Algorithm 1 Deep Q-Learning with Experience Replay

Initialize replay memory $\mathcal{D}$ to capacity N
Initialize action-value function Q with two random sets of weights θ, θ'
for *episode* $= 1, M$ **do**
  for $t = 1, T$ **do**
    Select a random action a_t with probability ε.
    Otherwise, select $a_t = \arg\max_a Q(s_t, a; \theta)$
    Execute action a_t, collect reward r_{t+1} and observe next state s_{t+1}
    Store the transition $(s_t, a_t, r_{t+1}, s_{t+1})$ in $\mathcal{D}$
    Sample mini-batch of transitions $(s_j, a_j, r_{j+1}, s_{j+1})$ from $\mathcal{D}$
    Set $y_j = \begin{cases} r_{j+1}, & \text{if } s_{j+1} \text{ is terminal} \\ r_{j+1} + \gamma \max_{a'} Q(s_{j+1}, a'; \theta'), & \text{otherwise} \end{cases}$
    Perform a gradient descent step using targets y_j with respect to the online parameters θ
    Every C steps, set $\theta' \leftarrow \theta$
  end for
end for