**Assignment 5**

Reinforcement Learning: assignment #1

Davide Brinati

Matricola: 771458

1. **Define the MDP underling the problem.**

* S = {s1, s2, s3, s4, s5, s6, s7}
* A = {Destra, Sinistra}
* T = Matrice di transizione per un processo deterministico, T(s,a) 🡪 s’
  + T(s1, Destra) = s3
  + T(s1, Sinistra) = s2
  + T(s2, Destra) = s5
  + T(s2, Sinistra) = s4
  + T(s3, Destra) = s7
  + T(s3, Sinistra) = s6
* R = funzione reward o payoff, R(s)
  + R(s1, Destra) = 0
  + R(s1, Sinistra) = 0
  + R(s2, Destra) = 0
  + R(s2, Sinistra) = 4
  + R(s3, Destra) = 3
  + R(s3, Sinistra) = 2

1. **Compute the optimal value function V\* using the value iteration algorithm.**

Assumiamo che 𝛾 = 1, utilizzando l’algoritmo di iterazione del valore avremo che:

* V(s4) = 0
* V(s5) = 0
* V(s6) = 0
* V(s7) = 0
* V(s2) = max{4, 0} = 4
* V(s3) = max{2, 3} = 3
* V(s1) = max{0 + 4, 0 + 3} = 4

1. **Suppose to have an initial policy which chooses equally between right and left at each junction, and assume 𝛾 = 0.5:**
2. **What is the value function Vπ0 for the initial policy?**

La *initial policy* ci da i seguenti risultati:

* A s1 sceglie Destra
* A s2 sceglie Destra
* A s3 sceglie Sinistra

Quindi il valore della funzione Vπ0  sarà:

* Vπ0(s1) = max{0 + 0.5\*0; 0 + 0.5\*2} = max{0, 1} = 1 (va a destra)
* Vπ0(s2) = max{4, 0} = 4 (va a sinistra)
* Vπ0(s3) = max{2, 3} = 3 (va a destra)

1. **What is the improved policy π1 based on the value function Vπ0?**

**[Policy Improvement]**

* + A s1 sceglie di andare a Destra
  + A s2 sceglie di andare a Sinistra
  + A s3 sceglie di andare a Destra

1. **What is the new value function Vπ1? [Policy Evaluation]**

* Vπ1(s1) = max{0 + 0.5\*4; 0 + 0.5\*3} = max{2, 1.5} = 2 (va a sinistra)
* Vπ1(s2) = max{4, 2} = 4(va a sinistra)
* Vπ1(s3) = max{2, 3} = 3 (va a destra)

1. **What is the improved policy 𝜋2 based on the value function 𝑉𝜋1? [Policy Improvement]**

* A s1 sceglie di andare a sinistra
* A s2 sceglie di andare a sinistra
* A s3 sceglie di andare a destra