Chapter 6 solutions **Hanna Gábor**

1.

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

2. If I get lost a few times at the beginning, then I can see why TD updates can be better. When I get lost, only the first few states learn that it takes a long time for me to go home. After I reached the highway, the predictions won't change just because I don't know the way from the new office building to the highway.

3. The first episode must have terminated on the left with a reward of -1.

   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 $\alpha(R_{t+1} + \gamma V(S_{t+1}) - V(S_t)) = 0.1 * (-1 + 0 - 0.5) = -0.15$.

4.

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.

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

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

$$|V_{t-2}(S_i) - V_{t-2}(S_t)| + ((1-\alpha)^2 - \alpha + 2)|(V_{t-2}(S_o) - V_{t-2}(S_t))| > 0$$

□

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

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

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

□