## Theory Question:

Consider the infinite-horizon long-run average cost setting. The policy is given.

In the class, we have learned TD(0) learning for evaluating in the infinite-horizon discounted cost setting. Now derive how you would perform TD(0) learning in the long-run average cost setting. (You can assume the average cost from policy is estimated from a separate simulation and is given.)

## Numerical Question:

Continue with the blackjack game setting, but you can think the game runs in the infinite-horizon discounted cost setting, with discount factor .

In (1) and (2), still use the policy that sticks if the player’s sum is 20 or 21, and otherwise hits.

1. Implement the TD(0) learning method to evaluate the value function under the given policy. Try three discount factors: 0.999, 0.9, and 0.8. Then compare the value functions you obtained under the three discount factors to see if there is any difference.
2. Compare the value function you get from TD(0) learning with the one you got from Monte-Carlo in the last homework.
3. Use the Policy Improvement framework along with TD(0) learning to find an optimal policy (use 0.999 as the discount factor).