```
for $m = 1 \ldots \mathit{MAX\_STEPS}$ do
    Gather and store $h$ experiences $(s_i, a_i, r_i, s'_i)$ using the current policy
    for $b = 1 \ldots B$ do
        Sample a batch, $b$, of experiences from the experience replay memory
        for $u = 1 \ldots U$ do
            for $i = 1 \ldots N$ do
                # Calculate target $Q$-values for each example
                $y_i = r_i + \delta_{s'_i} \gamma Q^{\pi_\varphi}(s'_i, \max_{a'_i} Q^{\pi_\theta}(s'_i, a'_i))$ where $\delta_{s'_i} = 0$ if $s'_i$ is terminal, 1 otherwise
            end for
            # Calculate the loss, for example using MSE
            $L(\theta) = \frac{1}{N} \sum_i (y_i - Q^{\pi_\theta}(s_i, a_i))^2$
            # Update the network's parameters
            $\theta = \theta - \alpha \nabla_\theta L(\theta)$
        end for
    end for
    Decay $\tau$
    if $(m \mod F) == 0$ then
        # Update the target network
        $\varphi = \theta$
    end if
end for
```