Algorithm 1 Generic Q-learning (includes FQI and DQN as special cases

```
initialize $\phi_0$
initialize $\pi_0(\mathbf{a}|\mathbf{s}) = \epsilon\mathcal{U}(\mathbf{a}) + (1-\epsilon)\delta(\mathbf{a} = \arg\max_{\mathbf{a}} Q_{\phi_0}(\mathbf{s},\mathbf{a}))$ ▷ Use $\epsilon$-greedy exploration
initialize replay buffer $\mathcal{D} = \emptyset$ as a ring buffer of fixed size
initialize $\mathbf{s} \sim d_0(\mathbf{s})$
for iteration $k \in [0,\ldots,K]$ do
    for step $s \in [0,\ldots,S-1]$ do
        $\mathbf{a} \sim \pi_k(\mathbf{a}|\mathbf{s})$    ▷ sample action from exploration policy
        $\mathbf{s}' \sim p(\mathbf{s}'|\mathbf{s},\mathbf{a})$    ▷ sample next state from MDP
        $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s},\mathbf{a},\mathbf{s}',r(\mathbf{s},\mathbf{a}))\}$    ▷ append to buffer, purging old data if buffer too big
    end for
    $\phi_{k,0} \leftarrow \phi_k$
    for gradient step $g \in [0,\ldots,G-1]$ do
        sample batch $B \subset \mathcal{D}$    ▷ $B = \{(\mathbf{s}_i,\mathbf{a}_i,\mathbf{s}'_i,r_t)\}$
        estimate error $\mathcal{E}(B,\phi_{k,g}) = \sum_i \left(Q_{\phi_{k,g}} - (r_i + \gamma \max_{\mathbf{a}'} Q_{\phi_k}(\mathbf{s}',\mathbf{a}'))\right)^2$
        update parameters: $\phi_{k,g+1} \leftarrow \phi_{k,g} - \alpha\nabla_{\phi_{k,g}}\mathcal{E}(B,\phi_{k,g})$
    end for
    $\phi_{k+1} \leftarrow \phi_{k,G}$    ▷ update parameters
end for
```