Reinforcement Learning

Marshal Sinaga - marshal.sinaga@aalto.fi

Last update: September 3, 2024

This note aims to cover some materials on the reinforcement learning. The primary references are Reinforcement Learning: An Introduction (2nd edition) by Sutton & Barto and ELEC-E8125 by Joni Pajarinen.

1 Overview

- Reinforcement learning (RL) problem:
  - Denote that $\pi : O \rightarrow A$ is a policy that maps the observation to an action.
  - Determine a policy:

$$a = \pi(s) \tag{1}$$

  - s.t. the expected cumulative return is maximum, i.e.,

$$\pi^* = \arg\max_\pi \mathbb{E}[G] \tag{2}$$

$$G = \sum_t r_t \tag{3}$$

- Markov decision process (MDP):
  - We have an environment observable $z = s$, defined by a Markov dynamics defined as:

$$p(s_{t+1}|s_t, a_t) \tag{4}$$

    and a reward function

$$r_t = r(s_t, a_t) \tag{5}$$

  - The solution is formulated as follows:

$$a^*_{1,\ldots,T} = \arg\max_{a_1,\ldots,a_T} \sum_{t=1}^{T} r_t \tag{6}$$

    Represented as policy:

$$a = \pi(s) \tag{7}$$

- Connection between RL and MDP: RL is a MDP with unknown Markov dynamics $p(s_{t+1}|s_t, a_t)$, and unknown reward function r_t.
- Partially observable MDP (POMDP):
  - The environment is not directly observable.
  - Following MDP, POMDP is governed by a Markov dynamics $p(s_{t+1}|s_t, a_t)$ and reward function $r_t = r(s_t, a_t)$. In addition, we have an observation model $p(z_{t+1}|s_{t+1}, a_t)$.