Solutions to Reinforcement Learning by Sutton Chapter 10

Yifan Wang

Jan 2020

Exercise 10.1

There are few reasons of why book did not consider any Monte Carlo methods. First, the computational cost of Monte Carlo method is too high to make it practical in complex settings. In Mountain Car problem, Monte Carlo method will not start to learn unless it can complete the first full episode with a random policy. Even if it completes it, the amount of time steps would make the computation cost extremely high. On the other hand, SARSA and other TD methods can learn during progress and converge much faster to make it online.

Second, Monte Carlo method is just an extreme variation of more general group of TD methods. If we write down its pseudocode, it will not be so much different from n-step SARSA except $n = \mathrm{T}$.

■

Exercise 10.2

Repeat the pseudocode as in the box "Episodic Semi-gradient Sarsa for Estimating $\hat{q} \approx q_*$" on page 244. Then, replace the "$\gamma\hat{q}(S', A', w)$" in the 3rd line from bottom with "$\gamma \sum_a \pi(a|S')\hat{q}(S', a, w)$." Finally, swap this line with the line above. Though we can simplify a little bit more, it completes the on-policy Episodic Semi-gradient expected Sarsa.

■

Exercise 10.3

Larger n will tend to value longer history of rewards and thus amplify its variance effect in its first episodes. It is due to the strong dependence on its individual trajectory which usually is chaotic and random in its early stage. In another word, some 'strange' choices made during early steps will affect many other states before, and make the result less general. On the other hand, smaller n such as $n = 1$ only looks into next state S_{t+1}. Even with strange choices, the importance of its result S_{t+1} will be averaged out soon, since it only provides information to S_t.

However, when we have much longer episodes like in Figure 10.3 with 500 episodes, relatively large n will provide nicer result, since it looks into related local variations carefully (by a DP-like way) and thus avoids shaking around the convergence point.

■

Exercise 10.4

Recall the subtle difference between Q-learning and SARSA: Q-learning chooses the greedy action at each timing but SARSA provides the 'old guess' of its (possible) greedy action to the next step.

Thus, we can repeat the pseudocode in the box on page 251, but change the next few lines in the box as following:

Take action A, observe $R, S' \longrightarrow$
Choose A as a function of $\hat{q}(S', \cdot, \mathrm{w})$, observe R, S'.

Choose A' as a function of $\hat{q}(S', \cdot, \mathrm{w}) \longrightarrow$
Choose A' as $\arg\max_{A'} \hat{q}(S', \cdot, \mathrm{w})$.

A Fuller algorithm:

Input: a differentiable action-value function parameterization
$\hat{q} : \mathbb{S} \times \mathbb{A} \times \mathbb{R}^d \to \mathbb{R}$
Algorithm parameters: step sizes $\alpha, \beta > 0$
Initialize value-function weights $\mathrm{w} \in \mathrm{R}^d$ **arbitrarily (e.g.,** $\mathbf{w = 0}$**)**
Initialize average reward estimate $\bar{R} \in \mathbb{R}$ **arbitrarily (e.g.,** $\bar{R} = \mathbf{0}$**)**
Initialize state S (remove A initialization)
Loop for each step:

Choose A as a function of $\hat{q}$ (ϵ-greedy)
Take action A, observe R, S'
$\delta \leftarrow R - \bar{R} + \arg\max\, \hat{q}(s', a', w) - \hat{q}(s, a, w)$
$\bar{R} \leftarrow \bar{R} + \beta\delta$
$w \leftarrow w + \alpha\delta\nabla(\hat{q}(S, A))$
$S' \leftarrow S$

■

Exercise 10.5

Though I do not fully understand what this question is asking for, I expect to see the update formula for w**:**

$$\mathrm{w}_{\mathrm{t}+1} \doteq \mathrm{w}_t + \alpha\delta_t\nabla\hat{v}(S_t, \mathrm{w}_t)$$

Anyone who knows the meaning of this question please report as an issue, Thanks.

■

Exercise 10.6

The average reward, by (10.6), is 1/2.

$$
\begin{aligned}
v_\pi(s) &\doteq \lim_{\gamma\to 1}\lim_{h\to\infty}\sum_{t=0}^{h}\gamma^t(\mathbb{E}_\pi[R_{t+1}|S_0=s]-r(\pi)) \\
v_\pi(A) &= \lim_{\gamma\to 1}\lim_{h\to\infty}\sum_{t=0}^{h}\gamma^t(-\frac{1}{2})^t \\
&= \frac{1}{2}\lim_{\gamma\to 1}\lim_{h\to\infty}\sum_{t=0}^{h}(-\gamma)^t \\
&= \frac{1}{2}\lim_{\gamma\to 1}\lim_{h\to\infty}(1-\gamma)\frac{1-\gamma^h}{1-\gamma^2} \qquad \text{(summation of geometric series)} \\
&= \frac{1}{2}\lim_{\gamma\to 1}\frac{1}{1+\gamma} \\
&= \frac{1}{4} \\
v_\pi(B) &= -\gamma\lim_{\gamma\to 1}\lim_{h\to\infty}\sum_{t=0}^{h}\gamma^t(-\frac{1}{2})^t \\
&= -\frac{1}{4}
\end{aligned}
$$

■

Exercise 10.7

The average reward of policy is 1/3 based on definition of (10.6). Because the system is violating ergodicity, bellman function cannot be used.

For anyone who cannot be convinced by it, I implemented the SARSA algorithm here to check out. It shows average reward is indeed convergent to 1/3, and delta is converging to 0 but the value of state is changing with each run. Thus, the differential values are not well defined in this situation.

$$v_\pi(s) \doteq \lim_{\gamma \to 1} \lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t (\mathbb{E}_\pi[R_{t+1}|S_0 = s] - r(\pi))$$

$$\begin{aligned}
v_\pi(A) &= \lim_{\gamma \to 1} \lim_{h \to \infty} \big[\sum_{t=0}^{h} (-\frac{1}{3}\gamma^{3t}) + \gamma \sum_{t=0}^{h} (-\frac{1}{3}\gamma^{3t}) + \gamma^2 \sum_{t=0}^{h} (\frac{2}{3}\gamma^{3t}) \big] \\
&= \lim_{\gamma \to 1} \lim_{h \to \infty} \big[(-\frac{1}{3} - \frac{1}{3}\gamma + \frac{2}{3}\gamma^2) \frac{1-\gamma^{3h}}{1-\gamma^3} \big] \\
&= \lim_{\gamma \to 1} \big[(-\frac{1}{3} - \frac{1}{3}\gamma + \frac{2}{3}\gamma^2) \frac{1}{1-\gamma^3} \big] \\
&= \lim_{\gamma \to 1} \big[-\frac{2\gamma + 1}{3(\gamma^2 + \gamma + 1)} \big] \\
&= -\frac{1}{3}
\end{aligned}$$

$$\begin{aligned}
v_\pi(B) &= \lim_{\gamma \to 1} \lim_{h \to \infty} \big[\sum_{t=0}^{h} (-\frac{1}{3}\gamma^{3t}) + \gamma \sum_{t=0}^{h} (\frac{2}{3}\gamma^{3t}) + \gamma^2 \sum_{t=0}^{h} (-\frac{1}{3}\gamma^{3t}) \big] \\
&= \lim_{\gamma \to 1} \lim_{h \to \infty} \big[(-\frac{1}{3} + \frac{2}{3}\gamma - \frac{1}{3}\gamma^2) \frac{1-\gamma^{3h}}{1-\gamma^3} \big] \\
&= \lim_{\gamma \to 1} \big[(-\frac{1}{3} + \frac{2}{3}\gamma - \frac{1}{3}\gamma^2) \frac{1}{1-\gamma^3} \big] \\
&= \lim_{\gamma \to 1} \big[-\frac{\gamma - 1}{3(\gamma^2 + \gamma + 1)} \big] \\
&= 0
\end{aligned}$$

$$
\begin{aligned}
v_\pi(C) &= \lim_{\gamma\to 1}\lim_{h\to\infty}\big[\sum_{t=0}^{h}(\frac{2}{3}\gamma^{3t}) + \gamma\sum_{t=0}^{h}(-\frac{1}{3}\gamma^{3t}) + \gamma^2\sum_{t=0}^{h}(-\frac{1}{3}\gamma^{3t})\big] \\
&= \lim_{\gamma\to 1}\lim_{h\to\infty}\big[(\frac{2}{3} - \frac{1}{3}\gamma - \frac{1}{3}\gamma^2)\frac{1-\gamma^{3h}}{1-\gamma^3}\big] \\
&= \lim_{\gamma\to 1}\big[(\frac{2}{3} - \frac{1}{3}\gamma - \frac{1}{3}\gamma^2)\frac{1}{1-\gamma^3}\big] \\
&= \lim_{\gamma\to 1}\big[-\frac{\gamma+2}{3(\gamma^2+\gamma+1)}\big] \\
&= \frac{1}{3}
\end{aligned}
$$

■

Exercise 10.8

The sequence of $R_t - \bar{R}_t$ would be (from state C):

$$\frac{2}{3}, \quad -\frac{1}{3}, \quad -\frac{1}{3}, \ldots \text{(repeat the same pattern)}$$

However, the sequence of δ_t by (10.10) would be (from state C):

$$\frac{2}{3} + 0 - \frac{2}{3} = 0, \quad -\frac{1}{3} + \frac{1}{3} - 0 = 0, \quad -\frac{1}{3} + \frac{2}{3} - \frac{1}{3} = 0, \quad \ldots$$

We observe $\delta_t = 0$ will result no update because we have already have accurate estimate of $\hat{v}$ and $\bar{R}$. This method will give us more stable estimate since it will automatically decrease the speed of oscillation when we are close to convergence and increase the speed of change (by amount $\hat{v}(S_{t+1}) - \hat{v}(S_t)$) when we are in the opposite situation.

■

Exercise 10.9

Replacing all β with β_n. Adding following lines before δ get updated:

$$\beta_n \doteq \beta/\bar{o}_n,$$
$$\bar{o}_n \doteq \bar{o}_{n-1} + \beta(1 - \bar{o}_{n-1}), \text{for } n \geq 0, \text{ with } \bar{o}_0 \doteq 0.$$

■