

# Q-l ering

## INF100

**Odin Hoff Gard **

UNIVERSITY OF BERGEN



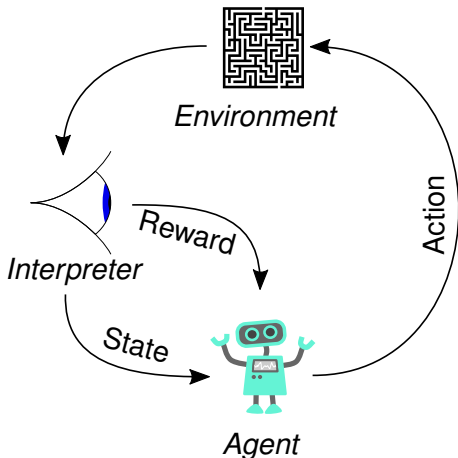
# Plan

- 1 Kort introduksjon til Q-læring.
- 2 Workshop: Implementer Q-læring for å løse en labyrint.



# Forsterkende L ring<sup>1</sup>

- Vi har en **agent** som **handler** i et **milj **.
- Agenten lærer gjennom **bel nning** basert p  **tilstand** og **handling**.
- **Q-l ring** er en form for forsterkende l ring.



<sup>1</sup>Engelsk: Reinforcement Learning

# Q-l ring: Oppsett

- Vi starter med:
  - En mengde  $\mathcal{S}$  av mulige **tilstander**
  - En mengde  $\mathcal{A}$  av mulige **handlinger**
  - Et par  $(s, a) \in \mathcal{S} \times \mathcal{A}$  kalles et **tilstand-handlings-par**
  - En **bel nningsfunksjon**  $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$



# Q-l ring: Oppsett

- Vi starter med:
  - En mengde  $\mathcal{S}$  av mulige **tilstander**
  - En mengde  $\mathcal{A}$  av mulige **handlinger**
  - Et par  $(s, a) \in \mathcal{S} \times \mathcal{A}$  kalles et **tilstand-handlings-par**
  - En **bel nningsfunksjon**  $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$
- Ved   la agenten utforske milj et  nsker vi   l re **Q-funksjonen**

$$Q: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$$

som gir oss en **Q-verdi**  $Q(s, a)$  til hvert par  $(s, a) \in \mathcal{S} \times \mathcal{A}$ .



# Q-læring: Oppsett

- Vi starter med:
  - En mengde  $\mathcal{S}$  av mulige **tilstander**
  - En mengde  $\mathcal{A}$  av mulige **handlinger**
  - Et par  $(s, a) \in \mathcal{S} \times \mathcal{A}$  kalles et **tilstand-handlings-par**
  - En **belønningsfunksjon**  $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$
- Ved å la agenten utforske miljøet ønsker vi å lære **Q-funksjonen**

$$Q: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$$

som gir oss en **Q-verdi**  $Q(s, a)$  til hvert par  $(s, a) \in \mathcal{S} \times \mathcal{A}$ .

- **Endelig mål:** For en  $s \in \mathcal{S}$  så ønsker vi at  $\arg \max_{a \in \mathcal{A}} Q(s, a)$  er den optimale handlingen for å maksimere forventet belønning.



# Gjennomgående Eksempel: $3 \times 3$ Labyrint

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)



Åpen rute



Vegg



Agent



Mål



# Gjennomgående Eksempel: $3 \times 3$ Labyrint

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)



Åpen rute



Vegg



Agent



Mål

**Læringsmål for agenten:** Fra en vilkårlig åpen rute, nå mål ved å bruke så få steg som mulig.





# Gjennomgående Eksempel: $3 \times 3$ Labyrint

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)



Åpen rute



Vegg



Agent



Mål

Mulige tilstander (agentens posisjon):

$$\mathcal{S} = \{(0,0), (1,0), (2,0), (0,1), (1,1), (2,1), (0,2), (1,2), (2,2)\}$$



# Gjennomgående Eksempel: $3 \times 3$ Labyrint

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)

	Åpen rute
	Vegg
	Agent
	Mål

Mulige tilstander (agentens posisjon):

$$\mathcal{S} = \{(0,0), (1,0), (2,0), (0,1), (1,1), (2,1), (0,2), (1,2), (2,2)\}$$

Mulige handlinger (retninger å gå):

$$\mathcal{A} = \{\text{venstre, høyre, opp, ned}\}$$



# Gjennomgående Eksempel: $3 \times 3$ Labyrint

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)

	Åpen rute
	Vegg
	Agent
	Mål

Noen eksempler på par  $(s, a) \in \mathcal{S} \times \mathcal{A}$ :

- $((2, 1), \text{opp})$
- $((2, 1), \text{høyre})$
- $((0, 0), \text{ned})$
- $((1, 1), \text{venstre})$
- $((2, 0), \text{opp})$
- $((1, 1), \text{ned})$



# Eksempel: Belønningsfunksjonen

La  $s'$  være ruten i retning  $a$  fra posisjon  $s$ .

Definer belønningsfunksjonen  $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  ved

$$R(s, a) = \begin{cases} -1.0 & \text{hvis } s' \text{ er en veggrute,} \\ -0.1 & \text{hvis } s' \text{ er en åpen rute og} \\ 1.0 & \text{hvis } s' \text{ er målruten.} \end{cases}$$

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)

■ **Q:** Hva er  $R((0, 1), \text{høyre})$ ?



# Eksempel: Belønningsfunksjonen

La  $s'$  være ruten i retning  $a$  fra posisjon  $s$ .

Definer belønningsfunksjonen  $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  ved

$$R(s, a) = \begin{cases} -1.0 & \text{hvis } s' \text{ er en veggrute,} \\ -0.1 & \text{hvis } s' \text{ er en åpen rute og} \\ 1.0 & \text{hvis } s' \text{ er målruten.} \end{cases}$$

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)

- **Q:** Hva er  $R((0, 1), \text{høyre})$ ?
- **A:**  $R((0, 1), \text{høyre}) = 1.0$



# Eksempel: Belønningsfunksjonen

La  $s'$  være ruten i retning  $a$  fra posisjon  $s$ .

Definer belønningsfunksjonen  $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  ved

$$R(s, a) = \begin{cases} -1.0 & \text{hvis } s' \text{ er en veggrute,} \\ -0.1 & \text{hvis } s' \text{ er en åpen rute og} \\ 1.0 & \text{hvis } s' \text{ er målruten.} \end{cases}$$

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)

- **Q:** Hva er  $R((0, 1), \text{høyre})$ ?
- **A:**  $R((0, 1), \text{høyre}) = 1.0$
- **Q:** Hva er  $R((1, 1), \text{venstre})$ ?



# Eksempel: Belønningsfunksjonen

La  $s'$  være ruten i retning  $a$  fra posisjon  $s$ .

Definer belønningsfunksjonen  $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  ved

$$R(s, a) = \begin{cases} -1.0 & \text{hvis } s' \text{ er en veggrute,} \\ -0.1 & \text{hvis } s' \text{ er en åpen rute og} \\ 1.0 & \text{hvis } s' \text{ er målruten.} \end{cases}$$

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)

- **Q:** Hva er  $R((0, 1), \text{høyre})$ ?
- **A:**  $R((0, 1), \text{høyre}) = 1.0$
- **Q:** Hva er  $R((1, 1), \text{venstre})$ ?
- **A:**  $R((1, 1), \text{venstre}) = -1.0$



# Eksempel: Belønningsfunksjonen

La  $s'$  være ruten i retning  $a$  fra posisjon  $s$ .

Definer belønningsfunksjonen  $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  ved

$$R(s, a) = \begin{cases} -1.0 & \text{hvis } s' \text{ er en veggrute,} \\ -0.1 & \text{hvis } s' \text{ er en åpen rute og} \\ 1.0 & \text{hvis } s' \text{ er målruten.} \end{cases}$$

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)

- **Q:** Hva er  $R((0, 1), \text{høyre})$ ?
- **A:**  $R((0, 1), \text{høyre}) = 1.0$
- **Q:** Hva er  $R((1, 1), \text{venstre})$ ?
- **A:**  $R((1, 1), \text{venstre}) = -1.0$
- **Q:** Hva er  $R((1, 1), \text{opp})$ ?





# Eksempel: Belønningsfunksjonen

La  $s'$  være ruten i retning  $a$  fra posisjon  $s$ .

Definer belønningsfunksjonen  $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  ved

$$R(s, a) = \begin{cases} -1.0 & \text{hvis } s' \text{ er en veggrute,} \\ -0.1 & \text{hvis } s' \text{ er en åpen rute og} \\ 1.0 & \text{hvis } s' \text{ er målruten.} \end{cases}$$

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)

- **Q:** Hva er  $R((0, 1), \text{høyre})$ ?
- **A:**  $R((0, 1), \text{høyre}) = 1.0$
- **Q:** Hva er  $R((1, 1), \text{venstre})$ ?
- **A:**  $R((1, 1), \text{venstre}) = -1.0$
- **Q:** Hva er  $R((1, 1), \text{opp})$ ?
- **A:**  $R((1, 1), \text{opp}) = -0.1$



# Q-tabell

Hvis vi har endelig mange tilstander og handlinger, kan vi representere Q-funksjonen  $Q: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  som en tabell (**Q-tabell**):

$s \backslash a$	venstre	høyre	opp	ned
(0,0)	1.0	-0.5	0.1	-0.3
(1,0)	0.2	-0.3	-1.0	0.6
(2,0)	-0.4	-0.1	0.3	0.7
(0,1)	0.4	-0.9	1.0	0.2
(1,1)	0.6	-0.1	0.4	0.8
(2,1)	1.0	0.5	0.8	-0.5
(0,2)	-0.2	0.5	-0.3	-0.7
(1,2)	0.7	-1.0	0.1	-0.5
(2,2)	0.1	0.0	0.8	0.1

**Spørsmål:**

- **Q:** Hva er  $Q((0,0), \text{høyre})$ ?



# Q-tabell

Hvis vi har endelig mange tilstander og handlinger, kan vi representere Q-funksjonen  $Q: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  som en tabell (**Q-tabell**):

$s \backslash a$	venstre	høyre	opp	ned
(0,0)	1.0	-0.5	0.1	-0.3
(1,0)	0.2	-0.3	-1.0	0.6
(2,0)	-0.4	-0.1	0.3	0.7
(0,1)	0.4	-0.9	1.0	0.2
(1,1)	0.6	-0.1	0.4	0.8
(2,1)	1.0	0.5	0.8	-0.5
(0,2)	-0.2	0.5	-0.3	-0.7
(1,2)	0.7	-1.0	0.1	-0.5
(2,2)	0.1	0.0	0.8	0.1

**Spørsmål:**

- **Q:** Hva er  $Q((0,0), \text{høyre})$ ?
- **A:**  $-0.5$



# Q-tabell

Hvis vi har endelig mange tilstander og handlinger, kan vi representere Q-funksjonen  $Q: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  som en tabell (**Q-tabell**):

$s \backslash a$	venstre	høyre	opp	ned
(0, 0)	1.0	-0.5	0.1	-0.3
(1, 0)	0.2	-0.3	-1.0	0.6
(2, 0)	-0.4	-0.1	0.3	0.7
(0, 1)	0.4	-0.9	1.0	0.2
(1, 1)	0.6	-0.1	0.4	0.8
(2, 1)	1.0	0.5	0.8	-0.5
(0, 2)	-0.2	0.5	-0.3	-0.7
(1, 2)	0.7	-1.0	0.1	-0.5
(2, 2)	0.1	0.0	0.8	0.1

## Spørsmål:

- **Q:** Hva er  $Q((0, 0), \text{høyre})$ ?
- **A:**  $-0.5$
- **Q:** Hva er  $Q((1, 1), \text{ned})$ ?



# Q-tabell

Hvis vi har endelig mange tilstander og handlinger, kan vi representere Q-funksjonen  $Q: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  som en tabell (**Q-tabell**):

$s \backslash a$	venstre	høyre	opp	ned
(0, 0)	1.0	-0.5	0.1	-0.3
(1, 0)	0.2	-0.3	-1.0	0.6
(2, 0)	-0.4	-0.1	0.3	0.7
(0, 1)	0.4	-0.9	1.0	0.2
(1, 1)	0.6	-0.1	0.4	0.8
(2, 1)	1.0	0.5	0.8	-0.5
(0, 2)	-0.2	0.5	-0.3	-0.7
(1, 2)	0.7	-1.0	0.1	-0.5
(2, 2)	0.1	0.0	0.8	0.1

## Spørsmål:

- **Q:** Hva er  $Q((0, 0), \text{høyre})$ ?
- **A:**  $-0.5$
- **Q:** Hva er  $Q((1, 1), \text{ned})$ ?
- **A:**  $0.8$



# Q-tabell

Hvis vi har endelig mange tilstander og handlinger, kan vi representere Q-funksjonen  $Q: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  som en tabell (**Q-tabell**):

$s \backslash a$	venstre	høyre	opp	ned
(0,0)	1.0	-0.5	0.1	-0.3
(1,0)	0.2	-0.3	-1.0	0.6
(2,0)	-0.4	-0.1	0.3	0.7
(0,1)	0.4	-0.9	1.0	0.2
(1,1)	0.6	-0.1	0.4	0.8
(2,1)	1.0	0.5	0.8	-0.5
(0,2)	-0.2	0.5	-0.3	-0.7
(1,2)	0.7	-1.0	0.1	-0.5
(2,2)	0.1	0.0	0.8	0.1

## Spørsmål:

- **Q:** Hva er  $Q((0,0), \text{høyre})$ ?
- **A:**  $-0.5$
- **Q:** Hva er  $Q((1,1), \text{ned})$ ?
- **A:**  $0.8$
- **Q:** Hva er  $\max_{a \in \mathcal{A}} Q(s, a)$  når  $s = (1,2)$ ?



# Q-tabell

Hvis vi har endelig mange tilstander og handlinger, kan vi representere Q-funksjonen  $Q: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  som en tabell (**Q-tabell**):

$s \backslash a$	venstre	høyre	opp	ned
(0,0)	1.0	-0.5	0.1	-0.3
(1,0)	0.2	-0.3	-1.0	0.6
(2,0)	-0.4	-0.1	0.3	0.7
(0,1)	0.4	-0.9	1.0	0.2
(1,1)	0.6	-0.1	0.4	0.8
(2,1)	1.0	0.5	0.8	-0.5
(0,2)	-0.2	0.5	-0.3	-0.7
(1,2)	0.7	-1.0	0.1	-0.5
(2,2)	0.1	0.0	0.8	0.1

## Spørsmål:

- **Q:** Hva er  $Q((0,0), \text{høyre})$ ?
- **A:**  $-0.5$
- **Q:** Hva er  $Q((1,1), \text{ned})$ ?
- **A:**  $0.8$
- **Q:** Hva er  $\max_{a \in \mathcal{A}} Q(s, a)$  når  $s = (1,2)$ ?
- **A:**  $0.7$



# Q-tabell

Hvis vi har endelig mange tilstander og handlinger, kan vi representere Q-funksjonen  $Q: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  som en tabell (**Q-tabell**):

$s \backslash a$	venstre	høyre	opp	ned
(0,0)	1.0	-0.5	0.1	-0.3
(1,0)	0.2	-0.3	-1.0	0.6
(2,0)	-0.4	-0.1	0.3	0.7
(0,1)	0.4	-0.9	1.0	0.2
(1,1)	0.6	-0.1	0.4	0.8
(2,1)	1.0	0.5	0.8	-0.5
(0,2)	-0.2	0.5	-0.3	-0.7
(1,2)	0.7	-1.0	0.1	-0.5
(2,2)	0.1	0.0	0.8	0.1

## Spørsmål:

- **Q:** Hva er  $Q((0,0), \text{høyre})$ ?
- **A:**  $-0.5$
- **Q:** Hva er  $Q((1,1), \text{ned})$ ?
- **A:**  $0.8$
- **Q:** Hva er  $\max_{a \in \mathcal{A}} Q(s, a)$  når  $s = (1,2)$ ?
- **A:**  $0.7$
- **Q:** Hva er  $\max_{a \in \mathcal{A}} Q(s, a)$  når  $s = (2,2)$ ?





# Q-tabell

Hvis vi har endelig mange tilstander og handlinger, kan vi representere Q-funksjonen  $Q: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  som en tabell (**Q-tabell**):

$s \backslash a$	venstre	høyre	opp	ned
(0,0)	1.0	-0.5	0.1	-0.3
(1,0)	0.2	-0.3	-1.0	0.6
(2,0)	-0.4	-0.1	0.3	0.7
(0,1)	0.4	-0.9	1.0	0.2
(1,1)	0.6	-0.1	0.4	0.8
(2,1)	1.0	0.5	0.8	-0.5
(0,2)	-0.2	0.5	-0.3	-0.7
(1,2)	0.7	-1.0	0.1	-0.5
(2,2)	0.1	0.0	0.8	0.1

## Spørsmål:

- **Q:** Hva er  $Q((0,0), \text{høyre})$ ?
- **A:**  $-0.5$
- **Q:** Hva er  $Q((1,1), \text{ned})$ ?
- **A:**  $0.8$
- **Q:** Hva er  $\max_{a \in \mathcal{A}} Q(s, a)$  når  $s = (1,2)$ ?
- **A:**  $0.7$
- **Q:** Hva er  $\max_{a \in \mathcal{A}} Q(s, a)$  når  $s = (2,2)$ ?
- **A:**  $0.8$



# Eksempel: Handling basert på Q-tabell

La  $\pi^*: \mathcal{S} \rightarrow \mathcal{A}$  være funksjonen gitt ved  $\pi^*(s) = \arg \max_{a \in \mathcal{A}} Q(s, a)$ .

$s \backslash a$	venstre	høyre	opp	ned
(0, 0)	1.0	-0.5	0.1	-0.3
(1, 0)	0.2	-0.3	-1.0	0.6
(2, 0)	-0.4	-0.1	0.3	0.7
(0, 1)	0.4	-0.9	1.0	0.2
(1, 1)	0.6	-0.1	0.4	0.8
(2, 1)	1.0	0.5	0.8	-0.5
(0, 2)	-0.2	0.5	-0.3	-0.7
(1, 2)	0.7	-1.0	0.1	-0.5
(2, 2)	0.1	0.0	0.8	0.1

**Spørsmål:**

- **Q:** Hva er  $\pi^*((0, 1))$ ?



# Eksempel: Handling basert på Q-tabell

La  $\pi^*: \mathcal{S} \rightarrow \mathcal{A}$  være funksjonen gitt ved  $\pi^*(s) = \arg \max_{a \in \mathcal{A}} Q(s, a)$ .

$s \backslash a$	venstre	høyre	opp	ned
(0, 0)	1.0	-0.5	0.1	-0.3
(1, 0)	0.2	-0.3	-1.0	0.6
(2, 0)	-0.4	-0.1	0.3	0.7
(0, 1)	0.4	-0.9	1.0	0.2
(1, 1)	0.6	-0.1	0.4	0.8
(2, 1)	1.0	0.5	0.8	-0.5
(0, 2)	-0.2	0.5	-0.3	-0.7
(1, 2)	0.7	-1.0	0.1	-0.5
(2, 2)	0.1	0.0	0.8	0.1

Spørsmål:

- **Q:** Hva er  $\pi^*((0, 1))$ ?
- **A:** opp



# Eksempel: Handling basert på Q-tabell

La  $\pi^*: \mathcal{S} \rightarrow \mathcal{A}$  være funksjonen gitt ved  $\pi^*(s) = \arg \max_{a \in \mathcal{A}} Q(s, a)$ .

$s \backslash a$	venstre	høyre	opp	ned
(0, 0)	1.0	-0.5	0.1	-0.3
(1, 0)	0.2	-0.3	-1.0	0.6
(2, 0)	-0.4	-0.1	0.3	0.7
(0, 1)	0.4	-0.9	1.0	0.2
(1, 1)	0.6	-0.1	0.4	0.8
(2, 1)	1.0	0.5	0.8	-0.5
(0, 2)	-0.2	0.5	-0.3	-0.7
(1, 2)	0.7	-1.0	0.1	-0.5
(2, 2)	0.1	0.0	0.8	0.1

## Spørsmål:

- **Q:** Hva er  $\pi^*((0, 1))$ ?
- **A:** opp
- **Q:** Hva er  $\pi^*((2, 1))$ ?



# Eksempel: Handling basert på Q-tabell

La  $\pi^*: \mathcal{S} \rightarrow \mathcal{A}$  være funksjonen gitt ved  $\pi^*(s) = \arg \max_{a \in \mathcal{A}} Q(s, a)$ .

$s \backslash a$	venstre	høyre	opp	ned
(0, 0)	1.0	-0.5	0.1	-0.3
(1, 0)	0.2	-0.3	-1.0	0.6
(2, 0)	-0.4	-0.1	0.3	0.7
(0, 1)	0.4	-0.9	1.0	0.2
(1, 1)	0.6	-0.1	0.4	0.8
(2, 1)	1.0	0.5	0.8	-0.5
(0, 2)	-0.2	0.5	-0.3	-0.7
(1, 2)	0.7	-1.0	0.1	-0.5
(2, 2)	0.1	0.0	0.8	0.1

## Spørsmål:

- **Q:** Hva er  $\pi^*((0, 1))$ ?
- **A:** opp
- **Q:** Hva er  $\pi^*((2, 1))$ ?
- **A:** venstre



# Eksempel: Handling basert på Q-tabell

La  $\pi^*: \mathcal{S} \rightarrow \mathcal{A}$  være funksjonen gitt ved  $\pi^*(s) = \arg \max_{a \in \mathcal{A}} Q(s, a)$ .

$s \backslash a$	venstre	høyre	opp	ned
(0, 0)	1.0	-0.5	0.1	-0.3
(1, 0)	0.2	-0.3	-1.0	0.6
(2, 0)	-0.4	-0.1	0.3	0.7
(0, 1)	0.4	-0.9	1.0	0.2
(1, 1)	0.6	-0.1	0.4	0.8
(2, 1)	1.0	0.5	0.8	-0.5
(0, 2)	-0.2	0.5	-0.3	-0.7
(1, 2)	0.7	-1.0	0.1	-0.5
(2, 2)	0.1	0.0	0.8	0.1

## Spørsmål:

- **Q:** Hva er  $\pi^*((0, 1))$ ?
- **A:** opp
- **Q:** Hva er  $\pi^*((2, 1))$ ?
- **A:** venstre
- **Q:** Hva er  $\pi^*((0, 2))$ ?



# Eksempel: Handling basert på Q-tabell

La  $\pi^*: \mathcal{S} \rightarrow \mathcal{A}$  være funksjonen gitt ved  $\pi^*(s) = \arg \max_{a \in \mathcal{A}} Q(s, a)$ .

$s \backslash a$	venstre	høyre	opp	ned
(0, 0)	1.0	-0.5	0.1	-0.3
(1, 0)	0.2	-0.3	-1.0	0.6
(2, 0)	-0.4	-0.1	0.3	0.7
(0, 1)	0.4	-0.9	1.0	0.2
(1, 1)	0.6	-0.1	0.4	0.8
(2, 1)	1.0	0.5	0.8	-0.5
(0, 2)	-0.2	0.5	-0.3	-0.7
(1, 2)	0.7	-1.0	0.1	-0.5
(2, 2)	0.1	0.0	0.8	0.1

## Spørsmål:

- **Q:** Hva er  $\pi^*((0, 1))$ ?
- **A:** opp
- **Q:** Hva er  $\pi^*((2, 1))$ ?
- **A:** venstre
- **Q:** Hva er  $\pi^*((0, 2))$ ?
- **A:** høyre



# Hvordan lære Q-funksjonen?

Vi starter med  $Q(s, a) = 0$  for alle par  $(s, a) \in \mathcal{S} \times \mathcal{A}$ . (En Q-tabell hvor alle verdiene er 0.)



# Hvordan lære Q-funksjonen?

Vi starter med  $Q(s, a) = 0$  for alle par  $(s, a) \in \mathcal{S} \times \mathcal{A}$ . (En Q-tabell hvor alle verdiene er 0.)

Vi har to **læringsparametere** (begge tall mellom 0 og 1):

- $\alpha$ : **læringsrate** (learning rate) og
- $\gamma$ : **rabattfaktor** (discount factor).

# Hvordan lære Q-funksjonen?

Vi starter med  $Q(s, a) = 0$  for alle par  $(s, a) \in \mathcal{S} \times \mathcal{A}$ . (En Q-tabell hvor alle verdiene er 0.)

Vi har to **læringsparametere** (begge tall mellom 0 og 1):

- $\alpha$ : **læringsrate** (learning rate) og
- $\gamma$ : **rabattfaktor** (discount factor).

**Q-læringsalgoritmen (én episode):**

- 1 Vi er i tilstanden  $s_t$  ved tid  $t$ . Velg en handling  $a_t$ . Ved å utføre  $a_t$  i tilstand  $s_t$  ender vi opp i tilstand  $s_{t+1}$ .

# Hvordan lære Q-funksjonen?

Vi starter med  $Q(s, a) = 0$  for alle par  $(s, a) \in \mathcal{S} \times \mathcal{A}$ . (En Q-tabell hvor alle verdiene er 0.)

Vi har to **læringsparametere** (begge tall mellom 0 og 1):

- $\alpha$ : **læringsrate** (learning rate) og
- $\gamma$ : **rabattfaktor** (discount factor).

**Q-læringsalgoritmen (én episode):**

- 1 Vi er i tilstanden  $s_t$  ved tid  $t$ . Velg en handling  $a_t$ . Ved å utføre  $a_t$  i tilstand  $s_t$  ender vi opp i tilstand  $s_{t+1}$ .
- 2 Vi oppdaterer Q-verdien  $Q(s_t, a_t)$  med følgende regel:

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha \left( R(s_t, a_t) + \gamma \max_{a \in \mathcal{A}} Q(s_{t+1}, a) \right).$$

# Hvordan lære Q-funksjonen?

Vi starter med  $Q(s, a) = 0$  for alle par  $(s, a) \in \mathcal{S} \times \mathcal{A}$ . (En Q-tabell hvor alle verdiene er 0.)

Vi har to **læringsparametere** (begge tall mellom 0 og 1):

- $\alpha$ : **læringsrate** (learning rate) og
- $\gamma$ : **rabattfaktor** (discount factor).

**Q-læringsalgoritmen (én episode):**

- 1 Vi er i tilstanden  $s_t$  ved tid  $t$ . Velg en handling  $a_t$ . Ved å utføre  $a_t$  i tilstand  $s_t$  ender vi opp i tilstand  $s_{t+1}$ .
- 2 Vi oppdaterer Q-verdien  $Q(s_t, a_t)$  med følgende regel:

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha \left( R(s_t, a_t) + \gamma \max_{a \in \mathcal{A}} Q(s_{t+1}, a) \right).$$

- 3 Gjenta fra steg 1 med  $s_{t+1}$  (stopp hvis  $s_{t+1}$  er en terminaltilstand).

# Oppdatering av Q-funksjonen

$$(1 - \alpha) \underbrace{Q(s_t, a_t)}_{\text{nåværende Q-verdi}} + \alpha \left( \overbrace{R(s_t, a_t)}^{\text{Umiddelbar belønning}} + \gamma \underbrace{\max_{a \in \mathcal{A}} Q(s_{t+1}, a)}_{\text{estimert høyeste fremtidig belønning}} \right).$$

- Gammel og ny erfaring kombineres ( $\alpha$  bestemmer balansen).
- Belønningen for å utføre  $a_t$  i tilstand  $s_t$  påvirker den nye Q-verdien.
- Hvor mye vi bryr oss om fremtiden bestemmes av  $\gamma$ .

# $\epsilon$ -grådig Q-læring: Hvordan velge $a_t$ ?

La  $\epsilon$  være et tall mellom 0 og 1. I tilstand  $s_t$ , velg  $a_t$  på følgende måte:

- 1 Med sannsynlighet  $\epsilon$ , velg  $a_t$  tilfeldig.
- 2 Med sannsynlighet  $1 - \epsilon$ , velg  $a_t = \pi^*(s_t) = \arg \max_{a \in \mathcal{A}} Q(s_t, a)$ .

Vi reduserer vanligvis verdien av  $\epsilon$  gjennom læringen slik at agenten utforsker mest i starten men gradvis baserer valgene på lært kunnskap.

# Eksempel: Læringsteg 1

Sett  $\alpha = 0.8$  og  $\gamma = 0.5$ . La  $s_t = (2, 1)$  og  $a_t = \text{opp}$ .

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)

A 3x3 grid representing a state space. The cell (2,1) is highlighted in yellow, and a purple arrow points to it from below. The cell (0,2) is highlighted in blue. Other cells are white or black.

s \ a	venstre	høyre	opp	ned
(0,1)	0.1	1.0	-0.8	0.2
(1,1)	-0.6	-0.8	0.9	0.2
(2,1)	-0.4	-1.0	0.3	-0.9

A 5x5 table representing a Q-table. The row for state (2,1) is highlighted in orange. The cell (2,1) for action 'opp' contains the value 0.3.

$$Q(s_t, a_t) \leftarrow \underbrace{(1 - \alpha)}_{0.2} \underbrace{Q(s_t, a_t)}_{0.3} + \underbrace{\alpha}_{0.8} \left( \underbrace{R(s_t, a_t)}_{-0.1} + \underbrace{\gamma}_{0.5} \underbrace{\max_{a \in \mathcal{A}} Q(\overset{(1,1)}{s_{t+1}}, a)}_{0.9} \right).$$

Ny Q-verdi:  $Q((2, 1), \text{opp}) = 0.2 \cdot 0.3 + 0.8(-0.1 + 0.5 \cdot 0.9) = \mathbf{0.34}$

## Eksempel: Læringsteg 2

Nå er  $s_t = (1, 1)$ . La  $a_t = \text{venstre}$ .

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)

a \ s	venstre	høyre	opp	ned
(0,1)	0.1	1.0	-0.8	0.2
(1,1)	-0.6	-0.8	0.9	0.2
(2,1)	-0.4	-1.0	0.34	-0.9

$$Q(s_t, a_t) \leftarrow \underbrace{(1 - \alpha)}_{0.2} \underbrace{Q(s_t, a_t)}_{-0.6} + \underbrace{\alpha}_{0.8} \left( \underbrace{R(s_t, a_t)}_{-1.0} + \underbrace{\gamma}_{0.5} \max_{a \in \mathcal{A}} \underbrace{Q(s_{t+1}, a)}_{\substack{(1,1) \\ 0.9}} \right).$$

Ny Q-verdi:  $Q((1, 1), \text{venstre}) = -0.56$



## Eksempel: Læringsteg 3

Vi har fortsatt  $s_t = (1, 1)$ . La  $a_t := \pi^*(s_t) = \text{opp}$ .

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)

A purple arrow points upwards from the cell (1,1) to the cell (0,1).

a \ s	venstre	høyre	opp	ned
(0,1)	0.1	1.0	-0.8	0.2
(1,1)	-0.56	-0.8	0.9	0.2
(2,1)	-0.4	-1.0	0.34	-0.9

$$Q(s_t, a_t) \leftarrow \underbrace{(1 - \alpha)}_{0.2} \underbrace{Q(s_t, a_t)}_{0.9} + \underbrace{\alpha}_{0.8} \left( \underbrace{R(s_t, a_t)}_{-0.1} + \underbrace{\gamma}_{0.5} \underbrace{\max_{a \in \mathcal{A}} Q(\overset{(0,1)}{s_{t+1}}, a)}_{1.0} \right).$$

Ny Q-verdi:  $Q((1, 1), \text{opp}) = \mathbf{0.5}$

## Eksempel: Læringsteg 4

Nå er  $s_t = (0, 1)$ . La  $a_t := \pi^*(s_t) = \text{høyre}$ .

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)

a \ s	venstre	høyre	opp	ned
(0,1)	0.1	1.0	-0.8	0.2
(1,1)	-0.56	-0.8	0.5	0.2
(2,1)	-0.4	-1.0	0.34	-0.9

$$Q(s_t, a_t) \leftarrow \underbrace{(1 - \alpha)}_{0.2} \underbrace{Q(s_t, a_t)}_{1.0} + \underbrace{\alpha}_{0.8} \left( \underbrace{R(s_t, a_t)}_{1.0} + \underbrace{\gamma}_{0.5} \underbrace{\max_{a \in \mathcal{A}} Q(\overset{(0,2)}{s_{t+1}}, a)}_{0.0} \right).$$

Ny Q-verdi:  $Q((0, 1), \text{høyre}) = 1.0$

# Workshop

Nå er det din tur til å implementere Q-læring!

- Gå til <https://github.com/odinhg/QL24>



- Spør en gruppeleder eller meg dersom du har spørsmål.