סודקו בלמידת חיזוקים

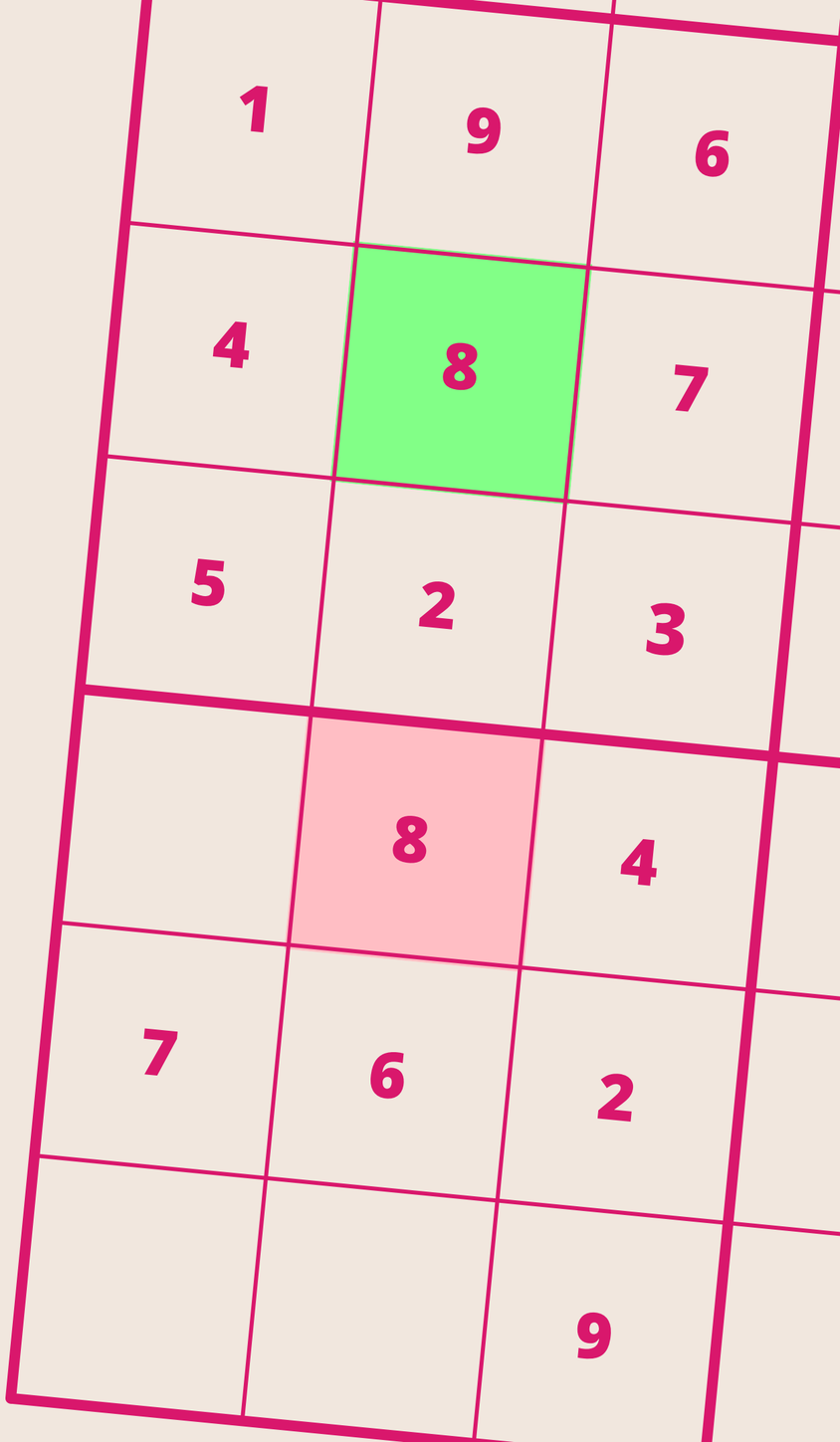
מגישים: אדר סבן, אליה זגורי ומתן וידל.

קורס: מבוא ללמידת חיזוקים.

מרצה: ד"ר טדי לזבניק.

תאריך: 08/06/25

התחל



S U D O K U

8	1		9	4	2	5	7	6
6	4	9	5		1			8
2	5				6	9	1	
7	9		1	3	4			5
		5		2			4	1
1	8	4		6		3	2	9
	2	8				1	9	7
		6	4	1	7	8	5	
						4	6	

רקע ומוטביציה לבעיה

הבעיה

סודוקו הוא פאזל 9×9 שבו יש למלא מספרים מ-1 עד 9 כך שכל שורה, עמודה ואזור 3×3 יהיו חוקיים. מרחב המצבים עצום, ורובם מובילים לטעויות. הסוכן פועל ללא מידע מלא, ומקבל חיזוק רק לפי תקינות הפתרון.

המטרה

המטרה היא לפתח סוכן חכם שלומד, באמצעות אינטראקציה עם הסביבה ותגמולים, כיצד למלא את לוח הסודוקו באופן חוקי ונכון. הסוכן פועל צעד אחר צעד ומקבל חיזוקים חיוביים או שליליים בהתאם להחלטותיו.

רלוונטיות

סודוקו מדמה בעיות אילוץ נפוצות בתחומים כמו לוגיקה, תזמון ואופטימיזציה. בלמידת חיזוקים האתגר ייחודי, כי הסוכן לא רואה את הלוח המלא בכל צעד – רק חיזוק לפי תקינות הפתרון הסופי.

היקף

הסוכן פועל על לוחות סודוקו מוכנים מראש, בוחר פעולות צעד-אחר-צעד, ומקבל תגמול לפי חוקיות המספר שהוזן. הסביבה נבנתה לבדוק כל צעד, והסוכן מתמודד עם מורכבות גבוהה בעזרת למידת חיזוקים.

ניסוח פורמלי של הבעיה

1. מרחב המצבים (States)

כל מצב מיוצג על ידי:

- לוח סודוקו מלא חלקית בגודל 9X9
- מיקום נוכחי של הסוכן (self.pos)
- בנוסף קיים BINARY_MAP המציינת אילו תאים נעולים.

2. מרחב הפעולות (Actions)

- פעולות 0–8: הכנסת ספרה 1–9 לתא הנוכחי.
- פעולת הסוכן מקודדת לפי מיקום בתא (X,Y)

3. פונקציית התגמול (Reward):

תגמולים משתנים לפי רמת הקושי.

לדוגמה ברמת "medium":

- הכנסת מספר חוקי: +8
- שגיאה בלוגיקה (כפילות בשורה/טור/ריבוע): -50
- ניסיון להכניס מספר לתא נעול: -12
- ניסיון פעולה לא חוקית: -12
- פתרון מלא ונכון של הלוח: +100
- מילוי נכון של שורה\עמודה +15.

4. דינמיקת הסביבה (Environment Dynamics):

- בכל צעד, הסוכן בוחר פעולה ומשנה את הלוח או את מיקומו.
- המצב הבא נקבע לפי הפעולה והחוקיות של המספר.
- הסביבה מסתיימת (done=True) רק כאשר הלוח נפתר בצורה תקינה או שנגמר מספר הצעדים שנקבע מראש.
- הסביבה מחזירה state, reward, ו-done.

סביבה ונתונים

נתונים

- נעשה שימוש בלוחות סודוקו מוכנים מראש בשלוש רמות קושי: קל, בינוני, קשה.
- כל לוח כולל פתרון תקני מלא – מאפשר לבדוק האם הסוכן הגיע לפתרון נכון.
- הלוחות מיוצגים כמחרוזות ומומרים למטריצות 9×9 בקוד.

הסביבה

- נבנתה סביבה מותאמת אישית בקוד (Sudoku class).
- הסוכן נע בין תאים ובוחר בין הכנסת ספרה או תזוזה.
- פונקציות פנימיות בודקות אם הפעולה חוקית ומחזירות תגמול בהתאם.
- רמות הקושי משפיעות על מערכת התגמולים (כגון ענישה על טעויות או בונוס לפתרון).

שיטת הלמידה והאלגוריתם

השתמשנו באלגוריתם Q-Learning טבלאי (Tabular), שבו נשמר ערך Q לכל פעולה בכל מצב אפשרי.
כל מצב מיוצג כטבלה חד־חד ערכית של מצבים ופעולות, והסוכן לומד לעדכן את הערכים על פי ניסיון בפועל.

טכניקות עיקריות:

- Tabular Q-Learning: שימוש במילון לאחסון ערכי Q
- Epsilon-Greedy: איזון בין חקירה לניצול עם דעיכה הדרגתית של ε
- Replay Buffer: שמירה של חוויות קודמות לצורך למידה חוזרת

תוצאות

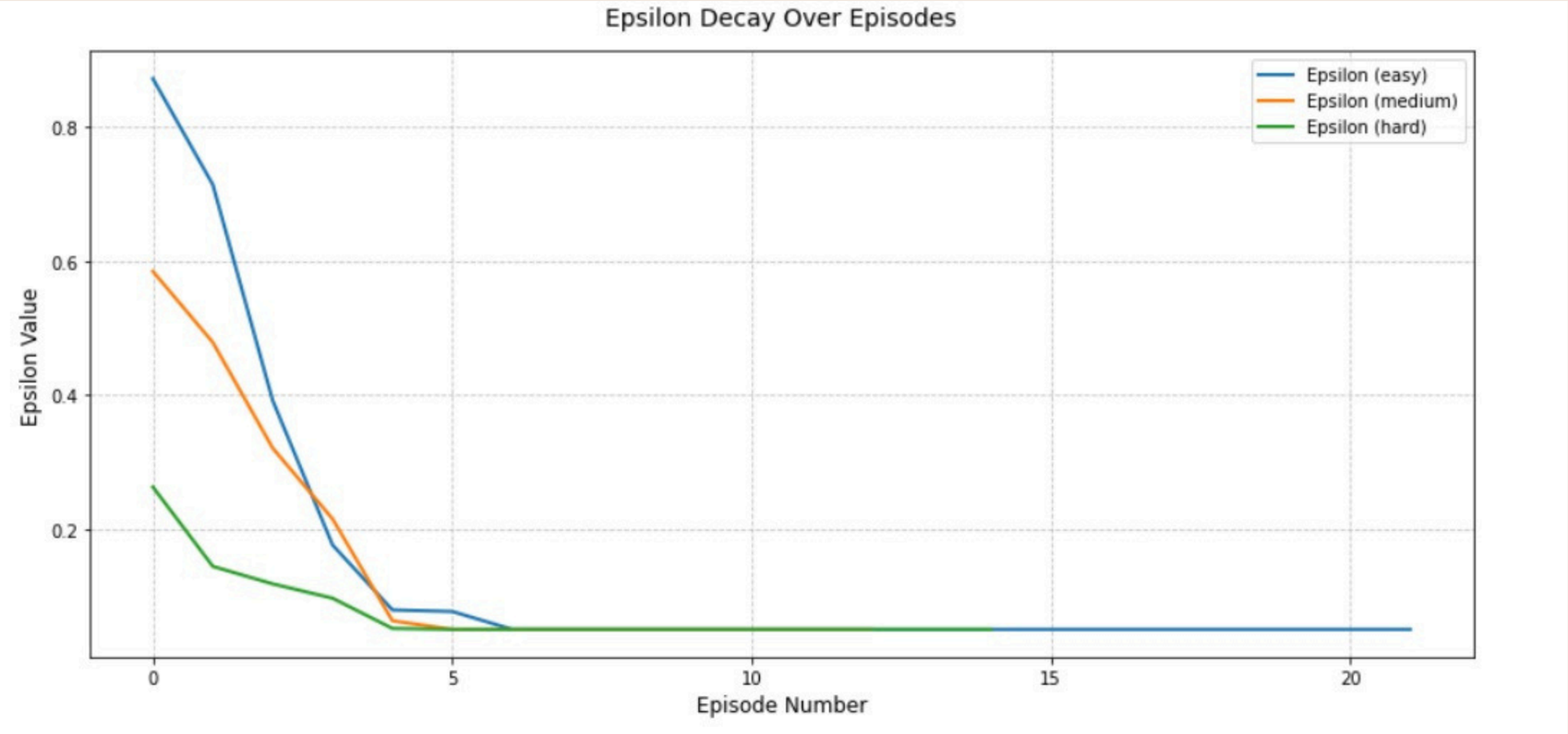
מטרה:
למדוד תוך כמה אפיזודות הסוכן מצליח לפתור לוח סודוקו שלם בכל רמת קושי.

מה רצינו לראות:
עדות לכך שהלמידה תלויה ברמת האתגר.

מה באמת ראינו:
- Easy: פתרון ראשון באפיזודה 14.
- Medium: פתרון ראשון רק באפיזודה 50.
- Hard: פתרון ראשון רק באפיזודה 50.
- כלומר הסוכן זקוק ליותר ניסיון ככל שהפאזל קשה יותר

למה זה חשוב:
מהגרף מוכיח שהסוכן לומד בקצב מותאם לרמת הקושי, ולא פותר מתוך ניחוש.
הצלחות מאוחרות ברמות קשות הן אינדיקציה ל־למידה יציבה ולא מקרית.

תוצאות



מטרה:

לבדוק האם הסוכן *עובר מחקירה (Exploration) לניצול (Exploitation)* לאורך האפיזודות, באמצעות ירידה בערך אפסילון (ε).

מה רצינו לראות:

קצב שונה בין רמות קושי – בהתאם לאתגר שבכל רמה. ירידה הדרגתית וקבועה באפסילון.

מה באמת ראינו:

- Easy: ירידה חדה ומהירה: מ־0.9 ל־0.05 תוך 5 אפיזודות
- Medium: ירידה מתונה יותר, אך כולן מתייצבות על 0.05
- Hard: ירידה מתונה יותר, אך כולן מתייצבות על 0.05

למה זה חשוב:

ירידת אפסילון תקינה מעידה על למידה הסתגלותית חכמה.
ככל שהרמה קשה יותר – כך הסוכן שומר על חקירה זהירה יותר.

תוצאות

```
--- Episode 39 (easy) ---

Initial Board:

Current Sudoku Board | Target Sudoku Board
-------------------------------------------------
. 2 . | 1 . . | . . .  | 5 2 7 | 1 3 9 | 8 6 4
6 3 . | 2 . . | 9 1 7  | 6 3 8 | 2 5 4 | 9 1 7
. . 4 | . 6 7 | 5 3 2  | 1 9 4 | 8 6 7 | 5 3 2
-------------------------------------------------
4 5 . | 9 . 3 | 6 . 1  | 4 5 2 | 9 7 3 | 6 8 1
. . . | . . 5 | 3 7 .  | 8 1 6 | 4 2 5 | 3 7 9
9 . 3 | . 1 8 | . 2 5  | 9 7 3 | 6 1 8 | 4 2 5
-------------------------------------------------
7 8 9 | 3 4 1 | . 5 6  | 7 8 9 | 3 4 1 | 2 5 6
2 4 1 | 5 8 6 | 7 . .  | 2 4 1 | 5 8 6 | 7 9 3
3 6 5 | 7 . 2 | . 4 8  | 3 6 5 | 7 9 2 | 1 4 8
-------------------------------------------------

Step 1: Action=(7, 7, 9), Reward=21.90

Current Sudoku Board | Target Sudoku Board
-------------------------------------------------
. 2 . | 1 . . | . . .  | 5 2 7 | 1 3 9 | 8 6 4
6 3 . | 2 . . | 9 1 7  | 6 3 8 | 2 5 4 | 9 1 7
. . 4 | . 6 7 | 5 3 2  | 1 9 4 | 8 6 7 | 5 3 2
-------------------------------------------------
4 5 . | 9 . 3 | 6 . 1  | 4 5 2 | 9 7 3 | 6 8 1
. . . | . . 5 | 3 7 .  | 8 1 6 | 4 2 5 | 3 7 9
9 . 3 | . 1 8 | . 2 5  | 9 7 3 | 6 1 8 | 4 2 5
-------------------------------------------------
7 8 9 | 3 4 1 | . 5 6  | 7 8 9 | 3 4 1 | 2 5 6
2 4 1 | 5 8 6 | 7 9 .  | 2 4 1 | 5 8 6 | 7 9 3
3 6 5 | 7 . 2 | . 4 8  | 3 6 5 | 7 9 2 | 1 4 8
-------------------------------------------------
```

כאן ניתן לראות פעולה מוצלחת של הסוכן באפיזודה **39.** הוא הכניס את המספר **9** לשורה **7** עמודה **7 –** פעולה שהייתה תקפה לפי חוקי הסודוקו**.** על כך הוא קיבל תגמול של **21.90.** זה מראה שהסוכן יודע לזהות תאים תקפים ולפעול בהתאם**,** וזה שלב חשוב בהתפתחות הלמידה שלו

סיכום ומסקנות

אתגרים עיקריים

תגמולים מרובים ולא סימטריים –
נדרש לאזן בין ענישה על טעויות לבין עידוד חקירה.
לדוגמה: הכנסת מספר שגוי לתא ריק צריכה להיענש פחות מאשר ניסיון לדרוס ערך קיים.

תיעדוף פעולות–
חשוב ללמד את הסוכן להעדיף טעויות "לגיטימיות" (כמו ניסיון בתא ריק) על פני טעויות חמורות (שיבוש תא קיים).

מורכבות לוגית–
בכל צעד מתבצעים כמה חישובים במקביל. תנועה, בדיקה אם התא חדש, בדיקה אם הערך חוקי – ולפעמים יש סכימה של כמה תגמולים בו-זמנית.

מסקנות עיקריות

סודוקו מציג אתגר אמיתי בלמידת חיזוקים בגלל עומס אילוצים ולוגיקה מורכבת.

תכנון מערכת תגמולים חכמה הוא קריטי – משפיע ישירות על איכות הלמידה.

גם תגמולים קטנים ולא מאוזנים עלולים להוביל את הסוכן להתנהגויות לא רצויות.