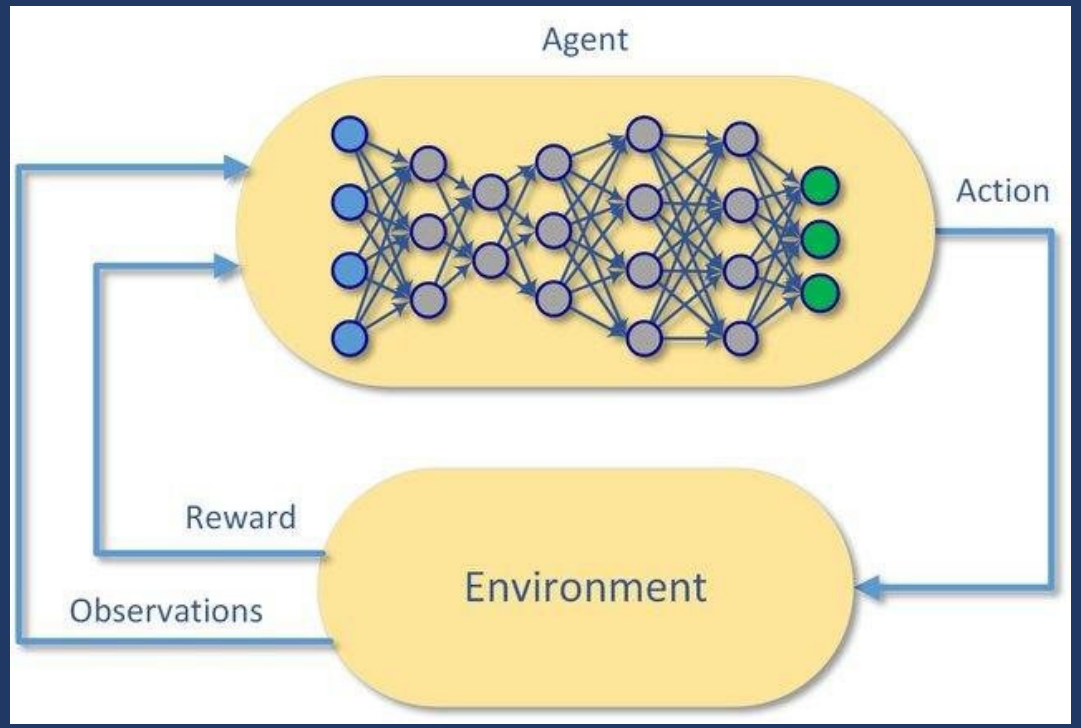
Policy Gradient - Intro

גלעד מרקמן

קריית החינוך
פארק המדע
בית לערכים
למצוינות ולחדשנות



חזרה - Q Learning

- האלגוריתמים שלמדנו התבססו על state value או V(s) state-action value או Q(s,a).
- הערך של V(s) הוגדר כסכום העתידי של התגמולים (rewards) שנקבל בהתאם למדיניות (Policy) נתונה, עם מקדם הפחתה gamma.

$$V_\pi(s_0) = \quad R_0 + \gamma R_1 + \gamma^2 R_2 + \gamma^3 R_3 + \cdots + \gamma^n R_n$$

- הערך של Q(s,a) הוגדר כסכום העתידי של התגמולים (rewards) שנקבל, לאחר שנבצע פעולה a ולאחר מכן נפעל בהתאם למדיניות.

חזרה - Q Learning

- אם ידוע ערך המצב, בהינתן פעולה מסויימת, ניתן לבחור את הפעולה שתיתן לנו את הערך המקסימלי:

$$V^*(s) = \max_a(Q(s,a))$$

$$a^* = argMax_a\big(Q(s,a)\big)$$

- הפעולה המיטבית נבחרה לאחר חישוב ערכי המצבים ומציאת הפעולה שתביא אותנו למצב הבא הטוב ביותר.

חזרה - Q Learning

- פיתוח משוואות בלמן אפשר לנו לעדכן את ערך פונקציית המצב-פעולה כדי לחשב את טבלת הערכים (Q) המיטבית:

- אלגוריתם מונטה קרלו:

$$Q(s,a) = Q(s,a) + \alpha\left(G_t - Q(s,a)\right)$$

$$G_t = R_0 + \gamma R_1 + \gamma^2 R_2 + \gamma^3 R_3 + \cdots + \gamma^n R_n$$

- Temporal Deference ו Bootstrapping:

$$Q(s,a) = Q(s,a) + \alpha\left(R + \gamma Q(s',a') - Q(s,a)\right)$$

$$a' = argmax(Q(s',a))$$

חזרה - Deep Q-Learning

- ברשת נוירונים - עדכון הפרמטרים נעשה באמצעות פונקציית הטעות הבאה, הנובעת ישירות מנוסחאות בלמן:

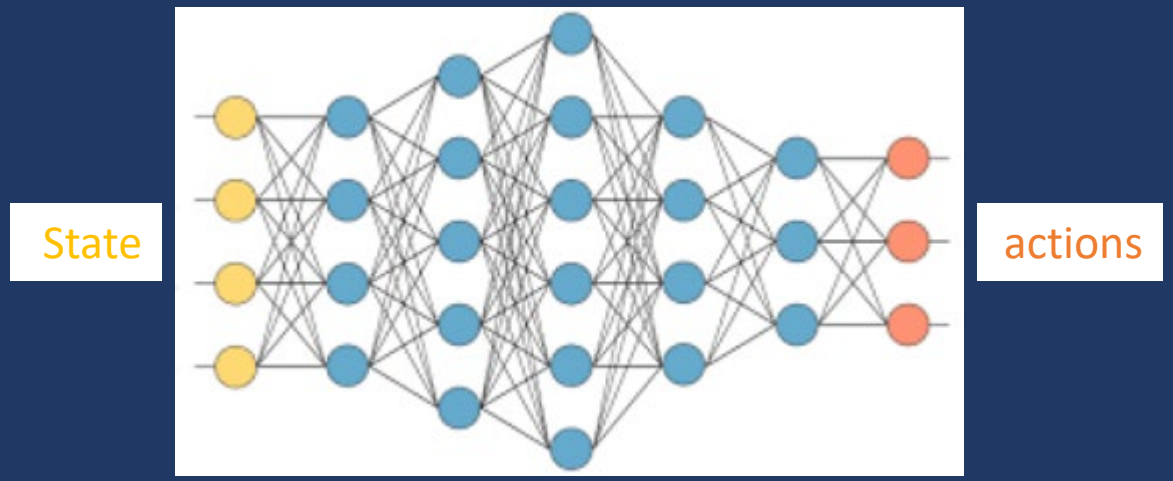
$$loss = \left(R + \gamma \cdot Q(s', a', w) - Q(s, a, w)\right)^2$$

- עדכון הפרמטרים של הרשת נעשה באמצעות SGD במטרה להקטין את הטעות:

$$w = w - \nabla loss * learning\_rate$$

Policy Gradient Methods (PGM)

- כאמור, ב Q-Learning אנחנו מצאנו את הפעולה המיטבית באמצעות חישוב ערכי המצבים, וחיפשנו מהי הפעולה שתביא אותנו למצב עם הערך הגבוה ביותר.
- באלגוריתם של Policy Gradient אנחנו מוצאים את הפעולה המיטבית ישירות באמצעות רשת הנוירונים.
- רשת הנוירונים היא המדיניות שלנו.



אלגוריתמים מבוססי Policy Gradient

בסדרת הרצאות זו נלמד את האלגוריתמים הבאים מבוססי Policy Gradient:

- REINFORCE Monte Carlo - האלגוריתם הבסיסי
  - REINFORCE MC with Entropy Regularization
  - REINFORCE MC Continuous Action Space
- Actor Critic Method – שילוב של Policy Gradient ו Value Base.
  - Actor-Critic – basic one step
  - Actor-Critic with mini batch - n_step
  - A2C – Advantage Actor Critic
- PPO – Proximal Policy Optimization.
- אלגוריתם מתקדם אשר פותח על ידי OpenAI ומשמש לאימון מודלי שפה גדולים כמו ChatGPT.

קריית החינוך
פארק המדע
בית לערכים
למצוינות ולחדשנות

יתרונות וחסרונות PGM
מדיניות אקראית

- PGM יכולים לייצג גם מדיניות אקראית.
- ישנן סביבות בהן מדיניות אקראית היא עדיפה.
- למשל במשחק אבן, נייר או מספריים המדיניות הטובה היותר היא מדיניות אקראית. כל מדיניות אחרת היריב ילמד וינצח אותנו תמיד.

- DQN מבוססת על מציאת הפעולה שתביא לערך המירבי, והיא מתקשה עם מציאת מדיניות אקראית, שלמשל תבחר בין שתי אפשרויות ביחס של 30%-70%.

קריית החינוך
פארק המדע
בית לערכים
למצוינות ולחדשנות

יתרונות וחסרונות PGM
מרחבי פעולה רציפים

- מרחב פעולה בדיד – זוהי סביבה שבה הפעולות הן ספציפיות, כגון: ימינה, שמאלה, ירי וכד'.
- מרחב פעולה רציף – זוהי סביבה שבה הפעולות הן מספריות, כגון: סיבוב הגה ב $31.25°$ או הוספת כוח בערכים שבין [100,100-], כולל שברים.

- אלגורתמים מבוססי PGM יכולים להתמודד עם מרחבי פעולה רציפים, ולבנות מודל שידע לתת לנו פעולה שנעה ברציפות בין ערכים שקבענו.

יתרונות וחסרונות PGM
המדיניות משתנה בצורה חלקה יותר

- בכל השיטות אנחנו מחפשים את המדיניות המיטבית.
- בשיטות מבוססי ערך אנחנו נוקטים בגישה עקיפה – מוצאים את הערך של המצב ולפיו מחפשים את הפעולה שתביא למצב המיטבי.
- בשיטות PGM אנחנו נוקטים בשיטה ישירה - מחשבים ישירות את המדיניות.

- האימון של PGM חלק יותר ורציף יותר.

קריית החינוך
פארק המדע
בית לערכים
למצוינות ולחדשנות

יתרונות וחסרונות PGM
שימוש בזיכרון

- ב DQN אנחנו משתמשים ב Replay Buffer ומנצלים דגימות שונות ומגוונות.
- לעומת זאת, ב PGM אנחנו משתמשים רק בדגימות האחרונות, דבר העלול לפגוע באימון הרשת (גורם לשונות Variance).

DQN – PGM

Q-Learning Methods	Policy Gradient Methods
מרחב פעולה בדיד (דיסקרטי) וקטן	מרחב פעולה רציף (Continuous) או גדול במיוחד
מדיניות דטרמיניסטית	מדיניות אקראית (סטוכסטית)
עדיף במשחקים וסביבות בדידות	עדיף ברובוטיקה ובפעולות הדורשות שליטה רציפה

- הפתרון האופטימלי – שילוב שתי השיטות שנעשה באלגוריתם Actor Critic

Policy Gradient Method פיתוח

- פיתוח רשת נוירונים המייצגת את המדיניות (מקבלת מצב ומחזירה פעולה מיטבית) מחייבת פתרון של הסוגיות הבאות:

1. מבנה רשת הנוירונים – כיצד נבנה את רשת הנוירונים המקבלת מצב (state) ומחזירה פעולה (action).
2. אימון הרשת – כיצד נעדכן את הפרמטרים של הרשת ומהי פונקציית הטעות.
3. כיצד נחקור את הסביבה - Exploration.