Assignment 2

1. **Explain clearly why V_pi is not useful in the MC development above?**

   The reason V_pi is not useful in this specific MC development is that it only provides information about the value of states under the fixed policy, but it does not provide information about the value of taking specific actions in those states.

2. **The MC algorithm so far (ref: p 99), requires an infinite number of episodes for Eval to converge on Q_pi_k (step k). We can modify this algorithm to the practical variant where Eval is truncated (c.f., DynProg GPI). In this case:**

   a. **Will we obtain Q_pi_k from eval?**

   We will get a valuable approximation Q_pi_k from the truncated evaluation, even if the Monte Carlo algorithm relies on averaging the returns obtained from multiple episodes to estimate the action values. With a finite number of episodes, the estimation is based on a limited set of samples and may not capture the true expected returns accurately. The more episodes we run in the evaluation phase, the closer the estimated Q_pi_k will be to the true values. Therefore, while we won't obtain the exact Q_pi_k from the truncated evaluation, the estimated Q_pi_k can still serve as a valuable approximation that can be used for policy improvement and decision-making.

   b. **If not why are we able to truncate Eval? Explain clearly.**

   The key idea behind truncating the evaluation step is that, even though we are not obtaining the exact value or action-value function, the estimates can still converge to a close approximation with enough episodes. As we run more episodes and collect more samples, the estimates tend to become more accurate and converge towards the true values.

   c. **Assuming ES (i.e., thorough sampling of the S x A space), and the above truncated Eval_trunc, is it possible to converge on a sub-optimal policy pi_c? Is this a stable fixed point of the GPI for MC? Explain clearly.**

   Yes, it is possible to converge on a sub-optimal policy pi_c using the Monte Carlo algorithm with truncated evaluation (Eval_trunc). However, a sub-optimal policy obtained through Monte Carlo methods is not a stable fixed point of the GPI framework. Due to the limitations of truncated evaluation, the estimated action-values might not accurately reflect the true values. As a result, the policy improvement step based on these estimates may not lead to

an optimal policy. It is possible to converge on a sub-optimal policy pi_c that is better than the initial policy but still not the globally optimal policy.

3. **Explain how you can synthesize a stochastic policy given what you know so far (you don't need to read ahead).**

   We can use the epsilon greedy policy to synthesize a stochastic policy. Even though the majority of the time the policy selects the action with the highest value, there is still a non-zero probability of selecting other actions randomly. Since we have a small probability of exploration policy, we can discover potentially better actions or states that it may not have encountered otherwise.