CPSC 422 — Intelligent Systems — Spring 2005
Assignment 3

Due: 2:00pm, Tuesday 22 February 2005.

The aim of this assignment is to learn about reinforcement learning. Note that the source is available for the reinforcement learning applets at:

```
http://www.cs.ubc.ca/spider/poole/demos/rl/q.html
http://www.cs.ubc.ca/spider/poole/demos/rl/qlambda.html
```

Please read and post to the bulletin board in the course WebCT site.

You can either do this assignment in groups of size one to two. There will be questions on the midterm about this assignment, so it is important that everyone in the group understands all of the solution.

Question 1

Consider four different ways to derive the value of α_k from k in Q-learning (note that in Q-learning there is a different k for each state-action pair).

i) Let $\alpha_k = 1/k$.

ii) Let $\alpha_k = 10/(9+k)$.

iii) Let $\alpha_k = 0.1$.

iv) Let $\alpha_k = 0.1$ for the first 10,000 steps, $\alpha_k = 0.01$ for the next 10,000 steps, $\alpha_k = 0.001$ for the next 10,000 steps, $\alpha_k = 0.0001$ for the next 10,000 steps, etc.

(a) Which of these will converge to the optimal Q-value in theory?

(b) Which converge in practice? Try it for the grid world. What would you expect to happen with other domains? [Explain how the example could be changed so that the methods break if they do break.]

(c) Which can adapt when the environment changes slowly?

Question 2

If you are working on a group of size 1, do either part (a) or part (b). If you are working in group of size 2, do both parts.

(a) Change the applet at

```
http://www.cs.ubc.ca/spider/poole/demos/rl/qlambda.html
```

So that it does $SARSA(\lambda)$ instead of $Q(\lambda)$. The only difference is that to compute the estimated value of the next state s', $Q(\lambda)$ uses $\max_{a'} Q(s', a')$, but SARSA uses which action the agent will do next using it's current policy. In particular, it before updating $Q[s, a]$ it determines its next action, a', and uses $Q[s', a']$ as the value for the next state.

Does this work better or worse that $Q\lambda)$?

(b) Consider an example where there are two states A, B. There is a reward of 10 coming into state A and no other rewards or penalties. There are two actions: left and right. These actions only make a difference in state B. Going left in state B goes directly to state A, but going right has a low probability of going into state A.

- $P(A|B, \mathit{left}) = 1$; reward is 10
- $P(A|B, \mathit{right}) = 0.01$; reward is 10. $P(B|B, \mathit{right}) = 0.99$; reward is 0
- $P(A|A, \mathit{left}) = P(A|A, \mathit{right}) = 0.999$ and $P(B|A, \mathit{left}) = P(B|A, \mathit{right}) = 0.001$. This is small enough so that the eligibility traces will be close enough to zero when state B is entered.
- γ and λ are 0.9 and α is 0.4

Suppose that your friend claimed that that $Q(\lambda)$ doesn't work in this example, as the eligibility trace for the action *right* in state B ends up being bigger than the eligibility trace for action *left* in state B and the rewards and all of the parameters are the same. In particular the eligibility trace for action *right* will be about 5 when it ends up entering state A, but it be one for action *left*. So that the best action will be go right in state B. Which is not correct.

What is wrong with your friend's argument? What does this example show?

Question 3

[Note that this question is worth marks, so don't forget to do it.]

(a) For each question in this assignment, say how long you spent on it. Was this reasonable? What did you learn?

(b) If there was more than one person in your group, say what each person did.

(c) Tell us every person you discussed this with, and every external source (e.g., web site, research paper, book) you referenced to do this assignment.