Offline Algorithm

```
1.Collecting-Data
    initialize Replay-Buffer $D$
    for $episode$ = 1 to $M$
        initialize Temp-Buffer $G$,  State $S$,  Behavioral-Policy $\mu$
        for $step$ = 1 to $episode$
            $r, S'$ = transition($S, \mu$)
            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 $\mu$ every $C'$ steps
        resample from $p$ every $C'$ steps
2.Training
    initialize Action-Value-Function($\theta$) $Q$
    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 $\theta$
        reset $\hat{Q} = Q$ every $C$ steps
```

Online Algorithm

```
Training
    initialize Replay-Memory($capacity\ N$) $D$
    initialize Action-Value-Function($random\ weight\ \theta$) $Q$
    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 $D$
            $a_t = \begin{cases} random\ action & \text{with probality } \epsilon \\ \underset{a}{\operatorname{argmax}} Q_\theta(s, g) & \text{otherwise} \end{cases}$
            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 $\theta$
            reset $\hat{Q} = Q$ every $C$ steps
        resample from $p$ every $C'$ steps
```