Reading Assignment Three

Amol Kapoor, Adam Cobb
Worcester College, Oxford University

June 4, 2017

1. **Question 6.5: Above we stated that the true values for the random walk task are $\frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}$, for states A through E. Describe at least two different ways that these could have been computed. Which would you guess we actually used? Why?**

   (a) : Analytically. The likelihood of going left or right is the same, thus the probability of ending on the far left or the far right is only dependent on the actual distance from a given state. Since there are a maximum of six transitions made to get from one non-terminal state to a terminal state, we can calculate all values by sixths. Specifically, the value of the state can therefore be defined as:

   $$\frac{N-d}{N}$$

   where N is the maximum number of transitions (6) and d is the distance to the far right. Note that this generalization only works when the probability of going left or right is the same, as one can imagine that the likelihoods of doubling back 'cancel' and so can be discounted.

   (b) : As a binomial distribution. This problem can be seen as a binomial distribution, or the number of times we succeed (in this case, go right) given N trials. This can also be interpreted as or a chain of bernoulli random variables each with some probability p of going left or right. The distribution for the binomial random variable is defined as follows:

$$\binom{N}{K} p^k (1-p)^{(N-K)}$$

where N is the total number of steps taken, K is the number of steps taken in a specific direction, and p is the Bernoulli likelihood of going in a specific direction (in this case, 0.5). To calculate the value of a state, we have to find the likelihood of an agent in that state reaching the far right terminal for every N value from 1 to infinity (or some arbitrarily large N where the numbers begin to converge). This is obviously much more tedious than the analytic solution.

2. **Question 6.6: Re-solve the windy gridworld task assuming eight possible actions, including the diagonal moves, rather than the usual four. How much better can you do with the extra actions? Can you do even better by including a ninth action that causes no movement at all other than that caused by the wind?**

   One can finish windy grid-world with 7 steps - 3 in the SW direction, followed by 4 in the W direction (three of which end up being NW due to the wind). The extra 9th action does not actually decrease the optimal solution, as 7 is the minimum and shortest distance between the start state S and the end state G.

3. **Question 6.9: Why is Q-learning considered an off-policy control method?**

   The definition of an off-policy control method is an algorithm that does not estimate the value function of the policy that generates the data. The Q learning algorithm estimates Q* from Q without taking into account the actual policy - rather, in the Q update step it updates each state value *as if* the policy being followed was greedy to the Q values. Thus the actual policy being followed can be anything, and the Q learning algorithm will still learn the optimal q*.