

## חישוביות וקוגניציה – תרגיל 7

להגשה עד: 30/12/2019

שימו לב: שאלה 1 היא שאלה שמערבת חישובים אנליטיים וסימולציות נומריות. הגישו קובץ ובו תשובות מילוליות, גרפים, וחישובים שנתבקשתם לערוך, וכן את קבצי הקוד שכתבתם

### שאלה 1

נתון Markov Decision Process (MDP) ובו שני מצבים – Home, Out ושתי פעולות Stay, Switch. סיכויי המעברים מוגדרים באופן הבא:

Stay	$H$	$O$
$H$	1	0
$O$	0	1

Switch	$H$	$O$
$H$	0.2	1
$O$	0.8	0

כאשר כל תא בטבלא מתאר את סיכויי המעבר מהמצב שמצויין בעמודה למצב שמצויין בשורה, כאשר לוקחים את הפעולה הרלבנטית. למשל

$$P[H|H, \text{Switch}] = 0.2$$

וכו'.

הגמול מוגדר באופן הבא (בצורה דטרמיניסטית):

$$\begin{aligned}r(H, \text{Stay}) &= 0 \\r(H, \text{Switch}) &= 1 \\r(O, \text{Stay}) &= 2 \\r(O, \text{Switch}) &= 0\end{aligned}$$

פרמטר ה discounting הוא  $\gamma = 0.5$   
שימו לב: השאלות מתחילות בעמוד הבא

## חלק א' – שיטות מבוססות מודל ומשוואות בלמן

1. נתון סוכן אשר מקבל החלטות באקראי, כלומר בכל מצב בוחר כל אחת מהפעולות בסיכוי שווה של 0.5. כתבו את משוואות בלמן עבור פונקציית הערך  $V^\pi$  של הסוכן, ופתרו אותן כדי למצוא את הפונקציה  $V^\pi$ .
2. מבלי לבצע חישובים, מהי לדעתכם המדיניות (Policy) האופטימלית?
3. וודאו שה Policy שניחשתם בסעיף 2 מקיימת את משוואות האופטימליות של בלמן.
4. כתבו פונקציה שמממשת את אלגוריתם Value Iteration עבור ה MDP הנתון. מהי  $V^*$ ? האם המדיניות האופטימלית  $\pi^*$  שמצאתם בעזרת הפונקציה זהה לזו שמצאתם בסעיף 2?

הערות:

- באלגוריתם ה Value Iteration, אתחלו את האלגוריתם לפונקציית ערך אקראית ( $V(s) = 0$  לכל  $s$ ), ובצעו איטרציות כל עוד  $|V^{(t+1)}(s) - V^{(t)}(s)| > \epsilon$  עבור  $\epsilon$  קטן (למשל  $1e-10$ ) וכל עוד מספר האיטרציות קטן מ  $T_{\max}$  עבור  $T_{\max} = 5000$  גדול (למשל  $T_{\max} = 5000$ ). שימו לב שבפועל ההתכנסות צפויה להיות מהירה בהרבה).
- השתמשו בכלל עדכון סינכרוני, כלומר חשבו את הערכים החדשים  $V^{(t+1)}$  של כל המצבים בהתבסס על הערכים הנוכחיים  $V^{(t)}$ .

## חלק ב' – שיטות Model-Free ולמידת הפרשים זמניים

1. כתבו פונקציה שתאפשר לכם להריץ סימולציה של סוכן בסביבה – כלומר, פונקציה שמקבלת מצב  $s$  ופעולה  $a$  ומחזירה (באופן הסתברותי במידת הצורך) את המצב הבא  $s'$ , ואת הגמול שהתקבל מהמעבר (כלומר  $R(s, a)$ , שימו לב שבמקרה שלנו הגמול הוא דטרמיניסטי).
2. השתמשו באלגוריתם TD-Learning ע"מ ללמוד את  $V^\pi$  עבור הסוכן האקראי מסעיף 1 בחלק א'. אתחלו את  $V$  בצורה שרירותית (למשל  $V(s) = 0$  לכל  $s$ ), ובצעו סימולציה של הסוכן בסביבה כאשר בחירת הפעולות נעשית לפי המדיניות שהוגדרה לעיל (בכל מצב, כל פעולה נבחרת באקראי בהסתברות שווה של 0.5). כאשר הסוכן עובר ממצב  $s$  למצב  $s'$  ומקבל גמול  $r$  עדכנו את הערך  $V(s)$  לפי כלל הלימוד:

$$V(s) \leftarrow V(s) + \eta(r + \gamma V(s') - V(s))$$

כאשר  $\eta$  קצב הלימוד (השתמשו ב  $\eta = 0.01$ ). הריצו את הלמידה עד  $T = 3000$  צעדים. הציגו גרף של  $V(\text{Home})$  ו  $V(\text{Out})$  כפונקציה של מספר הצעדים בסביבה שביצע הסוכן. סמנו בקווים מקווקווים על אותו גרף את הערכים האמיתיים של  $V^\pi$  שמצאתם בסעיף 1 של חלק א'. האם הלמידה מתכנסת לערך הנכון? מהי ההשפעה של קצב הלימוד על ההתכנסות?

3. השתמשו באלגוריתם Q-Learning ע"מ ללמוד את המדיניות האופטימלית באופן Off-Policy, בהתבסס על ההתנהגות של הסוכן האקראי. אתחלו את  $Q$  בצורה שרירותית (למשל  $Q(s, a) = 0$  לכל  $s, a$ ) ובצעו סימולציה של אותו סוכן. כאשר הסוכן במצב  $s$ , בוחר פעולה  $a$ , מקבל גמול  $r$  ועבור למצב  $s'$  עדכנו את  $Q(s, a)$  לפי כלל הלמידה הבא:

$$Q(s, a) \leftarrow Q(s, a) + \eta \left( r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right)$$

כאשר  $\eta$  קצב הלימוד (השתמשו ב  $\eta = 0.1$ ). הריצו את הלמידה עד  $T = 3000$  צעדים. הציגו גרף של  $\hat{V}^*(s) \equiv \max_a Q(s, a)$  עבור שני המצבים כפונקציה של מספר הצעדים בסביבה שביצע הסוכן. סמנו בקווים מקווקווים על אותו גרף את הערכים האמיתיים של  $V^*$  שמצאתם בסעיף 4 של חלק א'. האם הלמידה מתכנסת לערך הנכון?