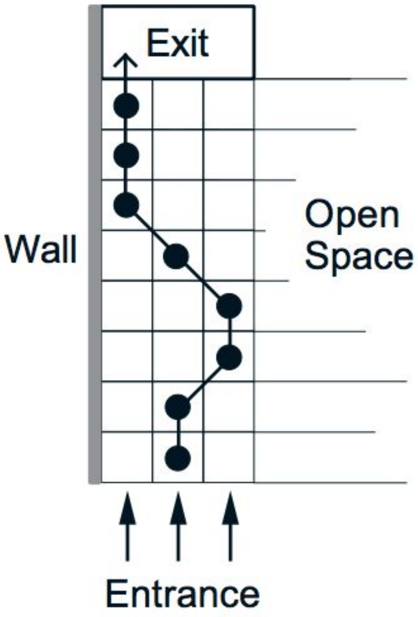
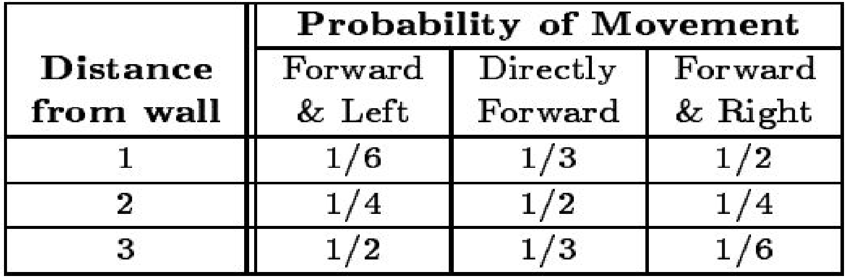
Comparison of different learning rates in planning with TD(0)

Sumana Basu and Charles C Onu

The wall-following domain



- Robot starts from one of the bottom three states
- Works its way up one row at a time
- Trial terminates if
  - robot runs into the wall
  - wanders off into the "open space"
  - successfully reaches the exit



Distance from wall	**Probability of Movement**		
	Forward & Left	Directly Forward	Forward & Right
1	1/6	1/3	1/2
2	1/4	1/2	1/4
3	1/2	1/3	1/6

Goal

- Apply TD(0) to evaluate the given policy
- Compare learning curves for different values of learning rate (α)

Procedure

1. Under the given policy generate 50,000 episodes
2. Estimate the “true” value function using TD(0) to learn from all 50,000 episodes (should converge to “true” in the limit)
3. Given a finite number of episodes (4,000), how long does it take TD(0) for α = {0.01, 0.1, 0.5} to find the true value function of the policy?

TD(0)

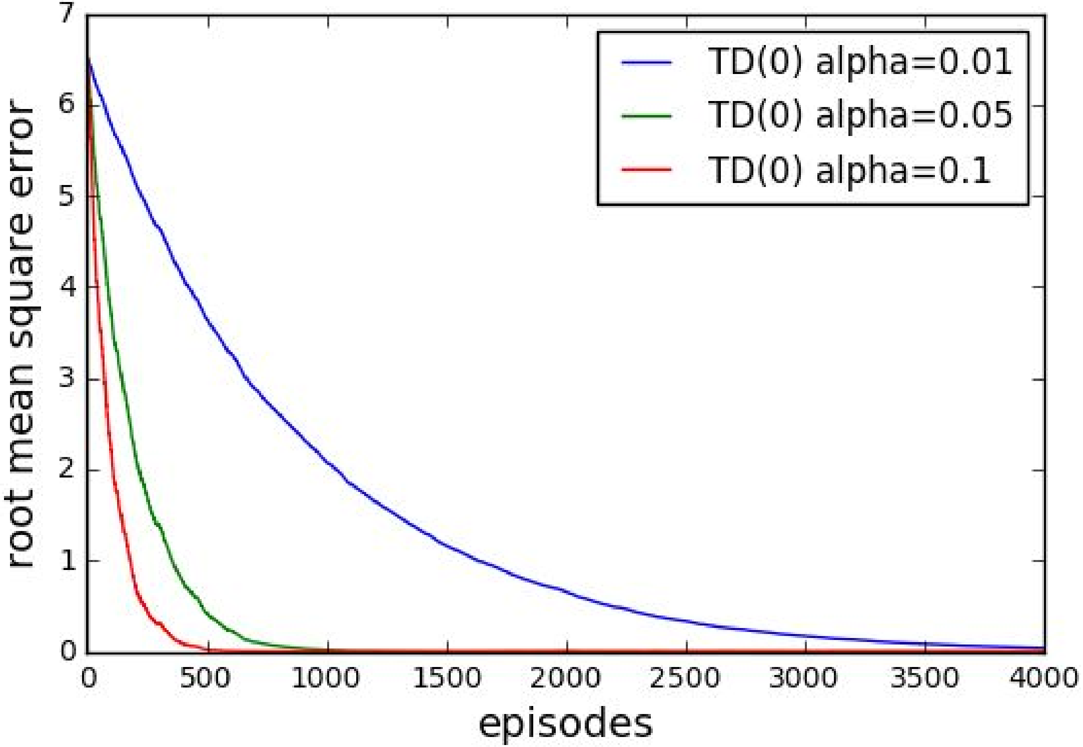
Update $V(S_t)$ towards a bootstrap estimate G_t:

- $V(S_t) \leftarrow V(S_t) + \alpha(G_t - V(S_t))$

Where G_t is:

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

Results



TD(0) alpha=0.01
TD(0) alpha=0.05
TD(0) alpha=0.1
root mean square error
0
1
2
3
4
5
6
7
0
500
1000
1500
2000
2500
3000
3500
4000
episodes