פתרון מטלה 6 — חישוביות וקוגניציה, 6119

24 בדצמבר 2025

שאלה 1

יצור חי מעל הגרף $(\{A,B,C,D,E\},\{(A,B),(B,C),(C,D),(D,E)\})$ יחד עם פעולה בינארית. הפעולה מזיזה קדימה או אחורה ביחס סדר הציון של הצמתים והגמול הוא אפס אלא אם בוצעה הפעולה a_1 ב־E, ערך ההנחה הוא $\gamma = 1$. נניח כי הייצור משתמש בהסתברות אחידה כרגע.

סעיף א'

נתאר את העולם בהגדרות MDP.

פתרון הסביבה היא צמתי הגרף, הפעולות $\mathcal{A} = \{a_0, a_1\}$, סיכויי המעבר $\mathbb{P}(s' \mid a, s) = \frac{1}{2}$ ופונקציית הגמול המיידית היא $r(s,a) = \mathbb{1}_{(E,a_1)}$. הסוכן הוא הייצור והמדיניות שלו היא כמתואר על־ידי פונקציית ההסתברות.

סעיף ב'

נמצא את ערכי המצבים בהינתן המדיניות של הסוכן $V_\pi(s)$.

פתרון נזכור כי מתקיים $V_\pi(C) = \mathbb{E}\left(\sum_{i=0}^{\infty} \gamma^i r_i \mid s_0 = C\right) = \mathbb{E}\left(\sum_{i=0}^{\infty} r_i \mid s_0 = C\right)$, אבל מטעמי סימטריה נוכל להסיק שבכל צעד הסיכוי שהסוכן יהיה במיקום $C + ca_i$ שווה לסיכוי ל־$C - ca_i$, ובהתאם בהכרח $V_\pi(C) = \frac{1}{2}(0+1)$ בלבד.

נשתמש במשוואת בלמן כדי לחשב את שאר המצבים,

$$V_\pi(B) = \sum_{i=0}^{1} \mathbb{P}(a_i \mid B)\left(r(B,a_i) + \gamma \sum_{s\in\{A,C\}} \mathbb{P}(s \mid B, a_i) V_\pi(s)\right) = \frac{1}{2}V_\pi(A) + \frac{1}{2}V_\pi(C) = \frac{1}{2}V_\pi(A) + \frac{1}{4}$$

באופן דומה נקבל שגם,

$$V_\pi(A) = \frac{1}{2}V_\pi(B) + 0.$$

ולכן,

$$V_\pi(B) = \frac{1}{4}V_\pi(B) + \frac{1}{4} \iff V_\pi(B) = \frac{1}{3}.$$

נבחין כי גם נובע,

$$V_\pi(C) = \frac{1}{2}V_\pi(B) + \frac{1}{2}V_\pi(D) \implies \frac{1}{2} = \frac{1}{6} + \frac{1}{2}V_\pi(D) \implies V_\pi(D) = \frac{2}{3}$$

ולבסוף,

$$V_\pi(E) = \frac{1}{2}V_\pi(D) + \frac{1}{2} = \frac{5}{6}$$

וקיבלנו את מפת הערך,

$$\left[\frac{1}{6}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{5}{6}\right]$$

קיבלנו תוצאה סימטרית באופן שמתכתב עם הסימטריה באסטרטגיה.

משמעות התוצאה היא שככל שהייצור עומד רחוק יותר מהריבוע השמאלי, כך הסיכוי שלבסוף הוא יגיע אליו הוא נמוך יותר, ולכן הסיכוי שהוא יקבל גמול חיובי נמוך אף הוא.

סעיף ג'

נחשב את הליך הלימוד של הייצור בשיטת TD כאשר קצב הלימוד שלו הוא $\eta = 0.1$ וכאשר הייצור עשה שני ניסויים שבשניהם פעל ב־a_1, a_1, a_1.

פתרון נחשב בעזרת טבלת מעקב. נשים לב שלאורך כל הניסוי הראשון הייצור לא לומד עד השלב האחרון שכן רק בשלב האחרון מתקבל גם גמול חיובי וגם V חיובי.

trial	step	location	$V_\pi(A)$	$V_\pi(B)$	$V_\pi(C)$	$V_\pi(D)$	$V_\pi(E)$
1	0	C	0	0	0	0	0
1	1	D	0	0	0	0	0
1	2	E	0	0	0	0	0
1	3	end	0	0	0	0	0.2
2	0	C	0	0	0	0	0.2
2	1	D	0	0	0	0	0.2
2	2	E	0	0	0	0.02	0.2
2	3	end	0	0	0	0.02	0.38

סעיף ד'

מצורף לשאלה גרף המתאר את הליך הלימוד של הייצור בשיטת TD עבור 100 הרצות, נבין מה היו הערכים ההתחלתיים של כל אחד מהמצבים ונבין איפה הסתיים הניסוי הראשון.

פתרון בגרף נראה שהניסוי האפס מתואר על־ידי מגמה קבועה ב־$\frac{1}{2}$, כלומר $V(s) = \frac{1}{2}$ לכל $s \in \{A, B, C, D, E\}$.

לפי הגרף לאחר הניסוי הראשון (ולפני השני) הערך של $\{B, C, D, E\}$ נשאר זהה, ולכן נסיק שלא השתנה, ובהתאם בעזרת האפקט שאנו רואים בסוף הניסוי הראשון בטבלה שחישבנו נוכל להסיק שהייצור לא הגיע לצד ימין בניסוי הראשון. נבחין כי גם מהגרף נתון ש־$\frac{2}{5} < V(A) < \frac{1}{2}$, ולכן נוכל להסיק שאכן היה שינוי רק במצב A, כלומר הייצור הגיע אליו והמשיך הלאה ובכך למד שאין גמול בביצוע מסלול זה.

שאלה 2

נתון MPD בעל שני המצבים Home, Out ושתי הפעולות Stay, Switch כשמתקיים $\mathbb{P}(x \mid x, \text{Stay}) = 1$ וכן $\mathbb{P}(H \mid H, \text{Switch}) = \frac{1}{5}, \mathbb{P}(H \mid O, \text{Switch}) = 1$. הגבול מוגדר חד־ערכית על־ידי $r(H, \text{Stay}) = 0, r(H, \text{Switch}) = 1, r(O, \text{Stay}) = 2, r(O, \text{Switch}) = 0$. פרמטר ההנחה הוא $\gamma = \frac{1}{2}$.

סעיף א'

יהי סוכן אקראי מתפלג אחיד, נכתוב את משוואות בלמן ונפתור אותן במטרה לחשב את V_π.

פתרון משוואת בלמן הכללית במקרה שלנו היא,

$$\begin{aligned} V_\pi(s) &= \sum_{a \in \{\text{Stay,Switch}\}} \mathbb{P}(a \mid X) \left(r(s,a) + \gamma \sum_{s' \in \{O,H\}} \mathbb{P}(s' \mid s,a) V_\pi(s') \right) \\ &= \frac{1}{2} \sum_{a \in \{\text{Stay,Switch}\}} r(s,a) + \frac{1}{2} \sum_{s' \in \{O,H\}} \mathbb{P}(s' \mid s,a) V_\pi(s') \end{aligned}$$

נציב ערכים בהתאם,

$$V_\pi(H) = \frac{1}{2}(0 + \frac{1}{2}(0 + 1 \cdot V_\pi(H)) + 1 + \frac{1}{2}(0.2 \cdot V_\pi(H) + 0.8 V_\pi(O))) = 0.5 + 0.3 V_\pi(H) + 0.2 V_\pi(O)$$

וכן מחישוב דומה,

$$V_\pi(O) = \frac{1}{2}(2 + \frac{1}{2}(1 \cdot V_\pi(O) + 0 \cdot V_\pi(H)) + 0 + \frac{1}{2}(1 \cdot V_\pi(H) + 0)) = 1 + 0.25 V_\pi(O) + 0.25 V_\pi(H)$$

מהעברת אגפים נסיק,

$$0.7 V_\pi(H) = 0.5 + 0.2 V_\pi(O), \quad 0.75 V_\pi(O) = 1 + 0.25 V_\pi(H)$$

נציב את המשוואה השנייה בראשונה,

$$\frac{7}{10} V_\pi(H) = \frac{1}{2} + \frac{4}{15}(1 + \frac{1}{4} V_\pi(H)) = \frac{23}{30} + \frac{1}{15} V_\pi(H) \implies \frac{19}{30} V_\pi(H) = \frac{23}{30} \implies V_\pi(H) = \frac{23}{19} \approx 1.21$$

בהתאם,

$$V_\pi(O) = \frac{4}{3}(1 + \frac{1}{4} \cdot \frac{23}{19}) = \frac{33}{19} \approx 1.736$$

סעיף ב'

ננסה לנחש את המדיניות האופטימלית.

פתרון ננחש שהמדיניות היא לנסות להחליף במצב של H ולהישאר במצב של O.

סעיף ג'

נבדוק אם המדיניות שניחשנו מקיימת את משוואת האופטימליות של בלמן.

פתרון המשוואה היא,

$$V^*(s) = \max \left\{ r(s,a) + \gamma \sum_{s'} \mathbb{P}(s' \mid s,a) V^*(s') \mid a \right\}$$

נציב ונקבל,

$$V^*(O) = \max \left\{ 2 + \frac{1}{2}(1 \cdot V^*(O)), 0 + \frac{1}{2}(1 \cdot V^*(H)) \right\} = 2 + \frac{1}{2} V^*(O) \implies V^*(O) = 4$$

וכן,

$$
\begin{aligned}
V^*(H) &= \max\left\{0+\frac{1}{2}(1\cdot V^*(H)+0\cdot V^*(O)), 1+\frac{1}{2}(0.2V^*(H)+0.8V^*(O))\right\} \\
&= \max\left\{\frac{1}{2}V^*(H), 1+0.1V^*(H)+0.4\cdot 4\right\} \\
&= 1+0.1V^*(H)+0.4\cdot 4
\end{aligned}
$$

ולכן $0.9V^*(H)=2.6 \implies V^*(H)=\frac{26}{9}\approx 2.888$.

נובע אם כך שהאסטרטגיה שבחרנו אכן מקבלת ערכים מקסימליים.

סעיף ד'

נממש את אלגוריתם Value Iteration עבור הבעיה ונמצא את ערכי V^* בהתאם לאלגוריתם.

פתרון הקוד נכתב והורץ בקובץ המצורף, ותוצאותיו הן,

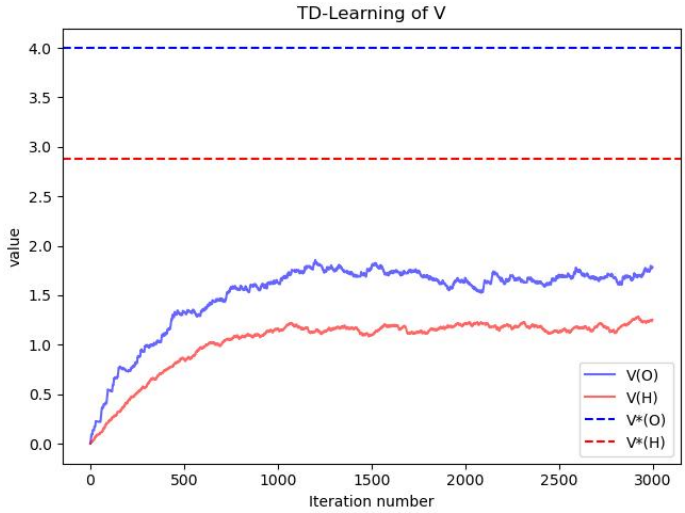
$$V^*(O)=4, \quad V^*(H)=3.333$$

נבחין כי הערכים שהתקבלו קרובים מאוד לערכים שנמצאו בסעיף הקודם, למעשה הערך $V^*(O)$ זהה לגמרי, בעוד יש פער ב־$V^*(H)$ שכנראה נובע מטעות חישוב.

סעיף ה'

נריץ את אלגוריתם הלימוד TD-Learning כדי ללמוד את V^π עבור המדיניות של סוכן אקראי בהתפלגות אחידה. נשתמש בקבוע הלימוד $\eta=0.01$ ונבצע $T=3000$ סבבים.

פתרון נריץ את האלגוריתם ונקבל את הגרף הבא המתאר של מהלך הריצה והשינוי ב־V המחושב לצד הערכים אשר קיבלנו בסעיפים הקודמים.

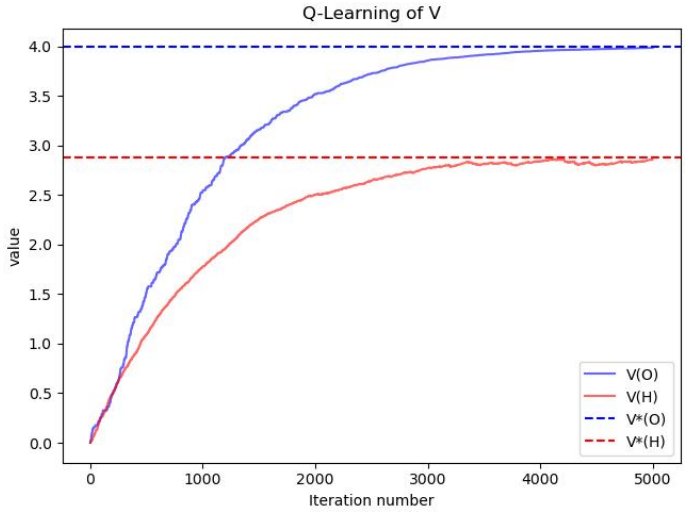


נבחין כי בעוד היחס בין V ל־V^* הוא דומה, הלמידה לא מתכנסת לערכים שמצאנו. לפי גודלי הערכים נשער שהסיבה לכך היא שאלגוריתם הלימוד מבצע איזושהי נורמליזציה על הערכים, ומשמר רק את היחסים ביניהם.

סעיף ו'

נשתמש הפעם באלגוריתם Q-Learning במטרה לחשב את המדיניות האופטימלית על־ידי שימוש בסוכן אקראי תוך הגדרה $\eta=0.01, T=5000$.

פתרון לאחר חישוב נקבל את מהלך הלמידה הבא,



נשים לב כי הפעם הערכים מתכנסים בדיוק, כלומר למדנו בהצלחה את המדיניות האופטימלית.