UCLRL Lecture 1 Notes

1 The Reinforcement Learning Problem

A reward R_t at time t is a scalar indicating how well an agent is doing. The goal of the agent is to maximize cummulative reward $\sum_t R_t$.

1.1 Reward Hypothesis

All goals can be described by the maximization of expected cummulative reward.

1.2 Sequential Decision Making

At each time step t,
Action: Execute A_t
Observation: Receive O_t
Reward Receive R_t
Goal: Select actions to maximize total future reward.

The **history** is the sequence of observations, actions, rewards

$$H_t = A_1, O_1, R_1, \ldots, A_t, O_t, R_t$$

The **state** is a function of history that determines what happens next.

$$S_t = f(H_t)$$

The **environment state** S_t^e is the private representation of the environment and produces the next observation after an action. It is not usually visible to the agent or if visible is full of irrelevant information. The **agent state** S_t^a is the agent's internal representation and is used to pick the next action and is the information used by the learning algorithms. Thus,

$$S_t^a = f^a(H_t)$$

In **full observability**, agent directly sees the environment state,

$$O_t = S_t^a = S_t^e$$

and in **partial observability**, $S_t^a \neq S_t^e$.

An **information state** contains all useful informmation from history.

Definition 1. *A state S_t is Markov if and only if*

$$\mathbb{P}\left[S_{t+1}|S_t\right] = \mathbb{P}\left[S_{t+1}|S_1, \ldots, S_t\right]$$

1.2.1 Constructing State Representation

- Complete history: $S_t^a = H_t$
- Beliefs: $S_t = (\mathbb{P}[S_t^e = s^1], \ldots, \mathbb{P}[S_t^e = s^n])$
- Recurrent Neural Network: $S_t^a = \sigma(S_{t-1}^a W_s + O_t W_o)$

2 RL Agent

Agent contains one or more of the components

- Policy : Determine what action to take next, or the agent's behavior. It is a map from state to action. **Determistic Policy** $a = \pi(s)$ and **Stochastic Policy** $\pi(a|s) = \mathbb{P}[A = a|S = s]$
- Value function : Prediction of future reward to evaluate the goodness/badness of state.
$$V_\pi(s) = \mathbb{E}_\pi \left[R_t + \gamma R_{t+1} + \gamma^2 R_{t+1} + \ldots | S_t = s\right]$$
- Model : Agent's model of the environment

2.1 Model

Predicts what the environment will do next. $\mathcal{P}$ predicts the next state (dynamics)

$$\mathcal{P}_{ss'}^a = \mathbb{P}\left[S' = s' | S = s, A = a\right]$$

and $\mathcal{R}$ predicts the next immediate reward

$$\mathcal{R}_s^a = \mathbb{E}\left[R | S = s, A = a\right]$$

2.2 Categorizing RL agents

- Value based - value function and no policy (implicit in the value function)
- Policy based - Only policy and no value function
- Actor Critic - Has a policy but also stores the value function
- Model Free - Do not make a model of the environment. Just create the policy or value function.