תעודת זהות:

מספר מחברת:

1	2	3	4	5	6	7	סהב

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

בית הספר למדעי המחשב, אוניברסיטת תל-אביב

מרצה: פרופ' ישי מנצור,

מתרגלים: מר נחמני אליה, מר פולק אדם.

9.8.2018

הוראות

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, **ומחשבון**.
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 grid דו-מימדי בגודל $(2k+1)\times(2k+1)$ כאשר המשבצות של ה-grid מסומנות ב- (i,j) כאשר $1 \le i,j \le 2k+1$. התא השמאלי התחתון מסומן ב- $(1,1)$.
המצב ההתחלתי הוא במרכז: $(k+1,k+1)$.
ישנם ארבע פעולות up, down, left, right שמתבצעות באופן דטרמיניסטי
(הפעולה right עוברת ממצב (i,j) למצב $(i+1,j)$, הפעולה left עוברת ממצב (i,j) למצב $(i-1,j)$
הפעולה down עוברת ממצב (i,j) למצב $(i,j-1)$ והפעולה up עוברת ממצב (i,j) למצב $(i,j+1)$).
אם אין מצב לעבור אליו, נשארים באותו מצב.

ה-MDP הוא אפיזודי, והריצה מסתיימת כאשר מגיעים לפינה השמאלית העליונה (מצב $(1,2k+1)$) או לפינה הימנית התחתונה (מצב $(2k+1,1)$).
עד לסיום הריצה כל צעד מקבל רווח של -1 (כלומר הפסד של 1)

דוגמא ל- MDP עבור $k=2$

סיום				
		התחלה		
				סיום

א. כתוב/י את ה- MDP של הבעיה באופן פורמלי.

S= $\{(i,j): 1 \le i \le 2k+1, 1 \le j \le 2k+1\}$

A = $\{up, down, left, right\}$

s_0 =$\{(k+1,k+1)\}$

R : $R(s,a) = -1$ for $s \notin \{(1,2k+1),(2k+1,1)\}$
$R(s,a) = 0$ otherwise

P : P((i',j')| (i,j),up)= I(i'=i and j' =min(2k+1,j+1)) similar for down, left, right

ב. מה המדיניות האופטימלית?(עבור k כללי)

$$\pi^*(i,j) = \begin{cases} up & if\ i \le j \le 2k \\ left & if\ j = 2k+1 \\ down & if\ i > j \ge 2 \\ right & if\ i = 1 \end{cases}$$

The optimal policy is not unique.

תעודת זהות:

מספר מחברת:

ג. מה ערך $V^*(s_0)$ ושל $Q^*((2k,2),left)$? (עבור k כללי)

$$V^*(s_0) = -2k$$

$$Q^*\big((2k,2),left\big) = -4$$

ד. מריצים אלגוריתם Value Iteration כאשר מתחילים שלכל המצבים ערך אפס.
לאחר האיטרציה השניה, מה הערך של מצבים $(2,2k+1)$ ושל $(2,2k-1)$?

$$V_2\big((2,2k+1)\big) = -1$$
$$V_2\big((2,2k-1)\big) = -2$$

ה. נניח שהפעולות אינם דטרמיניסטיות, אלה שבהסתברות 0.1 הולכים לכיוון רנדומלי ובהסתברות 0.9 לפי הפעולה הדטרמיניסטית.
האם המדיניות האופטימלית תשתנה? הסבר
האם פונקציית הערך האופטימלית $V^*(s)$ תשתנה? הסבר

האם המדיניות האופטימלית תשתנה? כן / לא

הסבר

Some optimal policies will become sub-optimal (like the one define in (b)). The optimal policy will try to go on the diagonal. It is also optimal in the deterministic case.

האם פונקציית הערך האופטימלית $V^*(s)$ תשתנה? כן / ~~לא~~

הסבר

תעודת זהות:

מספר מחברת:

שאלה 2 (20 נקודות).

נתונות שלוש הריצות הבאות של מדיניות π (E זה המצב המסיים את האפיזודה)

Episode 1: S0,a1,+1,s1,a1,-1,s2,a1,+2,s1,a1,+1,E

Episode 2: S0,a2,+2,s2,a2,-1,s3,a2,0,s1,a2,+1,s2,a1,0,E

Episode 3: S0,a1,0,s3,a1,0,s2,a2,1,s1,a2,-1,s3,a2,0,E

מריצים אלגוריתם Monte-Carlo עם פרמטר $\alpha=1/n(s)$ כאשר $n(s)$ זה מספר העדכונים למצב s (כולל העדכון הנוכחי).

א. עבור Monte-Carlo First Visit, מלא/י בעמודה Vi את הערך לאחר הריצה ה-i.

	V0	V1	V2	V3
S0	0	3	5/2	5/3
S1	0	2	3/2	2/3
S2	0	3	3/2	1
S3	0	0	1	½
E	0	0	0	0

ב. עבור Monte-Carlo Every Visit, מלא/י בעמודה Vi את הערך לאחר הריצה ה-i.

	V0	V1	V2	V3
S0	0	3	5/2	5/3
S1	0	3/2	4/3	¾
S2	0	3	1	¾
S3	0	0	1	1/3
E	0	0	0	0

תעודת זהות:

מספר מחברת:

שאלה 3 (20 נקודות).

עבור עדכון TD(0) ב- off-policy מתבוננים בשני מצבים s1 ו-s2 .
עבור ה- function approximation יש פרמטר $w \in \mathbb{R}$ (מספר ממשי).
ההערכה של הערך במצב s1 היא $\phi(s1) = w$ ובמצב s2 היא $\phi(s2) = 2w^2$.
מעדכנים את הפרמטר לפי TD(0) עם α=1 ו-$\gamma = 0.5$ בעזרת המעבר (s1,a,0,s2)
(כל העדכונים נעשים עם המעבר הזה!).

א. חשב/י את w_1 כאשר מאתחלים את הפרמטר לערך $w_0 = 0.1$.
(w_1 זה הערך של הפרמטר אחרי עדכון בודד).

$$w_1 = 0.01$$

ב. אם מתחילים עם w_0=10 לאיזה ערך w_t יתכנס כאשר t שואף לאין-סוף?
(w_t זה הערך של הפרמטר אחרי t עדכונים)

$$\lim_{t\to\infty} w_t = \infty$$

ג. תן נוסחה עבור w_t . לאיזה ערך w_t יתכנס?
(w_t זה הערך של הפרמטר אחרי t עדכונים)

נוסחה עבור w_t (גם כפונקציה של w_{t-1} וגם כפונקציה של w_0):

$$w_t = w_{t-1} + \alpha\left(r_{t-1} + \gamma\phi(s_2) - \phi(s_1)\right) = w_{t-1}^2 = w_0^{2^t}$$

הערך אליו יתכנס w_t כאשר t שואף לאין-סוף (כפונקציה של w_0):

$$\lim_{t\to\infty} w_t = \begin{cases} 0 & if\ w_0 \in (-1,1) \\ 1 & if\ w_0 \in \{-1,1\} \\ \infty & else \end{cases}$$

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 $V \in \mathbb{R}^{|S|}$:

$$(HV)(s) = \frac{1}{|A|}\sum_{a}(r(s,a) + \gamma E_{s' \sim p(\cdot|s,a)}[V(s')]$$

א. כתוב/י את ההגדרה מתי אופרטור H הוא $\gamma - contracting$.

הגדרה:

$$\forall v_1, v_2: \ ||Hv_1 - Hv_2||_\infty \leq \gamma: \ ||v_1 - v_2||_\infty$$

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

הוכחה:

$$\forall V_1, V_2: \ ||HV_1 - HV_2||_\infty = \frac{1}{|A|}\max_{s\in S}\left|\sum_{a\in A}\gamma E_{s' \sim p(\cdot|s,a)}\left(V_1(s') - V_2(s')\right)\right|$$

$$= \frac{\gamma}{|A|}\max_{s\in S}\left|\sum_{a\in A}\sum_{s'}p(s'|s,a)\left(V_1(s') - V_2(s')\right)\right| \leq$$

$$\leq \frac{\gamma}{|A|}\max_{s\in S}\left||A|\sum_{s'}\max_{s'}(|V_1(s') - V_2(s')|)\right|$$

$$= \gamma\max_{s\in S}|\left(V_1(s') - V_2(s')\right)| =$$

$$\gamma||V_1 - V_2||$$

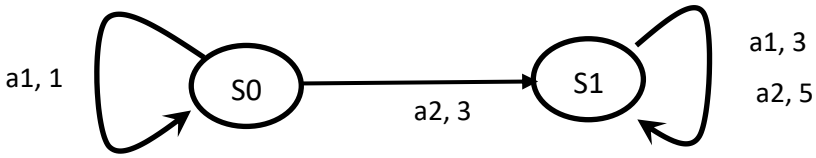
תעודת זהות:

מספר מחברת:

תעודת זהות:

מספר מחברת:

ג. עבור ה- MDP הבא:



חשב/י את HV את H(HV) ואת H(H(HV)) עבור V=(0,0) ו- $\gamma = 0.5$.
הקידוד של הוקטור V הוא $(V(s0), V(s1))$

HV = (2 , 4)

H(HV) = (3.5 , 6)

H(H(HV)) = (4.375 , 7)

ד. האם בהכרח קיימת מדיניות π (יתכן סטוכסטית) עבורה $V^c = V^\pi$ כאשר V^c היא נקודת השבת (fixed point) של H. הסבר.

קיימת מדיניות: כן / לא

הסבר:

מדיניות רנדומלית, לכל מצב s ופעולה a: $\pi(a|s) = \frac{1}{|A|}$