
Algorithm: MAPPO

- 1: Initialize: θ the parameters of individual actors $\pi_i(\cdot; \theta)$, ϕ the parameters of centralized critic $V(\cdot; \phi)$
- 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 the 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 $r^t, done^t$
- 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: Compute the advantages A^t and TD targets y
- 17: Compute the actor losses:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{B}|} \sum_b \frac{1}{L^b} \sum_t \frac{1}{n} \sum_i \min \left(\frac{\pi(a_i^t | o_i^t; \theta)}{\pi(a_i^t | o_i^t; \theta_{old})} A^t, \right. \\ \left. \text{clip} \left(\frac{\pi(a_i^t | o_i^t; \theta)}{\pi(a_i^t | o_i^t; \theta_{old})}, 1 - \varepsilon, 1 + \varepsilon \right) A^t \right)$$

- 18: Compute the entropy bonus

$$\mathcal{H}(\theta) = \frac{1}{|\mathcal{B}|} \sum_b \frac{1}{L^b} \sum_t \frac{1}{n} \sum_i \mathcal{H}_i(\theta)$$

- 19: Compute the critic loss:

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

- 20: Update θ and ϕ using:

$$\mathcal{L}(\theta, \phi) = -\mathcal{L}(\theta) + \alpha^{critic} \mathcal{L}(\phi) - \alpha^{entropy} \mathcal{H}(\theta)$$

- 21: **end while**
-