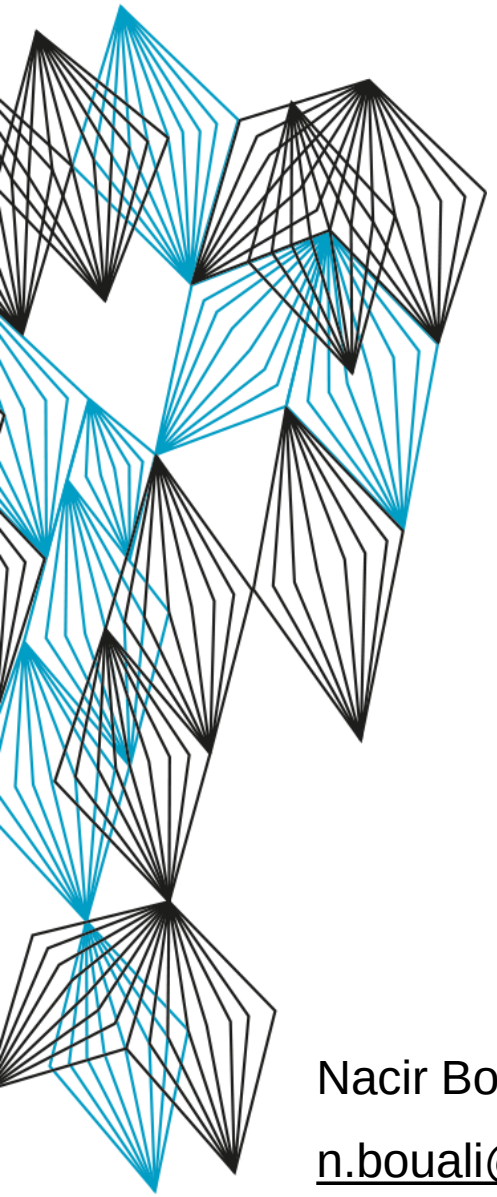

ARTIFICIAL INTELLIGENCE & CYBER SECURITY

REINFORCEMENT LEARNING

DIRECT UTILITY ESTIMATION

Nacir Bouali

n.bouali@utwente.nl

UNIVERSITY OF TWENTE.

REINFORCEMENT LEARNING

- We worked with two algorithms for MDPs;
  - Value Iteration
  - Policy Iteration
- Problem;
  - Transition model and rewards in an environment are not known a priori.

REINFORCEMENT LEARNING

- We worked with two algorithms for MDPs;
  - Value Iteration
  - Policy Iteration
- Problem;
  - Transition model and rewards in an environment are not known a priori.
- The real goal of RL is to find the optimal policy without knowing the transition model of the environment

REINFORCEMENT LEARNING - PASSIVE

- Given a fixed policy π, we want to learn the utility function without a prior knowledge of P(s'|s,a) or R(s)
- Since we don't know the transition model, we need to learn it from some trials;

$$(1,1) \xrightarrow[Up]{-.04} (1,2) \xrightarrow[Up]{-.04} (1,3) \xrightarrow[Right]{-.04} (1,2) \xrightarrow[Up]{-.04} (1,3) \xrightarrow[Right]{-.04} (2,3) \xrightarrow[Right]{-.04} (3,3) \xrightarrow[Right]{+1} (4,3)$$

$$(1,1) \xrightarrow[Up]{-.04} (1,2) \xrightarrow[Up]{-.04} (1,3) \xrightarrow[Right]{-.04} (2,3) \xrightarrow[Right]{-.04} (3,3) \xrightarrow[Right]{-.04} (3,2) \xrightarrow[Up]{-.04} (3,3) \xrightarrow[Right]{+1} (4,3)$$

$$(1,1) \xrightarrow[Up]{-.04} (1,2) \xrightarrow[Up]{-.04} (1,3) \xrightarrow[Right]{-.04} (2,3) \xrightarrow[Right]{-.04} (3,3) \xrightarrow[Right]{-.04} (3,2) \xrightarrow[Up]{-1} (4,2)$$

REINFORCEMENT LEARNING - PASSIVE

- We defined the utility as the (discounted) sum of rewards when a policy π is followed.

$$U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(S_t, \pi(S_t), S_{t+1})\right]$$

- Assuming for the grid world that γ=1 (we have final states).

$$(1,1) \xrightarrow[Up]{-.04} (1,2) \xrightarrow[Up]{-.04} (1,3) \xrightarrow[Right]{-.04} (1,2) \xrightarrow[Up]{-.04} (1,3) \xrightarrow[Right]{-.04} (2,3) \xrightarrow[Right]{-.04} (3,3) \xrightarrow[Right]{+1} (4,3)$$

$$(1,1) \xrightarrow[Up]{-.04} (1,2) \xrightarrow[Up]{-.04} (1,3) \xrightarrow[Right]{-.04} (2,3) \xrightarrow[Right]{-.04} (3,3) \xrightarrow[Right]{-.04} (3,2) \xrightarrow[Up]{-.04} (3,3) \xrightarrow[Right]{+1} (4,3)$$

$$(1,1) \xrightarrow[Up]{-.04} (1,2) \xrightarrow[Up]{-.04} (1,3) \xrightarrow[Right]{-.04} (2,3) \xrightarrow[Right]{-.04} (3,3) \xrightarrow[Right]{-.04} (3,2) \xrightarrow[Up]{-1} (4,2)$$

- How to estimate the utility of state (1,1)?
- $U^{\pi}(1,1)=(0.76+0.76-1.2)/3=0.32$
- $U^{\pi}(1,2)=(0.8+0.88+0.8-1.16)/3=0.44$