Course 2, Module 3 **TD for Prediction CMPUT 397** Fall 2019

A few clarifications

- Batch versus Non-batch: Batch means that we are given a dataset, and learn a value function/policy using that fixed data
- We have only talked about the online algorithm that constantly gets new data
 - Each update in TD is on a the most recent experience
- How would the algorithm change for a batch of data?
- Can we still say the V (in TD) will converge to vpi?

1. (*Exercise 6.3 S&B*) From the results shown in the left graph of the random walk example it appears that the first episode results in a change in only V(A). What does this tell you about what happened on the first episode? Why was only the estimate for this one state changed? By exactly how much was it changed?



Estimated





this last transition is **Tocketstic** water istently better than the MC method on this task.



The left graph above shows the values learned after valious numbers of episodes 2. Assume the agenting teraots with The simple states to tep MDP as how it declow as Every episode begins in state X, and ends the true values gravith to an stant step size parameter to the terminal state in this example), the values fluctuate indefinitely in response to the outcomes (denoted by grave box) recent spisedes. The right graph shows dearning our ves for the two ere is only one possible actionethodactor states solutioner as the possible possible possible possible possible action this NBC. Let's denote the set of actions $\mathcal{A} = \{(A, A)\}$ in state \mathcal{A} the value function learned and the true value set of actions \mathcal{A} and function, averaged over the five states, then averaged over 100 runs. In all cases the gets proximate value function was initialized to the states are value of $\mathcal{A} = \{(A, A)\}$. $\alpha = .03$ α=.05 75 100 Episodes bers of episodes bout as close as they ever composite the product $P(R = r|Y) = \begin{cases} 0.5 & \text{if } r = -1000 \text{ ameter } (\alpha = 0.1 \\ 0.5 & \text{if } r = +1000 \text{ o the outcomes} \end{cases}$ of the most recommensation of the most reco

Write downtight the people the feven of the feven of the feven of the several of MDP that start from the track was initialized to the intermediate value V(s) = 0.5, for all s. The TD method was consistently better than the MC method on this task.

Deterministic transitions (X to Y to terminal) 1 action Stochastic reward from Y



- (d) $V(Y) = v_{\pi}(Y)?$

(f) What is the expected TD-update, from state Y for the given V?

(g)state X?

 $P(R = r|Y) = \begin{cases} 0.5 & \text{if } r = -1000 \\ 0.5 & \text{if } r = +1000 \end{cases}$

Assume our estimate is equal to the value of π . That is $V(s) = v_{\pi}(s) \forall s \in \mathcal{S}$. Now compute the TD-error $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ for the transition from state Y to the terminal state, assuming $R_{t+1} = +1000$. Why is the TD-error not zero if we start with

(e) Based on your answer to (e), what does this mean for the TD-update, for constant $\alpha = 0.1$? Will $V(Y) = v_{\pi}(Y) = 0$ after we update the value? Recall the TD-update is $V(S_t) \leftarrow$ $V(S_t) + \alpha \delta_t$. What does this tell us about the updates TD(0) would make on this MDP?

Assume still that $V = V_{\pi} = 0$. What is the expectation and the variance of the TD update from state X? What is the expectation and the variance of the Monte-carlo update from

= 1