JMLR: Workshop and Conference Proceedings vol 49:1–13, 2016

Monte Carlo Markov Chain Algorithms for Sampling Strongly Rayleigh Distributions and Determinantal Point Processes

Nima Anari ANARI@BERKELEY.EDU
Department of Management Science and Engineering, Stanford University.

Shayan Oveis Gharan * SHAYAN@CS.WASHINGTON.EDU

Alireza Rezaei * AREZAEI@CS.WSHINGTON.EDU
Department of Computer Science and Engineering, University of Washington

Abstract

Strongly Rayleigh distributions are natural generalizations of product and determinantal probability distributions and satisfy the strongest form of negative dependence properties. We show that the "natural" Monte Carlo Markov Chain (MCMC) algorithm mixes rapidly in the support of a *homogeneous* strongly Rayleigh distribution. As a byproduct, our proof implies Markov chains can be used to efficiently generate approximate samples of a k-determinantal point process. This answers an open question raised by Deshpande and Rademacher (2010) which was studied recently by Kang (2013); Li et al. (2015); Rebeschini and Karbasi (2015).

Keywords: determinantal point processes sampling, strongly Rayleigh distributions, markov chains, MCMC algorithms

1. Introduction

Let $\mu : 2^{[n]} \to \mathbb{R}_+$ be a probability distribution on the subsets of the set $[n] = \{1, 2, \dots, n\}$. In particular, we assume that $\mu(.)$ is nonnegative and,

$$\sum_{S \subseteq [n]} \mu(S) = 1.$$

We assign a multi-affine polynomial with variables $z_1, \dots, z_n$ to μ,

$$g_\mu(z) = \sum_{S \subseteq [n]} \mu(S) \cdot z^S,$$

where for a set $S \subseteq [n]$, $z^S = \prod_{i \in S} z_i$. The polynomial g_μ is also known as the *generating polynomial* of μ. We say μ is *k-homogeneous* if g_μ is a homogeneous polynomial of degree k, i.e., if for any $S \in \text{supp}\{\mu\}$, we have $|S| = k$.

A polynomial $p(z_1, \dots, z_n) \in \mathbb{C}[z_1, \dots, z_n]$ is *stable* if whenever $\text{Im}(z_i) > 0$ for all $1 \leq i \leq m$, $p(z_1, \dots, z_m) \neq 0$. We say $p(.)$ is real stable, if it is stable and all of its coefficients are real. Real stable polynomials are considered to be a natural generalization of real rooted polynomials to the multivariate setting. In particular, as a sanity check, one can observe that any univariate polynomial with real coefficients is real stable if and only if it is real rooted. We say that μ is a *strongly Rayleigh* distribution if g_μ is a real stable polynomial.

[*] This material is based on work supported by the National Science Foundation Career award.

Strongly Rayleigh distributions are introduced and deeply studied in the work of Borcea et al. (2009). These distributions are natural generalizations of determinantal measures and random spanning tree distributions. It is shown in Borcea et al. (2009) that strongly Rayleigh distributions satisfy the strongest form of negative dependence properties. These negative dependence properties were recently exploited to design approximation algorithms for various problems Oveis Gharan et al. (2011); Pemantle and Peres (2014); Anari and Oveis Gharan (2015).

In this paper we show that the "natural" Monte Carlo Markov Chain (MCMC) defined on the support of a homogeneous strongly Rayleigh distribution μ *mixes* rapidly. Therefore, this Markov Chain can be used to efficiently draw an approximate sample from μ. Since determinantal point processes are special cases of strongly Rayleigh measures, our result implies that the same Markov chain efficiently generates random samples of a k-determinantal point process (see Section 1.1 for the details).

We now describe the *lazy* MCMC $\mathcal{M}_\mu$. The state space of $\mathcal{M}$ is supp$\{\mu\}$ and the transition probability kernel P_μ is defined as follows. We may drop the subscript if μ is clear from the context. For a set $S \subseteq [n]$ and $i \in [n]$, let

$$\begin{aligned} S - i &= S \setminus \{i\}, \\ S + i &= S \cup \{i\}. \end{aligned}$$

In any state S, choose an element $i \in S$ and $j \notin S$ uniformly and independently at random, and let $T = S - i + j$; then

i) If $T \in$ supp$\{\mu\}$, move to T with probability $\frac{1}{2}\min\{1, \mu(T)/\mu(S)\}$;

이러면 최대 50% 확률로 변화를 시킨다.

ii) Otherwise, stay in S.

It is easy to see that $\mathcal{M}_\mu$ is reversible and $\mu(.)$ is the stationary distribution of the chain. In addition, Brändén showed that the support of a (homogeneous) strongly Rayleigh distribution is the set of bases of a matroid (Brändén, 2007, Cor 3.4); so $\mathcal{M}_\mu$ is irreducible. Lastly, since we stay in each state S with probability at least $1/2$, $\mathcal{M}_\mu$ is a lazy chain.

In our main theorem we show that the above Markov chain is *rapidly mixing*. In particular, if we start $\mathcal{M}_\mu$ from a state S, then after $\mathrm{poly}(n, k, \log(\frac{1}{\epsilon \cdot \mu(S)}))$ steps we obtain an ϵ-approximate sample of the strongly Rayleigh distribution. First, we need to setup the notation. For probability distributions $\pi, \nu : \Omega \to \mathbb{R}_+$, the total variation distance of π, ν is defined as follows:

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

If X is a random variable sampled according to ν and $\|\nu - \pi\|_{\mathrm{TV}} \leq \epsilon$, then we say X is an ϵ-approximate sample of π.

Definition 1 (Mixing Time) *For a state $x \in \Omega$ and $\epsilon > 0$, the total variation mixing time of a chain started at x with transition probability matrix P and stationary distribution π is defined as follows:*

$$\tau_x(\epsilon) := \min\{t : \|P^t(x, .) - \pi\|_{\mathrm{TV}} \leq \epsilon\},$$

where $P^t(x, .)$ is the distribution of the chain at time t.

The following is our main theorem.

Theorem 2 *For any strongly Rayleigh k-homogeneous probability distribution $\mu : 2^{[n]} \to \mathbb{R}_+$, $S \in \text{supp}\{\mu\}$ and $\epsilon > 0$,*

$$\tau_S(\epsilon) \leq \frac{1}{C_\mu} \cdot \log\left(\frac{1}{\epsilon \cdot \mu(S)}\right),$$

where

$$C_\mu := \min_{S,T\colon P(S,T)>0} \max(P(S,T), P(T,S)) \tag{1.1}$$

is at least $\frac{1}{2kn}$ by construction.

Suppose we have access to a set $S \in \text{supp}\{\mu\}$ such that $\mu(S) \geq \exp(-n)$. In addition, we are given an oracle such that for any set $T \in \binom{n}{k}$, it returns $\mu(T)$ if $T \in \text{supp}\{\mu\}$ and zero otherwise. Then, by the above theorem we can generate an ϵ-approximate sample of μ with at most $\text{poly}(n, k, \log(1/\epsilon))$ oracle calls.

For a strongly Rayleigh probability distribution $\mu : 2^{[n]} \to \mathbb{R}_+$, and any integer $0 \leq k \leq n$, the *truncation* of μ to k is the conditional measure μ_k where for any $S \subseteq [n]$ of size k,

$$\mu_k(S) = \frac{\mu(S)}{\sum_{S':|S'|=k} \mu(S')}.$$

Borcea, Brändén, and Liggett showed that for any strongly Rayleigh distribution μ, and any integer k, μ_k is also strongly Rayleigh, Borcea et al. (2009). Therefore, if we have access to a set $S \subset [n]$ of size k, we can use the above theorem to generate a random sample of μ_k.

1.1. Determinantal Point Processes and the Volume Sampling Problem

A determinantal point process (DPP) on a set of elements $[n]$ is a probability distribution μ on the set $2^{[n]}$ identified by a positive semidefinite $L \in \mathbb{R}^{n\times n}$ where for any $S \subseteq [n]$ we have

$$\mathbb{P}[S] \propto \det(L_S),$$

where L_S is the principal submatrix of L indexed by the elements of S. The matrix L is called the *ensemble* matrix of μ.

DPPs are one of the fundamental objects used to study a variety of tasks in machine learning, including text summarization, image search, news threading, etc. For more information about DPPs and their applications we refer to a recent survey by Kulesza and Taskar (2013).

For an integer $0 \leq k \leq n$, and a DPP μ, the truncation of μ to k, μ_k is called a k-DPP. It turns out that the family of determinantal point processes is not closed under truncation. Perhaps, the simplest example is the k-uniform distribution over a set of n elements. Although the uniform distribution over subsets of n elements is a DPP, for any $2 \leq k \leq n-2$, the corresponding k-DPP is not a DPP (Kulesza and Taskar, 2013, Section 5).

Generating a sample from a k-DPP is a fundamental computational task with many practical applications Kannan and Vempala (2009); Deshpande and Rademacher (2010); Kulesza and Taskar (2013). This problem is also equivalent to the k-volume sampling problem Deshpande et al. (2006); Kannan and Vempala (2009); Boutsidis et al. (2009); Deshpande and Rademacher (2010) which has applications in low-rank approximation and row-subset selection problem. In the volume sampling problem, we are given a matrix $X \in \mathbb{R}^{n\times m}$ and we want to choose a set $S \subseteq [n]$ of k rows of X with probability proportional to $\det(X_{S,[m]} X_{S,[m]}^\intercal)$, where $X_{S,[m]}$ is the submatrix of X with rows indexed by elements of S. If L is the ensemble matrix of a given k-DPP μ, and $L = XX^\intercal$ is the Cholesky decomposition of L, then the k-volume sampling problem on X is equivalent to the problem of generating a random sample of μ.

In the past, several spectral algorithms were designed for sampling from a k-DPP Hough et al. (2006); Deshpande and Rademacher (2010); Kulesza and Taskar (2013), but these algorithms typically need to diagonalize a giant n-by-n matrix, so they are inefficient in time and memory [1]. It was asked by Deshpande and Rademacher (2010) to generate random samples of a k-DPP using Markov chain techniques. Markov chain techniques are very appealing in this context because of their simplicity and efficiency. There has been several attempts Kang (2013); Li et al. (2015); Rebeschini and Karbasi (2015) to upper bound the mixing time of the Markov chain $\mathcal{M}_\mu$ for a k-DPP μ; but, to the best of our knowledge this question is still open[2].

Here, we show that for a k-DPP μ, $\mathcal{M}_\mu$ can be used to efficiently generate an approximate sample of μ. Borcea et al. (2009) show that any DPP is a strongly Rayleigh distribution. Since strongly Rayleigh distributions are closed under truncation, any k-DPP is a strongly Rayleigh distribution. Therefore, by Theorem 2, for any k-DPP μ, $\mathcal{M}_\mu$ mixes rapidly to the stationary distribution.

Corollary 3 *For any k-DPP $\mu : 2^{[n]} \to \mathbb{R}_+$, $S \in \text{supp}\{\mu\}$ and $\epsilon > 0$,*

$$\tau_S(\epsilon) \leq \frac{1}{C_\mu} \cdot \log\left(\frac{1}{\epsilon \cdot \mu(S)}\right).$$

Given access to the ensemble matrix of a k-DPP, we can use the above theorem to generate ϵ-approximate samples of the k-DPP.

Theorem 4 *Given an ensemble matrix L of a k-DPP μ, for any $\epsilon > 0$, there is an algorithm that generates an ϵ-approximate sample of μ in time $\text{poly}(k)O(n\log(n/\epsilon))$.*

To prove the above theorem, we need an efficient algorithm to generate a set $S \in \text{supp}\{\mu\}$ such that $\mu(S)$ is bounded away from zero, perhaps by an exponentially small function of n, k. We use the greedy algorithm 1 to find such a set, and we show that, in time $O(n)\,\text{poly}(k)$, it returns a set S such that

$$\mu(S) \geq \frac{1}{k!|\text{supp}\{\mu\}|} \geq n^{-2k}. \tag{1.2}$$

Noting that each transition step of the Markov chain $\mathcal{M}_\mu$ only takes time that is polynomial in k, this completes the proof of the above theorem.

Algorithm 1 Greedy Algorithm for Selecting the Starting State of $\mathcal{M}_\mu$

```
input : The ensemble matrix, $L$, of a $k$-DPP $\mu$.
$S \leftarrow \emptyset$. for $i = 1$ to $k$ do
    Among all elements $j \notin S$ pick the one maximizing $\det(L_{S+j})$ and let $S \leftarrow S + j$.
end
```

It remains to analyze Algorithm 1. This problem is already studied by Çivril and Magdon-Ismail (2009) in the context of maximum volume submatrix problem. In the maximum volume submatrix problem, given

[1] We remark that the algorithms in Deshpande and Rademacher (2010) are almost linear in n; however they need access to the Cholesky decomposition of the ensemble matrix of the underlying DPP.

[2] We remark that Kang (2013) claimed to have a proof of the rapid mixing time of a similar Markov chain. As it is pointed out in Rebeschini and Karbasi (2015) the coupling argument of Kang (2013) is ill-defined. To be more precise, the chain specified in Algorithm 1 of Kang (2013) may not mix in a polynomial time of n. The chain specified in Algorithm 2 of Kang (2013) is similar to $\mathcal{M}_\mu$, but the statement of Theorem 2 which upper bounds its mixing time is clearly incorrect even when $k = 1$.

a matrix $X \in \mathbb{R}^{n\times m}$, we want to choose a subset S of k rows of X maximizing $\det(X_{S,[m]}X_{S,[m]}^\intercal)$. Equivalently, given a matrix $L = XX^\intercal$, we want to choose $S \subseteq [n]$ of size k maximizing $\det(L_S)$. Note that if L is an ensemble matrix of a k-DPP μ, then

$$\max_{|S|=k} \mu(S) = \frac{\max_{|S|=k} \det(L_S)}{\sum_{|S|=k} \det(L_S)} \geq \frac{1}{|\text{supp}\{\mu\}|} \geq n^{-k}.$$

The maximum volume submatrix problem is NP-hard to approximate within a factor c^k for some constant $c > 1$ Çivril and Magdon-Ismail (2013). Numerous approximation algorithm are given for this problem Çivril and Magdon-Ismail (2009, 2013); Nikolov (2015). It was shown in (Çivril and Magdon-Ismail, 2009, Thm 11) that choosing the rows of X greedily gives a $k!$ approximation to the maximum volume submatrix problem. Algorithm 1 is equivalent to the greedy algorithm of Çivril and Magdon-Ismail (2009); it is only described in the language of ensemble matrix L. Therefore, it returns a set S such that

$$\mu(S) \geq \frac{\max_{|T|=k} \det(L_T)}{k! \sum_{|T|=k} \det(L_T)} \geq \frac{1}{k!|\text{supp}\{\mu\}|},$$

as desired.

1.2. Proof Overview

In the rest of the paper we prove Theorem 2. To prove Theorem 2, we lower bound the spectral gap, a.k.a. the Poincaré constant of the chain $\mathcal{M}_\mu$. This directly upper bounds the mixing time in total variation distance. To lower bound the spectral gap, we use an extension of the seminal work of Feder and Mihail (1992). Feder and Mihail showed that the *bases exchange* graph of the bases of a *balanced matroid* is an *expander*. This directly lower bounds the spectral gap by Cheeger's inequality. A matroid is called balanced if the matroid and all of its minors satisfy the property that the uniform distribution of the bases is negatively associated (see Section 2.2 for the definition of negative correlation).

Our proof can be seen as a *weighted* variant of Feder and Mihail (1992). As we mentioned earlier, the support of a homogeneous strongly Rayleigh distribution corresponds to the bases of a matroid. Our proof shows that if a distribution μ over the bases of a matroid and all of its conditional measures are negatively associated, then the MCMC algorithm mixes rapidly. To show that μ satisfies the aforementioned property we simply appeal to the negative dependence theory of strongly Rayleigh distributions developed in Borcea et al. (2009). Although our proof can be written in the language of Feder and Mihail (1992), we work with the more advanced chain decomposition idea of Jerrum et al. (2004) to prove a tight bound on the Poincaré constant; see Section 2.3 for the details.

We remark that the decomposition idea of Jerrum et al. (2004) can be used to lower bound the *log-Sobolev* constant of $\mathcal{M}_\mu$. However, it turns out that in our case, the log-Sobolev constant may be no larger than $\frac{1}{-\log(\min_{S\in\text{supp}\{\mu\}} \mu(S))}$. Since the latter quantity is not necessarily lower-bounded as a function of k, n, the L_2 mixing time of the chain may be unbounded.

2. Background

2.1. Markov Chains and Mixing Time

In this section we give a high level overview of Markov chains and their mixing times. We refer the readers to Levin et al. (2006); Montenegro and Tetali (2006) for details. Let Ω denote the state space, P denote the

Markov kernel and $\pi(.)$ denote the stationary distribution of a Markov chain. We say a Markov chain is *lazy* if for any state $x \in \Omega$, $P(x, x) \geq 1/2$.

A Markov chain (Ω, P, π) is reversible if for any pair of states $x, y \in \Omega$, $\pi(x)P(x, y) = \pi(y)P(y, x)$. This is known as the *detailed balanced* condition. In this paper we only work with reversible chains. We equip the space of all functions $f : \Omega \to \mathbb{R}$ with the standard inner product for $L^2(\pi)$,

$$\langle f, g \rangle_\pi := \mathbb{E}_\pi [f \cdot g] = \sum_{x \in \Omega} \pi(x) f(x) g(x).$$

In particular, $\|f\|_\pi = \sqrt{\langle f, f \rangle_\pi}$. For a function $f \in L^2(\pi)$, the *Dirichlet form* $\mathcal{E}_\pi(f, f)$ is defined as follows

$$\mathcal{E}_\pi(f, f) := \frac{1}{2} \sum_{x,y \in \Omega} (f(x) - f(y))^2 P(x, y)\pi(x),$$

and the *Variance* of f is

$$\mathrm{Var}_\pi(f) := \|f - \mathbb{E}_\pi f\|_\pi^2 = \sum_{x \in \Omega} (f(x) - \mathbb{E}_\pi f)^2 \pi(x).$$

Next, we overview classical spectral techniques to upper bound the mixing time of Markov chains.

Definition 5 (Poincaré Constant) *The* Poincaré constant *of the chain is defined as follows,*

$$\lambda := \inf_{f:\Omega \to \mathbb{R}} \frac{\mathcal{E}_\pi(f, f)}{\mathrm{Var}_\pi(f)},$$

where the infimum is over all functions with nonzero variance.

It is easy to see that for any transition probability matrix P, the second largest eigenvalue of P is $1 - \lambda$. If P is a lazy chain, then $1 - \lambda$ is also the second largest eigenvalue of P in absolute value. In the following fact we see how to calculate the Poincaré constant of any reversible 2-state chain.

Fact 6 *The Poincaré constant of any reversible two state chain with $\Omega = 0, 1$ and $P(0, 1) = c \cdot \pi(1)$ is c.*

Proof Consider any function f. Since $\mathrm{Var}(f)$ is shift-invariant, we can assume $\mathbb{E}_\pi f = 0$, i.e., $\pi(0)f(0) = -\pi(1)f(1)$. Since $\frac{\mathcal{E}_\pi(f,f)}{\mathrm{Var}_\pi(f)}$ is invariant under the scaling of f, we can assume $f(0) = \pi(1)$ and $f(1) = -\pi(0)$. Since the chain is reversible $P(1, 0) = c \cdot \pi(0)$. Plugging this unique f into the ratio we obtain $\lambda = c$. ■

To prove Theorem 2 we simply calculate the Poincaré constant of the chain $\mathcal{M}_\mu$ and then we use the following classical theorem of Diaconis and Stroock to upper bound the mixing time.

Theorem 7 ((Diaconis and Stroock, 1991, Prop 3)) *For any reversible irreducible lazy Markov chain (Ω, P, π) with Poincaré constant λ, for any $\epsilon > 0$, and any state $x \in \Omega$,*

$$\tau_x(\epsilon) \leq \frac{1}{\lambda} \cdot \log \left(\frac{1}{\epsilon \cdot \pi(x)} \right)$$

Using the above theorem, one see that in order to prove Theorem 2, it is enough to lower bound the Poincaré constant of $\mathcal{M}_\mu$.

Theorem 8 *For any k-homogeneous strongly Rayleigh distribution $\mu : 2^{[n]} \to \mathbb{R}_+$, the Poincaré constant of the chain $\mathcal{M}_\mu = (\Omega_\mu, P_\mu, \mu)$ is at least*

$$\lambda \geq C_\mu.$$

It is easy to see that Theorem 2 follows by the above two theorems.

We also remark that the bound of Theorem 8 on λ is tight. To show this we give an example of the k-volume sampling problem with n vectors where the poincaré constant of the corresponding Markov chain is $O(\frac{1}{kn})$. Note that in this case $C_\mu = \frac{1}{2kn}$. Here is the example: Let $v_1, v_2, \ldots, v_n \in \mathbb{R}^k$ be n vectors where $v_1 = v_2$ and the rest of them are orthogonal to v_1. Note that each element of supp$\{\mu\}$ contains exactly one of v_1 or v_2 where μ is the stationary distribution. Now define A to be the collection of the elements containing v_1, and set $f :$ supp$\{\mu\} \to \{-1, 1\}$ to be 1 on A and -1 on $\overline{A}$. Then, it is easy to verify that $\text{Var}_\mu(f) = 1$ and $\mathcal{E}_\mu(f, f) = 1/kn$ which implies

$$\lambda_\mu \geq \frac{\mathcal{E}_\mu(f,f)}{\text{Var}_\mu(f)} = 1/kn.$$

2.2. Strongly Rayleigh Measures

A probability distribution $\mu : 2^{[n]} \to \mathbb{R}_+$ is pairwise *negatively correlated* if for any pair of elements $i, j \in [n]$,

$$\mathbb{P}_{S\sim\mu}[i \in S] \cdot \mathbb{P}_{S\sim\mu}[j \in S] \geq \mathbb{P}_{S\sim\mu}[i, j \in S].$$

Feder and Mihail (1992) defined *negative association* as a generalization of negative correlation. We say an event $\mathcal{A} \subseteq 2^{[n]}$ is increasing if it is upward closed under containment, i.e., if $S \in \mathcal{A}$, and $S \subseteq T$, then $T \in \mathcal{A}$. We say a function $f : 2^{[n]} \to \mathbb{R}_+$ is *increasing* if it is the indicator function of an increasing event. We say μ is negatively associated if for any pair of increasing functions $f, g : 2^{[n]} \to \mathbb{R}_+$ depending on disjoint sets of coordinates,

$$\mathbb{E}_\mu[f] \cdot \mathbb{E}_\mu[g] \geq \mathbb{E}_\mu[f \cdot g].$$

Building on Feder and Mihail (1992), Borcea, Brändén and Liggett proved that any strongly Rayleigh distribution is negatively associated.

Theorem 9 (Borcea et al. (2009)) *Any strongly Rayleigh probability distribution is negatively associated.*

As an example, the above theorem implies that any k-DPP is negatively associated. The negative association property is the key to our lower bound on the Poincaré constant of the chain $\mathcal{M}$.

For $1 \leq i \leq n$, let Y_i be the random variable indicating whether i is in a sample of μ. We use

$$\mu|_{\bar{i}} := \{\mu | Y_i = 0\}, \mu|_i := \{\mu | Y_i = 1\}.$$

In addition, they showed that these distributions are closed under conditioning.

Theorem 10 (Borcea et al. (2009)) *For any strongly Rayleigh distribution $\mu : 2^{[n]} \to \mathbb{R}_+$ and any $1 \leq i \leq n$, the distributions $\mu|_{\bar{i}}, \mu|_i$ are strongly Rayleigh.*

The above two theorems are the only properties of the strongly Rayleigh distributions that we use in the proof of Theorem 2. In other words, the statement of Theorem 2 holds for any homogeneous probability distribution $\mu : 2^{[n]} \to \mathbb{R}_+$ where μ and all of its conditional measures are negatively associated.

2.3. Decomposable Markov Chains

In this section we describe the decomposable Markov chain technique due to Jerrum, Son, Tetali and Vigoda Jerrum et al. (2004). This will be our main tool to lower bound the Poincaré constant of $\mathcal{M}_\mu$. Roughly speaking, they consider Markov chains that can be decomposed into "projection" and "restriction" chains. They lower bound the Poincaré constant of the original chain assuming certain properties of these projection/restriction chains.

Let $\Omega_0 \cup \Omega_1$ be a decomposition of the state space of a Markov chain (Ω, P, π) into two disjoint sets[3]. For $i \in \{0,1\}$ let

$$\bar{\pi}(i) = \sum_{x \in \Omega_i} \pi(x),$$

and let $\bar{P} \in \mathbb{R}^{2\times 2}$ be

$$\bar{P}(i,j) = \bar{\pi}(i)^{-1} \sum_{x \in \Omega_i, y \in \Omega_j} \pi(x) P(x,y).$$

The Markov chain $(\{0,1\}, \bar{P}, \bar{\pi})$ is called a projection chain. Let $\bar{\lambda}$ be the Poincaré constant of this chain.

We can also define a restriction Markov chain on each Ω_i as follows. For each $i \in \{0,1\}$,

$$P_i(x,y) = \begin{cases} P(x,y) & \text{if } x \neq y, \\ P(x,x) + \sum_{z \notin \Omega_i} P(x,z) & \text{if } x = y. \end{cases}$$

In other words, for any transition from x to a state outside of Ω_i, we remain in x. Observe that in the stationary distribution of the restriction chain, the probability of x is proportional to $\pi(x)$. Let λ_i be the Poincaré constant of the chain $(\Omega_i, P_i, .)$. Now, we are ready to explain the main result of Jerrum et al. (2004).

Theorem 11 ((Jerrum et al., 2004, Cor 3)) *If for any distinct $i,j \in \{0,1\}$, and any $x \in \Omega_i$,*

$$\bar{P}(i,j) = \sum_{y \in \Omega_j} P(x,y), \tag{2.1}$$

then the Poincaré constant of (Ω, P, π) is at least $\min\{\bar{\lambda}, \lambda_0, \lambda_1\}$.

3. Inductive Argument

In this section we prove Theorem 8. Throughout this section we fix a strongly Rayleigh distribution μ, and we let Ω, P be the state space and the transition probability matrix of $\mathcal{M}_\mu$.

We prove Theorem 8 by induction on $|\text{supp}\{\mu\}|$. If $|\text{supp}\{\mu\}| = 1$, then there is nothing to prove. To do the induction step, we will use Theorem 11. So, let us first start by defining the restriction chains. Without loss of generality, perhaps after renaming, let n be an element such that $0 < \mathbb{P}_{S\sim\mu}[n \in S] < 1$. Let $\Omega_0 = \{S \in \text{supp}\{\mu\} : n \notin S\}$ and $\Omega_1 = \{S \in \text{supp}\{\mu\} : n \in S\}$. Note that both of these sets are nonempty. Observe that the restricted chain $(\Omega_0, P_0, .)$ is the same as $\mathcal{M}_{\mu|_{\overline{n}}}$ and $(\Omega_1, P_1, .)$ is the same as $\mathcal{M}_{\mu|_n}$. In addition, by Theorem 10, $\mu|_{\overline{n}}$ and $\mu|_n$ are strongly Rayleigh, and also clearly $C_{\mu|_n}, C_{\mu|_{\overline{n}}} \geq C_\mu$. So, we can use the induction hypothesis to lower bound $\lambda_0, \lambda_1 \geq C_\mu$.

It remains to lower bound the Poincaré constant of the projection chain and to prove equation (2.1). Unfortunately, P does not satisfy (2.1). So, we use an idea of Jerrum et al. (2004). We construct a new

[3] Here, we only focus on decomposition into two disjoint sets, although the technique of Jerrum et al. (2004) is more general.

Markov kernel $\hat{P}$ satisfying (2.1) such that (i) $\hat{P}$ has the same stationary distribution. (ii) The Poincaré constant of $\hat{P}$, $\hat{\lambda}$ lower-bounds λ. Then we use Theorem 11 to lower bound $\hat{\lambda}$.

To make sure that $\hat{P}$ satisfies (i), (ii), it is enough that for all distinct states $x, y \in \Omega$,

$$\begin{aligned} \mu(x)\hat{P}(x,y) &= \mu(y)\hat{P}(y,x), \qquad (3.1) \\ \hat{P}(x,y) &\leq P(x,y). \qquad (3.2) \end{aligned}$$

Equation (3.1) implies (i), i.e., that μ is also the stationary distribution of $\hat{P}$. By an application of the comparison method Diaconis and Saloff-Coste (1993) (i) together with (3.2) implies (ii), i.e.,

$$\hat{\lambda} \leq \lambda. \tag{3.3}$$

So, to prove the induction step, it is enough to show that

$$\hat{\lambda} \geq C_\mu. \tag{3.4}$$

Lemma 12 *There is a transition probability matrix $\hat{P} : \Omega \times \Omega \to \mathbb{R}_+$ such that*

1) $\hat{P}$ satisfies (3.1), (3.2).

2) For any $i \in \{0, 1\}$ and any distinct states $x, y \in \Omega_i$, $\hat{P}(x,y) = P(x,y)$.

3) The Poincaré constant of the chain $(\Omega, \hat{P}, \mu)$ projected onto Ω_0, Ω_1 is at least $\bar{\hat{\lambda}} \geq C_\mu$,

4) For any state $x \in \operatorname{supp}\{\mu\}$ and distinct $i, j \in \{0, 1\}$,

$$\bar{\hat{P}}(i,j) = \sum_{y \in \Omega_j} \hat{P}(x,y).$$

Before proving the above lemma, we use it to finish the proof of the induction. By part (2), $\hat{P}$ agrees with P on the projection chains. Therefore, the Poincaré constants of the chains $(\Omega_0, \hat{P}_0, .)$ and $(\Omega_1, \hat{P}_1, .)$ are at least $\hat{\lambda}_0, \hat{\lambda}_1 \geq C_\mu$. So, by parts (3) and (4) we can invoke Theorem 11 for $\hat{P}$ and we get that

$$\hat{\lambda} \geq \min\{\bar{\hat{\lambda}}, \hat{\lambda}_0, \hat{\lambda}_1\} \geq C_\mu.$$

This proves (3.4). As we discussed earlier, part (1) implies (3.3) which completes the induction.

3.1. Proof of Lemma 12

In the rest of this section we prove Lemma 12. Note that the main challenge in proving the lemma is part (4). The transition probability matrix P already satisfies part (1)-(3). The key to proving part (4) is to construct a fractional perfect matching between the states of Ω_0 and Ω_1; see the following lemma for the formal definition. This idea originally was used in Feder and Mihail (1992) and it was later extended in Jerrum and Son (2002).

Lemma 13 *There is a function $w : \{\{x,y\} : x \in \Omega_0, y \in \Omega_1\} \to \mathbb{R}_+$ such that $w_{\{x,y\}} > 0$ only if $P(x,y) > 0$ and*

$$\begin{aligned} \sum_{y \in \Omega_1} w_{\{x,y\}} &= \frac{\mu(x)}{\mu(\Omega_0)} \qquad \forall x \in \Omega_0, \\ \sum_{x \in \Omega_0} w_{\{x,y\}} &= \frac{\mu(y)}{\mu(\Omega_1)} \qquad \forall y \in \Omega_1. \end{aligned} \tag{3.5}$$

We use the negative association property of the strongly Rayleigh distributions to prove the above lemma. But before that let us prove Lemma 12.

Proof of Lemma 12. We use w to construct $\hat{P}$. For any $i,j \in \{0,1\}$ and $x \in \Omega_i$ and $y \in \Omega_j$ where $x \neq y$, we let

$$\hat{P}(x,y) = \begin{cases} \frac{C_\mu}{\mu(x)} \mu(\Omega_i)\mu(\Omega_j) w_{\{x,y\}} & \text{if } i \neq j, \\ P(x,y) & \text{otherwise.} \end{cases}$$

We also set $\hat{P}(x,x) = 1 - \sum_{y \neq x \in \Omega} \hat{P}(x,y)$ for any $x \in \Omega$. Note that by definition part (2) is satisfied. First we verify part (1). If $i \neq j$, then

$$\hat{P}(x,y)\mu(x) = C_\mu \mu(\Omega_i)\mu(\Omega_j) w_{\{x,y\}} = \hat{P}(y,x)\mu(y),$$

and if $i = j$ the same identity holds because $\hat{P}(x,y) = P(x,y)$. This proves (3.1). To see (3.2), let $x \in \Omega_i, y \in \Omega_j$ be two distinct states. First note that WLOG we can assume $i \neq j$ and $P(x,y) \neq 0$; otherwise clearly $\hat{P}(x,y) = P(x,y)$. So we have

$$\begin{aligned} \hat{P}(x,y) &= \frac{C_\mu}{\mu(x)} \cdot \mu(\Omega_0)\mu(\Omega_1) w_{\{x,y\}} \\ &\leq \frac{\max(P(x,y), P(y,x))}{\mu(x)} \mu(\Omega_i)\mu(\Omega_j) w_{\{x,y\}} \\ &\leq \max(P(x,y), P(y,x)) \cdot \frac{\min(\mu(x), \mu(y))}{\mu(x)} \leq P(x,y). \end{aligned}$$

The first inequality follows by the definition of C_μ (see (1.1)), and the second inequality follows by the fact that $w_{\{x,y\}} \leq \frac{\mu(x)}{\mu(\Omega_0)}$ and $w_{\{x,y\}} \leq \frac{\mu(y)}{\mu(\Omega_1)}$, and the last inequality follows by the detailed balanced condition. This completes the proof of part (1).

Next, we prove part (3). By the definition of $\hat{P}$, for distinct $i,j \in \{0,1\}$ we have

$$\begin{aligned} \bar{\hat{P}}(i,j) &= \frac{1}{\mu(\Omega_i)} \sum_{x \in \Omega_i, y \in \Omega_j} \mu(x)\hat{P}(x,y) \\ &= \frac{C_\mu}{\mu(\Omega_i)} \sum_{x \in \Omega_i, y \in \Omega_j} \mu(\Omega_i)\mu(\Omega_j) w(x,y) \\ &= C_\mu \cdot \mu(\Omega_j) \sum_{x \in \Omega_i} \frac{\mu(x)}{\mu(\Omega_i)} = C_\mu \cdot \mu(\Omega_j), \end{aligned}$$

where the second to last equality follows by (3.5). By Fact 6, the Poincaré constant of $\bar{\hat{P}} = C_\mu$. This proves part (3).

Finally we prove part (4). Fix distinct $i, j \in \{0, 1\}$ and $z \in \Omega_i$. We have,

$$\sum_{y \in \Omega_j} \hat{P}(z, y) = \frac{C_\mu}{\mu(z)} \mu(\Omega_i)\mu(\Omega_j) \sum_{y \in \Omega_j} w_{\{z,y\}} = C_\mu \cdot \mu(\Omega_j),$$

where we used (3.5). On the other hand, by the definition of $\hat{P}$ we know that

$$\bar{\hat{P}}(i, j) = \frac{1}{\mu(\Omega_i)} \sum_{x \in \Omega_i, y \in \Omega_j} \mu(x)\bar{\hat{P}}(x, y) = C_\mu \cdot \mu(\Omega_j) \sum_{x \in \Omega_i} \frac{\mu(x)}{\mu(\Omega_i)} = C_\mu \cdot \mu(\Omega_j),$$

where the second equality follows by (3.5). This completes the proof of part (4) and Lemma 12. ■

It remains to prove Lemma 13. For a set $A \subseteq \Omega$ let

$$N(A) = \{y \in \Omega \setminus A : \exists x \in A, P(x, y) > 0\}.$$

To prove Lemma 13 we use a maximum flow-minimum cut argument. To prove the claim we need to show that the support graph of the transition probability matrix P_μ satisfies Hall's condition. This is proved in the following lemma using the negative association property of strongly Rayleigh measures. The proof is simply an extension of the proof of (Feder and Mihail, 1992, Lem 3.1).

Lemma 14 *For any $A \subseteq \Omega_1$,*

$$\frac{\mu(N(A))}{\mu(\Omega_0)} \geq \frac{\mu(A)}{\mu(\Omega_1)}.$$

Proof Let $R \sim \mu$ be a random set. Recall that $\Omega_0 = \{S \in \text{supp}\{\mu\} : n \notin S\}$ and $\Omega_1 = \{S \in \text{supp}\{\mu\} : n \in S\}$. Let g be a random variable indicating whether $n \in R$. Let f be an indicator random variable which is 1 if there exists $T \in A$ such that $R \supseteq T \setminus \{n\}$. It is easy to see that f and g are two increasing functions which depend on two disjoint sets of elements. By the negative association property, Theorem 9, we can write

$$\mathbb{P}_\mu \left[f(R) = 1 | g(R) = 0\right] \geq \mathbb{P}_\mu \left[f(R) = 1 | g(R) = 1\right].$$

The lemma follows by the fact that the LHS of the above inequality is $\frac{\mu(N(A))}{\mu(\Omega_0)}$ and the RHS is $\frac{\mu(A)}{\mu(\Omega_1)}$. ■

Proof of Lemma 13. Let G be a bipartite graph on $\Omega_0 \cup \Omega_1$ where there is an edge between $x \in \Omega_1$ and $y \in \Omega_0$ if $P(x, y) > 0$. We prove the lemma by showing there is a unit flow from Ω_1 to Ω_0 such that the amount of flow going out of any $x \in \Omega_1$ is $\frac{\mu(x)}{\mu(\Omega_1)}$, and the incoming flow to any $y \in \Omega_0$ is $\frac{\mu(y)}{\mu(\Omega_0)}$. Then, we simply let $w_{\{x,y\}}$ be the flow on the edge connecting x to y.

Add a source s and a sink t. For any $x \in \Omega_1$ add an arc (s, x) with capacity $c_{s,x} = \mu(x)/\mu(\Omega_1)$. Similarly, for any $y \in \Omega_0$ add an arc (y, t) with capacity $c_{y,t} = \mu(y)/\mu(\Omega_0)$. Let the capacity of any other edge in the graph be ∞. Since the sum of the capacities of all edges leaving s is 1, to prove the lemma, it is enough to show that the maximum flow is 1. Equivalently, by the max-flow min-cut theorem, it suffices to show that the value of the minimum cut separating s and t is at least 1. Let $B, \overline{B}$ be an arbitrary s-t cut, and assume that $s \in B$ and $t \in \overline{B}$. Let $B_0 = \Omega_0 \cap B$ and $B_1 = \Omega_1 \cap B$. For disjoint $X, Y \subseteq \Omega$, let

$c(X,Y) = \sum_{x\in X, y\in Y} c_{x,y}$. We have

$$\begin{aligned} c(B,\overline{B}) &\geq c(s,\Omega_1\setminus B_1) + c(B_0,t) \\ &= \frac{\mu(\Omega_1\setminus B_1)}{\mu(\Omega_1)} + \frac{\mu(B_0)}{\mu(\Omega_0)} \\ &= 1 - \frac{\mu(B_1)}{\mu(\Omega_1)} + \frac{\mu(B_0)}{\mu(\Omega_0)} \geq 1 - \frac{\mu(N(B_1))}{\mu(\Omega_0)} + \frac{\mu(B_0)}{\mu(\Omega_0)}, \end{aligned} \tag{3.6}$$

where the inequality follows by Lemma 13. If there are any edge from B_1 to $\Omega_0 \setminus B_0$, then $c(B,\overline{B}) = \infty$ and we are done. Otherwise, $N(B_1) \subseteq B_0$. Therefore, $\mu(N(B_1)) \leq \mu(B_0)$, and the RHS of the above inequality is at least 1. So, $c(B,\overline{B}) \geq 1$ as desired. ■

References

Nima Anari and Shayan Oveis Gharan. Effective-Resistance-Reducing Flows and Asymmetric TSP. In *FOCS*, pages 20–39, 2015.

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

Christos Boutsidis, Michael W Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *SODA*, pages 968–977, 2009.

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

Ali Çivril and Malik Magdon-Ismail. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoretical Computer Science*, 410(47):4801–4811, 2009.

Ali Çivril and Malik Magdon-Ismail. Exponential inapproximability of selecting a maximum volume submatrix. *Algorithmica*, 65(1):159–176, 2013.

Amit Deshpande and Luis Rademacher. Efficient volume sampling for row/column subset selection. In *FOCS*, pages 329–338. IEEE, 2010.

Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. In *SODA*, pages 1117–1126, 2006.

Persi Diaconis and Laurent Saloff-Coste. Comparison theorems for reversible markov chains. *The Annals of Applied Probability*, pages 696–730, 1993.

Persi Diaconis and Daniel Stroock. Geometric bounds for eigenvalues of markov chains. *The Annals of Applied Probability*, pages 36–61, 1991.

Tomás Feder and Milena Mihail. Balanced matroids. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of Computing*, pages 26–38, New York, NY, USA, 1992. ACM.

J.B. Hough, M. Krishnapur, Y. Peres, and B. Virág. Determinantal processes and independence. *Probability Surveys*, (3):206–229, 2006.

Mark Jerrum and Jung Bae Son. Spectral gap and log-sobolev constant for balanced matroids. In *FOCS*, pages 721–729, 2002.

Mark Jerrum, Jung-Bae Son, Prasad Tetali, and Eric Vigoda. Elementary bounds on poincaré and log-sobolev constants for decomposable markov chains. *Annals of Applied Probability*, pages 1741–1765, 2004.

Byungkon Kang. Fast determinantal point process sampling with application to clustering. In *NIPS*, pages 2319–2327, 2013.

R. Kannan and S. Vempala. Spectral algorithms. *Foundations and Trends in Theoretical Computer Science*, 4:157–288, 2009.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. 2013. URL http://arxiv.org/abs/1207.6083.

David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2006.

Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Efficient sampling for k-determinantal point processes. 2015. URL http://arxiv.org/abs/1509.01618.

Ravi Montenegro and Prasad Tetali. Mathematical aspects of mixing times in Markov chains. *Found. Trends Theor. Comput. Sci.*, 1(3):237–354, May 2006. ISSN 1551-305X.

Aleksandar Nikolov. Randomized rounding for the largest simplex problem. In *STOC*, pages 861–870, 2015.

Shayan Oveis Gharan, Amin Saberi, and Mohit Singh. A Randomized Rounding Approach to the Traveling Salesman Problem. In *FOCS*, pages 550–559, 2011.

Robin Pemantle and Yuval Peres. Concentration of Lipschitz Functionals of Determinantal and Other Strong Rayleigh Measures. *Combinatorics, Probability and Computing*, 23:140–160, 1 2014.

Patrick Rebeschini and Amin Karbasi. Fast mixing for discrete point processes. In *COLT*, pages 1480–1500, 2015.