Implementation: Policy Improvement

In the last lesson, you learned that given an estimate Q of the action-value function q_π corresponding to a policy π, it is possible to construct an improved (or equivalent) policy π', where $\pi' \geq \pi$.

For each state $s \in \mathcal{S}$, you need only select the action that maximizes the action-value function estimate. In other words,

$\pi'(s) = \arg\max_{a \in \mathcal{A}(s)} Q(s,a)$ for all $s \in \mathcal{S}$.

The full pseudocode for **policy improvement** can be found below.

Policy Improvement

Input: MDP, value function V
Output: policy π'
for $s \in \mathcal{S}$ **do**
  for $a \in \mathcal{A}(s)$ **do**
    $Q(s,a) \leftarrow \sum_{s' \in \mathcal{S}, r \in \mathcal{R}} p(s', r|s,a)(r + \gamma V(s'))$
  end
  $\pi'(s) \leftarrow \arg\max_{a \in \mathcal{A}(s)} Q(s,a)$
end
return π'

In the event that there is some state $s \in \mathcal{S}$ for which $\arg\max_{a \in \mathcal{A}(s)} Q(s,a)$ is not unique, there is some flexibility in how the improved policy π' is constructed.

In fact, as long as the policy π' satisfies for each $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$:

$\pi'(a|s) = 0$ if $a \notin \arg\max_{a' \in \mathcal{A}(s)} Q(s,a')$,

it is an improved policy. In other words, any policy that (for each state) assigns zero probability to the actions that do not maximize the action-value function estimate (for that state) is an improved policy. Feel free to play around with this in your