חטיבת מערכות חימוש
והמרכז הלאומי לטחוש וסיסמולוגיות
שדה הקרב האלקטרוני והקיברנטי - מסמ"א

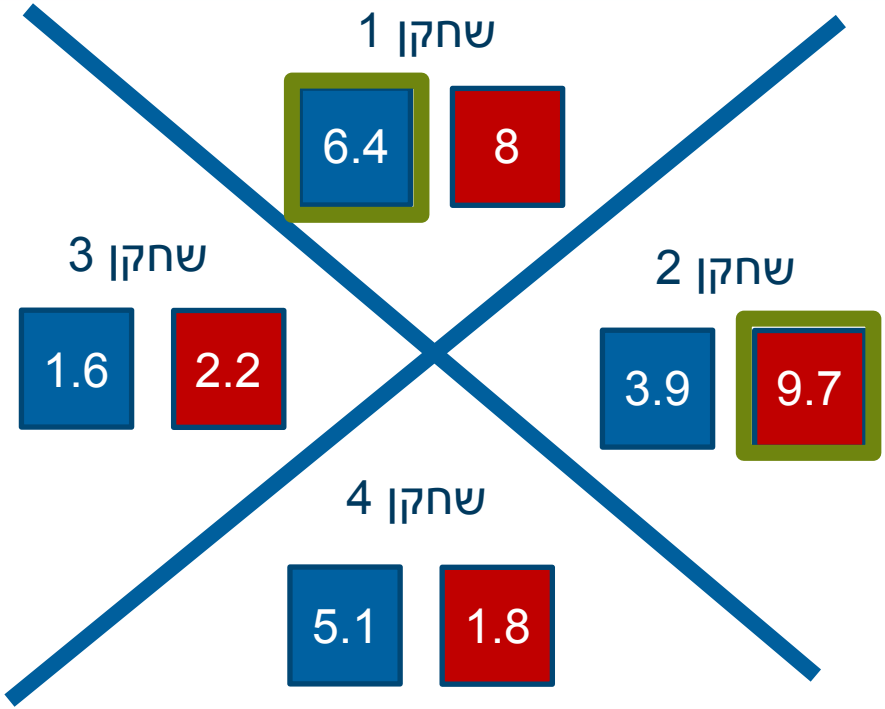
למידת חיזוקים עמוקה מרובת משתמשים ברשתות תקשורת

ד"ר אושרי נפרסטק

רפאל
מערכות לחימה מתקדמות בע"מ

בעיית צעצוע

אתגר לימוד אסטרטגיה במשחק

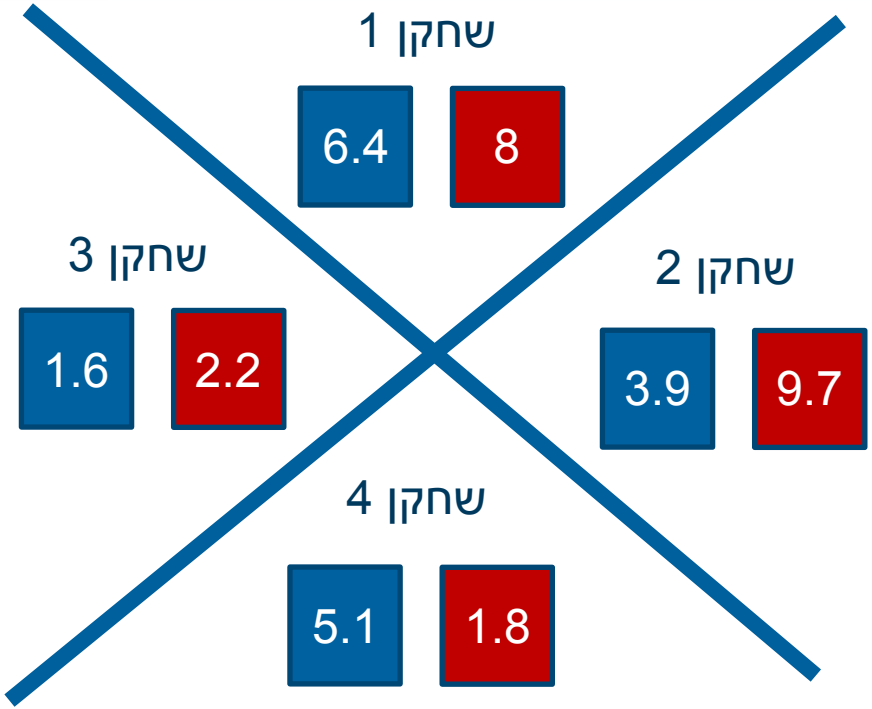


- ארבעה שחקנים ללא יכולת לתקשר ביניהם.
- לכל שחקן שני כפתורים, כחול ואדום.
- משחקים 20 פעמים את המשחק.
- לכל שחקן יש תגמול על כל כפתור שמפולג אחיד בין 0 ל 10.
- בכל שלב כל שחקן צריך להחליט האם ללחוץ או לא ללחוץ ועל איזה כפתור.
- אם רק שחקן אחד לוחץ על כפתור בצבע מסוים, הוא מקבל תגמול.
- אם יותר משחקן אחד לוחץ על צבע מסוים, אף אחד לא מקבל תגמול.
- בסוף המשחק התגמולים שהתקבלו מחולקים בצורה שווה.

אתגר לימוד פרוטוקול תקשורת פשוט

המטרה:

ללמוד אסטרטגיה שממקסמת את הרווח.



רעיונות?

איך נגשים לפתור את הבעיה?

אפשרות 1: לחשוב על אלגוריתם ולבדוק אותו או להוכיח שהוא טוב.

הבעיה: זה קשה ולא בטוח שנצליח.

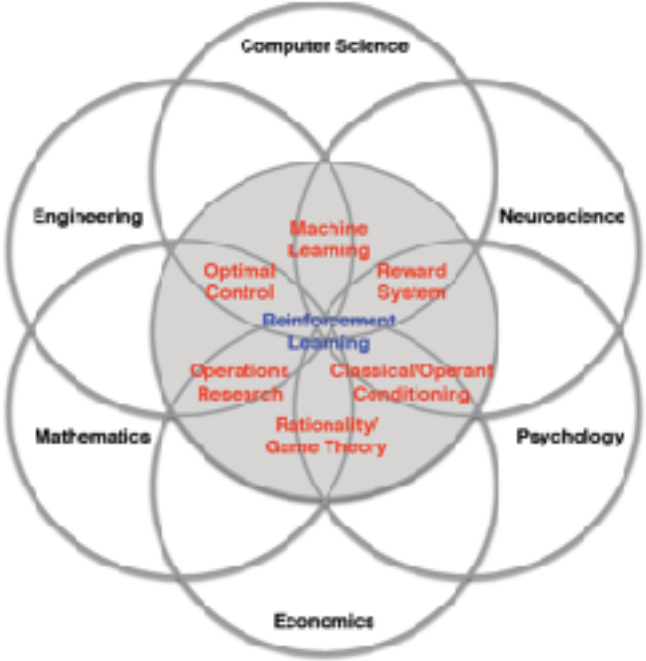
אפשרות 2: לבנות מערכת לומדת למציאת אסטרטגיות טובות

הבעיה: איך עושים את זה?

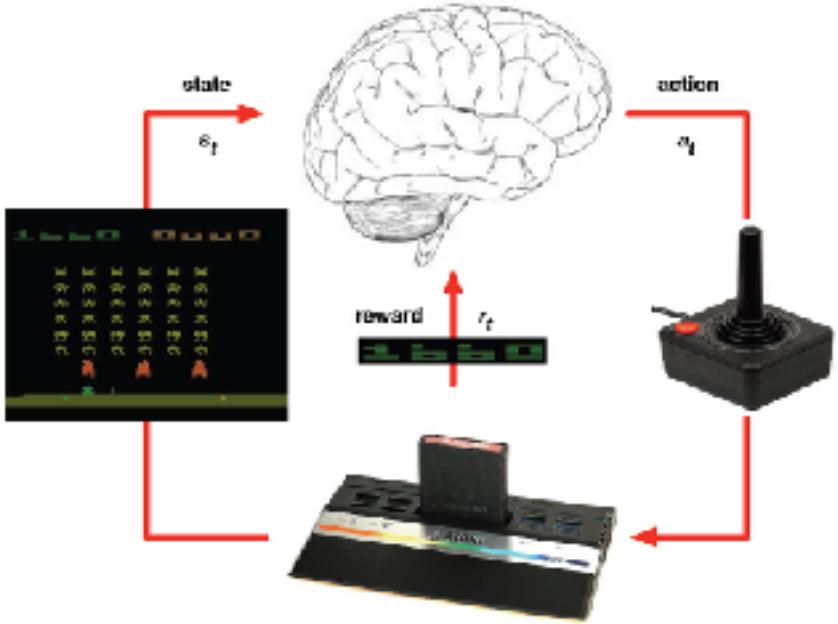
הפתרון: למידת חיזוקים עמוקה (Deep reinforcement learning)

למידת חיזוקים

למידת חיזוקים



למידת חיזוקים (David Silver)



- At each step t the agent:
  - Executes action a_t
  - Receives observation o_t
  - Receives scalar reward r_t
- The environment:
  - Receives action a_t
  - Emits observation o_{t+1}
  - Emits scalar reward r_{t+1}

למידת חיזוקים (David Silver)

- Experience is a sequence of observations, actions, rewards

$$o_1, r_1, a_1, \ldots, a_{t-1}, o_t, r_t$$

- The state is a summary of experience

$$s_t = f(o_1, r_1, a_1, \ldots, a_{t-1}, o_t, r_t)$$

- In a fully observed environment

$$s_t = f(o_t)$$

Major Components of an RL Agent

- An RL agent may include one or more of these components:
  - Policy: agent's behaviour function
  - Value function: how good is each state and/or action
  - Model: agent's representation of the environment

Policy

- A policy is the agent's behaviour
- It is a map from state to action:
  - Deterministic policy: $a = \pi(s)$
  - Stochastic policy: $\pi(a|s) = \mathbb{P}[a|s]$

Discounted Future Reward

Total Reward:

$$R = r_1 + r_2 + r_3 + \ldots\ldots + r_N$$

Total Future Reward from point t:

$$R_t = r_t + r_{t+1} + r_{t+2} + \ldots\ldots + r_N$$

ע"מ לתת עדיפות לטווח הקרוב והוודאי יותר, נבטא את R_t בצורה הבאה:

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \ldots\ldots + \gamma^N r_N = r_t + \gamma R_{t+1}$$

$Q_{\pi}(S_t,A_t) = E(R_{t+1})$ (or: How good is action A in state S?)

$\pi(S,A) = \mathrm{argmax}_A Q(S,A)$

https://www.intelnervana.com/demystifying-deep-reinforcement-learning/

Discounted Future Reward

Bellman's equation: $Q(S,A) = r + \gamma Q(S',A')$

Implementation: $Q(S,A) = Q(S,A) + \alpha \{r + \gamma \max_{A'} Q(S',A') - Q(S,A)\}$

קצב למידה

Target

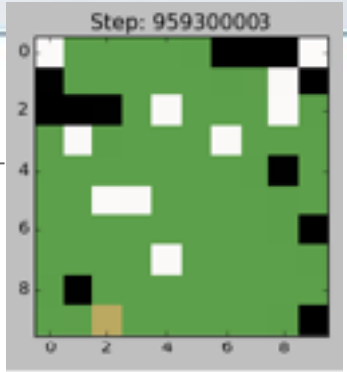
Prediction

CONVERGENCE OF Q-LEARNING: A SIMPLE PROOF

Francisco S. Melo
Institute for Systems and Robotics
Instituto Superior Técnico,
Lisboa, PORTUGAL
fmelotic.ist.utl.pt

Q(S,A)

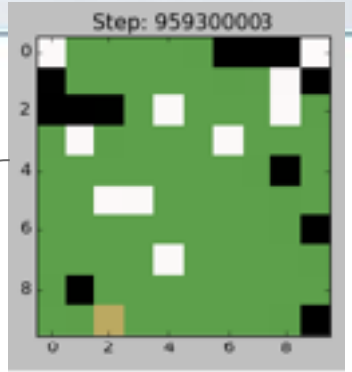
	⇐	0	⇒
State #1	0	0	0
→ State #2	0	0	0
→ State #3	0	0	0
→ State #4	0	0	0
→ ⋮	0	0	0
	0	0	0
State #N		0	0
→			

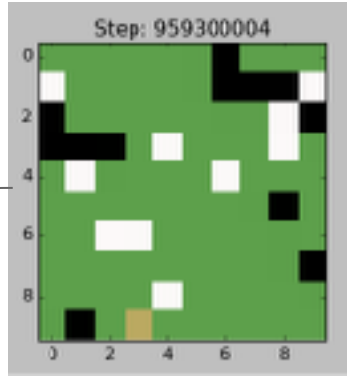


$Q(S,A) = Q(S,A) + \alpha \{r + \gamma \max_{A'} Q(S',A') - Q(S,A)\}$

Q(S,A)

	⇐	0	⇒
State #1	0	0	0
State #2	0	0	0
State #3	0	0	0
State #4	0	0	0
⋮	0	0	0
⋮	0	0	0
State #N		0	0

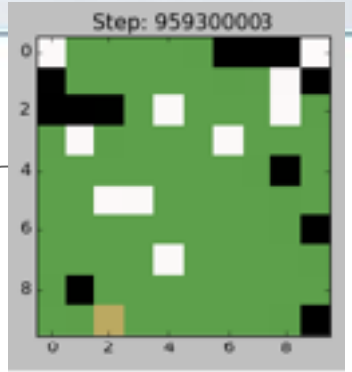




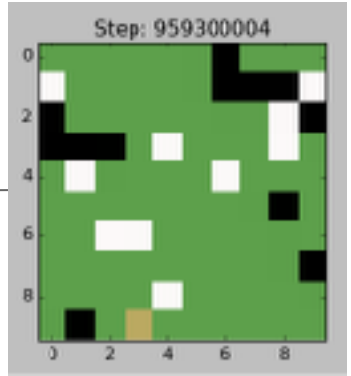
$Q(S,A) = Q(S,A) + \alpha \{r + \gamma \max_{A'} Q(S',A') - Q(S,A)\}$

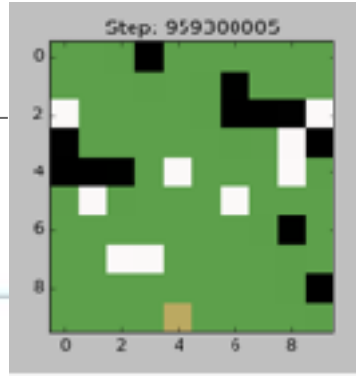
Q(S,A)

	⇐	0	⇒
State #1	0	0	0
→ State #2	0	0	0
→ State #3	0	0	0
→ State #4	0	0	0
→ ⋮	0	0	$\boxed{\alpha r}$
	0	0	0
~~State #N~~	0	0	0



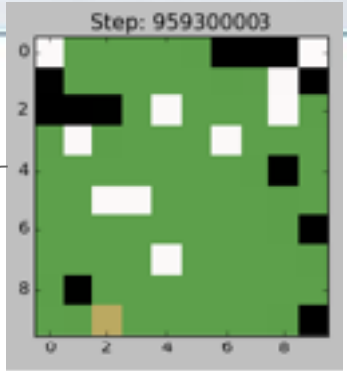
$Q(S,A) = Q(S,A) + \alpha\{r + \gamma \max_{A'} Q(S',A') - Q(S,A)\}$



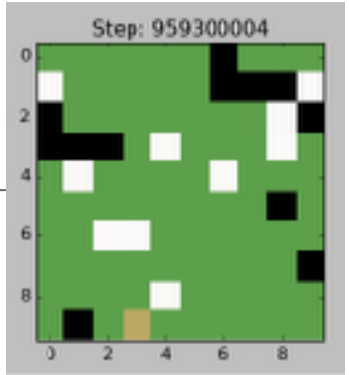


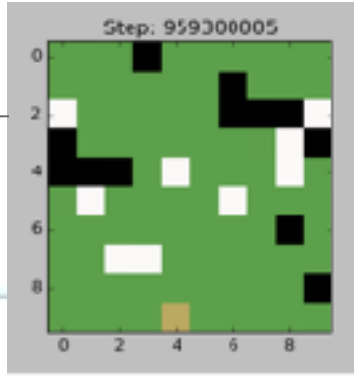
Q(S,A)

	⇐	0	⇒
State #1	0	0	0
→ State #2	0	0	0
→ State #3	0	0	0
→ State #4	0	0	0
→ ⋮	0	0	αr
⋮	0	0	0
State #N	0	0	0
→			



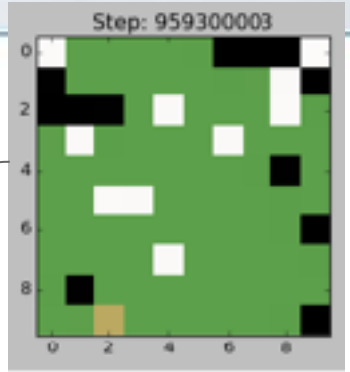
$Q(S,A) = Q(S,A) + \alpha\{r + \gamma \max_{A'} Q(S',A') - Q(S,A)\}$

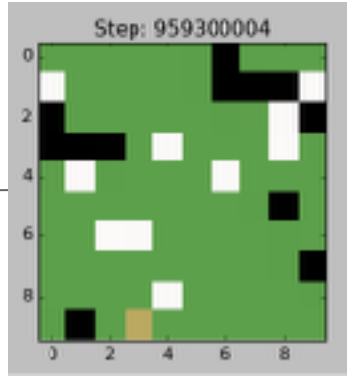


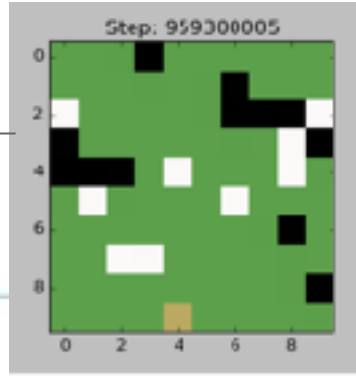


Q(S,A)

	⇐	0	⇒
State #1 →	0	0	$\boxed{\alpha^2\gamma r}$
State #2 →	0	0	0
State #3 →	0	0	0
State #4 →	0	0	0
⋮	0	0	αr
	0	0	0
State #N →	0	0	0



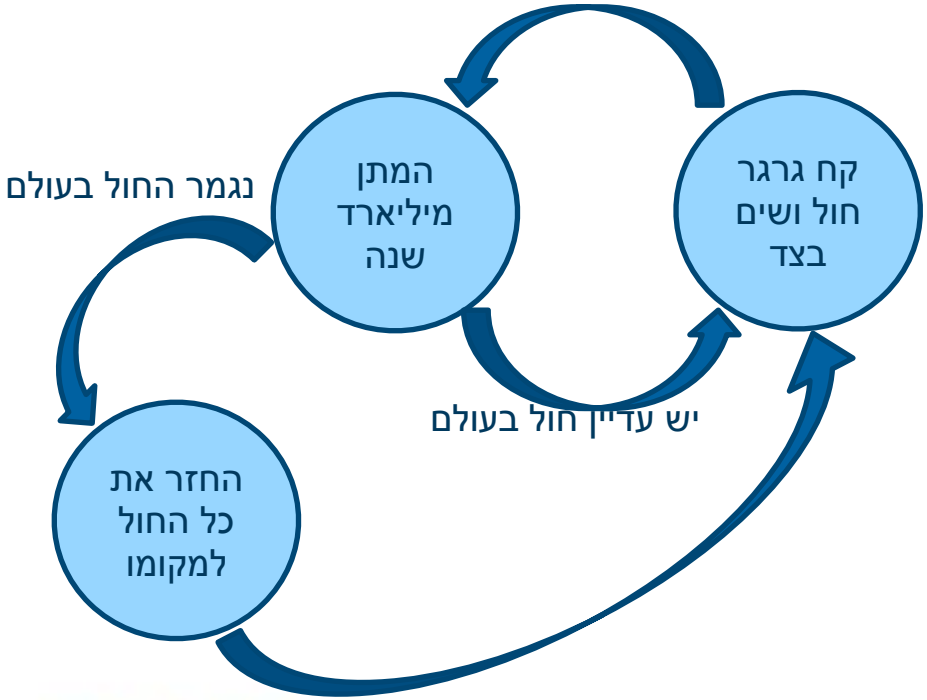




$Q(S,A) = Q(S,A) + \alpha\{r + \gamma \max_{A'} Q(S',A') - Q(S,A)\}$

בעיה גדולה

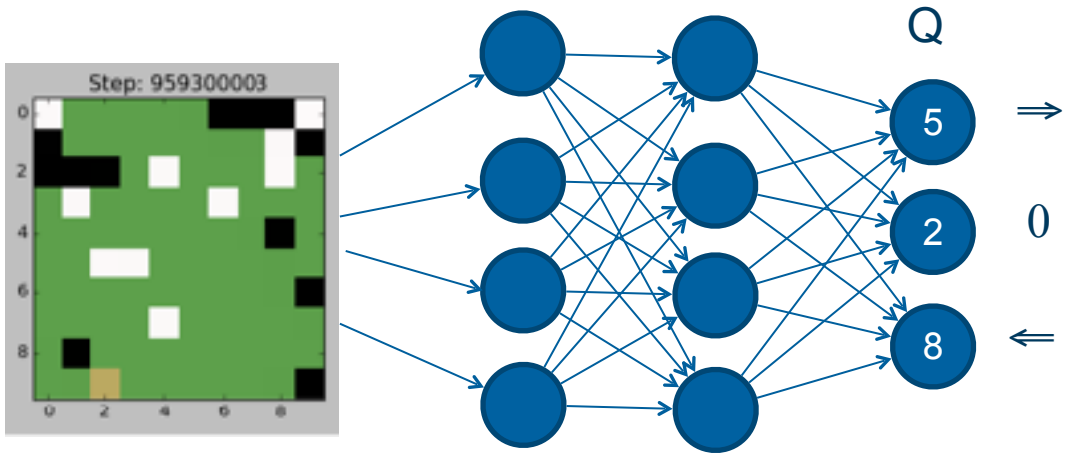
- מספר המצבים האפשריים בדוגמא הוא בערך 10^49.
- כמה גדול המספר הזה? כמה זמן זה בשניות?



X 10000

בעיה גדולה

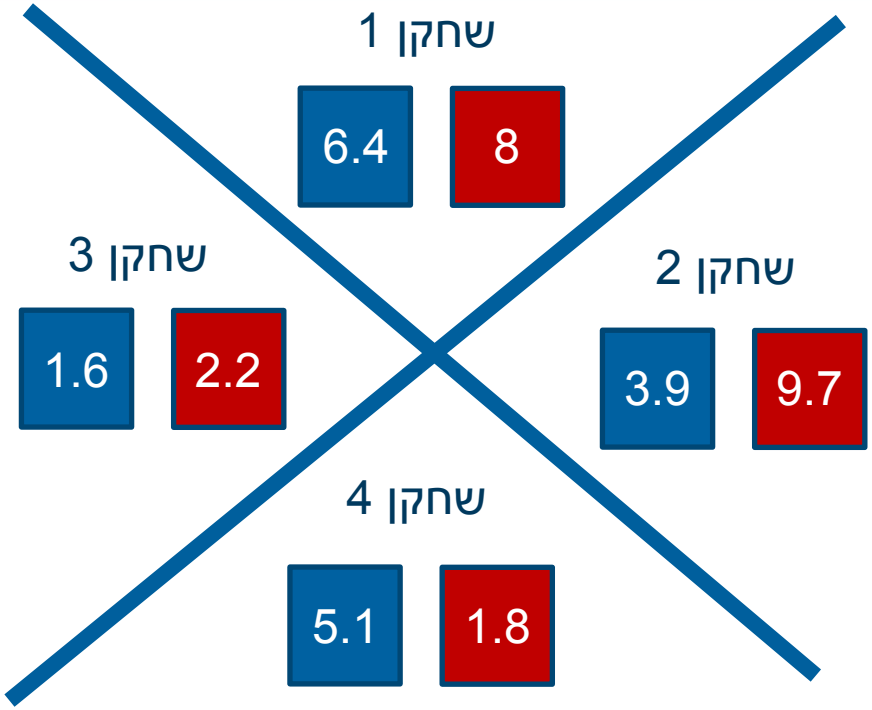
- הפתרון: נשתמש ברשת נוירונים בשביל ללמוד ייצוג של פעולות כתגובה למצבים.



חזרה לבעיה

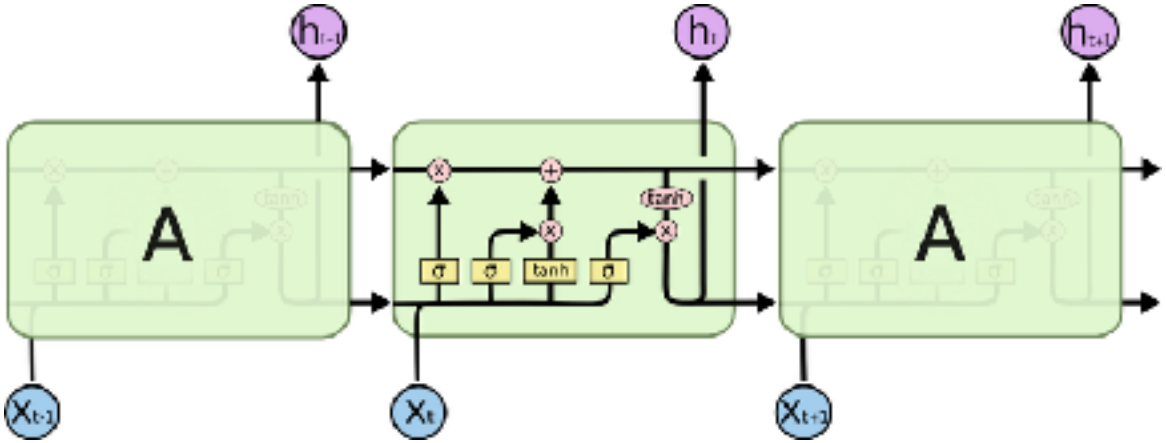
אתגר לימוד פרוטוקול תקשורת פשוט

- תזכורת



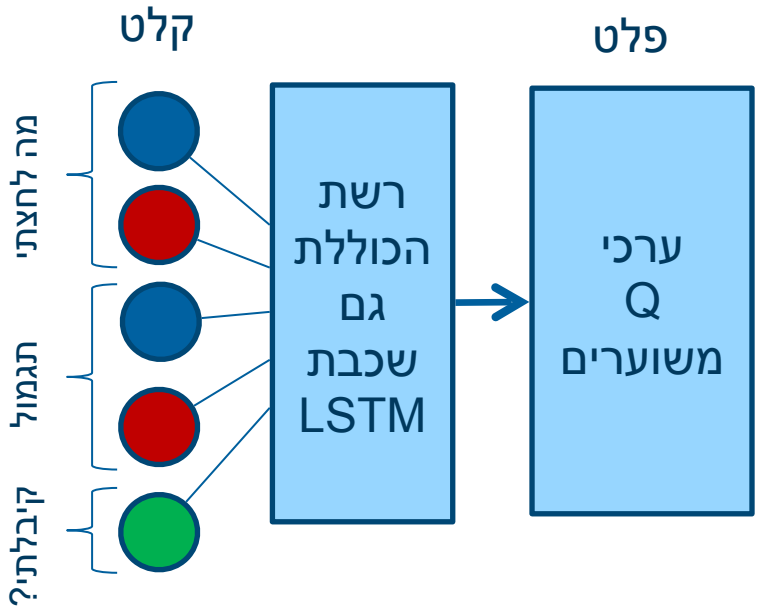
LSTM

- שכבת LSTM היא שכבה שבה כל נוירון כולל מצב פנימי שאותו הוא מעדכן על סמך הקלט.
- אנו משתמשים בה בכדי לפתור את בעיית ה partial observability וחוסר המרקוביות.



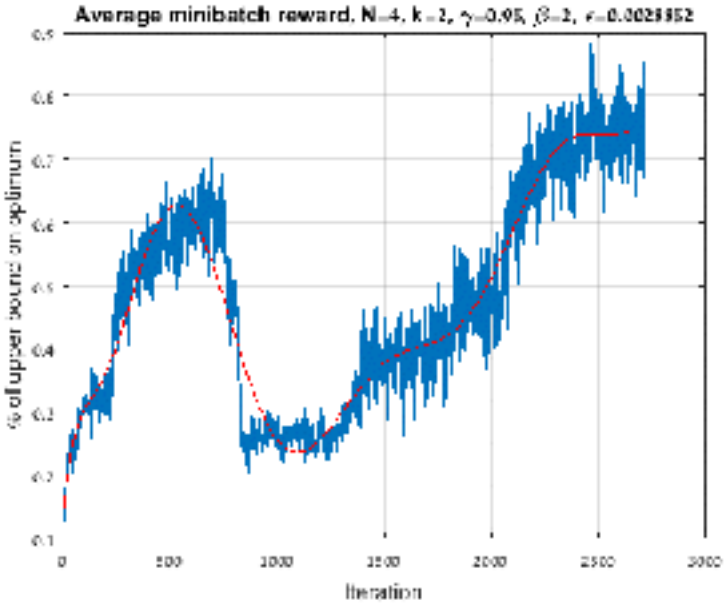
מימוש פתרון

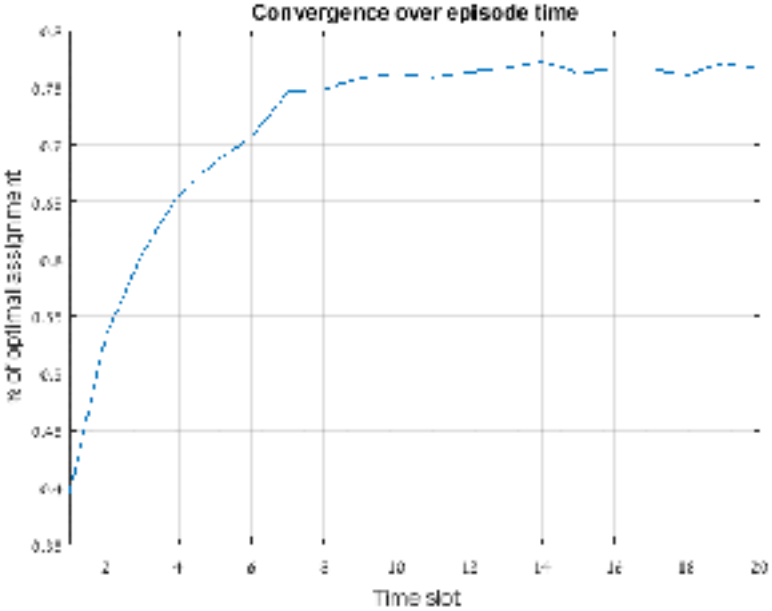
- הרשת מורכבת משכבת LSTM בכדי שתוכל ללמוד תלויות בזמן.
- לאחר ה LSTM ישנה שכבה שחוזה את ערכי ה Q – תוחלת הרווחים העתידיים.
- כל השחקנים ישתמשו תמיד באותה רשת.
- העדכון נעשה כל 30 משחקים של 20 שלבים.



תוצאות

- משיג כ 75% מהביצועים של חסם עליון על האופטימום.





בעיית תקשורת aloha

תיאור הבעיה

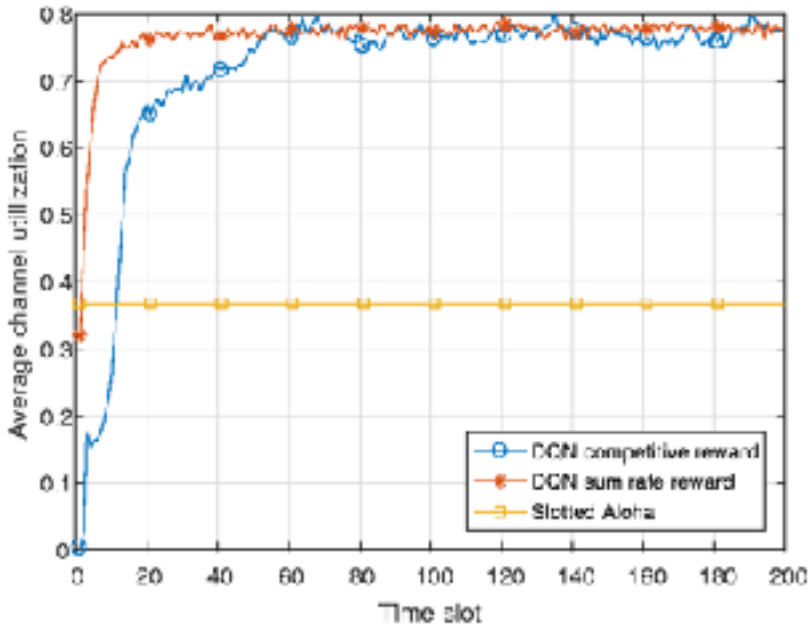
- זוגות משתמשים ללא יכולת העברת הודעות בין זוגות שונים מעוניינים להשתמש בערוץ N תקשורת ללא תיאום מראש.
- בכל תא זמן כל משתמש מחליט האם לשדר או לא ובאיזה ערוץ.
- אם משתמש שידר הודעה בהצלחה הוא מקבל ACK מבן הזוג שלו.
- אם יותר ממשתמש אחד משדר בחריץ זמן אז אנו מניחים כי כל ההודעות לא הגיעו.
- אנו מעוניינים לפתח שיטות מבוזרות למיקסום פונקציית תועלת ברשת כדוגמת utillization.

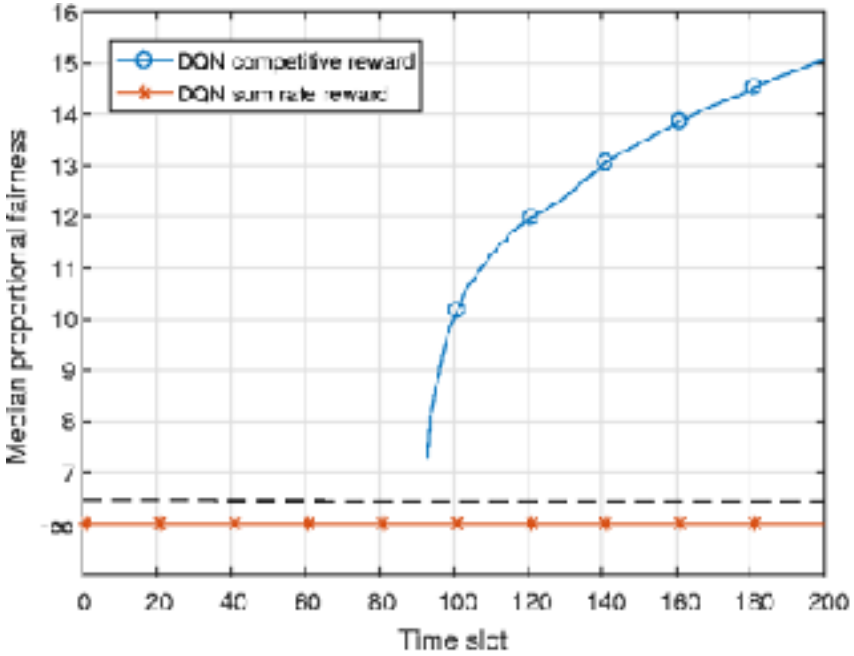
ניסויים

בדקנו את האלגוריתם עבור שיטות תגמול שונות:

- תגמול כללי - כל המשתמשים מקבלים תגמול על כל הודעה ששודרה בהצלחה.
- תגמול פרטי - רק המשתמש ששידר בהצלחה מקבל תגמול.
- מספר log כל שמשתמשים מקבלים תגמול שהוא סכום – proportional fairness ההודעות שכל משתמש הצליח לשדר.

השוואה בין תגמול פרטי לכללי

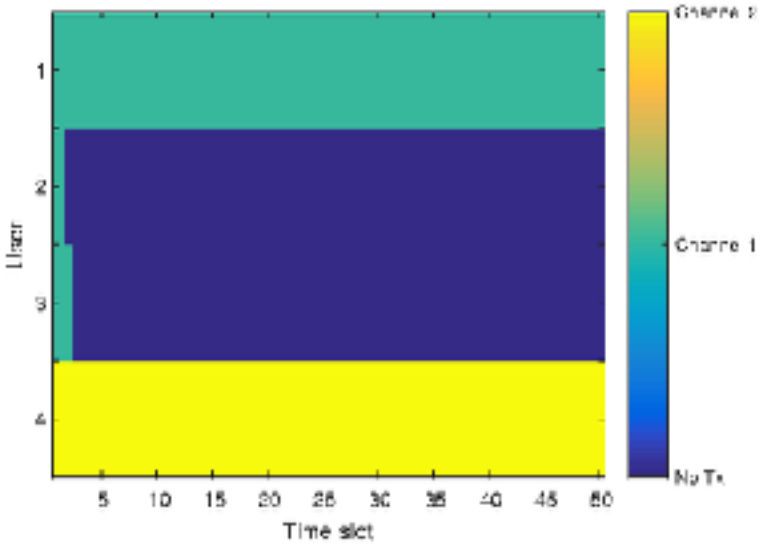




השוואה בין פונקציות התגמול השונות

- מכיוון שכל השחקנית מתוגמלים בצורה זהה עבור כל שידור מוצלח, האסטרטגיה האופטימלית היא להגיע למצב שתמיד שניים משדרים והשאר שותקים.

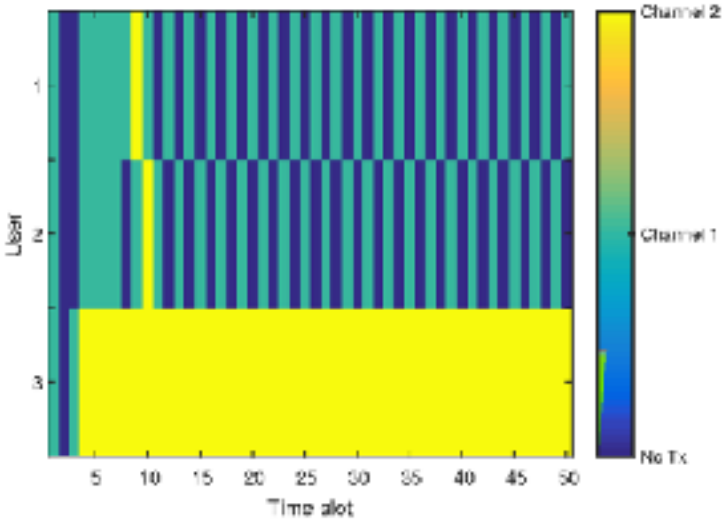
פעולות - תגמול כללי



השוואה בין פונקציות התגמול השונות

- השחקנים מתוגמלים רק על ההודעות שהם שידרו, תגמול זה מכריח אותם למצוא דרך להתחלק בזמן.

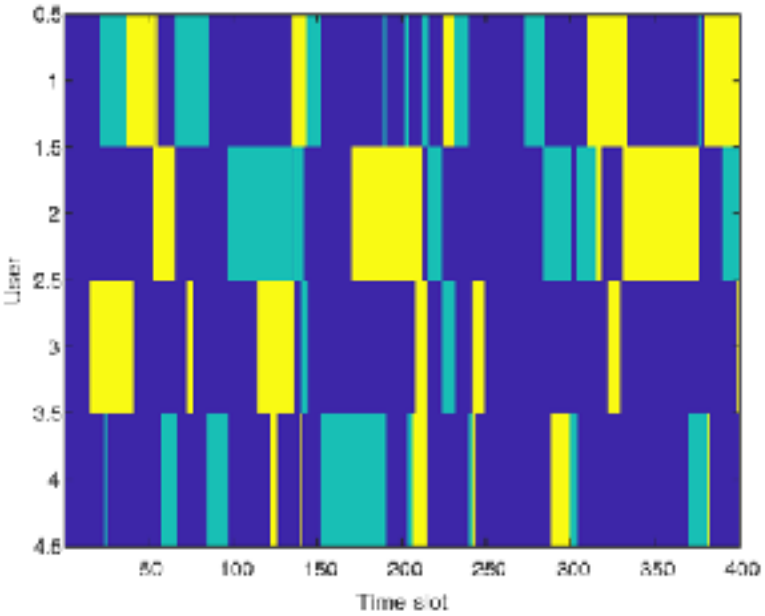
פעולות - תגמול פרטי



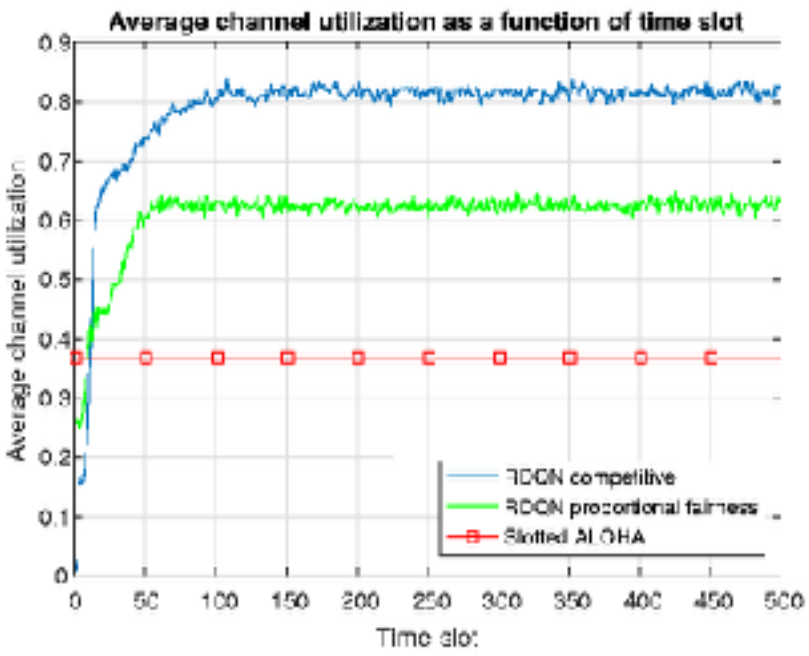
השוואה בין פונקציות התגמול השונות

- השחקנים מתוגמלים על ידי תגמול כללי הנותן יתרון לחלוקת משאבים שווה.
האסטרטגיה שנלמדה היא לשדר בערוץ עד שמישהו אחר רוצה לשדר ואז לוותר לו. כמו כן, לא נשארים זמן ארוך מידי בשידור למקרה שמישהו אחר ירצה לשדר.

פעולות - תגמול כללי עם הוגנות



השוואה בין תגמול פרטי ל fairness



- שני הסכמות התגמול התכנסו לפתרונות הוגנים.
- ניצולת הערוץ הייתה גבוהה יותר עבור התגמול הפרטי.

תודה!

לעוד פרטים ניתן לגשת למאמר ב arxiv:

Deep Multi-User Reinforcement Learning for Dynamic Spectrum Access in Multichannel Wireless Networks