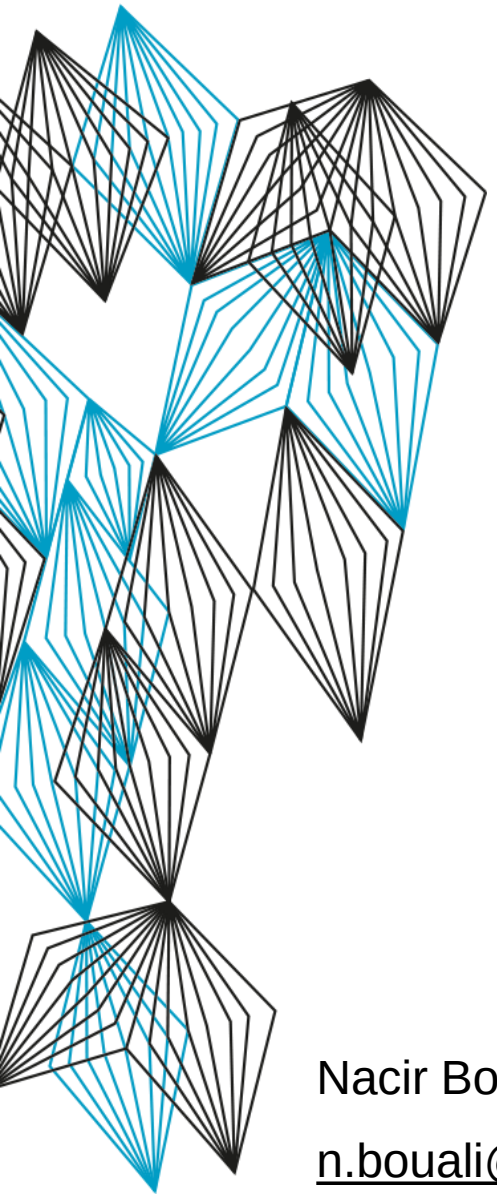

ARTIFICIAL INTELLIGENCE & CYBER SECURITY

REINFORCEMENT LEARNING

TEMPORAL DIFFERENCE

Nacir Bouali

n.bouali@utwente.nl

UNIVERSITY OF TWENTE.

REINFORCEMENT LEARNING

- ADP mandates frequent updates to the system of equations U(s) for every state s, after each transition/observation.
  - Not optimal if we have a large number of states.
  - Useless for states that are unreachable from a spefic state on which the transition was made.

TEMPORAL DIFFERENCE LEARNING

- If an observation confirms that transitioning from a state s to another s' occurs with a probability of 1, then

$$U(s) = R(s) + \gamma \sum_{s' \epsilon S} P(s'|s, \pi(s))U(s') = R(s) + \gamma\, U(s')$$

- There's no need to wait for the end of the trial to update U(s), we can update it using a factor or weight α;

$$U(s) = (1 - \alpha)U(s) + \alpha(R(s) + \gamma\, U(s'))$$

- α is a learning rate.

TEMPORAL DIFFERENCE LEARNING

- From $U(s) = (1 - \alpha)U(s) + \alpha(R(s) + \gamma\, U(s'))$
- We derive learning rule of temporal difference as;

$$U(s) = U(s) + \alpha\big(R(s) + \gamma\, U(s') - U(s)\big)$$

TEMPORAL DIFFERENCE LEARNING

- Initially, $U(s_0)= U(s_1)= U(s_2)= U(s_3)= 0$
- Agent wakes up at S_0: reward is -0.04 ($U(s_0)$=-0.04)
- ➢ Agent executes the action as mandated by the policy

$S_{0(-0.04)} \longrightarrow S_{0(-0.04)}$

$U(s_0)= U(s_0) + \alpha\left(R(s_0) + \gamma\, U(s_0) - U(s_0)\right)$

-0.04+0.2(-0.04+(1*-0.04)+0.04)=-0.048

$\boldsymbol{\alpha = 0.2 \text{ and } \gamma = 1}$

$$\boldsymbol{U(s) = U(s) + \alpha\left(R(s) + \gamma\, U(s') - U(s)\right)}$$