PEAR: Primitive Enabled Adaptive Relabeling for Boosting Hierarchical Reinforcement Learning

Utsav Singh[1], Vinay P Namboodiri[2]

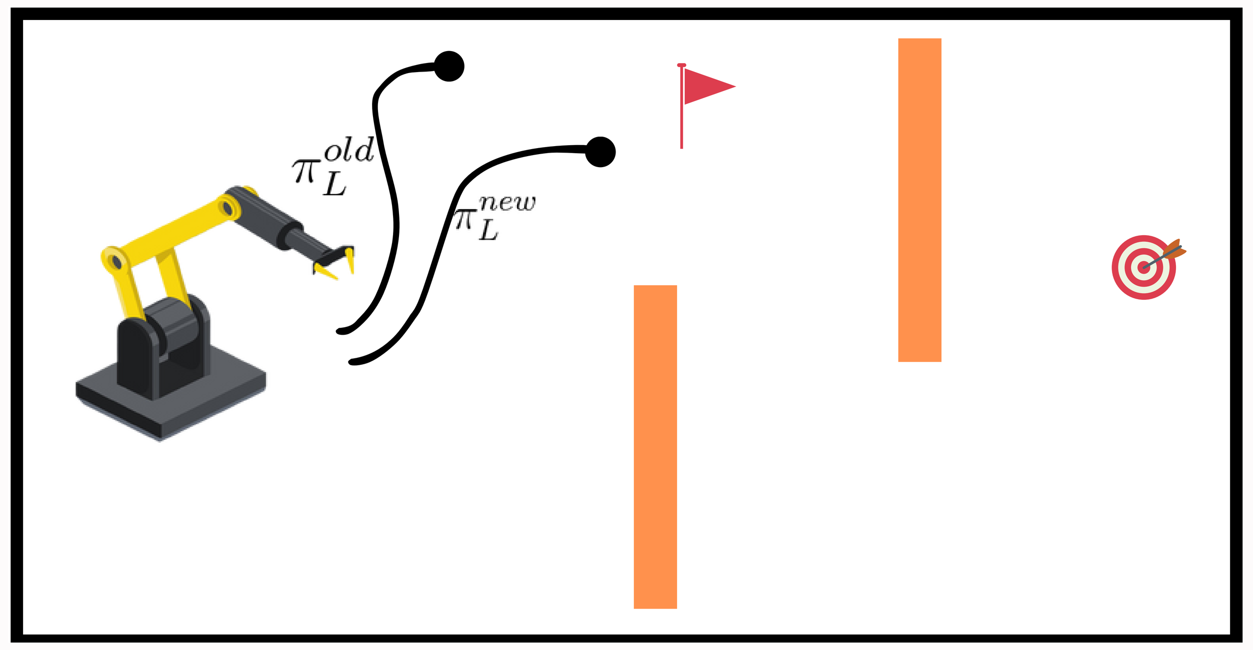
[1] IIT Kanpur [2] University of Bath

Motivation

- Off-policy Hierarchical Reinforcement Learning (HRL) suffers from non-stationarity.
- Can we leverage a few expert demonstrations to deal with non-stationarity in HRL?

Non-stationarity

- Off-policy HRL suffers from non-stationarity, due to non-stationary lower primitive behavior.
- Off-policy RL transitions for bi-level hierarchy:
  Higher level: $(s_t, g^*, g_t, \Sigma_{i=t}^{t+k-1} r_i, s_{t+k-1})$
  Lower level: $(s_t, g_t, a_t, r_t, s_{t+1})$
- Since the lower level primitive changes with training, the previously collected transitions become obsolete.



Convergence Bounds

:

- Sub-optimality Definition:

$$Subopt(\theta) = |J(\pi^*) - J(\pi)| \quad (1)$$

- Sub-optimality Analysis:

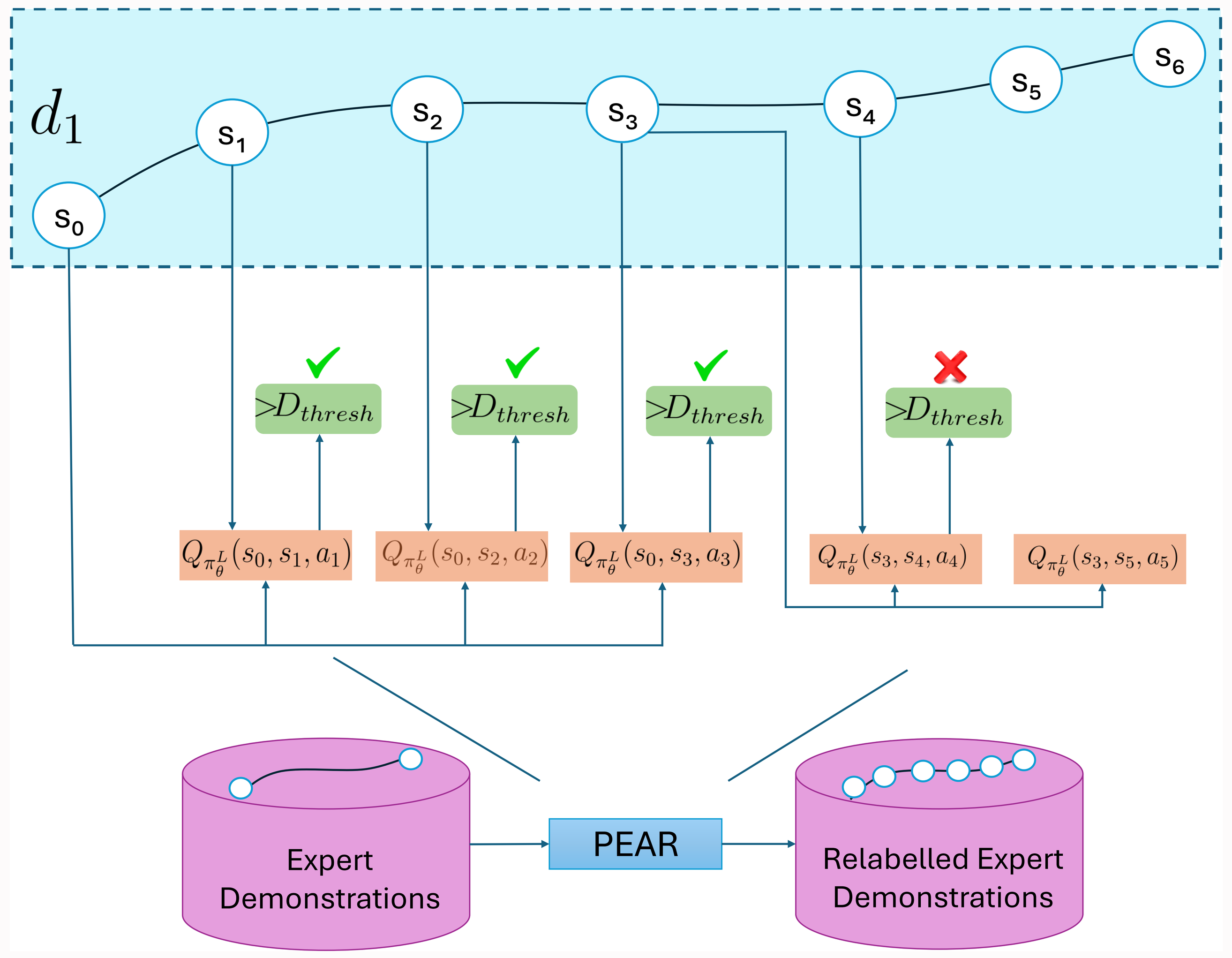
$$|J(\pi^*) - J(\pi_{\theta_H}^H)| \leq \lambda_H * \phi_D + \lambda_H * \mathbb{E}_{s\sim\kappa,\pi_D^H\sim\Pi_D^H,g\sim G}[D_{TV}(\pi_D^H(\tau|s,g)||\pi_{\theta_H}^H(\tau|s,g))] \quad (2)$$

where $\lambda_H = \frac{2}{(1-\gamma)(1-\gamma^c)} R_{max} \|\frac{d_c^{\pi^*}}{\kappa}\|_\infty$

PEAR

- **Main Idea**: Expert demonstrations are adaptively parsed and segmented using the current lower primitive, thereby mitigating non-stationarity.

Adaptive Relabeling Overview



Joint Optimization

- The higher level policy reinforcement learning term ($J_{\theta_H}^H$) is regularized using additional imitation learning term ($J_{BC}^H(\theta_H)$ or $J_D^H(\theta_H, \epsilon_H)$).

$$\min_{\theta_H} J_{BC}^H(\theta_H) = \min_{\theta_H} \mathbb{E}_{(s^e, s_g^e, s_{next}^e)\sim D_g, s_g\sim\pi_{\theta_H}^H(\cdot|s^e,g^e)} ||s_g^e - s_g||^2 \quad (3)$$

$$\max_{\theta_H} \min_{\epsilon_H} J_D^H(\theta_H, \epsilon_H) = \frac{1}{2}\mathbb{E}_{(s^e, s_g^e, \cdot)\sim D_g}[\mathbb{D}_{\epsilon_H}^H(s_g^e) - 1]^2 + \quad (4)$$

$$\max_{\theta_H} \min_{\epsilon_H} \frac{1}{2}\mathbb{E}_{(s^e,\cdot,\cdot)\sim D_g, s_g\sim\pi_{\theta_H}(\cdot|s^e,g^e)}[\mathbb{D}_{\epsilon_H}^H(\pi_{\theta_H}^H(\cdot|s^e, g^e)) - 0]^2$$

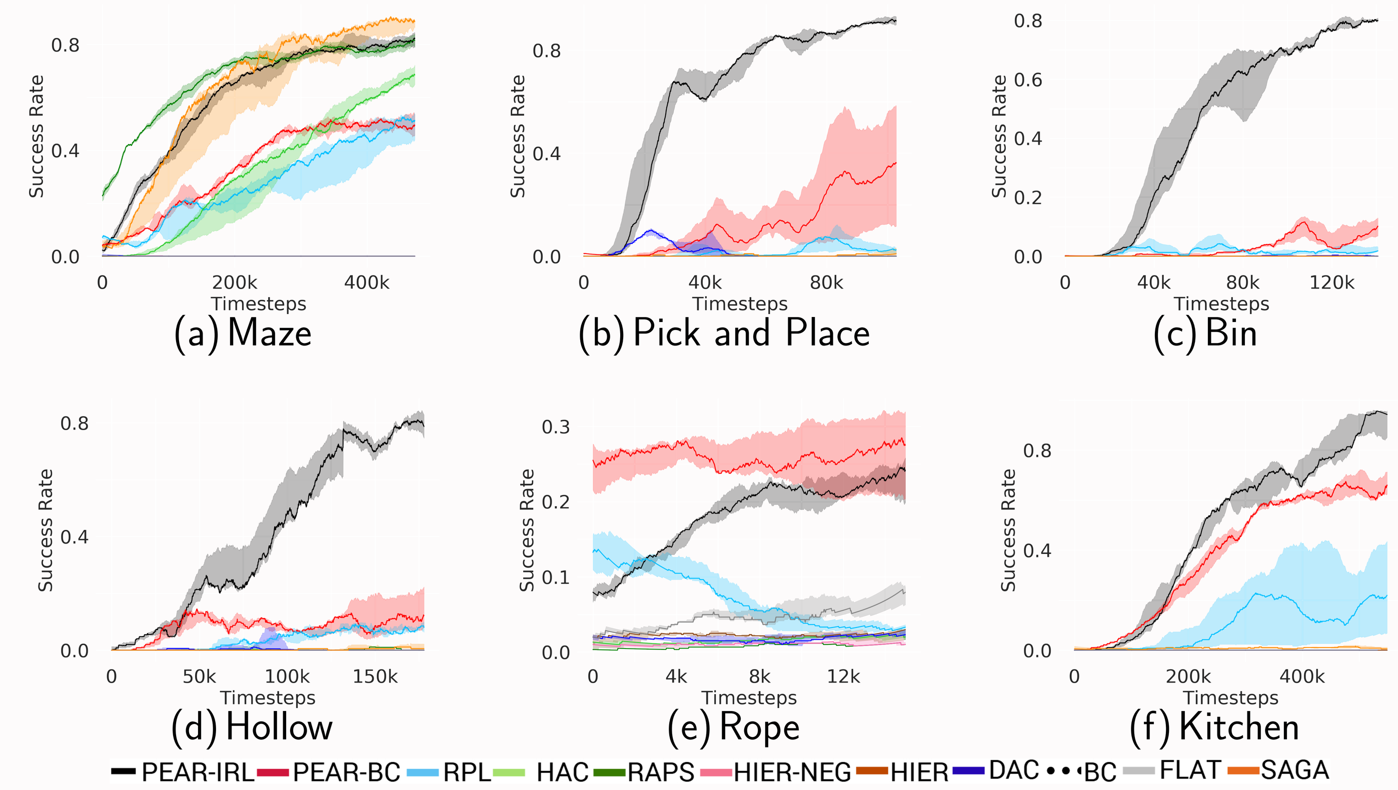
$$\max_{\theta_H}(J_{\theta_H}^H - \psi * J_{BC}^H(\theta_H)) \quad (5)$$

$$\min_{\epsilon_H} \max_{\theta_H}(J_{\theta_H}^H + \psi * J_D^H(\theta_H, \epsilon_H)) \quad (6)$$

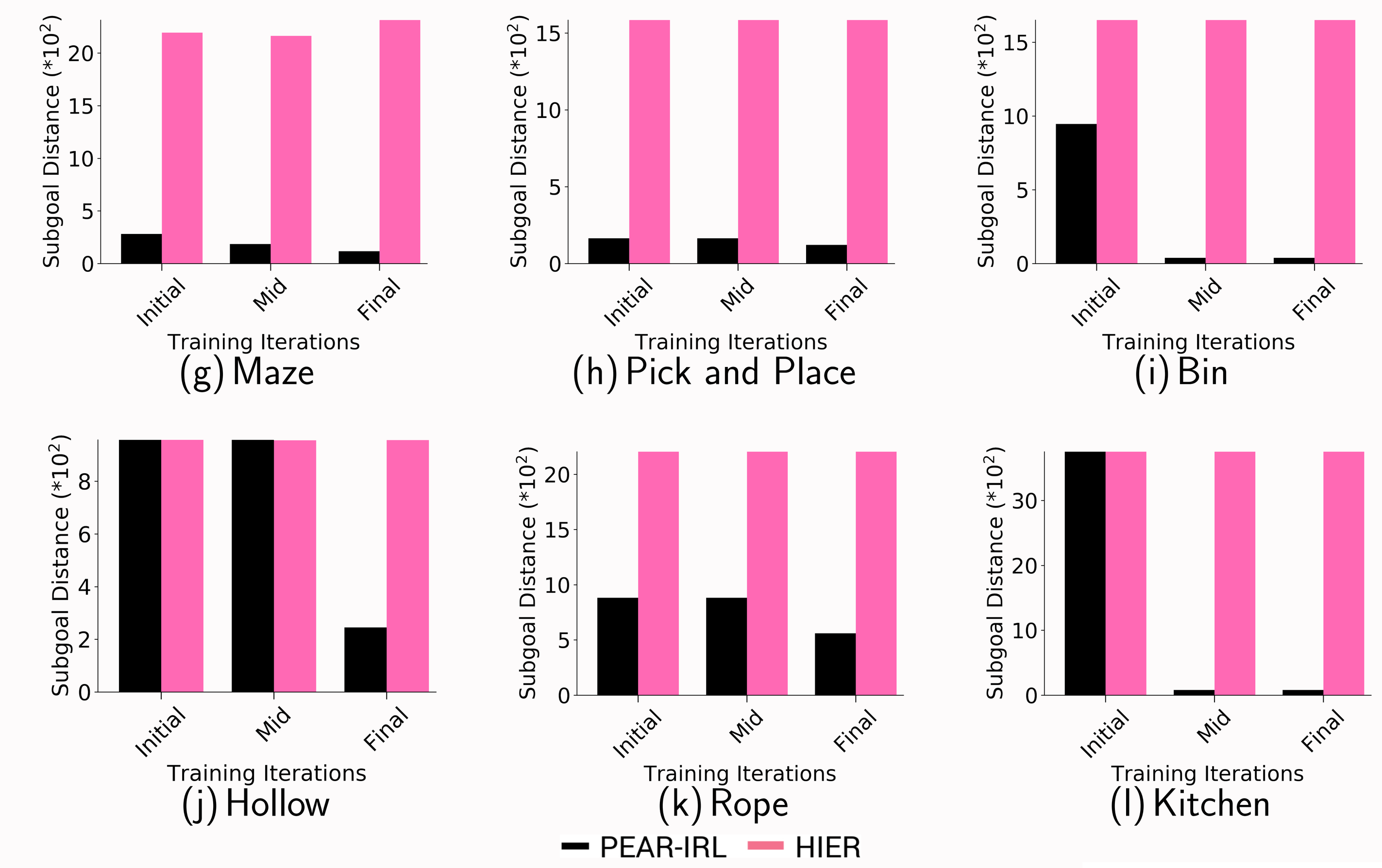
Results

Success Rate Comparison:

PEAR achieves > 80% success rates on all tasks:



Non-Stationarity Metric Comparison:



Key Insights:

- Adaptive Relabeling generates efficient subgoal supervision for higher-level policy.
- PEAR is able to mitigate non-stationarity in HRL.
- Sub-Optimality Bounds justify the importance of periodic re-population using adaptive relabeling.

SCAN ME

Scan to Read the Paper.