
Algorithm: COMA

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

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

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

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

- 19: Perform a gradient descent using:

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

- 20: **end while**
-