Untitled

Somebody

November 2024

1 Pseudo Code

Algorithm 1 DDPG For PRM Traning

1: **Initialize:** CriticNet: $Q_\omega(s,a)$, ActionNet: $\mu_\theta(s)$
2: **Initialize:** $Q_{\omega^-}(s,a) \leftarrow Q_\omega(s,a)$, $\mu_{\theta^-}(s) \leftarrow \mu_\theta(s)$
3: **for** $e = 0$ **to** E **do**
4:    Initialize initial state s_1 (Randomly pick a question from PRM800k)
5:    **for** $t = 1$ **to** T **and** $done == True$ **do**
6:        Choose an action $a_t \sim \mu_\theta(s_t)$
7:        $r_t \leftarrow Env(s_t, a_t)$; $s_{t+1} \leftarrow [s_t, a_t]$
8:        **if** Buffer is big enough **then**
9:            Randomly pick N touples $\{(s_i, a_i, r_i, s_{i+1})\}_{i=1,...,N}$
10:           Sample K actions: $a_{i+1}^m \sim \mu_{\theta^-}(s_{i+1})$, $(m = 1, ..., k)$
11:           Calculate for every tuples: $y_i = r_i + \gamma * \max_{m \in \{1,...,k\}} Q_{\omega^-}(s_{i+1}, a_{i+1}^m)$

12:           $L = \frac{1}{N}\sum_{i=1}^{N} y_i - Q_\omega(s_i, a_i)$ # Lossfunc for criticnet
13:           $J = \frac{1}{N}\sum_{i=1}^{N} Q_\omega(s_i, \mu_\theta(s_i))$ # Lossfunc for actornet
14:           **Update for Critic and Actor Network**
15:           **Soft update target Network**
16:           $\omega^- \leftarrow \tau\omega + (1-\tau)\omega$, $\theta^- \leftarrow \tau\theta + (1-\tau)\theta$
17:       **end if**
18:    **end for**
19: **end for**

---

2 Problems