Chapter 6 solutions **Hanna Gábor**

1. *If V changes during the episode, then (6.6) only holds approximately; what would the difference be between the two sides? Let V_t denote the array of state values used at time t in the TD error (6.5) and in the TD update (6.2). Redo the derivation above to determine the additional amount that must be added to the sum of TD errors in order to equal the Monte Carlo error.*

$$\begin{aligned}
G_t - V_t(S_t) &= R_{t+1} + \gamma G_{t+1} - V_t(S_t) + \gamma V_t(S_{t+1}) - \gamma V_t(S_{t+1}) + \gamma V_{t+1}(S_{t+1}) \\
&\quad - \gamma V_{t+1}(S_{t+1}) \\
&= (R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)) + (\gamma G_{t+1} - \gamma V_{t+1}(S_{t+1})) \\
&\quad + (\gamma V_{t+1}(S_{t+1}) - \gamma V_t(S_{t+1})) \\
&= \delta_t + \gamma(G_{t+1} - V_{t+1}(S_{t+1})) + \gamma\alpha\delta_{t+1} \\
&= \delta_t + \gamma\alpha\delta_{t+1} + \gamma\delta_{t+1} + \gamma^2\alpha\delta_{t+2} + \gamma^2(G_{t+2} - V_{t+2}(S_{t+2})) = \dots \\
&= \sum_{k=t}^{T-1} \gamma^{k-t}(1+\alpha)\delta_k
\end{aligned}$$

2. *This is an exercise to help develop your intuition about why TD methods are often more efficient than Monte Carlo methods. Consider the driving home example and how it is addressed by TD and Monte Carlo methods. Can you imagine a scenario in which a TD update would be better on average than a Monte Carlo update? Give an example scenario—a description of past experience and a current state—in which you would expect the TD update to be better. Here's a hint: Suppose you have lots of experience driving home from work. Then you move to a new building and a new parking lot (but you still enter the highway at the same place). Now you are starting to learn predictions for the new building. Can you see why TD updates are likely to be much better, at least initially, in this case? Might the same sort of thing happen in the original scenario?*

   I can't really see it.

   I have lots of experience, so the estimates that are made after I reached the highway are quite accurate, there will be only small updates for those. For the estimate before I reach the highway, the two update rules just give approximately the same. I can't see a big difference before that, either.

3. *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?*

   The first episode must have terminated on the left with a reward of 0.

   The other state values didn't change because the error in those cases was $R_{t+1} + \gamma V(S_{t+1}) - V(S_t) = 0 + 1 \cdot 0.5 - 0.5 = 0$.

   $v(A)$ changed by $\alpha(R_{t+1} + \gamma V(S_{t+1}) - V(S_t)) = 0.1(0 + 0 - 0.5) = -0.05$.

4. *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?*

   I don't think a wider range of α values would have a significant effect on the results.

   - **Monte Carlo methods.** The Monte Carlo version with $\alpha = 0.4$ starts to oscillate pretty fast and doesn't seem to decrease. I expect a Monte Carlo method with a bigger α value to oscillate around an error level which is at least as high as the error in the $\alpha = 0.4$ case. The smallest α for the MC method among the examples is 0.1. That is the only MC method which we can't see oscillate. This version does not achieve the performance of the TD methods. A smaller α value would make the learning progress even slower. It's possible that the error using Monte Carlo update with $\alpha =< 0.1$ goes under the TD error eventually, but it would still be pretty slow.
   - **TD methods.** The TD version with $\alpha = 1.5$ plateaued fast at a relatively high error. I expect bigger α values to make the convergence faster, but plateaue at a higher value. The version with $\alpha = 0.5$ is the best among all shown algorithms. A TD method with a smaller α value might reach an even lower error, but the $\alpha = 0.5$ case is already pretty good.

5. For now, let us treat the terminal points as states with values 0 and 1, respectively. Let all the rewards be 0. These changes together don't change the updates, but make the notation a bit easier. For a state $S \neq C$ let S^i denote the neighbor of C which is closer to C and S^o denote the neighbor of S which is further from C. Let $V_t(S)$ denote the estimate of $V(S)$ in time step t.

   For state C, the expected estimated value is 0.5 by symmetry. We will prove the following theorem.

   Theorem 1. *Let S be any state in $\{A, B, D, E\}$. Suppose we are in state S in time step t. Suppose also that*

   $$(1-\alpha)|V_t(S_t) - V_{t-2}(S^i)| < |V_{t-2}(S^o)) - V_t(S_t)|.$$

   Then $|\mathbb{E}(V_{t+1}(S_t)) - V_t(S^o)| < |V_t(S_t) - V_t(S^o)|$.

   For proving the theorem we need a few lemmas.

   Lemma 1. *If all the state values are initialized to* 0.5*, then*

   $$0 \leq V_t(A) \leq V_t(B) \leq V_t(C) \leq V_t(D) \leq V_t(E) \leq 1$$

   holds throughout the run of the TD algorithm.

   Proof. The proof goes by induction on the number of steps. The statement clearly holds in the beginning. Suppose you are in state S at time t and take a step to the right, arriving to S'. Then $V_{t+1}(S)$ equals to

   $$(1-\alpha)V_t(S) + \alpha V_t(S') \leq (1-\alpha)V_t(S') + \alpha V_t(S') = V_t(S') = V_{t+1}(S').$$

If you step to the left, then

$$(1-\alpha)V_t(S) + \alpha V_t(S') \geq (1-\alpha)V_t(S') + \alpha V_t(S') = V_t(S') = V_{t+1}(S').$$

□

Lemma 2. *Let S be any state in $\{A, B, D, E\}$. Suppose we are in state S in time step t and the previous state was S^i. Then $|\mathbb{E}(V_{t+1}(S_t)) - V_t(S^o)| < |V_t(S_t) - V_t(S^o)|$ if and only if*

$$(1-\alpha)|V_t(S_t) - V_{t-1}(S^i)| < |V_t(S^o)) - V_t(S_t)|.$$

Proof. Using the update rule $V_t(S^i) = (1-\alpha)V_{t-1}(S^i) + \alpha V_{t-1}(S_t)$ and $V_{t-1}(S_t) = V_t(S_t)$ the following holds.

$$\begin{aligned}
\mathbb{E}(V_{t+1}(S_t)) &= V_t(S_t) + \frac{\alpha}{2}\Big(\big(V_t(S^i) - V_t(S_t)\big) + \big(V_t(S^o) - V_t(S_t)\big)\Big) \\
&= V_t(S_t) + \frac{\alpha}{2}\Big(\big((1-\alpha)V_{t-1}(S^i) + \alpha V_{t-1}(S_t) - V_t(S_t)\big) \\
&+ \big(V_t(S^o) - V_t(S_t)\big)\Big) \\
&= V_t(S_t) + \frac{\alpha}{2}\Big(((1-\alpha)\big(V_{t-1}(S^i) - V_t(S_t)\big) \\
&+ \big(V_t(S^o)) - V_t(S_t)\big)\Big)
\end{aligned}$$

Hence

$$|\mathbb{E}(V_{t+1}(S_t)) - V_t(S^o)| < |V_t(S_t) - V_t(S^o)|$$

holds if and only if

$$|(1-\alpha)\big(V_{t-1}(S^i) - V_t(S_t)\big)| < |V_t(S^o)) - V_t(S_t)|.$$

□

Note that if $S_t \in \{A, E\}$, then $S_{t-1} = S^i$.

Lemma 3. *Let S be a state in $\{B, D\}$. Suppose we are in state S in time step t and the previous state was S^o. Then $|\mathbb{E}(V_{t+1}(S_t)) - V_{t-2}(S^o)| < |V_{t-2}(S_t) - V_{t-2}(S^o)|$ if and only if*

$$|V_{t-2}(S_t) - V_{t-2}(S^i)| < ((1-\alpha)^2 - \alpha + 2)|(V_{t-2}(S^o) - V_{t-2}(S_t))|$$

Proof. First note that $S_{t-2} = S_t$ and $S_{t-1} = S^o$. We will use the update rules

$V_t(S^o) = (1-\alpha)V_{t-1}(S^o) + \alpha V_{t-1}(S_t)$ and $V_{t-1}(S_t) = (1-\alpha)V_{t-2}(S_t) + \alpha V_{t-2}(S^o)$.

$$\begin{aligned}
\mathbb{E}(V_{t+1}(S_t)) &= V_t(S_t) + \frac{\alpha}{2}\Big(\big(V_t(S^i) - V_t(S_t)\big) + \big(V_t(S^o) - V_t(S_t)\big)\Big) \\
&= V_t(S_t) + \frac{\alpha}{2}\Big(\big(V_t(S^i) - V_t(S_t)\big) \\
&+ \big((1-\alpha)V_{t-1}(S^o) + \alpha V_{t-1}(S_t) - V_t(S_t)\big)\Big) \\
&= V_{t-1}(S_t) + \frac{\alpha}{2}\Big(\big(V_{t-2}(S^i) - V_{t-1}(S_t)\big) \\
&+ \big((1-\alpha)(V_{t-1}(S^o) - V_{t-1}(S_t))\big)\Big) \\
&= (1-\alpha)V_{t-2}(S_t) + \alpha V_{t-2}(S^o) \\
&+ \frac{\alpha}{2}\Big(\big(V_{t-2}(S^i) - (1-\alpha)V_{t-2}(S_t) - \alpha V_{t-2}(S^o)\big) \\
&+ \big((1-\alpha)(V_{t-2}(S^o) - (1-\alpha)V_{t-2}(S_t) - \alpha V_{t-2}(S^o))\big)\Big) \\
&= (1-\alpha)V_{t-2}(S_t) + \alpha V_{t-2}(S^o) \\
&+ \frac{\alpha}{2}\Big(\big(V_{t-2}(S^i) - V_{t-2}(S_t)\big) + \alpha\big(V_{t-2}(S_t) - V_{t-2}(S^o)\big) \\
&+ \big((1-\alpha)^2(V_{t-2}(S^o) - V_{t-2}(S_t))\big)\Big) \\
&= V_{t-2}(S_t) + \frac{\alpha}{2}\Big(\big(V_{t-2}(S^i) - V_{t-2}(S_t)\big) \\
&+ \big(((1-\alpha)^2 - \alpha + 2)(V_{t-2}(S^o) - V_{t-2}(S_t))\big)\Big)
\end{aligned}$$

Hence

$$|\mathbb{E}(V_{t+1}(S_t)) - V_{t-2}(S^o)| < |V_{t-2}(S_t) - V_{t-2}(S^o)|$$

holds if and only if

$$|V_{t-2}(S^i) - V_{t-2}(S_t)| < ((1-\alpha)^2 - \alpha + 2)|(V_{t-2}(S^o) - V_{t-2}(S_t))|$$

□

Proof of Theorem 1. $((1-\alpha)^2 - \alpha + 2) = 3 - 3\alpha + \alpha^2 \leq \frac{1}{1-\alpha}$.

$$(3 - 3\alpha + \alpha^2)(1-\alpha) \leq 1$$

$$3 - 6\alpha - 2\alpha^2 + \alpha^3$$

□