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

Nima Anari  
Stanford University  
anari@cs.stanford.edu

Yang P. Liu  
Stanford University  
yangliu@stanford.edu

Thuy-Duong Vuong  
Stanford University  
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 $O(|V|^3)$ time per sample after an initial $O(|E|)$ 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 $O(k^3)$ time after an initial $O(nk^{w-1})$ time preprocessing, where $w < 2.372864$ is the matrix multiplication exponent. The time to compute just the weight of the output set is simply $k^w$, 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 $O(nk^{w-1})$ to $O(nk^{w-1})$.

In our main technical result, we achieve the optimal limit on domain sparsification for strongly Rayleigh distributions. In domain sparsification, sampling from a distribution $\mu$ on $\binom{n}{k}$ is reduced to sampling from related distributions on $\binom{\tilde{c}}{t}$ for $t \ll n$. We show that for strongly Rayleigh distributions, the domain size can be reduced to nearly linear in the output size $t = O(k)$, improving the state of the art from $t = O(k^2)$ for general strongly Rayleigh distributions and the more specialized $t = O(k^{3.5})$ for spanning tree distributions. Our reduction involves sampling from $O(1)$ domain-sparsified distributions, all of which can be produced efficiently assuming approximate overestimates for marginals of $\mu$ 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.

Index Terms—domain sparsification, strongly Rayleigh, sublinear sampling, entropic independence, Markov chains

I. INTRODUCTION

In this work, we study how far we can push the runtime of sampling algorithms for the widely used class of strongly Rayleigh distributions [1], which are distributions supported on size $k$ subsets of a ground set $[n] = \{1, \ldots, n\}$, denoted from here on as $(\binom{n}{k})$, which satisfy strong forms of negative correlation (see Section II-B 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 [2], [3], [4], [5], [6], [7] discovered that under certain regularity assumptions on the distribution $\mu$, one can sample from $\mu$ 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 [5], 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 [8], [9], [10]. The Kannan-Lovász-Simonovits (KLS) conjecture was formulated to answer how fast one can sample from isotropic log-concave densities [8].

Motivated by the analogy with continuous distributions, Anari and Dereziński [5] defined a notion of isotropy for discrete distributions $\mu$ on $\binom{[n]}{k}$: $\mu$ 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 II-F for details). The main question then becomes how fast can we sample from (approximately) isotropic distributions $\mu$ on $\binom{[n]}{k}$?

Prior works [2], [3], [5], [6] showed that the answer to this is $\leq \log(k, \log n)$ for strongly Rayleigh distributions, assuming oracle access to $\mu$. 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 $O\left(\frac{n}{k}\right)$.
for sampling from related distributions on \(\binom{n}{k}\) for \(t = \tilde{O}(k)\). In other words, isotropy allows us to pretend that \(n\) is only as large as \(\tilde{O}(k)\).

**Theorem 1** (Informal, see Theorem 2 for a formal statement). Suppose that the time to sample from a class of strongly Rayleigh distributions on \(\binom{n}{k}\) is \(T(n, k)\). Then we can sample from (approximately) isotropic distributions in this family in time \(\tilde{O}(1) \cdot T(\tilde{O}(k), k)\).

Below we expand on two classes of distributions that constitute the main applications of our result.

a) 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) \([11]\) they are a key component of the Best-of-Many Christofides algorithm used in recent TSP improvements \([12]\). Random spanning trees have found applications in the construction of graph sparsifiers \([13, 14]\). As another example application, the recent breakthrough of Karlin, Klein, Gharan, and Zhang \([15]\) 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, \(\text{e.g.},\) through the matrix-tree theorem. This has all motivated a long sequence of works on obtaining fast algorithms for sampling from this ubiquitous distribution \([16, 17, 18, 19, 20, 21, 22, 23, 24, 25]\). 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 \([24, 25]\) 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 \(\tilde{O}(|V|)\), after an \(\tilde{O}(|E|)\) preprocessing step.

b) 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 \(\mu\) 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, \(\text{e.g.},\) \([26, 27, 28]\)). 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 \([29, 30]\). Recent works have shown how to go even further, and obtain after a preprocessing step, \(\text{poly}(k, \log n)\) sampling times \([31, 32, 3, 4, 2, 5, 6]\); 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 \(\tilde{O}(kn^\omega)\), where \(\omega\) is the matrix multiplication exponent. Note that \(k^\omega\) 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 \(\tilde{O}(nk^{\omega-1})\). 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 \(\tilde{O}(nk^2)\) \([29, 30]\) or were based on linear algebraic primitives \([27, 33]\), which implemented with fast matrix multiplication, would take time \(\tilde{O}(n^\omega)\), see ??.

We remark that our improvement from \(\tilde{O}(\min\{nk^2, n^\omega\})\) to \(\tilde{O}(nk^{\omega-1})\) is only made possible by a fast preprocessing step which crucially is implemented by bootstrapping with the primitive of fast sampling from isotropic distributions.

A. Sampling algorithm

To obtain our optimal sublinear samplers, we use the framework established in prior works \([22, 3, 2, 5, 6, 7]\) of sparsifying the domain \([n]\) for isotropic distributions, \(i.e.,\) distributions with roughly balanced marginals \([5]\). The preprocessing step for our algorithm consists only of putting the distribution into (approximately) isotropic position (see Section II-F) by finding approximate overestimates for the marginals \(P_{S \sim \mu}[i \in S]\) and transforming \(\mu\) 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 Rayleigh distributions. In domain sparsification, we reduce the task of sampling from our distribution on \(\binom{[n]}{k}\) to distributions on \(\binom{[\tilde{n}]}{k}\); we show this can be done with \(t = \tilde{O}(k)\). Prior works on this problem either required \(t \approx k^2\) \([3, 2, 5, 6, 7]\) or for the specific case of spanning tree distributions required \(t \approx k^{1.5}\) \([22]\).

More formally, for an approximately isotropic \(\mu\), 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 \(\tilde{O}(1)\) steps. We output \(S'_0 \subset \binom{[\tilde{n}]}{k}\) as our approximate sample from \(\mu\). 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 \(\mu\)). 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.

B. Our results

To formally state our results on sampling from strongly Rayleigh distributions, it is useful to define \(T_0(t, k)\) for a distribution \(\mu \in \mathbb{R}^\binom{[n]}{k}\) as the time it takes to produce a sample from \(\mu\) conditional on all elements of the sample being a subset of a fixed set \(T\) of size \(|T| = t\). We use \(\tilde{O}(\cdot)\) to suppress \(\text{poly} \log n\) factors. Notice below that the
Algorithm 1: Down-up walk on the complement distribution

for $i = 0, 1, 2, \ldots$ do
  From all $i$-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 $\mathbb{P}[S_{i+1}] \propto \mu(S_i)$.

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 2 (Sampling using 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 $\mu$ in time bounded by $\tilde{O}(1)$ calls to $T_{\mu}(O(K), k)$.

We prove Theorem 2 using a local-to-global argument, which requires us to also show that random conditionals of $\mu$ are isotropic with high probability. This is similar in spirit to the recent analyses of the KLS conjecture using stochastic localization [9], [10] which show that an isotropic continuous distribution remains approximately isotropic over a stochastic evolution.

Theorem 3 (Informal, see Theorem 26 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 $\mu_T(S) := \mu(S)/\sum_{S \subseteq \binom{T}{k}} \mu(S)$.

Then with high probability over $T \in \binom{[n]}{k}$ for $t = \tilde{O}(k)$, $\mu_T$ is approximately isotropic.

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

Theorem 4 (Informal, see Theorem 33 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 $O(n/k)$ calls to a sampler for isotropic distributions on sets of size $O(k)$.

We remark that the preprocessing time of $\tilde{O}(|E|)$ 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-Lindenstrauss lemma [34], [35]. However, we give a self-contained method by bootstrapping the sampling algorithm (Theorem 4) 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 $O(|E|)$ time [25], or a k-DPP on $n$ elements in $O(n^2)$ time (see ??) 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 5 (Sampling spanning trees). For a graph $G = (V, E)$, possibly weighted with weights $\lambda \in \mathbb{R}_E$, 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 $\tilde{O}(|E| + s|V|)$.

Corollary 6 (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 $O(nk^{d-1} + sk^d)$.

Finally, we remark that our methods also show analogous mixing times of $\tilde{O}(k)$ steps for the Markov chain that uses small up-down steps, i.e., calls to $T_{\mu}(k+1,k)$, 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 [36], [25]. We formally state these results in Theorem 34 in Section VI.

C. 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., [37]). We establish that isotropy significantly improves the “expansion” of the high-dimensional-expander. We use the notion of expansion called entropic independence [38], 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 II for definition) on the complement/dual distribution associated with $\mu$; note that each step of this walk involves a sparsified sampling problem, where we only have to sample from a distribution on $\binom{T}{k}$. 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 naive matrix-multiplication-based sampler for DPPs.

Below we describe the main techniques we use.

a) 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 II for definition) [38]. 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 $\geq O((n-k)/t)$ for Algorithm 1, which has an unacceptable dependence on $n$. This is not surprising, as these black-box results do not incorporate isotropy of $\mu$. In this work, we show that whenever $\mu$ 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 naively expected (Theorem 23).

b) 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 [39]) 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 \( \mu \). Unfortunately, in the worst case, this can significantly imbalance the marginals of \( \mu \). 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 26) and combine this with an average local-to-global result adapted from [40] (Theorem 19) 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 [9], [41], [10].

c) 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 \( \text{poly}(k) \) factors from the runtime of marginal estimation (Theorem 33), and are essential for our faster \( O(nk^{-1}) \) time algorithm for sampling from a \( k \)-DPP. This is achieved by a recursive procedure that uses marginals of the restriction of \( \mu \) to roughly half the domain \( [n] \) as overestimates for the marginals of \( \mu \). In the end, marginal overestimation is roughly reduced to \( \tilde{O}(n/k) \) subtasks of marginal overestimation for distributions over domains of size \( O(k) \).

d) Barriers faced by prior approaches. In order to derive the tight sparsification of \( t = \tilde{O}(k) \) in Algorithm 1, we had to rethik the entire analysis technique. To emphasize the importance of tight bounds on \( t \), we note that prior results on general strongly Rayleigh measures [2], [6], [3], [5] 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 [22] is roughly speaking that for the regime \( t = \tilde{O}(k) \), subsets \( T_i \) are not good sparsifiers for partition functions. To appreciate this better, consider a simple distribution \( \mu \) on \( \binom{[n]}{k} \) defined as follows: first we partition \( [n] \) into disjoint sets \( U_1, \ldots, U_c \) 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 \( f \) 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 [22]) would show that \( c \approx \sqrt{k} \) 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 \( t = \tilde{O}(k) \), 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.

II. PRELIMINARIES

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

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

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

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

Definition 7 (Restricted distribution). For a distribution \( \mu \) defined over subsets of a ground set \( [n] \) and \( S \subseteq [n] \), define \( \mu_S \) to be the distribution of \( F \sim \mu \) restricted to the set \( S \), i.e., conditioned on the event \( F \subseteq S \).

Definition 8 (Conditional distribution). For a distribution \( \mu \) defined over subsets of a ground set \( [n] \) and \( T \subseteq [n] \), define \( \mu^T \) to be the distribution of \( F \sim \mu \) conditioned on the event \( F \supseteq T \).

A. Markov chains and functional inequalities

Let \( \mu \) and \( \nu \) be probability measures on a finite set \( \Omega \). The Kullback-Liebler divergence (or relative entropy) between \( \nu \) and \( \mu \) is given by

\[ D_{\text{KL}}(\nu \| \mu) = \sum_{x \in \Omega} \nu(x) \log \left( \frac{\nu(x)}{\mu(x)} \right), \]

with the convention that this is \( \infty \) if \( \nu \) is not absolutely continuous with respect to \( \mu \). The total variation distance between \( \mu \) and \( \nu \) is given by

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

A Markov chain on \( \Omega \) is specified by a row-stochastic non-negative transition matrix \( P \in \mathbb{R}^{\Omega \times \Omega} \). We refer the reader to [42] for a detailed introduction to the analysis of Markov chains. As is common, we will view probability distributions on \( \Omega \) as row vectors. Recall that a transition matrix \( P \) is said to be reversible with respect to a distribution \( \mu \) if for all \( x, y \in \Omega \), \( \mu(x)P(x, y) = \mu(y)P(y, x) \). In this case, it follows immediately that \( \mu \) is a stationary distribution for \( P \), i.e., \( \mu P = \mu \). If \( P \) is further assumed to be ergodic, then \( \mu \) is its unique stationary distribution, and for any probability distribution \( \nu \) on \( \Omega \), \( d_{\text{TV}}(\nu P^t, \mu) \rightarrow 0 \) as \( t \rightarrow \infty \). The goal of this paper is to investigate the rate of this convergence.

Definition 9. Let \( \mu \) be an ergodic Markov chain on a finite state space \( \Omega \) and let \( \mu \) denote its (unique) stationary distribution.
For any probability distribution \( \nu \) on \( \Omega \) and \( \epsilon \in (0,1) \), we define \( t_{\text{mix}}(P, \nu, \epsilon) \) to be
\[
\min \left\{ t \geq 0 : d_{\text{TV}}(\nu P^t, \mu) \leq \epsilon \right\},
\]
and let \( t_{\text{mix}}(P, \epsilon) \) denote
\[
\max \left\{ \min \left\{ t \geq 0 : d_{\text{TV}}(\mathbb{1}_x P^t, \mu) \leq \epsilon \right\} : x \in \Omega \right\},
\]
where \( \mathbb{1}_x \) is the point mass supported at \( x \).

We will drop \( P \) and \( \nu \) if they are clear from context. Moreover, if we do not specify \( \epsilon \), then it is set to 1/4. This is because the growth of \( t_{\text{mix}}(P, \epsilon) \) is at most logarithmic in 1/\( \epsilon \) (cf. [42]).

**Lemma 10.** Let \( \mu \) be a probability measure on the finite set \( \Omega \). Let \( P \) denote the transition matrix of an ergodic, reversible Markov chain on \( \Omega \) with stationary distribution \( \mu \). Suppose there exists some \( \alpha \in (0,1) \) such that for all probability measures \( \nu \) on \( \Omega \), we have
\[
\mathcal{D}_{\text{KL}}(\nu P \| \mu P) \leq (1 - \alpha) \mathcal{D}_{\text{KL}}(\nu \| \mu).
\]
Then \( t_{\text{mix}}(P, \epsilon) \leq \left[ \frac{1}{\alpha} \left( \log \left( \frac{\mathbb{1} \cdot \log \left( \min \{ \mu(x) : x \in \Omega \} \right)}{\log \left( \frac{1}{2^4} \right)} \right) \right) \right] .
\]

This is the standard argument for bounding mixing times via modified log-Sobolev inequalities and can be found in, e.g., [43].

**B. Strongly Rayleigh distributions**

For density function \( \mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0} \), the generating polynomial of \( \mu \) 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 11.** Consider the open half-plane \( H = \{ z : \text{Im}(z) > 0 \} \subset \mathbb{C} \). We say a polynomial \( g(z_1, \ldots, z_n) \in \mathbb{R}[z_1, \ldots, z_n] \) is real-stable if \( g \) does not have roots in \( H^n \).

A distribution \( \mu : 2^{[n]} \to \mathbb{R}_{\geq 0} \) is strongly Rayleigh iff its generating polynomial is real stable (see [1]). If \( \mu \) is strongly Rayleigh, then its conditional and restricted distributions (see Definitions 7 and 8) are also strongly Rayleigh. The key fact we use about strongly Rayleigh distributions is that they are negatively correlated (see, e.g., [1]), i.e., the marginals (of non-restricted elements) increase under restrictions (Definition 7):
\[
P_{S \sim \mu}[i \in S] \leq P_{S \sim \mu_T}[i \in S] \text{ for } i \notin T.
\]

**C. Down-up and up-down walks**

**Definition 12 (Down operator).** For \( \ell \leq k \) define the row-stochastic matrix \( D_{k \to \ell} \in \mathbb{R}^{\binom{[n]}{\ell} \times \binom{[n]}{k}} \) by
\[
D_{k \to \ell}(S, T) = \begin{cases}
0 & \text{if } T \not\subseteq S, \\
\frac{1}{\binom{|T|}{\ell}} & \text{otherwise}.
\end{cases}
\]

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

**Definition 13 (Up operator).** For \( \ell \leq k \) define the row-stochastic matrix \( U_{\ell \to k} \in \mathbb{R}^{\binom{[n]}{\ell} \times \binom{[n]}{k}} \) by
\[
U_{\ell \to k}(T, S) = \begin{cases}
0 & \text{if } T \not\subseteq S, \\
\frac{\mu(S)}{\sum_{T \supseteq S} \mu(T)} & \text{otherwise}.
\end{cases}
\]

As in [5], [7], we consider the following Markov chain \( M^\ell_\mu \) defined for any positive integer \( t \), with the state space \( \text{supp}(\mu) \). Starting from \( S_0 \in \text{supp}(\mu) \), one step of the chain is:
1. Sample \( T \in \binom{[n]}{t-k} \) uniformly at random.
2. Downsample \( S_1 \sim \mu_{S \setminus T}, \) where \( \mu_{S \setminus T} \) is \( \mu \) restricted to \( S_0 \cup T \) and update \( S_0 \) to \( S_1 \).

**Proposition 14.** The complement of \( S_1 \) is distributed according to \( \mu_0 D_{(n-k) \to (n-t)} U_{(n-t) \to (n-k)} \) where \( \mu_0 \) is the distribution of the set \( S_0 \).

**Proposition 15.** For any distribution \( \mu : \binom{[n]}{k} \to \mathbb{R}_{\geq 0} \) that is strongly Rayleigh, the chain \( M^\ell_\mu \) for \( \ell \geq k + 1 \) is irreducible, aperiodic and has stationary distribution \( \mu \).

**D. Entropic independence**

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

**Definition 16 (Entropic independence [38]).** A probability distribution \( \mu \) on \( \binom{[n]}{k} \) is said to be \( (1/\alpha) \)-entropically independent, if for all probability distributions \( \nu \) on \( \binom{[n]}{k} \),
\[
\mathcal{D}_{\text{KL}}(\nu_{D_{k \to 1}} \| \mu_{D_{k \to 1}}) \leq \frac{1}{\alpha k} \mathcal{D}_{\text{KL}}(\nu \| \mu).
\]

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 17 ([38, Theorem 4]).** Any strongly Rayleigh \( \mu \) is \( 1 \)-entropically independent. The conditional and restricted distributions of \( \mu \) are also strongly Rayleigh, and thus \( 1 \)-entropically independent.

**E. Average-case local-to-global method**

First, we define the notion of the link of the distribution \( \mu \) w.r.t. a set \( T \) (see, e.g., [37]). This is almost the same as the conditioned distribution \( \mu^T \), see Definition 8, except we remove the set \( T \).

**Definition 18.** 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 \( \mu^{-T} : \binom{[n]}{k-|T|} \to \mathbb{R}_{\geq 0} \) which describes the law of the set \( S - T \) where \( S \) is sampled from \( \mu \) 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 $\ell$. While this is by now a well-understood phenomenon, sometimes called the local-to-global method (see, e.g., [39]), we use an average case variant, adapted from [40], which only requires entropic independence for a "typical" link as opposed to a worst case link.

**Theorem 19.** Suppose that for every set $T$ of size $\leq k - 2$, $\mu^{-T}$ contracts KL-divergence in terms of factor parameterized by $p(T)$:

$$D_{KL}(\nu D_{k \to |T|-1} \| \mu^{-T} D_{k \to |T|-1}) \leq (1 - \rho(T)) D_{KL}(\nu \| \mu^{-T}).$$

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

$$\gamma_T := E\left[\left(\rho(\emptyset)\rho(\{e_1\})\cdots \rho(\{e_1,\ldots, e_{|T|-1}\})\right)^{-1}\right].$$

where $e_1, \ldots, e_{|T|}$ is a random permutation of $T$. Then the operator $D_{k \to \ell}$ has KL-divergence contraction

$$D_{KL}(\nu D_{k \to \ell} \| \mu D_{k \to \ell}) \leq (1 - \kappa) D_{KL}(\nu \| \mu),$$

with

$$\kappa := \min \left\{ \gamma_T \big| T \subset [n] \right\}.$$  

The proof is similar to [40, Theorem 46], [38, Theorem 5], and is deferred to the full version.

**Remark 20.** Similar to [40], if the KL-divergence is replaced by any other type of $f$-divergence, a common choice being $\chi^2$-divergence which roughly relates to the notion of spectral independence, Theorem 19 still remains valid.

**F. Isotropic transformation**

Anari and Dereziński [5] introduced the following subdivision process that takes marginal overestimates of an arbitrary distribution $\mu \in \mathbb{R}^{(n)^\alpha}$, and transforms sampling from $\mu$ to sampling from a distribution with nearly uniform marginals. In the following, we call $\mu'$ the isotropic transformation of $\mu$.

**Definition 21.** Let $\mu : \binom{\ell}{n} \to \mathbb{R}_{\geq 0}$ 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 $p_i \geq \Pr_{S \sim \mu}[i \in S]$ for all $i$. Let $t_i := \lceil \frac{n}{p_i} \rceil$. We will create a new distribution out of $\mu$: 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: $U = \bigcup_{i=1}^{n} \{i^{(1)}_i, \ldots, i^{(t_i)}_i\}$. Define the following distribution $\mu' : \binom{\ell}{n} \to \mathbb{R}_{\geq 0}$ from $\mu$:

$$\mu' \left( \{i^{(1)}_1, \ldots, i^{(\ell)}_k\} \right) := \frac{\mu(\{i_1, \ldots, i_k\})}{t_1 \cdots t_k}.$$  

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

As show in [7, Proposition 24], performing the isotropic transformation in Definition 21 at most only doubles the size of the universe $U$, but makes all marginals bounded by $O(K/n)$ now.

**Proposition 22.** Let $\mu : \binom{\ell}{n} \to \mathbb{R}_{\geq 0}$, and let $\mu' : \binom{\ell}{n} \to \mathbb{R}_{\geq 0}$ be the subdivided distribution from Definition 21. The following hold for $\mu'$:

1) Near-isotropy: For all $i \in [\ell]$, $\mu'(\{i\}) \geq \mu(\{i\})$.

2) Linear ground set size: $|U| \leq 2n$.

3) If $\mu$ is strongly Rayleigh then so is $\mu'$.

**III. Entropy contraction**

Our goal is to prove an entropy contraction inequality for the $(n - k) \to 1$ down operator. For strongly Rayleigh distributions $\bar{\mu} \in \mathbb{R}_+^{(n)}$, which are 1-entropically independent (Definition 16), the entropy contracts by $1/(n-k)$. Surprisingly, if $\mu$ also has nearly uniform marginals, the entropy contracts even more, by an extra $\sim \log(n/k)$ factor.

**Theorem 23 (Level one entropy contraction).** Let $\mu \in \mathbb{R}_+^{(n)}$ be a 1-entropically independent distribution with $p(\mu)^\alpha := \max_{i \in [n]} \Pr_{\mu}[i \in F] \leq \frac{1}{100}$. Then for any distribution $\nu \subseteq \mathbb{R}_+^{(n)}$,

$$D_{KL}(\nu D_{n-k} \| \mu D_{n-k}) \leq \frac{1}{(n-k) \log((\nu p(\mu)^\alpha)\max)} D_{KL}(\nu \| \mu).$$

We show this theorem by directly comparing the relative entropies of $p = \nu D_{n-k}$ and $q = \nu D_{(n-k)}$ with respect to $\mu D_{n-k}$ and $\bar{\mu} D_{(n-k)}$ respectively. We first show a single-variable instance of this, which we sum over to get the overall comparison in Lemma 25.

**Lemma 24.** 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 $C_{\alpha,\mu} \geq \frac{\alpha}{(1-\alpha)\log((\alpha p/\mu)\max)}$ and $K_{\alpha,\mu} := \frac{C_{\alpha,\mu}}{(1-\alpha)^{-1}}$.

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

$$f'(q) = -C_{\alpha,\mu} \cdot \frac{1 - \alpha}{\alpha} (\log(\alpha p/\mu) + 1) - \left(\log \left(\frac{(1 - \alpha)q}{1 - \mu}\right) + 1\right) + K_{\alpha,\mu} (1 - \alpha).$$
By our careful choice of $K_{\alpha,\mu}$, we have $f'(q) = 0$ for $q = \frac{1-\mu}{\alpha}$.
Additionally, we can calculate
\[
 f''(q) = p^{-1}q^{-1}\left(C_{\alpha,\mu}\left(\frac{1-\alpha}{\alpha}\right)^2q - \frac{1-(1-\alpha)\frac{q}{\alpha}}{\alpha}\right).
\]
(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
\[
 f''(q) = p^{-1}q^{-1}\left(C_{\alpha,\mu}\left(\frac{1-\alpha}{\alpha}\right)^2q - \frac{1-(1-\alpha)\frac{q}{\alpha}}{\alpha}\right)
\geq p^{-1}q^{-1}\left(\frac{1}{\log(1/(\alpha\mu))} - \frac{1}{\alpha} \frac{1-\mu}{\alpha} \frac{q}{\alpha}\right)
= p^{-1}q^{-1}\left(\frac{1}{\alpha\log(1/(\alpha\mu))} - \frac{1}{\alpha} \frac{1}{1+\log(1/(\alpha\mu))}\right)
= p^{-1}q^{-1}\left(\frac{1}{\alpha\log(1/(\alpha\mu))} + \frac{1}{\alpha} \log \mu \geq 0 \right.
\]
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 25.** Let $p, q$ be distributions on $[n]$ satisfying $\alpha p + (1-\alpha)q = 1/n$ for $\alpha = k/n$ and $\alpha < 1/100$. Then for any $\mu \in [1]$ with $\|\mu\|_1 = k$ and $\mu^{\text{max}} := \max_{i \in [n]} \mu_i < 1/100$,
\[
 D_{\text{KL}}\left(q \left\| \frac{\bar{\mu}}{n-k} \right\| \right) \leq C_{\alpha,\mu,\mu^{\text{max}}} D_{\text{KL}}\left(p \left\| \frac{\bar{\mu}}{k} \right\| \right)
\]
(2)
with $C_{\alpha,\mu,\mu^{\text{max}}} := \frac{\alpha}{(1-\alpha)\log(1/\alpha\mu^{\text{max}})}$ as in Lemma 24.

**Proof.** By Lemma 24 (for the choice $p = np_i$ and $q = nq_i$) and the fact that $C_{\alpha,\mu}$ is monotonically increasing in $\mu$, we deduce that
\[
 C_{\alpha,\mu,\mu^{\text{max}}} \sum_{i \in [n]} p_i \log(p_i/k/\mu_i) - q_i \log(q_i/(n-k)/(1-\mu_i)) \geq 1/n K_{\alpha,\mu,\mu^{\text{max}}} (\alpha p_i - \mu_i).
\]

Summing over all $i$ gives us
\[
 C_{\alpha,\mu,\mu^{\text{max}}} \sum_{i \in [n]} p_i \log(p_i/k/\mu_i) - q_i \log(q_i/(n-k)/(1-\mu_i)) \geq \frac{1}{n} K_{\alpha,\mu,\mu^{\text{max}}} \sum_{i \in [n]} (\alpha p_i - \mu_i).
\]
The r.h.s. equals $0$, so we get the desired inequality. \hspace{1cm} \Box

Now, Theorem 23 is an easy corollary of Lemma 25.

**Proof of Theorem 23.** Let $\nu, \mu$ be the complement of $\bar{\nu}, \bar{\mu}$ resp. Let $p = \nu D_{k-1}, q = \bar{\nu} D_{(n-k)-1}$, and $\bar{\mu}_i := p_{F^{-}\mu}[i \in F]$ for $i \in [n]$ in the setting of Lemma 25. Note that $\alpha p + (1-\alpha)q = 1/n$ for $\alpha = k/n$, and $\mu^{\text{max}} = p(\mu)^{\text{max}} \leq 1/100$.

Because $\mu$ is 1-entropically independent (Lemma 17),
\[
 D_{\text{KL}}\left(\nu \left\| \bar{\mu} \right\| \right) \leq \frac{1}{k} D_{\text{KL}}(\nu \left\| \mu \right\|) \leq \frac{1}{k} D_{\text{KL}}(\bar{\nu} \left\| \bar{\mu} \right\|).
\]
Combining this with Lemma 25 for $\alpha = k/n$ gives us
\[
 D_{\text{KL}}\left(q \left\| \frac{\bar{\mu}}{n-k} \right\| \right) \leq C_{\alpha,\mu,\mu^{\text{max}}} D_{\text{KL}}\left(p \left\| \frac{\bar{\mu}}{k} \right\| \right)
\leq \frac{\alpha}{(1-\alpha)\log((\alpha p(\mu))^{\text{max}}-1)} \leq \frac{1}{n-k} D_{\text{KL}}(\bar{\nu} \left\| \bar{\mu} \right\|).
\]

We can almost directly combine Theorem 23 and the average-case local-to-global principle Theorem 19 to deduce an entropy contraction for $D_{(n-k) \rightarrow (n-k+1)}$ and $D_{(n-k) \rightarrow (n-k-1)}$. The one remaining issue is that the local-to-global theorem requires that the marginals of conditionals of $\mu$ also have almost uniform marginals. This is the main result of Section IV, which we state here.

**Theorem 26.** Let $\mu, \bar{\mu} \in [n]$ be a strongly Rayleigh distribution, and let $T \subseteq [n]$ with $|T| = k$. For a sufficiently large constant $C$ and any $s \geq C(n(\mu)^{\text{max}} + k)\log n$, we have
\[
 P_{\bar{S}} \left[ \left( \mu_{[n]\setminus S} \right)^{\text{max}} \geq \left( 2(\mu(\mu)^{\text{max}})^{-1} \right) \right] \leq n^{-10}.
\]

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

**Theorem 27.** Let $\mu \in [n]$ be a strongly Rayleigh distribution with $\mu^{\text{max}} \leq 1/500$. Let $s := C(n(\mu)^{\text{max}} + k)\log n$ for $C$ be as in Theorem 26 and $k' = \Theta(n(\mu)^{\text{max}})$.

Then for any distribution $\bar{\nu} \subseteq [n]$, $k \geq k' + 2$
\[
 D_{\text{KL}}(\bar{\nu} D_{(n-k) \rightarrow (n-k+1)} \left\| \mu D_{(n-k+1) \rightarrow (n-k)} \right\|) \leq \frac{1}{(k-1)} D_{\text{KL}}(\bar{\nu} \left\| \bar{\mu} \right\|)
\]
with $k := \frac{k-k-1}{2s\log n}$. In particular,
\[
 D_{\text{KL}}(\bar{\nu} D_{(n-k) \rightarrow (n-k+1)} \left\| \mu D_{(n-k+1) \rightarrow (n-k)} \right\|) \leq \frac{1}{(k-1)} D_{\text{KL}}(\bar{\nu} \left\| \bar{\mu} \right\|)
\]
and
\[
 D_{\text{KL}}(\bar{\nu} D_{(n-k) \rightarrow (n-k-1)} \left\| \mu D_{(n-k-1) \rightarrow (n-k)} \right\|) \leq \frac{1}{(k-\kappa)^2} D_{\text{KL}}(\bar{\nu} \left\| \bar{\mu} \right\|)
\]
with $\kappa^{-1} = O(\log^2 n)$ and $\kappa_2^{-1} = O(n(\mu)^{\text{max}} \log^2 n)$.

**Proof.** Fix $k \geq k' + 2$ to be chosen later, and a set $\bar{T} \subseteq [n]$ of size $n-k$. Let $s := C(n(\mu)^{\text{max}} + k)\log n$ for $C$ be as in Theorem 26. In the context of Theorem 19, we want to bound $\gamma_{\bar{T}}$ with respect to $\bar{\mu}$. Theorem 23 implies that the link of $\emptyset$ is
1/\mu_0\text{-entropically independent with } u_0 = \log((\exp(\mu)_{\max})^{-1})\). Consider a random permutation \( e_1, \ldots, e_{n-k} \) of \( T \). Note that each set \( S_i := \{ e_1, \ldots, e_i \} \) is a randomly sampled size-\( i \) subsets of \( T \). By using Theorem 26 and taking a union bound over \( i \in [n-s] \), we have that except with probability \( n^{-10} \times n^{-9} = n^{-19} \), we have

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

for \( C \) sufficiently large. Suppose this event holds. Note that the complement of \( \mu_{[n]\setminus S_i} \) is exactly \( \mu^{S_i} \). Thus, Theorem 23 implies that this link is 1/\( u_i \)-entropically independent where \( u_i := \log(\frac{n-k-1}{s-k-1}) \).

As a result

\[
\rho(\theta)\rho(S_{n-s}) \geq \prod_{i=0}^{n-s} \left( 1 - \frac{1}{(n-k-i)u_i} \right)
\]

\[
\geq \exp \left( -\sum_{i=0}^{n-s} \left( \frac{1}{(n-k-i)u_i} + \frac{1}{(n-k-i)^2 u_i^2} \right) \right)
\]

\[
\geq \exp \left( -\sum_{i=0}^{n-s} \left( \frac{1}{(n-i)u_i} + \frac{k+1}{(n-k-i)^2} \right) \right)
\]

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

\[
\frac{1}{(n-k-i)u_i} \geq \frac{1}{(n-i)u_i} + \frac{k}{(n-k-i)^2}
\]

Next, let \( h := 2\exp(\mu)_{\max}n \).

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

where in the final equality we use \( (\log \log t)' = \frac{1}{t \log^2 t} \).

Similarly

\[
\sum_{i=0}^{n-s} \frac{k+1}{(n-k-i)^2} \leq \frac{k+1}{(n-k)(n-k)} \int_{s-k}^{n-k} \frac{1}{t^2} \, dt = \frac{(k+1)(n-s)}{(s-k)(n-k)}
\]

which is \( \leq O(1) \), where we use the fact that \( s \geq 2Ck \log n \gg k \). Thus, we have that the l.h.s. in Eq. (3) \( \geq \Omega \left( \frac{1}{(n-k)(n-s)} \right) \).

Moreover, for \( i \in \{ n-s+1, \ldots, n-k \} \), by 1-entropic independence of \( \mu \) and its links (Lemma 17) we get the trivial bound \( \rho(S_i) \geq 1 - \frac{1}{n-k-i+1} \). Thus

\[
\prod_{i=n-s+1}^{n-k} \rho(S_i) \geq \prod_{i=n-s+1}^{n-k} \left( 1 - \frac{1}{n-k-i} \right) = \frac{k-k-1}{s-k-1} \geq \frac{k-k-1}{s}
\]

Thus, with probability at least \( 1 - n^{-9} \), \( \prod_{i=0}^{n-k} \rho(S_i) \geq \frac{k-k-1}{s} \).

Otherwise we have a trivial lower bound of 1/n on the product of \( \rho \) due to 1-entropic independence of \( \mu \) and its links. Thus

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

We are done with the general entropy contraction statement. Next, we prove entropy contraction for specific values of \( k \).

Plugging in \( k = k' \) and noting that for our choice of \( k' \) and \( s \),

\[
\frac{k-k-1}{2s \log n} \geq \frac{1}{2} \geq \frac{1}{2} \geq \Omega \left( \frac{s}{\log n} \right)
\]

implies the first result. Similarly, setting \( k = k + 2 \) gives the second result for our choice of \( s \).

\begin{proof}

\end{proof}

**IV. CONCENTRATION OF MARGINALS**

The goal of this section is to show concentration of marginal upper bounds (Theorem 26) for random conditionals of strongly Rayleigh distributions. In Section III, this is applied in the context of an average-case local-to-global principle (Theorem 19) to deduce our main entropy contraction result (Theorem 27). The proof uses the following simple observation about covariances in a set-valued distribution \( \mu \) whose support contains only sets of identical size.

**Lemma 28** (Covariances of homogeneous distributions). For any distribution \( \mu \) defined over identically-sized subsets of a ground set of elements \( [n] \) and any element \( i \in [n] \) we have

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

We show Theorem 26 by analyzing the marginal of each coordinate \( i \in [n] \) conditioned on it remaining in \( \mu_{[n]\setminus S} \) via a stochastic process. Formally, fix \( T \) as in Theorem 26 and a coordinate \( i \in [n] \) (possibly in \( T \)).

**Definition 29** (Stochastic process). For fixed \( T \subseteq [n] \) with \( |T| = k \), \( i \in [n] \), and \( s \leq n - |T| - 1 \), let \( \sigma \) be a random permutation of \( [n]\setminus(T \cup \{i\}) \). For \( 0 \leq t \leq n-s \) define \( S_t = \{ \sigma(1), \sigma(2), \ldots, \sigma(t)) \}. \) Define \( S_{n-s} = S_t \) and \( \mu(t) := \mu_{[n]\setminus S_t} \).

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

**Lemma 30** (Expected drift). With the setup in Definition 29 and \( 0 \leq t < n-s \), we have

\[
\mathbb{E}_{S_{t+1}} \left[ p(\mu^{(t+1)})_i \bigg| S_{t} \right] - p(\mu(t))_i \leq \frac{(1 - p(\mu(t))_{\max})^{-1}}{n-k-1-t} \cdot p(\mu(t))_i.
\]
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

\[
p(\mu^{(t+1)})_i - p(\mu^{(t)})_i = \frac{P_{F-\mu^{(t)}}[i \in F, j \notin F]}{1 - p(\mu^{(t)})_j} - p(\mu^{(t)})_j
\]

\[
= \frac{p(\mu^{(t)})_j p(\mu^{(t)})_i - P_{F-\mu^{(t)}}[i, j \in F]}{1 - p(\mu^{(t)})_j}.
\]

(4)

Hence by Eq. (4),

\[
E_{S_{t+1}} \left[ p(\mu^{(t+1)})_i \mid S_t \right] - p(\mu^{(t)})_i
= \frac{1}{n - |T \cup \{ i \} \cup S_t|} \times
\sum_{j \in \{ n \} \backslash (T \cup \{ i \} \cup S_t)} \left( p(\mu^{(t)})_j p(\mu^{(t)})_i - P_{F-\mu^{(t)}}[i, j \in F] \right).
\]

(5)

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

\[
\frac{(1 - p(\mu^{(t)})_i)^{n-k-1}}{1 - n - k - 1 - t} \times
\sum_{j \in \{ n \} \backslash \{ i \}} \left( p(\mu^{(t)})_j p(\mu^{(t)})_i - P_{F-\mu^{(t)}}[i, j \in F] \right)
\]

\[
\leq \frac{(1 - p(\mu^{(t)})_i)^{n-k-1}}{1 - n - k - 1 - t} p(\mu^{(t)})_i.
\]

(6)

where (i) follows from Lemma 28. This completes the proof.

\( \square \)

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

Lemma 31 (Variance and maximum change). With the setup in Definition 29 and \( 0 \leq t < n - s \), we have with probability 1 conditioned on \( S_t \) that

\[
p(\mu^{(t+1)})_i - p(\mu^{(t)})_i \leq \frac{p(\mu^{(t)})_i p(\mu^{(t)})_i (1 - (n - k - 1 - t))^{n-k-1}}{|1 - p(\mu^{(t)})_i|^2}.
\]

(6)

Also, we have the variance bound

\[
E_{S_{t+1}} \left[ \left( p(\mu^{(t+1)})_i - p(\mu^{(t)})_i \right)^2 \mid S_t \right] \leq \frac{1}{n - k - 1 - t} p(\mu^{(t)})_i (1 - p(\mu^{(t)})_i)^{n-k-1} p(\mu^{(t)})_i.
\]

(7)

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

\[
p(\mu^{(t+1)})_i - p(\mu^{(t)})_i = \frac{p(\mu^{(t)})_i p(\mu^{(t)})_i - P_{F-\mu^{(t)}}[i, j \in F]}{1 - p(\mu^{(t)})_j} \leq \frac{p(\mu^{(t)})_i (1 - p(\mu^{(t)})_i)^{n-k-1}}{1 - p(\mu^{(t)})_i}.
\]

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

\[
E_{S_{t+1}} \left[ \left( p(\mu^{(t+1)})_i - p(\mu^{(t)})_i \right)^2 \mid S_t \right] \leq \frac{p(\mu^{(t)})_i (1 - p(\mu^{(t)})_i)^{n-k-1}}{1 - p(\mu^{(t)})_i} p(\mu^{(t)})_i E_{S_{t+1}} \left[ p(\mu^{(t+1)})_i - p(\mu^{(t)})_i \mid S_t \right].
\]

(7)

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 32 ([44, 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 \( \text{Var}[X^{(u)} \mid X^{(u-1)}] \leq \sigma^2 u \) for \( u \in [t] \). Then

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

Proof of Theorem 26. For the setup in Definition 29, define the random variables \( Y^{(t)} := \log p(\mu^{(t)})_i \), indexed by \( 0 \leq t < n - s \). Given this, we define the martingale \( X^{(0)} = Y^{(0)} \) and

\[
Y^{(t+1)} := X^{(t)} + Y^{(t+1)} - Y^{(t)} - E_{S_{t+1}} \left[ Y^{(t+1)} - Y^{(t)} \mid S_t \right],
\]

for \( 0 \leq t < n - s \).

We will bound the drift, maximum change, and variance of \( Y^{(t+1)} \) assuming that \( p(\mu^{(t)})_i \leq 2 p(\mu^{(t)})_i / (n - t) \). For the bound on the maximum increase, we get

\[
\frac{Y^{(t+1)} - Y^{(t)}}{p(\mu^{(t)})_i} \leq \frac{p(\mu^{(t+1)})_i - p(\mu^{(t)})_i}{p(\mu^{(t)})_i} \leq \frac{p(\mu^{(t)})_i (1 - p(\mu^{(t)})_i)^{n-k-1}}{1 - p(\mu^{(t)})_i}.
\]

(8)
by Lemma 31 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 32. Because \( \mu \) 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 Eq. (8) and sufficiently large \( C \) for \( s \geq C(n p(\mu)_{\max} + \hat{k}) \log n \) so we may take \( M = 1/(1000 \log n) \) in Theorem 32. Additionally, we have that

\[
\text{Var}[X(u+1) | X(u)] \leq \mathbb{E}_{S_{t+1}} \left[ \left( Y(t+1) - Y(t) \right)^2 \right] \leq \frac{8np(\mu)_{\max}}{(n - k - 1 - u)(n - u)} \leq \frac{8np(\mu)_{\max}}{(n - k - 1 - u)^2}.
\]

by Eq. (9) so we may take \( \sigma_u^2 = 8np(\mu)_{\max} / (n - k - 1 - u)^2 \) in Theorem 32. By Theorem 32 for \( \lambda = 1/10, M = 1/(1000 \log n) \), and \( \sigma_u^2 = 8np(\mu)_{\max} / (n - k - 1 - u)^2 \), we get that

\[
P[X(t) - X(0) \geq 1/10] \leq \exp \left( -\frac{100}{2 \sum_{u \in [t]} \sigma_u^2 + M/15} \right) \leq \exp \left( -\frac{100}{150000 \log n} \right) \leq \exp(-20 \log n) = n^{-20}
\]

for sufficiently large \( C \) in \( s \geq C(n p(\mu)_{\max} + \hat{k}) \log n \geq C\hat{k} \log n \). To finish, note that

\[
Y(t) - Y(0) = X(t) - X(0) + \sum_{u \in [t-1]} \mathbb{E}_{S_{u+1}} \left[ Y(u+1) - Y(u) \right | S_u]
\]

\[
\leq \frac{1}{10} X(t) - X(0) + \sum_{u \in [t-1]} \frac{4p(\mu)_{\max}}{(n - k - 1 - u)(n - u)}
\]

\[
\leq \frac{20p(\mu)_{\max}}{n - k - 1 - t}
\]

\[
\leq X(t) - X(0) + \log \left( \frac{n}{n - t} \right) + \frac{1}{100 \log n},
\]

where (i) follows from Eq. (7), (ii) follows from direction calculations with Riemann integrals, and (iii) follows from \( t \leq n - s \) and \( s \geq C(n p(\mu)_{\max} + \hat{k}) \log n \). To conclude, we write

\[
\mathbb{P}[p(\mu(t)) \leq 2p(\mu)_{\max} / n - t] \leq \mathbb{P}[Y(t) - Y(0) \leq \log \left( \frac{n}{n - t} \right) + 1]
\]

\[
\leq \mathbb{P}[X(t) - X(0) \geq 1/10] \leq n^{-20},
\]

where (i) follows from Eq. (10). Theorem 26 now follows from union-bounding over all times \( t \in [n - s] \) and all coordinates \( i \in [n] \).

V. 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 \( \mu \in \mathbb{R}^{(n \times 2)} \) into two smaller distributions (supported on \( S_1, S_2 \) for \( |n| = S_1 \cup S_2 \)), and recursively produce marginal overestimates for \( \mu_{S_1} \) and \( \mu_{S_2} \) that sum to at most \( 4k \). Now, we merge these groups. Because \( \mu \) is strongly Rayleigh, the marginal overestimates on \( \mu_{S_1} \) and \( \mu_{S_2} \) provide marginal overestimates for \( \mu \) summing to at most \( 8k \). Thus, we can cheaply take \( O(n \log n / k) \) samples from \( \mu \) to get marginal overestimates summing to at most \( 4k \) again.

Theorem 33 (Isotropic rounding from sampling). Let \( \mu \in \mathbb{R}^{(n \times 2)} \) be a strongly Rayleigh distribution. Assume that we can sample from restrictions \( \mu_S \) of \( \mu \) (Definition 7) in time \( \mathcal{A}_\mu(K) \) given marginal overestimates of \( \mu_S \) that sum to at most \( K^2 \). Then there is an algorithm that produces marginal overestimates \( \bar{q}_i \geq p(\mu) \), with \( \sum_i \bar{q}_i \leq 4k \) in time \( \tilde{O}(n/k \cdot \mathcal{A}_\mu(8k)) \).

Proof. We use a divide-and-conquer procedure. Precisely, given a set \( S \) we use the following algorithm to produce marginals of \( S \) summing to at most \( 4k \). If \( |S| \leq 4k \),

\footnote{We do not write an \( n \) dependence as it will be polylogarithmic in our algorithms.}
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 $q_i^{(1)} \geq p(\mu_{S_1})$, and $q_i^{(2)} \geq p(\mu_{S_2})$, with $\sum_{i \in S} q_i^{(1)} \leq 4k$ and $\sum_{i \in S} q_i^{(2)} \leq 4k$.

Because $\mu$ is strongly Rayleigh, in fact $q_i^{(1)} \geq p(\mu_{S_1})$, for all $i \in S_1$ and $q_i^{(2)} \geq p(\mu_{S_2})$, for all $i \in S_2$. Hence the vector $q_i$ defined as $q_i = q_i^{(1)}$ for $i \in S_1$ and $q_i = q_i^{(2)}$ for $i \in S_2$ are marginal overestimates for $\mu$. Additionally, $\sum_{i \in S} q_i \leq 8k$.

Set $s = \frac{100|S| \log n}{k}$, and let $F_1, F_2, \ldots, F_s$ be independent samples from $\mu_S$ generated in total time $\tilde{O}(s \cdot A_\mu(8k))$, by using the overestimates $q_i$. Define for $i \in S$

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

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

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

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

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

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 described algorithm on $\mu$, 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 $O(n/k \cdot \log n)$.

VI. PROOFS OF MAIN RESULTS

In this section, we combine our previous results to show Theorems 2 and 34 and Corollaries 5 and 6. All results will follow from our entropy contraction bound Theorem 27 combined with Lemma 10. Also, Theorem 34 requires Theorem 33.

Proof of Theorem 2. We first apply Proposition 22 to instead focus on sampling from a strongly Rayleigh distribution $\mu'$ with all marginals bounded by $K/n$. Let $\mu'$ be the complement of $\mu$.

For $\kappa_1$ as in Theorem 27 we run $O(\kappa_1^{-1} \log n)$ steps of a down-up operator on the complement of our set, to converge to $\mu'$, i.e., iterate the Markov chain $D_{(n-k) \rightarrow (n-k') \rightarrow \cdots \rightarrow (n-k) \rightarrow (n-k-1)} U_{(n-k-1) \rightarrow (n-k) \rightarrow \cdots \rightarrow (n-k-1)}$. Note that each $U_{(n-k-1) \rightarrow (n-k)}$ needs to be implemented via a baseline sampling algorithm which takes time $T(k' - 1, k)$. By Theorem 27 and Lemma 10, and the fact that the up step $U_{(n-k') \rightarrow (n-k)}$ 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 $T_{\mu}(O(K), k)$ as $k' = (\mu/np(\mu'))^{\text{max}} = O(K)$ from Theorem 27.

Proof of Corollary 5. Let $n = |E|$, $k = |V|$. By the results of [25] a spanning tree can be sampled in $O(|E|)$ time on a graph with edge set $E$. Hence applying Theorem 2 with $K = O(k) = O(|V|)$ shows that after obtaining the initial overestimates, each future sample requires $O(|V|)$ time. Obtaining the original overestimates takes time $\tilde{O}(n/k \cdot k) = O(|E|)$ by Theorem 33.

Proof of Corollary 6. Let $\mu$ be the $k$-DPP with ensemble matrix $L$. Sampling from a $k$-DPP over ground set of size $n$ can be done in $O(n^2)$ time, see [6]. This together with Theorem 33 shows that in $O(n/k \cdot k^2)$ we can get marginal overestimates that sum to $O(k)$, as well as one initial sample $S_0$ from $\mu$. We now apply Theorem 2, and note that each oracle call is equivalent to sampling from a $k$-DPP on a size-$O(K)$ subset of $[n]$, which takes $O(k^2)$ 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 $O(n)$ steps [36], we show that under isotropy of $\mu$ it mixes in $O(k)$ steps.

Theorem 34. Given a strongly Rayleigh distribution $\mu \in \mathbb{R}^{|V|}$ and marginal overestimates $q_i \geq \mathbb{P}_{T_{\mu}[i \in T]}$ 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 $\mu$ in time bounded by $O(K)$ calls to $T_{\mu}(k+1, k)$. Additionally, given a strongly Rayleigh distribution $\mu \in \mathbb{R}^{|V|}$, we can produce marginal overestimates $q_i \geq \mathbb{P}_{T_{\mu}[i \in T]}$ with sum $\sum_{i \in [n]} q_i \leq O(k)$ in time $O(n \cdot T_{\mu}(k+1, k))$.

Proof of Theorem 34. For $\kappa_2$ as in Theorem 27 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 $\mu'$, i.e., iterate the Markov chain $D_{(n-k-1) \rightarrow (n-k)} U_{(n-k-1) \rightarrow (n-k)}$. Note that each $U_{(n-k-1) \rightarrow (n-k)}$ needs to be implemented via a baseline sampling algorithm which takes time $T(k+1, k)$. By Theorem 27 and Lemma 10, and the fact that the up step $U_{(n-k) \rightarrow (n-k-1)}$ 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 $T_{\mu}(k+1, k)$.

Finally, by Theorem 33, we can obtain the desired overestimates $q_i$ in time $\tilde{O}(n/k \cdot 8k \cdot T_{\mu}(k+1, k)) = T_{\mu}(n \cdot T_{\mu}(k+1, k))$ as desired.

REFERENCES

[1] J. Borcea, P. Brändén, and T. Liggett, “Negative dependence and the geometry of polynomials,” Journal of the American Mathematical Society, vol. 22, no. 2, pp. 521–567, 2009.

[2] M. Dereziński, “Fast determinantal point processes via distortion-free intermediate sampling,” in Proceedings of the Thirty-Second Conference on Learning Theory, ser. Proceedings of Machine Learning Research, A. Beygelzimer and D. Hsu, Eds., vol. 99, Phoenix, USA, 2019, pp. 1029–1049.
