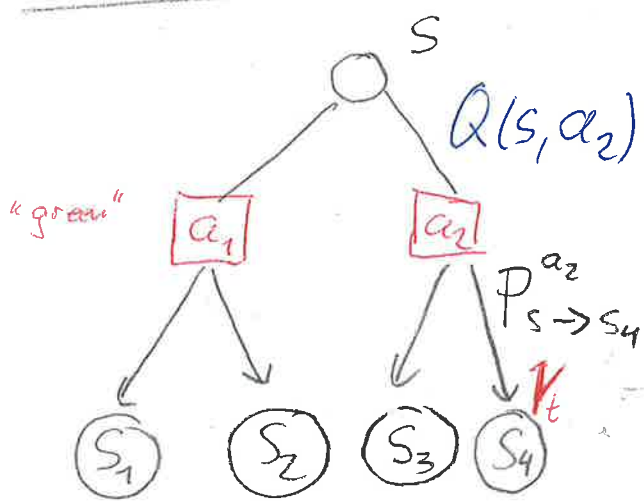
Blackboard (1)

Blackboard RL1 : Q - values



"branching ratio"

Transition probability

$$P^{a_2}_{s\to s_4} = P(s' = s_4 \mid a_2, s)$$

$\uparrow$ next state

- actual reward at time t : r_t
- expected reward for this "branch"

$$R^{a_2}_{s\to s_4} = E(r_t \mid s' = s_4, a_2, s)$$

reward received; end up in s_4; take a_2; start in s

- expected reward for action a_2

$$Q(s,a_2) = E(r_t \mid a_2, s)$$

$$= \sum_{s'} P^{a_2}_{s\to s'} \cdot R^{a_2}_{s\to s'}$$

(*Def Q)

s': all possible "next states"

- distinguish r_t from $R^{a_2}_{s\to s'}$! for example

$$r_t = \begin{cases} 4 & \text{with Prob } P_r(r=4 \mid s,a,s') \\ 1 & \text{" " } P_r(r=1 \mid s,a,s') \\ 0 & \text{" " } P_r(r=0 \mid s,a,s') \end{cases}$$

Blackboard RL1 - 3 : Stranges Proof [Blackboard ②]

Convergence in expectation

(perspective of "batch" update)

theorem (i) :

if $E[\Delta\hat{Q}(s,a)|s,a] = 0$ (H)

then $E[\hat{Q}(s,a)|s,a)] = \hat{Q}(s,a) = \sum_{s'} P^a_{s\to s'} R^a_{s\to s'} = Q(s,a)$

proof :

update formul

$\wedge$ = momentan estimate

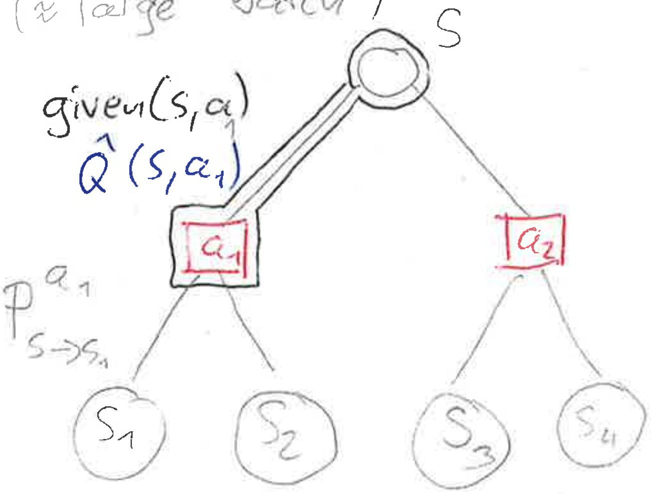
$$\underbrace{E}_{\text{possible futures with correct statistical weight ("large batch")}} [\Delta\hat{Q}(s,a)|\underbrace{s,a}_{\text{given s and a}}] = E[\eta\cdot(r_t - \hat{Q}(s,a))|s,a]$$

linear

$$= \eta\cdot E[r_t|s,a] - \eta E[\hat{Q}(s,a)|s,a]$$

$$= \eta\cdot\sum_{s'} P^a_{s\to s'}\cdot R^a_{s\to s'} - \eta \cancel{E}[\hat{Q}(s,a)]$$

$\hat{Q}(s,a)$ is fixed when conditioned on (s,a)
→ drop Expectat sign



with hypothesis H :

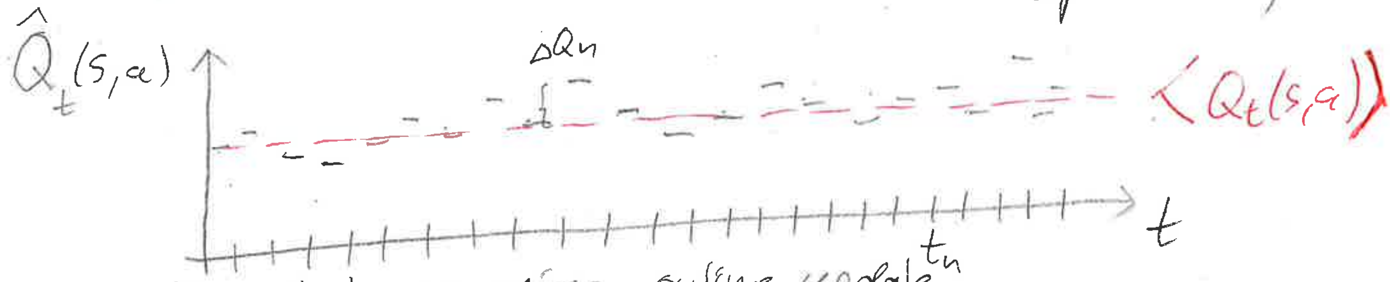
$$0 = \cancel{\eta}\cdot\sum_{s'} P^a_{s\to s'} R^a_{s\to s'} - \cancel{\eta} E[\hat{Q}(s,a)|s,a]$$

$$\cancel{E}[\hat{Q}(s,a)\cancel{|s,a}] = \sum_{s'} P^a_{s\to s'} R^a_{s\to s'} = Q(s,a)$$

↑ (* Def Q)

$$\boxed{\Longrightarrow \text{estimate } \hat{Q} \text{ equal "exact } Q\text{"}}$$

(ii) convergence of mean (perspective of "online" update) ③



mean over time while running online update

$$\langle \Delta \hat{Q}(s,a) | s,a \rangle_{t|s,a} := \lim_{N\to\infty} \frac{1}{N} \sum_{n=1} \Delta Q_n(s,a).$$

$t|s,a$: all time points with (s,a)

ΔQ_n: update in time step n

theorem (ii)

if $\langle \Delta \hat{Q}(s,a) | s,a \rangle_{t|s,a} = 0$

then $\langle \hat{Q}(s,a) \rangle_{t|s,a} = Q(s,a)$

proof: $\langle \Delta \hat{Q}(s,a) | s,a \rangle_{t|s,a} = \eta \langle (r_t - \hat{Q}_t(s,a)) | s,a \rangle_{t|s,a}$

$$\langle \Delta \hat{Q}(s,a) | s,a \rangle_{t|s,a} = \eta \langle r_t | s,a \rangle_{t|s,a} - \eta \langle \hat{Q}_t(s,a) \rangle_{t|}$$

automatic conditioning

$\Downarrow$ H

$$0 = \sum_{s'} P^a_{s\to s'} \cdot R^a_{s\to s'} - \langle \hat{Q}_t(s,a) \rangle$$

$$\Rightarrow \langle \hat{Q}_t(s,a) \rangle_t = \sum_{s'} P^a_{s\to s'} \cdot R^a_{s\to s'} \overset{(\text{* Def } Q)}{=} Q(s,a)$$

$\Rightarrow$ $\boxed{\text{mean over time } \langle Q_t \rangle_t = Q \text{ (exact Q-value)}}$

Q' Q_t Cee