TRAINING

initialize REPLAY-MEMORY(*capacity N*) D

initialize ACTION-VALUE-FUNCTION(*random weight* θ) Q

initialize TARGET-ACTION-VALUE-FUNCTION($\theta^- = \theta$) $\hat{Q}$

for *episode* $= 1$ **to** M

initialize STATE S, TEMP-BUFFER G

initialize COLLABORTOR-PROBABILITY-VECTOR(*Uniform*) p

for *step* $= 1$ **to** T

$\tilde{r}$ = discount sum of reward **from** S to S'

$\tilde{\gamma} = \gamma^{step}$

$\tilde{S} =$ CONCATENATE(S, p)

$\tilde{S'} =$ CONCATENATE(S', p)

store TRANSITION$(\tilde{S}, g, \tilde{r}, \tilde{S'})$ **in** D

$$a_t = \begin{cases} \textit{random action} & \text{with probality } \epsilon \\ \underset{a}{\operatorname{argmax}}\, Q_\theta(s, g) & \text{otherwise} \end{cases}$$

execute a_t, **observe** REWARD r_t, $S =$ NEXT-STATE S'

update p **with** normalized-BEHAVIOR-VECTOR

sample MINIBATCH $< \tilde{S}, g, \tilde{r}, \tilde{S'}, \tilde{\gamma} >$ **from** D

$$y = \tilde{r} + \tilde{\gamma} \cdot \max Q_\theta \left(\tilde{S'}, \underset{g}{\operatorname{argmax}}\, Q_{\theta^-}(\tilde{S}, g) \right)$$

perform SGD **on** $[y - Q_\theta(\tilde{S}, g)]^2$ with respect to θ

reset $\hat{Q} = Q$ every C steps

resample from p every C' steps