Reinforcement Learning - Homework 5

Hadi Askari

May 13 2023

1 Implementation

I have implemented all 3 algorithms, Q-Learning, SARSA and Expected SARSA. The code runs under a minute and plots all three lines for the learning rate of 0.5 as described in the book. The details of the algorithms are as follows:

1.1 Q Learning

This algorithm is defined in the function q_learning.

Initialize $Q(s,a)$ arbitrarily
Repeat (for each episode):
  Initialize s
  Repeat (for each step of episode):
    Choose a from s using policy derived from Q
    Take action a, observe r, s'
    Update
      $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$
    $s \leftarrow s'$;
  Until s is terminal

(a) Q-Learning Algorithm

1.2 SARSA

This algorithm is defined in the function sarsa.

Sarsa (on-policy TD control) for estimating $Q \approx q_*$

Initialize $Q(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily, and $Q(\textit{terminal-state}, \cdot) = 0$
Repeat (for each episode):
  Initialize S
  Choose A from S using policy derived from Q (e.g., ϵ-greedy)
  Repeat (for each step of episode):
    Take action A, observe R, S'
    Choose A' from S' using policy derived from Q (e.g., ϵ-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma Q(S',A') - Q(S,A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until S is terminal

(a) SARSA Algorithm

1.3 Expected SARSA

This algorithm is defined in the function expected sarsa.

Additionally I also defined the gridworld environment with the specified bounds and rewards according to the book in the function gridworld.

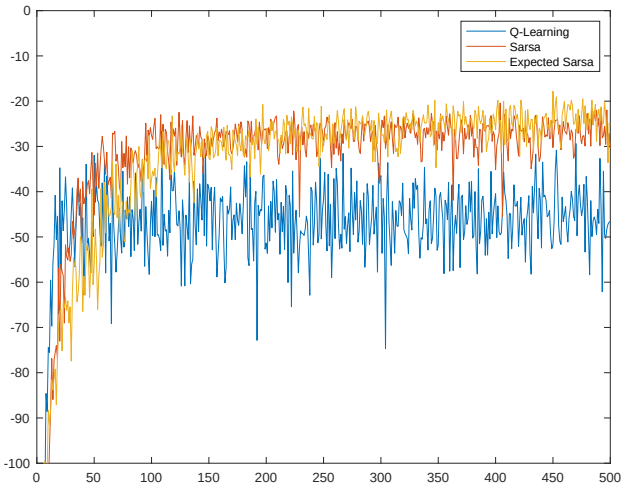
1. Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
2. Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(\text{ terminal}, \cdot) = 0$

3. Loop for each episode:
4. Initialize S
5. Loop for each step of episode:
6. Choose A from S using policy derived from Q (e.g., ε-greedy)
7. Take action A, observe R, S'
8. $Q(S, A) \leftarrow Q(S, A) + \alpha \left[R + \gamma \sum_a \pi \left(a \mid S_{t+1}\right) Q\left(S_{t+1}, a\right) - Q(S, A)\right]$
9. $S \leftarrow S'$
10. until S is terminal

(a) Expected SARSA

2 Results

The following is the plot I generated combining all three methods and plotting them on the same plot. Each line is averaged from a 100 different iterations. Each episode runs for 500 iterations. The learning rate was kept constant at 0.5.



(a) $\alpha = 0.5$