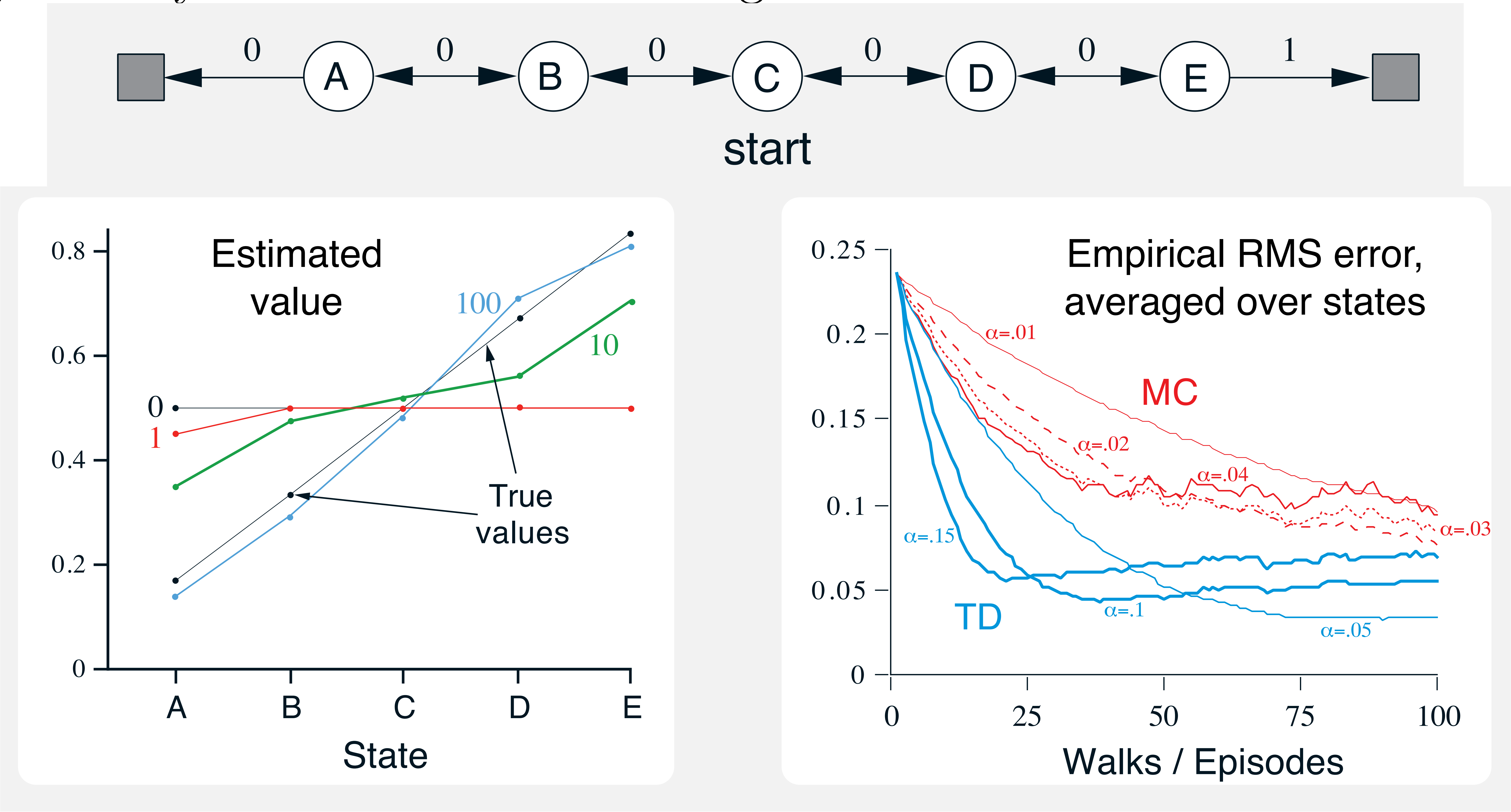
Course 2, Module 3
TD for Prediction

CMPUT 397
Fall 2019

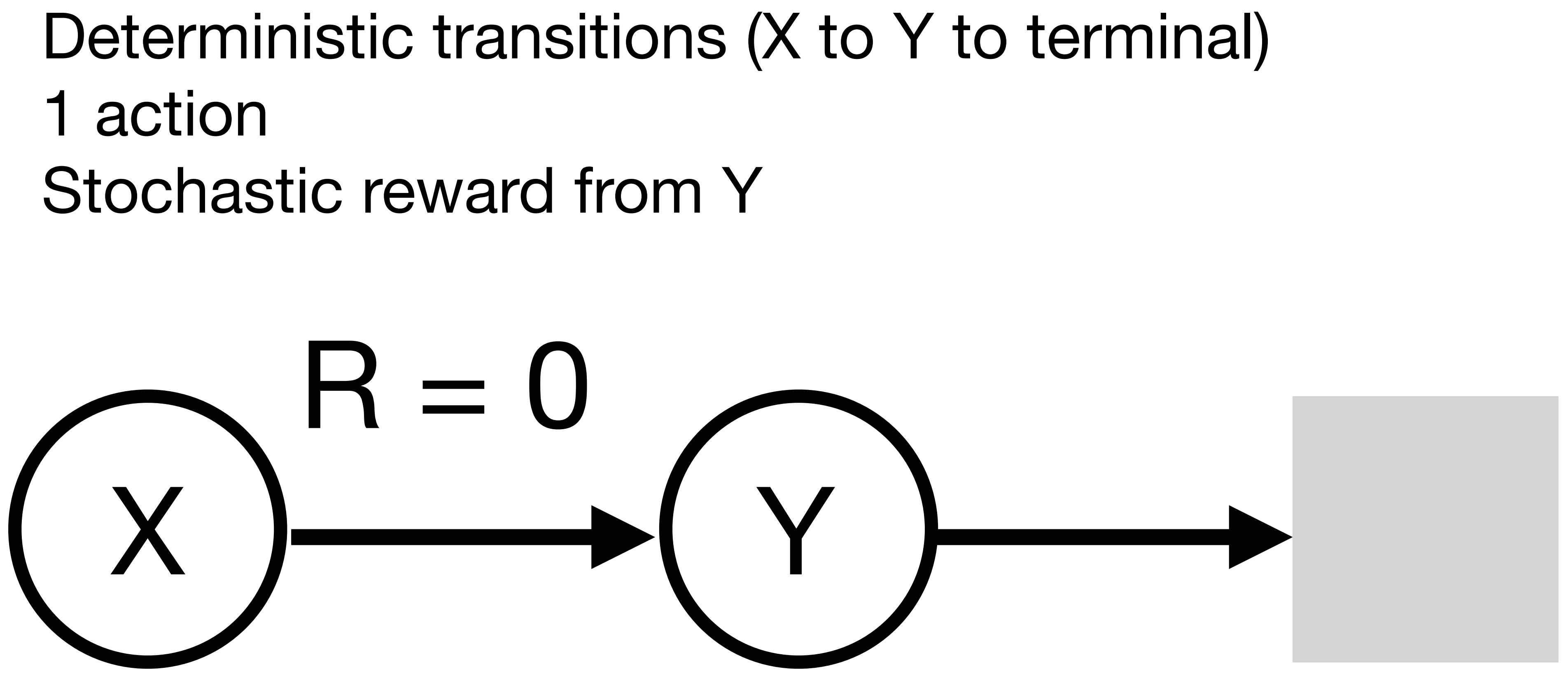
A few clarifications

- 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?



2. Assume the agent interacts with a simple two-state MDP shown below. Every episode begins in state X, and ends when the agent transitions from state Y to the terminal state (denoted by gray box). Let's denote the set of states as $\mathcal{S} = \{X, Y, T\}$. There is only one possible action in each state, so there is only one possible policy in this MDP. Let's denote the set of actions $\mathcal{A} = \{A\}$. In state Y the agent terminates when it takes action A and sometimes gets a reward of +1000, and sometimes gets a reward of -1000: the reward on this last transition is stochastic. Let $\gamma = 1.0$.



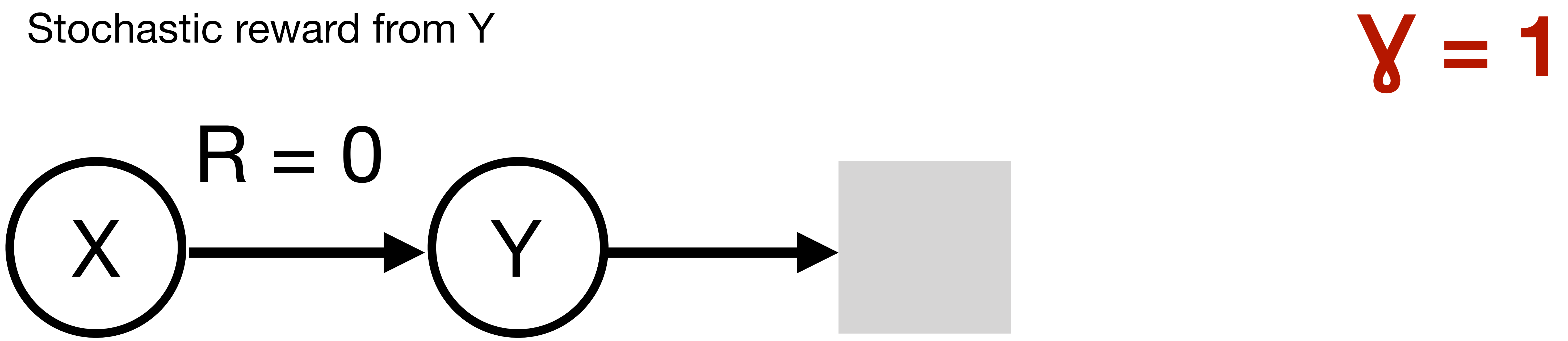
$$P(R = r|Y) = \begin{cases} 0.5 & \text{if } r = -1000 \\ 0.5 & \text{if } r = +1000 \end{cases}$$

(a) Write down $\pi(a|s)\ \forall\ s \in \mathcal{S}, a \in \mathcal{A}$

(b) Write down all the possible trajectories (sequence of states, actions, and rewards) in this MDP that start from state X?

(c) What the value of policy π (i.e. what is $v_\pi(X), v_\pi(Y)$)?

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



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

(d) 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 $V(Y) = v_\pi(Y)$?

(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?

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

(g) 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 state X?