3 (a)

Prove that if policy iteration terminates with the policy π^*, π^* is an optimal stationary policy.

We have that a policy $\pi(s)$ satisfies a corresponding value function
$V(s) = R(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s))V^{\pi}(s')$

Policy iteration solves the fixed point equation by beginning with any policy π_0, applying the right hand of the below equation to generate new policies until it converges:

Eq (1) $\pi'(s) = argmax_{a\in A} R(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s))V'(s')$

Since for each step of policy iteration, we take the argmax of the term on the right hand side, and at the last step, we will get policy π^* as the argmax of the above expression. However, we note that the Bellman equation for the value of an optimal policy is given by

Eq (2) $V^{\pi^*}(s) = max_{a\in A} R(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s))V^{\pi^*}(s')$

Noting that the right hand sides of Equations 1 and 2 above are almost the same other than the max and argmax, we note that the terminating policy generated from Equation 1, since it is the argmax of a set of actions, exactly maximizes the corresponding value for that policy.

It thus satisfies Bellman equation, Equation 2, and maximizes the Value equation so terminating policy π^* must be optimal.

3 (b)

We have for policy $\pi^{(1)}$, it has a corresponding value equation

$V_1(s) = R(s, \pi_1(a)) + \gamma \sum_{s'} P(s'|s, \pi(s))V_1^{\pi}(s')$
and for policy $\pi^{(2)}$ we have

(1) $\pi^{(2)} = argmax_{a\in A} R(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s))V_2(s')$
with corresponding value equation $V_2(s) = R(s, \pi^{(2)}(a)) + \gamma \sum_{s'} P(s'|s, \pi(s))V_2^{\pi}(s')$

When we step from V_1 to V_2, we are changing actions and also changing values.

To see our result we break this step into two parts, keeping one aspect constant per part.

We step from V_{1a} to V_{1b} by keeping the value in the right hand side of the equation constant:
$V_{1,1}(s) = R(s, \pi_1(s)) + \gamma \sum_{s'} P(s'|s, \pi_1(s)) V_1^{\pi}(s')$
$V_{1,2}(s) = R(s, \pi_2(s)) + \gamma \sum_{s'} P(s'|s, \pi_2(s)) V_1^{\pi}(s')$

and via the (policy iteration) update rule (1) we have that the update rule sets it to the argmax term, so we have that the value increases from V_{0a} to V_{0b}. We only need to update the action we take once in this step since it is argmax of the equation.
The next step is updating the value to complete the policy iteration step.

The next part of the iteration is updating V_{1b} to V_{1c}, keeping actions constant:

$V_{1,3}(s) = R(s, \pi_2(s)) + \gamma \sum_{s'} P(s'|s, \pi_2(s)) V_{1,2}^{\pi}(s')$

We observe $V_{1,3} - V_{1,2} = \gamma \sum_{s'} P(s'|s, \pi_2(s))(V_{1,3}^{\pi}(s) - V_{1,2}(s'))$,

This equation is satisfied inductively for any $V_{1,n}, V_{1,n+1}$ since the expression on the right hand side decreases by a factor of γ for each iteration because $0 \leq \gamma < 1$. This because a geometric series that will converge to $V_2(s)$. Because they all converge in a monotonically increasing series, it thus follows that $V_2(s) \geq V_1(s)$.