*Wednesday, October 7, 2020*

CS 59000_RL

Markov Decision Processes (MDPs)

- Agenda:
  - settings.
  - Policy classes
  - Value
  - Dynamic programming

---

Examples:
- Game of Casino
- Game of racing
- Continuous time Control system (plane)

---

MDPs can be continuous time, like the plane example, They can be discrete time, like the casino game. Or, time domain, where the process is defined can be general normed spaces (it need no be time any way)

---

For simplicity, we study discrete time MDPs in this class.

---

For the game of racing, we considered four possible states, and three possible actions. We could consider the case that states and actions are continuous, or some what general measurable spaces (e.g. Polish space)

In the casino night setting, both state and action spaces are finite spaces and categorical.

*Wednesday, October 7, 2020*

In this class we study tabular MDPs where both state and action spaces are finite sets. (Later we also talk about general ones)

---

MDP in a generic case is a tuple

$$\mathcal{M} = (\underbrace{\mathcal{X}}_{\text{state space}}, \underbrace{\mathcal{A}}_{\text{action space}}, \underbrace{P}_{\text{Transition Kernel}}, \underbrace{R}_{\text{Reward Kerel}}, \overbrace{P_1}^{\text{Initial state distribution}})$$

Protocol: At the first time step, $X_1$ is drawn from $P_1$. Observing $X_1$, i.e. the history up to time $t$: $h_1 := (X_1)$, the agent (the decision maker) makes decision by choosing an action $A_1$ and receives $r_1 \sim R_f(h_1, A_1)$

Given $h_1$, and $A_1$, the environment succeed to new state $X_2$.

*Wednesday, October 7, 2020*

At each time step $t$, the environment is at some state $X_t$, given $h_t : (X_1, A_1 \dots X_t)$, the agent makes a decision $A_t$ and receives reward

$$r_t \sim R_1(h_t, A_t)$$

and the environment succeeds to state

$$\underline{X_{t+1}} \sim \underline{P_t(h_t, A_t)}$$

(The reward could also be $r_t \sim R(h_{t+1})$ )

---

MDP is a controlled Markov Process. Therefore

$$P_t(h_t, A_t) = P_t(X_t, A_t)$$

In other words, $P_t(X_{t+1} | h_t, A_t) = P_t(X_{t+1} | X_t, A_t)$

similarly for the reward : $R(h_t, A_t) = R(X_t, A_t)$

$$r_t \sim R(X_t, A_t)$$

*Wednesday, October 7, 2020*

How the agent makes decision.

policy sets:

- History dependent and randomized policy: $\pi \in \Pi^{HR}$

$$A_t \sim \pi(h_t)$$

- History dependent and deterministic: $\pi \in \Pi^{HD}$

$$A_t = \pi(h_t)$$

- Markov and randomized: $\pi \in \Pi^{MR}$

$$A_t \sim \pi(X_t, t)$$

- Markov and deterministic: $\pi \in \Pi^{MD}$

(Memoryless policy is a special case of Markov policies, where $A_t \sim \pi(X_t)$ which is stationary.)
$A = \pi(X_t)$

(Note: Be careful about the notation since we used $\pi$ for any thing.)

Which policy class is the largest? $\Pi^{HR}$

*Wednesday, October 7, 2020*

Let's consider the case that we are interested in policies that provide us with large expected return:

Let's make it concrete.

---

– Infinite Horizon MDP

– Expected return : $\eta^{\pi} = E^{\pi}\left[\sum_{t=1}^{\infty} r_t\right]$
(Undiscounted reward)

Example: Collecting bit coin.

– Expected average return: $\eta^{\pi} = \lim_{T\to\infty} \frac{1}{T} E^{\pi}\left[\sum_{t=1}^{T} r_t\right]$

(A robot in assembling line keeps working forever)

– Expected discounted return:

which one you choose ?
– 1 grand today
– 1 grand in a year
– 1 grand and 5 bucks in two years.

$$\eta^{\pi} = E^{\pi}\left[\sum_{t=1}^{\infty} \lambda^{t-1} r_t\right] \quad 0 \leq \lambda < 1$$

Future reward worth less to us right now

*Wednesday, October 7, 2020*

Episodic:

- Expeted return (undiscounted)

$$- \eta^{\pi} = E\left[\sum_{t=1}^{\tau} r_t\right]$$

where $\tau$ is a termination (stopping) time, and is a random variable such that

$\tau < \infty$ a.s., i.e. the process terminates in finite time.

- Discounted reward

$$\eta^{\pi} = E^{\pi}\left[\sum_{t=1}^{\tau} \lambda^{t-1} r_t\right]$$

---

Note; there is a relationship between Episodic expected return setting and Infinite horizon discounted setting

For the latter $\eta^{\pi} = E^{\pi}\left[\sum_{t=1}^{\infty} \lambda^{t-1} r_t\right]$

*Wednesday, October 7, 2020*

Now, consider the case, where we terminate the process at time $t$ with probability $(1-\lambda)\lambda^{t-1}$.

$$\rightarrow \eta^\pi = E^\pi\left[\sum_{t\geq 1}^{\infty} r_t\right]$$

$$= E^\pi\left[\sum_{\tau=1}^{\infty}\left(\sum_{t=1}^{\tau} r(x_t, A_t)\right)(1-\lambda)\lambda^{\tau-1}\right]$$

if sumable and integrable
$$= E^\pi\left[\sum_{t=1}^{\infty} \underbrace{\sum_{\tau=t}^{\infty} r(x_t, A_t)(1-\lambda)\lambda^{\tau-1}}\right]$$

$$= E^\pi\left[\sum_{t=1}^{\infty} \lambda^{t-1} r(x_t, A_t)\right]$$

which is the expected discounted reward in infinite horizon setting.

---

Finite horizon MDP: The set of finite horizon MDPs is a subset of episodic

*Wednesday, October 7, 2020*

MDPs, where, there exists a finite number $H < \infty$, such that $\tau \leqslant H$

---

Fixed Horizon: is a subset of episodic MDPs where $\tau = H$. for $\cancel{t} \leqslant H < \infty$

---

Martin Putterman

MDP