

---

**Algorithm: MADDPG**

---

```
1: Initialize:  $\theta$  and  $\theta^-$  the parameters of individual policy and target policy  $\mu_i$ ,  $\phi$ 
   and  $\phi^-$  the parameters central Q-function  $Q$ .
2: Initialize replay buffer  $\mathcal{D}$  // (episode1, episode2, ...)
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  $(\mathbf{s}_t, \mathbf{o}^t, \mathbf{a}^t, r^t, \text{done}^t)$  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
```

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

```
15:   Set the targets
```

$$y^{t,b} = r^{t,b} + (1 - \text{done}^{t,b})Q(\mathbf{s}'^{t,b}, \mu(o_1^{t,b}; \theta^-), \dots, \mu(o_n^{t,b}; \theta^-); \phi^-)$$

```
16:   Update  $\phi$  using :
```

$$\mathcal{L}(\phi) = \frac{1}{|\mathcal{B}|} \sum_b \frac{1}{L^b} \sum_t \left( y^{t,b} - Q(\mathbf{s}^{t,b}, a_1^{t,b}, \dots, a_n^{t,b}; \phi) \right)^2$$

```
17:   Update  $\theta$  using:
```

$$\mathcal{L}(\theta) = -\frac{1}{|\mathcal{B}|} \sum_b \frac{1}{L^b} \sum_t \sum_i Q(\mathbf{s}^{t,b}, a_1^{t,b}, \dots, \mu(o_i^{t,b}; \theta), \dots, a_n^{t,b}; \phi)$$

```
18:   Every  $C$  steps, update  $\theta^- \leftarrow \theta$ ,  $\phi^- \leftarrow \phi$ 
```

```
19:   end if
```

```
20: end while
```

---