CS5180 Reinforcement Learning

Exercise 7: Function Approximation

Anway Shirgaonkar

Q1 On-policy Monte-Carlo control with approximation

Monte Carlo method is an unbiased method and requires for the episodes to be rolled out. In this case,

$$w_{t+1} = w_t + \alpha[G_t - \hat{v}(s, w)]\nabla\hat{v}(s, w)$$

Since G_t is an unbiased estimate, w_t is guaranteed to converge under the stochastic approximation conditions for decreasing α.

I think they have not given the pseudocode for Monte Carlo method since it is trivial. As far as the performance on the mountain car task is concerned, I do not think it is a good idea to use Monte Carlo method, because the learning will only happen at the end of the episode.

It is highly unlikely that the episode will end soon for a random choice of action selection (initially it will be random even if we follow a ϵ-soft policy). Hence, the learning would take much longer. It is best to use bootstrapped methods for mountain car task.

Q2 Semi-gradient expected SARSA and Q-learning

(a) Pseudocode for semi-gradient one-step Expected Sarsa for control.

Input: a differentiable action value function parameterization $\hat{q}: S \times A \times \mathbb{R}^d \to \mathbb{R}$
Algorithm parameters: step size $\alpha > 0$, small $\epsilon > 0$
Initialize value function weights $\mathbf{w} \in \mathbb{R}^d$ arbitrarily
Loop for each episode:
    $S, A \leftarrow$ initial state and action of episode (ϵ greedy)
    Loop for each step of episode:
        Take action A, observe R, S'
        If S' is terminal:
            $w \leftarrow w + \alpha[R - \hat{q}(S, A, w)]\nabla\hat{q}(S, A, w)$
            Go to next episode
        $w \leftarrow w + \alpha[R + \gamma\left[\sum_{a\in A}\pi_{\epsilon-greedy}(a|S')\,\hat{q}(S', a, w)\right] - \hat{q}(S, A, w)]\nabla\hat{q}(S, A, w)$
        $\pi(a|S')$ can be determined with ϵ
        $A \leftarrow A'$
        $S \leftarrow S'$

(b) For semi gradient Q-learning,

$$w \leftarrow w + \alpha[R + \gamma \max_{a\in A}\hat{q}(S', a, w) - \hat{q}(S, A, w)]\nabla\hat{q}(S, A, w)$$

Q5 Residual-gradient TD

$$\overline{VE}(w) \triangleq \sum_{s \in S} \mu(s)[v_\pi(s) - \hat{v}(s,w)]^2$$

(a) Substituting $v_\pi(S)$ with $R + \gamma\hat{v}(S', w)$ in the above equation,

$$\overline{VE}(w) \triangleq \sum_{s \in S} \mu(s)[R + \gamma\hat{v}(s',w) - \hat{v}(s,w)]^2$$

Taking the gradient of $\hat{v}$ w.r.t w,

$$\nabla[R + \gamma\hat{v}(s',w) - \hat{v}(s,w)]^2 = 2[R + \gamma\hat{v}(s',w) - \hat{v}(s,w)][\gamma\nabla\hat{v}(s',w) - \nabla\hat{v}(s,w)]$$