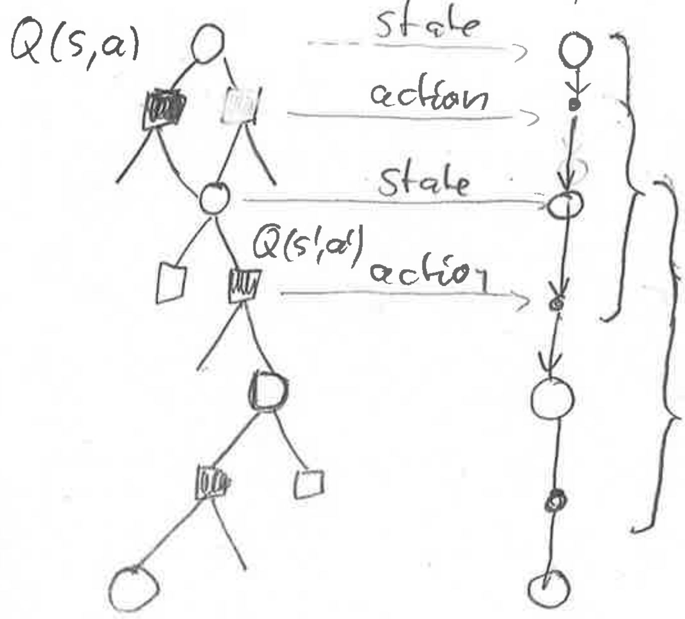
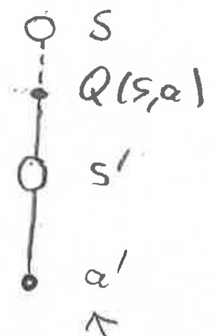
— RL2 ; Blackboard1: Backup Diagr

environment ⟶ path ⟶ update





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

earlier

$$\underset{\text{earlier action}}{Q(s,a)} \leftarrow Q(s,a) + \alpha\left[r_t + \gamma\, \underset{\text{next action}}{Q(s',a')} - Q(s,a)\right]$$

RL2, Exercise 1a + 1b   Blackboard 2

$(\hat{s},\hat{a})$ pair	encountered in trial	Monte Carlo average return $\langle R(\hat{s},\hat{a})\rangle$	Bootstrap Batch-SARSA from Bellman
(s', a_3)	2, 4, 8	$\frac{1}{3}[1+1+1]=1$	1
(s', a_4)	1, 3, 6, 7, 9	$\frac{1}{5}[0+0+0+0.5+0.5]=\frac{1}{5}$	$\frac{1}{5}$
(s, a_1)	5, 10	$\frac{1}{2}[0+0]=0$	0
(s, a_2)	1, 9	$\underline{\underline{\frac{1}{2}[0.2+0.7]=0.45}}$	$\langle r_t\rangle + \gamma\sum_{a'}\pi(s',a')Q(s',a')$ $\downarrow$ $0.2+\frac{1}{2}\left(\frac{1}{5}+1\right)=\underline{\underline{0.8}}$

only 2 trials are used

with same number of trials Bootstrap/Bellman/Batch-SARSA yields much better estimate than Monte-Carlo!

expected Batch-SARSA = Q from Bellman (without knowledge of branching ratio)

online SARSA-algo:

$$Q(s,a) \leftarrow Q(s,a) + \eta\cdot\left[r_t + \gamma\cdot \underbrace{\{Q(s',a')\}}_{\text{* compressed knowledge from previous trials}} - Q(s,a)\right]$$

expected batch-SARSA: n trials $(1\le k\le n)$ starting at (s,a)

$$Q(s,a_2) \leftarrow \underset{\downarrow\ 0}{Q(s,a)} + \underbrace{\frac{1}{n}\left[\sum_{k=1}^{n} r_t(k)\right]}_{\text{average}} + \gamma \underbrace{\sum_{a'}\pi(s',a')Q(s',a')}_{\text{* compressed knowledge from states close to target}} - \underset{\downarrow\ 0}{\gamma Q(s,a)}$$

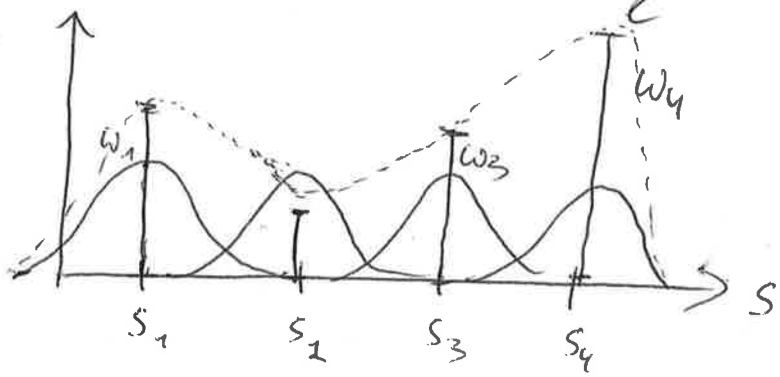
initialize: $Q=0$

since only one s', $\sum_{s'}$ dropped

$$Q(s,a_2) \leftarrow \langle r_t\rangle + \gamma\left[\sum_{a'}\pi(s',a')Q(s',a')\right]$$

here: $=1$ 0.5

Blackboard 3, RL2



$Q(a_2, s) = \sum_k w_k \phi(s - s_k)$

amplitudes

$w_1 \quad w_2 \quad w_3 \quad w_4$

$\Rightarrow$ smooth function with few parameters

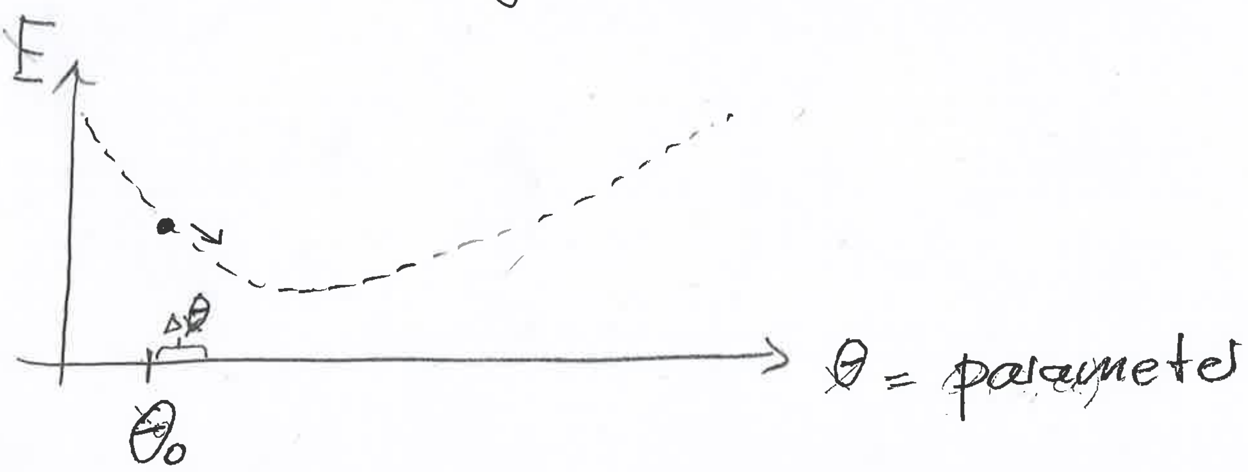
$Q(a_2, s)$: w_{21}, w_{22}, w_{23}, w_{24}

Blackboard 4 – RL 2 : Loss function

error (loss function)

$$E = \frac{1}{2}\Big[\underbrace{r + \gamma Q(s',a')}_{\text{target}} - \underbrace{Q(s,a)}_{\substack{\text{depends on parameters } \theta \\ \text{(the weights } w_1, w_2, \dots)}}\Big]^2$$

minimize error by gradient descent



$$\Delta\theta = -\eta \cdot \frac{\partial E}{\partial \theta} = +\eta \cdot \left[r + \gamma Q(s',a') - Q(s,a)\right] \frac{\partial Q(s,a)}{\partial \theta}$$