עבודת תכנות 2

1. חזרו לבעיית ה-black-jack והחליפו באלגוריתם שלכם את השימוש ב-policy-evaluation בשיטה האיטרטיבית המבוססת על משוואות בלמן ב:
   * שימוש ב-monte-carlo policy evaluation. הקוד שלכם צריך לאפשר לבחור בין first-visit MC לבין every-visit MC.
   * שימוש ב-SARSA. אתחלו את ערך SARSA באיטרציה הראשונה של המדיניות הראשונה ל-0. באטרציות הבאות השתמשו בערך של האטרציה הקודמת. השתמשו באלפא=0.1 ובגאמא=1 ובאפסילון = 0.3.

* בצעו 1000 סימולציות (אפיזודות) בכל שלב של הערכת מדיניות.
* התחילו עם המדיניות שהתחלתם בקוד המקורי.
* ציירו גרף שמתאר את הערך של כל מדיניות כפונקציה של 20 האטרציות הראשונות. הכוונה לאטרציה החיצונית שמעדכנת מדיניות. השתמשו בערך שקיבלתם עבורה בשיטה המתאימה
* כדי לחשב את הערך של מדיניות, השתמשו בערך הממוצע של כל הידיים שלשחקן יש קלפים בסכום 13-16 ואין לו אס, ולדילר יש 7 או 8.
* הגרף יכיל 3 עקומות – אחת לכ״א מהשיטות הנ״ל.

1. ממשו את אלגוריתם Q-learning על סביבת frozen-lake. השתמשו בגאמא=0.95. שאר הפרמטרים לבחירתכם על מנת לקבל התכנסות טובה.
   * ציירו גרף של ביצועי האלגוריתם. ציר ה-x יהיה מתאים לאטרציה (כל איטרציה = צעד יחיד במשחק) בעשרות אלפים. ציר ה-y יהיה ערך המדיניות החמדנית בצעד הנ״ל. שימו לב – חישוב ערך המדיניות יתבצע בנפרד. כלומר, כל 10000 צעדים אתם קוראים לפונקציה שתבצע את ההערכה עבור המדיניות החמדנית ביחס לערכי q הנוכחיים. הפונקציה הזאת תשתמש בשיטת first-visit monte carlo כדי להעריך את ערך הפונקציה על המצב ההתחלתי. (אין צורך לשמור ערכים על מצבים אחרים). כלומר, למעשה תריצו סימולציות מהמצב ההתחלתי ותמצאו את סכום הפרסים בכ״א מהן. יש להריץ לפחות 1000 צעדי סימולציה.
   * ציינו בצורה ברורה את ערכי הפרמטרים שבחרתם.

חשוב: הסבירו את האימפלמנטציה שלכם (אפשרי בעזרת ipynb למשל, או בקובץ pdf) ואיך זה מממש את השלבים השונים של האלגוריתם. הקפידו להוסיף את כל הגרפים בקובץ PDF כולל הסבר קצר על כל אחד.