Year: Month: Date: Subject:

سؤال ۱:

آ)

$$V_k^\pi(s) \leftarrow (1-\alpha)V^\pi(s) + \alpha\left[R(s,\pi(s),s') + \gamma V^\pi(s')\right]$$

V(A) ← ۰/۹(۰) + ۰/۱[-۳+۱(۰)] = -۰/۳

V(B) ← ۰/۹(۰) + ۰/۱[۴+۱(-۰/۳)] = ۰/۳۷

V(A) ← ۰/۹(-۰/۳) + ۰/۱[-۴+۱(-۰/۳)] = -۰/۷

-۰/۴۳

V(A) ← ۰/۹(-۰/۷) + ۰/۱[-۳+۱(۰/۳۷)] = -۰/۸۹۳

V(B) ← ۰/۹(۰/۳۷) + ۰/۱[۱+۱(۰)] = ۰/۴۳۳

ب)

Q₁(A,1) = ۰/۹(Q₀(A,1)) + ۰/۱(-3 + max Q(B)) =

۰/۹(۰) + ۰/۱(-3+۰) = -۰/۳

Q(B,1) = ۰/۹(Q₀(B,1)) + ۰/۱(4+۰) = ۰/4

Q(A,2) = ۰/۹(۰) + ۰/۱(-4+۰) = -۰/4

Q₂(A,1) = ۰/۹(-۰/۳) + ۰/۱(-3+۰/4) = -(۰/۲۷+۰/۲۴) = -۰/۵۳

Q(B,2) = ۰/۹(۰) + ۰/۱(۱+۰) = ۰/۱

ج) با توجه به مقدار Qvalue ها و بیشینه آن برای هر State هر بار در A اکشن ۲ و در B اکشن ۱ را انتخاب می کنیم.

د) استفاده از Qvalue نسبت به حرکت تصادفی می توانند دقیق تر باشد و پس از اعمال این مورد احتمالاً به مقادیر دقیق تری برسیم.

سوال ۲:

آ)

$\gamma = ۰/۷$

Q values = 1

$a = ۰/۵$

$$Q_1(A',\to) = \frac{1}{2} Q_0(A,\to) + \frac{1}{2}\left(2 + \gamma \max_{a'} Q(B,a')\right)$$

$$= \frac{1}{2}(1) + \frac{1}{2}(2 + ۰/۷(-2) = ۰/۵ + ۰/۳ = ۰/۸$$

$$Q_1(C,\leftarrow) = \frac{1}{2} Q_0(C,\leftarrow) + \frac{1}{2}(2 + ۰/۷(-2)) = ۰/۸$$

$$Q_1(B,\to) = \frac{1}{2} Q_0(B,\to) + \frac{1}{2}(-2 + ۰/۷(2)) = ۰/۵ - ۰/۳ = ۰/۲$$

$$Q_2(A,\to) = \frac{1}{2} Q_1(A,\to) + \frac{1}{2}(4 + ۰/۷(۰)) = 2,4$$

$Q(A,\to) = 2,4$

$Q(B,\to) = ۰/۲$

ب و ج) در این مثال با توجه به مقادیر Q value های نهایی از A به راست از B به راست و از C به چپ حرکت خواهیم کرد.

در سیاست حریصانه با توجه به مقدار Q value هایی که به دست می آوریم یک انتخاب حریصانه انجام می دهیم و دیگر جست و جو نمی کنیم. اشکال اینجا است که ممکن است تخمین های ما نامناسب بوده و با پافشاری بر این راه حریصانه نتوانیم بیشترین reward و راه بهینه را بیابیم در حالی که اگر کمی explore می کردیم شاید راه های بهتری پیدا می شد.

می توان از روش ε greedy استفاده کرد که هر بار با احتمال کمی explore و با بقیه آن راه حریصی را انتخاب می کند

د) الگوریتم های TD

۱) در هر دو مدل مسائل مورد استفاده است و به طور پیوسته update انجام می دهد حتی در هر اپیزود. TD(0) تنها به یک قدم جلوتر نگاه می کند برای مسائل episodic خوب است.

سؤال ۳: با استفاده از Q-Learning

آ)

$Q_1(A, Down) = \frac{1}{2}(0) + \frac{1}{2}(2 + \frac{1}{2}(0)) = 1$

$Q_1(B, Down) = \frac{1}{2}(0) + \frac{1}{2}(-4 + \frac{1}{2}(0)) = -2$

$Q_1(B, up) = \frac{1}{2}(0) + \frac{1}{2}(0 + \frac{1}{2}(0)) = 0$

$Q_2(B, up) = \frac{1}{2}(0) + \frac{1}{2}(3 + \frac{1}{2}(1)) = \frac{3}{2} + \frac{1}{4} = \frac{6+1}{4} = \frac{7}{4}$

$Q_1(A, Down) = 1$

$Q_1(B, up) = \frac{۷}{۴}$

ب)

$\hat{T}(A, Up, A) = 1/1$ $\hat{T}(A, Up, B) = 0$ $\hat{T}(B, Up, A) = 1/2$

$\hat{T}(B, Up, A) = 1/2$ $\hat{R}(A, Up, A) = -1$ $\hat{R}(A, Up, B) = n/a$

$\hat{R}(B, Up, A) = 3$ $\hat{R}(B, Up, B) = 0$

با توجه به تعداد تکرار هر کدام از شرایط $\hat{T}$ ها به دست آمده جوایز نیز از همان جدول به دست آمده .

سؤال ۴) $R=-1$

$$V^{\pi}(15) = E_{\pi}\{r_{t+1} + V^{\pi}(s_{t+1}) \mid s_t = s\}$$

$$= -1 + \frac{1}{4}V^{\pi}(12) + \frac{1}{4}V^{\pi}(13) + \frac{1}{4}V^{\pi}(14) + \frac{1}{4}V^{\pi}(15)$$

$$= -1 - \frac{1}{4}\times 22 - \frac{1}{4}\times 20 - \frac{1}{4}\times 14 - \frac{1}{4}V^{\pi}(15)$$

$$\frac{3}{4}V^{\pi}(15) = -15 \implies V^{\pi}(15) = -20$$

سؤال ۵ :

1) exploration یک راه سودمند در بلندمدت است که به agent اجازه می دهد که دانشش را در مورد action هایی که در طولانی مدت به سود می انجامد بیشتر کند.

در مقابل exploitation از ارزش تخمینی فعلی Agent استفاده کرده و راه حل حریصانه را انتخاب می کند تا به بیش ترین reward دست پیدا کند. از آنجایی که Agent تخمین های حریصانه می زند در واقع ممکن بیش ترین reward و بهترین راه نباشد.

ب) با کاهش دادن مقدار ε در ε-greedy الگوریتم و میل دادن آن به صفر به نوعی هر بار احتمال explore کردن را کم می کنیم و به الگوریتم greedy نزدیک تر می شویم که خیلی ایده ی مناسبی نیست.

د) تابع V value برای اندازه‌گیری میزان خوب بودن یک state خاص از نظر پاداش کلی سود انتظار برای
agentها است که طبق یک policy عمل می‌کند استفاده می‌شود و V^* حداکثر مقدار قابل دسترس
را برای هر حالت در فضای حالت داده شده به ما می‌دهد.

درحالیکه Q value به ما نشان می‌دهد که یک عمل خاص با توجه به وضعیت برای عاملی که از یک
policy پیروی می‌کند چقدر خوب است. تابع Q^* حداکثر بازده قابل دستیابی از یک جفت حالت
عمل معین توسط خط‌مشی به ما می‌دهد.

Q-value ها برای مسائلی که transition function در دسترس نیستند (مدل فری) ابزار جالبی که action_space زیادی
داریم اما Q ها خیلی راحت نیستند برای استفاده.
اگر transition function دسترسی داریم گاهی اوقات V مناسب است.