A Case Study of TD(0) and Monte Carlo Prediction

Charles C Onu
260663256

The environment: Random walk

- Always starts at C
- Policy: equiprobable random actions - left or right
- Rewards: +1 if terminate on right. 0 if terminate on left.
- Optimal state values: A, B, C, D, E are 1/6, 2/6, 3/6, 4/6, 5/6 respectively (if task is undiscounted)

Objectives

1. Compare performance of first-visit vs every-visit monte carlo in predicting state values
2. Observe impact of α in Monte Carlo prediction
3. Observe impact of α in alpha in TD(0) prediction
4. Compare performance of TD(0) and Monte Carlo prediction

Quick Reminder

Monte Carlo:

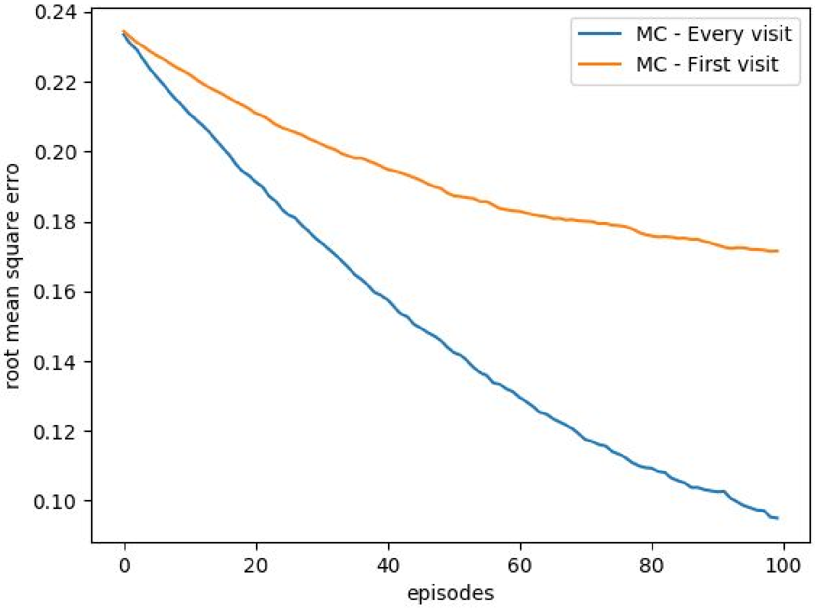
$$V(S_t) \leftarrow V(S_t) + \alpha\Big[G_t - V(S_t)\Big]$$

TD(0):

$$V(S_t) \leftarrow V(S_t) + \alpha\Big[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\Big]$$

Experiment 1: First-visit vs Every-visit Monte Carlo

- First-visit and Every-visit Monte Carlo prediction were implemented
- Both were compared for several values of α (only α = 0.01 shown)
- Every visit MC consistently performed better



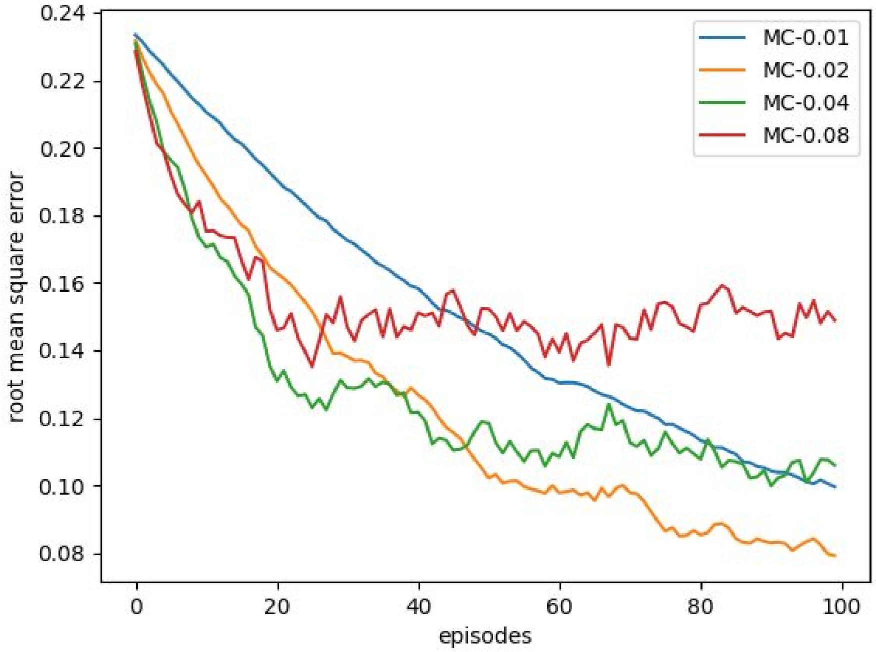
Values were averaged over 100 iterations

An episode:

[(C, 0, D) → (D, 0, C)→ (C, 0, B)→ (B, 0, C)→ (C, 0, B)→ (B, 0, C)→ (D, 0, E)→ (E, 1, T)]

Experiment 2: Monte Carlo for varying alpha values

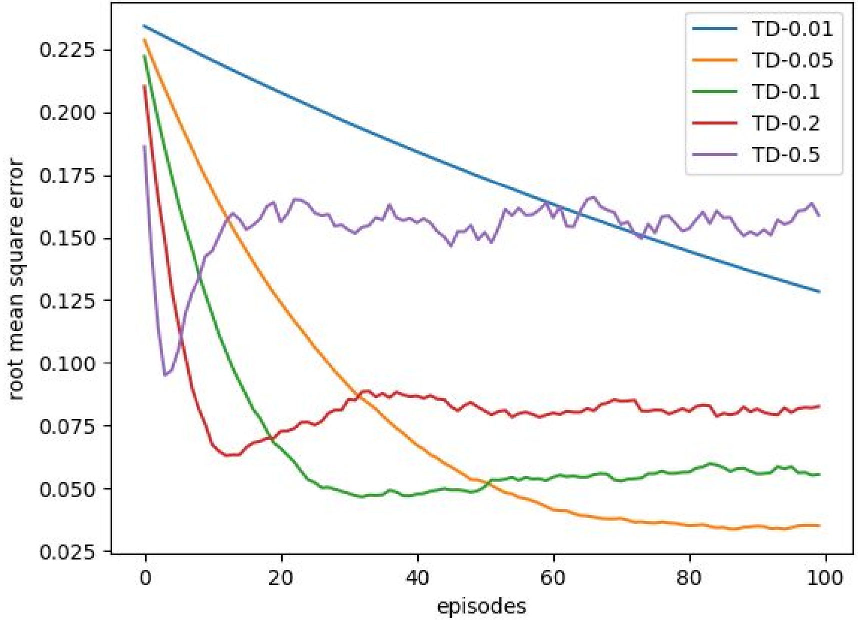
- Every-visit monte Carlo was compared for varying values of α.
- As α is decreased, it takes a longer time to converge. When it does, a better estimate is obtained.



Values were averaged over 100 iterations

Experiment 3: TD(0) for varying alpha

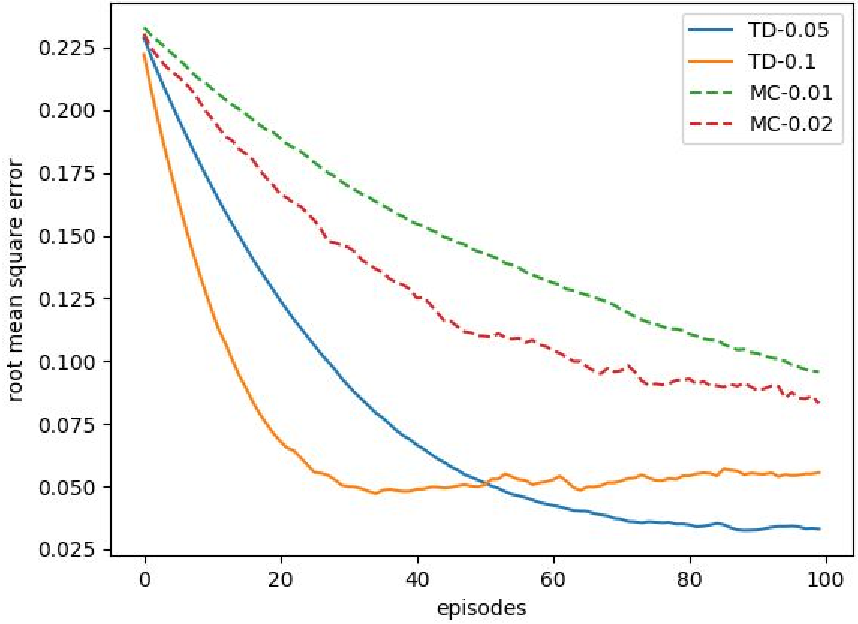
- TD(0) was implemented and evaluated for varying values of α.
- As α is decreased, it takes a longer time to converge. When it does, a better estimate is obtained.



Values were averaged over 100 iterations

Experiment 4: Monte Carlo vs TD(0)

- The most stable alpha values for (every-visit) Monte Carlo and TD(0) were compared
- TD(0) was found to perform consistently better than Monte Carlo



Values were averaged over 100 iterations

Conclusions

- Because each state could potentially be visited multiple times in an episode, every-visit Monte Carlo does a better job at estimating the state values
- As seen with TD and MC, as learning rate increases convergence is faster but not necessarily to a better estimate (and may become unstable)
- As with most empirical studies, TD(0) converges faster and obtains better estimate of the state value in the short-run.

Code:
https://github.com/rllabmcgill/rlcourse-february-24-onucharles