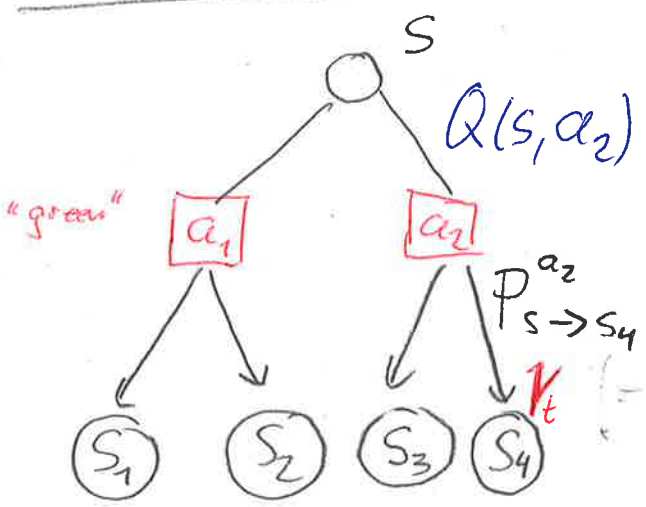
Blackboard RL1 : Q-values



"branching ratio"

Transition probability

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

(s': next state)

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

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

(r_t: reward received; $s'=s_4$: end up in s_4; a_2: take a_2; s: start in s)

- expected reward for action a_2

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

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

(s': all possible "next states")

($\star$ Def Q)

- 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 Simple Proof Convergence in expectation

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

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

note: expectation is given (s,a)!

---

proof:

(H) $E[\Delta\hat{Q}(s,a)|s,a] = 0$ ↓ updates average to 0 in expectations

Eq. (1) of slide ↓ $= \eta \cdot E[r_t - \hat{Q}(s,a)]$ ← estimate of Q, given (s,a)

$0 = E[r_t] - E[\hat{Q}(s,a)]$

$0 = \sum_{s'} P^{a}_{s\to s'} R^{a}_{s\to s'} - E[\hat{Q}(s,a)]$

$E[\hat{Q}(s,a)] = Q(s,a)$

(ii) Fluctations over time around empirical mean $\langle\hat{Q}\rangle$. role of η is qualitatively obvious.

smaller $\eta \Rightarrow$ smaller fluctuation

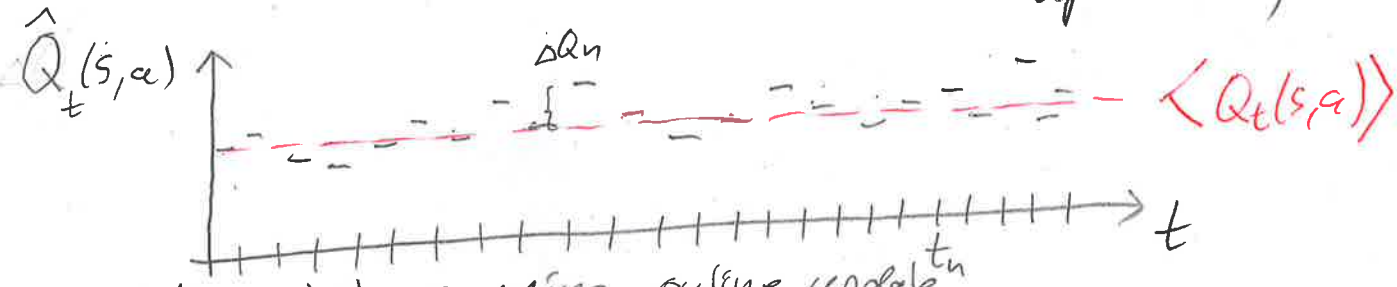
$\hat{Q}$ fluctuates around $\langle\hat{Q}(s,a)\rangle = Q(s,a)$

↑ long-term average expectation

repeat same calculation:

$\langle\Delta\hat{Q}(s,a)\rangle = 0 = \langle r_t\rangle - \langle\hat{Q}(s,a)\rangle$

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



mean over time while running online update

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

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

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

↑↑ automatic conditioning

$\downarrow$ H

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

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

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

Blackboard RL1-3 : Stranges Proof

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

$\wedge$ = momentary estimate

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

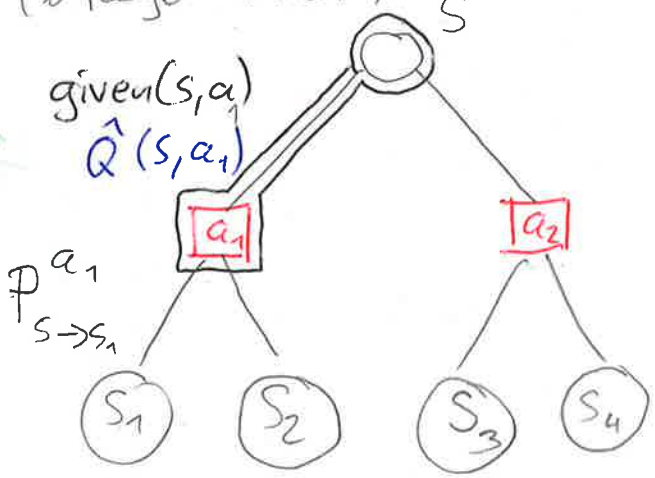
(linear)

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

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

possible futures with correct statistical weight ("large batch")

given s and a



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

→ drop Expectation sign

with hypothesis H :

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

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

(* Def Q)

⇒ estimate $\hat{Q}$ equal "exact Q"