Project 1: CS8803 - O03 Reinforcement Learning

Saad Khan (skhan315@gatech.edu)

June 19, 2016

1 Introduction

The purpose of this project report is to experimentally replicate temporal difference learning techniques put forward by Richard Sutton in his 'Learning to Predict by the Methods of Temporal Differences' paper published in 1988. The focus of this report will be to apply TD learning to the "Random Walk" example discussed by Sutton in the first half of the paper.

1.1 Temporal Difference (TD) Methods

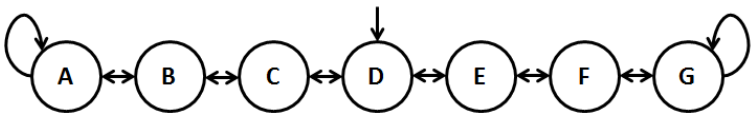
TD methods are a class of incremental learning procedures specifically geared towards solving prediction problems. These methods bridge the two ends of the spectrum where Monte Carlo methods TD(1) are at one end and one-step TD methods TD(0) are at the other. Unlike conventional learning methods (supervised learning) which are based on differences between predicted and actual results, TD methods learn based on differences between successive predictions over time. According to Sutton, this is the reason why TD methods make more efficient use of their experience than do supervised learning methods. Figure 1 shows the equations, from the paper, that were used to implement the TD methods. Equation '1' was used for the final weight update, while equations '4' and 'A' were used for intermediate weight calculations based on successive predictions while equation 'B' was used for error estimates.

$$w \leftarrow w + \sum_{t=1}^{m} \Delta w_t \textbf{(1)} \qquad \Delta w_t = \alpha(P_{t+1} - P_t) \sum_{k=1}^{t} \lambda^{t-k} \nabla_w P_k \textbf{(4)} \qquad P_t = w^T x_t \textbf{ (A)} \qquad RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \textbf{ (B)}$$

Figure 1: [Equations] weight update, TD(λ) prediction, prediction substitution, RMSE

1.2 Random Walk

As shown in Figure 2 [LEFT], random walk, a simple dynamical system, can be represented as a Markov decision process with states A and G as the termination states. The walk starts at state D in the center at time step t shown by the down arrow and then has equal chance of moving left or right on subsequent time steps, until it finishes in one of the termination states. In this system, the walk is designed in such a way that if it ends at A the outcome is $z = 0$ and if it ends at G then $z = 1$.




```
   A   B   C   D   E   F   G
[[ 0.  0.  0.  1.  0.  0.  0.]
 [ 0.  0.  0.  0.  1.  0.  0.]
 [ 0.  0.  0.  0.  0.  1.  0.]
 [ 0.  0.  0.  0.  0.  0.  1.]]
```

Figure 2: [Random Walk] LEFT: state representation, RIGHT: sequence vector representation

Typical examples of a walk can be [*DCDEFG, DCBCBA, DEFEFG, DEFG*]. If a walk *DEFG* (also called a single *sequence*) has occurred, the learning method will be provided with a series of vectors (x_D,x_E,x_F,*1*) representing a sequence in unit basis form as shown in Figure 2 [RIGHT]. According to Sutton, representing the sequence in this form simplifies the calculation as the prediction P at time t would just be $P_t = w^T x_t$, i.e. simply the i^{th} component of the weight vector at time t. This is the basis of all matrix calculations for replicating the experiments.

2 Experiments

2.1 Implementation

Based on the paper, two different experiments were designed to prove that TD methods converge more rapidly and make more accurate predictions along the way compared to supervised learning methods. For reliable outcome, 100 training sets each consisting of 10 sequences of walks were presented to the learning method and the root mean square (RMS) error was to be calculated between the asymptotic predictions from training sets and the ideal predictions where the ideal predictions for all 5 non-terminating were [B:1/6, C:2/6, D:3/6, E:4/6, F:5/6]. Data was supplied to the learning method, single training set at a time but the weight updates for the two experiments were performed at different instances.

The implementation of the random walks and the RMS error calculation for the experiments was done using code written in python. Random number generator was used to simulate the movement for a walk sequence. Equation 'A' from Figure 1 was

substituted in equation '4' and then equation '4' in tandem with equation '1' was used for weight updates. After the final weight vector was generated, error calculation formula similar to equation 'B' was used to calculate RMS error between ideal and predicted values. Finally, the results of the experiments were exported to MS Excel where the plots were generated.

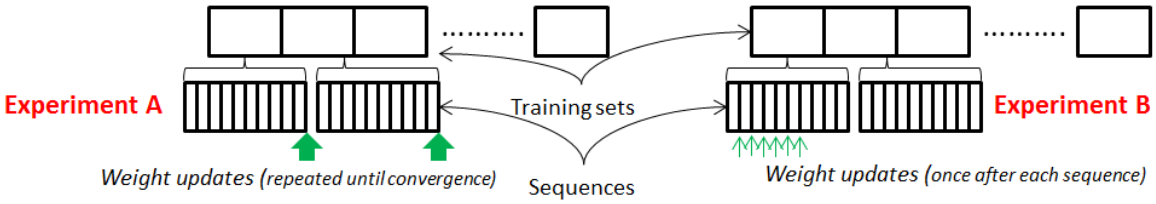


Figure 3: weight updates for the 2 experiments

2.2 Experiment A - Repeated Presentations

2.2.1 Description

The repeated presentations experiment, as evident by the name, was designed in a way to update weights based on TD(λ) equation (4) per each training set after all 10 sequences in a training set had been presented to the learning method. This is shown in Figure 3 [LEFT]. The method kept on iterating over a particular set of sequences until it observed no further changes in the weight vector (until the weights had converged), then it moved on to the next set. Once it had done so for all 100 training sets for one value of λ, it then moved on to the next value of λ. Once the final weight vector was obtained, its mean RMS difference was calculated with the ideal predictions and the results were plotted against λ.

2.2.2 Assumptions

As Sutton did not mention the criteria for convergence in his paper, there were few assumptions made in order for the replication of the results of [Sutton 1988 - Figure 3]. Different values of α, the learning rate, shown in equation (4) Figure 1 were used and finally the value which provided the best mean RMS error was retained. This was paired with different values of delta weight ($\triangle w$) threshold between weight vector at time t and $t + 1$ and updates were stopped when $\triangle w$ was less than that threshold.

2.2.3 Pitfall and Resolution

The hardest part of this experiment was to come up with the best ($\alpha/\triangle w$) pair. In order to explore which pair would yield the best performance by the learning method i.e. produce largest difference between the one-step estimate TD(0) and the Widrow-Hoff estimate TD(1), a number of pairs were presented to the repeated presentations method and results were noted as shown in Figure 4 [LEFT]. The pair producing the largest positive RMS error difference was then used as input for the final plot for experiment A.

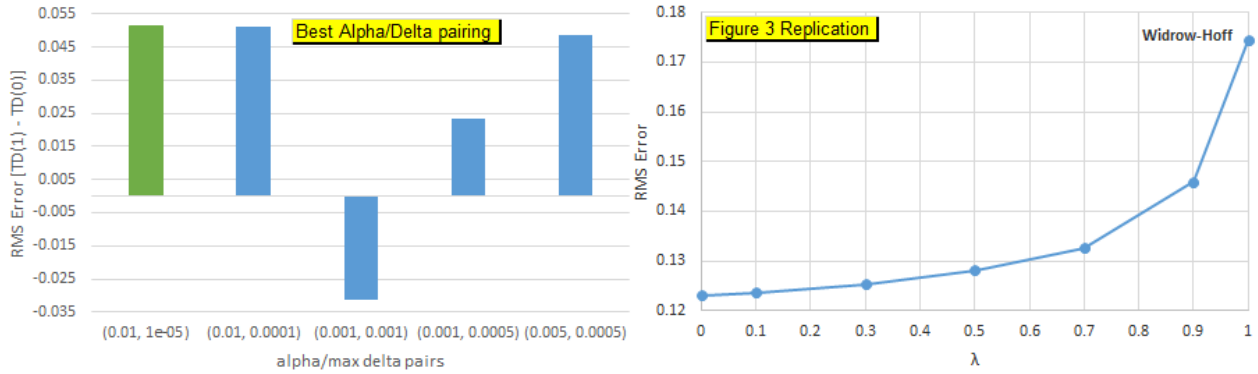


Figure 4: [LEFT]: Best alpha/delta pair selection, [RIGHT]: Experiment A: RMS Error vs. λ

2.2.4 Results and Observations

It can be seen that the best pair that yielded maximum difference was with $\alpha = 0.01$ and $\triangle w = 0.00001$ as highlighted by the bar in green. Figure 4 [RIGHT] shows the replicated plot using these values and it can be observed that the learning method performs much better for λ<1. This contradicts with what is expected of the TD(1) procedure (an equivalent of Widrow-Hoff procedure) as it encapsulates sum of all past values of $\triangle w$ P_t. However, this encapsulation only reduces error for the training set, causing the procedure to overfit, but for lower values of λ this is not the case because of the decay induced by λ^{t-k} term in equation (4). So the procedure tends to overfit less and less as λ approaches 0.

2.2.5 Significant Differences

The resulting plot for repeated presentations in Figure 4 [RIGHT] follows an asymptotic trend as λ approaches 1, which is similar to the original plot in Sutton's paper, however, this replication shows much less RMS error for both TD(0) and TD(1) at almost

0.07 units less less than the original plots. This could be due to randomization of the sequences, variations in alpha values or delta weight threshold. I also tried other pairs for α and $\triangle w$, however, those used in Figure 4 [LEFT] were the most meaningful because for higher value pairs, learning procedure started to diverge.

2.3 Experiment B - Sequential Updates

2.3.1 Description

This experiment can be divided in to two sub experiments (Experiment B(i) and B(ii)). In both these cases, in contrast to repeated presentations, the main difference is the weight update, which took place only once for each sequence, after a training set was presented to the learning method as illustrated in Figure 3 [RIGHT]. For the first part of this experiment, variants of the learning rate α were presented to the learning method to see the effect on RMS error and plots were constructed for 4 different values of λ. For the second part, for a particular value of λ, learning rate was chosen that gave the lowest RMS error between the ideal and predicted weight values and then plot was generate of minimum RMS error against λ.

2.3.2 Assumptions

Main assumption, in accordance with Sutton's paper was to setup an environment by selecting initial weights for non-terminating states a 0.5 each. This was to avoid a right or left side based termination for a sequence and neutralize any bias that could occur due to a single cycle of weight updates.

2.3.3 Pitfall and Resolution

Biggest problem was to handle weight updates for TD(0) as I was unable to produce this plot accurately at first using the regular TD(λ) equation '4'. This was resolved by implementing the straight forward TD(0) equation on page 16 of the paper.

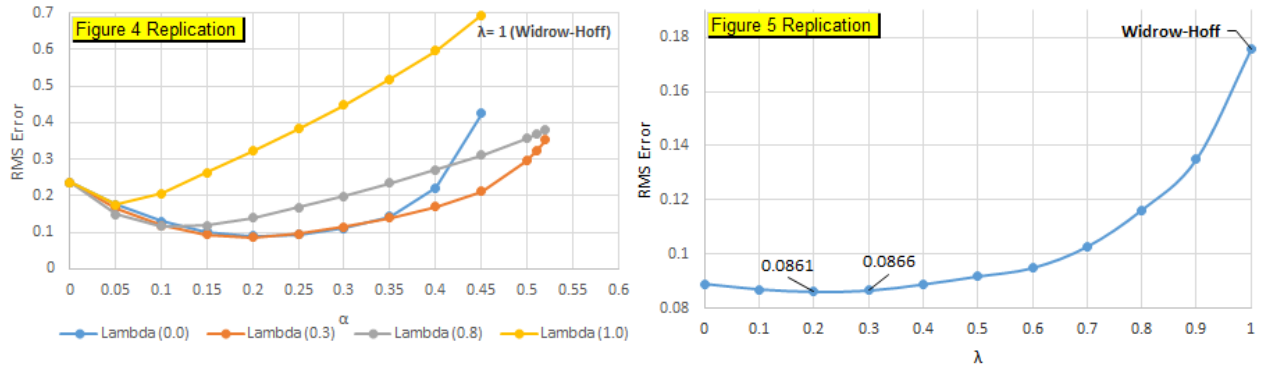


Figure 5: [LEFT]: Experiment B(i): Minimum RMS Error vs. α, [RIGHT]: Experiment B(ii): RMS Error vs. λ

2.3.4 Results and Observations

Figure 5 [LEFT] shows similar results for most parts when compared to Sutton - Figure 4. It can clearly be seen that intermediate values of α produce much less error. Again, similar to the repeated presentations experiment, TD(1) (where entire sequence is observed just like supervised learning methods) produced very poor results, however, on the other hand, for λ less than 1 results were very rewarding, specially for most part of the $\lambda = 0.3$ plot until it started diverging for higher values of α.

Figure 5 [RIGHT] shows the replication of Sutton - Figure 5, depicting least RMS error vs.λ and supports the claim that intermediate value of λ produce much better results. Here it can clearly be seen that ideal λ is some where between 0.2 and 0.3, rather than $\lambda = 0$ which was the case in repeated presentations experiment hence proving that single cycle of weight update with $\lambda = 0$ does not back propagate enough.

2.3.5 Significant Differences

Main difference for Figure 4 replication when compared to the original plot was that (i) all graphs for $\lambda<1$ started to diverge well before $\alpha = 0.6$, i.e. there is a constant negative x-axis offset observed. Specially, $\lambda = 0.3$ case, which diverges rapidly for $\alpha>0.5$.

Figure 5 [RIGHT], the plot replicated for Sutton - Figure 5 follows the same half concave trend with minor difference being the overall RMS error, For instance, RMS error for λ 0.2 and 0.3 in the original plot is around 0.11 while here it is below 0.09. This can again be attributed to the environment where Sutton was performing these tests, what method did he used generated the walk sequences, etc.

3 Conclusion

Overall this was a very unique and interesting project in the sense that we got to actually tinker with published results. It was a little hard to deal with randomization aspect of the walks at first. Overall things turned out quite well after studying the paper in depth and applying the TD(λ) equations in simialr fashion as were applied by Sutton in the paper.