CS 7624 Project 1 – Desperately Seeking Sutton

Arjun Chintapalli[1]
[1] Department of Computer Science, Georgia Institute of Technology, arjun.ch@gatech.edu
Github Repo: https://github.gatech.edu/achintapalli3/CS7642_Project1

ABSTRACT

This paper attempts to verify the conclusions achieved through Richard Sutton's eminent paper "Learning to Predict by the Method of Temporal Differences" (1988) [1] by reproducing his results. Sutton presents a relatively new class of incremental learners based on the difference in temporally successive predictions that had efficiency, accuracy, and computational advantages over conventional learning algorithms that only looked at the difference between predictions and outcomes. To verify Sutton's findings this report reproduces Suttons figures of RMSE Error vs Lambda after repeatedly presenting a 100 training sets each of 10 sequences, RMSE Error vs Alpha for various Lambdas after single presentation of a 100 training sets each of 10 sequences, and RMSE Error vs Lambda after single presentation of a 100 training sets each of 10 sequences.

Key Words: CS 7642, Project 1, Temporal Difference Learning, Sutton

1. INTRODUCTION TO TD(λ)

Temporal Difference Learning are a class of reinforcement learning algorithms, which learn from the differences in successive predictions as opposed to just the difference between predictions and the outcomes. This incremental nature gives Temporal Difference (TD) Learning Methods advantages with respect to information efficiency, memory utilization and computational time. Instead of simply comparing predictions to outcomes to learn the best predictor, TD methods can squeeze out more information by also comparing successive predictions in the learning process. The incremental nature of TD methods in that they only require pairs of successive predictions as opposed to the full history of predictions leads to lower memory and system requirements for these methods. Since calculations are also incremental and spread out, there are fewer computational bottlenecks.

TD Methods learn from successive predictions by calculating an incremental change in weights or predictors:

$$\Delta w_{\mathrm{t}} = \alpha(P_{t+1} - P_t)\sum_{k=1}^{t}\lambda^{t-k}\nabla_w P_k \tag{1}$$

Where Δw_{t} is the increment change in weight at time t, α is the learning rate parameter, P_t is the prediction at time t, λ is the parameter that weights with respect to recency such that $\lambda = 0$ equates to the weight increment being determined by only the most recent prediction and that $\lambda = 1$ equates to an equal weighting of all predictions, and $\nabla_w P_k$ is the vector of partial derivatives of P_t with respect to each component of w.

After all the Δw_{t}'s for a complete sequence have been calculated, they can be summed up to calculate a final predictor, w, as follows:

$$w \leftarrow w + \sum_{t=1}^{m}\Delta w_{\mathrm{t}} \tag{2}$$

2. INTRODUCTION TO BOUNDED RANDOM WALKS

To validate Sutton's methodology and conclusions, the TD Lambda Methods were tested on a Bounded Random Walk because it is a simple closed system that still involves a series of evolving predictions. The specific case explored by Sutton has states: [A,B,C,D,E,F,G], where the walk always starts at state D, has an equal chance of going either left or right, and ends at entering either state A (giving outcome 0) or G (giving outcome 1). The TD methods were implemented to predict the probability of a random walk ending in state G based on the current state (possibly being states: [B,C,D,E,F]). The final predictions can then be compared with the ideal predictions of [1/6,1/3,1/2,2/3,5/6] for states [B,C,D,E,F] respectively utilizing the root mean square error of the difference between the ideal predictions and the experimental predictions.

Since the Bounded Random Walk case has the predictors being a linear function of state x_{t}, the prior incremental weight change equation simplifies to:

$$\Delta w_{\mathrm{t}} = \alpha(w^T x_{t+1} - w^T x_t) \sum_{k=1}^{t} \lambda^{t-k} x_k \tag{3}$$

3. EXPERIMENT 1

For the first experiment, Sutton's figure 3 showing RMSE Error vs Lambda after repeatedly presenting a 100 training sets each of 10 sequences was replicated. Each training set was presented repeatedly till convergence was established and the average RMSE error of the 100 training sets was plotted against varied λ's. For this particular experiment, the weights were only updated after a training set was presented.

For the purposes of this experiment, the training sets were created once and repeated for the different Lambda trials so to reduce randomness. For each Lambda value, a multitude of alpha values were tried such that the one giving the smallest RMSE error was chosen. Each training set was repeatedly presented till there was no noticeable change in the sum of incremental changes.

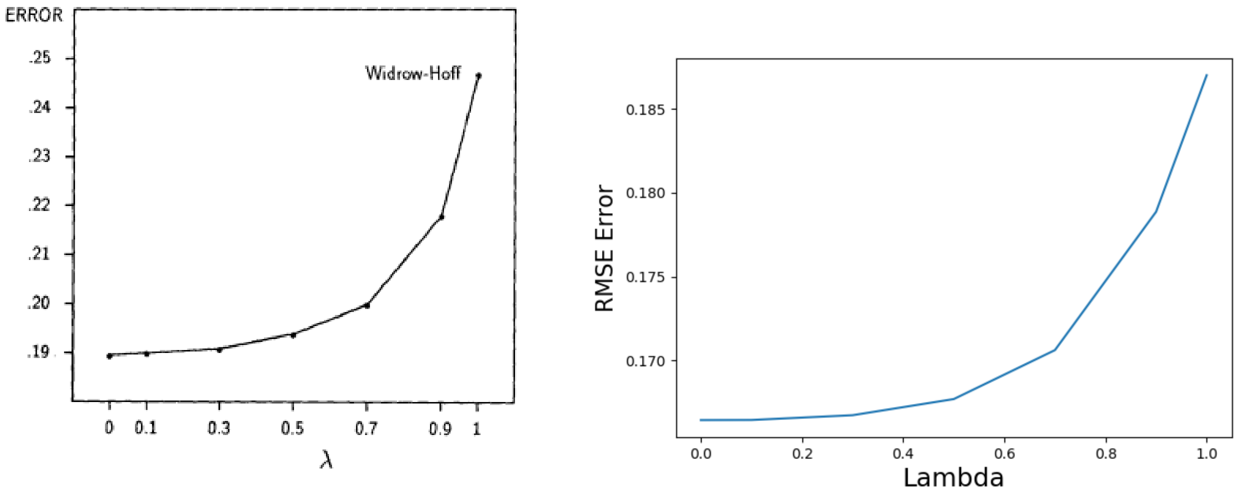

Figure 1: RMSE Error vs Lambda [Sutton on Left and Experimental on Right]

This experiment gives results in line with Sutton's and further fine tuning of the learning parameter, α, would give a graph with RMSE errors even closer to Sutton's results. These results further demonstrate the improved performance given by TD methods vis-à-vis supervised learning methods as shown by the spike in error at TD(1), which represents supervised methods. Additionally, the convergence criteria was set such that after a presentation of the training set leads to a $\sum_{t=1}^{m} abs(\Delta w_t)$ of less than .0001, the training set is not further presented. If Sutton's exact convergence criteria was implemented instead, the graph and errors could further resemble Sutton's exact results. Overall the errors are smaller than Sutton's results at every λ. The convergence criteria used for this experiment could have been too stringent leading to errors below what Sutton reported.

4. EXPERIMENT 2

In contrast to the prior experiment, the training set was only presented once, the weights are updated immediately after each sequence instead of after a training set, the learning rate was varied per λ and finally initial weights of all .5 were used. The plots are also qualitatively different in that they are now multiple overlaid curves of RMSE Error vs α for varied λ's.

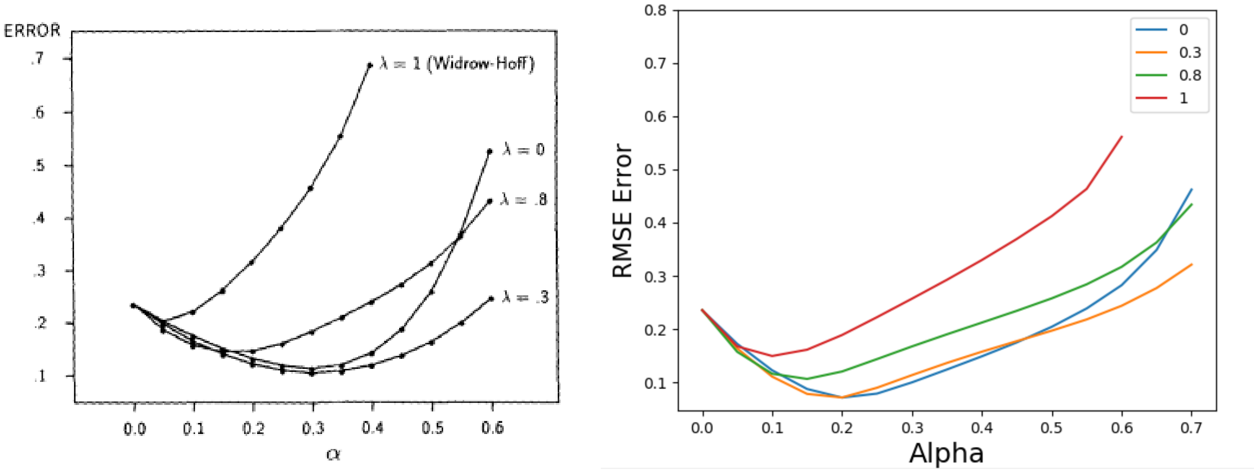

Figure 2: RMSE Error vs Alpha [Sutton on Left and Experimental on Right]

The experimental results are strikingly similar to Sutton's results. Similarly to Sutton's resuts, the learning rate had a tremendous effect on the error. Also, TD(1) (equivalent to conventional supervised methods) showed the worst results, further validating the benefits of using TD Methods. The cross-over of TD(0) to have the second highest error also occurs in the experimental results. As Sutton explained this is due to the fact that TD(0) is poor at propagating predictions back along a sequence because TD(0) equates to the weight increment being determined by only the most recent state. This wasn't an issue when training sets were repeatedly presented till convergence as in Experiment 1.

4. EXPERIMENT 3

Experiment 3 was the same as Experiment 2 except that the error for each λ was plotted directly using the best α.

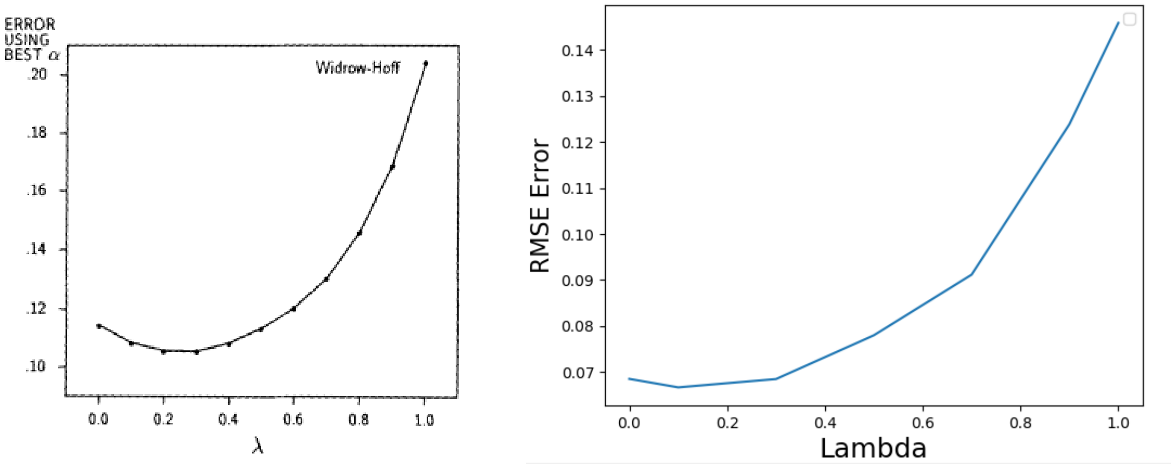

Figure 3: RMSE Error (Using Best α) vs Lambda [Sutton on Left and Experimental on Right]

Again, the experimental results are almost the same as Sutton's results albeit with reduced errors. Also, TD(1) (equivalent to conventional supervised methods) showed the worst results, further validating the benefits of using TD Methods.

Experiment 3 looks identical to Experiment 1 except for the bump in error in TD(0). As mentioned in Experiment 2 this is due to the fact that TD(0) performs poorly at propagating predictions back along a sequence. This bump in error is indeed captured in the experimental results. The experimental errors are lower than Sutton's probably due to not having the same α chosen as well as differences in the randomly generated training sets.

4. CONCLUSION

All three experiments are very similar to Sutton's results, thereby demonstrating first hand the benefits of using temporal difference methods as opposed to supervised methods. All three experiments confirmed that TD(1) performed the worst and low TD(λ) performed the best. The additional information gleaned by TD methods directly translated to increased accuracy.

A recurring theme throughout the experiments has been the slightly lower error achieved through the experimental results compared to Sutton's results. A variety of explanations have been posited to account for these improprieties including: not having Sutton's exact parameters, the random nature of the generation of the training sets, as well as possibly due to the dramatic increase in accuracy of modern computers in the past 30 years.

REFERENCES

[1] Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, *3*(1), 9-44.