Random Processes: Hidden Markov Models

Stochastic Processes for CS4070

Lecture 7

Contents

- Markov Chains:
  - Transition probabilities
  - State probabilities
  - Limiting state probabilities
  - State classification

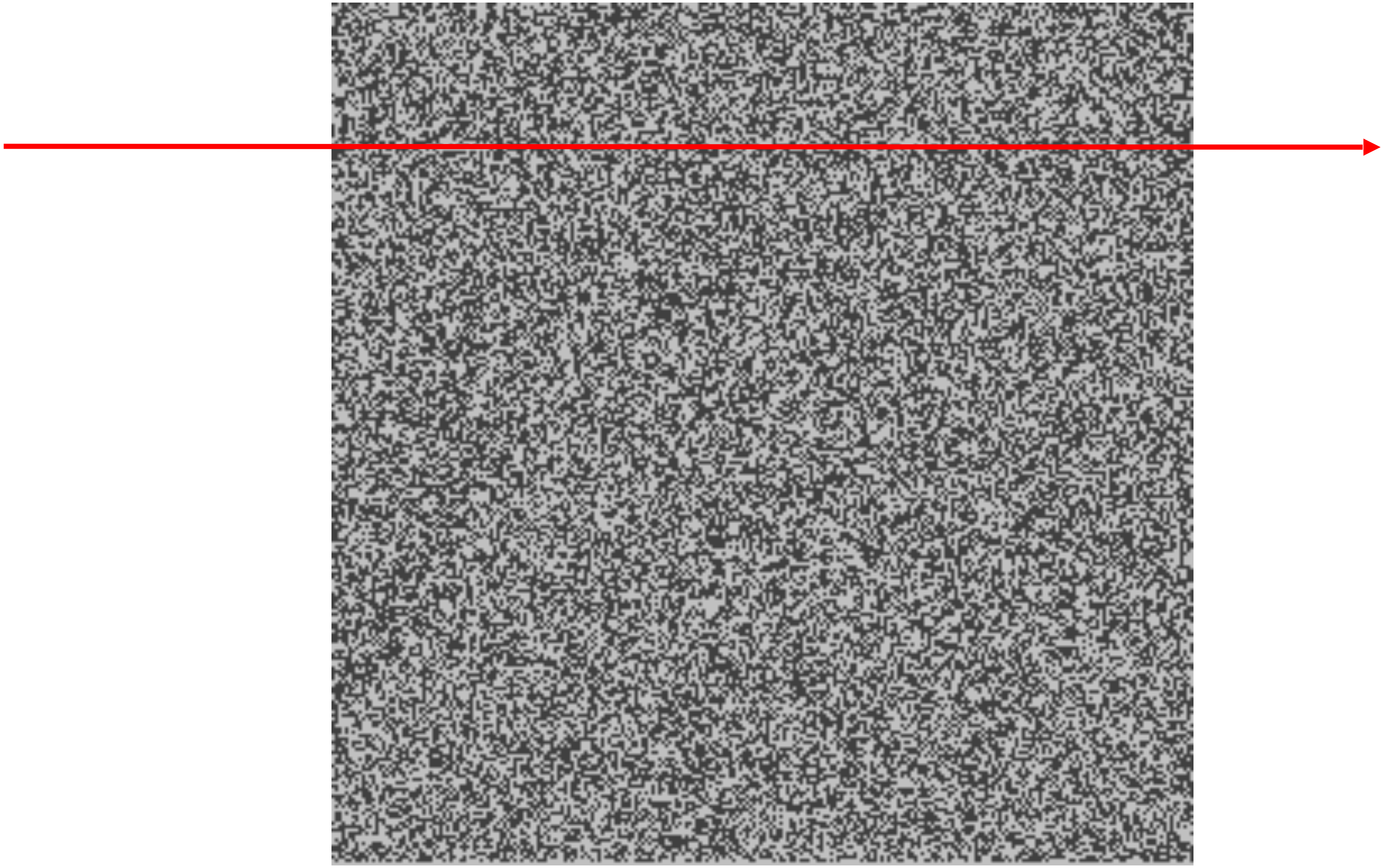
for exam

- Hidden Markov Models
  - Evaluation
  - Decoding
  - Learning

not for exam

- Questions

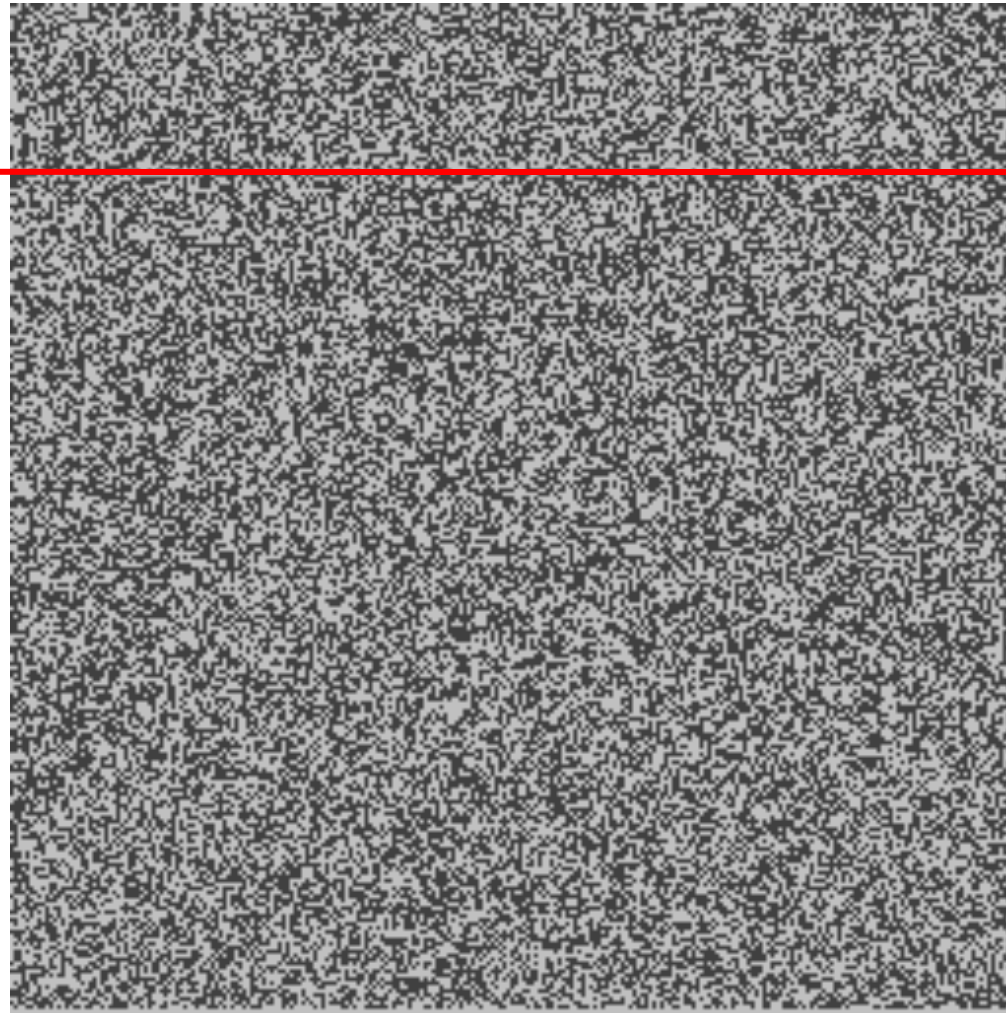
Markov Chains (1)



0010101101010110100001110101010010110111001000 11010

Markov Chains (1)

IID Bernoulli
Random process



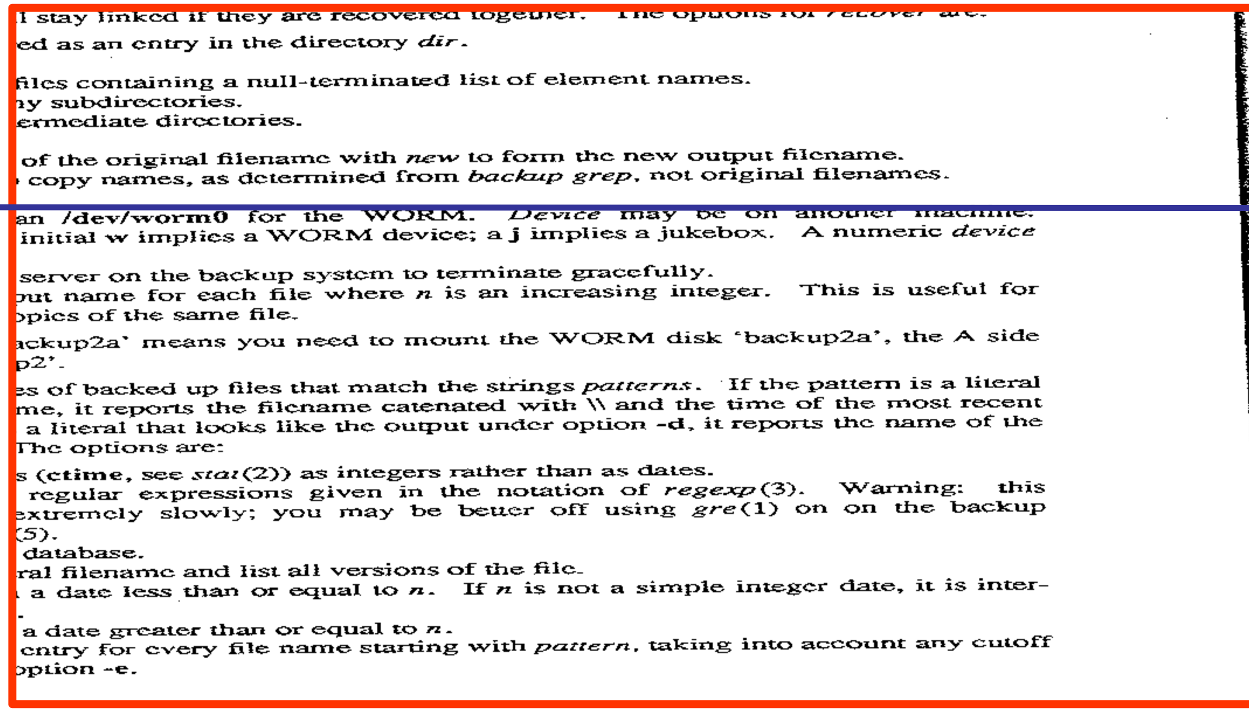
001010110101011010000111010101001011011100100011010

Markov Chains (2)

000000111111010111111111100000000000001111111000111

Markov Chains (3)

X_n



l stay linked if they are recovered together. The options for recover are:
ed as an entry in the directory *dir*.
files containing a null-terminated list of element names.
ny subdirectories.
ermediate directories.
of the original filename with *new* to form the new output filename.
copy names, as determined from *backup grep*, not original filenames.
an /dev/worm0 for the WORM. *Device* may be on another machine.
initial w implies a WORM device; a j implies a jukebox. A numeric *device*
server on the backup system to terminate gracefully.
put name for each file where *n* is an increasing integer. This is useful for
opies of the same file.
ackup2a' means you need to mount the WORM disk 'backup2a', the A side
p2'.
es of backed up files that match the strings *patterns*. If the pattern is a literal
me, it reports the filename catenated with \\ and the time of the most recent
a literal that looks like the output under option -d, it reports the name of the
The options are:
s (ctime, see *stat*(2)) as integers rather than as dates.
regular expressions given in the notation of *regexp*(3). Warning: this
extremely slowly; you may be better off using *gre*(1) on on the backup
(5).
database.
ral filename and list all versions of the file.
a date less than or equal to *n*. If *n* is not a simple integer date, it is inter-
a date greater than or equal to *n*.
entry for every file name starting with *pattern*, taking into account any cutoff
option -e.

X_n 000000110000000011111000000000000000111111000111

More structure than can be expressed with autocorrelation fnctn:

(1) if in "background" ($X_n = 0$) than very likely next value X_{n+1} is also in "background" ($X_{n+1} = 0$)

(2) long runs/subsequences of same value

Definition of Markov Chain

- Time discrete and amplitude discrete random process $\{X_n \mid n=0,1,2,\ldots\}$

- Property:

  The conditional PMF of X_{n+1} depends only on X_n and not on X_{n-1} , X_{n-2} , X_{n-3} ,… X_0

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0)$$
$$= P(X_{n+1} = j \mid X_n = i)$$
$$= P_{ij}$$

- This is called the Markov property

  (A process with the Markov property is called Markov Process)

Definitions

- The current value of the Markov chain X_n is called the "state"
- The conditional probabilities

$$P_{ij} = P(X_{n+1} = j | X_n = i)$$

time new state old state

- are called transition probabilities with

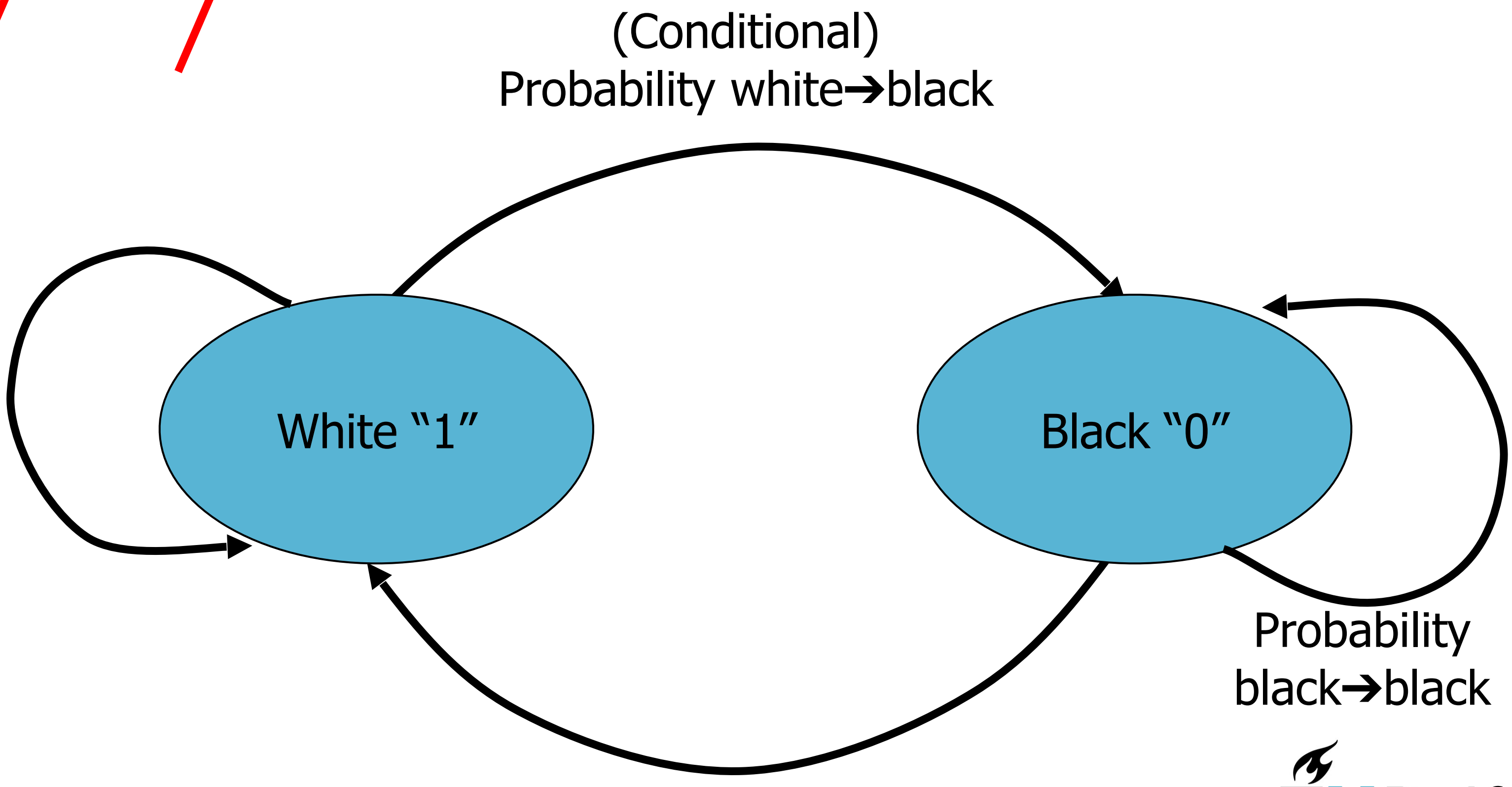
$$\sum_{j=0}^{\infty} P_{ij} = 1$$

TU Delft

Markov Model (Chain Diagram)

- For the image example:

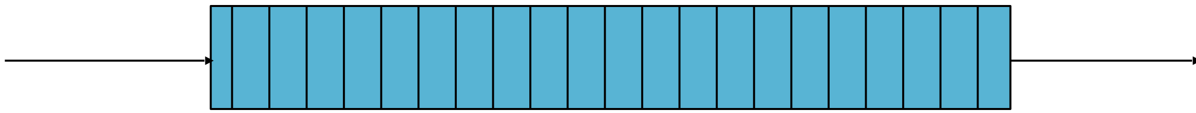
$P(X_{n+1}=1|X_n=1)$ $P(X_{n+1}=0|X_n=1)$

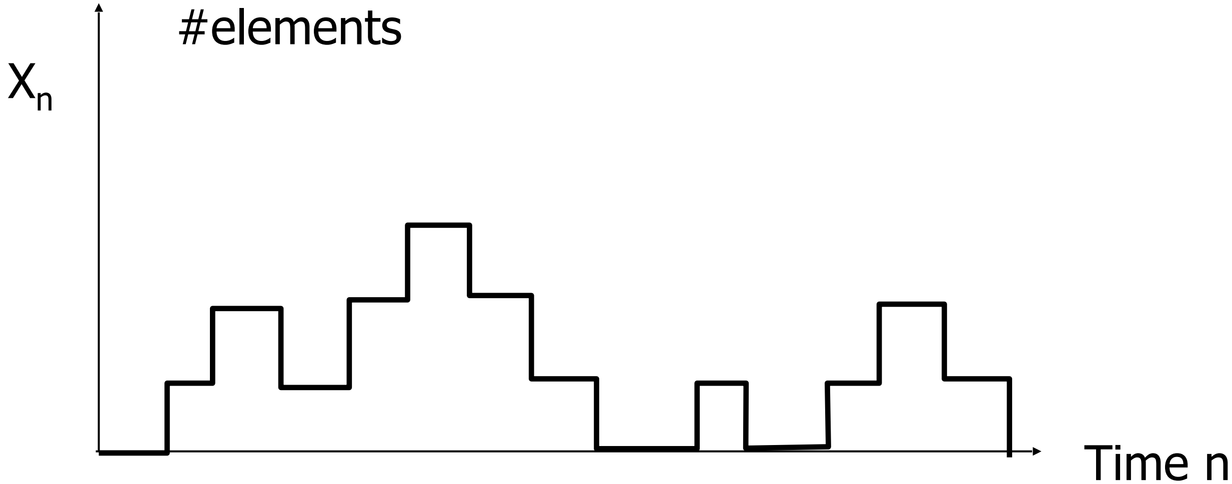


Number of Entries in Queue

Model for number of elements in a queue (read or write element at every time instance)

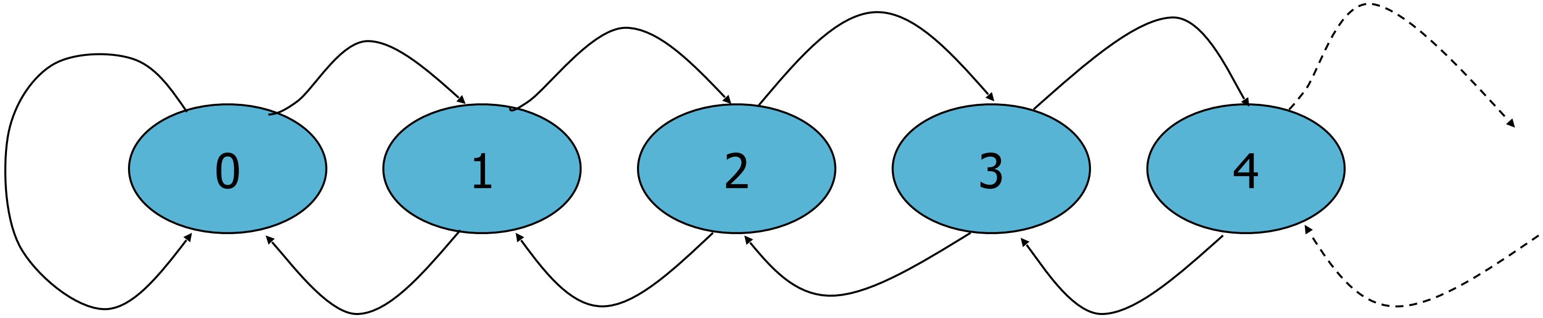
Birth-Death process





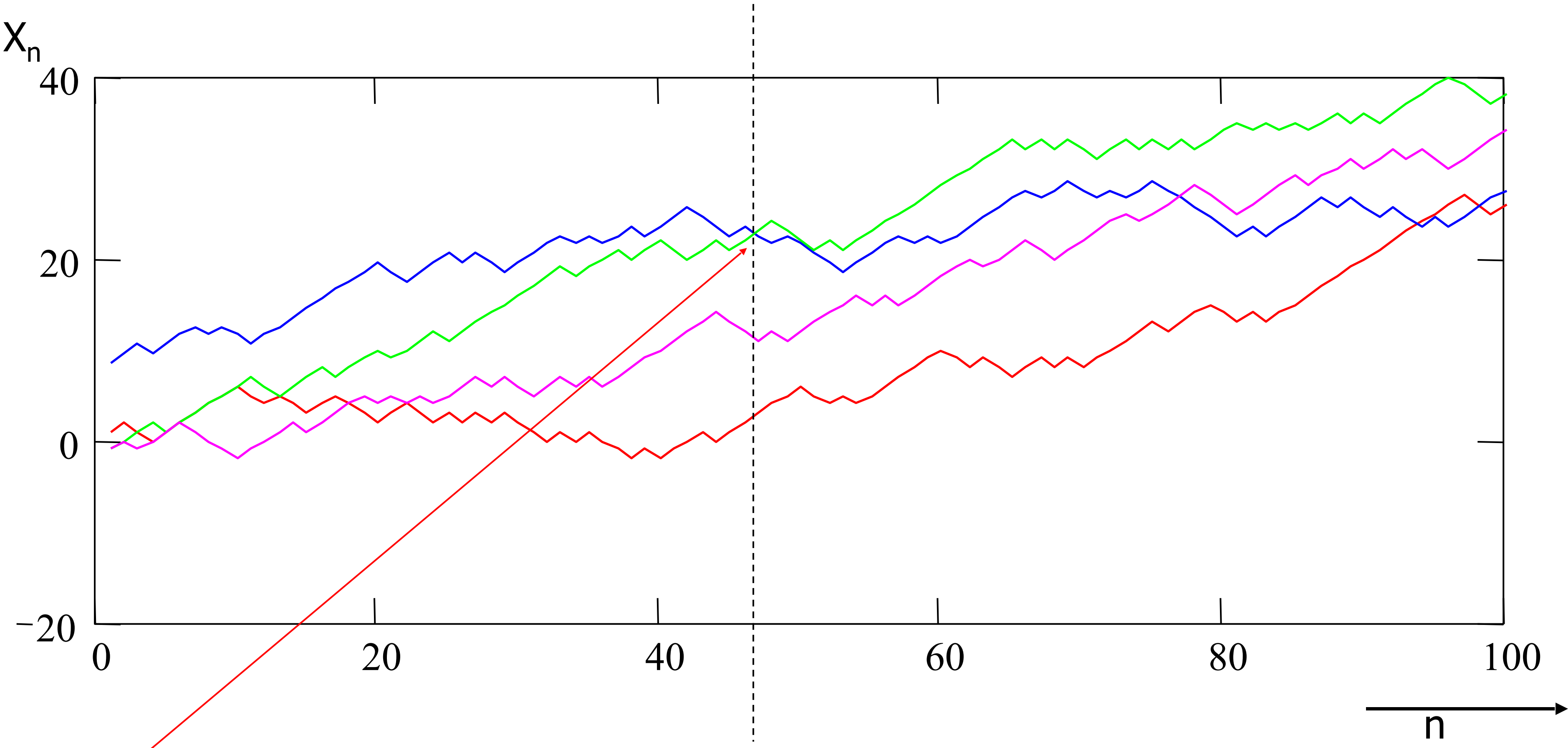
Chain Diagram

$$X_n = Z_0 + Z_1 + Z_2 + \ldots Z_n$$



(Number of states is infinite)

A Few Sample Functions



$$P(X_{49} = 21 \mid X_{48} = 20) = P(Z_{49} = +1)$$

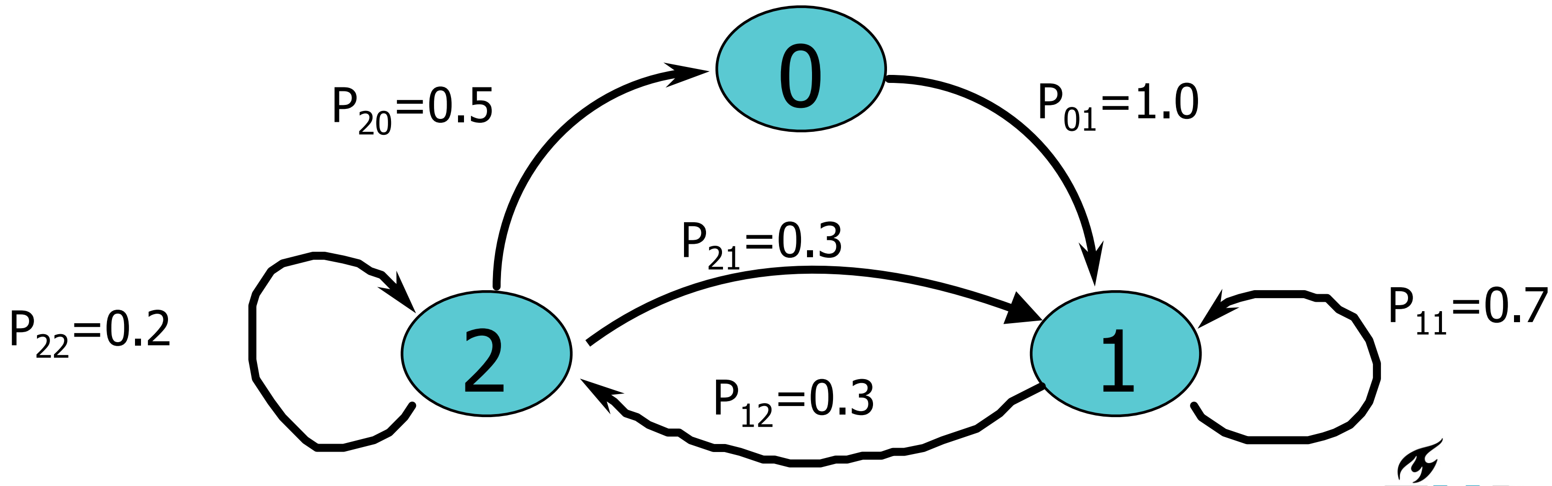
$$P(X_{49} = 19 \mid X_{48} = 20) = P(Z_{49} = -1)$$

Description Markov Chains

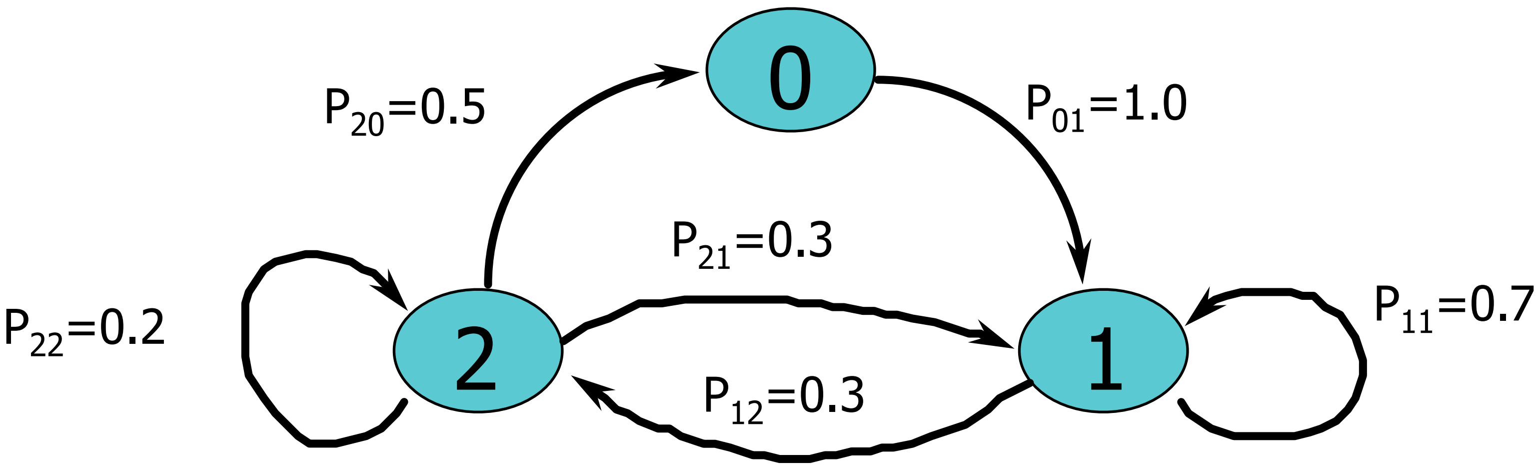
- State transition matrix (with transition probabilities)

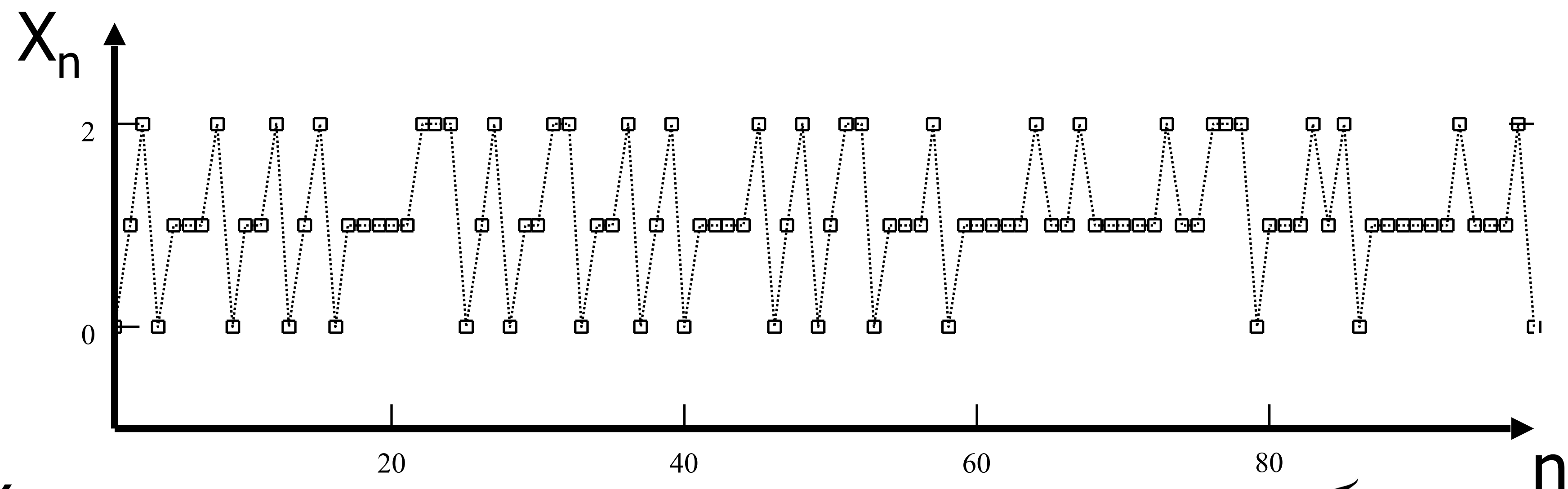
$$\mathbf{P} = \begin{bmatrix} P_{00} & P_{01} & P_{02} \\ P_{10} & P_{11} & P_{12} \\ P_{20} & P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} 0 & 1.0 & 0 \\ 0 & 0.7 & 0.3 \\ 0.5 & 0.3 & 0.2 \end{bmatrix}$$

- Transition diagram (chain diagram)



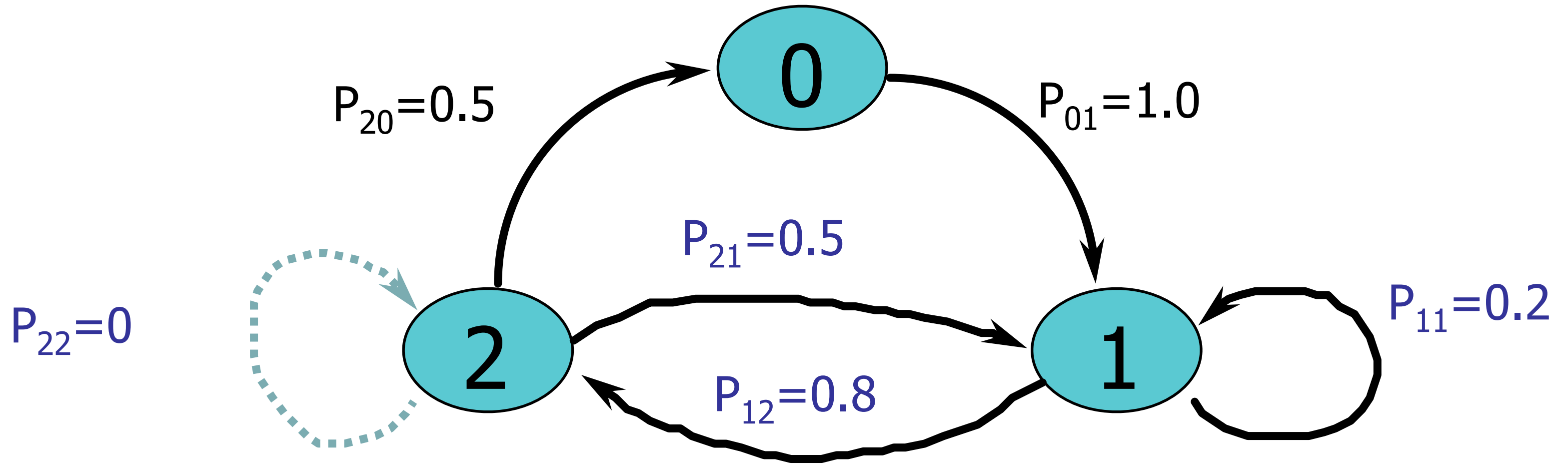
Markov Chain and One Realization

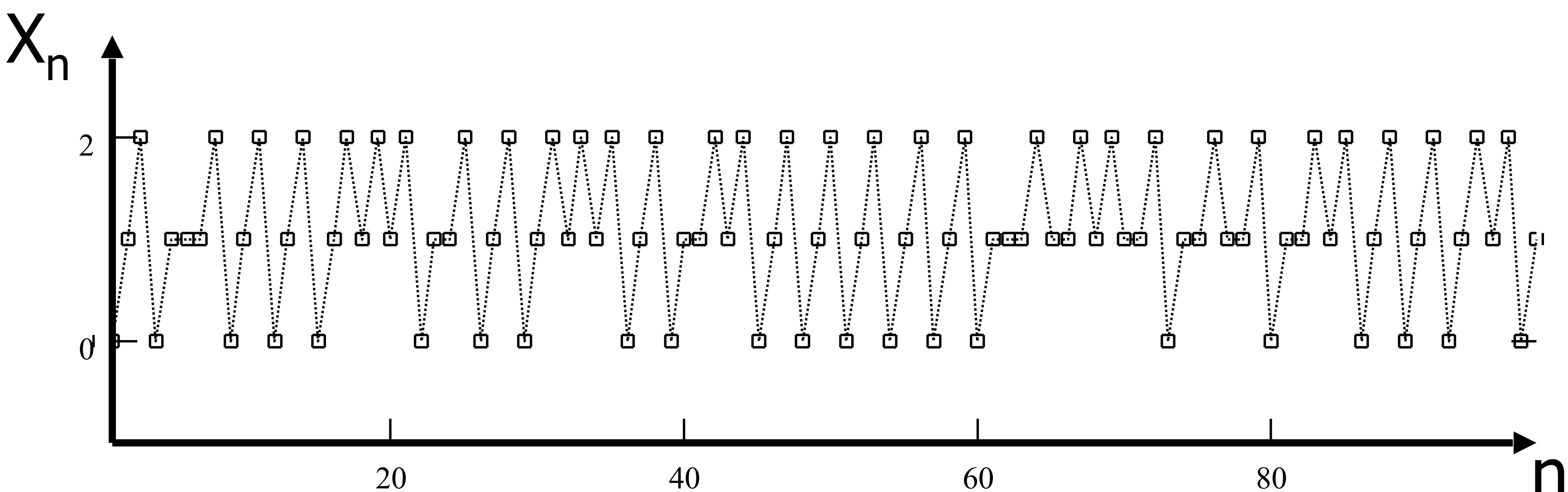




X_n 012011120112012011111222012011220112012

Example (change P_{22} and P_{11})



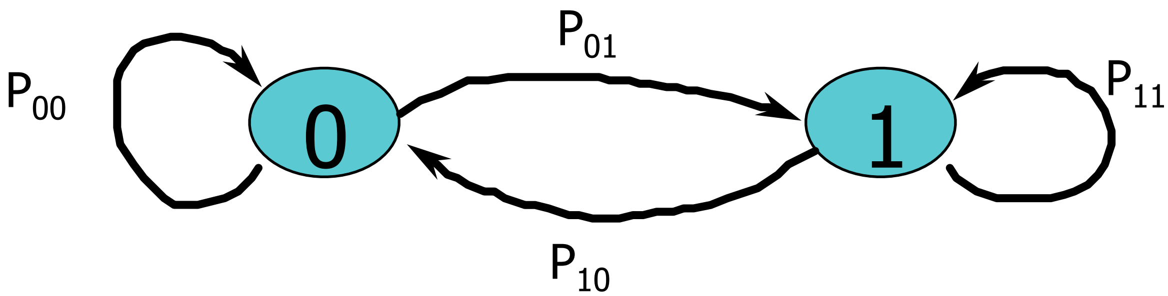


X_n 01201112012012012121201120120121212012001

Properties

- Probability of a particular sample function
- m-step transition probabilities
- State probabilities
- Limiting state probabilities

Probability of a Sample Function



$$\mathbf{P} = \begin{bmatrix} 0.8 & 0.2 \\ 0.9 & 0.1 \end{bmatrix}$$

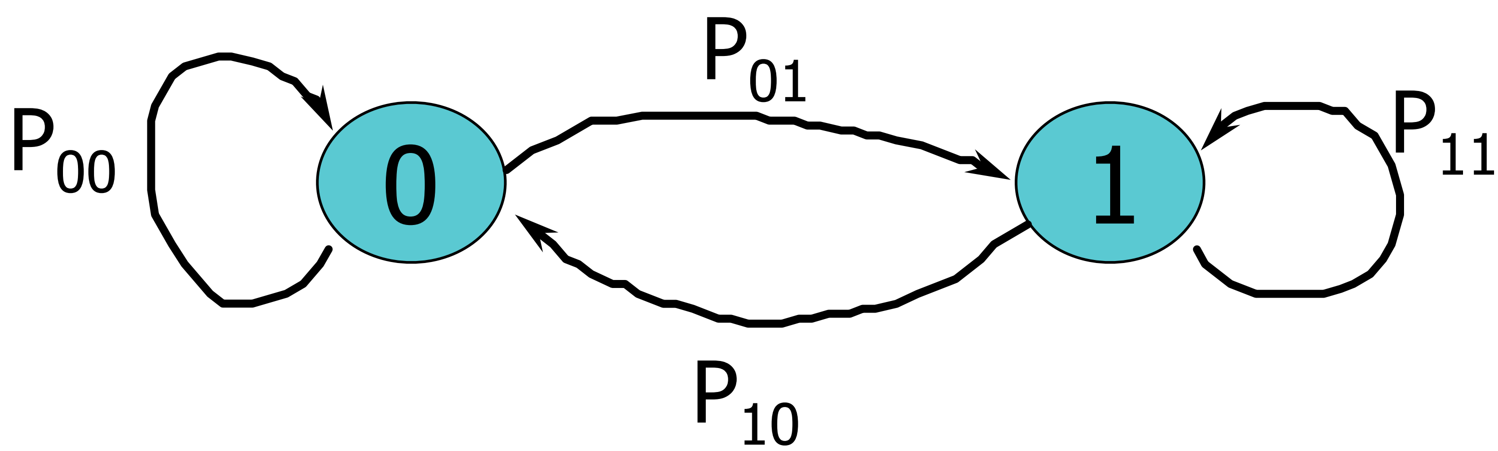
- What is probability of a particular sample function (realization)?

$$\begin{aligned} P[\text{"011"}] &= P[X_0 = 0, X_1 = 1, X_2 = 1] \\ &= P[X_2 = 1 \mid X_1 = 1 \cancel{, X_0 = 0}]P[X_1 = 1, X_0 = 0] \\ &= P[X_2 = 1 \mid X_1 = 1]P[X_1 = 1 \mid X_0 = 0]\underbrace{P[X_0 = 0]}_{\text{assume 1.0}} \\ &= 0.1 * 0.2 * 1 = 0.02 \end{aligned}$$

- Similarly for other or longer sample functions

m-Step Transition Probabilities

- The probabilities in the chain diagram are called one-step transition probabilities

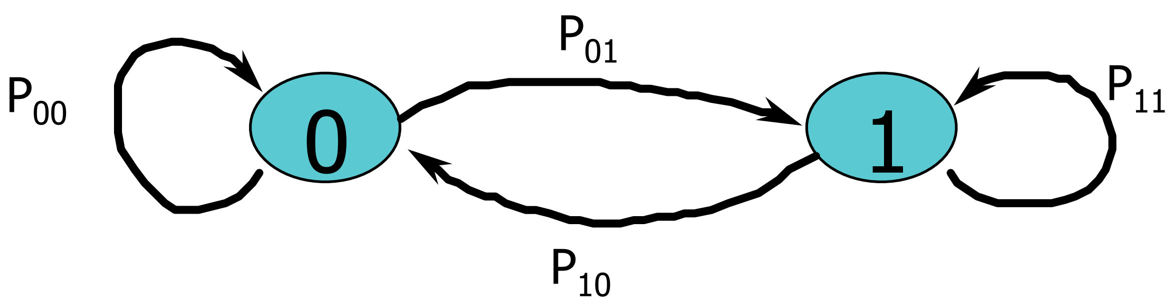


- X_n: 0 0 0 0 0 1 1 1 0 0 0 0 0 0 1 1 0 0 0 0 0 1 1 1 0 0

2-step 3-step

m-Step Transition Probabilities

- 2-step transition probabilities
- Example $P(X_{n+2} = j \mid X_n = i)$



$$
\begin{aligned}
&P(X_{n+2} = 1 \mid X_n = 0) \\
&= P(X_{n+2} = 1, X_{n+1} = 0 \mid X_n = 0) + P(X_{n+2} = 1, X_{n+1} = 1 \mid X_n = 0) \\
&= P(X_{n+2} = 1 \mid X_{n+1} = 0, \cancel{X_n = 0}) P(X_{n+1} = 0 \mid X_n = 0) \\
&+ P(X_{n+2} = 1 \mid X_{n+1} = 1, \cancel{X_n = 0}) P(X_{n+1} = 1 \mid X_n = 0) \\
&= P_{0,0} P_{0,1} + P_{0,1} P_{1,1}
\end{aligned}
$$

TU Delft

m-Step Transition Probabilities

- Similarly for

$$P(X_{n+2} = 1 | X_n = 1)$$
$$P(X_{n+2} = 0 | X_n = 0)$$
$$P(X_{n+2} = 0 | X_n = 1)$$

- All-in-one calculation

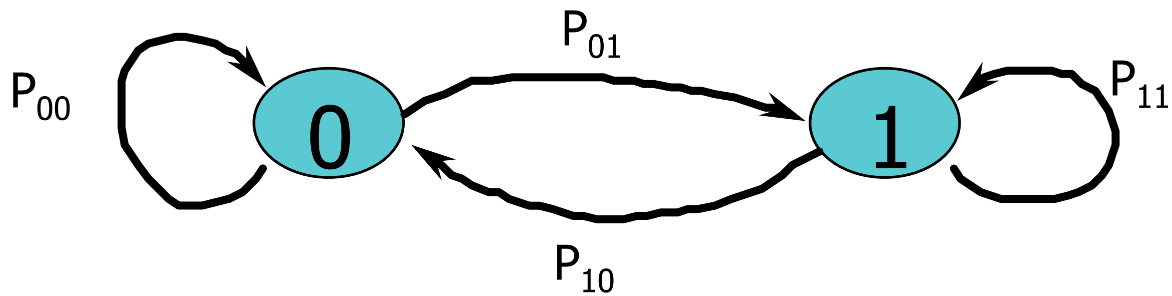
$$\mathbf{P}^2 = \mathbf{PP}$$

- m-Step:

$$P(X_{n+m} = j | X_n = i)$$

$$\mathbf{P}^m = \underbrace{\mathbf{PP}\ldots\mathbf{P}}_{\text{m times}}$$

State Probabilities (1)



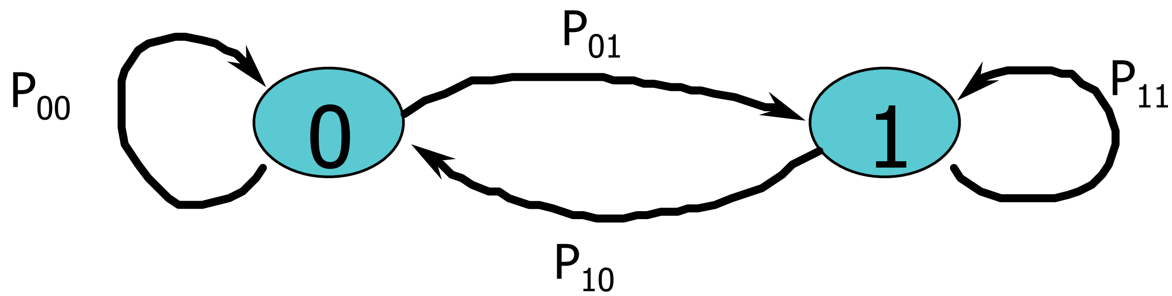
$$\mathbf{P} = \begin{bmatrix} 0.8 & 0.2 \\ 0.9 & 0.1 \end{bmatrix}$$

- State probabilities

$$\mathbf{p}(0) = (p_0(0), p_1(0)) = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

$$\begin{aligned} \mathbf{p}(1) &= \begin{pmatrix} p_0(1) & p_1(1) \end{pmatrix} \\ &= \begin{pmatrix} P_{00}p_0(0) + P_{10}p_1(0) & P_{01}p_0(0) + P_{11}p_1(0) \end{pmatrix} \\ &= \begin{pmatrix} p_0(0) & p_1(0) \end{pmatrix} \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix} \\ &= \mathbf{p}(0)\mathbf{P} \end{aligned}$$

State Probabilities (2)



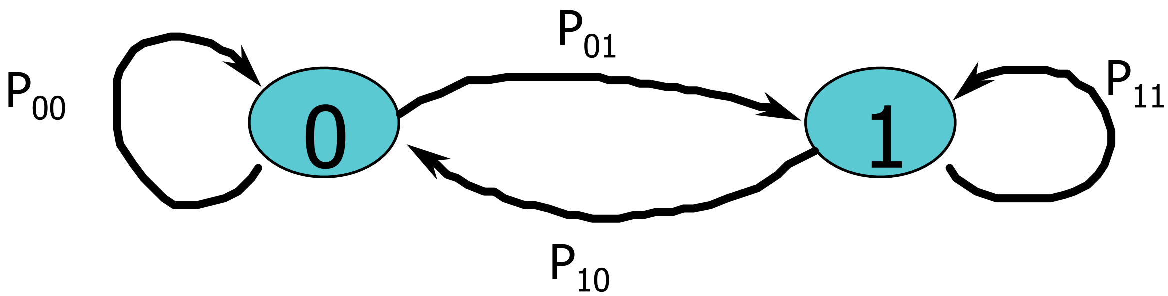
$$\mathbf{P} = \begin{bmatrix} 0.8 & 0.2 \\ 0.9 & 0.1 \end{bmatrix}$$

- State probabilities

$$\mathbf{p}(0) = (p_0(0), p_1(0)) = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

$$\begin{aligned}
\mathbf{p}(n) &= \begin{pmatrix} p_0(n) & p_1(n) \end{pmatrix} \\
&= \begin{pmatrix} P_{00}p_0(n-1) + P_{10}p_1(n-1) & P_{01}p_0(n-1) + P_{11}p_1(n-1) \end{pmatrix} \\
&= \begin{pmatrix} p_0(n-1) & p_1(n-1) \end{pmatrix} \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix} \\
&= \mathbf{p}(n-1)\mathbf{P}
\end{aligned}$$

State Probabilities (3)



$$\mathbf{P} = \begin{bmatrix} 0.8 & 0.2 \\ 0.9 & 0.1 \end{bmatrix}$$

- State probabilities

$$\mathbf{p}(0) = (p_0(0), p_1(0)) = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

$$\mathbf{p}(1) = \mathbf{p}(0)\mathbf{P} = \begin{pmatrix} 1 & 0 \end{pmatrix}\begin{pmatrix} 0.8 & 0.2 \\ 0.9 & 0.1 \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 \end{pmatrix}$$

$$\mathbf{p}(2) = \begin{pmatrix} 0.8 & 0.2 \end{pmatrix}\begin{pmatrix} 0.8 & 0.2 \\ 0.9 & 0.1 \end{pmatrix} = \mathbf{p}(0)\mathbf{P}^2 = \begin{pmatrix} 0.82 & 0.18 \end{pmatrix}$$

$$\mathbf{p}(8) = \begin{pmatrix} 1 & 0 \end{pmatrix}\begin{pmatrix} 0.8 & 0.2 \\ 0.9 & 0.1 \end{pmatrix}^8 = \begin{pmatrix} 0.818 & 0.182 \end{pmatrix}$$

State Probabilities (4)

- State probabilities are represented as a vector

$$\mathbf{p}(n) = [p_0(n), p_1(n), \ldots, p_K(n)]$$

- Two cases:
  1. Taking initial state $\mathbf{p}(0)$ into account

$$\mathbf{p}(n) = \mathbf{p}(n-1)\mathbf{P}$$
$$\mathbf{p}(n) = \mathbf{p}(0)\mathbf{P}^n$$

  2. Limiting state probabilities as $n \to \infty$

Limiting State Probabilities (1)

- Obtained when Markov chains runs for a long time
  - No effect of transients due to **p**(0)

$$\pi = \lim_{n \to \infty} [p_0(n), p_1(n), \ldots, p_K(n)]$$

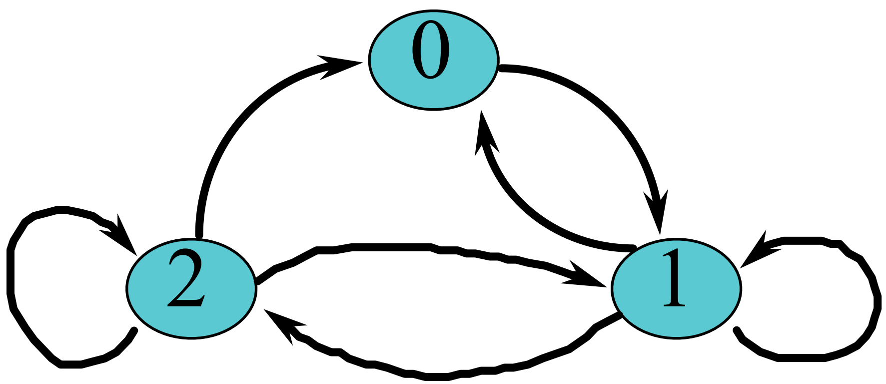
$$= [\pi_0, \pi_1, \ldots, \pi_K]$$

$$\pi = \lim_{n \to \infty} \mathbf{p}(n) = \lim_{n \to \infty} \mathbf{p}(0)\mathbf{P}^n$$

  - Sometimes a little hard to evaluate

Limiting State Probabilities (2)

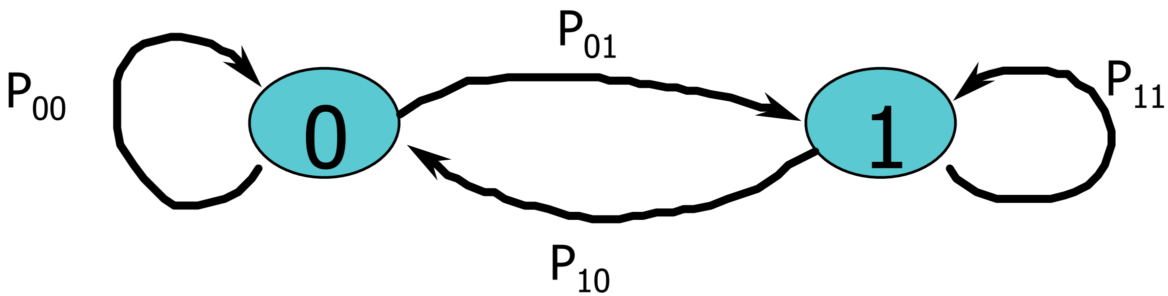
- Easier calculation: $\mathbf{p}(n) = \mathbf{p}(n-1)\mathbf{P}$ $\Rightarrow$ $\pi = \pi\,\mathbf{P}$



$$\begin{cases} \pi_0 = \pi_0 P_{0,0} + \pi_1 P_{1,0} + \pi_2 P_{2,0} \\ \pi_1 = \pi_0 P_{0,1} + \pi_1 P_{1,1} + \pi_2 P_{2,1} \end{cases}$$

$$\pi_0 + \pi_1 + \pi_2 = 1$$

Limiting State Probabilities (3)



$$\mathbf{P} = \begin{bmatrix} 0.8 & 0.2 \\ 0.9 & 0.1 \end{bmatrix}$$

$$\pi = \pi\,\mathbf{P}$$

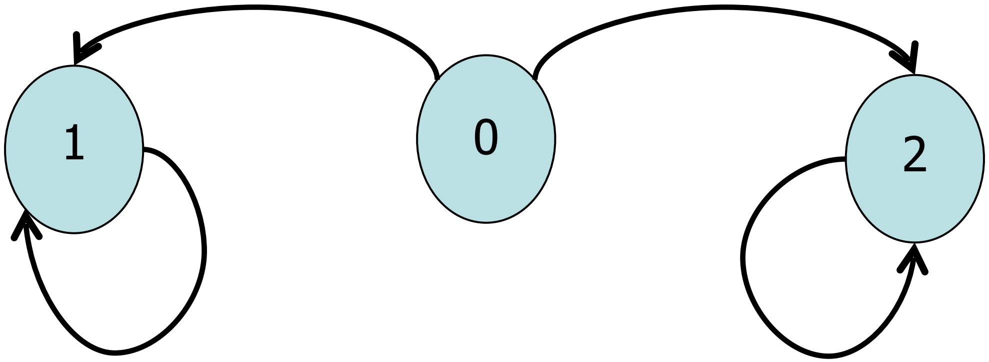
$$\pi(0) = \pi(0)P_{00} + \pi(1)P_{10} = 0.8\pi(0) + 0.9\pi(1)$$

$$\pi(0) + \pi(1) = 1$$

$$\Rightarrow \pi(0) = 0.8\pi(0) + 0.9[1 - \pi(0)]$$

$$\Rightarrow \pi(0) = \frac{0.9}{1.1} = 0.818$$

Limiting state probabilities?

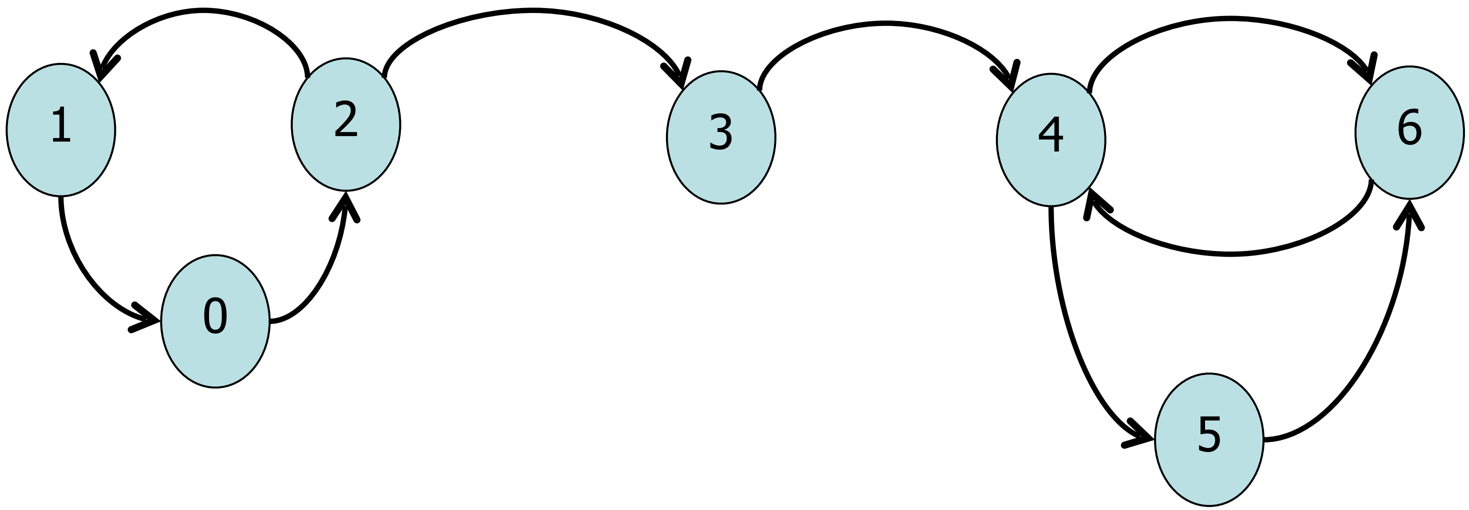


What are the limiting state probabilities here?

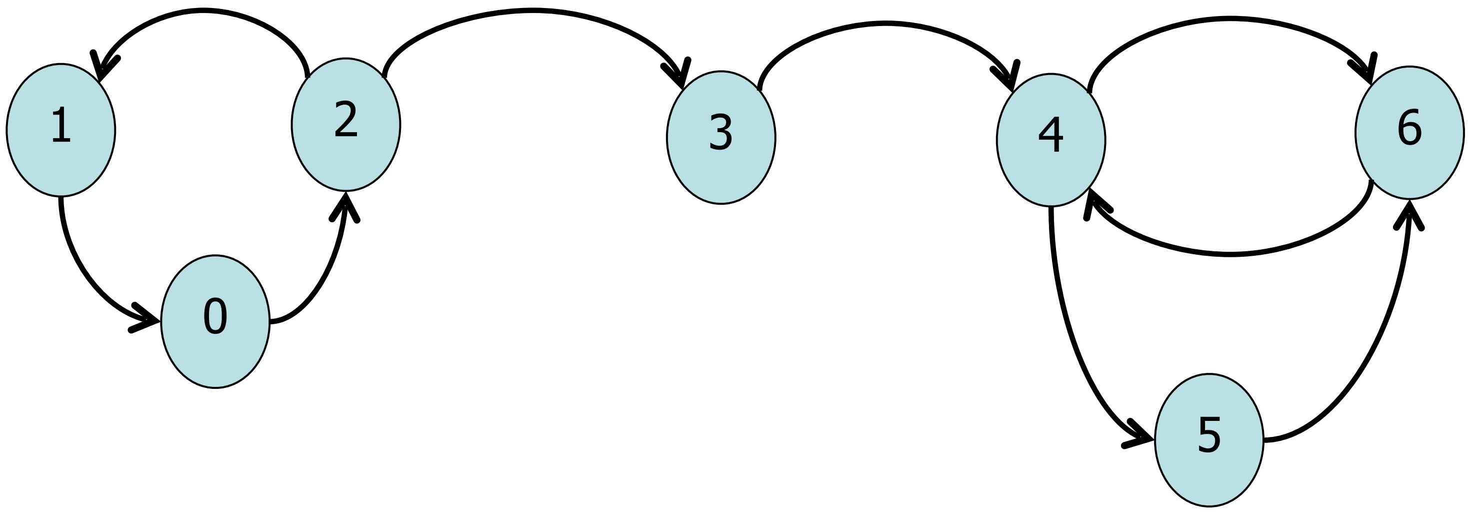
State Classification

- **Accessibility**: State j is accessible from state i if $P_{ij}(n) > 0$ for some $n>0$. ($i \rightarrow j$)
- **Communicating states**: States i and j communicate if $i \rightarrow j$ and $j \rightarrow i$. ($i \leftrightarrow j$)
- **Communicating Class**: A nonempty subset of states in which all states communicate

What are the communicating classes?

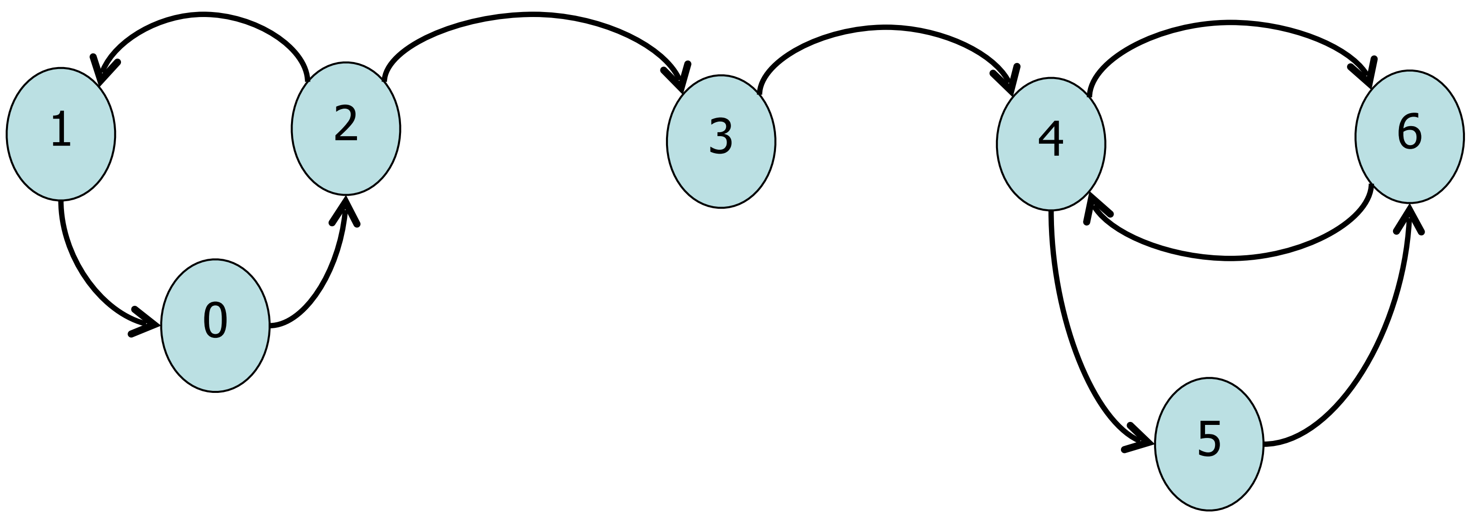


What are the communicating classes?



Classes: C1={0,1,2}, C2={3},C3={4,5,6}

What are the communicating classes?



Classes: C1={0,1,2}, C2={3},C3={4,5,6}

state i always communicates
with itself (i→i in zero steps)

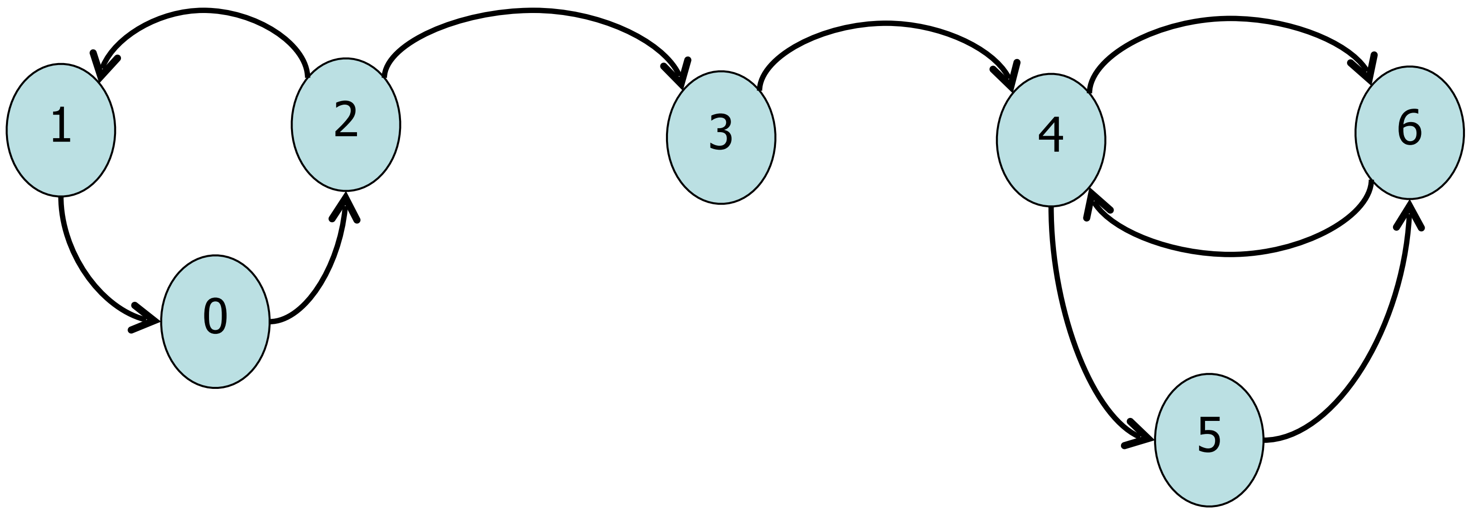
State Classification

- **Periodic**: State i has period d, if d is the largest integer that divides the length of all paths to state i. (gcd)

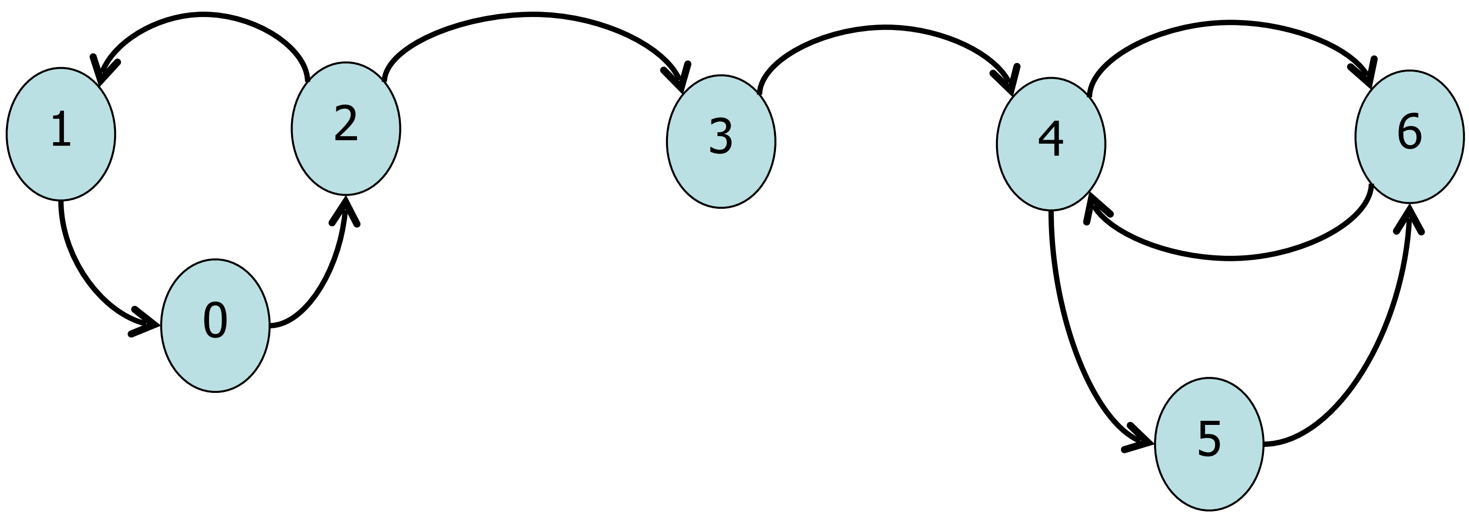
- **Aperiodic**: State i is aperiodic if d=1.

- All states in a communicating class have the same period.

Example



Example

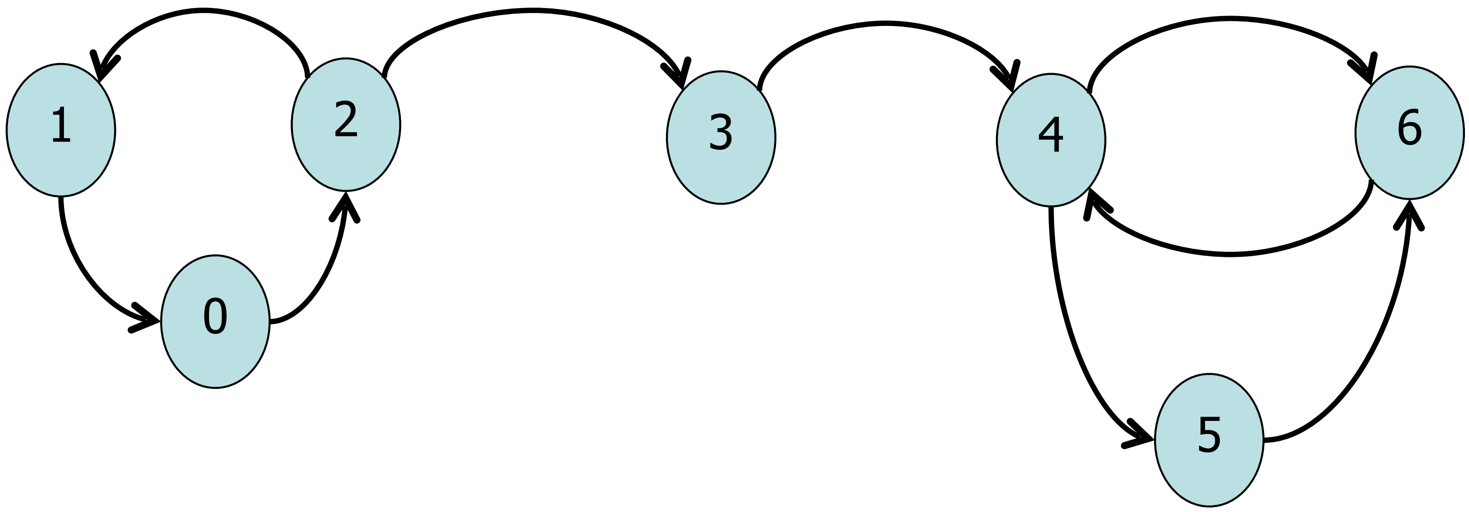


d=3 for C1
d=1 for C3 (aperiodic)

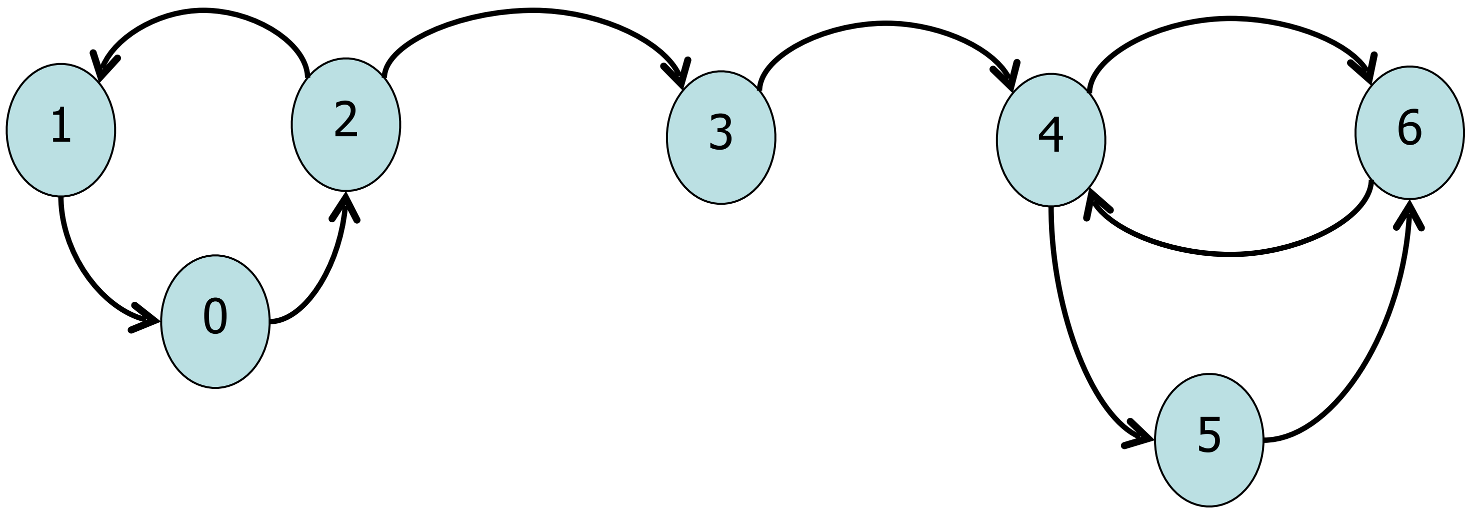
State Classification

- **Transient**: State i is transient if it is possible to leave the state and never return.
- **Recurrent**: State i is recurrent if it is possible to leave the state and return again.
- **Irreducible Markov Chain**: A Markov chain is irreducible if there is only one communicating class.

Example



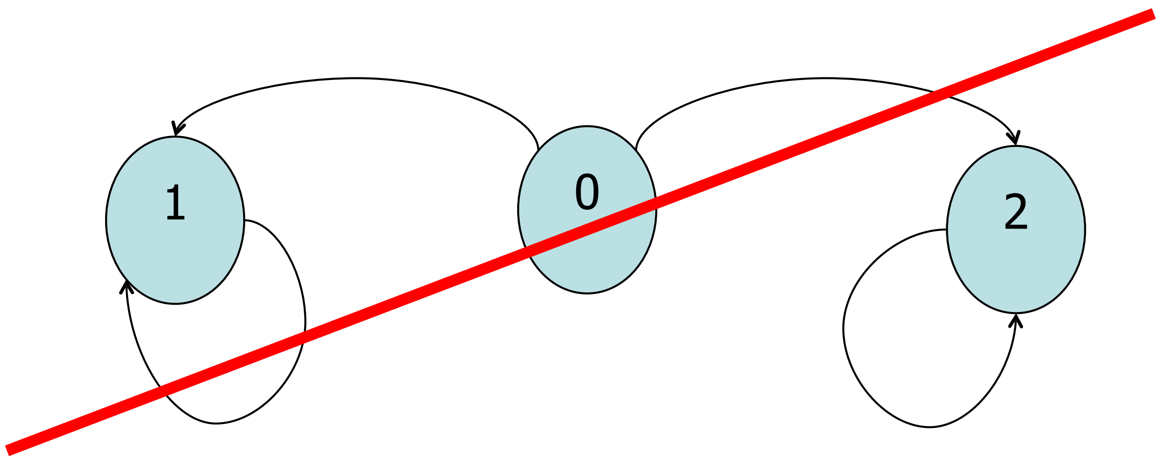
Example



C1, C3=recurrent, C2=transient

Limiting state probabilities

- Limiting state probabilities only exist for
  - Irreducible,
  - Aperiodic,
- Markov chains!



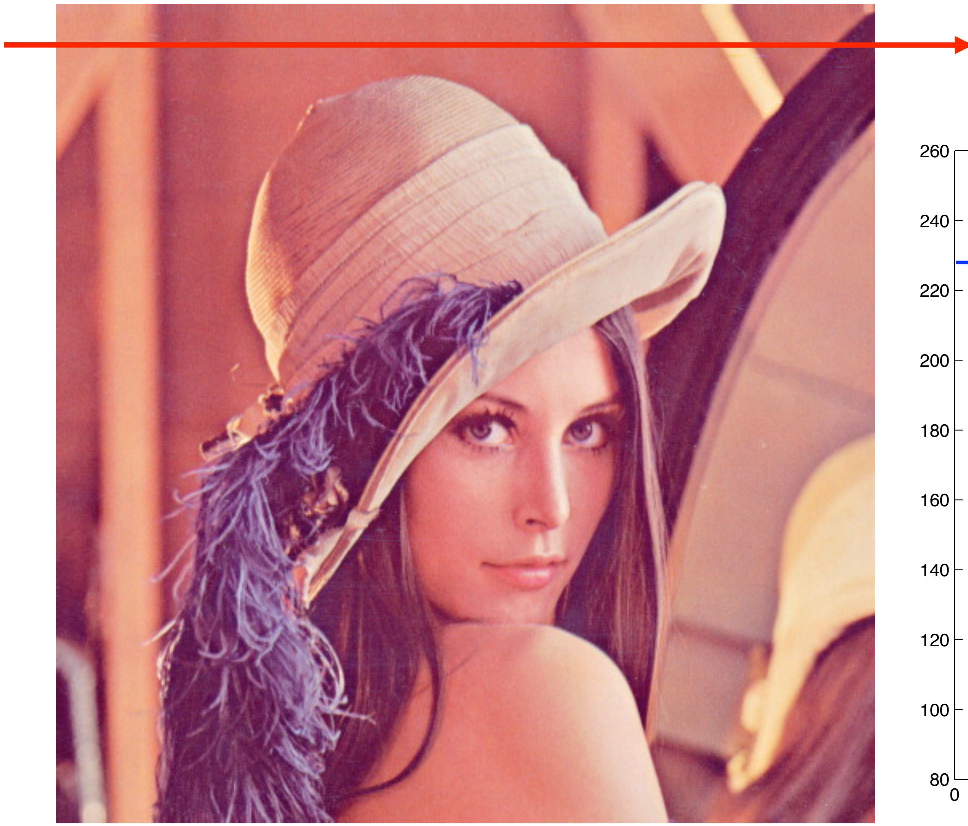
Random Processes:

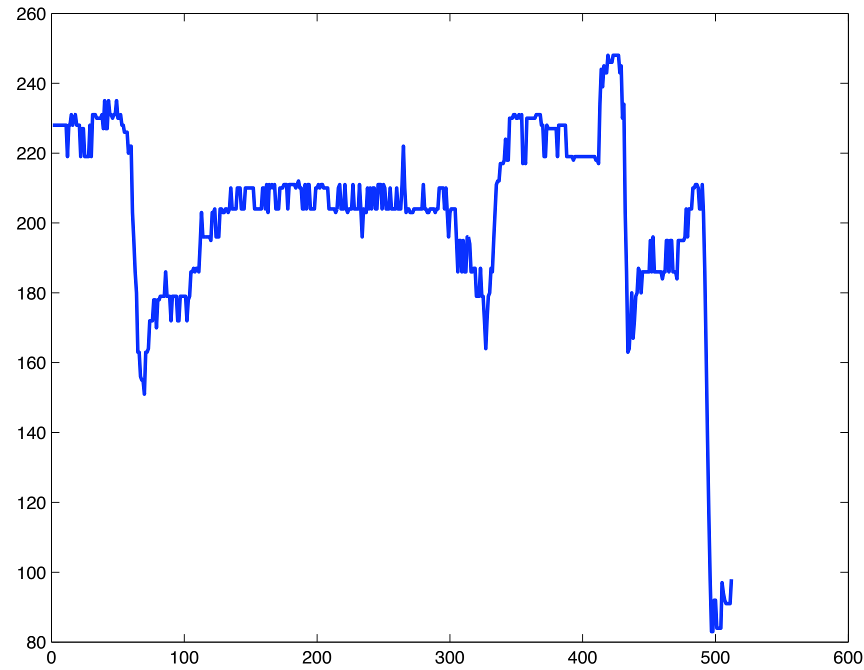
- Markov Chains

- Hidden Markov Models

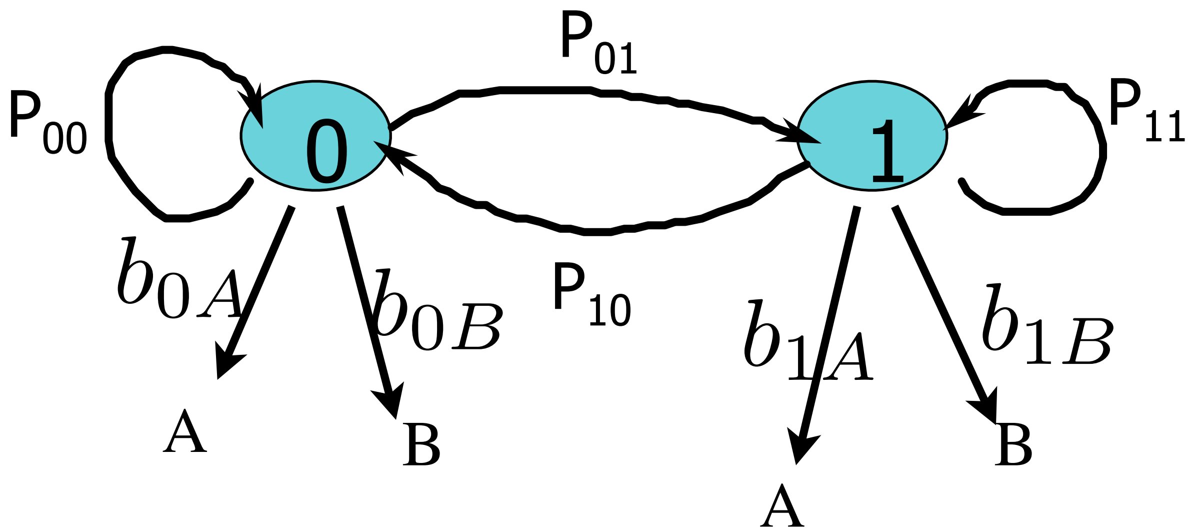
Hidden Markov Models

- In most cases the states cannot be observed directly:





Extend the model with observations



- States are called X
- Transition probabilities are called P
- Observations are called V (here: discrete obs.)
- Emission probabilities are called b

- States are **hidden**.

HMM with hidden states

- The states are not known, but it is assumed that each state has a different probability of generating observations
- Observations are here assumed to be discrete (continuous observations are also easily possible, but it is harder to explain)

- These types of models are often used in speech recognition:
  - the states are the phonemes,
  - the observations are (extracted) sound features

Hidden Markov Model

- Traditionally the following three central issues are discussed:

1. The **evaluation** problem

   compute the probability that a sequence of observations is generated by the HMM

2. The **decoding** problem

   derive the most likely sequence of hidden states, given a sequence of observations

3. The **learning** problem

   determine the probabilities, given sequence(s) of observations

Evaluation problem

- Given the transition probabilities
$$P_{ij} = P(X_{n+1} = j | X_n = i)$$
and the emission probabilities
$$b_{jk} = P(V_n = k | X_n = j)$$
can we estimate the probability that a certain sequence was generated?
$$\mathbf{V} = (V_1, V_2, ..., V_T)$$
- Yes:
$$P(\mathbf{V}) = \sum_{r=1}^{r_{max}} P(\mathbf{V}|\mathbf{X}_r)P(\mathbf{X}_r)$$
where
$$\mathbf{X}_r = (X_1, X_2, ..., X_T)$$
is a particular sequence.

Evaluation problem

- Given the transition probabilities
$$P_{ij} = P(X_{n+1} = j | X_n = i)$$
and the emission probabilities
$$b_{jk} = P(V_n = k | X_n = j)$$
can we estimate the probability that a certain sequence was generated?
$$\mathbf{V} = (V_1, V_2, ..., V_T)$$
- Yes:
$$P(\mathbf{V}) = \sum_{r=1}^{r_{max}} P(\mathbf{V}|\mathbf{X}_r)P(\mathbf{X}_r)$$
where
$$\mathbf{X}_r = (X_1, X_2, ..., X_T)$$
is a particular sequence.

NOTE: sum over all possible sequences!

Evaluation problem

- We assumed the Markov property, so

$$P(\mathbf{X}) = P(X_1) \prod_{n=2}^{T} P(X_n | X_{n-1})$$

- Further, we assumed that the observations only depend on the current hidden state:

$$P(\mathbf{V}|\mathbf{X}) = \prod_{n=1}^{T} P(V_n | X_n)$$

- Combined:

$$P(\mathbf{V}) = \sum_{r=1}^{r_{\max}} \prod_{n=1}^{T} P(V_n | X_n) P(X_n | X_{n-1})$$

HMM evaluation problem: trellis

- What is the probability to observe:

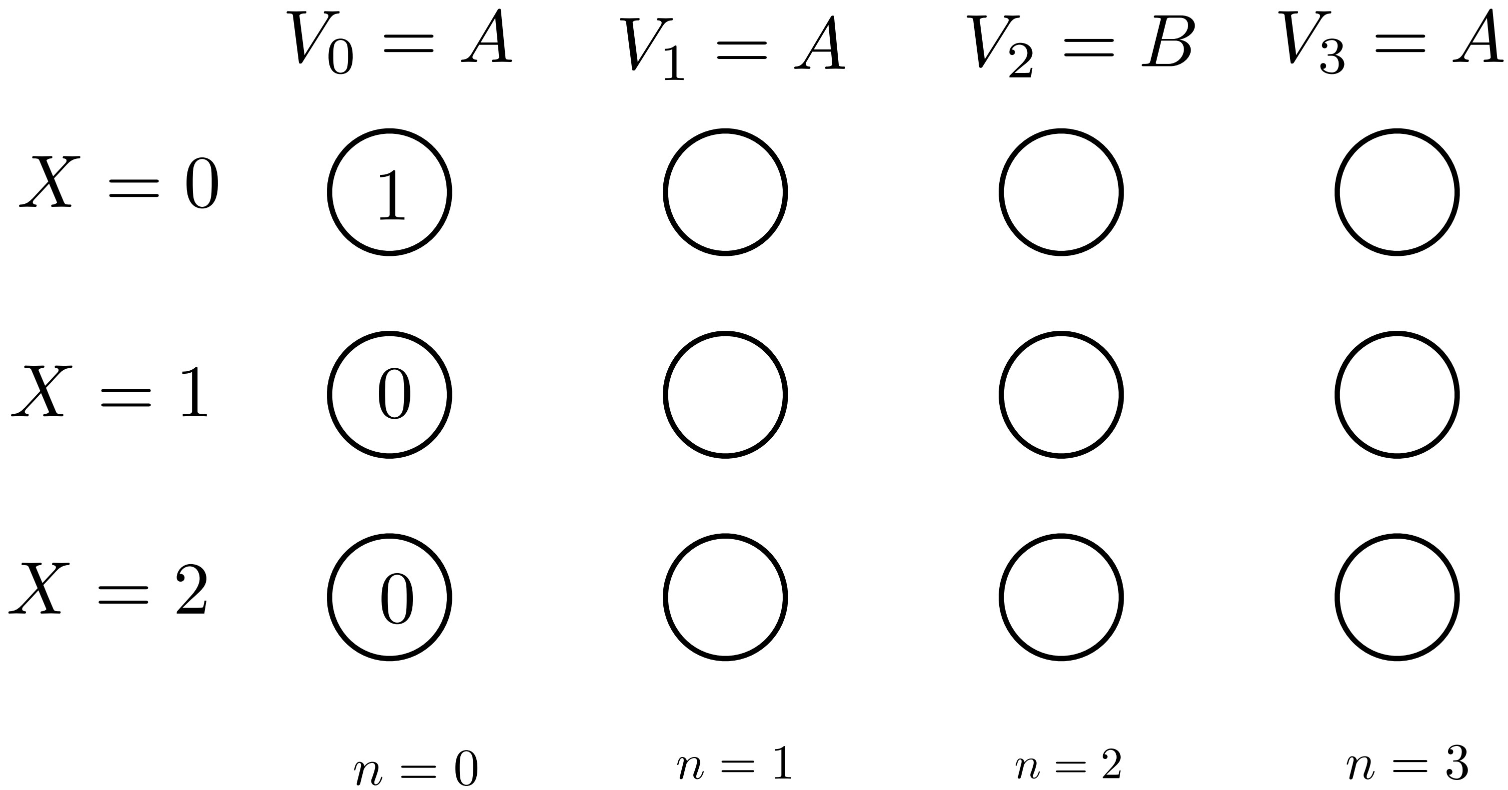
$$\mathbf{V} = (A, A, B, A)$$

- Assume 3 hidden states

$$X = 0,\ X = 1,\ X = 2$$

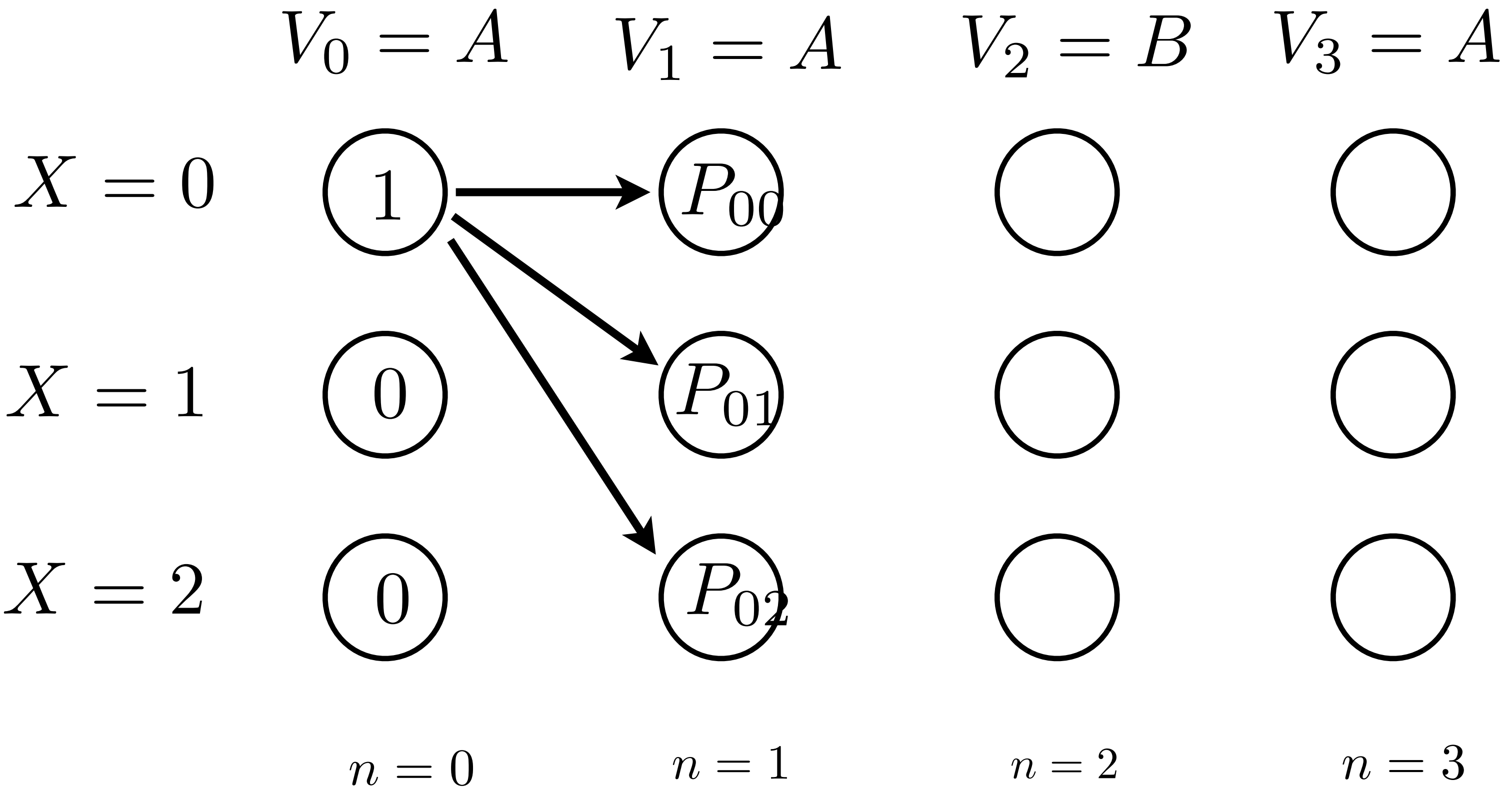
- Assume I know the transition probabilities, and emission probabilities
- Assume I start in $X = 0$

HMM evaluation problem: trellis



- Trellis contains all combinations of states and time points

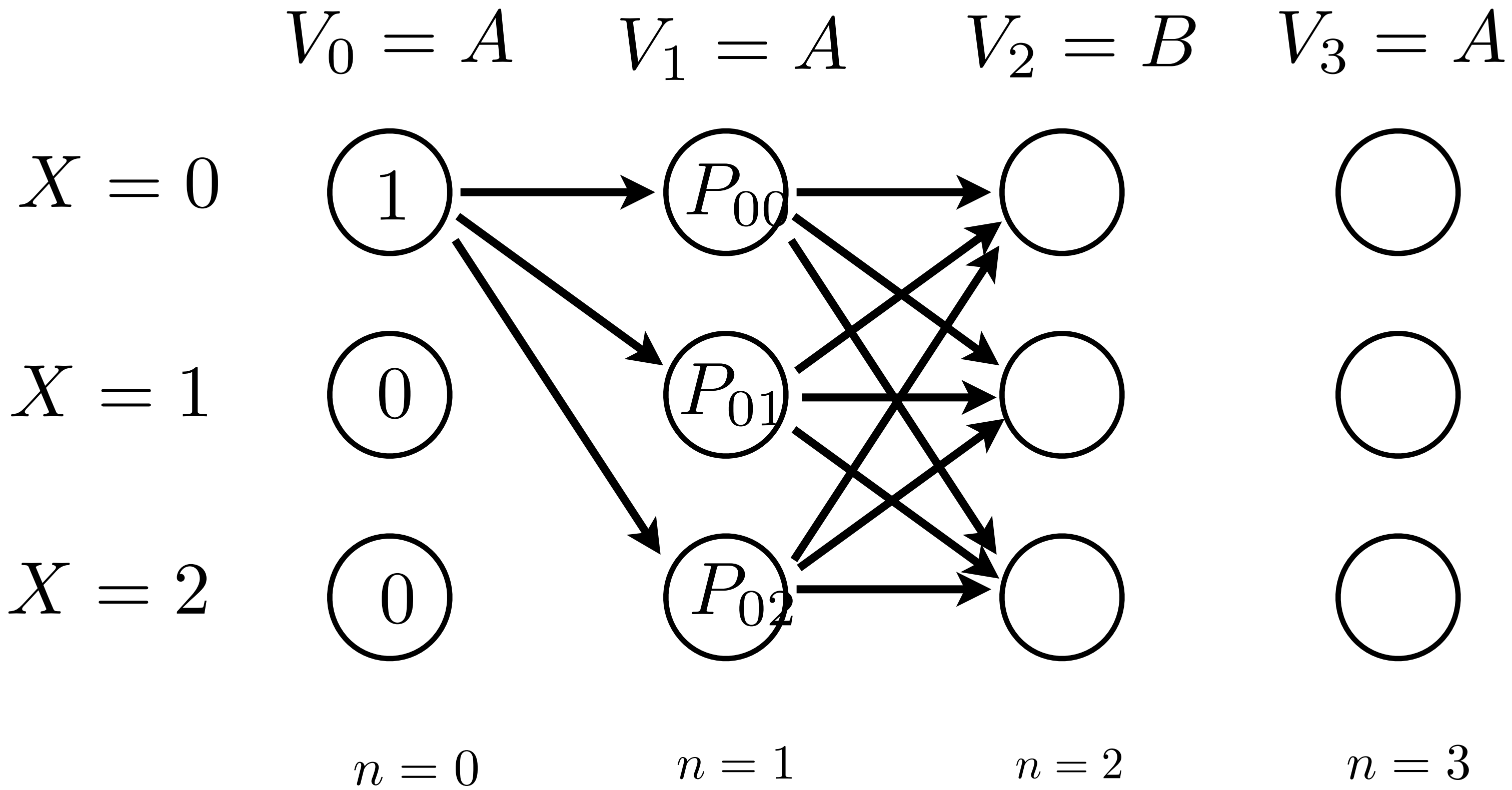
HMM evaluation problem: trellis



- The probability to observe A at n=1 is:

$$P_{00}b_{0A} + P_{01}b_{1A} + P_{02}b_{2A}$$

HMM evaluation problem: trellis



- The probability to observe B at n=2 is:

$$(P_{00}P_{00} + P_{01}P_{10} + P_{02}P_{20})b_{0B} + \ldots$$

HMM Forward algorithm

$$P(\mathbf{V}) = \sum_{r=1}^{r_{max}} \prod_{n=1}^{T} P(V_n|X_n)P(X_n|X_{n-1})$$

- We are given the observations $\mathbf{V} = (V_1, V_2, ..., V_T)$ and the probabilities $P(X_n|X_{n-1})$ $P(V_n|X_n)$
- Although the equation looks complicated, an efficient computation can be done using the forward algorithm

$$\alpha_i(n) = \begin{cases} 0 & n = 0, i \neq \text{initial state} \\ 1 & n = 0, i = \text{initial state} \\ \sum_j \alpha_j(n-1)P_{ij}b_{jk}V_n & \text{otherwise} \end{cases}$$

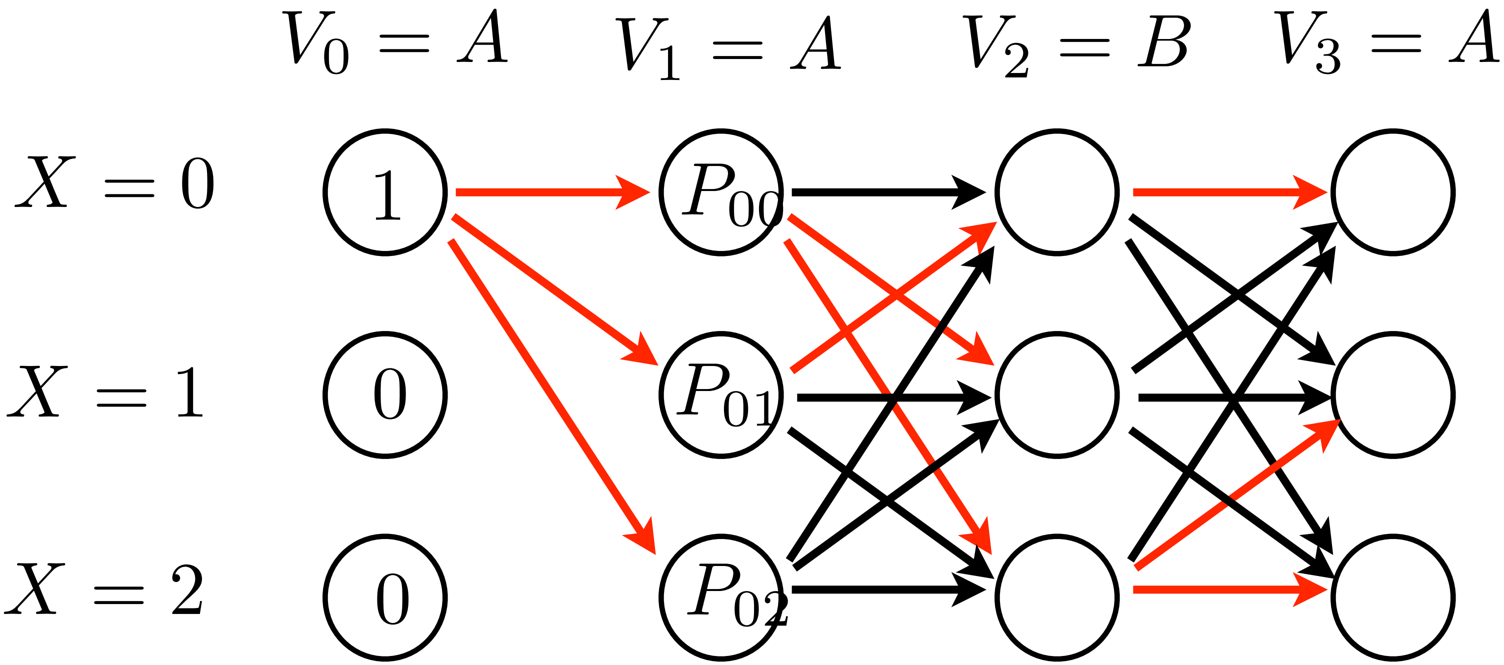
- The sequence probability becomes

$$P(\mathbf{V}) = \alpha_{V_T}(T)$$

HMM Decoding

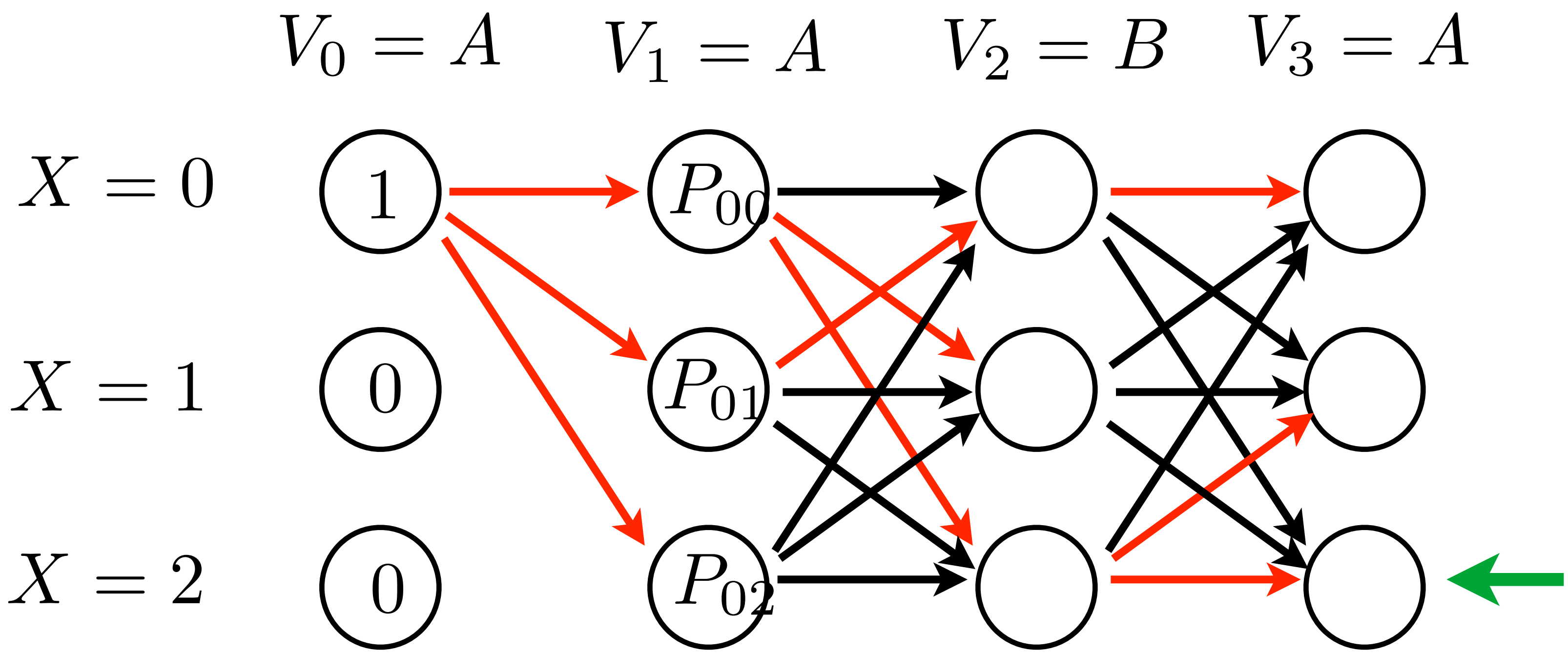
- We can now only find the probability of a sequence of observations, given a HMM model. But what is the most likely sequence of hidden states?
- In principle: try all possible sequences of hidden states, and compute $P(\mathbf{V}|\mathbf{X})$
- That is too much.
- But also for this a more efficient algorithm is possible, using the trellis that was also used in the forward algorithm

HMM decoding problem



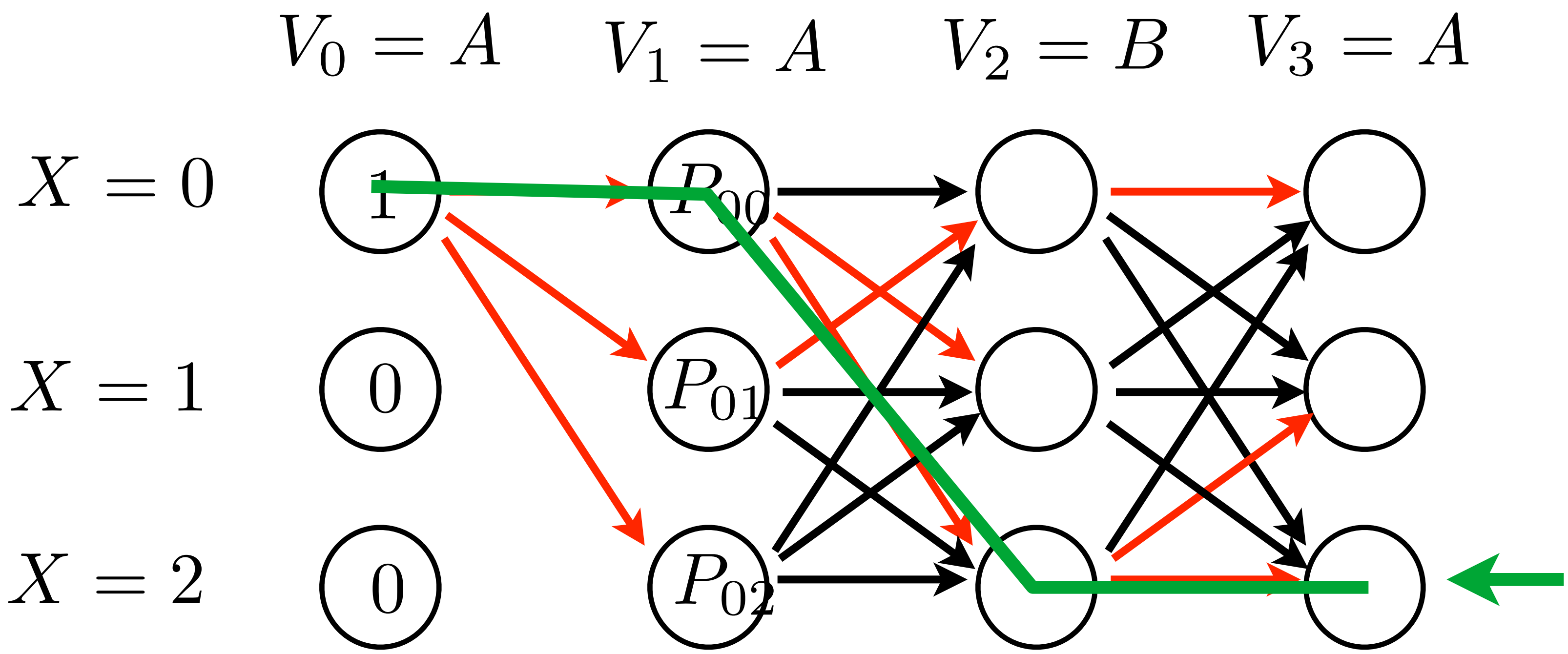
- Remember for each state the most likely previous state
- Backtrack from the final state and reconstruct the most likely sequence

HMM decoding problem



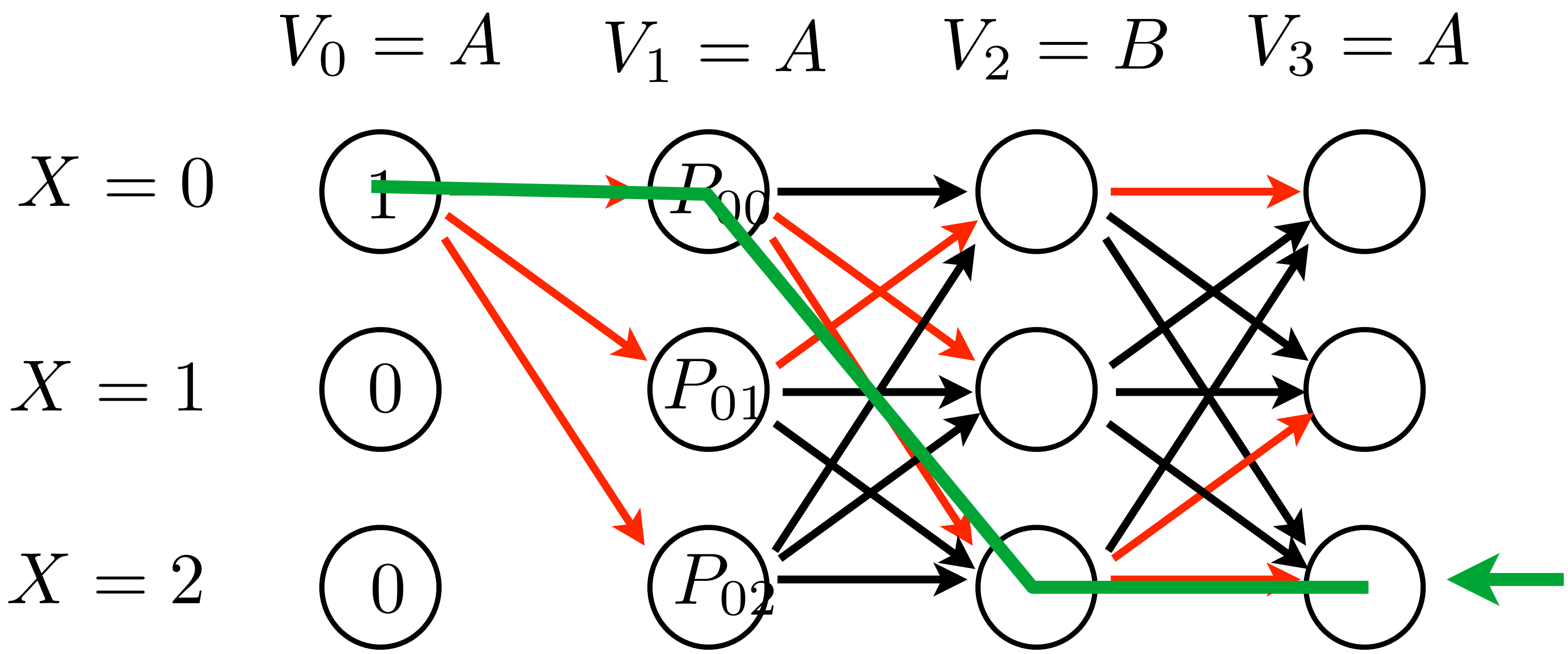
- Remember for each state the most likely previous state
- Backtrack from the final state and reconstruct the most likely sequence

HMM decoding problem



- Remember for each state the most likely previous state
- Backtrack from the final state and reconstruct the most likely sequence

HMM decoding problem



- Remember for each state the most likely previous state
- Backtrack from the final state and reconstruct the most likely sequence $\mathbf{X} = (0, 0, 2, 2)$

HMM learning problem

- In some cases (our exercises...) all probabilities are given. In most normal cases you have to **fit** them.
- Use Maximum Likelihood:

$$\max_{P_{ij},b_{jk}} P(\mathbf{V}|\mathbf{P},\mathbf{b}) = \max_{P_{ij},b_{jk}} \sum_{r=1}^{r_{max}} P(\mathbf{V},\mathbf{X}_r|\mathbf{P},\mathbf{b})$$

- Again, sum over exponentially many possible state sequences, AND no closed from solution
- Apply Expectation-Maximization:
  - Given $\mathbf{V}$ and P_{ij}, b_{jk} maximize $\mathbf{X}_r$
  - Given $\mathbf{V}$ and $\mathbf{X}_r$ maximize P_{ij}, b_{jk}
  - iterate

HMM learning problem

- To avoid the large sum, again the trellis is used
- Both a forward pass and a backward pass is required.
- I skip the technicalities

$$P(\mathbf{V}) = \sum_{r=1}^{r_{max}} \prod_{n=1}^{T} P(V_n|X_n)P(X_n|X_{n-1})$$

- Note that we are working with large products of probabilities: for long sequences the total probability goes to zero
- Normalization strategies are proposed, or the log-probabilities

TU Delft

HMM with continuous observations

- Fit the HMM on a single line, find the most likely state sequence in all other lines
- Observations are modelled by Mixture of Gaussians

Summary

- See file on Brightspace: ‘markov-chains_student.pdf’
- Markov Chains
  - Transition probabilities/diagram/matrix
  - Steady-state probabilities
  - Limiting probabilities
  - State classification

TU Delft

On the exam

- Some example exams on Brightspace
- A physical exam!
- For the final exam, you are allowed to use/make one A4 paper with your own notes (both sides can be used)
- (Graphical) calculator is allowed.

- Fourier transforms, math facts, distribution names and formulas will be provided (if needed)