Institut für Kartographie und Geoinformatik | Leibniz Universität Hannover

Simulation and Markov chain Monte Carlo II

Claus.Brenner@ikg.uni-hannover.de

Leibniz Universität Hannover

What we have learned so far…

- We want to compute expectations or maxima of distributions given by their probabilities *p(x)*

$$\mathrm{E}[f] = \int f(x)\,p(x)\,dx \qquad \text{or} \qquad x^* = \arg\max_x p(x)$$

- An approximate solution is drawing samples and computing the following terms

$$\mathrm{E}[f] \approx \frac{1}{m}\sum_{j=1}^{m} f(x_j) \qquad \text{or} \qquad x^* = \arg\max_{x_j} p(x_j)$$

- Therefore the goal is to efficiently draw samples $x_j \sim p(x)$
  - Accept/ Reject und Importance Sampling have drawbacks
- Now: Generation of samples using Markov chain Monte Carlo (MCMC)

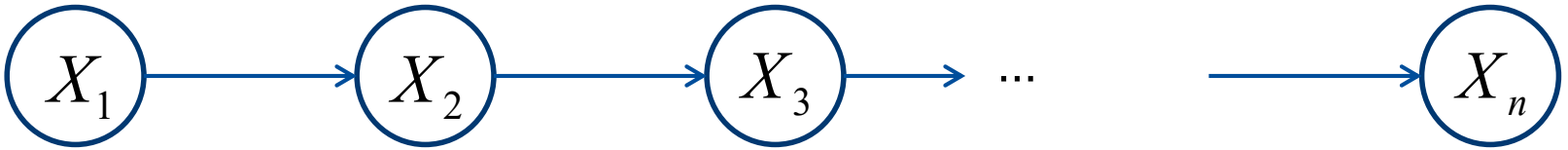
Simulation and Markov chain Monte Carlo

- Markov chains
  - Definition, homogeneity, transition kernel
  - Example: weather; convergence
- Properties of Markov chains
  - Irreducibility, periodicity, recurrence, ergodic chains
- Markov chain Monte Carlo
  - Principle: drawing of samples using a Markov chain
  - The algorithm of Metropolis-Hastings and Metropolis
  - Stationary distributions and convergence
- Examples
- Appendix:
  - A graphical model / Bayes network view.

Markov chains

Markov chains

- Andrei Andrejewitsch Markov, 1856-1922 (also: Markow, Markoff)
- A Markov chain is a Bayes network of random variables, where each variable X_i depends only on its immediate predecessor X_{i-1}

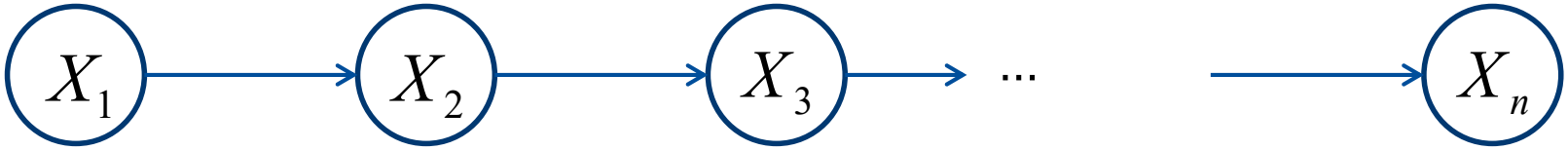


- Independence is expressed by:

$$P(X_i \in A \mid X_1, \ldots X_{i-1}) = P(X_i \in A \mid X_{i-1})$$

- Time discrete stochastic process: $i \in \{1,2,3\ldots\}$
- State space S with $X_i \in S$
  - may be discrete or continuous.

Markov chains



- State transitions are defined by a transition kernel:

$$T(X_i \mid X_{i-1})$$

  - discrete state space: conditional probability
  - continuous state space: conditional density

- If the transition kernel does not depend on the time (i.e., the index i) the chain is called homogeneous.

Example for a Markov chain: "weather"

- Example: discrete, finite state space

$$S = \{\text{sunny}, \text{cloudy}, \text{rainy}\}$$

- Definition of the transition kernel:
  - How will be the weather X_i tomorrow, if it is $X_{i-1} = \text{sunny}$ today?

$$T(X_i = \text{sunny} \mid X_{i-1} = \text{sunny}) = 0.7$$
$$T(X_i = \text{cloudy} \mid X_{i-1} = \text{sunny}) = 0.2$$
$$T(X_i = \text{rainy} \mid X_{i-1} = \text{sunny}) = 0.1$$

  - Values have to add up to 1.0, since $T(X_i \mid X_{i-1} = \text{sunny})$ is a (conditional) probability
  - Similarly, define:

$$T(X_i \mid X_{i-1} = \text{cloudy})$$
$$T(X_i \mid X_{i-1} = \text{rainy})$$

Example for a Markov chain: "weather"

- Overall definition of the (homogeneous, time discrete) transition kernel:

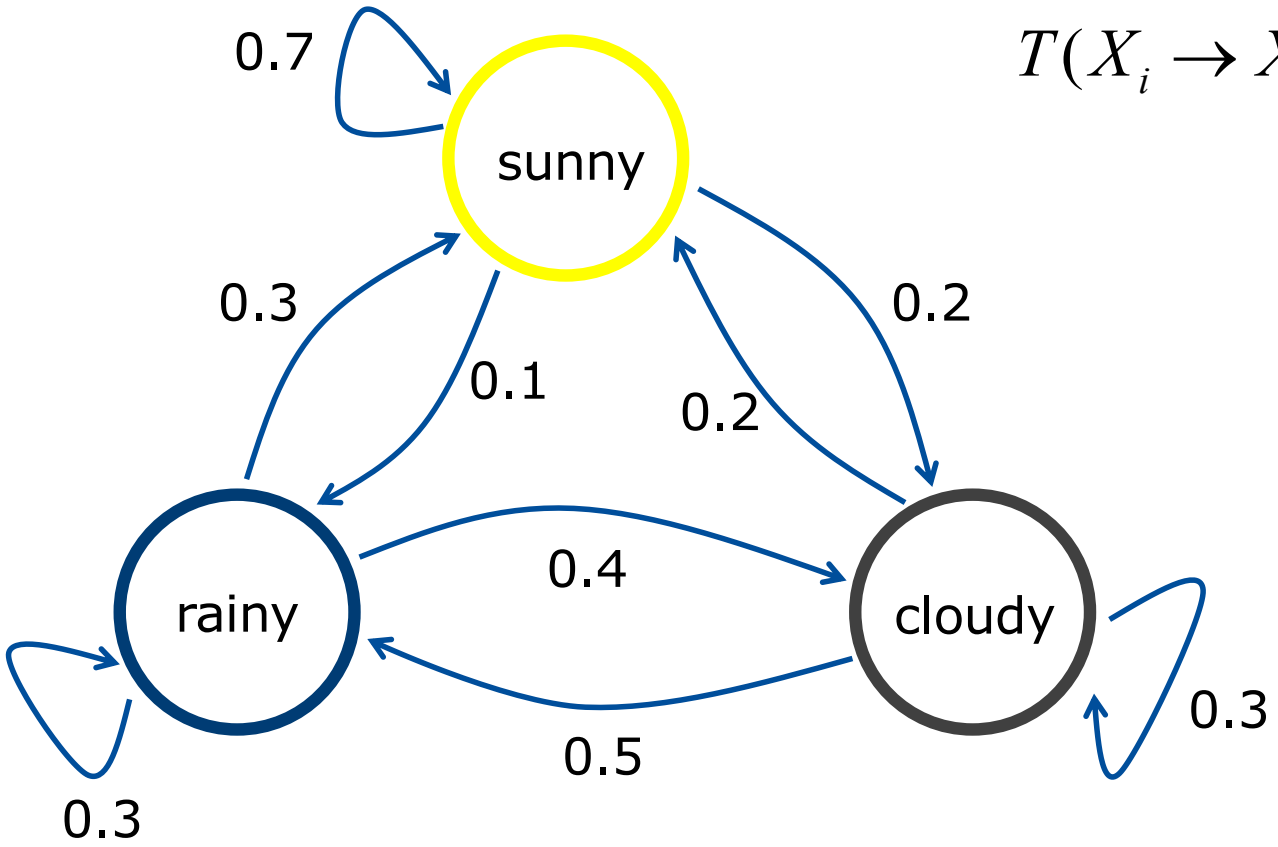
$$T(X_j \mid X_i) = T(X_i \to X_j) = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.2 & 0.3 & 0.5 \\ 0.3 & 0.4 & 0.3 \end{bmatrix}$$

$T(X_j \mid X_i = \text{sunny})$

- Each row must sum up to 1.0

Example for a Markov chain: “weather”

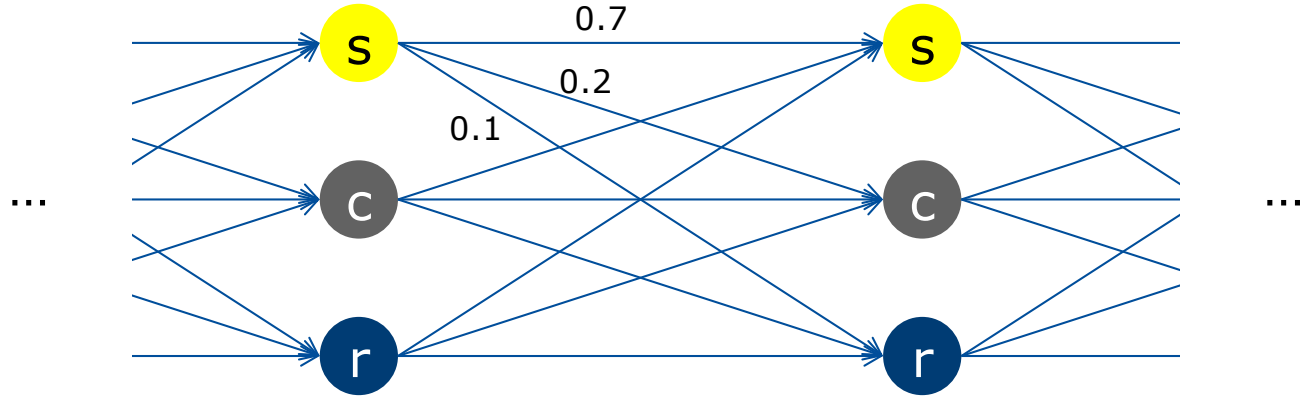
- Representation as a state/ transition diagram:



$$T(X_i \rightarrow X_j) = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.2 & 0.3 & 0.5 \\ 0.3 & 0.4 & 0.3 \end{bmatrix}$$

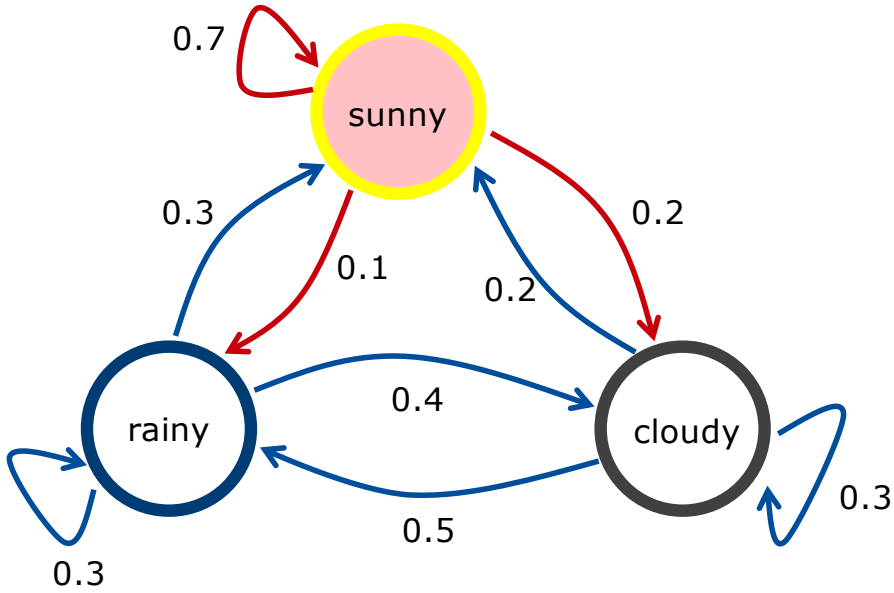
Illustration: Markov chain and states

- The state diagram is NOT a “Bayes network”!
- Illustration of the Bayes network and the states:



$P(X_i)$ $T(X_j \mid X_i)$ $P(X_j)$

"Weather" example



- Today, it is sunny.

$$P(X_i = \text{sunny}) = 1.0$$
$$P(X_i) = \begin{bmatrix} 1.0 & 0.0 & 0.0 \end{bmatrix}$$

- How will it be tomorrow?

$$P(X_{i+1}) = P(X_i) \cdot T = \begin{bmatrix} 1.0 & 0.0 & 0.0 \end{bmatrix} \cdot \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.2 & 0.3 & 0.5 \\ 0.3 & 0.4 & 0.3 \end{bmatrix}$$

$$= \begin{bmatrix} 0.7 & 0.2 & 0.1 \end{bmatrix}$$

“Weather” example: the day after tomorrow

- How will be the weather the day after tomorrow?

$$P(X_{i+2}) = P(X_{i+1}) \cdot T = P(X_i) \cdot T \cdot T$$

$$= \begin{bmatrix} 1.0 & 0.0 & 0.0 \end{bmatrix} \cdot \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.2 & 0.3 & 0.5 \\ 0.3 & 0.4 & 0.3 \end{bmatrix}^2$$

$$= \begin{bmatrix} 1.0 & 0.0 & 0.0 \end{bmatrix} \cdot \begin{bmatrix} 0.56 & 0.24 & 0.2 \\ 0.35 & 0.33 & 0.32 \\ 0.38 & 0.3 & 0.32 \end{bmatrix}$$

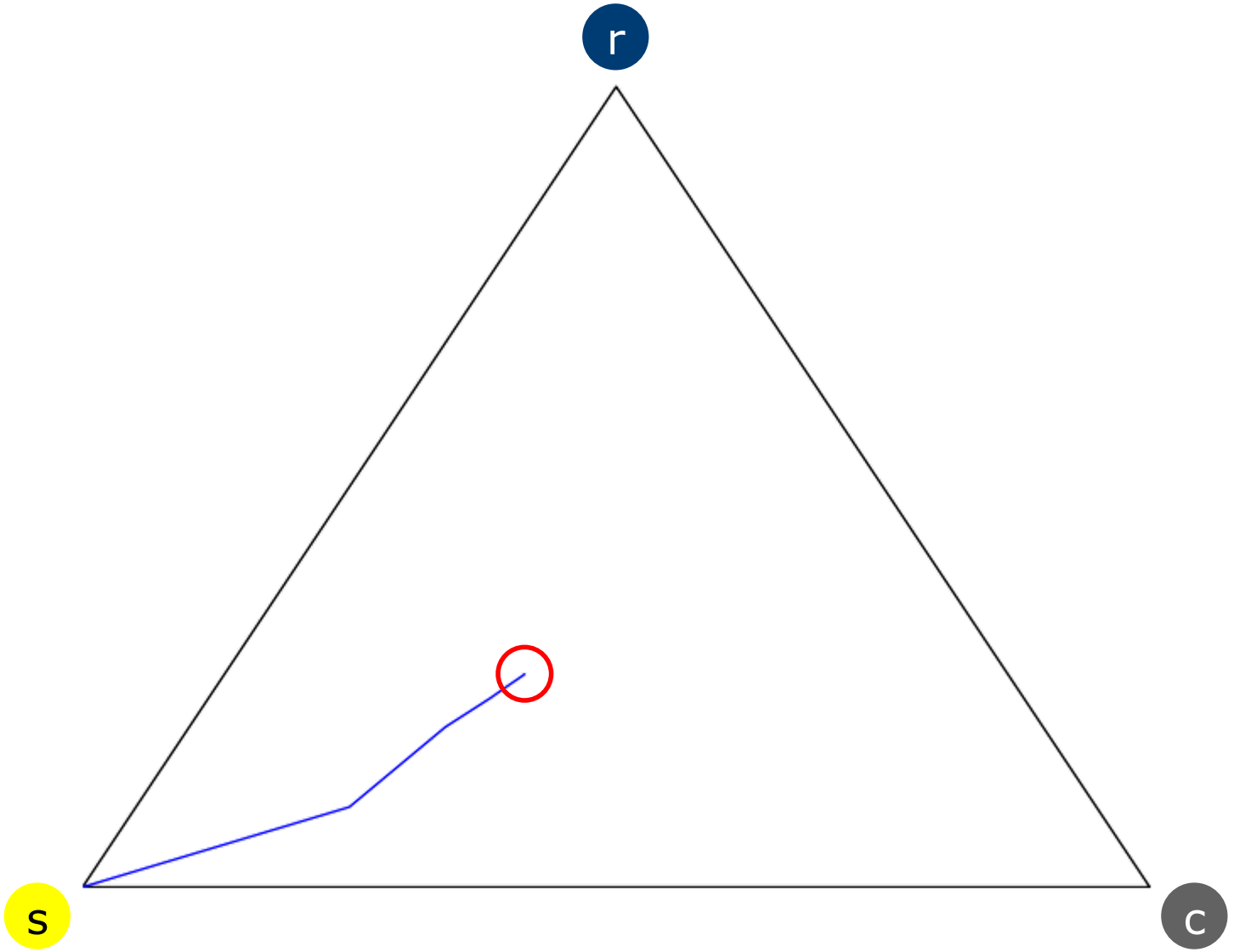
$$= \begin{bmatrix} 0.56 & 0.24 & 0.2 \end{bmatrix}$$

“Weather” example: convergence for [1 0 0]

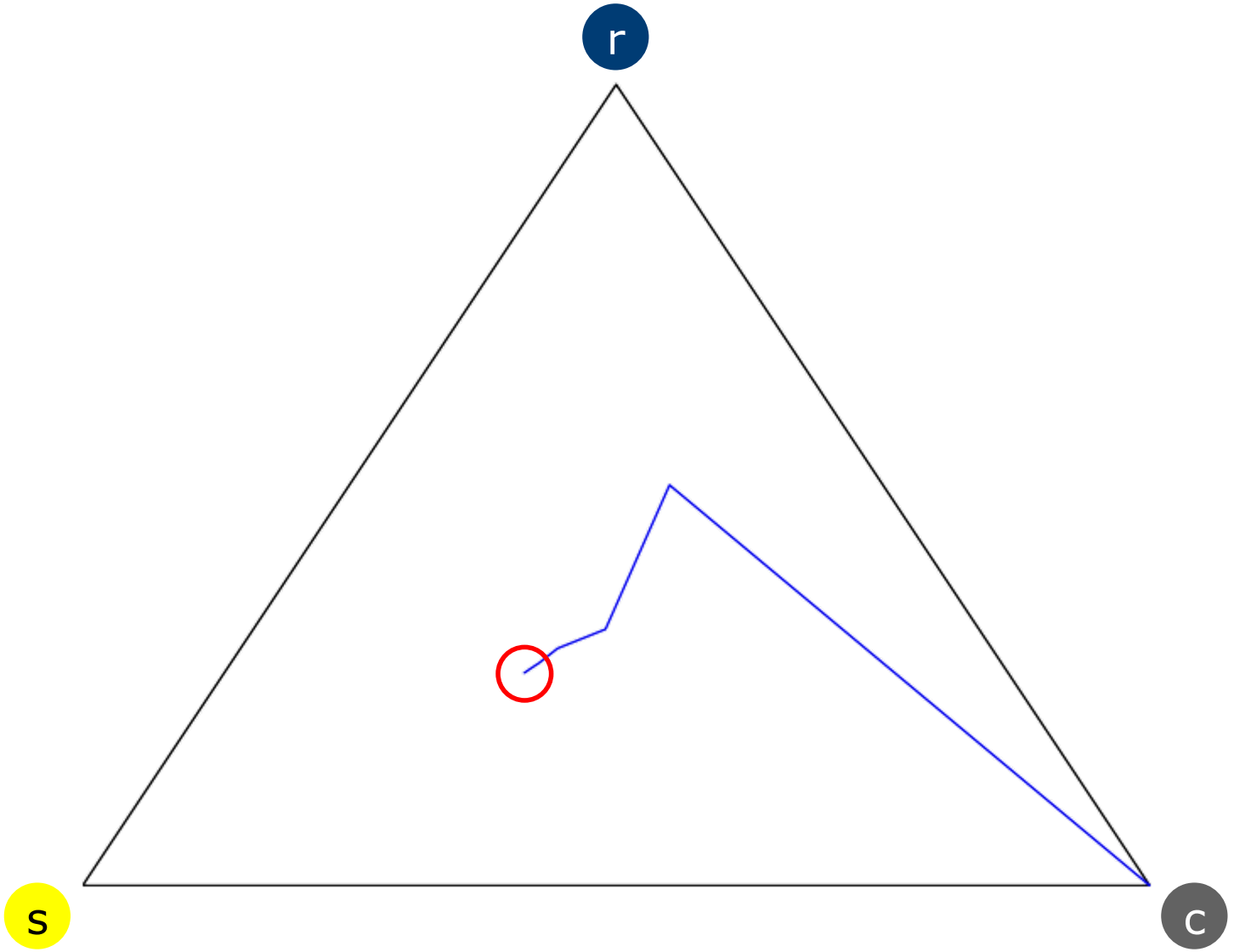
- Evolution of the distribution when starting with [1 0 0]
  - As can be seen: it converges!

```
 0: [ 1.0000  0.0000  0.0000 ]
 1: [ 0.7000  0.2000  0.1000 ]
 2: [ 0.5600  0.2400  0.2000 ]
 3: [ 0.5000  0.2640  0.2360 ]
 4: [ 0.4736  0.2736  0.2528 ]
 5: [ 0.4621  0.2779  0.2600 ]
 6: [ 0.4570  0.2798  0.2632 ]
 7: [ 0.4548  0.2806  0.2646 ]
 8: [ 0.4539  0.2810  0.2652 ]
 9: [ 0.4535  0.2811  0.2654 ]
10: [ 0.4533  0.2812  0.2655 ]
11: [ 0.4532  0.2812  0.2656 ]
12: [ 0.4532  0.2812  0.2656 ]
13: [ 0.4531  0.2812  0.2656 ]
14: [ 0.4531  0.2812  0.2656 ]
15: [ 0.4531  0.2812  0.2656 ]
```

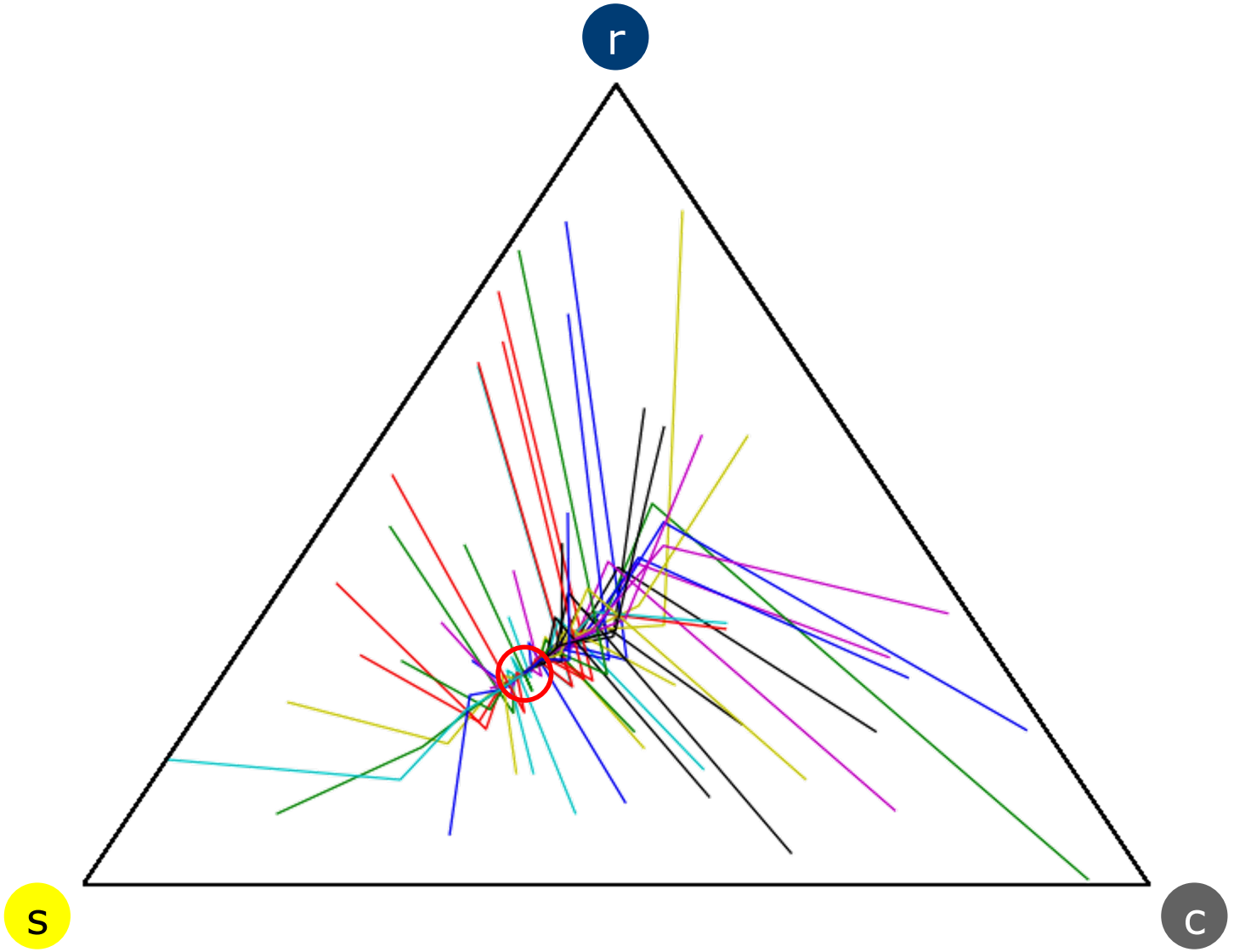
“Weather” example: convergence for [1 0 0]



“Weather” example: convergence for [0 1 0]



“Weather” example: convergence for a random choice of initial states



Convergence

- Does the distribution converge for arbitrary initial distributions?

$$P(X_1)$$
$$P(X_2) = P(X_1) \cdot T$$
$$P(X_3) = P(X_1) \cdot T^2$$
$$\vdots$$
$$P(X_n) = P(X_1) \cdot T^{n-1}$$

- Therefore, it is interesting how T^i evolves

Convergence: powers of the transition kernel

$T^1 =$
```
[[ 0.7  0.2  0.1]
 [ 0.2  0.3  0.5]
 [ 0.3  0.4  0.3]]
```

$T^5 =$
```
[[ 0.46208  0.27792  0.26   ]
 [ 0.44435  0.28449  0.27116]
 [ 0.44714  0.2835   0.26936]]
```

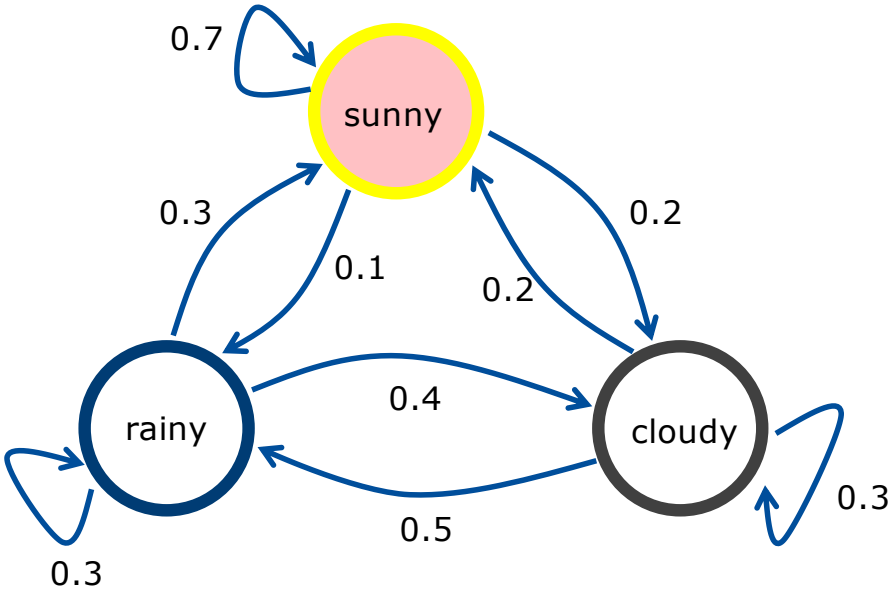
$T^{10} =$
```
[[ 0.45326808  0.28119673  0.26553519]
 [ 0.45298486  0.28130217  0.26571297]
 [ 0.45302931  0.28128562  0.26568507]]
```

$T^{25} =$
```
[[ 0.453125  0.28125   0.265625]
 [ 0.453125  0.28125   0.265625]
 [ 0.453125  0.28125   0.265625]]
```

- All rows are identical (and all elements are > 0)
- → Convergence: after a sufficient number of steps, the distribution is independent of the initial distribution.

What is the consequence for a single sequence of draws?

- Example sequence of 100 draws:



$$\sim P(X_1) = [1.0 \quad 0.0 \quad 0.0]$$

s, r, c, c, s, s, r, r, c, s, s, c, c, c, r, s, c, r, s, s, s, s, c, c, s, s, s, c, r, c, r, r, r, r, c, s, s, s, s, s, s, c, c, r, c, r, c, c, c, r, s, c, s, s, s, r, s, s, c, r, r, r, s, c, s, s, s, s, s, c, r, r, c, r, c, c, r, c, r, c, c, s, s, r, r, r, r, c, c, r, s, s, s, s, r, r, r, r, r, s

$$\sim P(X_i) \approx [0.4531 \quad 0.2812 \quad 0.2656] \text{ for } i \geq 13$$

$$\sim P(X_5) = [0.4621 \quad 0.2779 \quad 0.2600]$$

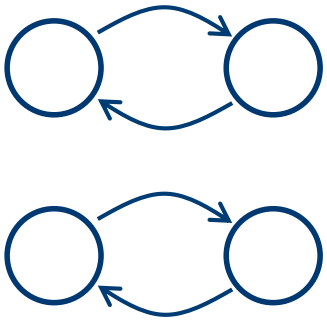
Properties of Markov chains

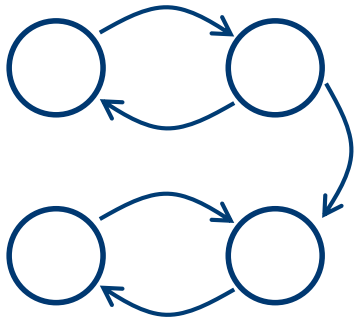
Properties of Markov chains

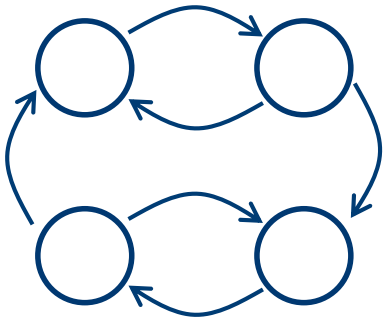
- We already know:
  - Homogeneity: the transition kernel is constant (in time/ in index)
- Which properties have to be fulfilled so that a stationary distribution exists to which the Markov chain converges?

Irreducibility

- A homogeneous Markov chain is termed irreducible, if for all states *u*, *v* the following holds:
  - There exists a sequence of states $u = s_1, s_2, \ldots s_n = v$ which satisfies: $T(s_{i+1} \mid s_i) > 0$
- An irreducible Markov chain can reach every state, when started from an arbitrary initial state, with positive probability
- = "probabilistically connected".





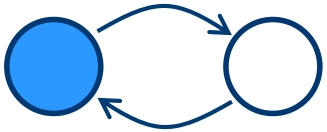


Periodicity

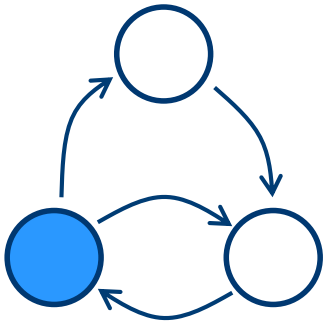
- The period of a state *u* is defined as follows:

$$\gcd\{m \geq 1 \mid T^m(u,u) > 0\}$$

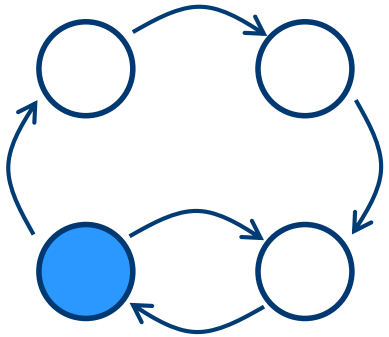
- I.e., the greatest common divisor of all "return times" to the original state *u*
- A homogeneous Markov chain is aperiodic, if all states have a period of 1.



$\gcd\{2, 4, 6, \ldots\} = 2$



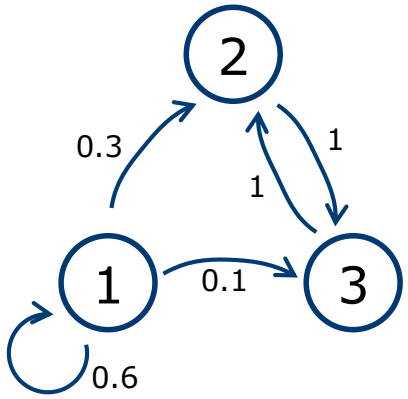
$\gcd\{2, 3, 4, 5, 6, \ldots\} = 1$

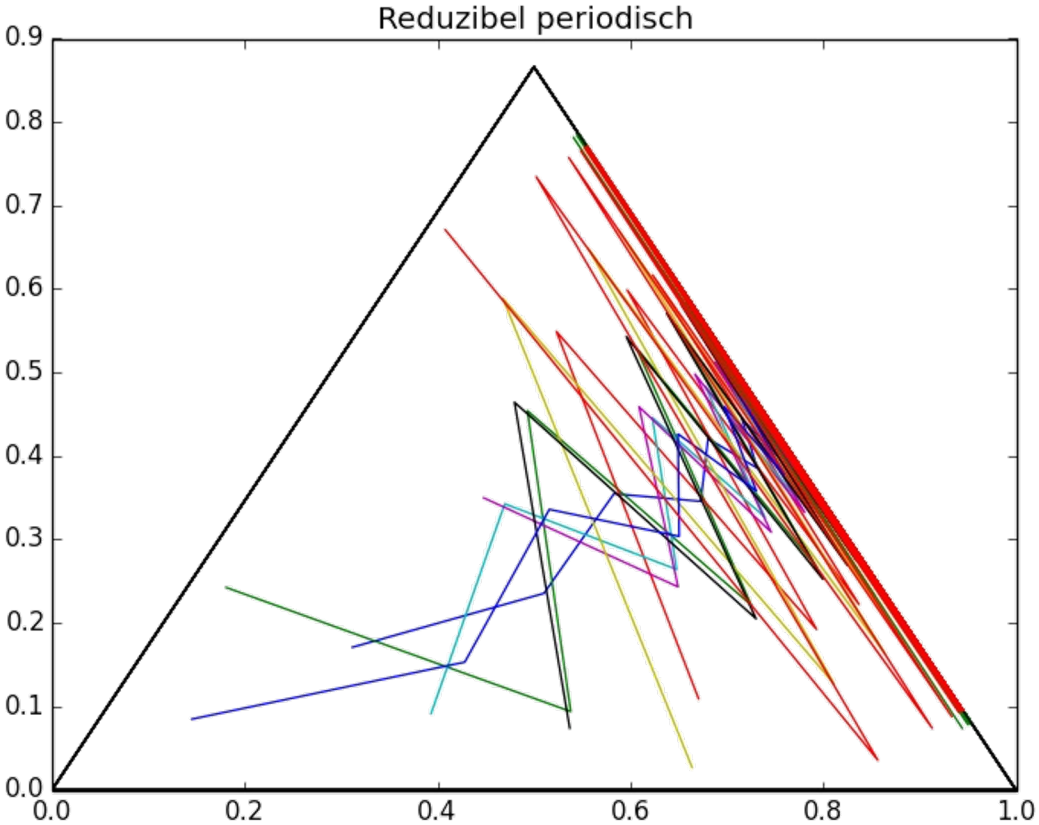


$\gcd\{2, 4, 6, 8, \ldots\} = 2$

Example: reducible, periodic

```
T
[[ 0.6  0.3  0.1]
 [ 0.   0.   1. ]
 [ 0.   1.   0. ]]
```





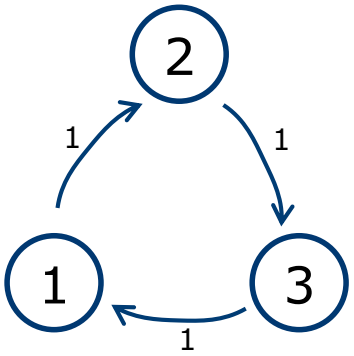

```
T**18
[[ 0.     0.437  0.562]
 [ 0.     1.     0.   ]
 [ 0.     0.     1.   ]]
```

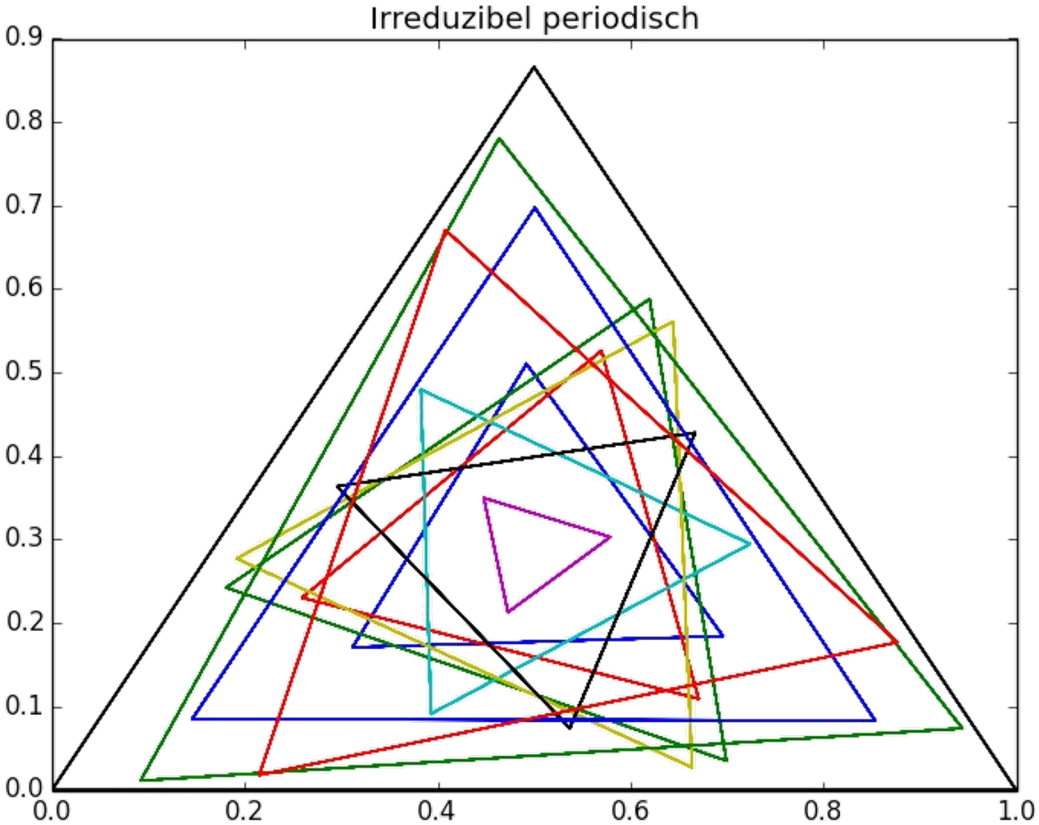
```
T**19
[[ 0.     0.562  0.437]
 [ 0.     0.     1.   ]
 [ 0.     1.     0.   ]]
```

```
T**20
[[ 0.     0.437  0.562]
 [ 0.     1.     0.   ]
 [ 0.     0.     1.   ]]
```

Example: irreducible, periodic

```
T
[[ 0.   1.   0.]
 [ 0.   0.   1.]
 [ 1.   0.   0.]]
```





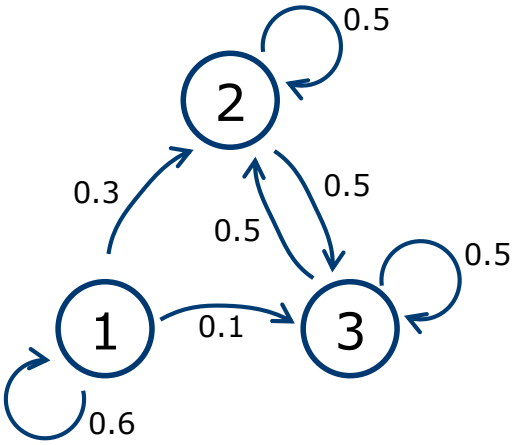

```
T**18
[[ 1.   0.   0.]
 [ 0.   1.   0.]
 [ 0.   0.   1.]]
```

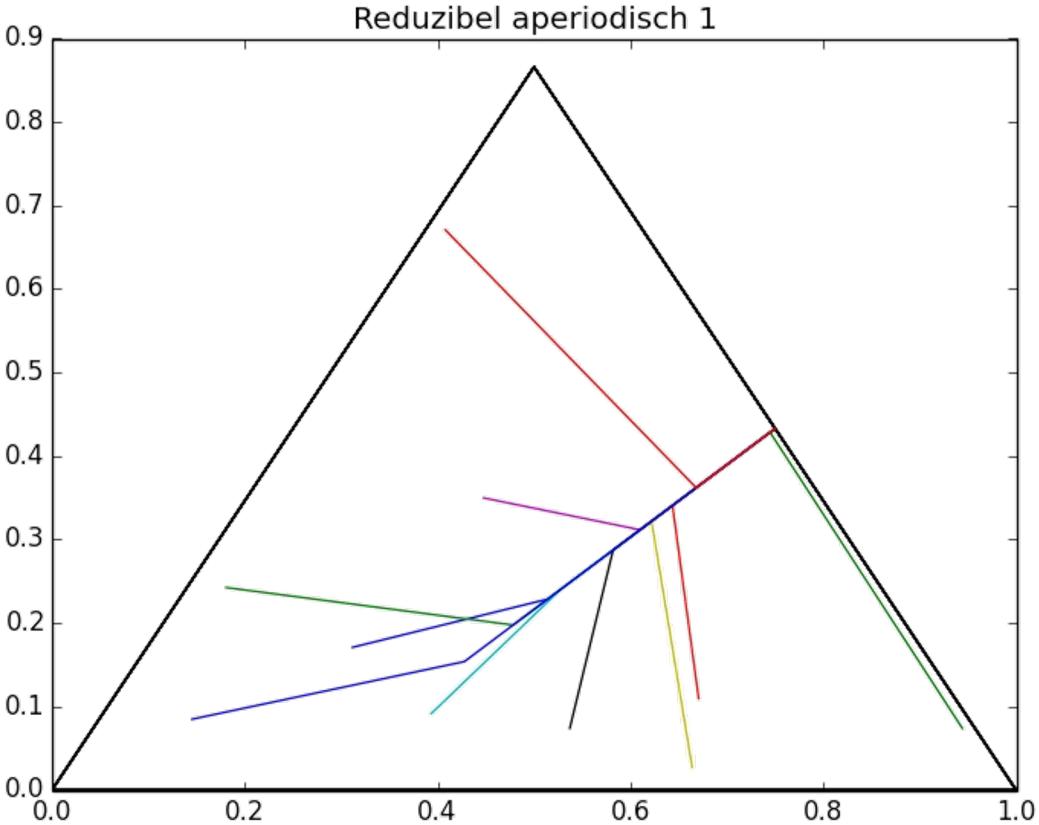
```
T**19
[[ 0.   1.   0.]
 [ 0.   0.   1.]
 [ 1.   0.   0.]]
```

```
T**20
[[ 0.   0.   1.]
 [ 1.   0.   0.]
 [ 0.   1.   0.]]
```

Example: reducible, aperiodic (1)

```
T
[[ 0.6  0.3  0.1]
 [ 0.   0.5  0.5]
 [ 0.   0.5  0.5]]
```





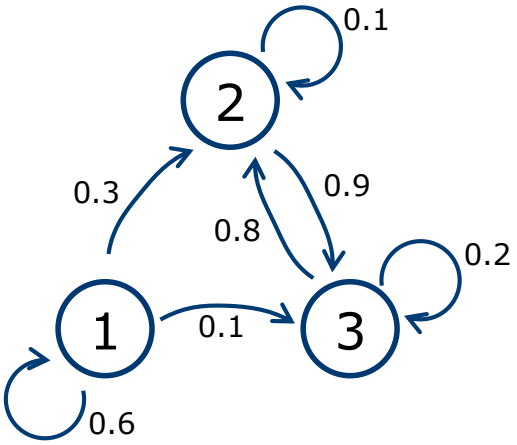

```
T**18
[[ 0.   0.5  0.5]
 [ 0.   0.5  0.5]
 [ 0.   0.5  0.5]]
```

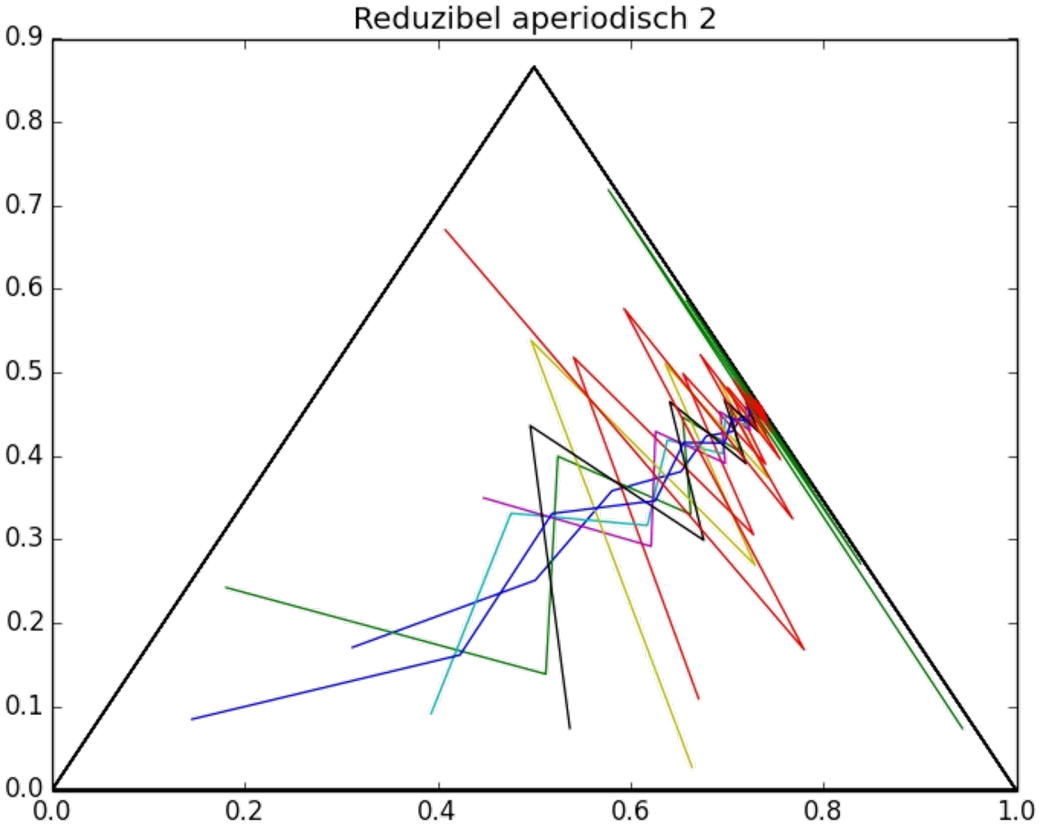
```
T**19
[[ 0.   0.5  0.5]
 [ 0.   0.5  0.5]
 [ 0.   0.5  0.5]]
```

```
T**20
[[ 0.   0.5  0.5]
 [ 0.   0.5  0.5]
 [ 0.   0.5  0.5]]
```

Example: reducible, aperiodic (2)

```
T
[[ 0.6   0.3   0.1]
 [ 0.    0.1   0.9]
 [ 0.    0.8   0.2]]
```





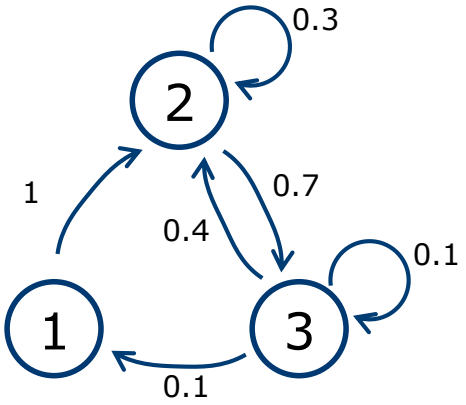

```
T**18
[[ 0.      0.47    0.529]
 [ 0.      0.471   0.529]
 [ 0.      0.47    0.53 ]]
```

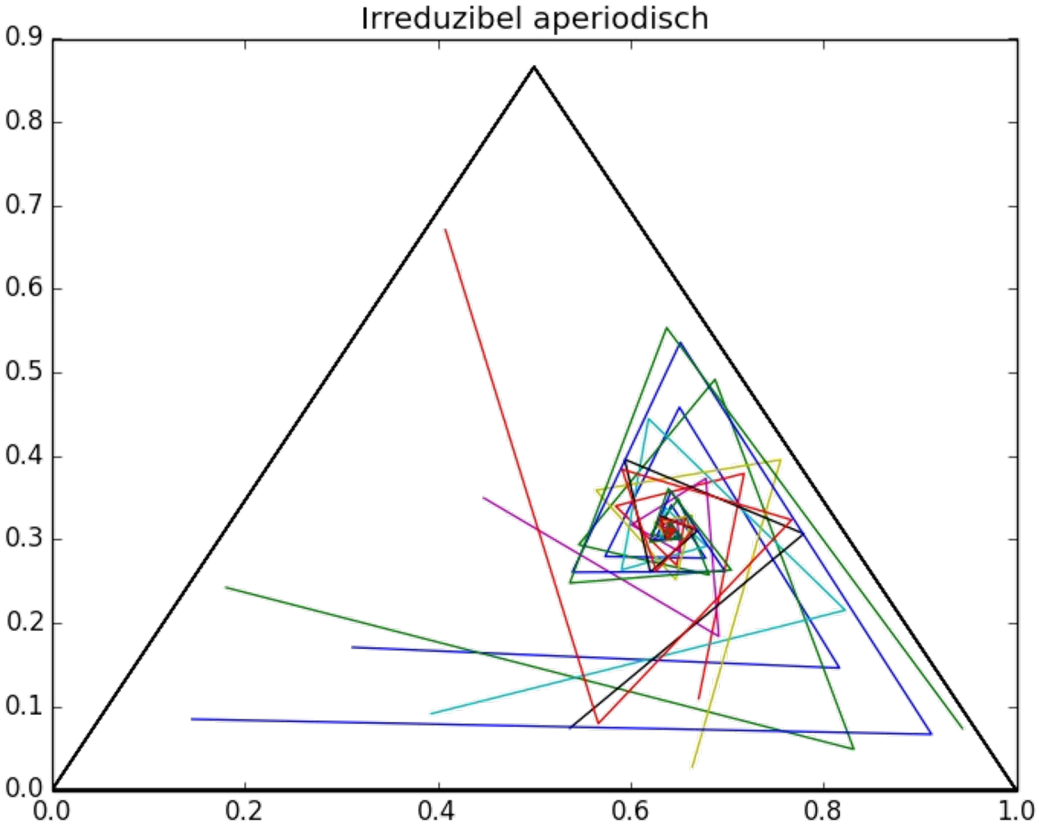
```
T**19
[[ 0.      0.471   0.529]
 [ 0.      0.47    0.53 ]
 [ 0.      0.471   0.529]]
```

```
T**20
[[ 0.      0.471   0.529]
 [ 0.      0.471   0.529]
 [ 0.      0.47    0.53 ]]
```

Example: irreducible, aperiodic

```
T
[[ 0.   1.   0. ]
 [ 0.   0.3  0.7]
 [ 0.5  0.4  0.1]]
```



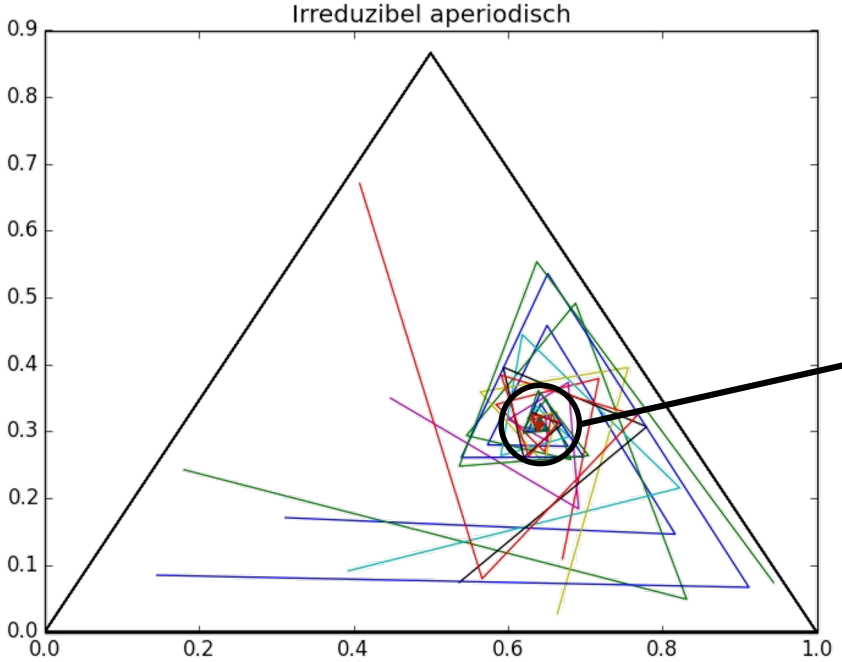



```
T**18
[[ 0.18   0.462  0.359]
 [ 0.179  0.462  0.359]
 [ 0.179  0.461  0.359]]
```

```
T**19
[[ 0.179  0.462  0.359]
 [ 0.179  0.462  0.359]
 [ 0.18   0.462  0.359]]
```

```
T**20
[[ 0.179  0.462  0.359]
 [ 0.18   0.462  0.359]
 [ 0.179  0.462  0.359]]
```

Convergence in distribution vs. in draws




```
T**20
[[ 0.179   0.462   0.359]
 [ 0.18    0.462   0.359]
 [ 0.179   0.462   0.359]]
```

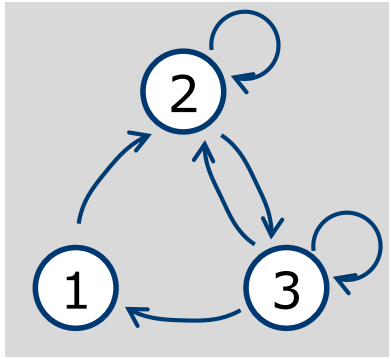
[0.179, 0.462, 0.359]

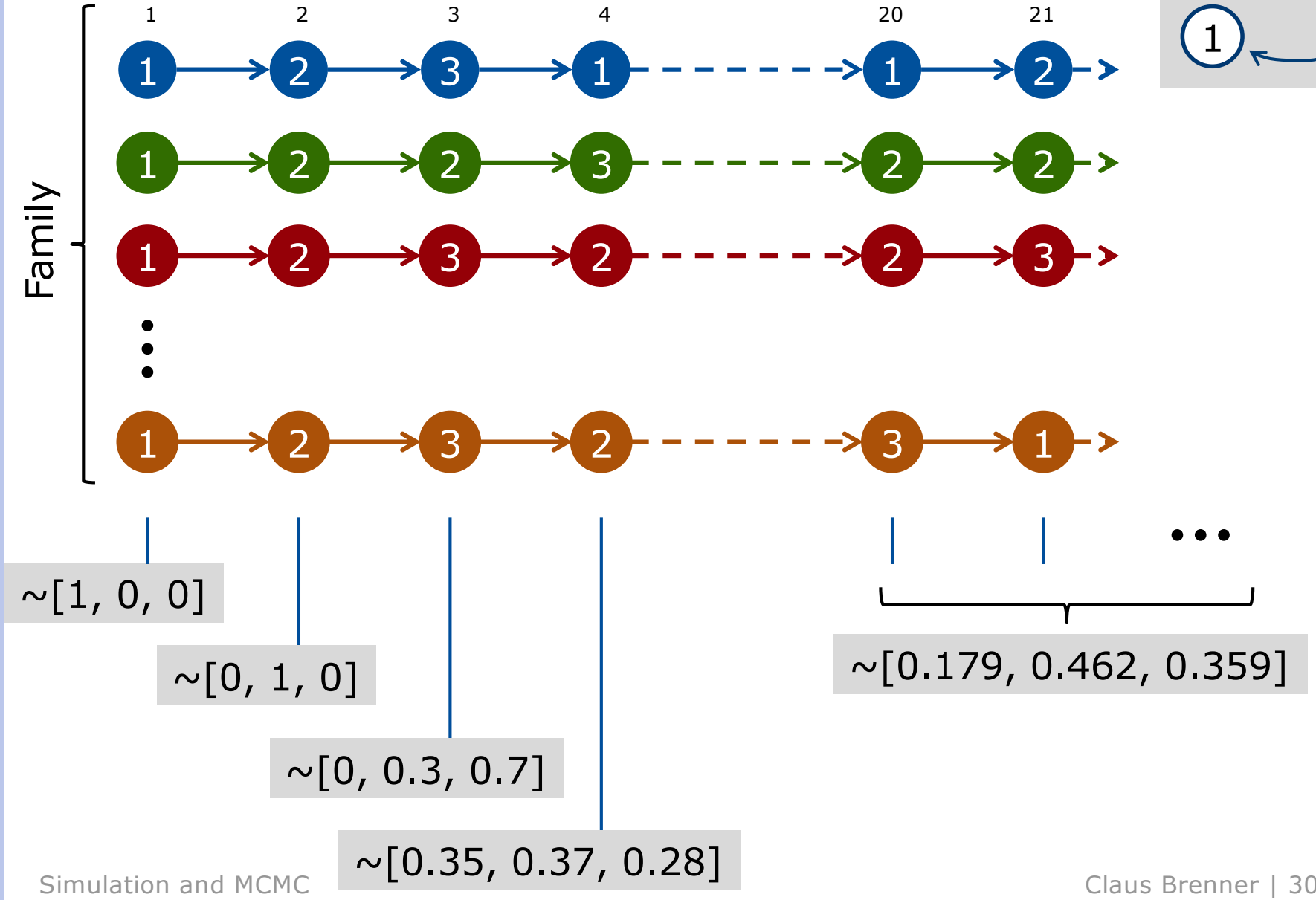
$$\begin{bmatrix} u & v & w \end{bmatrix} \cdot T^n = \begin{bmatrix} 0.179 & 0.462 & 0.359 \end{bmatrix}$$

u, v, w arbitrary with $u + v + w = 1$

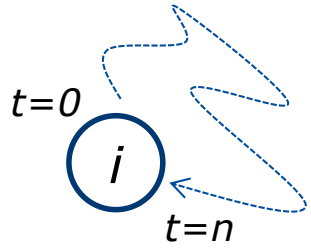
- Distribution:
  - Independently of the initial distribution, after some iterations n we have: $P(X_n=1)=0.179$, $P(X_n=2)=0.462$, $P(X_n=3)=0.359$
- Draw:
  - "Starting with the nth draw", all $x_k \sim P(X_n)$ for $k>n$

Convergence in distribution vs. in draws





Recurrent and positive recurrent chains



- Homogeneous, irreducible Markov chain
- Define: recurrence time

$$\tau_{ii} := \min\{n > 0 : X_n = i \mid X_0 = i\}$$

- A state is termed recurrent, if: $P(\tau_{ii} < \infty) = 1$
  - I.e., the state will be re-visited with probability 1 (almost surely)
- A recurrent state is termed positive recurrent, if: $\mathrm{E}[\tau_{ii}] < \infty$
  - The expected return time is finite
- A Markov chain is recurrent if each of its states is recurrent
- (If the state space is finite: irreducibility → positive recurrence).

Ergodic Markov chains

- A homogeneous, irreducible Markov chain has a stationary distribution

$$\pi \cdot T = \pi$$

  if and only if it is positive recurrent
- A homogeneous, irreducible, positive recurrent Markov chain converges to its (unique) stationary distribution, if it is aperiodic
- Such a chain is termed ergodic.

Properties of ergodic Markov chains

- Convergence in distribution:

$$\left(T^n\right)_i \to \pi \quad \text{for } n \to \infty \text{ and all } i$$

  - All rows of the transition kernel, raised to the power of n, converge to the stationary distribution
  - (and they do not contain zeros)

- Convergence of the ergodic mean:

$$\bar{f}_m := \frac{1}{m}\sum_{j=1}^{m} f(X_j)$$

(ergodic mean, average over time
not: average over family)

$$\text{if}: \ \mathrm{E}_\pi\left[\left|f(X)\right|\right] < \infty \ \text{ then } \ P\left[\bar{f}_m \to E_\pi\left[f(X)\right]\right] = 1$$

  - The ergodic mean converges to the expected value, with probability 1.

Further remarks

- From ergodicity follows convergence, but no implications w.r.t.
  - How many iterations are necessary until the samples are sufficiently well π distributed?
  - How large is the error of an estimate with given *m*?

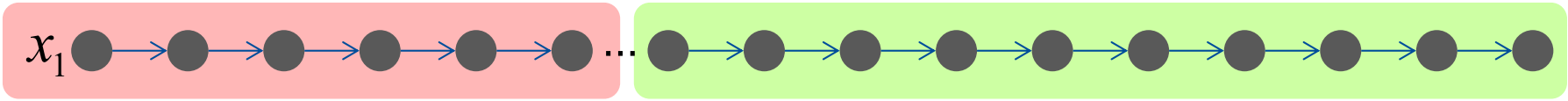
$$\left| \frac{1}{m} \sum_{j=1}^{m} f(x_j) - \mathrm{E}_{\pi}[f(X)] \right| \leq ?$$

- The notions and definitions can be extended to the continuous case (continuous densities instead of discrete distributions), not shown here.

Markov chain Monte Carlo

Markov chain Monte Carlo: drawing samples using a Markov chain

- Idea:
  - Given: a Markov chain (by the definition of a transition kernel)…
  - …which converges to a stationary distribution (as in the examples above)
  - Then, starting from an arbitrary initial state (from an arbitrary distribution)…
  - …after a sufficiently long "burn in" phase, during which one simply discards the samples…
  - …one obtains samples from the stationary distribution.



Burn-in, states are dependent on initial state: discard!

Draws from the stationary distribution: keep!

Markov chain Monte Carlo

- Definition:
  - **Any method**, which simulates a distribution f, by generating an ergodic Markov chain with stationary distribution f, is a "Markov chain Monte Carlo" method.

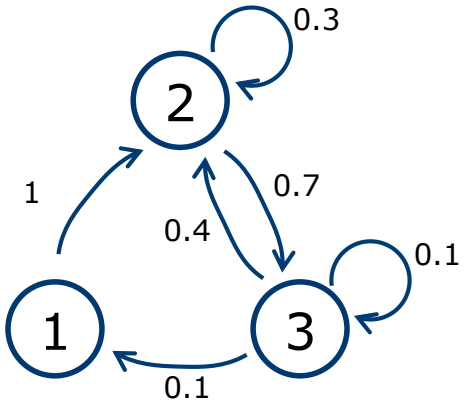
- Well known methods are e.g. :
  - Metropolis (Metropolis et al., 1953)
  - Metropolis-Hastings (extension by Hastings, 1970)
  - Gibbs sampling (Geman & Geman, 1984)
  - Slice sampling (Neal, 2003).

How do we have to choose the transition kernel?

- **So far:**

$$T = P(X_{i+1} \mid X_i)$$

Given



$\Rightarrow$

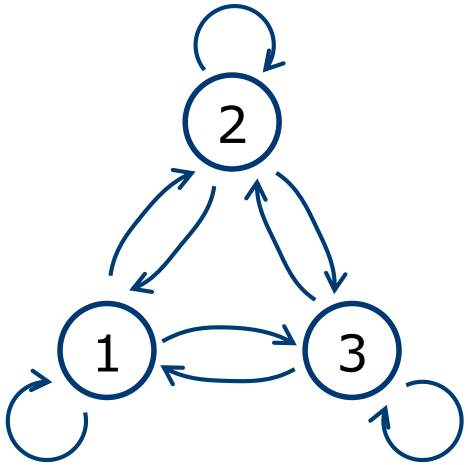
$$\pi = \begin{bmatrix} 0.179 & 0.462 & 0.359 \end{bmatrix}$$

Convergence to the stationary distribution

- **Now:**

$$T = ???$$

Sought for!



$\Leftarrow$

π

Desired stationary distribution

The algorithm of Metropolis-Hastings

$$P(X_{i+1} \mid X_i)$$

- Propose a candidate *Y* for a new state, based on a proposal density (e.g. the normal density) *q*:

$$Y \sim q(\cdot \mid X_i)$$

- Accept this proposal with the following probability

$$\alpha(X_i, Y) := \min\left(\frac{\pi(Y)\, q(X_i \mid Y)}{\pi(X_i) q(Y \mid X_i)}, 1\right)$$

  - if accepted, set: $X_{i+1} = Y$
  - else (if not accepted), set: $X_{i+1} = X_i$ (←repeat!)

The algorithm of Metropolis-Hastings

x = initial value

repeat m times:

draw $y \sim q(\cdot \mid x)$

compute $\alpha = \min\left(\frac{\pi(y)\,q(x \mid y)}{\pi(x)\,q(y \mid x)}, 1\right)$

draw $u \sim \text{uniform}(0,1)$

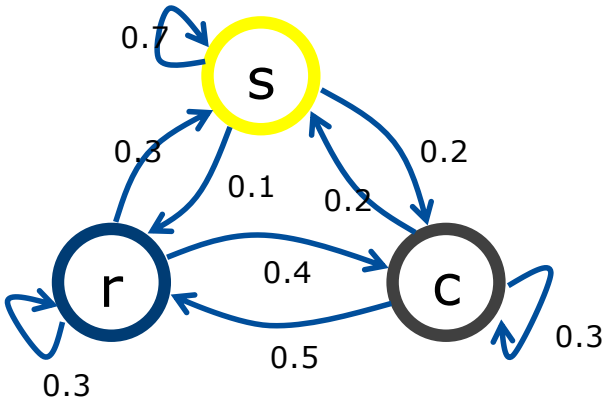
if $u \leq \alpha$:

$x = y$

output x # In any case.

Example sunny – cloudy – rainy

- We defined *T* as follows:



- From which we obtained this distribution:

```
 0: [ 1.0000   0.0000   0.0000 ]
 1: [ 0.7000   0.2000   0.1000 ]
 2: [ 0.5600   0.2400   0.2000 ]
 3: [ 0.5000   0.2640   0.2360 ]
 4: [ 0.4736   0.2736   0.2528 ]
 5: [ 0.4621   0.2779   0.2600 ]
 6: [ 0.4570   0.2798   0.2632 ]
 7: [ 0.4548   0.2806   0.2646 ]
 8: [ 0.4539   0.2810   0.2652 ]
 9: [ 0.4535   0.2811   0.2654 ]
10: [ 0.4533   0.2812   0.2655 ]
11: [ 0.4532   0.2812   0.2656 ]
12: [ 0.4532   0.2812   0.2656 ]
13: [ 0.4531   0.2812   0.2656 ]
14: [ 0.4531   0.2812   0.2656 ]
15: [ 0.4531   0.2812   0.2656 ]
```

Example sunny – cloudy – rainy

- We think this is too much rain!
- We want to generate samples such that:
  - P(sunny) = 80%
  - P(cloudy) = 15%
  - P(rainy) = 5%
- I.e., we want to sample from this distribution:

$$\pi = \begin{bmatrix} 0.8 & 0.15 & 0.05 \end{bmatrix}$$

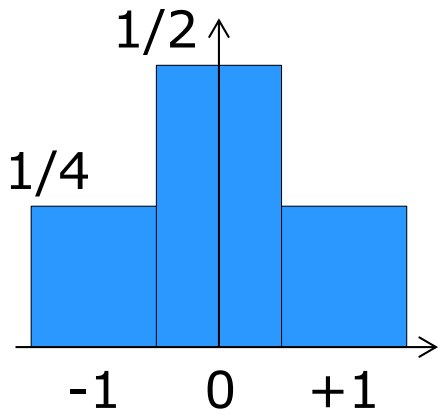
Example sunny – cloudy – rainy

- Definition of the proposal distribution
  - May be chosen "arbitrarily" (see later)
  - For example:
    - Stay in the current state with probability ½
    - Switch with probability of ¼ to one of the (two) other states

$q(y \mid x)$

	s	c	r
s	1/2	1/4	1/4
c	1/4	1/2	1/4
r	1/4	1/4	1/2

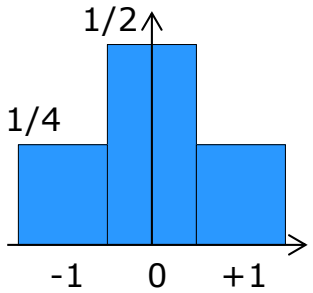
- (binomial distribution Binom(2, 0.5))



Example sunny – cloudy – rainy

- Note: the proposal distribution is symmetric:

$$q(y \mid x) = q(x \mid y)$$



- Therefore, we have:

$$\alpha(x, y) = \min\left(\frac{\pi(y)\, q(x \mid y)}{\pi(x) q(y \mid x)}, 1\right) = \min\left(\frac{\pi(y)}{\pi(x)}, 1\right)$$

- With this simplification, it is the algorithm proposed by Metropolis et al. (1953)

The algorithm of Metropolis (-Hastings) in the "sunny – cloudy – rainy" example

Use: 0=sunny, 1=cloudy, 2=rainy.

---

x = 0 # Or 1 or 2.

repeat m times:

    draw $\delta \sim \text{binom}(2, 0.5) - 1$ # In {-1,0,1}.

    proposal is: $y = (x + \delta) \,\%\, 3$ # Modulo*.

    compute $\alpha = \min(\pi(y)/\pi(x), 1)$

    draw $u \sim \text{uniform}(0,1)$

    if $u \leq \alpha$:

        $x = y$

    output x # In any case.

---

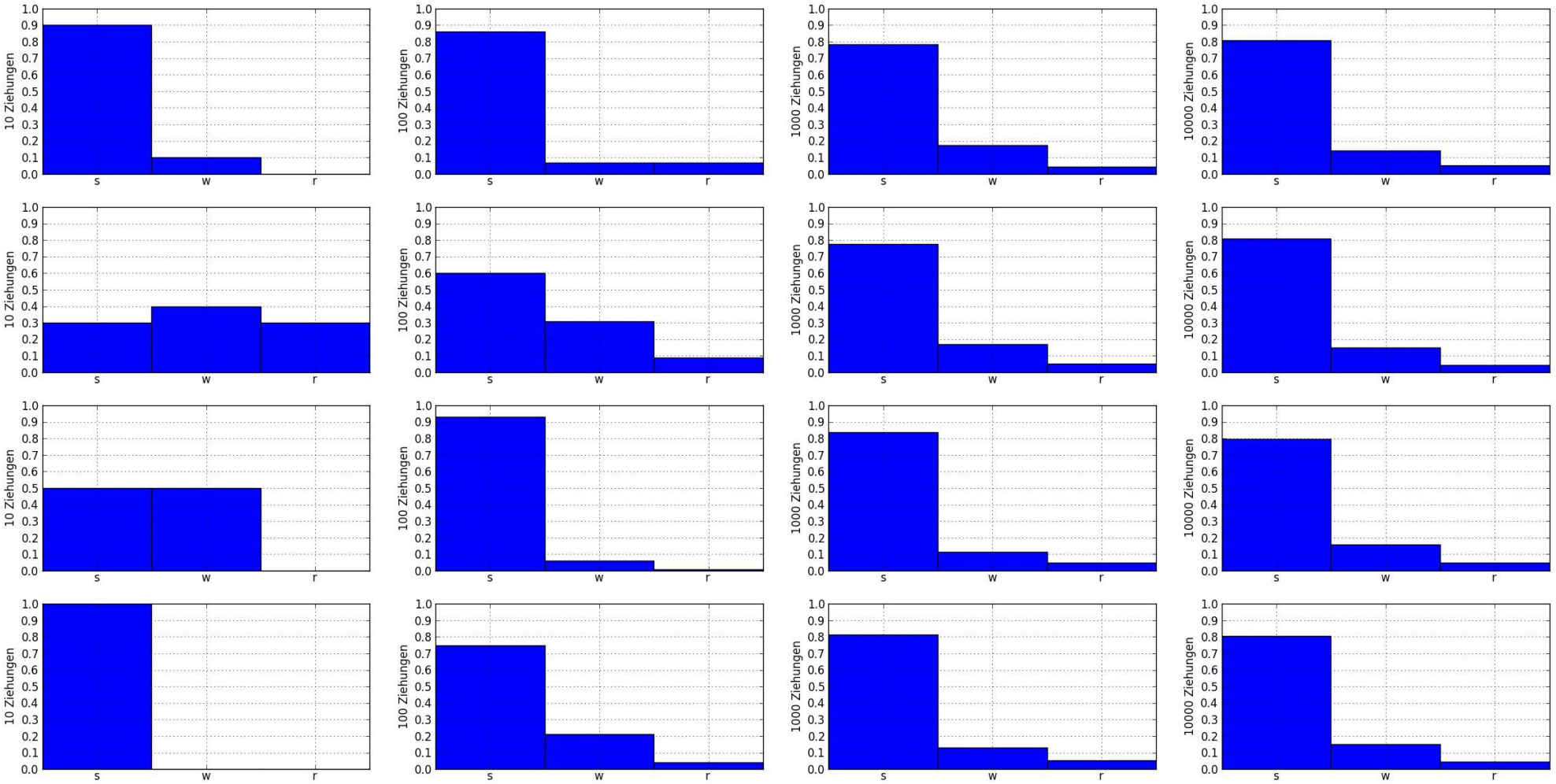
*Beware of modulo operations with negative operands.
If in doubt, use: (x+delta+3)%3.

Example code (real: Python)

- Direct implementation of the algorithm – very easy!

```
def simulate(start_state, stat_dist, steps):
    state = start_state
    result = [state]
    proposal_dist = binom(2, 0.5)
    acceptance_dist = uniform(0,1)
    for i in xrange(1, steps):
        state2 = (state + proposal_dist.rvs() - 1) % 3
        alpha = stat_dist[state2] / stat_dist[state]
        if acceptance_dist.rvs() <= alpha:
            state = state2
        result.append(state)
    return result
```

Example sunny – cloudy – rainy: Relative frequencies of the states



10 draws 100 draws 1000 draws 10000 draws

Why does the algorithm of Metropolis-Hastings produce the desired result?

- The transition kernel *T* is not (as in the previous examples) given by a simple matrix, but in terms of an algorithm, using proposal and acceptance
- We want to show:
  - The algorithm defines a transition kernel *T*, for which this holds:

    $$\pi \cdot T = \pi$$

  - Then, π is a stationary distribution of the generated chain
- The transition kernel is the probability:

  $$P(X_{i+1} \mid X_i)$$

  i.e., we want to show:

  $$\sum \pi(X_i) P(X_{i+1} \mid X_i) = \pi(X_{i+1})$$

Metropolis-Hastings: transition kernel

- The probability of a transition $X_i \rightarrow X_{i+1}$ consists of two terms:
  - Either: the proposal, made according to *q,* is accepted. This happens with probability:

$$q(X_{i+1} \mid X_i) \cdot \alpha(X_i, X_{i+1})$$

(This includes the case where *q* does not propose a change, and this "no-move" is accepted)

  - Or: a transition is proposed, which is not accepted. For this, the probability is:

$$1 - \sum_Y q(Y \mid X_i) \cdot \alpha(X_i, Y)$$

(This exclusively includes the case that no change takes place because the proposed change is not accepted)

Metropolis-Hastings: transition kernel

- Combining those two terms, we obtain:

$$\begin{aligned} P(X_{i+1} \mid X_i) &= q(X_{i+1} \mid X_i) \cdot \alpha(X_i, X_{i+1}) \\ &+ \mathrm{I}_{X_{i+1}=X_i} \left(1 - \sum_Y q(Y \mid X_i) \cdot \alpha(X_i, Y) \right) \end{aligned}$$

where:

$$\mathrm{I}_{X_{i+1}=X_i} := \begin{cases} 1 \text{ for } X_{i+1} = X_i \\ 0 \text{ else} \end{cases}$$

Auxiliary calculation: detailed balance

$$
\begin{aligned}
& \pi(X_i)q(X_{i+1} \mid X_i)\alpha(X_i, X_{i+1}) \\
= \quad & \pi(X_i)q(X_{i+1} \mid X_i)\min\left(1, \frac{\pi(X_{i+1})q(X_i \mid X_{i+1})}{\pi(X_i)q(X_{i+1} \mid X_i)}\right) \\
= \quad & \min\left(\pi(X_i)q(X_{i+1} \mid X_i), \pi(X_{i+1})q(X_i \mid X_{i+1})\right) \\
= \quad & \pi(X_{i+1})q(X_i \mid X_{i+1})\min\left(\frac{\pi(X_i)q(X_{i+1} \mid X_i)}{\pi(X_{i+1})q(X_i \mid X_{i+1})}, 1\right) \\
= \quad & \pi(X_{i+1})q(X_i \mid X_{i+1})\alpha(X_{i+1}, X_i)
\end{aligned}
$$

- The last row is equal to the first row with i ←→ i+1 flipped.

Auxiliary calculation: detailed balance

- What happens if we multiply π with the transition kernel?

$$\begin{aligned}\pi(X_i)P(X_{i+1}\,|\,X_i) &= \pi(X_i)q(X_{i+1}\,|\,X_i)\cdot\alpha(X_i,X_{i+1}) \\ &+ \pi(X_i)\cdot \mathrm{I}_{X_{i+1}=X_i}\left(1-\sum_Y q(Y\,|\,X_i)\cdot\alpha(X_i,Y)\right)\end{aligned}$$

  - 1. part, right: see previous slide, one may exchange i ←→ i+1
  - 2. part, right: here we can also exchange i ←→ i+1, since the term is only nonzero if $X_i=X_{i+1}$

- From this it follows:
  - The Markov chain fulfils the so-called detailed balance condition: There exists a distribution π, for which:

$$\pi(X_i)P(X_{i+1}\,|\,X_i)=\pi(X_{i+1})P(X_i\,|\,X_{i+1})$$

Metropolis-Hastings: stationary distribution

- From the detailed balance condition, it follows that π is a stationary distribution, since (summing on both sides):

$$\begin{aligned}\sum_{X_i}\pi(X_i)P(X_{i+1}\,|\,X_i) &= \sum_{X_i}\pi(X_{i+1})P(X_i\,|\,X_{i+1}) \\ &= \pi(X_{i+1})\cdot\sum_{X_i}P(X_i\,|\,X_{i+1}) \\ &= \pi(X_{i+1})\end{aligned}$$

- I.e., we have shown that the distribution π, which is used in the Metropolis-Hastings algorithm in the function α, **is** the stationary distribution of the generated chain.

Metropolis-Hastings: convergence

- For the convergence, we have to show in addition the following two conditions (see the slide “ergodic Markov chains” above)
- Aperiodicity: a sufficient condition is that the chain does not move ($X_{i+1}=X_i$) with a probability P>0
- Irreducibility: a sufficient condition is that the proposal distribution is > 0 everywhere (in the support)
  - This can be obtained easily, e.g. by using a normal distribution for the proposal

Metropolis-Hastings: remarks

- Improvements are always accepted (Metropolis), since

$$\alpha(x,y) = \min\left(\frac{\pi(y)}{\pi(x)}, 1\right) = 1$$

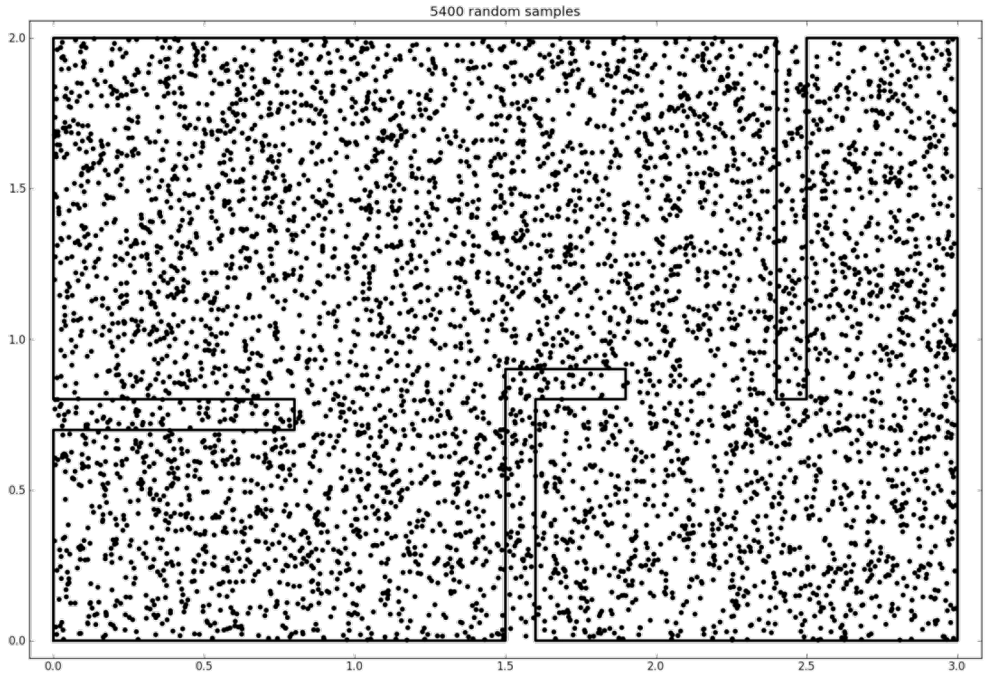
  - (similarly for Metropolis-Hastings, if the quotient improves)
- The proposal distribution is (astonishingly) "arbitrary"
  - BUT: it influences the convergence properties
  - So-called "mixing" of the chain (see later)
- Instead of the (target) distribution π one can use a (non-normalized) score function, since it is only used in the computation of the acceptance probability

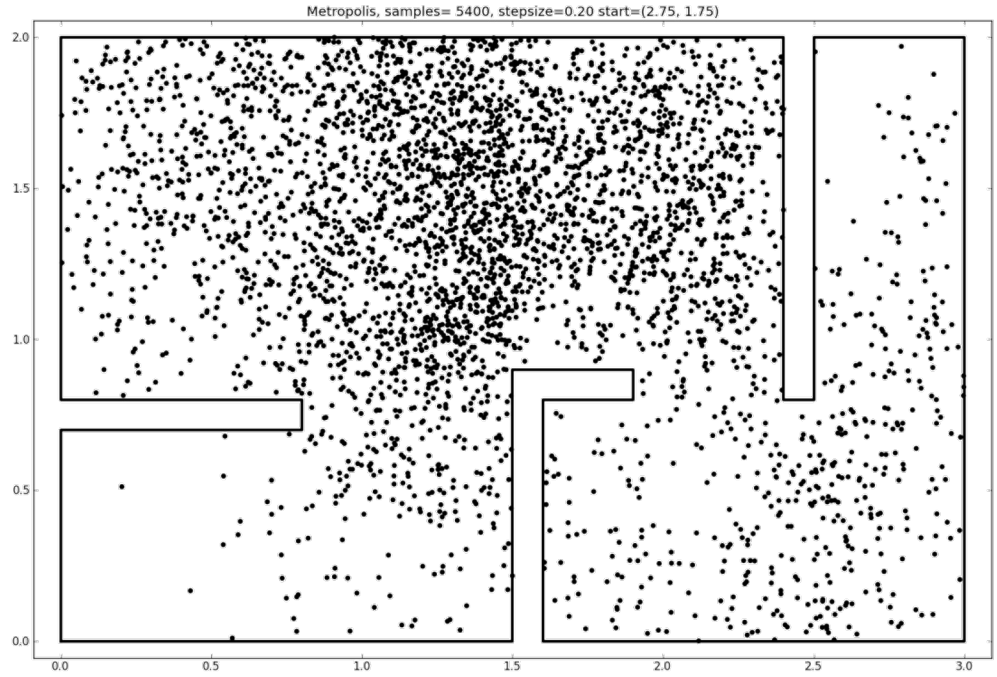
$$\alpha(x,y) = \min\left(\frac{\pi(y)\,q(x\,|\,y)}{\pi(x)q(y\,|\,x)}, 1\right)$$

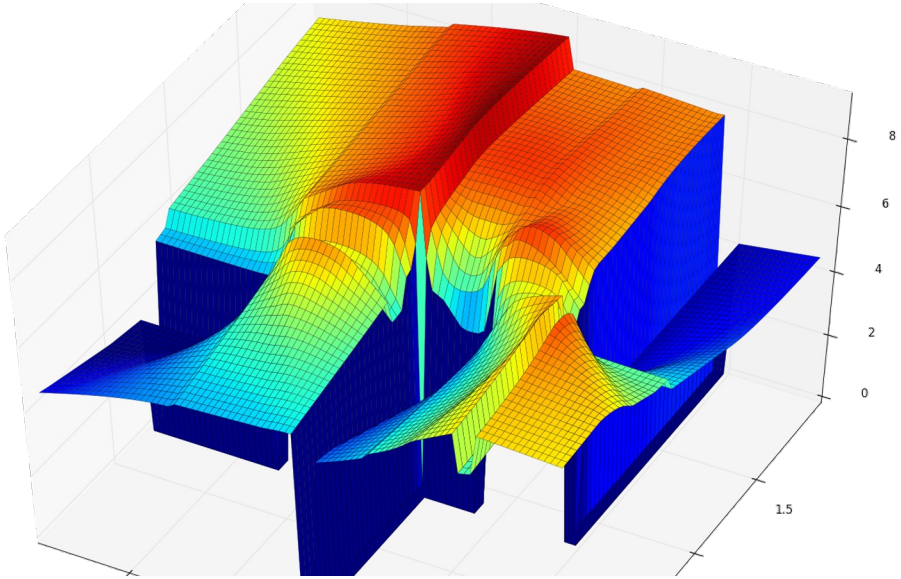
  where only the quotient $\pi(y)/\pi(x)$ takes part (so any constant factor cancels)

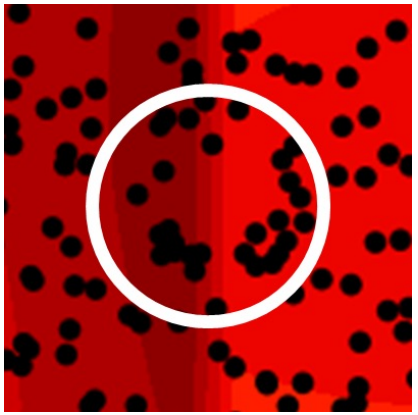
Simple examples

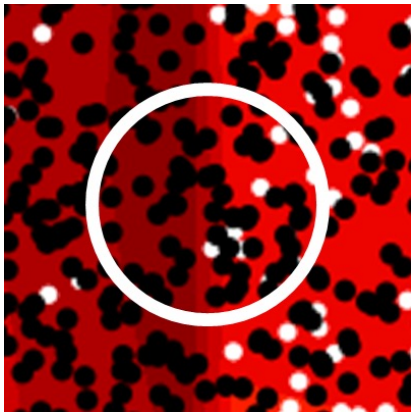
Art gallery problem (previous lecture): comparison uniform sampling → MCMC



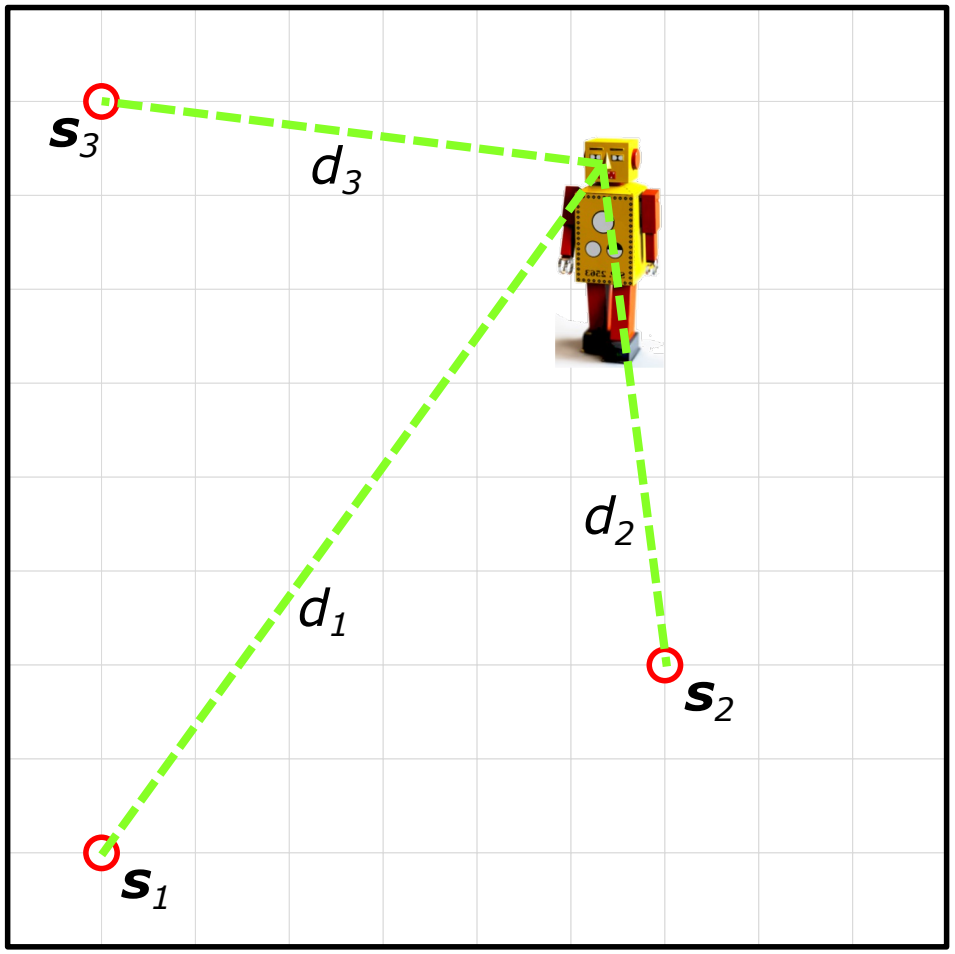


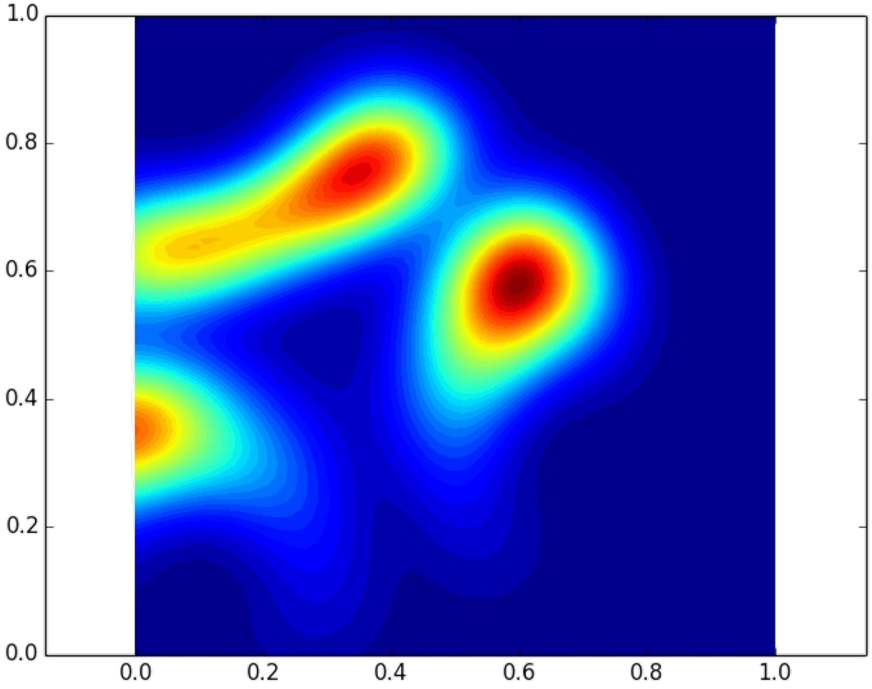




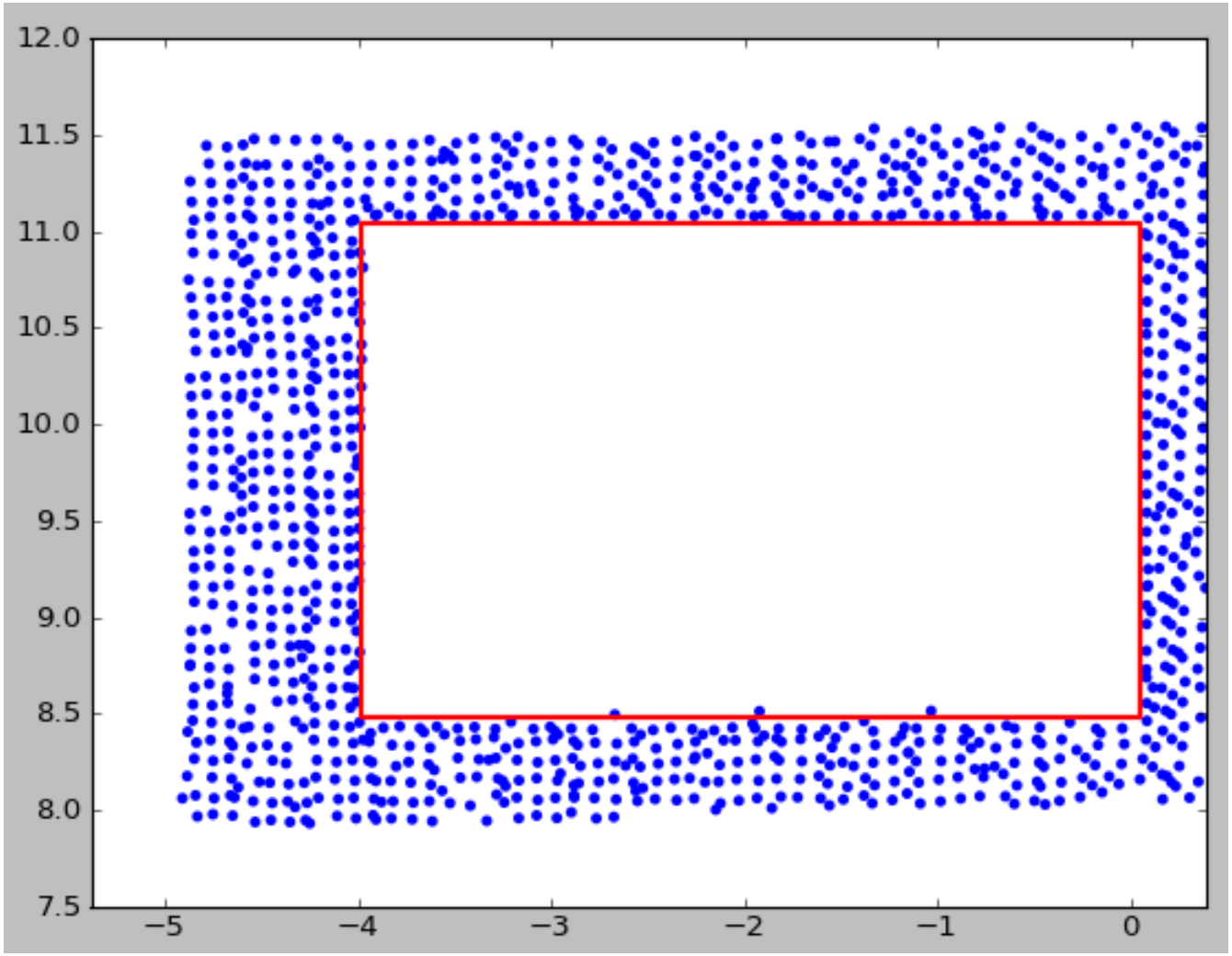


Robot localization (see exercises)



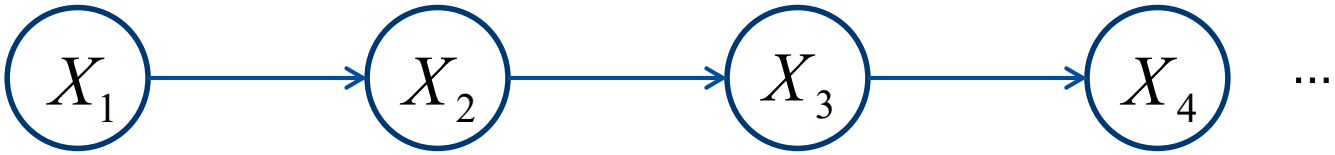


Find structures in a point cloud (e.g.: windows) (see exercises)



Appendix: Markov chain / Bayes network

Appendix: Markov chain / Bayes network



- Our (mis-) conception for i: "sequence of states" (in "time")
- But: the Markov chain only defines a special joint distribution, which is given by the Markov property (independence):

$$P(X_1, X_2, X_3, X_4) = P(X_4, X_3, X_2, X_1)$$

$$= \underbrace{P(X_4 \mid X_3, X_2, X_1)}_{=P(X_4|X_3)} \cdot P(X_3, X_2, X_1) \quad [\text{cond. prob.}]$$

$$= P(X_4 \mid X_3) \cdot P(X_3, X_2, X_1) \quad [\text{Markov property}]$$

$$= P(X_4 \mid X_3) \cdot \underbrace{P(X_3 \mid X_2, X_1)}_{=P(X_3|X_2)} \cdot P(X_2, X_1) \quad [\text{cond. prob.}]$$

$$= P(X_4 \mid X_3) \cdot P(X_3 \mid X_2) \cdot P(X_2, X_1) \quad [\text{Markov property}]$$

$$= P(X_4 \mid X_3) \cdot P(X_3 \mid X_2) \cdot P(X_2 \mid X_1) \cdot P(X_1)$$

Example.:
3 states:
$3^4-1=80$

6+6+6+2=20

Appendix: Markov chain / Bayes network

- After n "steps", a n-dimensional distribution is defined.
  E.g. for n=4:

$$P(X_1, X_2, X_3, X_4)$$

- We are interested in the marginal distribution:

$$P(X_4) = \sum_{X_1}\sum_{X_2}\sum_{X_3} P(X_1, X_2, X_3, X_4)$$

- The sums can be "moved inside":
  - ("sum-product", "message passing" algorithms)

$$\begin{aligned} & \sum_{X_1}\sum_{X_2}\sum_{X_3} P(X_4 \mid X_3)P(X_3 \mid X_2)P(X_2 \mid X_1)P(X_1) \\ = & \sum_{X_3} P(X_4 \mid X_3)\sum_{X_2} P(X_3 \mid X_2)\underbrace{\sum_{X_1} P(X_2 \mid X_1)P(X_1)}_{=P(X_1)\cdot T} \\ = & P(X_1)\cdot T^3 \end{aligned}$$

matrix x vector

Appendix: Markov chain / Bayes network

- A Markov chain with n steps defines a n-dimensional distribution
- So for any choice of x_i one can compute:

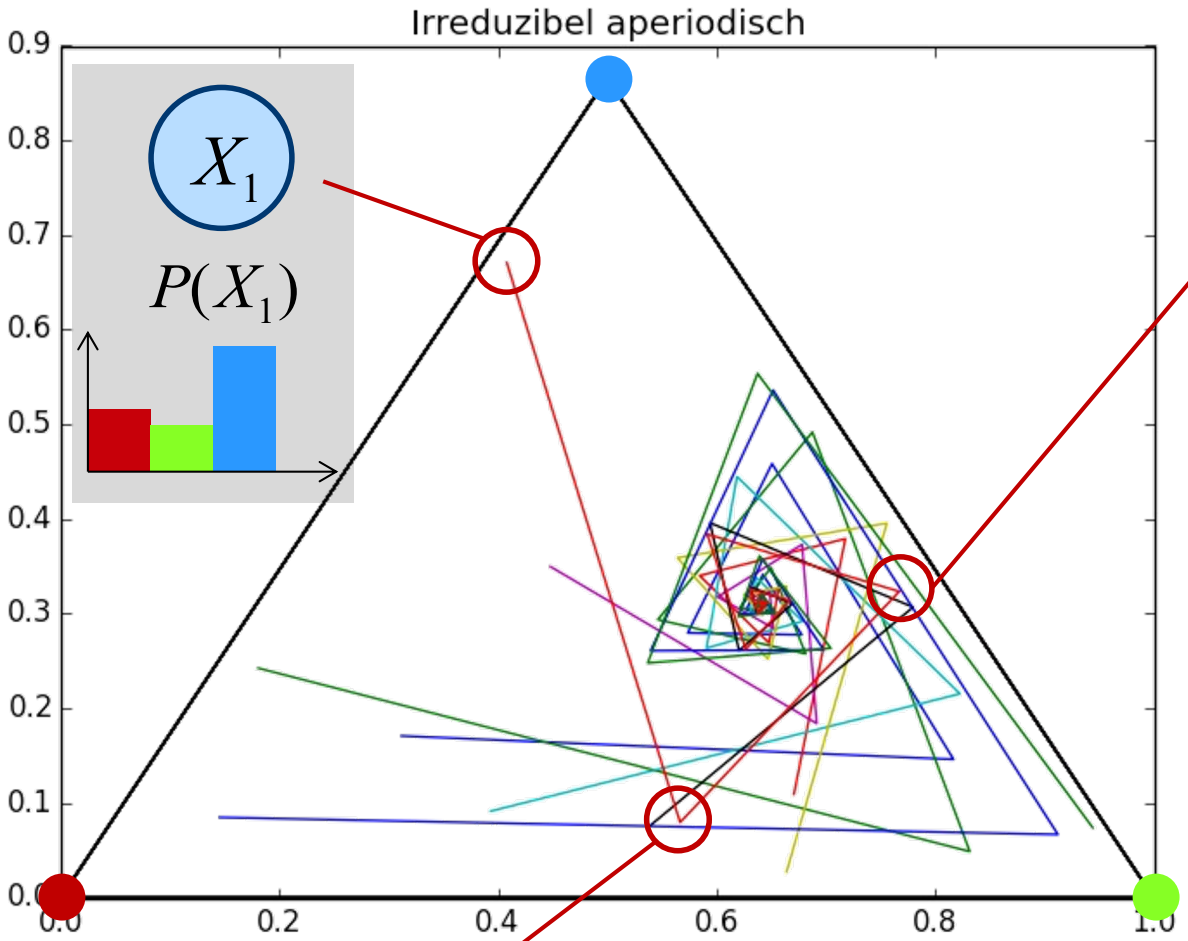
$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$$

- We are only interested in the distribution of the last, the nth, variable. Therefore, we compute the marginal distribution for variable n

$$P(X_4)$$

- This results from the initial distribution and n-1 applications of the transition kernel

$$P(X_4) = P(X_1) \cdot T^3$$



X_1

$P(X_1)$

$X_1 \rightarrow X_2 \rightarrow X_3$

$P(X_3) = \sum_{X_1, X_2} P(X_1, X_2, X_3) = P(X_1) \cdot T^2$

$X_1 \rightarrow X_2$

$P(X_2) = \sum_{X_1} P(X_1, X_2) = P(X_1) \cdot T$

$P(X_1, X_2) = P(X_2 \mid X_1) \cdot P(X_1)$

$P(X_1)$

References

- C. M. Bishop: Pattern Recognition and Machine Learning, Springer 2006 (Chapter 11)
- C. P. Robert, G. Casella: Monte Carlo Statistical Methods, Springer 2004
- W. R. Gilks, S. Richardson, D. J. Spiegelhalter: Markov Chain Monte Carlo in Practice, Chapman & Hall/CRC 1996
- Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller: Equations of State Calculations by Fast Computing Machines, In: The Journal of Chemical Physics, vol. 21, pp. 1087–1092, 1953
- W. Keith Hastings: Monte Carlo Sampling Methods Using Markov Chains and Their Applications, In: Biometrika, col. 57, pp. 97-109, 1970
- Radford M. Neal: Slice sampling, In: Annals of Statistics, vol. 31, no. 3, pp. 705-767, 2003