Paper Critique

Shuvrajeet Das, DA24D402

Course: DA7400, Fall 2024, IITM
Paper: [WHY DOES HIERARCHY WORK SO WELL IN REINFORCEMENT LEARNING?]
Date: [28-08-2024]

Make sure your critique Address the following points:
1. The problem the paper is trying to address
2. Key contributions of the paper
3. Proposed algorithm/framework
4. How the proposed algorithm addressed the described problem
Note: Be concise with your explanations. Unnecessary verbosity will be penalized. Please don't exceed 2 pages.

---

1 The problem the paper is trying to address

While HRL has been empirically successful in many challenging RL tasks, the underlying reasons for this success are unclear, especially in fully observed, Markovian settings where non-hierarchical policies could theoretically be optimal. The paper seeks to isolate and evaluate the commonly claimed benefits of HRL, such as improved exploration and easier policy learning, to understand which aspects of HRL contribute most to its empirical success. The authors find that the primary benefit of HRL in their experiments is improved exploration rather than the often-cited advantages of easier policy learning or imposed hierarchical structures. This insight leads them to propose simpler, exploration-focused techniques that can achieve performance competitive with HRL without the added complexity of hierarchical methods.

2 Key contributions of the paper

Key findings suggest that HRL's success is largely due to its ability to improve exploration, rather than other commonly cited benefits like temporal abstraction. The document discusses various HRL frameworks, including the options framework and goal-conditioned hierarchies, comparing them with non-hierarchical methods to isolate specific benefits.

- **Temporally Extended Training (H1):**
  - High-level actions span multiple environment steps.
  - Shortens effective episode length, speeding up reward propagation and improving learning.
- **Temporally Extended Exploration (H2):**
  - High-level actions lead to exploration that is temporally correlated across steps.
  - Enhances the efficiency of environment exploration.
- **Semantic Training (H3):**
  - High-level actions are semantically meaningful and correlated with future values.
  - Easier learning compared to training with low-level atomic actions.
- **Semantic Exploration (H4):**
  - Exploration strategies applied to meaningful actions are more effective.
  - In tasks like robot navigation, exploring at a high level (e.g., x-y coordinates) is more intuitive than at a low level (e.g., joint torques).

3 Proposed algorithm/framework

Algorithm 1 Hierarchical Reinforcement Learning Algorithm

Initialize critic networks $Q_{\theta_1^{lo}}, Q_{\theta_2^{lo}}, Q_{\theta_1^{hi}}, Q_{\theta_2^{hi}}$, actor networks $\mu_{\phi_1^{lo}}, \mu_{\phi_2^{lo}}, \mu_{\phi_1^{hi}}, \mu_{\phi_2^{hi}}$ with random θs and ϕs
Initialize target networks $\theta_1'^{lo} \leftarrow \theta_1^{lo}$, $\theta_2'^{lo} \leftarrow \theta_2^{lo}$, $\theta_1'^{hi} \leftarrow \theta_1^{hi}$, $\theta_2'^{hi} \leftarrow \theta_2^{hi}$, $\phi_1'^{lo} \leftarrow \phi_1^{lo}$, ...
Initialize replay buffer β_{lo}, β_{hi}
for $t = 1$ to T **do**
    Select action with explore-noise $a_t \sim \mu(s_t, g_t) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma)$
    Observe reward r and new state s_{t+1}, store tuple $(s_t, g_t, a_t, r, s_{t+1}, g_{t+1})$ in β_{lo}
    Select next goal via goal transition model with explore-noise $g_{t+1} \sim h(s_t, g_t, s_{t+1}) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma)$
    if $t \mod c$ **then**
        Generate next goal via μ^{hi} with explore-noise $g_{t+1} \sim \mu(s_{t+1}) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma)$
        Apply off-policy correction $\hat{g} = \text{correction}(g)$, store tuple $(s_{t-c+1}, g_{t-c+1}, r_{t-c+1:t}, s_{t+1})$ in β_{hi}
    end if
    Sample mini-batch of N steps $(s_t, g_t, a_t, r, s_{t+1}, g_{t+1})$ from β_{lo}
    $\hat{a} \leftarrow \mu_{\phi'}(s', g') + \epsilon, \epsilon \sim \text{clip}(\mathcal{N}(0, \hat{\sigma}), -k, k)$
    $y^{lo} \leftarrow r + \gamma \min_{i=1,2} Q_{\theta_i'}(s', g', \hat{a})$
    Update critics $\theta_i^{lo} \leftarrow \text{argmin}_{\theta_i^{lo}} N^{-1} \sum (y^{lo} - Q_{\theta_i^{lo}}(s, a, g))^2$
    if $t \mod d$ **then**
        Update ϕ_{lo} by the deterministic policy gradient
        Soft update target network
    end if
    if $t \mod c$ **then**
        Sample mini-batch of N steps $(s_{t-c+1}, g_{t-c+1}, r_{t-c+1:t}, s_{t+1})$ from β_{hi}
        $\hat{g'} \leftarrow \mu_{\phi'}(s') + \epsilon, \epsilon \sim \text{clip}(\mathcal{N}(0, \hat{\sigma}), -k, k)$
        $y^{hi} \leftarrow r + \gamma \min_{i=1,2} Q_{\theta_i'}(s', \hat{g'})$
        Update critics $\theta_i^{hi} \leftarrow \text{argmin}_{\theta_i^{hi}} N^{-1} \sum (y^{hi} - Q_{\theta_i^{hi}}(s, g, a))^2$
        if $t \mod d$ **then**
            Update ϕ_{hi} by the deterministic policy gradient
            Soft update target network
        end if
    end if
end for

4 How the proposed algorithm addressed the described problem

- **Temporally Extended Actions:** Speeds up learning by using high-level actions over multiple steps, leading to faster reward propagation.
- **Hierarchical Exploration:** Enhances exploration efficiency by operating at higher abstraction levels, covering the state space more effectively.
- **Multi-Level Learning:** Separates high-level goal setting from low-level control, enabling more efficient and robust learning.