Skolkovo Institute of Science and Technology

Stochastic methods in Mathematical Modelling

Lecture 7. Markov Chains

Markov Chains

A finite Markov chain is a process which moves among the elements of a finite set Ω in the following manner: when $x \in \Omega$, the next position is chosen according to a fixed probability distribution $P(x, \cdot)$.

More precisely, a sequence of random variables $(X_0, X_1, \ldots)$ is a Markov chain with state space Ω and transition matrix P if for all $x,y \in \Omega$, all $t \geq 1$ and all events $H_{t-1} = \{X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}\}$ satisfying $P(H_{t-1} = \{X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}\}) > 0$

$$P(X_{t+1} = y \mid H_{t-1} = \{X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}\} \cap \{X_t = x\}) = P(X_{t+1} = y \mid X_t = x) = P(x, y)$$

Markov property

The set of states is countable!!

Skoltech
Skolkovo Institute of Science and Technology

Markov Chains

Markov chain is a discrete random process with probability of an event depending on the state at the previous moment in time only

If S_n denotes a state at the moment *n* and *n* is integer then the Markov property is fulfilled if

$$P(S_n=i_n|S_{n-1}=i_{n-1},S_{n-2}=i_{n-2}...,S_0=i_0)=P(S_n=i_n|S_{n-1}=i_{n-1})$$

$$P(S_n=i_n,S_{n-1}=i_{n-1},S_{n-2}=i_{n-2}...,S_0=i_0)=$$

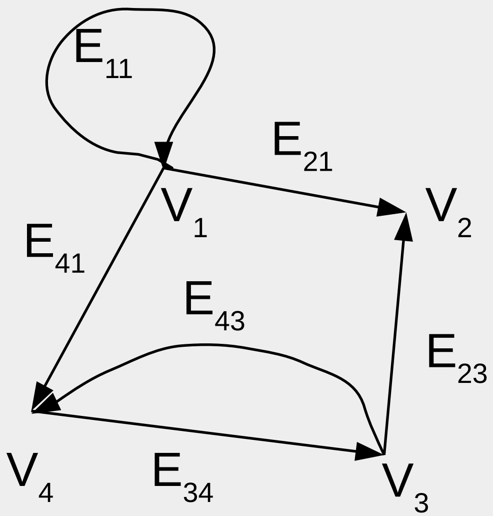
$$=P(S_n|S_{n-1}=i_{n-1})P(S_{n-1}|S_{n-2}=i_{n-2})...P(S_0=i_0)$$

transition probabilities

initial condition

The set of states is countable!!

Markov Chains



Transition probabilities

$$P(S_n = i | S_{n-1} = j) = p_{ij} = p_{i \leftarrow j}$$

$$\forall j: \sum_{i:(j \rightarrow i) \in E} p_{ij} = 1$$

all directed edges originating from vertex j

For *stationary* Markov chains p_{ij} do not depend on time

The set {*E*,*V*, *p*} defines the Markov chain

The particular trajectory will look as

$i_0(0)$, $i_1(t_1)$ … $i_n(t_n)$, where states i_0, … ,i_n $\in$ V

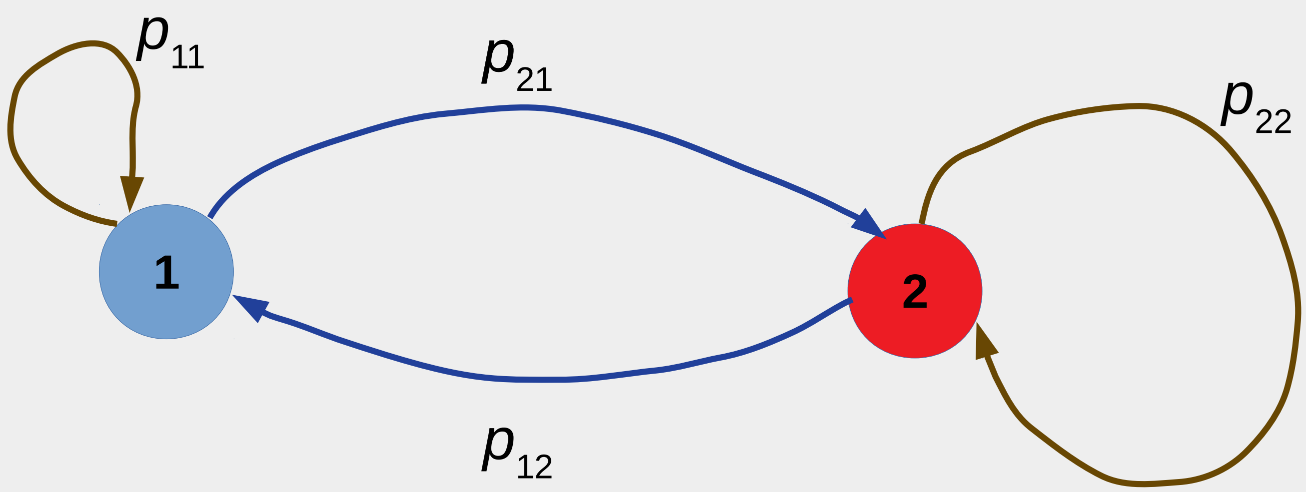
Set of states

https://setosa.io/ev/markov-chains/

Skoltech
Skolkovo Institute of Science and Technology

Markov Chains

Graphic representation: directed graph



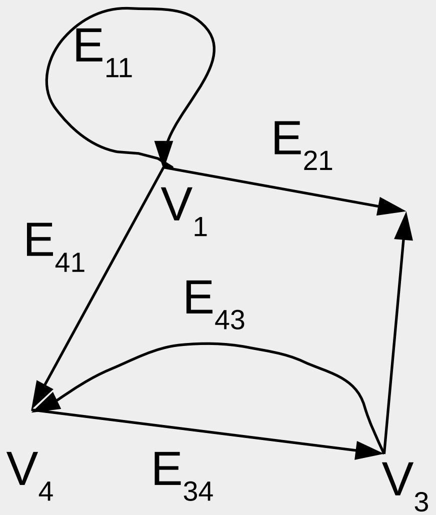
V={1,2}

Q: What can you tell about p_{ij} on the picture? Are there conditions on them?

https://setosa.io/ev/markov-chains/

Skoltech
Skolkovo Institute of Science and Technology

Markov Chains



One can sample many trajectories

$i_0^{(n)}(0),\ i_1^{(n)}(t_1)\ \dots\ i_k^{(n)}(t_k),\ n = 1,\ \dots\ ,\ N$

Alternatively, one can describe system in terms of "state" (stochastic) vectors $\pi = (\pi_1, \pi_2, \dots, \pi_M)$

$$\pi_i(t+1) = \sum_j p_{ij}\, \pi_j(t)$$

$$\hat{p} = \begin{pmatrix} p_{11} & p_{12} & p_{13} & \dots \\ p_{21} & p_{22} & p_{23} & \dots \\ p_{31} & p_{32} & p_{33} & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

In a vector (matrix form)

$$\pi(t+1) = \hat{p}\, \pi(t)$$

Transition (stochastic) matrix

The sum over columns gives 1

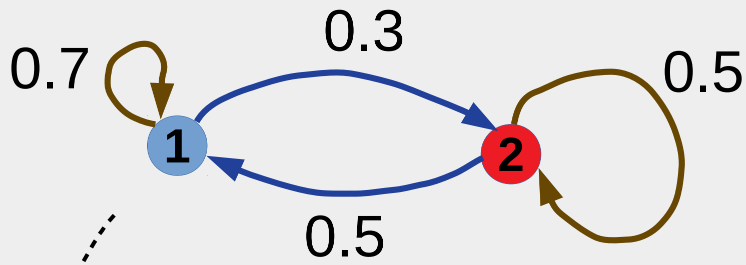
$$\boxed{1 = \sum_j p_{ij}}$$

$$\pi(t+k) = (\hat{p})^k\, \pi(t)$$

Visualisation: https://setosa.io/ev/markov-chains/

Markov Chains

Transition matrix and a steady state



$$p^1 = \begin{pmatrix} 0.7 & 0.5 \\ 0.3 & 0.5 \end{pmatrix}, \quad p^2 = \begin{pmatrix} 0.64 & 0.6 \\ 0.36 & 0.4 \end{pmatrix}, \quad p^{10} \approx p^{100} \approx \begin{pmatrix} 0.625 & 0.625 \\ 0.375 & 0.375 \end{pmatrix}$$

π^* is called a *stationary distribution* if

$$\pi^* = p\,\pi^*$$

$$\pi^* = \begin{pmatrix} 0.625 \\ 0.375 \end{pmatrix}$$

generally $\pi^* = \frac{e}{\sum_i e_i}$

Eigenvector with eigenvalue 1

Skoltech
Skolkovo Institute of Science and Technology

Markov Chains

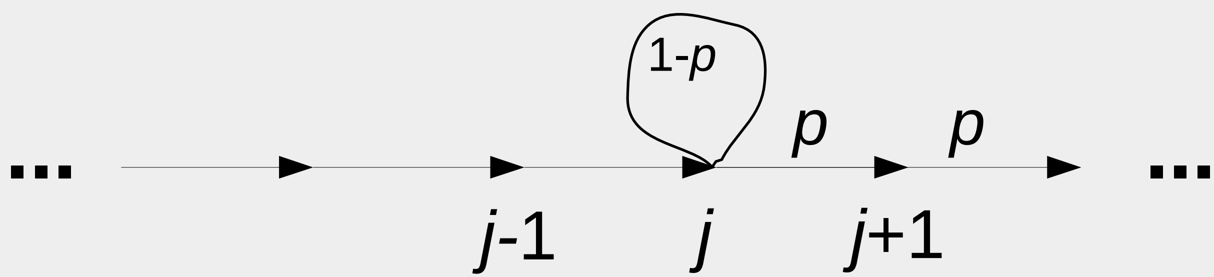
Example 1. Binomial Markov Chain

Let X_n be a number of successes in n trials of Bernoulli process with success probability p in each trial. The probability of k successes in n trials reads

$$P(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, 0 \leqslant k \leqslant n$$

Let's assume that $X_n = j$. Then X_{n+1} is equal $j+1$ and j with probabilities p and $1 - p$

Thus, X_n is a Markov chain with transition probabilities $p_{j+1,j} = p$, $p_{j,j} = 1-p$ and $p_{ij} = 0$ otherwise



Markov Chains

Example 2. Random walk as a Markov Chain

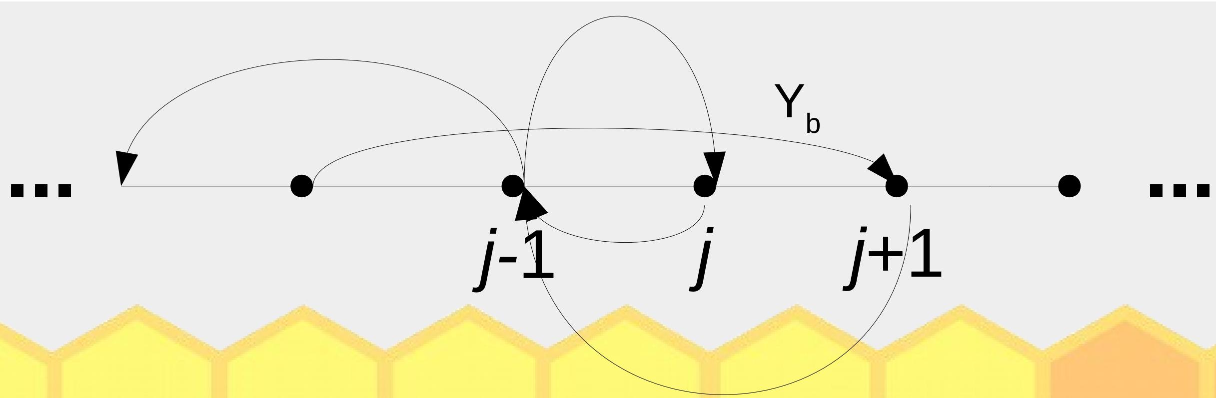
Assume increments $Y_1, \ldots, Y_n$ are i.i.d. integer valued random variables. Also suppose $X_0=0$ and

$$X_n = \sum_{m=1}^{n} Y_m, n \geq 1$$

The process X_n is a *random walk* on integers with Y_m being step lengths at step number m

$$X_{n+1} = X_n + Y_{n+1}$$

$$P(X_{n+1} = j \mid X_0, \ldots, X_{n-1}, X_n = i) = P(X_n + Y_{n+1} = j \mid X_n = i) = p_{ji}$$



Skoltech
Skolkovo Institute of Science and Technology

Markov Chains

Example 3. Coupon collecting

A company issues n different types of coupons. A collector desires a complete set. We suppose each coupon he acquires is equally likely to be each of the n types. How many coupons must he obtain so that his collection contains all n types?

Q How does one construct the Markov chain?

...  ...

Detour. Martingales

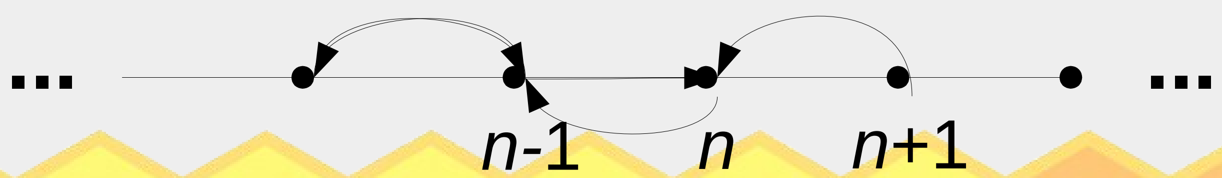
If M_n is an the amount of money at time *n* for a gambler betting on a fair game, and X_n as the outcomes of the gambling game we say that M_0, M_1 , . . . is a *martingale* with respect to X_0, X_1, . . . if for any $n \geq 0$ we have $E|M_n| < \infty$ and for any possible values x_n , . . . , x_0

$$\boxed{E\left(M_{n+1}-M_n \middle| X_n=x_n, X_{n-1}=x_{n-1}, \ldots X_0=x_0, M_0=m_0\right)=0}$$

$$A_v=\{X_n=x_n, X_{n-1}=x_{n-1}, \ldots X_0=x_0, M_0=m_0\}$$

Example: Simple random walk. X_1, X_2, . . . be i.i.d. with $E[X_i] = \mu$.
Let $S_n = S_0 + X_1 + \cdots + X_n$ be a random walk.
Then $M_n = S_n - n\mu$ is a martingale with respect to X_n
Proof: $M_{n+1} - M_n = X_{n+1} - \mu$ is independent of $X_1,\ldots,X_n$

hence $E(M_{n+1} - M_n | A_v) = E[X_{n+1}] - \mu = 0$



Skoltech

Skolkovo Institute of Science and Technology

Detour. Martingales

Martingale:

$$E(M_{n+1} - M_n | A_v) = 0$$

Unbiased random walks

Supermartingale:

$$E(M_{n+1} - M_n | A_v) \leq 0$$

Casino gambling
(expected winnings are negative)

Submartingale:

$$E(M_{n+1} - M_n | A_v) \geq 0$$

A walk biased towards higher values
(expected winnings are positive)

Detour. Martingales

Another example. “Stock market”. Products of independent random variables. To build a discrete time model of the stock market we let X_1 , X_2, . . . be independent ≥ 0 with $E[X_i] = 1$. Then $M_n = M_0 X_1 \cdot\cdot\cdot X_n$ is a martingale with respect to X_n .

Proof:

$E(M_{n+1} - M_n|A_v) = M_n E(X_{n+1} - 1|A_v) = 0$

The reason for a multiplicative model is that changes in stock prices are thought to be proportional to its value. Also, in contrast to an additive model, we are guaranteed that prices will stay positive.

$$E(M_{n+1} - M_n \mid A_n) = M_n E(X_{n+1} - 1 \mid A_n) = 0$$

$$E(X_{n+1} - 1 \mid A_n) = \underbrace{E(X_{n+1})}_{=1} - 1$$

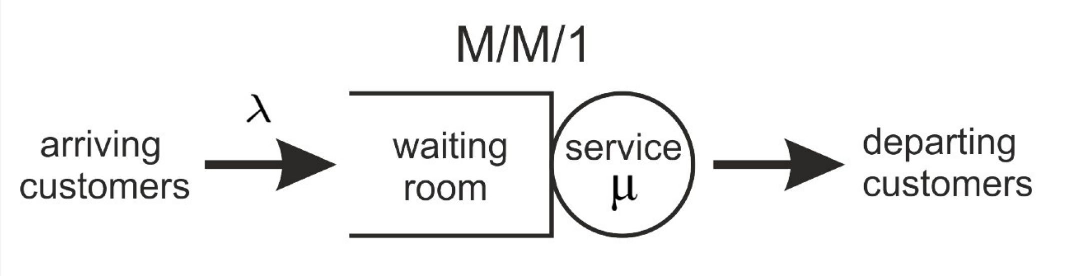
$$M_n = M_0 X_1 \cdots X_n$$

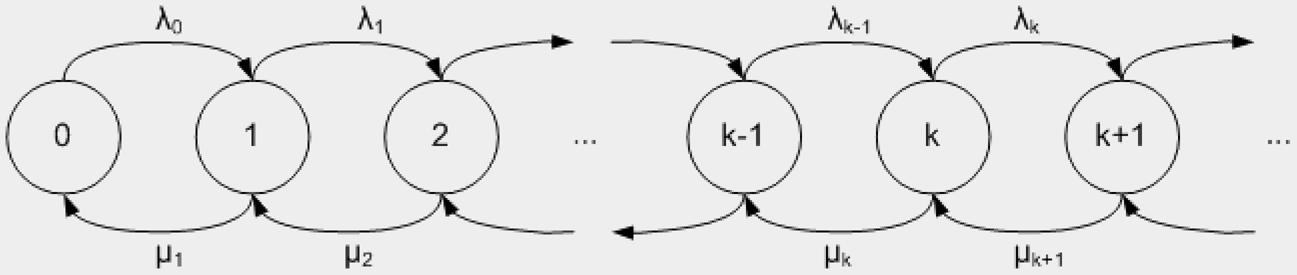
$$M_{n+1} = \underbrace{M_0 X_1 \cdots X_n}_{M_n} X_{n+1}$$

$$A_n = \{X_0, X_1, \ldots, X_n\}$$

Markov Chains

Example 4. Queuing systems

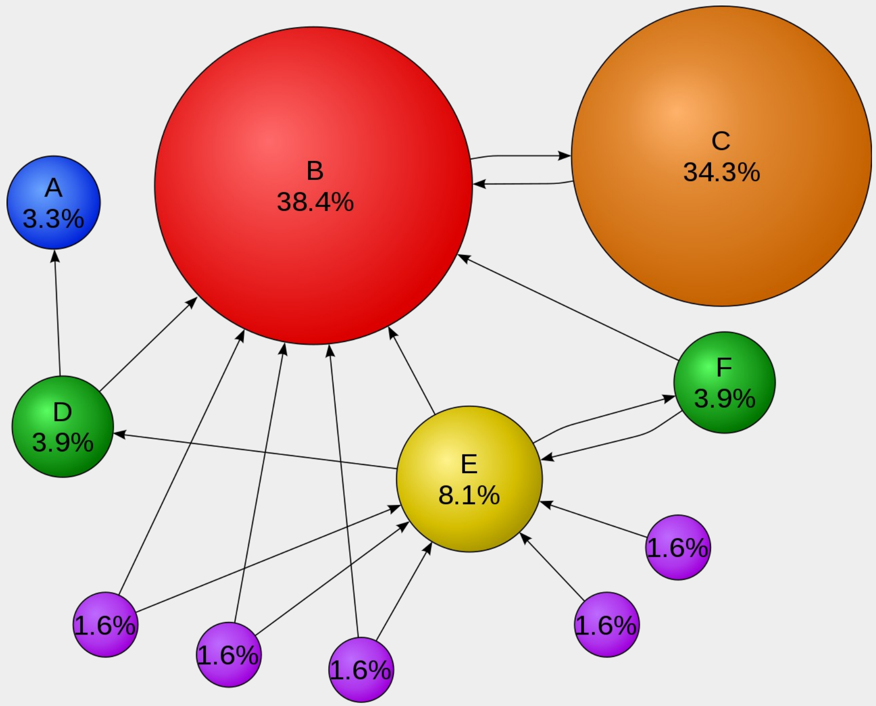




Markov Chains

Example 5. *Page rank algorithm* of Google

How to construct the algorithm which ranks pages from most to least popular/relevant ones?



The rank value indicates an importance of a particular page. A hyperlink to a page counts as a vote of support. A page that is linked to by many pages with high PageRank receives a high rank itself.

Model of a random surfer who reaches their target site after several clicks, then switches to a random page. The PageRank value of a page reflects the chance that the random surfer will land on that page by clicking on a link. It can be understood as a *Markov chain* in which the states are pages, and the transitions are the links between pages all of which are all equally probable.

If a page has no links to other pages, it becomes a sink and therefore terminates the random surfing process.

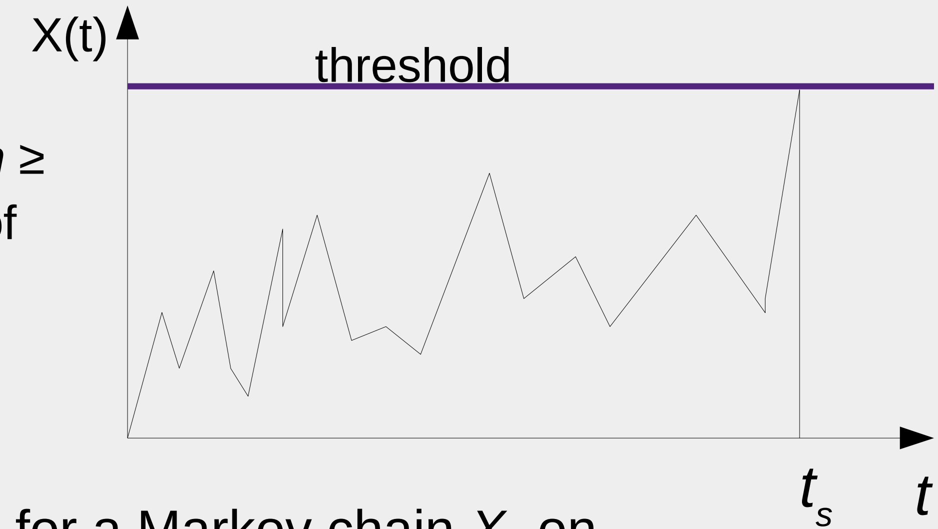
If the random surfer arrives at a sink page, it picks another URL at random and continues surfing again.

Markov Chains

Stopping time is a time when a random variable reaches certain value/behaviour

Def. A random variable τ that takes values in {0, 1, . . ., ∞} is a *stopping time* for a process $\{X_n: n \geq 0\}$ if, for any finite n, the event $\{\tau = n\}$ is a function of the history $X_0, \ldots, X_n$ up to time *n*

Example: hitting times



Strong Markov property

Suppose that τ is a finite-valued *stopping time* for a Markov chain X_n on S. Then, for any $i \in S$ and $i_1, i_2, \ldots, j_1, \ldots, j_m \in S$ and $m \geq 1$,

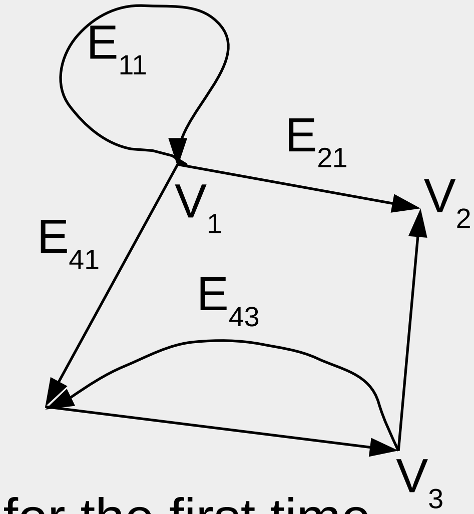
$$P(X_{\tau+1}=j_1,\ldots,X_{\tau+m}=j_m|X_0=i_0,\ldots,X_{\tau-1}=i_{\tau-1},X_\tau=i)=$$

$$=P(X_1=j_1,\ldots,X_m=j_m|X_0=i)$$

The strong Markov property roughly states that a Markov chain starts anew at a stopping time. The distribution of the future is equal to that of the original chain.

Markov Chains

Classification of states



(First) hitting times $\tau_j = min(n \geq 1, X_n = j)$

(First) hitting probability: The chain starting at *i* enters the state *j* for the first time at *n*th step

$$f_{ji}^{n} = P_i(\tau_j = n), n \geq 1$$

$$f_{ji}^{n} = \sum_{k \neq j} p_{ki} f_{jk}^{n-1}, n \geq 2, k \in S$$

(*)

overall hitting probability $f_{ji} = \sum_{n=1}^{\infty} f_{ji}^{n}$

From (*) => $f_{ji} = p_{ji} + \sum_{k \neq j} p_{ki} f_{jk}$

The latter is a linear system.
i.e. the hitting probabilities can be computed from the transition matrix!!!

Classification of states

1) Recurrent state $f_{ii} = 1$

hitting prob.

Transient state $f_{ii} < 1$

state i is positive recurrent if $E[\tau_i] < \infty$

null recurrent if $E[\tau_i] = \infty$

1D lattice

recurrent state $P_i(N_i = \infty) = 1$

transient $P_i(N_i < \infty) = 1$

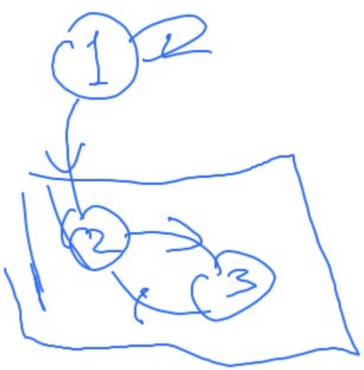
j is accessible from i if $i \to j$

i & j communicate if $i \to j$
$j \to i$

set C is recurrent if all ~~all~~ states are recurrent

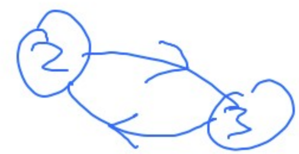
Reducible set

C is reducible



Irreducible

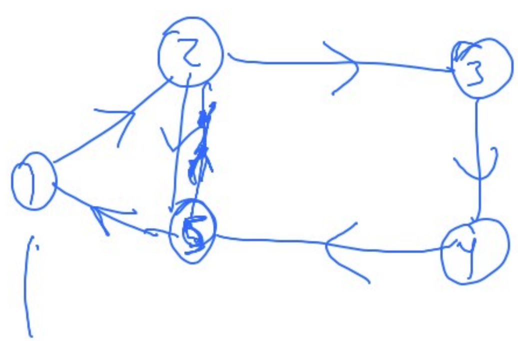
C is irreducible if $\forall i,j \in C$
$i \leftrightarrow j$



Period of a state

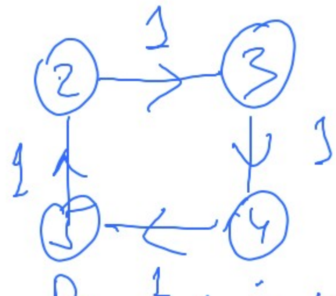
Period is k when the return takes k steps

$T_i \; \min_k \{ P_i(X_k = i) > 0 \}$

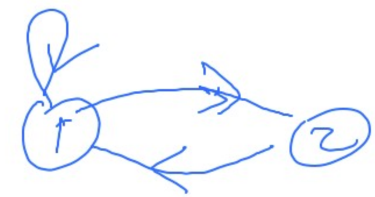


for 1,2,5 period is 3

3,4 period is 5



Period is 4



aperiodic Markov chains

GCD (T_i) = 1

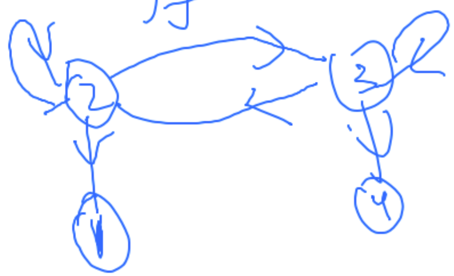
greatest common divisor

Proposition Irreducible Markov chain on a finite space is positive recurrent

Ergodicity of a Markov chain (MC)

MC is ergodic if it is irreducible and the states are positive recurrent and aperiodic

Theorem: An irreducible aperiodic Markov chain is ergodic iff it has a unique stationary state (distribution)



Markov Chains

Properties

Reducible/irreducible

Periodic/aperiodic

Transient/recurrent

ergodicity

Consequence of ergodicity is a unique and universal steady state

Remark For historical reasons, in the literature a positive recurrent and aperiodic Markov chain is sometimes called an ergodic chain.

The word ergodic, however, has a precise meaning in mathematics (ergodic theory) and this meaning has nothing to do

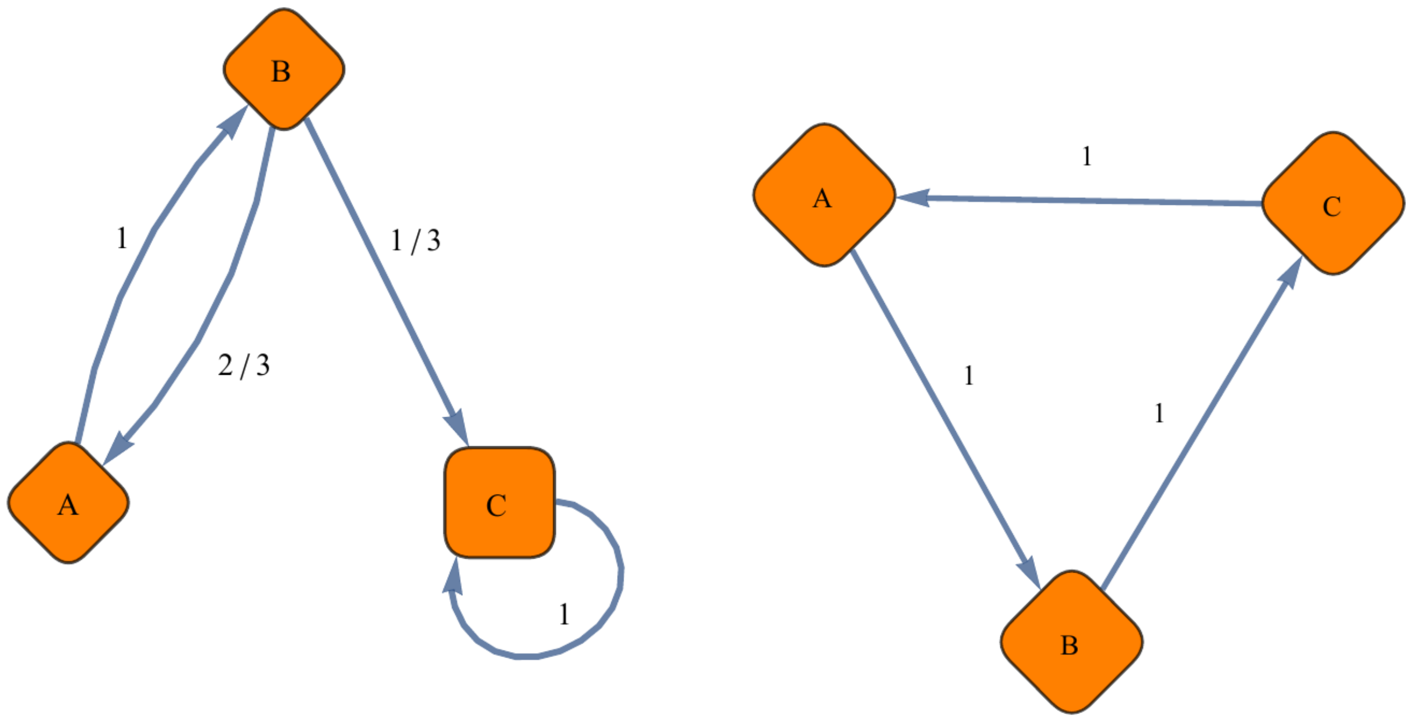
with aperiodicity! In fact any positive recurrent Markov chain is ergodic in this precise mathematical sense.

(So the historical use of ergodic in the context of aperiodic Markov chains is misleading and unfortunate.)

Skoltech
Skolkovo Institute of Science and Technology

Markov Chains

Examples



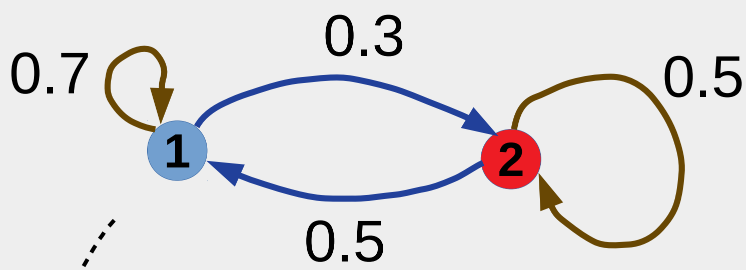
(a) Reducible, Periodic (b) Irreducible, Periodic

Skoltech

Skolkovo Institute of Science and Technology

Markov Chains

Transition matrix



$$p^1 = \begin{pmatrix} 0.7 & 0.5 \\ 0.3 & 0.5 \end{pmatrix}, \quad p^2 = \begin{pmatrix} 0.64 & 0.6 \\ 0.36 & 0.4 \end{pmatrix}, \quad p^{10} \approx p^{100} \approx \begin{pmatrix} 0.625 & 0.625 \\ 0.375 & 0.375 \end{pmatrix}$$

π* is called a *stationary distribution* if

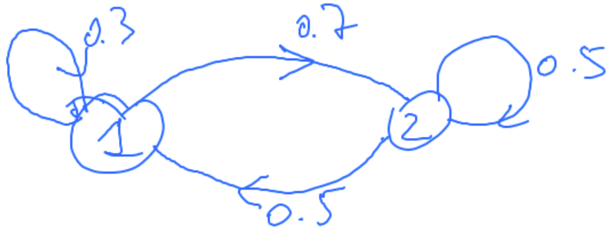
$$\pi^* = p\, \pi^*$$

$$\pi^* = \begin{pmatrix} 0.625 \\ 0.375 \end{pmatrix}$$

generally

$$\pi^* = \frac{e}{\sum_i e_i}$$

e is the eigenvector with eigenvalue 1

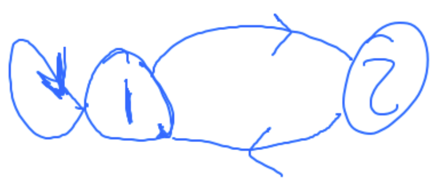


$$P = \begin{pmatrix} 0.7 & 0.5 \\ 0.3 & 0.5 \end{pmatrix} \qquad \pi = \begin{pmatrix} 0.625 \\ 0.375 \end{pmatrix}$$

reversible Markov chain

$$Q = \begin{pmatrix} 0.7 \cdot 0.625 & 0.5 \cdot 0.375 \\ 0.3 \cdot 0.625 & 0.5 \cdot 0.375 \end{pmatrix} = \begin{pmatrix} 0.4375 & 0.1875 \\ 0.1875 & 0.1875 \end{pmatrix}$$

irreducible MC of 2 states
is always reversible



Markov Chains

Skoltech

Skolkovo Institute of Science and Technology

Steady state analysis. The spectrum of the transition matrix

Let's assume the matrix p can be diagonalised and decompose it as

$$p = U^{-1} \Sigma U$$

$$\Sigma = diag(\lambda_1, \lambda_{2,\dots}, \lambda_n), 1 = |\lambda_1| \geq |\lambda_2| \geq |\lambda_3|, \dots, |\lambda_n|$$

U is a matrix of eigenvectors

u_i are normalised eigenvectors

Initial state of the system

$\pi^* = p\, \pi^*$

$$\pi^{(k)} = p^k \pi_0 = (U^{-1}\Sigma U)^k \pi_0 = U^{-1}\Sigma^k U \pi_0$$

$$\pi_0 = \sum_{i=1}^{n} a_i u_i$$

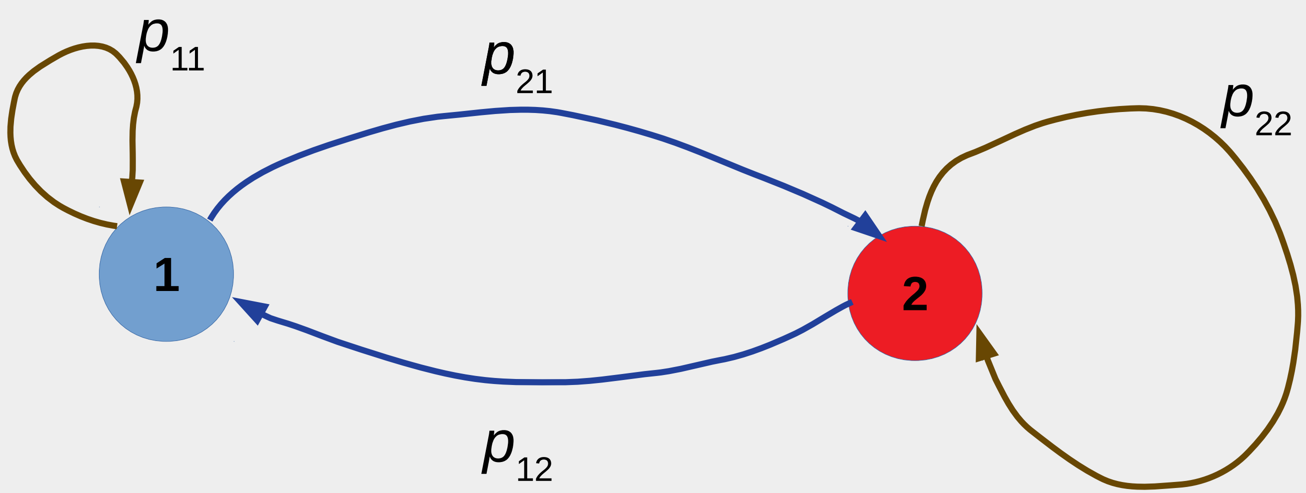
$$\pi^{(k)} = \lambda_1^k \left(a_1 u_1 + a_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k u_2 + \dots + a_n \left(\frac{\lambda_n}{\lambda_1}\right)^k u_n\right)$$

(cf. Perron-Frobenius theorem (our matrix is nonnegative and the largest eigenvalue is 1))

Markov Chains

Reversible and irreversible Markov chains. Detailed balance condition

Markov chain is called *reversible* if ∀{i, j} ∈ E: $p_{ji}\,\pi_i^* = p_{ij}\,\pi_j^*$



Ergodicity matrix $Q_{ji}=p_{ji}\,\pi_i^*$

Detailed balance is ensured by the symmetry of the matrix $\mathbf{Q}=\mathbf{Q}^T$!

If **Q** is asymmetric then Markov chain is *irreversible*

Markov Chains

Detailed vs global balance

Balance (global) balance condition

$$\sum_{j:\, j\leftarrow i\in E} p_{ji}\pi_i^* = \sum_{j:\, i\leftarrow j\in E} p_{ij}\pi_j^*$$

In the irreversible case

$$Q_{ij} - Q_{ji} = \sum_\alpha J_\alpha \left(C_{ij}^\alpha - C_{ji}^\alpha\right)$$

Enumerates cycles on the graph of states with adjacency matrices C^α.
J_α is a magnitude of the flux

Markov Chains

Exercise

1) Write down the transition matrix

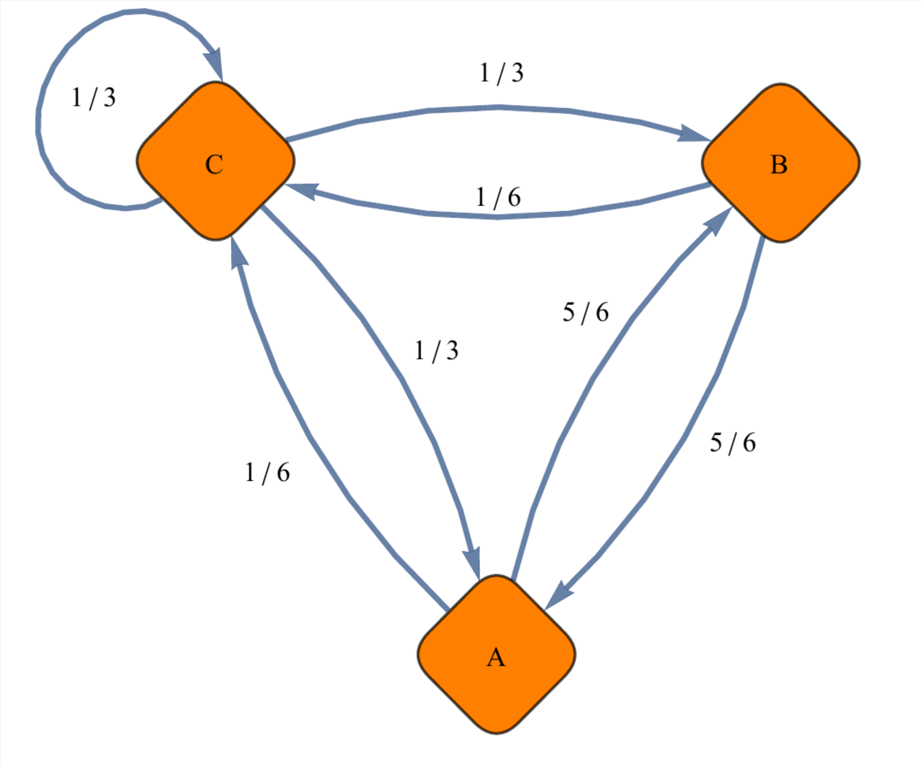
2) Check that the matrix is stochastic (values in every column add up to 1)

3) Find the eigenvectors and eigenvalues

4) Suppose that we start in the state “A”, i.e. $\pi(0)=(1,0,0)^T$. Write the initial state as a linear combination of eigenvectors and find $\pi(t)=\pi^t$

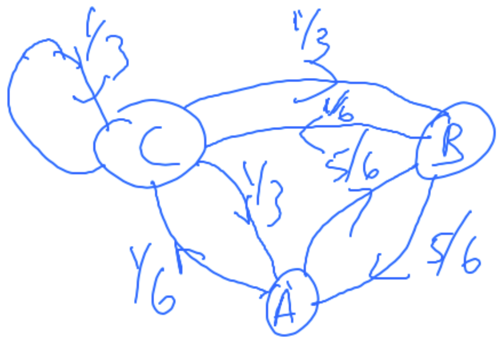
5) Check that the stationary distribution satisfies the detailed balance

$$p_{ji}\,\pi_i^* = p_{ij}\,\pi_j^*$$



$\pi^* = p\,\pi^*$

$$\pi^{(k)} = \lambda_1^k \left(a_1 u_1 + a_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k u_2 + \ldots + a_n \left(\frac{\lambda_n}{\lambda_1}\right)^k u_n \right)$$



Markov chain with 3 states

1) transition matrix $P = \begin{pmatrix} 0 & 5/6 & 1/3 \\ 5/6 & 0 & 1/3 \\ 1/6 & 1/6 & 1/3 \end{pmatrix}$

2) Obvious

3) EVs $\lambda_1 = 1, \ \lambda_2 = +1/6, \ \lambda_3 = -5/6$

eigenvectors

$$\pi^* = \left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right)^T, \ u_2 = \left(-\frac{1}{2}, -\frac{1}{2}, 1\right)^T, \ u_3 = (-1, 1, 0)^T$$

4) $\pi(0) = (1, 0, 0)^T = \pi^* - \frac{u_2}{5} - \frac{u_3}{2}$

$$\pi(t) = P^t \pi(0) = \pi^* - \frac{\lambda_2^t}{5} u_2 - \frac{\lambda_3^t}{2} u_3$$

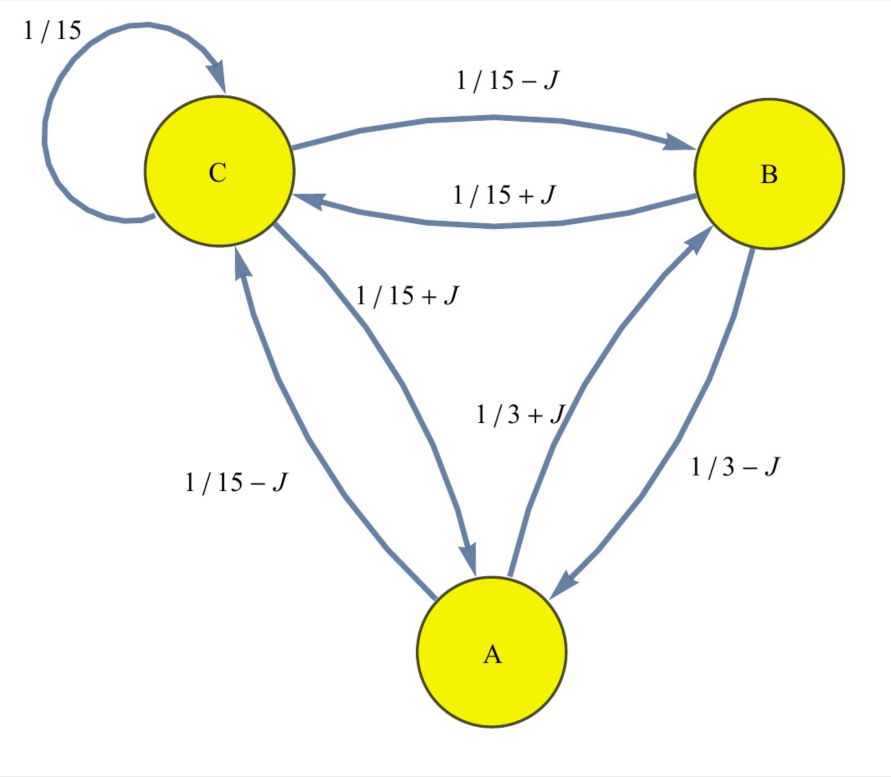
$$\lambda_1 = 1 \geq |\lambda_2| \geq |\lambda_3|$$

$$|\pi(t) - \pi^*| \underset{t \gg 1}{\simeq} \frac{|\lambda_2|^t}{5} |u_2| = \frac{|u_2|}{5} e^{-t \underbrace{\ln \frac{1}{|\lambda_2|}}_{1/\tau}}$$

Mixing time $\tau = \frac{1}{\ln \left|\frac{1}{\lambda_2}\right|} = -\frac{1}{\ln|\lambda_2|}$.

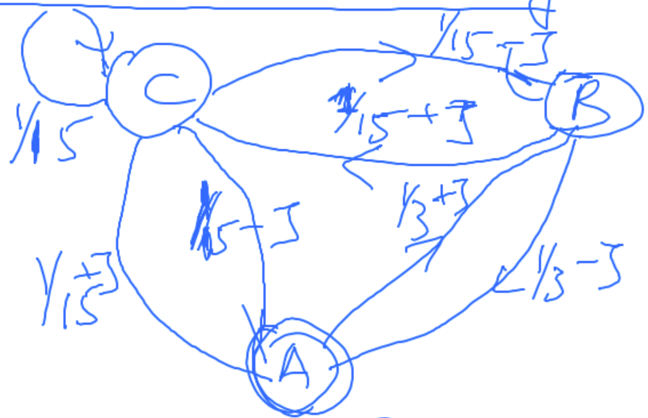
Markov Chains

Accelerated mixing



$$\pi^{(k)}=\lambda_1^k\left(a_1 u_1+a_2\left(\frac{\lambda_2}{\lambda_1}\right)^k u_2+\ldots+a_n\left(\frac{\lambda_n}{\lambda_1}\right)^k u_n\right)$$

Efficient mixing



$J = p\,\pi^*$

$$\bar{P} = \begin{pmatrix} 0 & \frac{5}{6} - \frac{5J}{2} & \frac{1}{3} + 5J \\ \frac{5}{6} + \frac{5J}{2} & 0 & \frac{1}{3} - 5J \\ \frac{1}{6} - \frac{5J}{2} & \frac{1}{6} + \frac{5J}{2} & 1/3 \end{pmatrix}$$

$Q_{ij} = p_{ij}\pi_j^*$

All $p_{ij} \geq 0$

$\Rightarrow J \leq 1/15$

eigenvalues of $\bar{P}$:

$$\lambda_1 = 1, \quad \lambda_{2,3} = \frac{1}{6}\left(-2 \pm 3\sqrt{1 - 125 J^2}\right)$$

$J_{opt}^2 = 1/125 \qquad |W_{opt}| = 1/3 = \min_J (|\lambda_2|, |\lambda_3|)$

Markov Chains

Exercise: The data-processing theorem

Data processing can only destroy information.

Prove this theorem by considering an ensemble WDR in which w is the state of the world, d is data gathered, and *r* is the processed data, so that these three variables form a Markov chain $w \to d \to r$,
that is, the probability P (w, d, r) can be written as

$$P(w, d, r) = P(w)P(d \mid w)P(r \mid d)$$

Show that the average information that R conveys about W, mutual information I(W;R), is less than or equal to the average information that D conveys about W, I(W;D).

The data processing theorem

$W \to D \to R$

$$P(w,d,r) = P(w)P(d|w)P(r|d)$$

$$I(W;R) \le I(W;D)$$

$\cancel{I(X;Y;Z)}$

$I(X;Y,Z)$

$$I(X;Y) = H(X) - H(X|Y)$$

$$I(X;Y|Z=c_k) = H(X|Z=c_k) - H(X|Y,Z=c_k)$$

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$$

$$I(X;Y,Z) = I(X;Y) + I(X;Z|Y)$$

$w \to d \to r$

$I(W;R|D) = 0$

$$\begin{cases} I(W;D,R) = I(W;D) + I(W;R|D) = 0 \\ I(W;D,R) = I(W;R) + I(W;D|R) \end{cases}$$

$$\Downarrow$$

$$I(W;R) - I(W;D) = -I(W;D|R) \le 0$$

Skoltech
Skolkovo Institute of Science and Technology

Markov Chains

Literature

1. Richard Serfozo, Basics of Applied Stochastic Processes, 2009

2. David A. Levin. Yuval Peres. Elizabeth L. Wilmer, Markov Chains and Mixing Times, 2009 (2nd edition 2017)

3. Zillion of other resources