Reinforcement Learning

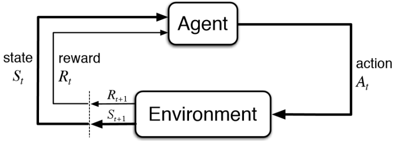
2. Markov Decision Processes

Olivier Sigaud

Sorbonne Université
http://people.isir.upmc.fr/sigaud

Markov Decision Processes

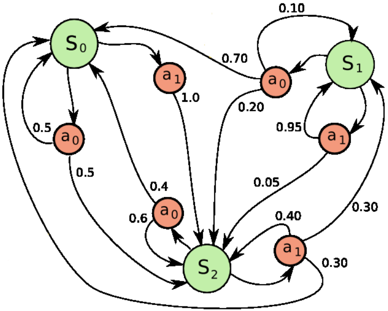
The agent tries an action and once the environment receives the action it gives the agent the next state and it's reward



- S: state space
- A: action space
- $T : S \times A \rightarrow \Pi(S)$: transition function probability
- $r : S \times A \rightarrow \mathbb{R}$: reward function

- An MDP describes a problem, not a solution to that problem

Stochastic transition function

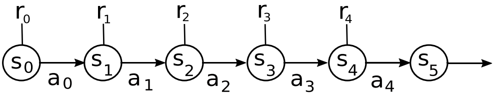


- Deterministic problem = special case of stochastic
- $T(s^t, a^t, s^{t+1}) = p(s'|s,a)$ next state given the curent state and action

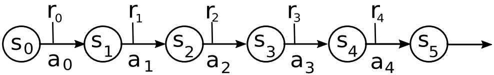
probability

Rewards: over states or action?

- Reward over states

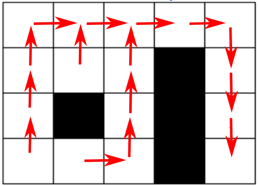


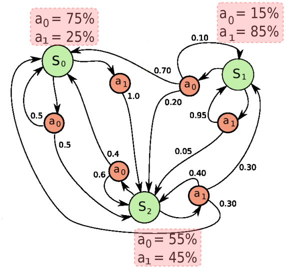
- Reward over actions in states



Deterministic versus stochastic policy

deterministic policy



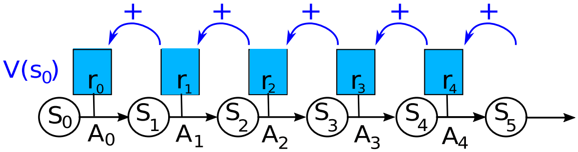


- Goal: find a policy $\pi : S \rightarrow A$ maximizing an agregation of rewards on the long run
  <=> behaviour
- Important theorem: for any MDP, there exists a deterministic policy that is optimal

Agregation criterion: mere sum

Goal of the agent is to maximize the reward

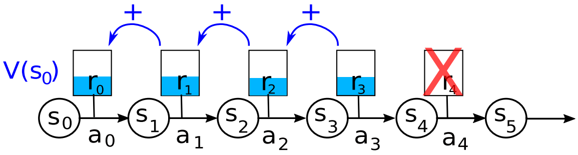
- The computation of value functions assumes the choice of an agregation criterion (discounted, average, etc.)



- The sum over a infinite horizon may be infinite, thus hard to compare
- Mere sum (finite horizon N): $V^{\pi}(S_0) = r_0 + r_1 + r_2 + \ldots + r_N$

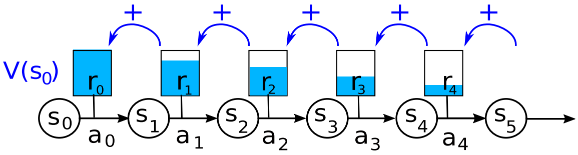
sum of rewards

Agregation criterion: average over a window



- Average criterion on a window: $V^{\pi}(S_0) = \frac{r_0+r_1+r_2}{3}\ldots$

Agregation criterion: discounted



- Discounted criterion: $V^{\pi}(s_{t_0}) = \sum_{t=t_0}^{\infty} \gamma^t r(s_t, \pi(s_t))$

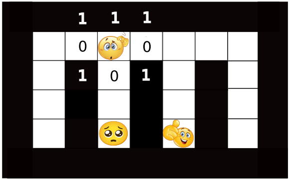
this sum won't diverge to infinity because of gamma between 0 and 1

- $\gamma \in [0, 1]$: discount factor
  - if $\gamma = 0$, sensitive only to immediate reward
  - if $\gamma = 1$, future rewards are as important as immediate rewards
- The discounted case is the most used

Markov Property

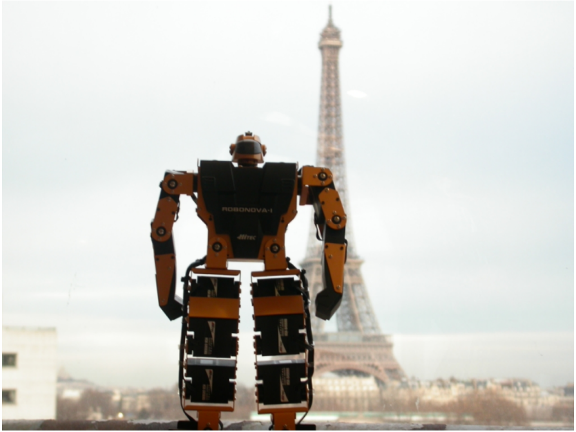
- An MDP defines s^{t+1} and r^{t+1} as $f(s_t, a_t)$
- Markov property : $p(s^{t+1}|s^t, a^t) = p(s^{t+1}|s^t, a^t, s^{t-1}, a^{t-1}, ...s^0, a^0)$
- In an MDP, a memory of the past does not provide any useful advantage
- Reactive agents $a_{t+1} = f(s_t)$, without internal states nor memory, can be optimal

Markov property: Limitations




- Markov property is not verified if:
  - the observation does not contain all useful information to take decisions (POMDPs)
  - or if the next state depends on decisions of several agents (Dec-MDPs, Dec-POMDPs, Markov games)
  - or if transitions depend on time (Non-stationary problems)

Any question?



Send mail to: `Olivier.Sigaud@upmc.fr`