CSC 449 Advanced Topics in Artificial Intelligence

Deep Reinforcement Learning
Exam 1
Fall, 2022

Name: ______________________________ ID#: __________________ Score: ______

Your solutions to these problems should be uploaded to D2L as a single pdf file by the deadline. As with problem sets, you may turn in the solution up to two days late, with a penalty of 10% per day, and you should only upload one version of your solutions.

This exam is individual and open book. You may consult any reference work. If you make specific use of a reference outside those on the course web page in solving a problem, include a citation to that reference.

You may discuss the course material in general with other students, but you should work on the solutions to the problems on your own.

It is difficult to write questions in which every possibility is taken into account. As a result, there may sometimes be "trick" answers that are simple and avoid addressing the intended problem. Such trick answers will not receive credit. As an example, suppose we said, use the chain rule to compute $\frac{\partial z}{\partial x}$ with $z = \frac{7}{y}$ and $y = x^2$. A trick answer would be to say that the partial deriviative is not well defined because y might equal 0. A correct answer might note this, but would then give the correct partial derivative when $y \neq 0$.

1. (30 pts) Consider the following pseudo-code for a faulty SARSA algorithm:

```
procedure SARSA( number of episodes $N \in \mathbb{N}$
                 discount factor $\lambda \in (0,1]$
                 learning rate $\alpha_n = \frac{1}{\log(n+1)}$ )
    Initialize matrices $Q(s,a)$ and $n(s,a)$ to $0, \forall s,a$
    for episode $k \in 1,2,3,\dots,n$ do
        $t \leftarrow 1$
        Initialize $s_1$
        Choose $a_1$ from a uniform distribution over the actions
        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 $\mu_t$: an $\varepsilon$-greedy policy with respect to $Q$
            if The current state is terminal then                ▷ Compute target value
                $$y_t = 0$$
            else
                $$y_t = r_t + \max_a Q(s_{t+1}, a)$$
            end if
            $n(s_t, a_t) \leftarrow n(s_t, a_t) + 1$
            Update Q function:
                $$Q(s_{t+1}, a_{t+1}) \leftarrow Q(s_t, a_t) - \alpha_{n(s_t,a_t)} (y_t - Q(s_t, a_t))$$
            $t \leftarrow t + 1$
        end while
    end for
end procedure
```

Find all of the mistakes in the algorithm. Explain why they are mistakes, and correct them.