# Latent Space Policies for Hierarchical Reinforcement Learning

**Manan Tomar : ED14B023**

## Abstract

This work (Haarnoja et al., 2018a) introduces a method for learning hierarchical policies using the entropy regularized objective. This allows not having restrictions on the lower levels to use higher level signals and still being able to learn diversified behavior at each level, making the task easier for the higher level. The experiments are done on the Mujoco continuous control tasks set, including the humanoid.

## 1. Approach

The Reinforcement Learning problem, when cast as inference over graphical models, results in the maximum entropy objective (Haarnoja et al., 2017). In such a case, each state action pair is assigned a binary random variable $\mathcal{O}_t$, called the optimality variable. This denotes if the time step was optimal or not and the problem is now of inferring the posterior given by $p(a_t|s_t, \mathcal{O}_{t:T} = true)$. Moreover, the reward function in this setting is incorporated as $p(\mathcal{O}_t|s_t, a_t) = \exp(r(s_t, a_t))$, where $r(s_t, a_t) < 0$. Using variational inference to determine the above mentioned posterior expression results in the following entropy formulation :

$$J(\pi) = E_{\tau \sim \rho_\pi(\tau)} \left[ \sum_t r(s_t, a_t) + \alpha \mathcal{H}(\pi(.|s_t)) \right] \quad (1)$$

The authors introduce latent variables in this setup. The base policy is written as $\pi(a_t|s_t, h_t)$, conditioned on the latent variable $h_t$, which is sampled from an assumed prior $p(h_t)$. Marginalizing out the actions $a_t$ from this results in $h_t$ acting as action variables for the higher level policy where the dynamics are shaped by the base policy and the true dynamics. This allows adding any number of layers to the model while keeping the original graphical structure intact.

Now the policy at each level, called as a sub-policy, is trained and frozen, with the upper level policy using the latent variables of the sub-policy as its actions. The training is done using the maximum entropy formulation where the soft actor-critic method (Haarnoja et al., 2018b) is used. The reward being used for each training each layer is in the designer's hands. Therefore, for tasks with explicit hierarchical structure, having shaping rewards for lower levels and the true task reward for higher levels can ensure that primitive skills are learnt first which can help solve the overall task. In other cases, all layers can be trained on the same reward, that given by the task in hand.

## 2. Experiments

The authors show results on Mujoco tasks including Half-Cheetah, Swimmer, Ant and Humanoid. Particularly, they first use a single layer policy to compare with popular methods such as PPO, DDPG, Soft Q-Learning and the original soft actor critic version which uses GMM policies. They show that their method is on par with these methods in all the environments considered. Then a two layer policy is implemented on the ant navigation task where the lower level is trained to learn general locomotion, while the higher level is trained to solve reaching the goal state in the given maze.

## 3. Critique

Although the method is motivated with the intention of training multiple layer policies, the authors never move beyond a two level policy in all their experiments, with the majority only using a single layer policy. Moreover, the current formulation still leaves unanswered questions in terms of what kind of reward should be used for each hierarchy layer, which has been observed to be highly task dependent.

## References

Haarnoja, Tuomas, Tang, Haoran, Abbeel, Pieter, and Levine, Sergey. Reinforcement learning with deep energy-based policies. *arXiv preprint arXiv:1702.08165*, 2017.

Haarnoja, Tuomas, Hartikainen, Kristian, Abbeel, Pieter, and Levine, Sergey. Latent space policies for hierarchical reinforcement learning. *arXiv preprint arXiv:1804.02808*, 2018a.

Haarnoja, Tuomas, Zhou, Aurick, Abbeel, Pieter, and Levine, Sergey. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018b.