# Paper Critique

Anirud N, CE21B014

**Course:** DA7400, Fall 2024, IITM
**Paper:** Why Does Hierarchy (Sometimes) Work So Well in Reinforcement Learning?
**Date:** 28 August 2024

Make sure your critique addresses 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 Problem Statement

This paper aims to understand why Hierarchies work well in Reinforcement Learning using a few experiments. Prior works have included arguments such as high-level actions give a more semantic sense to the required actions - operate at a lower frequency, and are easier in both learning and exploration. etc. This paper aims to find out why exactly are these, with a few experiments.

## 2 Key Contributions

The above questions as to why and when HRL works have been answered via empirical analysis with different tasks such as locomotion, navigation, and manipulation. Isolation and evaluation of the claimed benefits of HRL have been done. Most of the empirical benefits of hierarchies are attributed to improved exploration. Using this argument, the paper also imposes different exploration strategies inspired by HRL and easier to implement. The experiments revealed that only a few out of all the claimed benefits of HRL have been achieved so far.

## 3 Proposed Algorithm/Framework

For low level training Q function is learnt to minimise the error

$$(s_t, g_t, a_t, r_t, s_{t+1}, g_{t+1}) = ((s_t, g_t, a_t) - r_t - \gamma(s_{t+1}, g_{t+1}, (s_{t+1}, g_{t+1}))^2 , \tag{1}$$

over single-step transitions, where $g_t$ is the current goal (high-level action updated every $c$ steps) and $r_t$ is an intrinsic reward measuring negative L2 distance to the goal. The lower-level policy is then trained to maximize the Q-value $(s_t, g_t, (s_t, g_t))$.

For High level training

$$(s_t, g_t, R_{t:t+c-1}, s_{t+c}) = ((s_t, g_t) - R_{t:t+c-1} - \gamma(s_{t+c}, (s_{t+c})))^2 . \tag{2}$$

### 3.1 Hypothesis of benefits of Hierarchies

- **H1 Temporally extended training** Multiple steps in environment - faster learning
- **H2 Temporally extended exploration** Exploration at a level is mapped to environment exploration -correlated across steps. Efficient exploration
- **H3 Semantic training** - Training for Meaningful actions
- **H4 Semantic exploration** - Explorations on meaningful actions

## 4 Experiments

### 4.1 Benefits of Temporal Abstraction H1 ad H2

- During training, the higher-level policy is trained with respect to temporally extended transitions of the form $(s_t, g_t, R_{t:t+c-1}, s_{t+c})$
- During experience collection, a high-level action is sampled and updated every $c$ steps.

Average success rates and standard errors for 5 random seeds were calculated. In the training - for all environments, it was observed that having $c > 1$ gives significantly better results than $c = 1$. For Exploration - for ant maze environment alone - they had very little difference for different values of $c$. For other environments the differences were well seen.

### 4.2 Benefits of Hierarchial Training (H1 AND H3)

Disentangle exploration from action representation - shadow agent - standard non-HRL that is trained based on experiences of HRL along with an HRL. If exploration matters, shadow will work better, but if action representation is important, HRL works better.
Results show that learning atomic actions without higher-level action representation is good enough, and its performance is at par with the HRL framework. Except in Antmaze, where the small drop is observed. In amaze - temporally abstracted is more important, but multi-step rewards can do this easily, so even in this case, HRL is not compulsory.

### 4.3 Benefits of Hierarchial Exploration

This creates a new exploration strategy - explore and exploit and swicthing ensemble - inspired from HRL. One is trained to maximise rewards and other is trained to reach goals. similar to HRL.
Overall, all this suggests that HRL works well due to exploration. Option-based hierarchies are better in exploration as opposed to high-level representations. Using separate networks for Explore and Exploit is crucial.

### 4.4 Conclusions

Empirical analysis has limitations. Results and conclusions are restricted to a limited set of hierarchical designs. The use of other hierarchical designs may lead to different conclusions. The benefits of hierarchy may be different in other settings.
In addition, tasks with more complex environments and/or sparser rewards may benefit from other mechanisms for encouraging exploration (e.g., count-based exploration), which would be a complementary investigation to this study. An examination of different hierarchical structures and more varied settings is an important direction for future research.