به نام خدا

دانشکده مهندسی کامپیوتر

مبانی و کاربردهای هوش مصنوعی ترم پاییز ۱۴۰۰

پاسخ نامه تمرین سوم

سوال ۱)

الف) **مقدار مورد انتظار یک وضعیت** یا $V^*(s)$ عبارت است از **امتیاز نهایی مورد انتظار (میانگین)** چنانچه از وضعیت s شروع کنیم و MDP را طی کنیم.

ب) **سیاست بهینه در یک وضعیت** یا $\pi^*(s)$ عبارت است از **بهترین عملی** که در وضعیت s می توان انجام داد یا **عملی** که در وضعیت s انجام آن در نهایت به ما **بیشترین امتیاز مورد انتظار (میانگین)** را خواهد داد.

ج) از هر یک از دو رابطه بالا می‌توان به آن یکی رسید. در واقع هردوی آنها یک راه حل برای MDP مد نظر خواهند بود.

راه رسیدن از مقادیر (بهینه) به سیاست (بهینه) متناظر: از طریق فرمول زیر در یک مرحله قابل محاسبه است

$$\pi_{i+1}(s) = \arg\max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^{\pi_i}(s')\right]$$

راه رسیدن از سیاست (بهینه) به مقادیر (بهینه) متناظر: با استفاده از فرمول زیر و تا زمانی که مقادیر همگرا شوند باید آنها را به روز کنیم.

$$V^{\pi}_{k+1}(s) \leftarrow \sum_{s'} T(s,\pi(s),s')[R(s,\pi(s),s') + \gamma V^{\pi}_{k}(s')]$$

سوال ۲)

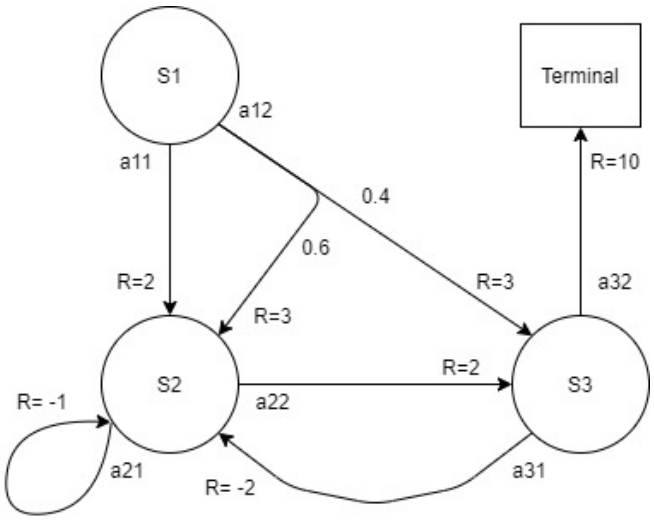
الف) **صحیح** – این ویژگی مارکو است که مستقل از گذشته است.

ب) **غلط** – با افزایش ضریب افق دید، تمرکز ما روی رویداد های قدیمی تر بیشتر خواهد شد.

ج) **غلط** – مقدار *value ربطی به پاداش بلافصل ندارد و باید مطابق فرمول محاسبه شود.

د) **صحیح** – با داشتن *Q می توان یک MDP را به نحو بهینه طی کرد که معادل حل شدن آن است.

سوال ۳)



الف) مقدار Value همه وضعیت‌ها (S) صفر است. مقادیر V وضعیت های مختلف را بر اساس فرمول به روز رسانی value (که در زیر آمده است) در **سه** مرحله به روز می کنیم.

$$V_{k+1}(s) \leftarrow \max_{a} \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s') \right]$$

$$V_0(s_1) = 0, V_0(s_2) = 0, V_0(s_3) = 0$$

———

$$V_1(s_1) = \max\big([2 + (0.5)(0)], [0.4\big(3 + (0.5)(0)\big) + 0.6\big(3 + (0.5)(0)\big)]\big) = 3$$

$$V_1(s_2) = \max([-1 + (0.5)(0)], [2 + (0.5)(0)]) = 2$$

$$V_1(s_3) = \max([-2 + (0.5)(0)], [10 + (0.5)(0)]) = 10$$

$$V_2(s_1) = \max\big([2 + (0.5)(2)], [0.4\big(3 + (0.5)(10)\big) + 0.6\big(3 + (0.5)(2)\big)]\big) = 5.6$$

$$V_2(s_2) = \max([-1 + (0.5)(2)], [2 + (0.5)(10)]) = 7$$

$$V_2(s_3) = \max([-2 + (0.5)(10)], [10 + (0.5)(0)]) = 10$$

$$V_3(s_1) = \max\big([2 + (0.5)(7)], [0.4\big(3 + (0.5)(10)\big) + 0.6\big(3 + (0.5)(7)\big)]\big) = 7.1$$

$$V_3(s_2) = \max([-1 + (0.5)(7)], [2 + (0.5)(10)]) = 7$$

$$V_3(s_3) = \max([-2 + (0.5)(10)], [10 + (0.5)(0)]) = 10$$

ب) **بله** همگرا خواهد شد – دلیل آن این است که MDP مد نظر عمق محدود دارد و دور هم ندارد و لذا اثبات می شود که با انجام این مراحل به صورت پی در پی در نهایت به مقادیر ثابتی خواهیم رسید. در واقع همین الان هم به نقطه همگرایی رسیدیم.

ج) بر اساس مقادیری که در بخش الف به دست آوردید، سیاست را تعیین می کنیم.

$$\pi\ (s) = \arg\max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V\ (s')]$$

$$\pi(S_1) = max\ \Big((2 + (0.5)(7)), 0.6\big(3 + (0.5)(7)\big) + 0.4\big(3 + (0.5)(10)\big)\Big)$$
$$\rightarrow \pi(S_1)\ = \boldsymbol{a12}$$

$$\pi(S_2) = max\ \Big((-1 + (0.5)(7)), (2 + (0.5)(10))\Big) \rightarrow\ \pi(S_2)\ = \boldsymbol{a22}$$

$$\pi(S_3) = max\ \Big((-2 + (0.5)(7)), (10 + 0)\Big) \rightarrow\ \pi(S_3)\ = \boldsymbol{a32}$$

د) طبق فرمول زیر، مقدار value هر وضعیت را تعیین می کنیم. دقت کنید که فقط عملی که سیاست فعلی تعیین می کند را بررسی می کنیم.

$$V_{k+1}^{\pi}(s) \leftarrow \sum_{s'} T(s, \pi(s), s')[R(s, \pi(s), s') + \gamma V_k^{\pi}(s')]$$

$V_0^{\pi}(s_1) = 0, V_0(s_2) = 0, V_0(s_3) = 0$

———

$V_1^{\pi}(s_1) = \big[0.4\big(3 + (0.5)(0)\big)\big] + \big[0.6\big(3 + (0.5)(0)\big)\big] = 3$

$V_1^{\pi}(s_2) = [2 + (0.5)(0)] = 2$

$V_1^{\pi}(s_3) = [10 + (0.5)(0)] = 10$

——

$V_2^{\pi}(s_1) = \big[0.4\big(3 + (0.5)(10)\big) + 0.6\big(3 + (0.5)(2)\big)\big] = 5.6$

$V_2^{\pi}(s_2) = [2 + (0.5)(10)] = 7$

$V_2^{\pi}(s_3) = [10 + (0.5)(0)] = 10$

——

$V_3^{\pi}(s_1) = \big[0.4\big(3 + (0.5)(10)\big) + 0.6\big(3 + (0.5)(7)\big)\big] = 7.1$

$V_3^{\pi}(s_2) = [2 + (0.5)(10)] = 7$

$V_3^{\pi}(s_3) = [10 + (0.5)(0)] = 10$

——

$V_4^{\pi}(s_1) = \big[0.4\big(3 + (0.5)(0)\big) + 0.6\big(3 + (0.5)(0)\big)\big] = 7.1$

$V_4^{\pi}(s_2) = [2 + (0.5)(10)] = 7$

$V_4^{\pi}(s_3) = [10 + (0.5)(0)] = 10$

همانطور که ملاحظه می شود در مرحله سوم همگرا شد. (همچنین دقت کنید که نتیجه با بخش (ب) نیز یکسان است و این تصادفی نیست چرا که همانطور که گفتیم از سیاست بهینه می توان مقادیر بهینه value را نیز به دست آورد)

ه) سیاست را (طبق معادله زیر) به روز می کنیم.

$$\pi_{i+1}(s) = \arg\max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^{\pi_i}(s') \right]$$

$$\pi(S_1) = max\left((2 + (0.5)(7)), 0.6(3 + (0.5)(7)) + 0.4(3 + (0.5)(10))\right)$$
$$\rightarrow \pi(S_1) = \boldsymbol{a12}$$

$$\pi(S_2) = max\left((-1 + (0.5)(7)), (2 + (0.5)(10))\right) \rightarrow \pi(S_2) = \boldsymbol{a22}$$

$$\pi(S_3) = max\left((-2 + (0.5)(7)), (10 + 0)\right) \rightarrow \pi(S_3) = \boldsymbol{a32}$$

همانطور که انتظار می رفت سیاستی که از روی این مقادیر بهینه به دست می آید با سیاستی که آنها را تولید کرده بود (بخش ج) یکسان است. همه این مراحل نشان می دهد که سیاست بهینه و مقادیر بهینه پاسخ نهایی یکسان با ظاهر های متفاوت برای یک مسئله هستند.

سوال ۴)

الف)

1) مسئله: رانندگی خودکار

عامل: ماشین خودران

محیط: خیابان شامل ماشین‌های دیگر و عابرین پیاده

2) مسئله: شطرنج

عامل: موتور شطرنج

محیط: بازی شطرنج شامل حریف مقابل

3) مسئله: معاملات سهام

عامل: معامله کننده سهام

محیط: قیمت‌ها و شاخص‌های سهام، اخبار مرتبط با سهام‌ها و ...

ب)

یادگیری مبتنی بر مدل و بدون مدل هر دو معایب و مزیت‌های خود را دارند. روش‌های مبتنی بر مدل به عامل اجازه برنامه‌ریزی و استنتاج بر اساس مدل را می‌دهند اما روش‌های بدون مدل تنها بر یادگیری تکیه می‌کنند. طور کلی با وجود آخرین پیشرفت‌های الگوریتم‌های یادگیری تقویتی، اگر همراه با تعریف مسئله (در دنیای واقعی) یک مدل دقیق ارائه نشده باشد، روش‌های بدون مدل برتری دارند زیرا کوچک‌ترین در خطا در مدل می‌تواند باعث ناپایداری عامل شود.

1) رانندگی خودکار: ساختن یک مدل دقیق دشوار است بنابراین **یادگیری بدون مدل** برتری دارد.

2) شطرنج: ساخت یک مدل دقیق آسان است پس **یادگیری با مدل** برتری دارد.

3) معاملات سهام: ساخت یک مدل دقیق دشوار است پس **یادگیری بدون مدل** برتری دارد.

ج)

1) رانندگی خودکار: عامل باید بین عبور از مسیرهایی که قبلا طی کرده (استخراج) و مسیرهای جدید (اکتشاف) انتخاب کند.

2) شطرنج: عامل باید بین موقعیت‌ها یا شروع بازی‌هایی که قبلا تجربه کرده (استخراج) و موقعیت‌های جدید و ناشناخته (اکتشاف) انتخاب کند.

3) معاملات سهام: عامل باید بین خرید سهام‌هایی که تجربه خرید آن‌ها را داشته (استخراج) و خرید سهام شرکت‌های جدید و نوپا (اکتشاف) انتخاب کند.

سوال ۵)

الف)

V(A1) = (0.2 + -1.5 + -1.5 + -1.7 + -1.8 + -2.0 + -1.6 + 0.3 + 0.0) / 9 = -1.067

V(A2) = (-1.3 + -1.4 + -1.3 + -1.4 + -1.6 + -1.8 + -1.9 + -1.3 + -1.4 + -1.5 + 0.4 + 0.2 + 0.1) / 13 = -1.092

V(A3) = (-1.2 + -1.2 + -1.3 + -1.5 + -1.7 + -1.2 + 0.5) / 7 = -1.086

V(A4) = (-1.1 + -1.1) / 2 = -1.100

V(B1) = (0.3) / 1 = 0.300

V(B3) = (-1.1 + -1.1 + -1.2 + -1.4 + -1.6 + 0.7 + 0.6) / 7 = -0.729

V(B4) = -1

V(C1) = (0.7 + 0.6 + 0.5 + 0.4) / 4 = 0.550

V(C2) = (0.8) / 1 = 0.800

V(C3) = (0.9 + 0.9 + 0.8) / 3 = 0.867

V(C4) = +1

ب)

Episode #1

Sample #1 = -0.1 + 0.000 = -0.100

Q(A1, U) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #2 = -0.1 + 0.000 = -0.100

Q(B1, U) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #3 = -0.1 + 0.000 = -0.100

Q(C1, W) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #4 = -0.1 + 0.000 = -0.100

Q(C1, U) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #5 = -0.1 + 0.000 = -0.100

Q(C1, E) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #6 = -0.1 + 0.000 = -0.100

Q(C1, E) = -0.050 + 1/2 * (-0.100 - -0.050) = -0.075

Sample #7 = -0.1 + 0.000 = -0.100

Q(C2, E) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #8 = -0.1 + 1 = 0.9

Q(C3, E) = 0.000 + 1/2 * (0.900 - 0.000) = 0.450

Episode #2

Sample #1 = -0.1 + 0.000 = -0.100

Q(A1, E) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #2 = -0.1 + 0.000 = -0.100

Q(A2, W) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #3 = -0.1 + 0.000 = -0.100

Q(A2, S) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #4 = -0.1 + 0.000 = -0.100

Q(A3, E) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #5 = -0.1 + -1 = -1.1

Q(A4, U) = 0.000 + 1/2 * (-1.100 - 0.000) = -0.550

Episode #3

Sample #1 = -0.1 + 0.000 = -0.100

Q(A1, S) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #2 = -0.1 + 0.000 = -0.100

Q(A1, E) = -0.050 + 1/2 * (-0.100 - -0.050) = -0.075

Sample #3 = -0.1 + 0.000 = -0.100

Q(A2, W) = -0.050 + 1/2 * (-0.100 - -0.050) = -0.075

Sample #4 = -0.1 + 0.000 = -0.100

Q(A1, E) = -0.075 + 1/2 * (-0.100 - -0.075) = -0.088

Sample #5 = -0.1 + 0.000 = -0.100

Q(A2, U) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #6 = -0.1 + 0.000 = -0.100

Q(A2, E) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #7 = -0.1 + 0.000 = -0.100

Q(A3, U) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #8 = -0.1 + -1 = -1.1

Q(B3, E) = 0.000 + 1/2 * (-1.100 - 0.000) = -0.550

Episode #4

Sample #1 = -0.1 + -0.050 = -0.150

Q(A1, E) = -0.088 + 1/2 * (-0.150 - -0.088) = -0.119

Sample #2 = -0.1 + -0.050 = -0.150

Q(A2, E) = -0.050 + 1/2 * (-0.150 - -0.050) = -0.100

Sample #3 = -0.1 + 0.000 = -0.100

Q(A2, E) = -0.100 + 1/2 * (-0.100 - -0.100) = -0.100

Sample #4 = -0.1 + 0.000 = -0.100

Q(A3, U) = -0.050 + 1/2 * (-0.100 - -0.050) = -0.075

Sample #5 = -0.1 + 0.000 = -0.100

Q(B3, S) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #6 = -0.1 + 0.000 = -0.100

Q(A3, U) = -0.075 + 1/2 * (-0.100 - -0.075) = -0.088

Sample #7 = -0.1 + 0.000 = -0.100

Q(B3, S) = -0.050 + 1/2 * (-0.100 - -0.050) = -0.075

Sample #8 = -0.1 + 0.000 = -0.100

Q(A3, U) = -0.088 + 1/2 * (-0.100 - -0.088) = -0.094

Sample #9 = -0.1 + 0.000 = -0.100

Q(B3, W) = 0.000 + 1/2 * (-0.100 - 0.000) = -0.050

Sample #10 = -0.1 + -1 = -1.1

Q(B3, E) = -0.550 + 1/2 * (-1.100 - -0.550) = -0.825

Episode #5

Sample #1 = -0.1 + -0.050 = -0.150

Q(A1, E) = -0.119 + 1/2 * (-0.150 - -0.119) = -0.134

Sample #2 = -0.1 + -0.050 = -0.150

Q(A2, S) = -0.050 + 1/2 * (-0.150 - -0.050) = -0.100

Sample #3 = -0.1 + -0.050 = -0.150

Q(A2, U) = -0.050 + 1/2 * (-0.150 - -0.050) = -0.100

Sample #4 = -0.1 + 0.000 = -0.100

Q(A2, E) = -0.100 + 1/2 * (-0.100 - -0.100) = -0.100

Sample #5 = -0.1 + 0.000 = -0.100

Q(A3, E) = -0.050 + 1/2 * (-0.100 - -0.050) = -0.075

Sample #6 = -0.1 + -1 = -1.1

Q(A4, U) = -0.550 + 1/2 * (-1.100 - -0.550) = -0.825

Episode #6

Sample #1 = -0.1 + -0.075 = -0.175

Q(A1, E) = -0.134 + 1/2 * (-0.175 - -0.134) = -0.155

Sample #2 = -0.1 + -0.075 = -0.175

Q(A2, S) = -0.100 + 1/2 * (-0.175 - -0.100) = -0.138

Sample #3 = -0.1 + 0.000 = -0.100

Q(A2, W) = -0.075 + 1/2 * (-0.100 - -0.075) = -0.088

Sample #4 = -0.1 + -0.088 = -0.188

Q(A1, E) = -0.155 + 1/2 * (-0.188 - -0.155) = -0.171

Sample #5 = -0.1 + 0.000 = -0.100

Q(A2, W) = -0.088 + 1/2 * (-0.100 - -0.088) = -0.094

Sample #6 = -0.1 + 0.000 = -0.100

Q(A3, U) = -0.094 + 1/2 * (-0.100 - -0.094) = -0.097

Sample #7 = -0.1 + 0.000 = -0.100

Q(B3, W) = -0.050 + 1/2 * (-0.100 - -0.050) = -0.075

Sample #8 = -0.1 + 0.450 = 0.350

Q(B3, S) = -0.075 + 1/2 * (0.350 - -0.075) = 0.137

Sample #9 = -0.1 + 0.450 = 0.350

Q(C3, S) = 0.000 + 1/2 * (0.350 - 0.000) = 0.175

Sample #10 = -0.1 + 1 = 0.9

Q(C3, E) = 0.450 + 1/2 * (0.900 - 0.450) = 0.675

Final Q-values

Q(A1, U) = -0.050

Q(A1, E) = -0.171

Q(A1, S) = -0.050

Q(A1, W) = 0.000

Q(A2, U) = -0.100

Q(A2, E) = -0.100

Q(A2, S) = -0.138

Q(A2, W) = -0.094

Q(A3, U) = -0.097

Q(A3, E) = -0.075

Q(A3, S) = 0.000

Q(A3, W) = 0.000

Q(A4, U) = -0.825

Q(A4, E) = 0.000

Q(A4, S) = 0.000

Q(A4, W) = 0.000

Q(B1, U) = -0.050

Q(B1, E) = 0.000

Q(B1, S) = 0.000

Q(B1, W) = 0.000

Q(B3, U) = 0.000

Q(B3, E) = -0.825

Q(B3, S) = 0.137

Q(B3, W) = -0.075

Q(C1, U) = -0.050

Q(C1, E) = -0.075

Q(C1, S) = 0.000

Q(C1, W) = -0.050

Q(C2, U) = 0.000

Q(C2, E) = -0.050

Q(C2, S) = 0.000

Q(C2, W) = 0.000

Q(C3, U) = 0.000

Q(C3, E) = 0.675

Q(C3, S) = 0.175

Q(C3, W) = 0.000

ج)

V(A1) = max(-0.050, -0.171, -0.050, 0.000) = 0.000

V(A2) = max(-0.100, -0.100, -0.138, -0.094) = -0.094

V(A3) = max(-0.097, -0.075, 0.000, 0.000) = 0.000

V(A4) = max(-0.825, 0.000, 0.000, 0.000) = 0.000

V(B1) = max(-0.050, 0.000, 0.000, 0.000) = 0.000

V(B3) = max(0.000, -0.825, 0.137, -0.075) = 0.137

V(C1) = max(-0.050, -0.075, 0.000, -0.050) = 0.000

V(C2) = max(0.000, -0.050, 0.000, 0.000) = 0.000

V(C3) = max(0.000, 0.675, 0.175, 0.000) = 0.675

نتایج با بخش (الف) بسیار متفاوت است زیرا در یادگیری تفاوت زمانی باید بر روی تعداد بسیار زیادی نمونه یادگیری صورت گیرد تا توابع ارزش و Q به سمت جواب درست همگرا شوند.

سوال ۶)

در یادگیری برون‌سیاست، سیاستی که برای به‌روزرسانی تابع ارزش استفاده می‌شود و سیاستی که برای انجام حرکات مورد استفاده قرار می‌گیرند، یکسان نیستند. یادگیری Q یک مثال از یادگیری برون‌سیاست می‌باشد. در یادگیری برسیاست، عامل از همان سیاستی برای انجام حرکات استفاده می‌کند که برای به‌روزرسانی تابع ارزش نیز مورد استفاده قرار می‌دهد. یک مثال از یادگیری برسیاست الگوریتم Sarsa می‌باشد.

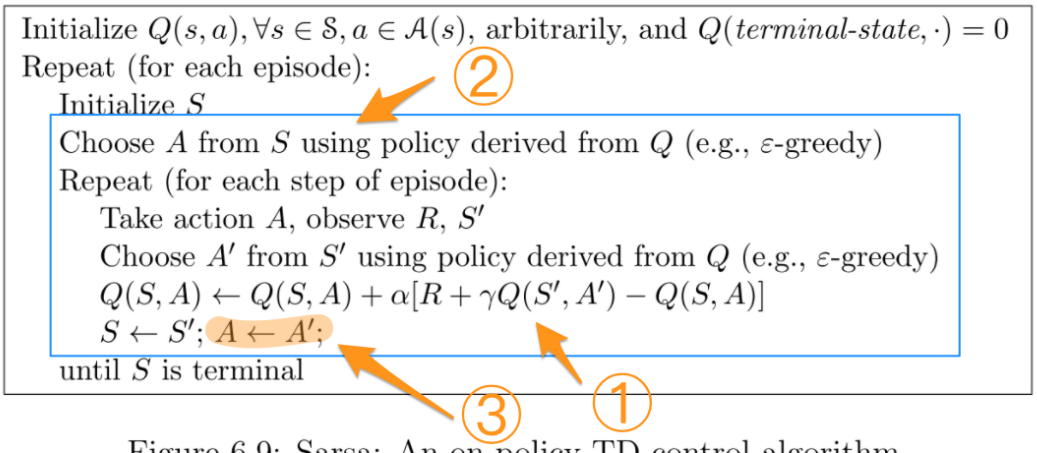


Figure 6.9: Sarsa: An on-policy TD control algorithm.

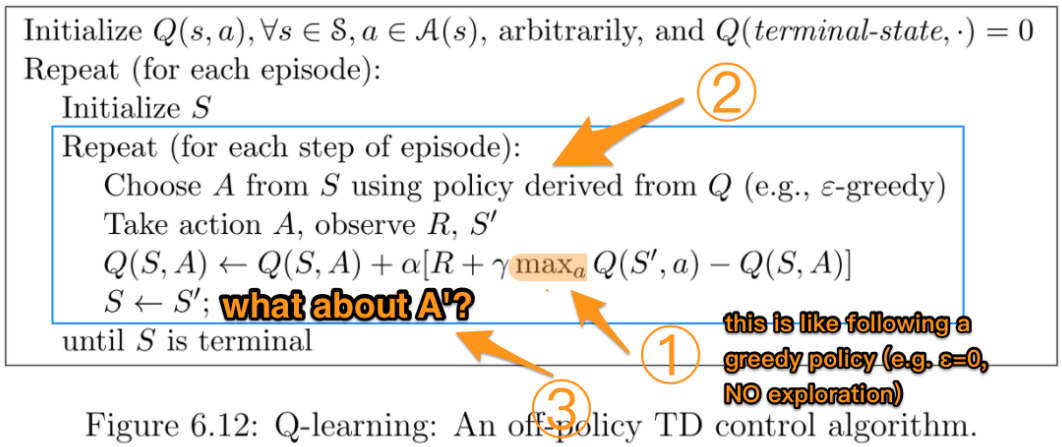


Figure 6.12: Q-learning: An off-policy TD control algorithm.

سوال ۷)

الف)

ویژگی‌هایی که انتخاب می‌شوند باید بتوانند تا حد امکان موقعیت‌های متفاوت را از یکدگیر تمیز دهند. یک حالت از انتخاب ویژگی‌ها می‌تواند به شکل زیر باشد:

1) موقعیت خودرو

2) سرعت خودرو

3) فاصله خودرو از نزدیک‌ترین جسم

4) سرعت نزدیک شدن به نزدیک‌ترین جسم

5) فاصله خودرو از حاشیه جاده

6) میزان ترافیک جاده

7) میزان بارندگی، رطوبت و یخ زدگی جاده

و .....

ب) حتی اگر تمام این ویژگی‌ها در تابع ارزش خطی مورد استفاده قرار بگیرند و ضرایب نیز به بهترین شکل آموخته شوند، هنوز عامل می‌تواند دچار خطاهای فاحش شود. برای مثال عامل نمی‌تواند یک خودرو که چراغ راهنمای خود را روشن کرده و قصد پیچیدن در جاده دارد را از یک خودروی معمولی تشخیص دهد و همین ممکن است باعث رخ دادن حادثه شود.