Dependence of Convergence of the Iterative Policy Evaluation on the Discount Factor

Hasib Aslam

November 2024

1 Introduction and Objective

In reinforcement learning theory, the Iterative Policy Evaluation (IPE) method is used to estimate the expected return (a.k.a value) from environment states, while following some specific strategy (a.k.a policy) to solve the Markov Decision Processes. The core idea is to start with a random guess for the state values, and then improve on iteratively.

In this context, a discount factor, denoted by γ, is often introduced, where $0 < \gamma \leq 1$, to regulate the agent's emphasis on long-term rewards. Our objective is to theoretically analyze the convergence of IPE with respect to the discount factor γ.

2 Formulation

Before actually deriving the dependence, let us formalize the notation first. Consider a system with n possible states represented by the set $S = \{s_1, s_2, \ldots, s_n\}$, and k possible actions represented by the set $A = \{a_1, a_2, \ldots, a_k\}$. Let the policy be denoted by $\pi(a \mid s)$, where $s \in S$ and $a \in A$.

Let the environment's behaviors be modeled by following two 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 that the 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 formally define the value of a state s as the expected return 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.

3 Iterative Policy Evaluation and its Dependence

Now let us focus on the iterative policy evaluation and its convergence. The method starts by assuming a random vector, say $V_g \in R^n$, 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 we assume the updates to be performed *synchronously* across all states; that is, a single estimate V_g is used to compute new values for all states simultaneously, and the update is applied only after all new values have been obtained.

The n equations defined above, one corresponding to each state, collectively form a system of n linear equations in n unknowns. The value iteration procedure can thus be interpreted as an instance of the *Jacobi Iteration Method* applied to this linear system.

For the Jacobi iteration method to converge, the underlying coefficient matrix must be singular. Since by *Levy–Desplanques Theorem* a **strictly diagonally dominant** matrix is guaranteed to be singular. Therefore, to establish the convergence of the value iteration method, it suffices to verify that the associated coefficient matrix satisfies the condition of strict diagonal dominance.

We can express the system of linear equations derived earlier in matrix form as follows:

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

The condition for strict diagonal dominance of the coefficient matrix is given by

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

To verify that our system satisfies this property, consider again the Bellman expectation equation:

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

The first term in the above equation does not involve any value variable and can thus be treated as a constant:

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

Substituting this back, we obtain:

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

Generalizing for all states $s_1, s_2, \ldots, s_n$, we obtain the following system:

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

Each equation above involves n value variables. Assuming that $V(s_i)$ does not appear on the right-hand side of the i-th equation (which is typically valid, as most actions result in transitions to different states), the diagonal entries of the coefficient matrix correspond to the coefficients of $V(s_i)$ and thus take the value 1.

Under this assumption, the strict diagonal dominance condition simplifies to

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

Since the transition probabilities satisfy

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

the above condition becomes

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

Furthermore, because a policy defines a valid probability distribution over actions at each state,

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

the convergence condition reduces to the simple form

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

Therefore, the convergence of the value iteration (policy evaluation) method is guaranteed only when the discount factor γ satisfies $0 < \gamma < 1$.