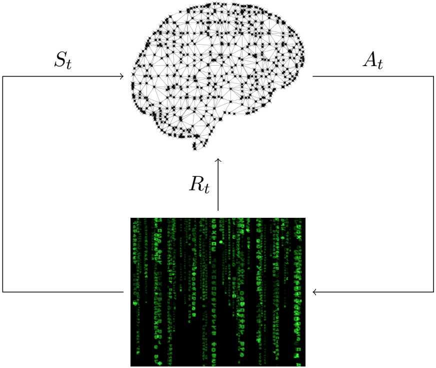
Partially observable Markov Decision Processes

Corrado Possieri

Machine and Reinforcement Learning in Control Applications

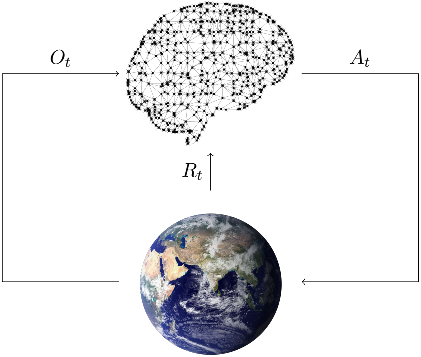
Introduction

- Up to now, we assumed to measure the state.



Partially observable environments

- In several environments, we have just observations.



Environmental interaction

- The interactions with the environment would be then

$$A_0, O_1, A_1, O_2, A_3, O_3, \ldots .$$

- We can introduce the notion of **history** up to time t

$$H_t = A_0, O_1, A_1, O_2, A_3, O_3, \ldots, A_{t-1}, O_t.$$

- The history is all we know about the past.

Notion of state

- The state should be a summary of the history

$$S_t = f(H_t).$$

- If the state retains all information about the history
  - S_t can be used to predict futures as accurately as from H_t;
  - S_t and f have the *Markov property*.

- Real agents may be non-Markov but may approach it as an ideal.

Test

- A *test* is a sequence of alternating actions and observations
  - *e.g.*, a 3-step test is $\tau = a_1 o_1 a_2 o_2 a_3 o_3$.
  - the probability of τ given an history h is

$$p(\tau|h) = \mathbb{P}[O_{t+1} = o_1, O_{t+2} = o_2, O_{t+3} = o_3 \\ |A_t = a_1, A_{t+1} = a_2, A_{t+2} = a_3].$$

- f is Markov if

$$f(h) = f(h') \implies p(\tau|h) = p(\tau|h'), \quad \forall \tau, \forall h, \forall h'.$$

- This implies that a Markov state summarizes all that is necessary to make predictions.

Compact representation

- The state should be small compared to the history.
- Actually we do not want to consider the whole history.
- We may think about a recursive update

$$S_{t+1} = u(S_t, A_t, O_{t+1}).$$

- Given f it is always possible to find u.

Strategy for finding a Markov state

- Actually, we want to make one-step predictions.
- If f is incrementally updatable, then

$$
\begin{aligned}
f(h) = f(h') \implies \mathbb{P}[O_{t+1} = o | H_t = h, A_t = a] \\
= \mathbb{P}[O_{t+1} = o | H_t = h', A_t = a].
\end{aligned}
$$

- If there is any error in the one-step predictions, then it can lead to inaccurate long-term predictions.

Partially observable MDP

- The environment is assumed to have a latent state

$$X_t \in \{1, 2, \ldots, d\}.$$

- X_t produces observations but is not available.
- The natural Markov state $\mathbf{s}_t \in [0, 1]^d$ is a *belief* about X_t

$$\mathbf{s}_t[i] = \mathbb{P}[X_t = i | H_t].$$

- Assuming complete knowledge of the environment
  - can be updated using Bayes' rule

$$u(\mathbf{s}, a, o)[i] = \frac{\sum_{x=1}^{d} \mathbf{s}[x] p(i, o|x, a)}{\sum_{x'=1}^{d} \sum_{x=1}^{d} \mathbf{s}[x] p(x', o|x, a)},$$

where

$$p(x', o|x, a) = \mathbb{P}[X_t = x', O_t = o | X_{t-1} = x, A_{t-1} = a].$$

Belief MDP

- The state s satisfies the Markov property.
- The resulting belief MDP is defined on a continuous space.
- the set $\mathcal{A}$ is the same as in the original POMDP.
- We can apply classical algorithms.

Approximation

- We must work with an approximate notion of state.
- The state S_t may not be Markov.
- Two possible selections of states are

$$
\begin{aligned}
S_t &= O_t, \\
S_t &= O_t, A_{t-1}, O_{t-1}, \dots, A_{t-k}.
\end{aligned}
$$

- This kth-order history approach is still very simple, but can greatly increase the agent's capabilities.