Markov Decision Process

The Markov Property

“The future is independent of the past given the present”

Markov Chain

A Markov process is a memoryless random process, i.e. a sequence of random states S1, S2, ... with the Markov property.

Definition

A *Markov Process* (or *Markov Chain*) is a tuple $\langle \mathcal{S}, \mathcal{P} \rangle$

- $\mathcal{S}$ is a (finite) set of states
- $\mathcal{P}$ is a state transition probability matrix, $\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s]$

Example: Birth Death processes (queues)

- Higher order Markov Chain

Markov Reward Process

A Markov reward process is a Markov chain with values.

Definition

A *Markov Reward Process* is a tuple $\langle \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma \rangle$

- $\mathcal{S}$ is a finite set of states
- $\mathcal{P}$ is a state transition probability matrix, $\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s]$
- $\mathcal{R}$ is a reward function, $\mathcal{R}_s = \mathbb{E}[R_{t+1} \mid S_t = s]$
- γ is a discount factor, $\gamma \in [0, 1]$

Return

Definition

The *return* G_t is the total discounted reward from time-step t.

$$G_t - R_{t+1} + \gamma R_{t+2} + \ldots - \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Discount

Value Function

Definition

The *state value function* $v(s)$ of an MRP is the expected return starting from state s

$$v(s) = \mathbb{E}[G_t \mid S_t = s]$$

Markov Decision Process

- Discrete time stochastic control process.
- A Markov decision process (MDP) is a Markov reward process with decisions. It is an environment in which all states are Markov.
-

Definition

A *Markov Decision Process* is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$

- $\mathcal{S}$ is a finite set of states
- $\mathcal{A}$ is a finite set of actions
- $\mathcal{P}$ is a state transition probability matrix, $\mathcal{P}^{a}_{ss'} = \mathbb{P}\left[S_{t+1} = s' \mid S_t = s, A_t = a\right]$
- $\mathcal{R}$ is a reward function, $\mathcal{R}^{a}_{s} = \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right]$
- γ is a discount factor $\gamma \in [0, 1]$.

- Fully Observable MDP
- Partially Observable MDP

Reinforcement Learning

- The reinforcement learning problem is meant to be a straightforward framing of the problem of learning from interaction to achieve a goal.
- The learner and decision- maker is called the agent.
- The thing it interacts with, comprising everything outside the agent, is called the environment. These

All reinforcement learning agents have explicit goals, can sense aspects of their environments, and can choose actions to influence their environments

Assume that the environment is modelled as a Fully Observable MDP.

Two Types of Value Functions

Bellman Equations

Bellman Optimality Equations

Grid problem