Admin

- Sample midterm and mini-essay available on website (schedule)
- **Late submissions of quizzes and graded items:**
  - Policy: no late submission. Late = 0
  - Typically I review the quiz every Monday during lecture!
    - Quiz is due at noon.
    - How could it make sense to submit it after I review the answers?

Quiz review

Worksheet Question

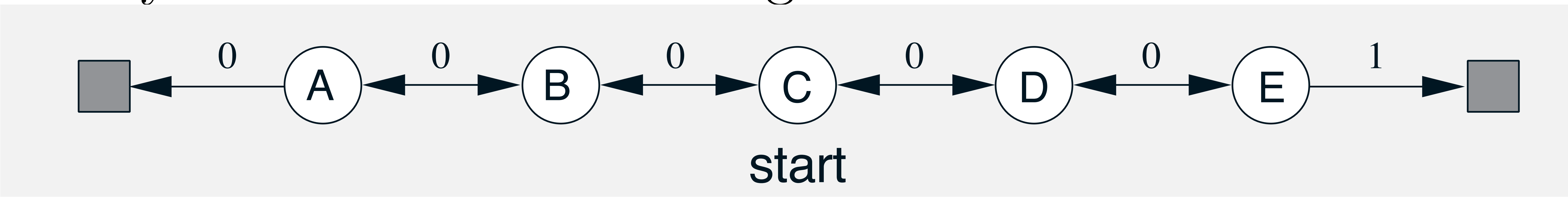
Modify the Tabular TD(0) algorithm for estimating v_π, to estimate q_π.

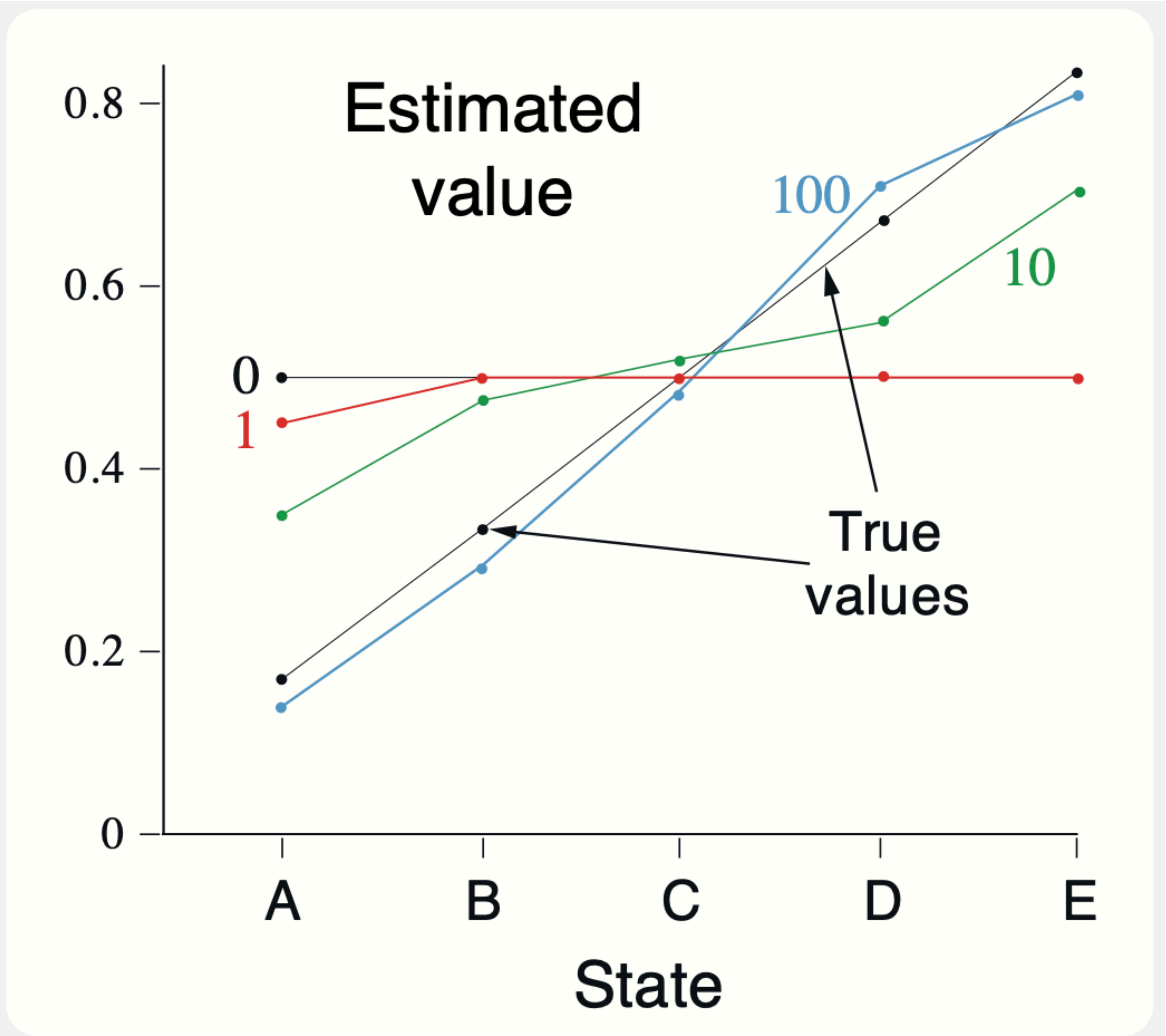
Tabular TD(0) for estimating v_π

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

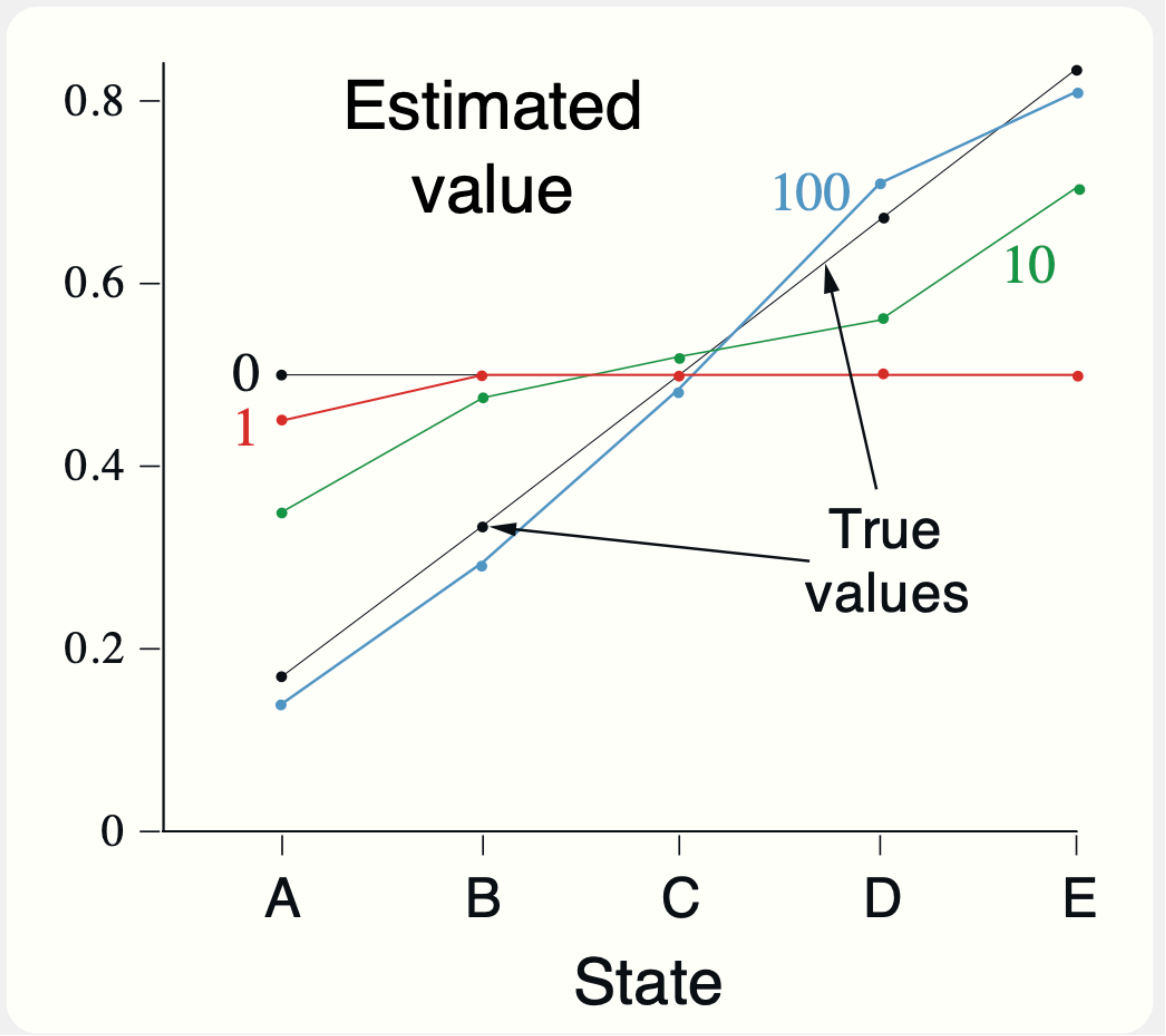
1. (*Exercise 6.3 S&B*) From the results shown in the left graph of the random walk example it appears that the first episode results in a change in only $V(A)$. What does this tell you about what happened on the first episode? Why was only the estimate for this one state changed? By exactly how much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

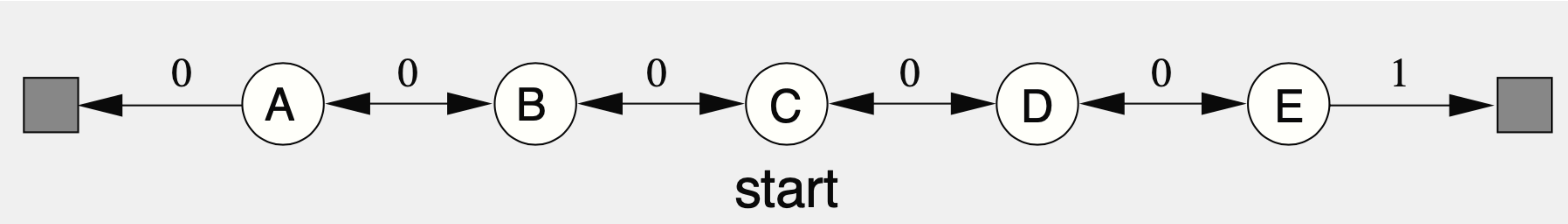




Exercise 6.3

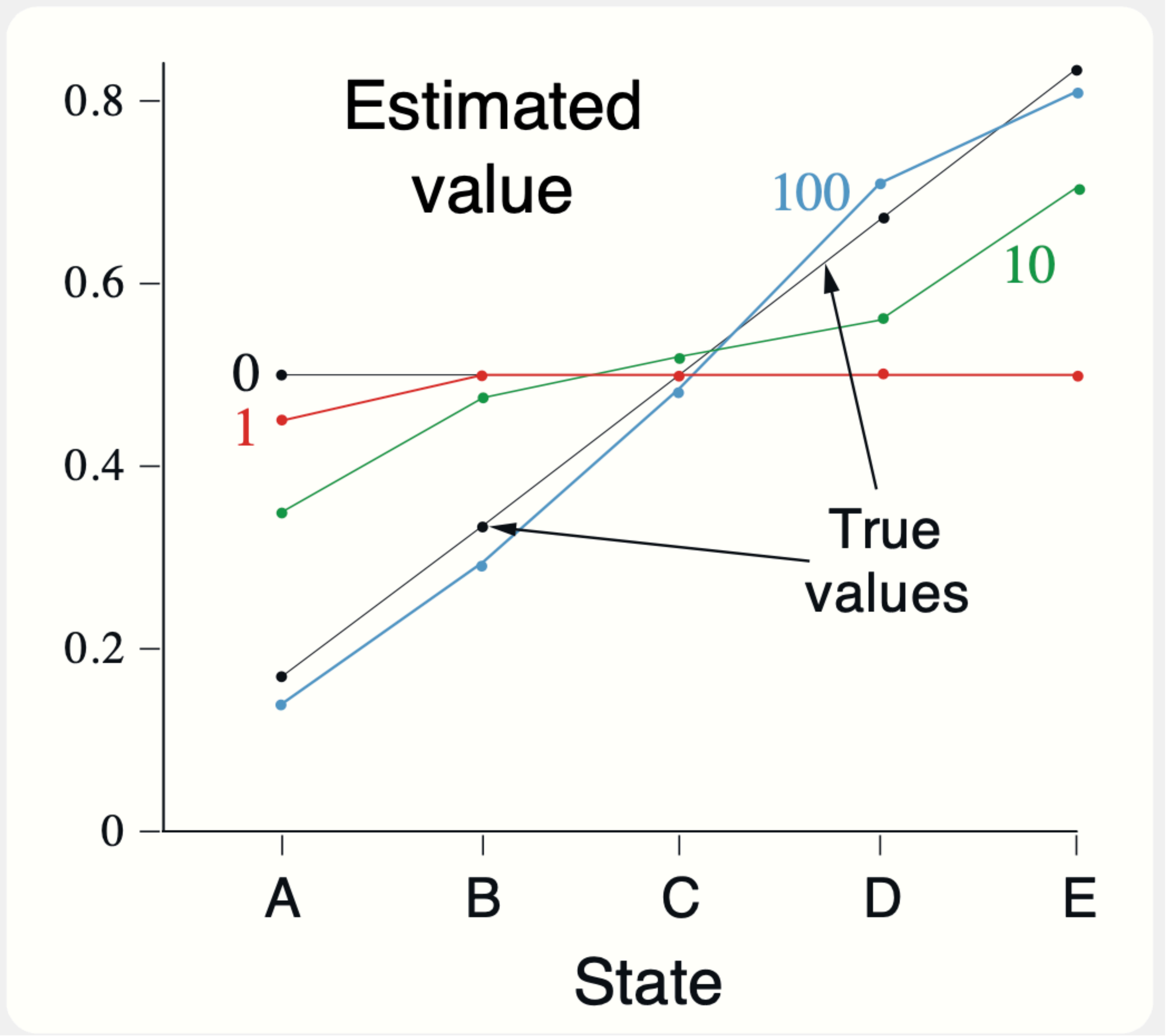
- What happened during the first episode? Why was only the estimate for this one state changed?

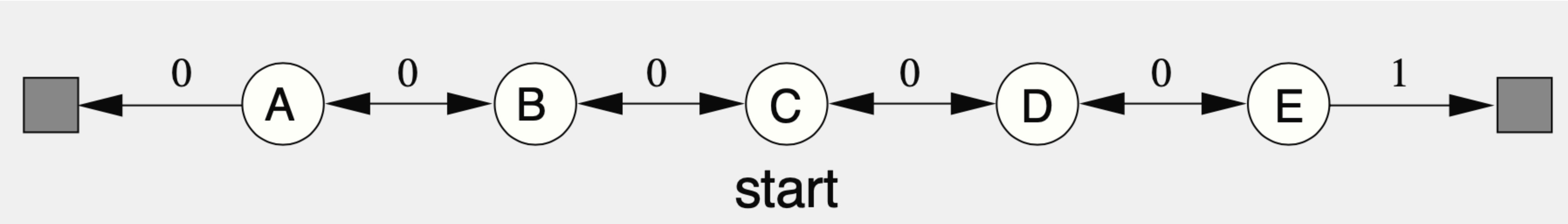




- Recall: V was set equal to 0.5 for all states

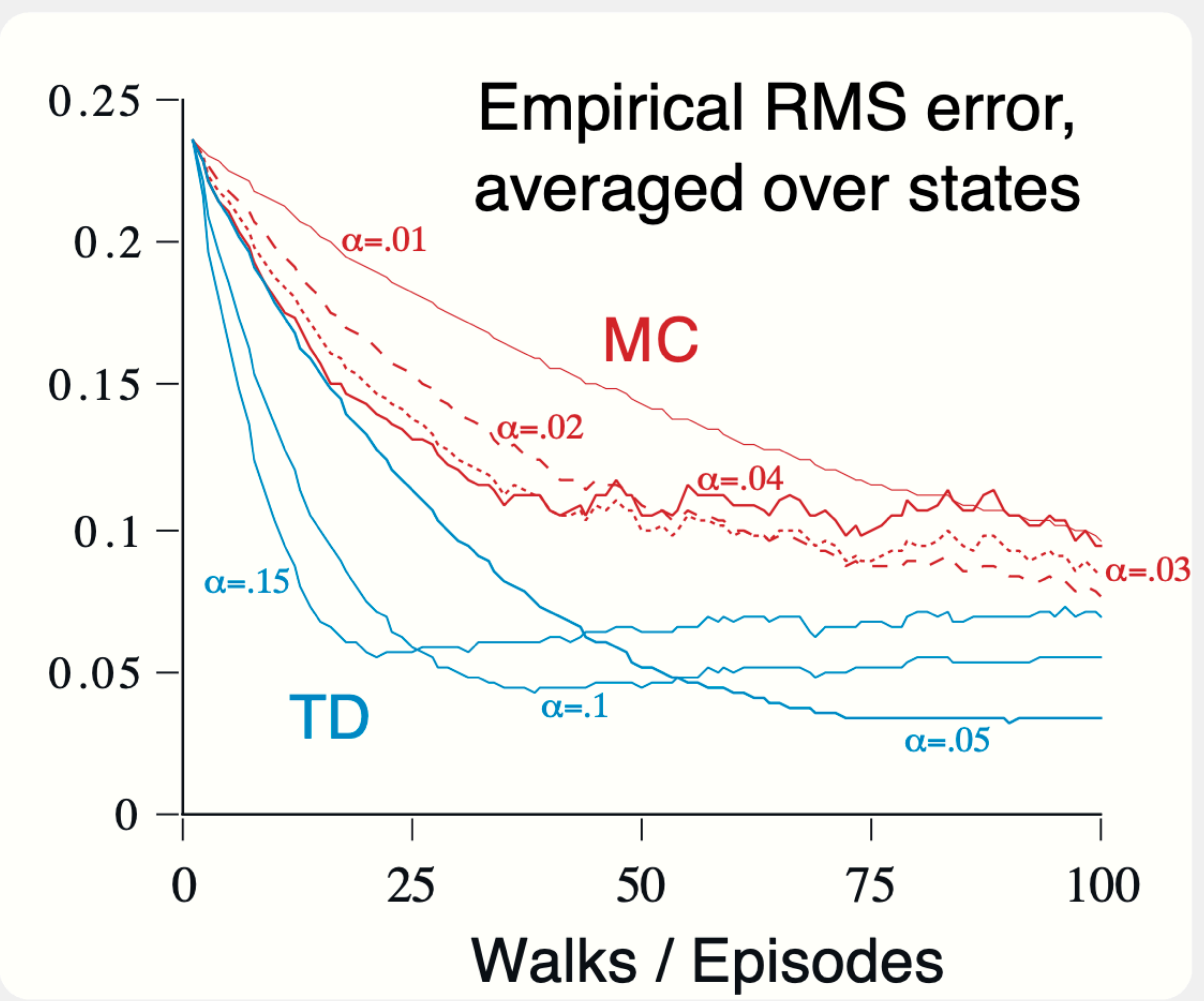
Exercise 6.3

- By exactly how much was V(A) changed?
- 



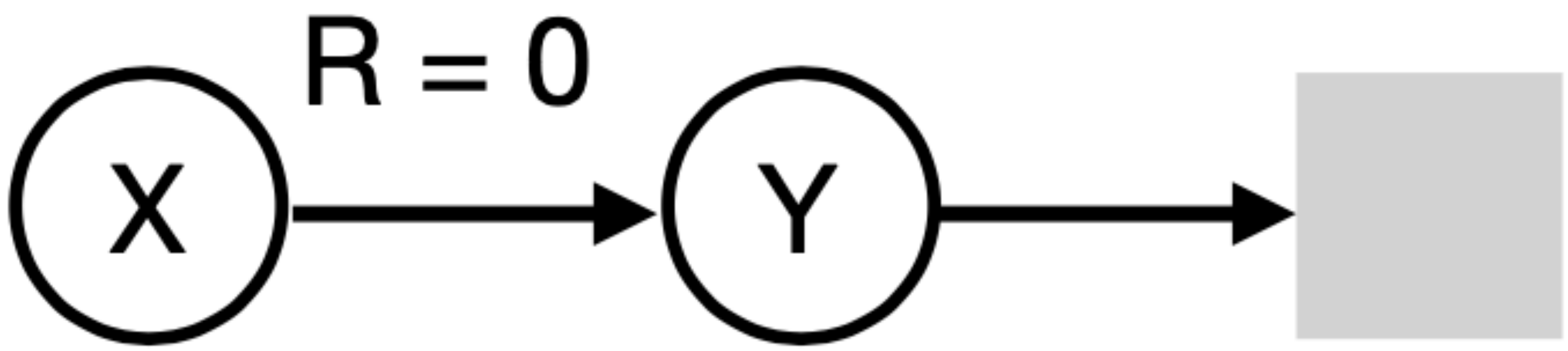
- Recall: V was set equal to 0.5 for all states

(b) (*Exercise 6.4 S&B*) The specific results shown in the right graph of the random walk example are dependent on the value of the step-size parameter, α. Do you think the conclusions about which algorithm is better would be affected if a wider range of α values were used? Is there a different, fixed value of α at which either algorithm would have performed significantly better than shown? Why or why not?



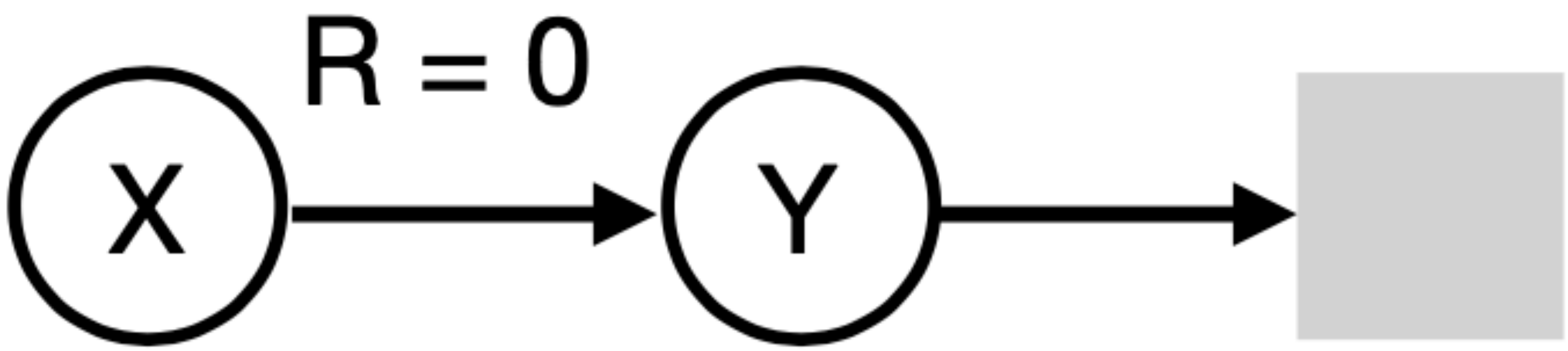
1. Assume the agent interacts with a simple two-state MDP shown below. Every episode begins in state X, and ends when the agent transitions from state Y to the terminal state (denoted by gray box). Let's denote the set of states as $\mathcal{S} = \{X, Y\}$. There is only one possible action in each state, 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 $\gamma = 1.0$.

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



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

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



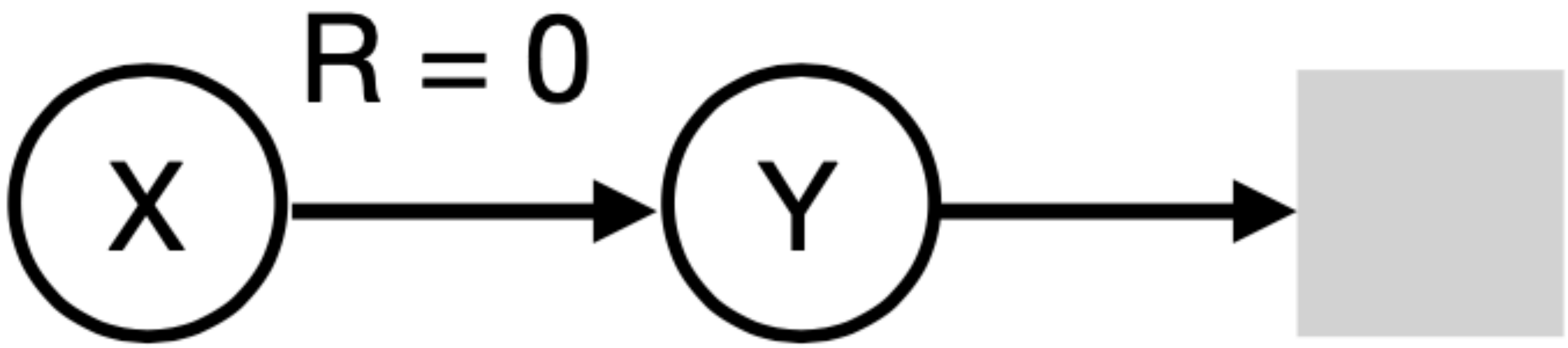
$$P(R = r|Y) = \begin{cases} 0.5 & \text{if } r = -1000 \\ 0.5 & \text{if } r = +1000 \end{cases}$$

(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 is 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

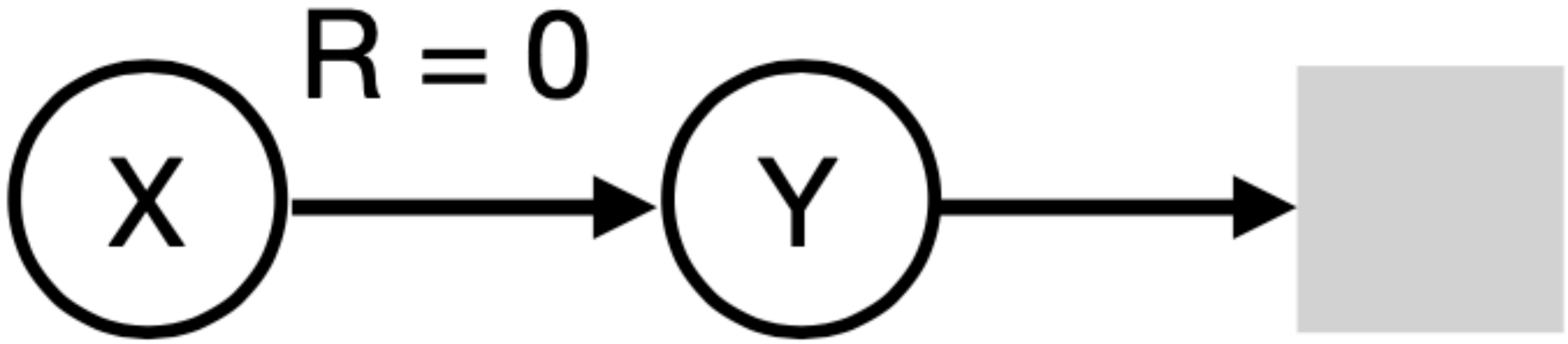


$$P(R = r|Y) = \begin{cases} 0.5 & \text{if } r = -1000 \\ 0.5 & \text{if } r = +1000 \end{cases}$$

(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 (d), 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 using TD? Recall the TD-update is $V(S_t) \leftarrow V(S_t) + \alpha\delta_t$.

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



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

(g) Assume again that $V = v_\pi$. 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?