

Reinforcement Learning

המטרה: policy למצוא את המדיניות הטובה ביותר $S \rightarrow A$ המביאה ל- R .

כדי לעשות זאת, נעזרים במדיניות $policy$ ובה פונקציה $reward$.

$$R: \bigcup (S \times A)^T \rightarrow \mathbb{R} : \vec{s} = ((s_1, a_1), \dots, (s_T, a_T))$$

ואם נרצה למצוא את המדיניות הטובה:

$$\max_{policy} \mathbb{E}_{\vec{s}} [R(\vec{s})]$$

↑
expectation over episodes \vec{s}

אפשר לייצג קצוות SL במקרה שבו יש לנו RL . אבל עדיין יש כמה קשיים:

1. ק-SL, פעולה לא מוגדרת על הסביבה

⇐ אפשר להשתמש ב- $Policy$ מלא ורק אז נאסוף נתונים

2. ק-SL, ההערכה של פעולה היא מקומית בלבד.

3. ק-SL, נרצה לנו (המלומד) היתרון (ב- RL) איתנו מן הקודם

$reward$ מסיים על הסביבה שלנו.

העולם בו יש לנו מומחה (expert):

אפשר להתחיל מזה כ- supervised ולתקן את המומחה...

קצוות: 1. אי אפשר להיות יותר טובים מהמומחה ממנו למדנו.

2. distribution drift:

איתנו על המערכת מסווגת אבל קודם

כרגע בסעיף, אולי יבואו מחולף לאולם המכונים

שלא כאלו וקצוות המקומית שאיתנו לא מוכרים בפעם!

העולם בו יורדים את החול ומוכרים איתנו איך לעשות בו:

MDP - מתימים להחליט קודם קצוות קטנים ודיווח.

* ואלו מקדמים פונקציות V, Q עוזרים למצוא אסטרטגיה

expected reward starting from current state with current action

expected reward starting from current state on

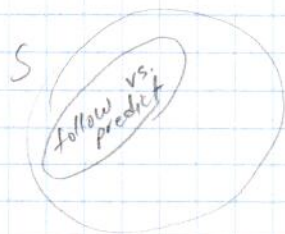
Value Iteration

אלו מלומדים קואלוריים איטרטיביים

מה המטרה? צריך לעבור על כל המצבים האפשריים באמצעות המדיניות

אלו הם מרחב המצבים גדול (נחיה קצוות)

⇐ Approximate Value Iteration



מכונה אל תהיה - חשבון

* Tree Search אל תהיה אל תהיה אל תהיה

הפסד אל תהיה אל תהיה אל תהיה אל תהיה אל תהיה

אל תהיה אל תהיה אל תהיה אל תהיה אל תהיה

אל תהיה אל תהיה אל תהיה אל תהיה אל תהיה

* AlphaZero אל תהיה אל תהיה אל תהיה אל תהיה

אל תהיה אל תהיה אל תהיה אל תהיה