CSE 521 Design and Analysis of Algorithms I Spring 2019

Summary of ”MCMC for Sampling Strongly Rayleigh Distribution”

Sean Yin 1771984 *June 10th*

Disclaimer: *The report has not been subjected to the usual scrutiny reserved for formal publications.*

A discrete point process is a probability measure over subset $S \subseteq [n]$, where $[n] = \{1, 2, ..., N\}$ can be represented as N items of objects. For any feature function g on any items in $[n]$, we defined $\mathbf{L}_{ij} := g(i)^T g(j)$. A determinantal point process (DPP) is a probability distribution $\mathbb{P}[S] \propto det(\mathbf{L}_S)$, which is equivalent to squared volume spanned by $g(i)$, where $i \in S$.

The negative-correlated property makes DPP a good choice for the sake of diversities in the context of machine learning [KT13]. In particular, it is useful in the field of robotics such as bandit optimization [DK16]. Even though the truncation of DPP to k-DPP is not DPP, it still has many practical application such as the k-volume sampling problem [VW06], which is related to the low-rank approximation problem. Efficiently generating approximate samples of a k-DPP therefore is studied recently [Sra15; Kan13; RK15].

In practice, lots of probability distribution are hard to sample from, but one can sample the target distribution with the help of a proposal distribution, which is easy to sample from. There are three main approaches to accomplish the task [FJ03]:

1. Rejection Sampling
2. Importance Sampling
3. Monte Carlo Markov Chain

In the final report, we will follow [GR16] and focus on the Monte Carlo Markov Chain (MCMC) for homogeneous strongly Rayleigh measures. We follow the paper to construct a transition kernel of the Markov chain sampling the distribution rapidly-mixing. As a byproduct, one can generate k-DPP under the same framework.

The material of this final report are mainly following [GR16] and based on [TV04; FM92].

1.1 Preliminaries

In this section, we give a high level overview of preliminaries knowledge. For interesting readers about Markov Chain and Mixing Time, please see [Lee19; PW06; MT06] for more details, and about Strongly Rayleigh Distribution, please see [Gha15; Gha19] for more details.

1.1.1 Markov Chain and Mixing Time

Consider a finite state space Ω and a transition kernel $\mathbf{P} : \Omega \times \Omega \to [0, 1]$ such that for every $x \in \Omega$, $\sum_{y\in\Omega} \mathbf{P}(x, y) = 1$.The Markov Chain corresponding to the kernel $\mathbf{P}$ is the sequence of random variables $\{X_0, X_1, X_2, ...\}$ such that for every $t \geq 0$, we have $\mathbb{P}[X_{t+1} = y|X_t = x] = \mathbf{P}(x, y)$.

The Markov chain described by $\mathbf{P}$ is said to be irreducible if for every x, $y \in \Omega$, there is some t such that $\mathbf{P}^t(x,y) > 0$; in words, there is always some way to reach any state form any other. The chain is aperiodic if for every $x, y \in \Omega$, $gcd(\{t : \mathbf{P}^t(x,y) > 0\}) = 1$, and is reversible with respect to the measure π if for every $x, y \in \Omega$, we have $\pi(x)\mathbf{P}(x,y) = \pi(y)\mathbf{P}(y,x)$, which is called detailed balanced condition. Note that reversible chains correspond precisely to random walks on (weighted) undirected graphs.

Theorem 1.1 (Fundamental Theorem of Markov Chain). *If a chain* $\mathbf{P}$ *is irreducible and aperiodic, then there is a unique probability measure* $\pi : \Omega \to [0,1]$*such that for every* $x, y \in \Omega$*, we have:*

$$\mathbf{P}^t(x,y) \to \pi(y) \textit{ as } t \to \infty.$$

In other words, the Markov chain forgets where it started and converges to a unique limiting stationary measure π*.*

Now we know that for any irreducible, aperiodic Markov chain on a finite state space Ω converges to a unique stationary measure π. We are not only concerned with the convergence, but the rate of convergence. First, we introduce the metric total variance distance of any two probabilities measures π, ν as followed:

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

Definition 1.2 (Mixing time). *For any* $t \geq 0$ *and* $x \in \Omega$*, define* $\Delta_x(t) = d_{TV}(\mathbf{P}^t(x,\cdot), \pi)$*, and we set* $\Delta(t) = \max_{x\in\Omega} \Delta_x(t)$*, the total variance mixing time is defined as follows:*

$$\tau(\epsilon) := \min\{t : \Delta(t) \leq \epsilon\},$$

where $\mathbf{P}^t(x,\cdot)$ *is the distribution of the chain started at* x *at time* t*.*

Let's fix an inner product in which the matrix $\mathbf{P}$ is self-adjoint:

$$\langle f, g\rangle_{L^2(\pi)} = \sum_{x\in\Omega} \pi(x)f(x)g(x),$$

and let $\|f\|_{L^2(\pi)} = \sqrt{\langle f, f\rangle_{L^2(\pi)}}$. For a function $f \in L^2(\pi)$, the Dirchlet form $\mathcal{E}_\pi(f,f)$ is defined as:

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

and the Variance of f is :

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

Poincaré's inequality gives us $\mathcal{E}_\pi(f,f) \geq \lambda Var_\pi(f)$, which holds uniformly over all $f : \Omega \to \mathbf{R}$ with $\lambda > 0$. It is well known the lower bound on λ translates the upper bound on mixing time. Hence, the Poincaré's constant is defined as follows:

Definition 1.3 (Poincaré's Constant). *The Poincaré's Constant of the chain is defined as follow:*

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

where the infimum is over all functions with nonzero variance.

We call a Markov chain is a lazy chain if it stays in each state S with probability $\frac{1}{2}$. It is well known that for a lazy chain, the second largest eigen-value in absolute value of kernel $\mathbf{P}$ is $1 - \lambda$.

Fact 1.4. *The Poincaré's constant of any reversible chain with two states $\Omega = 0, 1$ and $\mathbf{P}(0,1) = c \cdot \pi$ is c.*

We introduce a classic theorem bounded the mixing time of a chain with it's Poincaré's constant λ as follows:

Theorem 1.5 ([DS91], Prop3)**.** *For any reversible irreducible lazy Markov chain $(\Omega, \mathbf{P}, \pi)$ with Poincaré's constant λ, $\epsilon > 0$ and any state $x \in \Omega$,*

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

1.1.2 Natural Monte Carlo Markov Chain (MCMC)

MCMC is based on the fact that for a reversible, and irreducible Markov Chain, the $\pi(x)$ will converge to a fixed point, which is our target distribution, regardless of the initial distribution. Different choices of transition kernel lead to different algorithms. We introduce the most famous one called Metropolis Hasting Algorithm in Algorithm 1, and most of other MCMC algorithms are just variants of this "natural" MCMC algorithm.

Algorithm 1 Metropolis Hasting Algorithm.

```
Choose a symmetric stochastic proposal matrix $\mathbf{Q}$
Initilize $x_0$
for  $i = 0 \to N-1$ do
    Sample $u \sim uniform$
    Sample $x^* \sim \mathbf{Q}(x^*|x)$
    $\mathbf{P}(x_i, x^*) = \min\left\{1, \frac{p(x^*)\mathbf{Q}(x|x^*)}{p(x)\mathbf{Q}(x^*|x)}\right\}$
    if $u < \mathbf{P}(x_i, x^*)$ then
        $x_{i+1} \leftarrow x^*$
    else
        $x_{i+1} \leftarrow x_i$
    end if
end for
```

We can calculate the transition kernel $\mathbf{T}$ of the chain:

If $x_i \neq x$:

$$\mathbf{T}(x_i, x) = \mathbb{P}\left[x_{i+1} = x | x_i = x_i\right] = \mathbb{P}\left[x | x_i\right] \mathbb{P}\left[accept x | x_i\right] = \mathbf{Q}(x_i, x) \min\left\{1, \frac{\tilde{\pi}(x)}{\tilde{\pi}(x_i)}\right\},$$

else:

$$\mathbf{T}(x_i, x_i) = 1 - \sum_{x \neq x_i} \mathbf{T}(x_i, x).$$

The correctness of the Algorithm 1 follows the fact that the above chain has stationary distribution and is irreducible and aperiodic.

1.1.3 Strongly Rayleigh Distribution (SR)

We say a polynomial $p \in \mathbb{R}[z_1, \dots, z_n]$ is *homogeneous* if every monomial of p has the same degree.

Definition 1.6 (Stable Polynomials). *For $\Omega \subseteq \mathbb{C}^d$ we say a polynomial $p \in \mathbb{C}[z_1, \dots, z_d]$ is Ω-stable if no roots of p lies in Ω. In particular, we say p is $\mathcal{H}$-stable, or stable if no roots of p lies in the upper-half complex plane, i.e., $p(z) \neq 0$ for all points z in*

$$\mathcal{H}^n := \{v : \mathrm{Im}(v) > 0, \forall 1 \leq i \leq n\}.$$

We say p is real-stable if p is $\mathcal{H}$-stable and all of the coefficients of p are real.

Note that a real stable polynomial is not necessarily homogeneous.

Definition 1.7 (Strongly Rayleigh distributions). *Let $\mu : 2^{[n]} \to \mathbb{R}_+$, and $\sum_{S \subseteq [n]} \mu(S) = 1$, be a probability distribution. If the assigned multi-affine polynomial:*

$$g_\mu(z) := \sum_{S \subseteq [n]} \mu(S) \prod_{j \in S} z_j.$$

to μ is a real stable polynomial, we call μ is a strongly Rayleigh distribution.

Example 1.8 (Bernoulli). *Given two i.i.d. Bernoulli distributions $\mathcal{B}_1$, $\mathcal{B}_2$ with prior probabilities p_1, p_2 respectively. It is easy to see the generating function:*

$$g_\mu = p_1 p_2 z^1 z^2 + p_1(1-p_2)z^1 + p_2(1-p_1)z^2 + (1-p_1)(1-p_2),$$

which is real-stable and hence it is a strongly Rayleigh distribution.

Example 1.9 (Determinantal Point Process). *A determinantal point process (DPP) on a set of elements $[n]$ is a probability distribution on the set $2^{[n]}$ identified by a positive semi-definite ensemble matrix $\mathbf{L} \in \mathbb{R}^{n \times n}$ where for any $S \subseteq [n]$ we have:*

$$\mathbb{P}[S] \propto det(\mathbf{L}_S),$$

where $\mathbf{L}_S$ is the principal submatrix of $\mathbf{L}$ indexed by the elements of S.
To see it's real-stable we can let $\mathbf{L} = \mathbf{V}\mathbf{V}^T$, then the generating function becomes:

$$det\left(\mathbf{I} + \sum_{i \in S} z_i v_i v_i^T\right) = \sum_{S \subseteq [n]} det_{|S|}\left(\sum_{i \in S}\left(z_i v_i v_i^T\right)\right),$$

then factor z out, we can calculate and find DPPs is real-stable and hence it's a strongly Rayleigh distribution.

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, Brandén, and Liggett showed that for any strongly Rayleigh distribution μ, and any integer k, μ_k is also strongly Rayleigh distribution [BL09]. This is a nontrivial property because even if μ_k is not an independent probability distribution, we can still argue that μ_k has strongly Rayleigh distribution properties if we know μ is strongly Rayleigh distribution (Ex: Bernoulli distribution and it's truncation).

Example 1.10 (k-DPP). *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. Since strongly Rayleigh distributions are closed under truncation, even though any truncated k-DPP is not a DPP, any k-DPP is a strongly Rayleigh distribution.*

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,

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

Negatively associated is a generalization of negatively correlated, and we have the following fact:

Fact 1.11. *Any strongly Rayleigh probability distribution is negatively associated.*

1.2 Main contributions of the paper

Consider the following Markov Chain $\mathcal{M}_\mu$:
In a state $S \subseteq [n]$ and choose an element $i \in [n]$ and $i \in S$, and an element $j \in [n]$ but $j \notin S$ uniformly and independent respectively, and $T = S \setminus \{i\} \cup \{j\}$,

$$\mathbf{P}(S, \cdot) = \begin{cases} \frac{1}{2} \min\left\{1, \frac{\mu(T)}{\mu(S)}\right\}, & \text{move to T if T} \in supp\{\mu\} \\ \text{stay in S,} & \text{otherwise} \end{cases}.$$

It is easy to see $\mathcal{M}_\mu$ has the following properties:

1. reversible
2. irreducible
3. lazy chain
4. $\mu(\cdot)$ is the stationary distribution of the chain.

The main contribution in [GR16] is to prove the following theorem, which means the above chain is rapidly-mixing.

Theorem 1.12 (Main Theorem)**.** *For any strongly Rayleigh k-homogeneous probability distribution $\mu : 2^{[n]} \to \mathbb{R}_+$, $S \in supp\{\mu\}$ and $\epsilon > 0$,*

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

where

$$C_\mu := \min_{S,T \in supp\{\mu\}} \max\{\mathbf{P}_\mu(S,T), \mathbf{P}_\mu(T,S)\}$$

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

Theorem 1.12 implied that we can generate ϵ-approximate k-DDP in time $\frac{1}{C_\mu} \cdot \log\left(\frac{1}{\epsilon \cdot \mu(S)}\right)$.

1.3 Overview of the proof

To prove the main theorem, we lower bound the spectral gap, which is the Poincaré's constant of the chain $\mathcal{M}_\mu$. The proof in the paper can be regarded as a weighted version of [FM92] but work with more advanced idea of [TV04]. Before we get into the details of the proof, we first introduce two main techniques in the proof:

1.3.1 Closure property of SR measures under conditioning

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

$$\mu|i := \{\mu|Y_i = 1\},$$

to denote the conditional measure on sets that contain i and

$$\mu|i := \{\mu|Y_i = 0\},$$

to denote the conditional measure on sets that do not contain i. Borcea, Branden and Ligett showed that strongly Rayleigh distributions are closed under conditioning.

Theorem 1.13. *For any strongly Rayleigh distribution $\mu : 2^{[n]} \to \mathbb{R}_+$ and any $1 \leq i \leq n$, $\mu|i$, $\mu|i$ are strongly Rayleigh.*

In other words, the statement of Theorem 1.13 holds for any homogeneous probability distribution $\mu : 2^{[n]} \to \mathbb{R}_+$ where μ and all of its conditional measures are negatively associated.

1.3.2 Decomposable Markov Chains

This section we describe the decomposable Markov chain technique due to Jerrum, Son, Tetali and Vigoda [TV04]. Consider an ergodic Markov chain $(\Omega, \mathbf{P}, \pi)$ is reversible, in words, satisfies the detailed balance condition.

Let $\Omega_0 \cup \Omega_1$ be a decomposition of the state space of the chain into two disjoint sets. We use $[2] := \{0, 1\}$.

For each $i \in \{0, 1\}$, $\bar{\pi} : [2] \to [0, 1]$ is defined by:

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

and define $\bar{\mathbf{P}} : [2] \times [2] \to [0, 1]$ by

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

The Markov chain on state space $[2]$ with transition probabilities $\bar{\mathbf{P}}$ is called a projection chain induced by the partition $\{\Omega_i\}$. Since the original chain is reversible, so is the projection chain. Let $\bar{\lambda}$ be the Poincaré's constant of the projection chain.

We can also define a restriction Markov chain on each i as follows. For each $i \in \{0, 1\}$, $\mathbf{P}_i : \Omega_i \times \Omega_i \to [0, 1]$ is defined by:

$$\mathbf{P}_i(x, y) = \begin{cases} \mathbf{P}(x, y), & \text{if } x \neq y, \\ 1 - \mathbf{P}(x, x) - \sum_{z \notin \Omega_i} \mathbf{P}(x, z), & \text{if } x = y \end{cases}.$$

In other words, for any transition from x to a state outside of i, we remain in x. Observe that the restriction chain is reversible and so its stationary distribution, the probability of x is proportional to $\pi(x)$. Let λ_i be the Poincaré's constant of the chain $(i, \mathbf{P}_i, \cdot)$. The following is the main result of [TV04].

Theorem 1.14 ([TV04], Cor 3)**.** *If for any distinct i, $j \in \{0, 1\}$, and any $x \in i$,*

$$\bar{\mathbf{P}}(i, j) = \sum_{y \in \Omega_j} \mathbf{P}(x, y), \tag{1.1}$$

then the Poincaré's constant of $(\Omega, \mathbf{P}, \pi)$ is at least $\min\{\bar{\lambda}, \lambda_0, \lambda_1\}$.

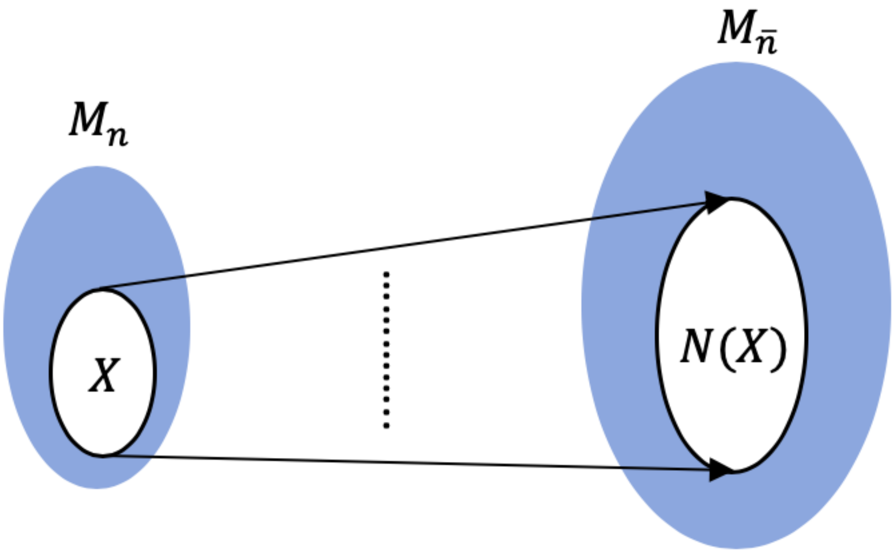

Figure 1.1: Break $\mathbf{M}$ into $\mathbf{M}_n$ and $\mathbf{M}_{\bar{\mathbf{n}}}$.

1.3.3 Overview of the proof

In the section, we will give a big picture of the proof:

To prove our main theorem, we fix a strongly Releigh distribution μ, and then prove Theorem 1.19 by induction on $|supp\{\mu\}|$. Figure 1.1 illustrates the general framework of the proof with advantage of using SR measures, which using induction to bound the mixing time of $\mathbf{M}_n$ and $\mathbf{M}_{\bar{n}}$ respectively. In the induction step, we use decomposable Markov chains techniques (see subsection 1.3.2). However, even though the restriction part can be directly got from the closure property under conditioning, the projection chain does not satisfy (1.1), so we construct a new transition kernel satisfying the following properties:

$$\mu(x)\hat{\mathbf{P}}(x,y) = \mu(y)\hat{\mathbf{P}}(y,x), \tag{1.2}$$

$$\hat{\mathbf{P}}(x,y) \leq \mathbf{P}(x,y). \tag{1.3}$$

(1.2) implies $\hat{\mathbf{P}}$ has same distribution μ, while with (1.3) implies $\hat{\lambda}$ lower bounded λ.

We construct $\hat{\mathbf{P}}$ by Lemma 1.15. To prove Lemma 1.15, first we prove that the support graph of the transition probability matrix $\mathbf{P}_\mu$ satisfies Hall's condition by Theorem 1.13. Then we create a bipartite graph $\mathbf{G}$ to construct a function $w_{\{x,y\}}$, which is used to construct $\hat{\mathbf{P}}$.

1.4 Details of the proof

We use the following lemma to inductively prove our Theorem 1.12. Hence before we get into inductively arguments, let's start from proving Lemma 1.15 first and then complete the proof of Theorem 1.12.

Lemma 1.15. *There is a transition probability matrix $\hat{\mathbf{P}} : \Omega \times \Omega \to \mathbb{R}_+$ such that:*
(1) $\hat{\mathbf{P}}$ satisfies (1.3), (1.2).
(2) For any $i \in \{0,1\}$ and states $x,\ y \in i$, $\hat{\mathbf{P}}(x,y) = \mathbf{P}(x,y)$.

(3) The Poincaré's constant of the chain $(\Omega, \hat{\mathbf{P}}, \mu)$ *projected onto* Ω_0, Ω_1 *is at least* $\bar{\hat{\lambda}} \geq C_\mu$,
(4) For any state $S \in supp\{\mu\}$ *and distinct* i, $j \in \{0, 1\}$,

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

1.4.1 Proof of Lemma 1.15

Before we prove Lemma 1.15, we have to prove the following two lemma first. The first lemma says that the support graph of the transition probability matrix $\mathbf{P}_\mu$ satisfies Hall's condition.

First, let us remind Hall's condition:

Fact 1.16. *If a bipartite graph* $\mathbf{G}$ *with bipartitions* X *and* Y *has a perfect matching, then* $|N(S)| \geq |S|$ *for all subsets* $S \subseteq \mathbf{X}$, *where* S *is the subset of vertices* $N(S)$ *is the set of nodes adjacent to nodes in* S.

Lemma 1.17. *For any* $A \subseteq \Omega_1$,

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

where $\mu(\Omega_i) := \sum_{z \in \Omega_i} \mu(z)$ *is the probability in the set* Ω_i *and* $N(A) := \{y \in \Omega \setminus A : \exists x \in A, \mathbf{P}(x, y) > 0\}$.

Proof. Let $R \sim \mu$, g is an indicator whether $n \in R$, and f is an indicator whether there exists $T \in A$ such that $R \supseteq T \setminus \{n\}$.
Since f, g are two increasing functions, by negative association property:

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

and we know LHS is $\frac{\mu(N(A))}{\mu(\Omega_0)}$ and RHS is $\frac{\mu(A)}{\mu(\Omega_1)}$, which completes the proof of above lemma. □

Now, we are prepared to prove the following lemma:

Lemma 1.18. *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* $\mathbf{P}(x, y) > 0$ *and*

$$\sum_{y \in \Omega_1} w_{\{x,y\}} = \frac{\mu(x)}{\mu(\Omega_0)} \quad \forall x \in \Omega_0,$$

$$\sum_{x \in \Omega_0} w_{\{x,y\}} = \frac{\mu(y)}{\mu(\Omega_1)} \quad \forall y \in \Omega_1.$$

Proof. Let $\mathbf{G}$ be a bipartite graph on $\Omega_0 \cup \Omega_1$, where there is an edge between $x \in \Omega_0$ and $y \in \Omega_1$ if $\mathbf{P}(x, y) > 0$. Figure 1.2 shows the graph $\mathbf{G}$ where $c_{s,x} = \frac{\mu(x)}{\mu(\Omega_1)}$ and $c_{y,t} = \frac{\mu(y)}{\mu(\Omega_0)}$ respectively.
To prove Lemma 1.18, it is enough to show that the maximum flow is 1, and by maximum flow-minimum cut theorem, it is equivalently to show that the minimum cut is at least 1:

Let $c(X, Y) = \sum_{x \in X, y \in Y} c_(x, y)$, then we get:

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

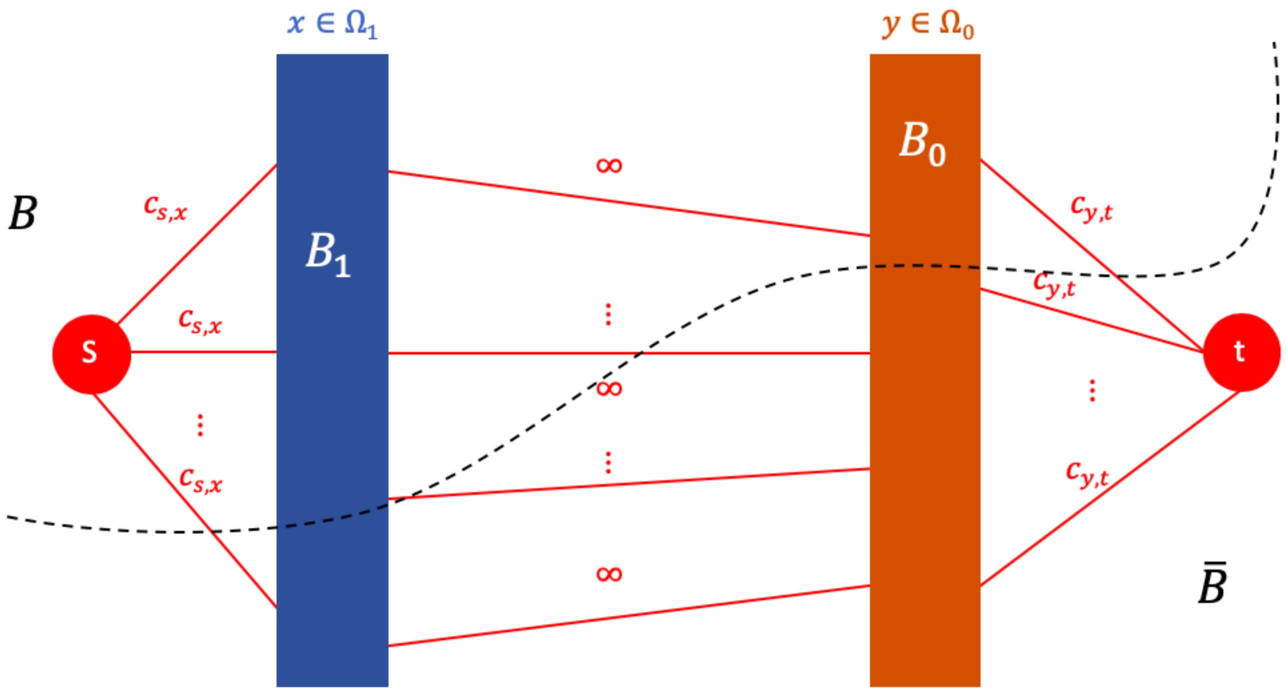

Figure 1.2: **G**: a bipartite graph on $\Omega_0 \cup \Omega_1$ (This looks like the final homework of 421).

the second last inequality follows from Lemma 1.17. and the last inequality follows the following reason:

$$\begin{cases} c(B,\bar{B}) = \infty, & \text{if there is any edge from } B_1 \text{ to } \Omega_0 \setminus B_0 \\ N(B_1) \subseteq B_0, & \text{otherwise} \end{cases},$$

indicated $\mu(N(B_1)) \leq \mu(B_0)$.
Hence, we can construct $w_{\{x,y\}}$ be such a unit flow connecting x and y. □

With the above two lemma at hand, we are ready to prove Lemma 1.15:

Proof. For any i,$j \in \{0,1\}$, $x \in \Omega_i$,$y \in \Omega_j$:

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

(1) For any $i \neq j$,$x \in \Omega_i$, $y \in \Omega_j$:

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

and

$$\begin{aligned}
\hat{\mathbf{P}}(x,y) &= \frac{C_\mu}{\mu(x)}\mu(\Omega_i)\mu(\Omega_j)w_{x,y} \\
&\leq \frac{\max\left(\mathbf{P}(x,y),\mathbf{P}(y,x)\right)}{\mu(x)}\mu(\Omega_i)\mu(\Omega_j)w_{\{x,y\}} \\
&\leq \max\left(\mathbf{P}(x,y),\mathbf{P}(y,x)\right)\frac{\min\left(\mu(x),\mu(y)\right)}{\mu(x)} \\
&\leq \mathbf{P}(x,y) \quad (\because \text{detailed balanced condition}).
\end{aligned}$$

(2) By definition.
(3) By definition of $\hat{\mathbf{P}}$ and Lemma 1.18 for distinct i,$j \in \{0,1\}$:

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

so the Poincaré 's constant of $\bar{\hat{\mathbf{P}}} = C_\mu$ by Fact 1.4.
(4) Fix distinct i,$j \in \{0,1\}$, $z \in \Omega_i$:

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

this completes the proof with (3). □

1.4.2 Inductive argument to complete the proof

Theorem 1.19. *For any k-homogeneous strongly Raleigh distribution $\mu : 2^{[n]} \to \mathbb{R}_+$, the Poincaré's constant λ the chain $\mathcal{M}_\mu = (\Omega_\mu, \mathbf{P}_\mu, \mu)$ is at least:*

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

Proof. We prove by induction on $|supp\{\mu\}|$:
[Base Case] $|supp\{\mu\}| = 1$, clearly.
[Induction Hypothesis] Theorem 1.19 holds for some $|supp\{\mu\}| = n$.
[Induction Step]
Wlog, let $\{n|0 < \mathbb{P}_{s\sim\mu}[n \in S] < 1\}$, and $\Omega_0 = \{S \in supp\{\mu\} : n \notin S\}$, $\Omega_1 = \{S \in supp\{\mu\} : n \in S\}$ both are non-empty.

$(\Omega_0, \mathbf{P}_0, \cdot)$, which is the same as $\mathcal{M}_{\mu|n}$, and $(\Omega_1, \mathbf{P}_1, \cdot)$, which is the same as $\mathcal{M}_{|\bar{n}}$ are both strongly Rayleigh by Theorem 1.13. and clearly $C_{\mu|n}$, $C_{\mu|\bar{n}} \geq C_\mu$, hence I.H. on lower bound λ_0, $\lambda_1 \geq C_\mu$.

Since $\mathbf{P}$ does not satisfy Theorem 1.14, we construct a new $\hat{\mathbf{P}}$ from Lemma 1.15 such that satisfies (1.3) and (1.2), then $\hat{\lambda}_0, \hat{\lambda}_1 \geq \lambda_0, \lambda_1 \geq C_\mu$by Lemma 1.15 (3), and we get $\lambda \geq min\{\bar{\hat{\lambda}}, \lambda_0\}$, hence $\hat{\lambda} \geq C_\mu$, and $\hat{\lambda} \leq \lambda$ (from Lemma 1.15 (3)), this completes the proof. □

Finally, let's complete the proof of the Theorem 1.12:

Proof. It is easy to see Theorem 1.12 followed by Theorem 1.19 and Theorem 1.5, this competes the proof of our main theorem. □

1.5 Conclusion

In the final report, we overview the properties of Markov chains, and Strongly Rayleigh distribution. Then we follow [GR16], which introduce a transition kernel and give it a proof. We will introduce the main byproduct from [GR16] as the conclusion of the final report.

As a byproduct of Theorem 1.12, for any k-DPP μ, $\mathcal{M}_\mu$ mixes rapidly to the stationary distribution.

Corollary 1.20. *For any k-DPP $\mu : 2^{[n]} \to \mathbb{R}_+$, $S \in supp\{\mu\}$ 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 corollary to generate ϵ-approximate samples of the k-DPP.

Theorem 1.21. *Given an ensemble matrix $\mathbf{L}$ of a k-DPP μ there is an algorithm for any $\epsilon > 0$ generates an ϵ-approximate sample of μ in time $poly(k)O(nlog(\frac{n}{\epsilon}))$.*

Algorithm 2 Greedy Algorithm for Selecting the Start State of $\mathcal{M}_\mu$.

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

To prove the above theorem, the author of the paper generate a set $S \in supp\{\mu\}$ such that $\mu(S)$ is bounded away from zero by an exponentially small function of n, k by Algorithm 2, which returns a set S such that

$$det(\mathbf{L}_S) \geq \frac{1}{k! \cdot |supp\{\mu\}|} \geq \frac{1}{k! \cdot \binom{n}{k}} \geq n^{-k},$$

in time $O(n)$poly(k).

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.

To sum up, in stead of trying to directly sample from DPP, the authors put DPP under the SR measures framework and take advantage of closure properties of SR measures. Theorem 1.21 shows an ϵ-approximate sample can be generated by running the chain for $\mathrm{O}\big(nk^2 \log(\frac{n}{\epsilon})\big)$ steps. The total running time is $\mathrm{O}\big(nk^4 \log(\frac{n}{\epsilon})\big)$. In the context of low rank approximation, we want to find $\mathbf{A}_k$ of rank at most k minimize $\|\mathbf{A} - \mathbf{A}_k\|_F$. Sampling $\mathrm{O}\big(\frac{k}{\epsilon}\big)$-DPP with kernel $\mathbf{A}\mathbf{A}^T$ gives $(1+\epsilon)$-approximate to $\|\mathbf{A} - \mathbf{A}_k\|_F$. Hence, Theorem 1.21 gives us a $(1-\delta)$-approximation in $\tilde{O}(n)\cdot$poly$(k, \frac{1}{\delta})$ for low rank approximation [Rez16].

1.6 Acknowledgement

I would like to express my special thanks of gratitude to Shayan Oveis Gharan, who is one of the authors of the paper and is my algorithms and graph theory teacher as well. I would also like to thank Archis Ghate, who is my random processes and optimization teacher. Without the above two professors, I would not have a chance to explore the field in optimization algorithms design.

References

[BL09] J. B. P. Branden, and T. M. Liggett. “Negative dependence and the geometry of polynomials”. In: *Journal of American Mathematical Society* 22 (2009), pp. 521–567 (cit. on p. 1-4).

[DK16] T. K. A. Deshpande, and P. Kohli. “Batched Gaussian Process Bandit Optimization via Determinantal Point Processes”. In: *Neural Information Processing Systems (NIPS)* (2016) (cit. on p. 1-1).

[DS91] P. Diaconis and D. Stroock. “Geometric bounds for eigenvalues of markov chains”. In: *The Annals of Applied Probability* (1991), pp. 36–61 (cit. on p. 1-3).

[FJ03] C. A. N. de Freitas, Arnaud Doucet and M. I. Jordan. “An Introduction to MCMC for Machine Learning”. In: *Machine Learning* (2003), pp. 5–43 (cit. on p. 1-1).

[FM92] T. Feder and M. Mihail. “Balanced matroids”. In: *Proceedings of the twenty- fourth annual ACM symposium on Theory of Computing* (1992), pp. 26–38 (cit. on pp. 1-1, 1-5).

[Gha15] S. O. Gharan. *Real Stable and Hyperbolic Polynomials*. Apr. 2015. URL: https://homes.cs.washington.edu/~shayan/courses/cse599/adv-approx-10.pdf (cit. on p. 1-1).

[Gha19] S. O. Gharan. *Real Stable Polynomials, Strongly Rayleigh Distributions, and Applications, Part II*. Jan. 2019. URL: https://simons.berkeley.edu/talks/tba-16 (cit. on p. 1-1).

[GR16] N. A. S. O. Gharan, and A. Rezaei. “Monte Carlo Markov Chain Algorithms for Sampling Strongly Rayleigh Distributions and Determinantal Point Processes”. In: *Proceedings of Machine Learning Research* 49 (2016), pp. 103–105 (cit. on pp. 1-1, 1-5, 1-11).

[Kan13] B. Kang. “Fast determinantal point process sampling with application to clustering”. In: *NIPS* (2013), pp. 2319–2327 (cit. on p. 1-1).

[KT13] A. Kulesza and B. Taskar. *Determinantal point processes for machine learning*. 2013 (cit. on p. 1-1).

[Lee19] J. R. Lee. *Markov chains and mixing times*. May 2019. URL: https://homes.cs.washington.edu/~jrl/teaching/cse525sp19/notes/lecture16.pdf (cit. on p. 1-1).

[MT06] R. Montenegro and P. Tetali. “Mathematical aspects of mixing times in Markov chains”. In: *Found. Trends Theor. Comput. Sci.* 1.3 (2006), pp. 237–354 (cit. on p. 1-1).

[PW06] D. A. L. Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2006 (cit. on p. 1-1).

[Rez16] A. Rezaei. *Monte Carlo Markov Chain Algorithms for Sampling Strongly Rayleigh distributions*. 2016. URL: https://www.youtube.com/watch?v=RHW5KxC5HCs (cit. on p. 1-11).

[RK15] P. Rebeschini and A. Karbasi. “Fast mixing for discrete point processes”. In: *COLT* (2015), pp. 1480–1500 (cit. on p. 1-1).

[Sra15] C. L. S. J. S. Sra. “Efficient sampling for k-determinantal point processes”. In: (2015) (cit. on p. 1-1).

[TV04] J.-B. S. P. Tetali, and E. Vigoda. “Elementary bounds on poincaré and log-sobolev constants for decomposable markov chains”. In: *Annals of Applied Probability* (2004), pp. 1741–1765 (cit. on pp. 1-1, 1-5, 1-6).

[VW06] A. D. L. R. S. Vempala and G. Wang. “Matrix approximation and projective clustering via volume sampling”. In: *SODA* (2006), pp. 1117–1126 (cit. on p. 1-1).