EE 290 Theory of Multi-armed Bandits and Reinforcement Learning Lecture 17 - 3/16/2021

Lecture 17: Bellman Operators, Policy Iteration, and Value Iteration

Lecturer: Jiantao Jiao *Scribe: Ryan Moughan*

In this lecture we introduce the Bellman Optimality Operator as well as the more general Bellman Operator. We then introduce Policy Iteration and prove that it gets no worse on every iteration of the algorithm. Lastly we introduce Value Iteration and give a fixed horizon interpretation of the algorithm. [1]

1 Bellman Operator

We begin by defining the Bellman Optimality Operator: $\mathcal{T} : \mathbb{R}^{\mathrm{SxA}} \to \mathbb{R}^{\mathrm{SxA}}$, $f \in \mathbb{R}^{\mathrm{SxA}}$,

$$(\mathcal{T}f)(\mathrm{s,a}) \triangleq \mathrm{R(s,a)} + \gamma \langle \mathrm{P}(\cdot|\mathrm{s,a}), V_f \rangle$$

Where V_f is defined as:

$$V_f(\mathrm{s}) \triangleq \max_{\mathrm{a} \in \mathrm{A}} f(s,a)$$

We can further define the Bellman Operator for a policy, $\mathcal{T}^\pi$, as:

$$\mathcal{T}^\pi : (\mathcal{T}^\pi f)(\mathrm{s,a}) \triangleq \mathrm{R(s,a)} + \gamma \langle \mathrm{P}(\cdot|\mathrm{s,a}), V_f^\pi(s) \rangle$$

Where here we define $V_f^\pi(\mathrm{s})$ as:

$$V_f^\pi(\mathrm{s}) \triangleq \mathbb{E}_{a \sim \pi(\cdot|\mathrm{s})} \left[f(\mathrm{s,a}) \right]$$

There are two important properties about this operator:

1. Let Q^* denote the optimal Q function. Then for a Q^* that represents the optimal Q function $\Leftrightarrow Q^* = \mathcal{T}Q^*$. Similarly, we can define Q^π: Q function for policy $\pi \Leftrightarrow Q^\pi = \mathcal{T}^\pi Q^\pi$
2. γ-contraction property: for any $f, f' \in \mathbb{R}^{\mathrm{SxA}}$,

$$||\mathcal{T}f - \mathcal{T}f'||_\infty \leq \gamma ||f - f'||_\infty$$

Note that this property also holds for $\mathcal{T}^\pi$ with any π. What this is saying is that for two functions f and f' where we are iterating to a fixed point, we are getting there exponentially fast.

2 Policy Iteration Analysis

Recall the Policy Iteration Algorithm:

Algorithm 1 Policy Iteration Algorithm

1: Start with a policy π_0
2: **for** k = 1, 2, 3, ... **do**
3: $\quad \pi_k = \pi_{Q^{\pi_{\mathrm{k-1}}}}$
4: **end for**

Where we define a policy π using a particular function f, π_f, as:

$$\pi_f(\mathrm{s}) \triangleq \arg\max_{\mathrm{a} \in \mathrm{A}} f(\mathrm{s,a})$$

Theorem 1. *Policy Iteration gets no worse on every iteration*

$$||Q^* - Q^{\pi_{k+1}}||_\infty \leq \gamma ||Q^* - Q^{\pi_k}||_\infty$$

Before proving this theorem, we will first establish two properties.

Property 1: for any π,

$$\mathcal{T}^{\pi_{\mathrm{k}+1}} Q^{\pi_{\mathrm{k}}} \geq \mathcal{T}^{\pi} Q^{\pi_{\mathrm{k}}}$$

where $Q^{\pi_{\mathrm{k}}}$ is the Q function for the policy π_k and $\mathcal{T}^{\pi_{\mathrm{k}+1}}$ is the corresponding Bellman operator for the operator $\pi_{\mathrm{k}+1}$. Note that this and future inequalities over vectors are entry-wise unless otherwise specified.

Property 2:

$$\mathcal{T}^{\pi_{\mathrm{k}+1}} Q^{\pi_{\mathrm{k}}} \leq Q^{\pi_{\mathrm{k}+1}}$$

Given these two properties, we can can prove Theorem 1:

Proof

$$0 \leq Q^* - Q^{\pi_{\mathrm{k}+1}} \leq (Q^* - \mathcal{T}^{\pi_{\mathrm{k}+1}} Q^{\pi_{\mathrm{k}}}) + (\mathcal{T}^{\pi_{\mathrm{k}+1}} Q^{\pi_{\mathrm{k}}} - Q^{\pi_{\mathrm{k}+1}})$$

Note here that we were able to arrive at this equation by adding and subtracting the same term, $\mathcal{T}^{\pi_{\mathrm{k}+1}} Q^{\pi_{\mathrm{k}}}$. Then, noting that the second term in this inequality, $(\mathcal{T}^{\pi_{\mathrm{k}+1}} Q^{\pi_{\mathrm{k}}} - Q^{\pi_{\mathrm{k}+1}})$, must be less than 0 by Property 2, we get

$$Q^* - Q^{\pi_{\mathrm{k}+1}} \leq (Q^* - \mathcal{T}^{\pi_{\mathrm{k}+1}} Q^{\pi_{\mathrm{k}}})$$

Next, noting that $Q^* = \mathcal{T} Q^*$,

$$Q^* - Q^{\pi_{\mathrm{k}+1}} \leq (\mathcal{T} Q^* - \mathcal{T}^{\pi_{\mathrm{k}+1}} Q^{\pi_{\mathrm{k}}})$$

Recall that Q^* represents the optimal policy, and as such we can represent the Bellman operator as $\mathcal{T}^{\pi^*}$:

$$Q^* - Q^{\pi_{\mathrm{k}+1}} \leq (\mathcal{T}^{\pi^*} Q^* - \mathcal{T}^{\pi_{\mathrm{k}+1}} Q^{\pi_{\mathrm{k}}})$$

Then we can use Property 1 where $\pi = \pi^*$ to arrive at

$$Q^* - Q^{\pi_{\mathrm{k}+1}} \leq (\mathcal{T}^{\pi^*} Q^* - \mathcal{T}^{\pi^*} Q^{\pi_{\mathrm{k}}})$$

Finally we can use the contraction property to complete the proof:

$$\leq \gamma ||Q^* - Q^{\pi_{\mathrm{k}}}||_\infty$$

□

We now discuss the interpretations of the quantities in these properties. First, recall that we define $Q^{\pi_{\mathrm{k}}}(\mathrm{s},\mathrm{a})$ as:

$$Q^{\pi_{\mathrm{k}}}(\mathrm{s},\mathrm{a}) \triangleq \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} \mathrm{r}_t \middle| \mathrm{s}_1 = \mathrm{s}, \mathrm{a}_1 = \mathrm{a}, \mathrm{a}_{2:\infty} \sim \pi_{\mathrm{k}}\right]$$

Using this we can then define $V^{\pi}_{Q^{\pi_{\mathrm{k}}}}(s)$ as:

$$V^{\pi}_{Q^{\pi_k}}(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^{\pi_k}(s,a)\right]$$

Then by the definition of the Bellman Optimality Operator with f as Q^{π_k}:

$$(\mathcal{T}^{\pi}Q^{\pi_k})(s,a) \triangleq R(s,a) + \gamma\langle P(\cdot|s,a), V^{\pi}_{Q^{\pi_k}}(\cdot)\rangle$$

With these definitions in hand, we now discuss the quantities used in Properties 1 and 2.

A $(\mathcal{T}^{\pi_{k+1}}Q^{\pi_k})(s,a) = \mathbb{E}\left[\sum_{h=1}^{\infty}\gamma^{h-1}r_h \middle| s_1 = s, a_1 = a, a_2 \sim \pi_{k+1}, a_{3:\infty} \sim \pi_k\right]$

The interpretation of this quantity is as follows. The initial state and action are given as inputs to the Q function. But after that, the next step we take is according to π_{k+1} by the Bellman operator. Then all future steps follow π_k.

B $(\mathcal{T}^{\pi}Q^{\pi_k})(s,a) = \mathbb{E}\left[\sum_{h=1}^{\infty}\gamma^{h-1}r_h \middle| s_1 = s, a_1 = a, a_2 \sim \pi, a_{3:\infty} \sim \pi_k\right]$

Note that this is just a generalization of (A), where $\pi = \pi_k$.

C $Q^{\pi_{k+1}}(s,a) = \mathbb{E}\left[\sum_{h=1}^{\infty}\gamma^{h-1}r_h \middle| s_1 = s, a_1 = a, a_2 \sim \pi_{k+1}, a_{3:\infty} \sim \pi_{k+1}\right]$

This is just the general Q function without any Bellman Operator.

Using Property 1, A $\geq$ B. This is intuitive given that A is using the more improved policy π_{k+1} for a_2.

Using Property 2, C $\geq$ A. This is true by similar logic, C is always using the improved policy π_{k+1} and A is not. We will now prove this:

Proof

First, consider the monotonic property of $T/\mathcal{T}^{\pi}$: For any $Q \leq Q' \rightarrow \mathcal{T}Q \leq \mathcal{T}Q'$. Then it follows that

$$Q^{\pi_k} = \mathcal{T}^{\pi_k}Q^{\pi_k} \leq \mathcal{T}Q^{\pi_k} = \mathcal{T}^{\pi_{k+1}}Q^{\pi_k}$$

So we can show

$$Q^{\pi_k} \leq \mathcal{T}^{\pi_{k+1}}Q^{\pi_k} \leq \mathcal{T}^{\pi_{k+1}}(\mathcal{T}^{\pi_{k+1}}Q^{\pi_k}) \leq ... \leq (\mathcal{T}^{\pi_{k+1}})^{\infty}Q^{\pi_k} = Q^{\pi_{k+1}}$$

Noting here that:

$$\lim_{h\rightarrow\infty}(\mathcal{T}^{\pi_{k+1}})^{h}Q^{\pi_k} = Q^{\pi_{k+1}}$$

□

3 Value Iteration

Value iteration follows a different strategy than policy iteration. It seeks to directly compute the value of a given state and not the policy:

Algorithm 2 Value Iteration Algorithm

1: Set $Q^{*,0} = 0 \in \mathbb{R}^{\mathrm{S,A}}$
2: **for** h = 1,2,3, ..., H **do**
3: $\quad Q^{*,\mathrm{h}} = \mathcal{T}Q^{*,\mathrm{h}-1}$
4: **end for**

This algorithm is simple because of its use of the contraction policy:

$$||Q^{*,\mathrm{h}} - Q^*||_\infty = ||\mathcal{T}Q^{*,\mathrm{h}-1} - \mathcal{T}Q^*||_\infty \leq \gamma ||Q^{*,\mathrm{h}-1} - Q^*||_\infty \leq \gamma^{\mathrm{H}} \frac{\mathrm{R}_{\max}}{1-\gamma} \leq \epsilon$$

Since $||Q^*||_\infty \leq \frac{\mathrm{R}_{\max}}{1-\gamma}$ For some R(s,a) $\in [0, \mathrm{R}_{\max}]$ and some ϵ. This implies that:

$$\mathrm{H} \geq \frac{\log \frac{\mathrm{R}_{\max}}{\epsilon(1-\gamma)}}{1-\gamma}$$

Fixed Horizon Interpretation: Value iteration from time 1 to time H is solving a finite horizon problem. Here we define $Q^{*,\mathrm{H}}$(s,a) as the output with corresponding value function:

$$V^{*,\mathrm{H}}(\mathrm{s}) = \max_{\text{All policies}} \mathbb{E}\left[\sum_{t=1}^{\mathrm{H}} \gamma^{t-1} r_t \middle| \mathrm{s}_1 = \mathrm{s}\right]$$

Claim: The Value Iteration outputs $(\pi_{\mathrm{Q}^{*,\mathrm{H}}}, \pi_{\mathrm{Q}^{*,\mathrm{H}-1}}, ..., \pi_{\mathrm{Q}^{*,1}})$ exactly achieves the maximum in the definition of $V^{*,\mathrm{H}}$(s).

The proof was originally planned for a homework assignment, so think of the following as an exercise. We will use the claim next lecture to argue the rate of convergence of the value function.

Proof Let $\pi : \mathcal{S} \times \{1, \ldots, H\} \to \mathcal{A}$ denote a non-stationary, deterministic policy, written as $\pi(s,t)$, so the action depends on the current time in addition to state. We want to show that the policy derived from H steps of value iteration, $\pi^{VI,H}(s,t) := \pi_{Q^{*,H+1-t}}(s)$, is the maximizer of $V^{*,H}(s)$.

We proceed by induction. Let $H = 1$, then

$$V^{\pi,H}(s) = V^{\pi,1}(s) = \mathbb{E}[r_1 | s_1 = s] = R(s, \pi(s))$$

hence the maximizing policy is

$$\pi(s,1) = \arg\max_a R(s,a)$$

while $\pi^{VI}(s,1)$ is derived from taking an argmax of

$$Q^{*,1} = \mathcal{T}Q^{*,0} = R(s,a)$$

so they are identical.

Suppose now it is true for horizon length H, we will show it holds for length $H+1$ that $\pi^{VI,H+1}(s,t)$ is the maximizing policy.

We expand

$$V^{\pi,H+1}(s) = \mathbb{E}\left[\sum_{t=1}^{H+1} \gamma^{t-1} r_t \middle| s_1 = s, \pi\right] = R(s,\pi(s,1)) + \gamma \mathbb{E}\left[\sum_{t=2}^{H+1} \gamma^{t-2} r_t \middle| s_1 = s, \pi\right]$$

$$= R(s,\pi(s,1)) + \gamma \mathbb{E}\left[\sum_{t=1}^{H} \gamma^{t-1} r_{t+1} \middle| s_1 = s, \pi\right]$$

By tower law, we can introduce an inner expectation conditioned on the second state:

$$= R(s, \pi(s,1)) + \gamma \sum_{s'} P(s'|s, \pi(s,1)) \mathbb{E}\left[\sum_{t=1}^{H} \gamma^{t-1} r_{t+1} \middle| s_1 = s, s_2 = s', \pi \right]$$

By Markov Property

$$= R(s, \pi(s,1)) + \gamma \sum_{s'} P(s'|s, \pi(s,1)) \mathbb{E}\left[\sum_{t=1}^{H} \gamma^{t-1} r_{t+1} \middle| s_2 = s', \pi(\cdot,2), \ldots, \pi(\cdot, H+1) \right]$$

where the notation shows that the expectation only depends on π through its values taken at time steps 2 through $H+1$.

Then

$$V^{*,H+1}(s) = \max_{\pi} \left\{ R(s, \pi(s,1)) + \gamma \sum_{s'} P(s'|s, \pi(s,1)) \mathbb{E}\left[\sum_{t=1}^{H} \gamma^{t-1} r_{t+1} \middle| s_2 = s', \pi(\cdot,2), \ldots, \pi(\cdot, H+1) \right] \right\}$$

$$= \max_{\pi(\cdot,1)} \left\{ \max_{\pi(\cdot,2),\ldots,\pi(\cdot,H+1)} R(s, \pi(s,1)) + \gamma \sum_{s'} P(s'|s, \pi(s,1)) \mathbb{E}\left[\sum_{t=1}^{H} \gamma^{t-1} r_{t+1} \middle| s_2 = s', \pi(\cdot,2), \ldots, \pi(\cdot, H+1) \right] \right\}$$

Now

$$\mathbb{E}\left[\sum_{t=1}^{H} \gamma^{t-1} r_{t+1} \middle| s_2 = s', \pi(\cdot,2), \ldots, \pi(\cdot, H+1) \right] = V^{\pi_{(-1)}, H}(s')$$

where $\pi_{(-1)}$ is the policy that such that $\pi_{(-1)}(s,t) = \pi(s, t+1)$, i.e. it is shifted just so that it starts on time-step 1. Then the choice for $\pi_{(-1)}$ fixes a choice for $\pi(\cdot,2), \ldots, \pi(\cdot, H+1)$. So

$$V^{*,H+1}(s) = \max_{\pi(\cdot,1)} \left\{ \max_{\pi_{(-1)}} R(s, \pi(s,1)) + \gamma \sum_{s'} P(s'|s, \pi(s,1)) V^{\pi_{(-1)}, H}(s') \right\}$$

By the induction assumption, the inner maximum is achieved by choosing $\pi_{(-1)}(s,t) = \pi^{VI,H}(s,t)$, which implies the choice

$$\pi^*(s,t) = \pi^{VI,H}(s, t-1) = \pi_{Q^{*,(H+1)+1-t}}(s) = \pi^{VI,H+1}(s,t)$$

for $t \geq 2$.

Now plugging this in,

$$V^{*,H+1}(s) = \max_{\pi(\cdot,1)} \left\{ R(s, \pi(s,1)) + \gamma \sum_{s'} P(s'|s, \pi(s,1)) V^{*,H}(s') \right\}$$

To finish we just have to show the result for $t = 1$; that is, that the outer maximum is achieved at $\pi(s,1) = \pi_{Q^{*,H+1}}(s)$.

Observe that

$$V^{*,H}(s) = \max_{a} Q^{*,H}(s,a)$$

because by optimality of $V^{*,H}$ over any policy (including the policy that maximizes the RHS in the first step),

$$V^{*,H}(s) \geq \max_{a} Q^{*,H}(s,a)$$

and by rewriting the left-hand side,

$$V^{*,H}(s) = \mathbb{E}_{a \sim \pi^*(\cdot|s)}[Q^{*,H}(s,a)] \leq \max_{a} Q^{*,H}(s,a)$$

(where in this case $\pi^*(a|s)$ is just deterministic, equal to $\pi^{VI,H}(s,1)$). Then

$$V^{*,H+1}(s) = \max_{\pi(\cdot,1)} \left\{ R(s,\pi(s,1)) + \gamma \sum_{s'} P(s'|s,\pi(s,1)) \max_{a'} Q^{*,H}(s',a') \right\}$$

$$= \max_{a} \left\{ R(s,a) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q^{*,H}(s',a') \right\}$$

$$= \max_{a} \left\{ \mathcal{T} Q^{*,H}(s,a) \right\}$$

$$= \max_{a} \left\{ Q^{*,H+1}(s,a) \right\}$$

so the maximizing action is certainly $\pi^*(s,1) = \arg\max_a \left\{ Q^{*,H+1}(s,a) \right\} = \pi_{Q^{*,H+1}}(s)$, which completes the proof.

The intuition embedded in this proof is that as time goes on, the policy gets more and more greedy, eventually reducing to just greedily maximizing the reward at the last time step—as we reach the end of the horizon, there is no need to plan for the future, so short-sighted optimization is best.

□

References

[1] N. Jiang, "Mdp preliminaries," https://nanjiang.cs.illinois.edu/files/cs598, 2020.