1 Equations

Optimal policy:

$$\pi^*(s) = argmax_\pi V^\pi(s) = argmax_{a \in A(s)} \sum_{s'} P(s'|s,a)V(s')$$

Bellman Equation:

$$V(s) = R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a)V(s')$$

Bellman Update:

$$V_{i+1}(s) \leftarrow BV_i = R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a)V_i(s')$$

Bellman=Contraction:

$$\|V\| = \max_s |V(s)|$$

$$\|BV_i - BV_i'\| \leq \gamma \|V_i - V_i'\|$$

$$\|BV_i - V\| \leq \gamma \|V_i - V\|$$

Error of the estimate V_i:

$$\|V_i - V\|$$

$$\|V_0 - V\| \leq 2R_{max}/(1-\gamma)$$

Bound on state values (utilities):

$$V(s) \leq \pm R_{max}/(1-\gamma)$$

To get $\|V_i - V\| \leq \epsilon$:

$$\gamma^N 2R_{max}/(1-\gamma) \leq \epsilon$$

$$N = \left\lceil \frac{log(2R_{max}/(\epsilon(1-\gamma)))}{log(1/\gamma)} \right\rceil$$

If $\|V_{i+1} - V_i\| \leq \epsilon(1-\gamma)/\gamma$ then $\|V_{i+1} - V\| < \epsilon$

Policy loss is $\|V^{\pi_i} - V\|$ and is connected to V_i:

$$\text{if } \|V_i - V\| < \epsilon \text{ then } \|V^{\pi_i} - V\| < w\epsilon\gamma/(1-\gamma)$$

Policy Iteration:

Policy Evaluation: Given a policy π_i, calculate $V_i = V^{\pi_i}$, the utility of each state if π_i were t be executed.

Policy Improvement: Calculate a new MEU policy π_{i+1}, using one-step look-ahead based on V_i.

Terminate when policy improvement yields no change in the utilities.

function POLICY-ITERATION(*mdp*) **returns** a policy
  inputs: *mdp*, an MDP with states S, actions $A(s)$, transition model $P(s' \mid s, a)$
  local variables: U, a vector of utilities for states in S, initially zero
    π, a policy vector indexed by state, initially random

  repeat
    $U \leftarrow$ POLICY-EVALUATION(π, U, *mdp*)
    unchanged? $\leftarrow$ true
    for each state s **in** S **do**
      if $\max_{a \in A(s)} \sum_{s'} P(s' \mid s, a)\ U[s'] > \sum_{s'} P(s' \mid s, \pi[s])\ U[s']$ **then do**
        $\pi[s] \leftarrow \operatorname{argmax}_{a \in A(s)} \sum_{s'} P(s' \mid s, a)\ U[s']$
        unchanged? $\leftarrow$ false
  until *unchanged*?
  return π