2			+5
1	S		-5
	1	2	3

$$V_{i+1}(S)=\max_{a}\left[\sum_{S'}T(S,a,S')\left(R(S,a,S')+\gamma V_i(S')\right)\right]$$ .1 (سوال اول

MDP

$\gamma=0.9$

$$\boxed{\forall S_{ij}: V_0(S_{ij})=0}$$

احتمال حرکت مورد نظر = 0.8

احتمال حرکت عمود = 0.1

$V_1(S_{11})=$

$a=\text{up}: 0.8\times\overset{\text{up}}{(0+0.9\times0)}+0.1\times\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{right}}{(0+0.9\times0)}$

$a=\text{down}: 0.8\times\overset{\text{down}}{(0+0.9\times0)}+0.1\times\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{right}}{(0+0.9\times0)}=0$

$a=\text{left}: 0$

$a=\text{right}: 0$

$\Rightarrow \boxed{V_1(S_{11})=0}$

$V_1(S_{12})=$

$a=\text{up}: 0.8\overset{\text{up}}{(0+0.9\times0)}+0.1\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{right}}{(-5+0.9(-5))}=-0.95$

$a=\text{down}: 0.8\overset{\text{down}}{(0+0.9\times0)}+0.1\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{right}}{(-5+0.9\times(-5))}=-0.95$

$a=\text{left}: 0.8\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{up}}{(0+0.9\times0)}+0.1\overset{\text{down}}{(0+0.9\times0)}=0$

$a=\text{right}: 0.8\underset{\text{right}}{(-5+0.9\times(-5))}+0.1\overset{\text{up}}{(0+0.9\times0)}+0.1\overset{\text{down}}{(0+0.9\times0)}=-7.6$

$\Rightarrow \boxed{V_1(S_{12})=0}$

$\boxed{V_1(S_{13})=-5}$

مثل $V_1(S_{11})$

$\boxed{V_1(S_{21})=0}$

$V_1(S_{22})=$

$a=\text{up}: 0.8\overset{\text{up}}{(0+0.9\times0)}+0.1\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{right}}{(+5+0.9(5))}=+0.95$

$a=\text{down}: 0.8\overset{\text{down}}{(0+0.9\times0)}+0.1\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{right}}{(5+0.9(5))}=+0.95$

$a=\text{left}: 0.8\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{up}}{(0+0.9\times0)}+0.1\overset{\text{down}}{(0+0.9\times0)}=0$

$a=\text{right}: 0.8\overset{\text{right}}{(5+5\times0.9)}+0.1\overset{\text{up}}{(0+0\times0.9)}+0.1\overset{\text{down}}{(0+0.9\times0)}=+7.6$

$\boxed{V_1(S_{22})=+7.6}$

$\boxed{V_1(S_{23})=+5}$

$V_2(S_{11})$

$a=\text{up}: 0.8\overset{\text{up}}{(0+0.9\times0)}+0.1\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{right}}{(0+0.9\times0)}=0$

$a=\text{down}: 0.8\overset{\text{down}}{(0+0.9\times0)}+0.1\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{right}}{(0+0.9\times0)}=0$

$a=\text{left}: 0.8\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{up}}{(0+0.9\times0)}+0.1\overset{\text{down}}{(0+0.9\times0)}=0$

$a=\text{right}: 0.8\overset{\text{right}}{(0+0.9\times0)}+0.1\overset{\text{down}}{(0+0.9\times0)}+0.1\overset{\text{up}}{(0+0.9\times0)}=0$

$\boxed{V_2(S_{11})=0}$

$V_2(S_{12})$

$a=\text{up}: 0.8\overset{\text{up}}{(0+0.9\times7.6)}+0.1\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{right}}{(-5+0.9(-5))}=+4.522$

$a=\text{down}: 0.8\overset{\text{down}}{(0+0.9\times0)}+0.1\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{right}}{(-5+0.8(-5))}=-0.95$

$a=\text{left}: 0.8\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{down}}{(0+0.9\times0)}+0.1\overset{\text{up}}{(0+0.9\times7.6)}=+0.684$

$a=\text{right}: 0.8\underset{\text{right}}{(-5+(-5)0.9)}+0.1\overset{\text{down}}{(0+0.9\times0)}+0.1\overset{\text{up}}{(0+0.9\times5)}=-3.62$

$\Rightarrow \boxed{V_2(S_{12})=+4.522}$

$\boxed{V_2(S_{13})=-5}$

$V_2(S_{21})$

$a=\text{up}: 0.8\overset{\text{up}}{(0+0.9\times0)}+0.1\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{right}}{(0+0.9\times7.6)}=+0.684$

$a=\text{down}: 0.8\overset{\text{down}}{(0+0.9\times0)}+0.1\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{right}}{(0+0.9\times7.6)}=+0.684$

$a=\text{left}: 0.8\overset{\text{left}}{(0+0.9\times0)}+0.1\overset{\text{up}}{(0+0.9\times0)}+0.1\overset{\text{down}}{(0+0.9\times0)}=0$

$a=\text{right}: 0.8\times(0+0.9\times7.6)+0.1\overset{\text{up}}{(0+0.9\times0)}+0.1\overset{\text{down}}{(0+0.9\times0)}=+5.472$

$\Rightarrow \boxed{V_2(S_{21})=5.472}$

$V_۲(S_{۲۲}) =$

a=up : ۰,۸(۰+۰,۹×۷,۶) + ۰,۱×(۰+۰,۹×۰) + ۰,۱(+۵+۰,۹×(۵)) = +۶,۴۴۲

a=down : ۰,۸(۰+۰,۹×۰) + ۰,۱×(۰+۰,۹×۰) + ۰,۱(+۵+۰,۹(۵)) = +۰,۹۵

a=left : ۰,۸(۰+۰,۹×۰) + ۰,۱×(۰+۰,۹×۷,۶) + ۰,۱(۰+۰,۹×۰) = +۰,۶۸۴

a=right : ۰,۸(+۵+۰,۹×۵) + ۰,۱(۰+۰,۹×۷,۶) + ۰,۱(۰+۰,۹×۰) = +۸,۲۸۴

⟹ $\boxed{V_۲(S_{۲۲}) = ۸,۲۸۴}$

$\boxed{V_۲(S_{۲۳}) = +۵}$

S	(۱,۱)	(۱,۲)	(۱,۳)	(۲,۱)	(۲,۲)	(۲,۳)
$V_۰$	۰	۰	-۵	۰	۰	+۵
$V_۱$	۰	۰	-۵	۰	۷,۶	+۵
$V_۲$	۰	۴,۵۲۲	-۵	۵,۴۷۲	۸,۲۸۴	+۵

.2

$$\pi^*(s) = \arg\max_a Q^*(s,a)$$

$$Q^*(s,a) = \sum_{s'} T(s,a,s')\{R(s,a,s') + \gamma V^*(s')\}$$

$\pi^*(S_{۱۱}) =$

a=up : $\overbrace{۰,۸(۰+۰,۹×۵,۴۷۲)}^{۳,۹۳۹}$ + ۰,۱×(۰+۰,۹×۰) + $\overbrace{۰,۱×(۰+۰,۹×۴,۵۲۲)}^{۰,۴۰۶۹}$ = ۴,۳۴

a=down : ۰,۸(۰+۰,۹×۰) + ۰,۱×(۰+۰,۹×۰) + ۰,۱(۰+۰,۹×۴,۵۲۲) = ۰,۴۰۶۹

a=left : ۰,۸(۰+۰,۹×۰) + ۰,۱×(۰+۰,۹×۰) + ۰,۱(۰+۰,۹×۵,۴۷۲) = ۰,۴۹۲۴

a=right : ۰,۸(۰+۰,۹×۴,۵۲۲) + ۰,۱(۰+۰,۹×۰) + ۰,۱(۰+۰,۹×۵,۴۷۲) = ۳,۷۴۹

$\boxed{\text{up}}$

$\pi^*(S_{۱۲}) =$

a=up : ۰,۸(۰+۰,۹×۸,۲۸۴) + ۰,۱(۰+۰,۹×۰) + ۰,۱(-۵+۰,۹×(-۵)) = ۵,۰۱

a=down : ۰,۸(۰+۰,۹×۴,۵۲۲) + ۰,۱(۰+۰,۹×۰) + ۰,۱(۰+۰,۹×(-۵)) = —

a=left : ۰,۸(۰+۰,۹×۰) + ۰,۱(۰+۰,۹×۸,۲۸۴) + ۰,۱×(۰+۰,۹×۴,۵۲۲) = —

a=right : ۰,۸(-۵+۰,۹(-۵)) + ۰,۱(۰+۰,۹×۴,۵۲۲) + ۰,۱(۰+۰,۹×۸,۲۸۴) = —

$\boxed{\text{up}}$

$\pi^*(S_{۲۱}) \xRightarrow{\text{محاسبات را انجام نمی‌دهم}} \boxed{\text{right}}$

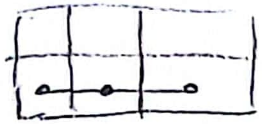
$\pi^*(S_{۲۲}) \Longrightarrow \boxed{\text{right}}$

S	(۱,۱)	(۱,۲)	(۱,۳)	(۲,۱)	(۲,۲)	(۲,۳)
$\pi^*(S)$	↑	↑	—	→	→	—

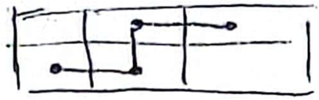
3

$$V^{\pi}(s) = E\left\{\sum_i \gamma^i R_i\right\}$$

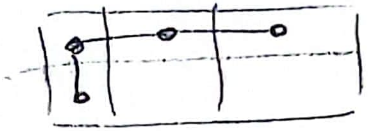
I) $(1,1) \xrightarrow{0} (1,2) \xrightarrow{-5} (1,3)$



II) $(1,1) \xrightarrow{0} (1,2) \xrightarrow{0} (2,2) \xrightarrow{5} (2,3)$



III) $(1,1) \xrightarrow{0} (2,1) \xrightarrow{0} (2,2) \xrightarrow{5} (2,3)$



$$V^{\pi}(S_{11}) = \frac{(0+(-5)\times 0.9) + (0+0\times 0.9+0.9^2(+5)) + (0+0\times 0.9+5\times(0.9)^2)}{3} = 1.2$$

$$V^{\pi}(S_{22}) = \frac{5+5}{2} = 5$$

$$\text{Sample} = R(S, \pi(S), S') + \gamma V^{\pi}(S')$$

$$V^{\pi}(s) = (1-\alpha)V^{\pi}(s) + (\alpha)[\text{Sample}]$$

$\alpha = 0.1$

$\gamma = 0.9$

π-Random

0	0	+5
0	-0.95	5

سوال ۳ (امتیازی)

$$V^{\pi}(S_{11}) = (1-0.1)\times 0 + 0.1[0+0.9\times 0] = 0$$

$$V^{\pi}(S_{12}) = (1-0.1)\times 0 + 0.1[-5+0.9\times -5] = -0.95$$

DQN : DQN یک الگوریتم قوی در زمینه یادگیری تقویتی است. این الگوریتم، اصول شبکه‌های عصبی عمیق را با Q-learning ترکیب میکند و agent ها را قادر می‌سازد تا optimal policy ها را در شرایط و محیط‌های پیچیده یاد بگیرند.

این الگوریتم از یک رویکرد مبتنی بر شبکه‌های عصبی عمیق برای یادگیری و بهینه‌سازی توابع action value پیروی می‌کند. رویکردها را می‌توان به این شکل خلاصه کرد:

1) State representation : وضعیت فعلی محیط را به یک نمایش عددی مناسب، مانند مقادیر خام پیکسل‌ها و ویژگی‌های از پیش پردازش شده، تبدیل میکنند.

2) Neural network architecture : طراحی یک شبکه عصبی، معمولا یک شبکه عصبی کانولوشنی، که state را به‌عنوان ورودی بگیرد و برای هر action یک بردار action-values خروجی دهد.

3) Experience Replay : تجربیات agent شامل state ها، action ها و reward ها، next state ها را در replay memory ذخیره میکنند.

4) Q-learning updates : mini-batches از تجربیات را از بخش replay memory buffer نمونه میگیرد تا برای آپدیت کردن وزن‌های شبکه‌های عصبی استفاده کند. این آپدیت‌ها به کمک توابع LOSS که از معادلات Bellman بدست آمده‌اند انجام می‌شود که اختلافات بین مقادیر پیش‌بینی شده (predicted action value) و هدف را به حداقل برساند.

5) Exploration and Exploitation : بالانس کردن بین exploration و exploitation به کمک انتخاب action ها که براساس policy یا رندوم انجام می‌شود.

6) Target network : استفاده از یک شبکه هدف جداگانه مشابه شبکه اصلی برای ثبات فرایند یادگیری. به‌طور دوره‌ای با کپی کردن وزن‌ها از شبکه اصلی، شبکه هدف را آپدیت می‌کنیم.

7) Repeat steps 1 to 6 : با محیط تعامل کنیم، تجربیات جمع‌آوری کنیم، شبکه را بروز کنیم و policy را بهبود کنیم تا به همگرایی برسیم.

مفاهیم اصلی:

Q-learning : از الگوریتم Q-learning برای DQN استفاده می‌شود که هدف آن تخمین تابع ارزش action value (Q-function) است که state ها را به پاداش‌های مورد انتظار شده map میکند.

Experience Replay : این قسمت کمک میکند تا با ذخیره کردن آن‌ها در replay memory buffer ارتباط بین تجربیات متوالی کمک کند. این memory buffer نمونه تصادفی در طول بروزرسانی شبکه برای شکستن وابستگی‌های زمانی و ثبات یادگیری نمونه‌برداری می‌شود.