Graduate Artificial Intelligence
CS 640
Temporal Difference and Q-Learning

Jeffrey Considine

jconsidi@bu.edu

Plan for Today

- Assumptions of Iteration Methods
- Temporal difference learning
- Q-learning

What is the biggest assumption of Value and Policy Iteration?

???

What Can We Do Without That Assumption?

???

Any Questions?

???

Model-Free Reinforcement Learning

- Model-free = no explicit model of rewards and transitions
- Reinforcement learning
  - Catchall term for learning optimal actions…
  - Mostly behavior reinforcement from rewards (like the phrase in psychology)
  - Estimates for some states and/or actions based on other estimates…
    - Previous value and state iteration methods could be considered examples.

Temporal Difference (TD) Learning

- For a given policy , learn without a model of the environment.
  - Cannot calculate analytically or use value iteration.

- Key idea:
  - and are closely related.

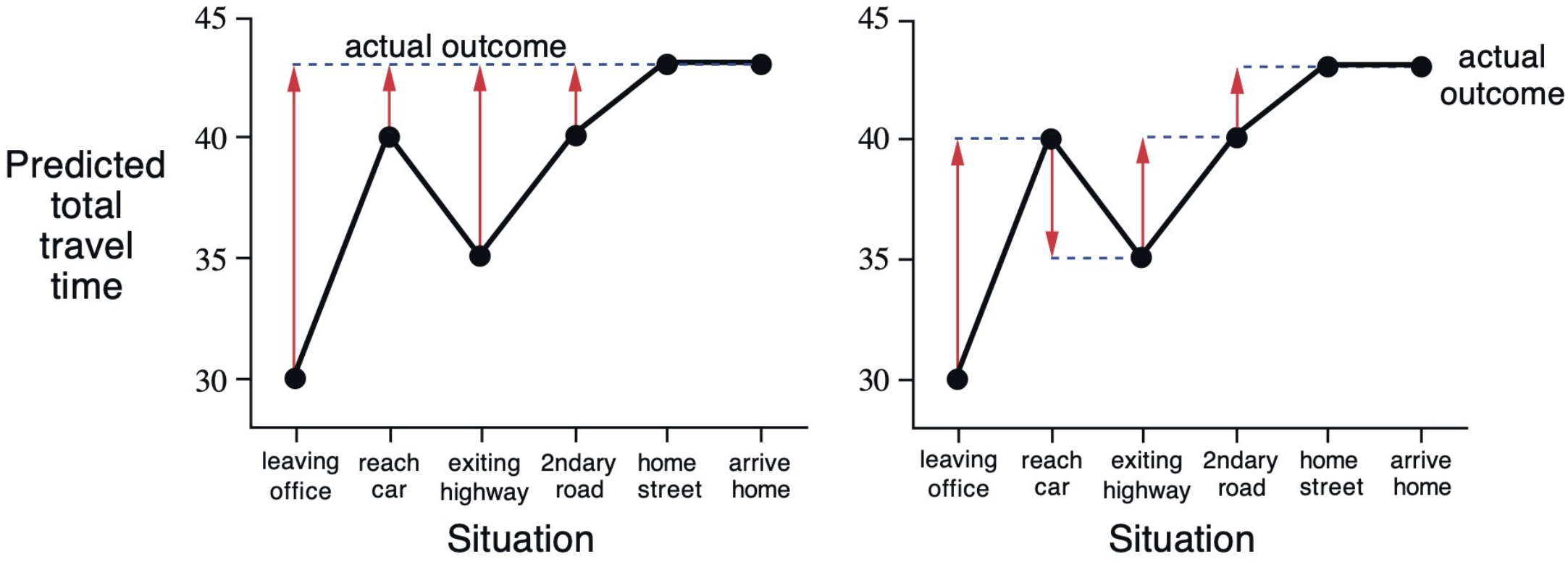
  - Try to optimize away the “temporal difference”.

- Works for Markov decision processes with a policy, but same algorithm works for Markov reward processes.

Driving Example (Sutton and Barto, 2020)

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

How Early Can a Prediction by Updated?
(Sutton and Barto, 2020)



What if You Drive Home From a New Office?
(Sutton and Barto, 2020)

???

Intuition for TD Learning

- For a particular policy , there is a known relationship between the current state value and the expected next state value.

- Every time we observe a transition from to and receive reward , did we just get a sample of the righthand side?

Inconsistencies in a Value Function

- While working on a value function ,

- What if we update every time we get a sample transition?

  - Goal is to compute the expectation of the righthand side.
  - Typically use exponential moving average.

The Temporal Difference Learning Algorithm

- Sample state transitions following policy .
  - Usually this is testing a policy.

- Pick learning rate .
- For each observed state transition from to receiving reward ,

  - Update

Why the exponential moving average?

???

Temporal Difference Learning Reformulated

Previously update was table/assignment oriented.

Incremental / learning version:

Any Questions?

???

What is Q-Learning?

- Q-learning is any process learning the optimal state-action value function.

- Not explicit, but this is usually model-free.

Q-Learning vs Temporal Difference-Learning

- Q-learning learns optimal state-action values ().
- Temporal difference learning only learns optimal state values ().
- More state-action values to learn.

Why Q-Learning?

Assuming there is a known and finite set of action choices, then

- Sufficient to pick optimal actions (separate not needed).
- Sufficient to calculate optimal state values .

Can we do those with just optimal state values?

Recursive Definition of Q Values

$$q_*(s,a) = E\left[R_{t+1} + \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a\right]$$

Will Exponential Moving Average Work Again?

???

What Order Should Q Values be Estimated?

???

Final States

- Characterized by a reward and immediately stopping.
  - Typical implementation would be fixed reward and transition to do nothing sink state.
- Handle these first.
  - Usually identified by problem definition.

The Q-Learning Algorithm (initialization) (Watkins et al, 1992)

- Initialize a table with entries for every state and action .
- For all final states without any allowed actions, set to the reward associated with for all actions *a* .
- Initialize the remaining entries to 0.

The Q-Learning Algorithm (learning) (Watkins et al, 1992)

- For i =0,1,2,…
  a) Sample non-final state .
  b) Pick action .
  c) Observe the resulting payoff and next state .
  d) Update

Learning Rate for Q-Learning

- If the system is deterministic, then should just be set to 1.
- If the system is probabilistic, then should slowly decline with the number of times that and have been sampled.
  - See (Watkins et al, 1992) for details to guarantee convergence…
- In practice, many practioners ignore this technical condition and set to a constant such as 0.1.

How to Sample for Q-Learning

???

So How Did TD- and Q-Learning Work Around the Lack of Model?

???

Are TD- and Q-Learning Solving or Learning?

???

Any Questions?

???