Policy Iteration Algorithm

Algorithm 1 Policy Iteration

Initialize a random policy π
repeat
  Policy Evaluation:
  Initialize $V(s) = 0$ for all states s
  repeat
    for all states s **do**
      $V_{\text{new}}(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$
    end for
    $V \leftarrow V_{\text{new}}$
  until convergence
  Policy Improvement:
  Policy-stable $\leftarrow$ **true**
  for all states s **do**
    $old\_action \leftarrow \pi(s)$
    $\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$
    if $old\_action \neq \pi(s)$ **then**
      Policy-stable $\leftarrow$ **false**
    end if
  end for
until Policy-stable