Algorithm 1: Policy Evaluation

```
Input: MDP, policy $\pi$, small positive number $\theta$
Output: $V \approx v_\pi$
Initialize $V$ arbitrarily (e.g., $V(s) = 0$ for all $s \in \mathcal{S}^+$)
repeat
    $\Delta \leftarrow 0$
    for $s \in \mathcal{S}$ do
        $v \leftarrow V(s)$
        $V(s) \leftarrow \sum_{a \in \mathcal{A}(s)} \pi(a|s) \sum_{s' \in \mathcal{S}, r \in \mathcal{R}} p(s', r|s, a)(r + \gamma V(s'))$
        $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
    end
until $\Delta < \theta$;
return $V$
```

Algorithm 2: Estimation of Action Values

```
Input: MDP, state-value function $V$
Output: action-value function $Q$
for $s \in \mathcal{S}$ do
    for $a \in \mathcal{A}(s)$ do
        $Q(s, a) \leftarrow \sum_{s' \in \mathcal{S}, r \in \mathcal{R}} p(s', r|s, a)(r + \gamma V(s'))$
    end
end
return $Q$
```

Algorithm 3: Policy Improvement

```
Input: MDP, value function $V$
Output: policy $\pi'$
for $s \in \mathcal{S}$ do
    for $a \in \mathcal{A}(s)$ do
        $Q(s,a) \leftarrow \sum_{s' \in \mathcal{S}, r \in \mathcal{R}} p(s', r|s,a)(r + \gamma V(s'))$
    end
    $\pi'(s) \leftarrow \arg\max_{a \in \mathcal{A}(s)} Q(s,a)$
end
return $\pi'$
```

Algorithm 4: Policy Iteration Estimation of Action Values

```
Input: MDP, small positive number $\theta$
Output: policy $\pi \approx \pi_*$
Initialize $\pi$ arbitrarily (e.g., $\pi(a|s) = \frac{1}{|\mathcal{A}(s)|}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$)
$policy\text{-}stable \leftarrow false$
repeat
    $V \leftarrow$ Policy_Evaluation$(\text{MDP}, \pi, \theta)$
    $\pi' \leftarrow$ Policy_Improvement$(\text{MDP}, V)$
    if $\pi = \pi'$ then
        $policy\text{-}stable \leftarrow true$
    end
    $\pi \leftarrow \pi'$
until $policy\text{-}stable = true$;
return $\pi$
```

Algorithm 5: Truncated Policy Evaluation

```
Input: MDP, policy $\pi$, value function $V$, positive integer $max\_iterations$
Output: $V \approx v_\pi$ (if $max\_iterations$ is large enough)
$counter \leftarrow 0$
while $counter < max\_iterations$ do
    for $s \in \mathcal{S}$ do
        $V(s) \leftarrow \sum_{a \in \mathcal{A}(s)} \pi(a|s) \sum_{s' \in \mathcal{S}, r \in \mathcal{R}} p(s', r|s,a)(r + \gamma V(s'))$
    end
    $counter \leftarrow counter + 1$
end
return $V$
```

Algorithm 6: Truncated Policy Iteration

```
Input: MDP, positive integer $max\_iterations$, small positive number $\theta$
Output: policy $\pi \approx \pi_*$
Initialize $V$ arbitrarily (e.g., $V(s) = 0$ for all $s \in \mathcal{S}^+$)
Initialize $\pi$ arbitrarily (e.g., $\pi(a|s) = \frac{1}{|\mathcal{A}(s)|}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$)
repeat
    $\pi \leftarrow$ Policy_Improvement(MDP, $V$)
    $V_{old} \leftarrow V$
    $V \leftarrow$ Truncated_Policy_Evaluation(MDP, $\pi$, $V$, $max\_iterations$)
until $\max_{s \in \mathcal{S}} |V(s) - V_{old}(s)| < \theta$;
return $\pi$
```

Algorithm 7: Value Iteration

```
Input: MDP, small positive number $\theta$
Output: policy $\pi \approx \pi_*$
Initialize $V$ arbitrarily (e.g., $V(s) = 0$ for all $s \in \mathcal{S}^+$)
repeat
    $\Delta \leftarrow 0$
    for $s \in \mathcal{S}$ do
        $v \leftarrow V(s)$
        $V(s) \leftarrow \max_{a \in \mathcal{A}(s)} \sum_{s' \in \mathcal{S}, r \in \mathcal{R}} p(s', r|s, a)(r + \gamma V(s'))$
        $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
    end
until $\Delta < \theta$;
$\pi \leftarrow$ Policy_Improvement(MDP, $V$)
return $\pi$
```