Markov Decision Processes

We define a controlled system as a labeled Markov decision process.

[Labeled Markov Decision Process] A (labeled) Markov decision process (MDP) is a tuple $M = (S, A, \mathcal{A}, P, s_{init}, AP, L)$, where S is a finite set of states, A is a finite set of actions, $\mathcal{A} : S \to 2^A$ is a mapping that maps each state ability such that $\sum_{s' \in S} P(s'|s,a) = 1$ for any state $s \in S$ and any action $a \in \mathcal{A}(s)$, $s_{init} \in S$ is the initial state, AP is a finite set of atomic propositions, and $L : S \times A \times S \to 2^{AP}$ is a labeling function that assigns a set of atomic propositions to each transition $(s, a, s') \in S \times A \times S$.

In the MDP M, an infinite path

action-value function $Q^{\pi}(s,a)$ under the policy π as follows.

$$Q^{\pi}(s,a) = E^{\pi}[\sum_{n=0}^{\infty}\gamma^n \mathcal{R}(S_n, A_n, S_{n+1})|S_0 = s, A_0 = a].$$

We have the following recursively equation for the state-value function and the action-value function.

$$\begin{aligned}
V^{\pi}(s) &= E^{\pi}[\sum_{n=0}^{\infty}\gamma^n \mathcal{R}(S_n, A_n, S_{n+1})|S_0 = s] \\
&= \sum_{a\in\mathcal{A}(s)}\pi(s,a)\sum_{s'\in S}P(s'|s,a)E^{\pi}[\sum_{n=0}^{\infty}\gamma^n\mathcal{R}(S_n, A_n, S_{n+1})|S_0 = s, A_0 = a, S_1 = s'] \\
&= \sum_{a\in\mathcal{A}(s)}\pi(s,a)\sum_{s'\in S}P(s'|s,a)\{\mathcal{R}(s,a,s') + \gamma E^{\pi}[\sum_{n=0}^{\infty}\gamma^n\mathcal{R}(S_n, A_n, S_{n+1})|S_1 = s']\} \\
&= \sum_{a\in\mathcal{A}(s)}\pi(s,a)\sum_{s'\in S}P(s'|s,a)\{\mathcal{R}(s,a,s') + \gamma V^{\pi}(s')\},
\end{aligned}$$

by the definition of the action-value function, it holds that

$$\begin{aligned}
Q^{\pi}(s,a) &= \max_{a\in\mathcal{A}(s)} V^{\pi}(s) \\
&= \sum_{s'\in S}P(s'|s,a)\{\mathcal{R}(s,a,s') +
\end{aligned}$$

overall procedure TD-learning for a state-value function is given by Algoritym .

TD-learning methods for an action-value function are classified as two main learning methods that are referred to Q-learning and SARSA.

Stochastic Discrete Event Systems

We represent a stochastic discrete event system (DES) as an MDP.

[Stochastic discrete

ative frequency of occurrence of each event does not depend on the control pattern.

We define a reward function $\mathcal{R} : S \times 2^E \times E \times S \to R$ and the reward $\mathcal{R}$ can be decomposed into $\mathcal{R}_1$ and $\mathcal{R}_2$. The first reward $\mathcal{R}_1 : S \times 2^E \to R$ is determined by the control pattern selected by the supervisor, which