

# **Advanced Machine Learning**

## **(AIE 322)**

### **Lecture 8: Temporal Difference Learning Methods for Prediction**

# What is Temporal Difference (TD) learning?

- Temporal difference (TD) learning refers to a class of model-free reinforcement learning methods which learn by bootstrapping from the current estimate of the value function.
- These methods sample from the environment, like Monte Carlo methods, and perform updates based on current estimates, like dynamic programming methods.
- That means this algorithm can form a Monte Carlo estimate without saving lists of returns.
- In this case we want to be able to learn incrementally before the end of the episode.

# TD Prediction

$$v_{\pi}(s) \doteq \mathbb{E}_{\pi}[G_t | S_t = s]$$

$$V(S_t) \leftarrow V(S_t) + \alpha [G_t - V(S_t)] \quad \leftarrow \text{Incremental update rule (as in the bandit)}$$

$$V(S_t) \leftarrow V(S_t) + \alpha \boxed{G_t} - V(S_t)]$$

In MC we need to take samples of full trajectories to compute  $G_t$

# Bootstrapping

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi[G_t | S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma \boxed{G_{t+1}} | S_t = s] \\ &= R_{t+1} + \gamma v_\pi(S_{t+1}) \end{aligned}$$

# TD Prediction

$$v_{\pi}(s) \doteq \mathbb{E}_{\pi}[G_t | S_t = s]$$

We need to take samples of full trajectories to compute  $G_t$

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

$$V(S_t) \leftarrow V(S_t) + \alpha \boxed{G_t} - V(S_t)]$$

$$G_t \approx R_{t+1} + \gamma V(S_{t+1})$$

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

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t) \quad \text{TD Error}$$

# In Dynamic Programming

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

## Iterative Policy Evaluation, for estimating $V \approx v_\pi$

Input  $\pi$ , the policy to be evaluated

$V \leftarrow \vec{0}, V' \leftarrow \vec{0}$

Loop:

$\Delta \leftarrow 0$

Loop for each  $s \in \mathcal{S}$  :

→  $V'(s) \leftarrow \sum_a \pi(a | s) \sum_{s', r} p(s', r | s, a) [r + \gamma V(s')]$

$\Delta \leftarrow \max(\Delta, |V'(s) - V(s)|)$

$V \leftarrow V'$

until  $\Delta < \theta$  (a small positive number)

Output  $V \approx v_\pi$

# 1-Step TD

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

$$S_t \leftarrow S_{t+1}$$

**S, A, R, S, A, R, S, A, R, S, A, R, ...**

# 1-Step TD

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

$$S_t \leftarrow S_{t+1}$$

update

S, A, **R, S**, A, R, S, A, R, S, A, R, ...

# TD(o) Algorithm

## Tabular TD(0) for estimating $v_\pi$

Input: the policy  $\pi$  to be evaluated

Algorithm parameter: step size  $\alpha \in (0, 1]$

Initialize  $V(s)$ , for all  $s \in \mathcal{S}^+$ , arbitrarily except that  $V(\text{terminal}) = 0$

Loop for each episode:

    Initialize  $S$

    Loop for each step of episode:

$A \leftarrow$  action given by  $\pi$  for  $S$

        Take action  $A$ , observe  $R, S'$

$V(S) \leftarrow V(S) + \alpha[R + \gamma V(S') - V(S)]$

$S \leftarrow S'$

    until  $S$  is terminal

## Example: Driving Home

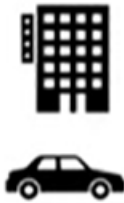
Driving Home Each day as you drive home from work, you try to predict how long it will take to get home. When you leave your office, you note the time, the day of week, the weather, and anything else that might be relevant. Say on this Friday you are leaving at exactly 6 o'clock, and you estimate that 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. The sequence of states, times, and predictions is thus as follows:

## Example: Driving Home

<i>State</i>	<i>Elapsed Time (minutes)</i>	<i>Predicted Time to Go</i>	<i>Predicted Total Time</i>
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

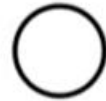
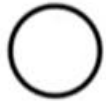
# Example: Driving Home

Driving Home



# Example: Driving Home

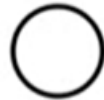
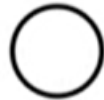
leave



0 mins  
elapsed

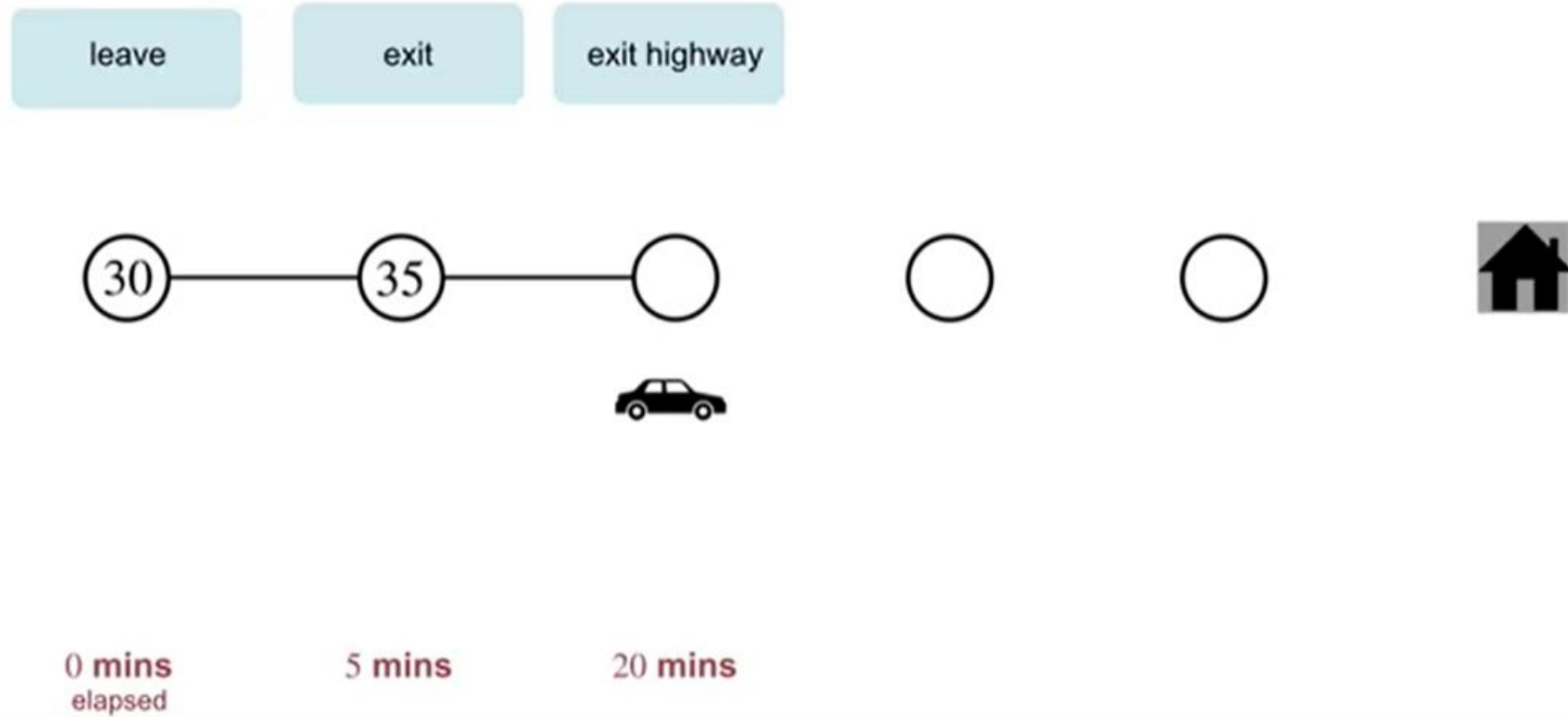
# Example: Driving Home

leave

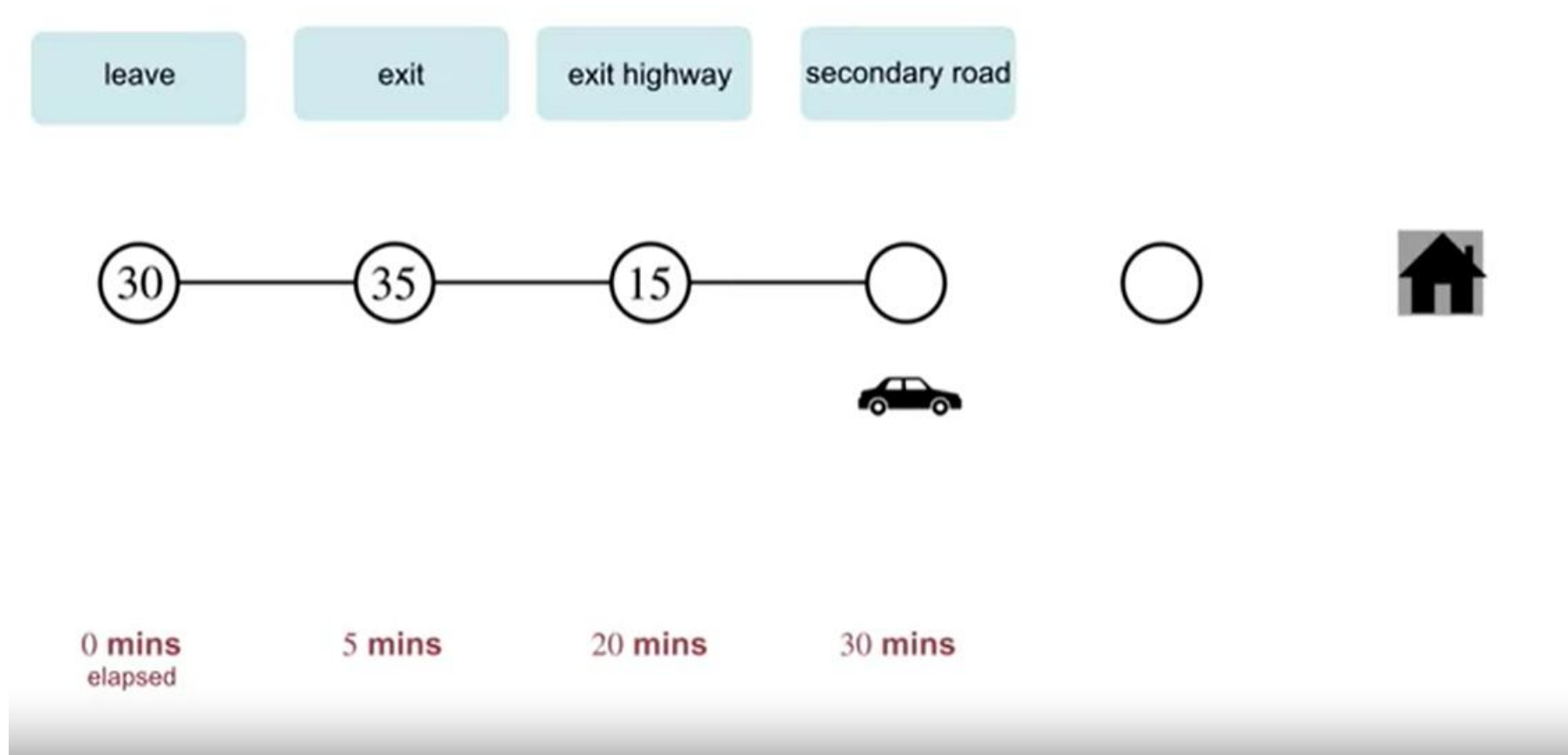


0 mins  
elapsed

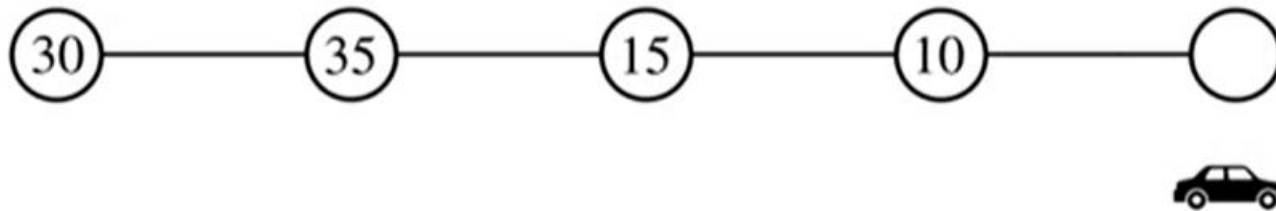
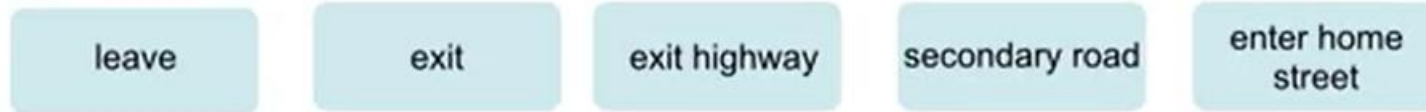
# Example: Driving Home



# Example: Driving Home



# Example: Driving Home



0 mins  
elapsed

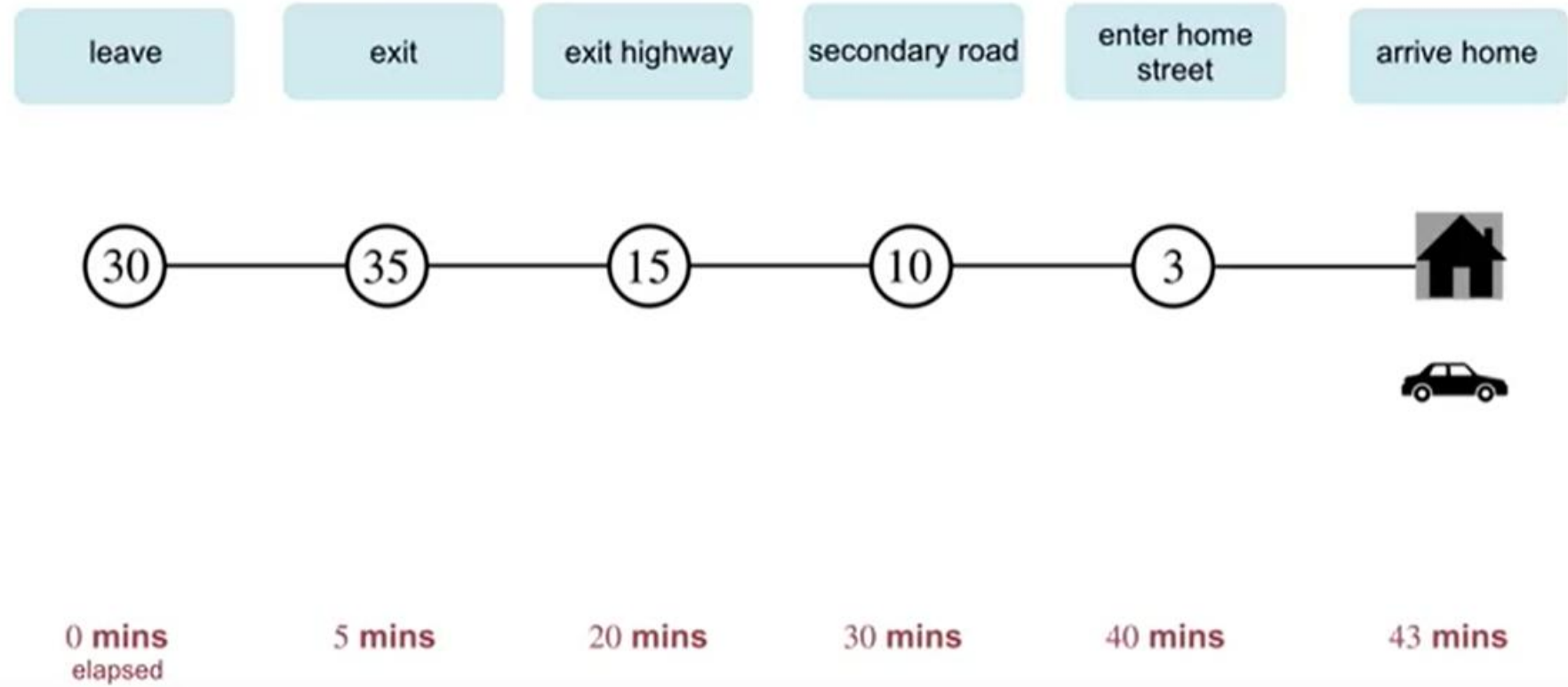
5 mins

20 mins

30 mins

40 mins

# Example: Driving Home



# Example: Driving Home



0 mins  
elapsed

5 mins

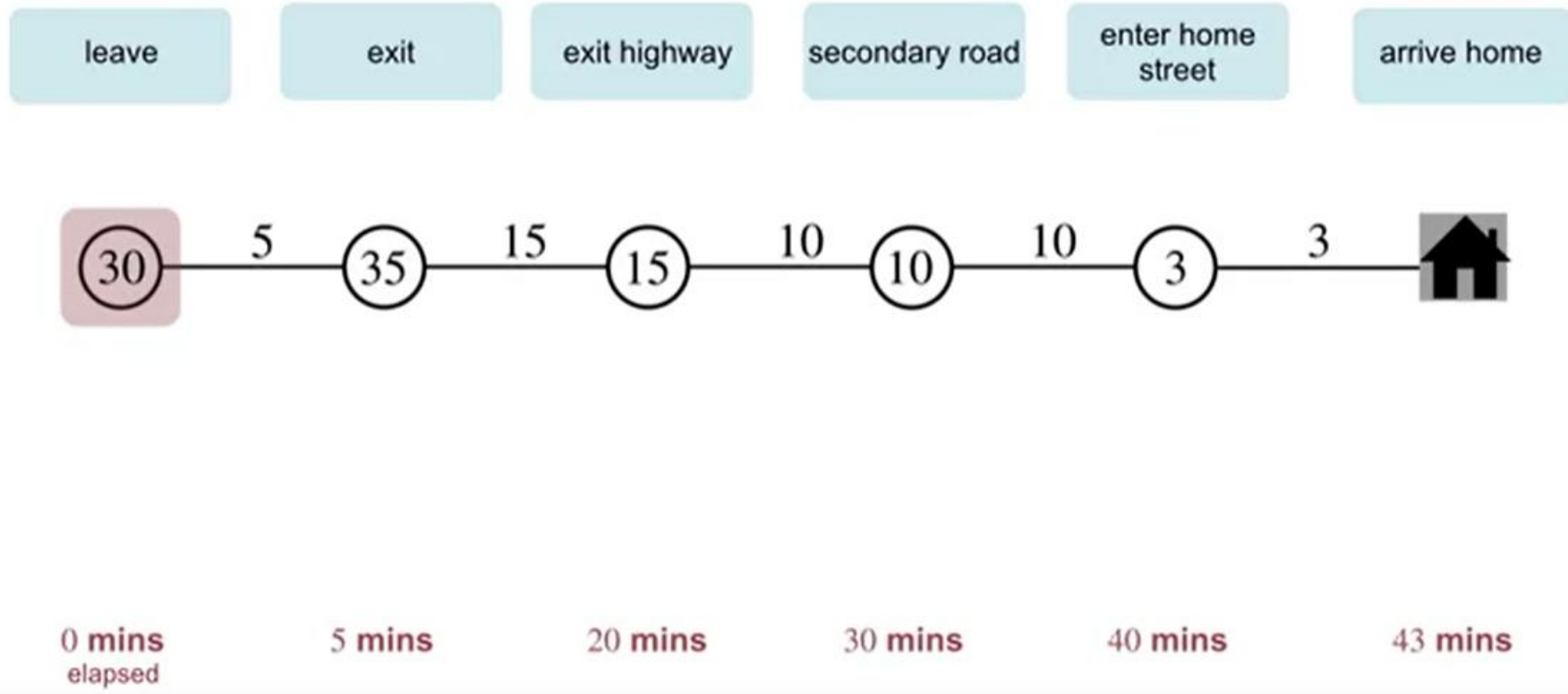
20 mins

30 mins

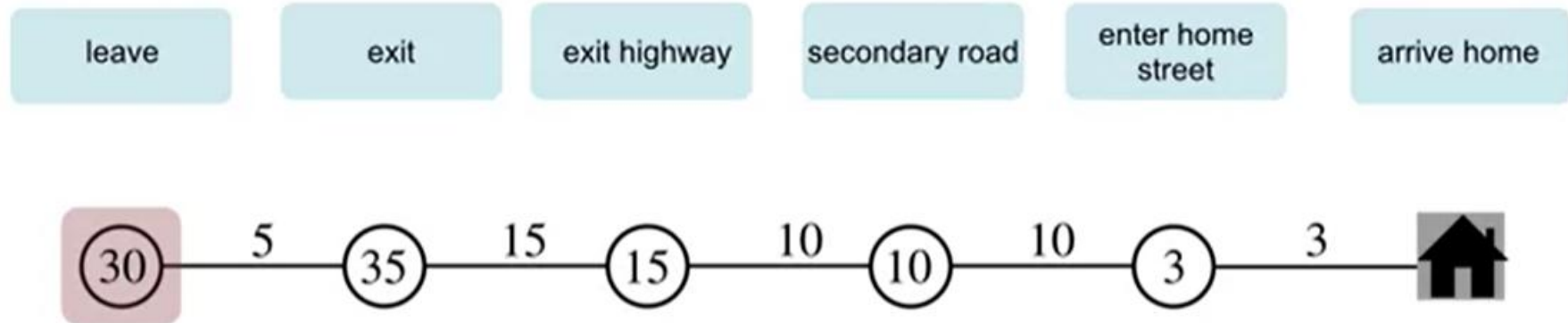
40 mins

43 mins

# Example: Driving Home

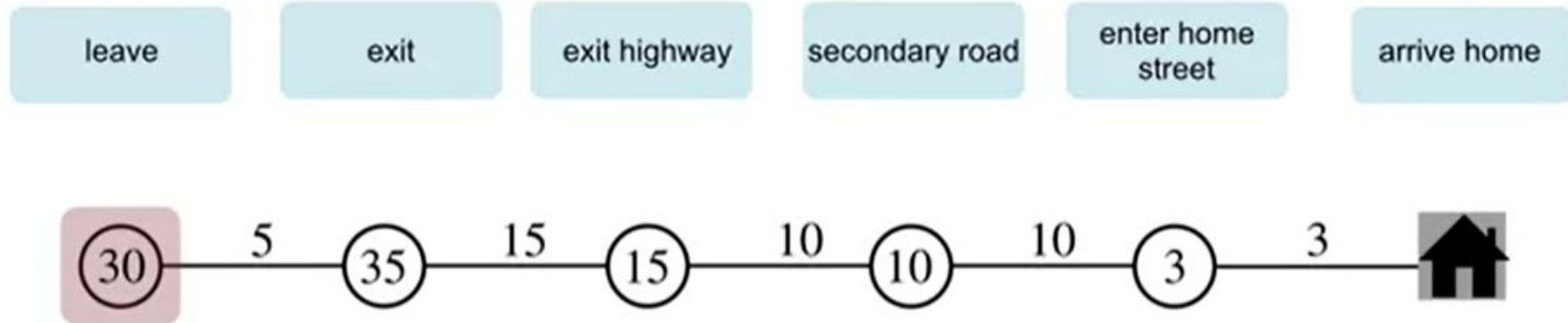


# Example: Driving Home (Using MC)



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

# Example: Driving Home (Using MC)



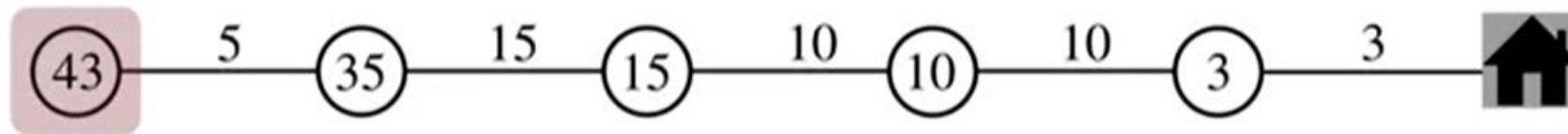
$$G_0 = 5 + 15 + 10 + 10 + 3 = 43$$

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

$$\alpha = 1 \quad \gamma = 1$$

$$V(\text{leave}) \leftarrow V(\text{leave}) + \alpha[G_0 - V(\text{leave})]$$

# Example: Driving Home (Using MC)



$$G_0 = 5 + 15 + 10 + 10 + 3 = 43$$

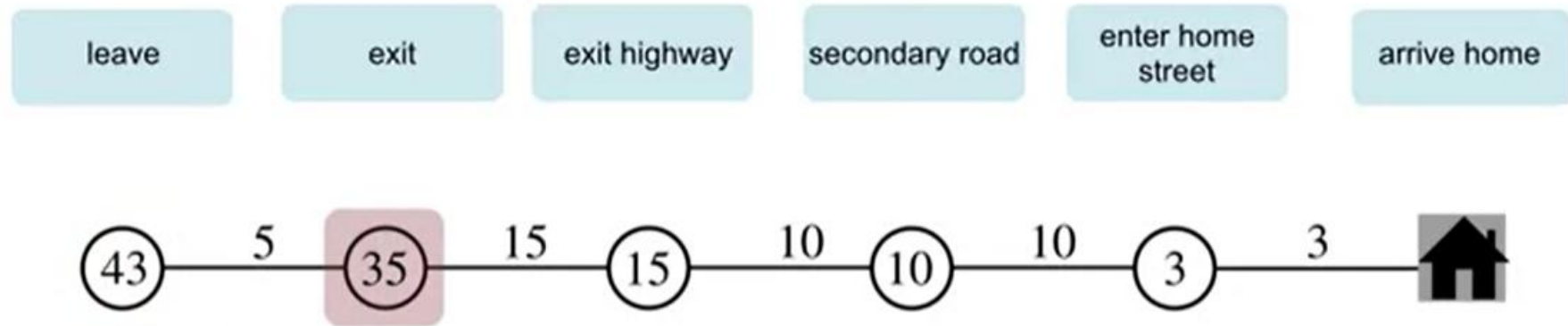
$$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$$

$$\alpha = 1 \quad \gamma = 1$$

$$V(\text{leave}) \leftarrow V(\text{leave}) + \alpha[G_0 - V(\text{leave})]$$

30                  43                  30

# Example: Driving Home (Using MC)

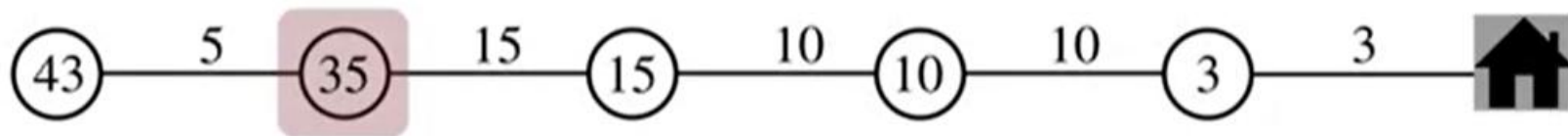


$$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$$

$$V(\mathbf{exit}) \leftarrow V(\mathbf{exit}) + \alpha[G_1 - V(\mathbf{exit})]$$

$$\alpha = 1 \quad \gamma = 1$$

# Example: Driving Home (Using MC)



$$G_1 = 15 + 10 + 10 + 3 = 38$$

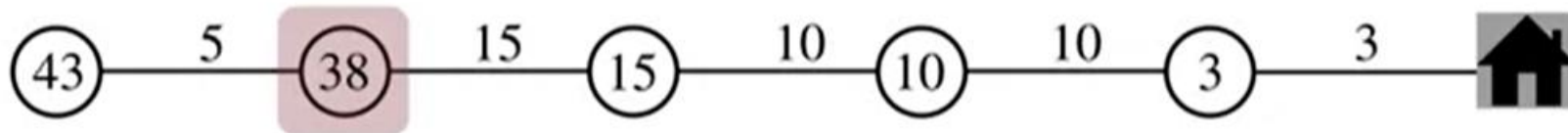
$$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$$

$$\alpha = 1 \quad \gamma = 1$$

$$V(\mathbf{exit}) \leftarrow V(\mathbf{exit}) + \alpha[G_1 - V(\mathbf{exit})]$$

35          38          35

# Example: Driving Home (Using MC)



$$G_1 = 15 + 10 + 10 + 3 = 38$$

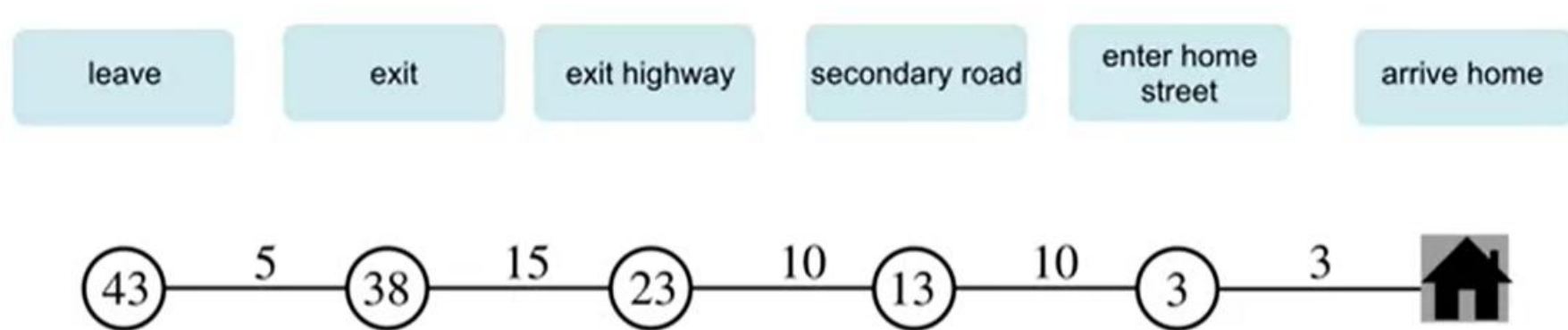
$$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$$

$$\alpha = 1 \quad \gamma = 1$$

$$V(\mathbf{exit}) \leftarrow V(\mathbf{exit}) + \alpha[G_1 - V(\mathbf{exit})]$$

35      38      35

# Example: Driving Home (Using MC)



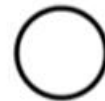
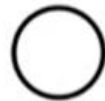
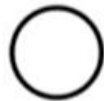
$$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$$

$$\alpha = 1 \quad \gamma = 1$$

**Must wait until the end of the episode  
before learning can begin!**

# Example: Driving Home (Using TD)

leave



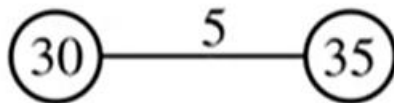
$$V(S_t) \leftarrow V(S_t) + \alpha [R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

$$\alpha = 1 \quad \gamma = 1$$

# Example: Driving Home (Using TD)

leave

exit



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

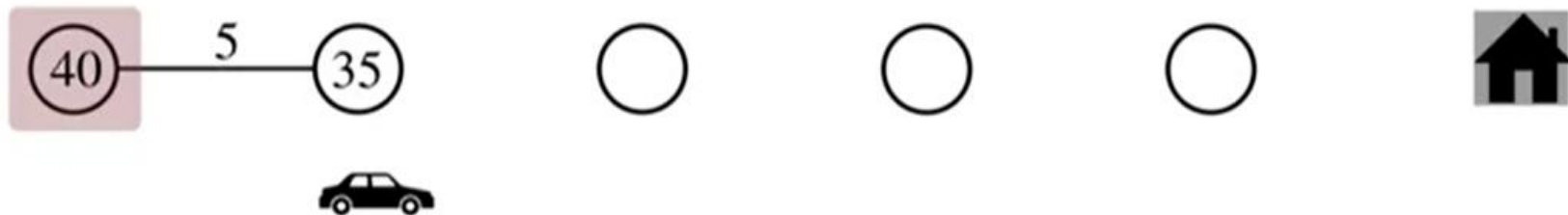
$$\alpha = 1 \quad \gamma = 1$$

$$V(\text{leave}) \leftarrow V(\text{leave}) + \alpha [R_1 + \gamma V(\text{exit}) - V(\text{leave})]$$

# Example: Driving Home (Using TD)

leave

exit



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

$$\alpha = 1 \quad \gamma = 1$$

$$V(\text{leave}) \leftarrow V(\text{leave}) + \alpha [ R_1 + \gamma V(\text{exit}) - V(\text{leave}) ]$$

30            5            35            30

# Example: Driving Home (Using TD)

leave

exit

exit highway

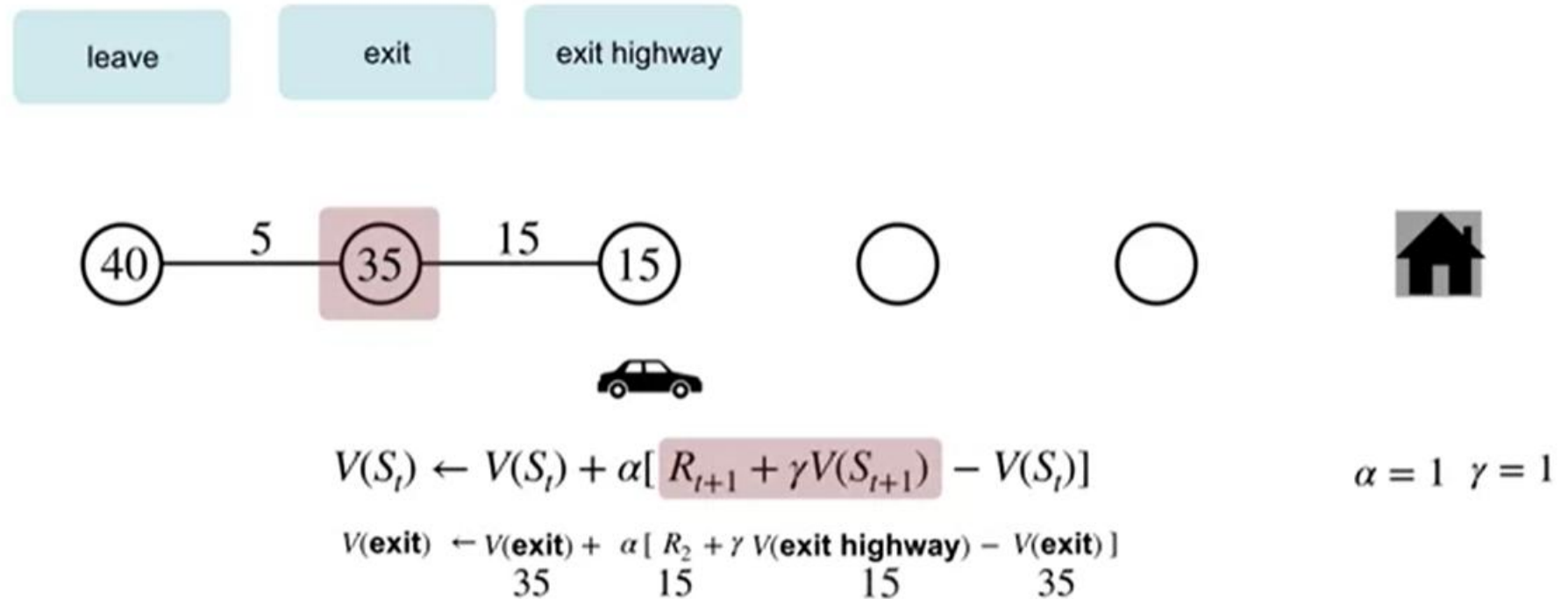


$$V(S_t) \leftarrow V(S_t) + \alpha [ R_{t+1} + \gamma V(S_{t+1}) - V(S_t) ]$$

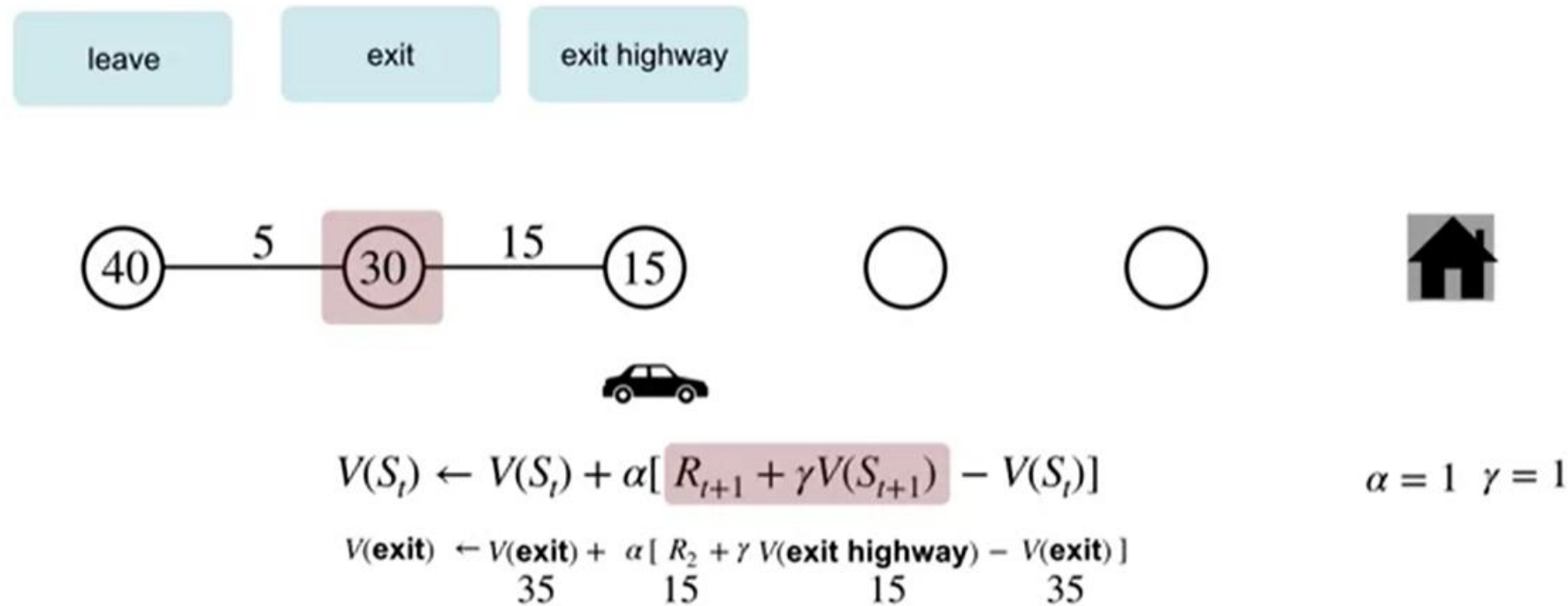
$$\alpha = 1 \quad \gamma = 1$$

$$V(\text{exit}) \leftarrow V(\text{exit}) + \alpha [ R_2 + \gamma V(\text{exit highway}) - V(\text{exit}) ]$$

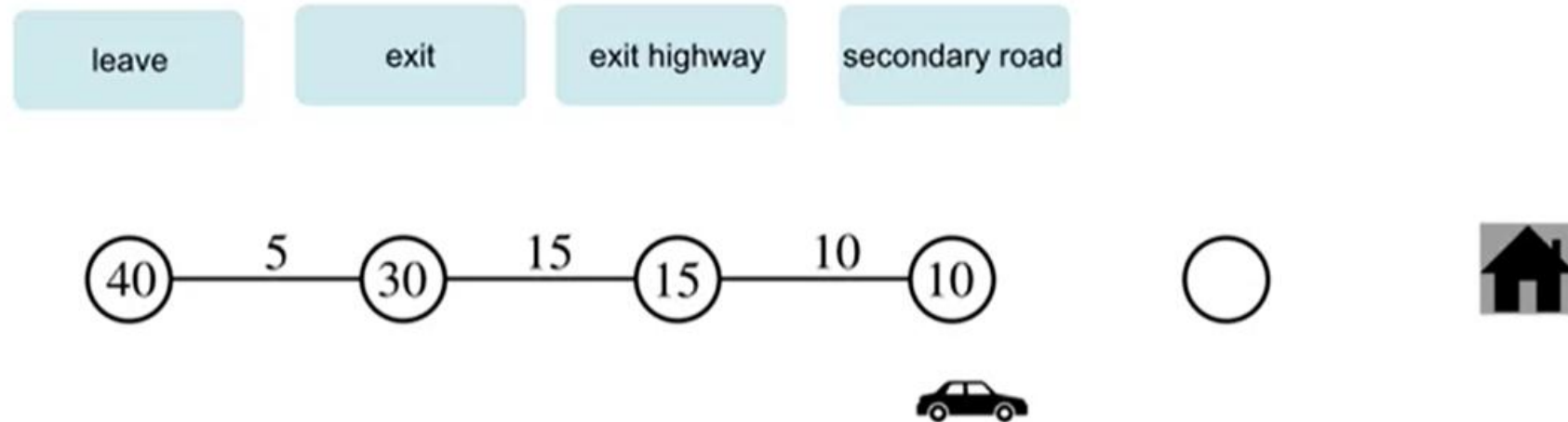
# Example: Driving Home (Using TD)



# Example: Driving Home (Using TD)



# Example: Driving Home (Using TD)

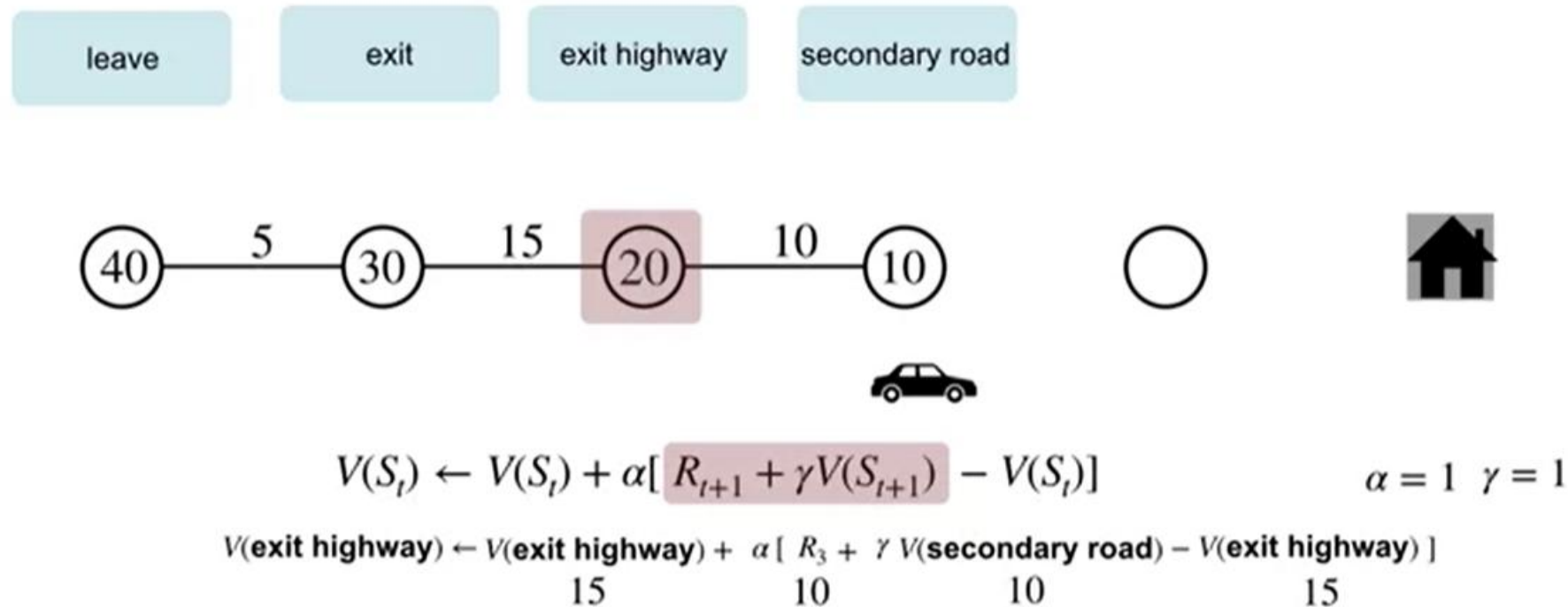


$$V(S_t) \leftarrow V(S_t) + \alpha [ R_{t+1} + \gamma V(S_{t+1}) - V(S_t) ]$$

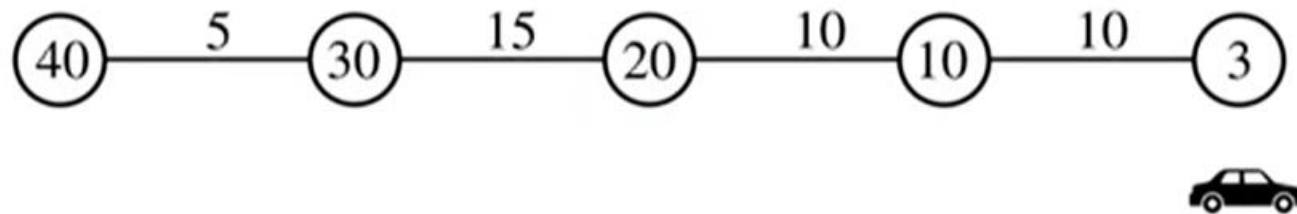
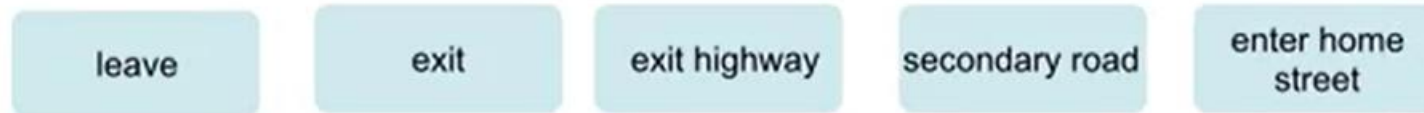
$$\alpha = 1 \quad \gamma = 1$$

$$V(\text{exit highway}) \leftarrow V(\text{exit highway}) + \alpha [ R_3 + \gamma V(\text{secondary road}) - V(\text{exit highway}) ]$$

# Example: Driving Home (Using TD)



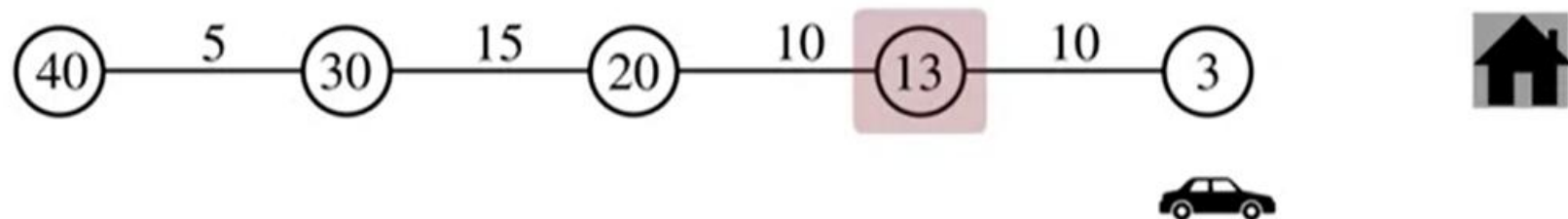
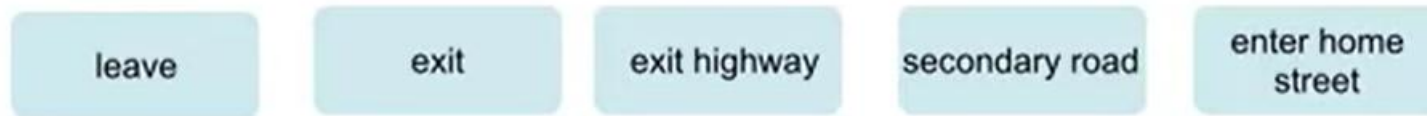
# Example: Driving Home (Using TD)



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

$$V(\text{secondary road}) \leftarrow V(\text{secondary road}) + \alpha [ R_4 + \gamma V(\text{enter home st.}) - V(\text{secondary road}) ]$$

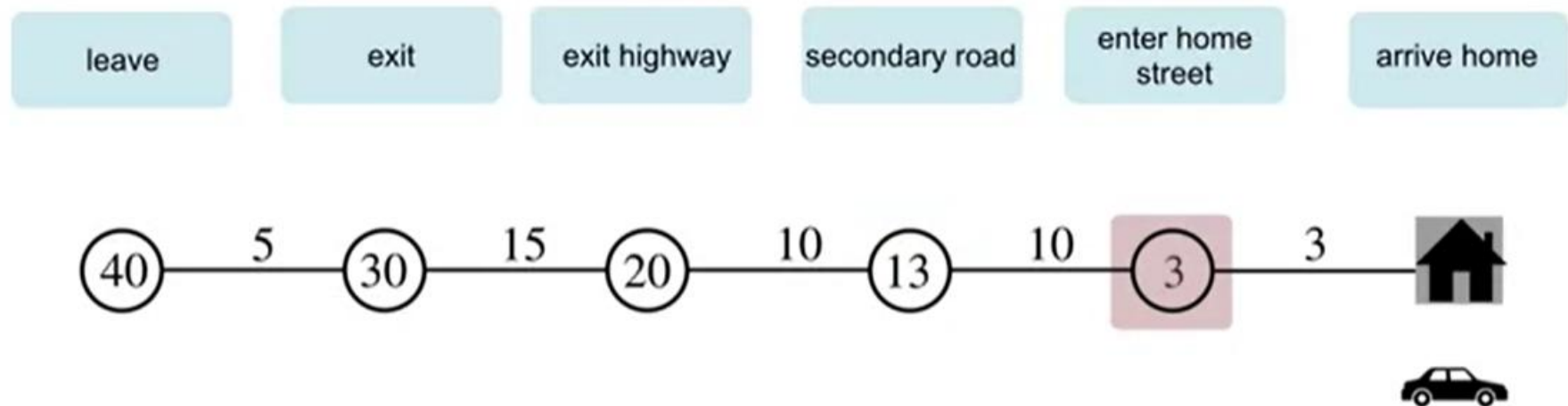
# Example: Driving Home (Using TD)



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

$$V(\text{secondary road}) \leftarrow V(\text{secondary road}) + \alpha \left[ \underset{10}{R_4} + \underset{10}{\gamma} \underset{3}{V(\text{enter home st.})} - \underset{10}{V(\text{secondary road})} \right]$$

# Example: Driving Home (Using TD)

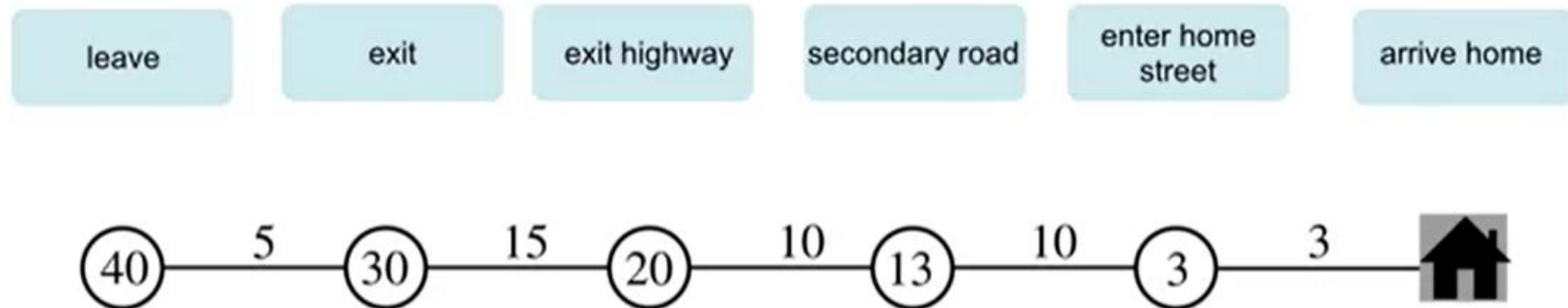


$$V(S_t) \leftarrow V(S_t) + \alpha [R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

$$\alpha = 1 \quad \gamma = 1$$

$$V(\text{enter home st.}) \leftarrow V(\text{enter home st.}) + \alpha \left[ \underset{3}{R_5} + \underset{3}{\gamma} \underset{0}{V(\text{home})} - \underset{3}{V(\text{arrive home st.})} \right]$$

# Example: Driving Home (Using TD)



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

$$\alpha = 1 \quad \gamma = 1$$

**We can learn online without waiting  
for the episode to end!**

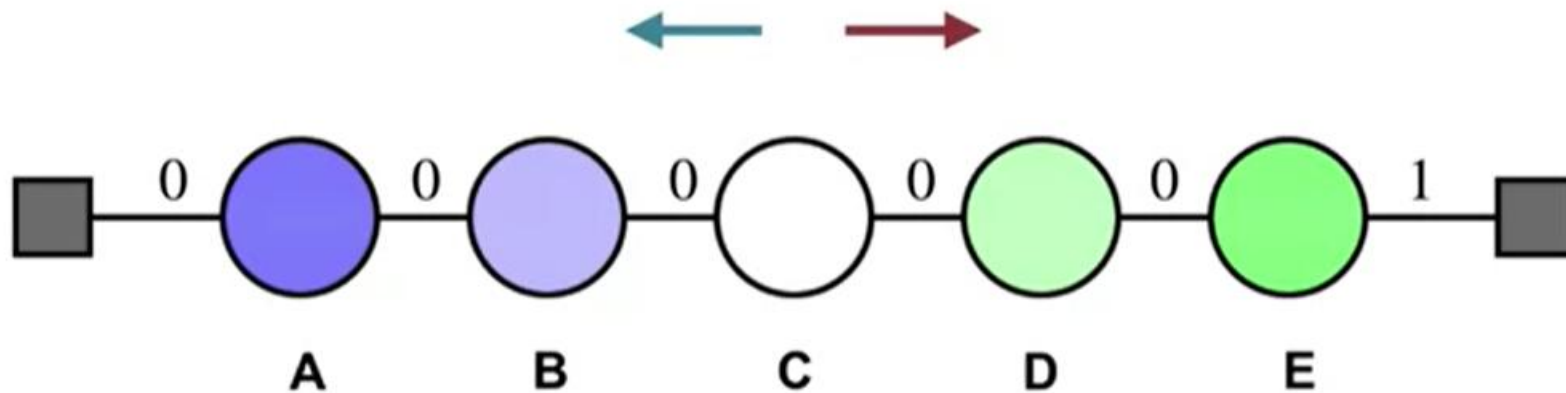
**Unlike MC**

# The advantages of TD

- Do not require a model of the environment (Unlike Dynamic Programming)
- Updates estimates based on other estimates on every step (Unlike Monte Carlo)
- Converges faster than Monte Carlo methods

# Comparing TD and MC

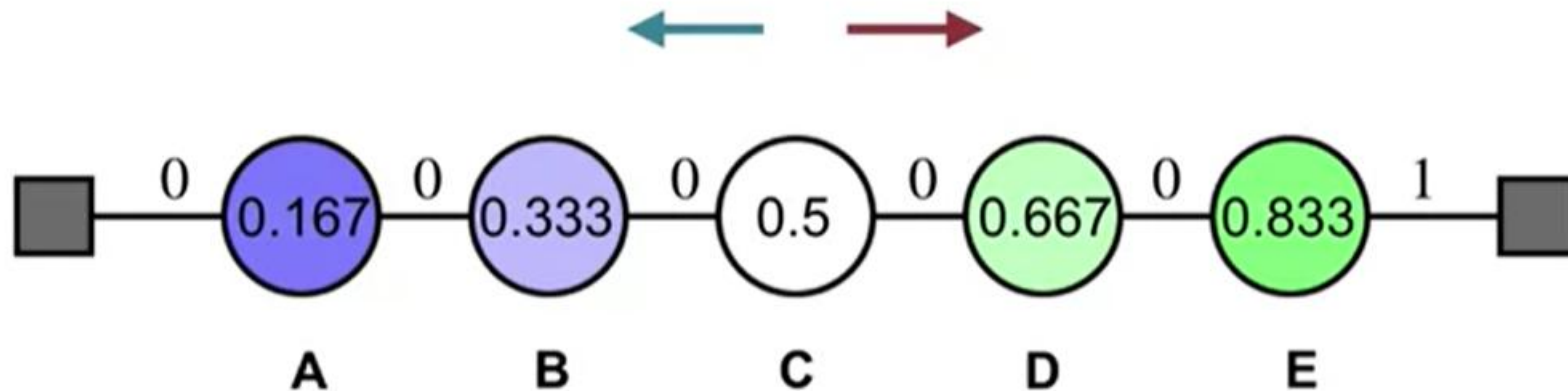
## Random Walk



$$\pi(. | s) = 1/2 \quad \forall s \in \mathcal{S} \quad \gamma = 1$$

# Comparing TD and MC

## Random Walk



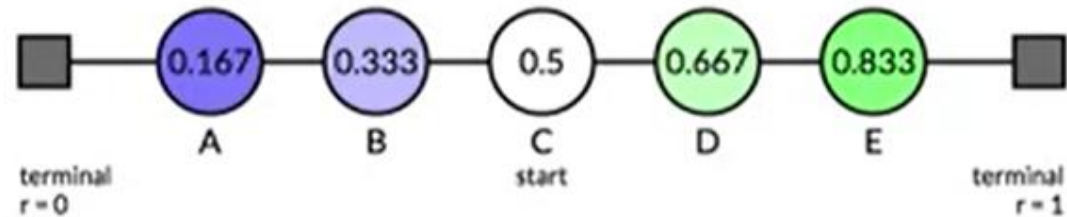
$$\pi(. | s) = 1/2 \quad \forall s \in \mathcal{S} \quad \gamma = 1$$

# Comparing TD and MC

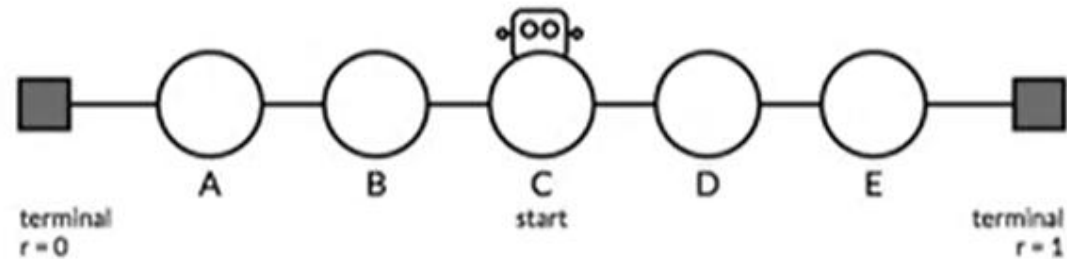
$\alpha=0.5$



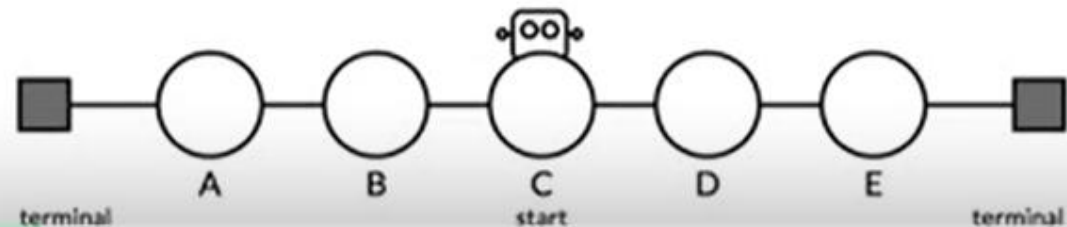
Target / Exact Values



Updates using TD Learning



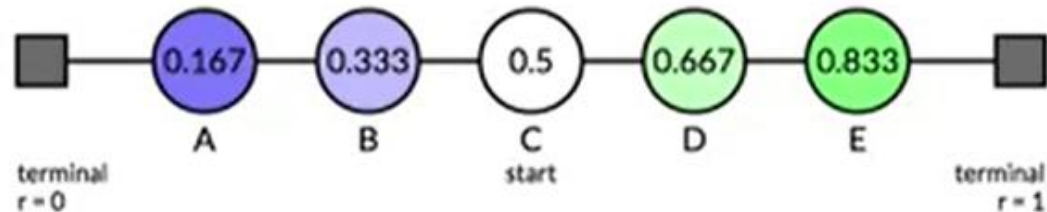
Updates using Monte Carlo



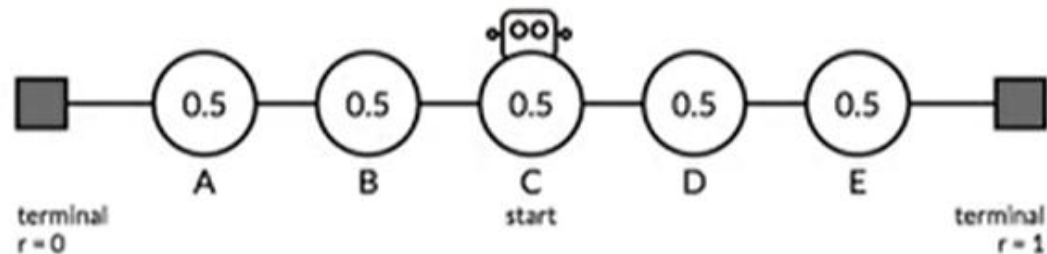
# Comparing TD and MC

$\alpha=0.5$

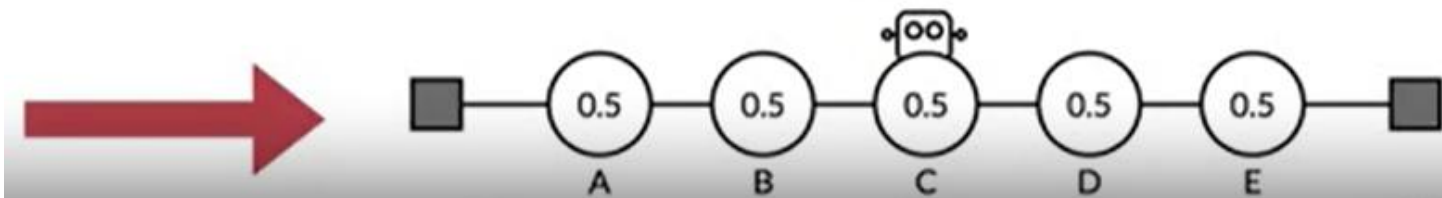
Target / Exact Values



Updates using TD Learning  $V(S_t) \leftarrow V(S_t) + \alpha [R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$



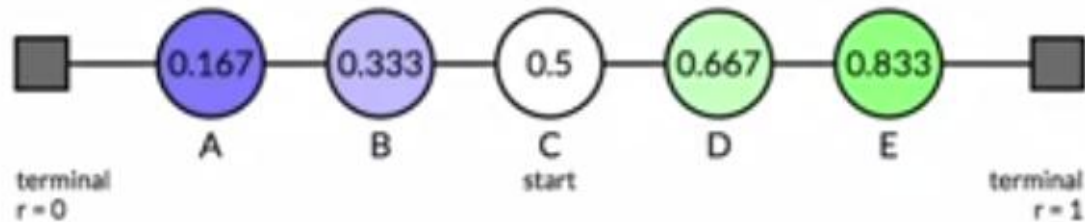
Updates using Monte Carlo



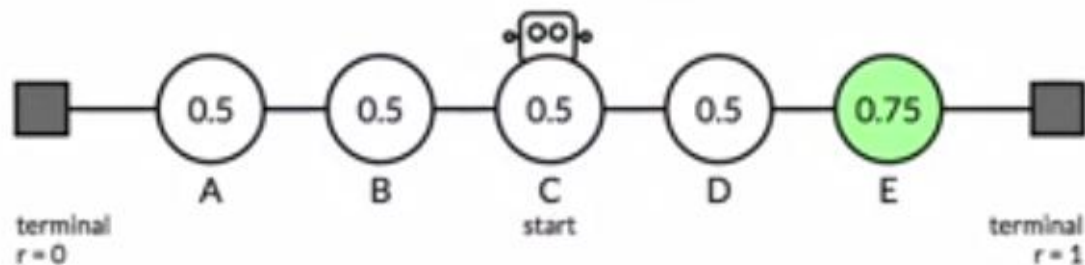
# Comparing TD and MC

$\alpha=0.5$

Target / Exact Values



Updates using TD Learning  $V(S_t) \leftarrow V(S_t) + \alpha [R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$



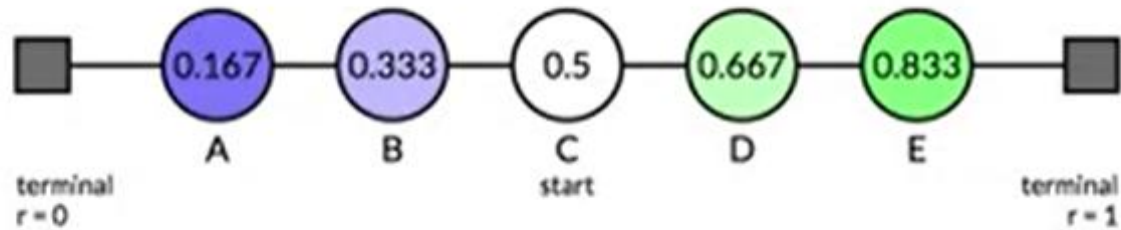
Updates using Monte Carlo  $V(S_t) \leftarrow V(S_t) + \alpha [G_t - V(S_t)]$



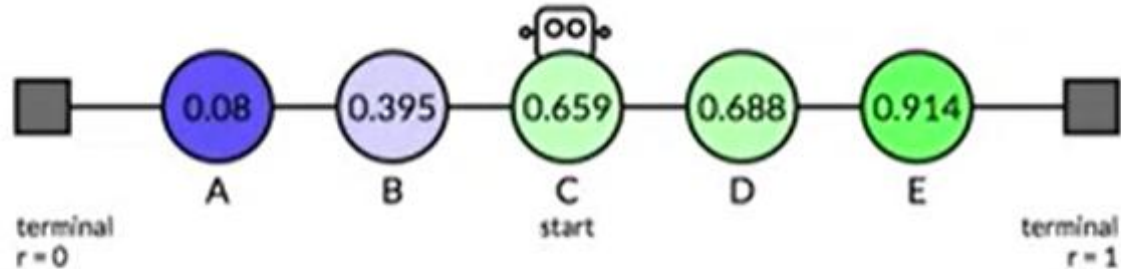
# Comparing TD and MC

$\alpha=0.5$

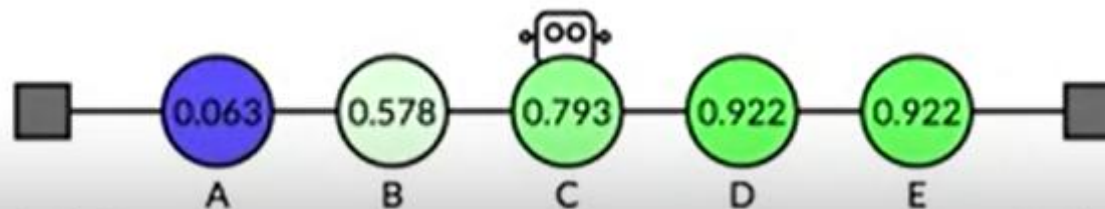
Target / Exact Values



Updates using TD Learning



Updates using Monte Carlo



# Comparing TD and MC

