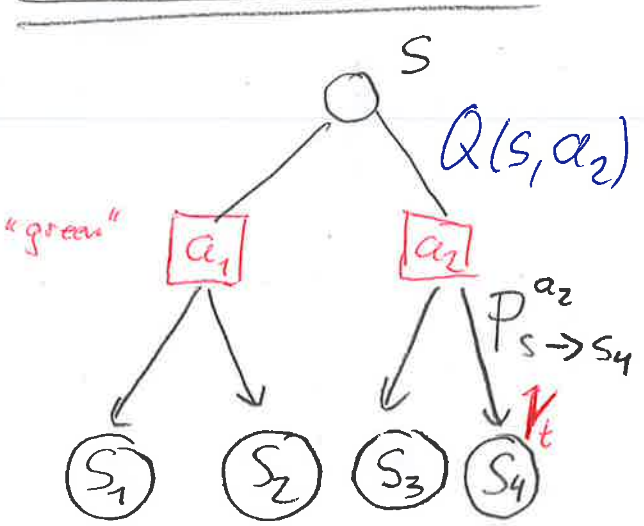
Blackboard 8.1 : Q - values



"branching ratio"

Transition probability

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

next state

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

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

reward received

end up in s_4

take a_2

start in s

- expected reward for action a_2

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

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

all possible "next states"

Blackboard 8.2 = Exercise 1

Q = expected reward $\approx$ empirical mean r. = $\hat{Q}$

$\hat{Q}^{(k-1)}(s,a)$ after $k-1$ trials (playing action a)

$$\hat{Q}^{(k-1)}(s,a) = \frac{1}{k-1}\left(r_1 + r_2 + \dots r_{k-1}\right)$$

↑ 2nd time action a

---

after k trials

$$\hat{Q}^{(k)}(s,a) = \frac{1}{k}\left(\underbrace{r_1 + r_2 + \dots r_{k-1}}_{} + r_k\right)$$

$$= \frac{k-1}{k} \cdot \hat{Q}^{(k-1)}(s,a) + \frac{1}{k} r_k$$

$$= \cancel{\frac{k}{k}}\, \hat{Q}^{(k-1)}(s,a) + \frac{1}{k} r_k - \frac{1}{k}\hat{Q}^{(k-1)}(s,a)$$

$$\Delta \hat{Q}(s,a) = \hat{Q}^{(k)}(s,a) - \hat{Q}^{(k-1)}(s,a) = \frac{1}{k}\left[r_k - \hat{Q}^{(k-1)}\right]$$

$$\Longrightarrow \boxed{\eta = \frac{1}{k}}$$

Blackboard 8.3 Convergence in expectation

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

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

---

proof:

$$\underbrace{E[\Delta Q(s,a)]}_{\text{fluctuates around zero}} \overset{(H)}{=} 0 \overset{\text{Eq.(1)}}{=} E[r_t - Q(s,a)]$$

$$0 = E[r_t] - E[Q(s,a)]$$

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

∎

(ii) Fluctuations: role of η is qualitatively obvious.

Blackboard 8.4 – Exercise 2

update with $\Delta Q(s,a) = 0.2 \cdot [r_t - Q(s,a)]$ (*)

2.1. initialise $Q(s,a_1) = Q(s,a_2) = 0$

$t=1$, action a_1; $r_t = 1 \Rightarrow Q(s,a_1) = 0.2$

$t=2$, action a_2; $r_t = 0.4 \Rightarrow \boxed{Q(s,a_2) = 0.08}$

2.2. $t=3$, best action $= a_1$; $r_t = 0$

$Q(s,a_1) \leftarrow Q(s,a_1) + 0.2[0 - 0.2]$; $Q(s,a_1) = 0.16$

$t=4$, best action a_1; $r_t = 0$

$Q(s,a_1) \leftarrow Q(s,a_1) + 0.2[0 - 0.16]$

$0.16 - 0.032$ $\quad Q(s,a_1) = 0.128$

$t=5$, best action a_1; $r_t = 0$

$Q(s,a_1) \leftarrow 0.128 - 0.2 \cdot 0.128$; $Q(s,a) \cong 0.102.$

$\Rightarrow$ a_1 remains "best action" for several steps!

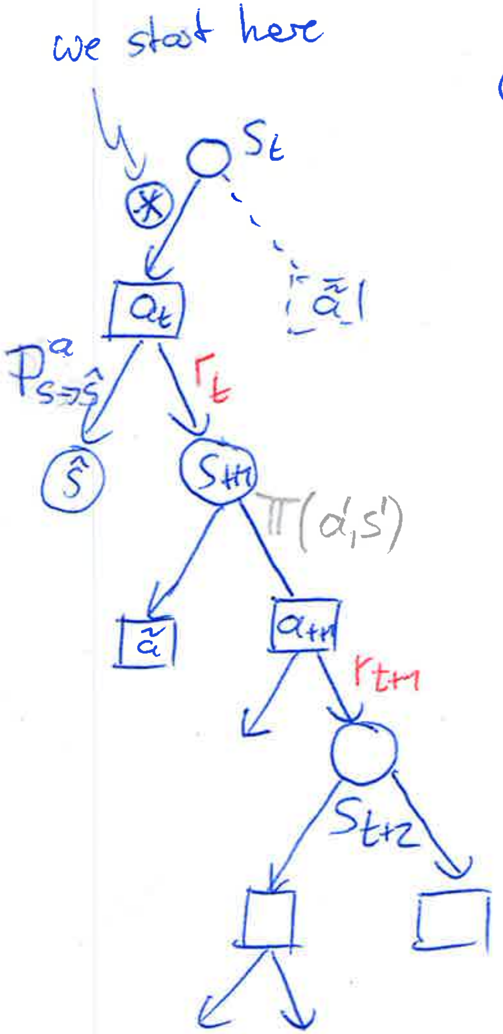
2.3 actual values

$$\left.\begin{aligned} Q(s,a_1) &= 0.25 \\ Q(s,a_2) &= 0.30 \end{aligned}\right\} \Rightarrow a_2 \text{ is best action}$$

Blackboard 8.5

Bellman equation

week 8 /



⊛ total reward collected in single trial starting in s with action a_t

$$R(s_t,a_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \dots$$

$$= r_t + \gamma \left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots \right]$$

$$= r_t + \gamma \cdot R(s_{t+1}, a_{t+1})$$

total reward (single trial) starting from $s' = s_{t+1}$ with a_{t+1}

now we look at diagram to calculate expectation

$$E(R(s_t,a_t)) = E(r_t + \gamma R(s_{t+1},a_{t+1}))$$

$$= \sum_{s'} P^{a_t}_{s\to s'} \left[R^{a_t}_{s\to s'} + \gamma E(R|s') \right]$$

↑ starting in s'

$$= \sum_{s'} P^{a_t}_{s\to s'} \left[R^{a_t}_{s\to s'} + \gamma \cdot \sum_{a'} \pi(a',s') \underbrace{E[R(s',a')]}_{Q(s',a')} \right]$$

$$Q(s_t,a_t) = \sum_{s'} P^{a_t}_{s\to s'} \left[R^{a_t}_{s\to s'} + \gamma \sum_{a'} \pi(a',s')\, Q(s',a') \right]$$

Blackboard 8.6 — SARSA

from diagram

$$Q(s,a) \approx r_t + \overset{\text{discount}}{\gamma} \cdot Q(s',a')$$

$$0 \approx r_t + \gamma \cdot Q(s',a') - Q(s,a)$$

proposed update

$$\Delta Q(s,a) = \eta \left[r_t + \gamma \cdot Q(s',a') - Q(s,a) \right]$$

check:

if $\underset{\substack{\uparrow \\ \text{actual reward}}}{r_t} > \underbrace{\gamma Q(s',a') - Q(s,a)}_{\text{expected reward for this transition}} \Rightarrow$ increase $Q(s,a)$

Blackboard 8.7 – Convergence

SARSA update

$$\Delta Q(s,a) = \eta \left[r_t + \gamma\, Q(s',a') - Q(s,a) \right]$$

hypothesis

$$E\left[\Delta Q(s,a)\right] \overset{\downarrow}{=} 0 = \underset{\uparrow}{E}\left[r_t + \gamma\, Q(s',a') - Q(s,a) \right]$$

starting in s with a

$$0 = \sum_{s'} P^{a}_{s \to s'} \left[R^{a}_{s \to s'} + \gamma \sum_{a'} \pi(s',a')\, Q(s',a') \right] - Q(s,a)$$

$\Rightarrow$ Bellman ✓

in order to evaluate expectations:

- look at graph!
- if I am in s, all remaining expectations are "given s"
- if I am in a branch (s, a) all remaing expectation are given s and a