Naam: .....

Collegekaart-nummer: ..... Versie opgavenblad (omcirkelen): A | B | C | D

Dit tentamen duurt 3 uur. Er zijn 16 vragen, waarvan 6 open en 10 meerkeuze. Het is verboden literatuur, aantekeningen, een programmeerbare rekenmachine, of een telefoon te gebruiken. Het gebruik van een standalone niet-programmeerbare rekenmachine is toegestaan. Veel succes!

Open vragen

Antwoorden op alle vragen worden ingevuld op deze bladen. Berekeningen worden gedaan op gelijnd papier. Berekeningen worden ook ingeleverd. Antwoorden op open vragen zonder berekening of motivatie leveren geen punten op. Twee punten voor elk onderdeel, tenzij anders aangegeven.

1. (Speltheorie.) Gegeven is het volgende twee-persoons niet-nulsom spel, G, op basis van volledige informatie met simultane zetten en kwantitatieve beloningen.

$$G = \begin{array}{c} \\ \text{T} \\ \text{B} \end{array} \begin{array}{c} \begin{array}{cc} \text{L} & \text{R} \end{array} \\ \begin{pmatrix} 0,1 & 0,1 \\ -1,-1 & 0,-1 \end{pmatrix} \end{array}$$

   (a) Bepaal alle Pareto-optimale strategie-profielen.

   Antw. $\{(T,L),(T,R)\}$.

   (b) Bepaal alle pure Nash evenwichten.

   Antw. $\{(T,L),(T,R),(B,R)\}$.

   (c) Bepaal alle (pure en gemixte) Nash evenwichten.

   Antw. $\{(p,q) \mid p = 0 \text{ of } q = 1\}$, waarbij de rij-speler met kans p actie B, en de kolom-speler met kans q actie R speelt.

2. (Speltheorie.) Beschouw het volgende (symmetrische) spel.

		Speler B		
		b_1	b_2	b_3
Speler A	a_1	$4(4)$	$0(5)$	$0(6)$
	a_2	$5(0)$	$3(3)$	$0(1)$
	a_3	$6(0)$	$1(0)$	$2(2)$

   (a) Bepaal alle Pareto-optimale strategie-profielen.

Antw. Pareto-optimale strategie-profielen zijn

$$\{\ (a_1, b_1),\ (a_3, b_1),\ (a_1, b_3)\ \}$$

(Let op dat je strategie-profielen vermeld en geen uitbetalings-profielen.) De resterende strategie-profielen worden gedomineerd door deze drie Pareto-optimale profielen.

(b) Bepaal alle pure Nash evenwichten.

Antw. Pure Nash evenwichten zijn

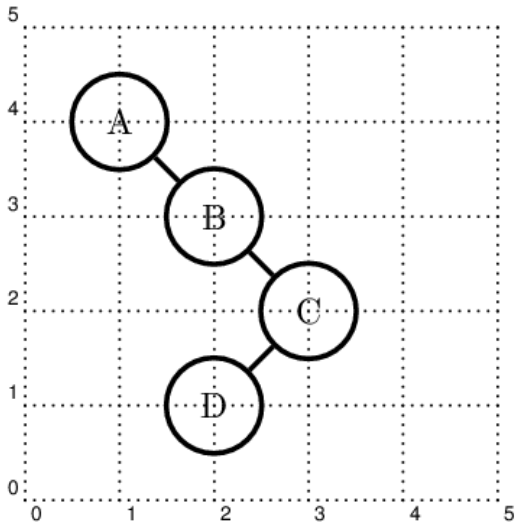
$$\{\ (a_2, b_2),\ (a_3, b_3)\ \}$$

(Let weer op dat je strategie-profielen vermeld en geen uitbetalings-profielen.)

(c) Leg uit waarom het voor Speler A nooit interessant is a_1 te spelen. (En waarom het voor Speler B nooit interessant is b_1 te spelen.)

Antw. Alle uitbetalingen van a_1 worden, onafhankelijk van wat B doet, gedomineerd door uitbetalingen van a_2 (of uitbetalingen van a_3). Analoog worden alle uitbetalingen van b_1, onafhankelijk van wat A doet, gedomineerd door uitbetalingen van b_2 (of uitbetalingen van b_3).

3. Gegeven is een 1-dimensionaal Kohonen netwerk



met buurfunctie

$$g(x,y) = \begin{cases} 1 & \text{als } |x-y| = 0, \\ 0.5 & \text{als } |x-y| = 1, \\ 0 & \text{anders.} \end{cases}$$

Bepaal de ligging van dit netwerk na updates met de volgende twee voorbeelden, waarbij de leerfactor, α, gelijk is aan 0.5. Voorbeeld 1: $(3, 4)$. Voorbeeld 2: $(4, 2)$. De voorbeelden worden aangeboden in deze volgorde.

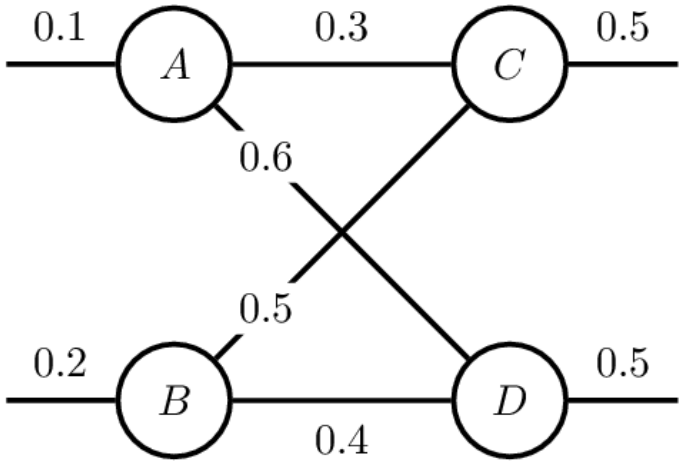
Schrijf de tussenresultaten van je berekening op in de volgende tabel.

	A	B	C	D
0.	$(1, 4)$	$(2, 3)$	$(3, 2)$	$(2, 1)$
Na Voorbeeld 1.				
Na Voorbeeld 2.				

Antw.

	A	B	C	D
0.	$(1, 4)$	$(2, 3)$	$(3, 2)$	$(2, 1)$
Na Voorbeeld 1.	$(1.5, 4)$	$(2.5, 3.5)$	$(3, 2.5)$	onveranderd
Na Voorbeeld 2.	onveranderd	$(2.875, 3.125)$	$(3.5, 2.25)$	$(2.5, 1.25)$

4. Gegeven is het volgende neurale netwerk:



Alle knopen zijn voorzien van een sigmoïde activatiefunctie. De gewichten tussen AC, AD, BC en BD zijn gelijk aan resp. 03, 0.6, 0.5 en 0.4. Bij input $(A, B) = (0.1, 0.2)$ is de gewenste output $(C, D) = (0.5, 0.5)$.

(a) Bereken bij input $(A, B) = (0.1, 0.2)$ de output o_A, o_B, o_C,... van alle knopen. Schrijf de resultaten in de volgende tabel.

	A	B	C	D
Output				

Antw. Output $o_A = 0.525$, output $o_B = 0.545$, output $o_C = 0.430$, output $o_D = 0.533$.

(b) Bereken de fout $t_C - o_C$,... bij knopen C en D. Schrijf de resultaten in de volgende tabel.

	A	B	C	D
Fout	—	—		

Antw. Fout $t_C - o_C = 0.5 - 0.430 = 0.07$, fout $t_D - o_D = 0.5 - 0.533 = -0.033$,

(c) Bereken voor alle knopen A, B,... de correctiefactor δ_A, δ_B,... bij terug-propagatie. Schrijf de resultaten in de volgende tabel.

	A	B	C	D
Correctiefactor				

Antw. correctiefactor $\delta_A = 0.525(1 - 0.525)[0.3 \times 0.01716 + 0.6 \times -0.00821] = 0.0000553$, correctiefactor $\delta_B = 0.545(1 - 0.545)[0.5 \times 0.01716 + 0.4 \times -0.00821] = 0.00131$, Correctiefactor $\delta_C = 0.43(1 - 0.43)0.07 = 0.01716$, correctiefactor $\delta_D = 0.533(1 - 0.533) - 0.033 = -0.00821$.

5. (Reinforcement leren.)

   (a) Teken een abstracte (i.e., algemene) graaf van een toestand-backupdiagram, waarbij in elke toestand drie acties kunnen worden uitgevoerd, en elke actie kan leiden naar twee mogelijk nieuwe toestanden.

   (b) Schrijf (de algemene versie van) de Bellman-vergelijking op.
   - Tip: doe eerst even op klad.
   - Voor $A(s)$, $S(a,s)$, $P(s'|s,a)$ en $R(s,a,s')$ mag ook A_s, S^a_s, $P^a_{ss'}$ resp. $R^a_{ss'}$ worden geschreven.

   Antw.

$$V(s) = \sum_{a \in A(s)} \pi(s,a) \sum_{s' \in S(s,a)} P(s'|s,a) \left[R(s,a,s') + \gamma V(s') \right].$$

Antw.

6. (Reinforcement leren.)

   (a) Beschouw het Markov beslis-probleem voor bobben (Fig. 1). Ellipsen representeren toestanden, trapezia representeren acties, en stippellijnen representeren onzekere overgangen van acties naar nieuwe toestanden. Het label naast een overgang representeert de geschatte kans op die overgang, tussen haakjes gevolgd door het geschatte nut van die overgang. Om aan te geven dat een individu voor 04:00 uur 's ochtends maar een beperkt aantal toestand kan doorlopen, gebruiken we een verdisconteringsfactor $\gamma = 0.9$.

   (b) Teken in het vak hieronder een toestand-backupdiagram voor toestand B. **Antw.** Als platte boom: $(B, (s, (BP)), (d, (H, A)), (c, (H, A)))$.

   (c) Gegeven is de volgende policy:

π

A	B			H	P	
c	c	d	s	d	s	d
1.0	0.2	0.1	0.7	1.0	0.9	0.1

Bereken $V(B)$ als de waarde (het verwachte nut) van alle overige toestanden gelijk is aan 10. **Antw.**

$$\begin{aligned}
V(B) &= \pi_s(\text{reward } B \text{ en } P) + \pi_d(\text{reward } H \text{ en } A) + \pi_c(\text{reward } H \text{ en } A) \\
&= 0.7(0.9(6 + 0.9V(B)) + 0.1(7 + 0.9V(P))) + \\
&\quad + 0.1(0.7(2 + 0.9V(H)) + 0.3(-99 + 0.9V(A))) + \\
&\quad + 0.2(0.99(1 + 0.9V(H)) + 0.01(-99 + 0.9V(A))) \\
&= 0.7(0.9(6 + 0.9V(B)) + 0.1(7 + 0.9\cdot 10)) + \\
&\quad + 0.1(0.7(2 + 0.9\cdot 10) + 0.3(-99 + 0.9\cdot 10)) + \\
&\quad + 0.2(0.99(1 + 0.9\cdot 10) + 0.01(-99 + 0.9\cdot 10)).
\end{aligned}$$

Vereenvoudigen geeft

$$V(B) = 0.567V(B) + 4.77.$$

Dit geeft

$$V(B) = 11.0162.$$

(d) De aanname in het vorige onderdeel (de aanname dat de waarde van alle overige toestanden gelijk is aan 10) is niet juist. Leg uit hoe de onjuistheid van deze aanname kan worden aangetoond. **Antw.** Alle V-waarden mogen bij opnieuw uitreken niet veranderen. Door $V(B)$ met het toestand backup-diagram voor B opnieuw te berekenen met als oude waarde $V(B) = 11.0162$. Als de nieuwe waarde $V(B) \neq 11.0162$, dan weten we dat de oude waarden niet klopt en dus dat de aanname in het vorige onderdeel onjuist was.

(e) Gegeven zijn de volgende twee runs:

Run 1: $H, d \rightarrow P, s \rightarrow P, s \rightarrow B, s \rightarrow B, c \rightarrow H.$
Run 2: $H, d \rightarrow P, s \rightarrow B, d \rightarrow A, c \rightarrow H.$

Geef van beide runs de ter plekke, en feitelijk ontvangen, immediate rewards. (Schrijf in de tabellen.) **Antw.**

	H	P	P	B	B	H
Immediate reward run 1.	0	2	8	9	6	1

	H	P	B	A	H
Immediate reward run 2.	0	2	9	-99	1

	H	P	P	B	B	H
Immediate reward run 1.	—					

	H	P	B	A	H
Immediate reward run 2.	—				

(f) Bereken van beide runs bij Monte-Carlo simulatie de reward-to-go van alle toestanden. Bereken daarna de gemiddelde reward-to-go, first visit, voor elke toestand. **Antw.**

Run 1:

H	0
B	$1 + 0.9 \cdot 0 = 1$
B	$6 + 0.9 \cdot 1 = 6.9$
P	$9 + 0.9 \cdot 6.9 = 15.21$
P	$8 + 0.9 \cdot 15.21 = 21.689$
H	$2 + 0.9 \cdot 21.689 = 21.520$

Run 2:

H	0
A	$1 + 0.9 \cdot 0 = 1$
B	$-99 + 0.9 \cdot 1 = -98.1$
P	$9 + 0.9 \cdot -98.1 = -79.29$
H	$2 + 0.9 \cdot -79.29 = -69.361$

	H	P	B	A
1st visit RtG na Run 1.	21.520	21.689	6.9	—
1st visit RtG na Run 2.	-69.361	-79.29	-98.1	1
Gemiddeld.	-23.921	-28.801	-45.6	1

	H	P	B	A
1st visit RtG na Run 1.				
1st visit RtG na Run 2.				
Gemiddeld.				

(g) Bereken de temporal difference update van alle toestanden na de eerste run. Neem leerfactor $\alpha = 0.2$. **Antw.**

Run 1:

$$
\begin{array}{ll}
H \to P: & V(H) = 0.8 \cdot 0 + 0.2(2 + 0.9 \cdot 0) = 0.4 \\
P \to P: & V(P) = 0.8 \cdot 0 + 0.2(8 + 0.9 \cdot 0) = 1.6 \\
P \to B: & V(P) = 0.8 \cdot 1.6 + 0.2(9 + 0.9 \cdot 0) = 3.08 \\
B \to B: & V(B) = 0.8 \cdot 0 + 0.2(6 + 0.9 \cdot 0) = 1.2 \\
B \to H: & V(B) = 0.8 \cdot 1.2 + 0.2(1 + 0.9 \cdot 0.4) = 1.232
\end{array}
$$

	H	P	B	A
	0	0	0	0
TDU na Run 1.	0.4	3.08	1.232	0

	H	P	B	A
	0	0	0	0
TDU na Run 1.				

Antw.

Einde van de open vragen. Dit tentamen bevat ook nog meerkeuze vragen.

Meerkeuze vragen

Bij elke meerkeuzevraag is steeds precies één antwoord het juiste.

Antwoorden hier invullen:

	A	*B*	*C*	*D*
1.				
2.				
3.				
4.				
5.				
6.				
7.				
8.				
9.				
10.				

1. We bekijken de volgende beweringen over het prisoner's dilemma. Welke daarvan zijn waar?

   i) Een nulsom spel.

   ii) Een twee-persoons spel.

   iii) Een symmetrisch spel.

   iv) Een spel op basis van volledige informatie.

   (a) *i)*, *ii)* en *iii)*.

   (b) *i)*, *ii)* en *iv)*.

   √ *ii)*, *iii)* en *iv)*.

   (d) Het goede antwoord staat er niet bij.

2. Welk paar van de volgende twee-persoons competitieve symmetrische spelen met simultane zetten en kwantitatieve beloningen bezitten onderling dezelfde pure Nash equilibria én Pareto-optima?

   i) Chicken.

   ii) Stag hunt.

   iii) Snowdrift.

   iv) Matching pennies.

   Let op:

   – Er wordt gevraagd welke twee spelen qua Nash equilibria en Pareto-optima overeenkomen.

   – Er wordt **niet** gevraagd voor welke spelen de pure Nash equilibria en Pareto-optima samenvallen.

   √ *i)* en *iii)*.

   (b) *ii)* en *iii)*.

   (c) *ii)* en *iv)*.

   (d) Het goede antwoord staat er niet bij.

3. *i)* Vector-kwantisatie is een combinatie van ongesuperviseerd en gesuperviseerd leren.

   ii) Vector-kwantisatie kan online en in batch worden uitgevoerd.

   (a) Alleen *i)* is waar.

   (b) Alleen *ii)* is waar.

   √ Zowel *i)* als *ii)* is waar.

   (d) Beide uitspraken zijn onwaar.

   Antw. Het gesuperviseerde deel kan in batch worden uitgevoerd.

4. Hier volgen een aantal termen. Welke daarvan zijn gerelateerd aan ongesuperviseerd leren?

   i) Competitive learning.

   ii) K-means leren.

   iii) Q-leren.

   iv) Leaky leren.

   v) Backpropagation.

   vi) Kohonen netwerk.

   √ *i)*, *ii)*, *iv)* en *vi)*.

   (b) *i)*, *iii)*, *v)* en *vi)*.

   (c) *ii)*, *iv)*, *v)* en *vi)*.

   (d) Het goede antwoord staat er niet bij.

5. Hier volgen een aantal uitspraken over reinforcement leren. Welke daarvan is onwaar?

   (a) Een discount-factor kan op twee manieren begrepen worden: als kans om te overleven, en als ontwaarding van nut op de lange termijn.

   (b) Een policy geeft voor elke mogelijke toestand aan, hoe er moet worden gehandeld.

   (c) Een waardefunctie voor een policy, V, geeft voor elke toestand de verwachte toekomstige beloning.

   √ Een kwaliteitsfunctie voor een policy geeft voor elke actie de verwachte toekomstige beloning.

   Antw. Een kwaliteitsfunctie voor een policy geeft voor elk toestand/actie-paar de verwachte toekomstige beloning.

6. We passen kwalitatief gesuperviseerd leren met vector-kwantisatie toe met drie cluster-representanten, te weten $A = (0, 2)$, $B = (2, 0)$ en $C = (3, 3)$. Bepaal de bewegingen van deze cluster-representanten als het leervoorbeeld
$$l = (0, 1) \rightarrow B$$
wordt aangeboden.

   (a) Er gebeurd niets.

   (b) Repr-B beweegt naar l toe.

   $\surd$ Repr-B beweegt naar l toe en Repr-A beweegt van l af.

   (d) Repr-B beweegt naar l toe en Repr-A en Repr-C bewegen van l af.

7. Benoem het verschil tussen value iteration en temporal difference learning.

   (a) Bij value iteration worden alle values van alle toestanden ge-update, bij temporal difference learning worden alle policies van alle toestanden ge-update.

   (b) Bij value iteration worden alle values van alle toestanden in een epoch ge-update, bij temporal difference learning worden alle policies van alle toestanden in een epoch ge-update.

   $\surd$ Bij value iteration worden alle states ge-update, bij temporal difference learning alleen de states in een epoch.

   (d) Bij value iteration worden alle states in een epoch ge-update, bij temporal difference learning alle states.

8. Zoals bekend kan elke willekeurige functie worden benaderd door een neuraal netwerk bestaande uit hoogstens drie lagen (laag 1: verbindingen tussen input en hidden1; laag 2: verbindingen tussen hidden1 en hidden2; laag 3: verbindingen tussen hidden2 en output-laag).

   i) Boolese functies kunnen exact worden gerepresenteerd door een 2-lagig netwerk.

   ii) Begrensde continue functies kunnen willekeurig dicht worden benaderd door een 2-lagig netwerk.

   (a) Beide uitspraken zijn onwaar.

   (b) Alleen uitspraak *i)* is waar.

   (c) Alleen uitspraak *ii)* is waar.

   $\surd$ Beide uitspraken zijn waar.

9. *i)* Q-leren is een vorm van temporal difference leren.

   ii) Kies actie a in toestand s met kans ϵ op exploratie. Voer a uit. We arriveren in opvolger-toestand s' met beloning r.
   Kies vervolg-actie a' in opvolger-toestand s' met kans ϵ op exploratie. Update $Q(s, a)$ volgens
   $$Q(s, a) := (1 - \alpha)Q(s, a) + \alpha\,[r + \gamma Q(s', a')]$$
   Voer tenslotte a' uit.

   iii) Kies actie a in toestand s met kans ϵ op exploratie. Voer a uit. We arriveren in opvolger-toestand s' met beloning r.
   Bepaal in toestand s' de waarde $\max_{a'} Q(s', a')$. Update $Q(s, a)$ volgens
   $$Q(s, a) := (1-\alpha)Q(s, a)+\alpha\left[r + \gamma \max_{a'} Q(s', a')\right]$$
   Kies pas daarna in opvolger-toestand s' de daadwerkelijke (niet noodzakelijk optimale) vervolg-actie a'' uit met kans ϵ op exploratie, en voer deze uit.

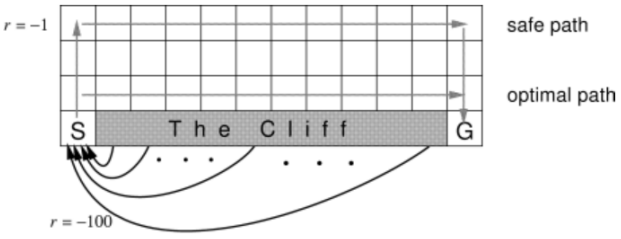
   (a) Uitspraak *i)* is waar en Q-leren wordt geformuleerd in *ii)*.

   $\surd$ Uitspraak *i)* is waar en Q-leren wordt geformuleerd in *iii)*.

   (c) Uitspraak *i)* is niet waar en Q-leren wordt geformuleerd in *ii)*.

   (d) Het goede antwoord staat er niet bij.

10. Beschouw de volgende niet-verdisconteerde episodische taak met starttoestand S en doeltoestand G, en de gebruikelijke acties links, rechts, omhoog, omlaag.



   Alle toestandovergangen leveren een beloning van -1 op, met uitzondering van die in de regio aangegeven met "The Cliff". Een stap in deze regio levert een beloning van -100 op en stuurt de agent direct terug naar S.

   Er is het optimale pad en er is het veilige pad. Het optimale pad is het kortst, maar bij exploratie kan de robot in de klif lopen. Het veilige pad is iets langer maar behoedt de robot voor een val in de klif bij exploratie.

   (a) Q-leren is niet geschikt voor deze leertaak.

   (b) Q-leren levert het veilige pad op.

   $\surd$ Q-leren levert het optimale pad op.

   (d) Q-leren kan beide paden opleveren.

Einde van de meerkeuzevragen.

Einde van alle opgaven. Heb je je antwoorden gecontroleerd? Blad 7/8 mag worden meegenomen. Vergeet niet de onderwijsenquete in te vullen. Bedankt voor je deelname en een prettige vakantie.

Figuren

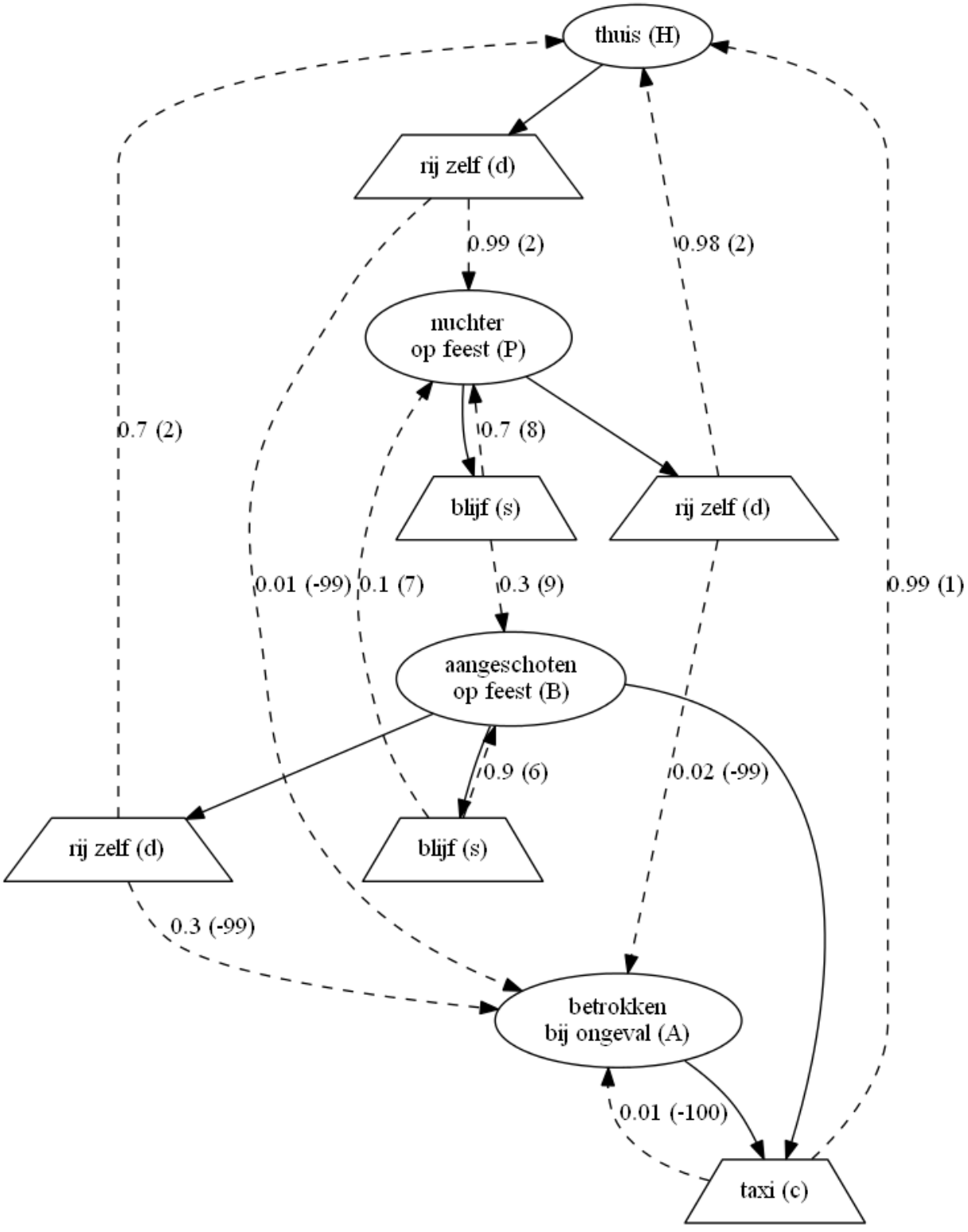


Fig. 1: Toestandengraaf voor bobben.