# Lecture 16

Wednesday, October 14, 2020

CS 59000_RL

MDP

Agenda

- Bellman equation for Fixed Horizon MDPs
- Discounted infinite horizon MDPs.

---

For a policy $\pi \in \Pi^{MR}$, $\forall t, \forall x$

$$V_t^{\pi}(x_t) = \sum_a \pi(a|x_t)\, \bar{r}_t(x_t, a) + \sum_a \sum_x P(X_{t+1} = x | x_t, a)\, \pi(a; x_t)\, V_{t+1}^{\pi}(x)$$

$$V_t^{\pi} = T^{\pi} V_{t+1}^{\pi}$$

This is known as Bellman equation

or Bellman consistency equation.

$T^{\pi}$ is Bellman operator.

$$V_t^*(x_t) = \max_{a \in \mathcal{A}} \left( \bar{r}(x_t, a) + \sum_x P(X_{t+1} = x | x_t, a)\, V_{t+1}^*(x) \right)$$

is known as Bellman optimality equation.

and $\quad V_t^* = T^*(V_{t+1})$

where $T^*$ is Bellman optimality operator.

Q function: For a policy $\pi \in \Pi^{MR}$

$$Q^{\pi}_t(x, a) = E^{\pi}\left[\sum_{k=t}^{H} r_k \,\middle|\, X_t = x, A_t = a\right]$$

$$= \bar{r}_t(x,a) + \sum_{x'} p(X_{t+1} = x' | x, a)\, V^{\pi}_{t+1}(x)$$

and for the optimal policy.

$$Q^*_t(x, a) = \bar{r}_t(x, a) + \sum_{x'} p(X_{t+1} = x' | x, a)\, V^*_{t+1}(x')$$

$$\pi^* : \quad \arg\max_{a \in \mathcal{A}} Q^*_t(x, a)$$

---

Tabular discounted infinite horizon MDPs

(Finite state-action spaces

Consider stationary setting where we have

$p(x'|x,a)$ $\quad R(x,a)$ $\quad$ independent of time

For each state $x$, we have:

$$\Rightarrow V^{\pi}(x) = E\left[\sum_{t=1}^{\infty} \lambda^{t-1} r_t \,\middle|\, X_1 = x\right]$$

Optimal: $V^*(x) = \max_{\pi \in \Pi^{HR}} V^{\pi}(x)$

Wednesday, October 14, 2020

Assume $|\bar{r}(x,a)| \leq M < \infty$.

* With somewhat different argument, There is a stationary memory-less Markovian policy $\pi^*$ that is optimal (Also deterministic)

Therefore we focus on this policy class.
For the proof: Chapter 6 of MDP book]

---

Let's use vector and matrix notation

$$r_\pi \in R^{|X|}, \quad r_{\pi,i} = \sum_a \bar{r}(x=i, a)\,\pi(a; x=i)$$

$$P_\pi \in R^{|X| \times |X|}, \quad (P_\pi)_{i,j} = \sum_a P(j|x_i, a)\,\pi(a; i)$$

---

Now

$$V^\pi = \sum_{t=1}^{\infty} \lambda^{t-1} P_\pi^{t-1} r_1$$

$$= r_\pi + [\lambda P_\pi r_\pi + \lambda^2 P_\pi^2 r_\pi + \cdots]$$

$$= r_\pi + \lambda P_\pi V_\pi$$

$$= T^\pi(V^\pi) \quad \text{which is Bellman operator.}$$

Cool thing: $T^\pi(V) = V \to$ has a unique solution, and $V^\pi$ is that one.

Wednesday, October 14, 2020

Let's play with this equation:

$$V^{\pi} = T^{\pi}(V^{\pi}) = T^{\pi} V^{\pi} = r_{\pi} + \lambda \underbrace{P_{\pi}} V^{\pi}$$

$$\underbrace{(I - \lambda P_{\pi})} V^{\pi} = r$$

If $(I - \lambda P_{\pi})$ is full rank, $\Rightarrow V^{\pi}$ is the solution to $T^{\pi}(V) = V$

proof:

Note that $\|P_{\pi}\|_{\infty} = 1$, and $\|\lambda P_{\pi}\|_{\infty} = \lambda < 1$

Therefor, $(I - \lambda P_{\pi})$ is full rank and $(I - \lambda P_{\pi})^{-1}$ exists.

Another way of proving it. For $\alpha \in R^{|X|} \neq 0$

$$\underline{\|(I - \lambda P^{\pi})\alpha\|_{\infty}} = \|\alpha - \lambda P_{\pi}\alpha\|_{\infty}$$

Triangle inequality $\geq \|\alpha\|_{\infty} - \lambda \|P_{\pi}\alpha\|_{\infty}$

$$\geq \|\alpha\|_{\infty} - \lambda \|\alpha\|_{\infty}$$

$$= (1 - \lambda)\|\alpha\|_{\infty} > 0$$

Wednesday, October 14, 2020

Now consider the Bellman optimality equation

$$V = \max_{\pi} \left( r_\pi + \lambda P_\pi V \right)$$
$$= T V$$

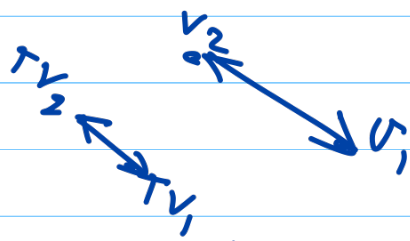
Theorem: There exist a unique $V$ that satisfies the Bellman optimality equation and it is the optimal $V^*$.

---

In the first step we need to show a solution exists. Then show it is unique.

Ten, show, such solution is $V^*$.

Lemma: The $T$ operator is contraction. under $\|\cdot\|_\infty$ norm.




proof: Consider $U, V \in \mathbb{R}^{|X|}$. For a state $x$,

$$a^*_x \in \arg\max_{a \in \mathcal{A}} \left( \bar{r}(x,a) + \lambda \sum_{x'} P(x'|x,a) V(x') \right)$$

an optimal action if $V$ was the value.

Now: Assume $TV(x) \geq TU(x)$, then

*Wednesday, October 14, 2020*

$$0 \leq TV(x) - TU(x)$$

$$\leq \bar{r}(x, a_x^*) + \lambda \sum_{x' \in \mathcal{X}} P(x'|x, a_x^*) V(x')$$

$$- \left(\bar{r}(x, a_x^*) + \lambda \sum_{x' \in \mathcal{X}} P(x', x, a_x^*) U(x')\right)$$

$$\leq \lambda \sum_{x' \in \mathcal{X}} P(x'|x, a_x^*) \left( V(x') - U(x') \right)$$

$$\leq \lambda \sum_{x' \in \mathcal{X}} P(x'|x, a_x^*) \|V - U\|_\infty = \lambda \|V - U\|_\infty$$

we can do the same when $TV(x) \leq TU(x)$

$$\Rightarrow \quad 0 \leq TU(x) - TV(x) \leq \lambda \|V - U\|_\infty$$

Making the same argument for all $x \in \mathcal{X}$

$$\longrightarrow \|TV - TU\|_\infty \leq \lambda \|V - U\|_\infty$$

---

Theorem: (Banach Fixed-Point)