מבוא לבינה מלאכותית - תשפ"ד - תרגיל 5

אחמד דלאשה 324059856 | עבד נסאר 207063108

24 ביולי 2024

שאלה מספר 1.1.1 :

להלן התיאור של רכיבי ה־MDP ותהליך איטרציית הערכים כולו, כולל החישובים :

רכיבי ה־MDP :

- 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.
- פונקציית התגמול (R) :

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

$$r(s_5) = -10$$

$$r(s_7) = 15$$

$$r(s_9) = 30$$

$$s_5, s_7, s_9 \;\; are\ absorbing\ states.$$

אתחול :

ערך התחלתי לכל המצבים $V_0(s) = 0$

נוסחת איטרציית הערכים :

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

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
0	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

איור 1:

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

שאלה מספר 1.1.2 :

חישוב 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

איור 2:

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

←	→	←
↓	←	↑
→	→	↑

איור 3:

שאלה מספר 1.2 :

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.

רכיבי ה־MDP :

- 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) :

- $P\left(s' \mid s,a\right) = 0.9$ עבור הפעולה המוצלחת
- $P\left(s' \mid s,a\right) = \frac{0.1}{3}$ עבור כל תא סמוך כאשר N הוא מספר התאים הסמוכים

- פונקציית התגמול (R) :

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

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

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

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

$$s_5, s_7, s_9 \;\; are \; absorbing \; states.$$

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

איור 4:

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

איור 5:

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 :

	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

איור 6:

ניתוח השוואתי

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

- s_1 : 14.39 נמוך מ־ 24.011
- s_2 : 24.08 נמוך מ־ 24.61
- s_4 : 13.53 נמוך מ־ 21.55

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.

שאלה מספר 2 :

Q_2 : POMDP

1) נתון e - prior belief שהלקח מהאוכלוסיה, לכן מתקיים:

$$p(TL) = p(TR) = 0.5$$

- $$p(TL \mid GL, d=1) \underset{\text{bayes rule}}{=} \frac{p(GL \mid TL, d=1) \cdot p(TL)}{p(GL \mid TL, d=1) \cdot p(TL) + p(GL \mid TR, d=1) \cdot p(TR)}$$

$$= \frac{0.9 \cdot \frac{1}{2}}{0.9 \cdot \frac{1}{2} + 0.1 \cdot \frac{1}{2}} = \frac{0.45}{0.5} = \boxed{0.9}$$

- $$p(TR \mid GL, d=1) \underset{\text{bayes rule}}{=} \frac{p(GL \mid TR, d=1) \cdot p(TR)}{p(GL \mid TR, d=1) \cdot p(TR) + p(GL \mid TL, d=1) \cdot p(TL)}$$

$$\frac{0.1 \cdot \frac{1}{2}}{0.1 \cdot \frac{1}{2} + 0.9 \cdot \frac{1}{2}} = \frac{0.05}{0.5} = \boxed{0.1}$$

- $$p(TL \mid GL, d=2) \underset{\text{bayes rule}}{=} \frac{p(GL \mid TL, d=2) \cdot p(TL)}{p(GL \mid TL, d=2) \cdot p(TL) + p(GL \mid TR, d=2) \cdot p(TR)}$$

$$\frac{0.6 \cdot \frac{1}{2}}{0.6 \cdot \frac{1}{2} + 0.4 \cdot \frac{1}{2}} = \frac{0.3}{0.5} = \boxed{0.6}$$

$$- \; p(TR \mid GL, d=2) \underset{\text{bayes rule}}{=} \frac{p(GL \mid TR, d=2) \cdot p(TR)}{p(GL \mid TR, d=2) \cdot p(TR) + p(GL \mid TL, d=2) \cdot p(TL)}$$

$$\frac{0.4 \cdot \frac{1}{2}}{0.4 \cdot \frac{1}{2} + 0.6 \cdot \frac{1}{2}} = \frac{0.2}{0.5} = \boxed{0.4}$$

לכן נקבל:

$d=1$: $\langle 0.9, 0.1 \rangle$

$d=2$: $\langle 0.6, 0.4 \rangle$

2) The posterior belief about the tiger's location after hearing roar from the left door calculation :-

$$- \; p(TL \mid GL) \underset{\text{Law of total probability}}{=} p(TL \mid GL, d=1) \cdot p(d=1) + p(TL \mid GL, d=2) \cdot p(d=2)$$

$$\Rightarrow 0.9p + 0.6(1-p) = \boxed{0.6 + 0.3p}$$

$$- \; p(TR \mid GL) \underset{\text{Law of total probability}}{=} p(TR \mid GL, d=1) \cdot p(d=1) + p(TR \mid GL, d=2) \cdot p(d=2)$$

$$\Rightarrow 0.1p + 0.4(1-p) = \boxed{0.4 - 0.3p}$$

לכן נקבל:

$$\boxed{\langle 0.3p + 0.6, \; 0.4 - 0.3p \rangle}$$

3.)

a.) Move to $d=1$, listen if GR, then OL:

$$E(R) = R(\text{move}) + R(\text{listen}) + \frac{1}{2} \cdot \big(p(TL \mid GR, d=1) \, R(TL, OL, d=1) + p(TR \mid GR, d=1) \cdot R(TR, OL, d=1) \big)$$

$R(\text{move}) = -2$, $R(\text{listen}) = -1$

$$p(TL \mid GR, d=1) \cdot R(TL, OL, d=1) \underset{\text{from symmetry}}{=} p(TR \mid GL, d=1) \cdot R(TL, OL, d=1)$$

b.) listen. if GL, then listen. Else if GR, then oL:

$$E(R) = R(\text{listen}) + \frac{1}{2}\cdot\left(p(TL|GL,d=2) + p(TR|GL,d=2)\cdot R(\text{listen})\right)$$

$$+\frac{1}{2}\cdot\left((p(TL|GR,d=2)\cdot R(TL,oL,d=2) + p(TR|GR,d=2)\cdot R(TR,oL,d=2))\right)$$

- $R(\text{listen}) = -1$
- $p(TL|GL,d=2)\cdot R(\text{listen}) = 0.6\cdot(-1) = -0.6$
- $p(TL|GL,d=2)\cdot R(\text{listen}) = 0.4\cdot(-1) = -0.4$
- $p(TL|GR,d=2)\cdot R(TL,oL,d=2) = p(TR|GL,d=2)\cdot R(TL,oL,d=2) = 0.4\cdot(-50) = \boxed{-20}$ (מהמודל)
- $p(TR|GR,d=2)\cdot R(TR,oL,d=2) = p(TL|GL,d=2)\cdot R(TL,oR,d=2) = 0.6\cdot 10 = \boxed{6}$ (מהמודל)

עכשיו נסכום ונקבל:

$$E(R) = 0.5\cdot(-1) + 0.5\cdot(-14) = \boxed{-8.5}$$

4.) א' נלקח 4 אפשרויות לסוכן ב- observations:

$\{GL,GL\}$ $\{GL,GR\}$ $\{GR,GL\}$ $\{GR,GR\}$

ב.) עבור $\{GL,GL\}$:

מאחר מנוקם ב- $p(TL|GL,d=1) = 0.9$

$p(TR|GR,d=1) = 0.1$

לכן נקבל:

$$p(TL \mid GL, GL, d=1) = \frac{p(GL \mid TL, GL, d=1) \cdot p(TL \mid GL, d=1)}{\underbrace{p(GL \mid GL, d=1)}}$$

$$p(GL \mid GL, d=1) = p(GL \mid GL, TL, d=1) \cdot p(TL \mid GL, d=1) + p(GL \mid GL, TR, d=1) \cdot p(TR \mid GL, d=1)$$

$$\Rightarrow p(GL \mid TL, d=1) \cdot p(TL \mid GL, d=1) + p(GL \mid TR, d=1) \cdot p(TR \mid GL, d=1) = 0.9 \cdot 0.9 + 0.1 \cdot 0.1 = \boxed{0.82}$$

$$= \frac{p(GL \mid LT, d=1) \cdot p(TL \mid GL, d=1)}{0.82} = \frac{0.9 \cdot 0.9}{0.82} = \boxed{0.988}$$

לכן beliefs agent יהיה:

$d=1$	:	0.988
$d=2$	:	$1-0.988 = 0.012$

II) עבור $\{GL, GR\}$:

נתונים מקומם :-

$$P(TL \mid GL, d=1) = 0.9$$

$$P(TR \mid GR, d=1) = 0.1$$

ערכו נחשב :-

$$P(TL \mid GL, GR, d=1) = \frac{P(GR \mid TL, GL, d=1) \cdot P(TL \mid GL, d=1)}{P(GR \mid GL, d=1)}$$

$$P(GR \mid GL, d=1) = P(GR \mid GL, TL, d=1) \cdot P(TL \mid GL, d=1) + P(GR \mid GL, TR, d=1) \cdot P(TR \mid GL, d=1)$$

$$\Rightarrow P(GR \mid TL, d=1) \cdot P(TL \mid GL, d=1) + P(GR \mid TR, d=1) \cdot P(TR \mid GL, d=1)$$

$$= 0.1 \cdot 0.9 + 0.9 \cdot 0.1 = \boxed{0.18}$$

ולכן נקבל :-

$$\frac{P(GR \mid LT, d=1) \cdot P(TL \mid GL, d=1)}{0.18} = \frac{0.1 \cdot 0.9}{0.18} = \boxed{0.5}$$

לכן נקבל :-

$$\boxed{\begin{array}{lcl} d=1 & : & 0.5 \\ d=2 & : & 1-0.5 = 0.5 \end{array}}$$

לסיכום נקבל 4 הצורות המאוחדות :-

$$\boxed{\begin{array}{l} \{GL, GL\} = \langle 0.988, 0.012 \rangle \\ \{GR, GL\} = \langle 0.5, 0.5 \rangle \\ \{GL, GR\} = \langle 0.5, 0.5 \rangle \\ \{GR, GR\} = \langle 0.012, 0.988 \rangle \end{array}}$$

לפי הנתונים לעיל :

! בנוסף לחוק הנשלים ההסתברות.

שאלה מספר 3.1 :

Question 3:

1.) we know that policy determines how the agent behaves

a.) μ^t - policy parameter, i.e. the location the robot will pick

b.) y^t - the action of the robot $\Rightarrow$ $y^t = \mu^t + Z$, $Z \sim N(0,\sigma^2)$

מכיוון ש $Z \sim N(0,\sigma^2)$ אזי נקבל ש- $y^t \sim N(\mu^t,\sigma^2)$

ואם מהתנהגות פונ ש- agent's policy -:

$$\pi(y|\mu) = L(\mu|y) = f_\mu(X=y) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(y-\mu)^2}{2\sigma^2}}$$

שאלה מספר 3.2 :

2) update rule :

לפי הגדרת הגרדיאנט נקבל :-

$$\mu = \mu + \alpha * G_t * \nabla_\mu \log \pi_\mu(y^t)$$

כאשר $G^t = \sum_{i=t+1}^{T} r_i$

* For the policy, we get that the log-likelihood equals:

$$\log \pi_\mu(y^t) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(y-\mu)^2}{2\sigma^2}}\right) = \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y-\mu)^2}{2\sigma^2}$$

$$= \log(1) - \log\left(\sqrt{2\pi\sigma^2}\right) - \frac{(y-\mu)^2}{2\sigma^2}$$

$$= -\log\left(\sqrt{2\pi\sigma^2} - \frac{(y-\mu)^2}{2\sigma^2}\right)$$

ונקבל גרדיאנט לפי הפרמטר הוא:

$$\nabla_\mu \log \pi_\mu(y^t) = -\frac{-2(y-\mu)}{2\sigma^2} = \boxed{\frac{y-\mu}{\sigma^2}}$$

לכן בסוף נקבל:

$$\boxed{\mu = \mu + \alpha_t G_t * \frac{y^t - \mu}{\sigma^2}}$$

שלב באופן הנגזרת לפי הפרמטרים

בתאור המלבן מה פתם ל. האיבר הוא אבל יותר מובן לדגמים שבלי יותר. תאור חיובי מכוון לכיוון הגרדיאנט ותאור שלילי הלך לכיוון הפוך

מודדים כ"כ איתנה ה- likelihood לפי הפלה y בים לפרמטר μ - לפי כיוון הסטיה שאבל ההתפלגות

שאלה מספר 3.3 :

3.) נתון לנו $y = \mu + z$ ו- $z \sim N(0,\sigma)$ לכן אם- $y \sim N(\mu,\sigma)$

ערך ערכך m אנו יכולים לקבל המקסימום של פו' התוחלת המצופה.

נחשב תוחלת ונקבל:

$$E(r) = E[-(m-y)^2] = -E[(m-y)^2] = -E[m^2-2my+y^2]$$

$$= -E[m^2] + E[2my] - E[y^2] = -m^2 + 2m\,E[y] + E[y^2]$$

$$= -m^2 + 2m\mu - \mu^2 - \sigma^2$$

$E[y] = \mu$ כי

$E[y^2] = \mu^2 - \sigma^2$

נגזור ונשווה לאפס פי לקבל המקסימום:

$$\left(-m^2 + 2m\mu - \mu^2 - \sigma^2\right)' = 2m - 2\mu = 0$$

$$\Rightarrow \boxed{m = \mu}$$

שאלה מספר 3.4 :

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

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

פתרון שלב אחר שלב:

הגדרת שלב העדכון:

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

לקיחת התוחלת:

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

החלפת $y_t = \mu + z$ **כך ש-** $z \sim N(0,\sigma^2)$:

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

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

פיתוח הריבוע:

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

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

הפרדת התוחלת:

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

הערכת התוחלות:

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

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

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

לכן :

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

$$= 2\alpha \cdot (m-\mu)$$

הסבר :

כלל העדכון מראה ששיעור השינוי הממוצע של μ הוא יחסי ישיר להבדל בין היעד m והערכת הנוכחית μ. קצב הלמידהα מגדיר את מהירות השינוי. זה אומר שככל שהסוכן מעדכן שוב ושוב את μ, הוא יכוון בהדרגה את μ לכיוון היעד m , והמהירות של התאמה זו מושפעת מה־ α.

פתרון זה מראה כיצד אלגוריתם $REINFORCE$ מתאים את פרמטר המדיניות μ עם הזמן, ומבטיח שהרובוט משתפר בקליעת המטרה על ידי הפחתת ההבדל בין μ ל־ m.

שאלה מספר 3.5 :

5) צריך לחשב:- $\nabla_\sigma \log \pi(y|\mu,\sigma)$

$$\nabla_\sigma \log \pi(y|\mu,\sigma) = \nabla_\sigma\left(\log \frac{1}{\sigma\sqrt{2\pi}}\right) + \log\left(e^{-\frac{(y-\mu)^2}{2\sigma^2}}\right)$$

$$= \nabla_\sigma\left(-\log(\sigma\sqrt{2\pi}) - \frac{(y-\mu)^2}{2\sigma^2}\right) = \frac{-\sqrt{2\pi}}{\sigma\sqrt{2\pi}} \cdot \frac{(y-\mu)^2}{\sigma^2} = \frac{(y-\mu)^2}{\sigma^2} - \frac{1}{\sigma}$$

$$= \frac{(y-\mu)^2 - \sigma^2}{\sigma^2}$$

לכן כלל העדכון:

$$\begin{bmatrix}\mu^t \\ \sigma^t\end{bmatrix} = \theta^t = \theta^{t-1} - \alpha(m-y)^2 \begin{bmatrix} \frac{(y-\mu)^{t-1}}{(\sigma^{t-1})^2} \\ \frac{(y-\mu^{t-1})^2 - (\sigma^{t-1})^2}{(\sigma^{t-1})^3} \end{bmatrix}$$

מסעיף הקודם קיבלנו כי $\mathbb{E}[\Delta\mu] = 2\alpha(m-\mu)$, לכן מאבלים על כך שהעדכון הינו פרופורציונלי להבדל בין m ל- μ.

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

$y \sim N(\mu,\sigma)$

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

$(m-y) \sim N(m-\mu,\sigma)$

$$= \frac{-\alpha}{\sigma^2}\left((m-\mu)^2\sigma^2 + 3\sigma^4 - \sigma^2(\sigma^2 + (m-\mu)^2)\right) = \boxed{-2\alpha\sigma}$$

* אנו יכולים לראות שאם שאף את קטן מ- m, הדפדפן יצליח את מן, ואם אף שלא מ- m, הדפדפן יקטין אותו מן ובכל ג- מקרב ל-m הדפדפנים קטנים יותר עד שדרפדפ (m-מן) יגיע לטופס.

=> עכשיו אחרי החישובים הגענו ל- $\sum[\Delta\sigma]=-2\alpha\sigma$.

לכן אפשר להסיק ש- σ קטן עד הפמן בגלל הדפדפן עד שיתאפס.

כי הגיוני מכיוון ש- magnet פוגע במטרה, כלומר הוא יעבוד על האנרגיה ומפחית אותה, וכך מגיעים לפיוק יותר.