**מבוא לבינה מלאכותית**

**236501**

**דו"ח הגשה – תרגיל בית רטוב 3**

**מתן צחורי 208936989**

**אלון פנפיל 318598166**

**חלק א' – MDP ו-RL**

**חלק א' – חלק יבש**

שאלה 1:



**function** VALUE-ITERATION(, ) **returns** a utility function

**inputs:** - an MDP with states , actions , transition model ,

rewards , discount

- the maximum error allowed in the utility of any state

**local variables:** , - vectors of utilities for states in , initially zero

the maximum change in the utility of any state in an iteration

**Repeat**

**for each** state **in****do**

**if** **then**

**until**

**return**

**function** POLICY-ITERATION() **returns** a policy

**inputs:** - an MDP with states , actions , transition model ,

rewards , discount

**local variables:** a vector of utilities for states in , initially zero

a policy vector indexed by state, initially random

**Repeat**

**for each** state **in****do**

curr\_eval =

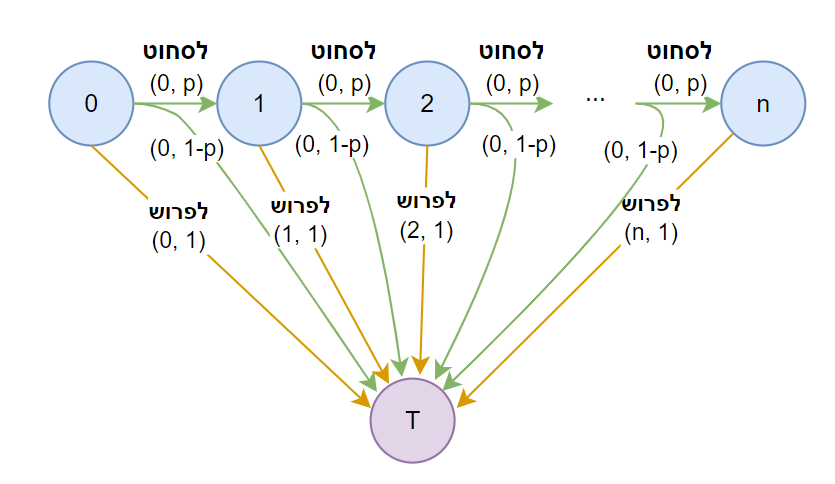
**if** **then**

**until**

**return**

בסעיפים ג' ו-ד', כאשר , אנחנו כבר לא נמצאים במצב של , והתועלת עלולה להיות אינסופית. במצב זה, לא ניתן להבדיל בין איכות של מדיניות.  
על מנת לפתור זאת, יש לדרוש שבסביבה יהיה קיים מצב סופי, ויש לדרוש על ה- שהמדיניות תבטיח שהסוכן יגיע למצב סופי.

שאלה 2:

1. נגדיר את הבעיה כבעיית באופן הבא:  
   נגדיר את קבוצת המצבים:  
   כאשר , .  
     
   נגדיר את הפעולות לכל מצב:  
   נגדיר את מודל המעבר:  
   נגדיר את פונקציית התגמולים:  
   נציג את ה- בתרשים מצבים:  
   כאשר קשתות ירוקות הן פעולת סחיטה, קשתות כתומות הן פעולת פרישה, והצמד על הקשת הוא:  
   (<הסתברות>,<תגמול פעולה>)
2. לא ניתן לנסח את הבעיה עם מצב יחיד ומצב סיום, מכיוון שהתגמול משתנה בהתאם למספר הפעמים שהסוחט הצליח לסחוט והקורבן לא קרא למשטרה.
3. כן, היינו יכולים לנסח את הבעיה בצורה שונה בה התגמול על סחיטה מוצלחת הוא 1 במקום 0, והתגמול על סחיטה לא מוצלחת הוא בהינתן שנמצאים במצב במקום 0. התגמול על פרישה הוא 0.
4. בהינתן , לפי פונקציית העדכון של בלמן מתקיים:

נציב את ל-, ונקבל שנעדיף לבחור "לסחוט" כאשר .

עבור מקרה זה, נקבל שגם ב- נעדיף "לסחוט".

במקרה ש- :  
נקבל שנעדיף לבחור "לסחוט" במצב 1 (ולפרוש ב-2) כאשר .

במקרה ש- :

נקבל שנעדיף "לסחוט" במצב 0.

במקרה ש- :  
נקבל שנעדיף "לפרוש" במצב 1. במצב 0 נעדיף "לסחוט", ובמצב 2, 3 לא משנה מה נעשה כי לא נגיע לשם.

|  |  |  |
| --- | --- | --- |
| תועלת | מדיניות | ערכי |
|  |  |  |
|  |  |  |
|  |  |  |

**חלק ב' – מבוא ללמידה**

**חלק א' – חלק יבש**