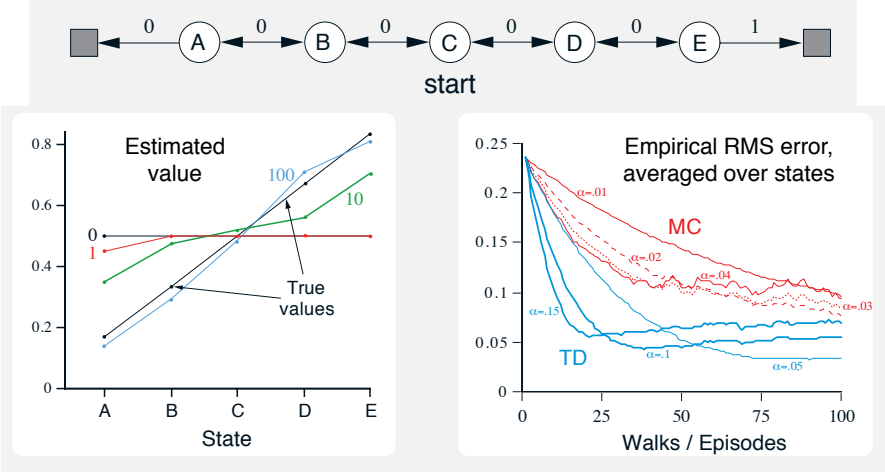
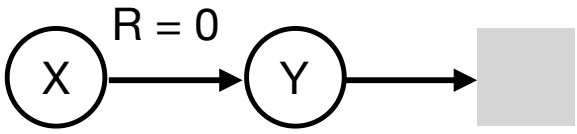
Worksheet 7

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 $\gamma = 1.0$.

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

Worksheet 7

(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?

3. (*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?

4. **(Challenge Question)** (*Exercise 6.5 S&B*) In the right graph of the random walk example, the RMS error of the TD method seems to go down and then up again, particularly at high α's. What could have caused this? Do you think this always occurs, or might it be a function of how the approximate value function was initialized?

5. (*Exercise 6.7 S&B*) Design an off-policy version of the TD(0) update that can be used with arbitrary target policy π and covering behavior policy b, using at each step t the importance sampling ratio $\rho_{t:t}$ (5.3).

6. 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 $V(\textit{terminal}) = 0$

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

Worksheet 7

7. (**Challenge Question**) In this question we consider the variance of the TD target, $R_{t+1} + \gamma V(S_{t+1})$ compared to the variance of the Monte Carlo target, G_t. Let's assume an idealized setting, where we have found a V that exactly equals v_π. We can show that, in this case, the variance of the Monte Carlo target is greater than or equal to the variance of the TD target. Note that variance of the targets is a factor in learning speed, where lower variance targets typically allow for faster learning. Show that the Monte Carlo target has at least as high of variance as the TD target, using the following decomposition, called the Law of Total Variance

$$\text{Var}(G_t) = \mathbb{E}[\text{Var}(G_t|S_{t+1})] + \text{Var}(\mathbb{E}[G_t|S_{t+1}]).$$

Note that the above means the variance of the return given state $S_t = s$: $\text{Var}(G_t|S_t = s)$. We omit the given $S_t = s$ from the above, to make it more readable, but implicitly it is there. Further, to better understand the double expectations,

$$\mathbb{E}[G_t] = \mathbb{E}[\mathbb{E}[G_t|S_{t+1}]] = \sum_{s' \in \mathcal{S}} p(s'|S_t = s)\mathbb{E}[G_t|S_{t+1}].$$

The outer expectation is over S_{t+1}, with the inner expectation computed given S_{t+1}. Similarly,

$$\mathbb{E}[\text{Var}(G_t|S_{t+1})] = \mathbb{E}[\mathbb{E}[G_t|S_{t+1}]] = \sum_{s' \in \mathcal{S}} p(s'|S_t = s)\text{Var}[G_t|S_{t+1}].$$

To answer the above question, try to simplify $\text{Var}(\mathbb{E}[G_t|S_{t+1}])$ by showing that $\mathbb{E}[G_t|S_{t+1}] = R_{t+1} + \gamma V(S_{t+1})$. Notice that this is only true because $V(S_{t+1}) = v_\pi(S_{t+1})$. Then conclude that this implies

$$\text{Var}(G_t) \geq \text{Var}(R_{t+1} + \gamma V(S_{t+1})).$$