תרגיל בית 3 – MDP ומבוא ללמידה

עברו על כלל ההנחיות לפני תחילת התרגיל.

הנחיות כלליות:

- תאריך ההגשה: לחלק א' של התרגיל (MDP) – עד ליום האחרון של הסמסטר - **08/04/2024** ב23:59
  לחלק ב' של התרגיל (מבוא ללמידה) – עד לסוף מועדי א' - **17/05/2024** ב23:59
- את המטלה יש להגיש **בזוגות בלבד.**
- יש להגיש מטלות מוקלדות בלבד. פתרונות בכתב יד לא ייבדקו.
- ניתן לשלוח שאלות בנוגע לתרגיל בפיאצה בלבד.
- המתרגל האחראי על תרגיל זה: **דניאל אלגריסי.**
- בקשות דחיה מוצדקות (מילואים, אשפוז וכו') יש לשלוח למתרגל האחראי (**ספיר טובול**) בלבד.
- במהלך התרגיל ייתכן שנעלה עדכונים, למסמך הנ"ל – תפורסם הודעה בהתאם.
- העדכונים הינם מחייבים, ועליכם להתעדכן עד מועד הגשת התרגיל.
- שימו לב, העתקות תטופלנה בחומרה.
- התשובות לסעיפים בהם מופיע הסימון ✍️ צריכים להופיע בדוח.
- לחלק הרטוב מסופק שלד של הקוד.
- אנחנו קשובים לפניות שלכם במהלך התרגיל ומעדכנים את המסמך הזה בהתאם. גרסאות עדכניות של המסמך יועלו לאתר. הבהרות ועדכונים שנוספים אחרי הפרסום הראשוני יסומנו כאן בצהוב. ייתכן שתפורסמנה גרסאות רבות – אל תיבהלו מכך. השינויים בכל גרסה יכולים להיות קטנים.

שימו לב שאתם משתמשים רק בספריות הפייתון המאושרות בתרגיל (מצוינות בתחילת כל חלק רטוב)
לא יתקבל קוד עם ספריות נוספות

מומלץ לחזור על שקפי ההרצאות והתרגולים הרלוונטיים לפני תחילת העבודה על התרגיל.

חלק א' – MDP (44 נק')

רקע

בחלק זה נעסוק בתהליכי החלטה מרקובים, נתעניין בתהליך עם **אופק אינסופי** (מדיניות סטציונרית).

חלק א' - חלק היבש ✍️

1. בתרגול ראינו את משוואת בלמן כאשר התגמול ניתן עבור המצב הנוכחי בלבד, כלומר $R: S \to \mathbb{R}$, למתן תגמול זה נקרא "תגמול על הצמתים" מכיוון שהוא תלוי בצומת שהסוכן נמצא בו.

   בהתאם להגדרה זו הצגנו בתרגול את האלגוריתמים Value iteration ו-Policy Iteration למציאת המדיניות האופטימלית.

   כעת, נרחיב את ההגדרה הזו, לתגמול המקבל את המצב הנוכחי והמצב אליו הגיע הסוכן, כלומר: $R: S \times S \to \mathbb{R}$, למתן תגמול זה נקרא "תגמול תוצאתי". לצורך שלמות ההגדרה, נגדיר שאם לכל $a \in A$ מתקיים - $P(s'|s,a) = 0$ אז $R(s,s') = -\infty$.

   א. (1 נק') התאימו את הנוסחה של התוחלת של התועלת מהתרגול, עבור התוחלת של התועלת המתקבלת במקרה של " תגמול תוצאתי ", אין צורך לנמק.

$$U^{\pi}(s) = E_{\pi}[\Sigma_{t=0}^{\infty} \gamma^t R(s,s')|s_0 = s]$$

   ב. (1 נק') כתבו מחדש את נוסחת משוואת בלמן עבור המקרה של " תגמול תוצאתי ", אין צורך לנמק.

$$U(s) = \max_{a \in A(s)} \left[\sum_{s'} P(s'|s,a)[R(s,s') + \gamma U(s')] \right]$$

בסעיפים הבאים התייחסו גם למקרה בו $\gamma = 1$, והסבירו מה לדעתכם התנאים שצריכים להתקיים על הסביבה\mdp על מנת שתמיד נצליח למצוא את המדיניות האופטימלית.

עבור $\gamma = 1$ נרצה לוודא שהתועלת לא תשאף לאינסוף, כדי שהסוכן לא יתקע בלולאה אינסופית ורק יגדיל את התועלת שלו כל הזמן.

כדי לוודא שנוכל למצוא מדיניות אופטימלית ניתן לדרוש על הסביבה להכיל מצב סופי, או שהמדיניות האופטימלית תגיע למצב סופי, או שהתגמול על המצבים יהיה שלילי או אפס, כדי שלא ישתלם לסוכן להסתובב בלי סוף.

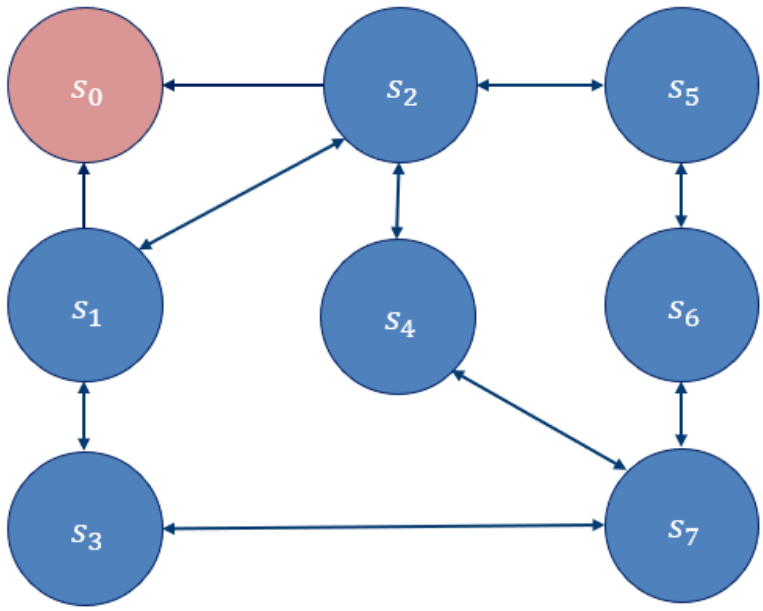
ג. (2 נק') נסחו את אלגוריתם Value Iteration עבור המקרה של " תגמול תוצאתי ".

```
VALUE ITERATION:
initialize: U, U', δ
Repeat
        U ← U'; δ ← 0
        for each state s in S do:
                U'(s) = max_{a ∈ A(s)} [Σ_{s'} P(s'|s,a)[R(s,s') + γU(s')]]
                If |U'(s) − U(s)| > δ:
                        δ ← |U'(s) − U(s)|
Until δ < ε(1−γ)/γ, incase γ < 1. Or δ = 0, incase γ = 1.
Return U
```

ד. (2 נק') נסחו את אלגוריתם Policy Iteration עבור המקרה של " תגמול תוצאתי ".

```
POLICY ITERATION:
initialize: U(initially zero), π(initially random)
Repeat
        U ← POLICY − EVALUTAION(π, U, mdp)
        unchanged? ← true
        for each state s in S do:
                if max_{a ∈ A(s)} Σ_{s'} P(s'|s,a)[R(s,s') + γU(s')] > Σ_{s'} P(s'|s,π(s))[R(s,s') + γU(s')] then do
                        π(s) ← argmax_{a ∈ A(s)} max_{a ∈ A(s)} Σ_{s'} P(s'|s,a)[R(s,s') + γU(s')]
                        unchanged? ← false
Until unchanged?
Return π
```

נתון הגרף הבא:



נתונים:

- $\gamma = 0.5$ (Discount factor).
- אופק אינסופי.
- $S = \{s_0, s_1, s_2, s_3, s_4, s_5, s_6, s_7\}$ – קבוצת המצבים – מתארים את מיקום הסוכן בגרף.
- $S_G = \{s_0\}$ – קבוצת המצבים הסופיים.
- קבוצת הפעולות לכל מצב (על פי הגרף), לדוגמא: $A(s_3) = \{\uparrow, \rightarrow\}$.
- תגמולים ("תגמול תוצאתי"):
- $\forall s \in S, s' \in S \backslash S_G:\ \ R(s, s') = -1,\ \ R(s_1, s_0) = 5,\ \ R(s_2, s_0) = 7$
- מודל המעבר הוא דטרמיניסטי, כלומר כל פעולה מצליחה בהסתברות אחת.

ה. (יבש 2 נק') הרץ את האלגוריתם Value iteration שכתבת על הגרף הנתון. ומלא את הערכים בטבלה הבאה, כאשר $\forall s \in S \cancel{\setminus S_G}: U_0(s) = 0$. (ייתכן שלא צריך למלא את כולה).

לא ברור לי מהו אפסילון, אבל כנראה לא צריך כי באיטרציה 4 זה מתכנס ולא משתנה יותר

	$U_0(s_i)$	$U_1(s_i)$	$U_2(s_i)$	$U_3(s_i)$	$U_4(s_i)$	$U_5(s_i)$	$U_6(s_i)$	$U_7(s_i)$	$U_8(s_i)$
s_1	0	5	5	5	5				
s_2	0	7	7	7	7				
s_3	0	-1	1.5	1.5	1.5				
s_4	0	-1	2.5	2.5	2.5				
s_5	0	-1	2.5	2.5	2.5				
s_6	0	-1	-1.5	0.25	0.25				
s_7	0	-1	-1.5	0.25	0.25				

ו. (יבש 2 נק') הרץ את האלגוריתם Policy iteration שכתבת על הגרף הנתון. ומלא את הערכים בטבלה הבאה, כאשר המדיניות ההתחלתית π_0 מופיעה בעמודה הראשונה בטבלה. (ייתכן שלא צריך למלא את כולה).

	$\pi_0(s_i)$	$\pi_1(s_i)$	$\pi_2(s_i)$	$\pi_3(s_i)$	$\pi_4(s_i)$	$\pi_5(s_i)$	$\pi_6(s_i)$	$\pi_7(s_i)$	$\pi_8(s_i)$
s_1	$\downarrow$	$\uparrow$	$\uparrow$	$\uparrow$					
s_2	$\downarrow$	$\leftarrow$	$\leftarrow$	$\leftarrow$					
s_3	$\rightarrow$	$\rightarrow$	$\uparrow$	$\uparrow$					
s_4	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$					
s_5	$\leftarrow$	$\leftarrow$	$\leftarrow$	$\leftarrow$					
s_6	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$					
s_7	$\uparrow$	$\uparrow$	$\nwarrow$	$\nwarrow$					

ז. (יבש 2 נק') חיזרי על הסעיף הקודם. הפעם עם **אופק סופי כאשר $N = 2$** (שימי לב, המדיניות לא חייבת להסתיים במצב מסיים, ישנם מצבים שלא יכולים להגיע למצב מסיים עם אופק זה. ישנם צמתים עם מספר תשובות נכונות, נקבל את כולם).

כאשר האופק של תהליך ההחלטה הוא סופי, אז הסוכן לא יכול לבצע יותר מ2 צעדים, ואחרי 2 צעדים המשחק נגמר. במקרה הזה, הרצת האלגוריתם על הגרף מניב את **אותה תוצאה שקיבלנו בסעיף הקודם**. וזאת מכיוון שהיה לנו מקדם דעיכה, והreward על מעבר לצומת שהיא לא צומת סופית הוא -1. לכן העדפנו להגיע כמה שיותר מוקדם לצומת הסופית. במקרה ש N = 2, אנחנו עדיין נעדיף להגיע כמה שיותר מוקדם לצומת הסופית, ולכן אין שוני בהרצת של האלגוריתם ובתוצאות שלו.

נשים לב כי הצמתים s_6, s_7 לא יגיעו לצומת היעד, וזאת מכיוון שהם במרחק יותר גדול מ2.

ח. (1 נק') ללא תלות בשינוי של הסעיף הקודם. אם $\gamma = 0$, מה מספר המדיניות האופטימליות הקיימות? נמקו.

עבור צמתים שרחוקים מרחק צעד אחד מצומת היעד, יש מדיניות אופטימלית יחידה, וזאת מכיוון שתמיד נעדיף ללכת ישר לצומת היעד, אחרת לא היינו מקבלים את הreward עליה, אם לא הגענו אליה בצעד הראשון.

עבור הצמתים האחרים, זה לא משנה לאן נלך, הreward הסופי תמיד יהיה 1-. מכיוון שצמתים אלה רחוקים יותר מצעד אחד לצומת היעד, נצטרך לעבור קודם לצומת אחרת, ולקבל את ה reward שהוא 1-. אחר כך, כל צעד שנעשה נקבל עליו 0 reward, לכן לא משנה מהו הצעד, ולכן כל צעד הוא צעד אופטימלי.

לs_1, s_2 יש מדיניות אופטימלית אחת לכל אחד

לצומת אחר יש מספר מדיניות אופטימליות שהוא זהה למספר הקשתות היוצאות ממנו.

לכן בסה"כ יש לנו 48 מדיניות אופטימליות $\underbrace{1}_{s_1} * \underbrace{1}_{s_2} * \underbrace{2}_{s_3} * \underbrace{2}_{s_4} * \underbrace{2}_{s_5} * \underbrace{2}_{s_6} * \underbrace{3}_{s_7} = 48$

ט. (1 נק') ללא תלות בשנוי של הסעיף הקודם, הסבירי מה היה קורה אם

$$R(s_1, s_2) = R(s_2, s_1) = 2, \qquad \gamma = 1$$

בתשובתך, התייחסי גם לערכי התועלות של כל צומת וגם לשינוי במדיניות, אין צורך לחשב.

עקב השינוי הנתון, כעת המדיניות עבור כל צומת שאינו s_1, s_2 לא תשתנה. אבל המדיניות של שני צמתים אלה כן תשתנה. כעת במקום לעבור מהן לצומת היעד, נעדיף לעבור לצומת השנייה, כלומר מs_1 אל s_2 ולהיפך. וזאת מכיוון שמקדם הדעיכה שלנו הוא 1, ואין הפסד על הreward שמקבלים. ולכן ניתן לעבור כל הזמן בין שני הצמתים הללו ולקבל תועלת אינסופית. בנוסף, ערכי התועלת של שאר הצמתים גם ישתנו, וגם בהם נקבל תועלת אינסופית.

צילומים מהרטוב:

get_all_policies עבור וקטור U הנתון:

```
| →      | →      | →      | +1     |
| ↑      | WALL   | ←      | -1     |
| ↑      | ←      | ←      | ↓      |

1
```

:get_policy_for_different_rewards

```
-5.0 <= R(s) < -1.46
| R     | R     | R     | +1    |
| U     | WALL  | R     | -1    |
| R     | R     | R     | U     |

-1.46 <= R(s) < -1.30
| R     | R     | R     | +1    |
| U     | WALL  | U     | -1    |
| R     | R     | R     | U     |

-1.3 <= R(s) < -0.61
| R     | R     | R     | +1    |
| U     | WALL  | U     | -1    |
| R     | R     | U     | U     |

-0.61 <= R(s) < -0.60
| R     | R     | R     | +1    |
| U     | WALL  | U     | -1    |
| RU    | R     | U     | U     |

-0.6 <= R(s) < -0.37
| R     | R     | R     | +1    |
| U     | WALL  | U     | -1    |
| U     | R     | U     | U     |
```

```
-0.37 <= R(s) < -0.05
| R      | R      | R      | +1     |
| U      | WALL   | U      | -1     |
| U      | R      | U      | L      |

-0.05 <= R(s) < -0.04
| R      | R      | R      | +1     |
| U      | WALL   | U      | -1     |
| U      | L      | U      | L      |

-0.04 <= R(s) < 0.01
| R      | R      | R      | +1     |
| U      | WALL   | U      | -1     |
| U      | L      | U      | D      |

0.01 <= R(s) < 0.09
| R      | R      | R      | +1     |
| U      | WALL   | L      | -1     |
| U      | L      | U      | D      |

0.09 <= R(s) < 0.10
| R      | R      | R      | +1     |
| U      | WALL   | L      | -1     |
| U      | L      | LU     | D      |

0.1 <= R(s) < 0.11
| DLRU   | DLRU   | DLRU   | +1     |
| DLRU   | WALL   | L      | -1     |
| DLRU   | DLRU   | DLRU   | D      |
```

```
0.11 <= R(s) <= 5.0
| DLRU   | DLRU   | L      | +1     |
| DLRU   | WALL   | L      | -1     |
| DLRU   | DLRU   | DLRU   | D      |
```

```
[-1.46, -1.3, -0.61, -0.6, -0.37, -0.05, -0.04, 0.01, 0.09, 0.1, 0.11]
```