Input: initial policy parameters θ_1, θ_2, Q-function parameters ϕ_1, ϕ_2, empty replay buffer $\mathcal{D}$

Set target parameters equal to main parameters $\theta_{\text{targ},1} \leftarrow \theta_1$, $\theta_{\text{targ},2} \leftarrow \theta_2$, $\phi_{\text{targ},1} \leftarrow \phi_1$, $\phi_{\text{targ},2} \leftarrow \phi_2$

repeat

Observe state s and select actions for both players $a_1 = \text{clip}(\mu_{\theta_1}(s) + \epsilon_1, a_{Low}, a_{High})$ and $a_2 = \text{clip}(\mu_{\theta_2}(s) + \epsilon_2, a_{Low}, a_{High})$, where $\epsilon_1 \sim \mathcal{N}$ and $\epsilon_2 \sim \mathcal{N}$

Execute actions (a_1, a_2) in the environment.

Observe next state s', reward pair (r_1, r_2) for both players, and done signal d

Store $(s, a_1, a_2, r_1, r_2, s', d)$ in replay buffer $\mathcal{D}$

If s' is terminal, reset environment state.

if it's time to update **then**

for j in range (however many updates) **do**

Randomly sample a batch of transitions, $B = \{(s, a_1, a_2, r_1, r_2, s', d)\}$ from $\mathcal{D}$

Compute targets for both players:

$$y_1 = r_1 + \gamma(1-d)Q_{\phi_{\text{targ},1}}(s', \mu_{\theta_{\text{targ},1}}(s'), \mu_{\theta_{\text{targ},2}}(s'))$$

$$y_2 = r_2 + \gamma(1-d)Q_{\phi_{\text{targ},2}}(s', \mu_{\theta_{\text{targ},1}}(s'), \mu_{\theta_{\text{targ},2}}(s'))$$

Update Q-functions for both players by one step of gradient descent:

$$\nabla_{\phi_1} \frac{1}{|B|} \sum_{B} \left(Q_{\phi_1}(s, a_1, a_2) - y_1(r_1, s', d)\right)^2$$

$$\nabla_{\phi_2} \frac{1}{|B|} \sum_{B} \left(Q_{\phi_2}(s, a_1, a_2) - y_2(r_2, s', d)\right)^2$$

Update policies for both players by one step of gradient ascent:

$$\nabla_{\theta_1} \frac{1}{|B|} \sum_{s \in B} Q_{\phi_1}(s, \mu_{\theta_1}(s), \mu_{\theta_2}(s))$$

$$\nabla_{\theta_2} \frac{1}{|B|} \sum_{s \in B} Q_{\phi_2}(s, \mu_{\theta_1}(s), \mu_{\theta_2}(s))$$

Update target networks for both players:

$$\begin{aligned}
\phi_{\text{targ},1} &\leftarrow \rho\phi_{\text{targ},1} + (1-\rho)\phi_1 \\
\phi_{\text{targ},2} &\leftarrow \rho\phi_{\text{targ},2} + (1-\rho)\phi_2 \\
\theta_{\text{targ},1} &\leftarrow \rho\theta_{\text{targ},1} + (1-\rho)\theta_1 \\
\theta_{\text{targ},2} &\leftarrow \rho\theta_{\text{targ},2} + (1-\rho)\theta_2
\end{aligned}$$

end for

end if

until convergence