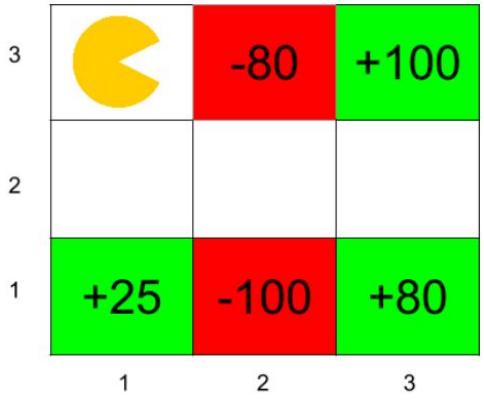
Aufgabe

- Reinforcement Learning -

Betrachten Sie Pacman, der gerade versucht, die optimale Strategie π^* im dargestellten $3x3$-Gitter zu erlernen. Wenn eine Aktion dazu führt, dass Pacman auf einen der schattierten Flächen landet, erhält er die angezeigte Belohnung. Alle schattierten Zustände sind absorbierend, d.h. der Markov-Entscheidungsprozess terminiert, sobald Pacman sie erreicht. Für die anderen Zustände sind die Aktionen North (N), East (E), South (S) oder West (W) verfügbar, die Pacman deterministisch in den entsprechenden Nachbarzustand bewegen. Nehmen Sie einen Diskontierungsfaktor $\gamma = 0.5$ an. Pacman startet in Zustand (1,3).



Pacman-Welt

a) Berechnen Sie den Wert der optimalen Wertefunktion V^{π^*} für folgende Zustände:

$V^{\pi^*}(3,2) =$ _____ $V^{\pi^*}(2,2) =$ _____ $V^{\pi^*}(1,3) =$ _____

Hinweis: Verwenden Sie das Optimalitätskriterium $V^{\pi^*}(s) \geq V^{\pi}(s)$ aus der Vorlesung.

Lösung

$V^{\pi}(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots$ (abgeschwächte Belohnung)

$V^{\pi^*}(s) \geq V^{\pi}(s)$ (Die optimale Strategie wird jeweils dem Schaubild entnommen)

$V^{\pi^*}(3,2) = 100$ (Der Zustand (3,3) ist absorbierend, d.h. es gibt keinen Nachfolgezustand)

$$\begin{aligned} V^{\pi^*}(2,2) &= 0 + \gamma * 100 \\ &= 50 \\ V^{\pi^*}(1,3) &= 0 + \gamma * 25 \\ &= 0 + \gamma * 0 + \gamma^2 * 0 + \gamma^3 * 100 \\ &= 12.5 \quad \text{(beide Strategien sind optimal)} \end{aligned}$$

b) Der Agent startet von der oberen linken Ecke und durchläuft folgende Episoden der Pacman-Welt. Jeder Eintrag einer Episode stellt ein Tuple mit (s, a, s', r) dar.

Episode 1	Episode 2	Episode 3
(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0
(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0
(2,2), S, (2,1), -100	(2,2), E, (3,2), 0	(2,2), E, (3,2), 0
	(3,2), N, (3,3), +100	(3,2), S, (3,1), +80

Berechnen Sie die Q-Werte mit der Iterationsvorschrift für Q-Lernen aus der Vorlesung:

$Q((3,2),N) =$ _____ $Q((1,2),N) =$ _____ $Q((2,2),N) =$ _____

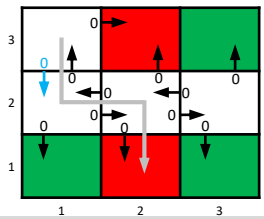
Lösung

$Q((3,2),N) =$ 100 $Q((1,2),N) =$ 0 $Q((2,2),N) =$ 0

Episode 1

- Pacman durchläuft die erste Episode und aktualisiert nach jeder Aktion den entsprechenden Wert der ursprünglich mit 0-en initialisierten Q-Tabelle
- Für jede Aktion erhält Pacman ein Feedback in Form einer Belohnung.
- Die schattierten Zustände sind absorbierend, daher sind die jeweiligen Pfeile nur in eine Richtung eingezeichnet (im Gegensatz zum Beispiel des Krabbelroboters aus der Vorlesung, der jeden Zustand beliebig oft hintereinander einnehmen kann).

Schritt E1.1 Übergang vom Zustand $s = (1,3)$ in den Zustand $\delta((1,3),S) = (1,2)$ durch die Aktion $S = \text{South}$:



Mit der Iterationsvorschrift aus der Vorlesung erhalten wir:

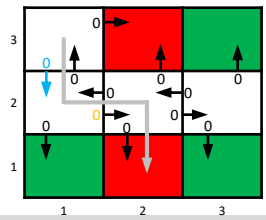
$$\hat{Q}(s,a) = r(s,a) + \gamma \max_{a'} \hat{Q}(\delta(s,a),a')$$

$$\hat{Q}((1,3),S) = r((1,3),S) + 0{,}5 \max_{a'} \hat{Q}(\delta((1,3),S),a')$$
$$= 0 + 0{,}5 \max_{a'} \hat{Q}((1,2),a')$$
$$= 0{,}5\ max[\hat{Q}((1,2),S), \hat{Q}((1,2),E), \hat{Q}((1,2),N)]$$
$$= 0{,}5 \max[0,0,0]$$

Erläuterung
$(\hat{Q}((1,2),S), \hat{Q}((1,2),E), \hat{Q}((1,2),N) = 0$ folgt durch Ablesen des $\hat{Q}$-Werts aus der bisher gelernten $\hat{Q}$-Tabelle)

Ergebnis für das Durchlaufen der Iterationsvorschrift
Der initiale $\hat{Q}$-Wert 0 der Q-Tabelle erhält das Update auf denselben Wert 0.

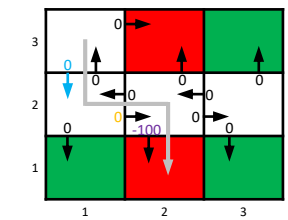
Schritt E1.2 Übergang vom Zustand $s = (1,2)$ in den Zustand $\delta((1,2),E) = (2,2)$ durch die Aktion $E = \text{East}$:



$$\hat{Q}((1,2),E) = r((1,2),E) + 0{,}5 \max_{a'} \hat{Q}(\delta((1,2),E),a')$$
$$= 0 + 0{,}5 \max_{a'} \hat{Q}((2,2),a')$$
$$= 0{,}5\ max[\hat{Q}((2,2),S), \hat{Q}((2,2),E), \hat{Q}((2,2),N), \hat{Q}((2,2),W)]$$
$$= 0{,}5 \max[0,0,0,0]$$
$$= 0$$

Ergebnis für das Durchlaufen der Iterationsvorschrift
Der initiale $\hat{Q}$-Wert 0 der Q-Tabelle wird wieder auf denselben Wert 0 geupdated.

Schritt E1.3 Übergang vom Zustand $s = (2,2)$ in den Zustand $\delta((2,2),S) = (2,1)$ durch die Aktion $S = \text{South}$:



$$\hat{Q}((2,2),S) = r((2,2),S) + 0{,}5 \max_{a'} \hat{Q}(\delta((2,2),S),a')$$
$$= -100 + 0{,}5 \max_{a'} \hat{Q}((2,1),a')$$
$$= -100 + 0{,}5 * 0$$
$$= -100$$

Die schattierten Zustände sind absorbierend:

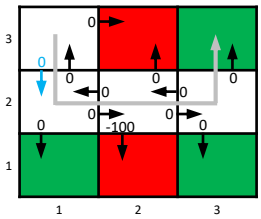
$$\forall a' \colon \hat{Q}((2,1),a') = 0$$

Ergebnis für das Durchlaufen der Iterationsvorschrift
Der initiale $\hat{Q}$-Wert 0 der Q-Tabelle wird auf den Wert -100 geupdated.

Episode 2

Episode 2
Pacman durchläuft die zweite Episode und aktualisiert dabei die bisher gelernten Werte der Q-Tabelle:

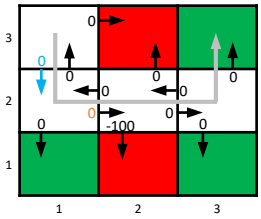
Schritt E2.1 analog zu Schritt E1.1



$$\hat{Q}((1,3),S) = 0$$

Ergebnis für das Durchlaufen der Iterationsvorschrift
Der $\hat{Q}$-Wert 0 aus der Q-Tabelle erhält das Update auf denselben Wert 0.

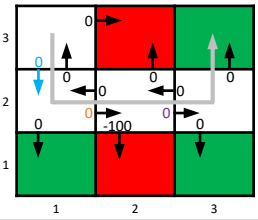
Schritt E2.2 Übergang vom Zustand $s = (1,2)$ in den Zustand $\delta((1,2),E) = (2,2)$ durch die Aktion $E = \text{East}$:



$$\hat{Q}((1,2),E) = r((1,2),E) + 0{,}5 \max_{a'} \hat{Q}(\delta((1,2),E),a')$$
$$= 0 + 0{,}5 \max_{a'} \hat{Q}((2,2),a')$$
$$= 0{,}5\ max[\hat{Q}((2,2),S), \hat{Q}((2,2),E), \hat{Q}((2,2),N), \hat{Q}((2,2),W)]$$
$$= 0{,}5 \max[-100,0,0,0]$$
$$= 0$$

Ergebnis für das Durchlaufen der Iterationsvorschrift
Der $\hat{Q}$-Wert 0 aus der Q-Tabelle erhält das Update auf denselben Wert 0.

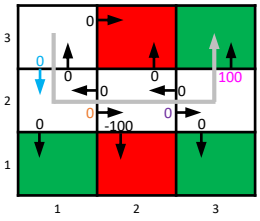
Schritt E2.3 Übergang vom Zustand $s = (2,2)$ in den Zustand $\delta((2,2),E) = (3,2)$ durch die Aktion $E = \text{East}$:



$$\hat{Q}((2,2),E) = r((2,2),E) + 0{,}5 \max_{a'} \hat{Q}(\delta((2,2),E),a')$$
$$= 0 + 0{,}5 \max_{a'} \hat{Q}((3,2),a')$$
$$= 0{,}5\ max[\hat{Q}((3,2),S), \hat{Q}((3,2),W), \hat{Q}((3,2),N)]$$
$$= 0{,}5\ max[0,0,0]$$
$$= 0$$

Ergebnis für das Durchlaufen der Iterationsvorschrift
Der $\hat{Q}$-Wert 0 aus der Q-Tabelle erhält das Update auf denselben Wert 0.

Schritt E2.4 Übergang vom Zustand $s = (3,2)$ in den Zustand $\delta((3,2),N) = (3,3)$ durch die Aktion $N = \text{North}$:



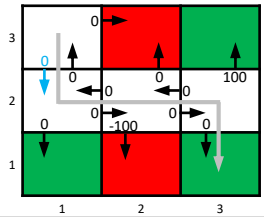
$$\hat{Q}((3,2),N) = r((3,2),N) + 0{,}5 \max_{a'} \hat{Q}(\delta((3,2),N),a')$$
$$= 100 + 0{,}5 \max_{a'} \hat{Q}((3,3),a')$$
$$= 100 + 0$$
$$= 100$$

Ergebnis für das Durchlaufen der Iterationsvorschrift
Der $\hat{Q}$-Wert 0 aus der Q-Tabelle erhält das Update auf den Wert 100.

Episode 3

Episode 3
Pacman durchläuft die dritte Episode und aktualisiert dabei die bisher gelernten Werte der Q-Tabelle:

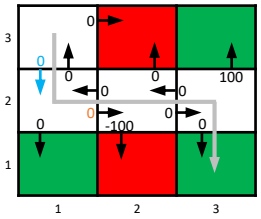
Schritt E3.1 analog zu Schritt E1.1



$$\hat{Q}((1,3),S) = 0$$

Ergebnis für das Durchlaufen der Iterationsvorschrift
Der $\hat{Q}$-Wert 0 aus der Q-Tabelle erhält das Update auf denselben Wert 0.

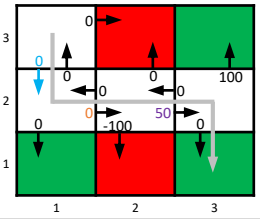
Schritt E3.2 analog zu Schritt E2.2



$$\hat{Q}((1,2),E) = 0$$

Ergebnis für das Durchlaufen der Iterationsvorschrift
Der $\hat{Q}$-Wert 0 aus der Q-Tabelle erhält das Update auf denselben Wert 0.

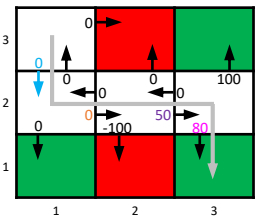
Schritt E3.3 Übergang vom Zustand $s = (2,2)$ in den Zustand $\delta((2,2),E) = (3,2)$ durch die Aktion $E = \text{East}$:



$$\hat{Q}((2,2),E) = r((2,2),E) + 0{,}5 \max_{a'} \hat{Q}(\delta((2,2),E),a')$$
$$= 0 + 0{,}5 \max_{a'} \hat{Q}((3,2),a')$$
$$= 0{,}5\ max[\hat{Q}((3,2),S), \hat{Q}((3,2),W), \hat{Q}((3,2),N)]$$
$$= 0{,}5\ max[0,0,100]$$
$$= 50$$

Ergebnis für das Durchlaufen der Iterationsvorschrift
Der $\hat{Q}$-Wert 0 aus der Q-Tabelle erhält das Update auf den Wert 50.

Schritt E3.4 Übergang vom Zustand $s = (3,2)$ in den Zustand $\delta((3,2),S) = (3,1)$ durch die Aktion $S = \text{South}$:



$$\hat{Q}((3,2),S) = r((3,2),S) + 0{,}5 \max_{a'} \hat{Q}(\delta((3,2),S),a')$$
$$= 80 + 0{,}5 \max_{a'} \hat{Q}((3,1),a')$$
$$= 80 + 0$$
$$= 80$$

Ergebnis für das Durchlaufen der Iterationsvorschrift
Der $\hat{Q}$-Wert 0 aus der Q-Tabelle erhält das Update auf den Wert 80.