אוניברסיטת
אריאל
בשומרון

קורס מבוא ללמידת חיזוקים

מבוך פורמולה עם מכשולים

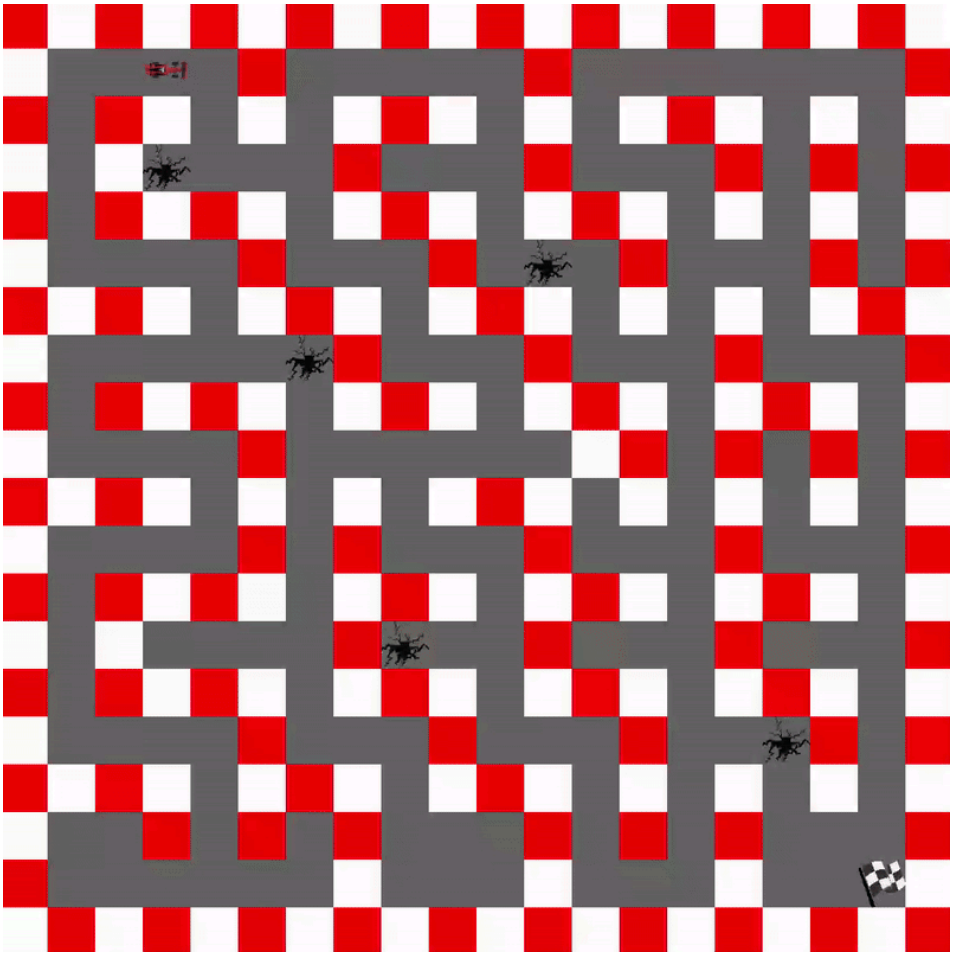
מגישים:

רון שטיינמץ 315578575

שחר תורג'מן 316223307

מנחה:

ד"ר טדי לזבניק

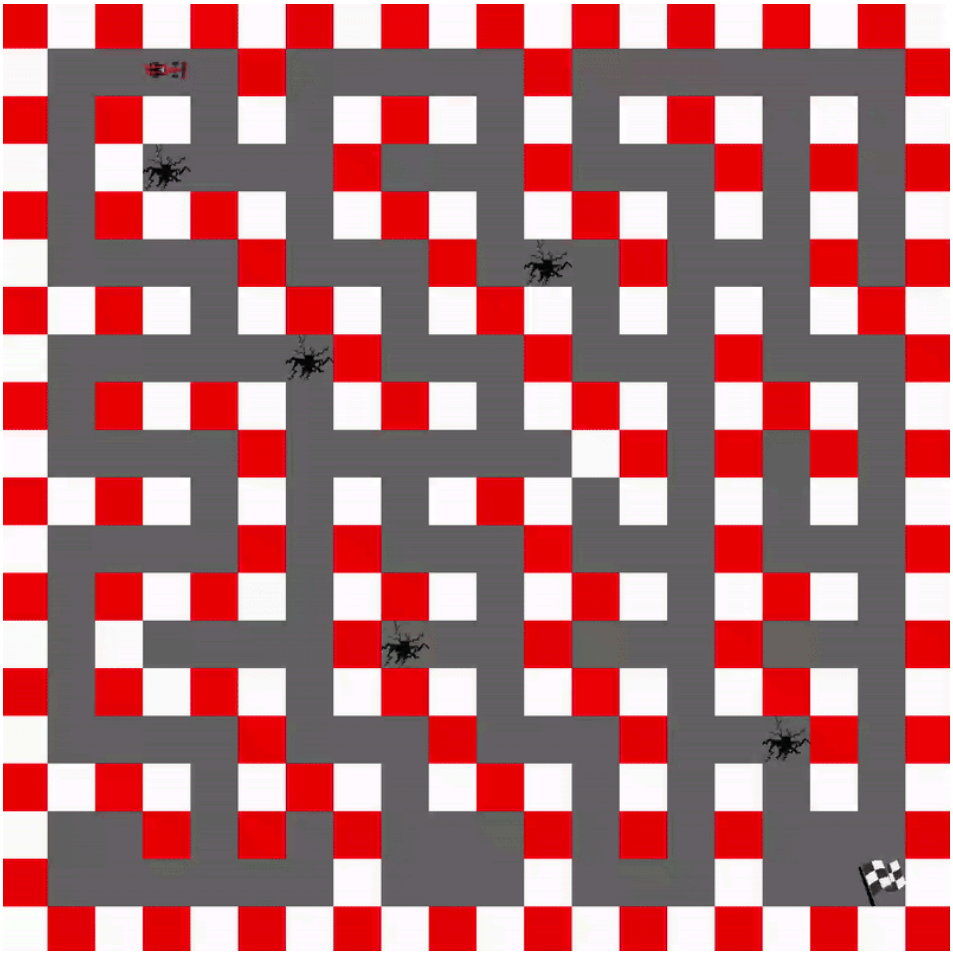


סקירה כללית של הפרויקט

במסגרת פרויקט זה, יישמנו אלגוריתם למידת חיזוקים מסוג Q-Learning, כדי ללמד סוכן רובוטי לנווט במבוך דו־ממדי, תוך התמודדות עם אתגרי תנועה, מכשולים, ומטרה קבועה.

מטרת הפרויקט הייתה כפולה: מצד אחד, להבין לעומק את אופן פעולתו של אלגוריתם Q-Learning ואת ההיגיון שמאחוריו ומצד שני, לבחון כיצד ניתן ליישם את האלגוריתם בסביבה גראפית אינטראקטיבית שפיתחנו בעזרת ספריית pygame.

במהלך הפרויקט, בנינו סביבה שבה הסוכן מתחיל כל אפיזודה במקום קבוע (צד שמאל למעלה), כאשר עליו למצוא את הדרך הקצרה והבטוחה ביותר לנקודת הסיום, שהיא יעד קבוע בסביבה(צד ימין למטה). לאורך הזמן, בעזרת למידה חוזרת, אמור הסוכן לשפר את בחירת הפעולות שלו על מנת להגיע ליעד בצורה יעילה יותר. סביבת המבוך כוללת קירות שמפריעים לתנועה החופשית ובורות במסלולים (שמכשילים את הסוכן במידה והגיע לשם), מה שמאלץ את הסוכן ללמוד איך לפעול בצורה חכמה ולא לבחור פעולות בצורה אקראית או פשוטה.



הבעיה/מטרה

הסוכן (מכונית המרוץ במקרה שלנו) נמצא בסביבת המבוך שיצרנו בספריית pygame, הסוכן צריך למצוא את הדרך היעילה ביותר מנקודת הפתיחה שלו בצד העליון השמאלי של המבוך ,לצד התחתון ביותר בצד הימני של המבוך (דגל משבצות).
בדרך ישנן אופציות נרחבות לאיפה שהסוכן יכול ללכת ,הקירות (ריבועים בצבע לבן ואדום) מגבילות את הסוכן והוא לא יכול לעבור דרכם, לא כל הדרכים מובילות ליעד ,חלקן אפילו כוללות בורות כדי לאתגר את הסוכן ולכן אם הוא מגיע לבור,הסוכן נפסל ומתחיל מחדש את המשחק.

- **מרחב המצבים**:

כל מצב במערכת מייצג את מיקומו הנוכחי של הסוכן בתוך המבוך, אשר מיוצג כמטריצה דו ממדית של תאים. כל תא במטריצה מהווה מצב (state) ייחודי, והזהות של כל מצב נקבעת על פי מיקומו בקואורדינטות של המפה. לצורך ייעול תהליך הסיווג, קואורדינטות אלו הומרו למספרים סידוריים ייחודיים: התא שבו הסוכן מתחיל מוגדר כ־State מספר 0, ואם הסוכן נוקט בפעולה של צעד אחד ימינה, הוא עובר ל־State מספר 1, וכן הלאה. מבנה זה מאפשר ייצוג חד משמעי של כל מצב באמצעות אינדקס מספרי. בסך הכול, קיימים 186 מצבים שבהם הסוכן עשוי להימצא במהלך תנועתו, כולם תואמים לתאים האפורים במפה, אשר מייצגים אזורים פתוחים ונגישים במבוך.

- **מרחב הפעולות :**

הסוכן פועל במרחב המוגדר כרשת דו ממדית, והוא מסוגל לבחור בין ארבע פעולות תנועה בסיסיות: למעלה, למטה, ימינה ושמאלה. עם זאת, חשוב להדגיש כי לא בכל סטייט קיימת אפשרות לבצע את כל ארבע הפעולות. מגבלות אלו נובעות ממבנה הסביבה. כך למשל, אם הסוכן ממוקם בצמוד לקיר בצדו הימני, האפשרות לנוע ימינה לא תהיה זמינה עבורו באותו מצב. מגבלה זו מדגישה את חשיבות ההתאמה בין הפעולה לבין המצב הנוכחי, ומהווה אתגר נוסף בתהליך הלמידה של הסוכן.

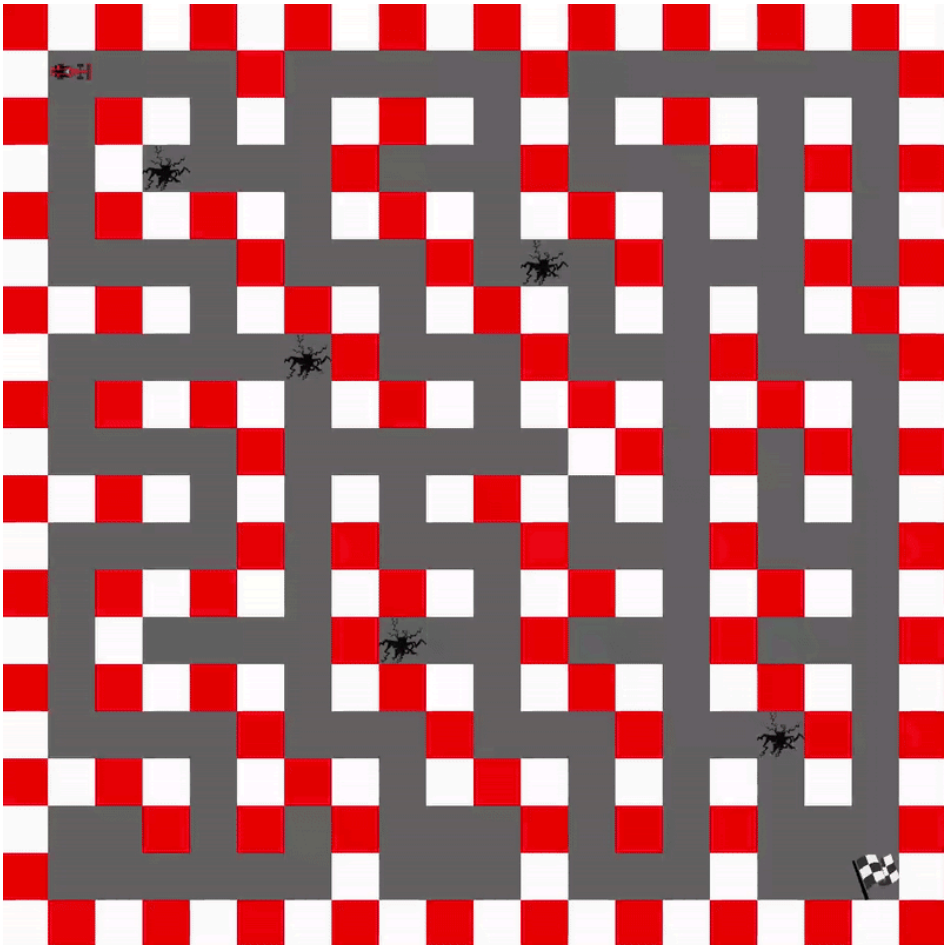
- **פונקציית התגמול**

על כל צעד שהסוכן מבצע במהלך תנועתו במבוך, הוא מקבל תגמול של 0.1-. מטרתו של תגמול שלילי זה היא לעודד את הסוכן לבחור במסלול הקצר והיעיל ביותר האפשרי אל היעד, על מנת לצמצם את מספר הצעדים הננקטים. בנוסף, במרחב המבוך פזורים בורות תאים

המוגדרים כמצבים מסוכנים. כאשר הסוכן מגיע לאחד מהבורות או מגיע למספר המקסימלי לצעדים שהגדרנו לו (150), הפרק מסתיים באופן מיידי והוא מקבל תגמול של 10-. תגמול שלילי גבוה זה נועד להרתיע את הסוכן מלהיכנס לאזורים אלה, וללמד אותו להימנע מהם לאורך הלמידה. מנגד, כאשר הסוכן מצליח להגיע ליעד הסופי, הוא מתוגמל ב +10, תגמול חיובי משמעותי שמטרתו לחזק את הקישור בין פעולות שהובילו להצלחה לבין השגת היעד בפועל. באמצעות מערכת תגמולים זו, הסוכן מפתח מדיניות אופטימלית המעדיפה מסלולים קצרים, בטוחים ומובילים ליעד.

- **דינמיקת הסביבה**

הסביבה שבה פועל הסוכן היא דיסקרטית, כלומר הן מרחב המצבים והן קבוצת הפעולות האפשריות מוגדרים מראש ובעלי גודל סופי. כל מצב מייצג מיקום מסוים במבוך, וכל פעולה מביאה את הסוכן, בתנאים מסוימים, למצב חדש. מעבר בין מצבים מתרחש בצורה דטרמיניסטית. כאשר הסוכן מבצע פעולה חוקית (שאינה חסומה על ידי קיר או בור), הוא עובר למצב החדש בהתאם לאופי הפעולה.



*דוגמה לסוכן שנכשל במשימה ונוסע על בור(נפסל במשחק)

מתודולוגיה

האלגוריתם שנבחר לפרויקט הוא Q-Learning, מהסיבה העיקרית שהסביבה שלנו היא דיסקרטית ונתנת לפתירה ע"י האלגוריתם הטבלאי Q-Learning .

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$$

תהליך הלמידה ומדיניות הפעולה

בתחילת האימון, הסוכן מתחיל עם טבלת־Q מאופסת, כלומר כל ערכי ה־Q (עבור כל צירוף של מצב ופעולה) שווים לאפס. במהלך האינטראקציה עם הסביבה, הסוכן מעדכן את ערכי הטבלה בהתאם לתגמולים שהוא מקבל, תוך שימוש בנוסחת עדכון ה־Q.
עם הזמן, ככל שמספר האפיזודות גדל, טבלת ה־Q נעשית מדויקת יותר, ומשקפת בצורה טובה יותר את המדיניות האופטימלית כלומר, את הפעולה הטובה ביותר שיש לבצע בכל מצב נתון.

	0	1	2	3
0	-1.549710	-1.572225	-1.549710	-1.526011
1	-1.526011	-1.526011	-1.549710	-1.501064
2	-1.501064	-1.501064	-1.526011	-1.474804
3	-1.474804	-1.447162	-1.501064	-1.474804
4	-1.431771	-1.387447	-1.418903	-1.451654
...	...	...	...	...
181	0.446882	0.446882	0.324538	0.575665
182	0.711226	0.575665	0.446882	0.575665
183	7.225384	0.000000	3.436790	0.000000
184	8.808076	5.618955	3.415790	7.000000
185	0.000000	0.000000	0.000000	0.000000

*טבלת ה-Q האופטימלית

כדי לאפשר למידה אפקטיבית, נעשה שימוש במדיניות חמדנית עם הסתברות אפסילון לחקירה. בתחילת הלמידה, ערך האפסילון נקבע ל־1, כך שהסוכן בוחר פעולות באקראי (exploration), דבר המאפשר לו לחקור את הסביבה ולגלות את התגמולים האפשריים. בהמשך, ערך האפסילון יורד בהדרגה לפי נוסחת דעיכה קבועה, עד שהוא מתייצב על ערך סופי של 0.05. בשלב זה, הסוכן ממעט לחקור ומרבה לנצל (exploitation) את הידע שצבר כלומר, לבחור את הפעולה עם ערך ה־Q הגבוה ביותר בכל מצב.

לאורך כל תהליך הלמידה, טבלת ה־Q מתעדכנת באופן רציף. ככל שהסוכן ממשיך לפעול בסביבה, הטבלה מתכנסת לערכים יציבים, והמדיניות שנגזרת ממנה מובילה להשגת תגמולים מצטברים מרביים דבר המעיד על כך שהסוכן פועל בצורה אופטימלית בסביבה הנתונה.

היפר-פרמטרים שנבחרו לניסוי היו:

- **מספר אפיזודות:** 1500
- **מספר צעדים מקסימלי:** 150
- **α (learning rate):** 0.7
- **γ (discount factor):** 0.95
- **ε התחלתי:** 1.0
- **ε מינימלי:** 0.05
- **קצב דעיכה של** 0.0005 **:ε**

בתחילת אימון הסוכן, ניסינו לאמן אותו לאורך של 200 אפיזודות, אך כפי שניתן לראות בתרשים מספר 2 הסוכן עדיין לא מספיק להתייצב מבחינת התגמולים המצטברים שהוא מצליח לקבל פר אפיזודה.

לכן ניסינו כל פעם לאמן אותו למספר אפיזודות גדול יותר עד שהגענו למספר של 1500 שצלח במשימה . קצב הלמידה אלפא היה לא נמוך מידי ולא גבוה מידי ולכן בחרנו להשאיר אותו 0.7 ואכן גילינו שזהו הקצב המתאים לסוכן בסביבה שלנו.

הוספנו גם פרמטר של מספר צעדים מקסימלי(150) כדי למנוע מצב שהסוכן נתקע בלופ אינסופי עם עצמו ולא מסיים את האפיזודה עם צעדים יעילים.

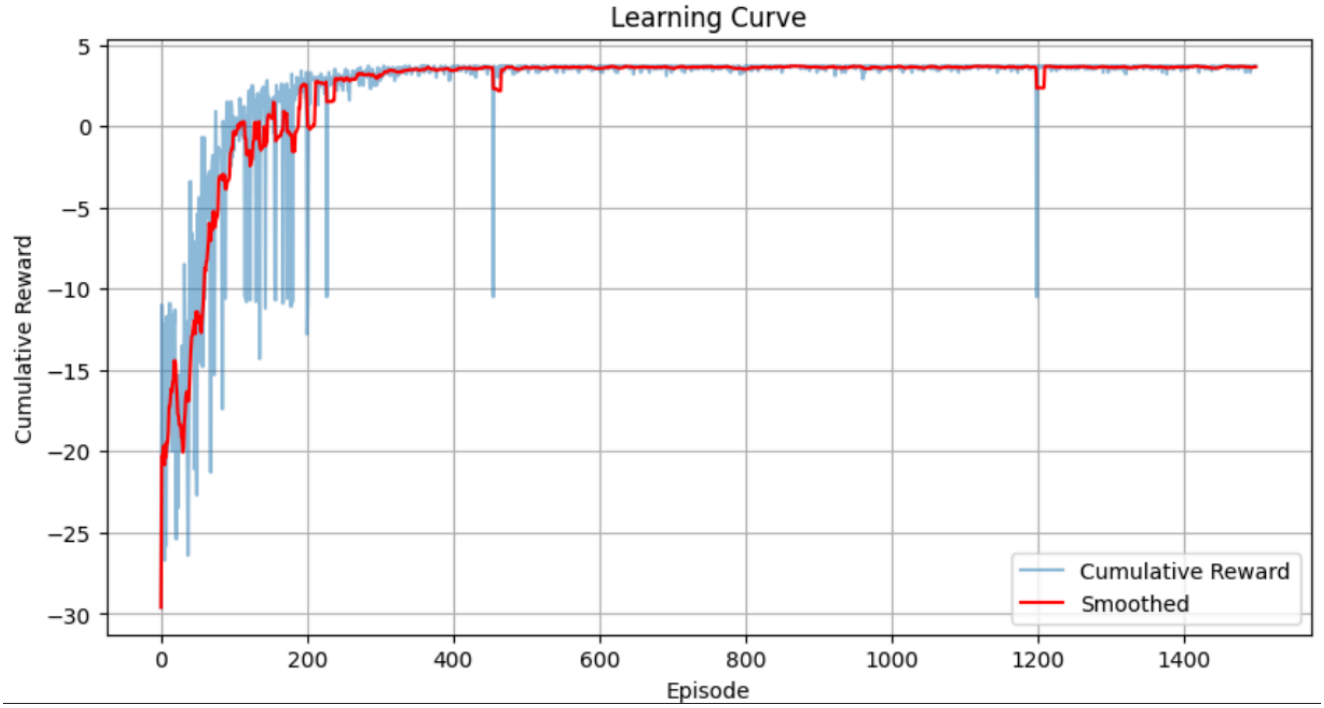
תוצאות

לאחר אימון של 1500 אפיזודות, ניתן היה לראות שיפור משמעותי בביצועי הסוכן. בתחילת הדרך, כאשר טבלת ה־Q עדיין הייתה ריקה או כמעט ריקה, הסוכן ביצע פעולות אקראיות, נתקל בקירות, חזר על אותם מסלולים ואף לעיתים לא הצליח להגיע ליעד. ככל שהתהליך התקדם, הביצועים הלכו והשתפרו הסוכן התחיל לפתח מדיניות פעולה שמביאה אותו ליעד במספר צעדים נמוך יותר, תוך הימנעות ממכשולים ידועים.

תצפיות כמותיות:

אחת הדרכים למדוד את התקדמות הסוכן היא מעקב אחר התגמול המצטבר בכל אפיזודה. בתחילה, התגמולים היו שליליים מאוד לעיתים אף מתחת ל25-, מה שמעיד על פעולות לא יעילות, התקלות בבורות ואי הגעה ליעד. בהמשך האימון, התגמולים השתפרו באופן ניכר, והחלו להופיע אפיזודות עם תגמול חיובי סימן לכך שהסוכן הגיע ליעד ואף עשה זאת בדרך יעילה .

ניתן לראות בתרשים שבהתחלה התגמולים שליליים מאוד ואז לאט לאט הם עולים ומתייצבים מה שמעיד על למידה של הסוכנים למספרים צעדים אופטימלי ומספר צעדים יעילים.



*תרשים -2 התגמולים המצטברים לאורך זמן (אפיזודות)

תצפיות איכותיות: במהלך הניסויים שערכנו, ניתן היה להבחין בבירור כי לאחר מספר אפיזודות למידה, הסוכן הצליח לפתור את המבוך בצורה עקבית ויעילה. התנהגותו הפכה ממקרית ומבולבלת בתחילת הדרך, לאסטרטגית וממוקדת עם התקדמות האימון. הסוכן פיתח מדיניות פעולה המעדיפה מסלולים קצרים ככל האפשר, תוך הימנעות מהגעה לבורות והגעה מהירה אל היעד. התנהגות זו מעידה על כך שהסוכן הצליח להפנים את חוקי הסביבה ואת מבנה התגמולים, ולפעול בהתאם לאסטרטגיה שממקסמת את התגמול המצטבר. ניתן לומר כי הלמידה אכן הייתה אפקטיבית, והביאה להיווצרות של דפוס פעולה רציונלי ויעיל.

דיון

התוצאות שהוצגו מראות שהסוכן מתייצב לאורך זמן כאשר עד 1500 אפיזודות הוא מצליח לפתור את המבוך בצורה היעילה והמהירה ביותר .
ניתן לראות שמספר התגמולים המצטברים מתייצב על 4-5 נקודות תגמול וזה אומר שהסוכן מצליח לפתור בהצלחה את המבוך .

במהלך אימון המודל הבחנו שכאשר הוא מאומן על מספר יחסית קטן של אפיזודות (200) עם קצב למידה מהיר יחסית (0.7) הסוכן לא מצליח לפתור את בעיית המבוך, הרבה פעמים הוא נתקע בקירות, נתקע בבורות והסתובב במעגלים.
ברגע שהבנו שמספר האפיזודות קטן מידי, העלנו את מספר האפיזודות ובכך גרמנו לסוכן להשתפר ועכשיו בכל פעם שנפעיל אותו מחדש, הוא ידע לפתור את המבוך.

במהלך העבודה ניסינו גם לשנות את פרמטרי האימון, ולמדתי עד כמה בחירה שגויה של ערכים יכולה לפגוע בתהליך. לדוגמה, ערך גבוה מדי של γ גורם לסוכן להעדיף את העתיד הרחוק בצורה מוגזמת, בעוד שערך נמוך מדי של α עלול למנוע ממנו לעדכן את הידע בצורה יעילה. לכן יש חשיבות רבה לא רק לאלגוריתם עצמו, אלא גם לבחירה מדויקת של פרמטרים ושל סביבה מתאימה.

מסקנות

הפרויקט אפשר לנו להבין לעומק את עקרונות למידת החיזוקים, הן ברמה התיאורטית והן ברמה הפרקטית. הבחירה באלגוריתם Q-Learning התגלתה כנכונה מאוד לסביבה הדיסקרטית שנבנתה. דרך התהליך ראינו איך מדיניות פעולה אופטימלית משתפרת עם הזמן, ואיך טעויות חוזרות ונמנעות לאחר מספיק ניסיונות.

ההישג המרכזי של הפרויקט הוא היכולת ללמד סוכן רובוטי לנוע בצורה מושכלת, מבלי לתת לו "מתכון" מראש, אלא רק באמצעות תגמולים. כמו כן, למדנו את חשיבות הגיוון בפעולה את הצורך באיזון נכון בין חקירה לניצול, ואת ההשפעה של מבנה הסביבה על יכולת הלמידה.