

A Research on Card Shuffling and MCMC Paradigm

Stochastic Process (2016) Course Project

CHEN, You; WANG, Shu; YANG, Ze; FENG, Biaojian

June 23, 2016

1 Introduction

This project is inspired by course 6.896, 2011 Spring in MIT.[1] This project aims to study the theory and implementation of *Markov Chain Monte Carlo* algorithm. We discuss general MCMC paradigm through the famous card-shuffling problem arisen in combinatorics. We will model the card shuffle by markov chain, and further discuss its stationary distribution, variation distance and mixing time. Numerical computations will also be performed to validate the theory.

2 Illustration of the Problem and Related Theories

2.1 The MCMC Paradigm

Consider probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We define a real, positive *weight function* $w : \Omega \rightarrow \mathbb{R}^+$. The goal is to sample an element $x \in \Omega$, with probability distribution $\pi(x) = \mathbb{P}(X = x)$ defined as

$$\pi(x) := \frac{w(x)}{Z_w}; \quad Z_w := \sum_{x \in \Omega} w(x)$$

Where Z_w is called the partition function of weight function w in Ω .

In MCMC paradigm, we establish a general iterative solution approach to this sampleability problem, where we construct a sequence of RV $\mathcal{X} = \{X_t\}$ that converges to the target probability, in the sense that

$$\lim_{t \rightarrow \infty} \mathbb{P}(X_t = y | X_0 = x) = \pi(y)$$

Example 1. We now illustrate the **Card Shuffling Problem** as an example of MCMC paradigm. Namely, we set

- The probability space $\Omega := \{\text{All **Permutations** of the deck.}\}$, $x \in \Omega$ is a permutation of the deck. In a 3-cards deck [AS, KH, QC]¹, the probability space has cardinality $|\Omega^3| = 3! = 6$, namely $\Omega^3 = \{\text{AKQ, AQK, KAQ, KQA, QAK, QKA}\}$.
- $w(x) = 1$ for every permutation $x \in \Omega$. Hence $\pi(x) = \frac{1}{|\Omega|}$. That is, by assigning equal probability mass to each permutation, we want to sample a uniformly distributed $x \sim U(\Omega)$.
- Denote X_0 the initial configuration of deck, which is arbitrary; and X_t is the state of deck after t^{th} shuffle. Note that the state of deck at t only depends on that of $t - 1$ and how it is shuffled, we conclude that $\{X_t\}$ is a **markov chain**.
- We want to consider a random variable instead of permutations. Thus we define $X : \Omega \rightarrow \mathbb{N}^+$ be the **lexicographic order** of the deck permutation. For example on deck {AKQ} we can define a partial order as $Q < K < A$. Then: $X(\{\text{QKA}\}) = 1, X(\{\text{KQA}\}) = 2, X(\{\text{AKQ}\}) = 6$, etc. Moreover, $X \sim \text{Uniform}(\{1, 2, \dots, n!\})$

¹We use two letters to represent a card, for example, AS means spade A.

2.2 Markov Chains as Graphs

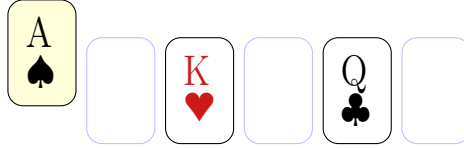
Definition 1. (*Transition Matrix*) $\forall x, y \in \Omega$, we denote one-step transition probability, t -step transition probability and corresponding transition matrix as $P(x, y), P^{[t]}(x, y), \mathbf{P}, \mathbf{P}^{[t]}$.² We refer to matrices like \mathbf{P} as *stochastic matrix* if it satisfies

- *Non-negativity*: $P(x, y) \geq 0, \forall x, y \in \Omega$. And
- *Stochasticity*: $\sum_{y \in \Omega} P(x, y) = 1$, for any fixed $x \in \Omega$.

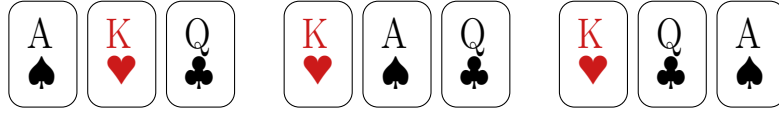
We can establish the combinatorial flavored theory of Markov chains by viewing them as graphs. In particular, for a given markov chain $\{X_t\}$, we define a weighed, directed graph $G(\{X_t\}) = (V, E)$ as follows

- $V = \Omega$, i.e. the vertices are the states in Ω .
- $e_{x \rightarrow y} \in E$ if $P(x, y) > 0$, with weight $w(e_{x \rightarrow y}) = P(x, y)$. That is, \mathbf{P} is the *adjacency matrix* of graph G .

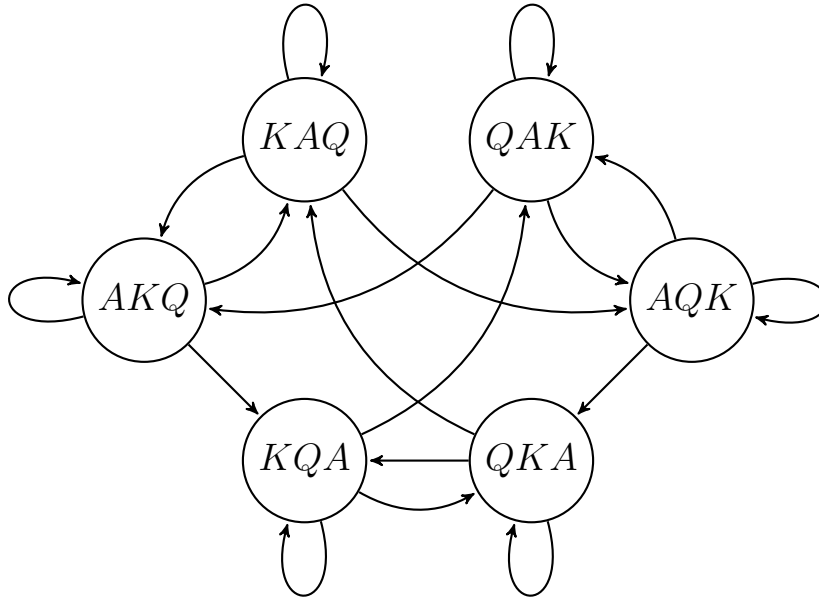
Example 2. (*3-cards Deck Top-in Shuffle*) We consider a 3-cards deck [AS, KH, QC] and shuffled by top-in method. That is, we take one card from the top and insert it at a random position in the deck. If initially, $X_0 = (AS, KH, QC)$, we have



one out of three slots is picked to insert AS with equal probability, hence there are three outcomes w.p. $\frac{1}{3}$



And $G(\{X_t\}) = (V, E)$:



²Slightly deviates from the notations in textbook [5], because the x -es are not indexed by integers.

The graph above represents the markov chain associated with top-in shuffle. All edges have weight $\frac{1}{3}$, and its adjacency matrix gives the one-step transition matrix of this markov process.

Definition 2. (*Undirected Graph*) Graph of markov chain $G(\mathcal{X})$ is called undirected if $P(x, y) > 0 \iff P(y, x) > 0$. The exact value of weight doesn't matter, we only care about positivity.

2.3 Basic Properties

Definition 3. (*Irreducible*) \mathcal{X} is called irreducible if (TFAE)

- $\forall x, y \in \Omega, \exists t \in \mathbb{N}$ such that $P^{[t]}(x, y) > 0$.
- The graph $G(\mathcal{X})$ is strongly connected.

Definition 4. (*Aperiodic, Period*) \mathcal{X} is called aperiodic if $\forall x, y \in \Omega$, we have

$$\gcd\{t : P^{[t]}(x, y) > 0\} = 1$$

For any vertex $x \in \Omega$, define the period of x as $\gcd\{t : P^{[t]}(x, x) > 0\}$.

Lemma 1. If \mathcal{X} is irreducible, the period of every $x \in \Omega$ is the same. Further, if $G(\mathcal{X})$ is undirected, the period is either 1 or 2.

Lemma 2. If \mathcal{X} is irreducible and $G(\mathcal{X})$ is undirected, then \mathcal{X} is aperiodic $\iff G(\mathbf{X})$ is *non-bipartite*. I.e. there is **Not** a partition $V = V \cup V'$ ($V \cap V' = \emptyset$), such that every edge connects a vertex in V to one in V' .

Proof. (\Rightarrow) it suffices to show its converse-negative, i.e. bipartite \Rightarrow periodic. Suppose $G(\mathcal{X})$ is bipartite, pick $x, y \in V$ belongs to same side, then every path from x to y requires even # of steps. So $\gcd\{t : P^{[t]}(x, y) > 0\} \geq 2 > 1$.

(\Leftarrow) Suppose $G(\mathcal{X})$ is bipartite $\iff \exists C \subset G(\mathcal{X})$ is an odd cycle. Pick any $x, y \in \Omega$ and $c \in C$, we can always go through $x \rightarrow c \rightarrow y$ because the chain is irreducible. Moreover, denote $l(x, y)$ be the path from x to y , we can go either

- $l(x, c) \rightarrow C \rightarrow l(c, y)$. $t = |l(x, c)| + |C| + |l(c, y)|$
- $l(x, c) \rightarrow$ go on C until the $\frac{|C|-1}{2}$ vertex on C , then go back to $c \rightarrow l(c, y)$. $t' = |l(x, c)| + |C| - 1 + |l(c, y)|$

We see $\gcd(t, t')$ is already 1, because they are consecutive integers. So it is aperiodic. \square

Lemma 3. If \mathcal{X} is irreducible, then \mathcal{X} is aperiodic \iff there exists some t such that $P^{[t]}(x, y) > 0$ for all $x, y \in \Omega$.

Lemma 4. If \mathcal{X} is irreducible and has at least one self-loop, i.e. $P(x, x) > 0$ for some x ; then \mathcal{X} is aperiodic.

Proof. Suppose this loop is on z . From x to y we can either

- $l(x, z) \rightarrow l(z, y)$. $t = |l(x, z)| + |l(z, y)|$
- $l(x, z) \rightarrow \text{loop}(z) \rightarrow l(z, y)$. $t' = |l(x, z)| + 1 + |l(z, y)|$

Same argument as lemma 2. \square

2.4 Mixing of Markov Chains

Definition 5. (*Stationary Distribution*) Let $\mathcal{X} = \{X_t\}$ be a Markov chain. A probability distribution $\pi(\cdot) : \Omega \rightarrow \mathbb{R}$ over Ω is a stationary distribution of \mathcal{X} if

$$\pi P = \pi$$

Where π is $1 \times |\Omega|$ vector that stacks the values of $\pi(\cdot)$ together, i.e. $\pi = (\pi(x))_{x \in \Omega}$.

Theorem 1. (*Ergodic Thm.*) If a Markov chain \mathcal{X} is **Irreducible** and **Aperiodic**, then it has a unique stationary distribution $\pi(\cdot)$. In particular

- π is the unique left eigenvector of P corresponding to eigenvalue $\lambda = 1$.
- $\lim_{t \rightarrow \infty} P^{[t]}(x, y) = \pi(y), \forall x, y \in \Omega$.

Definition 6. (*Reversibility*) Let $\mathcal{X} = \{X_t\}$ be a Markov chain. And $\pi(\cdot) : \Omega \rightarrow \mathbb{R}$ is a probability distribution over Ω . \mathcal{X} is called reversible with respect to π if, for any $x, y \in \Omega$

$$\pi(x)P(x, y) = \pi(y)P(y, x)$$

Remark.(i) Intuitively, a markov chain is reversible w.r.t π means that the probability mass that flows from node $x \rightarrow y$ is same as that from $y \rightarrow x$.

Remark.(ii) If P is symmetric, i.e. $P(x, y) = P(y, x)$, then \mathcal{X} is trivially reversible with respect to uniform distribution on Ω , $\pi(\cdot) \sim U(\Omega)$.

Lemma 5. If \mathcal{X} is reversible wrt. π , then π is a stationary distribution of \mathcal{X} .

Proof. For any fixed $y \in \Omega$:

$$\sum_{x \in \Omega} \pi(x)P(x, y) = \sum_{x \in \Omega} \pi(y)P(y, x) = \pi(y)$$

Implies $\pi P = \pi$. □

Definition 7. (*Ergodic Flow*) The ergodic flow between nodes x and y wrt. a probability distribution π is defined to be the amount of probability mass flowing between x, y :

$$Q(x, y) := \pi(x)P(x, y) = \pi(y)P(y, x)$$

Enables us to write transition probability of reversible markov chain.

2.5 Mixing Time and Rate of Convergence

We hope to construct a metric to capture how statistically “close” two distributional objects are from each other.

Definition 8. (*Total Variation Norm*) For two probability mass function π, μ , define total variation as

$$\|\mu - \pi\|_{TV} := \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \pi(x)|$$

In particular, if $\mu(\cdot) = P^{[t]}(x_0, \cdot)$ is conditional distribution of deck $\{X_t\}$, given initial deck x_0 . $\pi(\cdot) = \frac{1}{n!}$, then we define

$$\Delta(t) := \left\| P^{[t]} - \pi \right\|_{TV} = \frac{1}{2} \sum_{x \in \Omega} \left| P^{[t]}(x_0, x) - \frac{1}{n!} \right|$$

We may apply Monte Carlo simulation method, draw large a sample from $P^{[t]}$ (i.e. the deck after t shuffles), then the empirical distribution of sample appoximates $P^{[t]}$. Then, $\Delta(t)$, as a function of t , captures the rate by which the Markov chain converges to uniform $\pi(\cdot)$. Moreover, we can set a **threshold** to $\Delta(t)$ and calculate time needed for the **deviation of X_t from the limit is bounded by this threshold**.

Definition 9. (*Mixing Time*) Set the threshold as constant s , the time needed is called Mixing Time of Markov chain associated with threshold s , given by

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

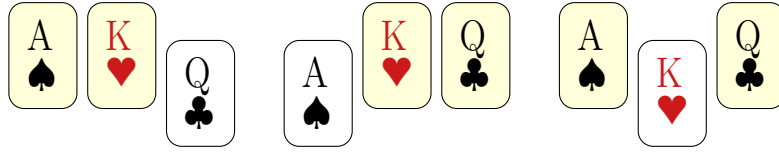
Equipped with the metrics and definitions above, we may able to consider the rate in which difference suffling methods (markov chain) converges to stationary distribution, i.e. the uniform distrobution.

3 Types of Shuffles

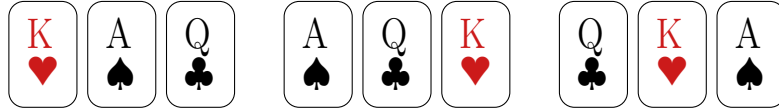
3.1 Random Transposition

In the method of random transposition, in each iteration, we pick uniformly at random two cards and switch them.

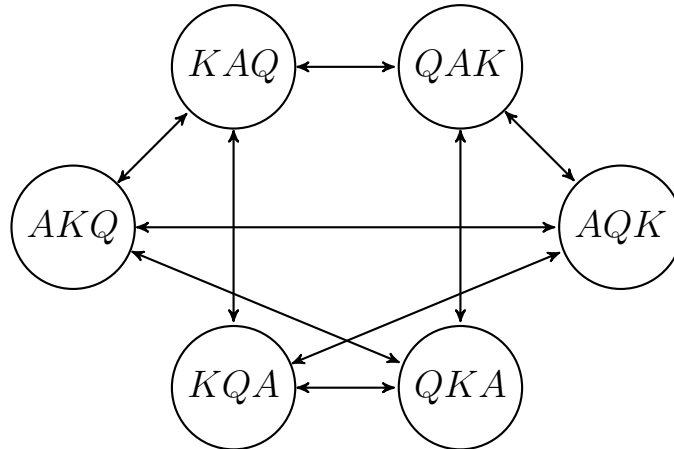
Example 3. (*3-cards Deck Random Transposition*) We still consider our 3-cards deck [AS, KH, QC] and shuffled Random Transposition. Initially $X_0 = (AS, KH, QC)$, the first transposition has $\binom{3}{2} = 3$ possible cases,



And after the first transposition, the deck becomes



$G(\mathcal{X})$ is undirect, given by

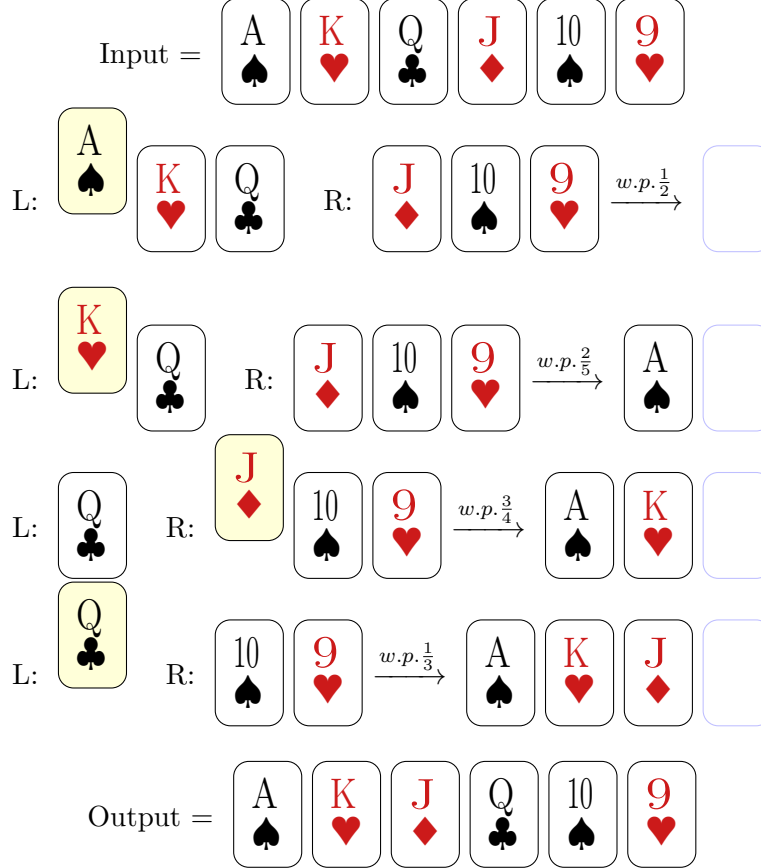


3.2 Riffle Shuffle

The name Riffle Shuffle refers to the following procedures:

1. Split the deck into two parts in each hand.
2. Drop the cards one after another to merge into a new deck. The next card is dropped from left hand with probability $\frac{\# \text{ of cards in left hand}}{\# \text{ of cards remaining}}$.

Example 4. (*Riffle Shuffle*) One complete riffle shuffle process is illustrated as below:



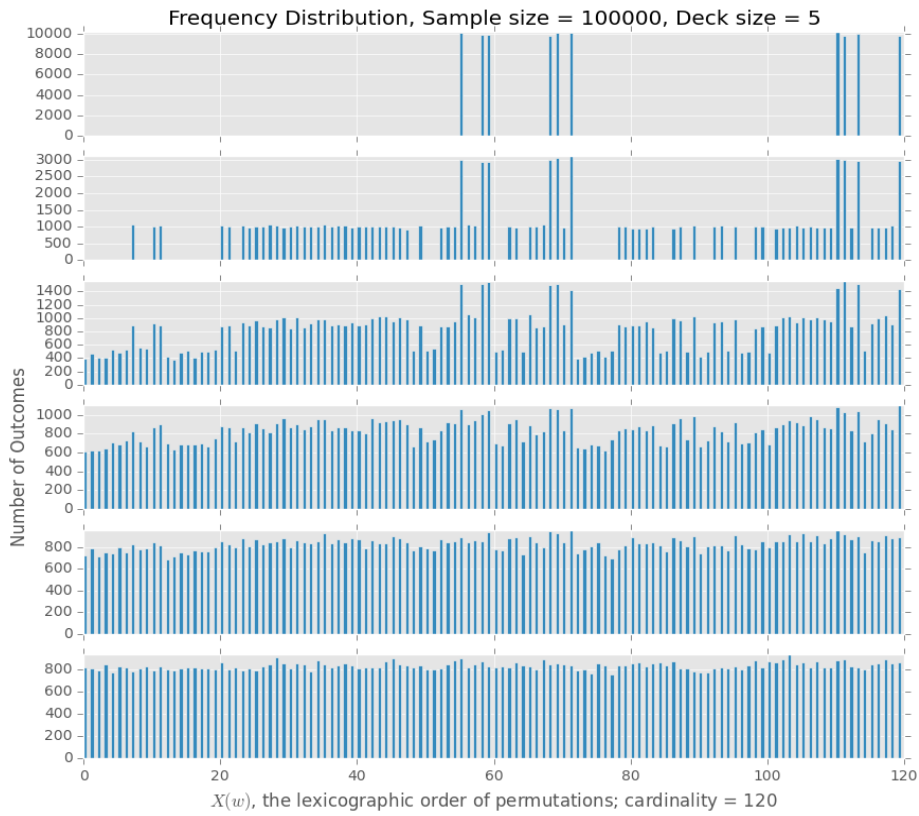
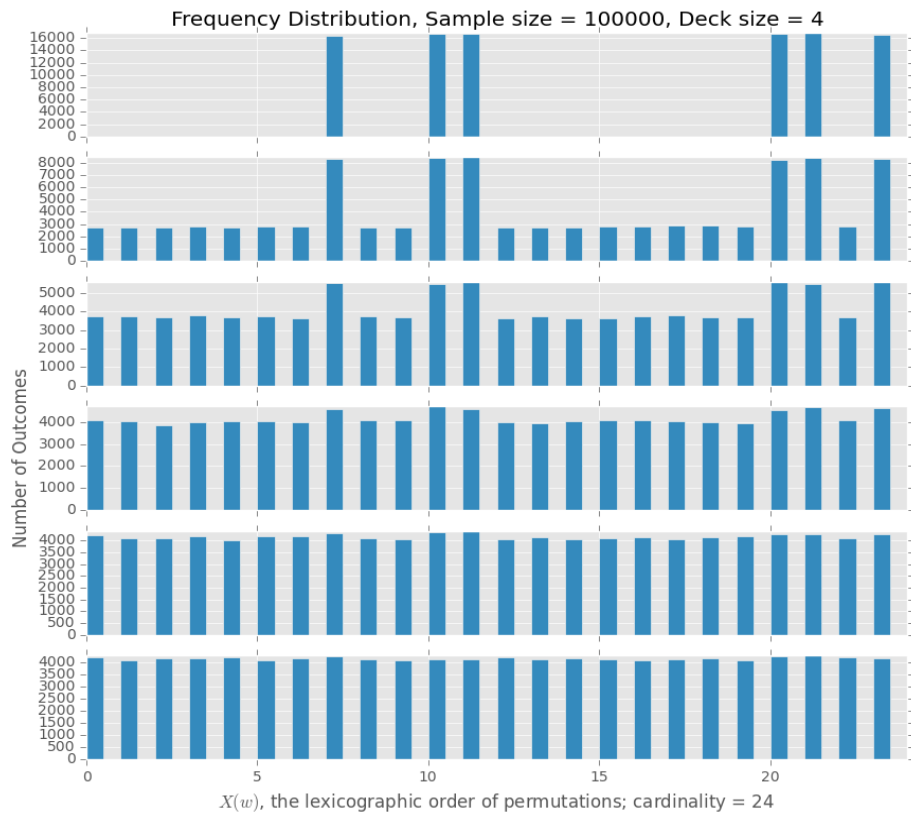
4 Simulated Results for Riffle Shuffle

It has been shown by Bayer, Diaconis and Persi[3] that riffle shuffle has significantly higher rate of convergence, that is, it has short mixing time given a required threshold (distance from output empirical distribution to uniform distribution.) It is beyond the scope of this project to look into the details of their proofs, instead, we view the result by computer simulation.

A pure C program is written to finish the shuffling process and compute final distribution (the computation relevant to permutations and the lexicographic orders). The C code is exported to a Python package. It is Python's job to draw the statistics and plots.

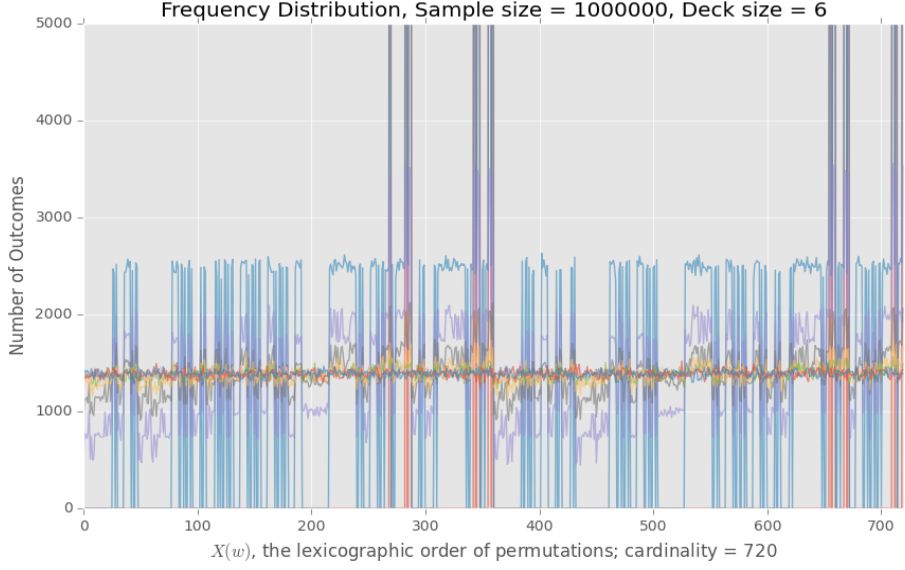
4.1 Mixing of $P^{[t]}$

The two figures below represents the mixing of $P^{[t]}$. The x axis lists different lexicographic order of permutation instances, when simulating with 4 cards there are $4! = 24$ bars, 5 cards deck has 120 bars. y axis is the counting within totally 100000 samples. The first subplot depict the outcome after one shuffle, and the sixth subplot is that after 6 shuffles.



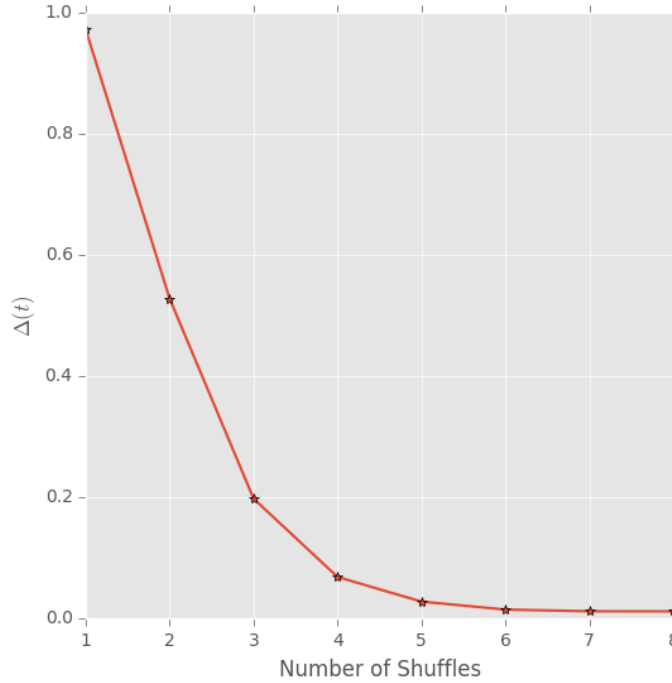
When deck size $n = 6$, there is no space for bars, so we plot the out come in one plot. Lines

of different colors represents (unnormalized) pmf after increasing numbers of shuffles.



4.2 Rate of Convergence

We now using the simulated data to compute Total Variation Norm from uniform distribution. When choosing a deck $n = 6$ cards, shuffle t times for $t = 1, 2, \dots, 8$, we obtain $\Delta(7) \approx 0.011$, which is fairly close to the targeting distribution.



Hence, with picking appropriate threshold, we can conclude that 6 to 7 riffle shuffles are sufficient to make a 6-cards-deck close enough to as if uniformly distributed in the sense that each permutations arises with equal probability $\frac{1}{n!}$. In fact, due to Bayer and his fellows[3], we still obtain a small quantity (namely 10 to 15) for the number of riffle shuffles required when it comes to a real deck of 52 cards. Since $52!$ is large, simulating this requires heavy computing efforts and optimization on algorithms, which we leave for further discussion.

References

- [1] Constantinos Daskalakis: 6.896: Probability and Computation, MIT
<http://people.csail.mit.edu/costis/6896sp11/>
- [2] Harald Hammarström. *Card-Shuffling Analysis with Markov Chains*. 2005.
- [3] Bayer, Dave and Diaconis, Persi, *Trailing the Dovetail Shuffle to its Lair*. Annals of Applied Probability, 2(2), 294-313, 1992.
- [4] Brad Mann. *How many times should you shuffle a deck of cards*. Topics in Contemporary Probability and Its Applications, Harvard University, 1995.
- [5] Sheldon M. Ross. *Introduction to Probability Models, Eleventh Edition*. ISBN: 978-0-12-407948-9, 2015