# UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Corso di laurea magistrale in Data Science

Decision Models

# Assignment 5

**Federico Manenti**
Matr. 790032

# 1 Introduzione

Nel *Problema Decisonale di Markov* proposto un uomo deve percorrere un tragitto dal punto di partenza (stato 0) al punto di arrivo (casa) cercando di ottenere la *reward* più alta possibile. Gli stati intermedi possibili sono 6, ad ogni mossa l'agente deve scegliere se procedere a destra o a sinistra e ogni percorso ha una *reward* propria.
Un *Problema Decisonale di Markov* è definito da:

- **S**: lo spazio degli Stati
- **A**: lo spazio delle Azioni
- **T**: la matrice delle transizioni (il caso in questione è un problema deterministico), $T(s,a) \rightarrow s'$
- **R(s,a)** : funzione di *payoff* o *reward*

Una *Policy* $\pi$ invece mappa uno stato ad un azione, $\pi : S \rightarrow A$.
Risolvere un *problema di Markov* significa trovare la *Policy* ottimale $\pi^*$ che massimizza (o minimizza) la *reward* (o i *costi*).
La funzione di stato è data dall'equazione di *Bellman*:

$$V(s)^{\pi} = R(s,\pi(s)) + \gamma \sum_{s' \in S} T(s,\pi(s),s') \cdot V^{\pi}(s') \tag{1}$$

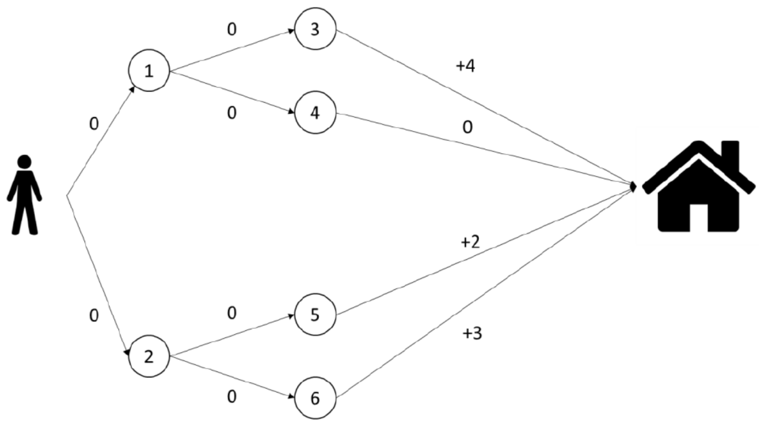


Figura 1: Percorso

La *policy* iniziale $\pi_0$ data è:

- $0 \to 2$
- $2 \to 5$
- $1 \to 4$
- Da 3, 4, 5, 6 è possibile andare solo allo stato finale

Le richieste da risolvere sono:

1. Calcolare la funzione di stato $V^{\pi_0}$ per ogni stato
2. Aumentare la *Policy* per ottenere la nuova $\pi_1$
3. Calcolare la nuova funzione di stato $V^{\pi_1}$ per ogni stato
4. Aumentare ancora la *policy* per ottenere $\pi_2$
5. Dopo ciò la soluzione converge alla *Policy* ottimale?

## 2 Risoluzione

La *Policy* ottimale $\pi^*$ è quella per cui $V^{\pi^*}(s) \geq V^{\pi}(s)$ $\forall s$ e $\forall \pi$. Quindi:

$$V^*(s) = \max_{a \in A}\big[R(s,a) + \gamma \sum_{s' \in S} T(s, \pi(s), s') \cdot V^*(s')\big] \tag{2}$$

In questo problema è stato considerato $\gamma = 1$.
Si calcola ora la funzione di stato per la *Policy* $\pi_0$:

- $V^{\pi_0}(0) = \max_{a \in A}\{0+0+0; 0+0+2\} = \max_{a \in A}\{0; 2\} = 2 \Rightarrow 0 \to 2$
- $V^{\pi_0}(1) = \max_{a \in A}\{0+4; 0+0\} = \max_{a \in A}\{4; 0\} = 4 \Rightarrow 1 \to 3$
- $V^{\pi_0}(2) = \max_{a \in A}\{0+2; 0+3\} = \max_{a \in A}\{0; 3\} = 3 \Rightarrow 2 \to 6$
- Per $3, 4, 5, 6$ non vengono calcolate perché non c'è una possibile scelta, ma valgono rispettivamente $4, 0, 2, 3$

La nuova *Policy* $\pi_1$ quindi diventa:

- $0 \to 2$
- $2 \to 6$

- $1 \to 3$
- Da 3, 4, 5, 6 è possibile andare solo allo stato finale

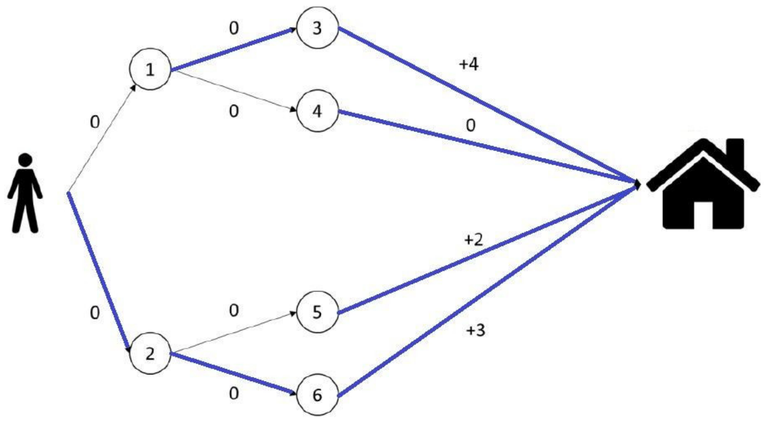


Figura 2: *Policy* $\pi_1$

Si calcola la nuova funzione di stato per $\pi_1$:

- $V^{\pi_1}(0) = \max\limits_{a \in A}\{0+0+4; 0+0+3\} = \max\limits_{a \in A}\{4; 3\} = 4 \Rightarrow 0 \to 1$
- $V^{\pi_1}(1) = \max\limits_{a \in A}\{0+4; 0+0\} = \max\limits_{a \in A}\{4; 0\} = 4 \Rightarrow 1 \to 3$
- $V^{\pi_1}(2) = \max\limits_{a \in A}\{0+2; 0+3\} = \max\limits_{a \in A}\{0; 3\} = 3 \Rightarrow 2 \to 6$
- Per $3, 4, 5, 6$ non vengono calcolate perché non c'è una possibile scelta, ma valgono rispettivamente $4, 0, 2, 3$

La nuova *Policy* $\pi_2$ quindi diventa:

- $0 \to 1$
- $2 \to 6$
- $1 \to 3$
- Da 3, 4, 5, 6 è possibile andare solo allo stato finale

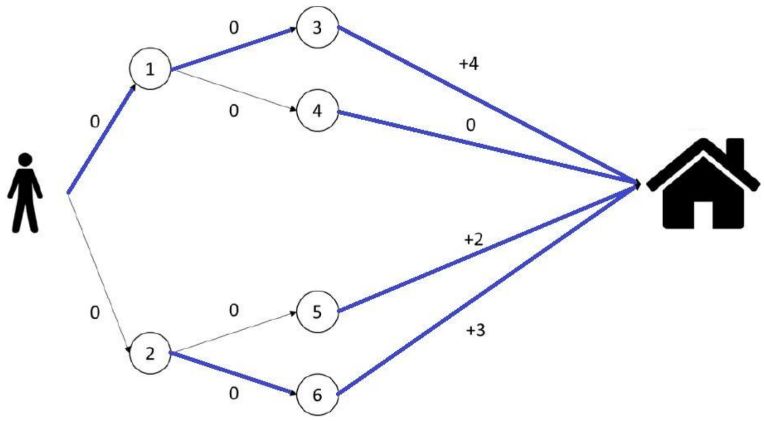


Figura 3: *Policy* $\pi_2$

Si verifica ora se la *policy* $\pi_2$ è quella ottimale, per farlo viene applicato lo stesso processo e se $\pi_3$ risulta uguale a $\pi_2$ allora è la *Policy* ottimale.
La funzione di stato è:

- $V^{\pi_2}(0) = \max\limits_{a\in A}\{0+0+4; 0+0+3\} = \max\limits_{a\in A}\{4; 3\} = 4 \Rightarrow 0 \to 1$
- $V^{\pi_2}(1) = \max\limits_{a\in A}\{0+4; 0+0\} = \max\limits_{a\in A}\{4; 0\} = 4 \Rightarrow 1 \to 3$
- $V^{\pi_2}(2) = \max\limits_{a\in A}\{0+2; 0+3\} = \max\limits_{a\in A}\{0; 3\} = 3 \Rightarrow 2 \to 6$
- Per $3, 4, 5, 6$ non vengono calcolate perché non c'è una possibile scelta, ma valgono rispettivamente $4, 0, 2, 3$

La *Policy* $\pi_3$ quindi sarebbe:

- $0 \to 1$
- $2 \to 6$
- $1 \to 3$
- Da 3, 4, 5, 6 è possibile andare solo allo stato finale

Per cui $\pi_2 = \pi^*$ e il percorso ottimale è $0 \to 1 \to 3 \to$ *home*.