# (044137) מבוא ללמידה MDP | תרגיל בית 3

## חלק א'

1. תגמול על הקשתות:

(א) תוחלת על התועלת עבור מקרה של תגמול בקשתות:

$$U^{\pi}(S) = E_{\pi}[\sum_{t=0}^{\infty} \gamma^t R(S_t, \pi[S_t], S_{t+1}) | S_0 = S]$$

(ב) משוואת בלמן עבור מקרה של תגמול בקשתות:

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

(ג) האלגוריתם עבור מקרה של תגמול על הקשתות:

**repeat**

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

**for each** state $s$ **in** $S$ **do**

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

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

**until** $\delta < \epsilon(1-\gamma)/\gamma$

**return** $U$

(ד) האלגוריתם עבור מקרה של תגמול על הקשתות:

**repeat**

$U \leftarrow$ POLICY-EVALUATION($\pi$, $U$, $mdp$)

$unchanged? \leftarrow$ true

**for each** state $s$ **in** $S$ **do**

**if** $\max_{a \in A(s)} \left[ \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma U(s')] \right] > \sum_{s'} P(s' \mid s, \pi[s]) \overbrace{[R(s,a,s') + \gamma U(s')]}$ **then do**

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

$unchanged? \leftarrow$ false

**until** $unchanged?$

**return** $\pi$

במידה ו $\gamma = 1$ והתועלות חיוביות לא קיימת למעשה מדיניות אופטימלית שכן מדיניות זו תמקסם את התועלת, אך במקרה זה התועלת אינה חסומה ועל כן לא קיים לה מקסימום. ככל שהסוכן ימשיך לשחק מכיוון שהתועלת חיובית ואין דעיכה בערכי התועלות הוא יעדיף לא לסיים את המשחק ולהמשיך ולהגדיל את התועלת שלו. לכן במצב זה על מנת שתמיד נצליח למצוא את המדיניות האופטימילית צריך להתקיים תנאים שמבטחים שהסוכן יסיים את המשחק, כלומר תועלת שלילית שתגרום לכך שלא משתלם לסוכן לשחק לאורך זמן אלא עדיף לו לסיים את המשחק. או מצב סופי ותתנאים במשחק שיבטיחו שהסוכן יגיע לאותו מצב סופי לבסוף. אם מתקיימים תנאים על הסביבה (כמו אלו שציינתי) שמבטיחים שכל מדינות תגיע למצב סופי כלשהו (כי תועלת חסומה) נוכל תמיד למצוא את המדיניות האופטימלית. בסעיף ג' תנאי העצירה ישתנה כך שהאלגוריתם יעצור כאשר $\delta = 0$.

(ה)

| | $U_0(S_i)$ | $U_1(S_i)$ | $U_2(S_i)$ | $U_3(S_i)$ | $U_4(S_i)$ | $U_5(S_i)$ | $U_6(S_i)$ |
|---|---|---|---|---|---|---|---|
| $S_1$ | 0 | 1- | 1- | 1- | 1- | 1- | 1- |
| $S_2$ | 0 | 1- | 2- | 2- | 2- | 2- | 2- |
| $S_3$ | 0 | 1- | 2- | 3- | 3- | 3- | 3- |
| $S_4$ | 0 | 1- | 2- | 3- | 4- | 4- | 4- |
| $S_5$ | 0 | 1- | 2- | 3- | 4- | 5- | 5- |
| $S_6$ | 0 | 1- | 2- | 3- | 4- | 4- | 4- |
| $S_7$ | 0 | 1- | 2- | 3- | 3- | 3- | 3- |

(ו)

| | $\pi_0(S_i)$ | $\pi_1(S_i)$ | $\pi_2(S_i)$ | $\pi_3(S_i)$ |
|---|---|---|---|---|
| $S_1$ | $\uparrow$ | $\uparrow$ | $\uparrow$ | $\uparrow$ |
| $S_2$ | $\uparrow$ | $\uparrow$ | $\uparrow$ | $\uparrow$ |
| $S_3$ | $\leftarrow$ | $\leftarrow$ | $\leftarrow$ | $\leftarrow$ |
| $S_4$ | $\uparrow$ | $\uparrow$ | $\uparrow$ | $\uparrow$ |
| $S_5$ | $\rightarrow$ | $\rightarrow$ | $\rightarrow$ | $\rightarrow$ |
| $S_6$ | $\rightarrow$ | $\rightarrow$ | $\uparrow$ | $\uparrow$ |
| $S_7$ | $\downarrow$ | $\rightarrow$ | $\rightarrow$ | $\rightarrow$ |

## חלק ג'

- סיבה אחת שמקרה זה ייתכן זה אם $\alpha = 0$ במצב זה כלל העידכון יתבסס כולו על העבר ומכיוון שכל המצבים אותחלו לאפס בהתחלה אז הם גם יישארו כך. מצב זה מתאר ריצה מנוונת שבה כל הערכים בטבלה יהיו עם 0.

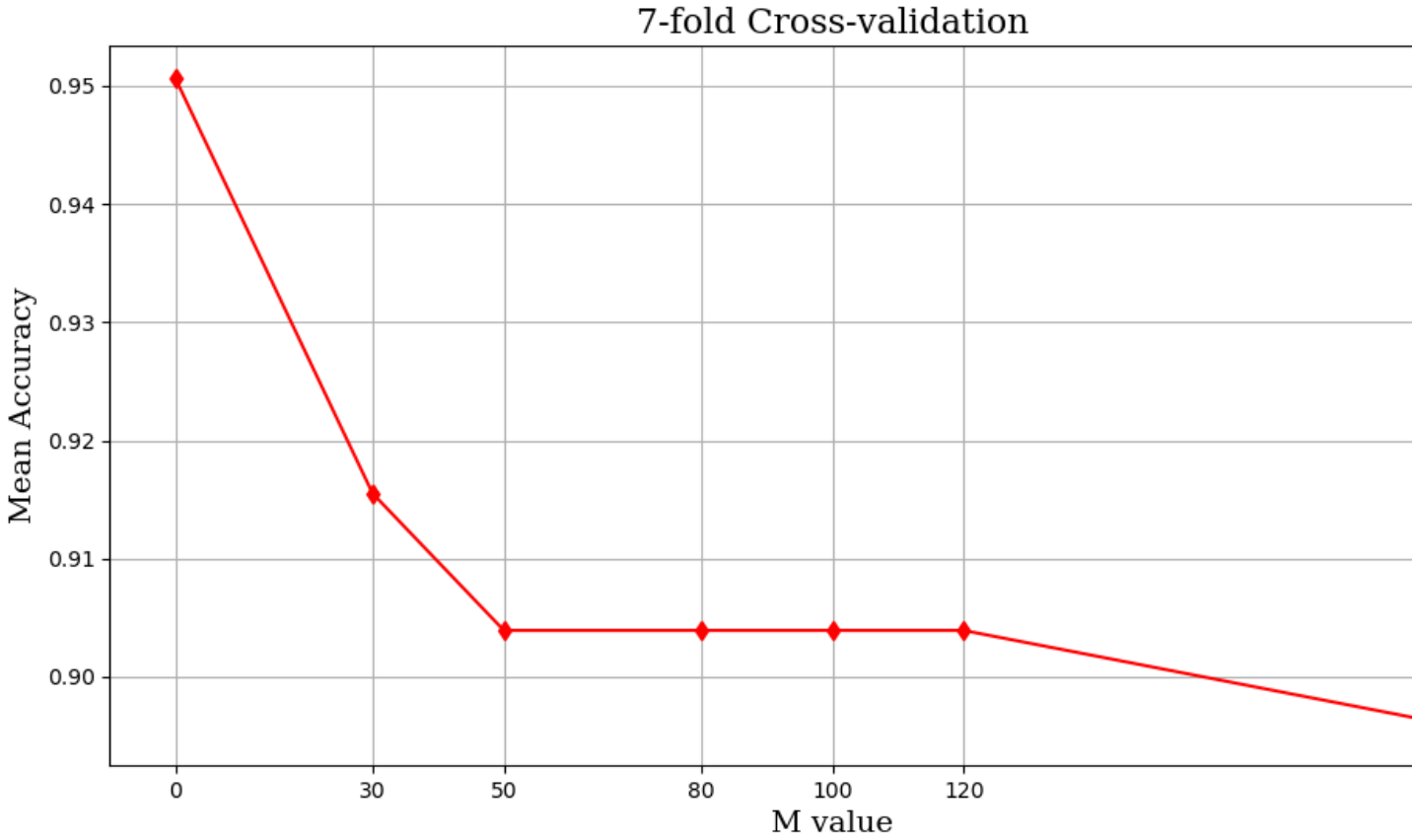
  סיבה נוספת היא מצב בו $\alpha \neq 0$ אך קיימים מצבים שכחלק מריצת האלגוריתם הוא לא עובר בהם, לכן כלל העידכון לא פועל על אותם מצבים ולמעשה הם לא מעודכנים ונשארים עם הערך 0 איתו הם אותחלו. זה עלול לקרות עבור מצבים שיש הסתברות נמוכה מאוד (או בכלל לא) להגיע לאליהם, או כי הלאגוריתם סיים את הריצה בלי להגיע לאותם מצבים (באף אחת מהאפיזודרות) בגלל הגבלת זמן למשל.

## חלק ג'

.6

.a המטרה של גיזום היא לשפר את יכולת ההכללה של המודל ובכך למנוע ( או להקטין) את תופעת ה overfitting. גיזום העץ גורם לכך שהוא לא יהיה עקבי בהכרח ולכן לא יתאים את עצמו בצורה מוחלטת לסט האימון שלנו שכולל גם דוגמאות רועשות ולא מייצגות. בסופו של דבר תוצאות המודל ימדדו על סט מבחן אחר שהמודל לא מכיר, כלומר החשיבות בהצלחה על סט המבחן גדולה יותר מהצלחה על סט האימון ועל ידי גיזום אנחנו מוכנים לשלם בשגיאת אימון גדולה יותר כדי לשפר יכולת הכללה במטרה להקטין את השגיאה על סט המבחן.

.c




- ניתן לראות בגרף שככל שM גדל כך הדיוק ההמומצע קטן, משום שהדיוק מחושב על סט האימון במקרה זה ובשל העובדה שככל שמגדלים את הפרמטר כך הגיזום מתבצע בצורה אגרסיבית יותר, זו תוצאה צפויה. עיקרון הפעולה של הגיזום הוא החלשה של מורכבות המודל מה שבא לידי ביטוי בדיוק נמוך יותר על סט האימון, כדי לשפר את היכולת הכללה של המודל. ככל ש M גדול יותר כך מתבצע גיזום בשלב מוקדם יותר בעץ ועל כן המודל פחות מורכב ומדייק פחות על סט האימון כפי שרואים בגרף בירידה בדיוק על הסט האימון.

.a הרצנו עם $M = 50$ וקיבלנו דיוק של 97.35% שזה שיפור לעומת הדיוק של 94.69% שקיבלנו ללא גיזום. ניתן לראות שכצפוי הגיזום הוביל לשיפור בביצועים