Bellman's equation has a unique solution

Our starting point is Bellman's equation, $V = TV$ (V is a vector over states), where the (nonlinear) operator, T, is given by

$$TV = \max_{\pi} T_\pi V \equiv T_{\pi(V)} V \,. \tag{20}$$

The policy, π, consists of a vector of actions (one for each state), $\pi \equiv (a_1, a_2, ...)$, and the i^{th} component of $T_\pi V$, denoted $(T_\pi V)_i$, is defined to be

$$(T_\pi V)_i \equiv \sum_j P^a_{ij} [\gamma V_j + R^a_{ij}] \,.$$

Here P^a_{ij} is the probability of going from state j to state i given action a (which really should be a_i, but we dropped the subscript for clarity), R^a_{ij} is the reward one receives going from state j to state i given action a, and γ is the discount factor.

We consider two types of problems: discounted and absorbing state. For the discounted problem, $0 < \gamma < 1$ and $\sum_j P_{ij} = 1$. For the absorbing state problem, $\gamma = 1$ and $\sum_j P_{ij} < 1$ for at least one i *and* all policies eventually lead to the absorbing state with probability 1.

The first issue we need to address is frustration: it is not clear that Bellman's equation, which we write

$$V = \max_{\pi} T_\pi V \tag{30}$$

makes sense. Consider, for example, a situation with there two states and two policies. It seems possible that V_1 could larger under policy 1 than policy 2, and V_2 could be larger under policy 2 than policy 1. This leads to frustration: you can maximize V_1 or V_2, but not both.

In fact, because of the structure of the above equations, this can never happens. That's because the actions in the different states decouple. Basically, an action in state i affects only row i of P_{ij} and (we assume) only row i of R_{ij}. Notice that this is a pretty strong constraint, and not always satisfied.

This rules out frustration, but now we want to know about convergence to a global maximum. First we show that Bellman's equation has at least one local maximum; then we show that the local maximum is a global one.

The first step is to show that policy iteration converges to a local maximum. We start by assuming that V^0 is the equilibrium set of values for some particular policy, π,

$$V^0 = T_\pi V^0 \,. \tag{40}$$

Now change the policy from π to $\pi + \delta$ to increase the one-step value,

$$V^1 = T_{\pi+\delta} V^0 \geq V^0 \tag{50}$$

where the greater-than-or-equal sign means: $V \geq U$ if $V_i \geq U_i$ for all i. Next, iterate,

$$V^2 = T_{\pi+\delta} V^1 = T_{\pi+\delta}(V^0 + V^1 - V^0)\,. \tag{60}$$

Defining the operator P_π as

$$(P_\pi V)_i \equiv \sum_j P^a_{ij} V_j \tag{70}$$

where a is the action corresponding to π, we have, after a small amount of algebra,

$$V^2 = T_{\pi+\delta}(V^0 + V^1 - V^0) = T_{\pi+\delta} V^0 + \gamma P_{\pi+\delta}(V^1 - V^0) = V^1 + \gamma P_{\pi+\delta}(V^1 - V^0)\,. \tag{80}$$

Since $V^1 > V^0$ and all the elements of P are positive, it follows that $V^2 \geq V^1$. Using the same analysis, it is not hard to show that V increases or stays the same on every iteration of $T_{\pi+\delta}$. Since the values are bounded, V must converge to a local maximum.

We now show that the local maximum is a global maximum. For that we have to work a little harder, and also take a couple of steps. The first one is to show that if $U \leq V$ (shorthand, as above, for $U_i \leq V_i \ \forall i$), then $TU \leq TV$. This follows from a string of inequalities,

$$\begin{aligned} TU &= T_{\pi(U)} U \\ &= T_{\pi(U)}(V - (V - U)) \\ &= T_{\pi(U)} V - \gamma P_{\pi(U)}(V - U) \\ &\leq T_{\pi(U)} V \\ &\leq T_{\pi(V)} V \\ &= TV\,. \end{aligned} \tag{90}$$

The second line is easy, the third requires a very small amount of algebra (it's what we did in Eq. (80)), the fourth follows because both P and $(V-U)$ consist of all non-negative elements, the fourth because, by definition, $T_{\pi(V)}$ is the maximum over π of $T_\pi V$, and the last line is a definition (see Eq. (20)).

For the next step – whose rationale will not be immediately obvious – we consider only the discounted case, for which $\sum_j P^a_{ij} = 1$ no matter what the policy. What we show is that if two value functions differ by a constant amount, then T brings them closer together. Letting e be a vector of all 1s ($e_i = 1\,\forall i$) and r be a scalar, we see that

$$T(V+er) = T_{\pi(V+re)}(V+re) = T_{\pi(V+re)}V + r\gamma P_{\pi(V+re)}e\,. \tag{100}$$

Because e is a constant vector and $\sum_j P^a_{ij} = 1$, the last term is simply $r\gamma e$, independent of policy. The right hand side is, therefore, largest when $\pi = \pi(V)$, and we have

$$T(V+er) = T_{\pi(V)}V + \gamma re = TV + \gamma re\,. \tag{110}$$

We are now almost done. Consider a vector V^* that satisfies

$$V^* = TV^*\,. \tag{120}$$

In other words, V^* is one of the local maximum that we know exists. Consider any vector U and choose r sufficiently large such that

$$V^* - re \le U \le V^* + re\,. \tag{130}$$

Operating on Eq. (130) k times with T, using Eqs. (90) and (110), and noting that $TV^* = V^*$, we see that

$$V^* - \gamma^k re \le T^K U \le V^* + \gamma^k re\,. \tag{140}$$

Letting $k \to \infty$ and using the fact that $0 \le \gamma < 1$, Eq. (140) tells us that all vectors U converge to V^*. Thus, V^* must be a global maximum.

This proves that if we iterate T an infinite number of times, we will converge to a global maximum. The proof basically relied on two facts; T is a contraction operator *and* there is at least one local maximum (the latter is necessary to rule out loops).

The extension to the absorbing-state case is not totally trivial, but it can be done.