MS&E338 Reinforcement Learning	Lecture 2 - April 4, 2018

Algorithms for MDPs and Their Convergence

Lecturer: Ben Van Roy *Scribe: Matthew Creme and Kristen Kessel*

1 Bellman operators

Recall from last lecture that we define two Bellman operators. The first,

$$(TV)(s) = \max_{a \in \mathcal{A}} \left\{ \bar{R}(s,a) + \sum_{s' \in \mathcal{S}} P_{s,a}(s')V(s') \right\}.$$

The second,

$$(T_\pi V)(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left[\bar{R}(s,a) + \sum_{s' \in \mathcal{S}} P_{s,a}(s')V(s') \right],$$

which is associated with policy π. We refer to

$$V^* = TV^* \quad \text{and} \quad V^\pi = T_\pi V$$

as the Bellman equations. The Bellman operators T and T_π satisfy two properties: monotonicity and contraction.

Proposition 1. *For all V, V', π, with $V \leq V'$, $TV \leq TV'$ and $T_\pi V \leq T_\pi V'$.*

Proof. We refer the reader to Lecture 1 for this proof. □

In proving the second property, that T and T_π are contraction operators, we first introduce two key concepts: maximum survival time and weighted max-norm.

Definition 2. *The maximum survival time of state s, denoted $\tau(s)$, is defined as*

$$\tau(s) = \max_\pi \mathbb{E}_\pi \left[\tau(s) | s_0 = s \right] < \infty.$$

Intuitively, $\tau(s)$ can be thought of as how far we must look ahead in order to plan well in the worst case, given that we start at state s. Note that we can equivalently express $\tau(s)$ as

$$\tau(s) = \max_\pi \mathbb{E}_\pi \left[\sum_{t=1}^{\tau(s)} 1 | s_0 = s \right] = \max_{a \in \mathcal{A}} \left\{ 1 + \sum_{s' \in \mathcal{S}} P_{s,a}(s')\tau(s') \right\}.$$

We use this last equation in Propositions 4 and 5.

Definition 3. *The weighted max-norm of V is defined as*

$$\|V\|_{\infty,1/\tau} = \max_{s \in \mathcal{S}} \frac{|V(s)|}{\tau(s)}.$$

Proposition 4. *For all V, V',*

$$\|TV - TV'\|_{\infty,1/\tau} \leq \alpha \|V - V'\|_{\infty,1/\tau},$$

where

$$\alpha = \max_{s \in \mathcal{S}} \frac{\tau(s) - 1}{\tau(s)} < 1.$$

Proof. First, recall from Definition 2 that the maximum survival time of state s, denoted $\tau(s)$, satisfies

$$\begin{aligned}\tau(s) &= \max_{a \in \mathcal{A}} \left\{ 1 + \sum_{s' \in \mathcal{S}} P_{s,a}(s')\tau(s') \right\} \\ &= 1 + \max_{a \in \mathcal{A}} \left\{ \sum_{s' \in \mathcal{S}} P_{s,a}(s')\tau(s') \right\}.\end{aligned}$$

Rearranging this expression yields

$$\begin{aligned}\max_{a \in \mathcal{A}} \left\{ \sum_{s' \in \mathcal{S}} P_{s,a}(s')\tau(s') \right\} &= \tau(s) - 1 \\ &= \frac{\tau(s) - 1}{\tau(s)} \tau(s) \\ &\leq \tau(s) \max_{s \in \mathcal{S}} \frac{\tau(s) - 1}{\tau(s)} \\ &= \alpha \tau(s),\end{aligned}$$

by definition of α. Then, by definition of T and $\|\cdot\|_{\infty,1/\tau}$, we have

$$\begin{aligned}\|TV - TV'\|_{\infty,1/\tau} &= \left\| \max_{a \in \mathcal{A}} \left\{ \bar{R}(s,a) + \sum_{s' \in \mathcal{S}} P_{s,a}(s')V(s') \right\} - \max_{a \in \mathcal{A}} \left\{ \bar{R}(s,a) + \sum_{s' \in \mathcal{S}} P_{s,a}(s')V'(s') \right\} \right\|_{\infty,1/\tau} \\ &= \max_{s \in \mathcal{S}} \frac{1}{\tau(s)} \left| \max_{a \in \mathcal{A}} \left\{ \bar{R}(s,a) + \sum_{s' \in \mathcal{S}} P_{s,a}(s')V(s') \right\} - \max_{a \in \mathcal{A}} \left\{ \bar{R}(s,a) + \sum_{s' \in \mathcal{S}} P_{s,a}(s')V'(s') \right\} \right| \\ &\leq \max_{s \in \mathcal{S}} \frac{1}{\tau(s)} \max_{a \in \mathcal{A}} \left| \bar{R}(s,a) + \sum_{s' \in \mathcal{S}} P_{s,a}(s')V(s') - \bar{R}(s,a) - \sum_{s' \in \mathcal{S}} P_{s,a}(s')V'(s') \right| \\ &= \max_{s \in \mathcal{S}} \frac{1}{\tau(s)} \max_{a \in \mathcal{A}} \left| \sum_{s' \in \mathcal{S}} P_{s,a}(s')V(s') - \sum_{s' \in \mathcal{S}} P_{s,a}(s')V'(s') \right| \\ &= \max_{s \in \mathcal{S}} \frac{1}{\tau(s)} \max_{a \in \mathcal{A}} \left| \sum_{s' \in \mathcal{S}} P_{s,a}(s') \left(V(s') - V'(s') \right) \right| \\ &= \max_{s \in \mathcal{S}} \frac{1}{\tau(s)} \max_{a \in \mathcal{A}} \left| \sum_{s' \in \mathcal{S}} P_{s,a}(s') \frac{V(s') - V'(s')}{\tau(s')} \tau(s') \right| \\ &\leq \max_{s \in \mathcal{S}} \frac{1}{\tau(s)} \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P_{s,a}(s') \frac{|V(s') - V'(s')|}{\tau(s')} \tau(s') \\ &\leq \max_{s \in \mathcal{S}} \frac{1}{\tau(s)} \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P_{s,a}(s')\tau(s') \max_{s'' \in \mathcal{S}} \frac{|V(s'') - V'(s'')|}{\tau(s'')} \\ &\leq \max_{s \in \mathcal{S}} \frac{1}{\tau(s)} \alpha \tau(s) \|V - V'\|_{\infty,1/\tau} \\ &= \alpha \|V - V'\|_{\infty,1/\tau},\end{aligned}$$

as required. □

Proposition 5. *For all* V, V', π,

$$\|T_\pi V - T_\pi V'\|_{\infty,1/\tau} \leq \alpha \|V - V'\|_{\infty,1/\tau},$$

where
$$\alpha = \max_{s \in \mathcal{S}} \frac{\tau(s) - 1}{\tau(s)} < 1.$$

Proof. The proof is analogous to the proof for Proposition 4 above. $\square$

Proposition 6. *T has a unique fixed point.*

Note that Proposition 6 is true for any contraction and is known as the contraction mapping theorem or the Banach fixed-point theorem.

Proof. Existence: Consider the sequence $\{V_i\}$ where $V_i = TV_{i-1}$ beginning with V_0 (i.e. the sequence V, TV, T^2V, $\ldots$). By Proposition 4,
$$\begin{aligned}
\left\|TV - T^2V\right\|_{\infty,1/\tau} &\leq \alpha \left\|V - TV\right\|_{\infty,1/\tau} \\
\left\|T^2V - T^3V\right\|_{\infty,1/\tau} &\leq \alpha \left\|TV - T^2V\right\|_{\infty,1/\tau} \leq \alpha^2 \left\|V - TV\right\|_{\infty,1/\tau} \\
&\vdots \\
\left\|T^kV - T^{k+1}V\right\|_{\infty,1/\tau} &\leq \alpha^k \left\|V - TV\right\|_{\infty,1/\tau}.
\end{aligned}$$
Since $\alpha < 1$, the sequence V, TV, T^2V, $\ldots$ is a Cauchy sequence, and as it is over a Euclidean space, which is complete, the sequence must converge. As the sequence converges, there exists some final $\bar{V}$ such that $\bar{V} = T\bar{V}$.

Uniqueness: Suppose some other $\hat{V} \neq \bar{V}$ is a fixed point of T. We then have $\hat{V} = T\hat{V}$ and
$$\begin{aligned}
\left\|\bar{V} - \hat{V}\right\|_{\infty,1/\tau} &= \left\|T\bar{V} - T\hat{V}\right\|_{\infty,1/\tau} \\
&\leq \alpha \left\|\bar{V} - \hat{V}\right\|_{\infty,1/\tau},
\end{aligned}$$
where the inequality follows from Proposition 4. As $\alpha < 1$, we must have $\bar{V} = \hat{V}$ and therefore $\hat{V}$ is unique. $\square$

Proposition 7. *For all π, T_π has a unique fixed point.*

Proof. The proof is analogous to the proof for Proposition 6 above. $\square$

Note that Propositions 6 and 7 are not necessarily true for MDP's with infinite state spaces as there can be value functions with infinite norm.

With these two properties, monotonicity and contraction, of T and T_π in hand, we can next prove that both policy and value iteration converge.

2 Planning Algorithms

2.1 Value Iteration

Algorithm 1 Value Iteration

```
Initialize V_0
for k = 0, 1, 2, ... do
    V_{k+1} = T V_k
end for
```

Proposition 8. *V_k converges to V^*.*

Proof. This is a Corollary of Proposition 6. $\square$

2.2 Policy Iteration

Algorithm 2 Policy Iteration

```
Initialize π_0
for k = 0, 1, 2, ... do
    Solve V_k = T_{π_k} V_k
    Select π_{k+1} s.t T_{π_{k+1}} V_k = T V_k
end for
```

Note that if $V_k = V^*$ then π_{k+1} is optimal and if π_k is optimal then $V_k = V^*$.

Proposition 9. *In policy iteration, there exists k_0 such that $V_k = V^*$ for all $k \geq k_0$.*

Proof. First, $V_k = T_{\pi_k} V_k$ by Algorithm 2 and $T_{\pi_k} V_k \leq TV_k$ by definition of T_{π_k} and T. Furthermore, $TV_k = T_{\pi_{k+1}} V_k$ by Algorithm 2, and therefore

$$\begin{aligned} V_k &\leq T_{\pi_{k+1}} V_k \\ &\leq T^2_{\pi_{k+1}} V_k \quad \text{by the monotonicity of } T_{\pi_{k+1}} \\ &\leq T^3_{\pi_{k+1}} V_k \quad \text{by the monotonicity of } T_{\pi_{k+1}} \\ &\vdots \\ &\leq V_{k+1}, \end{aligned}$$

where the last inequality follows from the fact that V_{k+1} is the fixed point of $T_{\pi_{k+1}}$ and therefore the sequence must converge to V_{k+1}. Thus, the sequence $\{V_0, V_1, \dots\}$ is a monotonically increasing sequence bounded above by V^*, and the sequence must therefore converge. As there are finitely many deterministic policies, this convergence must happen in finite time.

Now suppose the sequence of V's converges to V_k. We then have $V_{k+1} = T_{\pi_{k+1}} V_{k+1}$ and $V_k = T_{\pi_{k+1}} V_k$. This implies $V_k = TV_k$, and therefore $V_k = V^*$. Thus, the sequence $\{V_0, V_1, \dots\}$ converges to V^*, as required. □

2.3 Linear Programming

We look to solve the following linear program (LP), which is the dual of the LP presented in Lecture 1:

$$\begin{aligned} \min_V \sum_{s \in S} \rho(s) V(s) \\ \text{s.t. } V \geq TV, \end{aligned}$$

where $\rho(s)$ is the initial state distribution.

Proposition 10. *If $\rho(s) > 0$ for all $s \in \mathcal{S}$, then V^* is the unique optimum.*

Note that since V^* does not depend on $\rho(s)$, if necessary we can enforce $\rho(s) > 0$ for all $s \in \mathcal{S}$ by simply replacing $\rho(s)$ in the objective with another nonzero function without changing the optimal V.

Proof. First, V^* is feasible as $V^* = TV^*$, thus satisfying the constraint. Now, for any feasible V,

$$V \geq TV \geq T^2 V \cdots \geq T^k V$$

by the monotonicity of T. Furthermore, since T is a contraction with fixed point V^* (i.e. $TV^* = V^*$), the sequence V, TV, ... must converge to V^*. Therefore, for all feasible V, V dominates V^* (i.e. $V \geq V^*$). Thus, V^* minimizes the objective $\sum_{s \in S} \rho(s) V(s)$, and V^* is the optimal solution, as required. □

2.4 Asynchronous Value Iteration

Algorithm 3 Asynchronous Value Iteration

```
Initialize V_0
for k = 0, 1, 2, ... do
    Select some state s_k ∈ S
    V(s_k) := (TV)(s_k)
end for
```

Note that unlike in Algorithm 1 where we update each state synchronously, here we update only a single state at a time.

Proposition 11. *If for all $s \in \mathcal{S}$, s appears in the sequence $(s_0, s_1, \ldots)$ infinitely often, then V converges to V^*.*

Proof. Homework 1 $\square$