we compute the state value function $V^\pi(s)$ for the policy π. This is called the *policy evaluation* problem. We can write the Bellman equation for the state value function as

$$\begin{aligned} V^\pi(s) &= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \mid s_t = s\right] \\ &= \mathbb{E}_\pi\left[G_t \mid s_t = s\right] \\ &= \mathbb{E}_\pi\left[R_{t+1} + \gamma G_{t+1} \mid s_t = s\right] \\ &= \mathbb{E}_\pi\left[R_{t+1} + \gamma V^\pi(s_{t+1}) \mid s_t = s\right] \\ &= \sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma V^\pi(s')\right] \end{aligned} \tag{1}$$

If enviroment dynamics are known, then $V^\pi(s)$ is a simultaneous linear equations in $|S|$ unknowns. Iterative solutions are most suitable. Assume v_0 as the initial approximation for the state value function. Then update rule is given by the bellman equation

$$v_{k+1}(s) = \sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma v_k(s')\right] \tag{2}$$

This iterative approach converges to $V^\pi(s)$. The policy evaluation problem is solved when $v_k(s) = V^\pi(s)$ for all $s \in S$.

Iterative Policy Evaluation Algorithm

Algorithm 1 Iterative Policy Evaluation

```
Input policy $\pi$, initial approximation $v_0(s) \forall s \in S$,
Set discount factor $\gamma$, $k = 0$, $V(terminal) = 0$
$v_k(s) \leftarrow v_0(s)$ for all $s \in S$
while $|v_k - v_{k-1}| < \theta$ do
    $v_{k+1}(s) \leftarrow \sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma v_k(s')\right]$ for all $s \in S$
    $k \leftarrow k + 1$
end while
return $v_k(s)$ for all $s \in S$
```

State-Action Value Function

The state-action value function $Q^\pi(s, a)$ is defined as the expected return starting from state s, taking action a, and then following policy π thereafter. We can write the Bellman equation for the state-action value function as

$$
\begin{aligned}
Q^{\pi}(s,a) &= \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t} r_{t+1} \mid s_t = s, a_t = a\right] \\
&= \mathbb{E}_{\pi}\left[G_t \mid s_t = s, a_t = a\right] \\
&= \mathbb{E}_{\pi}\left[R_{t+1} + \gamma G_{t+1} \mid s_t = s, a_t = a\right] \\
&= \mathbb{E}_{\pi}\left[R_{t+1} + \gamma Q^{\pi}(s_{t+1}, a_{t+1}) \mid s_t = s, a_t = a\right] \\
&= \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma \sum_{a'} \pi(a' \mid s') Q^{\pi}(s', a')\right] \\
&= \mathbb{E}_{\pi}\left[R_{t+1} + \gamma Q^{\pi}(s_{t+1}, a_{t+1}) \mid s_t = s, a_t = a\right] \\
&= \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma V^{\pi}(s')\right]
\end{aligned}
\tag{3}
$$

policy improvement theorem helps us in updating the policy once we found out the value function of a policy by policy evaluation. The theorem states that if we have a policy π and a value function $V^{\pi}(s)$, then there exists a policy π' such that $V^{\pi'}(s) \geq V^{\pi}(s)$ for all $s \in S$. This means that we can always improve the value of a policy by following a greedy policy with respect to the value function. The greedy policy with respect to the value function is defined as

$$
\pi'(s) = \arg\max_{a \in \mathcal{A}} Q^{\pi}(s,a) \tag{4}
$$

which leads to $V^{\pi'}(s) \geq V^{\pi}(s)$ for all $s \in S$.

Policy Improvement Algorithm

Assuming a deterministic policy $\pi(s) = a$

Algorithm 2 Policy Improvement

```
Input policy $\pi$, value function $v(s)$ for all $s \in S$,
Set discount factor $\gamma$
Evaluate the policy $\pi$ to get $v(s)$ for all $s \in S$ using Algorithm 1 Policy Evaluation.
Improve the policy by following a greedy policy with respect to the value function.$\pi \leftarrow \arg\max_{a \in \mathcal{A}} Q^{\pi}(s,a)$ for all $s \in S$
```

However, policy improvement algorithm involves policy evaluation. So, we need to repeat the policy evaluation and policy improvement until the policy converges. So, Value iteration algorithm updates the value function by acting greedily with respect to the value function and then improves the policy by following a greedy policy with respect to the value function. The algorithm is given below.

Value Iteration Algorithm

Assuming a deterministic policy $\pi(s) = a$

Algorithm 3 Value Iteration

```
Input policy $\pi$, initial approximation $v_0(s) \forall s \in S$,
Set discount factor $\gamma$, $k = 0$, $V(terminal) = 0$
$v_k(s) \leftarrow v_0(s)$ for all $s \in S$
while $|v_k - v_{k-1}| < \theta$ do
    $v_{k+1}(s) \leftarrow \max_{a \in \mathcal{A}} \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma v_k(s')\right]$ for all $s \in S$
    $k \leftarrow k + 1$
end while
return $v_k(s)$ for all $s \in S$
```