Algorithm 1: Optimistic Actor-Critic (OAC)

```
Input: $w_1, w_2, \theta$    // Initial parameters of the critics and target policy $\pi_{\mathcal{T}}$
Initialization: $\tilde{w}_1 \leftarrow w_1$, $\tilde{w}_2 \leftarrow w_2$, $\mathcal{D} \leftarrow \emptyset$    // Initialize target networks and replay pool
for each iteration do
    for each environment step do
        Sample action $a_t \sim \pi_E(a_t|s_t)$    // Exploration policy from Eq. (9);
        Sample transition $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$;
        Store transition: $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, R(s_t, a_t), s_{t+1})\}$;
    end
    for each training step do
        for $i \in \{1, 2\}$ do
            Update $w_i$ with:
                $\nabla_{w_i} \left\| \hat{Q}^i_{\text{LB}}(s_t, a_t) - R(s_t, a_t) - \gamma \min(\hat{Q}^1_{\text{LB}}(s_{t+1}, a), \hat{Q}^2_{\text{LB}}(s_{t+1}, a)) \right\|_2^2$
        end
        Update $\theta$ with:
            $\nabla_\theta \hat{J}^\alpha_{\hat{Q}_{\text{LB}}}$
        Update target critics:
            $\tilde{w}_1 \leftarrow \tau w_1 + (1 - \tau)\tilde{w}_1, \quad \tilde{w}_2 \leftarrow \tau w_2 + (1 - \tau)\tilde{w}_2$
    end
end
Output: $w_1, w_2, \theta$    // Optimized parameters
```