Haoming Liang
1430396

Approximation are used not only on state value functions, but also can be used on action value functions. Episodic problems can directly apply the weight vector on the value functions, but continuing problems require to rewind backwards to review the previous discounting factors for calculating optimal policies.

We can now answer the question: Why doesn't approximation apply to Monte Carlo controls? I will think the answer is because Monte Carlo methods calculate returns and do updates at the end of episode, which goes against the idea that continuing problems rewind few time steps behind to calculate optimal policies.

Previously we consider an update target $S_t \rightarrow U_t$, where $U_t$ represent the update target approximating $v_\pi(S_t)$, now the update target is $S_t$, $A_t \rightarrow U_t$ where $U_t$ is now approximating $q_\pi(S_t, A_t)$. $U_t$ can be approximated in different methods, such as the back-up values from full Monte Carlo return, or the n-step Sarsa returns.

Turn the focus to problem settings, discounting settings has been proven useful in tabular methods, however approximating methods are different. As in the example described, averaging discounted returns over a time interval produces a result that's only proportional to the average reward, and also using a discount factor changes nothing about the policy compared to averaging the rewards. The use of discount factor in approximation is to be questioned. Alternatively we could use another setting which is average reward setting, without discounting factor. Average reward setting is useful for continuing tasks without start or end states.