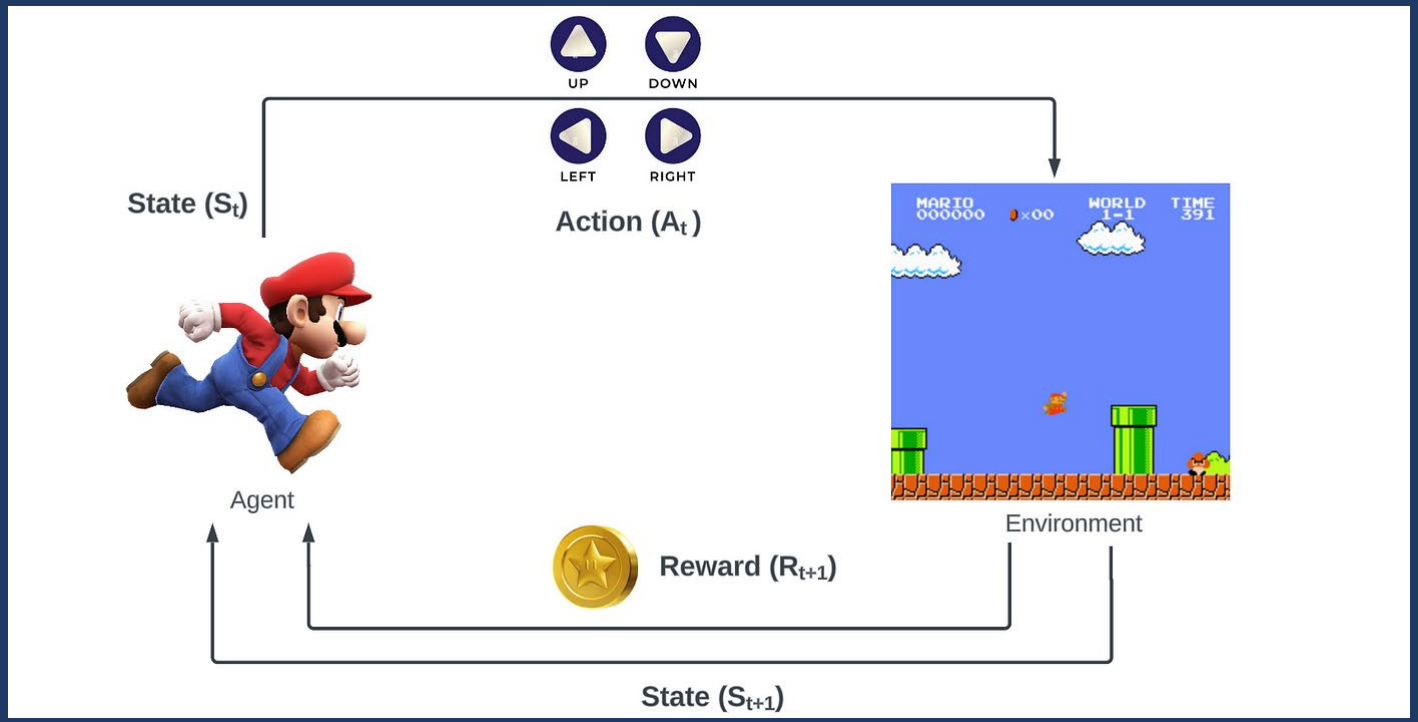
Policy Gradient Method

גלעד מרקמן

קריית החינוך
פארק המדע
בית לערכים
למצוינות ולחדשנות



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.

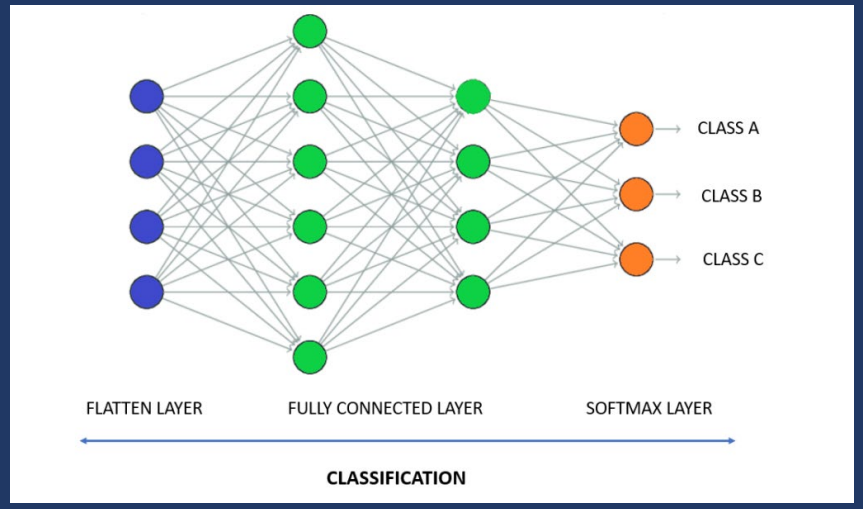
מבנה רשת הנוירונים – פעולות בדידות

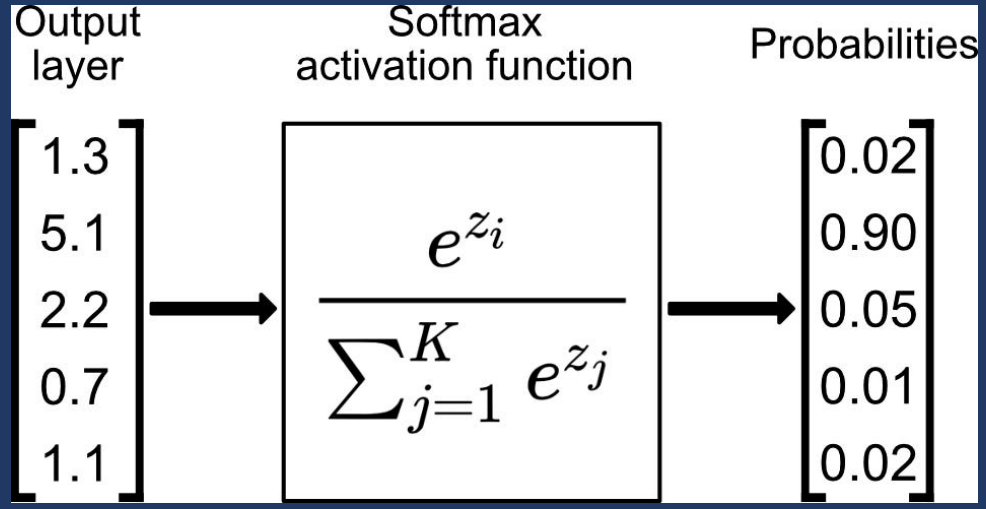
- בהרצאה זו נדון בסביבה בה הפעולות הן בדידות Discrete Actions בהן לסוכן יש מספר סופי של פעולות:
  - כגון: תנועה ימינה, שמאלה, למעלה, למטה, ירי.

- זאת בהבדל מסביבה בה מרחב פעולות רציף Continuous Actions:
  - למשל: תנועה של הגה במעלות בין $[-180^0, 180^0]$, לחיצה על דוושת הגז בעוצמה משתנה וכד'.
  - פתרון לפעולות רציפות יילמד בהרצאות הבאות.

מבנה רשת הנוירונים – מרחב בדיד

- מבנה רשת הנוירונים במקרה של פעולות בדידות הינה רשת קלאסית לסיווג.
  - הקלט הוא המצב (s), לעיתים מכונה observation.
  - הפלט הוא כמספר הפעולות בסביבה, כאשר לכל פעולה מחושבת ההסתברות לבחור בה (באמצעות Softmax) – action_probs.
  - נסמן את הפרמטרים של הרשת באות θ.





Exploration Vs. Explotation

קריית החינוך
פארק המדע
בית לערכים
למצוינות ולחדשנות

- פלט רשת הנוירונים הוא וקטור (טנסור) של הסתברויות action probs, כאשר כל ערך מייצג את ההסתברות לבחור בפעולה מסוימת.

אינדקס	פעולה	הסתברות
0	ימינה	0.5
1	שמאלה	0.25
2	למעלה	0.15
3	למטה	0.05
4	ירי	0.05

- אנחנו נדגום את הסביבה אקראית כך שנבחר בפעולה ימינה ב 50% מהמקרים, בפעולה שמאלה ב 25% וכך הלאה.
- בדרך זו ביצענו Exploration של הסביבה.
- בהמשך נלמד כיצד להגביר את החקר, באמצעות Entropy.

פונקציית הטעות ועדכון הרשת

קריית החינוך
פארק המדע
בית לערכים
למצוינות ולחדשנות

- הבעיה העיקרית שעלינו לפתור היא כיצד נוכל להעריך אם הפעולה שנבחרה על ידי הרשת היא פעולה טובה או רעה, וכיצד נעדכן את רשת הנוירונים על מנת לשפר את בחירת הפעולות.
- הפתרון לבעיה זו ניתן לנו על ידי The Policy Gradient Theorem.
- התאוריה מגדירה פונקציה למדידת ביצועי רשת הנוירונים.
$J(\theta)$
- הפונקציה מקבלת את הפרמטרים של פונקציית המדיניות שלנו (θ) ומחזירה את ממוצע התגמולים שיתקבלו אם נפעל בהתאם למדיניות.

The Policy Gradient Theorem

קריית החינוך
פארק המדע
בית לערכים
למצוינות ולחדשנות

- הפונקציה למדידת ביצועי רשת הנוירונים שווה למעשה לערך המצב ההתחלתי s לפי המדיניות הנתונה.

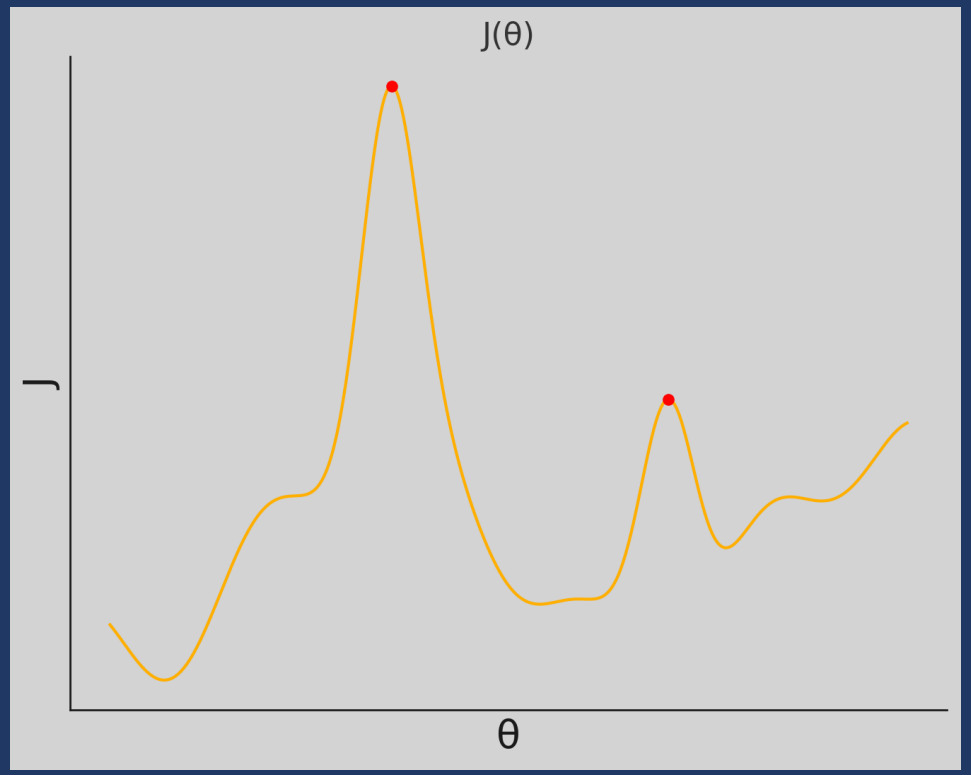
$$J(\theta) = V_{\pi_\theta}(s)$$

- ערך המצב, כיודע, הוא תוחלת התגמולים העתידית (אם פועלים לפי המדיניות):

$$J(\theta) = \mathbb{E}_\pi(R_0 + \gamma R_1 + \gamma^2 R_2 + \gamma^3 R_3 + \cdots + \gamma^n R_n)$$

- כלומר, אם נפעל בסביבה פעמים רבות עד למצב סופי (סיום המשחק), בהתאם למדיניות π_θ נקבל בממוצע סך תגמולים של $J(\theta)$.

המטרה – מקסום $J(\theta)$



- המטרה שלנו למקסם את הביצוע $J(\theta)$, כלומר למצוא את θ אשר תביא את הפונקציה $J(\theta)$ למקסימום.
- אנחנו מחפשים את הפרמטרים θ של רשת הנוירונים, שקובעת את המדיניות, אשר תביא אותנו לתגמולים הכי טובים.
- לצורך כך עלינו לדעת כיצד לחשב את הנגזרת של $J(\theta)$.
- די לנו בנוסחה לחישוב הנגזרת ואיננו זקוקים לפונקציה $J(\theta)$ עצמה.

The Policy Gradient Theorem

קריית החינוך
פארק המדע
בית לערכים
למצוינות ולחדשנות

- The Policy Gradient Theorem קובעת כי הנגזרת של פונקציית הביצוע J שווה לממוצע (תוחלת) של הנוסחה הבאה:

$$\nabla J(\theta) = \mathbb{E}_{\pi}(G_t * \nabla \ln \pi_{\theta}(a|s))$$

$$G_t = Q(s,a) = R_0 + \gamma R_1 + \gamma^2 R_2 + \gamma^3 R_3 + \cdots + \gamma^T R_T$$

- במילים: הנגזרת של פונקציית הביצוע שווה לממוצע של סכום ההחזרים (G_t) כפול הנגזרת של הלוגריתם של הערך (ההסתברות) של הפעולה a שנבחרה על ידי המדיניות π שלנו.
- G_t זהו למעשה חישוב סכום ההחזרים העתידי בהתאם לאלגוריתם Monte Carlo.

פונקציית הטעות ועדכון הרשת

קריית החינוך
פארק המדע
בית לערכים
למצוינות ולחדשנות

- קיבלנו: $\nabla J(\theta) = \mathbb{E}_{\pi}(G_t * \nabla \ln \pi_\theta(a|s))$
- המטרה שלנו למקסם את הביצוע $J(\theta)$, ולכן עלינו למצוא נקודת מקסימום.
- נשתמש באלגוריתם SGA – Stochastic Gradient Ascent. כאשר α הוא מקדם הלימדה.
- נקבל את נוסחת המחיר ונוסחת העדכון הבאים:

$$loss = G_t * \ln \pi_\theta(a|s)$$
$$\theta_{t+1} = \theta_t + \alpha * \nabla loss(\theta)$$

- להוכחה מלאה של הנוסחה ראו: Suton, Barto "Reinforcement Learning An Introduction, 2nd ed., p. 325.

REINFORCE Monte Carlo

קריית החינוך
פארק המדע
בית לערכים
למצוינות ולחדשנות

- האלגוריתם שפיתחנו מכונה REINFORCE Monte Carlo.
- אנחנו דוגמים את הסביבה עד לסיום כל משחק (episode) בהתאם לאלגוריתם Monte Carlo.
- במהלך הדגימה אנחנו שומרים: state, action_prob, reward.
- עם נתונים אילו אין כל בעיה לחשב את סכום ההחזרים בכל מצב t עד לסיום המשחקון (G_t), ולחשב את הטעות והנגזרת שלה.

$$loss = G_t * \ln \pi_\theta(a|s)$$

$$\nabla loss = G_t * \nabla \ln \pi_\theta(a|s)$$

- באמצעות SGA – Stochastic Gradient Ascent נמצא את הפרמטרים θ שיביאו למקסימום של תועלת:

$$\theta_{t+1} = \theta_t + \alpha * \nabla loss(\theta)$$

REINFORCE Monte Carlo

קריית החינוך
פארק המדע
בית לערכים
למצוינות ולחדשנות

- בלמידת מכונה מקובל להשתמש בפונקציות למציאת מינימום ולא מקסימום, לכן, אנחנו נכפיל את נוסחות הטעות במינוס וכך נוכל להשתמש ב SGD :

$$loss = -G_t * \ln \pi_\theta(a|s)$$
$$\theta_{t+1} = \theta_t - \alpha * \nabla loss(\theta)$$

מימוש האלגוריתם

- בהרצאה הבאה נדגים כיצד לממש את האלגוריתם לאימון סוכן הלומד לשחק איקס עיגול.

REINFORCE: Monte-Carlo Policy-Gradient Control (episodic) for π_*

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Algorithm parameter: step size $\alpha > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot,\boldsymbol{\theta})$
  Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
    $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ (G_t)
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \gamma^t G \nabla \ln \pi(A_t|S_t,\boldsymbol{\theta})$

- Sutton, Barto “Reinforcement Learning – An Introduction (2nd ed.), p. 328