April 19, 2019 Policy gradient HRL

g_i: Subgoal @ local i, Top agent considers local i subgoal (sub agent) as 'action'.

s_i: State @ local i

θ_i: policy parameter for local i

$\pi_i(g_i|s_i) \doteq \pi_{i\theta_i}(g_i|s_i)$: policy for local i $\quad$ $\pi_i(v_i, \theta_{iw}) \doteq \pi_{i\theta_{iw}}(v_i, \theta_{iw})$: policy for local i value weight

q_{it}: return, Sample of $q_{\pi_i}(s_i, g_i)$

$\pi_i(v_i, \theta_{iw})$: policy for local i value weight

notation ↑ ↓

θ_{iw}: policy parameter for local i value weight

$$\nabla_{\theta_i} J(\theta_i, \theta_{iw}) = \frac{\partial}{\partial \theta_i} J(\theta_i, \theta_{iw}) = \sum_{i=1}^{N} \frac{\partial}{\partial \theta_i} V_{i\pi_{\theta_i}}(s_{i0}) \cdot \pi_i(v_i, \theta_{iw}) = \sum_{i=1}^{N} \sum_{s} d^{\pi_i}(s_i) \sum_{g} \frac{\partial \pi_i(g_i|s_i)}{\partial \theta_i} q_{\pi_i}(s_i, g_i) \cdot \pi_i(v_i, \theta_{iw})$$

$$= \sum_{i=1}^{N} \underbrace{\sum_{s} d^{\pi_i}(s_i) \sum_{g} \pi_i(g_i|s_i)} \frac{1}{\pi_i(g_i|s_i)} \frac{\partial \pi_i(g_i|s_i)}{\partial \theta_i} q_{\pi_i}(s_i, g_i) \cdot \pi_i(v_i, \theta_{iw})$$

$$= \mathbb{E}_{\pi} \quad \underbrace{\frac{\partial}{\partial \theta_i} \{\ln[\pi_i(g_i|s_i)]\, q_{\pi_i}(s_i, g_i) \cdot \pi_i(v_i, \theta_{iw})}_{\text{↳ item for updating } \theta_i}$$

↑ global, system

$\because \quad \theta_i \leftarrow \theta_i + \alpha \nabla_{\theta_i} J(\theta_i, \theta_{iw})$

$\therefore \quad \theta_i \leftarrow \theta_i + \alpha \{\nabla_{\theta_i} [\ln \pi_i(g_i|s_i)\, q_{it}]\} \pi_i(v_i, \theta_{iw})$

$$\nabla_{\theta_{iw}} J(\theta_i, \theta_{iw}) = \frac{\partial}{\partial \theta_{iw}} J(\theta_i, \theta_{iw}) = \sum_{i=1}^{N} V_{i\pi_{\theta_i}}(s_{i0}) \cdot \frac{\partial}{\partial \theta_{iw}} \pi_i(v_i, \theta_{iw}) = \sum_{i=1}^{N} V_{i\pi_{\theta_i}}(s_{i0}) \cdot \nabla_{\theta_{iw}} \pi_i(v_i, \theta_{iw})$$

$$= \underbrace{\sum_{i=1}^{N} \sum_{s} d^{\pi_i}(s_i) \sum_{g} \pi_i(g_i|s_i)} \{\pi_i(s_i, g_i) \cdot \nabla_{\theta_{iw}} \pi_i(v_i, \theta_{iw})$$

$$= \mathbb{E}_{\pi} \quad [q_{\pi_i}(s_i, g_i) \cdot \nabla_{\theta_{iw}} \pi_i(v_i, \theta_{iw})]$$

$\because \quad \theta_{iw} \leftarrow \theta_{iw} + \beta \nabla_{\theta_{iw}} J(\theta_i, \theta_{iw})$

$\therefore \quad \theta_{iw} \leftarrow \theta_{iw} + \beta [q_{\pi_i}(s_i, g_i) \cdot \nabla_{\theta_{iw}} \pi_i(v_i, \theta_{iw})]$