

**Exam Probabilistic Programming 2020**

**April 16, 08:45–11:45**

**General Information:**

- Mark every sheet with your **student number**.
- Check that your copy of the exam consists of **five exercises**.
- This is an **open exam**, i.e. all lecture material (slides, exercises, solutions to the exercises) is permitted.
- Write with blue or black ink; do **not** use a pencil or red ink.
- You are neither allowed the help of anyone to complete your exam, nor is it allowed to help anyone else in completing this exam.
- Any attempt at deception leads to failure for this exam, even if detected later.
- Your exam is only valid if the integrity statement on the next page is signed by ticking the box.
- You are supposed to send your solutions to the exam via e-mail to [m.gerhold@utwente.nl](mailto:m.gerhold@utwente.nl).
- Your solutions need to be send in a **single** pdf-file or jpg-file and must be received ultimately on **Thursday April 16, noon (12:00, CEST)**.
- During the exam, Marcuc Gerhold is on-line available for questions via Canvas (conference).

Please read the following paragraph carefully, and tick the box to acknowledge that you have done so. To find more information, please consult the Canvas page of the course Probabilistic Programming 2019–2020.

*By testing you remotely in this fashion, we express our trust that you will adhere to the ethical standard of behaviour expected of you. This means that we trust you to answer the questions and perform the assignments in this test to the best of your own ability, without seeking or accepting the help of any source that is not explicitly allowed by the conditions of this test.*

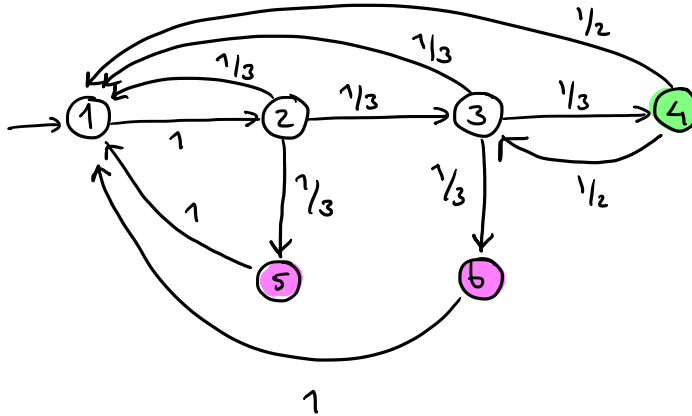
Please tick:

(If you are not able to tick this box, copy the above statement on your solution sheet, and tick the box drawn by you.)

Exercise 1 (Markov Chain Analysis)

12%

Consider the Markov chain depicted below:



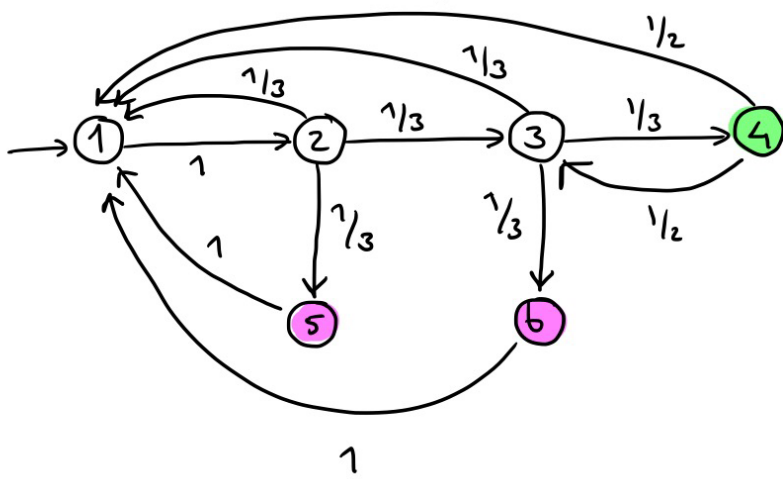
$$r(5) = r(6) = 10$$

$$r(1) = r(2) = r(3) =$$

$$r(4) = 0$$

- (a) [4%] Determine  $\Pr^M(\diamond\{4\})$ .
- (b) [4%] Determine  $ER^M(\diamond\{4\})$ .
- (c) [4%] Determine  $ER^M(\diamond\{4\} \mid \neg\diamond\{5,6\})$ .

Provide intermediate steps such that your computations are comprehensible.



$$r(5) = r(6) = 10$$

$$r(1) = r(2) = r(3) =$$

$$r(4) = 0$$

a  $R(\square 4)$

$x_s$  for every state

- $x_4 = 1$

all states can reach 4

there is no state  $s$   $x_s = 0$

$$S_? = \{1, 2, 3, 5, 6\} \quad |S_?| = 5$$

solve equation

$$(I - A) \cdot x = b$$

$$\begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & \\ & & & & & 1 \end{bmatrix}$$

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_5 \\ x_6 \end{pmatrix}$$

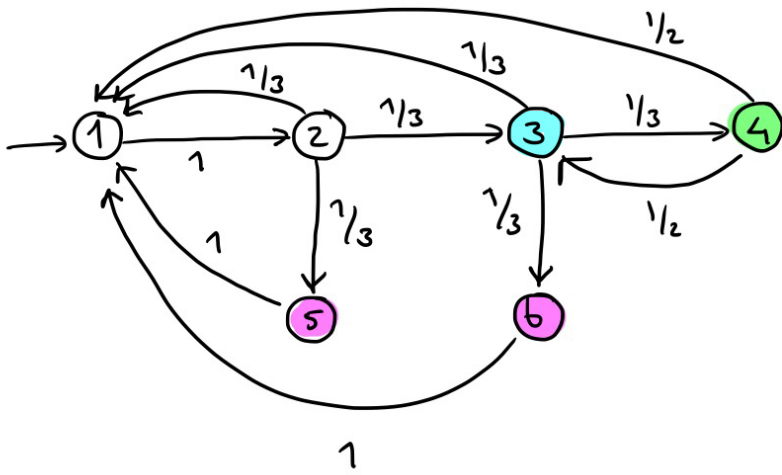
prob to reach 4 in one step

$$b = \begin{pmatrix} 0 \\ 0 \\ 1/3 \\ 0 \\ 0 \end{pmatrix}$$

	1	2	3	5	6
1	0	1	0	0	0
2	1/3	0	1/3	1/3	0
3	1/3	0	0	1/3	1/3
5	1	0	0	0	0
6	1	0	0	0	0

solve this leads to

$$x_1 = 1$$



$$\begin{aligned}
 r(5) &= r(6) = 10 \\
 r(1) &= r(2) = r(3) = \\
 r(4) &= 0
 \end{aligned}$$

$$r(4) = 1000$$

b. ER ( $\square 4$ ) associate variable  $y_i$  to states

$$y_4 = 0 \quad \leftarrow$$

$$y_3 = 0 + \underbrace{\frac{1}{3} y_4}_{=0} + \frac{1}{3} y_1 + \frac{1}{3} y_6$$

$$\left. \begin{aligned}
 y_6 &= 10 + y_1 \\
 y_5 &= 10 + y_1
 \end{aligned} \right\} y_5 = y_6$$

$$y_1 = 0 + y_2 \quad \longrightarrow \quad \boxed{y_1 = y_2}$$

$$y_2 = \underbrace{\frac{1}{3} y_1}_{=0} + \frac{1}{3} y_3 + \frac{1}{3} y_5$$

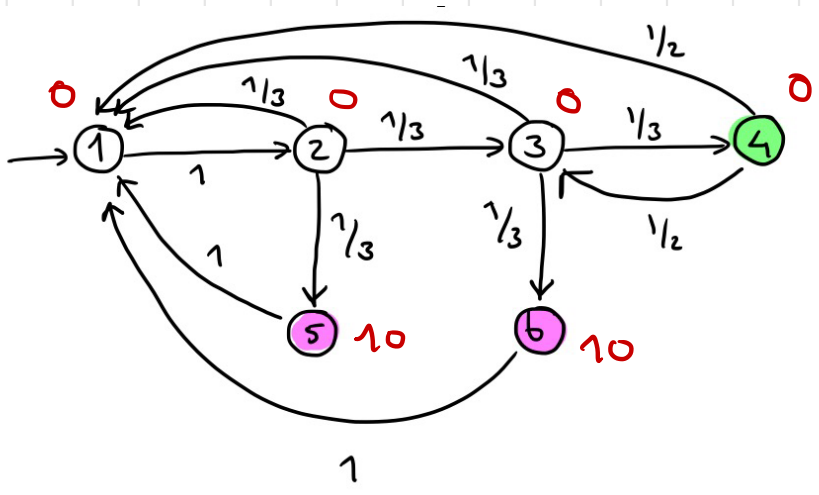
solve :  $\underline{y_1} = \boxed{40} = y_2$

$$y_3 = 30$$

$$y_4 = 0$$

$$y_5 = y_6 = 50$$

c.



$$r(5) = r(6) = 10$$

$$r(1) = r(2) = r(3) =$$

$$r(4) = 0$$

$ER(\square 4 \mid \neg \square \bullet)$  is undefined

$$\frac{ER(\square 4 \wedge \neg \square \bullet)}{Pr(\neg \square \bullet) = 0}$$

$$1 - Pr(\square \bullet) = 1$$

$Pr(\square \bullet)$

$x_5 = 1 \quad x_6 = 1 \quad x_1 = x_2$

$$x_2 = \frac{1}{3}x_5 + \frac{1}{3}x_3 + \frac{1}{3}x_1$$

$$\frac{1}{3} \qquad \frac{1}{3}x_2$$

solve:

$x_1 = x_2 = x_3 = x_4 =$   
 $x_5 = x_6 = 1$

$$x_3 = \frac{1}{3}x_4 + \frac{1}{3}x_1 + \frac{1}{3}x_6$$

$$\frac{1}{3}x_2 \qquad \frac{1}{3}$$

$$x_4 = \frac{1}{2}x_3 + \frac{1}{2}x_1$$

## Exercise 2 (Weakest Pre-expectations)

20%

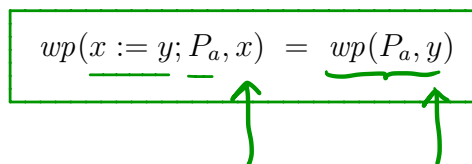
Consider the following pGCL program  $P_a$  where  $a$  is a rational number in the interval  $[0, 1]$ :

$P_a ::=$

$$\begin{aligned} & \{\ell := 1\} [1/4] \{\ell := 0\}; \\ & \{r := 1 - \ell\} [1/3] \{r := 0\}; \\ & \{t := 1 - (\ell + r)\} [a] \{t := 0\}; \\ & d := 1 - (r + \ell + t); \\ & x := x + r - \ell; \\ & y := y + t - d \end{aligned}$$

$0 \leq a \leq 1$   
 $a \in \mathbb{Q}$

Determine for which values of  $a$  the following statement holds:

$$\boxed{wp(x := y; P_a, x) = wp(P_a, y)}$$


$\{l := 1\} [1/4] \{l := 0\};$   
 $\{r := 1 - l\} [1/3] \{r := 0\};$   
 $\{t := 1 - (l+r)\} [a] \{t := 0\};$   
 $d := 1 - (r+l+t);$   
 $x := x + r - l;$   
 $y := y + t - d$

$P_a$

$$Q: \text{wp}(x=y; P_a, x) = \text{wp}(P_a, y)$$

(\*)  $\left\{ \begin{aligned} &\frac{1}{4} \left( \frac{1}{3}(x+1) + \frac{2}{3}(x-1) \right) \\ &+ \frac{3}{4} \left( \frac{1}{3}(x+1) + \frac{2}{3}x \right) \end{aligned} \right.$

$\text{wp}(P_a, x)$

$l := 1 \quad [1/4] \quad l := 0$

$$\frac{1}{3}(x + (1-l) + l) + \frac{2}{3}(x-l) = \frac{1}{3}(x+1) + \frac{2}{3}(x-l)$$

$r := (1-l) \quad [1/3] \quad r := 0$

$x+r-l$

$t := 1 - (l+r) \quad [a] \quad t := 0$

$x+r-l \leftarrow$

$d := 1 - (l+r+t)$

$x+r-l \leftarrow$

$x := x + r - l$

x

$y := y + t - d \leftarrow$

x

postexpectation

no a

simplify (\*)

$$\frac{1}{4} \left( x - \frac{1}{3} \right) + \frac{3}{4} (x+1)$$

$= \text{wp}(P_a, x)$

$\text{wp}(x=y; P_a, x)$

$$= \frac{1}{4} \left( y - \frac{1}{3} \right) + \frac{3}{4} (y+1)$$



in a similar way, compute  $w_p(P_a, y)$

$e_a =$  expression in terms of  $a$

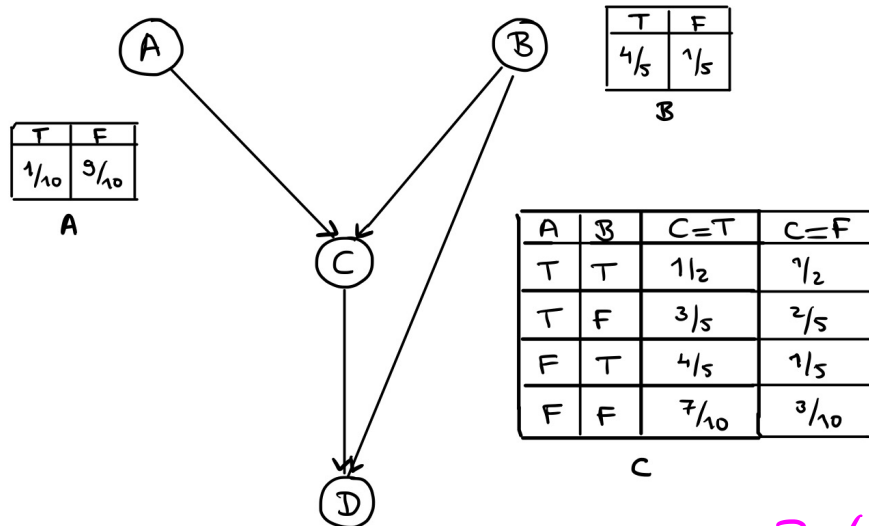
$$e_a = \frac{1}{4} \left( y - \frac{1}{3} \right) + \frac{3}{4} (y+1)$$



### Exercise 3 (Bayesian Networks)

18%

Consider the Bayesian network depicted below:



T	F
1/10	9/10

A

T	F
4/5	1/5

B

A	B	C=T	C=F
T	T	1/2	1/2
T	F	3/5	2/5
F	T	4/5	1/5
F	F	7/10	3/10

C

B	C	D=T	D=F
T	T	1/5	4/5
T	F	4/5	1/5
F	T	9/10	1/10
F	F	1/10	9/10

$Pr(A, B, C, D)$

$$= Pr(D|B, C) \cdot Pr(C|A, B) \cdot Pr(A) \cdot Pr(B)$$

- [4%] Given an expression for the joint probability distribution of the Bayesian network.
- [4%] Determine  $Pr(B = T | D = F \wedge A = T)$
- [4%] Determine the pGCL program of the Bayesian network for the evidence  $D = F \wedge A = T$
- [6%] Determine the expected run-time of your pGCL program using the following equation:

$$ert(\text{repeat } Seq \text{ until } G, f) = \frac{1 + ert(Seq, [G].f)}{wp(Seq, [G])}$$

where  $Seq$  stands for the sequential composition of the programs for the Bayesian network vertices.

$$b. \Pr(B=T | D=F, A=T)$$

$$= \frac{\Pr(B=T, D=F, A=T)}{\Pr(D=F, A=T)}$$

$$\Pr(D=F, A=T)$$

$$\Pr(B=T, D=F, A=T) = \sum_{\substack{C \in \mathcal{C} \\ C=T \vee \\ C=F}} \Pr(A=T, B=T, C=c, D=F)$$

$$= \Pr(A=T, B=T, C=T, D=F) +$$

$$\Pr(A=T, B=\bar{T}, C=F, D=F) = \frac{3}{125}$$

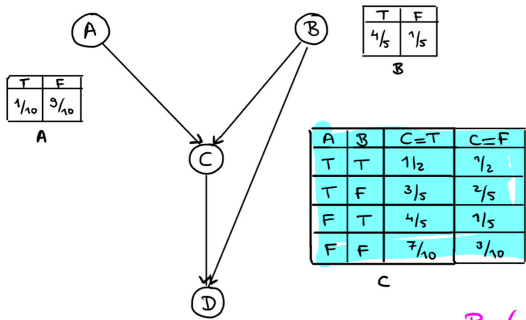
$$\Pr(D=F, A=T) = \sum_{b \in \mathcal{B}} \Pr(B=b, D=F, A=T)$$

$$= \Pr(A=T, B=T, D=F) +$$

$$\Pr(A=T, B=F, D=F)$$

$$\sum_{C \in \mathcal{C}} \Pr(A=T, B=F, C=c, D=F)$$

C.



$$D=F, A=T$$

PGCL program

topological sort

$$\begin{cases} A; B; C; D \quad (*) \\ B; A; C; D \end{cases}$$

B	C	D=T	D=F
T	T	1/5	4/5
T	F	4/5	1/5
F	T	3/10	7/10
F	F	1/10	9/10

$$\begin{aligned} &Pr(A, B, C, D) \\ &= Pr(D|B, C) \cdot \\ &Pr(C|A, B) \cdot \\ &Pr(A) \cdot Pr(B) \end{aligned}$$

PGCL program

repeat { prog A ; prog B ; prog C ; prog D }  
 until  $(D=F \wedge A=T)$   
 evidence

prog C:

if  $(A=T \wedge B=T)$   
 C := T [1/2] C := F  
 else if  $(A=T \wedge B=F)$   
 C := T [3/5] C := F  
 else if  $(A=F \wedge B=T)$   
 C := T [4/5] C := F  
 else C := T [7/10] C := F

d.

$$\text{ert}(\text{repeat Seq until } G, f) = \frac{1 + \text{ert}(\text{Seq}, [G].f)}{\text{wp}(\text{Seq}, [G])} = \frac{121}{12500} \quad f=0$$

$$\text{ert}(\text{prog A}; \text{prog B}; \text{prog C}; \text{prog D}, \underbrace{[G].f}_{=0})$$

$D=F \wedge A=T$

①  $\text{ert}(\text{prog D}, 0) = 5$

prog D has the same shape as prog C

}	if $(A=T \wedge B=T)$	}	1 +
	$C:=T$ [1/2] $C:=F$		$1 + \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1$
	else if $(A=T \wedge B=F)$	}	= 1
	$C:=T$ [3/5] $C:=F$		= 3
	else if $(A=F \wedge B=T)$		}
$C:=T$ [4/5] $C:=F$			
	else $C:=T$ [7/10] $C:=F$		

$$\text{ert}(\text{prog C}, ) = 5$$

$$\text{ert}(\text{prog B}, ) = 2 +$$

$$\text{ert}(\text{prog A}, ) = 2 +$$

$$\text{ert}(\text{seq}, 0) = 14$$

$$\text{ert}(\text{repeat} \dots) = \frac{1 + 14}{121/12500} \approx 310$$

Exercise 4 (Loops, Invariants, and Termination)

25%

Consider the following pGCL program  $P$ :

$P$ :

```

x := 1;
while (x = 1) {
  {x := 0}[1/3]{y := y + 1}
}
    
```

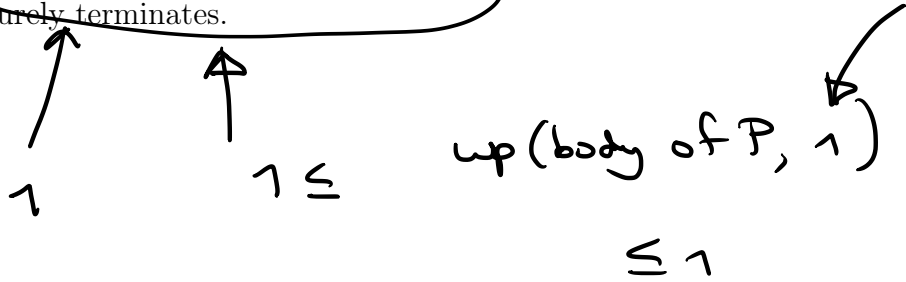
Let expectation  $I_n$  for natural number  $n$  be defined by:

$$I_n := [x \neq 1] + [x = 1] \cdot \sum_{k=0}^{n-1} 1/3 \cdot (2/3)^k,$$

where  $\sum_{k=0}^{-1} \dots = 0$ .

Let post-expectation  $f = 1$ .

- (a) [5%] Give the characteristic function  $\Phi_f$  of the loop in  $P$
- (b) [5%] Show that  $I_0 = \Phi_f(0)$
- (c) [8%] Show that for all  $n \geq 0$ , it holds:  $I_{n+1} = \Phi_f(I_n)$
- (d) [4%] Determine  $\lim_{n \rightarrow \infty} I_n$
- (e) [3%] Given that  $\lim_{n \rightarrow \infty} I_n \leq wp(\text{body of } P, f)$ , determine whether program  $P$  almost surely terminates.



$\forall$  pGCL program  $P$                        $0 \leq wp(P, 1) \leq 1$

termination prob.  $\uparrow$

• if  $wp(P, 1) = 1$                        $P$  is AST

• if  $wp(P, 1) < 1$                        $P$  is non-AST

the following pGCL program  $P$ :

$P$ :

```

x := 1;
while (x = 1) {
  {x := 0} [1/3] {y := y + 1}
}
    
```

a.  $f = 1$   
 $\Phi_f$

(\*)  $\Phi_f(x) = [x \neq 1] \cdot 1 + [x = 1] \left( \frac{1}{3} \times (x \leftarrow 0) + \frac{2}{3} \times (y \leftarrow y + 1) \right)$

b. show that  $\underline{I}_0 = \Phi_f(0)$

$I_n = [x \neq 1] + [x = 1] \cdot \sum_{k=0}^{n-1} \frac{1}{3} \cdot \left(\frac{2}{3}\right)^k$

$\Phi_f(0) = [x \neq 1] + [x = 1] \left( \underbrace{\frac{1}{3} \cdot 0}_{=0} (x \leftarrow 0) + \frac{2}{3} \cdot 0 (y \leftarrow y + 1) \right)$   
 $= [x \neq 1] = \underline{I}_0 = [x \neq 1] + [x = 1] \sum_{k=0}^{-1} \dots$

c. show  $\forall n \geq 0 : I_{n+1} = \Phi_f(I_n)$

$\Phi_f(I_n) = [x \neq 1] + [x = 1] \left( \frac{1}{3} I_n (x \leftarrow 0) + \frac{2}{3} I_n (y \leftarrow y + 1) \right)$   
 $= [x \neq 1] + [x = 1] \left( \frac{1}{3} + \frac{2}{3} I_n \right)$

$= [x \neq 1] + [x = 1] \left( \frac{1}{3} + \frac{2}{3} I_n \right)$

$= [x \neq 1] + [x = 1] \left( \frac{1}{3} + \frac{2}{3} \sum_{k=0}^{n-1} \frac{1}{3} \cdot \left(\frac{2}{3}\right)^k \right)$

$= [x \neq 1] + [x = 1] \left( \frac{1}{3} + \sum_{k=0}^{n-1} \frac{1}{3} \cdot \left(\frac{2}{3}\right)^{k+1} \right)$

$\frac{1}{3} \cdot \left(\frac{2}{3}\right)^0 + \sum_{k=0}^{(n+1)-1} \frac{1}{3} \cdot \left(\frac{2}{3}\right)^k$

$= I_{n+1}$



d. Determine  $\lim_{n \rightarrow \infty} I_n$

$$I_n = [x \neq 1] + [x = 1] \cdot \sum_{k=0}^{n-1} \frac{1}{3} \cdot \left(\frac{2}{3}\right)^k$$

$$\lim_{n \rightarrow \infty} I_n = [x \neq 1] + [x = 1] \underbrace{\sum_{k=0}^{\infty} \frac{1}{3} \left(\frac{2}{3}\right)^k}_{\text{geometric series}}$$

$$= [x \neq 1] + [x = 1] \underbrace{\frac{1/3}{1 - 2/3}}_{= 1}$$

$$= \underbrace{[x \neq 1]} + \underbrace{[x = 1]}$$

$$= 1$$

e.  $\lim_{n \rightarrow \infty} I_n \leq \underbrace{\text{wp}(\text{loop body of } P, f)}_{f=1} \quad (*)$

$$< 1$$

$$= \underbrace{\text{wp}(x := 1; \text{loop body}, 1)}$$

$$= 1 \quad \rightarrow \quad P \cup \text{AST}$$

**Exercise 5 (Probabilistic Databases)**

**25%**

The example data of Figure 2 was produced by a handwriting recognition software that analysed the example sentence of Figure 1. The table **words** has two attributes: **pos** is the position of the word and **word** is the word recognized in that position. As you can see, for some positions there are several possible words recognized. The second table **meaning** holds data from a dictionary with meanings of words. It has three attributes: **id** is a unique identifier for words, **word** is the word, and **meaning** is a description of a possible meaning of the word. Also here you see that there are several possible meanings for each word.

For simplicity, we only consider a part of the data produced, namely the first 5 words of the example sentence and the possible meanings of the second and third word. Note that our dictionary for meaning has no entry for the word “shute”.

- (a) [2%] How many possible worlds does the probabilistic database of Figure 2

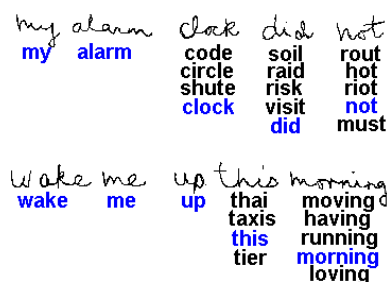


Figure 1: Example sentence for handwriting recognition (source <https://cedar.buffalo.edu/handwriting/HROverview.html>).

words		meaning		description	
$\langle \text{pos}, \text{word} \rangle$	$\varphi$	$\langle \text{id}, \text{word}, \text{meaning} \rangle$	$\varphi$	$\varphi$	
$\langle 1, \text{my} \rangle$	$\top$	$\langle 1, \text{alarm}, \text{warning of danger} \rangle$	$m_1 = 1$	$w_k = i$	$k$ th word is $i$ th alternative
$\langle 2, \text{alarm} \rangle$	$\top$	$\langle 1, \text{alarm}, \text{warning sound or device} \rangle$	$m_1 = 2$	$m_l = j$	word with id $l$ has meaning $j$
$\langle 3, \text{code} \rangle$	$w_3 = 1$	$\langle 2, \text{code}, \text{program instructions} \rangle$	$m_2 = 1$	All probabilities uniform	
$\langle 3, \text{circle} \rangle$	$w_3 = 2$	$\langle 2, \text{code}, \text{collection of laws} \rangle$	$m_2 = 2$		
$\langle 3, \text{shute} \rangle$	$w_3 = 3$	$\langle 2, \text{code}, \text{cryptographic system} \rangle$	$m_2 = 3$		
$\langle 3, \text{clock} \rangle$	$w_3 = 4$	$\langle 3, \text{circle}, \text{round figure} \rangle$	$m_3 = 1$		
$\langle 4, \text{soil} \rangle$	$w_4 = 1$	$\langle 3, \text{circle}, \text{group of people} \rangle$	$m_3 = 2$		
$\langle 4, \text{raid} \rangle$	$w_4 = 2$	$\langle 4, \text{clock}, \text{time measuring device} \rangle$	$m_4 = 1$		
$\langle 4, \text{risk} \rangle$	$w_4 = 3$	$\langle 4, \text{clock}, \text{spherical seed head} \rangle$	$m_4 = 2$		
$\langle 4, \text{visit} \rangle$	$w_4 = 4$	$\langle 4, \text{clock}, \text{person's face} \rangle$	$m_4 = 3$		
$\langle 4, \text{did} \rangle$	$w_4 = 5$				
$\langle 5, \text{rout} \rangle$	$w_5 = 1$				
$\langle 5, \text{hot} \rangle$	$w_5 = 2$				
$\langle 5, \text{riot} \rangle$	$w_5 = 3$				
$\langle 5, \text{not} \rangle$	$w_5 = 4$				
$\langle 5, \text{must} \rangle$	$w_5 = 5$				

Figure 2: Example probabilistic data corresponding with Figure 1.

contain? Explain your answer by giving the complete calculation of the answer.

**Solution:**

There are 7 random variables with varying numbers of alternatives, so by independence of random variables  $4 \times 5 \times 5 \times 2 \times 3 \times 2 \times 3 = 3600$  possible worlds.

**NB:** Even though there is no meaning for the word “shute” in the meaning table, the record does exist in the words table, so should simply be counted.

- (b) [2%] We define “correct contents” for the database as containing for each position the correctly recognized word in table **words**, and for each such correctly recognized word its correct meaning in table **meaning**. Give a sentence that selects the correct contents for the database, i.e., where the recognized words are “my alarm clock did not” with meanings “warning sound or device” for “alarm” and “time measuring device” for “clock”.

**Solution:**  $w_3 = 4 \wedge w_4 = 5 \wedge w_5 = 4 \wedge m_1 = 2 \wedge m_4 = 1$

- (c) [2%] How many of the possible worlds in the database have correct contents, i.e., have the abovedescribed recognized words and meanings? Explain your answer by giving the complete calculation of the answer.

**Solution:** The random variables  $w_3$ ,  $w_4$ ,  $w_5$ ,  $m_1$ , and  $m_4$  need to have their fixed value to arrive at the correct contents of the world; the other random variables ( $m_2$  and  $m_3$ ) can have any value, so  $3 \times 2 = 6$  possible worlds have correct contents.

- (d) [3%] Calculate the probability of  $(w_3 = 4 \wedge m_4 = 1) \vee (w_3 = 2 \wedge m_3 = 2)$ . Explain your answer by giving the complete calculation of the answer.

**Solution:**

$$\begin{aligned} &P((w_3 = 4 \wedge m_4 = 1) \vee (w_3 = 2 \wedge m_3 = 2)) = \\ &\quad \text{Left and right side of } \vee \text{ are mutually exclusive, because} \\ &\quad \text{\(} w_3 = 4 \text{ and } w_3 = 2 \text{ are different alternatives of the same variable.} \\ &= P(w_3 = 4 \wedge m_4 = 1) + P(w_3 = 2 \wedge m_3 = 2) = \\ &= P(w_3 = 4) \times P(m_4 = 1) + P(w_3 = 2) \times P(m_3 = 2) = \\ &= \frac{1}{4} \times \frac{1}{3} + \frac{1}{4} \times \frac{1}{2} = \frac{1}{12} + \frac{1}{8} = \frac{5}{24} \\ &\approx 0.2083 \end{aligned}$$

You need to explicitly observe the mutual exclusiveness either in words or in the calculation.

- (e) [3%] Suppose we have a table `meaning-raw` that contains the same data as table `meaning` in Figure 2 but without the sentences  $\varphi$ . The purpose of this table is to provide the raw data as a *certain table* from which we derive the *uncertain table* `meaning`.

Given the following attempt at creating the *uncertain table* `meaning`

```
DROP TABLE IF EXISTS meaning;
CREATE TABLE meaning AS
REPAIR KEY word IN meaning-raw
```

See <http://maybms.sourceforge.net/manual/index.html#x1-460006.2.1> for an explanation of `REPAIR KEY` from the official MayBMS manual.

For each statement, indicate whether or not it is true *and provide an explanation why*.

- The `REPAIR KEY` is a pure query, i.e., it has no side effects on the data in the database.**
- The attempt indeed creates 4 random variables  $m_1..m_4$ .**
- There is no mention of `WEIGHT BY` in the `REPAIR KEY` statement, so no probabilities are created.
- We could also have used `id` instead of `word` to arrive at the same result.**

**Solution:**

- REPAIR KEY is not updating the mentioned table (here `meaning_raw`), but is just returning a table just like a normal `SELECT-FROM-WHERE` query would. Therefore, we need the `CREATE-TABLE-AS` to store the result, otherwise we would just see the result, but it would be gone again afterwards, just like a normal `SELECT-FROM-WHERE` result.
- Since there are 4 different words in the table and each has multiple occurrences, for each word a new random variable is created to “turn word into a unique key”. So, indeed four random variables are created.
- The behavior of REPAIR KEY when no `WEIGHT BY` is given, is not that probabilities are omitted, but that the probabilities will be computed as uniformly distributed over the alternatives. For example,  $P(m_2 = j) = \frac{1}{3}$  for each  $j$ .
- Since the values of `id` and `word` coincide, i.e., there are also 4 different `id` values and each is in the same records as the associated `word`, indeed the same number of random variables with the same number of alternatives are created and the same sentences are associated with each record.

- (f) [8%] Given the probabilistic algebra expression  $E$  below (‘ $w.a$ ’ refers to attribute  $a$  of table ‘words’; ‘ $m.a$ ’ refers to attribute  $a$  of table ‘meaning’).

$$\pi_{w.pos,w.word,m.meaning}(\bowtie_{w.word=m.word}(\text{meaning}, \sigma_{w.pos=3}(\text{words})))$$

Since there are 4 words possible for position 3 and in total 8 possible meanings for these words, the query can have many results. Give exactly *one result* of  $E$ .

**NB:** I do not ask for a derivation, nor do I ask for all results, but I do ask for one of those many results *in the right form*, so take care to provide all components that a result of a probabilistic algebra expression should have and omit those components that such a result should not have.

**Solution:** The result of a probabilistic algebra expression, is a probabilistic table itself, i.e., a set of pairs containing a record and a sentence. One result, hence, is a pair of a record with a sentence. One such results of this query is  $(\langle 3, \text{clock, time measuring device} \rangle, w_3 = 4 \wedge m_4 = 1)$ .

- (g) [5%] In indeterministic duplicate detection, an  $M$ -graph is constructed from similarity match results of a duplicate detection tool which ran on tuples  $a, b, c, d$ . The tool determines the following similarities:  $s(a - b) = 0.3$ ,  $s(a - c) = 0.95$ ,  $s(a - d) = 0.2$ ,  $s(b - c) = 0.5$ ,  $s(b - d) = 0.8$ ,  $s(c - d) = 0.2$ . We set the upper threshold to 0.9 and the lower threshold to 0.4.

1. Draw the  $M$ -graph *after the thresholds have been applied*.
2. Which possible worlds does this produce? Use the following notation:  $\{\dots\}$  for the set of records comprising a possible world;  $ab$  for the merge of records  $a$  and  $b$  (other combinations analogously). Explain your answer.

**NB:** I do not ask for probabilities of possible worlds, so no need to compute them.

**Solution:** Enforcing the thresholds has as a consequence that we consider  $a - c$  as a certain edge,  $b - c$  and  $b - d$  as uncertain edges, and between all other combinations of nodes, there is no edge.

With two uncertain edges, there are  $2^2 = 4$  possible  $W$ -graphs. A  $W$ -graph can be inconsistent because equality is transitive. Two of the four  $W$ -graphs are inconsistent, namely the ones that have the  $b - c$  edge. So there are two consistent possible worlds, namely

- $\{ac, b, d\}$
- $\{ac, bd\}$