Algorithm: COMA

Initialize: θ the parameters of individual policies $pi_i(.;\theta)$, ϕ the parameters of the centralized Q-network $Q(;\phi)$, and ϕ^- the parameters the target Q network.
while $t < T$ **do**
    Initialize a rollout buffer $\mathcal{D}$ // $(episode_1, episode_2, \dots)$
    for a number of episodes **do**:
        $current\_episode = \{\}$
        while $\mathbf{o}_t$ is not *done* **do**
            Collect observations $\{o_1^t, \dots, o_n^t\}$ and state $\mathbf{s}_t$
            Sample an action $a_i^t \sim \pi_i(.|o_i^t)$ for each agent i
            Execute joint action $\mathbf{a}^t = (a_1^t, \dots, a_n^t)$
            Collect r^t, $done^t$
            Store $(\mathbf{o}^t, \mathbf{a}^t, r^t, done^t, \mathbf{o}^{t+1})$ in $current\_episode$
        end while
        Store $current\_episode$ in the rollout buffer $\mathcal{D}$
    end for
    Process the rollout buffer for batch training // Episodes with different lengths
    Train the centralized critic using TD(λ)

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{B}|} \sum_b \frac{1}{L^b} \sum_t \left(y^{t,b} - Q^{tot}(\mathbf{s}^{t,b}, \mathbf{o}^{t,b}, \mathbf{a}^{t,b}; \phi) \right)^2$$

    Every C training steps, update $\phi^- \leftarrow \phi$ //training step = one full buffer pass
    Compute the counterfactual advantages

$$A_i(\mathbf{s}, \mathbf{o}, \mathbf{a}) = Q(\mathbf{s}, \mathbf{o}, \mathbf{a}; \phi) - \sum_{a_i'} \pi_i(a_i'|o_i; \theta) Q(\mathbf{s}, \mathbf{o}, (\mathbf{a}_{-i}, a_i'); \phi)$$

    Perform a gradient descent using:

$$-\sum_i A_i(\mathbf{s}, \mathbf{o}, \mathbf{a}) \log(\pi(a_i, o_i; \theta))$$

end while