CS 4100/5100: Foundations of Artificial Intelligence (Fall 2015) Robert Platt
Student(s): Oct. 22, 2015

Reinforcement Learning

1 Odds and Ends

1. Can all MDPs be solved using expectimax search? If your answer is yes prove it, and if it is no provide a counterexample.

2. When learning with ϵ-greedy action selection, is it a good idea to decrease ϵ to 0 with time? Why or why not?

3. When using features to represent the Q-function is it guaranteed that the feature-based Q-learning finds the same optimal Q^* as would be found when using a tabular representation for the Q-function?

4. Does the temporal difference learning of optimal utility values (U) require knowledge of the transition probability tables? Why or why not?

5. Why is temporal difference (TD) learning of Q-values (Q-learning) superior to TD learning of values?

2 Learning with Feature-based Representations

We would like to use a Q-learning agent for Pacman, but the state size for a large grid is too massive to hold in memory. To solve this, we will switch to feature-based representation of Pacmans state.

1. Let's assume our two minimal features are the number of ghosts within 1 step of Pacman (F_g) and the number of food pellets within 1 step of Pacman (F_p). Calculate F_p and F_g for the following pacman board:

2. With Q Learning, we train off of a few episodes, so our weights begin to take on values. Right now $w_g = 70$ and $w_p = 5$. Calculate the Q value for the state above

3. We receive an episode, so now we need to update our values. An episode consists of a start state s, an action a, an end state s', and a reward R(s, a, s'). The start state of the episode is the state above (where you already calculated the feature values and the expected Q value). The next state has feature values $F_g = 0$ and $F_p = 4$ and the reward is 50. Assuming a discount of 0.4, calculate the new estimate of the Q value for s based on this episode.

4. With this new estimate and a learning rate (α) of 0.5, update the weights for each feature.

5. What values do we learn in this process (assuming features are defined)? When we have completed learning, how do we tell if Pacman does a good job?

6. In some sense, we can think about this entire process, on a meta level, as an input we control that produces an output that we would like to maximize. If you have a magical function (F(input)) that maps an input to an output you would like to maximize, what techniques can we use to search for the best inputs? Keep in mind that the magical function is a black box.