להגשה עד:
04/05/2021

חישוביות וקוגניציה
מרצה: ד"ר אורן שריקי, מתרגל: אביב דותן

דף תרגיל 4

1. (25 נק') ניר, סטודנט בקורס חישוביות וקוגניציה, עומד בכל יום בפני החלטה הרת גורל: האם ללכת לבריכה (b) או לנסות לפתור את שיעורי הבית (h). הסיכוי שניר ינסה לפתור את שיעורי הבית, p_h, נקבע באופן הבא:

$$p_h = g(w)\,, \qquad g(x) \equiv \frac{1}{1+e^{-x}}$$

כאשר w הוא פרמטר פנימי של ניר.

א. בהנחה שניר לומד באמצעות אלגוריתם REINFORCE, כתבו כלל עדכון on-line לפרמטר w כתלות בפעולה שנבחרה h/b, קצב למידה η וגמול רגעי התלוי בפעולה r_a.

נניח שהגמול מתפלג באופן הבא:

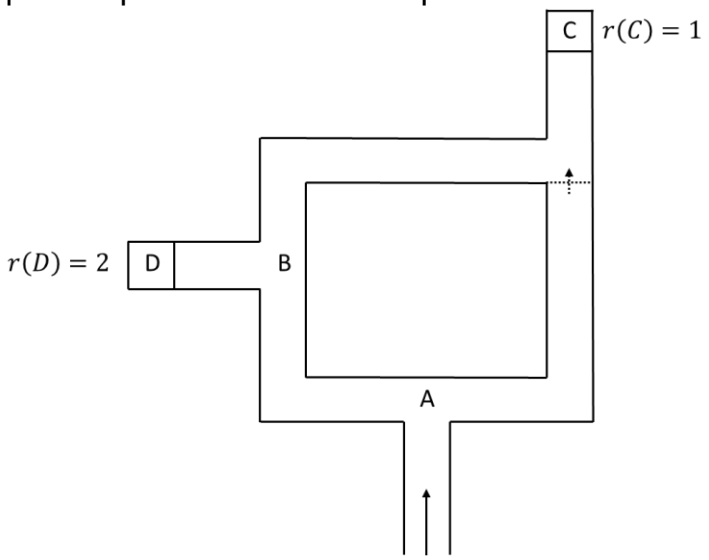
$$r_h = \begin{cases} 2\,, p = 0.75 \\ 0\,, p = 0.25 \end{cases}\,, \qquad r_b = \begin{cases} \dfrac{1}{1-g(w)}\,, p = 0.75 \\ 0 \qquad\quad , p = 0.25 \end{cases}$$

ב. האם האלגוריתם מסעיף א' יתכנס בממוצע? אם כן – חשבו את הגמול הממוצע עבור w שאליו יתכנס האלגוריתם, אם לא – מדוע?

ג. מהו הגמול הממוצע המקסימלי? האם קיים ערך של w עבורו מתקבל גמול זה? הסבירו.

ד. עבור כל אחד מהמקרים בסעיפים ב' ו-ג', קבעו מהי ההסתברות שבה ניר ינסה לפתור את שיעורי הבית. האם אלגוריתם REINFORCE מתכנס לפתרון האופטימלי?

2. (25 נק') רובוט מנסה ללמוד לנווט במבוך המתואר בתרשים ולאסוף גמול מקסימלי:



לשם כך, הרובוט משתמש באלגוריתם actor-critic, כפי שנלמד בשיעור.

א. קבעו בעזרת שיקולים פשוטים, מהי המדיניות האופטימלית. חשבו את ערכי המבקר $v(u)$ בכל מצב ואת ערכי השחקן $m_a(u)$ לכל מצב ופעולה עבור המדיניות שבחרתם. (הניחו כי פקטור הדעיכה הוא 1.)

ב. כעת הרובוט משתמש בפקטור דעיכה $\gamma = \frac{1}{2}$.

i. התאימו את כללי הלמידה להערכת המדיניות ולשיפורה שנלמדו בשיעור כך שיכללו את פקטור הדעיכה.

ii. כיצד יראו כללי הלמידה בממוצע על פני כל הפעולות האפשריות?

iii. בצעו שתי איטרציות למידה של אלגוריתם actor-critic באמצעות כללי הלמידה הממוצעים על פני הפעולות האפשריות. השתמשו בקצב למידה $\eta = 1$, ואתחלו את המבקר והשחקן לערכים שמצאתם בסעיף א' – למעט שני הערכים הבאים:

$$m_B(A) = 1 \ , \qquad m_C(A) = 1$$

דונו בתוצאות שקיבלתם.

<u>הדרכה:</u> התחילו את הלמידה מאיטרציה ללמידת המבקר, והמשיכו באיטרציה ללמידת השחקן. חזרו על התהליך פעם נוספת.

3. (50 נק') ישמו רשת עצבית הכוללת מספר שכבות, אשר לומדת לשחק איקס-עיגול באמצעות למידת חיזוקים. לשם כך, נתון לכם קוד חלקי (ראו קבצים באתר).

הקלט של הרשת מורכב מעשרה נוירונים. תשעת הנוירונים הראשונים מייצגים את המשבצות על הלוח, כאשר משבצת המאוכלסת ב-X תיוצג על ידי המספר 1 ואילו משבצת המאוכלסת ב-O תיוצג על ידי המספר -1. משבצת ריקה תיוצג על ידי המספר 0. הנוירון העשירי מייצג את השחקן הבא לבצע מהלך (1 לתור של X או -1 לתור של O).

הפלט של הרשת הוא נוירון יחיד. הגמול המתקבל הוא 0 לאורך כל המשחק, למעט הצעד האחרון. בצעד זה הגמול יהיה 1 אם X ניצח, -1 אם O ניצח או 0 אם המשחק הסתיים בתיקו.

הרשת מתאמנת על ידי משחק נגד יריב חצי-אקראי, הפועל לפי האסטרטגיה הבאה: אם ניתן לנצח בצעד אחד, נצח את המשחק. אחרת, אם היריב יכול לנצח בצעד אחד, חסום אותו. אחרת, בצע מהלך אקראי בהתפלגות אחידה. בסוף האימון, הרשת נבחנת על ידי משחק כנגד יריב זה.

א. יישמו את אלגוריתם TD(0). נסו לבחון מגוון ארכיטקטורות של הרשת, קצבי למידה וערכים שונים עבור פקטור הדעיכה. במידת הצורך, ניתן לשנות את ערכי הפרמטרים לאורך הלמידה.

ב. ממשו מדיניות softmax ונסו לאמן את הרשת באמצעות שימוש במדיניות שונה עבור השחקן בכל פעם (חמדנית, ϵ-חמדנית או softmax). איזו מדיניות אופטימלית לשלב האימון? האם תוצאה זו היתה משתנה באימון כנגד שחקן דטרמיניסטי?

הגמול הממוצע (על פני הרבה משחקים) נקבע, למעשה, רק לפי ההפרש בין ההסתברות לניצחון לבין ההסתברות להפסד (מדוע?). לכן נשתמש במדד זה להערכת ביצועי הרשת. ידוע כי שחקן מושלם שישחק נגד יריב חצי-אקראי ינצח ב-61.4% מהמשחקים ויסיים בתיקו בכל המשחקים הנותרים. לכן, מערכת שמשיגה הפרש של 61.4% בין הסתברות הנצחונות להסתברות ההפסדים נקראת מערכת בעלת ביצועים אופטימליים (במובן של הגמול הממוצע בלבד).

ג. האם הרשת בעלת ביצועים אופטימליים? אם לא - מדוע? (יש להסביר כל סטיה מעל אחוז בודד, גם אם התקבלו ביצועים טובים מהצפוי.)

<u>הדרכה:</u> נסו לשחק בעצמכם נגד הרשת לאחר האימון, ולהבין את היתרונות והחסרונות באסטרטגיית המשחק שלה.

<u>הבהרות:</u>

- מומלץ לערוך את קוד ה-MATLAB המוכן – הוראות מפורטות ניתן למצוא בקובץ overview.txt. עם זאת, מי שמעוניין לממש את הפתרון בגישה תכנותית שונה (כמו OOP) או בשפת תכנות שונה (כמו python), מוזמן לממש את הקוד הרלוונטי בעצמו.
- יש להגיש רק את הקוד שהגיע לתוצאות הטובות ביותר.