**חישוביות וקוגניציה**

**מרצה: ד"ר אורן שריקי, מתרגל: אביב דותן**

**25/04/2018**

**להגשה עד: 15/5/2018**

**דף תרגיל 4**

1. בנו רשת מצבי הדהוד (ESN) בעלת 500 נוירוני ביניים אשר מקבלת קלט יחיד ומחזירה 100 פלטים שונים באופן הבא:

הקלט הוא רעש לבן גאוסי בעל ממוצע 0 ושונות 1 (למשתמשי MATLAB - randn). המורה עבור נוירון הפלט ה- בצעד הזמן הוא הקלט עצמו, אבל לפני צעדי זמן: .

* 1. עבור כל נוירון פלט, חשבו את ריבוע מקדם הקורלציה () שלו עם המורה המתאים. הציגו את התוצאות בגרף המתאר את ריבוע מקדם הקורלציה כתלות באינדקס של הנוירון.
  2. *קיבולת הזיכרון* (Memory Capacity) של הרשת מוגדרת כסכום ריבועי מקדמי הקורלציה הנ"ל, והיא מהווה מדד מספרי ליכולת שלה לשחזר קלט מהעבר. חשבו אותה עבור הרשת שלכם.

הבהרה: כפי שלמדתם בשיעור, ישנם פרמטרים שונים שמשפיעים על התוצאות שיתקבלו: גודל הקשרים מהקלט לשכבת המשוב, גודל הקשרים בשכבת המשוב, אחוז הקשרים שמתאפסים ברשת, מקדם הזליגה ומקדם הרגולריזציה. נסו לקבוע את הפרמטרים השונים כך שקיבולת הזיכרון של הרשת תהיה מקסימלית (אינכם נדרשים לבדוק פרמטרים נוספים). יש להגיש את הקוד עם הפרמטרים הכי טובים שמצאתם.

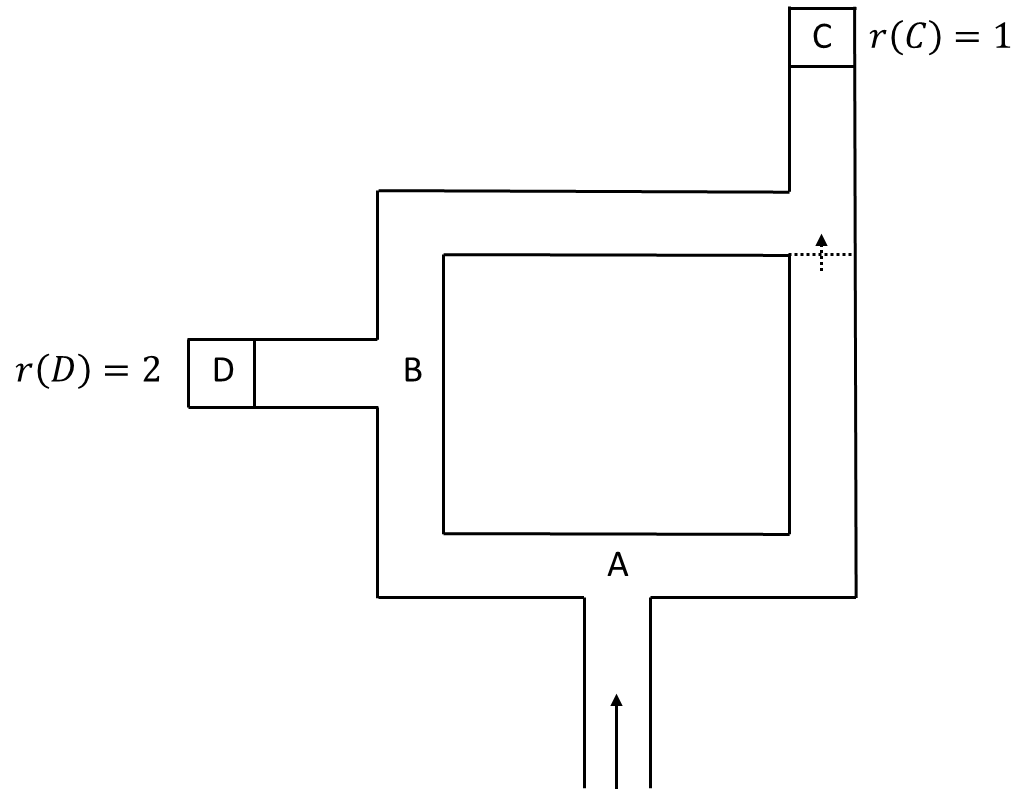
1. חוקר מבצע ניסוי צריכת דופמין בעכברים. הנחת החוקר היא שמטרת העכברים היא להביא לכדי מקסימום את רמת האושר שלהם. החוקר מציג בפני העכברים שני לחצנים, האחד גורם להזרקת דופמין רב והשני להזרקת מעט דופמין. במידה והעכבר ילחץ פעמיים ברצף על הלחצן שמוביל להזרקת דופמין רב הוא עלול לקבל התקף לב. החוקר מסיק כי לעכבר שלושה מצבים – אושר רב, אושר מועט והתקף לב. כאשר העכבר נמצא במצב של אושר רב הוא מקבל reward של 10 וכאשר העכבר נמצא במצב של אושר מועט הוא מקבל reward של 4. ניתן למדל את המחקר כ- MDP בעל 3 מצבים – אושר רב, אושר מועט והתקף לב. פונקציית המעבר (transition function) נתונה בטבלה. על סמך תצפיות החוקר, ניתן להניח כי ה discount factor הוא 0.8=γ. כמו כן, הניחו שכאשר העכבר חווה התקף לב, אין באפשרותו לקבל יותר reward.

[מקרא: אושר מועט – joy ; אושר רב – pleasure ; הזרקת מעט דופמין – s\_inject ; הזרקת הרבה דופמין –l\_inject ; התקף לב – heart]

|  |  |  |  |
| --- | --- | --- | --- |
| s | a | s' | P(s' |a, s) |
| joy | s\_inject | joy | 1 |
| joy | l\_inject | joy | 2/5 |
| joy | l\_inject | pleasure | 3/5 |
| pleasure | s\_inject | joy | 1/2 |
| pleasure | s\_inject | pleasure | 1/2 |
| pleasure | l\_inject | pleasure | 6/8 |
| pleasure | l\_inject | heart | 2/8 |

* 1. בהינתן מדיניות שמרנית, כאשר העכבר תמיד מזריק מעט דופמין, מה יהיה ה-value?
  2. מהי המדיניות האופטימאלית עבור כל מצב (אין צורך בחישוב מפורש – רשמו שיקולכם. שימו לב לvalue שחושב במדיניות שמרנית)?
  3. האם ניתן לשנות את ה discount factor כדי לקבל מדיניות אחרת? במידה וכן, תנו דוגמא, במידה ולא רשמו נימוק מפורט.
  4. האם ניתן לשנות את הrewards מבלי להשפיע על המדיניות האופטימאלית? במידה וכן, תנו דוגמא, במידה ולא רשמו נימוק מפורט.
  5. החוקר מנסה להעריך את ה – discount factor של עכבר אחר. הוא מוצא כי אותו עכבר מעדיף את רצף הפעולות {l\_inject,s\_inject,s\_inject} על-פני רצף הפעולות s\_inject,l\_inject,l\_inject}} ללא תלות במצב בו העכבר נמצא לפני ביצוע רצף הפעולות. תחת ההנחה שהפעולה בעלת הסבירות הגבוהה ביותר היא זו שמתבצעת, מהו הטווח של ערכי discount factor שמתאים להתנהגות זו? רשמו ביטוי מתאים ופתרו אותו.

1. רובוט מנסה ללמוד לנווט במבוך המתואר בתרשים ולאסוף גמול מקסימלי:



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

* 1. קבעו בעזרת שיקולים פשוטים, מהי המדיניות האופטימלית. חשבו את ערכי המבקר בכל מצב ואת ערכי השחקן לכל מצב ופעולה עבור המדיניות שבחרתם. (הניחו כי פקטור הדעיכה הוא 1.)
  2. כעת הרובוט משתמש בפקטור דעיכה .
     1. התאימו את כללי הלמידה להערכת המדיניות ולשיפורה שנלמדו בשיעור כך שיכללו את פקטור הדעיכה.
     2. כיצד יראו כללי הלמידה בממוצע על פני כל הפעולות האפשריות?
     3. בצעו שתי איטרציות למידה של אלגוריתם actor-critic באמצעות כללי הלמידה הממוצעים על פני הפעולות האפשריות. השתמשו בקצב למידה , ואתחלו את המבקר והשחקן לערכים שמצאתם בסעיף א' – למעט שני הערכים הבאים:

דונו בתוצאות שקיבלתם. הדרכה: התחילו את הלמידה מאיטרציה ללמידת המבקר, והמשיכו באיטרציה ללמידת השחקן. חזרו על התהליך פעם נוספת.

1. 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 ואילו משבצת המאוכלסת ב-O תיוצג על ידי המספר . משבצת ריקה תיוצג על ידי המספר . הנוירון העשירי מייצג את השחקן הבא לבצע מהלך ( לתור של X או לתור של O).

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

* 1. 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.
  2. 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).

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

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

הבהרות:

* 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.