**דוח מטלה 1**

**בקורס: מערכות מרובות סוכנים**

**מגישים:**

שי דונדיק 00000000

אנטון דמיטרוק 316940691

אליאב שללאשווילי 00000000

**תאריך הגשה:** 15/05/2020

**תוכן עניינים**

[שלב 1. הגדרת הבעיה 2](#_Toc37601417)

[שלב 2. משוואות בלמן ואלגוריתם Value Iteration 4](#_Toc37601418)

[שלב 3. מציאת מדיניות אופטימאלית ואלגוריתם Policy Iteration Algorithm 6](#_Toc37601419)

[שלב 4. מציאת V באמצעות אלגוריתם Temporal difference 6](#_Toc37601420)

[שלב 5. אלגוריתם Temporal difference state-action 6](#_Toc37601421)

[שלב 6. אלגוריתם Q-Learning algorithm עם וללא eligibility traces 6](#_Toc37601422)

[נספח – מטריצה A אשר חושבה בשלב 2 7](#_Toc37601423)

## – נעשה ע"י אנטון

## – נעשה ע"י אנטון

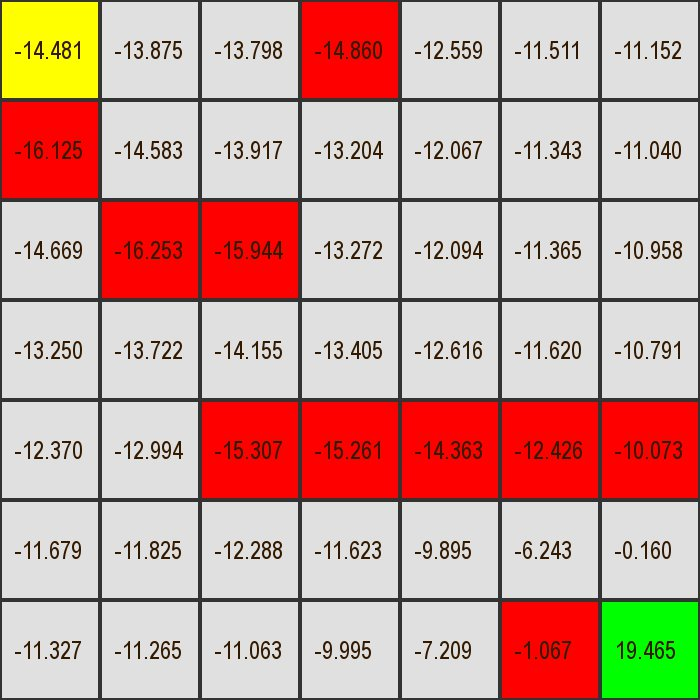
## – נעשה ע"י אנטון

## מציאת V באמצעות אלגוריתם Temporal difference

עבור המדיניות שנבחרה בשלב 2 (התפלגות אחידה על פני הפעולות האפשריות) ועבור הפרמטרים:

|  |  |
| --- | --- |
| Name | value |
| Alpha – learning rate | 0.001 |
| Gamma – discount rate | 0.9 |
| Number of episodes/games | 100,000 |

התוצאה המתקבלת של אלגוריתם TD(0) היא:



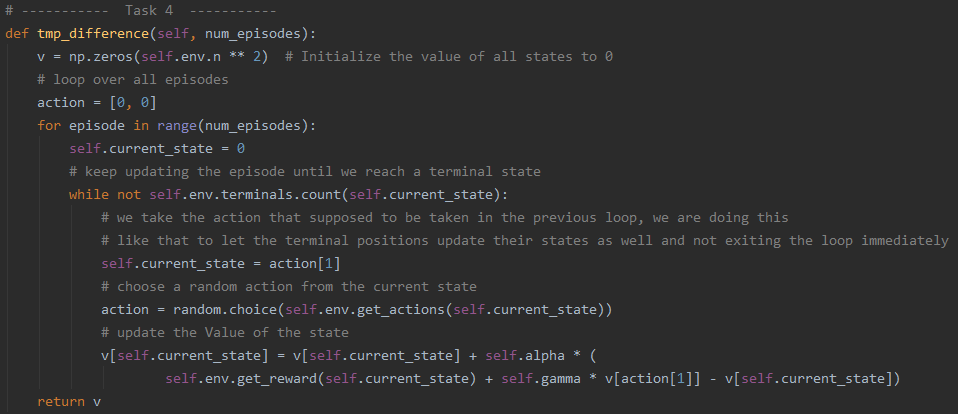
איור 5 - ערכי ה-V לכל מצב לאחר 1,000,000 איטרציות

באיור 5 ניתן לראות את ערכי שהתקבלו באמצעות שימוש ב-Temporal difference algorithm למשך של מיליון משחקים.

מסקנות:

* ניתן להגיד שבאופן כללי התוצאות לא קרובות להתכנסות גם לאחר מיליון צעדים באלגוריתם, יש לזכור שלכל אורך הדרך נבחרו פעולות מ"המדיניות האחידה" דבר שיכול לעקב מאוד התכנסות במקרים שבהם יש מרחב פעולות גדול או מורכבות שנובעת מהאופן שבו הסביבה נבנתה.
* מצד שני, ניתן להבחין באופי שונה בשני צידי לוח המשחק, בצד שבו הסוכן מתחיל את המשחק (שמאל למעלה) לעומת הצד שבו נגמר המשחק (ימין למטה). ההבדלים בגדלי הערכים נובע ממספר הפעמים שהסוכן ביקר בכל מצב, שכן, סביר שמספר המשחקים בהם הסוכן יבקר בתחילת הלוח (שמאל למעלה) יהיה גדול ביחס למספר הפעמים שיבקר בקצה הלוח (ליד נקודת הסיום).  
  ומכאן אפשר להסיק שיש תלות בין התכנסות (ביחס לערך האמיתי) ערך במצב מסוים ( עבור s כלשהו) למיקומו בלוח.

מימוש האלגוריתם:



## אלגוריתם Temporal difference state-action

אלגוריתם State-Action-Reward-State-Action (SARSA) מומש לפי הפרמטרים:

|  |  |
| --- | --- |
| Name | value |
| Alpha – learning rate | 0.001 |
| Gamma – discount rate | 0.9 |
| Epsilon – greedy policy | 0.1 |
| Number of episodes/games | 100,000 |

נבחן את תוצאותיו לפי טבלאות Q ההתחלתית כנגד הסופית.

* למען הסדר הטוב, שתי הטבלאות מובאות כנספחים: [נספח טבלה התחלתית](#_נספח_–_SARSA), [נספח טבלה סופית](#_נספח_–_SARSA_1)

[בטבלת Q ההתחלתית,](#_נספח_–_SARSA) אופן האתחול של ערכי הQ לכל זוג (state, action):

* עבור כל הפעולות הלא אפשריות בלוח (התקדמות הסוכן לתוך אחת מדפנות הלוח, לדוגמא ממצב '0' כלפי מעלה) אותחל הערך בטבלה במספר מאוד שלילי: -E05 .
* עבור כל שאר הפעולות ניתן לראות ערכי עדכון שנדגמו מתוך התפלגות אחידה בקטע (0, 1)

[בטבלת Q הסופית:](#_נספח_–_SARSA_1)

* ניתן לראות שכל מקום בטבלה שמייצג פעולה לא אפשרית ממצב מסוים, לא שינה את ערכו (ביחס [לערכים ההתחלתיים](#_נספח_–_SARSA)).
* ישנה הפרדה מובהקת של ערכי שליליים מאוד לעומת חיוביים מאוד בין מצבי הלוח ההתחלתיים לבין אלו הסופיים.

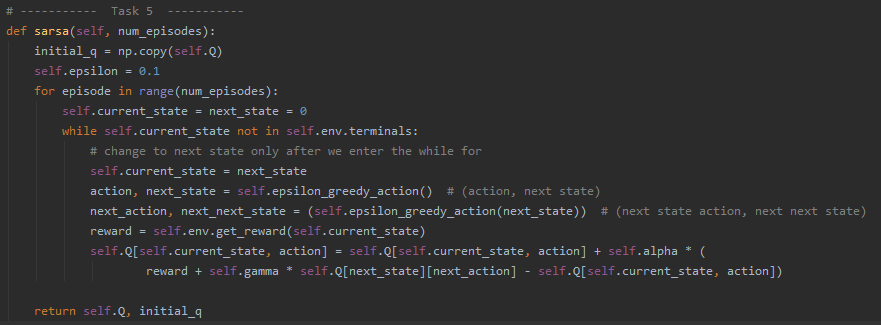
מסקנות:

* כמו שנאמר לעיל, [בטבלת Q הסופית](#_נספח_–_SARSA_1), כל מקום שמייצג פעולה אסורה ממצב מסוים, לא שינה את ערכו (כל ערכי ה -E05) ביחס [לטבלת Q ההתחלתית](#_נספח_–_SARSA).  
  ניתן להסביר זאת בקלות בכך שערכי יתעדכנו רק עבור פעולות אפשריות.
* במעבר על [טבלת Q הסופית](#_נספח_–_SARSA_1), ניתן להציג את המסלול האופטימלי (greedy), של מעברים בין מצבים אפשריים שהתקבל בסיום הלמידה (כאשר מתחילים ממצב '0'): (מימין לשמאל)

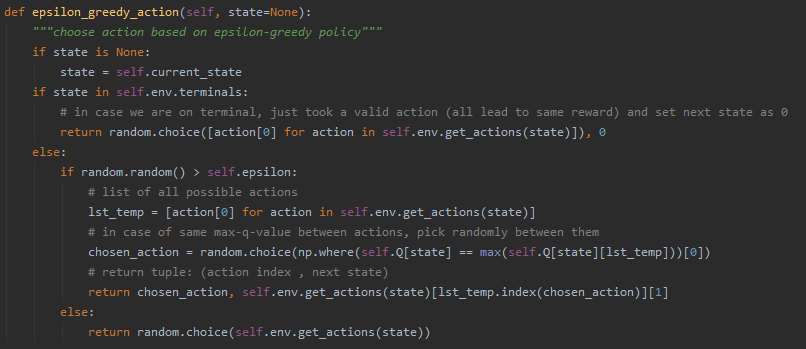
[0, 1, 8, 9, 10, 17, 24, 23, 22, 29, 36, 37, 38, 39, 40, 41, 48]

* בהמשך לסעיף הקודם, אם נתבונן בנקודה כלשהי ככל שמרחקה (לפי מטריקה אוקלידית / מנהטן) מסוף המסלול קטן, כך ערך ה שלה גדל, בממוצע. (הדבר גם תלוי במידת החריגה מהמסלול האופטימלי שכן, חריגות גדולות ממנו יזוכו בערך שלילי יותר, למרות "קירבה" מסוימת לסוף המסלול – לדוגמא – מצב '27')  
  הדבר נובע מהסיבה שהפרס (reward) החיובי היחיד במשחק מתקבל בהגעה למצב 48 שאפשרי רק דרך מצב 41 ובחירת הפעולה 'למטה' (או '1' לפי סדר האינדקסים של הפעולות בטבלה) ומכיוון שצורת העדכון באלגוריתם מקנה פרס לערך במצב נתון לפי ערך המצב הנוכחי ועוד ערך המצב הבא, יכול לקחת המון זמן (אם בכלל) עד שהמצבים "ההתחלתיים" בלוח "ירגישו" את אותו הפרס (שמתקבל רק עבור הגעה למצב הסופי) ובכך יתעדכנו לכיוון החיובי.

מימוש האלגוריתם:



פונקציית עזר למימוש מדיניות :



## נספח – SARSA – טבלת Q ההתחלתית



## נספח – SARSA – טבלת Q הסופית



## נספח – מטריצה A אשר חושבה בשלב 2

