Lecture 15

Monday, October 12, 2020

CS 59000-RL

MDPs

Agenda:

- Tabular Fixed Horizon MDPs
- Infinite Horizon Discounted reward MDP

Tabular fixed horizon MDPs with finite state and action:

$$M := (\mathcal{X}, \mathcal{A}, P, R, P_1, H)$$

$|\mathcal{X}| < \infty \quad |\mathcal{A}| < \infty$

at time t, state X_t, taking action A_t

result in reward $r_t \sim R_t(X_t, A_t)$

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

Simplicity $P_t(X_{t+1} | X_t, A_t)$ to denote

the probability of next state.

(It is fine when we deal with finite state and action spaces)

Let $\bar{r}_t(X_t, A_t) = E\left[r_t \mid \sigma(X_t, A_t)\right](X_t, A_t)$

Also, for the last state, we have

$r_H \sim R_H(X_H)$ and $\bar{r}_H$ is it expected value.

---

starting from state x, and following π, we have

$$V_1^{\pi}(x) = E^{\pi}\left[\sum_{k=1}^{H-1} r_k + r_H \,\middle|\, \sigma(X_1)\right](x)$$

$$= E^{\pi}\left[\sum_{k=1}^{H-1} \bar{r}_k(X_k, A_k) + r_H(X_H) \,\middle|\, \sigma(X)\right](x)$$

---

similarly for any $1 \leq t < H$

$$V_t^{\pi}(h_t) = E^{\pi}\left[\sum_{k=t}^{H-1} \bar{r}_k(X_k, A_k) + \bar{r}_H(X_H) \,\middle|\, \sigma(h_t)\right](h_t)$$

For simplicity we use the following notation

$$V_t^{\pi}(h_t) = E^{\pi}\left[\sum_{k=t}^{H-1} \bar{r}_k(X_k, A_k) + \bar{r}_H(X_H) \,\middle|\, h_t\right]$$

This is the value of h_t following π

Monday, October 12, 2020

note that

$$V_t^{\pi}(h_t) = E^{\pi}\left[\bar{r}_t(X_t, A_t) \mid h_t\right] + E^{\pi}\left[E^{\pi}\left[\sum_{k=t+1}^{H-1} \bar{r}_k(X_k, A_k) + \bar{r}_H(X_H) \,\middle|\, \sigma(h_{t+1})\right] \,\middle|\, h_t\right]$$

$\leftarrow V_{t+1}^{\pi}(h_{t+1})$

$$V_t^{\pi} = E^{\pi}\left[\bar{r}_t(X_t, A_t) \mid h_t\right] + E^{\pi}\left[V_{t+1}^{\pi}(h_{t+1}) \mid h_t\right]$$

Let's look at the first term:

$$E^{\pi}\left[\bar{r}_t(X_t, A_t) \mid h_t\right] = \sum_{a \in \mathcal{A}} \pi(a; h_t)\, \bar{r}_t(x_t, a)$$

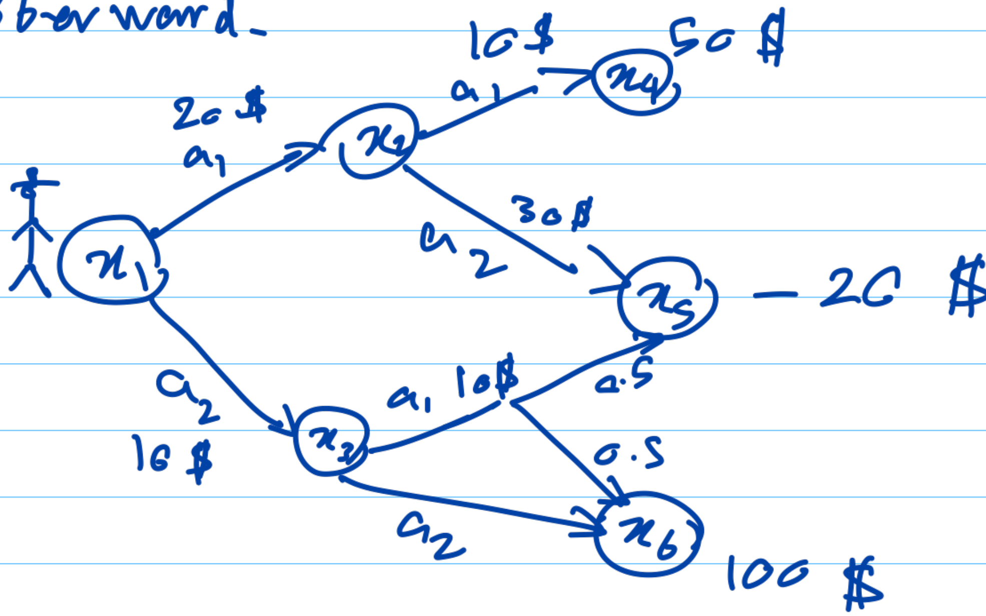
And the second term:

$$E^{\pi}\left[V_{t+1}^{\pi}(h_{t+1}) \mid h_t\right] = \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} \pi(a; h_t)\, P_t(X_{t+1} = x \mid x_t, a)\, V_{t+1}^{\pi}(h_t, a, x)$$

most likely chapter 2, 4, 5, 6, 7 of MDP book by Martin Puterman

Monday, October 12, 2020

Therefore to compute V_b^π, we look at what is the immediate reward, and what we expect afterward.



Consider this policy π: At x_1, choose a_1 and at any other state choose a_1, a_2 uniformly at random:

$$\left.\begin{aligned} V_3^\pi(x_4) &= 50 \\ V_3^\pi(x_5) &= -20 \\ V_3^\pi(x_6) &= 100 \end{aligned}\right\} \quad V_2^\pi(x_2) = \frac{1}{2}10 + \frac{1}{2}30 + \frac{1}{2}50 + \frac{1}{2}(-20) = 35$$

$$V_1(x_1) = 20 + 35 = 55$$

The thing we have done is called

Dynamic Programming

Monday, October 12, 2020

Optimality:

$$V_1^*(x) = \max_{\pi \in \Pi^{HR}} V_1^{\pi}(x) \quad x \in \mathcal{X}$$

Let's define $u_H(h_H) = \bar{r}_H(x_H)$

and $u_t(h_t) = \max_{a \in A} \left(\bar{r}_t(x_t, a) + \sum_{x \in \mathcal{X}} p_t(X_{t+1} = x \mid x_t, a) \, u_{t+1}(h_t, a, x) \right)$

Now we show $u_t = V_t^*$ and there a deterministic optimal policy.

Proof:

- we know that $u_H = V_H^*$ by definition
- we can write $u_H(h_t) \geq V_H^*(h_t)$ for all h_t.
- Now let's do some induction:

Let $u_{t+1}(h_{t+1}) \geq V_{t+1}^*(h_{t+1})$;

then we show it results in

$$u_t(h_t) > V_t^*(h_t)$$

Monday, October 12, 2020

We have

$$u_t(h_t) = \max_{a \in A} \left(\bar{r}_t(x_t, a) + \sum_{x \in \mathcal{X}} p_t(X_{t+1} = x \mid x_t, a)\, \underline{u_{t+1}(h_t, a, x)} \right)$$

$$\geq \max_{a \in A} \left(\bar{r}_t(x_t, a) + \sum_{x \in \mathcal{X}} p_t(X_{t+1} = x \mid x_t, a)\, V^{\pi}_{t+1}(h_t, a, x) \right)$$

$$\geq \sum_{a \in A} \pi(a \mid h_t) \left[\bar{r}_t(x_t, a) + \sum_{x \in \mathcal{X}} p_t(X_{t+1} = x \mid x_t, a)\, V^{\pi}_{t+1}(h_{t+1}, a, x) \right]$$

$$= V^{\pi}_t(h_t)$$

what we showed is

$$u_t(h_t) \geq V^{\pi}_t(h_t) \quad \text{for all } h_t, \text{ and } \pi \in \Pi^{HR}$$

$$\Rightarrow u_t(h_t) \geq V^*_t(h_t).$$

we just proved that following a deterministic policy which gives us u_t has value at least as good as the V^*_t. $\Rightarrow$ Therefore, that policy is optimal and $V^*_t = u_t$

Monday, October 12, 2020

Let's show $V_t^*(h_t)$ depends just on x_t, not the whole h_t. $\Rightarrow$ we have

$$V_t^*(h_t) = \max_{a \in A} \left(\bar{r}_t(x_t, a) + \sum_{x \in X} P_t(X_{t+1} = x \mid x_t, a)\, V_{t+1}^*(h_t, a, x) \right)$$

we also know that $V_H^*(h_H) = \bar{r}_H(x_H)$ is a function of h_H through x_H.

so we write $V_H^*(h_H) = V_H^*(x_H)$

$\boxed{\text{Induction}}$:

Let for all h_{t+1}, $V_{t+1}^*(h_{t+1}) = V_{t+1}^*(x_{t+1})$

Then we show, same is true for V_t^*

$$\underline{V_t^*(h_t)} = \underline{\max_{a \in A} \left(\bar{r}_t(x_t, a) + \sum_{x \in X} P_t(X_{t+1} = x \mid x_t, a)\, V_{t+1}^*(x) \right)}$$

The right hand side just depends on x_t in h_t.

Therefore, V_t^* is a function of x_t.

no matter what is h_t as a whole.

Monday, October 12, 2020

The action we choose at x_t is

$$a_t \in \underset{a \in A}{\operatorname{argmax}} \left(\bar{r}_t(x_t, a) + \sum_{x \in \mathcal{X}} P(X_{t+1} = x \mid x_t, a) V^*_{t+1}(x) \right)$$

which is a function of just x_t.

Therefore, there exist an optimal policy which is $\left\{ \begin{array}{l} \text{Deterministic} \\ \text{Markovian} \end{array} \right\} \in \Pi^{MD}$

(Note: These awesome results do not hold for general state-action spaces. In general the optimal policy may not even exist.)