

Input: policy π , positive integer $num_episodes$, small positive fraction α , GLIE $\{\epsilon_i\}$

Output: value function Q ($\approx q_\pi$ if $num_episodes$ is large enough)

Initialize Q arbitrarily (e.g., $Q(s, a) = 0$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$, and $Q(\text{terminal-state}, \cdot) = 0$)

for $i \leftarrow 1$ **to** $num_episodes$ **do**

$\epsilon \leftarrow \epsilon_i$

 Observe S_0

 Choose action A_0 using policy derived from Q (e.g., ϵ -greedy)

$t \leftarrow 0$

repeat

 Take action A_t and observe R_{t+1}, S_{t+1}

 Choose action A_{t+1} using policy derived from Q (e.g., ϵ -greedy)

$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$

$t \leftarrow t + 1$

until S_t is terminal;

end

return Q