**מסמך סוף סמסטר א׳ – פרויקט גמר**

פרויקט מס׳ 7 – פיתוח משחק סוכנים נגד מרגלים

מגישים:

|  |  |
| --- | --- |
| שם מלא | תעודת זהות |
| יואב הראל | 303168231 |
| סתיו הרצוג | 208852327 |
| משה מנע | 313266223 |

פיתוח המשחק – Catch me if you can

המשחק מיועד ל2 שחקנים, שחקן אחד יהיה מרגל ושחקן שני מפעיל הסוכנים.   
מטרת המרגל היא להגיע ליעד הסופי מבלי להיתפס, ואילו מטרת מפעיל הסוכנים היא לתפוס את המרגל לפני שיגיע ליעד הסופי.

לוח המשחק הוא מפת עולם עם 40 שדות תעופה, אשר מחוברים במידה וקיימת בניהם טיסה ורק כך השחקנים יכולים לנוע במפה.

2 השחקנים ישחקו ע"י המחשב באמצעות מודל של Reinforcement Learning.

תרשים Reinforcement Learning של מודל המרגל:

[ מיקומי השחקנים והיעד [

Map

Description automatically generated Diagram

Description automatically generated

[ 40 שדות תעופה [

[ שדה תעופה [

נתוני המשחק – טיסות

Chart

Description automatically generatedנתוני הטיסות נלקחו מקובץ JSON לכל שדה תעופה יש מערך המכיל את השדות אליהם מקושר ע"י טיסה.

**ארכיטקטורה:**

תרשים:

Diagram

Description automatically generated

הסבר:

Client – מומש ב React

Server – מומש ב Phyton ע״י FastAPI

Model – מומש ב Phyton בעזרת ספריית Gym Open AI

בקשת post מצד הclient – שליחת Observation כפרמטר שמורכב מ: Spy\_Position,Agent1\_Position,Agent2\_Position,Target\_Postion

השרת מעביר את ה Observation למודל, המודל מבצע Predict ומחזיר Action לשרת והוא מעביר אותו כתשובה ל Client.

**אלגוריתמים עיקריים:**

אלגוריתם PPO – Proximal Policy Optimization – בשימוש ע״י המרגל

לאלגוריתם PPO יש 2 רכיבים מרכזיים: רשת מדיניות אשר ממפה states ל actions, ורשת ערכים שמעריכה את ה reward הצפוי של כל state. רשת המדיניות ורשת הערכים מאומנות בעזרת הנתונים שנאספו ע״י agent לאחר מס׳ איטרציות בסביבה.

שלבים של אלגוריתם PPO:

1. איסוף נתונים: agent אוסף נתונים ע״י ביצוע המדיניות הנוכחית בסביבה ושומר את התוצאות.
2. חישוב יתרונות: רשת הערכים משמשת להערכת ה reward הצפוי של כל state שה agent היה בו. היתרונות של כל action שנעשה ע״י agent מחושבים כהפרש בין ה reward הצפוי ל state הבא לבין ה state הנוכחי.
3. עדכון מדיניות: רשת המדיניות מתעדכנת ע״י חישוב יחס ההסתברויות בין המדיניות הישנה לחדשה. המדיניות מתעדכנת בעזרת Gradient descent כדי למזער את ההפסד.
4. עדכון רשת הערכים: רשת הערכים מתעדכנת ע״י מזעור פונקציית ההפסד בין היתרונות הממשיים לחזויים.

חוזרים על שלבים 1-4 מס׳ איטרציות עד שהמדיניות מתקרבת לאופטימליות.

האלגוריתם מתוכנן לאזן בין חקירה (explore) של צעדים חדשים, לבין שימוש בצעדים שכבר סווגו כצעדים טובים (exploit).

אלגוריתם BFS – בשימוש ע״י הסוכנים

מציאת המסלול הקצר מהסוכנים למרגל.