CS410: Artificial Intelligence 2021 Fall
Homework 4: Reinforcement Learning & Regression
Due date: 23:59:59 (GMT +08:00), December 3 2021

1. **TD Learning**.

   (a) Consider an MDP with states A, B, C, D, E, and F, where state F is the terminal state. The agent will receive a reward $+1$ if it transits to the terminal state F or receive a reward 0 otherwise and the discount factor $\gamma = 1$. Assume the current estimates of V at time t is $V_t(A) = 0.2$, $V_t(B) = 0.4$, $V_t(C) = 0.6$, $V_t(D) = 0.8$, $V_t(E) = 1.0$, and $V_t(F) = 0$. Further, suppose that our agent is at state C at time t and it will experience the following transitions shown in Figure 1 from time $t+1$ to time $t+7$. Use temporal difference (TD) learning with learning rate $\alpha = 1/2$ to compute the new estimates of V values of corresponding states each time the agent experiences a new transition. Finish this exercise by explicitly showing your computation process.

   $V_{t+1}(C) =$ _0.5_, $V_{t+2}(B) =$ _0.3_, $V_{t+3}(A) =$ _0.25_, $V_{t+4}(B) =$ _0.4_, $V_{t+5}(C) =$ _0.65_, $V_{t+6}(D) =$ _0.9_, $V_{t+7}(E) =$ _1.0_.

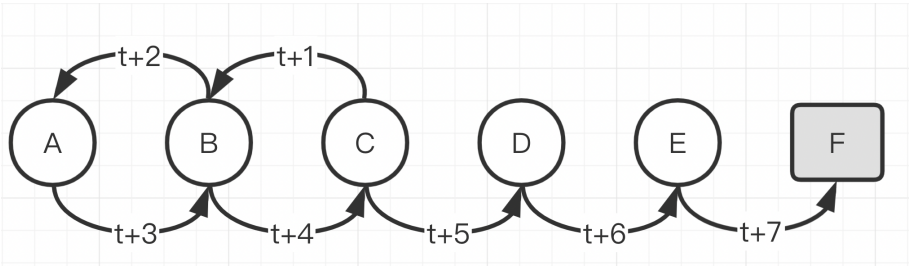

Figure 1: Problem 1 (a).

   (b) Update: Prove the statement in Lecture 7, Slide 38 that decreasing learning rate (α) can give converging averages in TD learning by verifying that the sequence $\{V_n\}$ with $V_n = (1 - \alpha_n)V_{n-1} + \alpha_n x_n$ is a Cauchy sequence under the assumption that $\forall n > 0$, $\alpha_n = \frac{1}{n^2}$, $|x_n| \leq C_1$ and $|V_n| \leq C_2$ for some constants $C_1 > 0$ and $C_2 > 0$.

Solution:

(a) $V_{t+1}(C) = V_t(C) + \alpha \cdot (r + \gamma V_{t+1}(B) - V_t(C))$ and other V values could be computed in the same way.

(b) *Proof.* $\forall n, m > 0$, it holds that

$$
\begin{aligned}
&|V_{n+m} - V_n| \\
&= \left| (1 - \prod_{i=n+1}^{n+m} (1 - \alpha_i)) V_n + \sum_{i=n+1}^{n+m} \alpha_i x_i \prod_{j=i+1}^{n+m} (1 - \alpha_j) \right| \\
&= \left| (1 - \prod_{i=n+1}^{n+m} (1 - \frac{1}{i^2})) V_n + \sum_{i=n+1}^{n+m} \frac{1}{i^2} x_i \prod_{j=i+1}^{n+m} (1 - \frac{1}{j^2}) \right| \\
&= \left| (1 - \prod_{i=n+1}^{n+m} (\frac{(i-1)(i+1)}{i^2})) V_n + \sum_{i=n+1}^{n+m} \frac{1}{i^2} x_i \prod_{j=i+1}^{n+m} (\frac{(j-1)(j+1)}{j^2}) \right| \\
&= \left| (1 - \frac{n(n+m+1)}{(n+1)(n+m)}) V_n + \sum_{i=n+1}^{n+m} \frac{1}{i^2} x_i \frac{i(n+m+1)}{(i+1)(n+m)} \right| \\
&\leq \left| (1 - \frac{n(n+m+1)}{(n+1)(n+m)}) C_2 \right| + \left| \frac{(n+m+1)C_1}{n+m} \sum_{i=n+1}^{n+m} \frac{1}{i(i+1)} \right| \\
&\leq \left| (1 - \frac{n(n+m+1)}{(n+1)(n+m)}) C_2 \right| + \left| \frac{(n+m+1)C_1}{n+m} (\frac{1}{n+1} - \frac{1}{n+m+1}) \right| ,
\end{aligned}
$$

which further implies that $\lim_{m>0, n\to\infty} |V_{n+m} - V_n| = 0$ and $\{V_n\}$ is a Cauchy sequence. □

2. **Q-Learning.** Recall the statement in Lecture 7, Slide 45 that Q-learning converges to optimal policy – even if you're acting suboptimally. The task in this exercise is to prove this statement step by step. Let us first cover some preliminaries. Denote by (X, d) the metric space, where X is a space and the metric $d : X \times X \to \mathbb{R}$ is defined on pairs of elements in X. For example, let $X = \mathbb{R}^n$ and $d(x_1, x_2) = \|x_1 - x_2\|_2$ for any $x_1, x_2 \in X$, where $\|\cdot\|_2$ denotes the 2-norm. Then $d(x_1, x_2)$ denotes the Euclidean distance between x_1 and x_2. Let $H : X \to X$ be a mapping between the space X and itself. If there exists some $0 < L < 1$, s.t. $\forall x, y \in X$, $d(H(x), H(y)) \leq L \cdot d(x, y)$, then the mapping H is defined as a contradiction mapping. If there exists a point $x \in X$ s.t. $H(x) = x$, then x is defined as the fixed-point of mapping H. If H is a contradiction mapping, then it could be shown that H admits a unique fixed-point x^* in X.

(a) Recall that the optimal Q^* satisfies the optimal Bellman equation

$$
Q^*(s, a) = \sum_{s'} T(s, a, s') \left[R(s, a, s') + \gamma \max_{a'} Q^*(s', a') \right] .
$$

If we treat X as the space of Q functions (i.e., each element q in X is a Q function, which further specifies a Q value $q(s,a)$ for a given state-action pair (s,a)), and treat H as

$$H : q(s,a) \mapsto \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} q(s',a') \right] .$$

Then it is clear that Q^* is the fixed-point of H since $Q^* = H(Q^*)$. Assume the discount factor $0 < \gamma < 1$. Prove that H is a contradiction mapping with respect to the metric

$$d(q_1, q_2) = ||q_1 - q_2||_\infty = \max_{s,a} |q_1(s,a) - q_2(s,a)|$$

for any two Q functions q_1 and q_2, where $\|\cdot\|_\infty$ is the maximum norm (i.e., prove that $\|H(q_1) - H(q_2)\|_\infty \leq \gamma \|q_1 - q_2\|_\infty$ for any two Q functions q_1 and q_2).

(b) Consider an finite MDP $(\mathcal{S}, \mathcal{A}, T, R, \gamma)$ with finite state space $\mathcal{S}$ and finite action space $\mathcal{A}$. Assume the reward function R is bounded and deterministic and $0 < \gamma < 1$. Recall that Q-learning updates Q function as

$$\begin{aligned} Q_{t+1}(s_t, a_t) &= Q_t(s_t, a_t) - \alpha_t (Q_t(s_t, a_t) - \text{sample}_t) \\ &= (1 - \alpha_t) Q_t(s_t, a_t) + \alpha_t \cdot \text{sample}_t \,, \end{aligned}$$

where $\text{sample}_t = R(s_t, a_t, s') + \gamma \max_{a'} Q_t(s', a')$, $0 \leq \alpha_t \leq 1$ is the learning rate at time t, and $\{s_t\}$ is the sequence of states obtained following policy π, which satisfies $\mathbb{P}_\pi [A_t = a \mid S_t = s] > 0$ for all state-action pairs (s,a). Prove that Q-learning converges to Q^* if $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$. First construct a sequence $\Delta_{t+1}(s,a) = Q_t(s,a) - Q^*(s,a)$ using the update rule of Q-learning and Q^*. Then verify that the three assumptions in Lemma 1 hold. Finally apply Lemma 1 to finish this exercise.

Update: Further assume that $Q^*(s,a)$ is bounded for all the state-action pair (s,a), and $Q_t(s,a)$ is bounded for all the state-action pair (s,a), $\forall t > 0$.

Lemma 1. *The random process $\{\Delta_t\}$ taking values in $\mathbb{R}$ and defined as*

$$\Delta_{t+1}(x) = (1 - \alpha_t) \Delta_t(x) + \alpha_t F_t(x)$$

converges to 0 under the following assumptions:

- $0 \leq \alpha_t \leq 1$, $\sum_t \alpha_t = \infty$ *and* $\sum_t \alpha_t^2 < \infty$;
- $\|\mathbb{E}[F_t \mid \mathcal{F}_t]\|_\infty \leq \gamma \|\Delta_t\|_\infty$ *with* $\gamma < 1$, *where* $\mathcal{F}_t = \{\Delta_t, \Delta_{t-1}, \ldots, \Delta_1, F_{t-1}, \ldots, F_1\}$ *stands for the past information at time t, and* $\mathbb{E}[F_t \mid \mathcal{F}_t]$ *denotes the conditional expectation of F_t given $\mathcal{F}_t$;*

- $\mathbb{V}\left[F_t(x) \mid \mathcal{F}_t\right] \leq C\left(1+\|\Delta_t\|_\infty\right)^2$ *for some constant* $C>0$*, where* $\mathbb{V}\left[F_t(x) \mid \mathcal{F}_t\right]$ *denotes the conditional variance of* $F_t(x)$ *given* $\mathcal{F}_t$*.*

Solution:

(a) *Proof.*

$$
\begin{aligned}
&\|H(q_1)-H(q_2)\|_\infty \\
&= \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| \sum_{s'\sim\mathcal{S}} T(s'|s,a) \left[\gamma \max_{a'\in\mathcal{A}} q_1(s',a') - \gamma \max_{a'\in\mathcal{A}} q_2(s',a') \right] \right| \\
&\leq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \gamma \sum_{s'\sim\mathcal{S}} T(s'|s,a) \left| \left[\max_{a'\in\mathcal{A}} q_1(s',a') - \max_{a'\in\mathcal{A}} q_2(s',a') \right] \right| \\
&\leq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \gamma \sum_{s'\sim\mathcal{S}} T(s'|s,a) \left| \max_{a'\in\mathcal{A}} (q_1(s',a') - q_2(s',a')) \right| \\
&\leq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \gamma \sum_{s'\sim\mathcal{S}} T(s'|s,a) \left| \max_{(s'',a')\in\mathcal{S}\times\mathcal{A}} (q_1(s'',a') - q_2(s'',a')) \right| \\
&= \gamma \|q_1-q_2\|_\infty \,,
\end{aligned}
$$

where the first inequality comes from Jensen's inequality. □

(b) *Proof.* Consider the Q-learning updates

$$
Q_{t+1}(s_t,a_t) = (1-\alpha_t) Q_t(s_t,a_t) + \alpha_t \left(R(s_t,a_t,s') + \gamma \max_{a'} Q_t(s',a') \right) .
$$

Substracting from both sides the quantity $Q^*(s_t,a_t)$ shows that

$$
\begin{aligned}
&Q_{t+1}(s_t,a_t) - Q^*(s_t,a_t) \\
&= (1-\alpha_t)\left(Q_t(s_t,a_t) - Q^*(s_t,a_t)\right) + \alpha_t \left(R(s_t,a_t,s') + \gamma \max_{a'} Q_t(s',a') - Q^*(s_t,a_t) \right) .
\end{aligned} \tag{1}
$$

Let $\Delta_t(s,a) = Q_t(s,a) - Q^*(s,a)$ and $F_t(s,a) = R(s,a,s') + \gamma \max_{a'\in\mathcal{A}}(Q_t(s',a') - Q^*(s,a))$, where $s' \sim T(\cdot|s,a)$. Then Eq.(1) could be reformulated as

$$
\Delta_{t+1}(s_t,a_t) = (1-\alpha_t)\Delta(s_t,a_t) + \alpha_t F_t(s_t,a_t) \,.
$$

Then Assumption 1 in Lemma 1 is satisfied by the given conditions in problem statement. Furthermore,

$$
\begin{aligned}
&\mathbb{E}[F_t(s_t,a_t)|\mathcal{F}_t] \\
&= \mathbb{E}_{s'\sim T(\cdot|s_t,a_t)} \left[R(s_t,a_t,s') + \gamma \max_{a'} Q_t(s',a') - Q^*(s_t,a_t) \right] \\
&= (H(Q_t))(s_t,a_t) - Q^*(s_t,a_t) \\
&= (H(Q_t))(s_t,a_t) - (H(Q^*))(s_t,a_t) \,,
\end{aligned}
$$

and hence $\|\mathbb{E}[F_t|\mathcal{F}_t]\|_\infty = \|H(Q_t) - H(Q^*)\|_\infty \leq \gamma\|Q_t - Q^*\|_\infty = \gamma\|\Delta_t\|_\infty$, which shows that Assumption 2 in Lemma 1 is satisfied. Finally,

$$
\begin{aligned}
&\mathbb{V}\left[F_t(s_t, a_t) \mid \mathcal{F}_t\right] \\
=&\mathbb{E}\left[(F_t(s_t, a_t) - \mathbb{E}[F_t(s_t, a_t) \mid \mathcal{F}_t])^2 \mid \mathcal{F}_t\right] \\
=&\mathbb{E}_{s'\sim T(\cdot|s_t,a_t)}\left[\left(R(s_t, a_t, s') + \gamma \max_{a'} Q_t(s', a') - Q^*(s_t, a_t) - (H(Q_t))(s_t, a_t) + Q^*(s_t, a_t)\right)^2\right] \\
=&\mathbb{E}_{s'\sim T(\cdot|s_t,a_t)}\left[\left(R(s_t, a_t, s') + \gamma \max_{a'} Q_t(s', a') - (H(Q_t))(s_t, a_t)\right)^2\right] \\
=&\mathbb{V}_{s'\sim T(\cdot|s_t,a_t)}\left[R(s_t, a_t, s') + \gamma \max_{a'} Q_t(s', a') \mid \mathcal{F}_t\right] \\
\leq& C' \leq C\left(1 + \|\Delta_t\|_\infty\right)^2 ,
\end{aligned}
$$

for some universal constants $C, C' > 0$ since the reward function R is bounded and $Q_t(s, a)$ is bounded for all the state-action pair (s, a), $\forall t > 0$. Then Assumption 3 in Lemma 1 is satisfied and Lemma 1 shows that $Q_t(s, a)$ converges to $Q^*(s, a)$. Hence Q-learning converges to Q^* since $\mathbb{P}_\pi\left[A_t = a \mid S_t = s\right] > 0$ for all state-action pairs (s, a). □

3. **Regression.** Recall the exercise in Lecture 8, Slide 21. Some economists say that the impact of GDP in 'current year' will have an effect on vehicle sales 'next year'. So whichever year GDP was less, the coming year sales were lower, and when GDP increased the next year vehicle sales also increased. Let's have the equation as $y = \theta_0 + \theta_1 x$, where y = number of vehicles sold in the year and x = GDP of the prior year. We need to find θ_0 and θ_1. Here is the data between 2011 and 2016.

   (a) What is the normal equation?

   (b) Suppose the GDP increasement in 2017 is 7%, how many vehicles will be sold in 2018?

   Solution:

   (a) We have the normal equation $X^\top X\theta = X^\top y$ with

$$
X = \begin{pmatrix} 1 & 6.2 \\ 1 & 6.5 \\ 1 & 5.48 \\ 1 & 6.54 \\ 1 & 7.18 \\ 1 & 7.93 \end{pmatrix} \quad y = \begin{pmatrix} 26.3 \\ 26.65 \\ 25.03 \\ 26.01 \\ 27.9 \\ 30.47 \end{pmatrix},
$$

   which leads to $\theta = (12.5494, 2.1859)$.

Year	GDP	Sales of vehicle
2011	6.2	
2012	6.5	26.3
2013	5.48	26.65
2014	6.54	25.03
2015	7.18	26.01
2016	7.93	27.9
2017		30.47
2018		

Figure 2: Problem 3.

(b) The GDP in 2017 will be $7.93 \times 1.07 = 8.4851$. Hence the number of vehicles sold in 2018 will be

$$12.5494 + 2.1859 \times 8.4851 = 31.0970.$$