**Algorithm 1** Deep Q-Learning with Experience Replay

Initialize replay memory 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 *at* with probability *ε*.

Otherwise, select *at* = argmax*aQ*(*st,a*;*θ*)

Execute action *at*, collect reward *rt*+1 and observe next state *st*+1

Store the transition (*st,at,rt*+1*,st*+1) in D

Sample mini-batch of transitions (*sj,aj,rj*+1*,sj*+1) from D

(

*rj*+1*,* if *sj*+1 is terminal

Set *yj* = *rj* + *γ* max ′ *Q*(*s ,a*′;*θ*′)*,* otherwise

+1 *a j*+1

Perform a gradient descent step using targets *yj* with respect to the online parameters *θ*

Every *C* steps, set *θ*′ ← *θ* **end for end for**

1