Value iteration

The algorithm

$$v_{k+1} = f(v_k) = \max_\pi (r_\pi + \gamma P_\pi v_k)$$

is called **value iteration**, which can be decomposed to two steps

- Step 1: policy update (PU).

  $$\pi_{k+1} = \arg\max_\pi (r_\pi + \gamma P_\pi v_k)$$

  where v_k is given.

- Step 2: value update (VU).

  $$v_{k+1} = r_{\pi_{k+1}} + \gamma P_{\pi_{k+1}} v_k$$

Procedure summary:

$$v_k(s) \rightarrow q_k(s,a) \rightarrow \text{greedy policy } \pi_{k+1}(a|s) \rightarrow \text{new value } v_{k+1} = \max_a q_k(s,a)$$

Pseudocode:

Initialization: The probability models $p(r|s,a)$ and $p(s'|s,a)$ for all (s,a) are known. Initial guess v_0.
Goal: Search for the optimal state value and an optimal policy for solving the Bellman optimality equation.

While v_k has not converged in the sense that $\|v_k - v_{k-1}\|$ is greater than a predefined small threshold, for the kth iteration, do
    For every state $s \in \mathcal{S}$, do
        For every action $a \in \mathcal{A}(s)$, do
            q-value: $q_k(s,a) = \sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a)v_k(s')$
        Maximum action value: $a_k^*(s) = \arg\max_a q_k(s,a)$
        Policy update: $\pi_{k+1}(a|s) = 1$ if $a = a_k^*$, and $\pi_{k+1}(a|s) = 0$ otherwise
        Value update: $v_{k+1}(s) = \max_a q_k(s,a)$

Policy iteration

Given a random initial policy π_0.

- Step 1: policy evaluation (PE)

  Get state value by

  $$v_{\pi_k} = r_{\pi_k} + \gamma P_{\pi_k} v_{\pi_k}$$

- Step 2: policy improvement (PI)

  $$\pi_{k+1} = \arg\max_\pi (r_\pi + \gamma P_\pi v_{\pi_k})$$

Procedure summary:

$$\pi_0 \xrightarrow{PE} v_{\pi_0} \xrightarrow{PI} \pi_1 \xrightarrow{PE} v_{\pi_1} \xrightarrow{PI} \pi_2 \xrightarrow{\dots}$$

Pseudocode:

Initialization: The system model, $p(r|s,a)$ and $p(s'|s,a)$ for all (s,a), is known. Initial guess π_0.
Goal: Search for the optimal state value and an optimal policy.

While v_{π_k} has not converged, for the kth iteration, do
  Policy evaluation:
  Initialization: an arbitrary initial guess $v_{\pi_k}^{(0)}$
  While $v_{\pi_k}^{(j)}$ has not converged, for the jth iteration, do
    For every state $s \in \mathcal{S}$, do
      $v_{\pi_k}^{(j+1)}(s) = \sum_a \pi_k(a|s) \left[\sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a) v_{\pi_k}^{(j)}(s')\right]$
  Policy improvement:
  For every state $s \in \mathcal{S}$, do
    For every action $a \in \mathcal{A}$, do
      $q_{\pi_k}(s,a) = \sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a) v_{\pi_k}(s')$
    $a_k^*(s) = \arg\max_a q_{\pi_k}(s,a)$
    $\pi_{k+1}(a|s) = 1$ if $a = a_k^*$, and $\pi_{k+1}(a|s) = 0$ otherwise

Truncated policy iteration

Based on $v_{\pi_1}^{(0)} = v_0 = v_{\pi_0}$, we can compare the policy iteration algorithm and the value iteration algorithm, getting truncated policy iteration algorithm.

$$\begin{aligned}
& v_{\pi_1}^{(0)} = v_0 \\
\text{value iteration} \leftarrow v_1 \leftarrow\ & v_{\pi_1}^{(1)} = r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(0)} \\
& v_{\pi_1}^{(2)} = r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(1)} \\
& \vdots \\
\text{truncated policy iteration} \leftarrow \bar{v}_1 \leftarrow\ & v_{\pi_1}^{(j)} = r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(j-1)} \\
& \vdots \\
\text{truncated policy iteration} \leftarrow v_{\pi_1} \leftarrow\ & v_{\pi_1}^{(\infty)} = r_{\pi_1} + \gamma P_{\pi_1} v_{\pi_1}^{(\infty)}
\end{aligned}$$

Initialization: The probability models $p(r|s,a)$ and $p(s'|s,a)$ for all (s,a) are known. Initial guess π_0.
Goal: Search for the optimal state value and an optimal policy.

While v_k has not converged, for the kth iteration, do
  Policy evaluation:
  Initialization: select the initial guess as $v_k^{(0)} = v_{k-1}$. The maximum number of iterations is set as j_{truncate}.
  While $j < j_{\text{truncate}}$, do
    For every state $s \in \mathcal{S}$, do
      $v_k^{(j+1)}(s) = \sum_a \pi_k(a|s) \left[\sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a) v_k^{(j)}(s')\right]$
  Set $v_k = v_k^{(j_{\text{truncate}})}$
  Policy improvement:
  For every state $s \in \mathcal{S}$, do
    For every action $a \in \mathcal{A}(s)$, do
      $q_k(s,a) = \sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a) v_k(s')$
    $a_k^*(s) = \arg\max_a q_k(s,a)$
    $\pi_{k+1}(a|s) = 1$ if $a = a_k^*$, and $\pi_{k+1}(a|s) = 0$ otherwise