תעודת זהות:

מספר מחברת:

1	2	3	4	5	6	7	סהב

<u>מבחן לדוגמא – 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.

x.x.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, **<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:

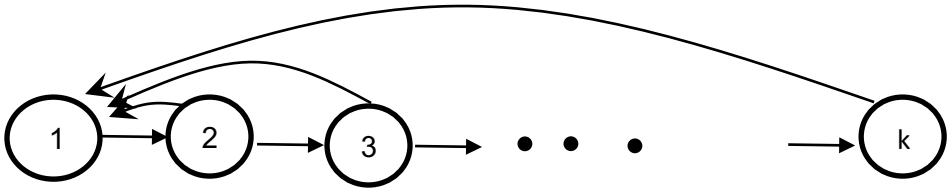
מספר מחברת:

<u>שאלה 1 (30 נקודות).</u>

נתון MDP שמוגדר ע"י קו מכוון בגודל k כאשר המצבים מסומנים במספרים 1 עד k.
המצב ההתחלתי הוא 1.
ישנם שתי פעולות advanced ו- return שמתבצעות באופן דטרמיניסטי.
הפעולה advanced עוברת ממצב i למצב $i+1$, והפעולה return עוברת ממצב i למצב 1
(במצב k יש רק פעולת return).

כל פעולה מקבלת רווח של 0 מלבד הפעולה return במצב k (שמקבלת רווח 1).

ההחזר (return) הוא discounted עם פרמטר γ.



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

S=

A =

s_0 =

R :

P :

ב. מה המדיניות האופטימלית ומה פונקצית הערך שלה?

תעודת זהות:

מספר מחברת:

ג. מה ערך $V^*(1)$ ו- $V^*(k)$?

$V^*(1) =$

$V^*(k) =$

ד. 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?

ה. עבור $k = 3$ (שלושה מצבים) חשב/י לכל מצב את $V^*(s)$ כאשר $\gamma = 0.5$.

$V^*(1) =$

$V^*(2) =$

$V^*(3) =$

תעודת זהות:

מספר מחברת:

שאלה 2 (20 נקודות).

נתונה הריצה הבאה של מדיניות π:

s0,a1,+2,s1,a1,-4,s2,a1,+3,s1,a1,+1,s1

מריצים אלגוריתם TD(0) עם פרמטר α=0.5 ופרמטר $\gamma = 0.5$.

א. עבור TD(0), מלא/י בעמודה את ערכי כל מצב בכל נקודת זמן (כולל ערכים שלא השתנו).

	t=0	t=1	t=2	t=3	t=4
s0	0				
s1	0				
s2	0				

ב. עבור TD(λ) חשב/י את ה eligibility traces של כל מצב כאשר $\lambda = 0.1$ ו-$\gamma = 1$.

	t=0	t=1	t=2	t=3	t=4
s0	0				
s1	0				
s2	0				

תעודת זהות:

מספר מחברת:

שאלה 3 (20 נקודות).

התפלגות פארטו מוגדרת עם פרמטר α כאשר פונקצית הצפיפת היא $\frac{\alpha}{x^{\alpha+1}}$ עבור $x \geq 1$.

משתמשים בהתפלגות פארטו לדגום פעולה $a \geq 1$ 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 $a \geq 1$ על ידי התפלגות פארטו עם פרמטר $\alpha = \left(\theta^\top \phi(s)\right)^2$.

א. מה ההסתברות שנדגום במצב s פעולה $a \in [1,10]$ כפונקציה של θ ו- $\phi(s)$?

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

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

תעודת זהות:

מספר מחברת:

<u>שאלה 4 (30 נקודות).</u>

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

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

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

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

הגדרה:

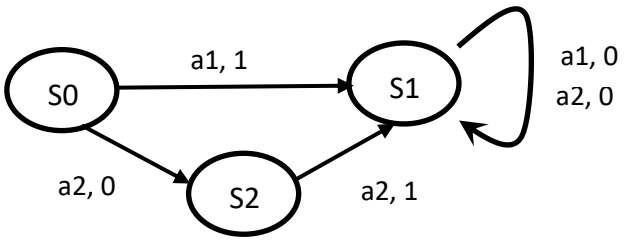
ב. הוכח שהאופרטור H הוא $\gamma - contracting$.

הוכחה:

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- $\gamma = 0.9$.

HQ = (, , ,)

H(HQ) = (, , ,)

H(H(HQ)) = (, , ,)

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

קיימת מדיניות: כן / לא

הסבר: