```
Initialize $Q_0(s,a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$ (e.g. all zero)
Set $k := 1$
Initialize sequence $(\alpha_l)_{l \in \mathbb{N}}$ with $\alpha_l \in [0,1]$
    **Repeat for each episode:**
        Initialize environment and get initial state $s_0 \in \mathcal{S}$
        Set $t := 0$
        **Repeat for each step of the episode (until terminal state):**
            Take action $a_t := \bar{\pi}_k(s_t)$
            ($\bar{\pi}_k$ is $\epsilon$-greedy policy wrt $Q_{k-1}$)
            Observe immediate reward $R_t$, observe new state $s_{t+1}$
            Set $Q_k := Q_{k-1}$
            Update $Q_k(s_t, a_t) = Q_{k-1}(s_t, a_t) + \alpha_k(R_t + Q_{k-1}(s_{t+1}, \pi_k(s_{t+1})) - Q_{k-1}(s_t, a_t))$
            ($\pi_k$ is greedy policy wrt $Q_{k-1}$)
            Increase $t$ by one, increase $k$ by one
```