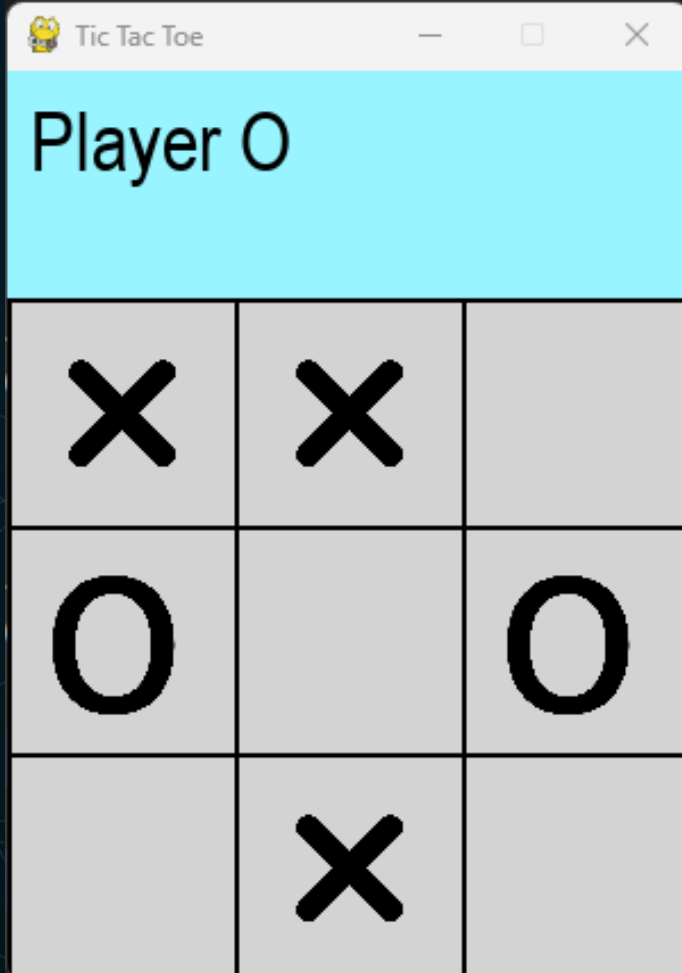




קריית החינוך  
פארק המדע  
בית לערכים  
למצוינות ולחדשנות



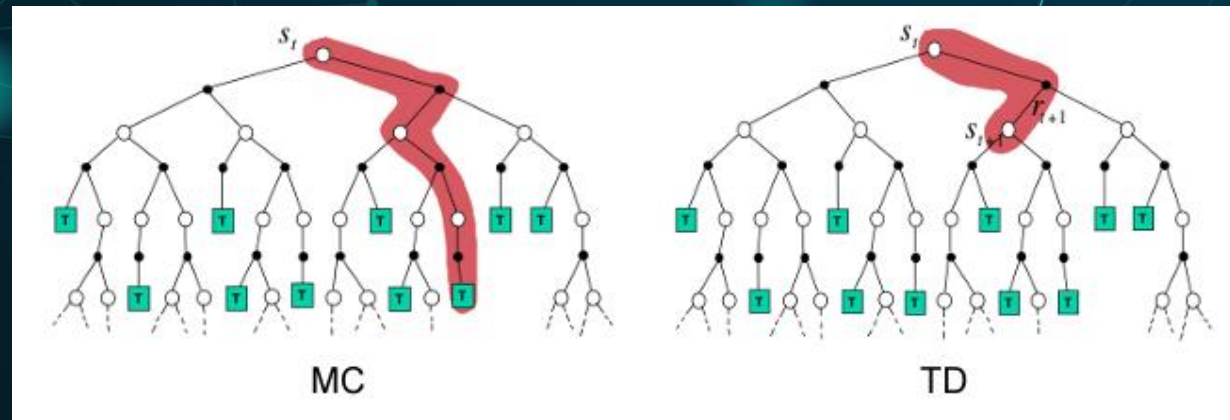
# Tic Tac Toe

Temporal Difference  
SARSA  
Q-Learning

# אלגוריתם SARSA

- הסוכן נמצא במצב  $S$  יבחר את הפעולה  $A$  בהתאם לטבלת  $Q$  ולמדיניות  $\epsilon$ -greedy.
- הסוכן ידגום את הסביבה ויקבל את התוצאות הבאות:  $S, A, R, S', A'$ .
- בחירת  $A'$  תעשה לפי המדיניות הנוכחית שלו  $\epsilon$ -greedy & Q-Table.
- אחרי כל צעד ישתמש הסוכן בנתונים כדי לעדכן את טבלת  $Q$  באמצעות הנוסחה:

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



# פסאודו קוד SARSA

Sarsa (on-policy TD control) for estimating  $Q \approx q_*$

Initialize  $Q(s, a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$ , arbitrarily, and  $Q(\text{terminal-state}, \cdot) = 0$

Repeat (for each episode):

Initialize  $S$

Choose  $A$  from  $S$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)

Repeat (for each step of episode):

Take action  $A$ , observe  $R, S'$

Choose  $A'$  from  $S'$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)

$Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma Q(S', A') - Q(S, A)]$

$S \leftarrow S'; A \leftarrow A';$

until  $S$  is terminal



# דגימת הסביבה

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

במהלך אימון הסוכן אנחנו דוגמים את הסביבה על מנת לקבל את SARSA'.

נחלק את הדגימה לשלבים:

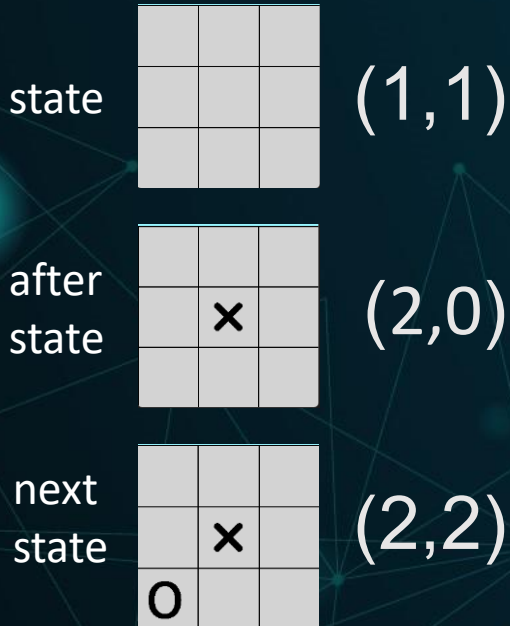
א- הסוכן מקבל מצב S ובוחר פעולה A

ב- הסביבה מבצעת את הפעולה ומחזירה:  $R'$ , afterState

ג- היריב מקבל מצב afterState ובוחר מצב afterAction

ד- הסביבה מבצעת את הפעולה ומחזירה:  $R''$ , nextState

ה- הסוכן מקבל מצב nextState ובוחר את A'



אילו נתונים נכנסים לנוסחה שלנו?

- תלוי אם אנחנו במצב סופי אם לאו?

# דגימת הסביבה – ללא מצב סופי

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



$R'=0$

ב- הסביבה מבצעת את הפעולה ומחזירה:  $R'$  afterState

$R''=0$

ד- הסביבה מבצעת את הפעולה ומחזירה:  $R''$  nextState

א- הסוכן מקבל מצב S ובוחר פעולה A

ג- היריב מקבל מצב afterState ובוחר מצב afterAction

ה- הסוכן מקבל מצב nextState ובוחר את A'

$$s=state; a =A; s' = nextState; a' = A'$$

# דגימת הסביבה – מצב סופי בשלב ב'

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

state (0,0)

|  |   |   |
|--|---|---|
|  |   | 0 |
|  | x | 0 |
|  |   | x |

א- הסוכן מקבל מצב S ובוחר פעולה A

after state

|   |   |   |
|---|---|---|
| x |   | 0 |
|   | x | 0 |
|   |   | x |

ב- הסביבה מבצעת את הפעולה ומחזירה:  $R'=1$  afterState, R'

~~ג- היריב מקבל מצב afterState ובוחר מצב afterAction (2,0)~~

~~ד- הסביבה מבצעת את הפעולה ומחזירה:  $R''=0$  nextState, R''~~

~~ה- הסוכן מקבל מצב nextState ובוחר את A' (2,2)~~

$$s = state; a = A; r = R'; \gamma Q(s', a') = 0$$



# דגימת הסביבה – מצב סופי בשלב ג'

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

|       |   |   |   |
|-------|---|---|---|
| state | 0 | x | x |
|       |   | 0 |   |
|       |   |   |   |

(2,1)

א- הסוכן מקבל מצב S ובוחר פעולה A

|             |   |   |   |
|-------------|---|---|---|
| after state | x | x | 0 |
|             |   | 0 | x |
|             |   |   |   |

(2,0)

ב- הסביבה מבצעת את פעולה ומחזירה:  $R' = 0$  afterState

ג- היריב מקבל מצב afterState ובוחר מצב afterAction

|            |   |   |   |
|------------|---|---|---|
| next state | x | x | 0 |
|            |   | 0 | x |
|            | 0 |   |   |

(2,2)

ד- הסביבה מבצעת את הפעולה ומחזירה:  $R'' = -1$  nextState

~~ה- הסוכן מקבל מצב nextState ובוחר את A'~~

$$s = state; a = A; r = R''; \gamma Q(s', a') = 0$$

# הדגמה - משחק איקס עיגול

• נדגים יישום של האלגוריתם SARSA במשחק איקס עיגול כשהשחקן הוא X.

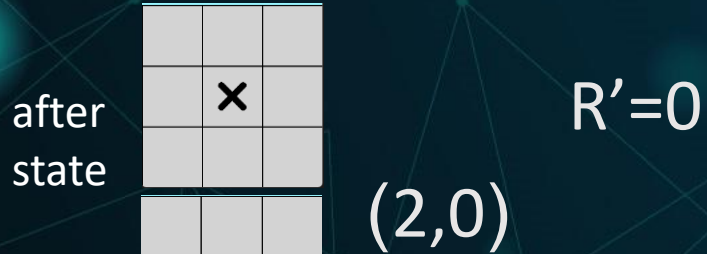
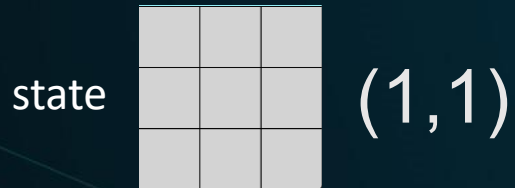
## תרגיל

- עליכם ליישם את האלגוריתם של Q-learning במשחק איקס עיגול.
- עליכם ליישם את אחד האלגוריתמים על שחקן O.



# afterState

במקרים בהם חלק מהמודל ידוע לנו ניתן להתאים את האלגוריתמים וליעל אותם.  
במשחקים ידוע לסוכן מה יהיה מצב הלוח לאחר הפעולה שהוא ביצע. זה אינו המצב הבא, אלא נכנה אותו afterState  
א- הסוכן מקבל מצב S ובוחר פעולה A



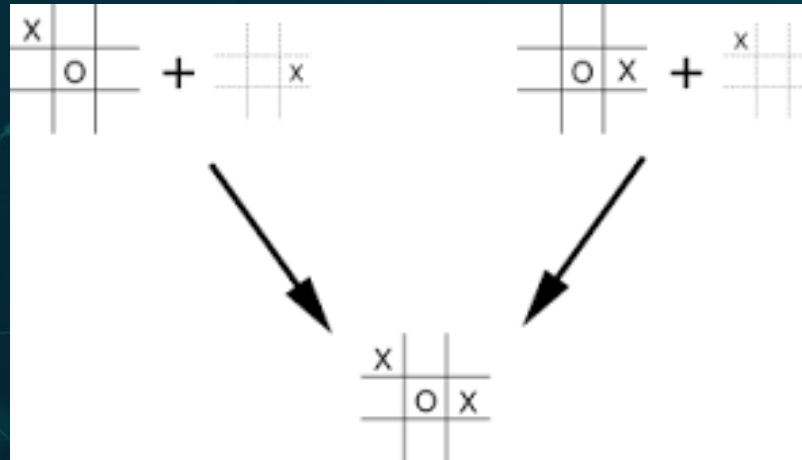
ב- הסביבה מבצעת את הפעולה ומחזירה: afterState,  $R'$   
ג- היריב מקבל מצב afterState ובוחר מצב afterAction  
ד- הסביבה מבצעת את הפעולה ומחזירה: nextState,  $R''$

afterState הוא למעשה (state, action) וניתן להשתמש בטבלה  $V(\text{afterState})$ .  
 $Q(S, A) = V(\text{afterState})$

# טבלת After State

$$V(\text{after}_S) = V(\text{after}_S) + \alpha(R + \gamma V(\text{after}_{S'}) - V(\text{after}_S))$$

- השימוש בטבלת after state הוא יעיל יותר מטבלה Q שכן מונע כפילויות.
- יתכן ששני זוגות (S,A) ו-(S',A') יתנו לנו את אותו מצב - afterState:



# תרגיל

$$V(\text{after}_S) = V(\text{after}_S) + \alpha(R + \gamma V(\text{after}_{S'}) - V(\text{after}_S))$$

עליכם ליישם את אלגוריתם SARSA באמצעות טבלת afterState  
במקום טבלת Q.  
לשים לב – מהו afterS' וכיצד ליצור אותו.