TME 11 : MADDPG

Algorithm 1: Multi-Agent Deep Deterministic Policy Gradient for N agents

---

for episode $= 1$ to M **do**
  Initialize a random process $\mathcal{N}$ for action exploration
  Receive initial state $\mathbf{x}$
  for $t = 1$ to max-episode-length **do**
    for each agent i, select action $a_i = \boldsymbol{\mu}_{\theta_i}(o_i) + \mathcal{N}_t$ w.r.t. the current policy and exploration
    Execute actions $a = \boxed{(a_1, \ldots, a_N)}$ and observe reward r and new state $\mathbf{x}'$
    Store $(\mathbf{x}, a, r, \mathbf{x}')$ in ~~replay buffer~~ $\mathcal{D}$
    $\mathbf{x} \leftarrow \mathbf{x}'$
    for agent $i = 1$ to N **do**
      Sample a random minibatch of S samples $(\mathbf{x}^j, a^j, r^j, \mathbf{x}'^j)$ from $\mathcal{D}$

policy target

      Set $y^j = r_i^j + \gamma\, Q_i^{\boldsymbol{\mu}'}(\mathbf{x}'^j, \boxed{a_1', \ldots, a_N'})|_{a_k' = \boldsymbol{\mu}_k'(o_k^j)}$
      Update critic by minimizing the loss $\mathcal{L}(\theta_i) = \frac{1}{S}\sum_j \left(y^j - Q_i^{\boldsymbol{\mu}}(\mathbf{x}^j, a_1^j, \ldots, a_N^j)\right)^2$
      Update actor using the sampled policy gradient:

$$\nabla_{\theta_i} J \approx \frac{1}{S}\sum_j \nabla_{\theta_i}\boldsymbol{\mu}_i(o_i^j)\nabla_{a_i} Q_i^{\boldsymbol{\mu}}(\mathbf{x}^j, a_1^j, \ldots, a_i, \ldots, a_N^j)\big|_{a_i = \boldsymbol{\mu}_i(o_i^j)}$$

x_i^j

    end for
    Update target network parameters for each agent i:

$$\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i'$$

  end for
end for

⚠ ⚠ ne fair que si DDPG déjà fait

• Actions / observations / reward
⇒ autant que d'agent
▷ env.n = # agents
▷ actions $\in \mathbb{R}^2$
▷ Taille de chaque espace d'observation :
obs = env.reset()

• μ (target) μ'
Q (target) $Q^{\mu'}$

⚠ Cas DDPG → pour les adversaires (env. 2/3)

DDPG

$\updownarrow$

MADDPG

$$\nabla_\theta \mu(o_j)\, \nabla_a Q(o_j, a)\big|_{a=\mu(o_j)} = \underbrace{\nabla_\theta\, Q(o_j, \mu(o_j))}$$

th. de derivatio des fonctions composés

Permet de calculer facilement

$$\nabla_{\theta_i} \mu_i(x_i^j)\, \nabla_{a_i} Q_i^{\mu}(x^j, a_1^j, \dots, a_N^j)\big|_{a_i = \mu_i(x_i^j)}$$

$$= \nabla_{\theta_i} \quad \dots$$