
Algorithm: FACMAC

- 1: Initialize: θ and θ^- the parameters of individual policy and target policy μ_i, ϕ and ϕ^- the parameters of individual and target Q-networks Q_i, ψ and ψ^- the parameters of the mixing and target-mixing network m .
- 2: Initialize replay buffer \mathcal{D} // ($episode_1, episode_2, \dots$)
- 3: **while** $t < T$ **do**
- 4: $current_episode = \{\}$
- 5: **while** \mathbf{o}_t is not *done* **do**
- 6: Collect observations $\{o_1^t, \dots, o_n^t\}$ and state \mathbf{s}_t
- 7: Select an action $a_i^t = \mu(o_i^t)$ for each agent i .
- 8: Execute the joint action $\mathbf{a}^t = (a_1^t, \dots, a_n^t)$
- 9: Collect $r^t, done^t$
- 10: Store $(\mathbf{s}_t, \mathbf{o}^t, \mathbf{a}^t, r^t, done^t)$ in $current_episode$
- 11: **end while**
- 12: Store $current_episode$ in the replay buffer \mathcal{D}
- 13: **if** t is a training step **then**
- 14: Sample batch of episodes

$$\mathcal{B} = \{\{\mathbf{s}^{t,b}, \mathbf{o}^{t,b}, \mathbf{a}^{t,b}, r^{t,b}, done^{t,b}, \mathbf{s}'^{t,b}, \mathbf{o}'^{t,b}\}_{t=1 \dots L^b}\}_{b=1, \dots, |\mathcal{B}|}$$

- 15: Set the targets

$$y^{t,b} = r^{t,b} + \gamma(1 - done^{t,b}) \times \max_{(a_1, \dots, a_n)} m(\mathbf{s}'^{t,b}, Q_1(\mathbf{o}_1^{t,b}, a_1; \phi^-), \dots, Q_n(\mathbf{o}_n^{t,b}, a_n; \phi^-); \psi^-)$$

- 16: Update ϕ and ψ using :

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

- 17: Update θ using:

$$\mathcal{L}(\theta) = -\frac{1}{|\mathcal{B}|} \sum_b \frac{1}{L^b} \sum_t \sum_i Q^{tot}(\mathbf{s}^{t,b}, \mu(o_1^{t,b}; \theta), \dots, \mu(o_i^{t,b}; \theta), \dots, \mu(o_n^{t,b}; \theta))$$

- 18: Every C steps, update $\theta^- \leftarrow \theta, \phi^- \leftarrow \phi, \psi^- \leftarrow \psi$
 - 19: **end if**
 - 20: **end while**
-