CSE7642 Project: Temporal Differences Learning by Sutton (1988)

Shaikh Shamid

Department of Computer Science, Georgia Institute of Technology

February 19, 2018

Abstract

This report tries to reproduce some of the findings presented in Richard S. Suttons paper, Learning to Predict by the Methods of Temporal Differences (1988). The article claimed that the performance of the temporal-difference methods is better compare to supervised learning methods with respect to memory usage, computation time and prediction accuracy for most of the prediction problems where both can be applied. We were able to reproduce some of his findings with some minor issues.

1 One Dimensional Random Walk

1.1 Introduction

1D random walk is a stochastic or random process that describes a path consisting sequence of random steps along a line. Sutton considered a bounded 1D random walk of seven states A through G where two end states (A and G) are terminal states, where the walk starts in the middle state D, has a 50-50 probability of moving either left or right. The outcome of each walk is then defined to be either 0 if it ended on the left (A) or 1 if it ended on the right (G). The ideal probability of a walk ending in state G from each state is: $T = 1/6, 2/6, 3/6, 4/6, 5/6$. The goal is to estimate these probabilities using Temporal Difference(TD) learning method, and assess the method's performance by calculating the root mean squared (RMS) error of estimated probabilities against the ideal predictions.

1.2 Create Data

We first create the data as a list of vectors of length 7 for seven states filling 1 for the current state otherwise zero. In this way, we create 100 training observations each having 10 complete sequences. Sutton only considered middle states for creating the data. For simplicity we created 7 states, but will update only weights for the middle states.

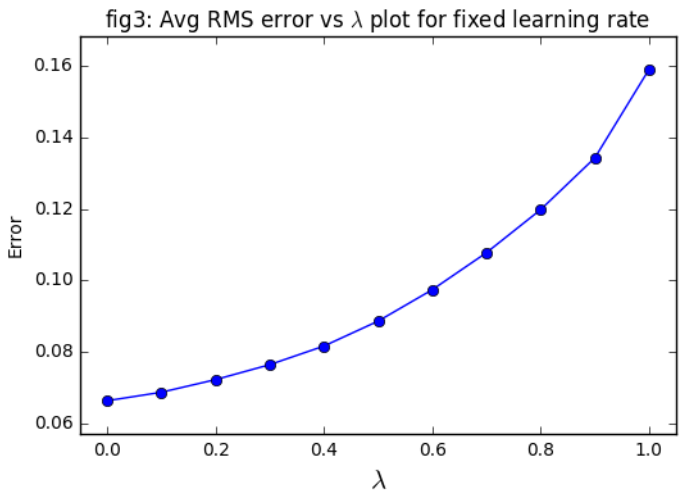


Figure 1: Average RMS plot for different λ values for a fixed learning rate of 0.01

2 TD Update Equation

Sutton defined the weight update as the following

$$\Delta w_t = \alpha \left(P_{t+1} - P_t\right) \sum_{k=1}^{t} \lambda^{t-k} \Delta_w P_k \tag{1}$$

where α is the learning rate, $P_t = w^T x_t$ is the linear prediction function at time t with observation vector $\mathbf{x_t}$, and $\Delta_w P_k$ is the vector differential of P_t with respect to weight vector w.

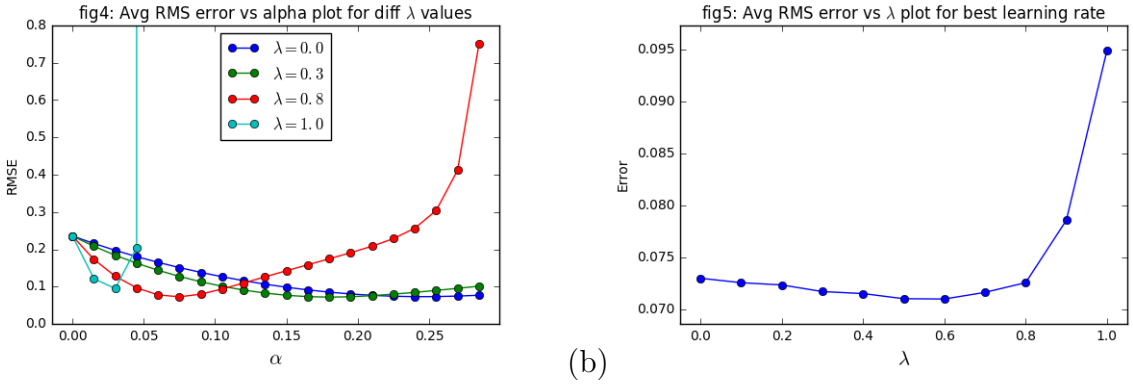
Using (1) we can update the weight vector w as

$$w = w + \sum_{t=1}^{m} \Delta w_t \tag{2}$$

3 Experiment 1

In this experiment (repeated presentation), the weight vector **w** was updated only after the algorithm was trained on the 10 sequences of a training set. In this repeated presentation experiment we repeat this process until convergence for each λ value. The convergence threshold was set to be 0.001.

Figure 1 shows that the experiment is a good match with what Sutton describes in figure 3 in his article except our average RMS error is lower compared to his plot. One reason could be the improved floating point precision compared to precision available back in 1988. But Overall, the learning is that the RMS error of TD(λ) batch method increases as λ increases.



(a) (b)

Figure 2: (a) Average RMS plot for different learning rates α for 4 different λ values (b) Average RMS plot for different λ values with best learning rates

4 Experiment 2

In this experiment (incremental update), the weight vector $\mathbf{w}$ was updated incrementally after each sequence. Also in this incremental update, each training set was fed to learning algorithm once, unlike experiment 1. In order to demonstrate the effect of learning rate (α) on the performance of TD method, multiple learning rates were applied to the TD algorithm.

Comparing fig. 2(a) and Sutton's fig 4, the overall pattern looks similar. The RMS value for $\alpha = 0$ looks same for all λ values in both figures. But for large λ ($\lambda = 1.0\&0.8$) the error increases very rapidly unlike Sutton. I think this is also attributed to the floating point precision issue.

When comparing fig. 2(b) (same plot as fig.1, but for best performing α) and Sutton's fig 5, this is a good match! We get a slight dip like Sutton's plot but for a little higher value of λ (around 0.6). Also, RMS error is lower than what Sutton reported.

5 Conclusion

We were able to replicate Sutton's original findings. In the repeated presentation experiment, the RMS error of TD method decreases as λ increases and $\lambda = 0$ has the best performance. As shown in fig. 2 (a), the learning rate affects the performance of TD method and it is different for different λ values. Figure 2 (b) shows the best RMS error for each λ with best performing learning rate α. After applying best learning rate parameter values, now TD(0.6) seems to have the lowest error unlike experiment 1 where TD(0) was the lowest.

References

[1] Richard S. Sutton. *J Learning to predict by the methods of temporal differences.* In MACHINE LEARNING, pages 9-44 Kluwer Academic Publishers, 1988.