformly random spanning trees from G in $\widetilde{O}(|V|)$[1] time per sample after an initial $\widetilde{O}(|E|)$ time preprocessing. This is the first nearly-linear runtime in the output size, which is clearly optimal. For a determinantal point process on k-sized subsets of a ground set of n elements, defined via an $n \times n$ kernel matrix, we show how to approximately sample in $\widetilde{O}(k^\omega)$ time after an initial $\widetilde{O}(nk^{\omega-1})$ time preprocessing, where $\omega < 2.372864$ is the matrix multiplication exponent. The time to compute just the weight of the output set is simply $\simeq k^\omega$, a natural barrier that suggests our runtime might be optimal for determinantal point processes as well. As a corollary, we even improve the state of the art for obtaining a single sample from a determinantal point process, from the prior runtime of $\widetilde{O}(\min\{nk^2, n^\omega\})$ to $\widetilde{O}(nk^{\omega-1})$.

In our main technical result, we achieve the optimal limit on domain sparsification for strongly Rayleigh distributions. In domain sparsification, sampling from a distribution μ on $\binom{[n]}{k}$ is reduced to sampling from related distributions on $\binom{[t]}{k}$ for $t \ll n$. We show that for strongly Rayleigh distributions, the domain size can be reduced to nearly linear in the output size $t = \widetilde{O}(k)$, improving the state of the art from $t = \widetilde{O}(k^2)$ for general strongly Rayleigh distributions and the more specialized $t = \widetilde{O}(k^{1.5})$ for spanning tree distributions. Our reduction involves sampling from $\widetilde{O}(1)$ domain-sparsified distributions, all of which can be produced efficiently assuming approximate overestimates for marginals of μ are known and stored in a convenient data structure. Having access to marginals is the discrete analog of having access to the mean and covariance of a continuous distribution, or equivalently knowing "isotropy" for the distribution, the key behind optimal samplers in the continuous setting based on the famous Kannan-Lovász-Simonovits (KLS) conjecture. We view our result as analogous in spirit to the KLS conjecture and its consequences for sampling, but rather for discrete strongly Rayleigh measures.

[1]Throughout, $\widetilde{O}(\cdot)$ hides polylogarithmic factors in n.

1 Introduction

Efficiently sampling from probability distributions is a fundamental algorithmic question whose study has been instrumental in revealing connections between many areas of mathematics and computer science. Markov chains are perhaps the single most utilized method in designing sampling algorithms. The study of Markov chains is an active area of research in both high-dimensional continuous settings and combinatorial/discrete settings [see, e.g., Jer98]. Unlike many other computational tasks, sampling is not in general "efficiently verifiable." This motivates a sharp theoretical understanding of the mixing time of Markov chains, because there is no general technique for knowing when to stop running Markov chains in practice without an a priori theoretical bound.

In this work, we study how far we can push the runtime of sampling algorithms for the widely used class of strongly Rayleigh distributions [BBL09], which are distributions supported on size k subsets of a ground set $[n] = \{1, \dots, n\}$, denoted from here on as $\binom{[n]}{k}$, which satisfy strong forms of negative dependence (see Section 2.2 for a formal definition). Examples of strongly Rayleigh distributions include uniformly random spanning trees in a graph (where n is the number of edges and $k+1$ is the number of vertices) and determinantal point processes.

Prior works [Der19; DCV19; Gil+19; AD20; CDV20; Ana+21a] discovered that under certain regularity assumptions on the distribution μ, one can sample from μ in sublinear ($\ll n$) time. Regularity assumptions are needed to prevent a scenario where an element $i \in [n]$ has an extremely high marginal $\mathbb{P}_{S\sim\mu}[i \in S]$; it is impossible to find out which element has this property (and output it as part of the sample) without examining roughly all the n elements. This is quite reminiscent of the problem of sampling from continuous log-concave densities on the Euclidean space, as was noted in prior works [AD20], where important directions in the space might be hard to find. The fastest algorithms for sampling from log-concave densities generally proceed by transforming the distribution into an "isotropic form" (a time-consuming part of the algorithm) which guarantees no particular direction accounts for a significant part of the variance, and proceed to obtain samples from isotropic log-concave densities [LV18; Che21; KL22]. The Kannan-Lovász-Simonovits (KLS) conjecture was formulated to answer how fast one can sample from isotropic log-concave densities [LV18].

Motivated by the analogy with continuous distributions, Anari and Dereziński [AD20] defined a notion of isotropy for discrete distributions μ on $\binom{[n]}{k}$: μ is isotropic when $\mathbb{P}_{S\sim\mu}[i \in S]$ is the same for all $i \in [n]$. A distribution can be *put in approximate isotropic position* via preprocessing (see Section 2.6 for details). The main question then becomes:

> How fast can we sample from (approximately) isotropic distributions μ on $\binom{[n]}{k}$?

Prior works [Der19; DCV19; AD20; CDV20] showed that the answer to this is $\leq \mathrm{poly}(k, \log n)$ for strongly Rayleigh distributions, assuming oracle access to μ. However, the optimal sampling runtime remained open. Our main result in this work shows that the optimal runtime for sampling from isotropic strongly Rayleigh distributions on $\binom{[n]}{k}$ is, roughly speaking, at most the runtime for sampling from related distributions on $\binom{[t]}{k}$ for $t = \widetilde{O}(k)$. In other words, isotropy allows us to pretend that n is only as large as $\widetilde{O}(k)$.

Theorem 1 (Informal, see Theorem 3 for a formal statement)**.** *Suppose that the time to sample from a class of strongly Rayleigh distributions on $\binom{[n]}{k}$ is $\mathcal{T}(n,k)$. Then we can sample from (approximately)*

isotropic distributions in this family in time $\widetilde{O}(1) \cdot \mathcal{T}(\widetilde{O}(k), k)$.

Remark 2. Our progress is analogous to the history of sampling algorithms for continuous distributions and the role of (continuous) isotropy [see, e.g., LV18]. Transforming a convex body or a log-concave density into isotropic position (defined as having covariance matrix $\simeq I$ instead of uniform marginals) is the standard preprocessing step, and the main challenge has been establishing properties of isotropic distributions that would then yield optimal bounds on mixing time of standard off-the-shelf Markov chains. A notable conjectured property of isotropic log-concave distributions is the KLS conjecture, which was recently nearly resolved [Che21; KL22]. We view our result as an analog, at least in spirit, of the KLS conjecture for discrete distributions: we establish optimal mixing time bounds (analogous to consequences of the KLS conjecture) for strongly Rayleigh (analogous to log-concave) discrete distributions in discrete-isotropic position (a natural analog of the continuous isotropic position). Interestingly, our proof technique also shows some resemblance to the common framework used in recent advances on the KLS conjecture [LV18; Che21; KL22]: a key technical result we prove is that isotropy is approximately preserved with high probability under a natural localization process (see Section 4). Chen and Eldan [CE22] observed recently that several localization processes used for continuous and discrete sampling problems can be, at least partially, unified under a single umbrella. We believe our results provide further justification for this unification.

In many applications of sampling, one needs not just one, but rather many independent samples from a distribution. A fundamental observation is that the amortized time of producing many samples can often be much less than the cost of producing one sample. As an example, consider the task of producing samples from a distribution on n points given explicitly by n numbers $p_1, \ldots, p_n \geq 0$ that sum to 1. The time to produce a single sample from this distribution is $\simeq n$, as one needs to look at all p_i. However, after reading through the whole input, it is easy to construct a data structure (such as a simple array of prefix sums) that allows subsequent samples to be obtained in $\widetilde{O}(1)$ time. Obtaining similar economies of scale for distributions supported on exponentially-sized state spaces is not possible with this naïve approach; instead, our results show how to obtain optimal economies of scale by a different method that preprocesses a strongly Rayleigh distribution and puts it into isotropic form.

We remark that in some natural scenarios, a preprocessing step might not be needed at all, and we can enjoy fast runtimes even for the first sample. For example, if the distribution is symmetric w.r.t. the ground set, see, e.g., [OR18] for examples of determinantal point processes on symmetric spaces, the distribution is automatically in isotropic form. Similarly, for random spanning trees in graphs, under mild expansion assumptions (roughly speaking, expanding mildly better than 2-dimensional grids) [Ale+18], no edge will have a large marginal and the distribution is automatically in approximately isotropic form. Below we expand on two classes of distributions that constitute the main applications of our result.

Random spanning trees. Random spanning trees of a graph $G = (V, E)$ have found many applications in theoretical computer science. In approximation algorithms for the Traveling Salesperson Problem (TSP) [GW17] they are a key component of the Best-of-Many Christofides algorithm used in recent TSP improvements [KKO21]. Random spanning trees have found applications in the construction of graph sparsifiers [GRV09; KS18]. As another example application, the recent breakthrough of Karlin, Klein, Gharan, and Zhang [Kar+21] on the k-edge connected multi-subgraph problem uses $\Theta(k)$ independent samples of random spanning trees, which demonstrates how economies of scale for sampling can lead to faster algorithms for some natural problems.

The distribution of random spanning trees is also deeply connected with spectral graph theory and Laplacians of graphs, e.g., through the matrix-tree theorem. This has all motivated a long sequence of works on obtaining fast algorithms for sampling from this ubiquitous distribution [Ald90; Bro89; Wil96; CMN96; KM09; MST14; Dur+17a; Dur+17b; Sch18; Ana+21c]. Many of these works have used random spanning trees as a testing ground for novel algorithm design techniques, in particular techniques originating in the study of Laplacian solvers, and more recently high-dimensional expanders. The latest works on sampling from spanning trees [Sch18; Ana+21c] obtained, using two very different approaches, nearly-linear time sampling algorithms. In this work we show how to push even further and get optimal sublinear sampling algorithms with runtime $\widetilde{O}(|V|)$, after an $\widetilde{O}(|E|)$ preprocessing step.

Determinantal point processes Another important example of strongly Rayleigh distributions is a k-sized determinantal point process, or k-DPP for short. A k-DPP μ is a distribution on $\binom{[n]}{k}$ defined with the help of an $n \times n$ positive semidefinite matrix L, where probabilities are given by $k \times k$ principal minors:

$$\mu(S) \propto \det(L_{S,S}).$$

DPPs have found many applications in machine learning, recommender systems, and randomized linear algebra [see, e.g., DR10; KT12; DM21]. In most applications of k-DPPs, the size of a sample is small compared to the domain $[n]$, i.e., $k \ll n$, and the primary goal of sampling algorithms is to minimize the runtime's dependence on n. A nearly-linear dependence on n can be achieved for example via Markov chains [AOR16; HS19]. Recent works have shown how to go even further, and obtain after a preprocessing step, $\text{poly}(k, \log n)$ sampling times [DWH18; DWH19; DCV19; Gil+19; Der19; AD20; CDV20]; however, the dependence on k remained suboptimal. We push the runtime to what we believe is the natural barrier for this sampling problem, and obtain a sampling algorithm with runtime $\widetilde{O}(k^\omega)$, where ω is the matrix multiplication exponent. Note that k^ω is the time needed to just compute $\mu(S)$ for one set S, which is a natural barrier and suggests our result might be optimal.

We further show that the preprocessing step for DPPs can be implemented in time $\widetilde{O}(nk^{\omega-1})$. This, surprisingly, leads to an improvement for obtaining even a single sample from DPPs. The best prior algorithms were either based on MCMC and had a runtime of $\widetilde{O}(nk^2)$ [AOR16; HS19] or were based on linear algebraic primitives [KT12; Pou20], which implemented with fast matrix multiplication, would take time $\widetilde{O}(n^\omega)$, see Lemma 36. We remark that our improvement from $\widetilde{O}(\min\{nk^2, n^\omega\})$ to $\widetilde{O}(nk^{\omega-1})$ is only made possible by a fast preprocessing step which crucially is implemented by bootstrapping with the primitive of fast sampling from isotropic distributions.

1.1 Sampling algorithm

To obtain our optimal sublinear samplers, we use the framework established in prior works [Dur+17a; DCV19; Der19; AD20; CDV20; Ana+21a] of sparsifying the domain $[n]$ for isotropic distributions, i.e., distributions with roughly balanced marginals [AD20]. The preprocessing step for our algorithm consists only of putting the distribution into (approximately) isotropic position (see Section 2.6) by finding approximate overestimates for the marginals $\mathbb{P}_{S\sim\mu}[i \in S]$ and transforming μ by splitting elements with large marginals. One of our novel contributions is the design of new schemes for bootstrapping very fast (and likely optimal) preprocessing steps.

For our main contribution, we obtain an optimal nearly-linear-in-k domain sparsification for isotropic strongly Rayeligh distributions. In domain sparsification, we reduce the task of sampling

from our distribution on $\binom{[n]}{k}$ to distributions on $\binom{[t]}{k}$; we show this can be done with $t = \widetilde{O}(k)$. Prior works on this problem either required $t \simeq k^2$ [DCV19; Der19; AD20; CDV20; Ana+21a] or for the specific case of spanning tree distributions required $t \simeq k^{1.5}$ [Dur+17a].

More formally, for an approximately isotropic μ, we generate a sample by starting from some set $S_0 \in \binom{[n]}{k}$ and following the random walk defined by Algorithm 1 for $\widetilde{O}(1)$ steps. We output $S_{\widetilde{O}(1)}$ as our approximate sample from μ. Note that this random walk has an easy step (choosing T_i uniformly at random from supersets of S_i) and a challenging step (choosing S_{i+1} from subsets of T_i with law induced by μ). The challenging step is an instance of a similar sampling problem but with a smaller domain size t, so we can use a problem-specific baseline sampling algorithm.

Algorithm 1: Down-up walk on the complement distribution

for $i = 0, 1, 2, \ldots$ **do**
  From all t-sized supersets of S_i, select one uniformly at random and name it T_i.
  Select among k-sized subsets of T_i a random set S_{i+1} with $\mathbb{P}[S_{i+1}] \propto \mu(S_i)$.

We remark that the sparsification algorithm (Algorithm 1) is not new and very similar variants of it have been used by almost all mentioned prior works. However, our analysis of Algorithm 1 is entirely different. A departure from prior methods of analysis is not for convenience, but rather necessary. Domain sparsification looks fundamentally different below $t \simeq k^{1.5}$. All prior works used in some shape or form the fact that the partition function of T_i, i.e. $\sum_{S \subseteq T_i} \mu(S)$ concentrates for a random T_i. Indeed, Durfee, Kyng, Peebles, Rao, and Sachdeva [Dur+17a] used this to design algorithms for not just sampling, but also counting spanning trees. Below the threshold of $t \simeq k^{1.5}$, the partition function no longer concentrates (see Section 1.3). Surprisingly, we still show that while T_i's are not good representatives of the ground set $[n]$ for partition functions or counting purposes, they still are good sparsifiers for sampling.

1.2 Our results

To formally state our main results on sampling from strongly Rayleigh distributions, it is useful to define $\mathcal{T}_\mu(t, k)$ for a distribution $\mu \in \mathbb{R}^{\binom{[n]}{k}}$ as the time it takes to produce a sample from μ conditional on all elements of the sample being a subset of a fixed set T of size $|T| = t$. We use $\widetilde{O}(\cdot)$ to suppress $\operatorname{poly}\log n$ factors. Notice below that the sum of marginals $\sum_i \mathbb{P}_{S \sim \mu}[i \in S]$ is always equal to k for a distribution supported on $\binom{[n]}{k}$.

Theorem 3 (Sampling via marginal overestimates)**.** *Given a strongly Rayleigh distribution $\mu \in \mathbb{R}^{\binom{[n]}{k}}$ and marginal overestimates $q_i \geq \mathbb{P}_{T \sim \mu}[i \in T]$ for $i \in [n]$ which sum to $K := \sum_{i \in [n]} q_i$, there is an algorithm that produces a sample from a distribution with total variation distance $n^{-O(1)}$ from μ in time bounded by $\widetilde{O}(1)$ calls to $\mathcal{T}_\mu(O(K), k)$.*

We prove Theorem 3 using a local-to-global argument, which requires us to also show that random conditionals of μ are isotropic with high probability. This is similar in spirit to the recent analyses of the KLS conjecture using stochastic localization [Che21; KL22] which show that an isotropic continuous distribution remains approximately isotropic over a stochastic evolution.

Theorem 4 (Informal, see Theorem 27 for a formal statement)**.** *Let $\mu \in \mathbb{R}^{\binom{[n]}{k}}$ be an isotropic strongly Rayleigh distribution. For $T \subseteq [n]$ and $S \in \binom{T}{k}$, let $\mu_T(S) := \mu(S) / \sum_{S \in \binom{T}{k}} \mu(S)$. Then with high*

probability over $T \in \binom{[n]}{t}$ *for* $t = \widetilde{O}(k)$, μ_T *is approximately isotropic.*

Our input distributions may not be isotropic, so we also design an efficient preprocessing step to obtain marginal estimates to transform μ into an isotropic distribution.

Theorem 5 (Informal, see Theorem 34 for a formal statement)**.** *Given access to a strongly Rayleigh distribution* $\mu \in \mathbb{R}^{\binom{[n]}{k}}$, *we can obtain overestimates of the marginals* $\mathbb{P}_{T\sim\mu}[i \in T]$ *summing to* $O(k)$ *in time proportional to* $\widetilde{O}(n/k)$ *calls to a sampler for isotropic distributions on sets of size* $\widetilde{O}(k)$.

We remark that the preprocessing time of $\widetilde{O}(|E|)$ for estimating marginals of a random spanning tree can be alternatively achieved by estimating effective resistances of the graph using Laplacian solvers and the Johnson-Lindenstraus lemma [ST04; SS11]. However, we give a self-contained method by bootstrapping the sampling algorithm (Theorem 5) that covers not only random spanning trees, but also k-DPPs.

We can apply these results along with known algorithms that sample a random spanning tree in $\widetilde{O}(|E|)$ time [Ana+21c], or a k-DPP on n elements in $\widetilde{O}(n^\omega)$ time (see Lemma 36) to achieve faster runtimes for sampling from these distributions. We note that our algorithm for sampling a k-DPP is faster than previously known runtimes, even in the case of generating a single sample.

Corollary 6 (Sampling spanning trees)**.** *For a graph* $G = (V, E)$, *possibly weighted with weights* $\lambda \in \mathbb{R}^E_{>0}$, *we can output* s *independent spanning trees with* $n^{-O(1)}$ *total variation distance from the distribution* $\mu(T) \propto \prod_{e\in T} \lambda_e$ *in time* $\widetilde{O}(|E| + s\,|V|)$.

Corollary 7 (Sampling DPPs)**.** *Given an* $n \times n$ *positive semidefinite matrix* L, *there is an algorithm that outputs* s *independent approximate samples from the* k*-DPP defined by* L *in time* $\widetilde{O}(nk^{\omega-1} + sk^\omega)$.

Finally, we remark that our methods also show analogous mixing times of $\widetilde{O}(k)$ steps for the Markov chain that uses small up-down steps, i.e., calls to $\mathcal{T}_\mu(k+1, k)$, when sampling isotropic strongly Rayleigh distributions. Such steps are easy to implement in practice, and were used to efficiently sample from general strongly Rayleigh and logconcave distributions [CGM19; Ana+21c]. We formally state these results in Theorem 35 in Section 6.

1.3 Techniques and comparison to prior work

We depart from previous analyses of Algorithm 1 and take the new approach of viewing the sparsification procedure as a down-up random walk on high-dimensional expanders [see, e.g., KO18]. We establish that isotropy significantly improves the "expansion" of the high-dimensional-expander. We use the notion of expansion called entropic independence [Ana+21b], which is one of the few able to yield modified log-Sobolev inequalities and tight mixing times for down-up walks.

The random walk in Algorithm 1 can be seen as the down-up walk (see Section 2 for definition) on the complement/dual distribution associated with μ; note that each step of this walk involves a sparsified sampling problem, where we only have to sample from a distribution on $\binom{T_i}{k}$. For this we use a baseline sampling algorithm, a Markov chain based on a clever link-cut tree data structure for spanning trees, and a naïve matrix-multiplication-based sampler for DPPs.

Below we describe the main techniques we use.

Boosted entropic independence under isotropy. The main tool we use to bound the mixing time of the random walk is the notion of entropic independence (see Section 2 for definition) [Ana+21b]. While standard results about strongly Rayleigh distributions give an out-of-the-box factor 1 entropic independence for the complement distribution $\bar{\mu}$, this is too weak for our purposes as it only implies a mixing time of $\simeq \widetilde{O}((n-k)/t)$ for Algorithm 1, which has an unacceptable dependence on n. This is not surprising, as these black-box results do not incorporate isotropy of μ. In this work, we show that whenever μ has entropic independence and its marginals are not too large, the complement distribution $\bar{\mu}$ has to have a boosted entropic independence, better by a logarithmic factor over what is naïvely expected (Theorem 24).

Average case local-to-global and concentration of marginals The standard machinery for establishing mixing times using entropic independence (i.e., the so-called local-to-global method [AL20]) needs entropic independence of not just the distribution $\bar{\mu}$, but all of its conditionings as well. Conditioning $\bar{\mu}$ on a set of elements is the same as throwing those elements out of the ground set for μ. Unfortunately, in the worst case, this can significantly imbalance the marginals of μ. As an example, consider the spanning tree distribution on a complete graph, which is by symmetry isotropic. Throwing edges out, we can create any graph as a subgraph of the complete graph; for example, we can throw out all but one edge in a cut to make the marginal of that edge equal to 1. To overcome this obstacle, we show that with high probability, i.e., in an average sense over the choice of elements in the conditioning, the marginals remain balanced (Theorem 27) and combine this with an average local-to-global result adapted from [Ali+21] (Theorem 20) to establish the tight mixing time. As far as we know, this is the first application of an average local-to-global theorem. Our strategy of showing average-case isotropy under conditionings is reminiscent of the strategy employed in works on the KLS conjecture which show approximate isotropy holds under an appropriate localization process [Che21; CE22; KL22].

Improved marginal estimation Our main focus is on the time per sample after preprocessing, but we also obtain fast algorithms that improve the preprocessing runtime compared to prior works. Our improved procedures are able to shave off $\mathrm{poly}(k)$ factors from the runtime of marginal estimation (Theorem 34), and are essential for our faster $\widetilde{O}(nk^{\omega-1})$ time algorithm for sampling from a k-DPP. This is achieved by a recursive procedure that uses marginals of the restriction of μ to roughly half the domain $[n]$ as overestimates for the marginals of μ. In the end, marginal overestimation is roughly reduced to $\simeq \widetilde{O}(n/k)$ subtasks of marginal overestimation for distributions over domains of size $\widetilde{O}(k)$.

Barriers faced by prior approaches In order to derive the tight sparsification of $t = \widetilde{O}(k)$ in Algorithm 1, we had to rethink the entire analysis technique. To emphasize the importance of tight bounds on t, we note that prior results on general strongly Rayleigh measures [Der19; CDV20; DCV19; AD20] had at least a quadratic dependence on the output size k, which made them moot for random spanning trees (where k^2 is always larger than the total number of edges in the graph). The barrier faced by the aforementioned works, and also that of [Dur+17a] is roughly speaking that for the regime $t = \widetilde{O}(k)$, subsets T_i are *not* good sparsifiers for partition functions. To appreciate this better, consider a simple distribution μ on $\binom{[n]}{k}$ defined as follows: first we partition $[n]$ into disjoint sets $U_1, \ldots, U_k$ of size n/k each, and then define our distribution as uniform over sets which pick exactly one element from each U_i. Clearly this distribution is isotropic. Now suppose that we select a uniformly random ck-sized set T from $[n]$. The intersection of T with each U_i has expected size c. For small values of c, the distribution of this intersection size is

well-approximated by a Poisson distribution. The count / partition function of the distribution restricted to T is $\prod_{i=1}^{k} |T \cap U_i|$.

The fluctuations of each $T \cap U_i$ are on the order of $\sqrt{c}$. These fluctuations make the above product typically very far from its mean, unless c is growing at least polynomially with k. A careful analysis (similar to [Dur+17a]) would show that $c \simeq \sqrt{k}$ is the threshold after which the count concentrates around the mean. To overcome this barrier, we do not use counts in our analysis at all. Rather, we show that *marginals* do concentrate all the way down to the threshold $t = \widetilde{O}(k)$, using a martingale argument. We combine this concentration of marginals with the fact that isotropy improves entropic independence to show that isotropic strongly Rayleigh distributions are extremely good high-dimensional expanders in an average sense.

1.4 Organization

In Section 2 we collect preliminary notions relating to distributions, conditionals, and Markov chains. We additionally introduce entropic-independence and local-to-global theorems that we use to analyze the down-up walk that our sampling algorithms are based on. In Section 3 we show our main bound on the entropy contraction of a down step of the complement distribution of a strongly Rayleigh distribution with bounded marginals. In Section 4 we show that random marginals of strongly Rayleigh distributions stay bounded with high probability, which is essential to applying the average-case local-to-global principle. In Section 5 we give a simple and efficient procedure for estimating marginal overestimates based on recursive sampling. In Section 6 we combine the previous sections to prove our main results about sampling spanning trees, DPPs, and strongly Rayleigh distributions in general. Finally, deferred proofs are given in Section 7.

Acknowledgments

We thank Michal Dereziński and Elizabeth Yang for useful discussions.

Nima Anari and Thuy-Duong Vuong are supported by NSF CAREER Award CCF-2045354, a Sloan Research Fellowship, and a Google Faculty Research Award. Yang P. Liu was supported by the Department of Defense (DoD) through the National Defense Science and Engineering Graduate Fellowship, and NSF CAREER Award CCF-1844855 and NSF Grant CCF-1955039.

2 Preliminaries

We use $[n]$ to denote the set $\{1, \ldots, n\}$. We view distributions/measures defined over a finite ground set Ω interchangeably as either (probability mass) functions $\mu : \Omega \to \mathbb{R}_{\geq 0}$ or just row vectors $\mu \in \mathbb{R}^{\Omega}$.

For a distribution $\mu \in \mathbb{R}^{\binom{[n]}{k}}$, let $p(\mu) \in \mathbb{R}^n$ denote the marginals of μ, i.e., $p(\mu)_i := \mathbb{P}_{S \sim \mu}[i \in S]$. Denote $p(\mu)^{\max} := \max\{p(\mu)_i \mid i \in [n]\}$. When μ is clear from context, we write p instead of $p(\mu)$. We define $\bar{\mu} : \binom{[n]}{n-k} \to \mathbb{R}_{\geq 0}$ as the *complement distribution* associated to μ, defined as

$$\bar{\mu}(S) := \mu([n] \backslash S).$$

Our analysis (in particular for applying a local-to-global principle) requires looking at restrictions of μ to specific subset of the ground set $[n]$ of elements. In the complement, this corresponds to conditioning that $\bar{\mu}$ contains certain elements.

Definition 8 (Restricted distribution)**.** For a distribution μ defined over subsets of a ground set $[n]$ and $S \subseteq [n]$, define μ_S to be the distribution of $F \sim \mu$ restricted to the set S, i.e., conditioned on the event $F \subseteq S$.

Definition 9 (Conditional distribution)**.** For a distribution μ defined over subsets of a ground set $[n]$ and $T \subseteq [n]$, define μ^T to be the distribution of $F \sim \mu$ conditioned on the event $F \supseteq T$.

2.1 Markov chains and functional inequalities

Let μ and ν be probability measures on a finite set Ω. The Kullback-Leibler divergence (or relative entropy) between ν and μ is given by

$$\mathcal{D}_{\mathrm{KL}}(\nu \parallel \mu) = \sum_{x \in \Omega} \nu(x) \log\left(\frac{\nu(x)}{\mu(x)}\right),$$

with the convention that this is ∞ if ν is not absolutely continuous with respect to μ. By Jensen's inequality, $\mathcal{D}_{\mathrm{KL}}(\nu \parallel \mu) \geq 0$ for any probability measures μ, ν. The total variation distance between μ and ν is given by

$$d_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|.$$

A Markov chain on Ω is specified by a row-stochastic non-negative transition matrix $P \in \mathbb{R}^{\Omega \times \Omega}$. We refer the reader to [LP17] for a detailed introduction to the analysis of Markov chains. As is common, we will view probability distributions on Ω as row vectors. Recall that a transition matrix P is said to be reversible with respect to a distribution μ if for all $x, y \in \Omega$, $\mu(x)P(x,y) = \mu(y)P(y,x)$. In this case, it follows immediately that μ is a stationary distribution for P, i.e., $\mu P = \mu$. If P is further assumed to be ergodic, then μ is its unique stationary distribution, and for any probability distribution ν on Ω, $d_{\mathrm{TV}}(\nu P^t, \mu) \to 0$ as $t \to \infty$. The goal of this paper is to investigate the rate of this convergence.

Definition 10. Let P be an ergodic Markov chain on a finite state space Ω and let μ denote its (unique) stationary distribution. For any probability distribution ν on Ω and $\epsilon \in (0,1)$, we define $t_{\mathrm{mix}}(P, \nu, \epsilon)$ to be

$$\min\{t \geq 0 \mid d_{\mathrm{TV}}(\nu P^t, \mu) \leq \epsilon\},$$

and let $t_{\mathrm{mix}}(P, \epsilon)$ denote

$$\max\left\{\min\{t \geq 0 \mid d_{\mathrm{TV}}(\mathbb{1}_x P^t, \mu) \leq \epsilon\} \;\middle|\; x \in \Omega\right\},$$

where $\mathbb{1}_x$ is the point mass supported at x.

We will drop P and ν if they are clear from context. Moreover, if we do not specify ϵ, then it is set to $1/4$. This is because the growth of $t_{\mathrm{mix}}(P, \epsilon)$ is at most logarithmic in $1/\epsilon$ [cf. LP17].

Lemma 11. *Let μ be a probability measure on the finite set Ω. Let P denote the transition matrix of an ergodic, reversible Markov chain on Ω with stationary distribution μ. Suppose there exists some $\alpha \in (0,1]$ such that for all probability measures ν on Ω, we have*

$$\mathcal{D}_{\mathrm{KL}}(\nu P \parallel \mu P) \leq (1 - \alpha)\, \mathcal{D}_{\mathrm{KL}}(\nu \parallel \mu).$$

Then $t_{\mathrm{mix}}(P, \epsilon) \leq$

$$\left\lceil \frac{1}{\alpha} \cdot \left(\log\log\left(\frac{1}{\min\{\mu(x) \mid x \in \Omega\}}\right) + \log\left(\frac{1}{2\epsilon^2}\right)\right)\right\rceil.$$

This is the standard argument for bounding mixing times via modified log-Sobolev inequalities and can be found in, e.g., [BT06].

2.2 Strongly Rayleigh distributions

For density function $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$, the generating polynomial of μ is the multivariate k-homogeneous polynomial defined as follows:

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

Definition 12. Consider the open half-plane $H = \{z \mid \mathrm{Im}(z) > 0\} \subseteq \mathbb{C}$. We say a polynomial $g(z_1, \cdots, z_n) \in \mathbb{R}[z_1, \cdots, z_n]$ is real-stable if g does not have roots in H^n. For convenience, the zero polynomial is taken to be real-stable.

A distribution $\mu : 2^{[n]} \to \mathbb{R}_{\geq 0}$ is strongly Rayleigh iff its generating polynomial is real stable [see BBL09]. If μ is strongly Rayleigh, then its conditional and restricted distributions (see Definitions 8 and 9) are also strongly Rayleigh. The key fact we use about strongly Rayleigh distributions is that they are negatively correlated [see, e.g., BBL09], i.e., the marginals (of non-restricted elements) increase under restrictions (Definition 8):

$$\mathbb{P}_{S \sim \mu}[i \in S] \leq \mathbb{P}_{S \sim \mu_T}[i \in S] \text{ for } i \notin T.$$

2.3 Down-up and up-down walks

Definition 13 (Down operator)**.** For $\ell \leq k$ define the row-stochastic matrix $D_{k \to \ell} \in \mathbb{R}_{\geq 0}^{\binom{[n]}{k} \times \binom{[n]}{\ell}}$ by

$$D_{k \to \ell}(S, T) = \begin{cases} 0 & \text{if } T \not\subseteq S, \\ \frac{1}{\binom{k}{\ell}} & \text{otherwise.} \end{cases}$$

Note that for a distribution μ on size k sets, $\mu D_{k \to \ell}$ will be a distribution on size ℓ sets. In particular, $\mu D_{k \to 1}$ will be the vector of normalized marginals of μ: $(\mathbb{P}[i \in S]/k)_{i \in [n]}$, i.e., $p(\mu)/k$.

Definition 14 (Up operator)**.** For $\ell \leq k$ define the row-stochastic matrix $U_{\ell \to k} \in \mathbb{R}_{\geq 0}^{\binom{[n]}{\ell} \times \binom{[n]}{k}}$ by

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

As in [AD20; Ana+21a], we consider the following Markov chain M_μ^t 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:

1. Sample $T \in \binom{[n] \setminus S_0}{t-k}$ uniformly at random.
2. Downsample $S_1 \sim \mu_{S_0 \cup T}$, where $\mu_{S_0 \cup T}$ is μ restricted to $S_0 \cup T$ and update S_0 to be S_1.

Proposition 15. *The complement of S_1 is distributed according to $\bar{\mu}_0 D_{(n-k) \to (n-t)} U_{(n-t) \to (n-k)}$ where μ_0 is the distribution of the set S_0.*

Proposition 16. *For any distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ that is strongly Rayleigh, the chain M_μ^t for $t \geq k+1$ is irreducible, aperiodic and has stationary distribution μ.*

2.4 Entropic independence

We say that a distribution μ is entropically independent if the down operator $D_{k\to 1}$ significantly contracts the relative entropy between ν and μ for any distribution ν.

Definition 17 (Entropic independence [Ana+21b])**.** A probability distribution μ on $\binom{[n]}{k}$ is said to be $(1/\alpha)$-entropically independent, if for all probability distributions ν on $\binom{[n]}{k}$,

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

Any distribution with a log-concave generating polynomial (e.g., uniform on bases of a matroid) is 1-entropically independent. This includes all strongly Rayleigh distributions.

Lemma 18 ([Ana+21b, Theorem 4])**.** *Any strongly Rayleigh μ is 1-entropically independent. The conditional and restricted distributions of μ are also strongly Rayleigh, and thus 1-entropically independent.*

2.5 Average-case local-to-global method

First, we define the notion of the link of the distribution μ w.r.t. a set T [see, e.g., KO18]. This is almost the same as the conditioned distribution μ^T, see Definition 9, except we remove the set T.

Definition 19. For a distribution $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ and a set $T \subseteq [n]$ of size at most k, we define the *link of T* to be the distribution $\mu^{-T} : \binom{[n]-T}{k-|T|} \to \mathbb{R}_{\geq 0}$ which describes the law of the set $S - T$ where S is sampled from μ conditioned on the event $S \supseteq T$.

We show that entropic independence for links, i.e., contraction of KL-divergence by $D_{k\to 1}$ operators, results in the contraction of KL-divergence by $D_{k\to \ell}$ operators for larger ℓ. While this is by now a well-understood phenomenon, sometimes called the local-to-global method [see, e.g., AL20], we use an *average case* variant, adapted from [Ali+21], which only requires entropic independence for a "typical" link as opposed to a worst case link. We use that we can imagine deleting conditioned out elements in a *random order*. This is essential for our result, where it is *not true* that for every permutation the resulting product of $\rho(\cdot)$ below is large enough.

Theorem 20. *Suppose that for every set T of size $\leq k-2$, μ^{-T} contracts KL-divergence in terms of a factor parameterized by $\rho(T)$:*

$$\mathcal{D}_{\mathrm{KL}}(\nu D_{k-|T|\to 1} \parallel \mu^{-T} D_{k-|T|\to 1}) \leq (1-\rho(T))\,\mathcal{D}_{\mathrm{KL}}(\nu \parallel \mu^{-T}).$$

In other words assume that μ^{-T} is $(k-|T|)(1-\rho(T))$-entropically independent. For a set T, define the harmonic mean

$$\gamma_T := \mathbb{E}_{e_1,\ldots,e_{|T|} \text{ uniformly random permutation of } T}\left[\left(\rho(\emptyset)\rho(\{e_1\})\rho(\{e_1,e_2\})\cdots\rho(\{e_1,\ldots,e_{|T|-1}\})\right)^{-1}\right]^{-1}.$$

Then the operator $D_{k\to\ell}$ has KL-divergence contraction

$$\mathcal{D}_{\mathrm{KL}}(\nu D_{k\to\ell} \parallel \mu D_{k\to\ell}) \leq (1-\kappa)\,\mathcal{D}_{\mathrm{KL}}(\nu \parallel \mu),$$

with

$$\kappa := \min\left\{\gamma_T \;\middle|\; T \in \binom{[n]}{\ell}\right\}.$$

The proof is similar to [Ali+21, Theorem 46], [Ana+21b, Theorem 5], and is deferred to Section 7.

Remark 21. Similar to [Ali+21], if the KL-divergence is replaced by any other type of f-divergence, a common choice being χ^2-divergence which roughly relates to the notion of spectral independence, Theorem 20 still remains valid.

2.6 Isotropic transformation

Anari and Dereziński [AD20] introduced the following subdivision process that takes marginal overestimates of an arbitrary distribution $\mu \in \mathbb{R}^{\binom{[n]}{k}}$, and transforms sampling from μ to sampling from a distribution with nearly uniform marginals. In the following, we call μ' the *isotropic transformation* of μ.

Definition 22. Let $\mu : \binom{n}{k} \to \mathbb{R}_{\geq 0}$ be an arbitrary probability distribution, and assume that for some constant $c \geq 1$, we have marginal overestimates $p_1, \ldots, p_n$ of the marginals with $p_1 + \cdots + p_n \leq K$ and $p_i \geq \mathbb{P}_{S\sim\mu}[i \in S]$ for all i. Let $t_i := \lceil \frac{n}{K} p_i \rceil$. We will create a new distribution out of μ: For each $i \in [n]$, create t_i copies of the element i and let the collection of all these copies be the new ground set: $U = \bigcup_{i=1}^{n} \{i^{(1)}, \ldots, i^{(t_i)}\}$. Define the following distribution $\mu' : \binom{U}{k} \to \mathbb{R}_{\geq 0}$ from μ:

$$\mu'\left(\left\{i_1^{(j_1)}, \ldots, i_k^{(j_k)}\right\}\right) := \frac{\mu(\{i_1, \ldots, i_k\})}{t_1 \cdots t_k}.$$

Another way we can think of μ' is that to produce a sample from it, we can first generate a sample $\{i_1, \ldots, i_k\}$ from μ, and then choose a copy $i_m^{(j_m)}$ for each element i_m uniformly at random.

As show in [Ana+21a, Proposition 24], performing the isotropic transformation in Definition 22 at most only doubles the size of the universe U, but makes all marginals bounded by $O(K/n)$ now. For the convenience of the reader, we present the proofs of the following in Section 7.

Proposition 23. *Let $\mu : \binom{n}{k} \to \mathbb{R}_{\geq 0}$, and let $\mu' : \binom{U}{k} \to \mathbb{R}_{\geq 0}$ be the subdivided distribution from Definition 22. The following hold for μ':*

1. *Near-isotropy: For all $i^{(j)} \in U$, the marginal $\mathbb{P}_{S\sim\mu'}[i^{(j)} \in S] \leq \frac{K}{n} \leq \frac{2K}{|U|}$.*
2. *Linear ground set size: $|U| \leq 2n$.*
3. *If μ is strongly Rayleigh then so is μ'.*

3 Entropy contraction

Our goal is to prove an entropy contraction inequality for the $(n-k) \to 1$ down operator. For strongly Rayleigh distributions $\bar{\mu} \in \mathbb{R}^{\binom{[n]}{k}}$, which are 1-entropically independent (Definition 17), the entropy contracts by $1/(n-k)$. Surprisingly, if μ also has nearly uniform marginals, the entropy contracts even more, by an extra $\sim \log(n/k)$ factor.

Theorem 24 (Level one entropy contraction)**.** *Let $\mu \in \mathbb{R}^{\binom{[n]}{k}}$ be a 1-entropically independent distribution with $p(\mu)^{\max} := \max_{i\in[n]} \mathbb{P}_{F\sim\mu}[i \in F] \leq \frac{1}{100}$. Then for any distribution $\bar{\nu} \subseteq \mathbb{R}^{\binom{[n]}{n-k}}$,*

$$\mathcal{D}_{\mathrm{KL}}(\bar{\nu}D_{(n-k)\to 1} \parallel \bar{\mu}D_{(n-k)\to 1}) \leq \frac{1}{(n-k)\log((ep(\mu)^{\max})^{-1})}\,\mathcal{D}_{\mathrm{KL}}(\bar{\nu} \parallel \bar{\mu}).$$

We show this theorem by directly comparing the relative entropies of $p = \nu D_{k\to 1}$ and $q = \bar{\nu} D_{(n-k)\to 1}$ with respect to $\mu D_{k\to 1}$ and $\bar{\mu} D_{(n-k)\to 1}$ respectively. We first show a single-variable instance of this, which we sum over to get the overall comparison in Lemma 26.

Lemma 25. *Let $p, q, \alpha \in \mathbb{R}_{\geq 0}$ be such that $\alpha p + (1-\alpha)q = 1$. If $\alpha < 1/100$ then for any $\mu \in (0, 1/100)$,*

$$C_{\alpha,\mu}\, p \log(\alpha p/\mu) - q \log((1-\alpha)q/(1-\mu)) \geq K_{\alpha,\mu}\,(\alpha p - \mu)$$

for any constants $C_{\alpha,\mu} \geq \frac{\alpha}{(1-\alpha)\log(1/(e\mu))}$ and $K_{\alpha,\mu} := C_{\alpha,\mu}/\alpha + (1-\alpha)^{-1}$.

We explain some intuition behind this claim. First, both sides vanish when $p = \mu/\alpha$ and $q = (1-\mu)/(1-\alpha)$. The constants $C_{\alpha,\mu}$ and $K_{\alpha,\mu}$ are chosen so that the inequality is tight up to the second order at this point where $p = \mu/\alpha$ and $q = (1-\mu)/(1-\alpha)$.

Proof of Lemma 25. Define $f(q) := C_{\alpha,\mu}\, p \log(\alpha p/\mu) - q \log((1-\alpha)q/(1-\mu)) - K_{\alpha,\mu}\,(\alpha p - \mu)$. This is defined for $q \in [0, 1/(1-\alpha)]$. Note that $\frac{d}{dq} p = -\frac{1-\alpha}{\alpha}$. Hence

$$f'(q) = -C_{\alpha,\mu} \cdot \frac{1-\alpha}{\alpha}\left(\log(\alpha p/\mu) + 1\right) - \left(\log\left(\frac{(1-\alpha)q}{1-\mu}\right) + 1\right) + K_{\alpha,\mu}(1-\alpha).$$

By our careful choice of $K_{\alpha,\mu}$, we have $f'(\bar{q}) = 0$ for $\bar{q} = \frac{1-\mu}{1-\alpha}$. Additionally, we can calculate

$$f''(q) = p^{-1} q^{-1} \left(C_{\alpha,\mu} \left(\frac{1-\alpha}{\alpha}\right)^2 q - \frac{1-(1-\alpha)q}{\alpha} \right). \tag{1}$$

Note that the $f''(q) = 0$ at exactly one value of q, which we denote by q_2. Observe that $f''(q) \geq 0$ if and only if $q \geq q_2$, because the coefficient of q in Eq. (1) is positive. Hence

$$\begin{aligned}
f''(\bar{q}) &= p^{-1}\bar{q}^{-1}\left(C_{\alpha,\mu}\left(\frac{1-\alpha}{\alpha}\right)^2 \bar{q} - \frac{1-(1-\alpha)\bar{q}}{\alpha} \right) \\
&\geq p^{-1}\bar{q}^{-1}\left(\frac{\alpha}{(1-\alpha)\log(1/(e\mu))}\left(\frac{1-\alpha}{\alpha}\right)^2 \frac{1-\mu}{1-\alpha} - \frac{1-(1-\alpha)\frac{1-\mu}{1-\alpha}}{\alpha} \right) \\
&= p^{-1}\bar{q}^{-1}\left(\frac{1}{\log(1/(e\mu))} \cdot \frac{1-\mu}{\alpha} - \frac{\mu}{\alpha} \right) \\
&= p^{-1}\bar{q}^{-1} \frac{1}{\alpha \log(1/(e\mu))}\left(1 - \mu(1 + \log(1/(e\mu)))\right) \\
&= p^{-1}\bar{q}^{-1} \frac{1}{\alpha \log(1/(e\mu))}\left(1 + \mu \log \mu\right) \geq 0
\end{aligned}$$

because $\mu \log \mu \geq -1$ for $\mu \leq 1/100$.

Note that $f'(0) = f'(1/(1-\alpha)) = +\infty$. Recall that $f''(q) < 0$ for any $q < q_2$ and $f''(q) > 0$ for $q > q_2$. The above calculation implies $\bar{q} \geq q_2$. Thus $0 = f'(\bar{q}) \geq f'(q_2)$ and $f'(q) \geq f'(\bar{q}) = 0$ for $q \geq \bar{q}$. Since f' is decreasing in $[0, q_2]$, there is no $q_1 \in (0, q_2)$ such that $f'(q_1) = 0$. Thus f must increase in $[0, q_1]$ for $q_1 < q_2 < \bar{q}$, decrease in $[q_1, \bar{q}]$, and increase on $[\bar{q}, 1/(1-\alpha)]$. In particular, $f(q) \geq f(\bar{q})$ for all $q \in [0, 1]$. Since $f(\bar{q}) = 0$, we get the desired inequality. □

Lemma 26. *Let p, q be distributions on $[n]$ satisfying $\alpha p + (1-\alpha)q = \vec{1}/n$ for $\alpha = k/n$ and $\alpha < 1/100$. Then for any $\mu \in \mathbb{R}^n_{[0,1]}$ with $\|\mu\|_1 = k$ and $\mu^{\max} := \max_{i\in[n]} \mu_i < 1/100$,*

$$\mathcal{D}_{\mathrm{KL}}\left(q \,\middle\|\, \frac{\vec{1}-\mu}{n-k}\right) \le C_{\alpha,\mu^{\max}}\, \mathcal{D}_{\mathrm{KL}}\left(p \,\middle\|\, \frac{\mu}{k}\right) \tag{2}$$

with $C_{\alpha,\mu^{\max}} := \frac{\alpha}{(1-\alpha)\log(1/(e\mu^{\max}))}$ as in Lemma 25.

Proof. By Lemma 25 (for the choice $p = np_i$ and $q = nq_i$) and the fact that $C_{\alpha,\mu}$ is monotonically increasing in μ, we deduce that

$$C_{\alpha,\mu^{\max}} p_i \log(p_i k/\mu_i) - q_i \log(q_i(n-k)/(1-\mu_i)) \ge \frac{1}{n} K_{\alpha,\mu^{\max}}(\alpha n p_i - \mu_i).$$

Summing this over all i gives us

$$C_{\alpha,\mu^{\max}} \sum_{i\in[n]} p_i \log(p_i k/\mu_i) - \sum_{i\in[n]} q_i \log(q_i(n-k)/(1-\mu_i)) \ge \frac{1}{n} K_{\alpha,\mu^{\max}} \sum_{i\in[n]} (\alpha n p_i - \mu_i).$$

The r.h.s. equals 0, so we deduce the desired inequality. □

Now, Theorem 24 is an easy corollary of Lemma 26.

Proof of Theorem 24. Let ν, μ be the complement of $\bar{\nu}, \bar{\mu}$ resp. Let $p = \nu D_{k\to 1}$, $q = \bar{\nu} D_{(n-k)\to 1}$, and $\widehat{\mu}_i := \mathbb{P}_{F\sim\mu}[i \in F]$ for $i \in [n]$ in the setting of Lemma 26. Note that $\alpha p + (1-\alpha)q = \vec{1}/n$ for $\alpha = k/n$, and $\widehat{\mu}^{\max} = p(\mu)^{\max} \le \frac{1}{100}$.

Because μ is 1-entropically independent (Lemma 18),

$$\mathcal{D}_{\mathrm{KL}}\left(p \,\middle\|\, \frac{\widehat{\mu}}{k}\right) = \mathcal{D}_{\mathrm{KL}}(\nu D_{k\to 1} \,\|\, \mu D_{k\to 1}) \le \frac{1}{k}\, \mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu) = \frac{1}{k}\, \mathcal{D}_{\mathrm{KL}}(\bar{\nu} \,\|\, \bar{\mu}).$$

Combining this with Lemma 26 for $\alpha = k/n$ gives us

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}\left(q \,\middle\|\, \frac{\vec{1}-\widehat{\mu}}{n-k}\right) &\le C_{\alpha,\widehat{\mu}^{\max}}\, \mathcal{D}_{\mathrm{KL}}\left(p \,\middle\|\, \frac{\widehat{\mu}}{k}\right) \\
&\le \frac{\frac{\alpha}{k}\, \mathcal{D}_{\mathrm{KL}}(\bar{\nu} \,\|\, \bar{\mu})}{(1-\alpha)\log((ep(\mu)^{\max})^{-1})} \\
&= \frac{\mathcal{D}_{\mathrm{KL}}(\bar{\nu} \,\|\, \bar{\mu})}{(n-k)\log((ep(\mu)^{\max})^{-1})}.
\end{aligned}$$

□

We can almost directly combine Theorem 24 and the average-case local-to-global principle Theorem 20 to deduce an entropy contraction for $D_{(n-k)\to(n-k'+1)}$ and $D_{(n-k)\to(n-k-1)}$. The one remaining issue is that the local-to-global theorem requires that the marginals of *conditionals of* μ also have almost uniform marginals. This is the main result of Section 4, which we state here.

Theorem 27. *Let $\mu \in \mathbb{R}^{\binom{[n]}{k}}$ be a strongly Rayleigh distribution, and let $T \subseteq [n]$ with $|T| = \bar{k}$. For a sufficiently large constant C and any $s \ge C(np(\mu)^{\max} + \bar{k})\log n$, we have*

$$\mathbb{P}_{\substack{S\sim[n]\setminus T \\ |S|=n-s}}\left[p(\mu_{[n]\setminus S})^{\max} \ge \frac{2p(\mu)^{\max} n}{s}\right] \le n^{-10}.$$

We prove this in Section 4. We now have all the pieces to show our main technical result.

Theorem 28. *Let $\mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0}$ be a strongly Rayleigh distribution with $p(\mu)^{\max} \leq 1/500$. Let $s := C(np(\mu)^{\max} + \bar{k}) \log n$ for C be as in Theorem 27 and $k' = \Theta(np(\mu)^{\max})$. Then for any distribution $\bar{\nu} \subseteq \mathbb{R}^{\binom{[n]}{n-k}}$ and $\bar{k} \geq k+2$*

$$\mathcal{D}_{\mathrm{KL}}(\bar{\nu} D_{(n-k)\to(n-\bar{k}+1)} \parallel \bar{\mu} D_{(n-k)\to(n-\bar{k}+1)}) \leq (1-\kappa)\, \mathcal{D}_{\mathrm{KL}}(\bar{\nu} \parallel \bar{\mu})$$

with $\kappa = \frac{\bar{k}-k-1}{2s \log n}$. In particular,

$$\mathcal{D}_{\mathrm{KL}}(\bar{\nu} D_{(n-k)\to(n-k'+1)} \parallel \bar{\mu} D_{(n-k)\to(n-k'+1)}) \leq (1-\kappa_1)\, \mathcal{D}_{\mathrm{KL}}(\bar{\nu} \parallel \bar{\mu})$$

and

$$\mathcal{D}_{\mathrm{KL}}(\bar{\nu} D_{(n-k)\to(n-k-1)} \parallel \bar{\mu} D_{(n-k)\to(n-k-1)}) \leq (1-\kappa_2)\, \mathcal{D}_{\mathrm{KL}}(\bar{\nu} \parallel \bar{\mu})$$

with $\kappa_1^{-1} = O(\log^2 n)$ and $\kappa_2^{-1} = O(np(\mu)^{\max} \log^2 n)$.

Proof. Fix $\bar{k} \geq k+2$ to be chosen later, and a set $\bar{T} \subseteq [n]$ of size $n - \bar{k}$. Let $s := C(np(\mu)^{\max} + \bar{k}) \log n$ for C be as in Theorem 27. In the context of Theorem 20, we want to bound $\gamma_{\bar{T}}$ with respect to $\bar{\mu}$. Theorem 24 implies that the link of $\emptyset$ is $1/u_0$-entropically independent with $u_0 = \log((ep(\mu)^{\max})^{-1})$. Consider a random permutation $e_1, \ldots, e_{n-\bar{k}}$ of elements of $\bar{T}$. Note that each set $S_i := \{e_1, \ldots, e_i\}$ is a randomly sampled size-i subsets of $\bar{T}$. By using Theorem 27 and taking a union bound over $i \in [n-s]$, we have that except with probability $n^{-10} \times n = n^{-9}$, we have

$$p(\mu_{[n]\setminus S_i})^{\max} \leq \frac{2p(\mu)^{\max} n}{|[n] \setminus S_i|} \leq \frac{2p(\mu)^{\max}}{s} \leq \frac{2}{C \log n} < \frac{1}{100}$$

for C sufficiently large. Suppose this event holds. Note that the complement of $\mu_{[n]\setminus S_i}$ is exactly $\bar{\mu}^{S_i}$. Thus, Theorem 24 implies that this link is $1/u_i$-entropically independent where $u_i := \log(\frac{n-i}{2ep(\mu)^{\max} n})$.

As a result

$$\begin{aligned}
\rho(\emptyset)\rho(S_1)\cdots\rho(S_{n-s}) &\geq \prod_{i=0}^{n-s} \left(1 - \frac{1}{(n-k-i)u_i}\right) \\
&\overset{(i)}{\geq} \exp\left(-\sum_{i=0}^{n-s}\left(\frac{1}{(n-k-i)u_i} + \frac{1}{(n-k-i)^2 u_i^2}\right)\right) \\
&\overset{(ii)}{\geq} \exp\left(-\sum_{i=0}^{n-s}\left(\frac{1}{(n-i)u_i} + \frac{k+1}{(n-k-i)^2}\right)\right)
\end{aligned} \tag{3}$$

where in (i) we use $1 - x \geq \exp(-x - x^2)$ for $x \leq 1/2$, in (ii) we use

$$\frac{1}{(n-k-i)u_i} = \frac{1}{(n-i)u_i} + \frac{k}{(n-i)(n-k-i)u_i} \leq \frac{1}{(n-i)u_i} + \frac{k}{(n-k-i)^2}$$

Next, let $h := 2ep(\mu)^{\max} n$.

$$\sum_{i=0}^{n-s} \frac{1}{(n-i)u_i} \leq \int_s^n \frac{1}{t \log \frac{t}{h}} dt = \int_{s/h}^{n/h} \frac{1}{t \log t} dt = \log\log(n/h) - \log\log(s/h).$$

where in the final equality we use $(\log\log t)' = \frac{1}{t\log t}$. Similarly

$$\sum_{i=0}^{n-s} \frac{k+1}{(n-k-i)^2} \le (k+1)\int_{s-k}^{n-k} \frac{1}{t^2}dt = \frac{(k+1)(n-s)}{(s-k)(n-k)}$$

which is $\le O(1)$, where we use the fact that $s \ge 2Ck\log n \gg k$. Thus, we have that the l.h.s. in Eq. (3) $\ge \Omega(\frac{1}{\log(n/h)})$.

Moreover, for $i \in \{n-s+1, \cdots, n-\bar{k}\}$, by 1-entropic independence of $\bar{\mu}$ and its links (Lemma 18) we get the trivial bound $\rho(S_i) \ge 1 - \frac{1}{n-k-i}$. Thus

$$\prod_{i=n-s+1}^{n-\bar{k}} \rho(S_i) \ge \prod_{i=n-s+1}^{n-\bar{k}} \left(1 - \frac{1}{n-k-i}\right) = \frac{\bar{k}-k-1}{s-k-1} \ge \frac{\bar{k}-k-1}{s}$$

Thus, with probability at least $1-n^{-9}$, $\prod_{i=0}^{n-\bar{k}} \rho(S_i) \ge \frac{\bar{k}-k-1}{s\log n}$. Otherwise we have a trivial lower bound of $1/n$ on the product of ρs due to 1-entropic independence of $\bar{\mu}$ and its links. Thus

$$\gamma_T = \mathbb{E}\left[\prod_{i=0}^{n-\bar{k}} \rho(S_i)^{-1}\right]^{-1} \ge \left((1-n^{-9})\cdot\frac{s\log n}{\bar{k}-k-1} + n^{-9}\cdot n\right)^{-1} \ge \frac{\bar{k}-k-1}{2s\log n}.$$

We are done with the general entropy contraction statement. Next, we prove entropy contraction for specific values of $\bar{k}$. Plugging in $\bar{k} = k'$ and noting that for our choice of k' and s,

$$\bar{k}-k-1 \ge \frac{1}{2}\bar{k} \ge \Omega\left(\frac{s}{\log n}\right)$$

implies the first result. Similarly, setting $\bar{k} = k+2$ gives the second result for our choice of s. □

4 Concentration of marginals

The goal of this section is to show concentration of marginal upper bounds (Theorem 27) for random conditionals of strongly Rayleigh distributions. In Section 3, this is applied in the context of an average-case local-to-global principle (Theorem 20) to deduce our main entropy contraction result (Theorem 28). The proof uses the following simple observation about covariances in a set-valued distribution μ whose support contains only sets of identical size.

Lemma 29 (Covariances of homogeneous distributions)**.** *For any distribution μ defined over identically-sized subsets of a ground set of elements $[n]$ and any element $i \in [n]$ we have*

$$\sum_{j\in[n]} \left(p(\mu)_i p(\mu)_j - \mathbb{P}_{F\sim\mu}[i,j\in F]\right) = 0.$$

We show Theorem 27 by analyzing the marginal of each coordinate $i \in [n]$ conditioned on it remaining in $[n]\backslash S$ via a stochastic process. Formally, fix T as in Theorem 27 and a coordinate $i \in [n]$ (possibly in T).

Definition 30 (Stochastic process)**.** For fixed $T \subseteq [n]$ with $|T| = \bar{k}$, $i \in [n]$, and $s \le n - |T| - 1$, let σ be a random permutation of $[n]\backslash(T\cup\{i\})$. For $0 \le t \le n-s$ define $S_t = \{\sigma(1), \sigma(2), \ldots, \sigma(t)\}$. Define $S = S_{n-s}$ and $\mu^{(t)} := \mu_{[n]\backslash S_t}$.

Note that S_t is generated from S_{t-1} by removing a random element in $[n]\backslash(T \cup \{i\} \cup S_{t-1})$. Now we can analyze $p(\mu_{[n]\backslash S})_i = p(\mu^{(n-s)})_i$ by analyzing the stochastic process $p(\mu^{(t)})_i$. We start by analyzing its drift.

Lemma 31 (Expected drift)**.** *With the setup in Definition 30 and $0 \le t < n - s$, we have*

$$\mathbb{E}_{S_{t+1}}\left[p(\mu^{(t+1)})_i \;\middle|\; S_t\right] - p(\mu^{(t)})_i \le \frac{(1 - p(\mu^{(t)})^{\max})^{-1}}{n - \bar{k} - 1 - t} p(\mu^{(t)})_i.$$

Proof. By definition, we know that if $S_{t+1} = S_t \cup \{j\}$ for some $j \in [n]\backslash(T \cup \{i\} \cup S_t)$, then

$$\begin{aligned} p(\mu^{(t+1)})_i - p(\mu^{(t)})_i &= \frac{\mathbb{P}_{F\sim\mu^{(t)}}[i \in F, j \notin F]}{1 - p(\mu^{(t)})_j} - p(\mu^{(t)})_j \\ &= \frac{p(\mu^{(t)})_i p(\mu^{(t)})_j - \mathbb{P}_{F\sim\mu^{(t)}}[i, j \in F]}{1 - p(\mu^{(t)})_j}. \end{aligned} \tag{4}$$

Hence by Equation (4),

$$\begin{aligned} &\mathbb{E}_{S_{t+1}}\left[p(\mu^{(t+1)})_i \;\middle|\; S_t\right] - p(\mu^{(t)})_i \\ = \;& \frac{1}{n - |T \cup \{i\} \cup S_t|} \sum_{j \in [n]\backslash(T \cup \{i\} \cup S_t)} \frac{p(\mu^{(t)})_i p(\mu^{(t)})_j - \mathbb{P}_{F\sim\mu^{(t)}}[i, j \in F]}{1 - p(\mu^{(t)})_j}. \end{aligned} \tag{5}$$

Because μ and hence $\mu^{(t)}$ is strongly Rayleigh, each numerator of the fractions in Equation (5) is nonnegative for $j \neq i$, hence the expression in Equation (5) is at most

$$\begin{aligned} &\frac{(1 - p(\mu^{(t)})^{\max})^{-1}}{n - \bar{k} - 1 - t} \sum_{j \in [n]\backslash\{i\}} \left(p(\mu^{(t)})_i p(\mu^{(t)})_j - \mathbb{P}_{F\sim\mu^{(t)}}[i, j \in F]\right) \\ = \;& \frac{(1 - p(\mu^{(t)})^{\max})^{-1}}{n - \bar{k} - 1 - t} \left(p(\mu^{(t)})_i - p(\mu^{(t)})_i^2 + \sum_{j \in [n]} \left(p(\mu^{(t)})_i p(\mu^{(t)})_j - \mathbb{P}_{F\sim\mu^{(t)}}[i, j \in F]\right)\right) \\ \overset{(i)}{\le} \;& \frac{(1 - p(\mu^{(t)})^{\max})^{-1}}{n - \bar{k} - 1 - t} p(\mu^{(t)})_i, \end{aligned}$$

where (i) follows from Lemma 29. This completes the proof. □

Now we analyze the variance/maximum change in $p(\mu^{(t)})_i$.

Lemma 32 (Variance and maximum change)**.** *With the setup in Definition 30 and $0 \le t < n - s$, we have with probability 1 conditioned on S_t that*

$$p(\mu^{(t+1)})_i - p(\mu^{(t)})_i \le \frac{p(\mu^{(t)})^{\max}}{1 - p(\mu^{(t)})^{\max}} p(\mu^{(t)})_i. \tag{6}$$

Also, we have the variance bound

$$\mathbb{E}_{S_{t+1}}\left[\left(p(\mu^{(t+1)})_i - p(\mu^{(t)})_i\right)^2 \;\middle|\; S_t\right] \le \frac{1}{n - \bar{k} - 1 - t} \cdot \frac{p(\mu^{(t)})^{\max}}{(1 - p(\mu^{(t)})^{\max})^2} p(\mu^{(t)})_i^2.$$

Proof. By the formula in Equation (4) we get

$$p(\mu^{(t+1)})_i - p(\mu^{(t)})_i = \frac{p(\mu^{(t)})_i p(\mu^{(t)})_j - \mathbb{P}_{F\sim\mu^{(t)}}[i,j\in F]}{1-p(\mu^{(t)})_j} \leq \frac{p(\mu^{(t)})^{\max}}{1-p(\mu^{(t)})^{\max}} p(\mu^{(t)})_i.$$

Because μ and hence $\mu^{(t)}$ is strongly Rayleigh, $p(\mu^{(t+1)})_i \geq p(\mu^{(t)})_i$ so

$$\begin{aligned}
&\mathbb{E}_{S_{t+1}}\left[\left(p(\mu^{(t+1)})_i - p(\mu^{(t)})_i\right)^2 \;\middle|\; S_t\right] \\
\overset{(i)}{\leq}\; &\frac{p(\mu^{(t)})^{\max}}{1-p(\mu^{(t)})^{\max}} p(\mu^{(t)})_i \,\mathbb{E}_{S_{t+1}}\left[p(\mu^{(t+1)})_i - p(\mu^{(t)})_i \;\middle|\; S_t\right] \\
\overset{(ii)}{\leq}\; &\frac{1}{n-\bar{k}-1-t}\cdot\frac{p(\mu^{(t)})^{\max}}{(1-p(\mu^{(t)})^{\max})^2} p(\mu^{(t)})_i^2,
\end{aligned}$$

where (i) follows from Equation (6) and (ii) follows from Lemma 31. □

Our desired concentration bound now essentially follows from a careful application of Bernstein's inequality for martingales to the sequence $\log p(\mu^{(t)})_i$.

Theorem 33 ([CL06, Theorem 20]). *Let $X^{(0)}, X^{(1)}, \ldots, X^{(t)}$ be a martingale such that $X^{(u)} - X^{(u-1)} \leq M$ with probability 1 and $\mathrm{Var}\left[X^{(u)} \;\middle|\; X^{(u-1)}\right] \leq \sigma_u^2$ for $u \in [t]$. Then*

$$\mathbb{P}\left[X^{(t)} - X^{(0)} \geq \lambda\right] \leq \exp\left(-\frac{\lambda^2}{2\sum_{u\in[t]}\sigma_u^2 + 2M\lambda/3}\right).$$

Proof of Theorem 27. For the setup in Definition 30, define the random variables $Y^{(t)} := \log p(\mu^{(t)})_i$ indexed by $0 \leq t \leq n-s$. Given this, we define the martingale $X^{(0)} = Y^{(0)}$ and

$$X^{(t+1)} := X^{(t)} + Y^{(t+1)} - Y^{(t)} - \mathbb{E}_{S_{t+1}}\left[Y^{(t+1)} - Y^{(t)} \;\middle|\; S_t\right],$$

for $0 \leq t < n-s$.

We will bound the drift, maximum change, and variance of $Y^{(t+1)}$ assuming that $p(\mu^{(t)})^{\max} \leq 2p(\mu)^{\max}n/(n-t) < 1/10$ (which we want to show holds with high probability). We may assume this, because we can just prematurely stop the stochastic process whenever this condition breaks. For the drift term, we bound

$$\begin{aligned}
&\mathbb{E}_{S_{t+1}}\left[Y^{(t+1)} - Y^{(t)} \;\middle|\; S_t\right] = \mathbb{E}_{S_{t+1}}\left[\log(p(\mu^{(t+1)})_i / p(\mu^{(t)})_i) \;\middle|\; S_t\right] \\
\leq\; &\mathbb{E}_{S_{t+1}}\left[\frac{p(\mu^{(t+1)})_i - p(\mu^{(t)})_i}{p(\mu^{(t)})_i} \;\middle|\; S_t\right] \overset{(i)}{\leq} \frac{(1-p(\mu^{(t)})^{\max})^{-1}}{n-\bar{k}-1-t} \\
\overset{(ii)}{\leq}\; &\frac{1}{n-\bar{k}-1-t} + \frac{4p(\mu)^{\max}n}{(n-\bar{k}-1-t)(n-t)},
\end{aligned} \tag{7}$$

where (i) follows from Lemma 31 and (ii) follows from our assumption on $p(\mu^{(t)})^{\max}$. For the bound on the maximum increase, we get

$$\begin{aligned}
Y^{(t+1)} - Y^{(t)} &= \log\left(p(\mu^{(t+1)})_i / p(\mu^{(t)})_i\right) \leq \frac{p(\mu^{(t+1)})_i - p(\mu^{(t)})_i}{p(\mu^{(t)})_i} \\
&\leq p(\mu^{(t+1)})_i - p(\mu^{(t)})_i \leq \frac{p(\mu^{(t)})^{\max}}{1-p(\mu^{(t)})^{\max}} \leq \frac{4p(\mu)^{\max}n}{n-t}.
\end{aligned} \tag{8}$$

by Lemma 32 Equation (6) and our assumption on $p(\mu^{(t)})^{\max}$. For the variance term, we first bound

$$\begin{aligned}
&\mathbb{E}_{S_{t+1}}\left[\left(Y^{(t+1)}-Y^{(t)}\right)^2 \;\middle|\; S_t\right] = \mathbb{E}_{S_{t+1}}\left[\log\left(\mu^{(t+1)})_i/p(\mu^{(t)})_i\right)^2 \;\middle|\; S_t\right] \\
\leq\; &\mathbb{E}_{S_{t+1}}\left[\left(\frac{p(\mu^{(t+1)})_i - p(\mu^{(t)})_i}{p(\mu^{(t)})_i}\right)^2 \;\middle|\; S_t\right] \\
\leq\; &\frac{1}{n-\bar{k}-1-t}\cdot\frac{p(\mu^{(t)})^{\max}}{(1-p(\mu^{(t)})^{\max})^2} \leq \frac{8np(\mu)^{\max}}{(n-\bar{k}-1-t)(n-t)} \qquad (9)
\end{aligned}$$

by Lemma 32 and our assumption on $p(\mu^{(t)})^{\max}$. Our next goal is to prove that $X^{(t)} \leq X^{(0)} + 1/10$ with high probability by using Theorem 33. Because μ and hence $\mu^{(u)}$ is strongly Rayleigh for all $0 \leq u \leq t$,

$$X^{(u+1)} - X^{(u)} \leq Y^{(u+1)} - Y^{(u)} \leq \frac{4p(\mu)^{\max}n}{n-u} \leq \frac{4p(\mu)^{\max}n}{s} \leq \frac{1}{1000\log n}$$

by Equation (8) and sufficiently large C for $s \geq C(np(\mu)^{\max} + \bar{k})\log n$ so we may take $M = 1/(1000\log n)$ in Theorem 33. Additionally, we have that

$$\begin{aligned}
\mathrm{Var}\left[X^{(u+1)} \;\middle|\; X^{(u)}\right] &\leq \mathbb{E}_{S_{t+1}}\left[\left(Y^{(t+1)}-Y^{(t)}\right)^2 \;\middle|\; S_t\right] \\
&\leq \frac{8np(\mu)^{\max}}{(n-\bar{k}-1-u)(n-u)} \leq \frac{8np(\mu)^{\max}}{(n-\bar{k}-1-u)^2},
\end{aligned}$$

by Equation (9) so we may take $\sigma_u^2 = 8np(\mu)^{\max}/(n-\bar{k}-1-u)^2$ in Theorem 33. By Theorem 33 for $\lambda = 1/10$, $M = 1/(1000\log n)$, and $\sigma_u^2 = 8np(\mu)^{\max}/(n-\bar{k}-1-u)^2$, we get that

$$\begin{aligned}
&\mathbb{P}\left[X^{(t)} - X^{(0)} \geq 1/10\right] \leq \exp\left(-\frac{1/100}{2\sum_{u\in[t]}\sigma_u^2 + M/15}\right) \\
\leq &\exp\left(-\frac{1/100}{16np(\mu)^{\max}\sum_{u\leq n-s}\frac{1}{(n-\bar{k}-1-u)^2} + \frac{1}{15000\log n}}\right) \leq \exp\left(-\frac{1/100}{\frac{50np(\mu)^{\max}}{s-\bar{k}-1} + \frac{1}{15000\log n}}\right) \\
\leq &\exp(-20\log n) = n^{-20}
\end{aligned}$$

for sufficiently large C in $s \geq C(np(\mu)^{\max} + \bar{k})\log n \geq C\bar{k}\log n$. To finish, note that

$$\begin{aligned}
Y^{(t)} - Y^{(0)} &= X^{(t)} - X^{(0)} + \sum_{u\in[t-1]}\mathbb{E}_{S_{u+1}}\left[Y^{(u+1)} - Y^{(u)} \;\middle|\; S_u\right] \\
&\overset{(i)}{\leq} X^{(t)} - X^{(0)} + \sum_{u\in[t-1]}\frac{1}{n-\bar{k}-1-u} + \frac{4p(\mu)^{\max}n}{(n-\bar{k}-1-u)(n-u)} \\
&\overset{(ii)}{\leq} X^{(t)} - X^{(0)} + \log\left(\frac{n-\bar{k}-1}{n-\bar{k}-1-t}\right) + \frac{20p(\mu)^{\max}n}{n-\bar{k}-1-t} \\
&\overset{(iii)}{\leq} X^{(t)} - X^{(0)} + \log\left(\frac{n}{n-t}\right) + \frac{1}{100\log n}, \qquad (10)
\end{aligned}$$

where (i) follows from Equation (7), (ii) follows from direction calculations with Riemann integrals, and (iii) follows from $t \le n - s$ and $s \ge C(np(\mu)^{\max} + \bar{k}) \log n$. To conclude, we write

$$\begin{aligned}\mathbb{P}\left[p(\mu^{(t)})_i \ge \frac{2p(\mu)_i n}{n-t}\right] &= \mathbb{P}\left[Y^{(t)} - Y^{(0)} \ge \log\left(\frac{n}{n-t}\right) + \log 2\right] \\ &\overset{(i)}{\le} \mathbb{P}\left[X^{(t)} - X^{(0)} \ge 1/10\right] \le n^{-20},\end{aligned}$$

where (i) follows from Equation (10). Theorem 27 now follows from union-bounding over all times $t \in [n-s]$ and all coordinates $i \in [n]$. □

5 Isotropic rounding

We give a reduction which estimates marginals of a distribution given an algorithm that samples using marginal overestimates. At a high level, we split our original strongly Rayleigh distribution $\mu \in \mathbb{R}^{\binom{[n]}{k}}$ into two smaller distributions (supported on S_1, S_2 for $[n] = S_1 \sqcup S_2$), and recursively produce marginal overestimates for μ_{S_1} and μ_{S_2} that sum to at most $4k$. Now, we merge these groups. Because μ is strongly Rayleigh, the marginal overestimates on μ_{S_1} and μ_{S_2} provide marginal overestimates for μ summing to at most $8k$. Thus, we can cheaply take $O(n \log n/k)$ samples from μ to get marginal overestimates summing to at most $4k$ again.

Theorem 34 (Isotropic rounding from sampling)**.** *Let $\mu \in \mathbb{R}^{\binom{[n]}{k}}$ be a strongly Rayleigh distribution. Assume that we can sample from restrictions μ_S of μ (Definition 8) in time $\mathcal{A}_\mu(K)$ given marginal overestimates of μ_S that sum to at most K[2]. Then there is an algorithm that produces marginal overestimates $q_i \ge p(\mu)_i$ with sum $\sum_{i\in[n]} q_i \le 4k$ in time $\widetilde{O}\big(n/k \cdot \mathcal{A}_\mu(8k)\big)$.*

Proof. We use a divide-and-conquer procedure. Precisely, given a set S we use the following algorithm to produce marginals overestimates of S summing to at most $4k$. If $|S| \le 4k$, then we let all our overestimates be 1. Otherwise, partition $S = S_1 \sqcup S_2$ into equally sized pieces, and recursively produce marginal overestimates $q_i^{(1)} \ge p(\mu_{S_1})_i$ and $q_i^{(2)} \ge p(\mu_{S_2})_i$ with $\sum_{i\in S_1} q_i^{(1)} \le 4k$ and $\sum_{i\in S_2} q_i^{(2)} \le 4k$.

Because μ is strongly Rayleigh, in fact $q_i^{(1)} \ge p(\mu_S)_i$ for all $i \in S_1$ and $q_i^{(2)} \ge p(\mu_S)_i$ for all $i \in S_2$. Hence the vector $\bar{q} \in \mathbb{R}^S$ defined as $\bar{q}_i = q_i^{(1)}$ for $i \in S_1$ and $\bar{q}_i = q_i^{(2)}$ for $i \in S_2$ are marginal overestimates for μ_S. Additionally, $\sum_{i\in S} \bar{q}_i \le 8k$.

Set $s = \frac{100\,|S| \log n}{k}$, and let $F_1, F_2, \ldots, F_s$ be independent samples from μ_S generated in total time $\widetilde{O}(s \cdot \mathcal{A}_\mu(8k))$, by using the overestimates $\bar{q}$. Define for $i \in S$

$$q_i = \max\left\{\frac{k}{|S|}, \frac{2\,|\{s' \in [s] \mid i \in F_{s'}\}|}{s}\right\}.$$

We claim that q_i are marginal overestimates for μ_S with high probability and sum to at most $4k$. The sum follows because

$$\sum_{i\in S} q_i \le \sum_{i\in S}\left(\frac{k}{|S|} + \frac{2\,|\{s' \in [s] \mid i \in F_{s'}\}|}{s}\right) \le k + \sum_{s'\in[s]} \frac{2\,|F_{s'}|}{s} = 3k.$$

[2]We do not write an n dependence as it will be polylogarithmic in our algorithms.

Now we show that q_i are marginal overestimates of μ_S. The case $p(\mu_S)_i \leq k/|S|$ is trivial. Otherwise, by a Chernoff bound,

$$\mathbb{P}_{F_1,\ldots,F_s}\left[\left|\{s' \in [s] \mid i \in F_{s'}\}\right| \leq p(\mu_S)_i s/2\right] \leq \exp(-p(\mu_S)_i s/8) \leq n^{-100}$$

by the choice $s = \frac{100|S|\log n}{k}$ and $p(\mu_S)_i \geq k/|S|$.

The final runtime claim follows from recursively calling the above describde algorithm on μ, and using that there are $O(\log n)$ layers, each with total size n. Precisely, the total number of samples taken in a layer is at most the sum over $s = \frac{100|S|\log n}{k}$ in a layer, which is $O(n/k \cdot \log n)$. □

6 Proofs of main results

In this section, we combine our previous results to show Theorems 3 and 35 and Corollaries 6 and 7. All results will follow from our entropy contraction bound Theorem 28 combined with Lemma 11. Also, Theorem 35 requires Theorem 34.

Proof of Theorem 3. We first apply Proposition 23 to instead focus on sampling from a strongly Rayleigh distribution μ' with all marginals bounded by K/n. Let $\bar{\mu}'$ be the complement of μ'.

For κ_1 as in Theorem 28 we run $O(\kappa_1^{-1}\log n)$ steps of a down-up operator on the complement of our set, to converge to $\bar{\mu}'$, i.e., iterate the Markov chain $D_{(n-k)\to(n-k'+1)}U_{(n-k'+1)\to(n-k)}$. Note that each $U_{(n-k'+1)\to(n-k)}$ part needs to be implemented via a baseline sampling algorithm which takes time $\mathcal{T}(k'-1,k)$. By Theorem 28 and Lemma 11, and the fact that the up step $U_{(n-k'+1)\to(n-k)}$ cannot increase entropy, the chain mixes in $O(\kappa_1^{-1}\log n) = O(\log^3 n)$ steps. Thus the runtime is bounded by making $O(\log^3 n)$ calls to $\mathcal{T}_\mu(O(K),k)$ as $k' = \Theta(np(\mu')^{\max}) = O(K)$ from Theorem 28. □

Proof of Corollary 6. Let $n = |E|, k = |V|$. By the results of [Ana+21c] a spanning tree can be sampled in $\widetilde{O}(|E|)$ time on a graph with edge set E. Hence applying Theorem 3 with $K = O(k) = \widetilde{O}(|V|)$ shows that after obtaining the initial overestimates, each future sample requires $\widetilde{O}(|V|)$ time. Obtaining the original overestimates takes time $\widetilde{O}(n/k \cdot k) = \widetilde{O}(|E|)$ by Theorem 34. □

Proof of Corollary 7. Let μ be the k-DPP with ensemble matrix L. Sampling from a k-DPP over ground set of size n can be done in $\widetilde{O}(n^\omega)$ time, see Lemma 36. This together with Theorem 34 shows that in $\tilde{O}(n/k \cdot k^\omega) = \widetilde{O}(nk^{\omega-1})$ time, we can get marginal overestimates that sum to $\tilde{O}(k)$, as well as one initial sample S_0 from μ. We now apply Theorem 3, and note that each oracle call is equivalent to sampling from a k-DPP on a size-$O(K)$ subset of $[n]$, which takes $\tilde{O}(k^\omega)$ time. □

We now state a result on sampling strongly Rayleigh distributions using up-down steps. While it is known that this Markov chain normally mixes in $\widetilde{O}(n)$ steps [CGM19], we show that under isotropy of μ it mixes in $\widetilde{O}(k)$ steps.

Theorem 35. *Given a strongly Rayleigh distribution $\mu \in \mathbb{R}^{\binom{[n]}{k}}$ and marginal overestimates $q_i \geq \mathbb{P}_{T\sim\mu}[i \in T]$ for $i \in [n]$ which sum to $K := \sum_{i\in[n]} q_i$, there is an algorithm that samples from a distribution with total variation distance $n^{-O(1)}$ from μ in time bounded by $\widetilde{O}(K)$ calls to $\mathcal{T}_\mu(k+1,k)$. Additionally, given a strongly Rayleigh distribution $\mu \in \mathbb{R}^{\binom{[n]}{k}}$, we can produce marginal overestimates $q_i \geq \mathbb{P}_{T\sim\mu}[i \in T]$ with sum $\sum_{i\in[n]} q_i \leq O(k)$ in time $\widetilde{O}(n \cdot \mathcal{T}_\mu(k+1,k))$.*

Proof of Theorem 35. For κ_2 as in Theorem 28 we run $O(\kappa_2^{-1} \log n)$ steps of the one level down-up operator on the complement of our set, to converge to μ', i.e., iterate the Markov chain $D_{(n-k)\to(n-k-1)} U_{(n-k-1)\to(n-k)}$. Note that each $U_{(n-k+1)\to(n-k)}$ part needs to be implemented via a baseline sampling algorithm which takes time $\mathcal{T}(k+1,k)$. By Theorem 28 and Lemma 11, and the fact that the up operator $U_{(n-k-1)\to(n-k)}$ cannot increase entropy, the chain mixes in $O(\kappa_2^{-1} \log n) = O(K \log^3 n)$ steps. Thus the runtime is bounded by making $O(K \log^3 n)$ calls to $\mathcal{T}_\mu(k+1,k)$.

Finally, by Theorem 34, we can obtain the desired overestimates q_i in time

$$\widetilde{O}(n/k \cdot 8k \cdot \mathcal{T}_\mu(k+1,k)) = \widetilde{O}(n \cdot \mathcal{T}_\mu(k+1,k))$$

as desired. □

7 Deferred proofs

Proof of Theorem 20. Let ν be an arbitrary distribution. Let $f(x) := x \log x$ and note that

$$\mathcal{D}_{\mathrm{KL}}(\nu \parallel \mu) = \mathbb{E}_{S\sim\mu}[f(\nu(S)/\mu(S))] - f(\mathbb{E}_{S\sim\mu}[\nu(S)/\mu(S)]).$$

Consider the following process: We sample a set $S \sim \mu$ and uniformly at random permute its elements to obtain $X_1, \ldots, X_k$. Define the random variable

$$\tau_i = f\left(\mathbb{E}\left[\frac{\nu(S)}{\mu(S)} \;\middle|\; X_1, \ldots, X_i\right]\right) = f\left(\frac{\sum_{S' \ni X_1,\ldots,X_i} \nu(S')}{\sum_{S' \ni X_1,\ldots,X_i} \mu(S')}\right) = f\left(\frac{\nu D_{k\to i}(\{X_1,\ldots,X_i\})}{\mu D_{k\to i}(\{X_1,\ldots,X_i\})}\right).$$

Note that τ_i is a "function" of $X_1, \ldots, X_i$. It is not hard to see that

$$\mathcal{D}_{\mathrm{KL}}(\nu \parallel \mu) = \mathbb{E}[\tau_k] - \mathbb{E}[\tau_0] = \sum_{i=0}^{k-1} \mathbb{E}[\tau_{i+1} - \tau_i].$$

Conveniently, we obtain $\mathcal{D}_{\mathrm{KL}}(\nu D_{k\to\ell} \parallel \mu D_{k\to\ell})$ by just summing over the first ℓ terms:

$$\mathcal{D}_{\mathrm{KL}}(\nu D_{k\to\ell} \parallel \mu D_{k\to\ell}) = \mathbb{E}[\tau_\ell] - \mathbb{E}[\tau_0] = \sum_{i=0}^{\ell-1} \mathbb{E}[\tau_{i+1} - \tau_i].$$

Our goal is to show that the sum of the last $k - \ell$ terms are at least κ times the entire sum. Applying the assumption of local contraction to the link of the set $T = \{X_1, \ldots, X_i\}$, we get

$$\mathbb{E}[\tau_{i+1} - \tau_i \mid X_1, \ldots, X_i] \le (1 - \rho(T)) \cdot \mathbb{E}[\tau_k - \tau_i \mid X_1, \ldots, X_i],$$

which we rewrite as

$$\mathbb{E}[\tau_k - \tau_{i+1} \mid X_1, \ldots, X_i] \ge \rho(T) \cdot \mathbb{E}[\tau_k - \tau_i \mid X_1, \ldots, X_i].$$

Define the random variable

$$Z_i := \frac{\tau_k - \tau_i}{\rho(\emptyset)\rho(\{X_1\}) \cdots \rho(\{X_1, \ldots, X_{i-1}\})},$$

and note that our previous inequality simplifies to $\mathbb{E}[Z_{i+1} \mid X_1, \ldots, X_i] \geq \mathbb{E}[Z_i \mid X_1, \ldots, X_i]$. Chaining these inequalities we get that $\mathbb{E}[Z_\ell] \geq \mathbb{E}[Z_0] = \mathcal{D}_{\mathrm{KL}}(\nu \parallel \mu)$. We can further simplify $\mathbb{E}[Z_\ell]$ by noting that the numerator $\tau_k - \tau_\ell$ remains the same if we permute $X_1, \ldots, X_\ell$.

$$\mathbb{E}[Z_\ell \mid \{X_1, \ldots, X_\ell\}, \{X_{\ell+1}, \ldots, X_k\}] = \frac{\tau_k(\{X_1, \ldots, X_k\}) - \tau_\ell(\{X_1, \ldots, X_\ell\})}{\gamma(\{X_1, \ldots, X_\ell\})}.$$

Taking a further expectation, we get

$$\mathbb{E}[Z_\ell] \leq \frac{\mathbb{E}[\tau_k - \tau_\ell]}{\min\left\{\gamma(T) \;\middle|\; T \in \binom{[n]}{\ell}\right\}}.$$

This together with $\mathbb{E}[Z_\ell] \geq \mathbb{E}[Z_0] = \mathbb{E}[\tau_k - \tau_0]$ completes the proof. □

Proof of Proposition 23. Clearly, $\mathbb{P}_{S \sim \mu'}[i^{(j)} \in S] \leq p_i / t_i \leq K/n$. Also,

$$|U| = \sum_{i \in [n]} t_i \leq \sum_{i \in [n]} \left(1 + \frac{n}{K} p_i\right) \leq n + n \cdot \frac{\sum_i p_i}{K} \leq 2n.$$

For the third property, if μ has the generating polynomial $g_\mu(z_1, \ldots, z_n)$, then the distribution μ' obtained by subdividing element i into t_i copies has generating polynomial

$$g_{\mu'}(z_1^{(1)}, \ldots, z_n^{(t_n)}) = g_\mu\left(\frac{z_1^{(1)} + \ldots + z_1^{(t_1)}}{t_1}, \ldots, \frac{z_n^{(1)} + \ldots + z_n^{(t_n)}}{t_n}\right).$$

Clearly, if g_μ is real-stable then so is $g_{\mu'}$. This is because if z_i^j are chosen from the upper half plane $\{z \in \mathbb{C} \mid \mathrm{Im}(z) > 0\}$, their averages also lie in the upper half plane. □

Lemma 36. *Given an $n \times n$ positive semidefinite matrix L and an integer $k \leq n$, we can sample from the k-DPP defined by L in time $\widetilde{O}(n^\omega)$.*

We remark that variants of this statement where slow (but more practical) matrix multiplication algorithms are used, which result in cubic $\widetilde{O}(n^3)$ runtimes, already exist in the literature. Here, we simply formalize the observation that these algorithms can be adapted to take advantage of fast matrix multiplication and thus the runtime can be reduced to $\widetilde{O}(n^\omega)$.[3]

Proof. Kulesza and Taskar [KT12] reduce the task of sampling from a k-DPP defined by L to sampling from a (size-unconstrained) DPP. This is achieved by performing a spectral decomposition of the kernel matrix, choosing a subset of exactly k eigenvectors, each subset chosen with probability proportional to the product of the corresponding eigenvalues and forming a new kernel matrix just from the chosen eigenvectors. For details, see [KT12]. We simply remark that an approximate spectral decomposition of L is the most expensive operation here (while choosing the subset of eigenvectors can be done in $O(n^2)$ time). Thus, this part of the algorithm takes time $\widetilde{O}(n^\omega)$ using fast matrix multiplication [YL93; Ban+20].

Now, for sampling from a (size-unconstrained) DPP, Kulesza and Taskar [KT12] presented a somewhat slow $O(n^4)$-time algorithm, which was subsequently refined to $O(n^3)$, see, e.g., [Pou20].

[3] See p.18 in `https://buildmedia.readthedocs.org/media/pdf/dppy/latest/dppy.pdf` for details on various algorithms for sampling from DPPs.

The same algorithm can be improved by switching linear algebraic operations it uses to those that employ fast matrix multiplication. The factorization-based algorithm presented by Poulson [Pou20] arranges the ground set of n elements as leaves of a balanced binary tree, where the final sample from the DPP is produced at the root of the tree. Each node of this binary tree with m leaves in its subtree is associated with an $m \times m$ kernel matrix. Roughly speaking, a node with m leaves first computes an $m/2 \times m/2$ submatrix for its left child (the marginal of its DPP on the first half of the elements), produces a sample from the left subtree, and then produces another $m/2 \times m/2$ submatrix for its right child (the conditional DPP, conditioned on choices made by the first child). These submatrices are produced simply by Schur complements and matrix multiplication, all of which take time $\widetilde{O}(m^\omega)$ using fast matrix multiplication. Summing over all levels of the binary tree results in an overall runtime of $\widetilde{O}(n^\omega)$. □

References

[AD20] Nima Anari and Michał Dereziński. "Isotropy and Log-Concave Polynomials: Accelerated Sampling and High-Precision Counting of Matroid Bases". In: *Proceedings of the 61st Annual Symposium on Foundations of Computer Science*. 2020.

[AL20] 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*. 2020, pp. 1198–1211.

[Ald90] David J Aldous. "The random walk construction of uniform spanning trees and uniform labelled trees". In: *SIAM Journal on Discrete Mathematics* 3.4 (1990), pp. 450–465.

[Ale+18] Vedat Levi Alev, Nima Anari, Lap Chi Lau, and Shayan Oveis Gharan. "Graph Clustering using Effective Resistance". In: *9th Innovations in Theoretical Computer Science Conference, ITCS*. Vol. 94. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018, 41:1–41:16. DOI: 10.4230/LIPIcs.ITCS.2018.41.

[Ali+21] Yeganeh Alimohammadi, Nima Anari, Kirankumar Shiragur, and Thuy-Duong Vuong. "Fractionally Log-Concave and Sector-Stable Polynomials: Counting Planar Matchings and More". In: *arXiv preprint arXiv:2102.02708* (2021).

[Ana+21a] Nima Anari, Michał Dereziński, Thuy-Duong Vuong, and Elizabeth Yang. *Domain Sparsification of Discrete Distributions using Entropic Independence*. 2021. arXiv: 2109.06442 [cs.DS].

[Ana+21b] 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". In: *arXiv (to appear)* (2021).

[Ana+21c] 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*. 2021, pp. 408–420.

[AOR16] 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*. PMLR. 2016, pp. 103–115.

[Ban+20] Jess Banks, Jorge Garza Vargas, Archit Kulkarni, and Nikhil Srivastava. "Overlaps, eigenvalue gaps, and pseudospectrum under real Ginibre and absolutely continuous perturbations". In: *arXiv preprint arXiv:2005.08930* (2020).

[BBL09] 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.

[Bro89] Andrei Z Broder. "Generating random spanning trees". In: *FOCS*. Vol. 89. Citeseer. 1989, pp. 442–447.

[BT06] Sergey G Bobkov and Prasad Tetali. "Modified logarithmic Sobolev inequalities in discrete settings". In: *Journal of Theoretical Probability* 19.2 (2006), pp. 289–336.

[CDV20] Daniele Calandriello, Michał Dereziński, and Michal Valko. "Sampling from a k-DPP without looking at all items". In: *arXiv preprint arXiv:2006.16947* (2020).

[CE22] Yuansi Chen and Ronen Eldan. "Localization schemes: A framework for proving mixing bounds for Markov chains". In: *arXiv preprint arXiv:2203.04163* (2022).

[CGM19] 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.

[Che21] Yuansi Chen. "An almost constant lower bound of the isoperimetric coefficient in the KLS conjecture". In: *Geometric and Functional Analysis* 31.1 (2021), pp. 34–61.

[CL06] Fan Chung and Linyuan Lu. "Concentration inequalities and martingale inequalities: a survey". In: *Internet Mathematics* 3.1 (2006), pp. 79–127.

[CMN96] Charles J Colbourn, Wendy J Myrvold, and Eugene Neufeld. "Two algorithms for unranking arborescences". In: *Journal of Algorithms* 20.2 (1996), pp. 268–281.

[DCV19] Michał Dereziński, Daniele Calandriello, and Michal Valko. "Exact sampling of determinantal point processes with sublinear time preprocessing". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 11542–11554.

[Der19] Michał Dereziński. "Fast determinantal point processes via distortion-free intermediate sampling". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Vol. 99. Proceedings of Machine Learning Research. Phoenix, USA, 2019, pp. 1029–1049.

[DM21] Michał Derezinski and Michael W Mahoney. "Determinantal point processes in randomized numerical linear algebra". In: *Notices of the American Mathematical Society* 68.1 (2021), pp. 34–45.

[DR10] Amit Deshpande and Luis Rademacher. "Efficient volume sampling for row/column subset selection". In: *2010 ieee 51st annual symposium on foundations of computer science*. IEEE. 2010, pp. 329–338.

[Dur+17a] David Durfee, Rasmus Kyng, John Peebles, Anup B Rao, and Sushant Sachdeva. "Sampling random spanning trees faster than matrix multiplication". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. 2017, pp. 730–742.

[Dur+17b] David Durfee, John Peebles, Richard Peng, and Anup B Rao. "Determinant-preserving sparsification of SDDM matrices with applications to counting and sampling spanning trees". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 926–937.

[DWH18] 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.

[DWH19] 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*. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 944–953.

[Gil+19] Jennifer Gillenwater, Alex Kulesza, Zelda Mariet, and Sergei Vassilvtiskii. "A tree-based method for fast repeated sampling of determinantal point processes". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2260–2268.

[GRV09] Navin Goyal, Luis Rademacher, and Santosh Vempala. "Expanders via random spanning trees". In: *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2009, pp. 576–585.

[GW17] Kyle Genova and David P Williamson. "An experimental evaluation of the best-of-many Christofides' algorithm for the traveling salesman problem". In: *Algorithmica* 78.4 (2017), pp. 1109–1130.

[HS19] Jonathan Hermon and Justin Salez. "Modified log-Sobolev inequalities for strong-Rayleigh measures". In: *arXiv preprint arXiv:1902.02775* (2019).

[Jer98] Mark Jerrum. "Mathematical foundations of the Markov chain Monte Carlo method". In: *Probabilistic methods for algorithmic discrete mathematics*. Springer, 1998, pp. 116–165.

[Kar+21] Anna R Karlin, Nathan Klein, Shayan Oveis Gharan, and Xinzhi Zhang. "An Improved Approximation Algorithm for the Minimum k-Edge Connected Multi-Subgraph Problem". In: *arXiv preprint arXiv:2101.05921* (2021).

[KKO21] Anna R Karlin, Nathan Klein, and Shayan Oveis Gharan. "A (slightly) improved approximation algorithm for metric TSP". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 2021, pp. 32–45.

[KL22] Bo'az Klartag and Joseph Lehec. "Bourgain's slicing problem and KLS isoperimetry up to polylog". In: *arXiv preprint arXiv:2203.15551* (2022).

[KM09] Jonathan A Kelner and Aleksander Madry. "Faster generation of random spanning trees". In: *2009 50th Annual IEEE Symposium on Foundations of Computer Science*. IEEE. 2009, pp. 13–21.

[KO18] 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.

[KS18] Rasmus Kyng and Zhao Song. "A matrix chernoff bound for strongly rayleigh distributions and spectral sparsifiers from a few random spanning trees". In: *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2018, pp. 373–384.

[KT12] Alex Kulesza and Ben Taskar. "Determinantal point processes for machine learning". In: *arXiv preprint arXiv:1207.6083* (2012).

[LP17] David A Levin and Yuval Peres. *Markov chains and mixing times*. Vol. 107. American Mathematical Soc., 2017.

[LV18] Yin Tat Lee and Santosh S Vempala. "The Kannan-Lovász-Simonovits Conjecture". In: *arXiv preprint arXiv:1807.03465* (2018).

[MST14] Aleksander Madry, Damian Straszak, and Jakub Tarnawski. "Fast generation of random spanning trees and the effective resistance metric". In: *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2014, pp. 2019–2036.

[OR18] Shayan Oveis Gharan and Alireza Rezaei. "A Polynomial Time MCMC Method for Sampling from Continuous DPPs". In: *arXiv e-prints* (2018), arXiv–1810.

[Pou20] Jack Poulson. "High-performance sampling of generic determinantal point processes". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 378.2166 (2020). ISSN: 1471-2962. DOI: 10.1098/rsta.2019.0059.

[Sch18] 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.

[SS11] Daniel A Spielman and Nikhil Srivastava. "Graph sparsification by effective resistances". In: *SIAM Journal on Computing* 40.6 (2011), pp. 1913–1926.

[ST04] Daniel A. Spielman and Shang-Hua Teng. "Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems". In: *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, STOC 2004*. 2004, pp. 81–90.

[Wil96] David Bruce Wilson. "Generating random spanning trees more quickly than the cover time". In: *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*. 1996, pp. 296–303.

[YL93] Shing-Tung Yau and Ya Yan Lu. "Reducing the Symmetric Matrix Eigenvalue Problem to Matrix Multiplications". In: *SIAM J. Sci. Comput.* 14.1 (Jan. 1993), pp. 121–136. ISSN: 1064-8275. DOI: 10.1137/0914008.