Convergence of Iterative Policy Evaluation

Hasib Aslam

November 2024

1 Introduction and Formulation

The iterative policy evaluation method is used to estimate the values of the state subject to some specific policy when solving Markov Decision Processes. The core idea is to start with a random guess for the values of states, and then improve that guess iteratively.

To understand clearly, let us assume a system with n possible states represented by set S and set A representing the possible actions. Let the policy be denoted by $\pi(a \mid s)$, where $s \in S$ and $a \in A$.

Let us model our environment using two simple functions:

1. The **reward function** $R(s, a, s')$ defines the reward obtained when action a is performed in state s, resulting in a transition to state s'. *Note: Alternative formulations of the reward function exist, but they do not affect the core discussion here.*

2. The **transition probability function** $P(s' \mid s, a)$ specifies the probability of transitioning from state s to state s' when action a is taken.

Lastly, let us define the value of a state as the expected return from s when following the policy π.

$$V_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a) \left[R(s, a, s') + \gamma V_\pi(s')\right] \tag{1}$$

Where λ is the discount factor, which is used to control the agent's perception of long-term rewards.

2 Iterative Policy Evaluation

Now let us focus on the iterative policy evaluation and its convergence. The method starts by assuming a random vector, say V_g belongs to R^n, and each entry of V_g represents the initial guess for the value of the corresponding state. The new guess is then obtained by using the known equations and the old guess. The following equations are used in this case :

$$
\begin{aligned}
V_\pi(s_1) &= \sum_a \pi(a \mid s_1) \sum_{s'} P(s' \mid s, a) \left[R(s, a, s') + \gamma V_g(s') \right], \\
V_\pi(s_2) &= \sum_a \pi(a \mid s_2) \sum_{s'} P(s' \mid s, a) \left[R(s, a, s') + \gamma V_g(s') \right], \\
&\vdots \\
V_\pi(s_n) &= \sum_a \pi(a \mid s_n) \sum_{s'} P(s' \mid s, a) \left[R(s, a, s') + \gamma V_g(s') \right]. \qquad (2)
\end{aligned}
$$

Note that the updates are usually considered synchronous i.e. V_g first the same guess is used to find new values for all states and then it is updated at once.

We can see that the above n equations, one for each state, form a system of linear equations with n variables, and the value iteration method is very similar to *Jacobi Iteration Method*. Since for the convergence of the Jacobi iteration method, a necessary condition is the strict diagonal dominance of the coefficient matrix, we must show that our coefficient matrix is also strictly diagonally dominant, to prove the convergence of the Value Iteration Method. We can rewrite our system of linear equations in the following matrix form:

$$
\begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{n1} & a_{n2} & \cdots & a_{nn}
\end{pmatrix}
\begin{pmatrix}
V(s_1) \\ V(s_2) \\ \vdots \\ V(s_n)
\end{pmatrix}
=
\begin{pmatrix}
k_1 \\ k_2 \\ \vdots \\ k_n
\end{pmatrix}
$$

The condition of strict diagonal dominance can be written as:

$$
|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad \forall i \in \{1, 2, \ldots, n\}.
$$

To prove this property, let us have another look at the (2). We can expand these equations as follows.

$$
V_\pi(s_i) = \sum_a \pi(a \mid s_i) \sum_{s'} P(s' \mid s, a) R(s, a, s') + \sum_a \pi(a \mid s_i) \sum_{s'} P(s' \mid s, a) \gamma V_g(s') \qquad (3)
$$

Note that the first part of equation (3) does not involve any variables, so it is a constant say, k_i

$$
k_i = \sum_a \pi(a \mid s_i) \sum_{s'} P(s' \mid s, a) R(s, a, s') \qquad (4)
$$

Plugging (4) in (3) gives us

$$V_\pi(s_i) = k_i + \gamma \sum_a \pi(a \mid s_i) \sum_{s'} P(s' \mid s, a) V_g(s') \tag{5}$$

Generalizing the above for all states in the environment results in :

$$\begin{aligned}
V_\pi(s_1) &= k_1 + \gamma \sum_a \pi(a \mid s_1) \sum_{s'} P(s' \mid s, a) V_g(s'), \\
V_\pi(s_2) &= k_2 + \gamma \sum_a \pi(a \mid s_2) \sum_{s'} P(s' \mid s, a) V_g(s') \\
&\vdots \\
V_\pi(s_n) &= k_n + \gamma \sum_a \pi(a \mid s_n) \sum_{s'} P(s' \mid s, a) V_g(s')
\end{aligned} \tag{6}$$

Note that each equation contains n variables, and if we keep the arrangement as it is then we are essentially picking V_{s_1} from the first, V_{s_2} from the second one, and so on. If we assume that V_{si} does not appear in the right-hand side of ith then the diagonal entry from row i of the coefficient matrix are the coefficients of V_{s_i}. This is fairly natural, in most of cases. As it is very unlikely that we have an action that does not result in the change of the state at all.

Under these conditions, all diagonal entries of the coefficient matrix become 1. Since the non-diagonal entries of the coefficient matrix are contained in the leftmost term, the convergence condition takes the following form:

$$1 > \gamma \sum_a \pi(a \mid s_i) \sum_{s'} P(s' \mid s, a), \quad \forall i \in \{1, 2, \ldots, n\}.$$

Since the terms on the left hand side of the inequality represent the probabilities we have ignored the absolute symbol.

Note that upon taking a specific action, say a from a state s, there must be a transition to some state s'. Therefore:

$$1 = \sum_{s'} P(s' \mid s, a) \tag{7}$$

So using 7 we can re-write the above condition as:

$$1 > \gamma \sum_a \pi(a \mid s_i), \quad \forall i \in \{1, 2, \ldots, n\}.$$

Since the policy must define some action for every state i.e.; the sum of probabilities of all actions at a given state must be 1. We can write:

$$1 = \sum_a \pi(a \mid s_i) \tag{8}$$

Using this 8, our simplified condition for convergence becomes:

$$1 > \gamma \tag{9}$$

Therefore, the convergence of the value iterative policy evaluation method depends strictly on the discount factor. It is only guaranteed to converge if a positive less than 1 discount factor is used.