Desperately Seeking Sutton

Tiantian Guo
tguo60@gatech.edu

***Abstract*—This report will reproduce Richard Sutton's 1988 paper [1] three figures in order to augments the understanding of the paper and extend existing literature.**

***Index Terms*—reinforcement learning, temporal differences, incremental learning**

I. Introduction

The paper "Learning to Predict by the Methods of Temporal Differences" [1] proved that for most real-world prediction problems, temporal-difference methods require less memory and less peak computation than conventional supervised learning methods and they even produce more accurate predictions. This report implements [1] experiments to better understanding TD-learning.

II. Problem Statement

A bounded random walk in a state sequence generated by taking random steps to the right or to the left until a boundary is reached. Fig. 1 shows a system that generates such state sequences. There are seven states (A, B, C, D, E, F, G) in random walks. Every walk begins in the center state D. At each step the walk moves to the neighboring state, either to the right or to the left with equal probability. If either edge state (A or G) is entered, the walk terminates.

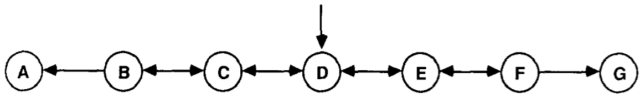

Fig. 1. A generator of bounded random walk.

A walk's outcome was defined to be $z = 0$ for a walk ending on the left at A and $z = 1$ for a walk ending on the right at G. The random walk learning methods estimated the expected value of outcome z, the expected value is equal to the probability of a right-side termination. Ideal predictions for non-terminal states are $\{\frac{1}{6}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{5}{6}\}$. The random walk outcome is the expected value of z, in this problem, its expected value is equal to the probability of a right-side termination.

For each nonterminal state i, the observation vector $\mathbf{x}_i$, which will set the walk position as 1, others as 0, e.g. $\mathbf{x}_D$ = $(0, 0, 1, 0, 0)^T$. The prediction can be calculated as $P_t = \omega^T x_t$ and $P_{m+1} = z$. When termination is state A, $z = 0$ while $z = 1$ when termination is state G.

TD(λ) is used to estimate probability of right-side termination. In TD(λ), learning procedure is summarized as follows:

$$\omega \leftarrow \omega + \sum_{m}^{t=1} \Delta\omega_t \tag{1}$$

$$\Delta\omega_t = \alpha(P_{t+1} - P_t) \sum_{t}^{k=1} \lambda^{t-k} \bigtriangledown_\omega P_k \tag{2}$$

$$P_t = \omega^T x_t \tag{3}$$

(1) is the updated process of the weight ω. (2) is the calculation of the change of ω, α is a positive parameter affecting the rate of learning and λ is an exponential weighting with recency, in which alterations to the predictions of observation vectors occurring k steps in the past are weighted according to λ^k for $0 \leq \lambda \leq 1$. At $\lambda = 1$, TD(1), the algorithm simplifies to supervised learning. In (3), P_t was simply the value of the i^{th} component of ω. To treat this problem simpler, $\bigtriangledown_\omega P_k = x_t$. Predictions or weight vectors are learnt by accumulating changes in weights per each time step.

This project is to evaluate the relationship between weights ω estimation error, α and λ under different experiment settings in III, IV and V.

III. Experiment 1 - Repeated Presentation

A. Settings

This experiment computes average error on the random-walk problem under repeated training. In order to obtain statistically reliable results, 100 training sets, each consisting of 10 random walk sequences, were constructed for all learning procedures. For all procedures, weight increments were computed according to TD(λ), as in (1). Seven values were used for λ. $\lambda = 1$ is refering to Widrow-Hoff supervised-learning, $\lambda = 0$ is TD(0), and $\lambda = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$. ω weight vector is initialized to [0.5,0.5,0.5,0.5,0.5].

In this experiment, the weight vector was not updated after each sequence as indicated by (1). The $\Delta\omega$'s were accumulated over sequences and only used to update the weight vector after a training set. Each training set was presented repeatedly to each learning procedure until the $\Delta\omega$ converge. If total $\Delta\omega$ is greater than 0.001, same training set is used until weights converge. α values are chosen from 0.001 to 0.01 with 0.0005 interval. α produces smallest $\Delta\omega$ is finally used in plotting Fig. 2.

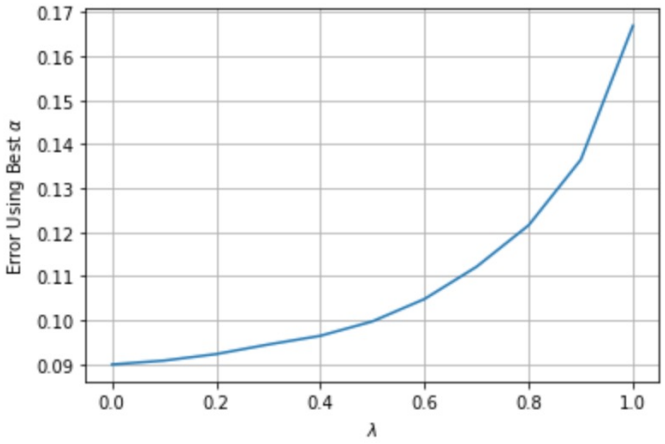

Fig. 2. Average error on the random-walk problem under repeated presentations. All data are from TD(λ) with different values of λ

λ	Best α
0.0 - 0.3	0.002
0.4 - 0.7	0.0015
0.8 - 1.0	0.001

TABLE I
BEST α IN REPEATED TRAINING

B. Result

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. Results are generated as Fig. 2 (with α generates best RMSE among 100 training set each with 10 sequences). The α produces best error is listed in Table.I.

As shown, result is similar to Figure 3 in Sutton paper, in which RSME error exponentially increases as λ reaches 1. TD(0) is converging to optimal predictions, outperforming Widrow-Hoff or TD(1).

However, the RMSE value in this experiment is slightly lower than Sutton's. The difference is mainly due to convergence limit settings or other settings that not explicitly explained.

C. Pitfalls

Sutton's paper does not explicitly define convergence condition and learning rate. Also the limit in number of steps or training sequences also remain unknown.

IV. Experiment 2 - Learning Rate Analysis

A. Settings

This experiment setting is similar as the experiment 1, but concerns of learning rate when the training set is presented just once rather than repeatedly until convergence. Weights ω are updated after each sequence but will not training till convergence. α and λ are both variables in this experiment.

λ	Best α
0.0 - 0.4	0.2
0.5 - 0.7	0.15
0.8 - 0.9	0.1
0.1	0.005

TABLE II
BEST α IN NON-REPEATED TRAINING

B. Result

As shown in Fig. 3, RMSE curves are shifted left from Sutton's work by about 0.1α, but the shape and trend are similar. This difference maybe because of various length of training sequences. Larger training sequences maybe can produce more similar plot as Sutton's result.

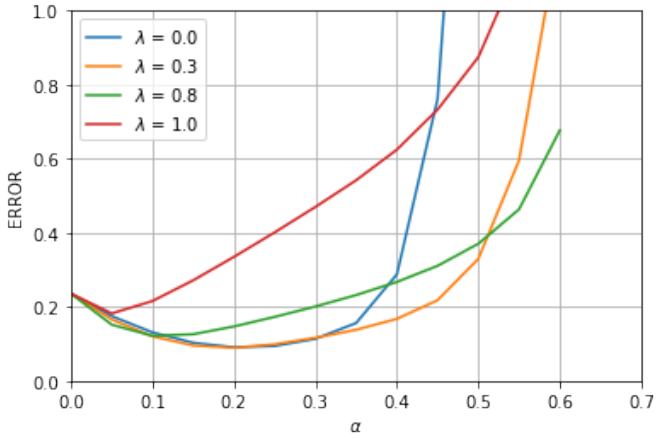

Fig. 3. Average error on random walk problem after experiencing 10 sequences.

C. Pitfalls

This experiment does not place a steps limitations in random walk sequences.

V. Experiment 3 - Best α Analysis

A. Settings

Same settings as Section IV, this experiments examine minimal RSME for each λ using best α determined. Each data point represents teh 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 horizontal coordinate. The α value was selected from those shown in Fig.3 to yield the lowest error for that λ value.

B. Result

The best α can be calculated as in Table.II. This experiment produces similar result, as seen in Fig.4 to Sutton's paper but a little difference in the optimal λ. This result demonstrate that the single training has slower learning as best λ is not the same as in Fig.2.

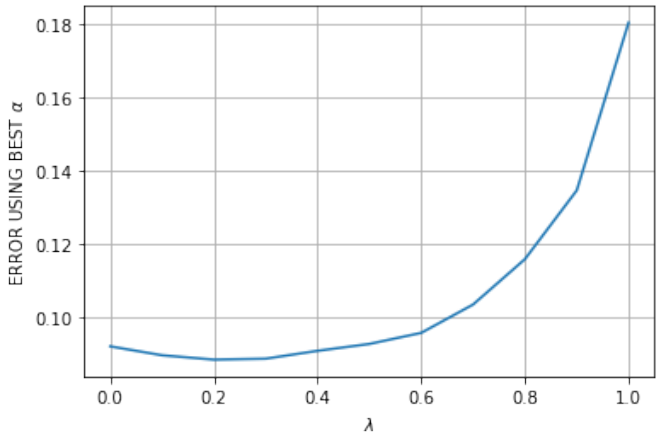

Fig. 4. Average error at best α value on random-walk problem.

C. Pitfalls

Sutton does not explicity present initial values of weight vector. Also as previously, the experiment assumes linear prediction.

VI. Conclusion

Three implementation results are very similar to Sutton's paper. However, the differences between Sutton's plots and this report's plot maybe due to initial value of weights ω, weight updating procedure, training set repeatation will significantly affects performance.

References

[1] Sutton, R.S , "Learning to predict by the methods of temporal differences," Machine Learning, 3, pages9–44(1988), 1988.