Desperately seeking Sutton:

Shailesh Tappe: stappe3 **Git Hash: e80e930ca568a38010c88ee4b5aae20e2df2f2f8**

Objective: The objective of this project report is to understand experimentally incremental learning procedures specialize for prediction i.e. *temporal difference* (TD) learning techniques authored by Richard Sutton in his paper “Learning to Predict by the Methods of Temporal Difference” published in 1988. The experiment is to apply TD Learning mechanism to the “Random Walk” example discussed by Sutton in his paper.

Introduction: TD methods are learning methods geared towards solving prediction problem. TD uses past experiences with unknown or incompletely know system to predicts its future behavior. Sutton in his paper suggest and proved hypothesis that supervised learning are prediction problem that can be resolved with temporal difference (TD) method. Conventional learning methods (supervised learning) are based on difference between predicted and actual result (Figure 1 equation (2)), whereas TD methods learn based on successive learning over time (Figure 1 equation (4)). This is why Sutton suggests that TD methods make more efficient use of experience than supervised learning procedure, the Widrow-Hoff rule.

$$w \leftarrow w + \sum_{t=1}^{m} \Delta w_t. \quad (1) \qquad \Delta w_t = \alpha(z - P_t)\nabla_w P_t, \quad (2) \qquad \Delta w_t = \alpha(P_{t+1} - P_t)\sum_{k=1}^{t} \lambda^{t-k}\nabla_w P_k. \quad (4)$$

Figure 1:

Random Walk: Random walk is simple dynamical system, that generates bounded random walks which can be represented as a Markov decision process within states A through G. The walk starts at state D at time step t and has 50% chance of moving either right or to the left. With any of the edge states i.e. A or G is entered, then walk terminates. The outcome is 0 for terminal state A and 1 for state G. For each not terminal state outcome at time step is 1 for sequence vector.

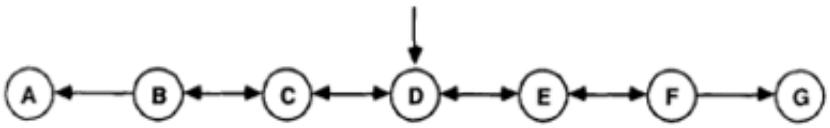

Figure 2

For walk sequence DEFG, the learning method provided with series of vectors (X_D,X_E,X_F,1) represented as in figure 3. Sutton suggest that representing sequences in this form simplifies calculation for prediction P at time t would just be value of *ith* component for weight vector i.e. $P_t = w^T x_t$

```
  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 3

Experiments: Basing on Sutton's paper, experiments were performed using observation-outcome sequence as described above. Experiments were performed to prove that TD methods converge more rapidly and with more accurate prediction than supervised learning procedures, the Widrow-Hoff rule. To reach out reliable outcome; 100 training sets, each consisting 10 sequences (walk matrix) presented to the learning methods. RMSE (root mean square error) calculated between asymptotic predictions from training sets and ideal predictions for all non-terminating state [1/6,1/3,1/2,2/3,5/6]. A walk matrix data supplied to experiments with varying weight update implementation for each experiment.

1. Experiment - Random walk under repeated presentation:

1) Experiment Synopsis:

Random walk under repeated representation was designed to update weights using TD(λ) equation (Figure 1: equation 4). Observation-outcome sequence generated by experiment setup and was presented to calculate weights for each training set. In this experiment weight vector is not updated for each sequence, but instead delta weight was accumulated until complete presentation of training set. The converged value of weight is then used to calculate RMS difference of weight with ideal weight of each non terminal state [1/6,1/3,1/2,2/3,5/6]. The learning procedure is repeated for 7 values of learning rates λ i.e. [0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0]

2) Issues and Resolution:

Sutton has specified need of convergence for repeated presentation, but had not mentioned criteria for it. For conversation gradient descent is used and tested with varying values of epsilon and compared it with difference in sum of weights for each iteration of training sets.

While experimenting TD learning with this experiment, it was not producing optimal result suggested in figure 3 of Sutton paper. One of the main pitfalls was finding best value for learning rate α. Different values for α were tested to find optimal result for TD suggested in figure 3 of Sutton's paper.

3) Observations:

Figure 4 shows replicated plot of Figure 3 in Sutton's paper. It is observed that performance improved with decreasing value of λ and was best at λ = 0, as suggested in Sutton's paper.

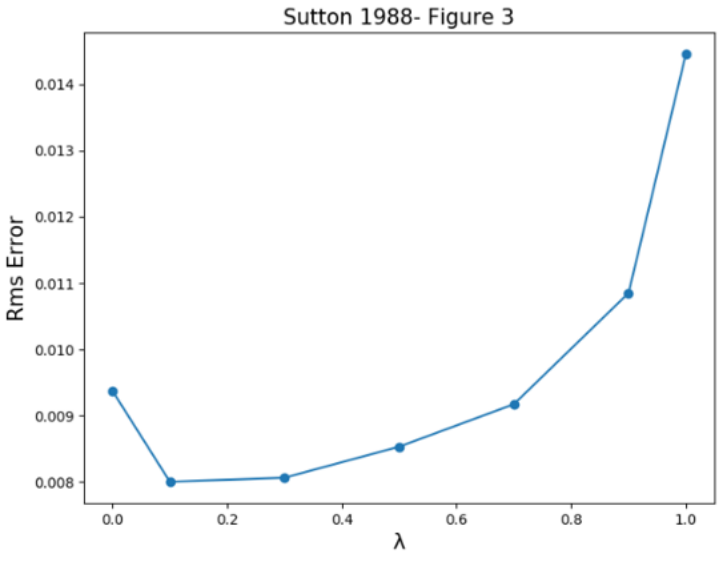

Figure 4

The result contradicts conventional wisdom that Widrow-Hoff (λ = 1) minimizes RMS Error between prediction and actual result. But experiment result shows that λ = 1 performs worse than λ < 0, that means Widrow-Hoff only minimizes error on training set and not necessarily on future experience.

2. Experiment - Random walk under repeated presentation:

1) Experiment Synopsis:

This experiment was carried out in 2 parts and concerns question of learning rate when training set is presented just once rather than repeated until converges. For the first part (Figure 4 of Sutton's paper) varying learning rates α were presented to learn and see RMS Error plotted against different λ [1,0,0.8,0.3]. Second part (Figure 4 of Sutton's paper) of experiment was provided with best learning rate α for a value of λ that gives lowest RMS Error between predicted and ideal weights of each non terminal state [1/6,1/3,1/2,2/3,5/6]. The plot was generated of minimum RMS Error against λ.

2) Issues and Resolution:

As suggested in Sutton's paper, initial weights for non-terminal states with 0.5 each. This is done to neutralize any bias because of single cycle weight updates.

3) Observations:

The outcome of experiment (Figure 5 and 6) shows similar result as in comparison with Sutton's result for Figure 4 and 5 in the paper.

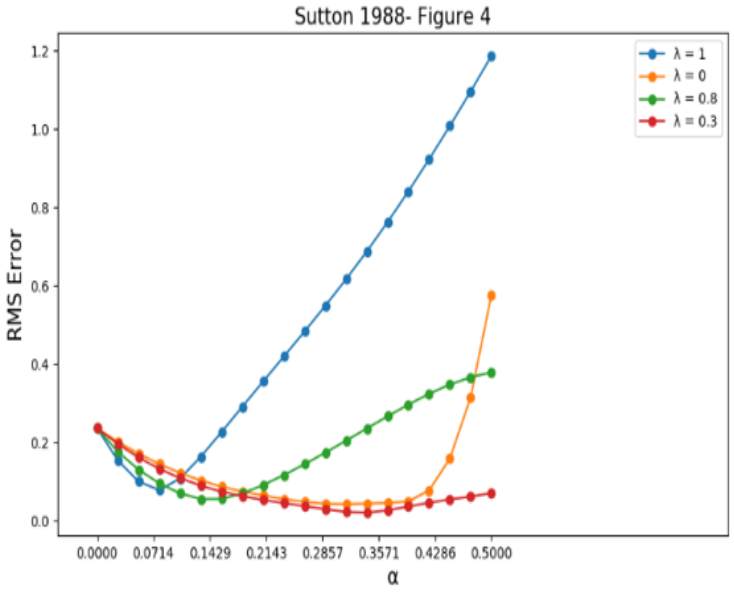

Figure 5

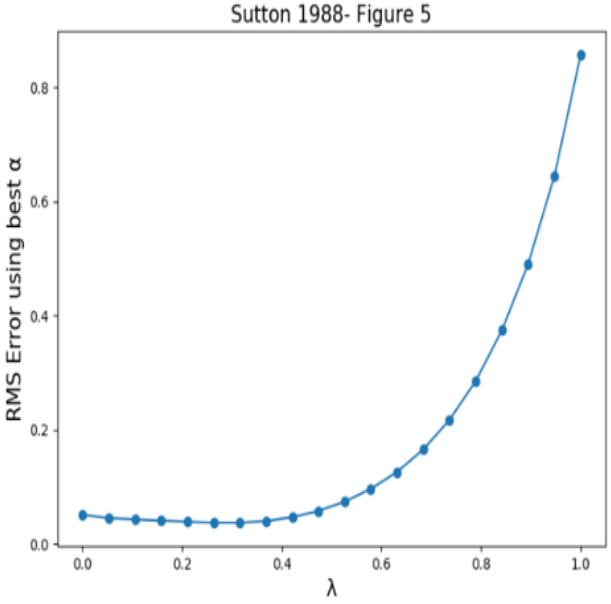

Figure 6

In Figure 5 (Figure 4 in Sutton paper), the result shows that learning rate α have significant effect on performance and best result is observed with λ between 0 and 1. The Widrow-Hoff procedure (λ=1) produces worst estimates whereas λ < 1 performed well in absolute term and as well as wider range of α values than with supervised learning.

In Figure 5 (Figure 4 in Sutton paper), rendering RMS Error for best α vs λ, supports that intermittent values of λ produces better result. It can be seen from figure that ideal value for lambda is somewhere between 0.3 and 0.4 and not for λ = 0. This therefore proves that single cycle weight update of λ = 0 does propagate predictions level back to sequence.

Conclusion: Overall project experience was very unique and interesting and outcome of experiment matches with what Sutton has suggested in his paper. Initially it was challenging to deal with randomization and adjusting learning rates, but later it turned out well after reading paper well and applying TD(λ) equation to estimate underlying Markov process.

The analysis and experiments suggest that TD methods may evolve as choice for many real-world learning problems as compare to supervised learning. TD method are more tailored to temporal structure and as a result it computes outcome incrementally and requires significantly less memory and peak computation. Empirically TD methods learn faster than supervised learning (Widrow-Hoff rule) and produces more optimal result for training sets that are presented repeatedly.

References:

Richard Sutton: Learning to Predict by the Methods of Temporal Difference

Dr Charles Isabel and Dr Michael Littman: Lesson 4: TD and Friends on Reinforcement Learning video lectures

TA's during office hour and piazza forum.