Q-Learning Algorithm

```
Initialize $Q(s, a)$ to 0 for all $\alpha \in A$ in each $s \in S$
Initialize learning rate $\alpha \in (0, 1]$
Initialize discount factor $\gamma \in [0, 1]$
Initialize exploration rate $\epsilon \in [0, 1]$
while not converged do
    $s \leftarrow s_0$
    while $s$ not terminal do
        Observe current state $s$
        if $explore()$ then
            $a \leftarrow$ random action
        else
            $a \leftarrow \arg\max_a Q(s, a)$
        end if
        Take action $a$, observe reward $R(s, a)$ and next state $s'$
        Update Q-value:
            $$Q(s, a) \leftarrow (1 - a) \cdot Q(s, a) + \alpha \left[ R(s, a) + \gamma \cdot \max_a Q(s', a) \right]$$
        $s \leftarrow s'$
    end while
end while
```