DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF COPENHAGEN

Deterministiske og stokastiske tilstandsmaskiner, grammatikker og regulære udtryk – del II

Kim Steenstrup Pedersen

Plan for denne uge

- Tilstandsmaskiner
  - Deterministiske endelige automata (DFA)
  - Som generel programstruktur
- Kontekstfri grammatikker og beregningsudtryk
- Regulære udtryk
- Stokastiske tilstandsmaskiner
  - Fra DFA til stokastiske tilstandsmaskiner
  - Markov kæder
  - Lidt sandsynlighedsregning

Regulære udtryk

- Regulære udtryk er et simpelt programmeringssprog til at finde mønstre i strenge.
- Et mønster er angivet med en streng som kan indeholde bogstaver, tal etc. (terminaler) og specialtegn og -notation (nonterminaler).
- Eksempel:
  Mønster: `hat`
  Streng og match: `Katten med hatten`
- Avanceret eksempel:
  Mønster: `[abt]+`
  Streng og begge match: `Katten med hatten`

Regulære udtryk i Python – `re` modulet

- Python har et standard modul til regulære udtryk kaldet `re`
- Virker både med ASCII (`str`) og `unicode` strenge
- Der er både en funktionsgrænseflade i `re` og et mønster objekt

```
import re
# Mønster objekt
patt = re.compile(r'hat')
mobj = patt.search('Katten med hatten')
print mobj.group(0), 'at', mobj.span()
# Funktionskald
findlist = re.findall(r'[abt]+', 'Katten med
hatten')
```

Regulære udtryk – `re` syntaks

- Et regulært mønster udtryk kan skrives ved hjælp af
  - Tegn som ikke er specialtegn matches som de er.
  - Tegn indenfor `[]` angiver en mængde af tegn som skal matches. `[abc1-6]`
  - Parenteser `()` gruppere mønstre. `(hat)`
  - + betyder match 1 eller flere gange efter hinanden. `(hat)+`
  - `*` betyder match 0 eller flere gange efter hinanden. `(hat)*`
  - ? betyder match 0 eller 1 gang efter hinanden. `(hat)?`
  - Punktum `.` betyder match alle tegn undtaget linjeskift. `.*`
  - | kan bruges som 'eller' tegn. `(hat)|(kat)`
  - Prædefineret tegn mængder `\n, \s, \w`
  - Mønsterstrenge bør angives som raw strenge (med `r` foran) ellers skal `\n` skrives som `\\n`. `r'hej\w'` ekvivalent med `'hej\\w'`
- Og der er mere – læs `re` manualen.

Regulære udtryk i Python – `re` modulet

- Find et match i starten af strengen

```
mobj=re.match(r'Hello[ \t]*(.*)’, 'Hello World!')
```

- Find det første match et sted i strengen

```
mobj = re.search(r'[abt]+', 'Katten med hatten')
```

- Find alle matches

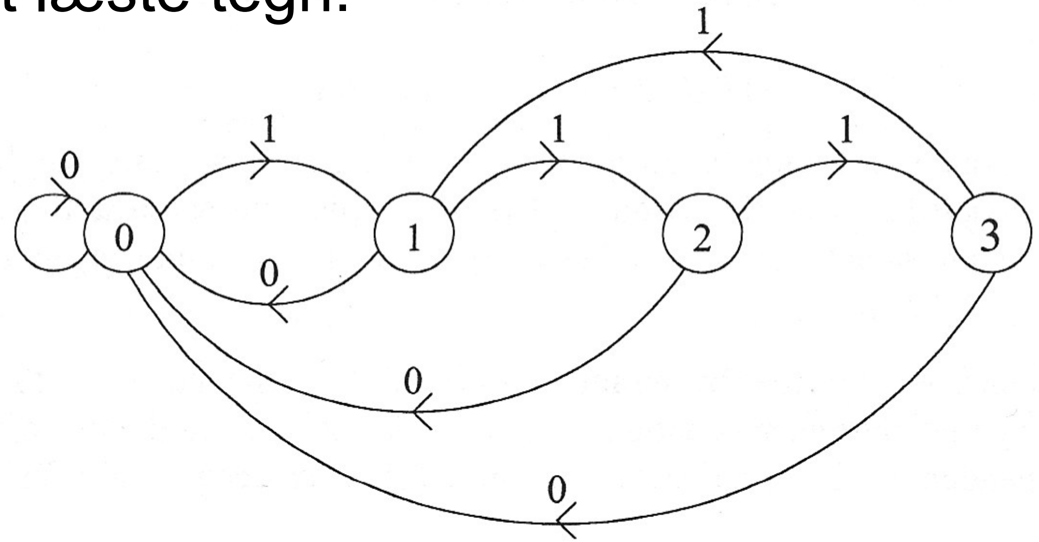
```
dkstreng = u'En å på Ærø'
findlist = re.findall(ur'[rÆøå]+', dkstreng)
```

- Split strengen i delstrenge ved match

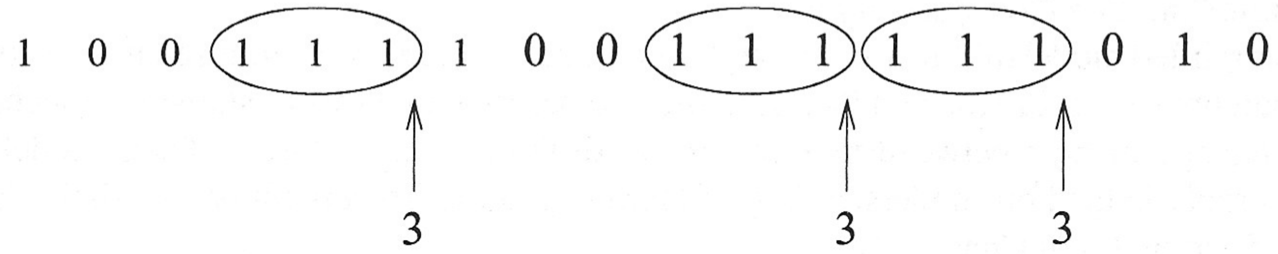
```
Splitlist = re.split(r'[ ]’, 'Katten med hatten’)
```

Automata igen
DFA kan simulerer nogle regulære udtryk, men ikke alle

Maskinen læser tegn fra input streng og skifter tilstande ud fra det sidst læste tegn.

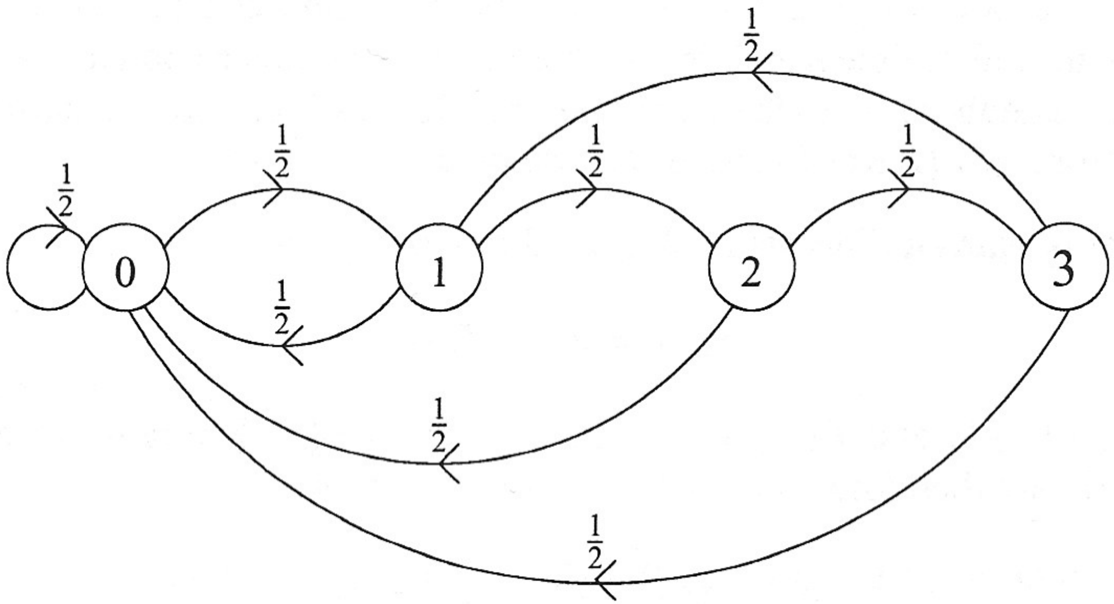


a



Stokastiske tilstandsmaskiner

- Lad os droppe input strengen! Tegn = Tilstande
- Hvad nu hvis tilstandsovergangene bliver valgt tilfældigt – lad os sige vi slår plat-og-krone om det?



- Nu kan vi generere nye tilfældige tegnsekvenser ved at slå plat-og-krone for hver tilstandsovergang i sekvensen.

Stokastiske tilstandsmaskiner

- Hvad nu hvis tilstandsovergangene bliver valgt tilfældigt – lad os sige vi slår plat-og-krone om det?
- Nu kan vi generere nye tilfældige tilstands-/ tegnsekvenser ved at slå plat-og-krone for hver ny tilstandsovergang.
- Eksempel: Vores plat-og-krone gav
  Generede tilstande / streng: 00120123
- Denne model er dog ikke så interessant – vi vil gerne kunne vælge tilstandsovergange, således at der er forskellige sandsynligheder for de efterfølgende tilstande.
- Vi skal bruge lidt sandsynlighedsregning!

Sandsynlighedsregning

- I sandsynlighedsregning arbejder vi med stokastiske variable X som tager værdier fra en mængde Ω
- Stokastiske variable er ikke som andre variable – hver gang vi læser fra den antager den en ny værdi
- Intuitiv analogi: Variablen X er en skuffe
- Kig i skuffen for at læse variablens værdi $\omega \in \Omega$
- Hver gang vi kigger, får vi en ny værdi
- De forskellige værdier forekommer med en vis sandsynlighed $p(X = \omega)$

Sandsynlighedsregning

- Vi er kun interesseret i diskrete stokastiske variable, dvs. som tager værdier fra tællelige mængder Ω
- Eksempler: $\Omega = \{1,2,3,4,5,6\} \quad \Omega = \{a,b,c\}$

- Sandsynlighederne for alle værdier i Ω er en funktion som kaldes sandsynlighedsmassefunktionen $p(X)$

- Sandsynlighedsmassefunktionen har disse egenskaber

$$\sum_{\omega \in \Omega} p(X=\omega) = 1$$ (normaliseret over alle værdier i mængden)

$$p(X=\omega) \geq 0 \quad , \quad \forall \omega \in \Omega$$ (ikke-negativ for alle værdier i mængden)

Eksempel: Terningekast

- Med en fair 6-sidet terning er der lige stor sandsynlighed for at slå 1 som at slå 6 (eller en af de andre sider)

$$p(X=1)=\frac{1}{6} \quad p(X=2)=\frac{1}{6} \quad p(X=3)=\frac{1}{6}$$

$$p(X=4)=\frac{1}{6} \quad p(X=5)=\frac{1}{6} \quad p(X=6)=\frac{1}{6}$$

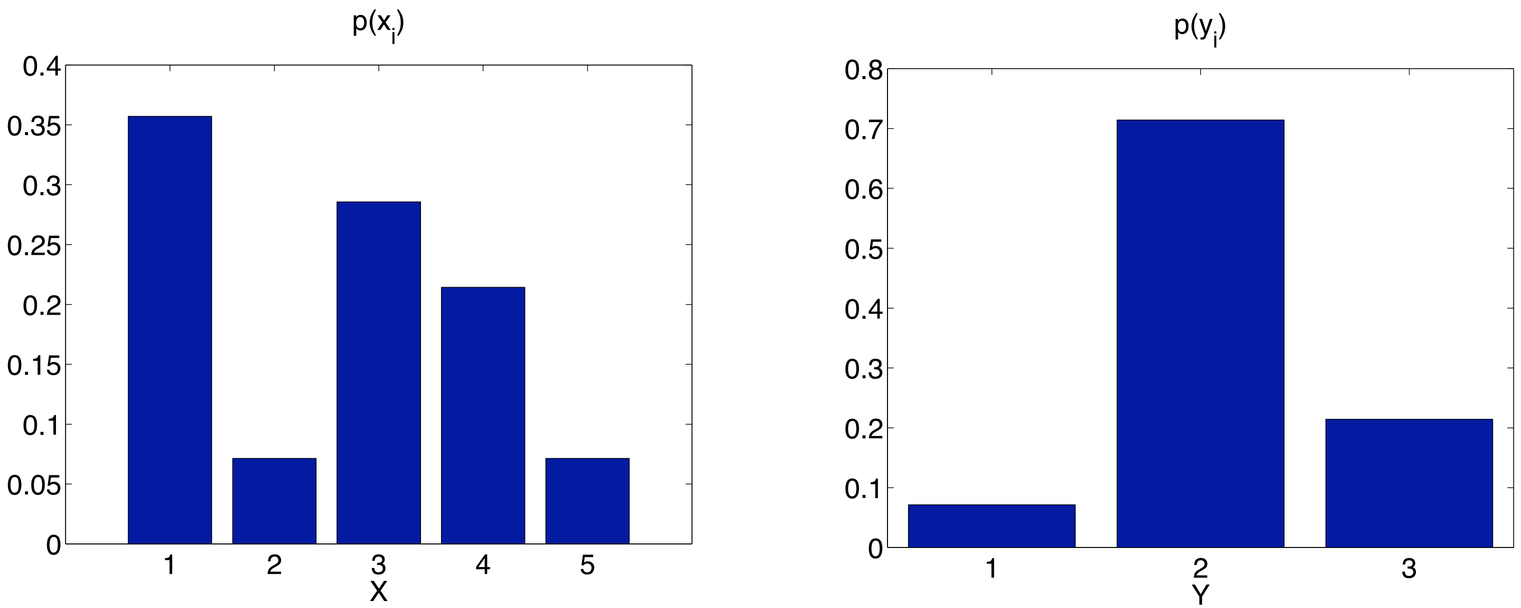
- Vi kan skrive sandsynlighedsmassefunktionen kompakt som en vektor eller tabel $p(X)=(\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6})$

p(X)	X=1	X=2	X=3	X=4	X=5	X=6
	1/6	1/6	1/6	1/6	1/6	1/6

- Vi ser at denne funktion summer til 1 (er normaliseret) og er ikke-negativ for alle værdier af *X*.

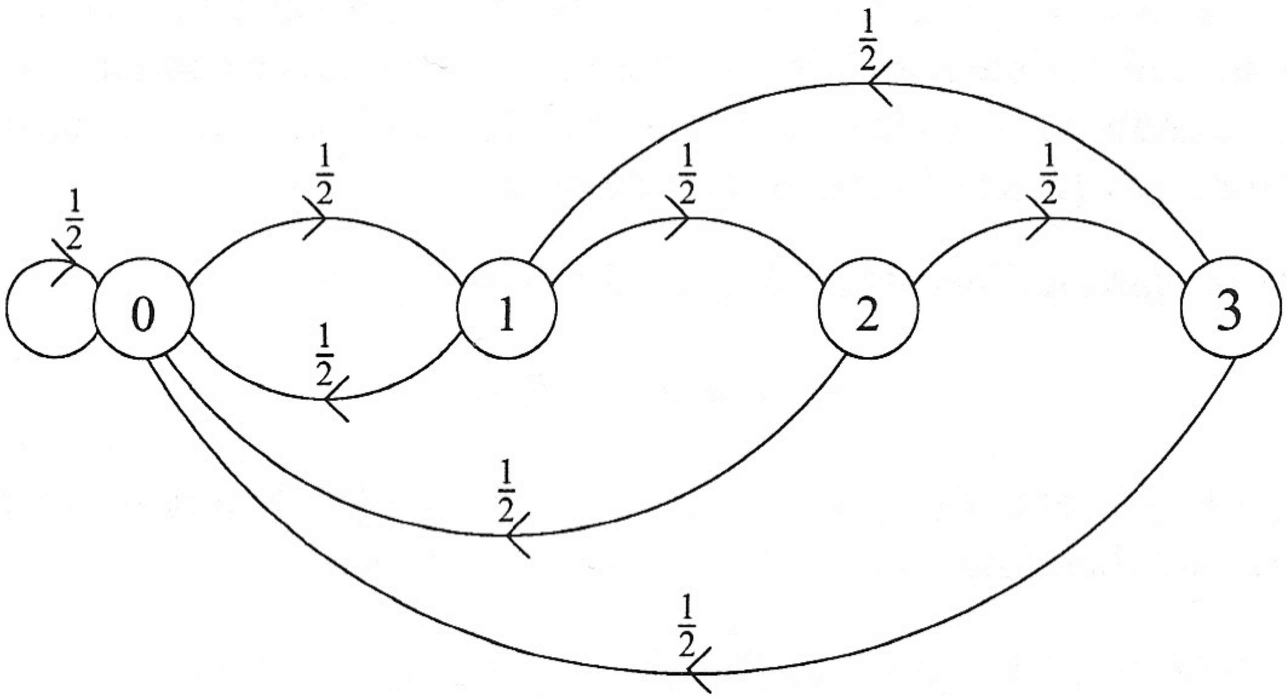
Visualisering af sandsynlighedsmassefunktioner

- Sandsynlighedsmassefunktioner vises oftest grafisk som histogrammer (pinde diagrammer)
- Her er to eksempler:



Tilbage til eksemplet

- Den nuværende tilstand er en stokastisk variabel $X \in \{0,1,2,3\}$

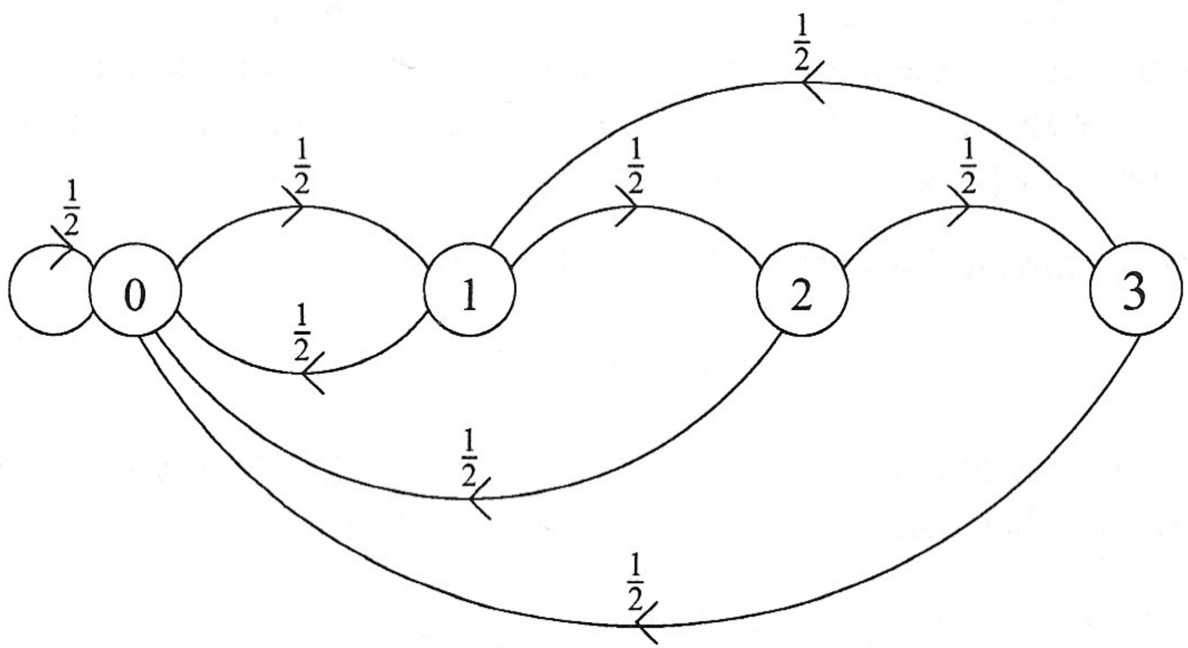


Kæder af stokastiske variable

- Når vi bevæger os rundt mellem tilstande så kan vi tænke på det som en kæde af stokastiske variable X_t
- Vores skridt indekseres med *t*
- Hvis vi tager 2 skridt fra start-tilstanden så er vores kæde 3 stokastiske variable lang $\{X_0, X_1, X_2\}$
- Til næste skridt

  $\{X_0, X_1, X_2, X_3\}$

- Dvs. Kæden vokser for hvert skridt



Lidt mere sandsynlighedsregning

- Hvis to stokastiske variable *X* og *Y* afhænger af hinanden kan deres relation bl.a. udtrykkes ved betingede sandsynligheder $p(Y = \omega_y \mid X = \omega_x)$
- Skal læses som sandsynligheden for *Y*'s værdi givet *X*'s værdi
- Vi kan også tale om den betingede sandsynlighedsmassefunktion $p(Y \mid X = \omega_x)$
- I vores tilstandsmaskine afhænger den nye tilstand af den forrige tilstand

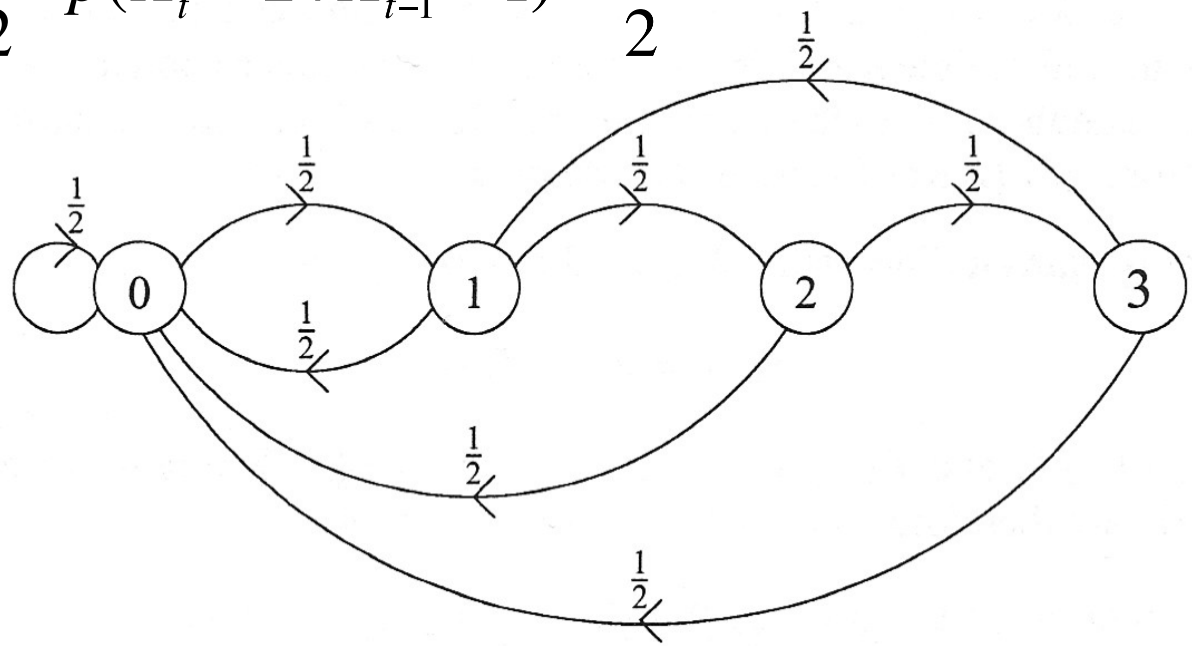
Tilbage til eksemplet

- Sandsynligheden for tilstandsovergangene afhænger af forrige tilstand i kæden

$$p(X_t = 0 \mid X_{t-1} = 0) = \frac{1}{2} \quad p(X_t = 1 \mid X_{t-1} = 0) = \frac{1}{2}$$

$$p(X_t = 0 \mid X_{t-1} = 1) = \frac{1}{2} \quad p(X_t = 2 \mid X_{t-1} = 1) = \frac{1}{2}$$

- osv.



Stokastiske tilstandsmaskiner - opsummering

- En stokastisk tilstandsmaskine er beskrevet ved
  - En mængde af tilstande Ω
  - En kæde af stokastiske variable der beskriver tilstanden i et givet skridt $X_t \in \Omega$ for $t \geq 0$
  - En sandsynlighedsmassefunktion for start-tilstanden i kæden $q(X_0)$
  - En samling af betingede sandsynlighedsmassefunktioner der beskriver sandsynligheder for de forskellige tilstandsovergange $p(X_t \mid X_{t-1} = \omega)$ for alle $\omega \in \Omega$

Stokastiske tilstandsmaskiner – Markov kæder

- Antag at overgangssandsynlighedsmassefunktionerne ikke afhænger af indekset *t,* så
  $p(X_t \mid X_{t-1} = \omega)$ for alle $\omega \in \Omega$ og alle $t \geq 0$
- Så afhænger næste tilstand kun af den foregående tilstand og ikke hvor i kæden af tilstande vi er nået til.
- Denne type af stokastiske tilstandsmaskiner kaldes for en Markov kæde.

Eksempel på en Markov kæde model

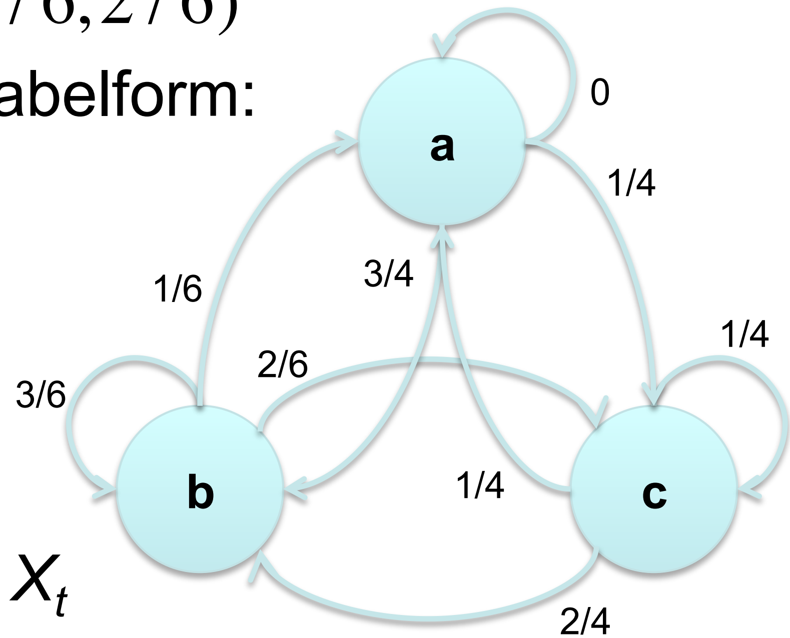
- Tilstande: $\Omega = \{a, b, c\}$ og $X_t \in \Omega$
- Start-tilstandssandsynligheder:

$$q(X_0 = a) = \frac{1}{6} \quad q(X_0 = b) = \frac{3}{6} \quad q(X_0 = c) = \frac{2}{6}$$

- eller på kort form $q(X_0) = (1/6, 3/6, 2/6)$
- Overgangssandsynligheder på tabelform:

$p(X_t \mid X_{t-1})$	$X_t=a$	$X_t=b$	$X_t=c$
$X_{t-1}=a$	0	3/4	1/4
$X_{t-1}=b$	1/6	3/6	2/6
$X_{t-1}=c$	1/4	2/4	1/4

Rækkerne er sandsynligheder for X_t



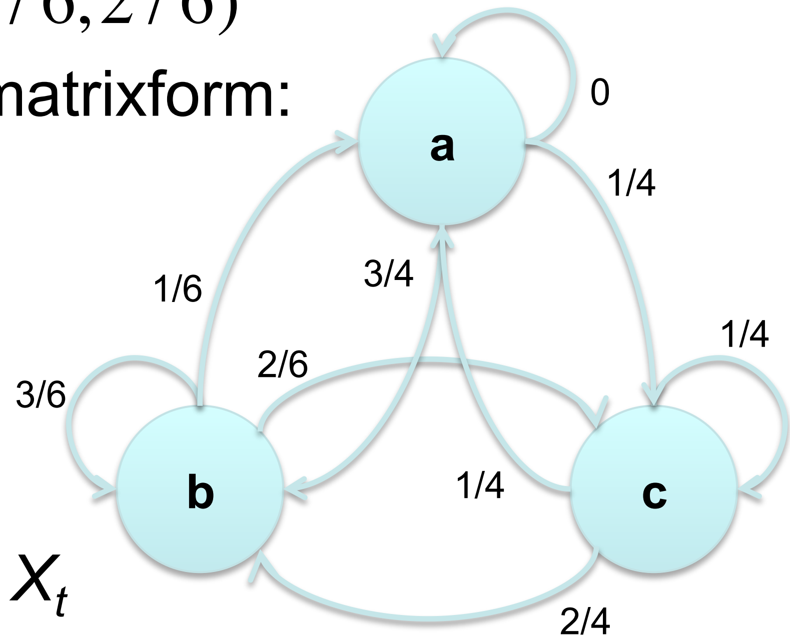
Eksempel på en Markov kæde model

- Tilstande: $\Omega = \{a, b, c\}$ og $X_t \in \Omega$
- Start-tilstandssandsynligheder:

$$q(X_0 = a) = \frac{1}{6} \quad q(X_0 = b) = \frac{3}{6} \quad q(X_0 = c) = \frac{2}{6}$$

- eller på kort form $q(X_0) = (1/6, 3/6, 2/6)$
- Overgangssandsynligheder på matrixform:

$$p(X_t \mid X_{t-1}) = \begin{bmatrix} 0 & 3/4 & 1/4 \\ 1/6 & 3/6 & 2/6 \\ 1/4 & 2/4 & 1/4 \end{bmatrix}$$

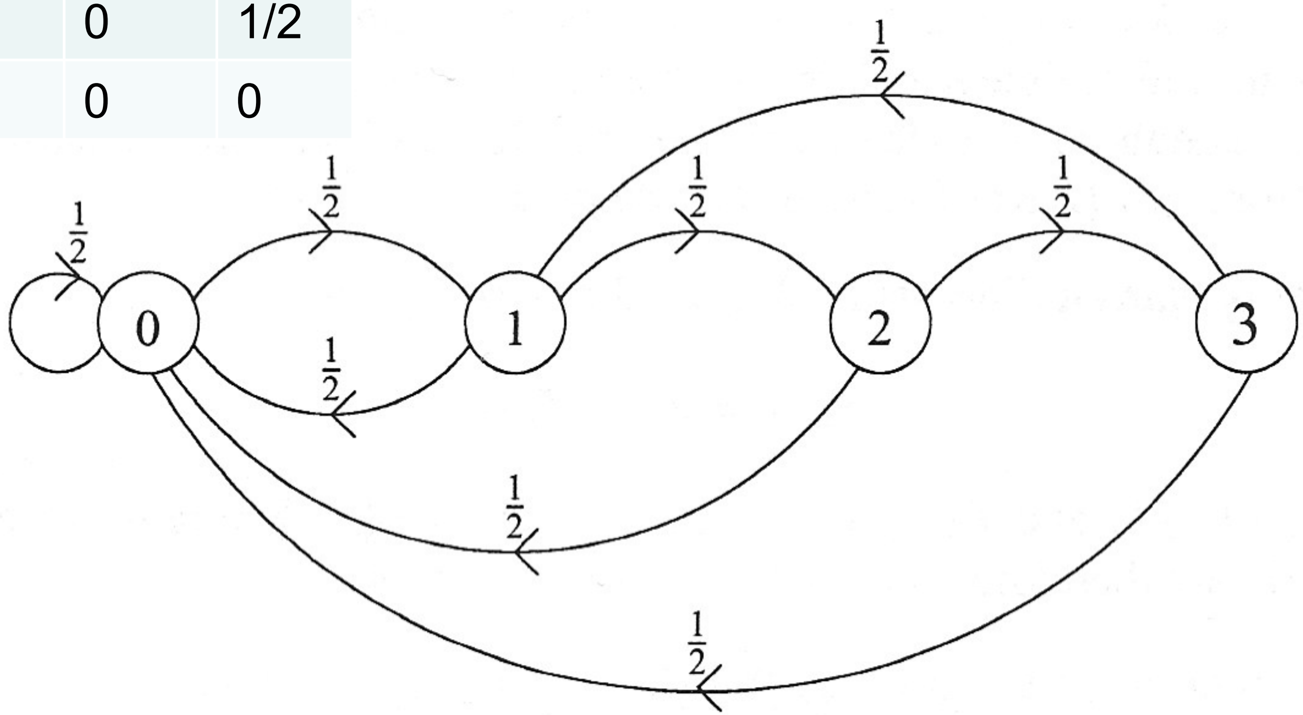


Rækkerne er sandsynligheder for X_t

Markov kæder og DFA

Vores model fra før er en Markov kæde. (Det er faktisk muligt at formulere en DFA som en Markov kæde model)

$p(X_t \mid X_{t-1})$	X_t=0	X_t=1	X_t=2	X_t=3
X_{t-1}=0	1/2	1/2	0	0
X_{t-1}=1	1/2	0	1/2	0
X_{t-1}=2	1/2	0	0	1/2
X_{t-1}=3	1/2	1/2	0	0



Hvad kan vi bruge Markov kæde modeller til?

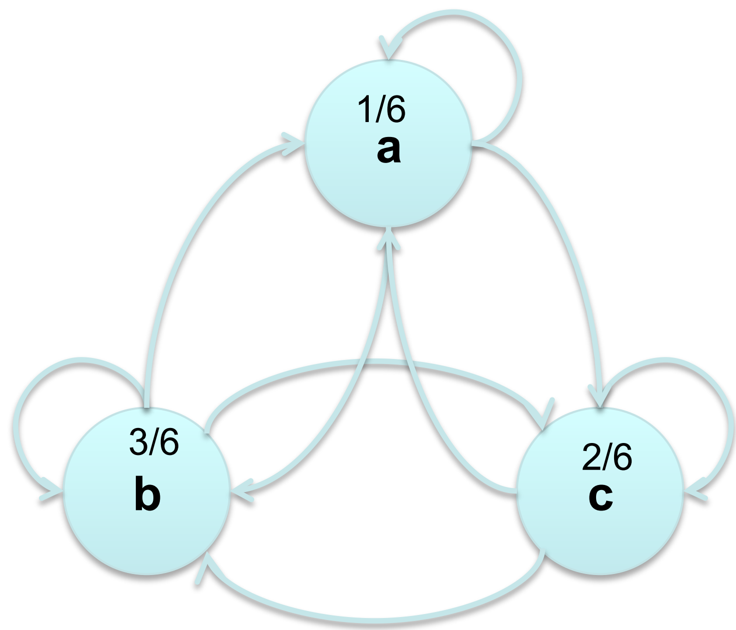
- Vi kan estimere sandsynlighederne i modellen fra data – dette kaldes indenfor machine learning for at lære en model.
- Vi kan simulere Markov modellen ved at generere nye tilfældige sekvenser af tilstande som efterligner data.
- Med den indlærte model kan vi også genkende mønstre i data ved at estimere den mest sandsynlige sekvens af tilstande som Markov modellen skal gennemgå for at generere data.
- (Hvordan genkendelsen foretages er avanceret stof og er udenfor dette kursus)

Simulering af en Markov kæde – et eksempel

- Udtræk en tilfældig start-tilstand (sampling) fra sandsynlighedsmassefunktionen $q(X_0)$

$q(X_0)$	X_0=a	X_0=b	X_0=c
	1/6	3/6	2/6

Resultat:

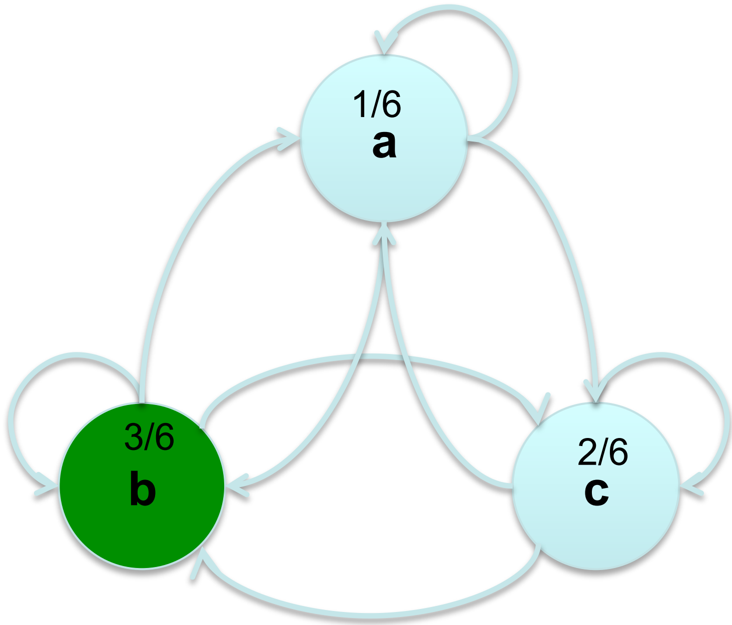


Simulering af en Markov kæde – et eksempel

- Udtræk en tilfældig start tilstand (sampling) fra sandsynlighedsmassefunktionen $q(X_0)$

$q(X_0)$	X_0=a	X_0=b	X_0=c
	1/6	3/6	2/6

Resultat:
b



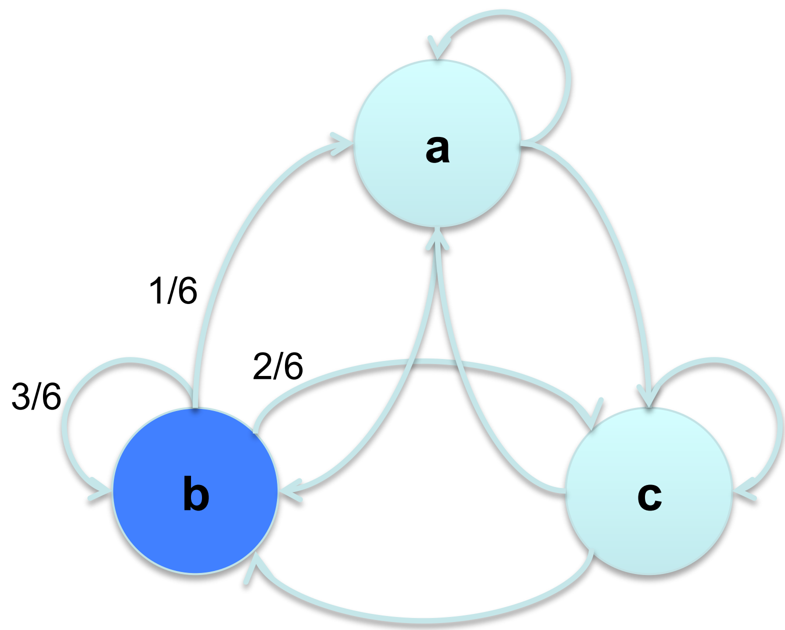
Simulering af en Markov kæde – et eksempel

- Sample en tilfældig næste tilstand fra overgangssandsynlighedsmassefunktionen $p(X_1 \mid X_0 = \mathbf{b})$

$p(X_t \mid X_{t-1})$	$X_t=a$	$X_t=b$	$X_t=c$
$X_{t-1}=a$	0	3/4	1/4
$X_{t-1}=b$	1/6	3/6	2/6
$X_{t-1}=c$	1/4	2/4	1/4

$p(X_1 \mid X_0)$	X_1=a	X_1=b	X_1=c
$X_0=b$	1/6	3/6	2/6

Resultat:
b

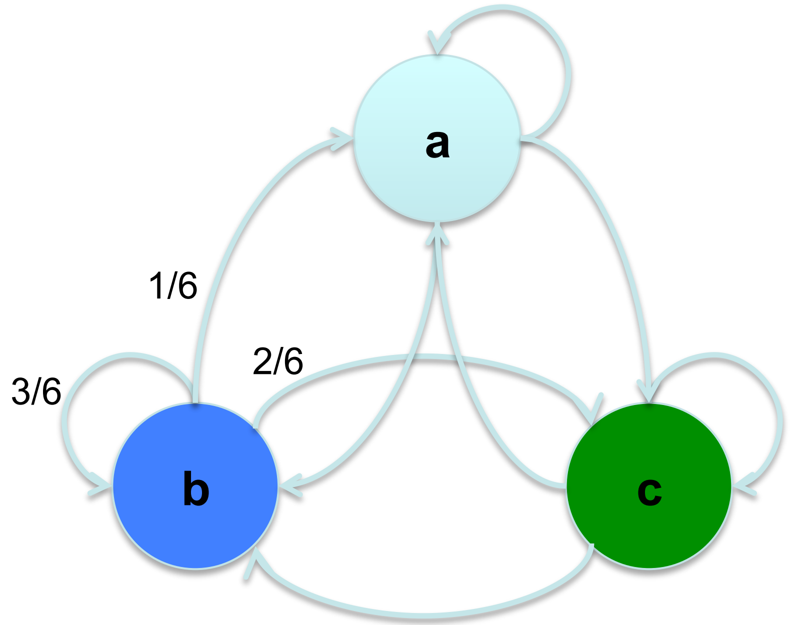


Simulering af en Markov kæde – et eksempel

- Sample en tilfældig næste tilstand fra overgangssandsynlighedsmassefunktionen $p(X_1 \mid X_0 = \mathbf{b})$

$p(X_t \mid X_{t-1})$	X_t=a	X_t=b	X_t=c
X_{t-1}=a	0	3/4	1/4
X_{t-1}=b	1/6	3/6	2/6
X_{t-1}=c	1/4	2/4	1/4

$p(X_1 \mid X_0)$	X_1=a	X_1=b	X_1=c
X_0=b	1/6	3/6	2/6



Resultat:

bc

Simulering af en Markov kæde – et eksempel

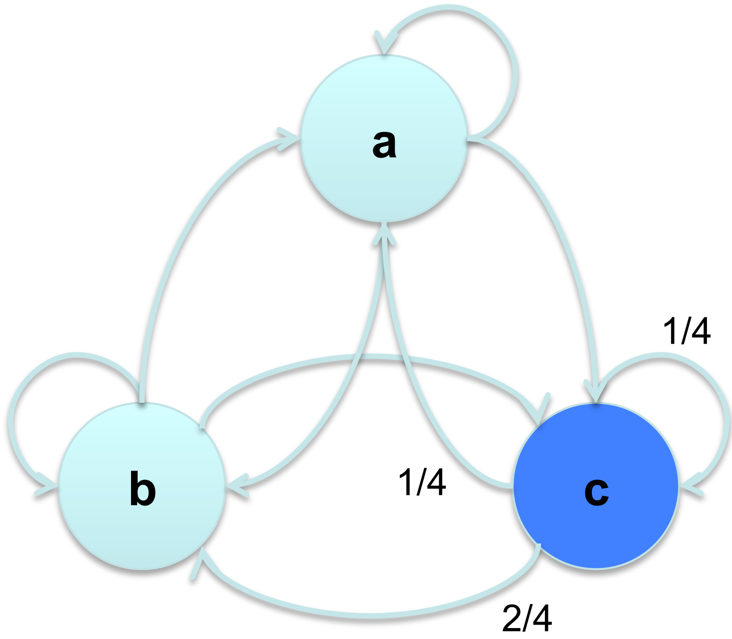
- Sample en tilfældig næste tilstand fra overgangssandsynlighedsmassefunktionen $p(X_2 \mid X_1 = \mathbf{c})$

$p(X_t \mid X_{t-1})$	X_t=a	X_t=b	X_t=c
X_{t-1}=a	0	3/4	1/4
X_{t-1}=b	1/6	3/6	2/6
X_{t-1}=c	1/4	2/4	1/4

$p(X_2 \mid X_1)$	X_2=a	X_2=b	X_2=c
X_1=c	1/4	2/4	1/4

Resultat:

bc



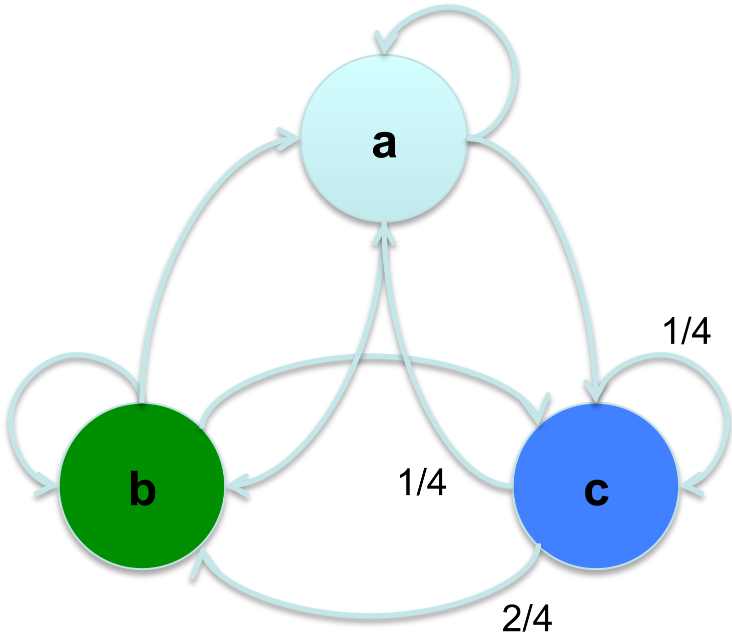
Simulering af en Markov kæde – et eksempel

- Sample en tilfældig næste tilstand fra overgangssandsynlighedsmassefunktionen $p(X_2 \mid X_1 = \mathbf{c})$

$p(X_t \mid X_{t-1})$	$X_t=a$	$X_t=b$	$X_t=c$
$X_{t-1}=a$	0	3/4	1/4
$X_{t-1}=b$	1/6	3/6	2/6
$X_{t-1}=c$	1/4	2/4	1/4

$p(X_2 \mid X_1)$	X_2=a	X_2=b	X_2=c
$X_1=c$	1/4	2/4	1/4

Resultat:
bcb



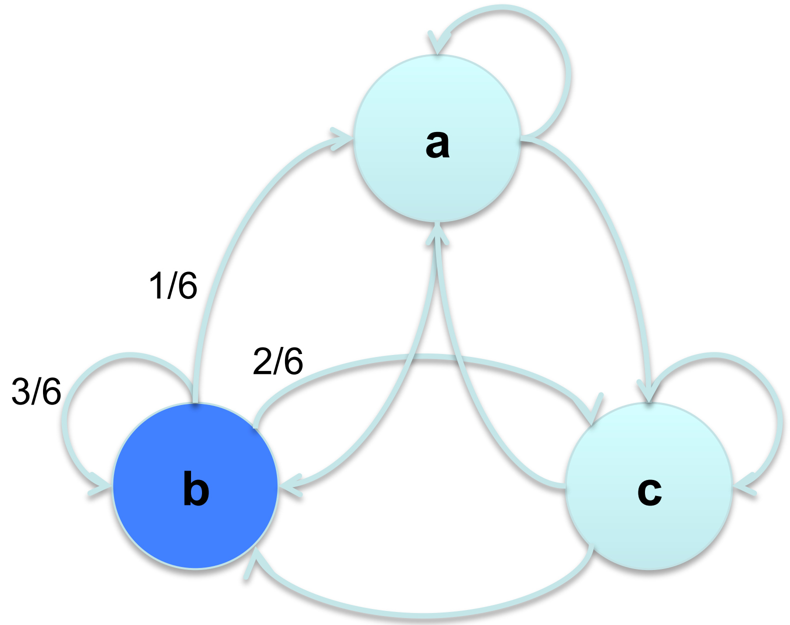
Simulering af en Markov kæde – et eksempel

- Sample en tilfældig næste tilstand fra overgangssandsynlighedsmassefunktionen $p(X_3 \mid X_2 = \mathbf{b})$

$p(X_t \mid X_{t-1})$	X_t=a	X_t=b	X_t=c
X_{t-1}=a	0	3/4	1/4
X_{t-1}=b	1/6	3/6	2/6
X_{t-1}=c	1/4	2/4	1/4

$p(X_3 \mid X_2)$	X_3=a	X_3=b	X_3=c
X_2=b	1/6	3/6	2/6

Resultat:
bcb



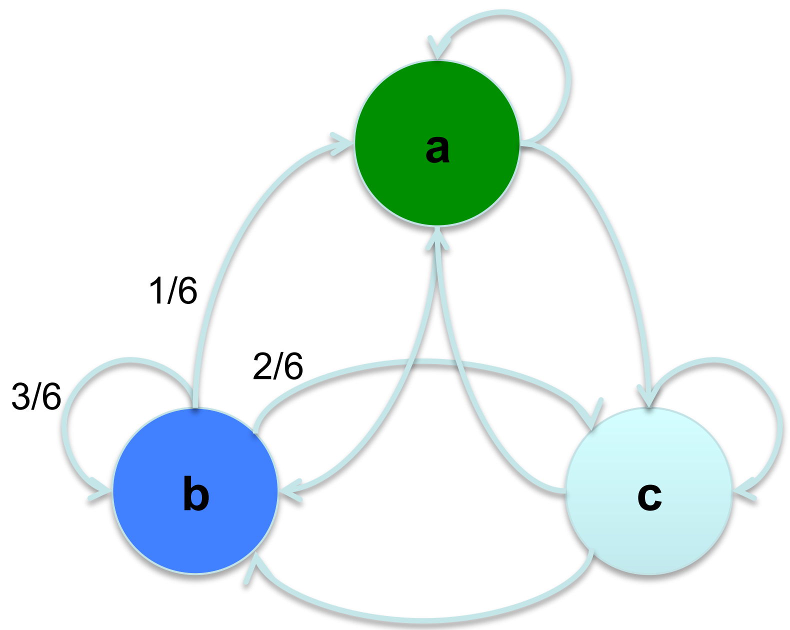
Simulering af en Markov kæde – et eksempel

- Sample en tilfældig næste tilstand fra overgangssandsynlighedsmassefunktionen $p(X_3 \mid X_2 = \mathbf{b})$

$p(X_t \mid X_{t-1})$	X_t=a	X_t=b	X_t=c
X_{t-1}=a	0	3/4	1/4
X_{t-1}=b	1/6	3/6	2/6
X_{t-1}=c	1/4	2/4	1/4

$p(X_3 \mid X_2)$	X_3=a	X_3=b	X_3=c
X_2=b	1/6	3/6	2/6

Resultat:
bcba

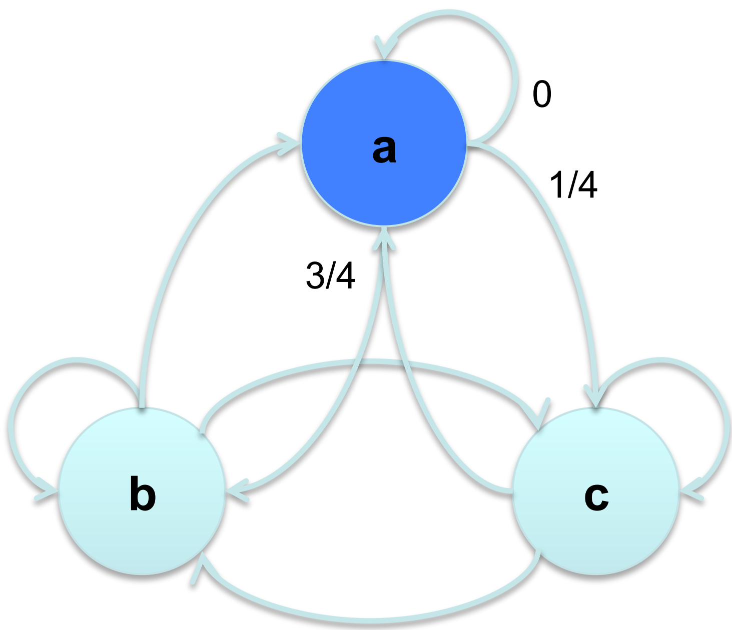


Simulering af en Markov kæde – et eksempel

- Sample en tilfældig næste tilstand fra overgangssandsynlighedsmassefunktionen $p(X_4 \mid X_3 = \mathbf{a})$

$p(X_t \mid X_{t-1})$	X_t=a	X_t=b	X_t=c
X_{t-1}=a	0	3/4	1/4
X_{t-1}=b	1/6	3/6	2/6
X_{t-1}=c	1/4	2/4	1/4

$p(X_4 \mid X_3)$	X_4=a	X_4=b	X_4=c
X_3=a	0	3/4	1/4



Resultat:
bcba

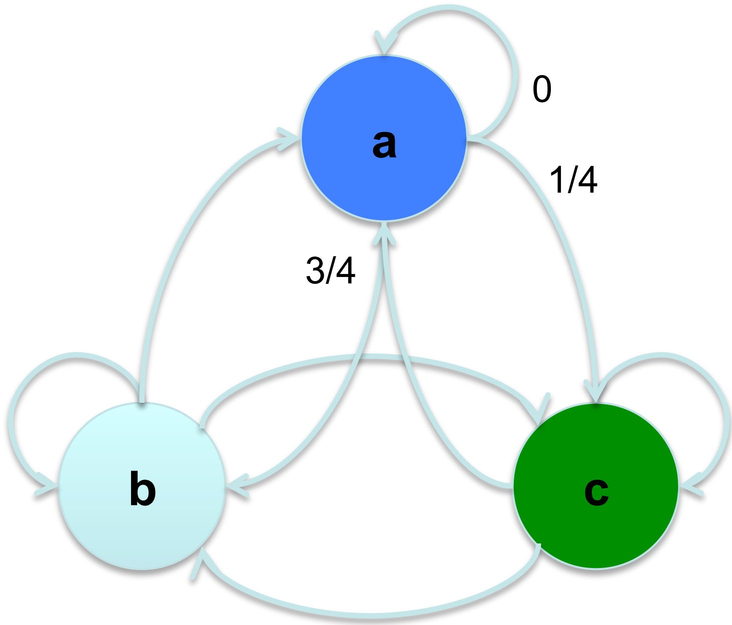
Simulering af en Markov kæde – et eksempel

- Sample en tilfældig næste tilstand fra overgangssandsynlighedsmassefunktionen $p(X_4 \mid X_3 = \mathbf{a})$

$p(X_t \mid X_{t-1})$	X_t=a	X_t=b	X_t=c
X_{t-1}=a	0	3/4	1/4
X_{t-1}=b	1/6	3/6	2/6
X_{t-1}=c	1/4	2/4	1/4

$p(X_4 \mid X_3)$	X_4=a	X_4=b	X_4=c
X_3=a	0	3/4	1/4

Resultat:
bcbac



Simulering af en Markov kæde model
Opsummering af simuleringsalgoritmen

- Markov modellen består af
  Tilstande: Ω
  sandsynlighedsmassefunktion for start tilstand: $q(X_0)$
  Overgangssandsynligheder: $p(X_t \mid X_{t-1})$

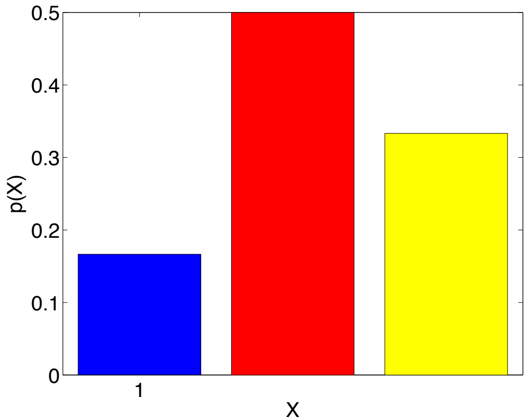
1. Sample en tilfældig start-tilstand fra sandsynlighedsmassefunktionen $q(X_0)$
2. For hvert skridt i kæden:
   – Sample en tilfældig næste tilstand fra overgangs-sandsynlighedsmassefunktionen $p(X_t \mid X_{t-1} = \omega_{t-1})$ baseret på forrige tilstand $X_{t-1} = \omega_{t-1}$

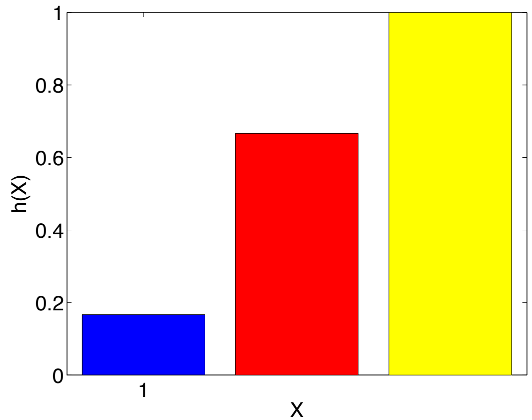
Sampling af en tilstand fra $q(X_0)$ og $p(X_t \mid X_{t-1})$

Transformationsmetoden

Sampling fra *p*(*X*) ved at

1. beregne den kumulative fordeling *h*(*X*) og læg ned på enheds-intervallet
2. træk et tilfældig tal fra $z \in [0,1]$
3. Slå op i *h* og se hvilket interval *z* falder i



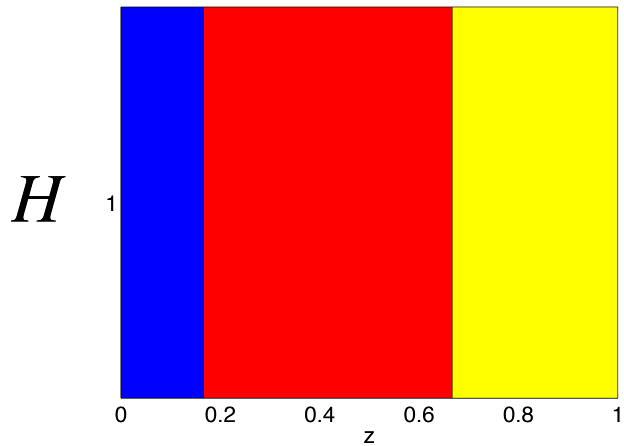


$z = 0.4 \Rightarrow$ Rød

$z = 0.8 \Rightarrow$ Gul

$z = 0.1 \Rightarrow$ Blå

$z = 0.6 \Rightarrow$ Rød



Sampling af en tilstand fra $q(X_0)$ og $p(X_t \mid X_{t-1})$

- Transformationsmetoden til sampling fra $p(X)$:

1. Beregn den cumulative fordeling
$$h(X=\omega^{(i)}) = \sum_{j=1}^{i} p(X=\omega^{(j)})$$
(anvend eks. `numpy.cumsum()`)
2. Sample et tilfældig tal $z \in [0,1]$
(anvend eks. `numpy.random.rand()`)
3. Find ved opslag i $H = [0, h(X=\omega^{(1)}), \ldots, h(X=\omega^{(K-1)})]$
det mindste indeks *i* så $z \geq H_i$
4. Vi har nu trukket tilstanden $\omega^{(i)}$ tilfældig fra $p(X)$

- Bemærk: Værdier med sandsynlighed 0 skal fjernes inden vi sampler

Estimering af en Markov model fra data

- Vi kan lære en Markov model fra data ved at estimere sandsynlighederne ud fra hyppighederne af tilstandene og af alle par af tilstande i data.

Data eksempel: **abbcacbbcb**

Optæl hyppigheder af tilstande og par af tilstande:

$g(X_0)$	X_0=a	X_0=b	X_0=c
	2	5	3

$$\sum_{\omega \in \Omega} g(X_t = \omega) = 10$$

$f(X_t \mid X_{t-1})$	X_t=a	X_t=b	X_t=c
X_{t-1}=a	0	1	1
X_{t-1}=b	0	2	2
X_{t-1}=c	1	2	0

$$\sum_{\omega \in \Omega} f(X_t = \omega \mid X_{t-1} = a) = 2$$

$$\sum_{\omega \in \Omega} f(X_t = \omega \mid X_{t-1} = b) = 4$$

$$\sum_{\omega \in \Omega} f(X_t = \omega \mid X_{t-1} = c) = 3$$

Estimering af en Markov model fra data

- Vi kan lære en Markov model fra data ved at estimere sandsynlighederne ud fra hyppighederne af tilstandene og af alle par af tilstande i data.

Data eksempel: **abbcacbbcb**

Sandsynligheder fra hyppigheder divideret med deres sum:

$q(X_0)$	X_0=a	X_0=b	X_0=c
	2/10	5/10	3/10

$$\sum_{\omega \in \Omega} g(X_t = \omega) = 10$$

$p(X_t \mid X_{t-1})$	X_t=a	X_t=b	X_t=c
X_{t-1}=a	0	1/2	1/2
X_{t-1}=b	0	2/4	2/4
X_{t-1}=c	1/3	2/3	0

$$\sum_{\omega \in \Omega} f(X_t = \omega \mid X_{t-1} = a) = 2$$

$$\sum_{\omega \in \Omega} f(X_t = \omega \mid X_{t-1} = b) = 4$$

$$\sum_{\omega \in \Omega} f(X_t = \omega \mid X_{t-1} = c) = 3$$

Eksempler på anvendelser af Markov modeller

- Bioinformatik:
  - Analyse af gensekvenser
  - Prædiktion af protein foldning (vigtigt for medicinal industrien)
- Processering af naturlige sprog:
  - Maskinoversættelse af naturlige sprog
  - Strukturel analyse af tekster
- Tracking:
  - Følg et mål over tid i radarbilleder (Eks. fiskeri overvågning, militære anvendelser)
  - 3D estimering af menneskelig bevægelse over tid
- Generaliseringer:
  - Skjulte Markov kæder
  - Gittermodeller også kendt som Markov random fields

Eksamen

- Eksamensopgaven bliver offentliggjort på mandag d. 12/1 kl. 10.00 og skal afleveres i Absalon den efterfølgende mandag d. 19/1 kl. 10.00.
- Opgaven er individuel, men i må godt tale sammen om forståelse af opgaven, men IKKE om løsning af opgaven.
- Kommunikation med undervisere og instruktorerne foregår på Absalon diskussionsforum - for at stille alle lige.
- Du kan forberede dig ved at sørge for at du har lavet alle øvelsesopgaver, udover de obligatoriske opgaver, og læse pensum (angivet på ugesedler) inklusiv slides.

Opsummering

- Regulære udtryk
- Stokastiske tilstandsmaskiner
  - Fra DEA til stokastiske tilstandsmaskiner
  - Markov kæder
  - Simulering af Markov kæder
  - Estimering af Markov kæder fra datasæt
  - Lidt sandsynlighedsregning

Læsestof

- Regulære udtryk: https://docs.python.org/2/howto/regex.html#regex-howto og dokumentationen for `re` modulet på https://docs.python.org/2/library/re.html (Begge findes også i PDF format på https://docs.python.org/2/download.html)

- Markov kæder: Læs ugesedlen.

Kursusevaluering

Husk at udfylde kursusevalueringen på KUnet (du har modtaget en e-mail eller besked på KUnet om dette)