Exercise 4.6, Sutton-Barto

Ricardo Montesinos Fernandez

Exercise 4.6, How would policy iteration be defined for action values? Give a complete algorithm for computing q_*, analogous to that on page 65 for computing v_*. Please pay special attention to this exercise, because the ideas involved will be used throughout the rest of the book.

1. Initialization $Q(s,a) \in \mathbb{R}$ and $\pi(s) \in A(s)$ arbitrarily for all $s \in S$.
2. Policy Evaluation

repeat
  $\Delta \leftarrow \emptyset$;
  foreach $s \in S$ *and* $a \in A$ **do**
    $q \leftarrow Q(s,a)$
    $Q(s,a) \leftarrow R(s,a) + \gamma \sum_{s'} p(s', r|s, a) q(s', \pi(s'))$
    $\Delta \leftarrow max(\Delta, |q - Q(s,a)|)$
  end
until $\Delta < 0$ *(small...)*;

3. Policy Improvement

begin
  $policy\_stable \leftarrow True$
  foreach (s,a) **do**
    $old\_action \leftarrow \pi(s)$
    $\pi(s) \leftarrow argmax_a \; q_\pi(s,a)$
    if $old\_action \neq \pi(s)$ **then**
      $policy\_stable \leftarrow False$
    end
  end
  if *policy_stable* **then**
    Stop and Return $q \approx q_*$ and $\pi \approx \pi_*$
  else
    Go to Step 2.
  end
end

Where: $q_\pi(s,a) = \sum_{r,s'} p(s', r|s, a)[r + \gamma q(s', \pi(s'))]$

Algorithm 1: Policy iteration (using iterative policy evaluation)