

function POLICY-ITERATION(*mdp*) **returns** a policy

inputs: *mdp*, an MDP with states S , actions $A(s)$, transition model $P(s')$

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 $\max_{a \in A(s)} \sum_{s'} P(s' | s, a) U[s'] > \sum_{s'} P(s' | s, \pi[s]) U[s']$ **the**

$\pi[s] \leftarrow \operatorname{argmax}_{a \in A(s)} \sum_{s'} P(s' | s, a) U[s']$

unchanged? \leftarrow false

until *unchanged?*

return π

Figure 17.7 The policy iteration algorithm for calculating an optimal r