This assignment consisted of implementing on-line TD algorithms with linear function approximation. The testbed for this assignment was an MDP popularly known as 'Baird's counterexample'.

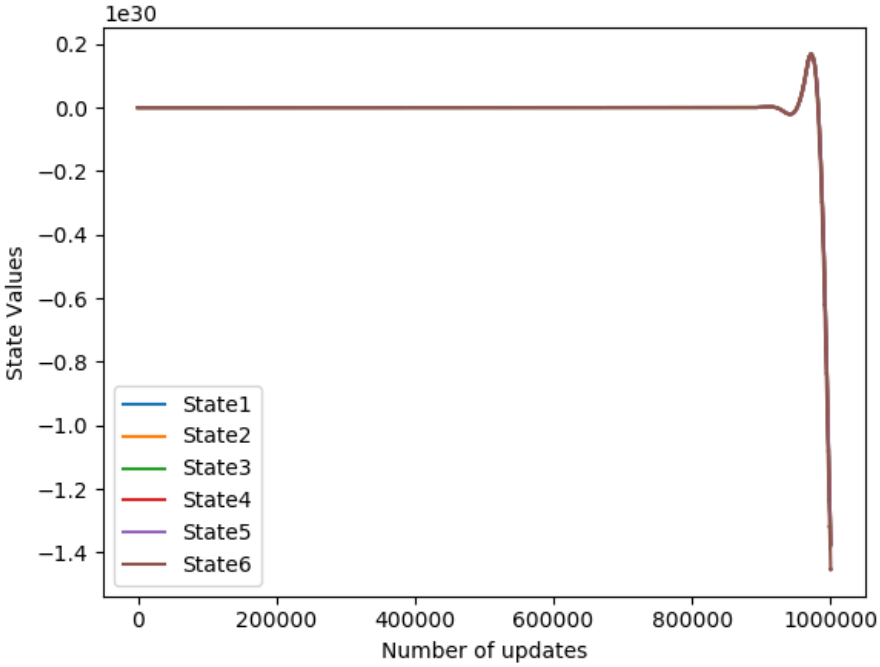

FIGURE 1. Experiment 1: Estimated states values vs Number of Updates for 1,000,000 steps. **Please note the scale on the y-axis which is 1e30**

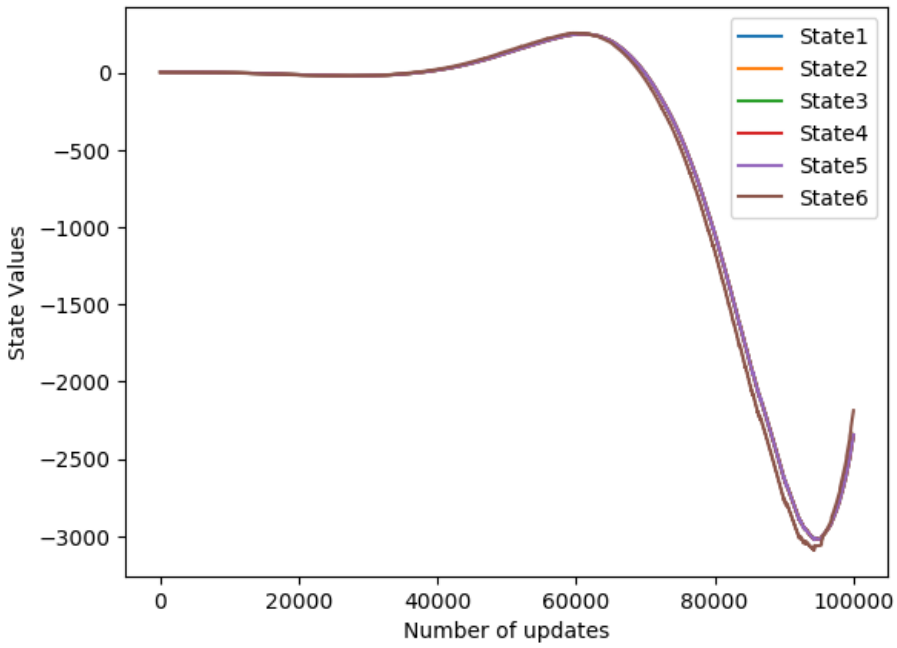

FIGURE 2. Experiment 1: Estimated states values vs Number of Updates for 100,000 steps

In experiment 1, as clear from figure 1 and 2, the state value estimates are diverging as we increase the number of updates. They first diverge to $+\infty$, then starts diverging to $-\infty$. The divergence is due to the fact that the distribution of updates does not match the on-policy distribution. Intuitively, as described in Sutton & Barto[1] in this case, for an action, the promise of

[1]Please refer to page 212 in the latest draft of Sutton and Barto

future reward can be made and then, after taking an action that the target policy never would, forgotten and forgiven. Due to symmetry, the value function estimate of the first 5 states are approximately the same and the 6^{th} state value is only somewhat different from other state values, because of the $\epsilon = 0.01$ probability of ending up at the terminal state.

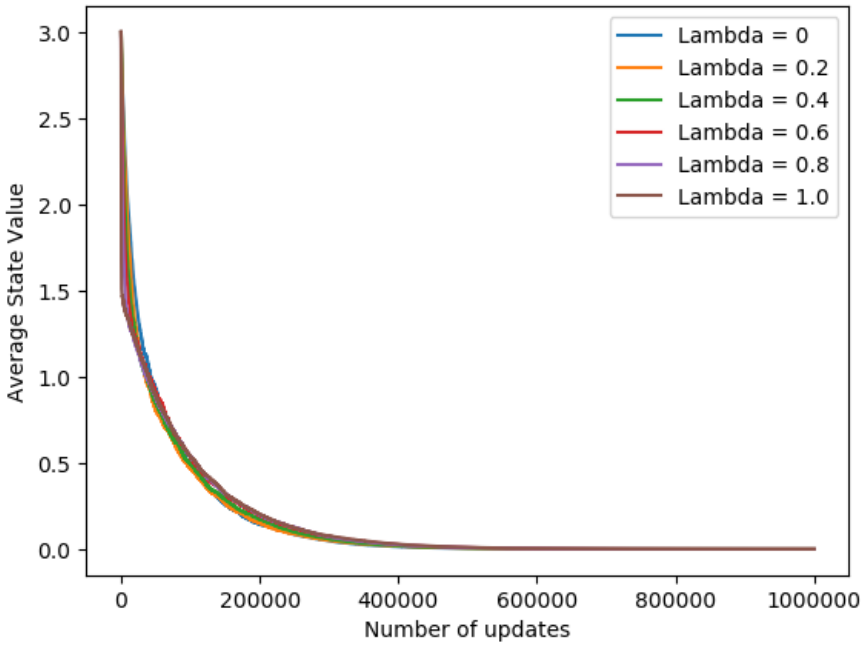

FIGURE 3. Average Estimated states values vs Number of Updates for TD(λ) for different values of λ.

In experiment 2, I observed that the magnitude of weights was constantly decreasing with number of updates and the weights finally converged quite close to zero, leading to a value function which is approximately zero for all the states. This was true for all the values of λ. Since, in this experiment, we are running on-policy TD(λ) which can be proved to converge to the TD *fixed point* similar to the proof of semi-gradient TD(0) [2].

In experiment 3, depending on the initialization, the number of updates required for convergence can change, however, the weights were converging to same values (which is approximately zero for all weights), for all initializations. This also implies that the state values were converging to zero for all initializations. This experiment is simply a on-policy semi-gradient TD(0), in which can be proved that the weights will always converge to the TD *fixed point*.

[2]Refer to page 168 in the latest draft of Sutton and Barto