Vince Petaccio
www.github.com/vjp23

TD(λ) Learning and Random Walks

In his 1988 paper entitled *Learning to Predict by the Methods of Temporal Differences*, Richard Sutton describes a class of learning procedures which seeks to make predictions using past experience with an incompletely known system. The method employs a prediction function with functional dependence on experienced input observations *x* and a set of learned weights *w*, and a learning process which updates this set of weights incrementally with experience using step-wise increments between predictions P:

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

where **λ** is a value between 0 and 1 inclusive which effectively acts as a discount rate on past observations; by varying the value of **λ**, the weighting of past observations relative to recent observations can be controlled. The gradient of P_k with respect to *w* represents the steepest rate of ascent along the derivative of $P(x_t, w)$, and in a linear function this is simply x_t. The value *t* represents the number of steps taken at runtime, and **α** is a learning rate.

In Sutton's 1988 paper, he describes an experiment designed to evaluate the effectiveness of this learning procedure by varying the value of **λ** while learning the predicted value of each step in a bounded random walk. In this model, uniformly random steps to the left or to the right are taken from an initial starting state until a terminal (absorbing) state is reached. If the terminal state on the right end is reached, there is a positive outcome of 1, and the terminal state on the left returns an outcome of 0. The walk task is depicted below in an image based upon figure 2 in Sutton's paper.

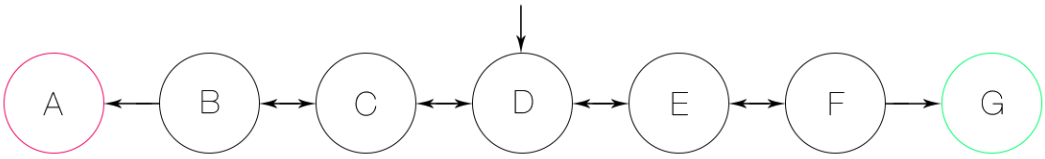

Figure 1- The bounded random walk task. For these experiments, all sequences began in state D, and then transitioned left or right with uniformly random steps (P(left) = P(right) = 0.5) until arriving at absorbing state A or absorbing state G. The outcome of state A was 0, while the outcome for G was 1. Based on Figure 2, Sutton (1988).

Three experiments employing the bounded random walk carried out by Sutton sought to explore the effects of varying both the learning rate **α** and the parameter **λ**, and of repeated versus single presentations of training data to the learner. In the first experiment, **α** was held constant and **λ** was varied during repeated presentation of 100 sets of training data, with each set containing 10 bounded random walk sequences. The training data was presented to the learner until the learner's weight update values converged, with convergence defined by a threshold value for the size of the updates. After collecting the weight vectors, the predictions that the learned weights produced were equal to the weights themselves, as the *x* values were vectors of zeros with a single one value, resulting in $P(x_t, w)$ as $w^T x_t$.

After generating the weights vectors with varied values of **λ**, the results were compared against the known actual probabilities from each state, which were [1/6, 1/3, ½, 2/3, 5/6] with the error measure reported as root mean squared error (RMSE) between the two. The second experiment explored the effect of four **λ** values across all values of **α**, again reporting error as RMSE. The third experiment showed the effect of varied values of **λ** for each value's best value of **α**. Experiments two and three functioned with only a single presentation of 10 sequences of training data, averaged over 100 trials. The results of each experiment are shown in the original figures from Sutton (1988) below:

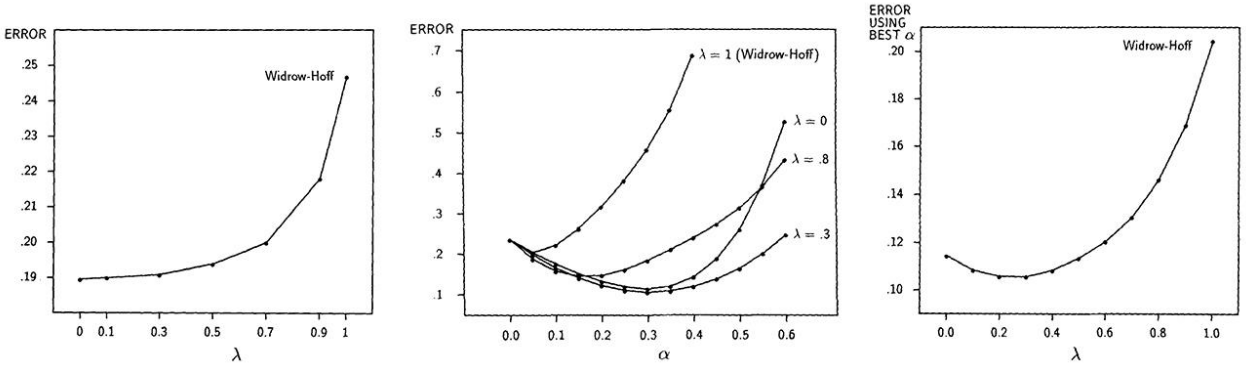

Figure 2- From left to right: Figures 3, 4, and 5 from Sutton (1988), depicting the results of three experiments performed using TD(λ) methods to learn the predicted values of each state in the bounded random walk task.

The goal of this work was to replicate the results that Sutton obtained in the figures shown above. In order to do this, a Python function was written to generate uniformly random steps for the training data. The TD(**λ**) procedure was then implemented in a second function which performed the updates to the weights vectors. As in Sutton's work, the weights were initialized to a vector with each value equal to 0.5. After performing the updates as described in Sutton's work, the figures shown above were replicated. The results are shown below.

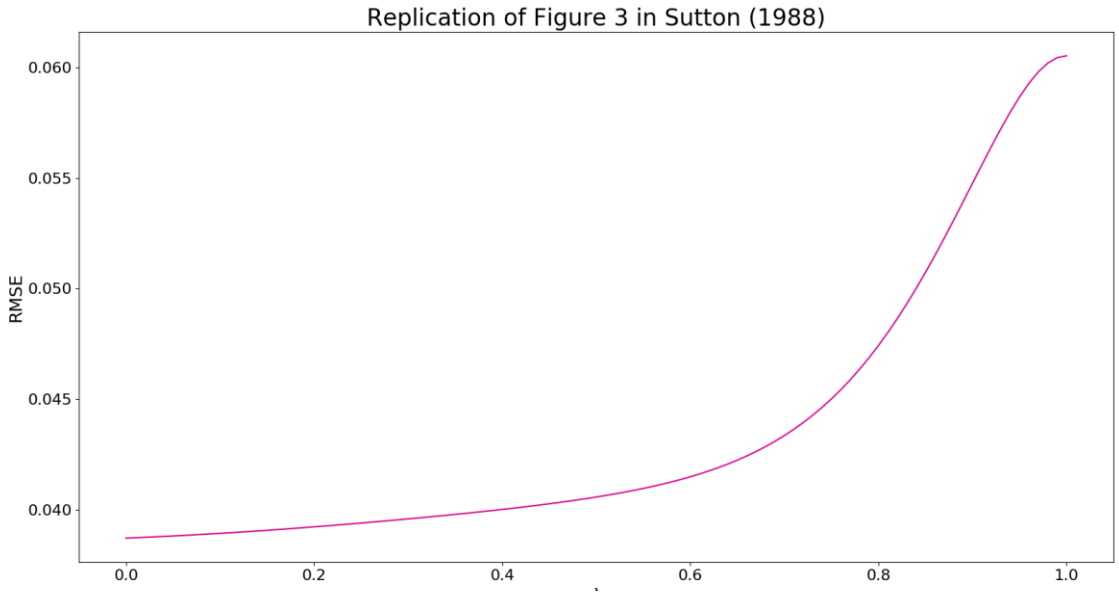

Figure 3- Replication of figure 3 in Sutton (1988)

Of note is the significant variation in RMSE between Sutton's figure 3 and that replicated below. Sutton's work does not indicate the stopping criteria used to identify convergence; in this work, the training set was presented 20 times. This variation may arise from early termination in Sutton's procedure, resulting in suboptimal RMSE values. In addition, since the TD(1) procedure learns the training data with little generalization, it is sensitive to overfitting and therefore to variance within the input training data sequences.

The replication of Sutton's figure 4 shows a similar pattern of growth for each value of **λ**, though the present results appear shifted by an **α** value of about 0.1. This may arise from a difference in implementation for the first ("0th") temporal difference calculation.

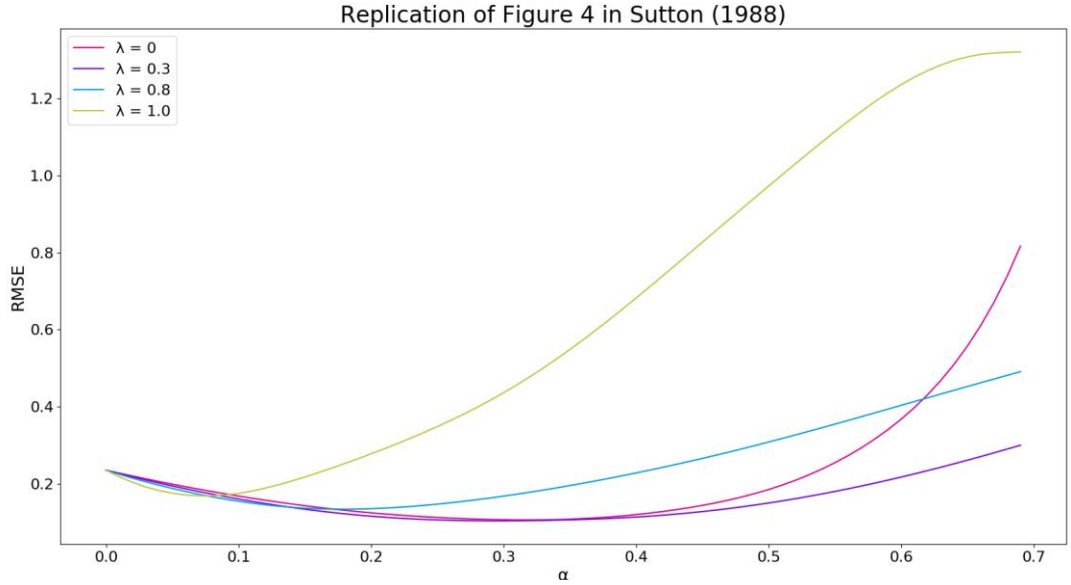

Figure 4- Replication of figure 4 in Sutton (1988)

Finally, the replication of Sutton's figure 5 below is nearly identical to the original plot. This suggests that both implementations represent a theoretical truth about the fundamental theory underlying the TD(**λ**) procedure.

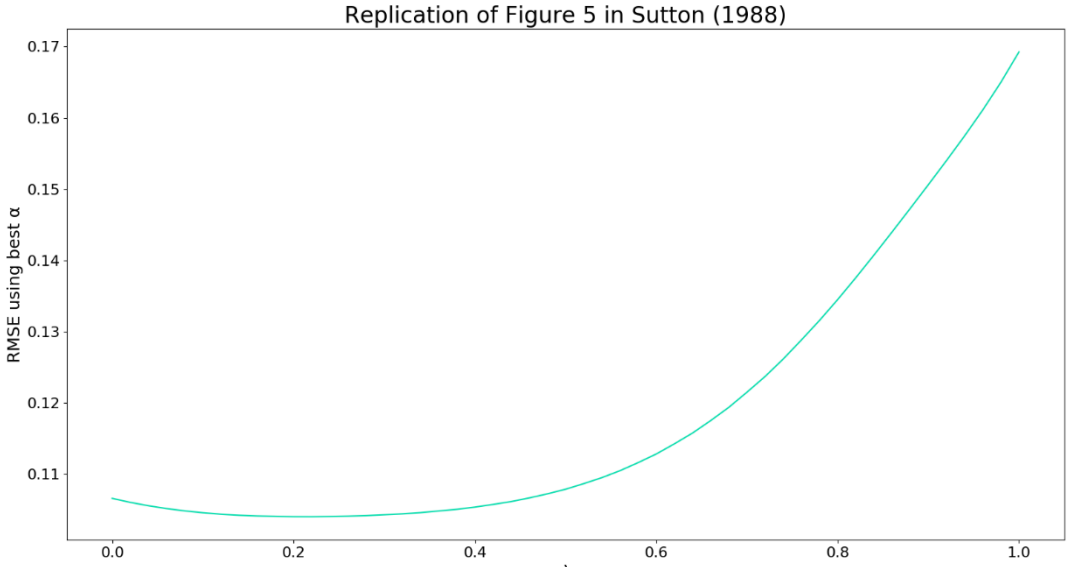

Figure 5- Replication of figure 5 in Sutton (1988)