Algorithm 1 Generic off-policy actor-critic

initialize ϕ_0
initialize θ_0
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**
        $\mathbf{a} \sim \pi_{\theta_k}(\mathbf{a}|\mathbf{s})$ ▷ sample action from current 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 $g \in [0, \ldots, G_Q - 1]$ **do**
        sample batch $B \subset \mathcal{D}$ ▷ $B = \{(\mathbf{s}_i, \mathbf{a}_i, \mathbf{s}'_i, r_t)\}$
        estimate error $\mathcal{E}(B, \phi_{k,g}) = \sum_i (Q_{\phi_{k,g}} - (r_i + \gamma \mathbb{E}_{\mathbf{a}' \sim \pi_k(\mathbf{a}'|\mathbf{s}')} Q_{\phi_k}(\mathbf{s}', \mathbf{a}')))^2$
        update parameters: $\phi_{k,g+1} \leftarrow \phi_{k,g} - \alpha_Q \nabla_{\phi_{k,g}} \mathcal{E}(B, \phi_{k,g})$
    end for
    $\phi_{k+1} \leftarrow \phi_{k,G_Q}$ ▷ update Q-function parameters
    $\theta_{k,0} \leftarrow \theta_k$
    for gradient step $g \in [0, \ldots G_\pi - 1]$ **do**
        sample batch of states $\{\mathbf{s}_i\}$ from $\mathcal{D}$
        for each $\mathbf{s}_i$, sample $\mathbf{a}_i \sim \pi_{\theta_{k,g}}(\mathbf{a}|\mathbf{s}_i)$ ▷ do not use actions in the buffer!
        for each $(\mathbf{s}_i, \mathbf{a}_i)$, compute $\hat{A}(\mathbf{s}_i, \mathbf{a}_i) = Q_{\phi_{k+1}}(\mathbf{s}_i, \mathbf{a}_i) - \mathbb{E}_{\mathbf{a} \sim \pi_{k,g}(\mathbf{a}|\mathbf{s}_i)}[Q_{\phi_{k+1}}(\mathbf{s}_i, \mathbf{a})]$
        $\nabla_{\theta_{k,g}} J(\pi_{\theta_{k,g}}) \approx \frac{1}{N} \nabla_{\theta_{k,g}} \log \pi_{\theta_{k,g}}(\mathbf{s}_i, \mathbf{a}_i) \hat{A}(\mathbf{s}_i, \mathbf{a}_i)$
        $\theta_{k,g+1} \leftarrow \theta_{k,g} + \alpha_\pi \nabla_{\theta_{k,g}} J(\pi_{\theta_{k,g}})$
    end for
    $\theta k + 1 \leftarrow \theta_{k,G_\pi}$ ▷ update policy parameters
end for