Introduction to Artificial Intelligence (236501)

תרגיל 3

מגישים : אייל אמדור, בארי זיטלני

ת.ז : 318849270, 209351626

חלק א' – MDP ו-RL

חלק א - יבש

שאלה 1

א. הנוסחה עבור התוחלת של התועלת המתקבלת במקרה של "תגמול על פעולה" היא:

$$U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t \cdot R(s_t, \pi(s_t)) | s_0 = s\right]$$

ב. הנוסחה עבור משוואת בלמן:

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

ג. הפונקציה value-iteration:

$repeat:$

$U \leftarrow U'; \delta \leftarrow 0$

$for\ each\ state\ s\ in\ S\ do:$

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

$if\ |U'[s] - U[s]| > \delta\ then\ \delta \leftarrow |U'[s] - U[s]|$

$$until\ \delta < \frac{\epsilon(1-\gamma)}{\gamma}$$

$return\ U$

נשים לב שעבור $\gamma > 1$ או $\gamma = 0,1$ האלגוריתם לא יתכנס.

ד. הפונקציה policy-iteration:

$repeat:$

$U \leftarrow Policy - Evaluation(\pi, U, mdp)$

$unchanged? \leftarrow true$

$for\ each\ state\ s\ in\ S\ do:$

$$if\ \max_{a \in \mathbb{A}(s)} \left(R(s,a) + \gamma \sum_{s' \in S} P(s'|a,s) \cdot U(s')\right) > R\big(s, \pi(s)\big) + \gamma \sum_{s' \in S} P(s'|\pi(s), s) \cdot U(s')$$

$then\ do:$

$$\pi[s] \leftarrow \operatorname{argmax}_{a \in \mathbb{A}(s)} \left(R(s,a) + \gamma \sum_{s' \in S} P(s'|a,s) \cdot U(s')\right)$$

$unchanged? \leftarrow false$

$until\ unchanged?$

$return\ \pi$

נשים לב שעבור אינסוף מצבים אפשריים ו-$\gamma > 1$ לא מובטח שהאלגוריתם ייתכנס.

שאלה 2

1. נגדיר את בעיית ה-MDP באופן הבא כאשר נתון:

$$p - probaility\ of\ accept\ blackmail$$

$$n - profit\ goal\ which\ achiveing\ it\ ends\ blackmailing$$

a. קבוצת מצבים S:

$$S = \{0,1,2 \dots, n, T |\ i = sum\ of\ money, \qquad T = Terminal\ state\}$$

b. קבוצת פעולות A:

$$A = \{B, L | B = Blackmail, L = Leave\}$$

c. מודל מעברים $P(s, a, s')$:

$action\ /\ state$	$i < n$	n	T
$BlackMail$	$P(i, B, i+1) = p$ $P(i, B, T) = 1 - p$ $P(i, B, j) = 0$ $\forall j \neq i + 1$	$Not\ defined$	$P(T, B, T) = 1$ $P(T, B, j) = 0$ $\forall j \in [n]$
$Leave$	$P(i, L, j) = 0\ \ \forall j \in [n]$ $P(i, L, T) = 1$	$P(n, L, i) = 0$ $\forall i \in [n]$ $P(n, L, T) = 1$	$P(T, L, T) = 1$ $P(T, L, j) = 0$ $\forall j \in [n]$

- הערה: מאופן הגדרת מודל המעברים T הוא מצב טרמינלי.

d. תגמול:

$$R(i, L) = 2i \quad \forall i \in [n] \qquad\qquad R(T, L) = 0$$

$$R(i, B) = 0 \quad \forall i < n \qquad\qquad R(T, B) = 0$$

e. מקדם דעיכה:

$$\gamma = 1$$

2. לא ניתן לנסח את הבעיה עם שני מצבים בלבד מאחר וכך לא ניתן יהיה להגדיר בצורה הנדרשת את פונקציית התגמול. בשני מצבים לא ניתן יהיה לדעת איך לאפס את התגמול במידה והקורבן בוחר להתקשר למשטרה (כלומר, כמה שלילי התגמול צריך להיות).
3. כן, ניתן לנסח את הבעיה עם ערכים שלילים, לכל מצב i זוגי נגדיר תגמול חיובי עם הערך 2, ולכל i אי זוגי נגדיר ערך שלילי $-(i-1)$. באופן הזה בכל פעם שהקורבן שילם סכמנו את התגמול וכאשר הוא מתקשר למשטרה נדע כמה להוריד מסך התגמול כדי לאפס אותו.
4. המדיניות האופטימלית מוגדרת באופן הבא:

$$\pi^*(s) = \underset{a \in \mathbb{A}(s)}{\operatorname{argmax}} \left(R(s, a) + \sum\nolimits_{s' \in S} P(s' | a, s) \cdot U^{\pi^*}(s') \right)$$

ואילו התועלת האופטימלית מוגדרת כך:

$$U^{\pi^*}(s) = \max_{a \in \mathbb{A}(s)} \left(R(s, a) + \sum\nolimits_{s' \in S} P(s' | a, s) \cdot U^{\pi^*}(s') \right)$$

כיוון ש T הוא מצב טרמינלי מתקיים:

$$U^{\pi^*}(T) = \sum\nolimits_{a \in \mathbb{A}(T)} R(T, a) = R(T, B) + R(T, L) = 0$$

על מנת למצוא את a, b נחשב:

- S=3

$$\pi^*(3) = \underset{a\in\mathbb{A}(s)}{\operatorname{argmax}}\left(R(3,a) + \sum\nolimits_{s'\in S} P(s'|a,3)\cdot U^{\pi^*}(s')\right) \overset{\mathbb{A}(3)=\{L\}}{\triangleq} L$$

$$U^{\pi^*}(3) = \max_{a\in\mathbb{A}(s)}\left(R(3,a) + P(T|L,3)\cdot U^{\pi^*}(T)\right) = R(3,L) + P(T|L,3)\cdot U^{\pi^*}(T) = 6$$

- S=2

$$\pi^*(2) = \underset{a\in\mathbb{A}(s)}{\operatorname{argmax}}\left(R(2,a) + \sum\nolimits_{s'\in S} P(s'|a,2)\cdot U^{\pi^*}(s')\right) =$$

$$= \underset{a\in\mathbb{A}(s)}{\operatorname{argmax}}\left(R(2,B) + P(3|B,2)\cdot U^{\pi^*}(3) + P(T|B,2)\cdot U^{\pi^*}(T), R(2,L) + P(3|L,2)\cdot U^{\pi^*}(3) + P(T|L,2)\cdot U^{\pi^*}(T)\right) =$$

$$= \underset{a\in\mathbb{A}(s)}{\operatorname{argmax}}(6p + 0{,}0 + 4) = \begin{cases} L, & p < \frac{2}{3} \\ B, & otherwise \end{cases}$$

$$U^{\pi^*}(2) =$$

$$= \max_{a\in\mathbb{A}(s)}\left(R(2,B) + P(3|B,2)\cdot U^{\pi^*}(3) + P(T|B,2)\cdot U^{\pi^*}(T), R(2,L) + P(3|L,2)\cdot U^{\pi^*}(3) + P(T|L,2)\cdot U^{\pi^*}(T)\right) =$$

$$= \max(0 + 6p + 0{,}4 + 0 + 0) = \begin{cases} 4, & p < \frac{2}{3} \\ 6p, & otherwise \end{cases}$$

- S=1

$$\pi^*(1) = \underset{a\in\mathbb{A}(s)}{\operatorname{argmax}}\left(R(1,a) + \sum\nolimits_{s'\in S} P(s'|a,1)\cdot U^{\pi^*}(s')\right) =$$

$$= \underset{a\in\mathbb{A}(s)}{\operatorname{argmax}}\left(R(1,B) + P(2|B,1)\cdot U^{\pi*}(2) + P(T|B,1)\cdot U^{\pi^*}(T), R(1,L) + P(2|L,1)\cdot U^{\pi*}(2) + P(T|L,1)\cdot U^{\pi^*}(T)\right) =$$

$$= \underset{a\in\mathbb{A}(s)}{\operatorname{argmax}}(0 + p\cdot(4+6p) + 0, 2 + 0 + 0) = \begin{cases} L, & p < \frac{1}{2} \\ B, & otherwise \end{cases}$$

$$U^{\pi^*}(1) = \max_{a\in\mathbb{A}(s)}\left(R(1,B) + P(2|B,1)\cdot U^{\pi^*}(2) + P(T|B,1)\cdot U^{\pi^*}(T), R(1,L) + P(2|L,1)\cdot U^{\pi^*}(2) + P(T|L,1)\cdot U^{\pi^*}(T)\right) =$$

$$= \begin{cases} \max(0 + 4p + 0{,}2 + 0 + 0), & p < \frac{2}{3} \\ \max(0 + 6p^2 + 0, 2 + 0 + 0), & otherwise \end{cases} = \begin{cases} 2, & p < \frac{1}{2} \\ 4\text{p}, & \frac{1}{2} < p < \frac{2}{3} \\ 6\text{p}^2, & otherwise \end{cases}$$

- S=0

$$\pi^*(0) = \underset{a\in\mathbb{A}(s)}{\operatorname{argmax}}\left(R(0,a) + \sum\nolimits_{s'\in S} P(s'|a,0)\cdot U^{\pi^*}(s')\right) =$$

$$= \underset{a\in\mathbb{A}(s)}{\operatorname{argmax}}\left(R(0,B) + P(1|B,0)\cdot U^{\pi*}(1) + P(T|B,0)\cdot U^{\pi^*}(T), R(0,L) + P(1|L,0)\cdot U^{\pi*}(1) + P(T|L,0)\cdot U^{\pi^*}(T)\right) =$$

$$= \begin{cases} \underset{a\in\mathbb{A}(s)}{\operatorname{argmax}}(0 + 2p + 0, 0 + 0 + 0), & p < \frac{1}{2} \\ \underset{a\in\mathbb{A}(s)}{\operatorname{argmax}}(0 + 4p^2 + 0, 0 + 0 + 0), & \frac{1}{2} < p < \frac{2}{3} \\ \underset{a\in\mathbb{A}(s)}{\operatorname{argmax}}(0 + 6p^3 + 0, 0 + 0 + 0), & otherwise \end{cases} = B$$

$$U^{\pi^*}(1) = \max_{a\in\mathbb{A}(s)}\left(R(0,B) + P(1|B,0)\cdot U^{\pi^*}(1) + P(T|B,0)\cdot U^{\pi^*}(T), R(0,L) + P(1|L,0)\cdot U^{\pi^*}(1) + P(T|L,0)\cdot U^{\pi^*}(T)\right) =$$

$$= \begin{cases} \max(0 + 2p + 0{,}0 + 0 + 0), & p < \frac{1}{2} \\ \max(0 + 4p^2 + 0{,}2 + 0 + 0), & \frac{1}{2} < p < \frac{2}{3} \\ \max(0 + 6p^3 + 0{,}2 + 0 + 0), & otherwise \end{cases} = \begin{cases} 2\text{p}, & p < \frac{1}{2} \\ 4\text{p}^2, & \frac{1}{2} < p < \frac{2}{3} \\ 6\text{p}^3, & otherwise \end{cases}$$

לכן מצאנו כי $a = \frac{1}{2}, b = \frac{2}{3}$.

נמלא את הטבלה כפי שמצאנו.

תועלות	מדיניות	ערכי p
$V^{\pi_1}(0) = 2p$	$\pi_1(0) = B$ $\pi_1(1) = L$ $\pi_1(2) = L$ $\pi_1(3) = L$	$0 < p < a$
$V^{\pi_2}(0) = 4p^2$	$\pi_2(0) = B$ $\pi_2(1) = B$ $\pi_2(2) = L$ $\pi_2(3) = L$	$a < p < b$
$V^{\pi_3}(0) = 6p^3$	$\pi_3(0) = B$ $\pi_3(1) = B$ $\pi_3(2) = B$ $\pi_3(3) = L$	$b < p < 1$

חלק ב' - היכרות עם הקוד

הכרנו.

חלק ג' - רטוב

- הרצת MC-algorithm :

לאחר הרצת האלגוריתם קיבלנו את התוצאות הבאות :

```
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@@@@    Reward matrix after 10 episodes    @@@@@@@
[[-0.40535053655877795 -0.24790982191738314 -0.1290573401071571 1.0]
 [-0.4034827501924484 None -0.7121016485221515 -1.0]
 [-0.4311555582885534 -0.6137157296822116 -0.7167045647738773
  -0.8338680652000001]]
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@

@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@@@    Reward matrix after 100 episodes    @@@@@@
[[-0.2743476501786998 -0.20927864237378604 0.09977869530627267 1.0]
 [-0.3600072379600251 None -0.4770938696860816 -1.0]
 [-0.4121217140204877 -0.48592468397863015 -0.5800697963377626
  -0.7110264730379243]]
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@

@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@@@    Reward matrix after 1000 episodes    @@@@@@
[[-0.28881981143370516 -0.1782581494140977 0.03303507962976523 1.0]
 [-0.3645060130669382 None -0.5139601033421183 -1.0]
 [-0.4109794548680666 -0.46353925157603043 -0.5367960206825328
  -0.7083068577637374]]
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
```

ניתן לראות שהתוצאת אינן זהות, ככל שמספר האפיסודות גדל כך התוצאה קרוב יותר לשיערוך האמיתי של המדינות.
ככל שיש יותר אפיסודות נדרשים יותר משאבים להרצת הסימולציה אך נוכל להגיע לשיערוך קרוב יותר למדיניות.
לעומת זאת עבור מספר קטן של אפיסדות נקבל תוצאה מהימנה פחות, אך נוכל לחסוך במשאבים.

- <u>נשלים את הנוסחה :</u>

$$V_{\infty}^{\pi}(s_0) = \frac{1}{N} \cdot \sum_{i=1}^{N} \sum_{t=0}^{\infty} \gamma^t R\left(s_t^i\right)$$

נשים לב שכל הסימולציות מתחילות ב s_0 , כלומר

$$\forall i: s_0 = s_0^i$$

- <u>נמצא חסם תחתון ל-T :</u>

ראשית נמצא חסם עליון הדוק ל- $|V_{\infty}^{\pi}(s_0) - V_T^{\pi}(s_0)|$:

$$|\mathrm{V}_{\infty}^{\pi}(\mathrm{s}_0) - \mathrm{V}_{\mathrm{T}}^{\pi}(\mathrm{s}_0)| = \left|\frac{1}{\mathrm{N}}\sum_{i=1}^{N}\sum_{t=0}^{\infty}\left(\gamma^t r\left(s_t^i\right)\right) - \frac{1}{\mathrm{N}}\sum_{i=1}^{N}\sum_{t=0}^{T}\left(\gamma^t r\left(s_t^i\right)\right)\right| =$$

$$\frac{1}{\mathrm{N}}\left|\sum_{i=1}^{N}\left(\sum_{t=0}^{\infty}\left(\gamma^t r\left(s_t^i\right)\right) - \sum_{t=0}^{T}\left(\gamma^t r\left(s_t^i\right)\right)\right)\right| = \left|\sum_{t=0}^{\infty}\left(\gamma^t r\left(s_t^i\right)\right) - \sum_{t=0}^{T}\left(\gamma^t r\left(s_t^i\right)\right)\right| =$$

$$\left|\sum_{t=T+1}^{\infty}\left(\gamma^t r\left(s_t^i\right)\right)\right| \le \left|\sum_{t=T+1}^{\infty}\left(\gamma^t R_{max}\right)\right| = R_{max} \cdot \sum_{t=T+1}^{\infty}\left(\gamma^t\right) = R_{max} \cdot \frac{\gamma^{T+1}}{1-\gamma}$$

כעת נמצא חסם תחתון הדוק עבור T ולכן נדרוש : $R_{max} \cdot \frac{\gamma^{T+1}}{1-\gamma} \le \epsilon$, ונחשב :

$$R_{max} \cdot \frac{\gamma^{T+1}}{1-\gamma} \le \epsilon \Longrightarrow \gamma^{T+1} \le \frac{\epsilon(1-\gamma)}{R_{max}} \Longrightarrow T \ge \log_{\gamma}\left(\frac{\epsilon(1-\gamma)}{R_{max}}\right) - 1$$

ולכן עבור $T \ge \log_{\gamma}\left(\frac{\epsilon(1-\gamma)}{R_{max}}\right) - 1$ נקבל :

$$|\mathrm{V}_{\mathrm{T}}^{\pi}(\mathrm{s}_0) - \mathrm{V}_{\pi}(\mathrm{s}_0)| = |\mathrm{V}_{\mathrm{T}}^{\pi}(\mathrm{s}_0) - \mathrm{V}_{\infty}^{\pi}(\mathrm{s}_0) + \mathrm{V}_{\infty}^{\pi}(\mathrm{s}_0) - \mathrm{V}_{\pi}(\mathrm{s}_0)| \le$$

$$|\mathrm{V}_{\mathrm{T}}^{\pi}(\mathrm{s}_0) - \mathrm{V}_{\infty}^{\pi}(\mathrm{s}_0)| + |\mathrm{V}_{\infty}^{\pi}(\mathrm{s}_0) - \mathrm{V}_{\pi}(\mathrm{s}_0)| \cong |\mathrm{V}_{\mathrm{T}}^{\pi}(\mathrm{s}_0) - \mathrm{V}_{\infty}^{\pi}(\mathrm{s}_0)| \le$$

$$R_{max} \cdot \frac{\gamma^{T+1}}{1-\gamma} \le \epsilon$$

- <u>אלגוריתם First-Visit Monte Carlo Anytime :</u>

נבנה אלגוריתם כך שהוא יעבוד בגישה איטרטיבית שבה הוא מתחיל עם פרמטר זמן התחלתי T=1 .

עבור כל צעד בסימולציה, האלגוריתם יריץ הרצת - First-Visit Monte Carlo ויעדכן את וקטור התועלות V_T^{π} .

התהליך הזה יימשך עד שיגמר הזמן שהוקצה לאלגוריתם.

אם הזמן הסתיים נחזיר את הערך עבור האיטרציה ה-T, כלומר האלגוריתם מחזיר את V_T^{π} .

אם הזמן נגמר באמצע איטרציה, הוא יחזיר את הוקטור מהאיטרציה הקודמת ת V_{T-1}^{π} .

אם האיטרציה שהוחזרה מקיימת : $T \ge \log_{\gamma}\left(\frac{\epsilon(1-\gamma)}{R_{max}}\right) - 1$ אז נדע $|\mathrm{V}_{\mathrm{T}}^{\pi}(\mathrm{s}_0) - \mathrm{V}_{\pi}(\mathrm{s}_0)| \le \epsilon$ כלומר הוחזר ערך קרוב מספיק לשיערוך האמיתי.

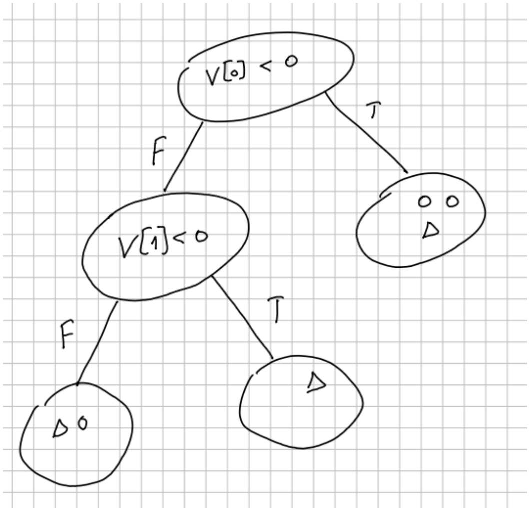
מכיוון שהאלגוריתם ממשיך לשפר את התוצאות ככל שניתן לו יותר זמן, הוא נחשב אלגוריתם Anytime.

חלק ב' – מבוא ללמידה

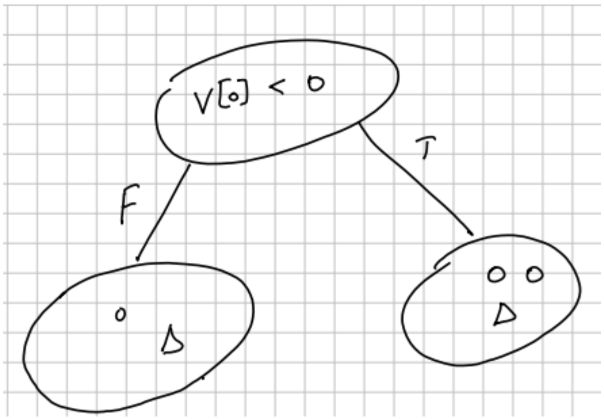
חלק א - יבש

- "מתפצלים ונהנים"

עבור הגרף T:



מתקבל הגרף T':



נגדיר: משולש – TRUE, עיגול – FALSE.

נניח בשלילה שקיים וקטור $\epsilon = \{\epsilon_1, \epsilon_2\} \in \mathbb{R}^2, \epsilon_i > 0$ כך שלכל דוגמת מבחן $x \in \mathbb{R}^2$ מתקיים שתוצאת ריצת הדוגמא בעץ T' זהה לתוצאת ריצת ϵ של x על T. נבחר דוגמת מבחן $x = \{\frac{\epsilon_1}{2}, -\frac{\epsilon_2}{2}\}$.

מתקיים $|x_i - v_i| = |x_i| < \epsilon_i$ ולכן ריצת ϵ של x על T תתפצל בכל צומת ותחזיר 3 עיגולים ו-2 משולשים כלומר עיגול (FALSE). בנוסף, ריצת x על עץ T' תחזיר עיגול ומשולש (כי $x_1 = \frac{\epsilon_1}{2} > 0$ כלומר תפנה שמאלה בצומת הראשון) ולכן תחזיר משולש (TRUE , כי הוגדר שבעת שיוויון חוזר TRUE).

קיבלנו שתוצאת הריצות אינה זהה ולכן מצאנו דוגמא נגדית לטענה.

חלק ב' - היכרות עם הקוד

הכרנו.

חלק ג' - חלק רטוב ID3

שאלה 1

מומש בקוד.

שאלה 2

לאחר שמימשנו את basic_experiment קיבלנו כי תוצאת דיוק המודל שלנו היא:

```
Test Accuracy: 96.46%
```