1. This is my final project for the reinforcement learning class. I applied TD-lambda method on five-step random walk problem to demonstrate the performance of TD-lambda.
2. The Random walk problem is like this: an agent always starts at state D. When the agent is in State B, C, D, E, or F, it has 50% chance of walking to the left and 50% chance of walking to the right. The walk ends at A or G. The agent receives a reward of 1 when it reaches state G and no reward for any other states. The goal here is to learn the probability of the agent ending up at state G for each states. The predication error is used for evaluating TD-lambda.
3. One thousand sequences were generated and grouped into a hundred training sets for two computational experiments. In the first experiment, the learning procedure would only update the weights after it saw all the 10 sequences in a given set, and this set was repeatedly presented to the learning procedure until the weights converge. The figure on the right side showed the averaged error of weights. As we can see, the error increases as lambda increases and TD(0) has the best performance.
4. This figure actually matches well with the figure 3 in the original paper. But the numerical values of error are not exactly the same. We will discuss this later.
5. In experiment 2, the learning procedure only saw each set once. It also updates the weights after seeing each sequence. Multiple learning rate were used. As shown in the figure, the impact of learning rate on the performance of TD(lambda) is evident. The best alpha seems to be around 0.2.
6. Again, the trends still match but the values don’t. the best alpha is around 0.3 in the paper.
7. After getting the best alpha for each lambda, the performance of TD-lambda was again evaluated. As we can see here in the figure, TD(0) is not the best performer anymore and TD(0.2) is kind of the best. And TD(1) still has the worst performance.
8. This figure does not match the one in the paper very well but the trends still show.
9. To summarize, my replication results supports the claims in the paper but the figures does not match perfectly. The differences, however, are mostly numerical. There might be two reason for the differences: first, the training sets were randomly generated. Second, parameters such as alpha and epsilon were not reported in the paper and I have to experiment with it. This concludes my presentation, thanks for watching.