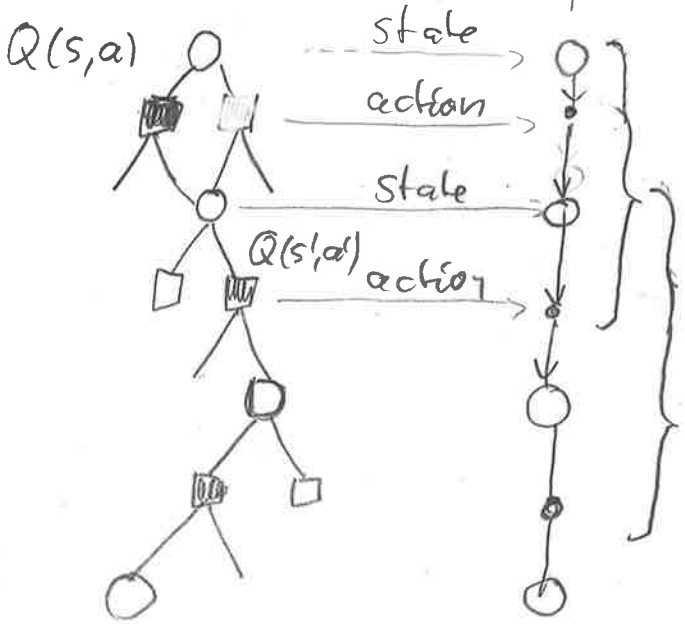
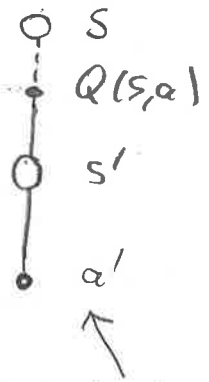
Week 9 — RL2 ; Blackboard1 : Backup Diagr

environment ⟶ path ⟶ update





need to know next action before you update Q(s,a) earlier

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r_t + \gamma Q(s',a') - Q(s,a)\right]$$

earlier — next

Exercise 1 (week 9)

	trials	Monte Carlo $\langle R(\tilde{s}, \tilde{a}) \rangle$ "average return"	Bootstrap $Q(\tilde{s}, \tilde{a})$ Q from Bellman
(s', a_3):	2, 4, 8	$\frac{1}{3}(1+1+1)$	$\frac{3}{3}(1+1+1)=1$
(s', a_4):	1, 3, 6, 7, 9	$\frac{1}{5}(0+0+0+0.5+0.5)$	$\frac{1}{5} \cdot 1$
(s, a_1):	5, 10	$\frac{1}{2} \cdot (0+0)$	0
(s, a_2):	1, 9	$\frac{1}{2}(0.2+0.7) = 0.45$; ↗	

$Q(s, a_2) = \langle r \rangle + \max_a Q(s', a')$

$= 0.2 + 1$

big difference for $Q(s, a_2)$ vs $R(s, a_2)$

$Q(s, a_2)$ much better!

Exercise 1 (continued) : relation online Q → batch Q

batch-Q = Q from Bellman

online Q: $Q(s,a) \leftarrow Q(s,a) + \alpha \cdot [\, r_t + \max_{a'} \{Q(s',a')\} - Q(s,a)]$

$\max_{a'} \{Q(s',a')\}$: fixed, compressed knowledge from previous trials

batch Q: n trials starting at s, a, trials $k = 1, \dots n$

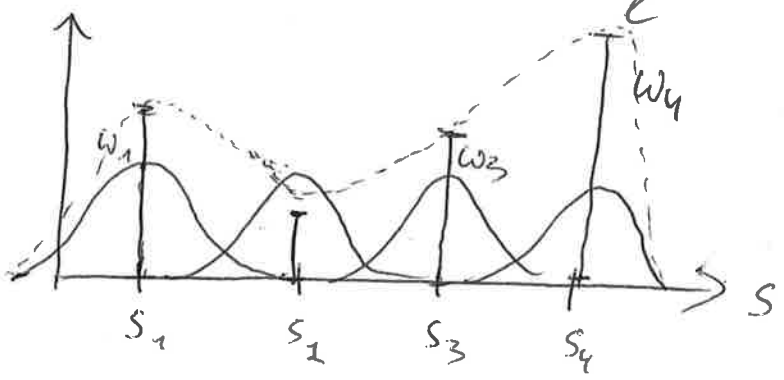
$$Q(s,a) \leftarrow Q(s,a) + \frac{1}{n} \left[\sum_{k=1}^{n} r_t(k) + \max_{a'} \{Q(s',a')\} - Q(s,a) \right]$$

initialize with $Q(s,a) = 0$ ↓ $=0$; ↓ $=0$

$$Q(s,a) \leftarrow \frac{1}{n} \left[\sum_{k=1}^{n} r_t(k) \right] + \max_{a'} \{Q(s',a')\}$$

$$Q(s,a) \leftarrow \langle r_t \rangle + \max_{a'} \{Q(s',a')\}$$

Blackboard 2

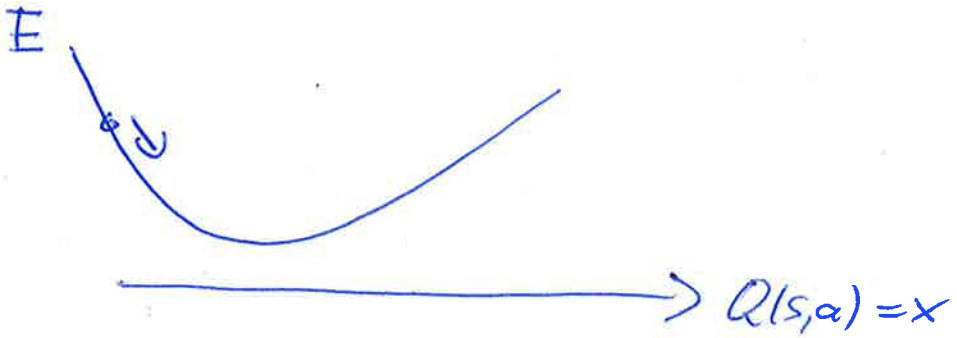


amplitudes

$\omega_1 \quad \omega_2 \quad \omega_3 \quad \omega_4$

$\Rightarrow$ smooth function with few parameters

$$E = \frac{1}{2}\left[\underbrace{r + Q(s',a')}_{\text{target}} - \underbrace{Q(s,a)}_{\text{variable } x}\right]^2$$



gradient descent

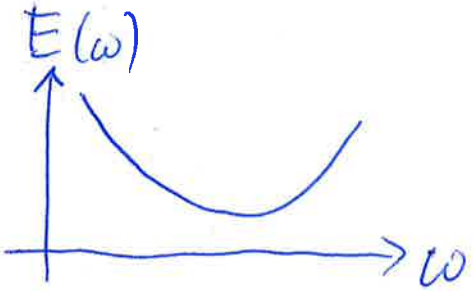
$$\Delta Q(s,a) = -\eta \frac{\partial E}{\partial x} = \underbrace{\left[r + Q(s',a') - Q(s,a)\right]}_{\delta} \cdot \underbrace{\delta_{ss_t}\delta_{aa_t}}_{\text{factor 1 if action } a \text{ was chosen in state } s}$$

→ SARSA update

Now with

$$Q(s,a) = Q_\omega(s,a)$$

↑ depends on weights ω



$$\Delta\omega = -\eta \cdot \frac{\partial E}{\partial \omega} = [\underbrace{r + Q(s',a')}_{\text{target}} - Q(s,a)] \cdot \frac{\partial Q_\omega(s,a)}{\partial \omega}$$

↑

if: radial basis function + linear output:

$$Q(s,a) = \sum_i \omega_{ai} \cdot \phi(s - s_i)$$

$$\frac{\partial Q}{\partial \omega_{bj}} = \phi(s - s_j) \cdot \delta_{ab}$$

↑ only in the action is $a = b$