CS 8803

Midterm Project

Replicating published results

Jacob Kilver (jkilver3@gatech.edu)

2-21-2016

1 Introduction

Reproducing results published in scientific journals is an important part of taking part in the research community. This project attempts to reproduce the results seen in Figures 3, 4, and 5 from Sutton (1988).

2 Description of experiments

Figures 3-5 from Sutton (1988) deal with the bounded random walk problem. Figure 1 shows the state diagram for this Markov Decision Process (MDP). Beginning in the center state D, each state has a single action, which 50% of the time results in a transition to the left and 50% of the time results in a transition to the right. Thus, states A and G are terminal states. The reward function is zero everywhere except for the transition from state F to G, which has a reward of 1.0.

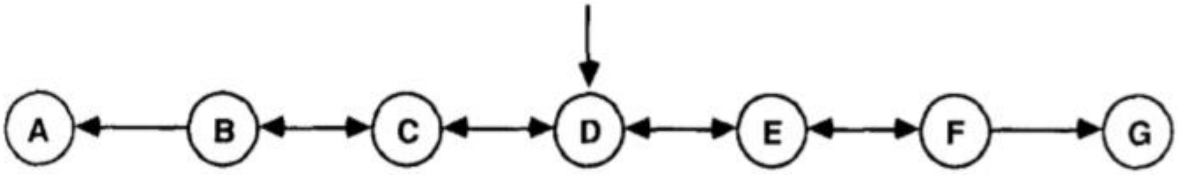

Figure 1: Bounded Random Walk State diagram

The TD Lambda algorithm was used to find the expected value of states B through F in this MDP. Two experiments were conducted, each with trainings sets of size 100, with each training set containing 10 sequences (10 walks through the MDP). The first experiment dealt with the “repeated presentation” paradigm – each training set was presented to the algorithm repeatedly until the state values converged. The second experiment dealt with how the learning rate effects error when each training set is presented only once.

3 Implementation details

BURLAP (http://burlap.cs.brown.edu/) was used to replicate these results. This tool was chosen so as to not duplicate existing work. The TD Lambda algorithm is readily available in this software package, along with tools to implement MDPs represented as graphs. Unfortunately using a pre-existing tool came with its own problems, which are detailed below.

While creating a training set for testing was rather straightforward, replaying that training set proved difficult. Specifically, when iterating over <state, action, next state> tuples, the state values (or weights as in Sutton) were not updated. This despite the set up being exactly the same as what was used in other places throughout the BURLAP library.

Since iterating over the training data was not successful, a different method was to generate the data used in these experiments. This method did not allow for the specific <state, action, next state> tuples to be used, so the <state, action, next state> tuples that were uses were generated randomly. This shortcoming limited the ability with which the results of Sutton could be recreated, most notably for the first experiment. Essentially, for each of the experiments above, the results were averaged over 100 training sets after the state values were calculated using the `planFromState()` method within BURLAP.

4 Results and Discussion

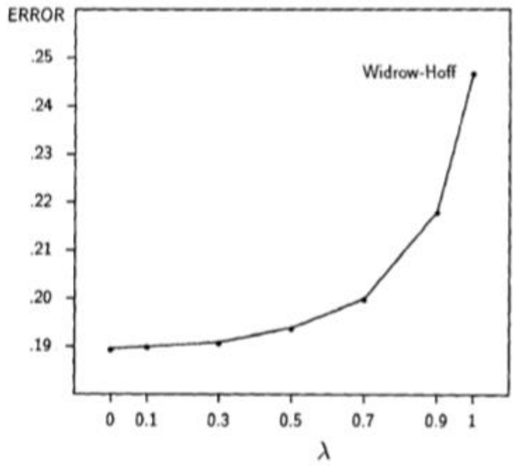

Figure 2: Results from Sutton (1988, Figure 3)

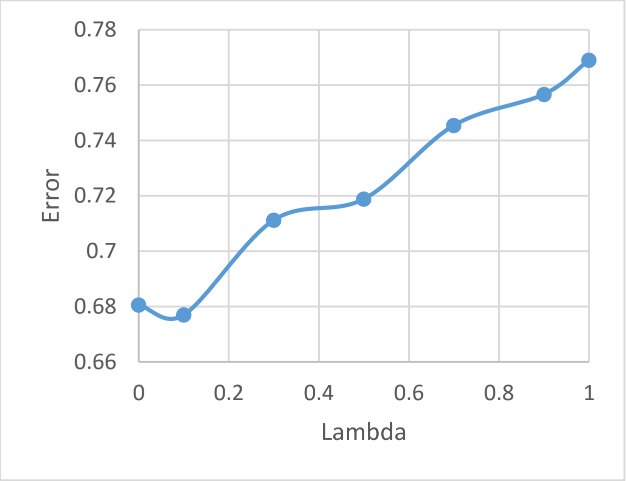

Figure 3: Reproduction of results from Sutton (1988, Figure 3)

Figure 2 and Figure 3 above present the results from Sutton (1988) and those from the reproduction attempted in this paper for the first experiment involving repeated presentations. Since the attempts to reproduce the results here were not able to replay specific training sets, the results from Sutton could not be reproduced exactly. Each new learning episode had no guarantee to be the same as any of the ones before, so potentially each new learning episode was unique, unlike in Sutton. Nevertheless, we see that the same general trend from lower accuracy using TD(0) to higher accuracy using TD(1) remains. While the lowest error was with TD(0.1), the value is close enough to TD(0) to be caused by the randomness that was unfortunately associated with this experiment. Additionally, the values for discount factor and learning rate would not have had an effect on these results since training sets were presented until convergence. These values would have only affected the rate of convergence, not the final value.

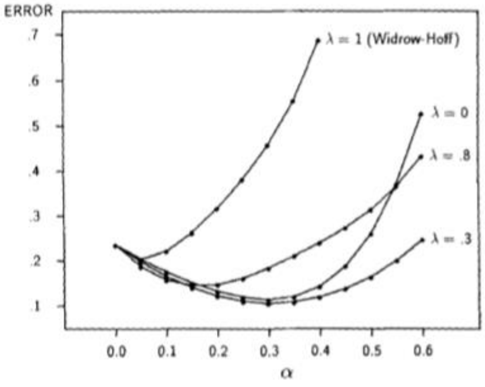

Figure 4: Results from Sutton (1988, Figure 4)

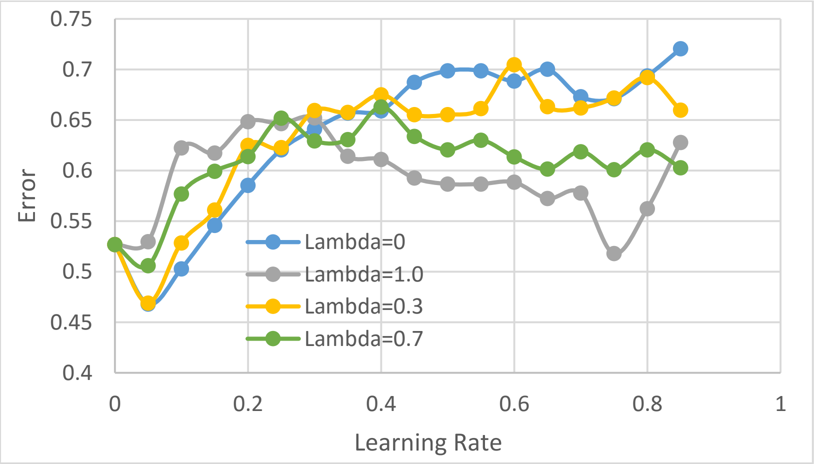

Figure 5: Reproduction of results from Sutton (1988, Figure 4)

Figure 4 and Figure 5 above present the results from the second experiment. It was hypothesized that the reproduced results from this experiment would be closer to those from Sutton (1988) than between those of Figure 2 and Figure 3. This was because in this experiment each training set was presented exactly once, so the uniqueness of the training set would have less effect on the results. This does not appear to be the case. As can be seen, the results differ significantly. While there is still the general trend from a starting error value to a lower error to an even higher error, they are not parabolic in form. Furthermore, the minimum error appears to be closer to 0.05 rather than 0.3 as in Sutton.

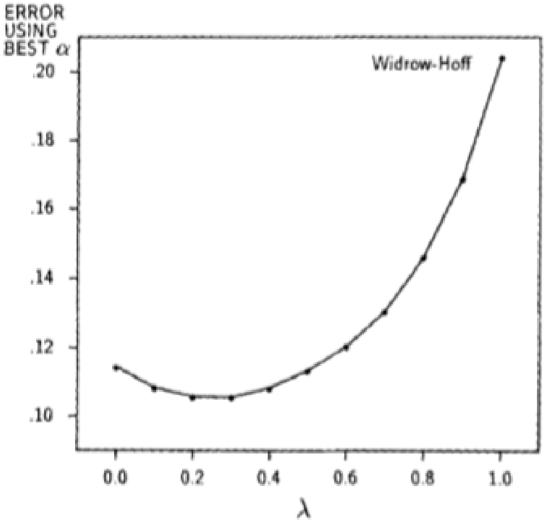

Figure 6: Results from Sutton (1988, Figure 5)

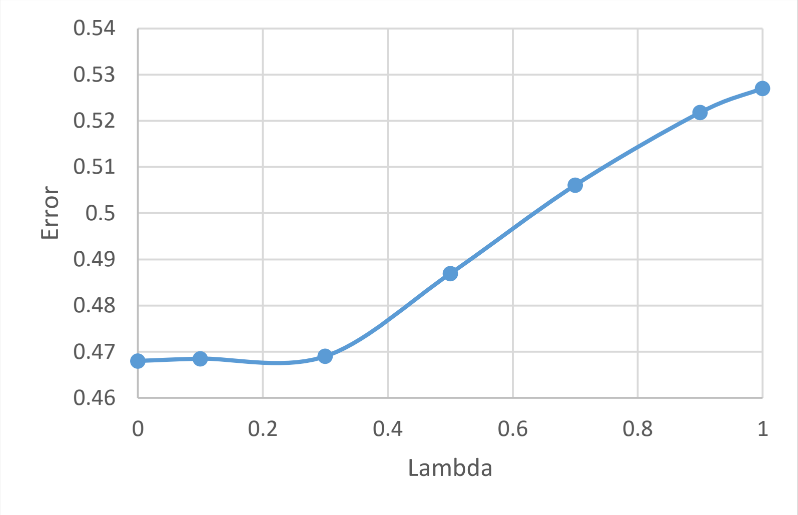

Figure 7: Reproduction of results from Sutton (1988, Figure 5)

Finally, Figure 6 and Figure 7 present the results of the second experiment but in a different format. Using the best learning rate from the second experiment, the total error is plotted for each value of λ. Here it can be seen that some of the results are confirmed. Once again, TD(1) had the worst error. Additionally, TD(0.3) has one of the lowest error, but then errors for all values of λ < 0.5 are about the same. The results reproduced in Figure 7 do align with those observed in Figure 3. In this case, it seems that BURLAP's implementation of TD(0) was able to propagate state information back faster than Sutton's implementation,, hence the results are very similar.

Sutton explores how learning rate and λ affect learning, but one parameter that is not explored is that of discount factor γ. In experiment 1 the value of γ does not affect results significantly since the training sets are presented repeatedly. The value for γ would only affect how quickly the values converge. For the second experiment when training sets are only presented once, it would have an effect on the state values output by the algorithm. However, assuming the value for γ is the same for each training set, the general shape of the graph should be preserved, although it may be shifted up or down. For these reasons Sutton most likely did not feel the need to specify these values in his results.

5 Conclusion and Future Work

These attempts to reproduce the results from Sutton (1988) were able to confirm several of his findings, but not all. From the first experiment, it was shown that TD(0) had the best estimate of the state values for the bounded random walk problem while TD(1) had the worst. Even when using the best learning rate it was shown that TD(1) had the worst prediction of the state values. However, several results were not confirmed or even found to be different. First, the best learning rate was found to be closer to 0.05 rather than 0.3 as in Sutton. Furthermore, even with single presentations of training sets, TD(0) was found to have the lowest error when predicting the values of the states. However, it should be noted that these results are known to be flawed. Using static training sets was unable to be fully implemented so each training set had the possibility of being unique. These results should be further confirmed using experiments truly identical to Sutton either using BURLAP or another tool such as the MDPToolbox for Python.