פרויקט סיום בקורס מבוא לבינה מלאכותית

3D multi-player connect 4

מגישים:

איתמר שרם, 206762551

שלום בלוי, 319144762

עבד נירוך, 213668700

כאן יהיה תוכן עניינים

מבוא

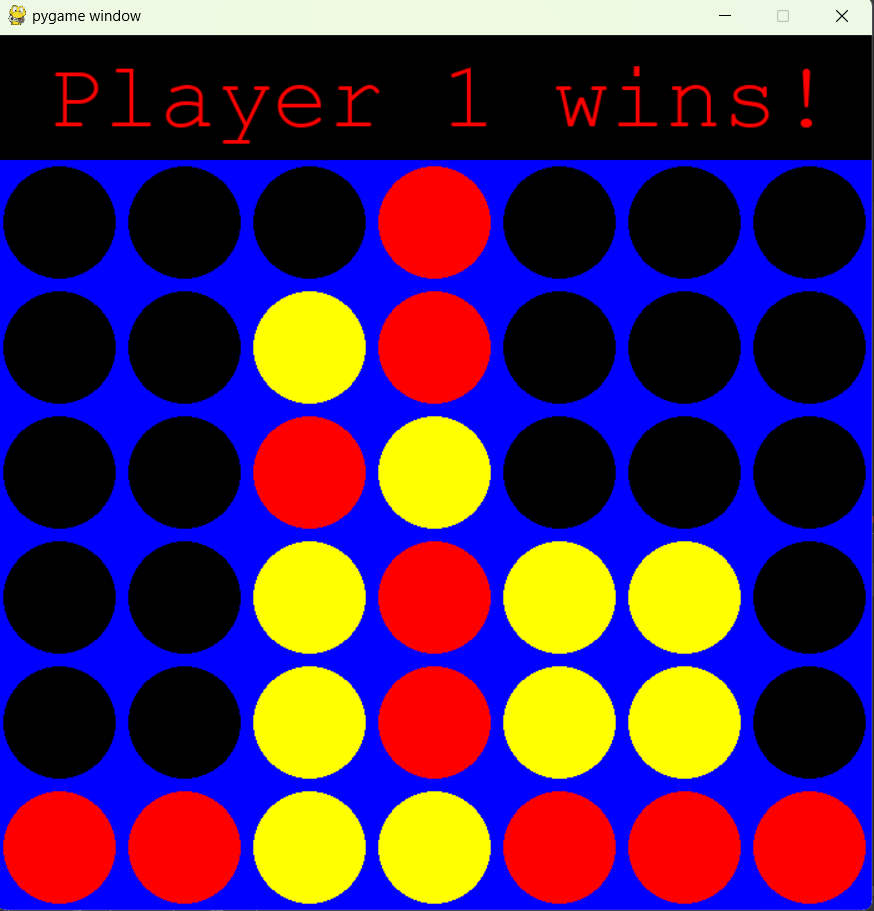
עבודות קודמות

מתודולוגיה

תוצאות

סיכום

ביביליוגרפיה(לא דרשו, אבל כדאי כדי לשים גם קישורים לUI ולעבודות קודמות וכו')

מבוא:

המשחק שלנו מבוסס על המשחק 4 בשורה. המשחק המקורי מתקיים בלוח דו ממדי בגודל קבוע, לשני שחקנים. כל שחקן בתורו משחיל דיסקית בצבע שלו לאחת העמודות בלוח. מטרתו של כל שחקן להגיע ראשון לרצף של 4 דיסקיות בצבע שלו. הרצף יכול להיות בשורה, בעמודה או באלכסון.

בפרויקט שלנו הרחבנו את המשחק, על מנת לאתגר את הסוכנים שלנו ולבחון אותם בסביבות מורכבות יותר. השדרוג חל בכמה אופנים:

* לוח המשחק מוגדר כעת ע"י הרביעייה הבאה: (R,C,D,W), כאשר R הוא מספר השורות, C הוא מספר העמודות, D הוא העומק וW הוא רצף הנצחון
* ניתן להריץ את המשחק עם יותר משני שחקנים, כל אחד מתחרה בשאר השחקנים על מנת לנצח.

עץ המצבים של המשחק גדל בצורה אקספוננציאלית- כאשר t הוא מספר התורות עד לניצחון. כלומר סוכנים שיממשו אלגוריתמי חיפוש פשוטים ירוצו למשך זמן רב! לפיכך צריך לחשוב על אלגוריתמים חכמים, וזו הסיבה שבחרנו במשחק זה.

לקחנו מימוש למשחק 4 בשורה שמצאנו באינטרנט, על מנת לאפשר לבן אדם לשחק נגד אחד הסוכנים בצורה אינטרקטיבית. שדרגנו משמעותית מימוש זה כדי שיתאים להרחבות שהוספנו למשחק. להוסיף קישור לgithub של המימוש, או רפרנס לנקודה מתאימה בביביליוגרפיה.

מימשנו סוכני AI שונים, לפי רעיונות שונים שלמדנו בכיתה, שמטרתם היא לנצח במשחק את השחקנים שיתמודדו מולם. הסוכנים שמימשנו הם:

1. minmax\_agent
2. alpha\_beta\_agent
3. Q-learning\_agent
4. רשת CNN

נשים לב שהמשחק המקורי הוא משחק סכום אפס- נצחונו של שחקן אחד הוא הפסדו של האחר. לפיכך חשבנו כי האלגוריתם הטבעי ביותר להתחיל איתו הוא אלגוריתם המינמקס, שמותאם לנצח בסוג זה של משחקים. אלגוריתם מינמקס במימוש נאיבי, יורד עד לעומק עץ המצבים, ומהעלים מתחיל לשערך את הציון למצבים. במשחק 4 בשורה, שבו עץ המצבים עמוק מאוד, דבר זה אינו ישים. לכן נצטרך להגביל את עומק החיפוש שלנו ולתת ציון למצב לוח למרות שהמשחק עדיין לא הסתיים. לשם כך השתמשנו בהיוריסטיקה/היוריסטיקות שהגדרנו, אותן נתאר בחלק המתודולוגיה.

סוכן האלפא-בטא הוא הסוכן ההגיוני הבא- הוא פועל כמו סוכן המינמקס מבחינת ההחלטות שבוחר במהלך המשחק, אך מדלג על ענפים מסוימים ובכך מקצר את זמן הריצה.

לסיום, רצינו לבחון את ביצועיו של סוכן reinforcement learning . דרך אפשרית להפעיל את האלגוריתם במקרה שלנו, היא לתת לשחקן למידת החיזוק לשחק מול שחקן מינימקס/רנדומי ודרכו ללמוד איך לשחק את המשחק בצורה הטובה ביותר.

סוכן reinforcement learning – סוכן זה לומד באמצעות משחקים רבים את הפעולות הכדאיות בהינתן מצבי הלוח, כאשר במהלך הלמידה הסוכן מעדכן את הציון על מצב לוח מסוים, באמצעות הרצת סימולציות עד לנצחון/הפסד(monte carlo).

אולי יהיה הסבר קצר על רשת CNN

עבודות קודמות:

בחנו עבודות קודמות(לשים קישורים לרלוונטיים בביביליוגרפיה)

בשתיהן, האלגוריתמים בהם השתמשו הם minmax ו-alpha beta pruning . כפי שאנחנו הסקנו, עץ המצבים במשחק גדול מאוד, ולכן החוקרים(?) בחרו להגביל את עומק החיפוש ולחשב ציון ללוח של משחק שטרם הסתיים. ההבדלים בין האלגוריתמים שהם בחרו הוא בפונקציית האבליואציה שמחשבת ציון ללוח. נציג את ההיוריסטיקות המוצלחות ביותר בכל אחת מהעבודות.

נסמן ב-N את אורך הרצף המוגדר עבור נצחון במשחק(ברירת המחדל היא 4)

פונקציה א:

במשחק 4 בשורה עם לוח דיפולטיבי(6 שורות על 7 עמודות) ישנן 69 פוזיציות בהן אפשר להגיע לנצחון(24 אופקיות, 21 אנכיות ו24 אלכסוניות). אם ברצף באורך N יש לשני השחקנים דיסקיות, נתעלם מרצף זה(שכן אף אחד מהם לא יכול לנצח שם). אחרת, לכל רצף בלוח ניתן ערך לפי כמות הדיסקיות שיש בו לשחקן מסוים. נסכום את כל הערכים הללו עבור השחקן הראשי, ונחסר מהם את סכום הערכים של היריב.

בהיוריסטיקה זו יש כמה חסרונות:

* הערכים שנבחרו מותאמים עבור לוח בגודל הדיפולטיבי בלבד, וקשה להכליל אותה למשחק עם לוח בגודל שונה, עם רצף נצחון שונה.
* ההיוריסטיקה יכולה לתת ערך זהה ללוח בו יש לנו הרבה רצפים קצרים, לעומת מעט רצפים ארוכים. אם נחשוב על הרלקסציה שבה ככל שיש לשחקן יותר דיסקיות ברצף, הוא יותר קרוב לנצחון, היוריסטיקה זו מתעלמת מעיקרון זה.

פונקציה ב(נקראה IBEF2): נספור רצפים לא חסומים(שלא מכילים דיסקיות של היריב).ונחזיר , עבור k מספר הדיסקיות ברצף זה. הפונקציה תסכום את הערכים הללו ותחסיר מערך זה את סכום הערכים של היריב.

חסרונות של היוריסטיקה זו:

בדומה להיוריסטיקה הקודמת, יכולה לתת ערך זהה ללוח בו יש לנו הרבה רצפים קצרים, לעומת מעט רצפים ארוכים.

מתודולוגיה:

מינמקס, אלפאבטא והיוריסטיקות

הרחבת אלגוריתמי minmax,alphabeta עבור משחק של יותר משני שחקנים:

במקרה זה, השחקן הראשי יהיה שחקן המקסימום, ומבחינתו כל שאר השחקנים יהיו שחקני המינימום. בשערוך הערך עבור כל קודקוד בעץ ששחקן המקסימום מגדיר(עד עומק מסוים), שחקן המקסימום יבחר את הערך המקסימלי מבין ערכי כל ילדיו. כל אחד משחקני המינימום יבחר את המינימום מבין ערכי כל ילדיו.

נשים לב כי בגישה זו, מטרתו של כל שחקן מינימום היא לוודא ששחקן המקסימום לא מנצח, גם אם זה גורר ששחקן מינימום אחר ינצח- מכיוון שההיוריסטיקות מחושבות במינוס על כלל היריבים, אזי מטרת כל אחד משחקני המינימום היא שכל השחקנים היריבים(כולל הם עצמם) יקבלו ניקוד גבוה, וכך יורידו מערכו של שחקן המקסימום. בעיני שחקן המקסימום, כלל יריביו עשו יד אחת נגדו.

נשים לב כי גישה זו נכונה גם עבור ה-pruning באלגוריתם האלפא-בטא: עבור כל אחד משחקני המינימום, הם יודעים שמעליהם קיים איפשהו שחקן מקסימום שייקח את המקסימלי מבין הערכים שיתנו, ולכן יגזמו בהתאם לערך זה. מכיוון שכל אחד משחקני המינימום לוקח מינימום על ילדיו, לשחקן המקסימום יפעפע מינימום על ערכי מינימום, דבר דומה ללקחת מינימום רק לאחר כל פעולותיהם של כל שחקני המינימום יחדיו, כאילו הוא מתמודד רק נגד שחקן מינימום אחד. לפיכך גם הוא יגזום בהתאם, ולא יפספס הסתעפויות רלוונטיות בעץ.

מימשנו מספר היוריסטיקות פשוטות לצורך בדיקת שאר השחקנים, עליהם נפרט בשלב התוצאות. כעת נציג את ההיוריסטיקה המרכזית ששימשה את אלגוריתם המינמקס:

היוריסטיקת complex (לשנות בקוד בarg\_parse לcomplex במקום allcomplex)

ההיוריסטיקה פועלת באופן הבא:

בהינתן לוח מסוים, ההיוריסטיקה תבחן את כל הרצפים השונים הקיימים על הלוח אשר לא חסומים, לכל לוח היא מחזירה שני ערכים, שמייצגים את הרצף הארוך ביותר שהושג בלוח, וכמות המופעים שלו. מאלו מחסרים את שקלול הערכים המתאימים ליריבים באופן הבא:

* מהרצף הארוך ביותר של שחקן המקסימום, נחסיר את ממוצע הרצפים המקסימליים של היריבים
* מכמות הרצפים הארוכים ביותר של השחקן המקסימלי, נחסיר רק את כמויות הרצפים המקסימליים של יריבים שהשיגו אותו רצף מקסימלי זהה לשל שחקן המקסימום- אנחנו לא רוצים לתת משקל ליריבים "חלשים" יותר בלוח נתון. רק יריבים שמאיימים על השחקן שלנו באים לידי ביטוי בציון הלוח.

לדוגמא:  
רצף הנצחון מוגדר להיות 4, שחקן המינמקס הוא השחקן הורוד. הרצף הכי ארוך שלו הוא 3, וכמות הרצפים באורך זה היא 1. לאדום יש רצף אחד באורך 3 (הרצף השני שלו באורך 3 חסום). לצהוב חמישה רצפים באורך 1 (ישנם רצפים שנספרים יותר מפעם אחת, עבור כל רצף נצחון בו הם יכולים להשתתף). ולכן פונקציית ההיוריסטיקה תחזיר את הציון הבא:

נשים לב כי בדרך זו, אנו נותנים עדיפות לאורך הרצף, ורק כאשר הרצפים זהים אנחנו מסתכלים על מספר המופעים. בכך אנו מתגברים על הבעיה שהצגנו בהיוריסטיקות מהעבודות הקודמות.

כאשר הרחבנו את המשחק ללוח תלת ממדי, נתקלנו בבעיה- על מנת לבדוק האם הנחה של דיסקית הובילה לנצחון, עלינו לבדוק רצפים רבים שיכולים להביא לנצחון. דבר זה פגע בזמני הריצה. לכן על מנת לשפר את זמני הריצה, אנחנו מחזיקים עבור כל לוח את מפת כל הרצפים המופיעים בו, וכאשר מוסיפים דיסקית, מעדכנים את מפות הרצפים רק באותו אזור מצומצם. הדבר שיפר משמעותית את זמני הריצה (למשל כדי למצוא את הרצף הארוך ביותר בלוח, מסתכלים במפת הרצפים)

הסוכן הכפול- סיכוי גבוה שכל החלק הזה יימחק

בarg parse להוציא את האפשרות לבחור בrl\_agent , ולשנות אתcnn\_agent להיקרא double\_agent

כדי שסוכן q-learning יציג תוצאות טובות, עליו להכיר הרבה מצבים במשחק, על מנת שידע לפעול בכל סיטואציה שייתקל בה. יתרה מזאת, אם יפגוש מצב שלא ראה קודם לכן, אפילו אם ראה מצב דומה(למשל תבניות דומות, אבל בהזזה מהמצב שכן ראה) הוא לא ידע כיצד לפעול.

לפיכך, היינו רוצים מרחב מצבים קטן ככל האפשר. מצד שני, ככל שייצוג המצב פשוט יותר, יותר מידע הולך לאיבוד.

ניתן היה לבחור ייצוג שמחזיק מידע חלקי לגבי הלוח, אך בחרנו להתמודד עם בעיה זו באופן הבא:

במשחק שלנו ישנה חשיבות רבה לתבניות, שיכולות להופיע במיקומים שונים בלוח, ולכן כדאי להשתמש בכלי שהוא translation invariance , שמסוגל לזהות תבנית ללא תלות במיקום. בכך, נתפוס את הטוב משני העולמות- מצד אחד נחזיק ייצוג מלא של הלוח, ומצד שני נוכל לזהות תבניות בלוחות שראינו על מנת להכליל ללוחות דומים.

לפיכך יצרנו את הסוכן הכפול- סוכן זה משתמש בסוכן Q-learning שילמד כמה שיותר מצבי לוח. בשל החסרון המוטמע שאינו יכול להכליל לוחות אלה, ישתמש בסוכן CNN, שהתאמן על הלוחות הנ"ל, ידע להכליל עליהן, ויבחר עבור הסוכן שלנו את המהלך הטוב ביותר.

סוכן Q-Learning

פונקציית הרווח:

כפי שלמדנו בקורס, על מנת שסוכן הRL ילמד לשחק במשחק, עלינו לספק פונקציית רווח שתיתן לו ציון על הפעולה שביצע. הפונקציה שבחרנו פועלת באופן הבא: ראשית תתגמל אותו על נצחון. אם לא ניצח אך הגענו לתיקו, תיתן ציון חיובי קטן יותר. אם הוא הולך להפסיד בתור הבא לאחד השחקנים, ניתן ציון שלילי, שערכו בערך מוחלט קטן ככל שהשחקן שרד למשך יותר משחקים.

מאופן פעולת האלגוריתם, ערכים אלה יפעפעו מטה עבור לוחות שרחוקים מהכרעה, ולשם כך דרושים משחקים רבים לצורך אימון הסוכן.

אופן האימון

אימנו סוכן זה לשחק מול \_\_\_\_\_ , מכיוון שאנו רוצים שיראה \_\_\_\_\_\_ . בפונקציית הרווח, כאשר נרצה לסמלץ מהלך של יריב, נסמלץ מהלך של יריב ה\_\_\_\_\_\_, כי אנחנו רוצים \_\_\_\_\_\_ .

סוכן CNN

פרסור ה-Q table

כזכור, סוכן ה-Q-learning מתחזק טבלה, שבה עבור כל מצב ופעולה בהם נתקל, קיים ציון. המרנו טבלה זו, למאגר מידע, בו לכל לוח יש ציון. הדרך שבה השגנו את הציון הזה, היא ע"י מעבר על הQ-table , לכל מצב עם ערך שאינו טריוואלי (שמעיד על כך שמידע לא פעפע לנקודה זו), יפעיל את הפעולה על הלוח, ונקבל לוח חדש, עם ציון זהה. נמצע ציונים של לוחות זהים. כלומר הדגימות שאנו מעבירים לרשת הם לוח, וציונו.

רשת ה-CNN

רשת זו תהיה מורכבת מקונבולוציות, עם קרנלים בגודל רצף הנצחון. הקרנלים יסרקו את לוחות המשחק שיקבלו, ונצפה שכל קרנל יחפש תבניות אחרות בלוח.

לאחר שכבות הקונבולוציה, נפעיל שכבות לינאריות שישקללו את כל הנתונים שאספו הקרנלים לצורך חיזוי ערך הלוח.

בחירת פעולה

בזמן משחק, כאשר סוכן ה-CNN ירצה לבצע פעולה, יחשב את כל הלוחות הישיגים ע"י הפעולות החוקיות, יפעיל על כל אחד מהם את הרשת, ויבחר בפעולה שהובילה ללוח שקיבל את הציון הגבוה ביותר.

סוכן Q-Learning – חדשש!!!!!!!!!!!!!!!!!!1

המשחק שלנו מכיל יותר משחקן אחד, ולכן אלגוריתם ה-Q-learning שונה מהאלגוריתם שראינו בכיתה: הסוכן שלנו מבצע פעולה, וכאשר באים לבחון את השלכות הפעולה שלו, כלומר לאיזה מצב הגענו, יש להסתכל לאיזה מצב הגענו אחרי הפעולה שלנו, אך גם של היריבים שלו.

החלטנו לממש את אלגוריתם ה-Q-learning בעצמנו, כך שנוכל להתאים את האלגוריתם לצרכנו: ריבוי משתתפים, שינוי שיטת האימון והתאמה ל-API שלנו כך שיוכל להתאמן ולשחק עם שאר השחקנים.

שיטת אימון

ראשית, היינו צריכים להחליט מול מי ישחק הסוכן שלנו בשלב הלמידה. מאחר ובמשחק שלנו קיימים מצבי נצחון והפסד רבים, וקיימות דרכים רבות להגיע למצבים זהים, יש בעיה בבחירת יריב רנדומלי- ייתכן מצב בו ניצחנו אותו, למרות שמדובר בסדרת מהלכים שמול שחקן רגיל הייתה מובילה להפסד ודאי. כך יפעפע ציון חיובי עבור סדרת מהלכים זו.

מאידך, אם היינו בוחרים לאמן נגד יריב חזק מדי, הסוכן שלנו כמעט ולא היה מנצח. שמנו לב שהאופן שבו האלגוריתם q-learning בנוי, הוא שכאשר מתעדכן ציון שלילי, לוקח זמן עד שהציון הזה יחלחל למצבים שקודמים לו. לכן בשיטה זו הלמידה של הסוכן מועטה, אם בכלל.

לעומת זאת, ציון חיובי מפעפע מהר- Good news travel fast!

לכן החלטנו לבחור ביריב שמבצע החלטות מושכלות, אך שיהיה חלש יותר מהסוכן שלנו במהלך הלמידה, וכך הסוכן יוכל להגיע יותר פעמים למשחקים בהם ניצח

על מנת לאפשר לסוכן לנצח את היריב, הוא התאמן בליווי "מאסטר": המאסטר הוא שחקן שמבצע החלטות מושכלות, וגם חזק יותר מהיריב. בתחילת דרכו שחקן הRL ייתן למאסטר לשחק בשבילו, ובאמצעות הקטנה של ה-exploration rate, המאסטר לאט לאט ישחרר את גלגלי העזר וייתן לסוכן לבצע יותר החלטות בעצמו.

בסופו של דבר בחרנו את המאסטר להיות שחקן אלפא בטא עם ההיוריסטיקה complex, עם עומק 2, והיריב הוא שחקן דומה שיוחלש ע"י כך שבסיכוי 0.2 הוא מגריל צעד.

פונקציית רווח

בחרנו לתת ציון של 500 עבור לוח שבו הסוכן ניצח. עבור לוח שבו הפסיד, בחרנו לתת -500 ועוד מספר הצעדים ששוחקו עד להפסד- כלומר ככל שהסוכן "שרד" יותר זמן, אנו מחשיבים את ההפסד שלו לקטן יותר, מכיוון שיכול להיות שרק המהלכים האחרונים שלו הביאו להפסד, אך הוא שיחק טוב בתחילה.

לסיום, עבור כל צעד שלא הביא לאיזושהי הכרעה, נתנו לו ציון של -1 , על מנת שיעדיף בנצחון המהיר ביותר.

ייצוג הלוח

כדי שסוכן q-learning יציג תוצאות טובות, עליו להכיר הרבה מצבים במשחק, על מנת שידע לפעול בכל סיטואציה שייתקל בה. יתרה מזאת, אם יפגוש מצב שלא ראה קודם לכן, אפילו אם ראה מצב דומה(למשל תבניות דומות, אבל בהזזה מהמצב שכן ראה) הוא לא ידע כיצד לפעול.

היינו רוצים מרחב מצבים קטן ככל האפשר. מצד שני, ככל שייצוג המצב פשוט יותר, יותר מידע הולך לאיבוד.

נבחין כי בבואנו לבחור את הפעולה, דיסקיות שאינן חלק מרצף שיכול להביא לנצחון/הפסד, ולא משנה לנו של מי הדיסקיות.

לפיכך בחרנו בייצוג הבא: תחילה מקודדים את מספר הדיסקיות בכל עמודה ולאחר מכן את אורכי הרצפים שאינם חסומים, ואת מיקומם בלוח, הן שלנו והן של היריב.

תוצאות

על מנת להעריך את ביצועיהם של הסוכנים שלנו, החלטנו לממש מספר סוכני baseline, שיהוו מדד להצלחה. הסוכנים שמימשנו לצורך בדיקות הם:

שחקן offensive – מטרתו היחידה היא להשלים רצפים שלו לרצפי נצחון

שחקן defensive- מטרתו היא לחסום את השחקן הקרוב ביותר לנצחון מלנצח

על מנת להימנע ממשחקים דטרמיניסטיים, שחקנים אלה יבחרו בפעולה שמניבה להם ציון גבוה ביותר בהסתברות גבוהה, אך בהסתברות נמוכה יכולים לבחור גם בפעולות אחרות. ההסתברות ממושקלת לפי כמה הפעולה טובה עבור השחקן ומטרתו.

שחקן IBEF2 – שחקן מינמקס עם ההיוריסטיקה החזקה ביותר שהצלחנו למצוא בעבודות קודמות

סוכני מינמקס, ואלפא בטא

ראשית, בדקנו סוכנים אלה במשחק עם שני שחקנים בלבד, בלוח תלת ממדי בגודל (6,7,5). רצינו להראות עד כמה ההיוריסטיקה שלנו חזקה, ולפיכך בחרנו להחליש את הסוכן שלנו ע"י כך שיבצע מהלכים אקראיים בהסתברות מסוימת. הגרף הבא מציג את ביצועי השחקן שלנו אל מול שחקני ה-offensive,defensive כתלות בסיכוי שלו לבצע מהלך אקראי:

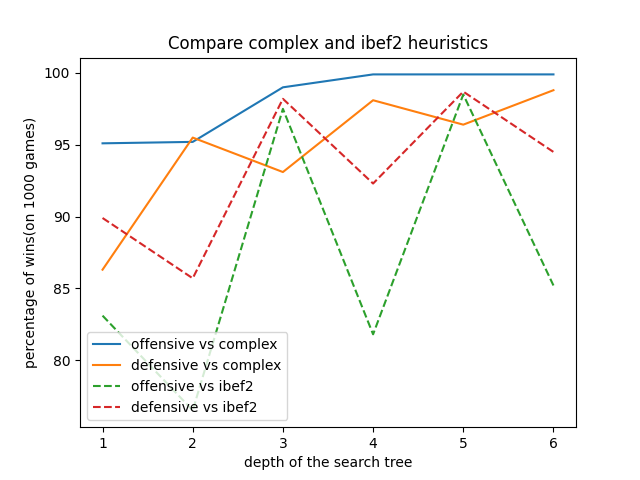
A graph of a performance

Description automatically generated with medium confidenceניתן לראות כי אפילו כאשר החלשנו את הסוכן שלנו עם הסתברות של 0.6 לפעולה רנדומית, עדיין הוא מנצח את היריבים שלו בהסתברות גבוהה מאוד! כמובן שככל שמקטינים ערך זה, הוא מנצח אותם אף יותר. מעניין לראות כי ביצועיו נגד שחקן ה-defensive גבוהים יותר מאשר נגד שחקן ה-offensive – הדבר הגיוני מפני שכאשר הסוכן שלנו מבצע פעולה רנדומית, שחקן ה-offensive שמטרתו היא רק להשלים רצפים ינצל זאת בשביל להתקדם, בעוד שחקן ה-defensive "יבזבז" את הצעד שלו בשביל לחסום, כאשר היה יכול בזמן זה להתקדם לנצחון.

כעת, השווינו את ההיוריסטיקה שלנו אל מול שחקן IBEF2. שחקן זה לא מותאם לשחק בלוח תלת ממדי ולכן לשם ההוגנות עברנו לשחק בלוח דו ממדי בגודל (6,7). בנוסף, השווינו כל אחת מההיוריסטיקות עבור עומקים שונים. מכיוון שהמהלכים שההיוריסטיקות משרות יוצאים יחסית דטרמיניסטיים(נבחר בפעולה שתקבל את הציון הגבוה היותר, ופעמים רבות במשחק יש רק מהלך אחד כזה), בחרנו להשוות כל אחת מההיוריסטיקות אל מול שחקני ה-baseline. להלן התוצאות:

ניתן לראות כי ביצועיו של הסוכן שלנו עקביים יותר, ואחוזי ההצלחה שלו עולים ככל שעומק עץ המינמקס גדל.

לעומתו סוכן ה-IBEF2 מציג תוצאות פחות טובות, ובנוסף פחות עקביות



לצורך קיצור זמני הריצה, ומעתה ועד סוף הדוח, ההשוואות שנערוך יהיו על לוחות דו ממדיים- עדכון מפות האקטיבציה עבור הרצפים בכל הכיוונים לוקחת זמן רב יותר בלוח תלת ממדי מפני שיש יותר רצפים להתחשב בהם. לאחר שבדקנו את נכונותן, היה לנו חשוב יותר להעריך את ביצועיהם של אלגוריתמי הסוכנים שלנו במספר גדול יותר של משחקים.

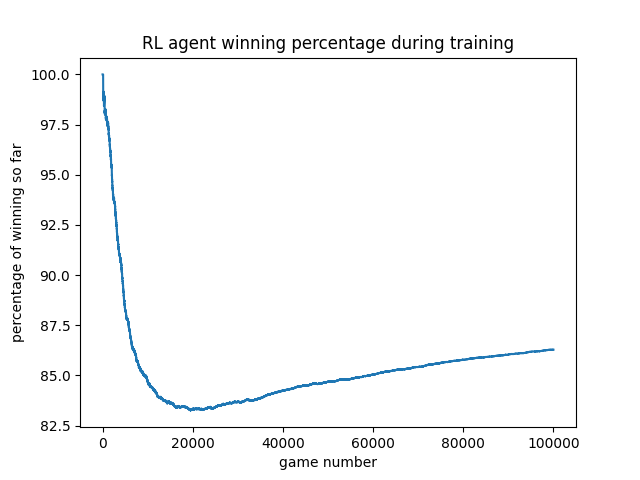
עתה, עברנו לבחון את ההיוריסטיקה שלנו במשחק עם 3 משתתפים:

להוסיף גרפים והתייחסויות לכך!

זמני ריצה:

סוכן ה-Q-learning

סוכן זה, בשונה מהסוכנים הקודמים, נדרש לשלב של למידה, בו ה-Q-table שלו תתעדכן. בחרנו לאמן את הסוכן שלנו לאורך 100000 משחקים, כאשר ה-exploration rate דועך למספר קטן ביותר לאורך האיטרציות- כלומר אנו רוצים שבתחילת האימון הסוכן יסתמך אך ורק על המאסטר שלו, כך הטבלה תתמלא בערכים שמייצגים בחירות מושכלות עבור מצבי לוח שונים. אך לאט לאט המאסטר ישחרר את המושכות וייתן לסוכן לבצע החלטות, ומכיוון שאנו מצפים שבטבלה הצטבר מספיק מידע משמעותי, הסוכן יבצע בהתחלה החלטות טובות יותר ויותר.

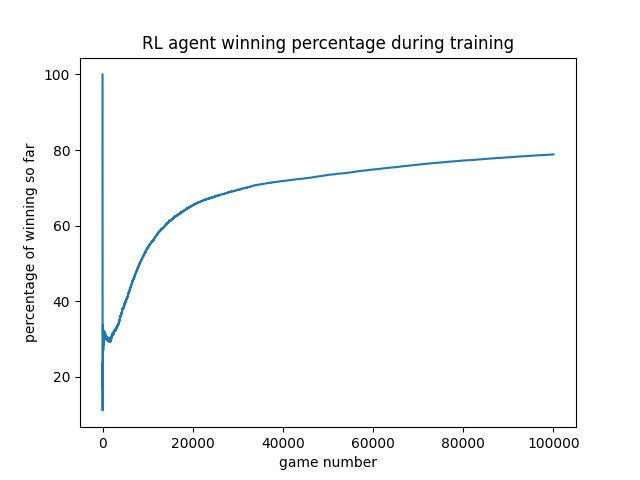
הגרף הבא מציג את את אחוז הנצחונות כתלות בשלב האימון:

ניתן לראות כי בתחילת האימון הסוכן מנצח בהסתברות גבוהה. זאת מכיוון שהסוכן משתמש במאסטר שלו על מנת לבצע מהלכים. מכיוון שבחרנו במאסטר שיהיה טוב יותר מהיריב של הסוכן, אכן אנו מנצחים בהסתברות גבוהה במשחקים.

לאורך 30000 המשחקים הראשונים רואים דעיכה באחוז הנצחונות המצטבר- מכיוון שה-exploration rate הולך וקטן, המאסטר עוזר לסוכן פחות ופחות, והסוכן מקבל החלטות בעצמו אבל הטבלה שלו עדיין לא מכילה מספיק ערכים לגבי לוחות מסוימים, וגם הערכים עצמם עדיין לא מספיק מייצגים נכונה את המציאות.

לאחר מכן, ניתן לראות עלייה באחוז הנצחונות- הטבלה של הסוכן כבר מכילה ערכים טובים יותר, ולכן הוא מצליח לנצח ביותר משחקים, ובנוסף ממשיך לעדכן את הטבלה שלו עם ערכים יותר ויותר מדויקים למדיניות האופטימלית.

כפי שהסברנו במתודולוגיה, הנחנו כי סוכן שהתאמן עם מאסטר ייתן ביצועים טובים יותר מאשר סוכן שהתאמן בלי מאסטר, מכיוון שהוא בונה את הטבלה שלו על סמך נתונים שמשקפים נכון יותר את המציאות, ואיזה לוחות טובים ואיזה לא. על מנת לבדוק את ההנחות שלנו בצורה מעשית, אימנו שני סוכנים עם פרמטרים זהים כמעט לחלוטין,למעט העובדה שאחד התאמן עם מאסטר ואחד ביצע במהלך האימון פעולות רנדומיות. נתנו להם לשחק נגד מספר יריבים שונים, להלן התוצאות:

קודם כל הוספתי גרף של אימון של RL עם מאסטר שהוא שחקן רנדומי, כנראה נמחק את הגרף הזה אבל סתם שיהיה לנו עכשיו:

ניתן לראות כי הסוכן שהתאמן עם מאסטר הצליח להשיג תוצאות גבוהות יותר, מול כל היריבים נגדם שיחק, בהשוואה לסוכן שהתאמן בלי מאסטר.

נקודה מעניינת נוספת היא שמול סוכן אלפא בטא עם היוריסטיקת complex בעומק 2 , הסוכן שהתאמן עם מאסטר ניצח בהסתברות גבוהה, בעוד הסוכן שהתאמן בלי מאסטר הפסיד ביותר מחצי מהמשחקים!

נגד סוכן אלפא בטא עם היוריסטיקת complex בעומק 4, ונגד שחקן ה-defensive , שני סוכני ה-Q-learning לא היוו יריב ראוי. הטבלה מתעדכנת בקלות כאשר הסוכן מנצח והמידע מפועפע מטה, אך יותר קשה לעדכן את הטבלה ולהבין שישנם שחקנים שינסו גם לחסום את הסוכן מלנצח.

מכך ניתן להבין שאימון סוכן Q-learning היא בעיה מורכבת, וכאשר משחקים מול שחקנים מורכבים יותר, הטבלה של הסוכן צריכה להכיר מספר גדול ביותר של מצבים, כדי לדעת איך להתמודד בכל תרחיש.

ביביליוגרפיה:

* <https://www.cs.huji.ac.il/course/2021/ai/projects/old/4InRow_2.pdf>
* <https://www.cs.huji.ac.il/course/2021/ai/projects/old/4InRow_1.pdf>
* <https://en.wikipedia.org/wiki/Connect_Four>
* <https://www.fierz.ch/strategy1.htm>

דברים להוסיף ולערוך:

1. מעבר לרשת CNN: צריך ליצור דאטה(נריץ מונטה קרלו מספר רב של פעמים ממצבי לוח, על מנת לקבל עבורם ערך. זה יהיה הלייבל של הלוח, שרשת הנוירונים תיתקל בה במהלך האימון.   
   נאמן רשת CNN שתקבל את הדאטה סט הנ"ל, ותפיק עבור כל לוח כזה ציון. סוכן הCNN בהנתן מצב לוח כלשהו, ישחק את כל המהלכים האפשריים עבורו, ויבחר במהלך שנותן לוח עם ציון מקסימלי.