מכון טכנולוגי חולון
Holon Institute of Technology

**51095 למידה מונחית חיזוקים** **סמסטר אביב תשפ"ו**

# פרויקט סופי-

## אלגוריתמים מתקדמים: PPO, MCTS ו-Dyna-Q

מטרת פרויקט זה היא לאפשר לכל סטודנט לבחור בעיית למידה מונחית חיזוקים, לממש אלגוריתמים מתאימים מאפס, לבצע ניסויים שיטתיים, ולהציג ניתוח ביקורתי של התוצאות. הפרויקט נועד לבחון לא רק את איכות התוצאה הסופית, אלא בעיקר את היכולת להגדיר בעיית RL, לבחור ייצוגים ומדדים מתאימים, לתכנן ניסוי, להסביר תופעות שנצפו בתהליך הלמידה ולהסיק מסקנות.

**<u>בחירת נושא הפרויקט</u>**

במסגרת פרויקט זה יש לבחור אחד משלושת האלגוריתמים הבאים, או להשוות בין שניים מהם באותה סביבה, בהתאם לרמת הקושי ולסביבת הניסוי:

- PPO - Proximal Policy Optimization: אלגוריתם Actor-Critic מודרני ממשפחת Policy Gradient. האלגוריתם נלמד בקצרה בכיתה, ולכן בחירה בו מחייבת לימוד עצמי והסבר ברור של הרעיון התאורטי והמימוש.
- MCTS - Monte Carlo Tree Search: אלגוריתם חיפוש מבוסס סימולציות, המתאים במיוחד לסביבות דיסקרטיות, משחקים, גרידים ובעיות תכנון.
- Dyna-Q: אלגוריתם Model-Based RL המשלב למידה מתוך ניסיון אמיתי עם תכנון על סמך מודל נלמד של הסביבה.

ניתן לבחור אלגוריתם אחד ולחקור אותו לעומק, או לבצע השוואה בין שני אלגוריתמים, למשל Dyna-Q מול Q-Learning, MCTS מול מדיניות אקראית/חמדנית, או PPO מול Actor-Critic בסיסי.

**<u>בחירת סביבת הניסוי</u>**

ניתן לבחור סביבות מוכרות מחבילות כגון Gymnasium, PettingZoo או MiniGrid, לבנות סביבת Grid World עצמאית, או להציע בעיה מקורית אחרת באישור המרצה. על הנושא להיות ברמת מורכבות המתאימה לפרויקט סופי בקורס.

- עבור PPO מומלץ לבחור סביבות כגון CartPole, MountainCar, LunarLander או סביבה פשוטה עם מרחב פעולות בדיד/רציף.
- עבור MCTS מומלץ לבחור סביבות דיסקרטיות כגון GridWorld, Tic-Tac-Toe, Connect-Four מוקטן, MiniGrid פשוט או משחק דטרמיניסטי אחר.

• עבור Dyna-Q מומלץ לבחור GridWorld, Windy GridWorld, Maze, Taxi-v3, FrozenLake או סביבה עצמאית שבה ניתן ללמוד מודל מעבר ותגמול.

בכל מקרה יש להגדיר באופן ברור: מצבים, פעולות, פונקציית תגמול, תנאי סיום וקריטריוני הצלחה.

**מסלולי עבודה אפשריים**

• מסלול א'- מימוש אלגוריתם מתקדם: בחירה באחד מהאלגוריתמים PPO, MCTS או Dyna-Q, מימושו מאפס וניתוח ביצועיו.

• מסלול ב'- השוואה מול baseline: השוואת האלגוריתם המתקדם מול אלגוריתם פשוט יותר שנלמד בקורס, כגון Q-Learning, SARSA, REINFORCE, Actor-Critic בסיסי או מדיניות אקראית/חמדנית.

• מסלול ג'- חקר רכיב באלגוריתם: לדוגמה השפעת מספר סימולציות ב-MCTS, השפעת מספר planning steps ב-Dyna-Q, או השפעת clip parameter, entropy coefficient ו-GAE ב-PPO.

**דוגמאות לפרויקטים אפשריים**

הדוגמאות הבאות נועדו להמחיש את רמת הציפייה בלבד. ניתן לבחור אחת מהן, לשנות אותה, או להציע פרויקט אחר.

| נושא | תיאור |
|---|---|
| MCTS במשחק פשוט | מימוש MCTS עבור Tic-Tac-Toe, Connect-Four מוקטן או GridWorld. יש להשוות מול מדיניות אקראית, greedy policy או minimax פשוט כאשר הדבר מתאים. יש לבחון את השפעת מספר הסימולציות ועומק ה-rollout. |
| Dyna-Q ב־GridWorld או Taxi | השוואת Dyna-Q מול Q-Learning או SARSA. יש לבחון את השפעת מספר צעדי התכנון, איכות המודל הנלמד, מהירות ההתכנסות וכמות הניסיון האמיתי הנדרשת. |
| MCTS + Value Function | פרויקט מתקדם: שימוש בפונקציית ערך פשוטה כדי להעריך מצבים בעלים של עץ החיפוש במקום rollout אקראי בלבד. |

| | |
|---|---|
| Dyna-Q עם מודל לא מושלם | בחינה של ביצועי Dyna-Q כאשר המודל הנלמד חלקי, רועש או מתעדכן באיחור. מתאים לדיון על Model-Based RL לעומת Model-Free RL. |

**דרישות מימוש**

- יש לממש את רכיבי האלגוריתם המרכזיים בעצמכם. אין להשתמש במימושים מוכנים של האלגוריתמים מספריות RL ייעודיות כגון Stable-Baselines3.
- מותר להשתמש בספריות כלליות כגון NumPy, PyTorch, Matplotlib ו-Gymnasium.
- יש להפריד בקוד בין הגדרת הסביבה, האלגוריתם, האימון, ההערכה וההצגה הגרפית.
- יש להריץ מספר ניסויים מספיק כדי לתמוך במסקנות. רצוי להשתמש במספר seeds ולהציג ממוצע וסטיית תקן כאשר הדבר רלוונטי.
- יש לתעד את הבחירות המרכזיות: ייצוג מצב, מרחב פעולות, פונקציית תגמול, hyper-parameters, תנאי עצירה וקריטריוני הצלחה.

**מדדים וגרפים נדרשים**

על כל פרויקט לכלול מדדים וגרפים המאפשרים להעריך את הלמידה ולהשוות בין חלופות. לכל הפחות יש לכלול:

- גרף reward מצטבר או ממוצע נע לאורך episodes.
- גרף או טבלה המתארים זמן התכנסות או מספר צעדים עד להשגת יעד מוגדר.
- השוואה כמותית בין האלגוריתמים, האסטרטגיות או ההגדרות שנבחנו.
- ניתוח יציבות הלמידה, למשל שונות בין ריצות, סטיית תקן, או תנודות ב-reward.
- דוגמה למדיניות שנלמדה: מסלול בגריד, סרטון קצר, גרף פעולות, או המחשה אחרת המתאימה לסביבה.

**מבנה מומלץ למחברת**

- תיאור הבעיה והסביבה שנבחרה.
- הגדרת מצבים, פעולות, תגמולים ותנאי סיום.
- תיאור האלגוריתמים או האסטרטגיות שנבחרו.
- פירוט hyper-parameters ותכנון הניסויים.
- קוד האימון וההערכה.
- גרפים וטבלאות תוצאות.
- ניתוח התוצאות, מגבלות העבודה ומסקנות.

- רעיונות לשיפור עתידי.

**הגשה, הצגה ובחינה**

- את הפרויקט ניתן לבצע ביחידים בלבד.
- יש להשתבץ, **עד ליום 10.6.2026**, לבחינה בעל-פה במועד שנקבע לשם כך (ניתן לתאם מועד חלופי).
- במועד אליו נרשמתם, תתבקשו להציג את הפרויקט בצורה תמציתית (20 דקות).
- לצד זאת, יכול ותישאלו שאלות תאורטיות ופרקטיות, לגבי כלל העבודות בקורס. לכן, עליכם להצטייד גם בקוד של מטלות 2-1 (וכמובן לשלוט בכל רזיו).
- יש להגיש במודול, עד לבוקר ההצגה, מחברת Jupyter הכוללת פתרון שלם ועובד. להלן הדגשים למחברת זו:
  - המחברת תכלול את כל הקוד שכתבתם. יש לארגן את הקוד בצורה שניתן יהיה לעקוב אחריו בנקל.
  - המחברת תכלול תאי טקסט להסברים של מה שנעשה, ניתוח התוצאות והמסקנות.
  - גרפים יכללו כותרות ומקרא, ובנוסף הסבר לגרף.
  - יש להגיש את המחברת לתיבת ההגשה המתאימה במודול. יש לרשום שם + מס' תעודת זהות בראש המחברת.
- קריטריונים לציון:

| מרכיב | דגשים |
|---|---|
| הגדרת הפרויקט והשאלה המחקרית | בהירות הבעיה, התאמת הסביבה, מקוריות ורמת מורכבות. |
| מימוש האלגוריתמים | מימוש עצמאי, קוד מסודר, טיפול נכון במצבים, פעולות ותגמולים. |
| תכנון ניסויים | בחירת מדדים, seeds, hyper-parameters והשוואה הוגנת בין חלופות. |
| ניתוח תוצאות | הצגת גרפים ברורים, הסבר לתופעות שנצפו, מסקנות מבוססות ולא רק תיאור טכני. |
| הבנה | יכולת להסביר את האלגוריתמים, את הקוד ואת ההחלטות שהתקבלו במהלך העבודה. |

**הערות חשובות**

- פרויקטים דומים מדי או שימוש בקוד מוכן ללא הבנה והסבר עלולים להיפסל.
- מותר להיעזר במקורות חיצוניים לצורך לימוד, אך יש לציין זאת במחברת ולהראות הבנה מלאה של הקוד והאלגוריתם.

- אם חסר נתון לצורך הגדרת הניסוי, יש לקבוע אותו באופן סביר, לנמק את הבחירה ולבדוק את השפעתו במידת האפשר.
- בכל מקרה של ספק לגבי התאמת הנושא, מומלץ לאשר את הרעיון מראש מול המרצה.