Control Robotics and Machine Learning Laboratory

המעבדה לבקרה רובוטיקה ולמידה חישובית

דו"ח אפיון פרויקט

שם הפרויקט: least Squares Methods dor DRL

מנחה: Tal Daniel

סטודנטים: Tom Norman, Zacharie Cohen

סמסטר: חורף 2019\2020

מטרת המסמך: סיכום תמציתי (כארבעה דפים) הכולל הגדרות מלאות ומעודכנות של הפרויקט ושל הסטאטוס הנוכחי שלו.

תוכן

לו"ז למשימות

מס'	משימה	תאור	משך	הערות
1	קורסי מבוא של Stanford ו David Silver	מעבר על קורסי בסיס ב DL ו RL של Stanford	חודש וחצי	
2	מאמרים מקצועיים	Q learning, DQN, Double and dual DQN, DDPG, D4PG	חודש	
3	מימוש חלק מהמאמרים	מימוש DQN, DDPG, D4PG	חודש וחצי	
4	מימוש LS-D4PG	מימוש האלגוריתם וניסוי על סביבות שונות		

Control Robotics and Machine Learning Laboratory

תרשים התקדמות (גאנט)

פעילות	מספר חודשים ממועד התחלת העבודה								
	9	8	7	6 feb	5 jan	4 dec	3 nov	2 oct	1 sep
1								V	V
2									V
3							V	V	
4									

רקע

בפרויק אנו נתעסק במימוש שיטה שמאמצת את שיטה של למידה בחיזוקים ומתאימה אותה לבחירת החלטה מתוך סט רציף של החלטות (לעומת השיטה הקלאסית שסט ההחלטות הינו בדיד וסופי). החומר רקע הנדרש הינו הבנה עמוקה של כל משפחות האלגורתמים ושיטות שעליהם מתבסס המאמר D4PG עם replay buffer כלומר הבנה עמוקה של כל האלגורתמים.

מטרת הפרויקט

מימוש LS-D4PG עם replay buffer

סביבת עבודה

אנו כותבים את הקוד בפייתון על google colab בסוף הפרויקט נריץ את הקוד על GPU של המעבדה.

תחומי ידע נדרשים

בסיס ב deep learning וב reinforcement learning

Q learning, DQN

שיטות אופטימיזציה

סיכום תמציתי של סקר הספרות

כל המאמרים שקראנו בסקר הספרות הינם שיטות המתבססות אחת על השנייה ללמידה מחיזוקים. כל השיטות משתמשות ב Q-Learning לשיערוך התגמול מפעולה אפשרית. מימוש מהשיטה הבסיסית ביותר (DQN) לשיטה המתקדמת ביותר (D4PG) מאפשר לנו לבנות את האלגוריתמים עם הבנה ומקצר את זמן ה-debugging.

מאמרים:

DQN: Human Level Control Through Deep Reinforcement Learning

אלגוריתם Q-Learning לפעולות ממרחב בדיד, משתמש ברשת נוירונים שהכניסות שלה הם ה state והיציאות הן התגמול הנחזה עבור כל פעולה. משתמש ב replay buffer, זיכרון שמשמש לסוכן ללימוד off-policy. הציג תוצאות מרשימות על משחקי אטרי.

DDQN: Deep Reinforcement Learning with Double Q-learning

אלגוריתם DQN היה "אופטימי" ולכן שינו את השיטה בה מחשבים את התגמול.

DUAL DDQN: Dueling Network Architectures for Deep Reinforcement Learning

שינו את מבנה הרשת ב DQN כדי שתחשב את ה advantage במקום את התגמול.

DDPG: Continuous control with deep reinforcement learning

אלגוריתם Policy Gradient לפעולות ממרחב רציף, משתמש בשיטת actor-critic: ה critic אומר לשחקן כמה התגמול הנחזה עבור כל פעולה, ה actor נותן את הגודל של אותה פעולה (למשל אם הסוכן יכול לשלוט על כמות הגז הנכנסת למנוע אז נקבל מה actor כמה גז להכניס).

LS-DQN: Shallow Updates for Deep Reinforcement Learning

מאמר של CRML, לקחו את השכבה האחרונה של הרשת נוירונים של DQN ועשו לה עידכון בשיטה לינארית ולא gradient descent.

D4PG: Distributed Distributional Deterministic Policy Gradients

אלגוריתם policy gradient מבוסס על DDPG, מבזרים את הסוכנים: כדי להשיג מידע באופן יעיל יותר מתחילים מספר סביבות שונות עם שחקנים, ונותנים להם למלא את ה replay buffer. שיטת חיזוי התגמול שונתה וכעת מחשבת את ההסתברות לתגמול כלשהו עבור פעולה כלשהי. התגמול הנחזה הוא התוחלת של ההסתברות הנ"ל.