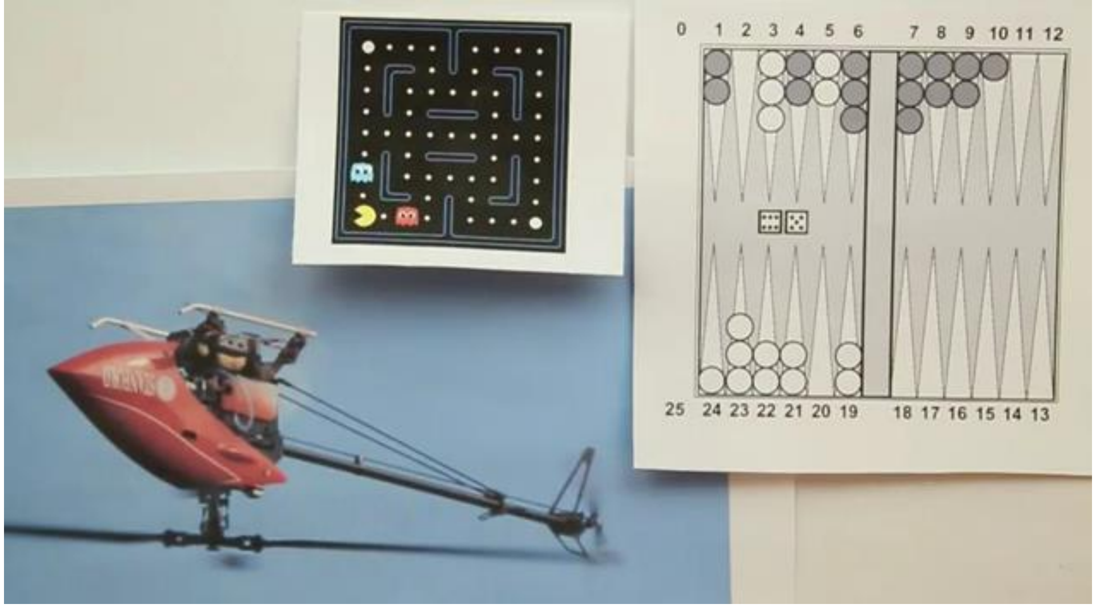
Computational Intelligence

Reinforcement Learning 11-2



Acknowledgement

- Several examples of this lecture have been taken from Stanford AI class and Stanford Machine Learning class.

Types of online learning

- Episodic
- Continuous

Types of learning

- *Episodic:* Collecting the rewards **at the end of the episode** and then calculating the **maximum expected future reward**.
- *Temporal Difference Learning*: Estimate **the rewards at each step**

Driving Home Example

- Each day as you drive home from work, you try to predict how long it will take to get home. When you leave your office, you note the time, the day of week, and anything else that might be relevant. Say on this Friday you are leaving at exactly 6 o'clock, and you estimate that it will take 30 minutes to get home. As you reach your car it is 6:05, and you notice it is starting to rain. Traffic is often slower in the rain, so you reestimate that it will take 35 minutes from then, or a total of 40 minutes. Fifteen minutes later you have completed the highway portion of your journey in good time. As you exit onto a secondary road you cut your estimate of total travel time to 35 minutes. Unfortunately, at this point you get stuck behind a slow truck, and the road is too narrow to pass. You end up having to follow the truck until you turn onto the side street where you live at 6:40. Three minutes later you are home.

R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction

Driving Home Example

- The sequence of states are as follows:

	Elapsed Time	*Predicted*	*Predicted*
State	*(minutes)*	*Time to Go*	*Total Time*
leaving office, friday at 6			
reach car, raining			
exiting highway			
2ndary road, behind truck			
entering home street			
arrive home			

R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction

Driving Home Example

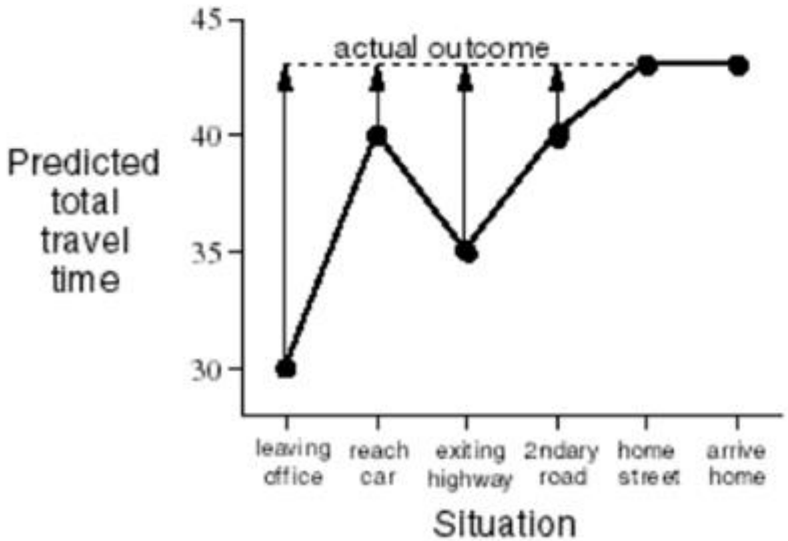
- The sequence of states, times, and predictions is thus as follows:

State	Elapsed Time (minutes)	Predicted Time to Go	Predicted Total Time
leaving office, friday at 6	0	30	30
reach car, raining	5	35	40
exiting highway	20	15	35
2ndary road, behind truck	30	10	40
entering home street	40	3	43
arrive home	43	0	43

R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction

Driving Home Example

Changes recommended by Monte Carlo methods (α=1)



Predicted total travel time

	Elapsed Time	*Predicted*	*Predicted*
State	*(minutes)*	*Time to Go*	*Total Time*
leaving office, friday at 6	0	30	30
reach car, raining	5	35	40
exiting highway	20	15	35
2ndary road, behind truck	30	10	40
entering home street	40	3	43
arrive home	43	0	43

R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction

Episodic Learning

- **Learning rate**
- The learning rate determines to what extent the newly acquired information will override the old information. A factor of 0 will make the agent not learn anything, while a factor of 1 would make the agent consider only the most recent information.

Monte Carlo

$$G_t = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{T-1} R_T$$

$$\underbrace{V(S_t)}_{\text{Maximum expected future reward starting at that state}} \leftarrow \underbrace{V(S_t)}_{\text{Former estimation of maximum expected future reward starting at that state}} + \underbrace{\alpha}_{\text{learning rate}}[\underbrace{G_t}_{\text{Discounted cumulative rewards}} - \underline{V(S_t)}]$$

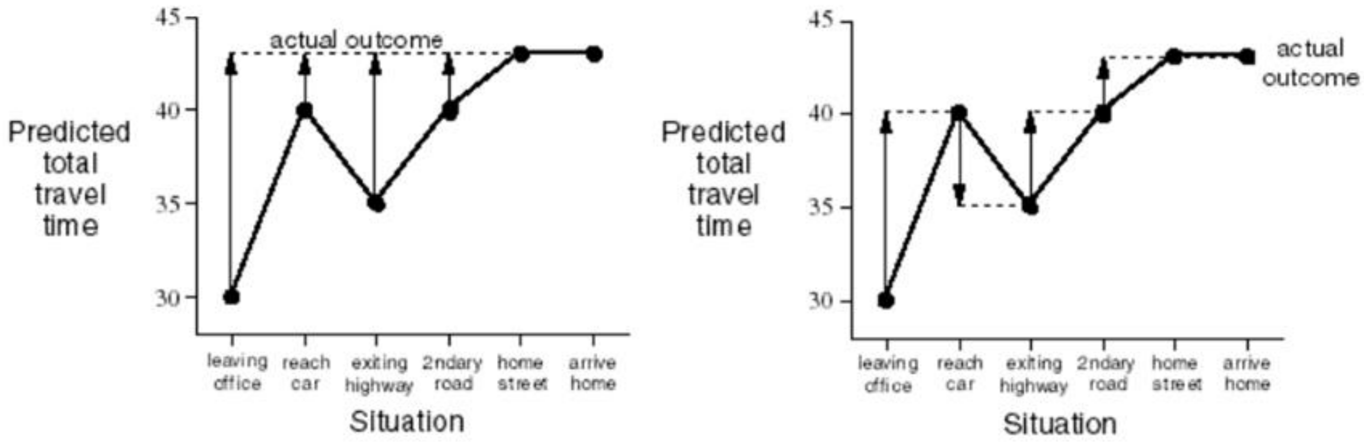
Temporal Difference Learning

- The basic idea of TD methods is that the learning is based on the difference between temporally successive predictions. In other words, the goal of learning is to make the learner's current prediction for the current input pattern more closely match the next prediction at the next time step.

Driving Home Example

Changes recommended by Monte Carlo methods (α=1)

Changes recommended by TD methods (α=1)



	Elapsed Time	*Predicted*	*Predicted*
State	*(minutes)*	*Time to Go*	*Total Time*
leaving office, friday at 6	0	30	30
reach car, raining	5	35	40
exiting highway	20	15	35
2ndary road, behind truck	30	10	40
entering home street	40	3	43
arrive home	43	0	43

R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction

Temporal Difference Learning

TD Learning $$V(S_t) \leftarrow \underbrace{V(S_t)}_{\text{Previous estimate}} + \alpha\Big[\underbrace{\underbrace{R_{t+1}}_{\text{Reward t+1}} + \underbrace{\gamma V(S_{t+1})}_{\text{Discounted value on the next step}}}_{\text{TD Target}} - \underline{V(S_t)}\Big]$$

Temporal Difference Learning

Initialize $V(s)$ arbitrarily, π to the policy to be evaluated
Repeat (for each episode):
  Initialize s
  Repeat (for each step of episode):
    $a \leftarrow$ action given by π for s
    Take action a; observe reward, r, and next state, s'
    $V(s) \leftarrow V(s) + \alpha\left[r + \gamma V(s') - V(s)\right]$
    $s \leftarrow s'$
  until s is terminal

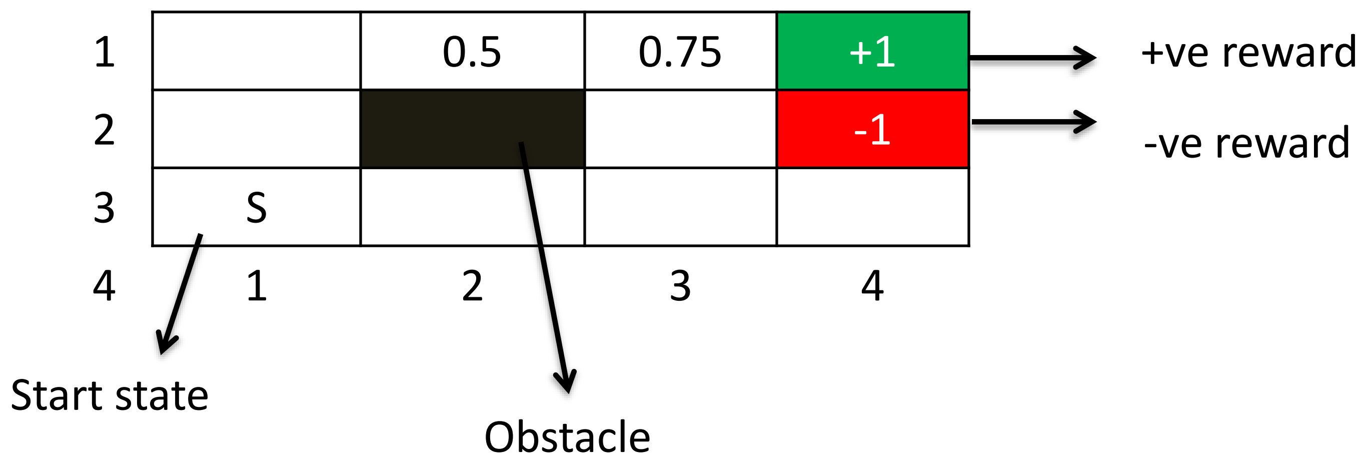
Figure 6.1: Tabular TD(0) for estimating V^{π}.

R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction

Temporal Difference Learning

- Temporal difference learning is faster but less stable and may converge to the wrong solution.

Temporal Difference Learning



$$V(s) \leftarrow V(s) + \alpha\left[r + \gamma V(s') - V(s)\right]$$

Temporal Difference Learning

Let alpha = 0.5 and gamma = 1
All initial values and rewards are zero.

$$V(s) \leftarrow V(s) + \alpha\left[r + \gamma V(s') - V(s)\right]$$

1	0	0	0.5	+1
2	0			-1
3	0			
	1	2	3	4

- The updated values as our RL agent moves are as follows:

V(S21) = 0 + 0.5[0 + 1(0) – 0)] = 0
Similarly,
V(11) = V(12) = 0
However,
V(13) = 0 + 0.5[0 + 1(1) – 0] = 0.5

Temporal Difference Learning

Let alpha = 0.5 and gamma = 1
All initial values and rewards are zero.

$$V(s) \leftarrow V(s) + \alpha\left[r + \gamma V(s') - V(s)\right]$$

1	0	0.25	0.75	+1
2	0			-1
3	0			
	1	2	3	4

- In next iteration, the following values will be updated:

V(12) = 0 + 0.5[0 + 1(0.5) – 0] = 0.25

V(13) = 0.5 + 0.5[0 + 1(1) -0.5] = 0.75

Thanks