Q-læring

INF100

Odin Hoff Gardå

UNIVERSITY OF BERGEN

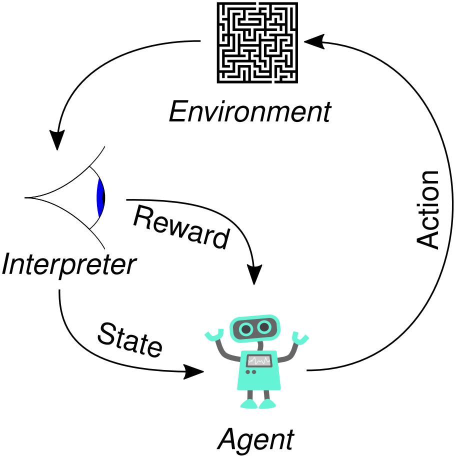
Plan

1. Kort introduksjon til Q-læring (15 - 30 min.)

2. Workshop: Implementer Q-læring for å løse en labyrint.

Forsterkende Læring[1]

- Vi har en **agent** som **handler** i et **miljø**.
- Agenten lærer gjennom **belønning** basert på agentens **tilstand** og **handling**.
- **Q-læring** er en form for forsterkende læring.



[1] Engelsk: Reinforcement Learning

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** $R\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

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** $R\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Ved å la agenten utforske miljøet ønsker vi å lære **Q-funksjonen**

$$Q\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$$

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

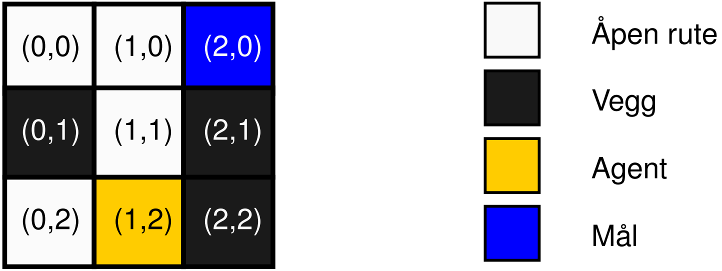
Q-læring: Oppsett

- 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** $R \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Ved å la agenten utforske miljøet ønsker vi å lære **Q-funksjonen**

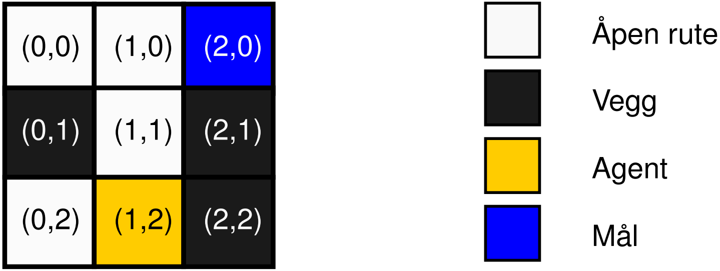
$$Q \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$$

  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: 3×3 Labyrint



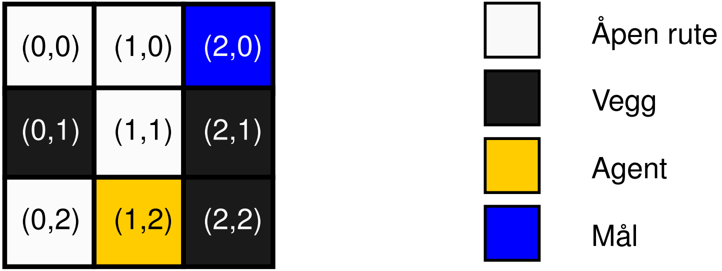
Gjennomgående Eksempel: 3×3 Labyrint



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: 3×3 Labyrint



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):

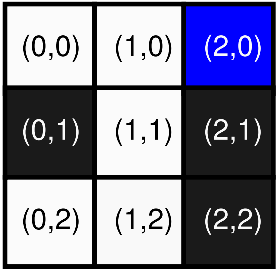
$$\mathcal{A} = \{\text{venstre}, \text{høyre}, \text{opp}, \text{ned}\}$$

Eksempel: Belønningsfunksjonen

La s' være posisjonen vi treffer ved å gå i retning a fra posisjon s.

Definer belønningsfunksjonen $R\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ ved

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



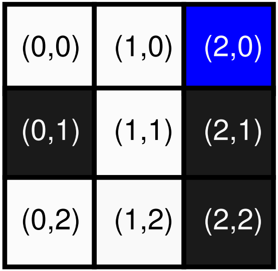
- **Q:** Hva er $R((1,0), \text{høyre})$?

Eksempel: Belønningsfunksjonen

La s' være posisjonen vi treffer ved å gå i retning a fra posisjon s.

Definer belønningsfunksjonen $R\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ ved

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



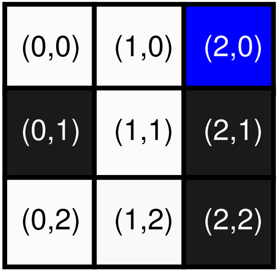
- **Q:** Hva er $R((1,0), \text{høyre})$?
- **A:** $R((1,0), \text{høyre}) = 1.0$

Eksempel: Belønningsfunksjonen

La s' være posisjonen vi treffer ved å gå i retning a fra posisjon s.

Definer belønningsfunksjonen $R\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ ved

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



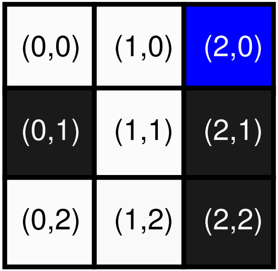
- **Q:** Hva er $R((1,0), \text{høyre})$?
- **A:** $R((1,0), \text{høyre}) = 1.0$
- **Q:** Hva er $R((1,1), \text{venstre})$?

Eksempel: Belønningsfunksjonen

La s' være posisjonen vi treffer ved å gå i retning a fra posisjon s.

Definer belønningsfunksjonen $R\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ ved

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



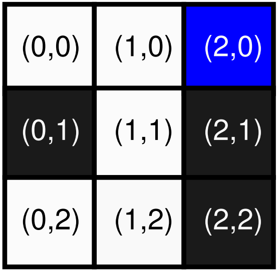
- **Q:** Hva er $R((1,0), \text{høyre})$?
- **A:** $R((1,0), \text{høyre}) = 1.0$
- **Q:** Hva er $R((1,1), \text{venstre})$?
- **A:** $R((1,1), \text{venstre}) = -1.0$

Eksempel: Belønningsfunksjonen

La s' være posisjonen vi treffer ved å gå i retning a fra posisjon s.

Definer belønningsfunksjonen $R\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ ved

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



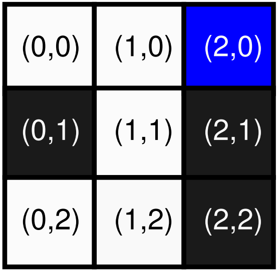
- **Q:** Hva er $R((1,0), \text{høyre})$?
- **A:** $R((1,0), \text{høyre}) = 1.0$
- **Q:** Hva er $R((1,1), \text{venstre})$?
- **A:** $R((1,1), \text{venstre}) = -1.0$
- **Q:** Hva er $R((1,1), \text{opp})$?

Eksempel: Belønningsfunksjonen

La s' være posisjonen vi treffer ved å gå i retning a fra posisjon s.

Definer belønningsfunksjonen $R\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ ved

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



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

s \ 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 som en tabell (**Q-tabell**):

s \ 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 som en tabell (**Q-tabell**):

s \ 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 som en tabell (**Q-tabell**):

s \ 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 som en tabell (**Q-tabell**):

s \ 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 som en tabell (**Q-tabell**):

s \ 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 som en tabell (**Q-tabell**):

s \ 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 som en tabell (**Q-tabell**):

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

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

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

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

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

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

s \ 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

Eksempel: ϵ-grådig Q-læring

La ϵ være et tall mellom 0 og 1. Med ϵ-grådig læring velger agenten å utføre

1. en tilfeldig handling med sannsynlighet ϵ, og
2. handlingen $\pi^*(s)$ med sannsynlighet $1 - \epsilon$.

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

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):

- α: **læringsrate** (learning rate) og
- γ: **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):

- α: **læringsrate** (learning rate) og
- γ: **rabattfaktor** (discount factor).

Q-læringsalgoritmen (én episode):

1. Agenten er i posisjon s_t ved tid t. Vi bruker den ϵ-grådige strategien for å velge en handling a_t. Ved å utføre a_t i s_t treffer vi 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):

- α: **læringsrate** (learning rate) og
- γ: **rabattfaktor** (discount factor).

Q-læringsalgoritmen (én episode):

1. Agenten er i posisjon s_t ved tid t. Vi bruker den ϵ-grådige strategien for å velge en handling a_t. Ved å utføre a_t i s_t treffer vi s_{t+1}.
2. Vi oppdaterer $Q(s_t, a_t)$ 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):

- α: **læringsrate** (learning rate) og
- γ: **rabattfaktor** (discount factor).

Q-læringsalgoritmen (én episode):

1. Agenten er i posisjon s_t ved tid t. Vi bruker den ϵ-grådige strategien for å velge en handling a_t. Ved å utføre a_t i s_t treffer vi s_{t+1}.
2. Vi oppdaterer $Q(s_t, a_t)$ 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).

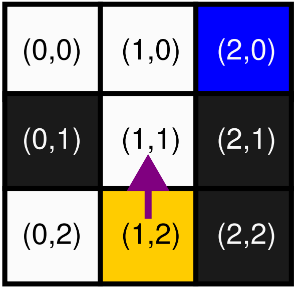
Oppdatering av Q-funksjonen

$$Q(s_t, a_t) \leftarrow (1-\alpha) \underbrace{Q(s_t, a_t)}_{\text{nåværende Q-verdi}} + \alpha \left(\overbrace{R(s_t, a_t)}^{\text{Belønning}} + \gamma \underbrace{\max_{a \in \mathcal{A}} Q(s_{t+1}, a)}_{\substack{\text{estimert beste} \\ \text{fremtidige Q-verdi}}} \right).$$

- Gammel og ny kunnskap kombineres (α bestemmer hvor mye av hver).
- Belønningen for å utføre a_t i tilstand s_t påvirker den nye Q-verdien.
- Hvor mye vi bryr oss om fremtiden bestemmes av γ.

Eksempel: Læringsteg 1

La $\alpha = 0.8$, $\gamma = 0.5$. Anta at $s_t = (1,2)$ og at agenten utfører $a_t = $ opp.



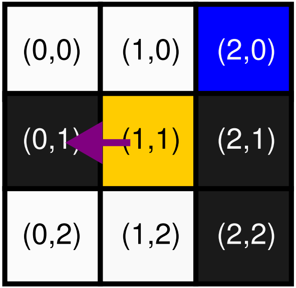
s \ a	venstre	høyre	opp	ned
⋮				
$(1,0)$	0.1	1.0	-0.8	0.2
$(1,1)$	-0.6	-0.8	0.9	0.2
$(1,2)$	-0.4	-1.0	0.3	-0.9
⋮				

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

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

Eksempel: Læringsteg 2

Nå er $s_t = (1, 1)$. Anta at agenten tilfeldig velger $a_t =$ venstre.



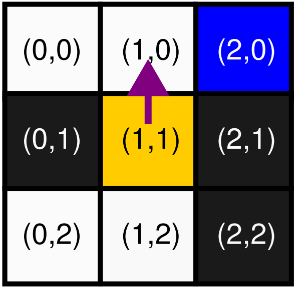
s \ a	venstre	høyre	opp	ned
⋮				
$(1, 0)$	0.1	1.0	-0.8	0.2
$(1, 1)$	-0.6	-0.8	0.9	0.2
$(1, 2)$	-0.4	-1.0	*0.34*	-0.9
⋮				

$$Q(s_t, a_t) \leftarrow \underbrace{(1-\alpha)}_{0.2}\underbrace{Q(s_t, a_t)}_{-0.6} + \underbrace{\alpha}_{0.8}\left(\underbrace{R(s_t, a_t)}_{-1.0} + \underbrace{\gamma}_{0.5}\underbrace{\max_{a \in \mathcal{A}} Q(\overbrace{s_{t+1}}^{(0,1)}, a)}_{0.0}\right).$$

Ny Q-verdi: $Q((1, 1), \text{venstre}) = \mathbf{-0.92}$

Eksempel: Læringsteg 3

Nå er $s_t = (1, 1)$. Anta at agenten velger $a_t := \pi^*(s_t) = \text{opp}$.



s \ a	venstre	høyre	opp	ned
⋮				
$(1,0)$	0.1	1.0	-0.8	0.2
$(1,1)$	*-0.92*	-0.8	0.9	0.2
$(1,2)$	-0.4	-1.0	*0.34*	-0.9
⋮				

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

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

Eksempel: Læringsteg 4

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

s \ a	venstre	høyre	opp	ned
⋮				
$(1, 0)$	0.1	1.0	-0.8	0.2
$(1, 1)$	*-0.92*	-0.8	*0.5*	0.2
$(1, 2)$	-0.4	-1.0	*0.34*	-0.9
⋮				

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

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

Workshop

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

- Gå til https://github.com/odinhg/Q-Learning-Tutorial (eller skann QR-koden)
- Spør en gruppeleder eller meg dersom du har spørsmål.
- Prosjekter som utmerker seg havner under "Wall of Fame".