Randomised Algorithms

Lecture 4: Markov Chains and Mixing Times

Thomas Sauerwald (tms41@cam.ac.uk)

Lent 2024

UNIVERSITY OF CAMBRIDGE

Outline

Recap of Markov Chain Basics

Irreducibility, Periodicity and Convergence

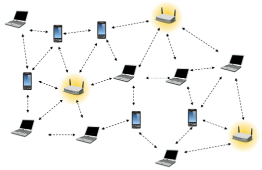
Total Variation Distance and Mixing Times

Application 1: Card Shuffling

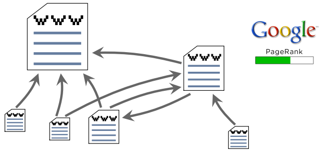
Application 2: Markov Chain Monte Carlo (non-examin.)

Appendix: Remarks on Mixing Time (non-examin.)

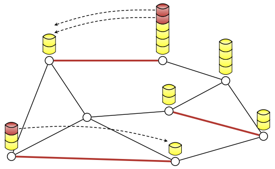
Applications of Markov Chains in Computer Science



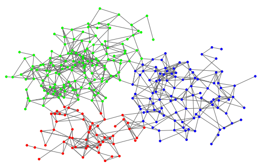
Broadcasting



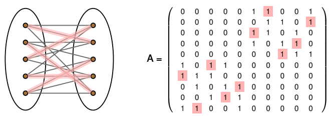
Ranking Websites



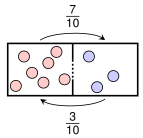
Load Balancing



Clustering



Sampling and Optimisation



Particle Processes

Markov Chains

Markov Chain (Discrete Time and State, Time Homogeneous)

We say that $(X_t)_{t=0}^{\infty}$ is a Markov Chain on State Space Ω with Initial Distribution μ and Transition Matrix P if:

1. For any $x \in \Omega$, $\mathbf{P}[\, X_0 = x \,] = \mu(x)$.
2. The Markov Property holds: for all $t \geq 0$ and any $x_0, \ldots, x_{t+1} \in \Omega$,

$$\mathbf{P}\left[\, X_{t+1} = x_{t+1} \,\middle|\, X_t = x_t, \ldots, X_0 = x_0 \,\right] = \mathbf{P}\left[\, X_{t+1} = x_{t+1} \,\middle|\, X_t = x_t \,\right]$$
$$:= P(x_t, x_{t+1}).$$

From the definition one can deduce that (check!)

- For all $t, x_0, x_1, \ldots, x_t \in \Omega$,

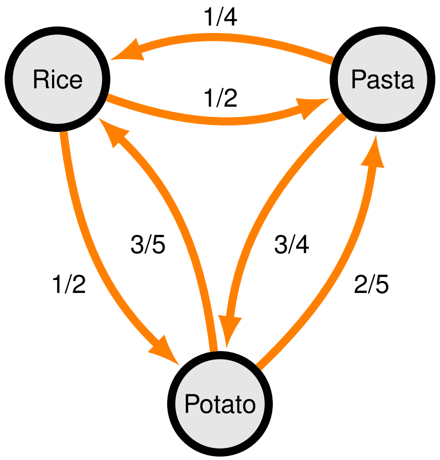
$$\mathbf{P}[\, X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0 \,]$$
$$= \mu(x_0) \cdot P(x_0, x_1) \cdot \ldots \cdot P(x_{t-2}, x_{t-1}) \cdot P(x_{t-1}, x_t).$$

- For all $0 \leq t_1 < t_2, x \in \Omega$,

$$\mathbf{P}[\, X_{t_2} = x \,] = \sum_{y \in \Omega} \mathbf{P}[\, X_{t_2} = x \mid X_{t_1} = y \,] \cdot \mathbf{P}[\, X_{t_1} = y \,].$$

What does a Markov Chain Look Like?

Example: the carbohydrate served with lunch in the college cafeteria.



This has transition matrix:

$$P = \begin{array}{cccc} \text{Rice} & \text{Pasta} & \text{Potato} & \\ 0 & 1/2 & 1/2 & \text{Rice} \\ 1/4 & 0 & 3/4 & \text{Pasta} \\ 3/5 & 2/5 & 0 & \text{Potato} \end{array}$$

Transition Matrices and Distributions

The Transition Matrix P of a Markov chain (μ, P) on $\Omega = \{1, \ldots n\}$ is given by

$$P = \begin{pmatrix} P(1,1) & \ldots & P(1,n) \\ \vdots & \ddots & \vdots \\ P(n,1) & \ldots & P(n,n) \end{pmatrix}.$$

- $\rho^t = (\rho^t(1), \rho^t(2), \ldots, \rho^t(n))$: state vector at time t (row vector).
- Multiplying ρ^t by P corresponds to advancing the chain one step:

$$\rho^t(y) = \sum_{x \in \Omega} \rho^{t-1}(x) \cdot P(x,y) \qquad \text{and thus} \qquad \rho^t = \rho^{t-1} \cdot P.$$

- The Markov Property and line above imply that for any $t \geq 0$

$$\rho^t = \rho \cdot P^{t-1} \qquad \text{and thus} \qquad P^t(x,y) = \mathbf{P}\left[X_t = y \mid X_0 = x \right].$$

Thus $\rho^t(x) = (\mu P^t)(x)$ and so $\rho^t = \mu P^t = (\mu P^t(1), \mu P^t(2), \ldots, \mu P^t(n))$.

- Everything boils down to deterministic vector/matrix computations

⇒ can replace ρ by any (load) vector and view P as a balancing matrix!

Stopping and Hitting Times

A non-negative integer random variable τ is a stopping time for $(X_t)_{t \geq 0}$ if for every $s \geq 0$ the event $\{\tau = s\}$ depends only on $X_0, \ldots, X_s$.

Example - College Carbs Stopping times:

✓ "We had rice yesterday" $\rightsquigarrow$ $\tau := \min\{t \geq 1 \colon X_{t-1} = \text{“rice”}\}$

× "We are having pasta next Thursday"

For two states $x, y \in \Omega$ we call $h(x,y)$ the hitting time of y from x:

$$h(x,y) := \mathbf{E}_x[\tau_y] = \mathbf{E}\left[\tau_y \mid X_0 = x \right] \quad \text{where } \tau_y = \min\{t \geq 1 : X_t = y\}.$$

Some distinguish between $\tau_y^+ = \min\{t \geq 1 \colon X_t = y\}$ and $\tau_y = \min\{t \geq 0 \colon X_t = y\}$

A Useful Identity

Hitting times are the solution to a set of linear equations:

$$h(x,y) \overset{\text{Markov Prop.}}{=} 1 + \sum_{z \in \Omega \setminus \{y\}} P(x,z) \cdot h(z,y) \qquad \forall x \neq y \in \Omega.$$

Outline

Recap of Markov Chain Basics

Irreducibility, Periodicity and Convergence

Total Variation Distance and Mixing Times

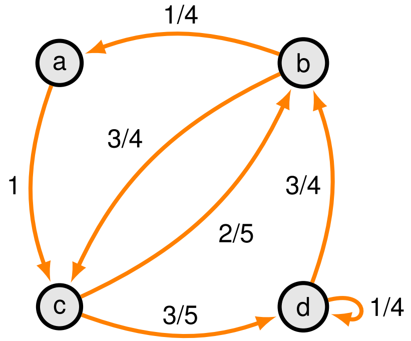
Application 1: Card Shuffling

Application 2: Markov Chain Monte Carlo (non-examin.)

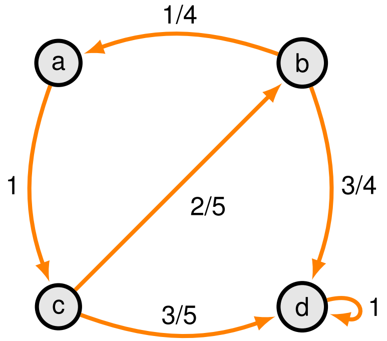
Appendix: Remarks on Mixing Time (non-examin.)

Irreducible Markov Chains

A Markov Chain is irreducible if for every pair of states $x, y \in \Omega$ there is an integer $k \geq 0$ such that $P^k(x, y) > 0$.



✓ irreducible



× not irreducible (thus reducible)

Finite Hitting Time Theorem

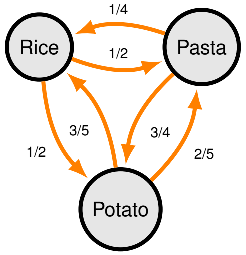
For any states x and y of a finite irreducible Markov Chain $h(x, y) < \infty$.

Stationary Distribution

A probability distribution $\pi = (\pi(1), \ldots, \pi(n))$ is the stationary distribution of a Markov Chain if $\pi P = \pi$ (π is a left eigenvector with eigenvalue 1)

College carbs example:

$$\underset{\pi}{\left(\frac{4}{13}, \frac{4}{13}, \frac{5}{13}\right)} \cdot \underset{P}{\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/4 & 0 & 3/4 \\ 3/5 & 2/5 & 0 \end{pmatrix}} = \underset{\pi}{\left(\frac{4}{13}, \frac{4}{13}, \frac{5}{13}\right)}$$



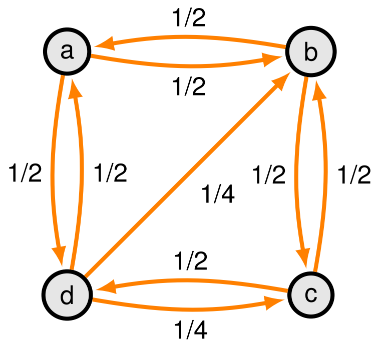
- A Markov Chain reaches stationary distribution if $\rho^t = \pi$ for some t.
- If reached, then it persists: If $\rho^t = \pi$ then $\rho^{t+k} = \pi$ for all $k \geq 0$.

Existence and Uniqueness of a Positive Stationary Distribution

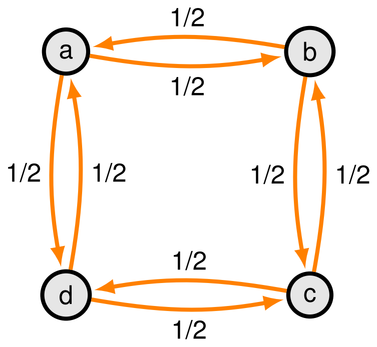
Let P be finite, irreducible M.C., then there exists a unique probability distribution π on Ω such that $\pi = \pi P$ and $\pi(x) = 1/h(x, x) > 0$, $\forall x \in \Omega$.

Periodicity

- A Markov Chain is aperiodic if for all $x \in \Omega$, $\gcd\{t \geq 1 : P^t(x, x) > 0\} = 1$.
- Otherwise we say it is periodic.



✓ Aperiodic



× Periodic

Question: Which of the two chains (if any) are aperiodic?

Convergence Theorem

Ergodic = Irreducible + Aperiodic

Convergence Theorem

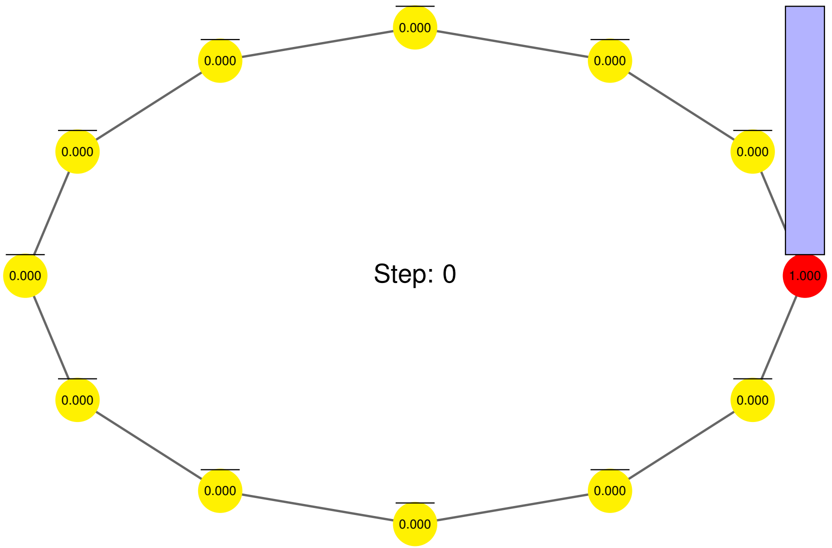
Let P be any finite, irreducible, aperiodic Markov Chain with stationary distribution π. Then for any $x, y \in \Omega$,

$$\lim_{t \to \infty} P^t(x, y) = \pi(y).$$

- mentioned before: For finite irreducible M.C.'s π exists, is unique and
$$\pi(y) = \frac{1}{h(y, y)} > 0.$$
- We will prove a simpler version of the Convergence Theorem after introducing Spectral Graph Theory.

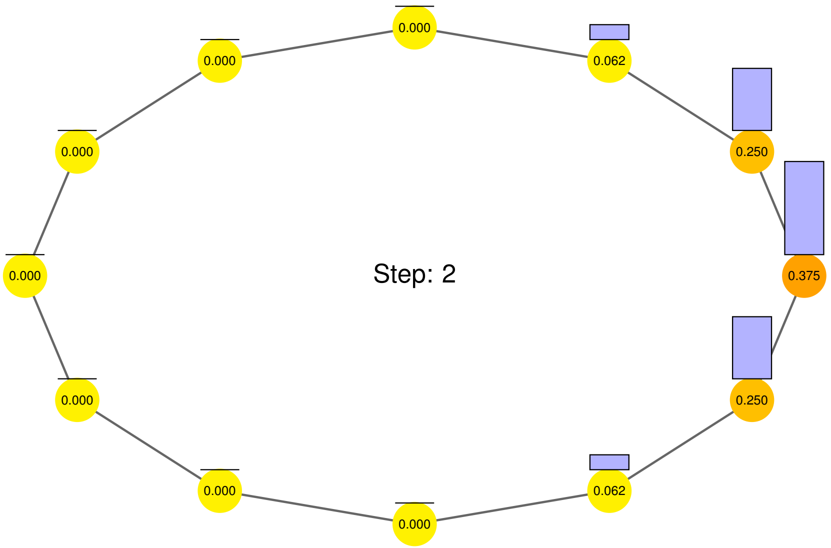
Convergence to Stationarity (Example)

- Markov Chain: stays put with $1/2$ and moves left (or right) w.p. $1/4$
- At step t the value at vertex $x \in \{1, 2, \dots, 12\}$ is $P^t(1, x)$.



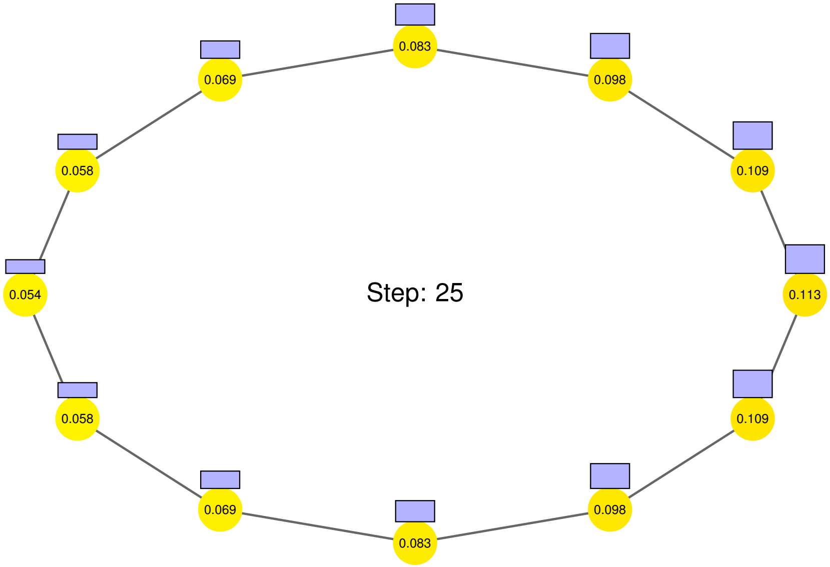
Convergence to Stationarity (Example)

- Markov Chain: stays put with $1/2$ and moves left (or right) w.p. $1/4$
- At step t the value at vertex $x \in \{1, 2, \dots, 12\}$ is $P^t(1, x)$.



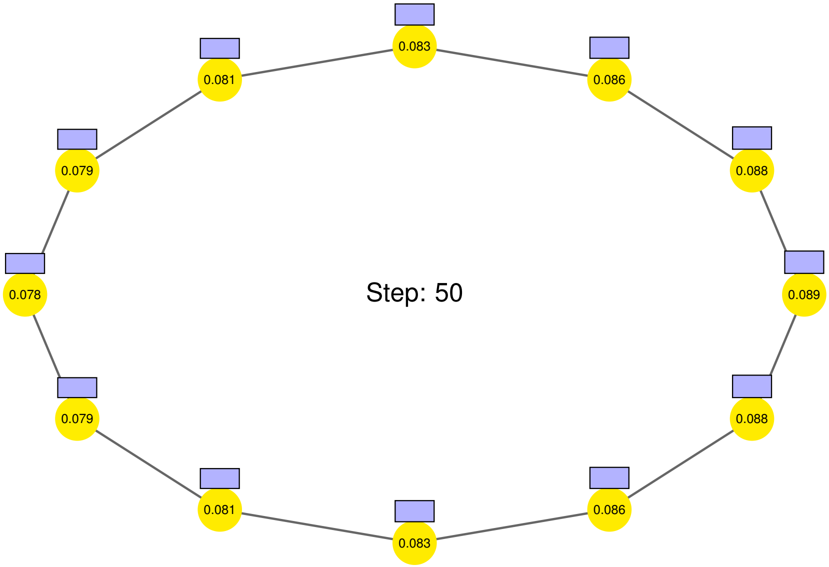
Convergence to Stationarity (Example)

- Markov Chain: stays put with $1/2$ and moves left (or right) w.p. $1/4$
- At step t the value at vertex $x \in \{1, 2, \dots, 12\}$ is $P^t(1, x)$.



Convergence to Stationarity (Example)

- Markov Chain: stays put with $1/2$ and moves left (or right) w.p. $1/4$
- At step t the value at vertex $x \in \{1, 2, \dots, 12\}$ is $P^t(1, x)$.



Outline

Recap of Markov Chain Basics

Irreducibility, Periodicity and Convergence

Total Variation Distance and Mixing Times

Application 1: Card Shuffling

Application 2: Markov Chain Monte Carlo (non-examin.)

Appendix: Remarks on Mixing Time (non-examin.)

How Similar are Two Probability Measures?

Loaded Dice

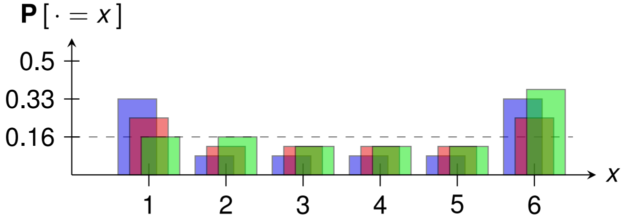
- You are presented three loaded (unfair) dice A, B, C:

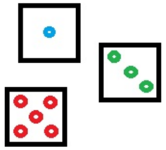
x	1	2	3	4	5	6
$\mathbf{P}[A = x]$	1/3	1/12	1/12	1/12	1/12	1/3
$\mathbf{P}[B = x]$	1/4	1/8	1/8	1/8	1/8	1/4
$\mathbf{P}[C = x]$	1/6	1/6	1/8	1/8	1/8	9/24

Question 1: Which dice is the least fair? Most choose A. Why?

Question 2: Which dice is the most fair? Dice B and C seem "fairer" than A but which is fairest?

We need a formal "fairness measure" to compare probability distributions!





Total Variation Distance

The Total Variation Distance between two probability distributions μ and η on a countable state space Ω is given by

$$\|\mu - \eta\|_{tv} = \frac{1}{2} \sum_{\omega \in \Omega} |\mu(\omega) - \eta(\omega)|.$$

Loaded Dice: let $D = Unif\{1,2,3,4,5,6\}$ be the law of a fair dice:

$$\|D - A\|_{tv} = \frac{1}{2}\left(2\left|\frac{1}{6} - \frac{1}{3}\right| + 4\left|\frac{1}{6} - \frac{1}{12}\right|\right) = \frac{1}{3}$$

$$\|D - B\|_{tv} = \frac{1}{2}\left(2\left|\frac{1}{6} - \frac{1}{4}\right| + 4\left|\frac{1}{6} - \frac{1}{8}\right|\right) = \frac{1}{6}$$

$$\|D - C\|_{tv} = \frac{1}{2}\left(3\left|\frac{1}{6} - \frac{1}{8}\right| + \left|\frac{1}{6} - \frac{9}{24}\right|\right) = \frac{1}{6}.$$

Thus

$$\|D - B\|_{tv} = \|D - C\|_{tv} \quad \text{and} \quad \|D - B\|_{tv}, \|D - C\|_{tv} < \|D - A\|_{tv}.$$

So A is the least "fair", however B and C are equally "fair" (in TV distance).

TV Distances and Markov Chains

Let P be a finite Markov Chain with stationary distribution π.

- Let μ be a prob. vector on Ω (might be just one vertex) and $t \geq 0$. Then

$$P_\mu^t := \mathbf{P}[X_t = \cdot \mid X_0 \sim \mu],$$

is a probability measure on Ω.

- *[Exercise 4/5.5]* For any μ,

$$\left\|P_\mu^t - \pi\right\|_{tv} \leq \max_{x \in \Omega} \left\|P_x^t - \pi\right\|_{tv}.$$

Convergence Theorem (Implication for TV Distance)

For any finite, irreducible, aperiodic Markov Chain

$$\lim_{t \to \infty} \max_{x \in \Omega} \left\|P_x^t - \pi\right\|_{tv} = 0.$$

We will see a similar result later after introducing spectral techniques (Lecture 12)!

Mixing Time of a Markov Chain

Convergence Theorem: “Nice” Markov Chains converge to stationarity.

Question: How fast do they converge?

Mixing Time

The mixing time $\tau_x(\epsilon)$ of a finite Markov Chain P with stationary distribution π is defined as

$$\tau_x(\epsilon) = \min \left\{ t \geq 0 \colon \left\| P_x^t - \pi \right\|_{tv} \leq \epsilon \right\},$$

and,

$$\tau(\epsilon) = \max_x \tau_x(\epsilon).$$

- This is how long we need to wait until we are “ϵ-close” to stationarity
- We often take $\epsilon = 1/4$, indeed let $t_{mix} := \tau(1/4)$

See final slides for some comments on why we choose $1/4$.

Outline

Recap of Markov Chain Basics

Irreducibility, Periodicity and Convergence

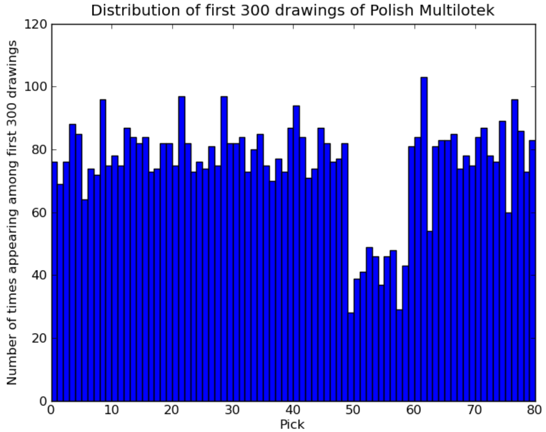
Total Variation Distance and Mixing Times

Application 1: Card Shuffling

Application 2: Markov Chain Monte Carlo (non-examin.)

Appendix: Remarks on Mixing Time (non-examin.)

Experiment Gone Wrong...



Thanks to Krzysztof Onak (pointer) and Eric Price (graph)

Source: Slides by Ronitt Rubinfeld

What is Card Shuffling?

Source: wikipedia

Here we will focus on one shuffling scheme which is easy to analyse.

How long does it take to shuffle a deck of 52 cards?

How quickly do we converge to the uniform distribution over all $n!$ permutations?

One of the leading experts in the field who has related card shuffling to many other mathematical problems.

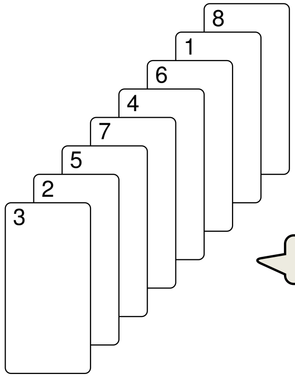
Persi Diaconis (Professor of Statistics and former Magician)

Source: www.soundcloud.com

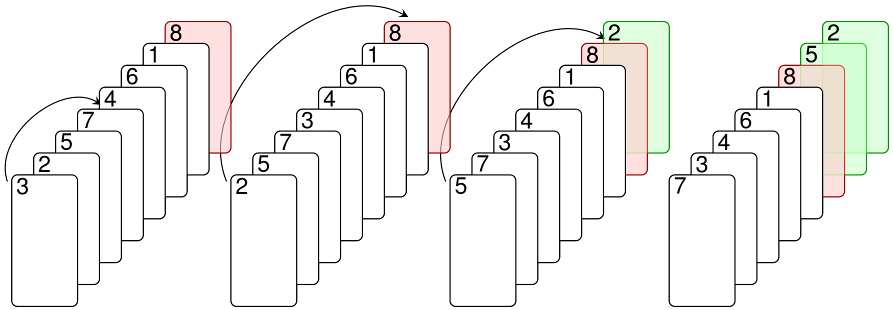
The Card Shuffling Markov Chain

```
TopToRandomShuffle (Input: A pile of $n$ cards)
For $t = 1, 2, \ldots$
    Pick $i \in \{1, 2, \ldots, n\}$ uniformly at random
    Take the top card and insert it behind the $i$-th card
```

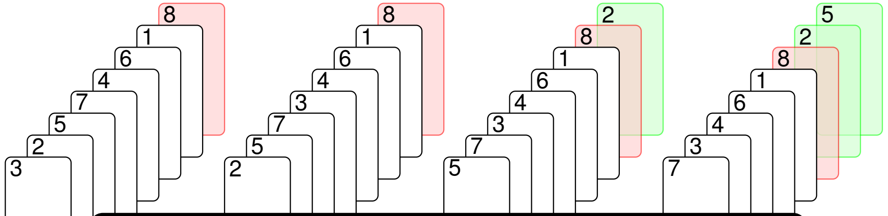
This is a slightly informal definition, so let us look at a small example...



We will focus on this “small” set of cards ($n = 8$)

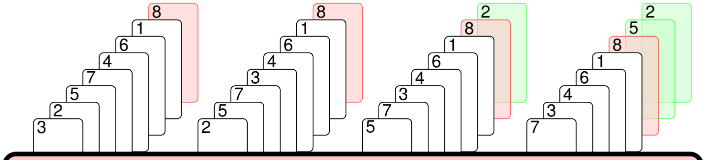


Even if we know which set of cards come after 8, every permutation is equally likely!



$\rightsquigarrow$ the deck of cards is perfectly mixed after the last card “8” reaches the top and is inserted to a random position!

Analysing the Mixing Time (Intuition)



$\rightsquigarrow$ deck of cards is perfectly mixed after the last card “8” reaches the top and is inserted to a random position!

- How long does it take for the last card “n” to become top card?
- At the last position, card “n” moves up with probability $\frac{1}{n}$ at each step
- At the second last position, card “n” moves up with probability $\frac{2}{n}$

$\vdots$

- At the second position, card “n” moves up with probability $\frac{n-1}{n}$
- One final step to randomise card “n” (with probability 1)

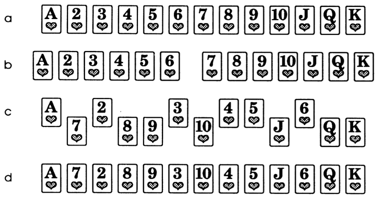
This is a **“reversed” coupon collector** process with n cards, which takes $n \log n$ in expectation.

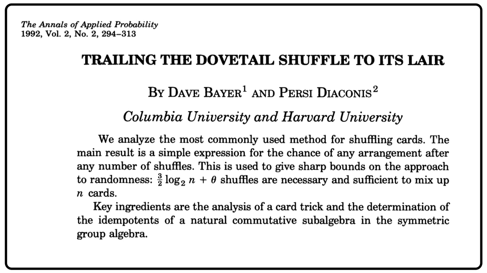
Using the so-called coupling method, one could prove $t_{mix} \leq n \log n$.

Riffle Shuffle

Riffle Shuffle

1. Split a deck of n cards into two piles (thus the size of each portion will be Binomial)
2. Riffle the cards together so that the card drops from the left (or right) pile with probability proportional to the number of remaining cards





The Annals of Applied Probability
1992, Vol. 2, No. 2, 294–313

TRAILING THE DOVETAIL SHUFFLE TO ITS LAIR

By Dave Bayer[1] and Persi Diaconis[2]

Columbia University and Harvard University

We analyze the most commonly used method for shuffling cards. The main result is a simple expression for the chance of any arrangement after any number of shuffles. This is used to give sharp bounds on the approach to randomness: $\frac{3}{2}\log_2 n + \theta$ shuffles are necessary and sufficient to mix up n cards.

Key ingredients are the analysis of a card trick and the determination of the idempotents of a natural commutative subalgebra in the symmetric group algebra.

t	1	2	3	4	5	6	7	8	9	10
$\|P^t - \pi\|_{tv}$	1.000	1.000	1.000	1.000	0.924	0.614	0.334	0.167	0.085	0.043

Figure: Total Variation Distance for t riffle shuffles of 52 cards.

Outline

Recap of Markov Chain Basics

Irreducibility, Periodicity and Convergence

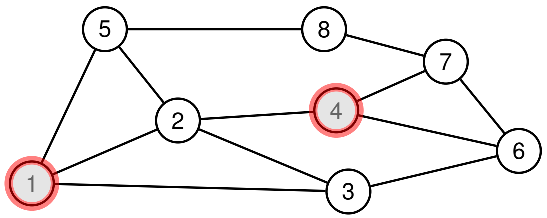
Total Variation Distance and Mixing Times

Application 1: Card Shuffling

Application 2: Markov Chain Monte Carlo (non-examin.)

Appendix: Remarks on Mixing Time (non-examin.)

Markov Chain for Sampling Independent Sets (1/2) (non-examin.)



$S = \{1, 4\}$ is an independent set ✓

Independent Set

Given an undirected graph $G = (V, E)$, an independent set is a subset $S \subseteq V$ such that there are no two vertices $u, v \in S$ with $\{u, v\} \in E(G)$.

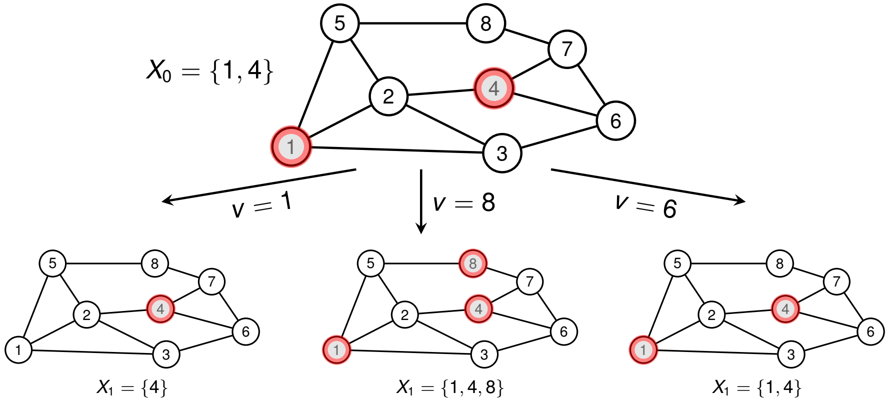
How can we take a sample from the space of all independent sets?

Naive brute-force would take an insane amount of time (and space)!

We can use a generic Markov Chain Monte Carlo approach to tackle this problem!

Markov Chain for Sampling Independent Sets (2/2) (non-examin.)

```
INDEPENDENTSETSAMPLER
Let $X_0$ be an arbitrary independent set in $G$
For $t = 0, 1, 2, \ldots$:
    Pick a vertex $v \in V(G)$ uniformly at random
    If $v \in X_t$ then $X_{t+1} \leftarrow X_t \setminus \{v\}$
    elif $v \notin X_t$ and $X_t \cup \{v\}$ is an independent set then $X_{t+1} \leftarrow X_t \cup \{v\}$
    else $X_{t+1} \leftarrow X_t$
```



Markov Chain for Sampling Independent Sets (2/2) (non-examin.)

```
INDEPENDENTSETSAMPLER
Let $X_0$ be an arbitrary independent set in $G$
For $t = 0, 1, 2, \ldots$:
    Pick a vertex $v \in V(G)$ uniformly at random
    If $v \in X_t$ then $X_{t+1} \leftarrow X_t \setminus \{v\}$
    elif $v \notin X_t$ and $X_t \cup \{v\}$ is an independent set then $X_{t+1} \leftarrow X_t \cup \{v\}$
    else $X_{t+1} \leftarrow X_t$
```

Remark

- This is a local definition (no explicit definition of P!)
- This chain is irreducible (every independent set is reachable)
- This chain is aperiodic (Check!)
- The stationary distribution is uniform, since $P_{u,v} = P_{v,u}$ (Check!)

Key Question: What is the mixing time of this Markov Chain?

not covered here, see the textbook by Mitzenmacher and Upfal

Outline

Recap of Markov Chain Basics

Irreducibility, Periodicity and Convergence

Total Variation Distance and Mixing Times

Application 1: Card Shuffling

Application 2: Markov Chain Monte Carlo (non-examin.)

Appendix: Remarks on Mixing Time (non-examin.)

Further Remarks on the Mixing Time (non-examin.)

- One can prove $\max_x \left\| P_x^t - \pi \right\|_{tv}$ is non-increasing in t (this means if the chain is "ϵ-mixed" at step t, then this also holds in future steps) *[Mitzenmacher, Upfal, 12.3]*
- We chose $t_{mix} := \tau(1/4)$, but other choices of ϵ are perfectly fine too (e.g, $t_{mix} := \tau(1/e)$ is often used); in fact, any constant $\epsilon \in (0, 1/2)$ is possible.

Remark: This freedom on how to pick ϵ relies on the sub-multiplicative property of a (version) of the variation distance. First, let

$$d(t) := \max_x \left\| P_x^t - \pi \right\|_{tv}$$

be the variation distance after t steps when starting from the worst state. Further, define

$$\overline{d}(t) := \max_{\mu,\nu} \left\| P_\mu^t - P_\nu^t \right\|_{tv}.$$

These quantities are related by the following double inequality

$$d(t) \leq \overline{d}(t) \leq 2d(t).$$

This 2 is the reason why we ultimately need $\epsilon < 1/2$ in this derivation. On the other hand, see *[Exercise (4/5).8]* why $\epsilon < 1/2$ is also necessary.

Further, $\overline{d}(t)$ is sub-multiplicative, that is for any $s, t \geq 1$,

$$\overline{d}(s+t) \leq \overline{d}(s) \cdot \overline{d}(t).$$

Hence for any fixed $0 < \epsilon < \delta < 1/2$ it follows from the above that

$$\tau(\epsilon) \leq \left\lceil \frac{\ln \epsilon}{\ln(2\delta)} \right\rceil \tau(\delta).$$

In particular, for any $\epsilon < 1/4$

$$\tau(\epsilon) \leq \left\lceil \log_2 \epsilon^{-1} \right\rceil \tau(1/4).$$

Hence smaller constants $\epsilon < 1/4$ only increase the mixing time by some constant factor.