# מבוא לבינה מלאכותית - תשפ"ד - תרגיל 5

207063108 עבד נסאר 324059856 אחמד דלאשה

2024 ביולי 24

### שאלה מספר 1.1.1

: ותהליך איטרציית הערכים כולו, כולל החישובים MDPותהליך איטרציית הערכים כולו

### : MDP-ה

: מצבים

$$(S): \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9\}$$

: פעולות

$$(A): \{Left, Right, Up, Down\}$$

- . מודל המעבר (T) דטרמיניסטי, כלומר התוצאה של כל פעולה היא ודאית. יציאה מהגריד מחזירה את הסוכן לאותו מצב.
  - :(R) פונקציית התגמול •

$$r(s) = -0.05 \ for \ s \notin \{s_5, s_7, s_9\}$$

$$r\left(s_5\right) = -10$$

$$r\left(s_7\right) = 15$$

$$r\left(s_9\right) = 30$$

 $s_5, s_7, s_9$  are absorbing states.

### : אתחול

 $V_{0}\left( s
ight) =0$  ערך התחלתי לכל המצבים

### : נוסחת איטרציית הערכים

$$V_{k+1}\left(s\right) = Max_{a}\left(\sum_{s^{'}} T\left(s^{'} \mid s, a\right) \left[R\left(s, a, s^{'}\right) + \gamma V_{k}\left(s^{'}\right)\right]\right)$$

### תהליך האיטרציה:

נבצע איטרציית ערכים למשך 7 איטרציות, ונעדכן את הערך של כל מצב בכל שלב.

- איטרציה 1: נחשב את הערכים עבור כל מצב בהתבסס על משוואת בלמן.
- . איטרציות 2 עד 7: נמשיך לעדכן את הערכים בהתבסס על תוצאות מהאיטרציה הקודמת.

להלן התוצאות של תהליך איטרציית הערכים למשך 7 איטרציות:

	Iteration	s1	s2	s3	s4	s5	s6	s7	s8	s9
o	1.0	-0.05	-0.05	-0.05	-0.05	-10.0	-0.05	15.0	-0.05	30.0
1	2.0	-0.0995	-0.0995	-0.0995	14.8	-10.0	29.65	15.0	29.65	30.0
2	3.0	14.602	-0.1485	29.3035	14.8	-10.0	29.65	15.0	29.65	30.0
3	4.0	14.602	28.9605	29.3035	14.8	-10.0	29.65	15.0	29.65	30.0
4	5.0	28.6209	28.9605	29.3035	14.8	-10.0	29.65	15.0	29.65	30.0
5	6.0	28.6209	28.9605	29.3035	28.2847	-10.0	29.65	15.0	29.65	30.0
6	7.0	28.6209	28.9605	29.3035	28.2847	-10.0	29.65	15.0	29.65	30.0

### :1 איור

ערכים אלו מייצגים את התגמולים המצופים המוזלים לכל מצב לאחר 7 איטרציות.

## : 1.1.2 שאלה מספר

וישוב הפעולות האופטימליות והצגת הטבלה:

נחשב את הפעולות האופטימליות עבור כל מצב בכל איטרציה ונציג את התוצאות בטבלה. הטבלה תכלול 7 שורות (אחת לכל איטרציה) ו־10 עמודות (מספר האיטרציה והפעולה האופטימלית לכל מצב).

	Iteration	s1	s2	s3	s4	s5	s6	s7	s8	s9
0	1	Left	Left	Left	Left	Left	Left	Left	Left	Left
1	2	Left	Left	Left	Up	Left	Up	Left	Right	Left
2	3	Up	Left	Up	Up	Left	Up	Left	Right	Left
3	4	Up	Right	Up	Up	Left	Up	Left	Right	Left
4	5	Right	Right	Up	Up	Left	Up	Left	Right	Left
5	6	Right	Right	Up	Down	Left	Up	Left	Right	Left
6	7	Right	Right	Up	Down	Left	Up	Left	Right	Left

 $The\ optimal\ policy\ grid\ correct\ :$ 

<b>←</b>	→	←	
<b>1</b>	<b>←</b>	1	
→	1	<b>↑</b>	

:3 איור

## : 1.2 שאלה מספר

בשלב הזה, חישבנו את ערכי המדינות עבור MDP סטוכסטי בו כל פעולה מצליחה עם הסתברות P=0.9 ובכשלון מעבירה את הסוכן לתא סמוך עם הסתברות אחידה.

## : MDP-רכיבי

: מצבים

$$(S): \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9\}$$

: פעולות

$$(A): \{Left, Right, Up, Down\}$$

:(T) מודל המעבר •

- עבור הפעולה המוצלחת עבור  $P\left(s^{'}\mid s,a\right)=0.9$  –
- עבור כל תא סמוך הוא מספר הוא עבור כל תא סמוך עבור כל  $P\left(s^{'}\mid s,a\right)=\frac{0.1}{3}$ 
  - :(R) פונקציית התגמול •

$$r(s) = -0.05 \text{ for } s \notin \{s_5, s_7, s_9\}$$
  
$$r(s_5) = -10$$

$$r\left(s_7\right) = 15$$

$$r\left(s_9\right) = 30$$

 $s_5, s_7, s_9$  are absorbing states.

## : להלן התוצאה

	s1	s2	s3	s4	s5	s6	s7	s8	s9
0	24.669	25.793	27.716	21.018	-10.0	28.19	15.0	27.761	30.0

:4 איור

השוואה בין הערכים של המדיניות האופטימלית בין המקרה הדטרמיניסטי והסטוכסטי

	State	Deterministic	Stochastic
0	s1	28.6209	24.011
1	s2	28.9605	24.61
2	s3	29.3035	26.904
3	s4	28.2847	21.55
4	s5	-10.0	-10.0
5	s6	29.65	27.52
6	s7	15.0	15.0
7	s8	29.65	26.93
8	s9	30.0	30.0

#### ניתוח:

במצב הסטוכסטי, הערכים לרוב נמוכים יותר מאשר במצב הדטרמיניסטי. ההסבר לכך הוא שבמצב הסטוכסטי יש סיכוי לכישלון של פעולה, מה שגורם להורדת הערכים של המדינות מכיוון שהסוכן יכול לעבור למצב פחות רצוי.

#### סיכום:

ההבדל בין הערכים הדטרמיניסטיים לסטוכסטיים מדגיש את החשיבות של מודל המעבר ואי הוודאות בניהול מדיניות אופטימלית בסביבות מורכבות. ההשוואה ממחישה כיצד אי ודאות משפיעה על הערכים הצפויים של המדינות ועל קבלת ההחלטות בסביבות שונות.

### שאלה מספר 1.3

Γ	s1	s2	s3	s4	s5	s6	s7	s8	s9
0	14.39	24.08	26.84	13.53	-10.0	27.51	15.0	26.92	30.0

#### :6 איור

### ניתוח השוואתי

ניתן לראות כי הערכים של המדיניות הנתונה נמוכים יותר בערכים מסוימים לעומת המדיניות האופטימלית:

- 24.011 ממוך מ־  $s_1:14.39$  •
- 24.61 ממוך מ־  $s_2:24.08$  •
- 21.55 ממוך מ־  $s_4:13.53$  •

המדיניות הנתונה מובילה לערכים נמוכים יותר מאשר המדיניות האופטימלית בחלק מהמצבים, דבר שמצביע על כך שהמדיניות הנתונה פחות אופטימלית בסביבה הסטוכסטית.

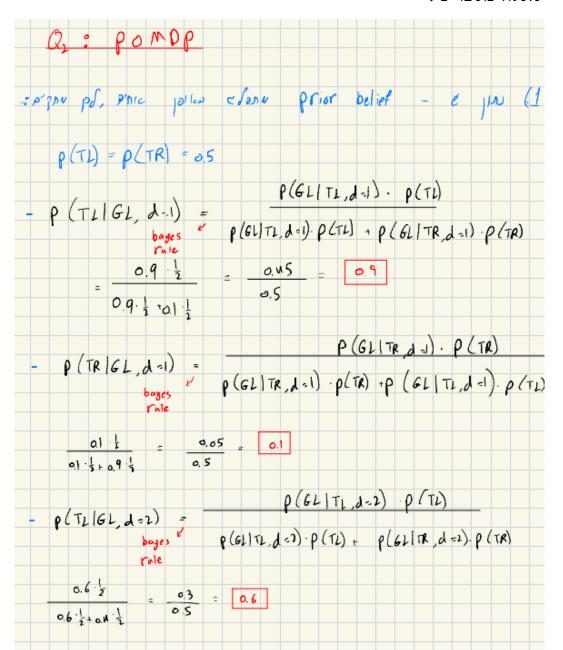
### מסקנות:

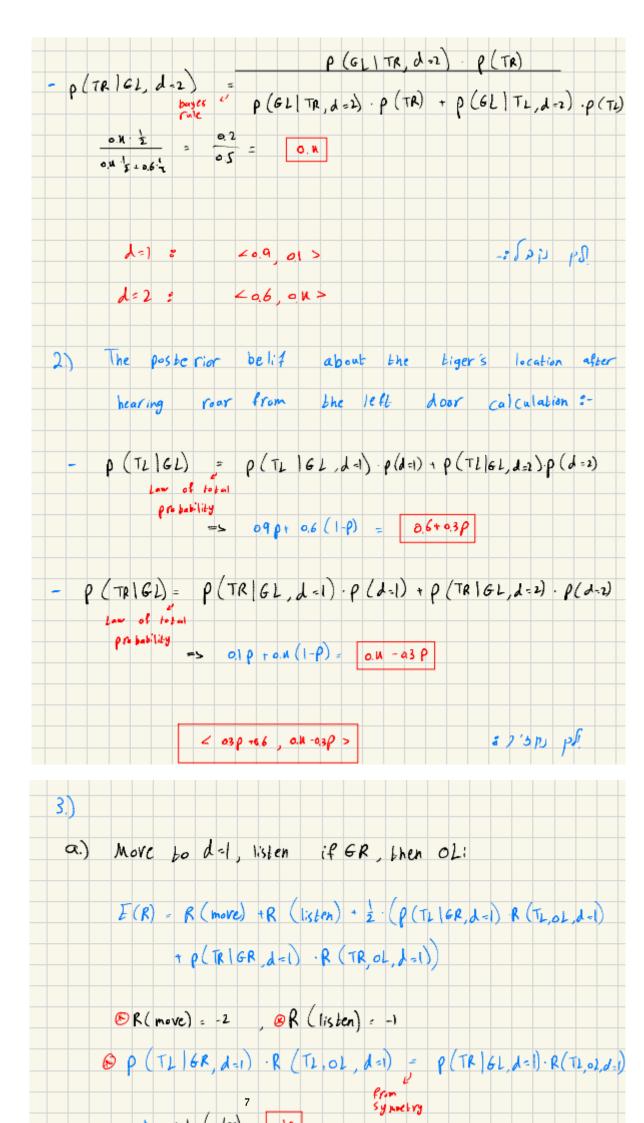
- השפעת המדיניות: המדיניות הנתונה משפיעה באופן ישיר על הערכים הצפויים של המצבים. מדיניות פחות אופטימלית תוביל לערכים נמוכים יותר.
- סביבה סטוכסטית: בסביבה סטוכסטית, קיימת אי־וודאות בתוצאות של פעולות. המדיניות האופטימלית מצליחה להתמודד טוב יותר עם אי־הוודאות הזו ומשיגה ערכים גבוהים יותר.
- החשיבות של אופטימיזציה: על מנת להשיג ערכים מקסימליים בסביבה סטוכסטית, חשוב לבצע אופטימיזציה למדיניות תוך התחשבות באי־ודאות.

### סיכום:

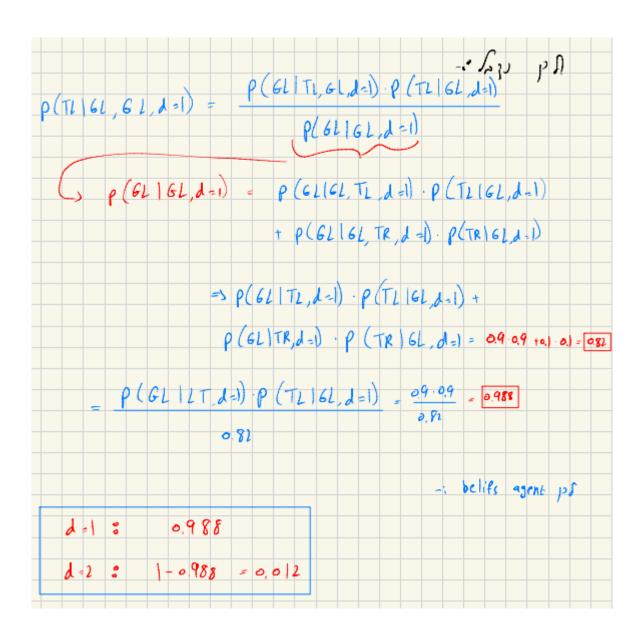
המדיניות הנתונה הובילה לערכים נמוכים יותר בהשוואה למדיניות האופטימלית בסביבה סטוכסטית. זה מדגיש את החשיבות של אופטימיזציה של מדיניות בסביבות עם אי־וודאות על מנת להשיג את הערכים הטובים ביותר עבור המצבים.

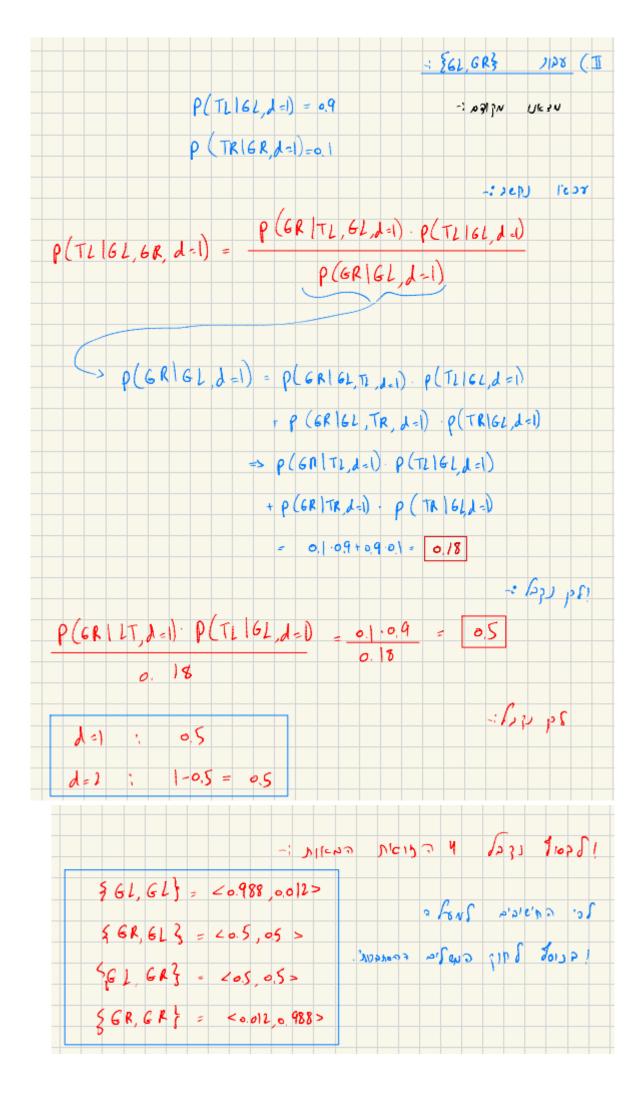
## : 2 שאלה מספר





6.)	listen if	GL, then	listen . F	1sc	if GR	then o	Li
	E(R)	= R (liston)	1. (p(1	TL (61,	d=1) + p(	TR   6L, d-2	) · R ( liston))
		1 / (-1-1		(	1 1	0/	,d=1) · R (TR,oL,d=1))
		1 2 ( (PC121	6K,4:2) - K	(11,01	,d=21+	PCIKIEK	,d=1) · K (  K,oL,d=2) )
	- R (listen)	) = -1					
	+ 0(TL	GL d=2) . R (1	sten) = 0.6.	(-1) -	-0.6		
		1					
	- P(T) (	6L,d=2) · R(1;	sten) o.	u · ( -1)	= -a,h		
	- P (T1)	GR, d=2) - R (TL,	oLd=2) -	P (TR	(CL, d=1).	R (TLpl, d	=2) = a.u.(50) = -20
							R, d=1)= 06.10 -6
		,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	4) CHOOLO P		,0	7	
						. 6	צכש ופכות ונק
						7.0	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
	E(R) < 0	i,5 · (-1) + as	·(-14) =	-8	5		
	e ah	er valians	2 16	٦.	n1'20'	a/c 4	115e' (4
	. 00%	CI VationS	- " );	, O ,	)///		150 6 (4
5.7	1 / 1 2	\$ GL, 6 R	١ 5	CR	C1 2	EGR, C	<b>(2)</b>
7 6	1,61}	502,61	5 2	٠,	023	100,0	
					_ 51	61,615	2124 (I
				,			Ι.) γ (.[
		PCT	L 16L, d=	1) =0	.9 _	6 0311N	BICSN
		0(1	RIGRA	-1) =	ا ا		
		0	1				

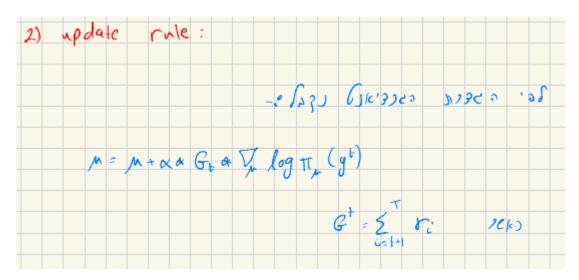


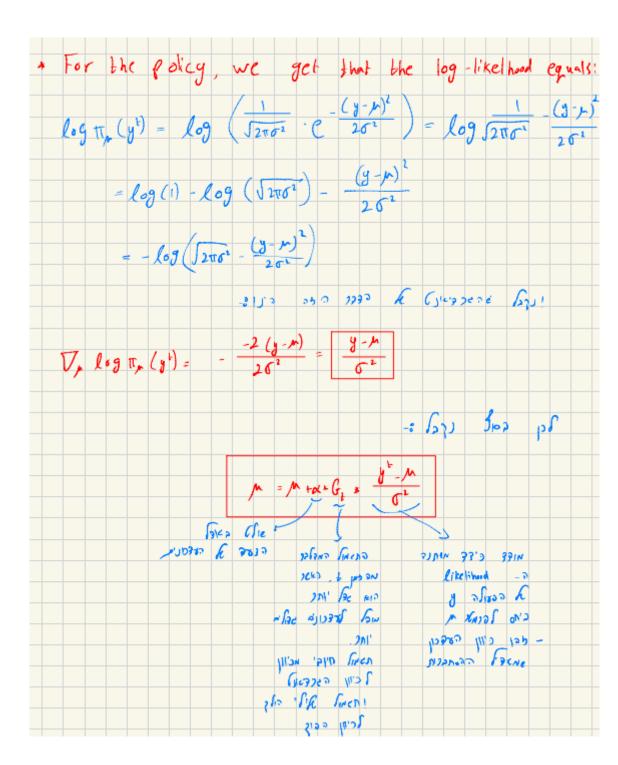


## : 3.1 שאלה מספר

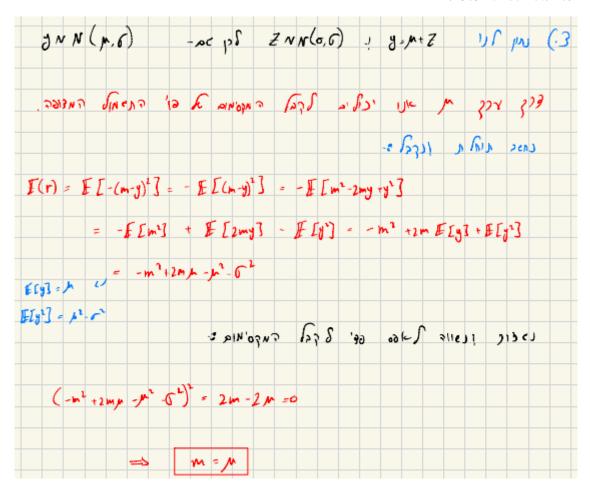
Question 32			
1) we know	bhat policy	determines how	the agent behaves
			tion therenot will sick
y to N (M, 09)	-e (23) !	vic ZNN(o,	אכיון נ
Jt (y lp		Sn ( y=y) =	1 - (y-n) 2710° C 20°

## : 3.2 שאלה מספר





### שאלה מספר 3.3:



## שאלה מספר 3.4:

כדי למצוא את שיעור השינוי הממוצע של עלינו לחשב את עלינו השינוי הממוצע של שלב העדכון:

$$\mu = \mu + \alpha \cdot \left(-\left(m - y_t\right)^2\right) \cdot \frac{y_t - \mu}{\sigma^2}$$

פתרון שלב אחר שלב:

הגדרת שלב העדכון:

$$\Delta \mu = \alpha \cdot \left( -\left( m - y_t \right)^2 \right) \cdot \frac{y_t - \mu}{\sigma^2}$$

לקיחת התוחלת:

$$\mathbb{E}\left[\Delta\mu\right] = \mathbb{E}\left[\alpha \cdot \left(-\left(m - y_t\right)^2\right) \cdot \frac{y_t - \mu}{\sigma^2}\right]$$

 $z\sim N\left(0,\sigma^{2}
ight)$  כך שי $y_{t}=\mu+z$  החלפת

$$\mathbb{E}\left[\alpha \cdot \left(-\left(m - (\mu + z)\right)^{2}\right) \cdot \frac{(\mu + z) - \mu}{\sigma^{2}}\right]$$
$$= \alpha \cdot \mathbb{E}\left[-\left(m - \mu - z\right)^{2} \cdot \frac{z}{\sigma^{2}}\right]$$

פיתוח הריבוע:

$$= \alpha \cdot \mathbb{E}\left[-\frac{(m-\mu-z)^2 \cdot z}{\sigma^2}\right]$$

$$= \alpha \cdot \mathbb{E}\left[-\frac{\left(\left(m-\mu\right)^{2}-2\left(m-\mu\right)+z+z^{2}\right) \cdot z}{\sigma^{2}}\right]$$

הפרדת התוחלת:

$$= \alpha \cdot \left[ -\frac{\left(m - \mu\right)^{2} \cdot \mathbb{E}\left[z\right]}{\sigma^{2}} + \frac{2\left(m - \mu\right) \cdot \mathbb{E}\left[z^{2}\right]}{\sigma^{2}} + \frac{\cdot \mathbb{E}\left[z^{3}\right]}{\sigma^{2}} \right]$$

#### הערכת התוחלות:

 $\mathbb{E}\left[z\right] = 0 \left(mean \ of \ azero - mean \ Gaussian\right)$ 

 $\mathbb{E}\left[z^{2}\right]\left(variance\ of\ Gaussian\right)$ 

 $\mathbb{E}\left[z^{3}\right]$  (skewness of Gaussian is zero)

: לכן

$$\mathbb{E}\left[\Delta\mu\right] = \alpha \cdot \left[0 + \frac{2(m-\mu)\sigma^2}{\sigma^2} - 0\right]$$
$$= 2\alpha \cdot (m-\mu)$$

### : הסבר

כלל העדכון מראה ששיעור השינוי הממוצע של  $\mu$  הוא יחסי ישיר להבדל בין היעד m והערכת הנוכחית  $\mu$ . קצב הלמידה  $\alpha$  מגדיר את מהירות השינוי.  $\alpha$  זה אומר שככל שהסוכן מעדכן שוב ושוב את  $\mu$ , הוא יכוון בהדרגה את  $\mu$  לכיוון היעד m, והמהירות של התאמה זו מושפעת מה־  $\mu$ , הוא יכוון בהדרגה את פרמטר המדיניות  $\mu$  עם הזמן, ומבטיח שהרובוט משתפר בקליעת המטרה על ידי  $\mu$  בפחתת ההבדל בין  $\mu$  ל־  $\mu$ .

## שאלה מספר 3.5:

