Desperately Seeking Sutton

----- Temporal Difference Learning

Manqing Mao
{GTID: mmao33}

1. Introduction

Temporal difference (TD) methods are a class of incremental learning procedures specialized for prediction problems. Whereas conventional prediction-learning methods are driven by the error between predicted and actual outcomes. TD methods are similarly driven by the error or difference between temporally successive predictions; with them, learning occurs whenever there is a change in prediction over time.

In this project, I first introduce the TD(λ) model with its weights update rule and a scenario – Bound Random Walk in Section 2. Next, I simulated TD(λ) with different weights update rules, which update weights (1) after each training set and (2) after each sequence in Section 3. I compare our results about Root Mean Square Error (RMSE) with results from Sutton's paper [1]. I also show how hyperparameters like α (learning rate) and convergence criterion have effect on RMSE. Different curves for different hyperparameters are also provided.

2. Background

2.1 TD(λ) Rule

Let us consider multi-step prediction problems in which experience comes in observation-outcome sequences of the form $x_1, x_2, x_3, ..., x_m, z$, where each x_t is a vector of observations available at time t in the sequence, and z is the outcome of the sequence. For each observation-outcome sequence, the learner produces a corresponding sequence of predictions P_1, P_2, P_3 .... , P_m, each of which is an estimate of z.

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

where α is learning rate. Since it is assumed that $P_t = w^T x_t$, so I can obtain that $\nabla_w P_k = x_k$. All learning procedures will be expressed as rules for updating w. For each observation, an increment to w, denoted Δw_t, is determined. So the update rule is:

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

2.2 Bounded Random Walks

All walks begin in state D. From states B, C, D, E, and F, the walk has a 50% chance of moving either to the right or to the left. If either edge state, A or G, is entered, then the walk terminates. The 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 ideal predictions for states B through F are then 1/6, 1/3, 1/2, 2/3 and 5/6,

respectively.

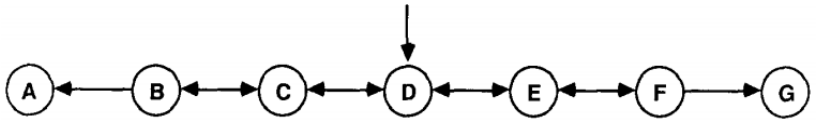


Fig. 1. A generator of bounded random walks.

3. Experimental Duplicates

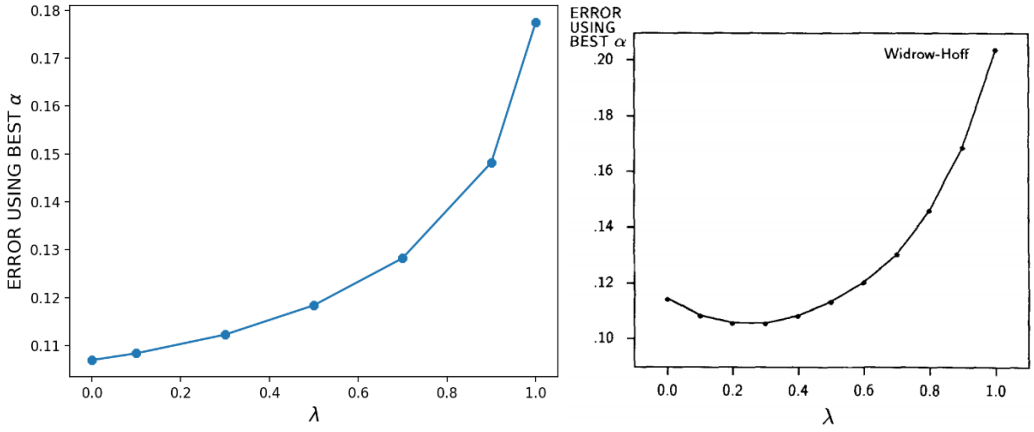
Two computational experiments were performed using observation-outcome sequences generated. In order to obtain statistically reliable results, 100 training sets with 10 sequences per training set were constructed.

3.1 Experimental I

A. Experiments Description

In the first experiment (*repeated presentation* training paradigm), I am going to find the relationship between the RMSE and λ. For each observation in a random walk sequence, I calculate Δw_t, which is accumulated over first all observations in one sequence and also for all sequences (=10) in a training set. Then, I update the weight vector after the complete presentation of a training set. Each training set was presented repeatedly to each learning procedure until the procedure no longer produced any significant changes in weight vector. In other words, I set the convergence criterion as the $\Delta w \leqslant 0.001$. Note that the Δw is a vector, so I can set the square root of its inner product $\leqslant 0.001$.

B. Results Comparison



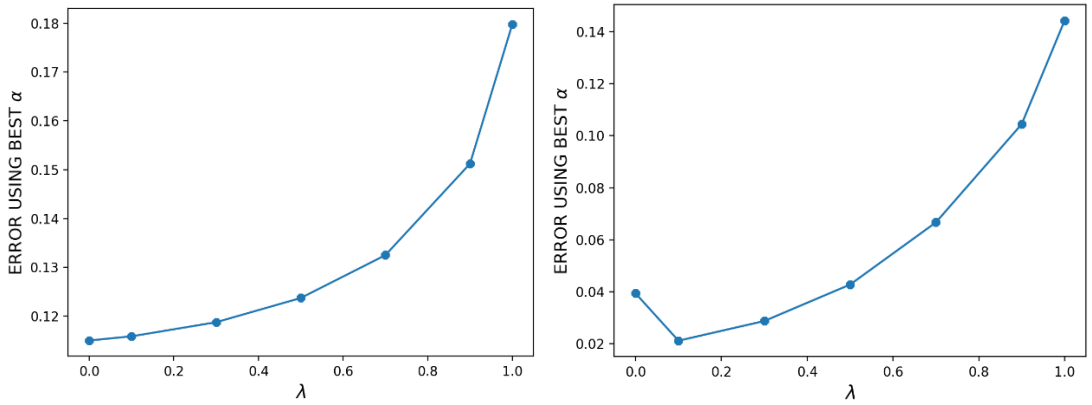
(a) (b)

Figure 2. Average RMSE as a function of λ in Experimental I. Comparison is shown between (a) My duplications and (b) Sutton's paper.

Figure 2 shows average RMSE as a function of λ. Fig. 2(a) is my duplication and Fig. 2(b) is the result from Sutton's paper. It shows that my duplication of the repeated presentations matches well the result from Sutton's paper, as shown in Fig. 2(b). However, I still find some minor differences between two figures above. (1) I obtain that average RMSE monotonically increases with increasing λ while the best λ corresponding to the lowest error is around 0.3 in Fig. 2(b). (2) Average RMS error in my duplication is lower compared with the Fig. 2(b).

Two possible reasons are the convergence criterion and the value of learning rate -- α. Those are called hyperparameters. Firstly, the convergence criterion determines the time to terminate the repeated presentations process. In paper [1], Dr. Sutton claims that the repeated presentation process terminates when there are no longer any significant changes in the weight vector. In our Experimental I, I set convergence criterion as the norm of vector w less than 0.001, which is expressed as the $\sqrt{\sum w_t^2} \leqslant 0.001$. It costs larger simulation time by using more stringent criterion. Secondly, the learning rate α determines whether the convergence or divergence. As mentioned in the paper [1], for small α, the weight vector always converged. So in my duplication, α is set as 0.005. After trying several values, I found that my algorithm cannot be converged when α > 0.02.

C. Hyperparameters Investigation & Pitfalls



(a) (b)

Figure 3. Average RMSE as a function of λ with different hyperparameters (a) larger α of 0.01 (b) looser convergence criterion ($\Delta w \leqslant 0.01$)

Figure 3 shows how different hyperparameters have effect on RMSE. Fig. 3(a) shows that larger α of 0.01 results in similar outcome as α of 0.005 did. So I can get similar results by picking small α when $\alpha \leqslant 0.02$. Thus, I am supposed to avoid the pitfalls of large α (> 0.02), which causes divergence. Fig. 3(b) shows that RMSE reduces by much and RMSE does not monotonically increase with increasing λ with looser convergence criterion ($\Delta w \leqslant 0.01$). So the convergence criterion is a more important factor to fit the result from Fig. 2(b). Also, I am supposed to avoid the pitfalls of too loose convergence criterion, which causes inaccurate results.

3.2 Experimental II

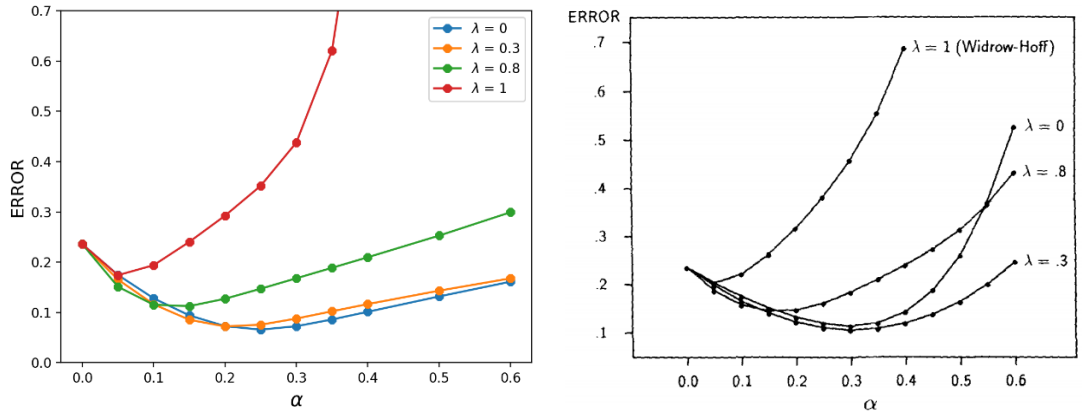
A. Experiments Description

In the second experiment, I am going to find the relationship between the RMSE and α with four varied λ values. I am going to set the value of λ to 0, 0.3, 0.8 and 1. Each training set was presented only once to each procedure instead of repeatedly until convergence. Weight updates are performed after each sequence instead of a whole training set. After Fig. 4 is obtained, I can get Fig. 5 by finding the smallest value of error with given λ.

B. Results Comparison

Figure 4 shows average RMSE as a function of α and λ. Fig. 4(a) is my duplication and Fig. 4(b) is the result from Sutton's paper. It shows that my duplication result matches very well with the result from Sutton's paper, as shown in Fig. 4(b). Some minor differences between two figures

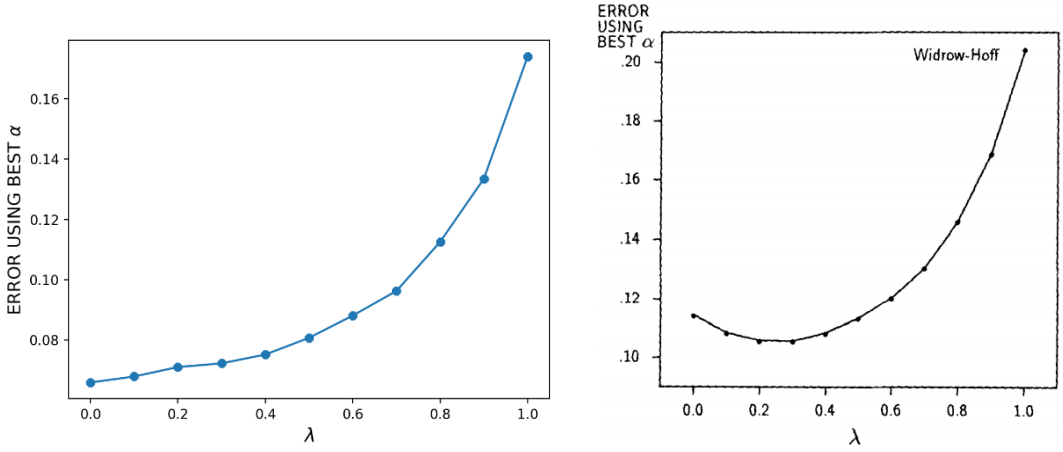
are shown as below: (1) With λ of 0, I did not get larger RMSE at large α compared with RMSE when λ is 0.3 or 0.8. (2) Average RMS error in my duplication is lower compared with the Fig. 4(b), which I got similar phenomenon in the Experimental I. It is worth to mention that RMSE is very large when α > 0.4 for TD(1). So we set the range of Y-axis from 0 to 0.7 to in Fig. 4(a).



(a) (b)

Figure 4. Average RMSE as a function of α and λ. Comparison is shown between (a) My duplications and (b) Sutton's paper.

Similarly, I got very similar trend in Fig. 5 (a) as I got in Fig. 2(a). From the Fig. 5(a), we still can find that (1) I obtain that average RMSE monotonically increases with increasing λ and (2) average RMS error in my duplication is lower compared with the Fig. 5(b).



(b) (b)

Figure 5. Average RMSE as a function of λ in Experimental II. Comparison is shown between (a) My duplications and (b) Sutton's paper.

Another two possible reasons beyond the hyperparameters are higher precision data type and limited number of training samples. Nowadays, we have higher precision data type (64-bit). Also, the samples are quite small – only 100 training sets maybe another reason to cause the difference.

Reference:

[1] R. S. Sutton. "Learning to predict by the methods of temporal differences", In Machine Learning, pp. 9-44, 1988.