מבוא למערכות לומדות
תרגיל 6

רן שחם - 203781000 - ransha

6 ביולי 2017

1 בעיית המסלולים הקצרים ביותר

1.1

יהי $G=(U,E)$ גרף מכוון עם משקולות $w:E\to\mathbb{R}$ וקדקוד מקור $u\in U$ כך שכל המעגלים ב־G הם עם משקל חיובי.

נגדיר: $\mathcal{S}=\mathcal{A}=U$ ולכל $s,a\in U$ נגדיר: $\rho(s,a)=\begin{cases}-w(a,s), & (a,s)\in E\\ 0, & a=s=u\\ -\infty, & \text{otherwise}\end{cases}$, כלומר ה־reward הוא מספר סופי אם נשארים ב־u או אם אפשר ללכת מ־a ל־s (כלומר הצלע $a\to s$ קיימת - או באופן שקול, אפשר ללכת בגרף ההפוך מ־s ל־a), והוא שווה למינוס משקל הצלע הזו. בנוסף, נגדיר $\tau(s,a)=a$ לכל $s,a\in U$.[1]

מדיניות π על המרחב הנ"ל היא סדרת קדקודים $(v_0,v_1,v_2,\dots)$ כך שהסוכן מתחיל בקדקוד v_0 ועובר בין קדקודים. נשים לב שמתקיים $V_\pi(s)=\mathbb{E}\left[\sum_{t=0}^{\infty}r(s_t,a_t)|s_0=s\right]=\sum_{t=0}^{\infty}\rho(s_t,a_t)$ לכל π ו־$s\in U$, שכן ρ,τ הן דטרמיניסטיות.

כעת, עבור π כלשהי וקדקוד $s\in U$, ברור שאם π הוא הילוך בגרף המסתיים ב־u (ואחריו ההילוך תמיד נשאר ב־u) מקיים $V_\pi(s)\neq-\infty$ - כי מתקיים $V_\pi(s)=\sum_{i=0}^{N}\rho(v_i,v_{i+1})=-\sum_{i=0}^{N}w(v_{i+1},v_i)>-\infty$.

בכיוון השני, נניח ש־$V_\pi(s)\neq-\infty$. ברור שההילוך האינסופי המוגדר לעיל (סדרת הקדקודים $v_0,v_1,\dots$) מקיים $(v_{i+1},v_i)\in E$ לכל $i\in\mathbb{N}$ - אחרת היה מתקיים $\rho(v_i,v_{i+1})=-\infty$ עבור i כלשהו ולכן הסכום המוגדר ב־$V_\pi(s)$ היה $-\infty$[2] בסתירה להנחה שהוא לא. אם כך סדרת הקדקודים מתאימה להילוך בגרף ההפוך כנדרש. נראה שסדרה זו חייבת להסתיים ב־u (ולהישאר שם), כלומר שקיים $N\in\mathbb{N}$ כך ש־$v_n=u$ לכל $n\geq N$. בשלילה שלא, אז קיימים אינסוף $n\in\mathbb{N}$ כך ש־$v_n\neq u$. מכאן ומעקרון שובך היונים, קיימים אינסוף מעגלים בהילוך $v_0,v_1\dots$. נבדוק את תרומתם ל־$V_\pi(s)$. נתבונן במעגל כלשהו $v_i,v_{i+1},\dots,v_{i+k}=v_i$. מתקיים ש־$\sum_{j=0}^{k-1}w(v_{i+k-j},v_{i+k-j-1})>0$ מההנחה שכל המעגלים חיוביים, ולכן $\sum_{j=0}^{k-1}\rho(v_{i+j},v_{i+j+1})=-\sum_{j=0}^{k-1}w(v_{i+j+1},v_{i+j})<0$, כלומר כל מעגל תורם ערך שלילי לסכום, ומכך שיש אינסוף כאלה נקבל ש־$V_\pi(s)=-\infty$ בסתירה. לכן קיימת נקודה שהחל ממנה ההילוך נשאר ב־u.

נתבונן בתת־הסדרה $v_0,\dots,v_N$ כך ש־$v_N=u$ ולכל $n\geq N$: $v_n=u$. כאמור, היא מתאימה להילוך בגרף ההפוך, כלומר מתאימה למסלול מ־u ל־s. כמובן מתקיים $V_\pi(s)=\sum_{i=0}^{N-1}\rho(v_i,v_{i+1})=-\sum_{i=0}^{N-1}w(v_{i+1},v_i)$ שזהו סכום המסלול הנ"ל מ־u ל־s. מכאן נסיק שהפונקצייה $V^*(s)$ מתאימה למינוס ערכו של המסלול הקצר ביותר מ־u ל־s, שכן היא ממקסמת את ערך הסכום השווה למינוס משקל המסלול.

[1] פורמלית, τ מחזיר התפלגות על המצבים. התבקשנו לספק פונקציה דטרמיניסטית, אז יש להסתכל על $\tau(s,a)$ כך: $\tau(s,a)(\{b\})=\begin{cases}1, & b=a\\ 0, & \text{otherwise}\end{cases}$ לכל $b\in U$.

[2] כי כל הערכים האפשריים ל־ρ הם אי־חיוביים והם היחידים שנסכמים

1.2

Value Iteration	אלגוריתם 1

$\forall s\in\mathcal{S}$ set $V_0\left(s\right)=-\infty$
Set $V_0\left(u\right)=0$
While V has not converged
$\quad V_{t+1}\left(s\right)=\max_{a\in\mathcal{A}}\rho\left(s,a\right)+\mathbb{E}_{s'\sim\tau(s,a)}\left[V_t\left(s'\right)\right]\ \forall s\in\mathcal{S}$

לא הספקתי לסיים ☹

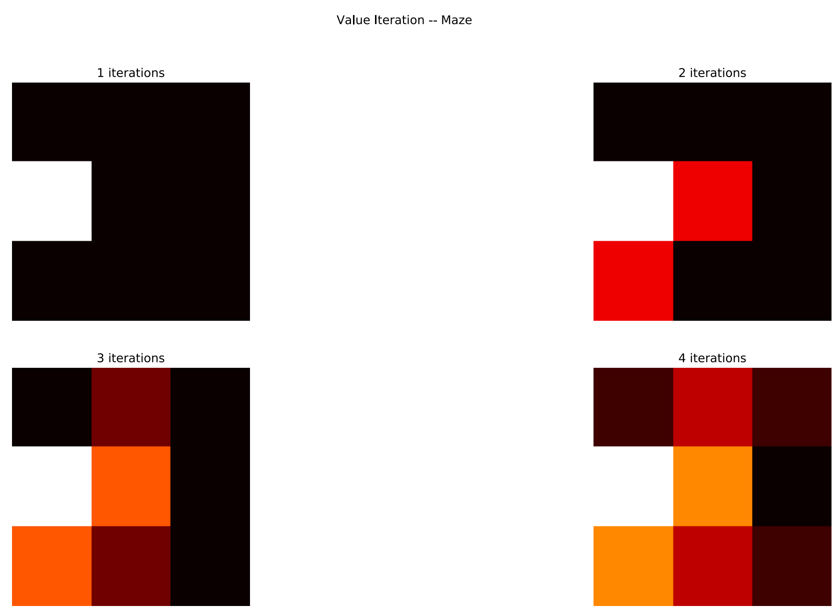
2 מבוך

האלגוריתם התכנס לערכים הבאים:

$$V\approx\begin{bmatrix}-5.186 & -4.013 & -5.186\\ 0 & -2.353 & -6.013\\ -2.353 & -4.013 & -5.186\end{bmatrix}$$

נשים לב שערכה של כל משבצת הוא נמוך יותר ככל שהמשבצת רחוקה מהמצב s_f - וכל 2 משבצות שנמצאות במרחק שווה מתכנסות לערך זהה.

מהרצת האלגוריתם עם מספר איטרציות שונה קיבלנו את התמונות הבאות:



מספר הצעדים הנדרש על מנת להגיע מכל משבצת ל-s_f במבוך הוא 4, לכן לאחר 4 איטרציות האלגוריתם יכול לחשב את הערך "האמיתי" של כל משבצת.

3 סבלנות

אלו ערכי V שהתקבלו מהרצת האלגוריתם עבור ערכי γ שונים:

$$V_{\gamma=0.5} \approx \begin{bmatrix} 6.99 & 3.49 & 1.749 & 0.99 & 1.99 \end{bmatrix}$$
$$V_{\gamma=0.75} \approx \begin{bmatrix} 8.99 & 6.749 & 5.062 & 3.79 & 3.99 \end{bmatrix}$$
$$V_{\gamma=0.85} \approx \begin{bmatrix} 11.66 & 9.916 & 8.429 & 7.164 & 6.66 \end{bmatrix}$$

כאשר ישנם 5 מצבים (השניים הקיצוניים לא באמת קיימים כי לא מגיעים אליהם, אלא חוזרים למצב ההתחלתי) והמצב ההתחלתי הוא הימני ביותר.

ההתנהגות האופטילית עבור $\gamma = 0.5$ היא ללכת ימינה שכן פעולה זו תביא את הערך $V\left[s_0\right] = 1.99$ לעומת הליכה שמאלה, שתביא את הערך 0.99. אם כך, הסוכן יעדיף תמיד ללכת ימינה מ־s_0. גם מהמצב השכן ל־s_0 (משמאלו) הסוכן יעדיף לצעוד ימינה כי שם ה־V גדול יותר. מהמצב הבא אחריו הסוכן כבר יעדיף ללכת שמאלה (אם כי הוא לא יגיע לשם, בהנחה שהוא מתחיל ב־s_0).

עבור $\gamma = 0.75$ ההתנהגות היא זהה מ־s_0 (כי $3.99 > 3.79$), אם כי הפער בין הערכים קטן יותר. לעומת זאת, מהשכן השמאלי של s_0 הסוכן כבר יעדיף ללכת שמאלה, וכך גם עבור כל מצב אחר. כלומר הסוכן עדיין ילך ימינה מיד אם יתחיל ב־s_0, אבל אם יגיע איכשהו לכל מצב אחר הוא יעדיף ללכת שמאלה.

עבור $\gamma = 0.85$ הסוכן כבר יעדיף ללכת שמאלה (מכל מצב), כלומר הוא למד שהגמול המירבי מתקבל מהליכה שמאלה.

מצורף הגרף עבור הסעיף האחרון: