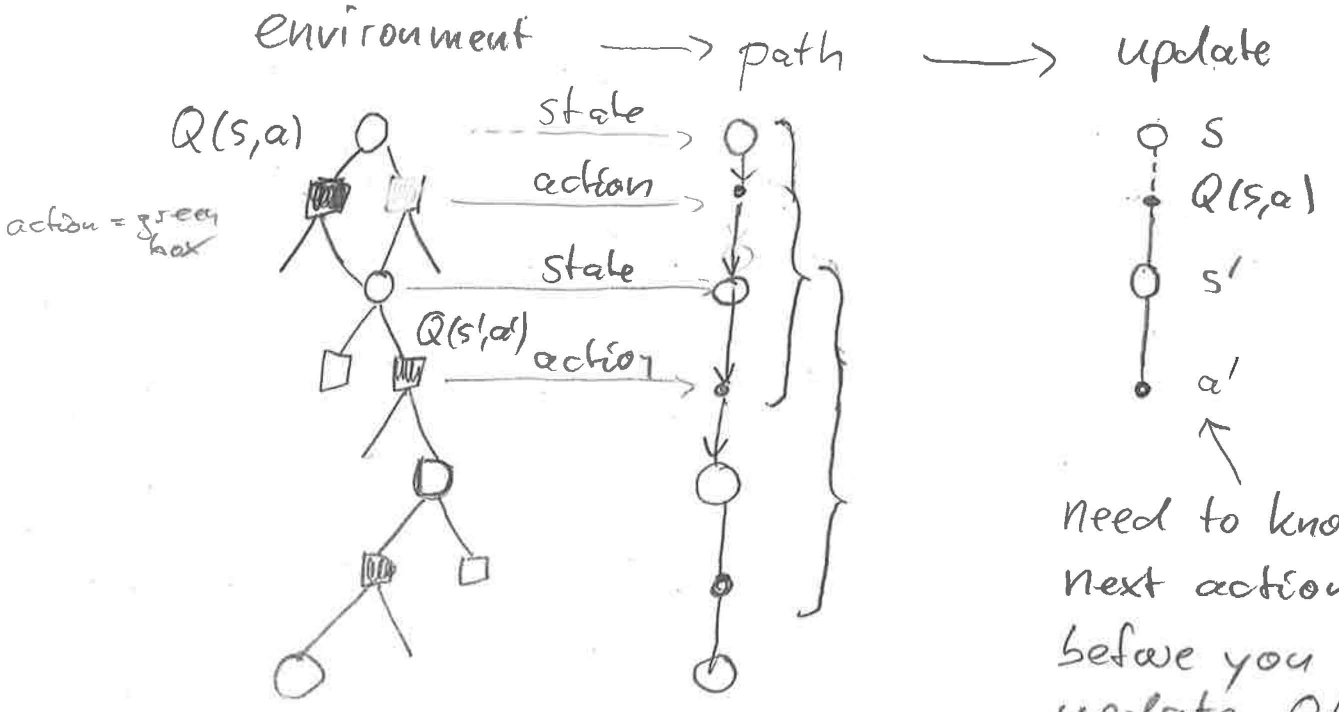
Week3 — RL2 ; Blackboard1 : Backup Diagr

(1)



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

earlier

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r_t + \gamma Q(s',a') - Q(s,a)]$$

earlier action

next action

Consistency of fluctuating online SARSA with Bellman equation [from RL 1]

right Blackboard

hypothesis: $0 = \langle \Delta \hat{Q}(s,a) \rangle = \eta \langle r_t + \gamma \hat{Q}(s',a') - \hat{Q}(s,a) \rangle$

cut

shift

$$\langle \hat{Q}(s,a) \rangle = \langle r_t + \gamma \cdot \hat{Q}(s',a') \rangle$$

fluctuates — fluctuates — temporal average over many "trials" $(N \to \infty)$

$$\langle \hat{Q}(s,a) \rangle = \sum_{s'} P^a_{s \to s'} \Big[R^a_{s \to s'} + \gamma \sum_{a'} \langle \pi^Q(s',a') \cdot \hat{Q}(s',a') \rangle$$

Problem: π^Q depends on $\hat{Q}$: $\pi(s',a' | \hat{Q}(s',a'))$
$\to$ slide (2-2)

Solution: ① if η is small, the fluctuations of $\hat{Q}$ are small and fluctuations of policy π^Q are "even smaller"*

② consider π^Q fixed for small enough Q
$\Rightarrow$ move π out: $\langle \pi^Q Q \rangle \rightsquigarrow \pi^Q \langle Q \rangle$

$$\langle \hat{Q}(s,a) \rangle = \sum_{s'} P^a_{s \to s'} \Big[R^a_{s \to s'} + \gamma \sum_{a'} \pi^Q(s',a') \langle \hat{Q}(s',a') \rangle \Big]$$

$\langle \hat{Q}(s,a) \rangle = Q(s,a)$ solves Bellman equation $\hat{Q}(s',a') = Q(s',a')$

remarks

① *example of π^Q "even smaller": ε-greedy
only rank-order of Q-values matters: best / 2nd best / ...
if fluctuations $|\Delta \hat{Q}(s',a')| \ll Q(s',\text{best}) - Q(s', 2^{\text{nd}}\text{best})$
then π^Q remains stable!

② evaluation of averages
- look at graph
- if in state s' all remaining averages are "given s'"

[see also slide]

$$(1) \quad \hat{Q}_t(s',a') = \langle \hat{Q}_t(s',a') \rangle_t + \delta Q_t(s',a')$$

$$\downarrow \text{defin} \qquad \downarrow \text{def}$$

$$= \bar{q} + \delta q$$

simplified notation

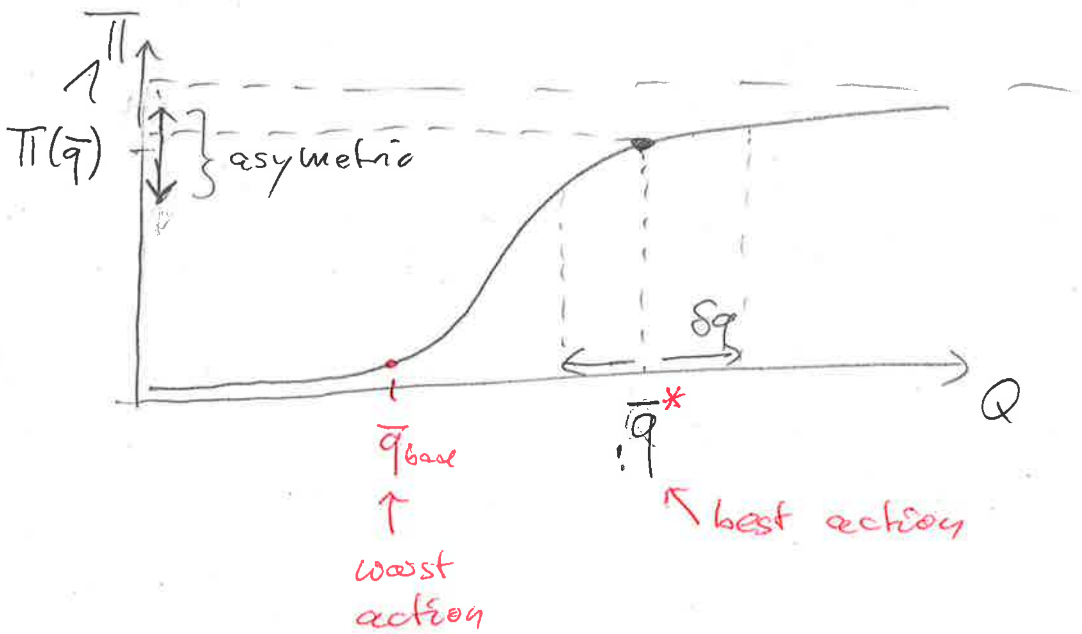
$$\bar{q} := \langle \hat{Q}_t(s',a') \rangle_t$$

$$\delta q := \delta Q_t(s',a') \quad \text{with} \quad \langle \delta q \rangle = 0$$

simplified notation

$$\pi'(s',a' \,|\, \hat{Q}_t(s',a')) = \pi(\hat{Q}_t) \text{ e.g. softmax}$$

$$(2) \quad \pi(s',a' \,|\, \hat{Q}_t(s',a')) = \pi(\bar{q}) + \frac{\partial \pi}{\partial \hat{Q}_t} \cdot \delta q + \frac{1}{2} \frac{\partial \pi}{\partial \hat{Q}_t} \cdot (\delta q)^2$$



center blackboard

We need to evaluate product $\Pi \cdot Q$

$$\Pi(s,a') \hat{Q}(s',a') \cdot \hat{Q}(s',a')$$

$$\stackrel{(1)+(2)}{=} \left[\Pi(\bar{q}) + \delta q \cdot \frac{\partial \Pi}{\partial Q} + \frac{1}{2}(\delta q)^2 \frac{\partial^2 \Pi}{\partial \hat{Q}^2}\right] \cdot \left[\bar{q} + \delta q\right]$$

collect terms according to $(\delta q)^n$ ← power

$$= \langle \Pi(\bar{q}) \cdot \bar{q} \rangle$$

drop $\langle \; \rangle$

$$+ \langle \delta q \rangle \left[\bar{q} \cdot \frac{\partial \Pi}{\partial \hat{Q}} + \Pi(\bar{q})\right]$$

vanishes since $\langle \delta q \rangle = 0$

$$+ \langle (\delta q)^2 \rangle \left[\bar{q} \cdot \frac{1}{2} \frac{\partial^2 \Pi}{\partial \hat{Q}^2} + \frac{\partial \Pi}{\partial \hat{Q}}\right]$$

"bias term"

now we average $\langle \; \rangle$:

$$\langle \Pi(\hat{Q}_t) \cdot \hat{Q}_t \rangle = \Pi(\bar{q}) \cdot \bar{q} + \underbrace{\langle (\delta q)^2 \rangle}_{>0} \left[\bar{q} \, \underbrace{\frac{1}{2} \frac{\partial^2 \Pi}{\partial \hat{Q}^2}}_{\text{unclear}} + \underbrace{\frac{\partial \Pi}{\partial \hat{Q}}}_{\geq 0}\right]$$

>0 for positive reward schemes

unclear: >0 for "bad" action; <0 for "best" action

term vanishes for $\eta \to 0$

for finite η: weighted Q-value of "bad" actions is overrated ($\to$ bias)

but for $\eta \to 0$ we have $\delta q \to 0$