תעודת זהות:

מספר מחברת:

1	2	3	4	5	6	7	סהב

<u>מבחן מועד א' – למידה מחיזוקים</u>
<u>סמסטר ב' תשע"ח (2017/8)</u>

בית הספר למדעי המחשב, אוניברסיטת תל-אביב

מרצה: פרופ' ישי מנצור,

מתרגלים: מר נחמני אליה, מר פולק אדם.

4.7.2018

<u>הוראות</u>

1. מומלץ לקרוא את כל ההנחיות והשאלות בתחילת המבחן, לפני תחילת כתיבת התשובות.
2. משך הבחינה – **שלוש שעות**. לא תינתן כל הארכה נוספת.
3. חומר עזר מותר: דף עזר אחד בגודל A4 דו-צדדי , **<u>ומחשבון</u>**.
4. **יש לענות על השאלות במקום המיועד לכך בטופס השאלון (טופס זה).** מחברות הבחינה לא ייקראו, וישמשו כטיוטה בלבד.
5. יש למלא בכל דף של השאלון מספר ת.ז. ומספר מחברת.
6. במבחן 4 שאלות:
   - הניקוד לכל שאלה מופיע לידה מספר השאלה.
   - יש לענות תשובות ברורות ענייניות ותמציתיות.
7. מותר להשתמש בכל טענה שהוכחה בכיתה (בהרצאה, בתרגול, או בתרגיל בית) בתנאי שמצטטים אותה במדויק. טענות אחרות (כאלה שהוכחו בספרים, בהרצאות מהסמסטר הקודם, וכו') יש להוכיח.

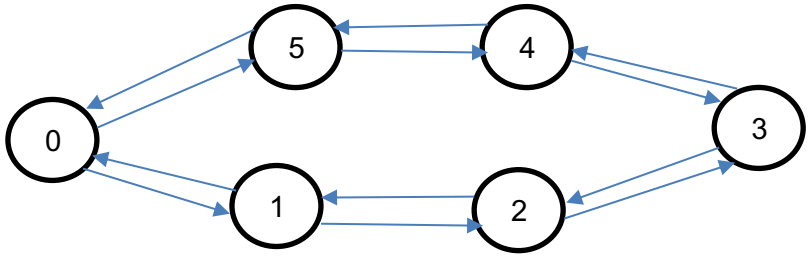
בהצלחה!

תעודת זהות:

מספר מחברת:

שאלה 1 (30 נקודות).

נתון MDP שמוגדר ע"י מעגל דו-כיווני בגודל $2k$ כאשר המצבים מסומנים במספרים 0 עד *2k-1*.
המצב ההתחלתי הוא k.
ישנם שתי פעולות CW (clockwise) ו-CCW (counter-clockwise) שמתבצעות באופן דטרמיניסטי (הפעולה clockwise עוברת ממצב i למצב $i+1 \ mod \ 2k$, והפעולה counter-clockwise עוברת ממצב i למצב $i-1 \ mod \ 2k$).
הרווח המיידי מכל מצב-פעולה (s,a) הוא 0 מלבד מהמצב 0 שמקבלים רווח מיידי 1 (עבור CW ו-CCW)
ההחזר (return) הוא discounted עם פרמטר γ.



א. כתוב/י את ה- MDP של הבעיה באופן פורמלי (עבור MDP בגודל 2k).

S= {0, … , 2k-1}

A = {CW, CCW}

s_0 = {0}

R : $r(s,a) = \begin{cases} 1 & s=0, a \in A \\ 0 & otherwise \end{cases}$

P : $p(j|i,CW) = \begin{cases} 1 & j = i+1 \ mod \ 2k \\ 0 & otherwise \end{cases}$ $\quad$ $p(j|i,CCW) = \begin{cases} 1 & j = i-1 \ mod \ 2k \\ 0 & otherwise \end{cases}$

ב. מה המדיניות האופטימלית π^*?

$$\pi^*(s) = \begin{cases} CCW & s \leq k \\ CW & s \geq k+1 \end{cases}$$

תעודת זהות:

מספר מחברת:

ג. מריצים אלגוריתם Value Iteration כאשר מתחילים שלכל המצבים ערך אפס.
לאחר האיטרציה הראשונה לאיזה מצבים ערך שונה מאפס? מה הערך שלהם?
לאחר האיטרציה השניה לאיזה מצבים ערך שונה מאפס? מה הערך שלהם?

לאחר האיטרציה הראשונה:

מצבים עם ערך שונה מאפס: 0

ערכי המצבים עם ערך שונה מאפס: $V_1(0) = 1$

לאחר האיטרציה השניה:

מצבים עם ערך שונה מאפס: 0,1,2k-1

ערכי המצבים עם ערך שונה מאפס:

$V_2(0) = 1;\ V_2(1) = \gamma; V_2(2k-1) = \gamma$

ד. עבור $k = 2$ (ארבעה מצבים) חשב/י לכל מצב את $V^*(s)$ (כפונקציה של γ).

$V^*(0) = \frac{1}{1-\gamma^2}$

$V^*(1) = \frac{\gamma}{1-\gamma^2}$

$V^*(2) = \frac{\gamma^2}{1-\gamma^2}$

$V^*(3) = \frac{\gamma}{1-\gamma^2}$

תעודת זהות:

מספר מחברת:

שאלה 2 (20 נקודות).

נתונה הריצה הבאה (trajectory):

s1,a1,-8,s1,a2,-16,s2,a1,+20,s1,a2,-10,s2

א. מריצים Q-learning עם $\gamma = 0.5$ ו- $\alpha = 0.5$, כאשר מאתחלים את הערכים לאפס, עבור זמנים $1 \leq t \leq 4$ חשב/י את הערכים של Q-learning.
(הכוונה בשאלה ל- Q-learning ללא ε-greedy).

t=1	s1	s2
a1	-4	
a2		

t=2	s1	s2
a1	-4	
a2	-8	

t=3	s1	s2
a1	-4	+9
a2	-8	

t=4	s1	s2
a1	-4	+9
a2	-6.75	

תעודת זהות:

מספר מחברת:

א. מריצים SARSA עם $\gamma = 0.5$ ו- $\alpha = 0.5$, כאשר מאתחלים את הערכים לאפס, עבור זמנים $1 \leq t \leq 4$ חשב/י את הערכים של SARSA.

בריצה, לאחר המצב האחרון מבוצעת פעולה a1, לכן הריצה היא:

s1,a1,-8,s1,a2,-16,s2,a1,+20,s1,a2,-10,s2,a1

t=1	s1	s2
a1	-4	
a2		

t=2	s1	s2
a1	-4	
a2	-8	

t=3	s1	s2
a1	-4	8
a2	-8	

t=4	s1	s2
a1	-4	+8
a2	-7	

תעודת זהות:

מספר מחברת:

שאלה 3 (20 נקודות).

התפלגות אקספוננציאלית מוגדרת עם פרמטר λ כאשר פונקצית הצפיפת היא $\lambda e^{-\lambda x}$ עבור $x \geq 0$.

משתמשים בהתפלגות אקספוננציאלית לדגום פעולות $a \geq 0$ באופן הבא:
לכל מצב s יש וקטור $\phi(s) \in \mathbb{R}^d$ שמאפיין אותו.
מדיניות π מאופיינת על ידי וקטור $\theta \in \mathbb{R}^d$.
במצב s דוגמים פעולה $a \geq 0$ על ידי התפלגות אקספוננציאלית עם פרמטר $\lambda = \exp(\theta^\top \phi(s))$.

א. מה ההסתברות שנדגום במצב s פעולה $a \in [0,1]$ כפונקציה של θ ו- $\phi(s)$?

$$\int_0^1 \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_0^1 = 1 - e^{-\lambda}$$

$$1 - \exp\left(-\exp\left(\theta^\top \phi(s)\right)\right)$$

ב. חשב/י את $\nabla_\theta \log \pi(a|s;\theta)$.

$$\log \pi(a|s,\theta) = \log \lambda - \lambda a = \theta^\top \phi(s) - a \exp(\theta^\top \phi(s))$$

$$\nabla_\theta \pi(a|s,\theta) = \phi(s) - a\, \phi(s) \exp(\theta^\top \phi(s))$$

ג. כתוב/י את העדכון של REINFORCE עבור ההתפלגות האקספוננציאלית:

$$\theta_{t+1} = \theta_t + \alpha G \phi(s)(1 - a \exp(\phi(s)\theta^\top))$$

Where G is the return

תעודת זהות:

מספר מחברת:

שאלה 4 (30 נקודות).

נתון MDP M המוגדר ע"י (S, A, p, s_0, R).

מגדירים את האופרטור הבא עבור $Q \in \mathbb{R}^{|S|\cdot|A|}$:

$$(HQ)(s,a) = r^2(s,a) + \gamma\, E_{s' \sim p(\cdot|s,a)}[\max_{b \in A} Q\,(s', b)]$$

א. כתוב/י את ההגדרה מתי אופרטור H כלשהו הוא $\gamma - contracting$, עבור נורמה $\|\cdot\|_\infty$

הגדרה:

$$\|HQ_1 - HQ_2\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty$$

ב. הוכח/י שהאופרטור H הוא $\gamma - contracting$.

הוכחה:

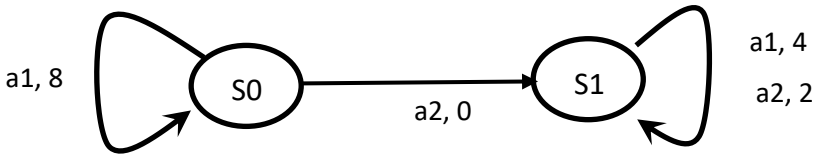
$$|(HQ_1)(s,a) - (HQ_2)(s,a)| = \gamma |E_{s' \sim p(\cdot|s,a)}[\max_{b_1} Q_1(s', b_1) - \max_{b_2} Q_2(s', b_2)|$$

$$\le \gamma \max_{s'} \max_{b} |Q_1(s', b) - Q_2(s', b)| \le \gamma \|Q_1 - Q_2\|_\infty$$

תעודת זהות:

מספר מחברת:

ג. עבור ה- MDP הבא:



חשב/י את HQ את H(HQ) ואת H(H(HQ)) עבור Q=(0,0,0,0) ו- $\gamma = 0.5$
הקידוד של הוקטור Q הוא $(Q(s0,a1), Q(s0,a2), Q(s1,a1), Q(s1,a2))$

HQ = (64 , 0 , 16 , 4)

H(HQ) = (96 , 8 , 24 , 12)

H(H(HQ)) = (112 , 12 , 28 , 16)

ד. הראה שלכל MDP=(S, A, p, s_0, R) קיימת פונקציית רווח R', כך שעבור (S, A, p, s_0, R'), מתקיים $Q^*(s,a;R') = Q^c(s,a;R)$ לכל מצב s ופעולה a, כאשר Q^c היא נקודת השבת (fixed point) של H. (H מוגדרת עם פונקציית הרווח r(s,a) המקורית!)

הפונקציה R':

R'(s,a)= $r^2(s,a)$

הסבר: