CS292F Statistical Foundation of Reinforcement Learning **Spring 2021**

Lecture 1: March 29

Lecturer: Yu-Xiang Wang *Scribe: Monsij Biswal*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1.1 Reinforcement learning: a general AI framework

Reinforcement learning is a technique wherein an agent learns to perform actions in a certain sequence such that it maximises a given reward over time. This generally occurs in an uncertain environment in which the agent can only interact with its surroundings to learn more about it. Though the applications of RL were limited in the past due to computational bottlenecks, RL is now being used in a variety of situations such as robotics, conversation systems, medical treatment and much more.

1.2 Problem setup

The key components which formulate any reinforcement learning problem are state space ($\mathcal{S}$), action space ($\mathcal{A}$), reward function (r), transition kernel (P), discount factor (γ) and initial state distribution (μ). The observation space ($\mathcal{O}$) is only relevant when all the states are not observable. Finally the interactions between the agent and the environment can be captured by a Markov Decision Process (MDP). Formally:

$$S_t \in \mathcal{S} \quad A_t \in \mathcal{A} \quad R_t \in \mathbb{R} \quad O_t \in \mathcal{O}$$

The policy π is defined in two ways:

- When the state is observable: $\pi : \mathcal{S} \to \mathcal{A}$
- When the state is not observable: $\pi_t : (\mathcal{O} \times \mathcal{A} \times \mathbb{R})^{t-1} \times \mathcal{O} \to \mathcal{A}$

The reward function r is defined over: $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as :

$$r(s,a) := \mathbb{E}[R_t | S_t = s, A_t = a], \text{where } r \in [0, R_{max}]$$

The goal of the problem is to find the optimal policy π^* that maximises the total expected reward over time. The horizon may be finite or infinite.

- Finite horizon (episodic) RL: $\pi^* = \underset{\pi \in \Pi}{\operatorname{argmax}} \mathbb{E}\big[\sum_{t=1}^{T} R_t\big]$
- Infinite horizon RL: $\pi^* = \underset{\pi \in \Pi}{\operatorname{argmax}} \mathbb{E}\big[\sum_{t=1}^{\infty} \gamma^{t-1} R_t\big]$

1.2.1 Policy, value function and goal of MDP

Policy. A policy determines the decision-making strategy for an agent using which it determines which actions to perform at a given state based on the history of observations (trajectories).
More formally, a trajectory $\mathcal{H}$ upto time t is given by the set:

$$\mathcal{H} = \bigcup_{t=1}^{\infty} (s_1, a_{1,1}), ...(s_{t-1}, a_{t-1}, r_{t-1}), s_t$$

Then the policy π_t is defined as: $\pi : \mathcal{H} \rightarrow \mathcal{A}$. This is any general policy. Special cases include stationary and stationary deterministic policies which are defined as below:

- Stationary: $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$
- Stationary, deterministic: $\pi : \mathcal{S} \rightarrow \mathcal{A}$

State value function. Long term expected reward when the initial state is s and the trajectory rolls out according to policy π.

$$V^{\pi}(s) = \mathbb{E}_{\pi}[R_1 + \gamma R_2 + ...\gamma^{t-1} R_t + ... |S_1 = s]$$

State-Action value function. Long term expected reward when the initial state is s, initial action is a and thereafter follows the policy π.

$$Q^{\pi}(s, a) = \mathbb{E}_{\pi}[R_1 + \gamma R_2 + ...\gamma^{t-1} R_t + ... |S_1 = s, A_1 = a]$$

Goal. Find π^* such that $V^{\pi^*}(s) = \max_{\pi} V^{\pi}(s)\ \forall s \in \mathcal{S}$ simultaneously.

In other words, we are looking for the ϵ-optimal policy $\tilde{\pi}$ that satisifes the inequality: $V^{\tilde{\pi}}(.) \geq V^*(.) - \epsilon \overrightarrow{1}$

Now we state and prove some interesting observations about MDPs

Proposition 1.1. *It suffices to consider stationary policies.*

Proof. It suffices to show that:
For any π, $\exists\ \pi^{'}$ such that $\pi^{'}$ is stationary and both π and $\pi^{'}$ have the same value : $V^{\pi}(\mu) = V^{\pi^{'}}(\mu)$

By definition:

$$\begin{aligned}
V^{\pi}(\mu) &= \mathbb{E}_s^{\pi}\big[\sum_{t=1}^{\infty} \gamma^{t-1} r(S_t, A_t)\big] \\
&= \sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{E}_{\mu}^{\pi}\big[r(S_t, A_t)\big] \\
&= \sum_{t=1}^{\infty} \gamma^{t-1} \sum_{s,a} P_{\mu}^{\pi}(S_t = s, A_t = a)\, r(s, a) \\
&= \sum_{s,a} \big[\sum_{t=1}^{\infty} \gamma^{t-1} P_{\mu}^{\pi}(S_t = s, A_t = a)\big]\, r(s, a) \\
&= \sum_{s,a} \nu_{\mu}^{\pi}(s, a)\, r(s, a)
\end{aligned}$$

Here, $\nu^{\pi}_{\mu}(s,a)$ is termed as the *occupancy measure* and the value function can alternatively be written in terms of *occupancy measure* as:

$$V^{\pi}_{\mu} = \langle\, \nu^{\pi}_{\mu}, r \,\rangle$$

From above, we conclude that it suffices to prove that the *occupancy measures* with both policies are equivalent: $\nu^{\pi}_{\mu} = \nu^{\pi^{'}}_{\mu}$.

Consider $\pi^{'}$ for any fixed π. Define:

$$\pi^{'}(a|s) = \left\{ \begin{array}{ll} \frac{\nu^{\pi}_{\mu}(s,a)}{\nu^{\pi}_{\mu}(s)} & \text{if } \nu^{\pi}_{\mu}(s) > 0 \\ \pi_0(a) & otherwise \end{array} \right\} \tag{1.1.1}$$

where $\nu^{\pi}_{\mu}(s) = \sum\limits_{a} \nu^{\pi}_{\mu}(s,a)$

It is clear that $\pi^{'}$ is stationary, $\pi^{'} : \mathcal{S} \to \Delta(\mathcal{A})$

By law of total probability and (1.1.1):

$$P^{\pi^{'}}(s^{'}|s) = \sum_{a} \pi^{'}(a|s) P(s^{'}|s,a) = \sum_{a} \frac{\nu^{\pi}_{\mu}(s,a)}{\nu^{\pi}_{\mu}(s)} P(s^{'}|s,a) \tag{1.1.2}$$

From the definition of occupancy measure, we had:

$$\begin{aligned}
\nu^{\pi}_{\mu}(s) &= \sum_{t=1}^{\infty} \gamma^{t-1} P^{\pi}_{\mu}(S_t = s) \\
&= \mu(s) + \gamma \sum_{t=2}^{\infty} \gamma^{t-2} P^{\pi}_{\mu}(S_t = s) \\
&= \mu(s) + \gamma \sum_{\widetilde{t}=1}^{\infty} \gamma^{\widetilde{t}-1} P^{\pi}_{\mu}(S_{\widetilde{t}+1} = s) && [\,\widetilde{t} = t-1\,] \\
&= \mu(s) + \gamma \sum_{t=1}^{\infty} \gamma^{t-1} \sum_{s^{'},a} P^{\pi}_{\mu}(S_t = s^{'}, A_t = a) P(s|s^{'},a) \\
&= \mu(s) + \gamma \sum_{s^{'},a} \big[\sum_{t=1}^{\infty} \gamma^{t-1} P^{\pi}_{\mu}(S_t = s^{'}, A_t = a) \big] P(s|s^{'},a) \\
&= \mu(s) + \gamma \sum_{s^{'},a} \nu^{\pi}_{\mu}(s^{'},a) P(s|s^{'},a)
\end{aligned}$$

By multiplying and dividing $\nu^{\pi}_{\mu}(s^{'})$ and using (1.1.2), we get:

$$\nu^{\pi}_{\mu}(s) = \mu(s) + \gamma \sum_{s^{'}} \nu^{\pi}_{\mu}(s^{'}) P^{\pi^{'}}(s|s^{'}) \implies \boldsymbol{\nu^{\pi}_{\mu} = (I - \gamma P^{\pi^{'}})^{-1} \mu}$$

Similarly,

$$\nu^{\pi^{'}}_{\mu}(s) = \mu(s) + \gamma \sum_{s^{'}} \nu^{\pi^{'}}_{\mu}(s^{'}) P^{\pi^{'}}(s|s^{'}) \implies \boldsymbol{\nu^{\pi^{'}}_{\mu} = (I - \gamma P^{\pi^{'}})^{-1} \mu}$$

The occupancy measures and hence the value function corresponding to any general policy and the constructed stationary policy are shown to be equal.

□

Exercise: Show that for all $\gamma < 1$, $I - \gamma P^{\pi^{'}}$ is invertible.

Corollary. There exists an optimal policy that is stationary.

Since we can find a policy π^s for each s by choosing such a policy that maximises the value function $v^\pi(s)$, we can subsequently form a stationary optimal policy.

Proposition 1.2. *There exists a stationary and deterministic policy that is optimal simultaneously for all initial state distributions.*

This will be proved in the next lecture.