

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.

$$V_{t+1}(C) = _, V_{t+2}(B) = _, V_{t+3}(A) = _, V_{t+4}(B) = _, \\ V_{t+5}(C) = _, V_{t+6}(D) = _, V_{t+7}(E) = _.$$

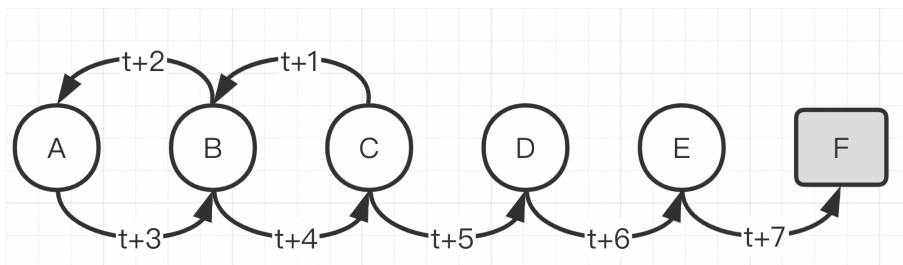


Figure 1: Problem 1 (a).

Solution.

The update algorithm provides that

$$\begin{aligned} \text{sample} &= R + \gamma \cdot V^\pi(s') \\ V^\pi(s) &\leftarrow V^\pi(s) + \alpha \cdot (\text{sample} - V^\pi(s)) \end{aligned} \tag{1}$$

So following this rule, we have

$$\begin{aligned}
V_{t+1}(C) &= 0.6 + 0.5 * (0.4 - 0.6) = 0.5 \\
V_{t+2}(B) &= 0.4 + 0.5 * (0.2 - 0.4) = 0.3 \\
V_{t+3}(A) &= 0.2 + 0.5 * (0.2 - 0.3) = 0.25 \\
V_{t+4}(B) &= 0.3 + 0.5 * (0.3 - 0.5) = 0.4 \\
V_{t+5}(C) &= 0.5 + 0.5 * (0.5 - 0.8) = 0.65 \\
V_{t+6}(D) &= 0.8 + 0.5 * (0.8 - 1.0) = 0.9 \\
V_{t+7}(E) &= 1 + 0.5 * (1 - 1) = 1
\end{aligned} \tag{2}$$

Thus we get the answer $V_{t+1}(C) = 0.5, V_{t+2}(B) = 0.3, V_{t+3}(A) = 0.25, V_{t+4}(B) = 0.4, V_{t+5}(C) = 0.65, V_{t+6}(D) = 0.9, V_{t+7}(E) = 1$

- (b) Prove the statement in Lecture 7, Slide 38 that decreasing learning rate (α) can give converging averages in TD learning.

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.

V_n will converges $\Leftrightarrow V_n$ is a Cauchy sequence, which means

$$\forall \epsilon > 0, \exists N \in \mathbb{N}, \forall n > m > N, |V_n - V_m| < \epsilon. \tag{3}$$

We know $x_n \leq C_1, |V_n| \leq C_2, \alpha_n = \frac{1}{n^2}$.

So for given ϵ , let $N = \lceil \frac{C_1 + C_2}{\epsilon} \rceil$. Then we have

$$\begin{aligned}
\forall t, |V_t - V_{t+1}| &= |\alpha_{t+1}V_t + \alpha_{t+1}x_{t+1}| \leq \frac{C_1 + C_2}{(t+1)^2} \\
\Rightarrow \forall n > m > N, |V_n - V_m| &= \sum_{k=m}^{n-1} |V_k - V_{k+1}| \\
&\leq (C_1 + C_2) \cdot \sum_{k=m}^{n-1} \frac{1}{(k+1)^2} \\
&< (C_1 + C_2) \cdot \sum_{k=m}^{n-1} \frac{1}{(k+1) \cdot k} \\
&\leq (C_1 + C_2) \cdot \sum_{k=m}^{n-1} \frac{1}{k} - \frac{1}{k+1} \\
&= (C_1 + C_2) \cdot \left(\frac{1}{m} - \frac{1}{n} \right) \\
&\leq (C_1 + C_2) \cdot \frac{1}{m} \\
&\leq (C_1 + C_2) \cdot \frac{1}{N} \leq \epsilon
\end{aligned} \tag{4}$$

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 $d : X \times X \rightarrow \mathbb{R}$ 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 $H : X \rightarrow X$ 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 contraction 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 contraction 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 contraction 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).

Solution.

First, since the definition of T , we have

$$\sum_{s'} T(s, a, s') = 1 \tag{5}$$

Then we have

$$\begin{aligned}
& \|H(q_1) - H(q_2)\|_\infty = \max_{s,a} |q_1(s,a) - q_2(s,a)| \\
& = \max_{s,a} \left| \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} q_1(s',a') \right] \right. \\
& \quad \left. - \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} q_2(s',a') \right] \right| \\
& = \max_{s,a} \left| \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} q_1(s',a') \right. \right. \\
& \quad \left. \left. - R(s,a,s') - \gamma \max_{a'} q_2(s',a') \right] \right| \\
& = \max_{s,a} \left| \sum_{s'} T(s,a,s') \left[\gamma \max_{a'} q_1(s',a') - \gamma \max_{a'} q_2(s',a') \right] \right| \tag{6} \\
& = \gamma \cdot \max_{s,a} \left| \sum_{s'} T(s,a,s') \left[\max_{a'} q_1(s',a') - \max_{a'} q_2(s',a') \right] \right| \\
& \leq \gamma \cdot \max_{s,a} \left| \sum_{s'} T(s,a,s') \|q_1 - q_2\|_\infty \right| \\
& \leq \gamma \|q_1 - q_2\|_\infty \cdot \max_{s,a} \sum_{s'} T(s,a,s') \\
& = \gamma \|q_1 - q_2\|_\infty
\end{aligned}$$

- (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 $\sum_t \alpha_t^2 < \infty$;
- $\|\mathbb{E}[F_t | \mathcal{F}_t]\|_\infty \leq \gamma \|\Delta_t\|_\infty$ with $\gamma < 1$, where $\mathcal{F}_t = \{\Delta_t, \Delta_{t-1}, \dots, \Delta_1, F_{t-1}, \dots, F_1\}$ stands for the past information at time t , and $\mathbb{E}[F_t | \mathcal{F}_t]$ denotes the conditional expectation of F_t given \mathcal{F}_t ;
- $\mathbb{V}[F_t(x) | \mathcal{F}_t] \leq C(1 + \|\Delta_t\|_\infty)^2$ for some constant $C > 0$, where $\mathbb{V}[F_t(x) | \mathcal{F}_t]$ denotes the conditional variance of $F_t(x)$ given \mathcal{F}_t .

Solution.

Firstly, construct $\Delta_{t+1}(s, a)$ that

$$\begin{aligned}
\Delta_{t+1}(s, a) &= Q_t(s, a) - Q^*(s, a) \\
&= (1 - \alpha_t) \cdot Q_t(s, a) + \alpha_t \left[R(s, a, s') + \gamma \max_{a'} q(s', a') \right] \\
&\quad - (1 - \alpha_t) \cdot Q^*(s, a) - \alpha_t \cdot Q^*(s, a) \\
&= (1 - \alpha_t) \cdot [Q_t(s, a) - Q^*(s, a)] \\
&\quad + \alpha_t \cdot \left[R(s, a, s') + \gamma \max_{a'} Q_t(s', a') - Q^*(s, a) \right] \\
&:= (1 - \alpha_t) \cdot \Delta_t(s, a) + F_t(s, a)
\end{aligned} \tag{7}$$

And for the **Lemma 1**:

- We set $\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$
- From above we know H is a contraction mapping and Q^* is the fixed-pointed, so we have

$$\begin{aligned}
\|\mathbb{E}[F_t | \mathcal{F}_t]\|_\infty &= \max_{s, a} \left| \sum_{s'} T(s, a, s') \left[R(s, a, s') + \gamma \max_{a'} Q^*(s', a') \right] - Q^*(s, a) \right| \\
&= \|H(Q_t(s, a)) - H(Q^*(s, a))\| \\
&\leq \gamma \|Q_t(s, a) - Q^*(s, a)\| \\
&= \gamma \|\Delta_t\|_\infty
\end{aligned} \tag{8}$$

- For the variance, and F_t is independent to Q^* , we have

$$\begin{aligned}
\mathbb{V}[F_t(s, a)] &= \mathbb{V}[F_t(s, a) + Q^*(s, a)] - \mathbb{V}[Q^*(s, a)] \\
&= \mathbb{V} \left[R(s, a, s') + \gamma \max_{a'} Q_t(s', a') \right] - 0 \\
&= \mathbb{E} \left[R(s, a, s') + \gamma \max_{a'} Q_t(s', a') \right]^2 - \mathbb{E}^2 \left[R(s, a, s') + \gamma \max_{a'} Q_t(s', a') \right]
\end{aligned} \tag{9}$$

Since for the conditional variance, and we know R, Q_t are both bounded, assume the bound is R_{max}, Q_{max} ,

so $\exists C = (R_{max} + \gamma \cdot Q_{max})^2$ that

$$\begin{aligned} \mathbb{V}[F_t | \mathcal{F}_t] &= \mathbb{E} \left[R(s, a, s') + \gamma \max_{a'} Q_t(s', a') | \mathcal{F}_t \right]^2 - \mathbb{E}^2 \left[R(s, a, s') + \gamma \max_{a'} Q_t(s', a') | \mathcal{F}_t \right] \\ &\leq \mathbb{E} \left[R(s, a, s') + \gamma \max_{a'} Q_t(s', a') | \mathcal{F}_t \right]^2 \leq C \end{aligned} \quad (10)$$

Finally, based on the **Lemma 1**, we know $\Delta_t \rightarrow 0$.

Since $\mathbb{P}_\pi[A_t = a | S_t = s] > 0$, we can conclude $Q_t(s, a) \rightarrow Q^*(s, a)$ for all state-action pairs.

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.

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.

- (a) What is the normal equation?

Solution. The formula for normal equation is

$$\mathbf{X}^T \mathbf{X} \theta = \mathbf{X}^T y \quad (11)$$

In this problem, we have

$$\mathbf{X} = \begin{bmatrix} 1 & 6.2 \\ 1 & 6.5 \\ 1 & 5.48 \\ 1 & 6.54 \\ 1 & 7.18 \\ 1 & 7.93 \end{bmatrix}, \quad y = \begin{bmatrix} 26.3 \\ 26.65 \\ 25.03 \\ 26.01 \\ 27.9 \\ 30.47 \end{bmatrix} \quad (12)$$

Then we have

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} 6.0 & 39.83 \\ 39.83 & 267.93 \end{bmatrix}, \quad \mathbf{X}^T y = \begin{bmatrix} 162.36 \\ 1085.5 \end{bmatrix} \quad (13)$$

So the solution of the equation is

$$\theta = \begin{bmatrix} 12.56 \\ 2.18 \end{bmatrix} \Rightarrow y = 12.56 + 2.18x \quad (14)$$

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

Solution. This time, we get $x = 7$, then we have $y = 12.56 + 2.18 \times 7 = 27.82$