למידה באמצעות "חיזוקים"

מטלת אמצע הקורס – תשפ"ב – סמסטר ב'

סטודנטים יקרים,
זוהי מטלת אמצע הקורס "למידה באמצעות "חיזוקים".

המטלה תבחן את השימוש בכלים ובידע התאורטי שרכשתם עד כה במהלך הקורס.
המטרה היא ש"תגעו" בנקודות חשובות שנלמדו עד כה, דרכם תפגינו את הידע שרכשתם ותחוו גם את הפרקטיקה של מימוש האלגוריתמים. כמו כן ש"תדברו" בשפת ה-RL, תנתחו את הבעיה וכמובן שגם תשיגו הצלחה בחלק של האימון.

המטלה בנויה משני חלקים, תאורטי ותכנותי.
להלן השאלה התיאורטית:

נתונה בעיית grid-world הבאה. המטרה להגיע לפינה שמאלית עליונה או לפינה ימנית תחתונה.

Action: UP, DOWN, RIGHT, LEFT
Rewards: -1 על כל תנועה (גם תנועה ש"נתקעת" בקיר ונשארת באותו מקום
Transition Model: דטרמיניסטי
Discount factor $\gamma = 1$
Random Policy: הסתברות של 0.25 לכל כיוון

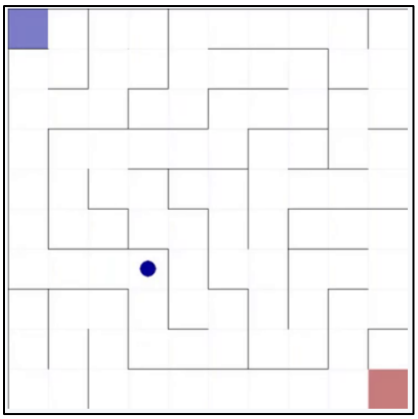
0	S1	-20	-22
-14	-18	-20	S2
-20	-20	-18	-14
-22	-20	-14	0

נתונים בצורה חלקית ערכי ה value למצבים מסוימים (לפי policy הנוכחי)
חשבו (ידנית) את ערכי ה Value (לפי policy הנוכחי)
עבור מצבים: S1, S2

פרטו את דרך החישוב וצרפו את הפתרון לתיבת ההגשה בקובץ ששמו grid_world.pdf.

להלן החלק התכנותי של המטלה:

הבעיה שתעסקו בה היא בעיית המבוך.
בהינתן סביבה של מבוך בגודל NXN, המטרה שלכם היא לאמן סוכן (הנקודה הכחולה) על הסביבה כך שימצא את הנתיב הקצר ביותר מנקודת ההתחלה (פינה שמאלית עליונה, מסומנת בריבוע כחול) לנקודת הסיום (פינה ימנית תחתונה, מסומנת בריבוע אדום).



Action space:

הסוכן שלנו יכול לבחור לנוע **למעלה**, **למטה**, **ימינה ושמאלה ("N", "S", "E", "W").** אם הדרך חסומה, הוא יישאר באותו המיקום.
הסביבה שתעבדו בה היא **סביבה סטוכטית, כאשר הסוכן נוקט פעולה מסוימת קיימת הסתברות של 0.9 שהסביבה תרשה לו להגיע למצב שהוא התכוון להגיע אליו.**
לדוגמא, אם הסוכן שלנו רוצה לפנות ימינה, קיימת הסתברות של 0.9 שהוא באמת יפנה ימינה אך קיימת גם הסתברות של 0.1 שהוא ינוע למעלה, למטה או שמאלה (בהסתברות שווה לכל אחד משלושת הכיוונים)
(הסביבה במחברת המצורפת הינה דטרמיניסטית, עליכם להוסיף במימוש את החלק הסטוכטי המתואר לעיל)

Observation space:

ה Observation space הוא הקואורדינטה (x, y) של הסוכן. התא השמאלי העליון הוא (0, 0).

Reward:

Reward של 1 ניתן כאשר הסוכן מגיע למטרה. עבור כל צעד במבוך, הסוכן מקבל Reward של: $\frac{-0.1}{\text{מספר התאים}}$

זאת הצעה ראשונית כמובן: אנו מצפים מכם "לבחון" השפעות ה Rewards

End condition:

המבוך מתאפס כאשר הסוכן מגיע לנקודת הסיום (פינה ימנית תחתונה, מסומנת בריבוע אדום).

הוראות המטלה:

המטלה בנויה משלושה חלקים:

1. בחלק הראשון תצטרכו לפתור סביבה בגודל 15X15, את הסביבה אתם תצטרכו לפתור עם שני אלגוריתמים שונים:
   a. Monte Carlo
   b. Q-Learning

   המטרה היא להפעיל Polices שונים וכמובן לפתור **בכמה שפחות Episodes.**

2. בחלק השני של המטלה תצטרכו לפתור סביבה בגודל 30X30, המשימה הזאת היא מעט יותר קשה, לכן ניתנת לכם האופציה להגדיר שני תאי Rewards (תאי "פרס") במיקומים אסטרטגיים על מנת שיכווינו את השחקן. מיקום התאים וערך ה Rewards נתון לשיקולכם. רק שימו לב, התאים הנ"ל צריכים להיות **קבועים** ולא יהיה ניתן להחליף אותם כל Episode.
   לטובת פתרון הסביבה, תוכלו לבחור בכל אלגוריתם שאתם רואים לנכון, כל עוד הוא נלמד במהלך ההרצאות ולא נכלל בתוך הקטגוריה של Deep reinforcement learning.
   המטרה גם פה היא לפתור את הסביבה **בכמה שפחות Episodes.**

פרטים טכניים:

התבנית של הסביבה כבר מוכנה עבורכם והיא קיימת על גבי מחברת Colab וקבצי npy (המבוכים) המצורפים לתיבת ההגשה.
לטובת טעינת המבוכים תצטרכו להוריד אותם למחשב שלכם ולהעלות אותם למחברת במקום הייעודי לכך.
**** שימו לב: לפעמים קיימות תקלות כאשר עושים שימוש בסביבות של gym בהם נדרשת גרפיקה שהיא מעבר לבסיסית במחברות של Google Colab. לכן אם אתם לא מצליחים להריץ תא מסוים והוא מראה שגיאה, תנסו להריץ אותו בשנית וזה אמור להסתדר.**

מימוש המטלה יתבצע על גבי מחברת Google Colab. שימו לב שאתם רושמים קוד נקי, מסודר ומפרידים בין תאי קוד.

קחו בחשבון כי **קיים איסור** לעשות שימוש בכל מני ספריות "אלגוריתמים מוכנים".

ההגשה תכלול דו"ח ומחברת גוגל קולב. **אנו שמים דגש גדול על שלב הניסויים והדו"ח, גם במידה והגעתם לתוצאות טובות, מחברת שתוגש ללא הסברים מפורטים תוביל להורדה משמעותית בציון.**

בדו"ח עצמו, מאוד חשוב לראות את ההבנה שלכם והיכולת שלכם בבניית הסביבה על פי ההוראות.
התייחסו בדוח להיבטים ולמושגים השונים אותם למדנו. התחילו בניתוח הבעיה (MDP, Stochastic/Deterministic transition model, Exploration-Exploitation, Episodic/Continuous, מס' מצבים וכו')

הסבירו בקצרה את האלגוריתם, התייחסו כמובן להיבטים השונים והציגו בדיווח תוצאות והשוואות מול הניסויים השונים.

הסבירו היכן הצבתם את תאי ה Rewards שמכווינים את השחקן (רק בסביבה של ה30X30 כמובן), הסבירו על ה policy, תוצאות עם Rewards שונים וכו'.. בקיצור, אל תחסכו בהסברים.

את הדוח תגישו בנוסף למחברת, שם הקובץ יהיה report.pdf.

עבור כל אלגוריתם הראו גרף שמראה התכנסות (Rewards לפי מספר Episodes) והציגו שני קטעי וידאו. קטע וידאו ראשון מראה את הסוכן באמצע תהליך האימון (אם לקח לכם 40 אפיזודות הראו איך הוא מנסה לפתור את הסביבה אחרי 20 אפיזודות) והקטע וידאו השני מראה את ה"תוצר המוגמר", לצורך העניין בדוגמא שלנו זה אחרי 40 אפיזודות.

להלן קישור למחברת Google Colab עליה מופיעה הסביבה. תכנסו לקישור ותעשו Copy to drive, כך תוכלו להעתיק את המחברת אליכם ולעבוד בסביבת הדרייב שלכם.
קישור למחברת ראשונית:
https://colab.research.google.com/drive/1Y0eDRyJizfrUlUI5SnG8pwOihCshU7GE?usp=sharing

לתיבת ההגשה מצורף קובץ rar המכיל שלושה קבצים:

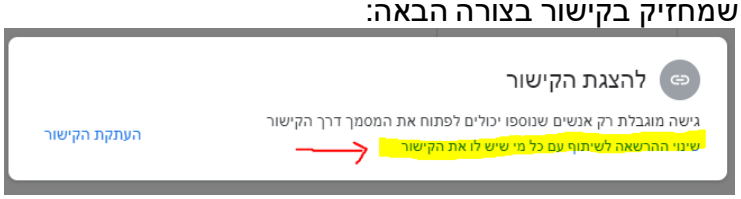
- maze2d_15x15.npy
- maze2d_30x30.npy
- submit.txt

השניים הראשונים הם קבצי npy המכילים את הסביבות של המבוכים שתעבדו עליהם, הקובץ הנוסף הוא קובץ טקסט (submit.txt), זה הקובץ עליו תגישו את המחברות המוכנות בצורה הבאה:

המחברות יהיו מחולקות בצורה מסודרת, יכילו תאי קוד נפרדים ותאי טקסט המסבירים על הפעולות שנעשו.

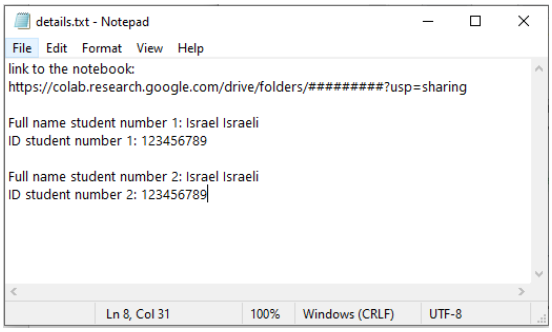
חשוב מאוד – בעת ההגשה, המחברות יכילו את **כל** הפלטים הרלוונטיים לתוצאות האימון**

את המחברת אתם תשתפו מתוך חשבון ה"Google Drive" שלכם, ניתן לייצר שיתוף לכל מי שמחזיק בקישור בצורה הבאה:



תאריך ההגשה הוא ה –. 8/5/2022

ההגשה תתבצע בזוגות כאשר **רק אחד מגיש את המטלה**. יש לציין בקובץ הטקסט ששמו submit.txt את הכתובת למחברת שלכם, את השמות ואת ת.ז המגישים בצורה הבאה:



לתיבת ההגשה תגישו שלושה קבצים בלבד.

1. grid_world.pdf
2. submit.txt
3. report.pdf

לשאלות נוספות:

aviv.german@post.idc.ac.il

בהצלחה!