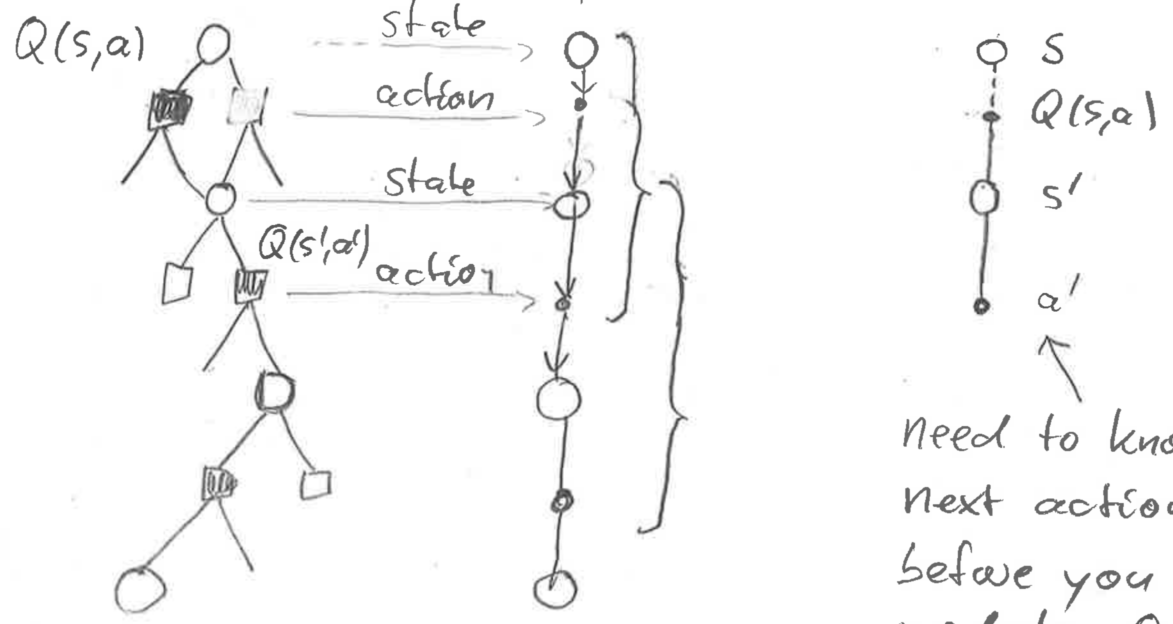
– RL2 ; Blackboard 1 : Backup Diagr

environment ⟶ path ⟶ update



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

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

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-Q 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	$\frac{1}{2}[0.2+0.7]=\underline{\underline{0.45}}$	$\langle r_t\rangle + \max_{a'} Q(s',a')$ → $0.2 + 1 = \underline{\underline{1.2}}$

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

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

online Q-learning

$$Q(s,a) \leftarrow Q(s,a) + \eta\cdot\Big[r_t + \underbrace{\max_{a'}\{Q(s',a')\}}_{}-\gamma Q(s,a)\Big]$$

* compressed knowledge from previous trials

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

$$Q(s,a) \leftarrow \underbrace{Q(s,a)}_{0} + \frac{1}{n}\Big[\sum_{k=1}^{n} r_t(k) + \underbrace{\max_{a'}\{Q(s',a')\}}_{} - \underbrace{\gamma Q(s,a)}_{0}\Big]$$

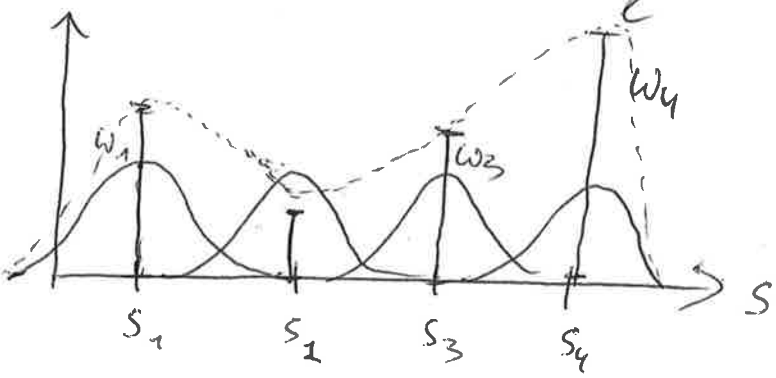
initialize: $Q=0$

average

* compressed knowledge from states closer to target

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

Blackboard 3, RL2



$$Q(a_1, s) = \sum_k w_k \, \phi(s - s_k)$$

amplitudes

w_1 w_2 w_3 w_4

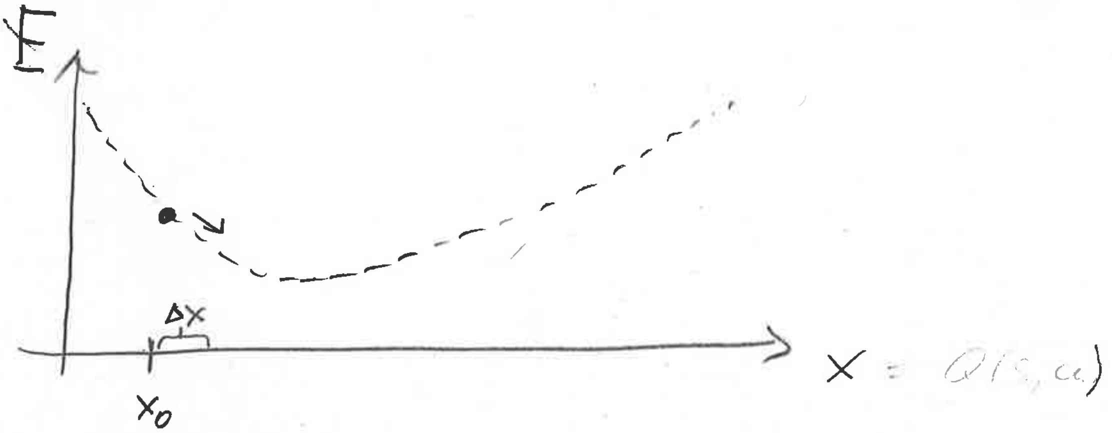
$\Rightarrow$ smooth function with few parameters

Blackboard 4 – RL 2 : Loss function

error

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

minimize error by gradient descent



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