CS 7642: Reinforcement Learning and Decision Making
Project 1: TD(λ)

Thomas Kim
tkim338
git hash: db3d913a8eff51d984a24a17430e3d83fda5a943

Problem

Conventional prediction-learning methods compare predictions and actual outcomes to assign weight to model parameters and refine its predictions. Temporal difference methods compare differences between temporally successive predictions (Sutton '88) to assign these weights. Temporal difference learning has been in use in real-world problems for a significant amount of time and has been shown in practice to produce more accurate predictions than conventional methods, but are not well understood. Sutton '88 provides a proof of temporal difference learning's convergence and optimality and presents a set of experiments to empirically support his argument as well as provide intuition to better understand the key concept. Here, we attempt to replicate these experiments and results to provide another perspective of his argument.

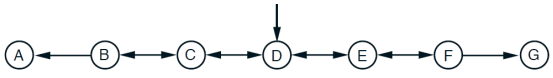

Figure 2. A generator of bounded random walks. This Markov process generated the data sequences in the example. All walks begin in state D. From states B, C, D, E, and F, the walk has a 50-50 chance of moving either to the right or to the left. If either edge state, A or G, is entered, then the walk terminates.

Figure 1: Bounded random walk diagram from Sutton '88.

These sets of experiments are conducted using this particular case of a random walk, shown in Figure 1. As described in the caption from Sutton '88, the process in this case is a sequence of states (A through G), which always begins in state D and terminates in state A or state G. In each state from B to F, the walk has a 50% chance of moving in either direction. In the terminal states A and G, the walk terminates with a reward of 0 for state A and 1 for state G.

Given a sequence of states, a conventional prediction-learning method could be implemented to compare each non-terminal state in the sequence to the final (terminal) state. For each state in a sequence, this method would increase the value of said state by some value determined by the difference between the reward/value of the terminal state of the sequence (1 for state G and 0 for state A) and the current state in the sequence.

In temporal difference learning, each non-terminal state in the sequence is compared with the final state along with the all of the non-terminal states that lead to the final state. For example, state F has a 50% chance of leading to state G, which has a value of 1, which means that state F has a value that isn't necessarily 1, but closer to 1 than to 0 (the minimum possible value in this model). In temporal difference learning, this value of F is used as a factor in determining the value of state E, which has a 50% change of leading to state F, a valuable state, although not as valuable as state G. In conventional learning methods, the value of state F would not influence the value of state E, and only the value of state G would have an effect on the value of state E.

This temporal difference learning logic is applied along the whole sequence of states, which results in a back-propagation of information through the sequence at a faster rate than conventional learning, which uses just the terminal state information and ignores sequential differences en route to the terminal state.

Temporal difference learning can be generalized to include conventional learning methods with the variable λ, as described in Equation 1:

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

where:
Δw_t is the increment to be applied to weight w_t
α is the learning rate
P is the prediction
t is the current time
k is the number of steps in the past
λ is a scaling factor from $0 \leq \lambda \leq 1$

In essence, a λ of 1 results in considering just the final state to assign values to intermediate states (equivalent to conventional learning), while a λ of 0 results in considering just directly neighboring states to assign values to each state (an extreme case of temporal difference learning). λ of values between 0 and 1 effectively scales how much consideration is placed on states between the current and final states to determine value. The following experiments demonstrate the performance of temporal difference models with varying λ in the special case of the bounded random walk described above.

Experiments

Ground truth for the bounded random walk was first verified before conducting experiments. Sutton '88 claims this to be [0, 1/6, 1/3, 1/2, 2/3, 5/6, 1]. To verify this, a maximum likelihood model was implemented and iterated until convergence was reached. This model simply iterated through each state and adjusted its value based on the value of each neighboring state, weighted by probability (in this case, two neighbors, each with 50% probability). The model reached final state values equal to Sutton's claim. The results of this verification test is shown in Figure 2.

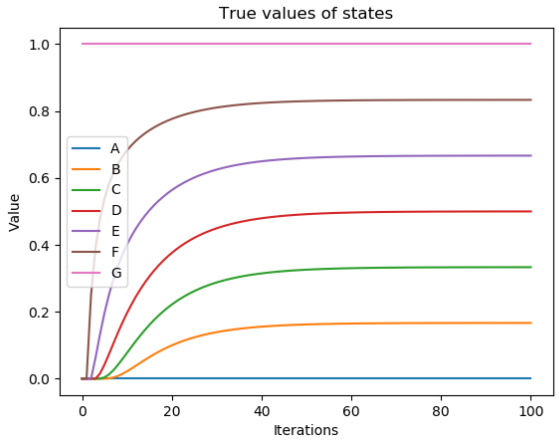

Figure 2: Empirical estimation of ground truth.

Experiment 1

The first experiment demonstrates a key difference between temporal difference learning and the Widrow-Hoff procedure (a conventional learning method): that although the Widrow-Hoff procedure does minimize error on the training set it's given, it's not necessarily the optimal solution for minimizing error for future experience. In this experiment, 100 sets of 10 sequences each of the bounded random walk are randomly generated. Each set of 10 sequences is applied to the learning model until convergence is reached. Sutton calls this a *repeated presentations* training paradigm. Estimated values of each state are updated only after the entire set of sequences are presented. RMS error is computed against the known ground truth after estimates converge for each set of sequences and the average of these 100 RMS errors is computed and plotted in Figure 3. The original figure from Sutton '88 is shown in Figure 4.

My replication of Sutton's figure matches fairly well in terms of overall trend, with a roughly exponential increase in error as λ increases, with minimum error when $\lambda = 0$ and maximum error when $\lambda = 1$. This figure illustrates that a temporal difference learning algorithm reaches predictions that have less error than the conventional Widrow-Hoff algorithm. Sutton explains this by stating that the Widrow-Hoff procedure minimizes error on the training set, but not necessarily for future experience. Intuitively, it can also be explained that by considering the values of non-terminal, adjacent states, the temporal difference makes use of more of the available information from the same set of data than does the Widrow-Hoff procedure.

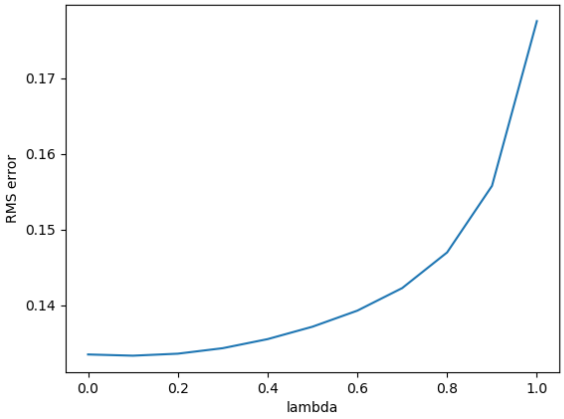

Figure 3: Replication of Figure 3 from Sutton '88.

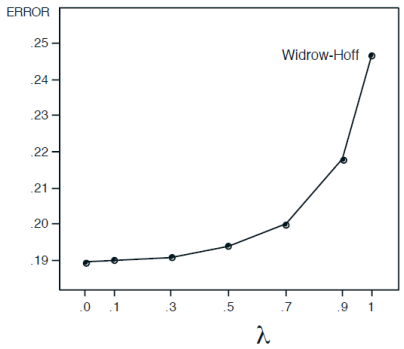

Figure 3. Average error on random walk problem under repeated presentations. All data are from TD(λ) with different values of λ. The error measure used is the RMS error between the ideal predictions and those found by the learning procedure after being repeatedly presented with the training set until 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.

Figure 4: Original Figure 3 from Sutton '88.

However, there is a difference in absolute error, where the error found in this replication was roughly 30% lower than Sutton's findings. Assumptions were made about the initial estimated values ([0, 0, 0, 0, 0, 0, 1]), the threshold for convergence (maximum change of 0.0001), and learning rate (0.01), and final error was found to be somewhat sensitive to these variables. Figure 5 serves to show the model's sensitivity to changes in these assumptions. Different sets of training data were tested by using different random number generator seeds and absolute error was found to be moderately sensitive to changes in training data, while the shape of the error curve was not. Figure 6 shows the effect of a

few sets of random training data on error. 100 sets of 10 sequences is a relatively small amount of training data, so it's possible that at least some of this difference in error is due to random chance.

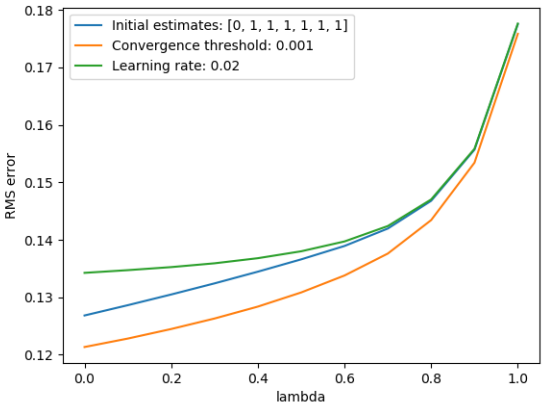

Figure 5: Replication of Figure 3 to gauge sensitivity to initial estimates, converge threshold, and learning rate.

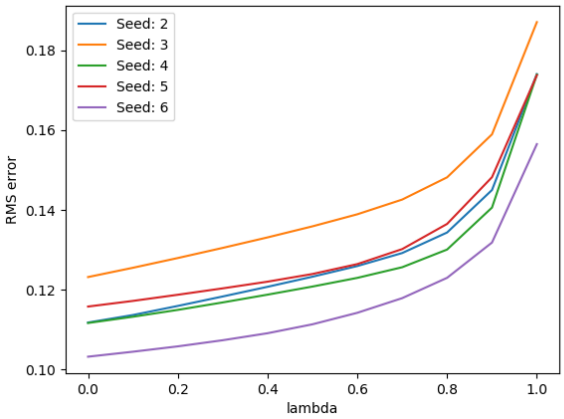

Figure 6: Replication of Figure 3 to gauge sensitivity to random training data.

This limitation on the size of training data used does serve a specific purpose, though, other than reducing computation time and space. With a large set of training data that are a closer representation of the random walk model, the Widrow-Hoff procedure would likely reach a set of predictions with less error than a temporal difference learning model (based on Sutton's earlier claim that the Widrow-Hoff procedure minimizes error on training data), even if the Widrow-Hoff procedure might take longer to converge on a solution.

Experiment 2

This second experiment examines the effect of α and λ on error. This experiment was conducted in a different manner than the first experiment: instead of using a repeated presentations training paradigm, this experiment used a more conventional single presentation of each set and sequence of states, while updating predictions after each sequence, rather than waiting until a full set was presented. Because sets of data were not presented repeatedly until predictions converged, this model was much more sensitive to learning rate α. A high learning rate causes predictions to change more quickly with each presentation of data, which can cause predictions to reach an optimum more quickly, but can also cause predictions to also diverge if learning rate is set too high. Sutton's textbook provides a helpful illustration of how a high learning rate can converge more quickly but result in more steady-state error than a low learning rate might (Figure 7).

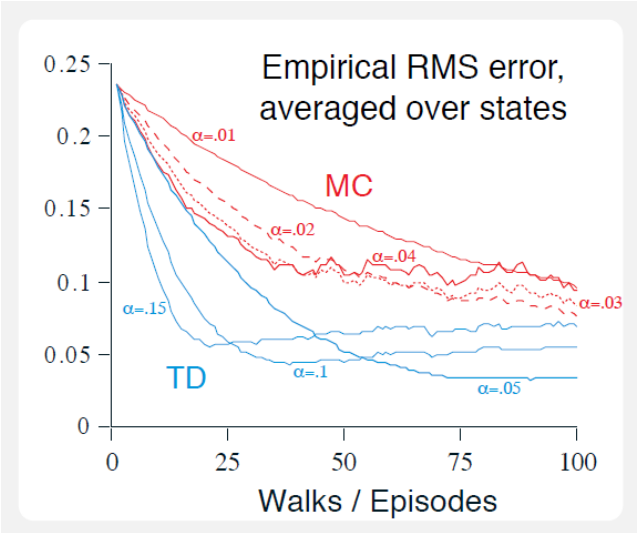

Figure 7: Illustration of effects of learning rate on error from Sutton and Barto '20.

The results of this experiment are shown in Figure 8. Four distinct values of λ were chosen and predictions were computed for a range of α. Error was computed as it was in Experiment 1, where the plotted error was calculated as the average error for the 100 sets of data presented to the model. The original figure from Sutton '88 is shown in Figure 9. The replication of the figure, again, matches fairly well in overall trend, but varies in how quickly error diverges at high α. The same λ values were used, along with the same range for α, and initial predictions of 0.5 for all states. This model show some sensitivity to the training data set used, as it did in Experiment 1.

Experiment 3

The third experiment builds on the results of the second experiment by taking the minimum errors produced by each combination of λ and α and plots them together, the results

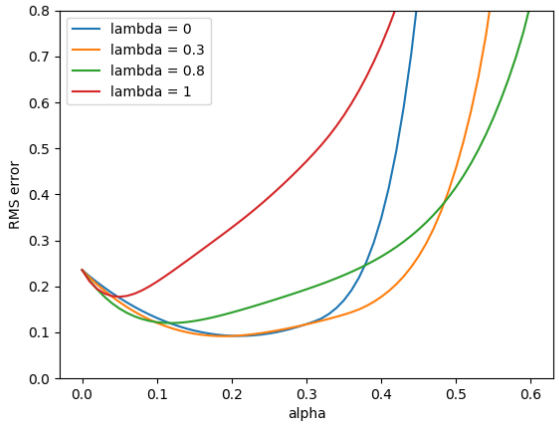

Figure 8: Replication of Figure 4 from Sutton '88.

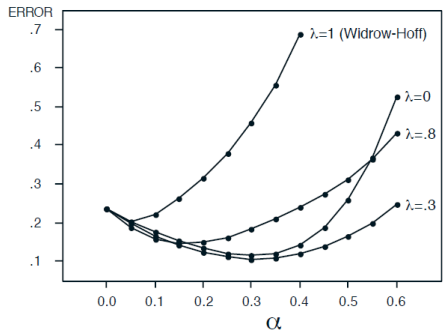

Figure 4. Average error on random walk problem after experiencing 10 sequences. All data are from TD(λ) with different values of α and λ. The error measure used is the RMS error between the ideal predictions and those found by the learning procedure after a single presentation of a training set. This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

Figure 9: Original Figure 4 from Sutton '88.

of which are shown in Figure 10. Figure 11 shows the original plot in Sutton '88. The same differences seen in Experiment 2 are reflected here, as they show the same data points. The overall trend between the results generated here and Sutton's original figure is still comparable, with a minimum error at a λ value between 0.1 and 0.4 and error increasing to a maximum when λ is 1.

There is a notable difference in the error at low values of λ. This is not particularly surprising, as this difference can be seen from Figure 8, where the minimum error for each curve falls lower than in the original figure from Sutton '88. This difference can be traced back to the first experiment as well, where the error found here was also less than what Sutton found in his paper.

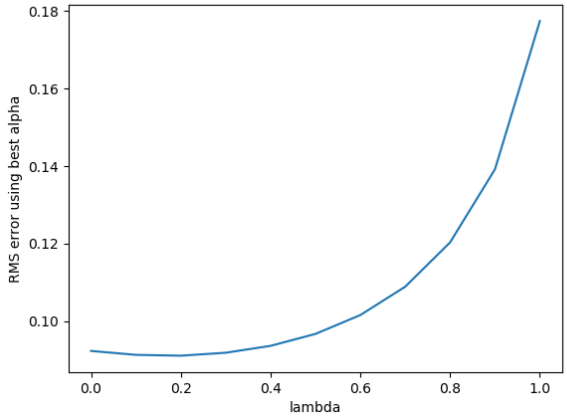

Figure 10: Replication of Figure 5 from Sutton '88.

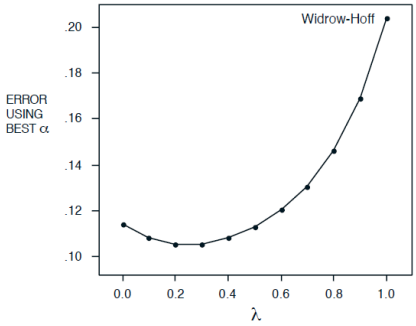

Figure 5. Average error at best α value on random walk problem. Each data point represents the average over 100 training sets of the error in the estimates found by TD(λ), for particular λ and α values, after a single presentation of a training set. The λ value is given by the horizontal coordinate.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

Figure 11: Original Figure 5 from Sutton '88.

References

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction.* 2nd Ed. MIT press, 2020. url: http://incompleteideas.net/book/the-book-2nd.html.

Richard Sutton. "Learning to Predict by the Method of Temporal Differences". In: *Machine Learning* 3 (Aug. 1988), pp. 9-44.