Paper Critique

Shuvrajeet Das, DA24D402
Course: DA7400, Fall 2024, IITM
Paper: [Generating Adjacency-Constrained Subgoals in Hierarchical Reinforcement Learning]
Date: [06-09-2024]

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

Problem Addressed:

Goal-conditioned hierarchical reinforcement learning (HRL) often faces training inefficiency due to the large high-level action space (goal space), making both high-level subgoal generation and low-level policy learning challenging. To mitigate this, the paper proposes restricting the high-level action space to a k-step adjacent region near the current state using an adjacency constraint, which can be expressed as:

$$GA(s, k) := \{g \in G \mid d_{st}(s, \phi^{-1}(g)) \leq k\}$$

where $d_{st}(s_1, s_2)$ is the shortest transition distance, minimizing the expected first hit time:

$$d_{st}(s_1, s_2) := \min_{\pi \in \Pi} \mathbb{E}[T_{s_1 s_2} \mid \pi] = \min_{\pi \in \Pi} \sum_{t=0}^{\infty} tP(T_{s_1 s_2} = t \mid \pi)$$

The constrained high-level objective is:

$$\max_{\theta_h} \mathbb{E}_{\pi_h^{\theta_h}} \left[\sum_{t=0}^{T-1} \gamma^t r_{hkt} \right] \text{ subject to } d_{st}(s_{kt}, \phi^{-1}(g_{kt})) \leq k,\ t = 0, 1, \ldots, T-1$$

This constraint aims to reduce the action space while preserving the optimal hierarchical policy in deterministic MDPs.

2 Key contributions of the paper

Key Contributions:

- **Adjacency-Constrained Subgoals:** Introduces a k-step adjacency constraint that restricts the high-level action space from the entire goal space to a k-step adjacent region, defined as:
$$GA(s, k) := \{g \in G \mid d_{st}(s, \phi^{-1}(g)) \leq k\}$$
- **Adjacency Network:** Proposes a practical implementation by training an adjacency network ψ_ϕ to approximate the shortest transition distance.
- **Optimization Formulation:** Introduces an unconstrained optimization objective incorporating adjacency loss:
$$\max_{\theta_h} \mathbb{E}_{\pi_h^{\theta_h}} \left[\sum_{t=0}^{T-1} \left(\gamma^t r_{hkt} - \eta \cdot \max\left(\|\psi_\phi(\phi(s_{kt})) - \psi_\phi(g_{kt})\|_2 - \epsilon_k, 0 \right) \right) \right]$$

3 Proposed algorithm/framework

Algorithm 1 HRAC

Input: High-level policy $\pi_h^{\theta_h}$ parameterized by θ_h, low-level policy $\pi_l^{\theta_l}$ parameterized by θ_l, adjacency network ψ_ϕ parameterized by ϕ, state-goal mapping function ϕ, goal transition function h, high-level action frequency k, number of training episodes N, adjacency learning frequency C, empty adjacency matrix $\mathcal{M}$, empty trajectory buffer $\mathcal{B}$.

Sample and store trajectories in the trajectory buffer $\mathcal{B}$ using a random policy.
Construct the adjacency matrix $\mathcal{M}$ using the trajectory buffer $\mathcal{B}$.
Pre-train ψ_ϕ using $\mathcal{M}$ by minimizing Equation (11).
Clear $\mathcal{B}$.
for $n = 1$ to N **do**
    Reset the environment and sample the initial state s_0.
    $t = 0$.
    repeat
        if $t \equiv 0 \pmod{k}$ **then**
            Sample subgoal $g_t \sim \pi_h^{\theta_h}(g|s_t)$.
        else
            Perform subgoal transition $g_t = h(g_{t-1}, s_{t-1}, s_t)$.
        end if
        Sample low-level action $a_t \sim \pi_l^{\theta_l}(a|s_t, g_t)$.
        Sample next state $s_{t+1} \sim P(s|s_t, a_t)$.
        Sample reward $r_t \sim R(r|s_t, a_t)$.
        Sample episode end signal done.
        $t = t + 1$.
    until done is true.
    Store the sampled trajectory in $\mathcal{B}$.
    Train high-level policy $\pi_h^{\theta_h}$ according to Equation (12) and (13).
    Train low-level policy $\pi_l^{\theta_l}$.
    if $n \equiv 0 \pmod{C}$ **then**
        Update the adjacency matrix $\mathcal{M}$ using the trajectory buffer $\mathcal{B}$.
        Fine-tune ψ_ϕ using $\mathcal{M}$ by minimizing Equation (11).
        Clear $\mathcal{B}$.
    end if
end for

4 How the proposed algorithm addressed the problem

The algorithm addresses the problem by restricting the high-level policy to generate subgoals within a k-step adjacent region:

$$GA(s,k) := \{g \in G \mid d_{st}(s, \phi^{-1}(g)) \leq k\}$$

It trains an adjacency network ψ_ϕ to estimate adjacency, enforcing the constraint with an adjacency loss. This reduces the high-level action space, improves learning efficiency, and preserves the optimal policy.