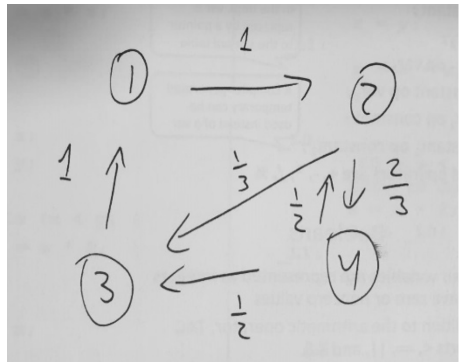
TEL AVIV UNIVERSITY אוניברסיטת תל אביב

<u>מגישים:</u>
דן מלכא 304773591
אופק ברנסקי 311170138
דן ברוך 207042391

<u>שאלה 1</u>

1. מכונת המצבים הינה:



2. הגרף מחלקת קשירות אחת, לכן הגרף *Irreducible*, וקיימת מחלקת קשירות אחת
3. נתבונן בגרף שקיבלנו, נחשב את הזמן מחזור של כל מצב:

$$p_{1,1}^{3}, p_{1,1}^{5} > 0 \to gcd(1 \to 1) = d_1 = 1$$
$$p_{2,2}^{2}, p_{2,2}^{3} > 0 \to \gcd(2 \to 2) = d_2 = 1$$
$$p_{3,3}^{3}, p_{3,3}^{4} > 0 \to \gcd(3 \to 3) = d_3 = 1$$
$$p_{4,4}^{2}, p_{4,4}^{5} > 0 \to \gcd(4 \to 4) = d_4 = 1$$

לפיכך ניתן לקבוע כי השרשרת היא *aperiodic*.

4. נחשב את ההתפלגות הסטוציונרית:

$$[x_1, x_2, x_3, x_4] \cdot \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1/3 & 2/3 \\ 1 & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix} = [x_1, x_2, x_3, x_4]$$

$$\Downarrow$$

$$\begin{cases} x_3 = x_1 \\ x_1 + \dfrac{x_4}{2} = x_2 \\ \dfrac{x_2}{3} + \dfrac{x_4}{2} = x_3 \\ \dfrac{2x_2}{3} = x_4 \end{cases} \to \begin{cases} x_1 + x_2 + x_3 + x_4 = 1 \\ x_3 = x_1 \\ x_1 + \dfrac{x_4}{2} = x_2 \\ \dfrac{x_2}{3} + \dfrac{x_4}{2} = x_3 \\ \dfrac{2x_2}{3} = x_4 \end{cases} \to \begin{cases} x_1 = \dfrac{2}{3} x_2 \\ x_3 = \dfrac{2}{3} x_2 = x_1 \\ x_1 + x_2 + x_3 + x_4 = 1 \\ \dfrac{2x_2}{3} = \left(\dfrac{2}{3}\right)\left(\dfrac{3}{2}\right) x_1 = x_4 \end{cases}$$

$$\to x_1 + \frac{3}{2} x_2 + x_1 + \left(\frac{2}{3}\right)\left(\frac{3}{2}\right) x_1 = 1 \to 4.5 x_1 = 1 \to x_1 = \frac{2}{9}$$

$$\Downarrow$$

$$x_2 = 1.5 \left(\frac{2}{9}\right) x_1 = \frac{1}{3}, \qquad x_3 = \frac{2}{9}, \qquad x_4 = \left(\frac{2}{3}\right)\left(\frac{1}{3}\right) = \frac{2}{9}$$

$$\boldsymbol{\pi} = \left[\frac{\mathbf{2}}{\mathbf{9}}, \frac{\mathbf{1}}{\mathbf{3}}, \frac{\mathbf{2}}{\mathbf{9}}, \frac{\mathbf{2}}{\mathbf{9}}\right]$$

5. נחשב את ה *expected return time* לכל מצב:
הראנו כי השרשרת אי פריקה ומחזורית, לכן המצבים הינם *Positive recurrent*, לפי הנלמד בשיעור, לכל מצב מתקיים:

$$\underbrace{\pi_i}_{\text{התפלגות סטוציונרית}} = \frac{1}{\mathbb{E}[T_i]}$$

לכן:

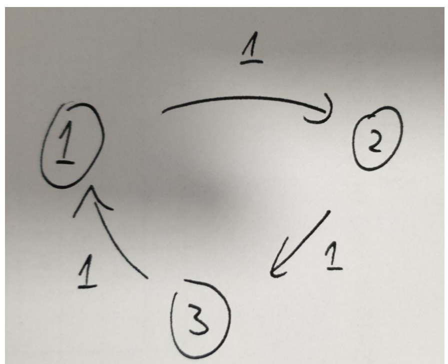
$$\mathbb{E}[T_1] = \mathbb{E}[T_3] = \mathbb{E}[T_4] = \frac{1}{\frac{2}{9}} = 4.5$$

$$\mathbb{E}[T_2] = \frac{1}{\frac{1}{3}} = 3$$

6. נגדיר מטריצה P' באופן הבא:

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

נקבל:



נחשב את ההתפלגות הסטוציונרית:

$$[x_1, x_2, x_3] \cdot \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = [x_1, x_2, x_3] \to \pi = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]$$

נקבל:

$$\mathbb{E}[T_1] = 3$$

וכמובן, מתקיים:

$$p_{1,1}^m = \begin{cases} 1 & \forall m \ s.t. \ m \bmod 3 = 0 \\ 0 & else \end{cases}$$

שאלה 2

1. נגדיר MDP לבעיה:

$$S = \{0,1,\dots,2k-1\}, S_0 = k$$
$$A = \{CW, CCW\}$$
$$P(s'|s,CW) = \begin{cases} 1, s' = s+1 \; mod \; 2k-1 \\ 0, \quad otherwise \end{cases}$$
$$P(s'|s,CCW) = \begin{cases} 1, s' = s-1 \; mod \; 2k-1 \\ 0, \quad otherwise \end{cases}$$
$$r(s,a) = \begin{cases} 1, s=0 \\ 0, otherwise \end{cases}$$

2. נבחין שעל מנת למקסם את הסיכוי לסיים ב-0, אנו רוצים להיות במצבים הקרובים ל-0 עד כמה שאפשר. לכן, המדיניות האופטימלית תהיה כזו ה"שואפת" להתקרב ל-0 עד כמה שאפשר, ולהישאר סביב 0 כאשר נמצאים במצבים הקרובים ל-0.

נוכל להגדיר באופן הבא:

$$\pi^*(s) = \begin{cases} CW, s > k \\ CCW, s \le k \end{cases}$$

עבור מדיניות זו, כל עוד אנו רחוקים ממצב 0, נתקדם לכיוון 0 במסלול הקצר ביותר, ואם אנו נמצאים במצב 0, נתקדם למצב *2k-1*, ובצעד הבא נחזור למצב 0.

3. נבצע איטרציה אחת של $value \; iteration$:

$$V_0(s) = 0, \forall s$$
$$V_1(s) = \max_{a\in A}\left\{r(s,a) + \gamma \sum_{\{s'\in S\}} p(s'|s,a)V_0(s')\right\} = \max_{a\in A}\{r(s,a)\}$$

מכיוון שלכל $s \neq 0$ מתקיים $r(s,a) = 0$ לכל a, ועבור $s = 0$ מתקיים $r(s,a) = 1$, נקבל:

$$V_1(s) = \begin{cases} 1, s=0 \\ 0, otherwise \end{cases}$$

4. נבצע איטרציה נוספת:

$$V_2(s) = \max_{a\in A}\left\{r(s,a) + \gamma \sum_{\{s'\in S\}} p(s'|s,a)V_1(s')\right\}$$

עבור $s \neq 2k-1, 0, 1$, מתקיים $p(s'|s,a) = 0$ או $V_1(s') = 0$ לכל s' (הם לא סמוכים למצב 0, לכן ההסתברות לעבור למצב s' עבורו $V_1(s')$ היא 0).

נתבונן במצבים שנותרו:

$$V_2(2k-1) = \max_{a\in A}\left\{r(2k-1,a) + \gamma \sum_{\{s'\in S\}} p(s'|2k-1,a)V_1(s')\right\}$$
$$= \max_{a\in A}\left\{\gamma \sum_{\{s'\in S\}} p(s'|2k-1,a)V_1(s')\right\} = \gamma p(s'|2k-1,CW) = \gamma$$

$$V_2(1) = \max_{a\in A}\left\{r(1,a) + \gamma \sum_{\{s'\in S\}} p(s'|1,a)V_1(s')\right\} = \max_{a\in A}\left\{\gamma \sum_{\{s'\in S\}} p(s'|1,a)V_1(s')\right\}$$
$$= \gamma p(s'|2k-1,CCW) = \gamma$$
$$V_2(0) = \max_{a\in A}\left\{r(0,a) + \gamma \sum_{\{s'\in S\}} p(s'|0,a)V_1(s')\right\} = r(0,CW) = 1$$

סה"כ:

$$V_2(s) = \begin{cases} 1, s = 0 \\ \gamma, s \in \{2k-1, 1\} \\ 0, otherwise \end{cases}$$

5. נחשב את $V^*(s)$ לכל $s \in \{0,1,2,3\}$:

חישבנו את V_0, V_1, V_2 לכל s, נמשיך בחישוב (אבחנה, משיקולי סימטריה, $V_i(1) = V_i(3)$, לכן אחשב רק את $V_i(1)$):

$$V_3(0) = \max_{a \in A}\left\{r(0,a) + \gamma \sum_{\{s' \in S\}} p(s'|0,a)V_2(s')\right\} = 1 + \gamma \cdot 1 \cdot \gamma = 1 + \gamma^2$$

$$V_3(1) = \max_{a \in A}\left\{r(1,a) + \gamma \sum_{\{s' \in S\}} p(s'|0,a)V_2(s')\right\} = 0 + \gamma \cdot 1 \cdot 1 = \gamma$$

$$V_3(2) = \max_{a \in A}\left\{r(2,a) + \gamma \sum_{\{s' \in S\}} p(s'|0,a)V_2(s')\right\} = 0 + \gamma \cdot 1 \cdot \gamma = \gamma^2$$

$$V_4(0) = \max_{a \in A}\left\{1 + \gamma \sum_{\{s' \in S\}} p(s'|0,a)V_3(s')\right\} = 1 + \gamma \cdot 1 \cdot \gamma = 1 + \gamma^2$$

$$V_4(1) = \max_{a \in A}\left\{0 + \gamma \sum_{\{s' \in S\}} p(s'|0,a)V_3(s')\right\} = \gamma \cdot 1 \cdot (1 + \gamma^2) = \gamma(1 + \gamma^2)$$

$$V_4(2) = \max_{a \in A}\left\{0 + \gamma \sum_{\{s' \in S\}} p(s'|0,a)V_3(s')\right\} = \gamma \cdot 1 \cdot \gamma = \gamma^2$$

$$V_5(0) = \max_{a \in A}\left\{1 + \gamma \sum_{\{s' \in S\}} p(s'|0,a)V_4(s')\right\} = 1 + \gamma \cdot 1 \cdot (\gamma(1 + \gamma^2) = 1 + \gamma^2(1 + \gamma^2)$$

$$V_5(1) = \max_{a \in A}\left\{0 + \gamma \sum_{\{s' \in S\}} p(s'|0,a)V_4(s')\right\} = \gamma \cdot 1 \cdot (1 + \gamma^2) = \gamma(1 + \gamma^2)$$

$$V_5(2) = \max_{a \in A}\left\{0 + \gamma \sum_{\{s' \in S\}} p(s'|0,a)V_4(s')\right\} = \gamma \cdot 1 \cdot \gamma(1 + \gamma^2) = \gamma^2(1 + \gamma^2)$$

ניתן לראות דפוס, ולחשב במקרה הכללי:

$$V_i(0) = 1 + \gamma \cdot V_{i-1}(1)$$
$$V_i(1) = \gamma \cdot V_{i-1}(0)$$

כאשר $V_{i+1} = V_i$ עבור i זוגי.

אזי:

$$V_i(0) = 1 + \gamma^2 \cdot V_{i-2}(0) = 1 + \gamma^2 + \gamma^4 \cdot V_{i-4}(0) \ldots$$

ולכן בגבול לאינסוף:

$$V^*(0) = \cdots = 1 + \gamma^2 + \gamma^4 + \cdots = \sum_{i=0}^{\infty}(\gamma^2)^i = \sum_{i=0}^{\infty}(\gamma^2)^i$$

זהו טור הנדסי ו- $\gamma^2 < 1$ לכן $V^*(0) = \frac{1}{1-\gamma^2}$

באופן דומה, נקבל:

$$V_i(1) = \gamma \cdot \left(1 + \gamma \cdot V_{i-2}(1)\right) = \gamma + \gamma^2 \cdot V_{i-2}(1) = \gamma + \gamma^3 + \gamma^5 + \cdots$$

$$V^*(1) = \sum_{i=0}^{\infty} \gamma^{2i+1} = \gamma \sum_{i=0}^{\infty} \gamma^{2i} = \frac{\gamma}{1-\gamma^2}$$

$$V^*(3) = \frac{\gamma}{1-\gamma^2}$$

עבור $V_i(2)$ הנוסחה הכללית היא:

$$V_i(2) = \gamma \cdot V_i(1) = \gamma \cdot (\gamma + \gamma^3 + \gamma^5 + \cdots) = \gamma^2 + \gamma^4 + \gamma^6 \ldots$$

$$V^*(2) = \sum_{i=0}^{\infty} \gamma^{2i+1} = \gamma^2 \sum_{i=0}^{\infty} \gamma^{2i} = \frac{\gamma^2}{1-\gamma^2}$$

TEL AVIV UNIVERSITY אוניברסיטת תל אביב

שאלה 3

1. נגדיר MDP:

$$\mathcal{S} = \{(x, d_i) |\ x \subseteq \{0,1,2..N\}, d_i \in \{0,1,..9\}\}$$

כאשר:

- x הינו קבוצת ה$slots$ הפנויים, שלא שובצו מספרים בהם (יכול לקבל כל תת קבוצה של $\{0,1..N\}$
- d_i הינו המספר שהוגרל, ונדרש לשיבוץ באחד ה $slots$

בנוסף:

$$\mathcal{A} = \{a_{x,j} | x \subseteq \{0,1,..N\}, j \in x\}$$

ז"א, בהינתן מצב $s' \coloneqq (x', d_i')$, קבוצת הפעולות האפשריות (ז"א הצבת מספר d_i שהוגרל באחד מהתאים הריקים מתוך s), הינו האוסף:

$$\mathcal{A}_{s'} = \{a_{x',j} | x' = x, j \in x'\}$$

ונגדיר את פונקציית הרווח עבור $s' = (x, d_i)$ ו $a_{x,j}$

$$\mathcal{R}(s', a_{x,j}) = d_i \cdot 10^j$$

ז"א, עבור הצבת המספר d_i בתא j, נקבל "תוספת" רווח למספר הסופי מהצורה הנ"ל.

נגדיר את מטריצת המעברים-
עבור פעולה $a_{x,j}$ המציבה את המספר d_i בתא j, יש לנו 10 מצבים שונים שניתן לעבור אליהם (בהתאם להגרלת המספר הבא) כל אחד בסיכוי $1/10$:

אז, עבור מעבר ממצב $s_t := (x, d_i)$, ביצענו את הפעולה $a_{x,j}$, נעבור למצב $s_{t+1} := (x_{t+1}, d_i')$ באופן הבא:

$$p_{s_t, s_{t+1}} = \begin{cases} 1/10 & x' \coloneqq x/\{j\} \\ 0 & else \end{cases}$$

לדוגמה, עבור:

$$s_t = (x_t, d_i | d_i \in \{0,..9\}), \quad a_t = a_{x_t,j}$$
$$\downarrow$$
$$s_{t+1} := (x_{t+1}, d_i') \ \coloneqq \big((x_t/\{j\}), d_i'\big)$$

ומצבי ההתחלה יהיו:

$$s_0 = \{\{0,1,..N\}, d_i | d_i \in \{0,1,..9\}\}$$

בנוסף הבעיה היא *finite horizon*, לכן $\gamma = 1$

2. למדנו בשיעור כי:

$$the\ tail\ of\ an\ optimal\ policy\ is\ optimal\ for\ the\ tail\ problem$$

עבור MDP עם *finite horizon*, לפיכך המדיניות האופטימלית עבור מצב s בזמן t איננו מושפע מהגרלת המספרים קודמת, ואיפה הם שובצו אלא רק מהמצב הנתון – קבוצת ה *slots* הריקים והמספר הרנדומי שדורש שיבוץ.

נתבונן על החלטה של מדיניות האופטימלית π^* עבור מצב s_t, אז לפי הטענה שציינו לעיל, המדיניות האופטימלית עבור מצב s_t תהיה זהה למדיניות כאילו s_t היה המצב ההתחלתי, עבור בעיה קטנה יותר.

נראה כי קיימת מדיניות אופטימלית יחידה-

עבור 2 מצבים שונים s_1, s_2 (כאשר הוצבו מספרים שהוגרלו בעבר, בתאים שונים, אך בשני המצבים נשארו m תאים ריקים), אנו נדרשים להציב מספר שהוגרל d_i באחד מהתאים הריקים מתוך m תאים.

הרי כי מצבים אלו שקולים, עד כדי הפעלת פונקציה מונוטונית על הרווח שהצטבר עד כה-

נניח כי הרווח המקסימלי עבור s_2 יתבצע ע״י הצבת המספר d_i שהוגרל בתא j, אז המדיניות עבור s_1 תניב רווח מקסימלי ע״י השמת המספר d_i בתא j', כמובן ברווח יהיה כפולה של המספר *10*.

נראה את המדיניות בעזרת אינדוקציה - על כמות המקומות הריקים – *slots* שנמצאים ב״משחק״:

<u>מקרה בסיס-</u>

נותר *slot* אחד פנוי, אז ברור כי המדיניות האופטימלית תבצע השמה של המספר i שהוגרל בתא הריק.

<u>הנחה-</u>

עבור m תאים פנויים, נניח כי בוצע מדיניות אופטימלית של השמת מספרים בתאים $N-M$, נסמן אותה כ a', והרווח שלה כ $r'(s', a')$ (כאשר s' מתאר את המצב בו אנו נמצאים, ואיזה מספר הוגרל).

<u>צעד-</u>

עבור *m+1* תאים פנויים, נראה כי המדיניות האופטימלית איננה תלויה בעבר, אלא רק בתאים הריקים והמספר שהוגרל:

$$\pi^* = \underset{a_{\{\dots\},j}}{argmax}\left[\left\{j \cdot d_i + \frac{1}{10} \cdot \left(\sum_0^9 r'(s', a')\right)\right\} \mid \forall j \in empty\ slots\right]$$

לפיכך ניתן לראות כי עבור פונקציית הרווח, היא איננה תלויה בשיבוצים הקודמים, כל עוד בוצע השמה אופטימלית, אלא רק בכמה תאים פנויים יש, ומה המספר שהוגרל.

3. נשתמש ב *backwards recursion* –

a. עבור הספרה האחרונה d_i, יש לנו בדיוק מקום אחד לשים אותה, לפיכך:

$$\forall d_i \in \{0,1,..9\}, \qquad j \in \{0,1,2\}$$
$$\downarrow$$
$$\pi^*(\{j\}, d_i\} = a_{\{\},j}, \quad r = d_i \cdot 10^j$$

b. עבור הספרה לפני אחרונה, ישנם 3 מצבים אפשריים

$\_\,\_\,X,\ \_\,X\,\_\,,\ X\,\_\,\_$

נחשב את הפעולה האופטימלית עבור כל מצב:

$X\_\,\_$:

$$\pi^*(\{0,1\}, d_i) = \underset{a}{argmax}\left\{100d_i + \frac{1}{10} \cdot \sum_{k=0}^{9} 10k, \qquad 10d_i + \frac{1}{10} \cdot \sum_{k=0}^{9} 100k\right\}$$
$$= \underset{a}{argmax}\{100d_i + 45, 10d_i + 450\}$$
$$\downarrow$$
$$a_{\{0,1\},0}\ if\ d_i > 4, \qquad a_{\{0,1\},1}\ if\ d_i \le 4$$

ז״א, אם המספר d_i שהוגרל קטן מ5, נשים אותו בתא האמצעי (אינדקס *1*), אחרת נשים אותו בתא השמאלי (אינדקס 0).

$\_\,X\,\_$:

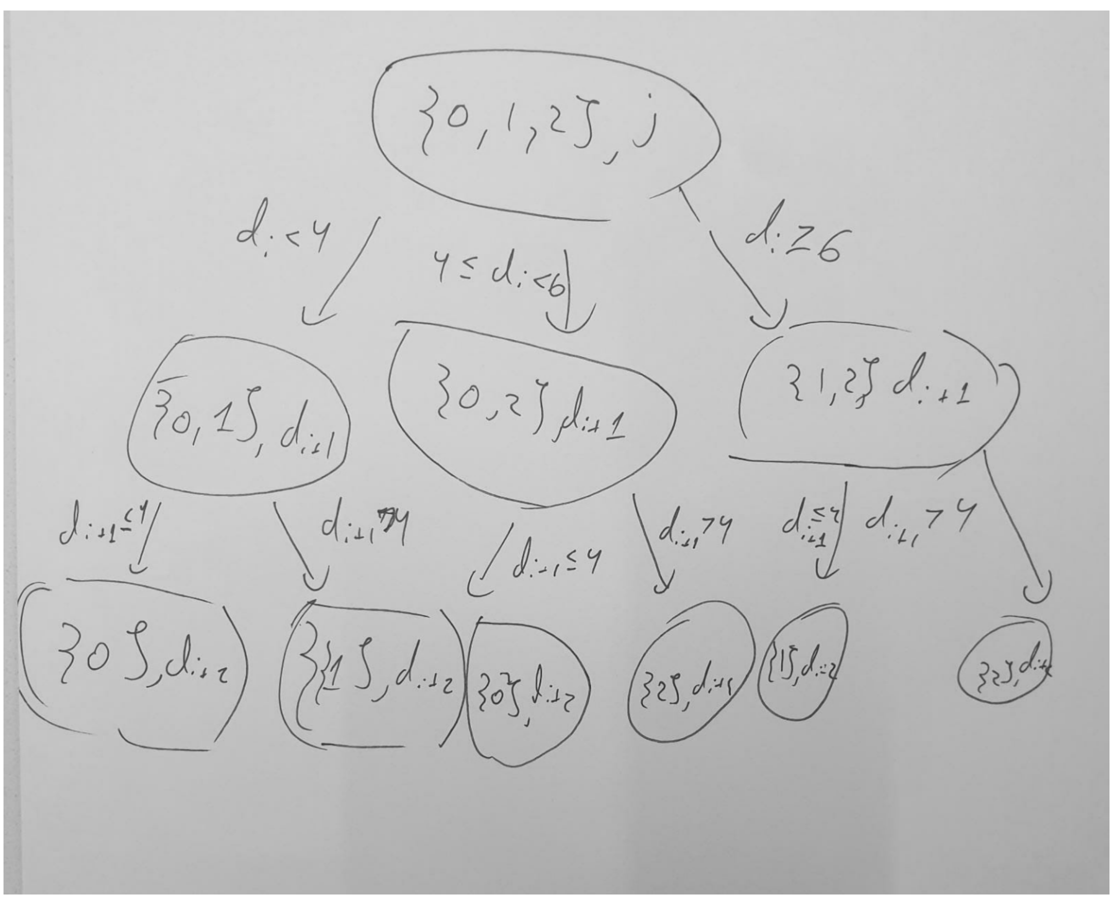
$$\pi^*(\{0,2\},d_i) = \underset{a}{argmax}\left\{100d_i + \frac{1}{10}\cdot\sum_{k=0}^{9}k, \qquad d_i + \frac{1}{10}\cdot\sum_{k=0}^{9}100k\right\}$$
$$= \underset{a}{argmax}\{100d_i + 45, d_i + 450\}$$
$$\downarrow$$
$$a_{\{0,2\},0}\ if\ d_i > 4, \qquad a_{\{0,2\},2}\ if\ d_i \le 4$$

: x _ _

$$\pi^*(\{1,2\},d_i) = \underset{a}{argmax}\left\{10d_i + \frac{1}{10}\cdot\sum_{k=0}^{9}k, \qquad d_i + \frac{1}{10}\cdot\sum_{k=0}^{9}10k\right\}$$
$$= \underset{a}{argmax}\{10d_i + 4.5, di + 45\}$$
$$\downarrow$$
$$a_{\{1,2\},1}\ if\ d_i > 4, \qquad a_{\{1,2\},2}\ if\ d_i \le 4$$

c. עבור הספרה הראשונה, נדרש להחליט היכן למקם אותה:

$$\pi^*(\{0,1,2\},d_i) = \underset{a}{argmax}\left\{\begin{matrix} 100i + \frac{1}{10}\left(5\cdot 45 + \sum_{k=0}^{4}k + 5\cdot 4.5 + \sum_{k=5}^{9}10k\right) \\ 10i + \frac{1}{10}\left(5\cdot 450 + \sum_{k=0}^{4}k + 5\cdot 4.5 + \sum_{k=5}^{9}100k\right) \\ i + \frac{1}{10}\left(5\cdot 450 + \sum_{k=0}^{4}10k + 5\cdot 45 + \sum_{k=5}^{9}100k\right) \end{matrix}\right\}$$
$$\downarrow$$
$$= \underset{a}{argmax}\,[100i + 60.75, 10i + 578.25, i + 607.5]$$
$$\downarrow$$
$$a_{\{0,1,2\},0}\ \ if\ d_i > 5$$
$$a_{\{0,1,2\},1}\ \ if\ 3 < d_i \le 5$$
$$a_{\{0,1,2\},2}\ \ if\ d_i \le 3$$



שאלה 4

$$V \in \mathbb{R}^{|S|};\ \ TV(s) = \frac{1}{|A|}\sum_{a\in A}\left(r(s,a) + \gamma \sum_{s\in S} p(s'|s,a)V(s')\right)$$

$Show\ that\ T\ is\ \gamma - contracting\ w.r.t\ Max - Norm$

Which is equivalent to:

$$\|T(V_1) - T(V_2)\|_\infty \le \gamma \cdot \|V_1 - V_2\|_\infty$$

$$\left\|\frac{1}{|A|}\sum_{a\in A}\left(\cancel{r(s,a)} + \gamma \sum_{s\in S} p(s'|s,a)V_1(s')\right) - \frac{1}{|A|}\sum_{a\in A}\left(\cancel{r(s,a)} + \gamma \sum_{s\in S} p(s'|s,a)V_2(s')\right)\right\|_\infty =$$

$$\left\|\frac{1}{|A|}\ \gamma \sum_{a\in A}\left(\sum_{s\in S} p(s'|s,a)(V_1(s') - V_2(s')\right)\right\|_\infty \underbrace{\le}_{\substack{Triangle\ ineq.\\ Max-Norm\\ not\ affecting\\ left\ term}}$$

$$\le \frac{1}{|A|}\gamma \cdot \sum_{a\in A}\left(\underbrace{\sum_{s\in S} p(s'|s,a)}_{\sum_p prob\ =1}\right)\|V_1(s') - V_2(s')\|_\infty = \frac{\cancel{1}}{\cancel{|A|}}\gamma \cdot \underbrace{\sum_{a\in A} 1}_{\cancel{|A|}} \cdot \|V_1(s') - V_2(s')\|_\infty =$$

$$= \gamma \cdot \|V_1 - V_2\|_\infty; \quad Q.E.D$$