Recreation of Sutton 1988

GIT submit id: 4fb584d848b8d047206a60aa43e22eddcbee598c
James Corwell
CS7642 OMSCS
jcorwell3@gatech.edu

Literature Review

This paper is investigating the topic of temporal difference learning of multi-step decision processes, such as Markov Decision Processes. In Sutton's paper, he proposes the idea of learning through temporal difference methods. The TD methods strength is operating on limited information, and should converge to the same answer when infinite information is available -- TD (1) or Monte Carlo. Generally, TD takes advantage of a temporal sequence, which while sometimes misleading, is generally helpful.

Sutton proposes that the only requirement for a system to be effectively predicted is that it be dynamical with states evolving over time (Sutton, 20). To typify this, Sutton uses a random-walk example for the TD methods.

Random walk is a type of markov chain, where there is a probability of an agent moving from one state to another. In Sutton's paper, he uses a 1-dimensional random walk to demonstrate the effect of the TD algorithm, where the agent had a 50/50 probability of moving to either the left or right. When the edge is reached, the walk is over.

Sutton Generated Figures

Figure 3

The random-walk experiment was implemented with 100 training sets of 10 sequences.

Weight increments were computed according to TD(lambda):

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

Sutton used several different values for lambda; 0,.1,.3,.5,.7, .9 and 1.

In the first experiment, the weight vector was not updated after each sequence, instead the weights were accumulated over sequences and only used to update the weight vector after the complete presentation of a training set. The learning rate was constant for each different value of lambda. As a measure of performance, Sutton uses the RMS between the procedure's asymptotic predictions using the training set and the ideal predictions. (Sutton 21)

There were several novel outcomes of this experiment. "For small alpha, the weight vector always converged to the same final value, independent of the initial value." (Sutton 21) This is known as the repeated presentations training paradigm. Best performance occurred at lower values of alpha.

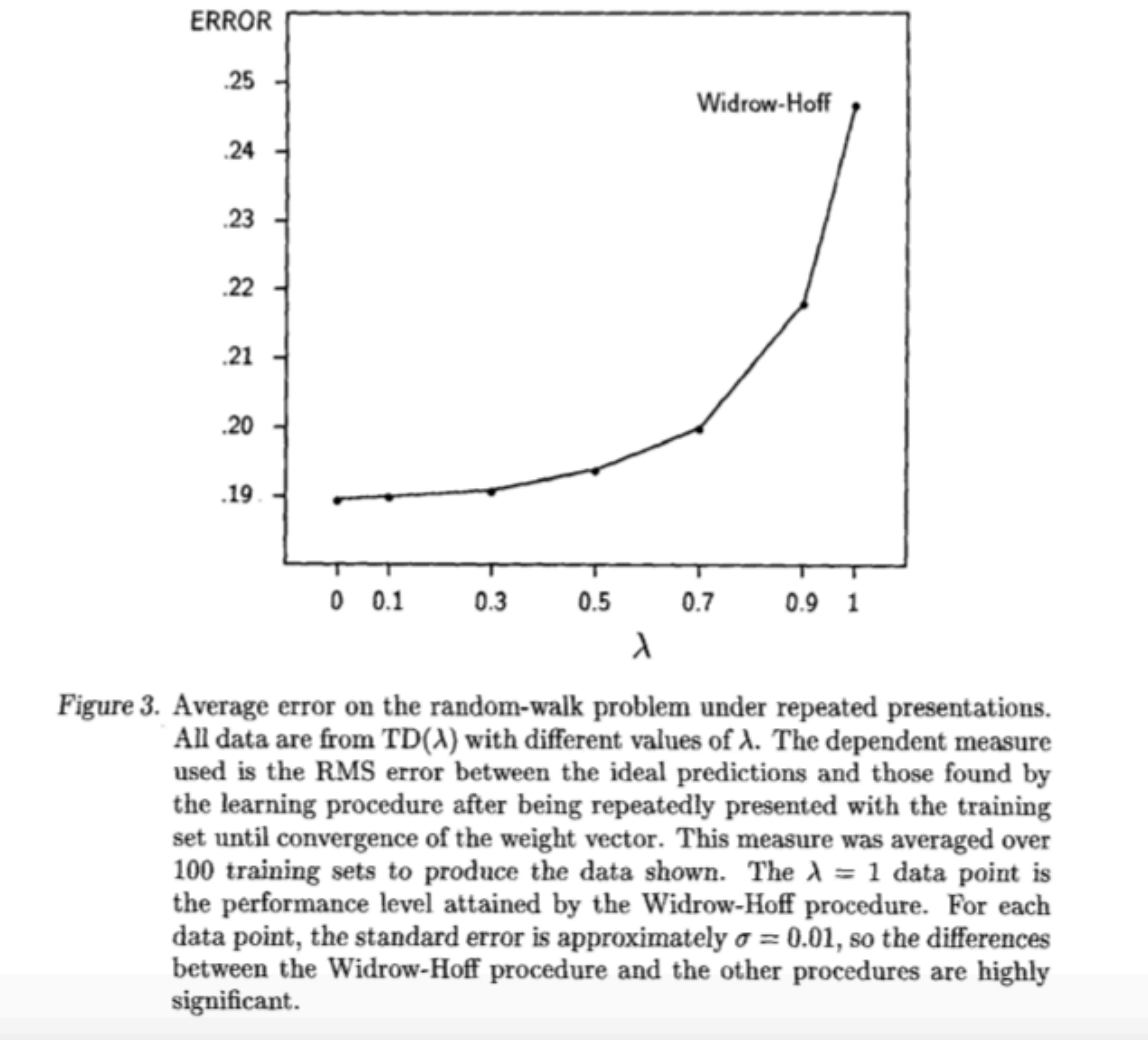


Figure 3. Average error on the random-walk problem under repeated presentations. All data are from TD(λ) with different values of λ. The dependent measure used is the RMS error between the ideal predictions and those found by the learning procedure after being repeatedly presented with the training set until convergence of the weight vector. This measure was averaged over 100 training sets to produce the data shown. The $\lambda = 1$ data point is the performance level attained by the Widrow-Hoff procedure. For each data point, the standard error is approximately $\sigma = 0.01$, so the differences between the Widrow-Hoff procedure and the other procedures are highly significant.

(Sutton, Erratum)

Figure 4

This experiment is concerned with learning rate when training set is available only once to the agent rather than continually. Namely, updates to weights were performed after every sequence, unlike in Figure 3 where updates were made after each training set. For this run, Sutton was concerned with supplying a number of different learning rates to compare.

From the experiment, we can find that the learning rate has as significant performance impact on the TD algorithm. TD(1) also produced the worst estimates independent of alpha. We also find significantly lower RMS values when the alpha is tuned properly, showing that updating after each sequence is generally superior.

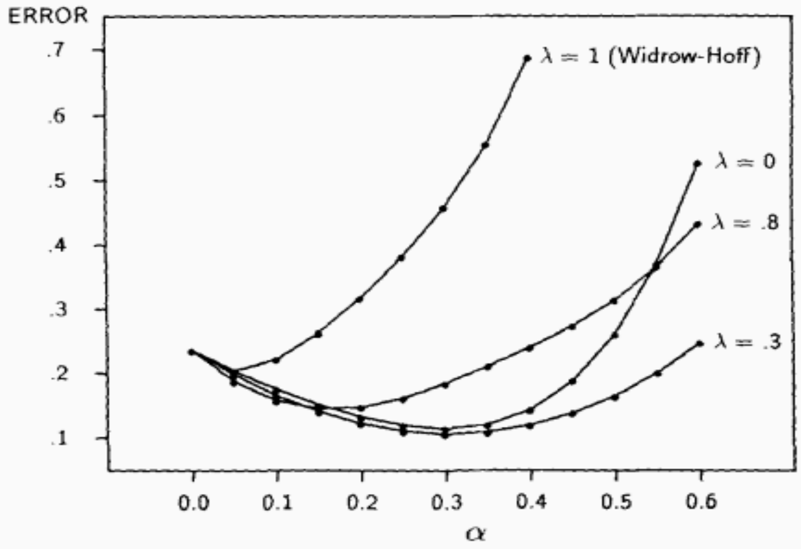


Figure 4. Average error on random walk problem after experiencing 10 sequences. All data are from TD(λ) with different values of α and λ. The dependent measure is the RMS error between the ideal predictions and those found by the learning procedure after a single presentation of a training set. This measure was averaged over 100 training sets. The λ = 1 data points represent performances of the Widrow-Hoff supervised-learning procedure.

Figure 5

Figure 5 was generated using the best alphas from figure 4. The outcome is a significantly lower error for every value of lambda compared to figure 3, this typifies that updating the weights after every sequence is superior than after each training set.

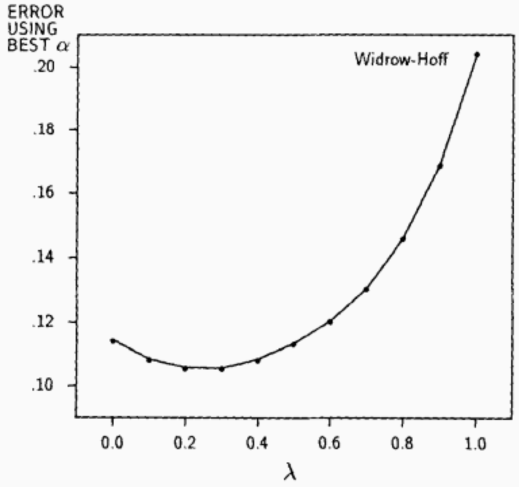
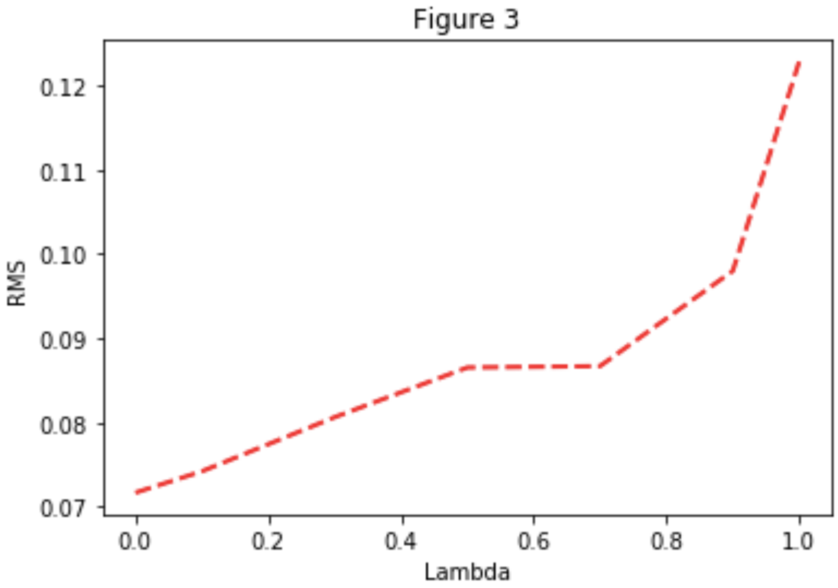
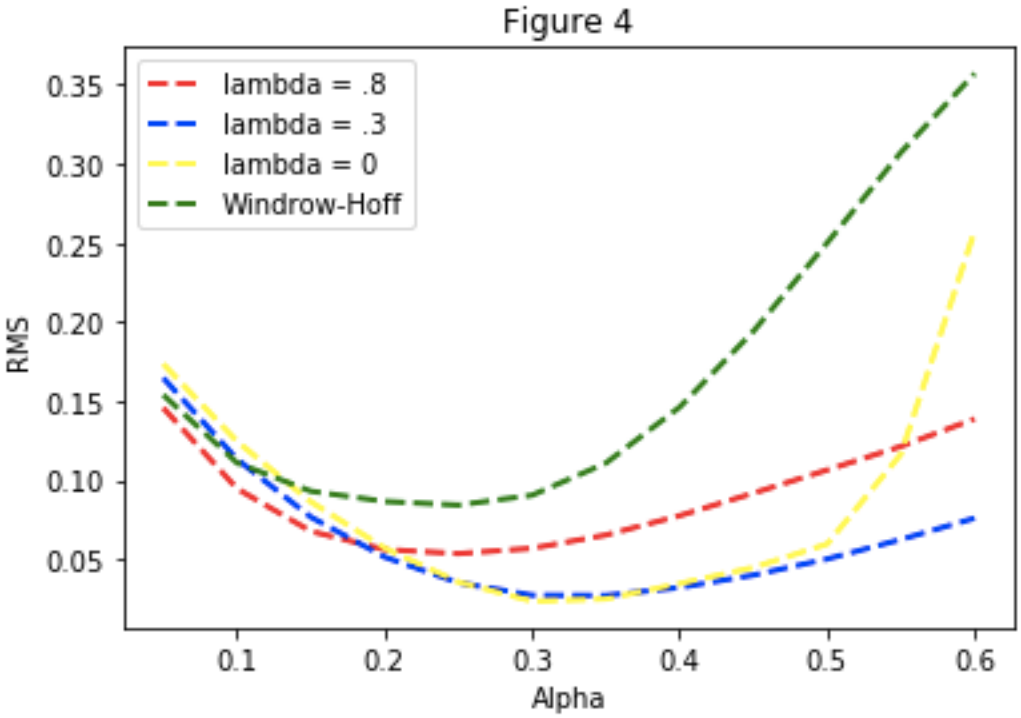


Figure 5. Average error at best α value on random-walk problem. Each data point represents the average over 100 training sets of the error in the estimates found by TD(λ), for particular λ and α values, after a single presentation of a training set. The λ value is given by the horizontal coordinate. The α value was selected from those shown in Figure 4 to yield the lowest error for that λ value.

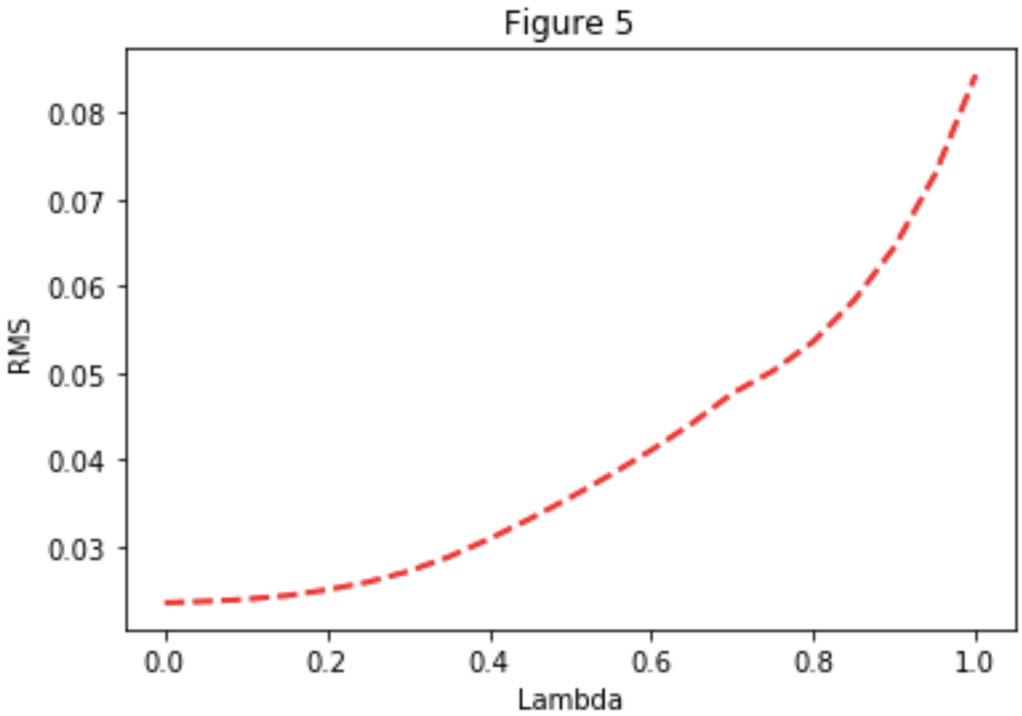
Sutton Figures Recreated



Recreated Sutton’s Figure 3



Recreated Sutton’s Figure 4



Recreated Sutton's Figure 5.

My results come very close to matching Sutton's figures, Figure 4's curves remarkably look near identical to those in Sutton's paper, clearly showing the efficiency of TD(lambda) over TD(0) and Windrow-Hoff, as well as showing how the learning rate effects the system. All of my figures have a significantly lower error than Sutton's figures, however changing the random seed –and consequently all of the random data— changes the RMS as well. In my run of the second experiment I was also able to get lambda = .3 as the value with the least RMS, just as in Sutton's paper.

No learning rate or convergence factor was given in the generation of figure 3. As per recommendations on piazza I went ahead with an arbitrary small value .01 for learning rate and .01 for convergence threshold. These worked well. Learning rates weren't explicitly given for the second experiment, but were inferable from the graph. Of course, the random generation of training sets may lead to random differences from Sutton's data.

References:

Sutton, Richard S., "Learning to Predict by the Methods of Temporal Differences" Machine Learning. 1988 9-44,377. Print.