# CS5180 Reinforcement Learning

## Exercise 6: Temporal-Difference Learning

Anway Shirgaonkar

## Q1 Q-learning vs. SARSA

(a) Q learning is indeed an off-policy method. Q learning selects the greedy action 100% of the time while updating the Q-values:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[R + \gamma \max_a Q(s',a) - Q(s,a)]$$

and follows a different $\epsilon$-soft policy which it uses as a behavior policy. Hence Q-learning is an off-policy method.

(b) Yes, if the SARSA algorithm uses a greedy action to update its Q values, then essentially it will behave the same way as that of Q learning.
Although the initialization of $Q(s,a) \in R$ and the behavior policy is not deterministic, but stochastic. Hence both algorithms may not exactly do the same weight updates.