

MDP

2.2.1 Markov Decision Processes A MDP is defined as the collection of the following elements:

- a finite set S of discrete states s of an agent ;
- a finite set A of discrete actions a ;
- a transition function $P : S \times A \rightarrow \Pi(S)$ where $\Pi(S)$ is the set of probability distributions over S . A particular probability distribution $Pr(s_{t+1}|s_t, a_t)$ indicates the probabilities that the agent reaches the different s_{t+1} possible states when he performs action a_t in state s_t ;
- a reward function $R : S \times A \rightarrow \mathbb{R}$ which gives for each (s_t, a_t) pair the scalar reward signal that the agent receives when he performs action a_t in state s_t .

The MDP formalism describes the stochastic structure of a problem faced by an agent, and does not tell anything about the behavior of this agent in its environment. It only tells what, depending on its current state and action, will be its future situation and reward.

MDP

The above definition of the transition function implies a specific assumption about the nature of the state of the agent. This assumption, known as the *Markov property*, stipulates that the probability distribution specifying the s_{t+1} state only depends on s_t and a_t , but not on the past of the agent. Thus $P(s_{t+1}|s_t, a_t) = P(s_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, \dots, s_0, a_0)$. This means that, when the Markov property holds, a knowledge of the past of the agent does not bring any further information on its next state.

The behavior of the agent is described by a policy π giving for each state the probability distribution of the choice of all possible actions.

MDP

In order to define the accumulated reward, we introduce the *discount factor* $\gamma \in [0, 1]$. This factor defines how much the future rewards are taken into account in the computation of the accumulated reward at time t as follows:

$$Rc_{\pi}(t) = \sum_{k=t}^{T_{max}} \gamma^{(k-t)} r_{\pi}(k)$$

where T_{max} can be finite or infinite and $r_{\pi}(k)$ represents the immediate reward received at time k if the agent follows policy π .

DP methods introduce a *value function* V^{π} where $V^{\pi}(s)$ represents for each state s the accumulated reward that the agent can expect if it follows policy π from state s . If the Markov property holds, V^{π} is solution of the Bellman equation (Bertsekas, 1995):

$$\forall s \in S, V^{\pi}(s) = \sum_a \pi(s_t, a_t) [R(s_t, a_t) + \gamma \sum_{s_{t+1}} P(s_{t+1} | s_t, a_t) V^{\pi}(s_{t+1})] \quad (1)$$

MDP

Rather than the value function V^π , it is often useful to introduce an action-value function Q^π where $Q^\pi(s, a)$ represents the accumulated reward that the agent can expect if it follows policy π after having done action a in state s . Everything that was said of V^π directly applies to Q^π , given that $V^\pi(s) = \max_a Q^\pi(s, a)$. The corresponding optimal functions are independent of the policy of the agent; they are denoted V^* and Q^* .

MDP

Let us consider the case of the average immediate reward. Its exact value after k iterations is

$$E_k(s) = (r_1 + r_2 + \cdots + r_k)/k$$

Furthermore,

$$E_{k+1}(s) = (r_1 + r_2 + \cdots + r_k + r_{k+1})/(k + 1)$$

thus

$$E_{k+1}(s) = k/(k + 1)E_k(s) + r_{k+1}/(k + 1)$$

which can be rewritten:

$$E_{k+1}(s) = (k + 1)/(k + 1)E_k(s) - E_k(s)/(k + 1) + r_{k+1}/(k + 1)$$

or

$$E_{k+1}(s) = E_k(s) + 1/(k + 1)[r_{k+1} - E_k(s)]$$

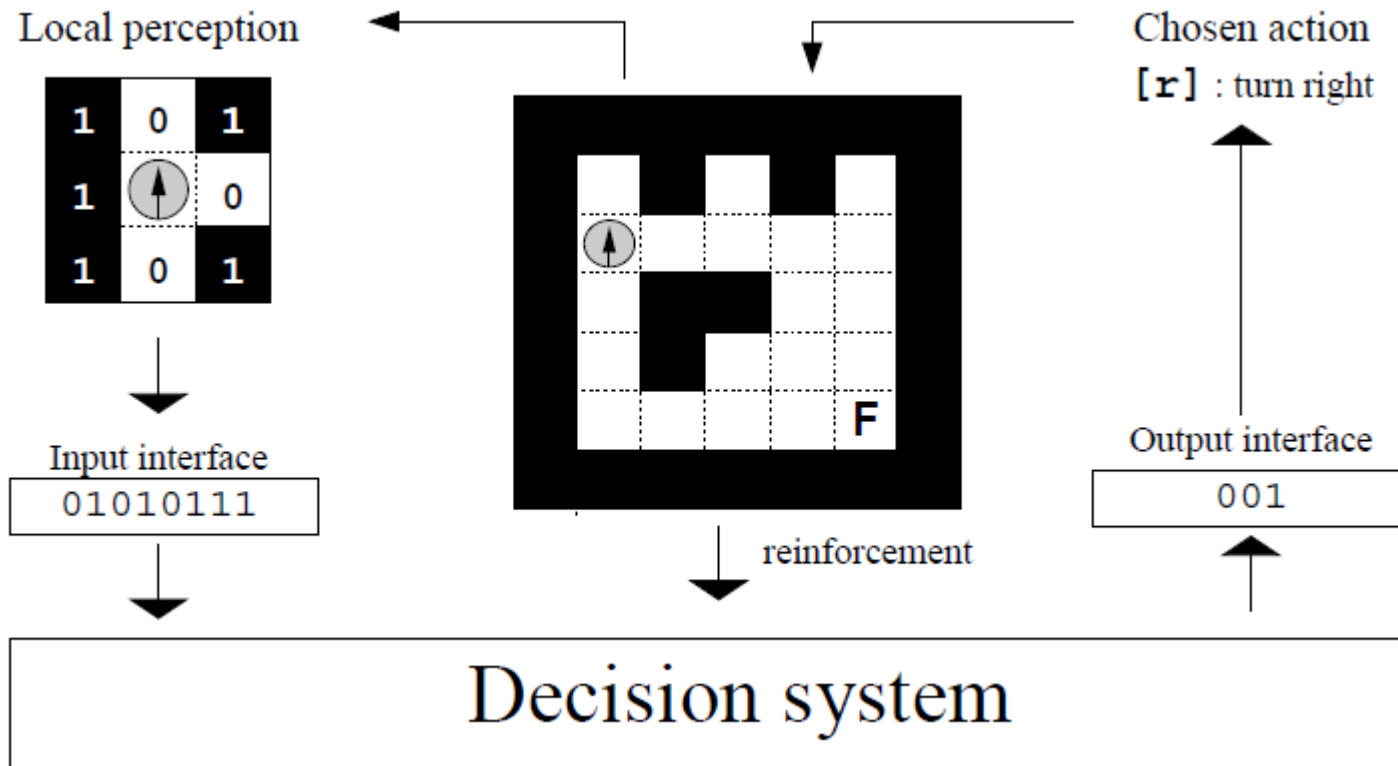
MDP

$$E_{k+1}(s) = E_k(s) + \alpha[r_{k+1} - E_k(s)]$$

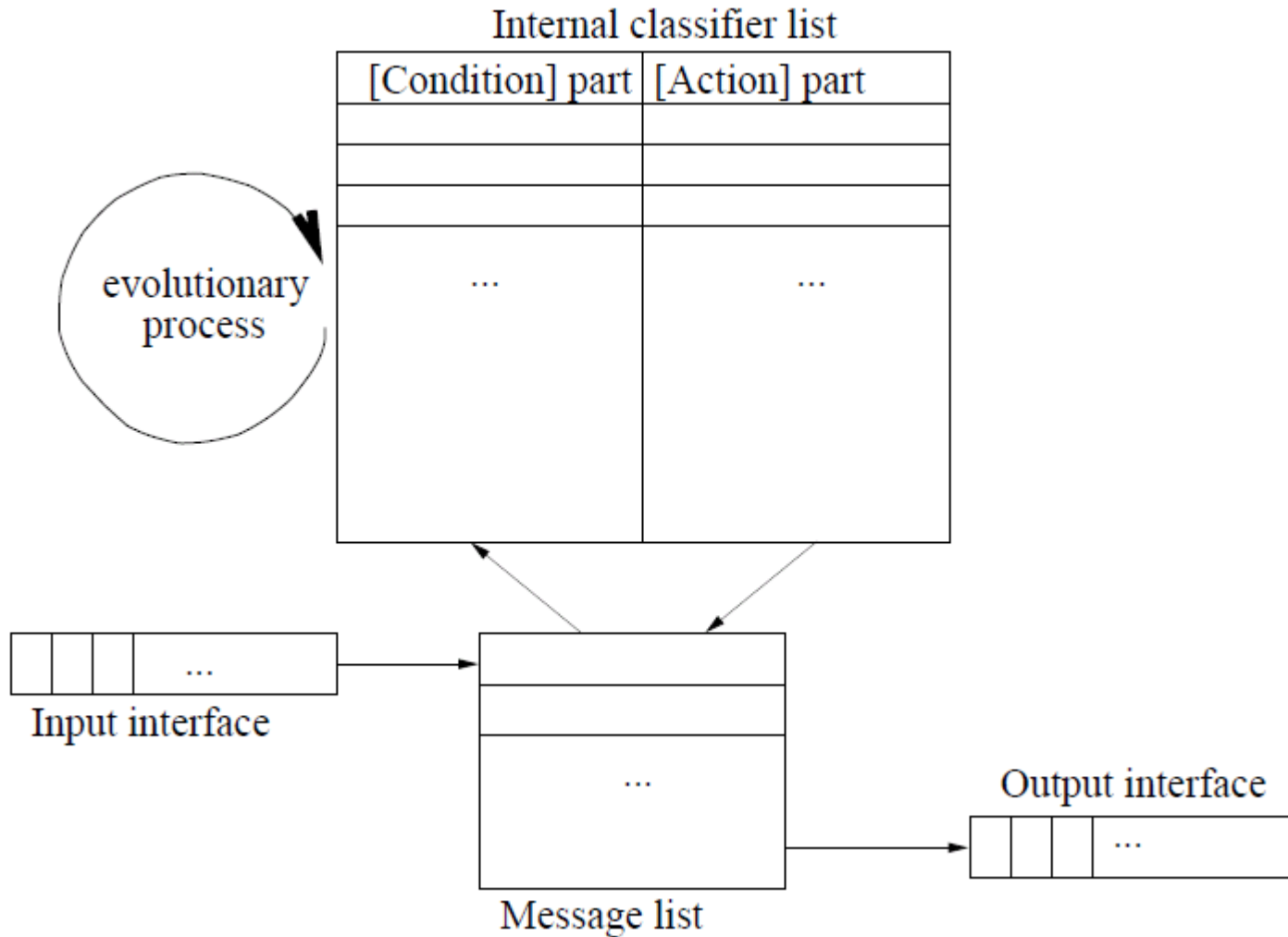
The parameter α , called *learning rate*, must be tuned adequately because it influences the speed of convergence towards the exact average.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$$

RL + GA = LCS



LCS



LCS

