1.Collecting-Data

initialize Replay-Buffer D

initialize Target-Action-Value-Function($\theta^- = \theta$) $\hat{Q}$

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

initialize Temp-Buffer G, State S, Behavioral-Policy μ

for *step* $= 1$ **to** *episode*

$r, S' =$ transition(S, μ)

store S **in** G

Goal $g = S'$

update p **with** normalized-Behavior-Vector

for *state* **in** G

$\tilde{r} =$ discount sum of reward **from** *state* **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

change Behavioral-Policy μ every C' steps

2.Training

initialize Action-Value-Function(θ) Q

initialize Target-Action-Value-Function($\theta^- = \theta$) $\hat{Q}$

for $t = 1$ **to** T

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