CSC 449 Advanced Topics in Artificial Intelligence Exam 2

90

Christian Olson

November 2022

1. **procedure** SARSA(number of episodes $N \in \mathbb{N}$, discount factor $\gamma \in (0,1]$[(1)], learning rate $\alpha_n = \frac{1}{log(n+1)}$, exploration rate $\epsilon > 0$ but small[(2)])
   Initialize matrices $Q(s,a)$ and $n(s,a)$ to 0, $\forall s, a$
   for episode $k \in 1, 2, 3, ..., n$ **do**
   $t \leftarrow 1$
   Initialize s_1
   Choose a_t from s_t using π_t: an ϵ-greedy policy with respect to Q[(3)]
   while Episode k is not finished **do**
   Take action a_t: observe reward r_t and next state s_{t+1}
   Choose a_{t+1} from s_{t+1} using $\pi_t^(4)$: an ϵ-greedy policy with respect to Q
   $y_t^{(5)(6)(7)} = r_t + \gamma Q(s_t + 1, a_t + 1)$
   $n(s_t, a_t) \leftarrow n(s_t, a_t) + 1$
   $Q(s_{t+1}, a_{t+1}) \leftarrow Q(s_t, a_t) + \alpha_{n(s_t,a_t)}(y_t - Q(s_t, a_t))$
   $t \leftarrow t + 1$
   end while
   end for
   end procedure

(1) The discount factor is typically represented by γ not λ.

(2) The exploration rate ϵ should be initialized as a parameter.

(3) The policy for choosing a_1 should be the same ϵ-greedy policy .

(4) The policy is usually typically represented by π not μ

(5) **if** The current state is terminal **then**$y_t = 0$
end if
is not necessary as the terminal state is caught by
while Episode k is not finished **do**
end while
and an update with $y_t = 0$ would update a state that is never actually visited.

and should not have max

(6) The update did not consider the discount factor γ.

(7) The update reward y_t was the *Q-learning* update $y_t = r_t + \max_a Q(s_{t+1}, a$ not the SARSA update $y_t = r_t + \gamma Q(s_{t+1}, a_{t+1})$.

2. a Expected SARSA removes effects of random action selection from Q-learning by considering the average of all possible next states. Like SARSA, this update learns the only the optimal policy making it on-policy. However, if the policy is is greedy and explores adding randomness back into the action selection, Expected SARSA inherits the behavior of Q-learning, learning from a policy other than the optimal policy (off-policy). For Expected SARSA to remain on-policy, the set of policies must be non-greedy policies which explore only according to the a non-random policy.

b One of the criteria for convergence is that all state action pairs are visited while learning. With a stationary policy, only one action will every be chosen at any one state. Therefore, SARSA will not explore the entire pairwise state-action space and will not converge as a result.

unless the environment is non-deterministic...