FoSML: Lecture 22

Thm (Bellman's optimality condition): A policy $\tilde{\pi}$ is optimal iff for any pair $(s,a) \in S \times A$ with $\tilde{\pi}(s)(a) > 0$, $a \in \underset{a' \in A}{\arg\max}\, Q_{\tilde{\pi}}(s,a')$

Proof: Recalling our policy improvement Theorem, we know

$$\left\{ \forall s \in S,\ \mathop{\mathbb{E}}_{a \sim \pi'(s)} \left[Q_{\pi}(s,a) \right] \geq \mathop{\mathbb{E}}_{a \sim \pi(s)} \left[Q_{\pi}(s,a) \right] \right\}$$

$$\Downarrow$$

$$\left\{ \forall s \in S,\ V_{\pi'}(s) \geq V_{\pi}(s) \right\}.$$

So, suppose that π is optimal. We show $\pi(s)(a) > 0 \Rightarrow a \in \underset{a' \in A}{\arg\max}\, Q_{\pi}(s,a')$.

Indeed, suppose not, that $\exists (s_0, a_0)$ s.t. $\pi(s_0)(a_0) > 0$ but $a_0 \notin \underset{a' \in A}{\arg\max}\, Q_{\pi}(s_0, a')$.

We will construct a higher value policy π' as follows.

Let $$\pi'(s) = \begin{cases} \pi(s), & s \neq s_0 \\ a^*, & s = s_0, \end{cases}$$

where $a^* \in \underset{a' \in A}{\arg\max}\, Q_{\pi}(s_0, a')$. Then clearly $\mathop{\mathbb{E}}_{a \sim \pi(s)} \left(Q_{\pi}(s,a) \right) = \mathop{\mathbb{E}}_{a \sim \pi'(s)} \left(Q_{\pi}(s,a) \right)$ for $s \neq s_0$, while $\mathop{\mathbb{E}}_{a \sim \pi'(s)} \left(Q_{\pi}(s_0,a) \right) > \mathop{\mathbb{E}}_{a \sim \pi(s)} \left(Q_{\pi}(s_0,a) \right)$. Hence, by our policy improvement Theorem, $V_{\pi'}(s^*) > V_{\pi}(s^*)$ for some s^*. This contradicts

π being optimal, and we conclude $\pi(s)(a) > 0 \Rightarrow a \in \arg\max_{a' \in A} Q_\pi(s,a')$, as desired. ②

To show the converse direction, let π be non-optimal; we want to show there exists a state (s_0, a_0) such that $\pi(s_0)(a_0) > 0$ and $a_0 \notin \arg\max_{a \in A} Q_\pi(s_0, a)$. Well, by non-optimality of π, $\exists$ policy π' and state $s^* \in S$ such that $V_{\pi'}(s^*) > V_\pi(s)$. By our policy improvement theorem, $\exists$ s' s.t. $\mathbb{E}_{a \sim \pi(s')}(Q_\pi(s',a)) \le \mathbb{E}_{a \sim \pi'(s')}(Q_\pi(s',a))$

In particular, there is some state-action pair (s_0, a_0) s.t. $\pi(s_0, a_0) > 0$ (so it contributes to the expectation above) and $a_0 \in \arg\max_{a \in A} Q_\pi(s, a)$. ∎

We now argue that MDP have optimal policies. In fact, they have optimal deterministic policies.

Thm (Optimal deterministic policies for MDP): Any finite MDP admits an optimal, deterministic policy.

Proof: Let π^* be a deterministic policy maximizing $\sum_{s \in S} V_\pi(s)$ among all deterministic policies. Such a maximizer exists because the set of deterministic policies is finite. We claim π^* is in fact optimal.

Indeed, suppose not. Then by the Bellman optimality condition, there exists a state-action pair (s_0, a_0) s.t. $\pi^*(s_0)(a_0) > 0$ and $a_0 \notin \arg\max_{a \in A} Q_{\pi}(s_0, a)$.

But π^* is deterministic, so this simplifies to saying there is a state-action pair (s_0, a_0) such that $\pi(s_0) \notin \arg\max_{a \in A} Q_\pi(s_0, a)$. By our policy improvement theorem, we could then improve π^* by defining π' as

$$\pi'(s) = \begin{cases} \pi^*(s), & s \neq s_0, \\ a^*, & s = s_0 \end{cases}$$

for $a^* \in \arg\max_{a \in A} Q_{\pi}(s_0, a)$. But then as in the proof of the Bellman conditions,

$V_{\pi^*}(s) \leq V_{\pi'}(s)$ for all s, with strict inequality for state s_0. Notice that π' is by construction deterministic. Hence, π', π^* are deterministic and

$$\sum_{s \in S} V_{\pi^*}(s) < \sum_{s \in S} V_{\pi'}(s),$$

which contradicts our claim that π^* maximized this sum. ■

- In light of the existence of deterministic optimal policies, we will in what follows consider only deterministic policies.

- Let π^* be a deterministic optimal policy, with associated state-action value function Q^* and value function V^*.

- We know that $\pi^*(s) = \arg\max_{a \in A} Q^*(s,a)$, i.e. the optimal action at state s is to maximize Q^*. So, it is enough to know Q^*.

- Recall $Q^*(s,a) = E(r(s,a) + \gamma V^*_{\pi}(s_1) \mid s_0 = s, a_0 = a)$

Then the optimal value for the policy π^* (optimal among all policies) is

$$V^*(s) = Q^*(s, \pi^*(s))$$

$$= \max_{a \in A} \Big[\mathbb{E}(r(a,s)) + \underbrace{\gamma \sum_{s' \in S} V^*(s') \cdot \mathbb{P}(s' \mid s,a)}_{\text{expectation of } \gamma V^*(s_1) \text{, given that we start at } s_0 = s,\ a_0 = a} \Big]$$

. This holds $\forall s \in S$; together these are the so-called Bellman equations, and we may re-formulate them as:

Proposition (Bellman conditions): The values $V_\pi(s)$ of an arbitrary policy π at states $s \in S$ for an infinite time horizon MDP satisfy the following linear system: $\forall s \in S$, $V_\pi(s) = \mathbb{E}_{a_1 \sim \pi(s)} [r(s,a)] + \gamma \sum_{s' \in S} V_\pi(s') \cdot \mathbb{P}(s' \mid s, \pi(s))$.

Proof: We compute: $V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot r(s_t, \pi(s_t)) \mid s_0 = s \right]$

$$= \mathbb{E}\left[\gamma^0 \cdot r(s_0, \pi(s_0)) \mid s_0 = s \right] + \mathbb{E}\left(\sum_{t=1}^{\infty} \gamma^t \cdot r(s_t, \pi(s_t)) \mid s_0 = s \right)$$

$$= \mathbb{E}(r(s, \pi(s))) + \gamma \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t \cdot r(s_{t+1}, \pi(s_{t+1})) \mid s_0 = s \right)$$

$$= \mathbb{E}(r(s, \pi(s))) + \gamma \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t \cdot r(s_{t+1}, \pi(s_{t+1})) \mid s_1 = \delta(s, \pi(s)) \right)$$

$$= \mathbb{E}(r(s, \pi(s))) + \gamma \mathbb{E}\left(V_\pi(\delta(s, \pi(s))) \right).$$

Now, the result follows from writing out the second expectation explicitly. ∎

- This system is linear (though we get a non-linear system when we want to consider optimal policies, due to the outer maximization).

- In matrix form, we may write $V = R + \gamma P V$, where:

  - V is the unknown value function; we would like to solve for it.
  - R is the reward matrix
  - P is the Markov transition matrix defined by $P_{s,s'} = \mathbb{P}(s' \mid s, \pi(s))$.

- Note that R, V are column matrices, with

$$R_s = \mathbb{E}(r(s, \pi(s)))$$
$$V_s = V_\pi(s).$$

- We can of course try to solve for V via matrix algebra:

$$V = R + \gamma P V$$
$$\Rightarrow V - \gamma P V = R$$
$$\Rightarrow (I - \gamma P) V = R$$
$$\Rightarrow V = \underbrace{(I - \gamma P)}_{\text{guaranteed invertible because } \gamma < 1}{}^{-1} R$$

expanding

- Indeed, because P is Markovian, its rows all sum to 1, and in particular has all eigenvalues ≤ 1. Thus, γP has all its eigenvalues $\leq \gamma < 1$, so $I - \gamma P$ has all positive eigenvalues, and in particular is invertible.