Course 2, Module 2
Temporal Difference Learning Methods for Prediction

CMPUT 397
Fall 2020

Comments

- Announcement about Slido question limits
- Announcement about projects. Email me if you want to do one of the projects
  - One additional option: compare Monte Carlo and Sarsa, in Mountain Car
- Slack group: TAs and I have joined. Come there to have useful discussions!

Review of Course 2, Module 2
TD Learning

Video 1: What is Temporal Difference Learning?

- One of the central ideas of Reinforcement Learning! We focus on policy evaluation first: learning v_π.
- Updating a guess from a guess: Bootstrapping. It means we can learning **during the episode. No waiting till the end of an episode!**
- Goals:
  - Define temporal-difference learning
  - Define the temporal-difference error
  - And understand the TD(0) algorithm.

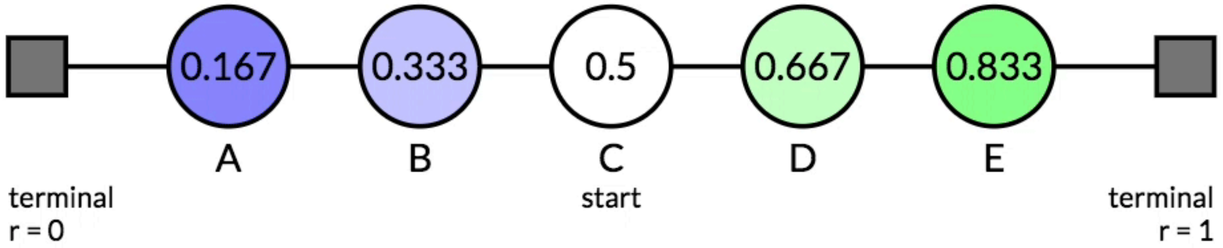
Video 2: The Advantages of Temporal Difference Learning

- How TD has some of the benefits of MC. Some of the benefits of DP. AND some benefits unique to TD
- Goals:
  - Understand the benefits of learning online with TD
  - Identify key **advantages of TD methods** over Dynamic Programming and Monte Carlo methods
    - do not need a model
    - update the value function on every time-step
    - typically learns faster than Monte Carlo methods

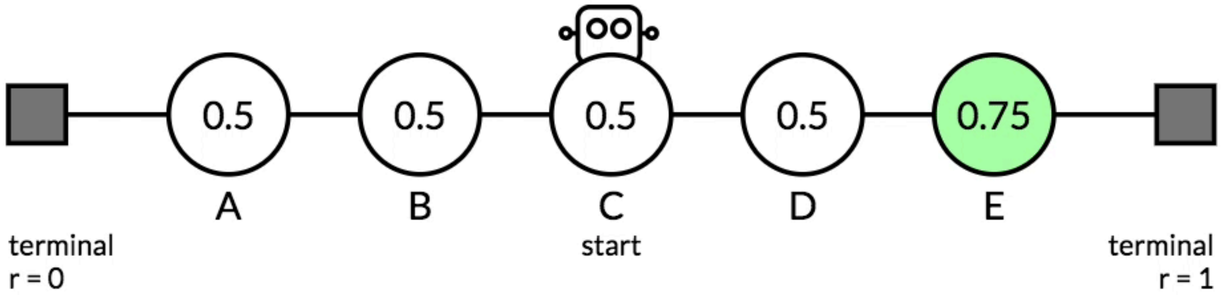
Video 3: Comparing TD and Monte Carlo

- Worked through an example using TD and Monte Carlo to learn v_π. We looked at how the updates happened on each step. And final performance via learning curves
- Goals:
  - Identify the empirical benefits of TD learning.

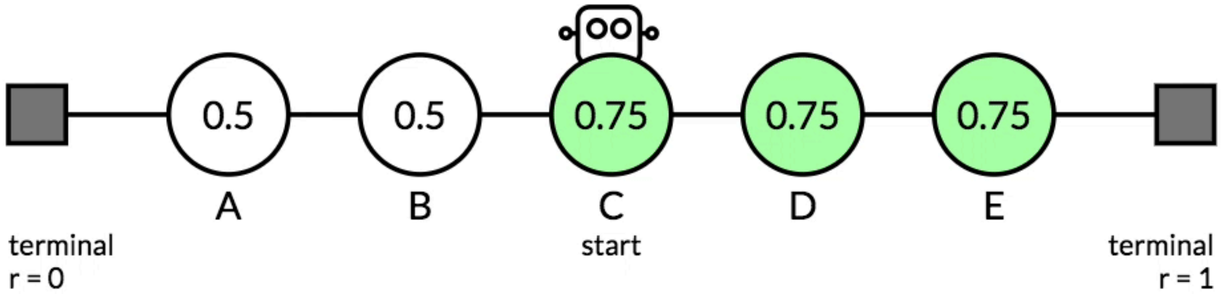
Target / Exact Values



Updates using TD Learning



Updates using Monte Carlo



Tabular TD(0) for estimating v_π

Input: the policy π to be evaluated
Algorithm parameter: step size $\alpha \in (0, 1]$
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(\mathit{terminal}) = 0$

Loop for each episode:
    Initialize S
    Loop for each step of episode:
        $A \leftarrow$ action given by π for S
        Take action A, observe R, S'
        $V(S) \leftarrow V(S) + \alpha\big[R + \gamma V(S') - V(S)\big]$
        $S \leftarrow S'$
    until S is terminal

Terminology Review

- In TD learning there are **no models, YES bootstrapping, YES learning during the episode**

- TD methods update the value estimates on a **step-by-step** basis. We **do not wait** until the end of an episode to update the values of each state.

- TD methods use **Bootstrapping**: using the estimate of the value in the next state to update the value in the current state: $V(S) \leftarrow V(S) + \alpha[R + \gamma V(S') - V(S)]$

  TD-error

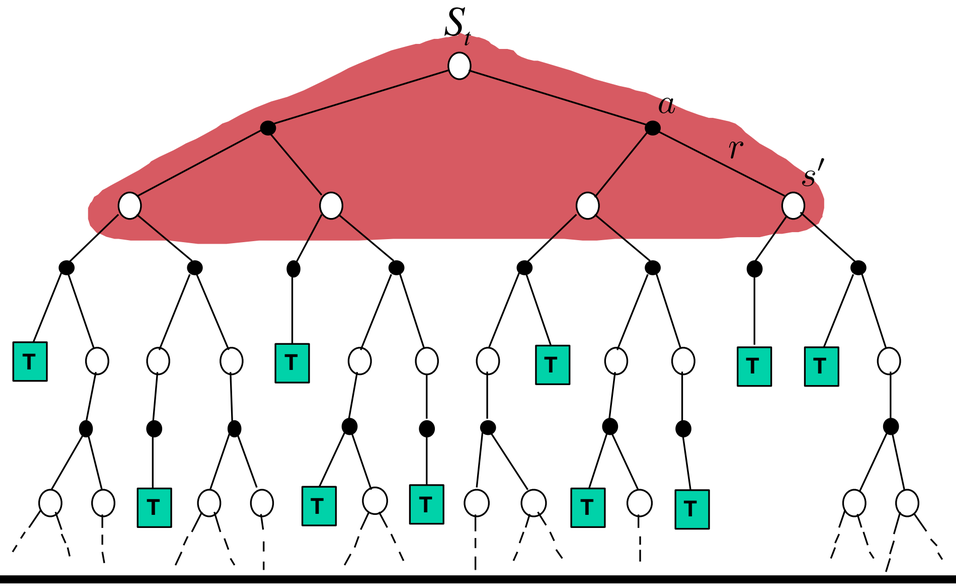
- TD is a **sample update** method: update involves the value of single sample successor state

- An **expected update** requires the complete distribution over all possible next states

- TD and MC are sample update methods. Dynamic programming uses expected updates

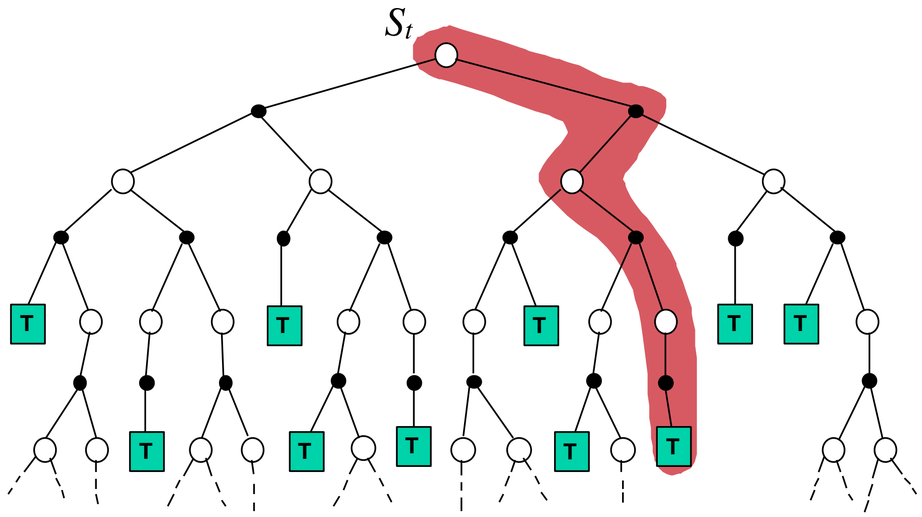
Dynamic programming

$$V(S_t) \leftarrow E_\pi\left[R_{t+1} + \gamma V(S_{t+1})\right] = \sum_a \pi(a|S_t) \sum_{s',r} p(s', r|S_t, a)[r + \gamma V(s')]$$



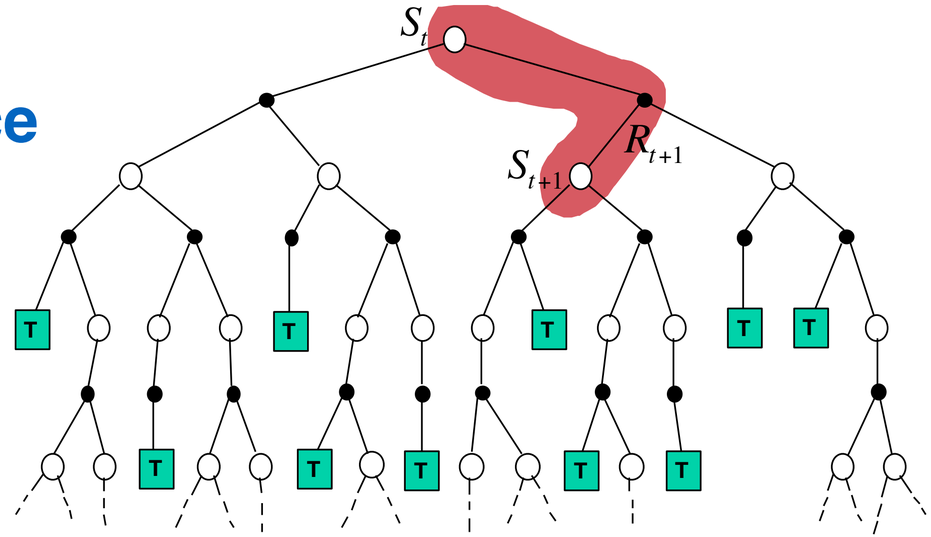
Simple Monte Carlo

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



Temporal Difference Learning

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



Worksheet Question

Modify the Tabular TD(0) algorithm for estimating v_π, to estimate q_π.

Tabular TD(0) for estimating v_π

Input: the policy π to be evaluated
Algorithm parameter: step size $\alpha \in (0, 1]$
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(\mathit{terminal}) = 0$

Loop for each episode:
    Initialize S
    Loop for each step of episode:
        $A \leftarrow$ action given by π for S
        Take action A, observe R, S'
        $V(S) \leftarrow V(S) + \alpha\big[R + \gamma V(S') - V(S)\big]$
        $S \leftarrow S'$
    until S is terminal

Slido Qs: Expected vs Sample Updates

- “Probably a question that I should have asked a few weeks ago... What is the difference between expected updates and sample updates?”
- Related: “What exactly does having a "high/low variance" mean and why is there a difference in variance between TD and MC?”
- —> Moving to my ipad to explain this, and to finish off the exercise from last week

Slido Q: TD vs MC

- “Why would we ever use monte carlo when we see that TD learning is significantly better in all cases? Not just do we not have to wait for the the episode to finish to learn but we get better results faster.”
- “Although it hasn't been proved in theory, in practice, TD learning tends to perform better than MC. Are there any shown scenarios where MC performs better than TD?”
- “What, if any, are the downsides to TD? In the book and the lectures, we learned about all the ways it was better than our previous methods. Are there any ways in which it is worse?”
- “Why does TD(0) have a lower variance than MC?”
- “In what situation would the Monte Carlo method be more practical to use than TD learning?”

Slido Qs: TD and MC clarifications

- “Why we use TD instead of MC in continues task?”
- “TD is a solution method for prediction where MC is a control learning method. What are the main differences between control and prediction methods and their usage/application?”
- “Why TD is online learning?”

Slido Q: Misc

- “The driving home from work example in video "The advantages of temporal difference learning" uses the amount of time taken as the reward. Doesn't this contradict the agent's goal of maximizing the reward and encouraging the agent to spent more time driving around rather than getting home quicker?”

- “Can you what is batch updating and how is it different than what happens normally? Is batch updating used only for short episodic problems or can it be applied to continuous problems as well? Will there be cases where the value function won't converge due to a finite amount of experience?”

- “Is there a way to make TD even more efficient?”

Slido Qs: Misc

- “Is that possible that Monte Carlo will outperform TD(0) if a vast number of previous experience(data) is provided? If so, it sparks another question if MC is similar to supervised learning where more data could essentially improve the future prediction?”

- “Are the TD(0) state-values really of the same quality compared to MC? Given that TD(0) updates based upon the next state, with a constant step size, it seems like in stochastic scenarios with high variance, the weighting of recent experiences would skew the values compared averaging returns.”

Slido Qs: TD in Practice

- “Is TD learning being used in practice?”
- “It seems (according to the textbook/videos? I could be wrong about this) that TD is maybe the best method of RL described so far. Is TD used most in industry? Or would it be something that hasn't been covered (yet)?”