UniTs - University of Trieste

Faculty of Scientific and Data Intensive Computing

Department of mathematics informatics and geosciences

Reinforcement Learning

Lecturer:
Prof. Antonio Celani

Authors:
Christian Faccio

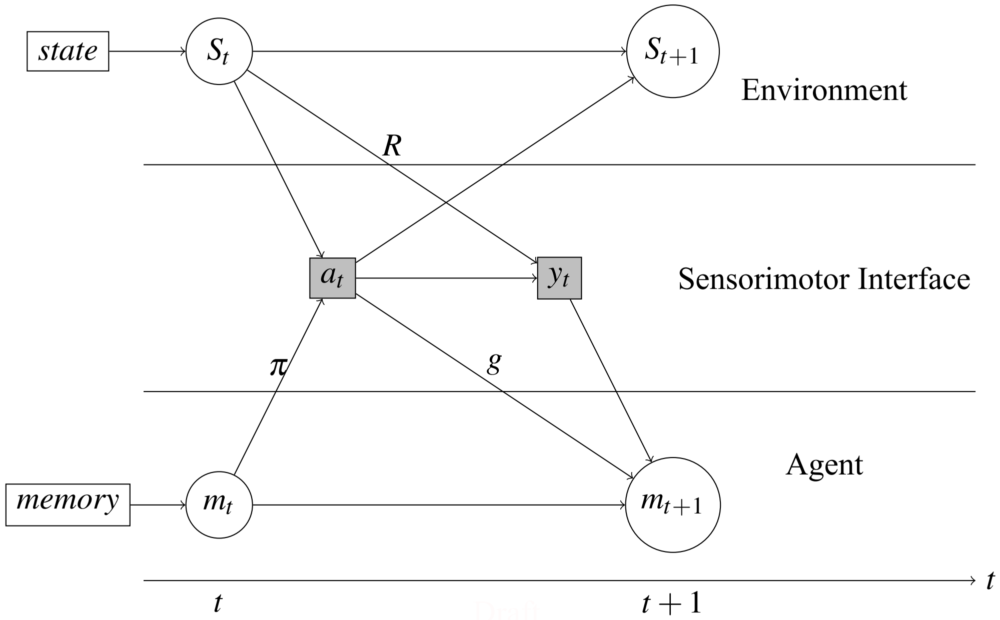
March 24, 2025

github.com/christianfaccio christianfaccio@outlook.it

Contents

1 Introduction

General scheme of a **Decision Process**:



- $\pi(a|m) \to$ policy
- $R(y) \to$ reward function
- $p(s'y|sa) \to$ model of the environment
- $g(m'|may) \to$ memory update

The goal is to find the optimal policy π^* that maximizes the expected return:

$$maximize_\pi \underbrace{\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(y_t)\right]}_{ExpectedReturn} \qquad 0 \leq \gamma < 1$$

with γ survival probability.
The expected survival time is:

$$\frac{1}{1-\gamma}$$

Specifications:

- **Perfect observability** $\to$ the agent knows the state of the environment ($y = S$) and $p(y|sas') = \infty(y = s')$

Observation:

$$p(s'y|sa) = p(s'|sa)p(y|sas')$$

- **Memory update** $\rightarrow$ the agent knows the state of the environment and the memory ($M = y$) and $g(m'|may) = \infty(m' = y)$

1.1 Markov Decision Process

Definition: ***Markov Decision Process***

A Markov Decision Process (MDP) is a fully observable set of tuples (S, A, R, P, γ) where:
- $s \in S$ is a finite set of states
- $a \in A$ is a finite set of actions
- $R : S \times A \rightarrow \mathbb{R}$ is the reward function
- $P : S \times A \times S \rightarrow [0, 1]$ is the transition probability function
- $\gamma \in [0, 1]$ is the discount factor
- $p(s'y|sa)$ is the model of the environment
- $p_0(s)$ is the initial state distribution
- $\pi(a|s)$ is the policy

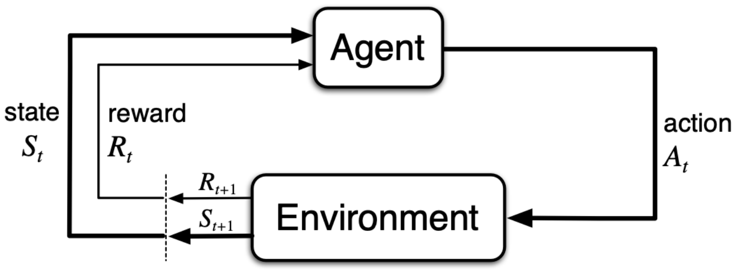


Figure 1.1: Markov Decision Process

$$\begin{aligned} G_\pi(\rho_0) &= \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t R(y_t)\right) \\ &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}\left[R(y_t)\right] \\ &= \sum_{t=0}^{\infty} \gamma^t \sum_{sa} \rho_t(s)\pi(a|S)p(s'y|sa)r(y) \\ &= \sum_{t=0}^{\infty} \gamma^t \sum_{s\in s'} \rho_t(s)\pi(a|s)p(s'|sa)r(sas') \end{aligned}$$

The difficulty here is that the dependence on π is non linear, but linear on the initial condition. Let's introduce now the **Chapman Kolmogorov equation**:

$$\rho_{t+1}(s') = \sum_{sa} \rho_t(s)\pi(a|s)p(s'|sa)$$

it basically tells us that the probability of being in state s' at time $t+1$ is the sum of the probabilities of being in state s at time t and then moving to state s' by taking action a.

$$\begin{aligned} G_\pi(\rho_0) &= \sum_s \rho_0(s) \quad \underbrace{V_\pi(s)}_{\text{value of the policy } \pi} \qquad \rho_0 = e_s \text{ and } G_\pi(e_s) = V_\pi(s) \\ &= \sum_{sas'} \rho_0(s)\pi(a|s)p(s'|sa)r(sas') + \gamma \underbrace{\sum_{t=1}^{\infty} \gamma^{t-1} \sum_{sas'} \rho_{t+1}(s)\pi(a|s)p(s'|sa)r(sas')}_{G_\pi(\rho_1)} \end{aligned}$$

The recursion equation is:

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{s'} p(s'|sa)[r(sas') + \gamma V_\pi(s')]$$

one can also prove that it has a unique solution. It is also the basis for evaluating the policy π. The problem is that we want to find the optimal policy π^* that maximizes the expected return seen before.

$$\pi^* = argmax_\pi G_\pi(\rho_0)$$

For this purpose we introduce the **Bellman equation**:

$$V^*(s) = \max_a \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^*(s') \right]$$

$$\bar{\pi} = \mathbb{1}(a = argmax_a \left[\sum_{s'} p(s'sa)(r(sas') + \gamma V^*(s')) \right])$$

1. $V^*(s) = V_{\bar{\pi}}(s)$
   Recursion equation:
   $$\begin{aligned} V_{\bar{\pi}}(s) &= \textstyle\sum_{as'} \bar{\pi}(a|s) p(s'|sa)[r(sas') + \gamma V_{\bar{\pi}}(s')] \\ V^*(s) &= \textstyle\sum_{as'} \bar{\pi}(a|s') p(s'|sa)[r(sas') + \gamma V^*(s')] \end{aligned}$$
   That leads to:
   $$(V_{\bar{\pi}}(s) - V^*(s)) = \sum_a \bar{\pi}(a|s) \sum_{s'} p(s'|sa)[r(sas') + \gamma V^*(s')] - \max_a \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^*(s') \right]$$
2. $G_{\bar{\pi}}(\rho_0) \geq G_\pi(\rho_0)$
   $$\begin{aligned} G_{\bar{\pi}}(\rho_0) &= \textstyle\sum_s \rho_0(s) V_{\bar{\pi}}(s) \\ &= \textstyle\sum_s \rho_0(s) V^*(s) \\ &= \textstyle\sum_s \rho_0(s) max_a \left[\sum_{s'} p(s'|sa)[r(sas') + \gamma V^*(s')]\right] \\ &\geq \textstyle\sum_{sa} \rho_0(s) \pi(a|s) \sum_{s'} p(s'|sa)[r(sas') + \gamma V^*(s')] \\ &= sum_{sas'} \rho_0(s) \pi(a|s) p(s'|sa) r(sas') + \underbrace{\gamma \sum_{sas'} \rho_0(s) \pi(a|s) p(s'|sa) V^*(s')}_{G_\pi(\rho_1)} \end{aligned}$$

Let's intrduce the **Bellman Operator**:

$$BW(s) = \max_a \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V(s') \right]$$

The Bellman Operator is a contraction mapping:

$$\begin{aligned} ||BW(s) - BW(s')|| &= || \max_a [r(s,a) + \gamma \textstyle\sum_{s'} p(s'|s,a) V(s')] - \max_a [r(s',a) + \gamma \sum_{s'} p(s'|s',a) V(s')] || \\ &\leq ||r(s,a) - r(s',a)|| + \gamma || \textstyle\sum_{s'} p(s'|s,a) V(s') - \sum_{s'} p(s'|s',a) V(s')|| \\ &\leq ||r(s,a) - r(s',a)|| + \gamma ||V(s) - V(s')|| \end{aligned}$$

γ controls the contraction.