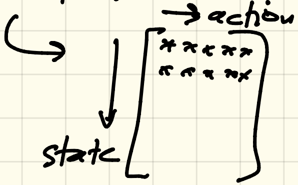
Q-learning with function approximation

tabular case $\longrightarrow$ $Q(s,a)$ $\forall s \in S$, $\forall a \in A$



$Q(s,a;w)$

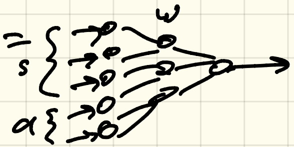
parameters (weights)

e.g. basis functions $f_1(s,a), \ldots, f_n(s,a)$

$$Q(s,a;w) = \sum_{k=1}^{n} w_k f_k(s,a)$$

$n \ll$ number of states

e.g. $Q(s,a;w) =$



$\rightarrow Q(s,a)$

Given w^- $\longrightarrow$ update to new w

$$L(w) = \frac{1}{2}\Big(\underbrace{r_{t+1} + \gamma \max_a Q(s_{t+1}, a; w^-)}_{\text{target}} - Q(s_t, a_t, w)\Big)^2$$

minimize $L(w)$: $\quad w \leftarrow w^- - \alpha \nabla_w L(w^-)$

$$\nabla_w L(w) = -\Big(\underbrace{r_{t+1} + \gamma \max_a Q(s_{t+1}, a; w^-) - Q(s_t, a_t; w)}_{\delta_t}\Big) \nabla_w Q(s_t, a_t; w)$$

$$= -\delta_t \nabla_w Q(s_t, a_t; w)$$

update rule : $\quad w \leftarrow w^- + \alpha \delta_t \nabla_w Q(s_t, a_t; w^-)$

$\delta_t(s_t, a_t, r_{t+1}, s_{t+1}; w^-)$

$\longrightarrow$ doesn't work due to correlated updates with global dependence on w

Mnih 2015 : experience replay → update for a collection of historical states.

$$L(w) = \mathbb{E}_{(s,a,r,s') \sim \underbrace{D}_{\text{history}}}\left[\frac{1}{2}\left(\underbrace{r + \gamma \max_a Q(s',a;w^-)}_{\text{target}} - Q(s,a;w)\right)^2\right]$$

$$= \sum_{(s,a,r,s') \in D} \frac{1}{N}\left[\left(\_\_\_\right)^2\right]$$

$$\nabla_w L(w) = -\mathbb{E}_{(s,a,r,s') \sim D}\left[\delta(s,a,r,s';w^-,w)\, \nabla_w Q(s,a;w)\right]$$

$$w \leftarrow w^- - \alpha \nabla_w L(w^-)$$

samples in history D were not generated with current Q, but with old Q policies → need off-policy Q-learning