Learning Options with Interest Functions

Khimya Khetarpal and Doina Precup

Mila-McGill University, Montréal (QC) Canada

Motivation

- How to create agents which efficiently represent, learn and use knowledge of the world in continual fashion just like humans?
- While we engage in a task, each skill employed is specialized in attending to only certain states. For example, a skill such as *'stop if the traffic light is red'* is only applicable in states in which a traffic light is present.
- **Learn options that represent specialized meaningful skills for lifelong learning**.
- **Hypothesis:** Knowing where to apply which skills results in specialization which is key to scaling up.

Key Contribution

- We introduce the notion of *interest functions* $I_\omega : S \times O \longrightarrow \mathbb{R}^+$, inspired by [3].
- The state-value function over options that have interest functions is defined as:

$$V_\Omega(s) = \sum_\omega \pi_{I_{\omega,z}}(\omega|s) Q_{\Omega,\theta}(s,\omega) \tag{1}$$

where $Q_{\Omega,\theta}$ is the option-value function parameterized by θ, and the probability of option ω being sampled in in state s is defined as:

$$\boldsymbol{\pi_{I_{\omega,z}}(\omega|s) = I_{\omega,z}(s)\pi_\Omega(\omega|s) \Big/ \sum_\omega I_{\omega,z}(s)\pi_\Omega(\omega|s)} \tag{2}$$

The Story So Far..

Temporally extended actions can be formalized as options [1]. A Markovian option $\omega \in \Omega$ is defined as $\langle I_\omega, \beta_\omega, \pi_\omega \rangle$

- Intra-option policy π_ω,
- Termination condition $\beta_\omega : S \rightarrow [0,1]$,
- Initiation set $I_\omega \subseteq S$.

Recent research has demonstrated that options can be learned automatically and end-to-end for a given task with *option-critic* architecture [2]. *What is missing?*

Interest Gradient Theorem

Given a set of Markov options with stochastic, differentiable interest functions $I_{\omega,z}$, the gradient of the expected discounted return with respect to z at (s,ω) is:

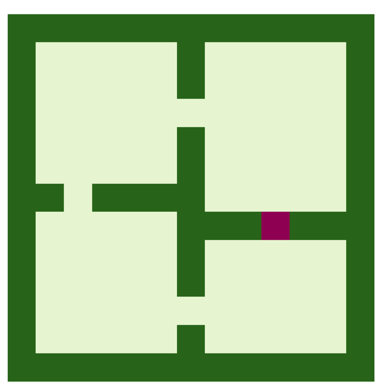
$$\boxed{\sum_{s',\omega'} \widehat{\mu}_\Omega(s',\omega'|s,\omega)\beta_{\omega,\nu}(s') \frac{\partial \pi_{I_{\omega,z}}(\omega'|s')}{\partial z} Q_\Omega(s',\omega')}$$

where $\widehat{\mu}_\Omega(s',\omega'|s,\omega)$ is the discounted weighting of the state-option pairs along trajectories starting from (s,ω) sampled from the sampling distribution determined by $I_{\omega,z}$.

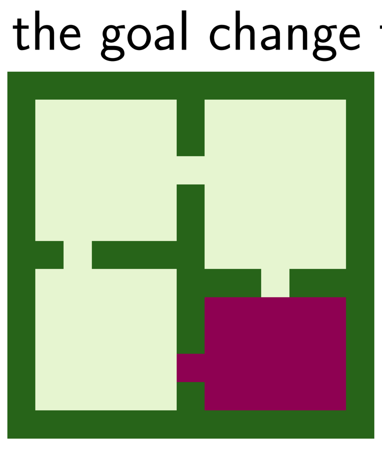
Interest Option Critic

Four Rooms Environment: Do options with interest help in transfer?

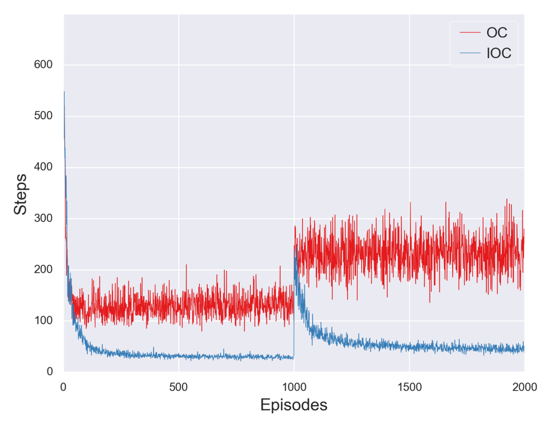
- After 1000 episodes, the goal is randomly moved to one of the cells in the lower right room (shown in red)
- The IOC agent performs better than OC in the initial stage, then is able to recover much faster after the goal change than the OC agent



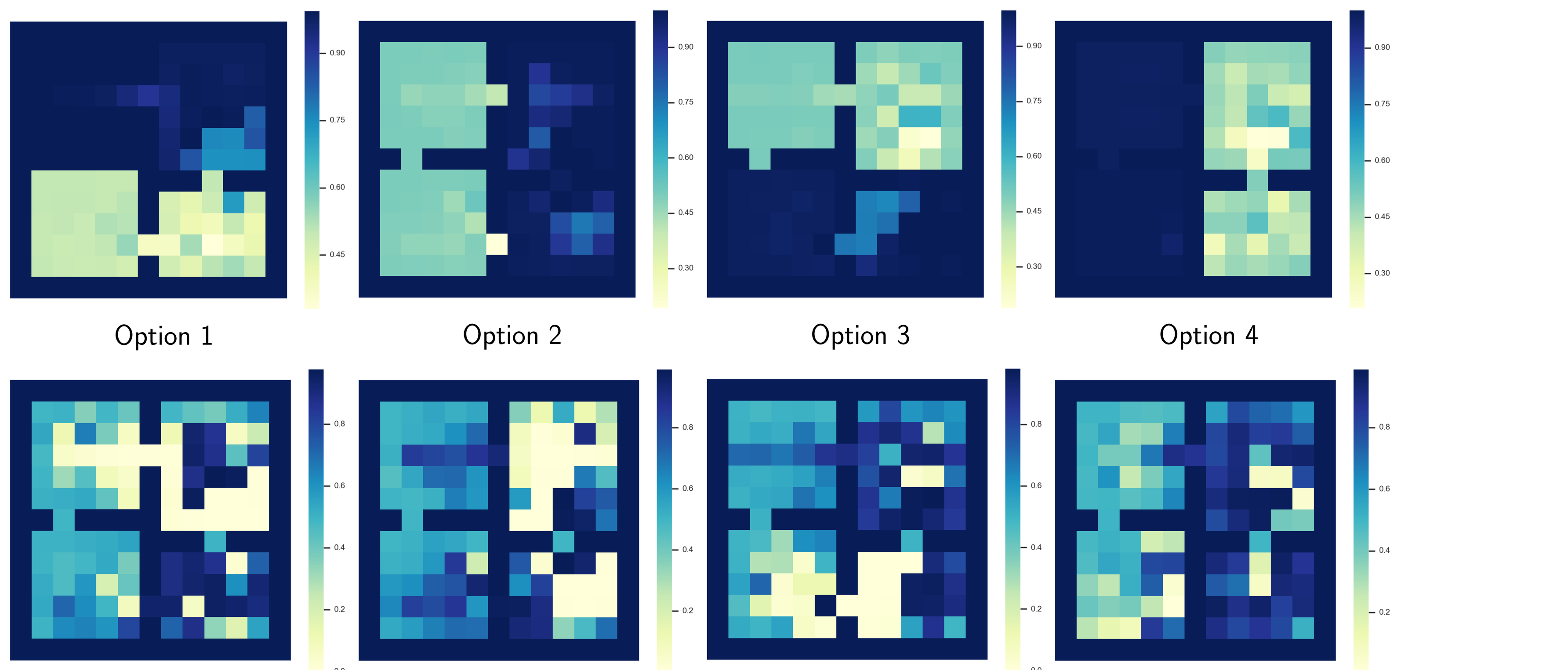
Task 1: Goal is in east hallway



Random Goal



Learning Options with Interest



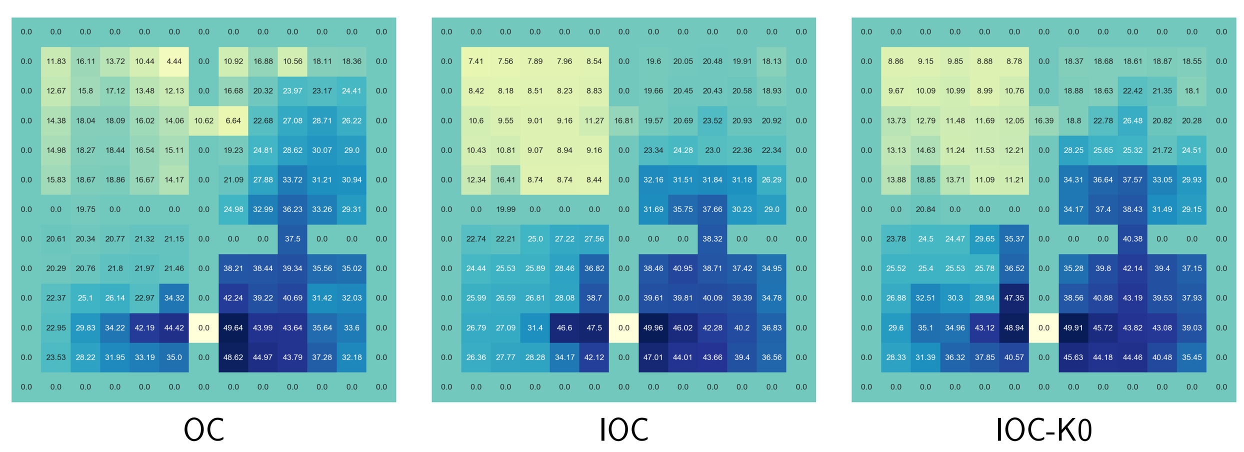
Interest Functions (top row) at the end of 500 episodes in task 1 for IOC with 4 options. Darker colors represent higher values of the interest function. **Termination Functions** (bottom row) of each option at the end of 500 episodes. Options learnt with interest functions are specialized in *different* regions of the state space.

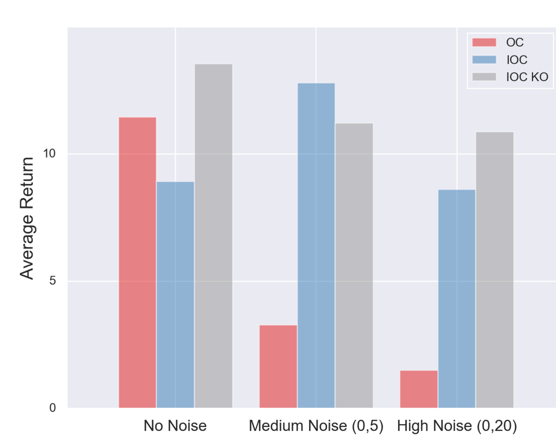
Few Shot Option Value Learning

Do learned interest functions help re-use of temporal abstraction?

- We then harvest the learned options and use them in the task of learning to navigate to the south hallway
- The policy over options ($\pi_\Omega(\omega|s)$) and option value function $Q(s,\omega)$ are being learnt from scratch
- We experiment with two conditions: using the interest function directly, or thresholding its value and choosing only among options whose interest at a state is higher than the threshold (indicated by a hyper parameter K).

Value Function Propagation





- We hypothesized that if the reward is affected by noise, knowing where to propagate would help IOC more. To test this hypothesis, we repeated the few-shot option value learning with varying degrees of noisy per-step reward.

Wrap up

- Our approach is capable of learning interest functions, leading to options that are reusable, interpretable, and specialized to different regions of state space.
- **Future Work:** Learn interest functions with function approximation in much larger, richer complicated environments.

References

- Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.
- Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *AAAI*, pages 1726–1734, 2017.
- Richard S. Sutton, Ashique Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research*, 17:73:1–73:29, 2016.