Options of Interest: Temporal Abstraction with Interest Functions

Khimya Khetarpal

in collaboration with Martin Klissarov, Maxime Chevalier-Boisvert, Pierre-Luc Bacon, Doina Precup

Reasoning and Learning Lab
Mila - McGill University

HRL Weekly Meeting - DeepMind, 2019

Research Goals

How should an AI agent efficiently represent, learn and use knowledge of the world in continual tasks?

Temporal Abstraction: Options Framework

- Definition

Let S, A be the set of states and actions. A Markov option $\omega \in \Omega$ is a triple:

$$\left(\mathbf{I}_\omega \subseteq \mathbf{S}\ ,\ \pi_\omega : \mathbf{S} \times \mathbf{A} \to [\mathbf{0}, \mathbf{1}]\ ,\ \beta_\omega : \mathbf{S} \to [\mathbf{0}, \mathbf{1}]\right)$$

Initiation set **Intra option policy** **Termination condition**

- I_ω set of states aka preconditions
- $\pi_\omega(s, a)$ probability of taking an action $a \in A$ in state $s \in S$ when following the option ω
- $\beta_\omega(s)$ probability of terminating option ω upon entering state s

with a policy over options $\pi_\Omega : S \times \Omega \to [0,1]$

- Example
  - Robot navigating in a house: when you come across a closed door (I_ω), open the door (π_ω), until the door has been opened (β_ω)

Can we learn such temporal abstractions?

- Bacon, Harb, and Precup, 2017 proposed the option-critic framework which provides the ability to ***learn*** a set of options

Can we learn such temporal abstractions?

- Bacon, Harb, and Precup, 2017 proposed the option-critic framework which provides the ability to ***learn*** a set of options

- Optimize directly the discounted return, averaged over all the trajectories starting at a designated state and option

$$J = E_{\Omega,\theta,\omega}[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \,|\, s_0, \omega_0]$$

Can we learn such temporal abstractions?

- Bacon, Harb, and Precup, 2017 proposed the option-critic framework which provides the ability to ***learn*** a set of options

- Optimize directly the discounted return, averaged over all the trajectories starting at a designated state and option

$$J = E_{\Omega,\theta,\omega}[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \mid s_0, \omega_0]$$

Assumption: All options are available in all states

Learning options with interest functions

Hypothesis:

Learning options which are specialized in situations of specific interest can be leveraged to learn meaningful, interpretable and reusable temporal abstractions.

Learning options with interest functions

- Break the assumption that all options are present in all states.
- Build-in a form of attention mechanism

Learning options with interest functions

- Break the assumption that all options are present in all states.
- Build-in a form of attention mechanism
- **Definition**: Interest Function $\mathbf{I}_{\omega,\mathbf{z}} : \mathbf{S} \times \mathbf{\Omega} \longrightarrow \mathbb{R}^{+}$ generalizes the notion of initiation sets, and is an indication of the extent to which an option ω is applicable in a state s.
- Here we consider differentiable interest functions parameterized with z.

Interest-Option-Critic

The value of $I_{\omega,z}(s)$ modulates the probability of option ω being sampled in state s by a policy over options $\pi_\Omega(\omega \mid s)$, resulting in an *interest policy over option* defined as:

$$\boxed{\pi_{I_{\omega,z}}(\omega \mid s) = I_{\omega,z}(s)\pi_\Omega(\omega \mid s) \Big/ \sum_{\omega'} I_{\omega',z}(s)\pi_\Omega(\omega' \mid s)}$$

$\pi_\Omega(\omega \mid s)$ is the policy over options

$I_{\omega,z}(s)$ is the Interest function

Interest-Option-Critic

The state-value function over options that have interest functions is now defined as:

$$\boxed{V_\Omega(s) = \sum_\omega \pi_{I_{\omega,z}}(\omega \mid s) Q_{\Omega,\theta}(s, \omega)}$$

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

$$\pi_{I_{\omega,z}}(\omega \mid s) = I_{\omega,z}(s)\pi_\Omega(\omega \mid s) \Big/ \sum_{\omega'} I_{\omega',z}(s)\pi_\Omega(\omega' \mid s)$$

$\pi_\Omega(\omega \mid s)$ is the policy over options

$I_{\omega,z}(s)$ is the Interest function

Interest-Option-Critic

The option value function is defined as

$$Q_{\Omega}(s,\omega) = \sum_{a} \pi_{\omega,\theta}(a \mid s) Q_U(s,\omega,a)$$

Interest-Option-Critic

The option value function is defined as

$$Q_\Omega(s,\omega) = \sum_a \pi_{\omega,\theta}(a\,|\,s)Q_U(s,\omega,a)$$

Taking the derivation w.r.t. z

$$\frac{\partial Q_\Omega(s,\omega)}{\partial z} = \frac{\partial}{\partial z}\left\{\sum_a \pi_{\omega,\theta}(a\,|\,s)Q_U(s,\omega,a)\right\}$$

Interest-Option-Critic

The option value function is defined as

$$Q_\Omega(s,\omega) = \sum_a \pi_{\omega,\theta}(a \mid s) Q_U(s,\omega,a)$$

Taking the derivation w.r.t. z

$$\frac{\partial Q_\Omega(s,\omega)}{\partial z} = \frac{\partial}{\partial z}\left\{ \sum_a \pi_{\omega,\theta}(a \mid s) Q_U(s,\omega,a) \right\}$$

where $Q_U : S \times \Omega \times A \rightarrow \mathbb{R}$ is the value of executing an action in the context of a state-option pair defined as:

$$Q_U(s,\omega,a) = r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) U(\omega,s')$$

Interest-Option-Critic

The option value function is defined as

$$Q_\Omega(s,\omega) = \sum_a \pi_{\omega,\theta}(a \mid s) Q_U(s,\omega,a)$$

Taking the derivation w.r.t. z

$$\frac{\partial Q_\Omega(s,\omega)}{\partial z} = \frac{\partial}{\partial z}\left\{ \sum_a \pi_{\omega,\theta}(a \mid s) Q_U(s,\omega,a) \right\}$$

where $Q_U : S \times \Omega \times A \to \mathbb{R}$ is the value of executing an action in the context of a state-option pair defined as:

$$Q_U(s,\omega,a) = r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) U(\omega,s')$$

where $U : S \times \Omega \to \mathbb{R}$ is the option-value function upon arrival in a state:

$$U(\omega,s') = (1 - \beta_{\omega,\nu}(s')) Q_\Omega(s',\omega) + \beta_{\omega,\nu}(s') V_\Omega(s')$$

Interest-Option-Critic

Main Result : Interest Function Gradient Updates

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

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

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

Interest-Option-Critic

Main Result : Interest Function Gradient Updates

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

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

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

Intuitively, the gradient update to z can be interpreted as increasing the interest in an option which terminates in states with good value. It links initiation and termination, which is natural.

Interest-Option-Critic

- The agent *initially* would consider that all options are available everywhere.

- As learning progresses, we would like the emerging options to be specialized over ***different*** state-space regions.

- We derive the policy gradient theorem for interest functions, intra-option policy and the termination function.

- **TL;DR** all three components of options are parameterized and learned

$$\Big(\mathbf{I}_{\omega,\mathbf{z}} : \mathbf{S} \times \boldsymbol{\Omega} \rightarrow \mathbb{R}^+, \ \pi_{\omega,\theta} : \mathbf{S} \times \mathbf{A} \rightarrow [\mathbf{0}, \mathbf{1}], \ \beta_{\omega,\nu} : \mathbf{S} \rightarrow [\mathbf{0}, \mathbf{1}]\Big)$$

Interest Functions **Intra option policy** **Termination condition**

Experimental Results

- Are options with interest functions useful in a single task?
- Do interest functions facilitate learning reusable options?
- Do interest functions lead to better interpretability of the learned options?

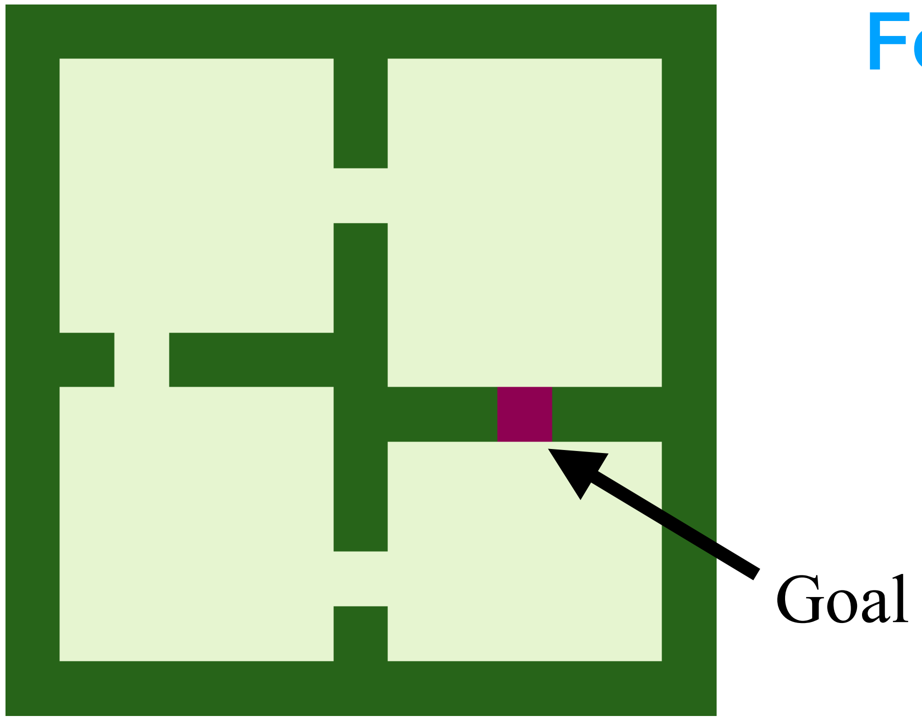
Experimental Results

- **Are options with interest functions useful in a single task?**

- Do interest functions facilitate learning reusable options?

- Do interest functions lead to better interpretability of the learned options?

Are interest functions useful in a single task?

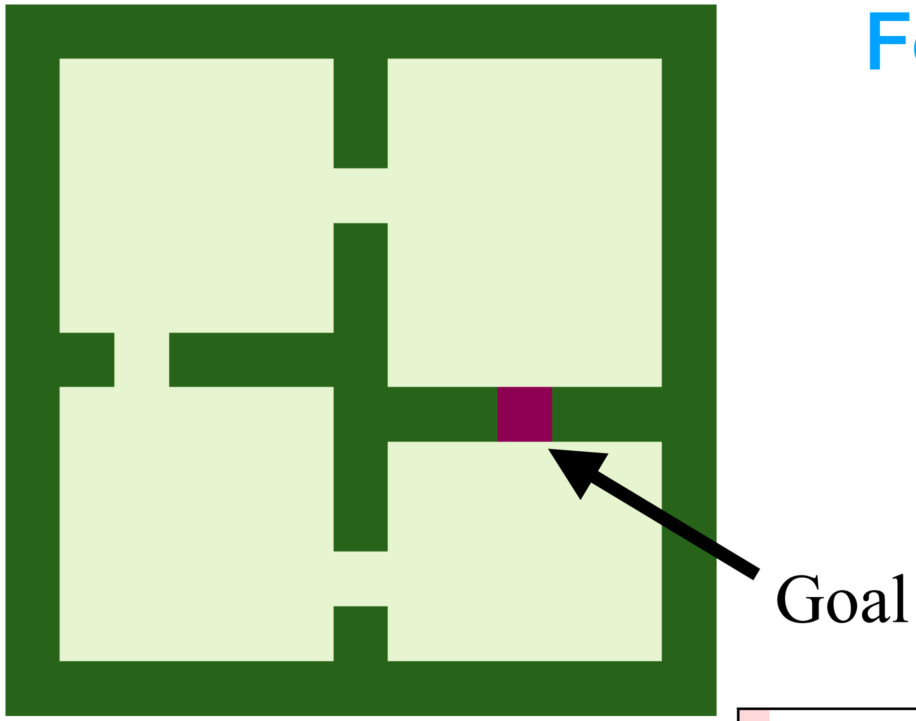


Four Rooms Domain

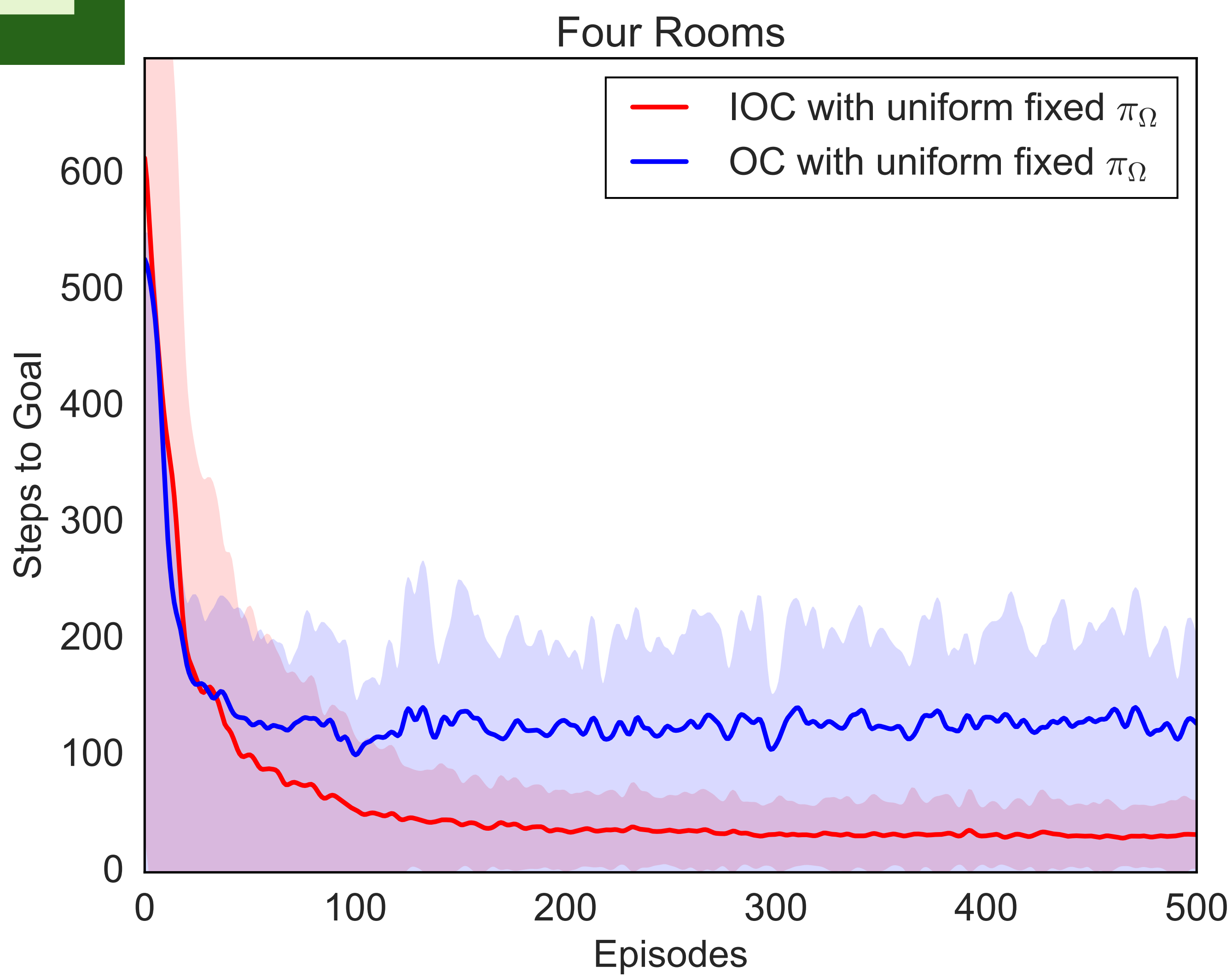
- 4 primitive actions, L, R, U, D
- Stochastic actions
- The discount factor is 0.99
- The reward is +50 at the goal and 0 otherwise.

Are interest functions useful in a single task?

Four Rooms Domain

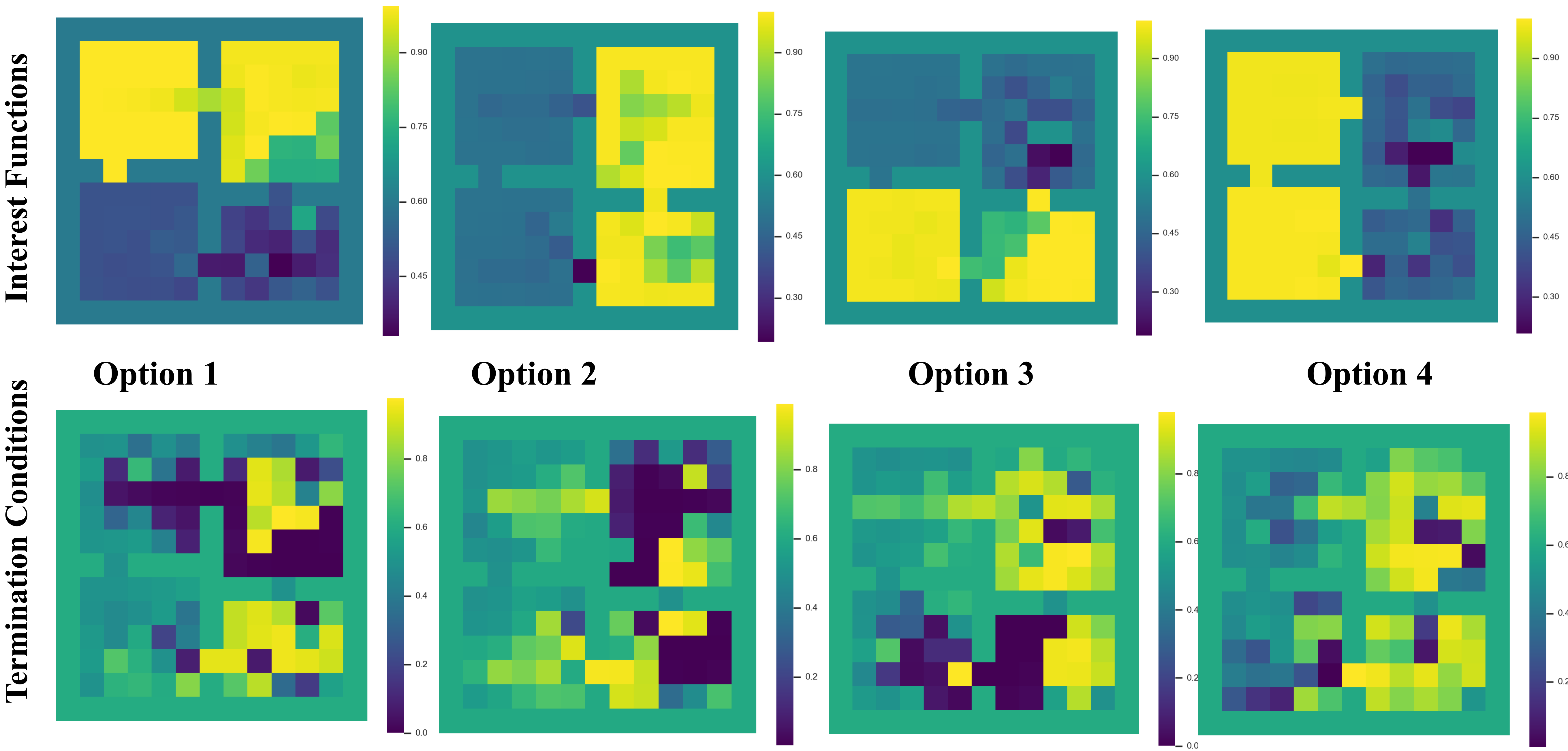


- 4 primitive actions, L, R, U, D
- Stochastic actions
- The discount factor is 0.99
- The reward is +50 at the goal and 0 otherwise.



Are interest functions useful in a single task?

Four Rooms Domain

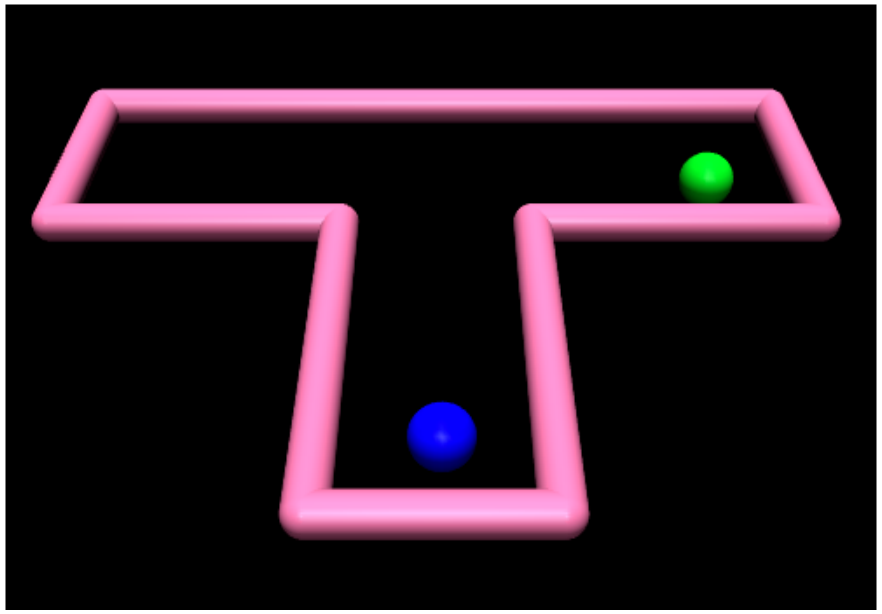


Visualization of Interest Functions at the end of 500 episodes in a task with the goal in the east hallway. Options learned with interest functions emerge with specific interest in different regions of the state space.

Visualization of Termination conditions shows that they emerge complimentary to interest of each options.

Are interest functions useful in a single task?

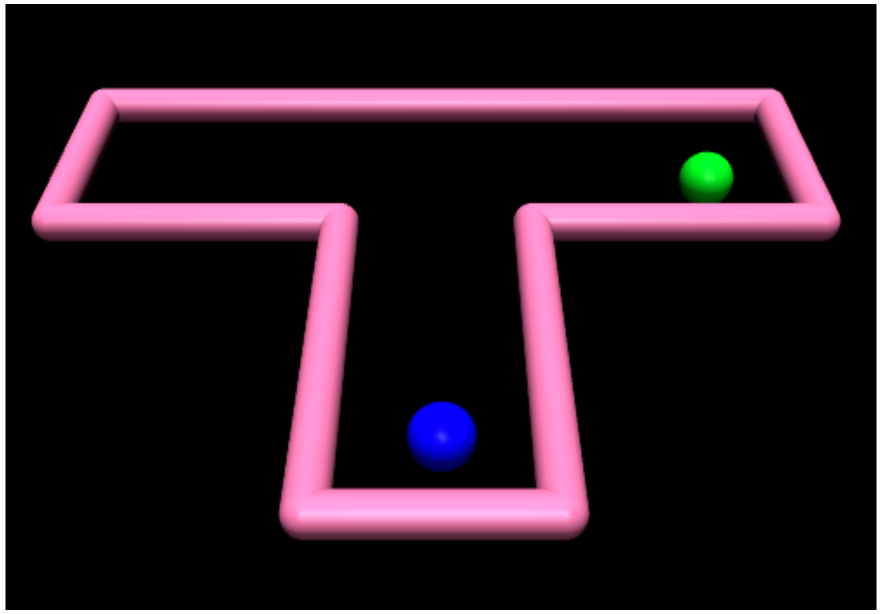
Continuous Control: Mujoco



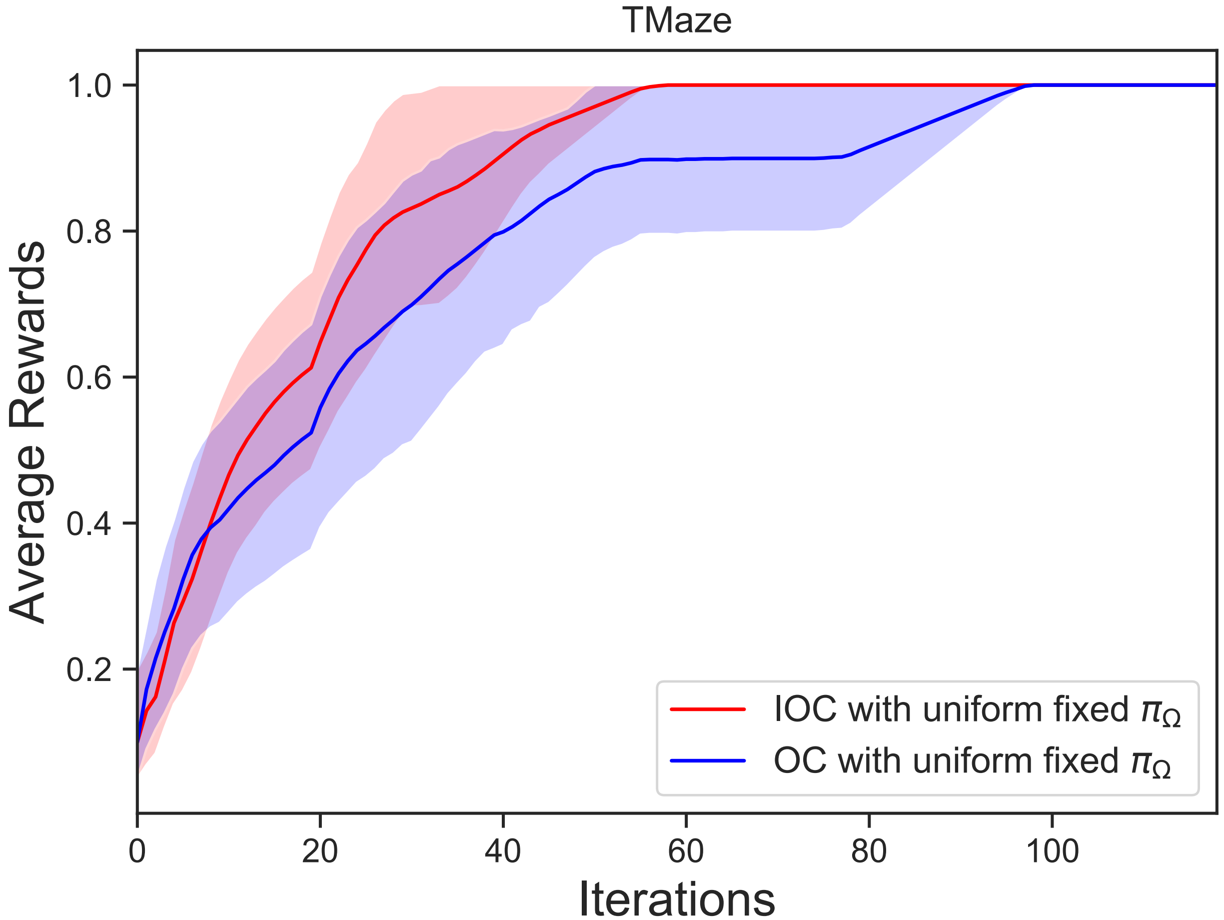
- Point mass agent (blue)
- Must navigate to the goal (green)
- State space: x, y coordinates of the agent
- Action space: Force applied in x, y directions
- Reward: +1 upon successful navigation to goal, 0 otherwise

Are interest functions useful in a single task?

Continuous Control: Mujoco

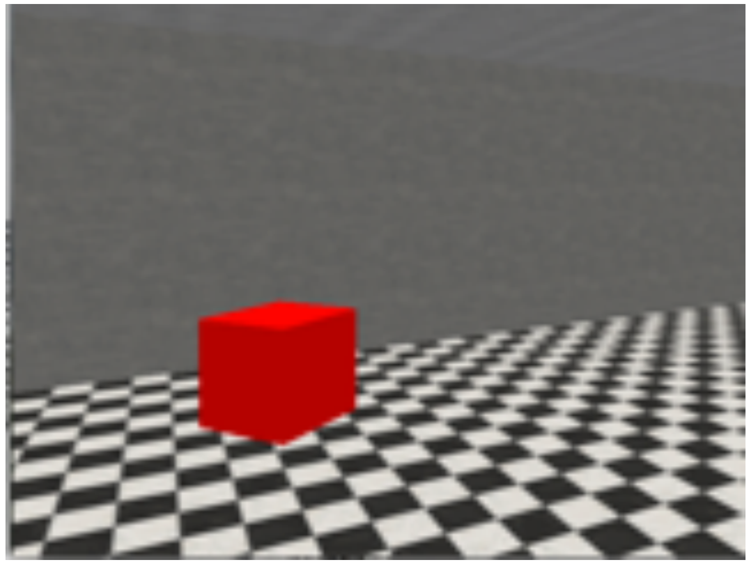


- Point mass agent (blue)
- Must navigate to the goal (green)
- State space: x, y coordinates of the agent
- Action space: Force applied in x, y directions
- Reward: +1 upon successful navigation to goal, 0 otherwise



Are interest functions useful in a single task?

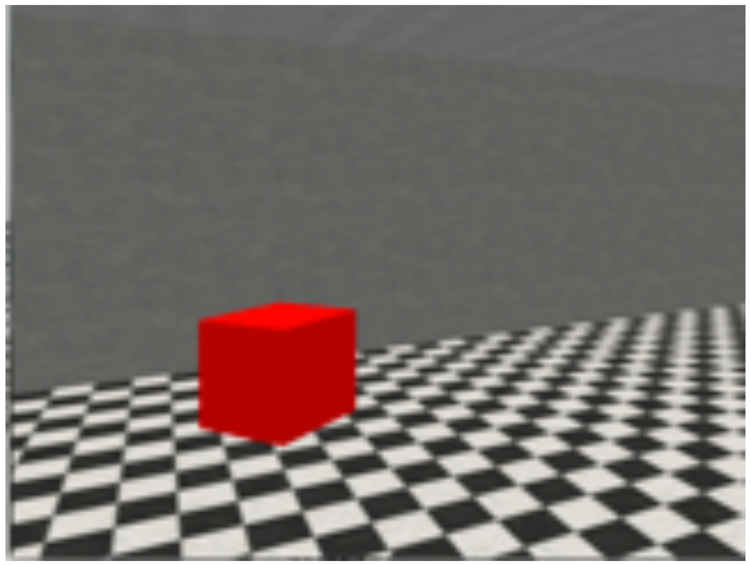
3D Visual Environment: MiniWorld



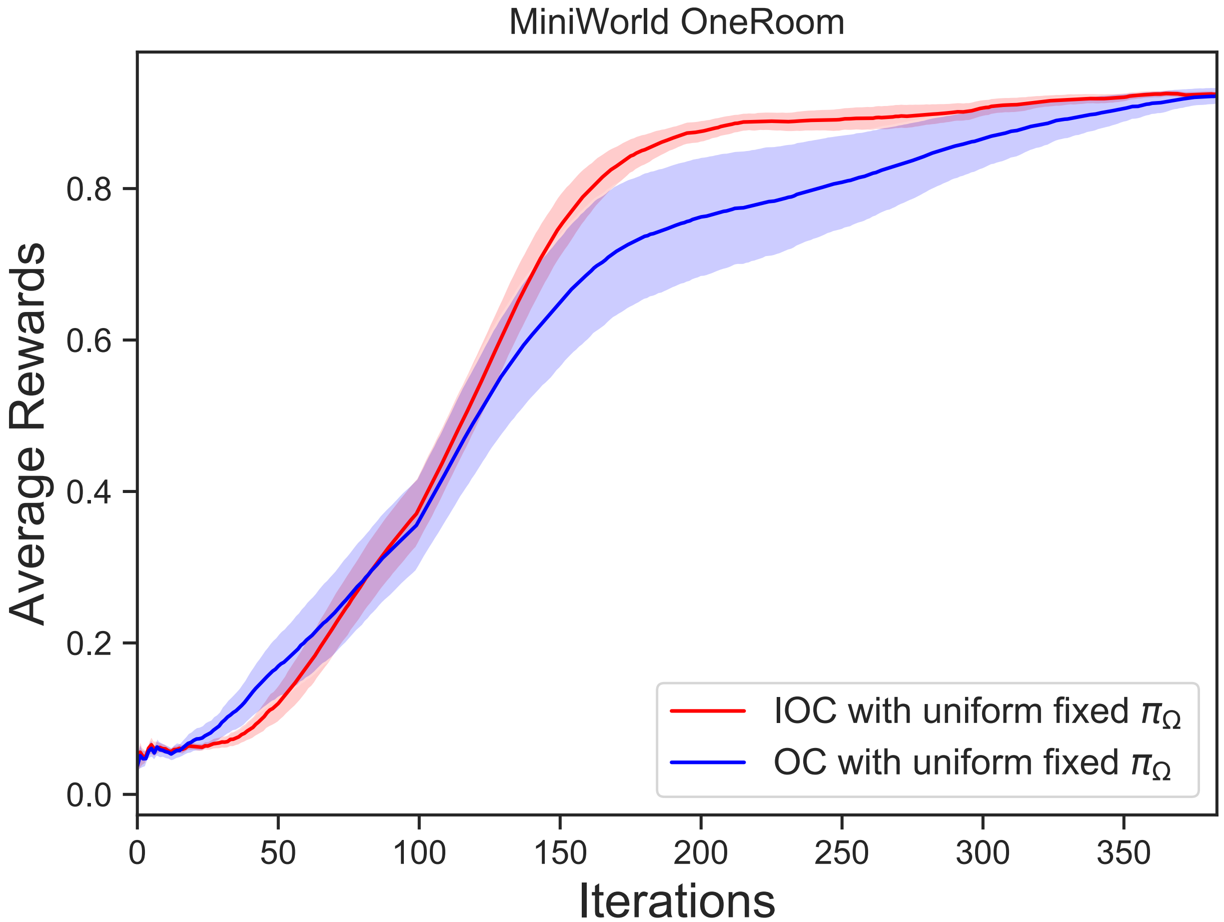
- *Oneroom* task
- Agent must navigate to a randomly placed red block in a closed room
- State space: 3-channel RGB image
- Action space: 8 discrete actions, max time steps per episode: 180
- Reward: 1.0 - 0.2 * (step_count / max_episode_steps)

Are interest functions useful in a single task?

3D Visual Environment: MiniWorld



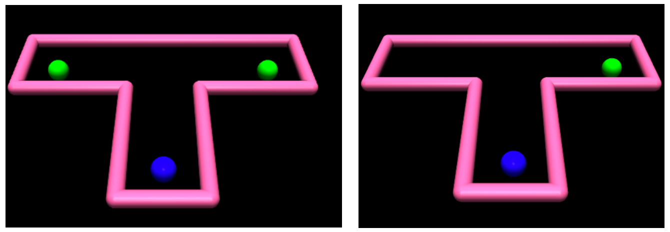
- *Oneroom* task
- Agent must navigate to a randomly placed red block in a closed room
- State space: 3-channel RGB image
- Action space: 8 discrete actions, max time steps per episode: 180
- Reward: 1.0 - 0.2 * (step_count / max_episode_steps)



Experimental Results

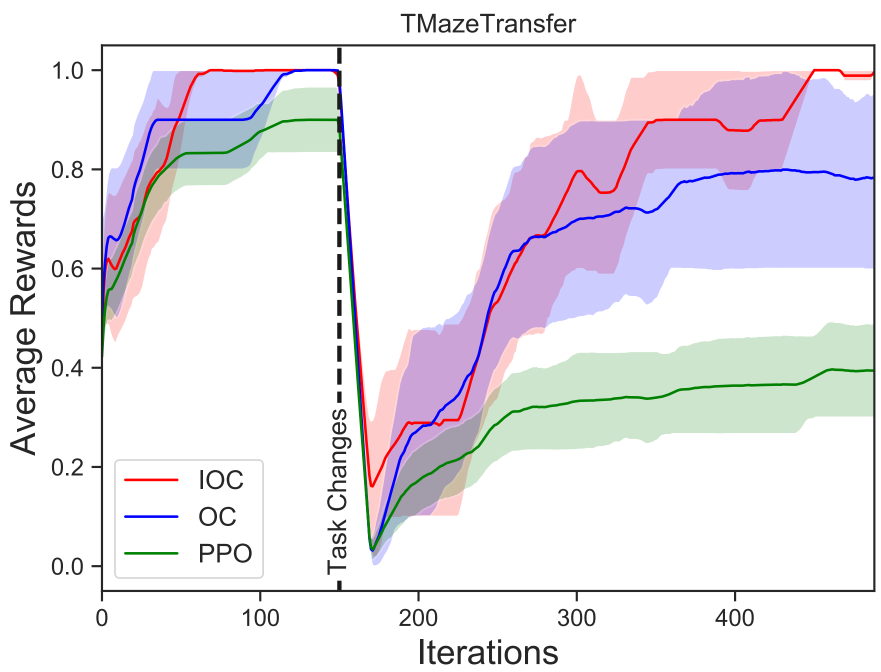
- Are options with interest functions useful in a single task?
- **Do interest functions facilitate learning reusable options?**
- Do interest functions lead to better interpretability of the learned options?

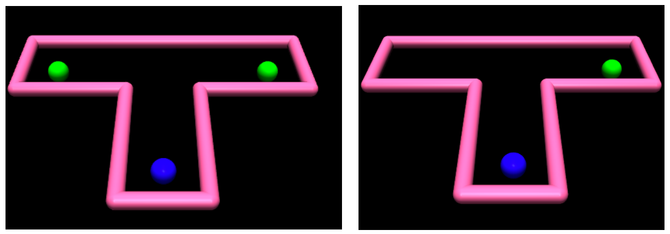
Do interest functions facilitate learning reusable options?



Reward: +1 upon successful navigation to goal, 0 otherwise, equi-rewarding goals, goal changes after 150 iterations

Do interest functions facilitate learning reusable options?

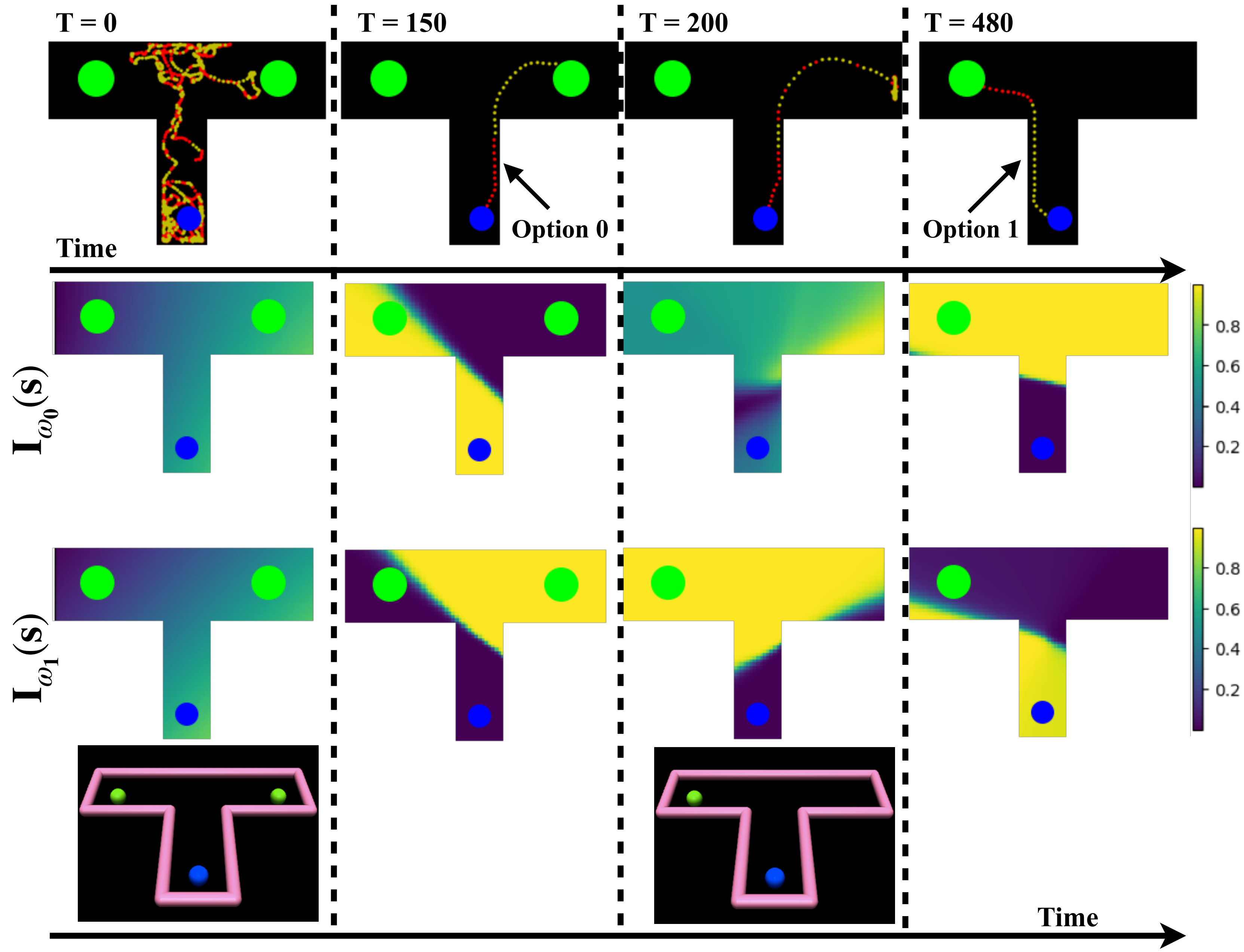


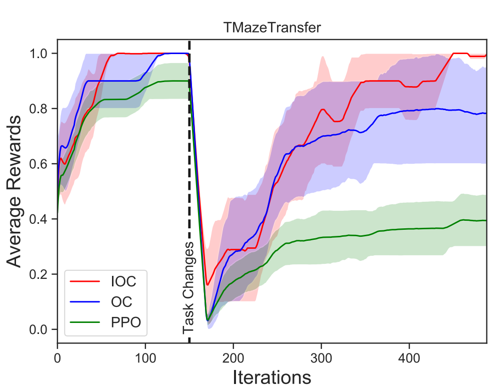


Reward: +1 upon successful navigation to goal, 0 otherwise, equi-rewarding goals, goal changes after 150 iterations

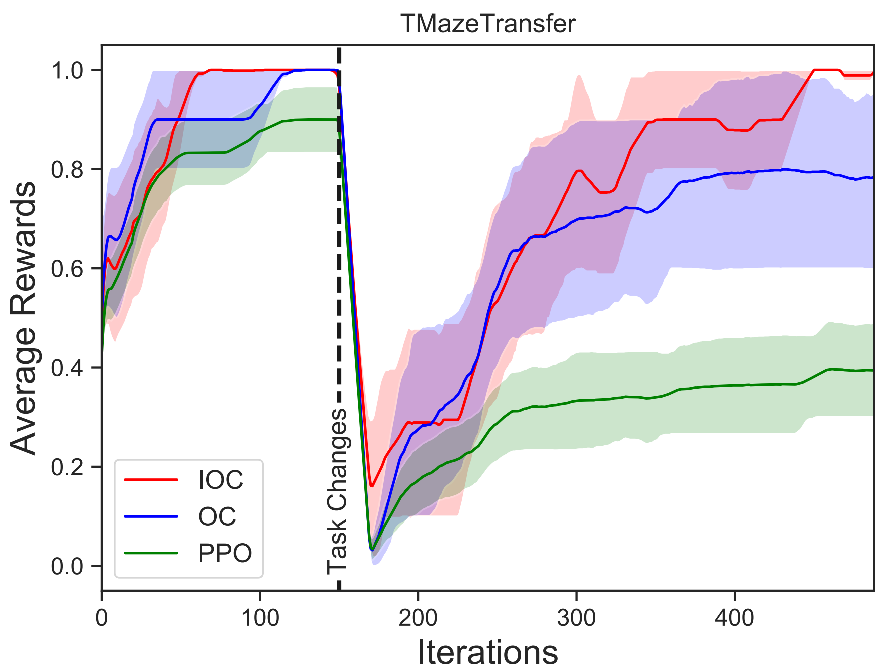
Do interest functions lead to better interpretability of learned options?

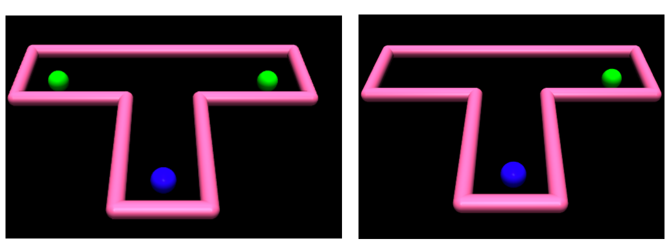
TMaze



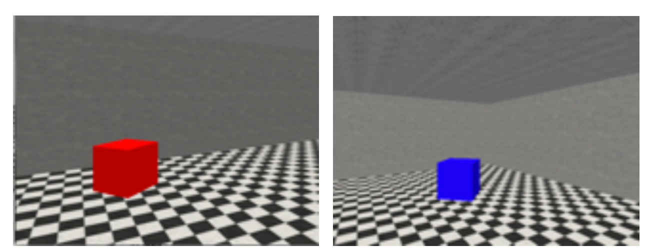


Do interest functions facilitate learning reusable options?



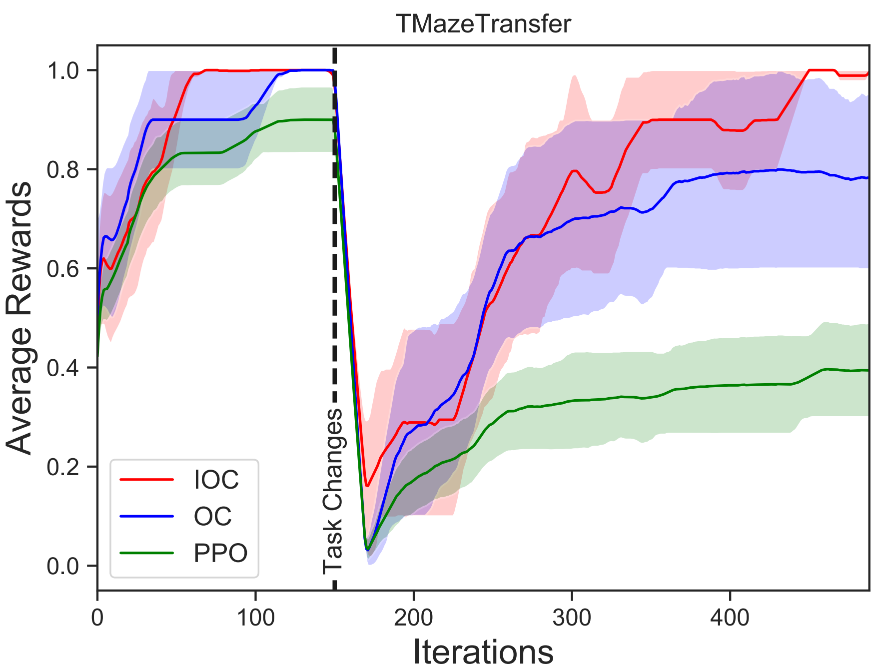


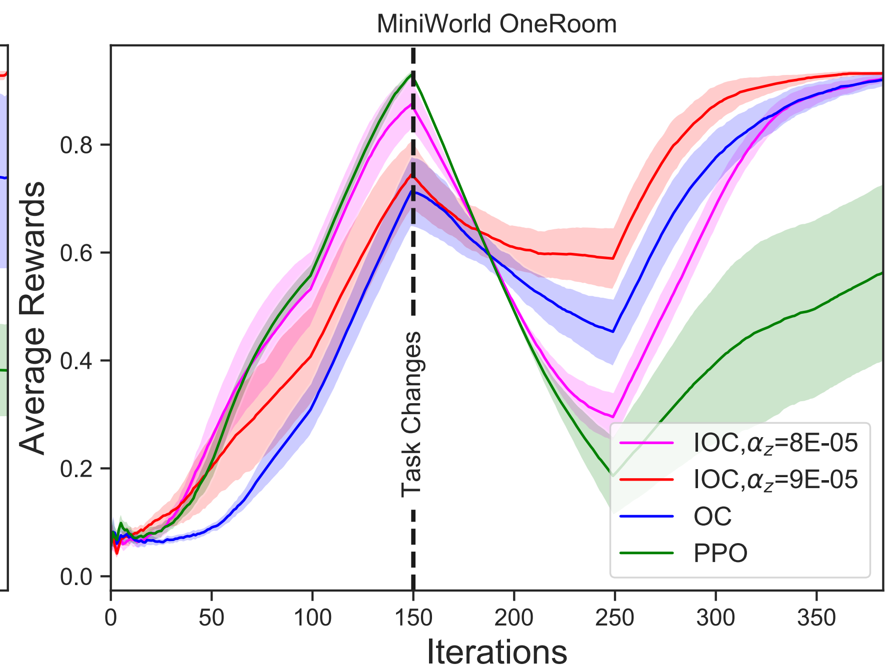
Reward: +1 upon successful navigation to goal, 0 otherwise, equi-rewarding goals, goal changes after 150 iterations

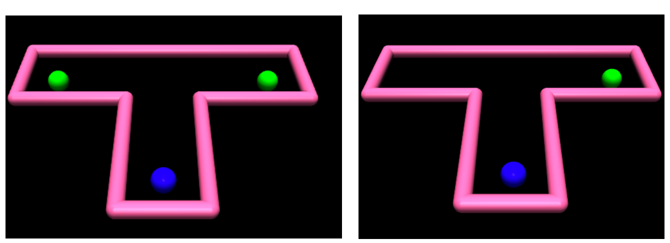


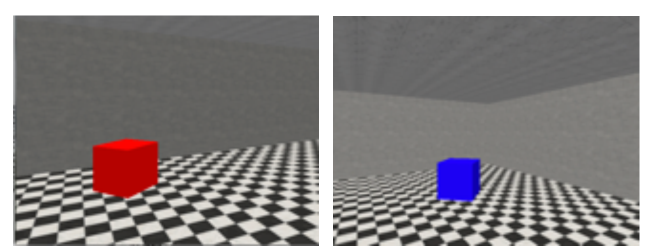
Reward: 1.0 - 0.2 * (step_count / max_episode_steps), agent needs to generalize to unseen blue box after 150 iterations

Do interest functions facilitate learning reusable options?







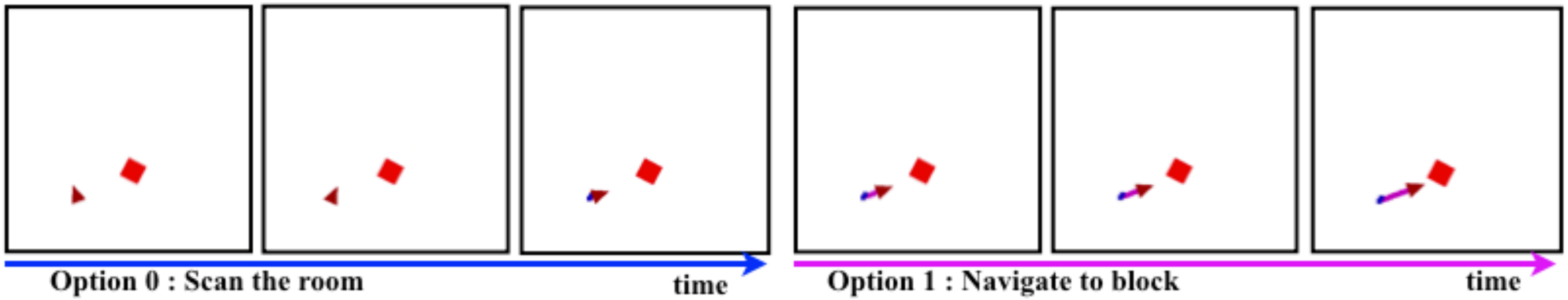


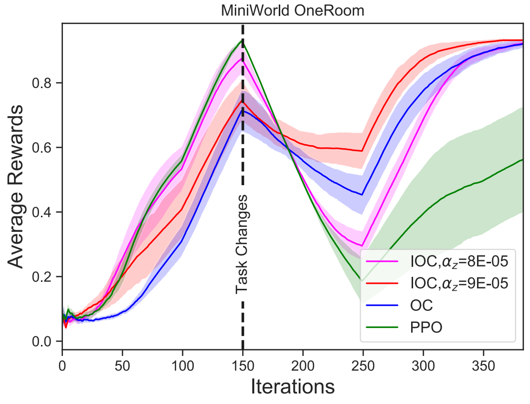
Reward: +1 upon successful navigation to goal, 0 otherwise, equi-rewarding goals, goal changes after 150 iterations

Reward: 1.0 - 0.2 * (step_count / max_episode_steps), agent needs to generalize to unseen blue box after 150 iterations

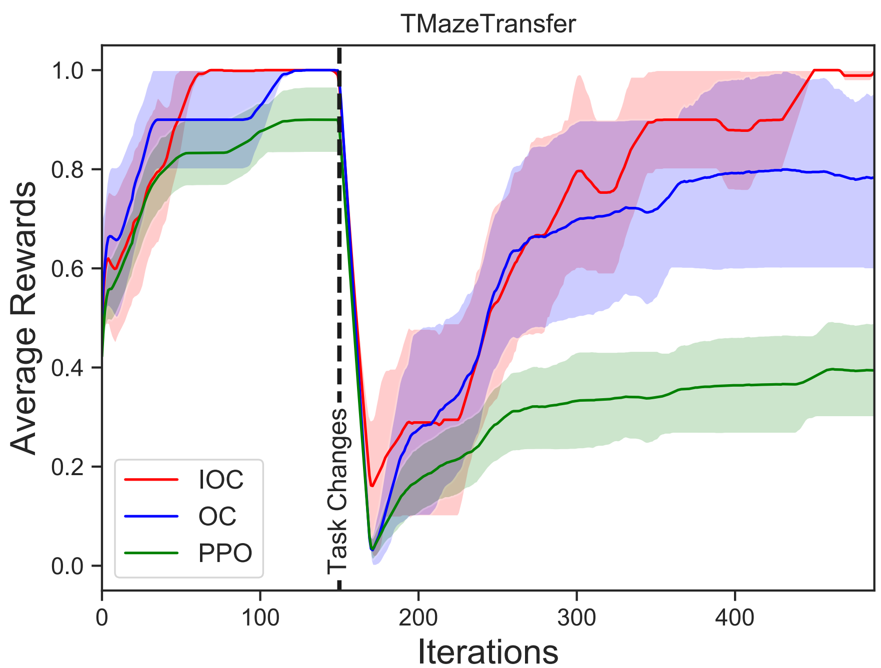
Do interest functions lead to better interpretability of learned options?

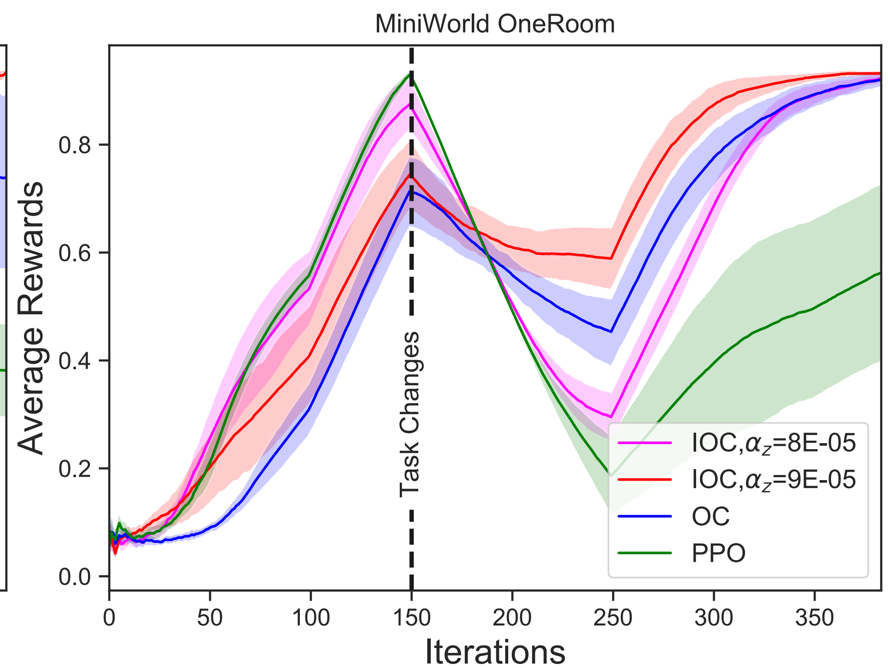
MiniWorld

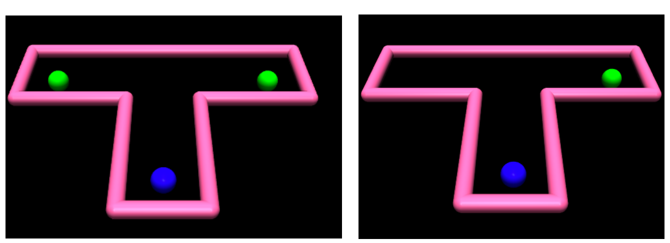




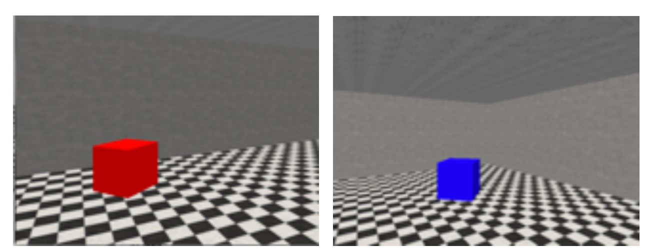
Do interest functions facilitate learning reusable options?



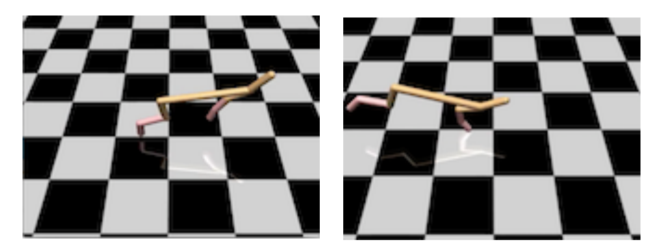




Reward: +1 upon successful navigation to goal, 0 otherwise, equi-rewarding goals, goal changes after 150 iterations

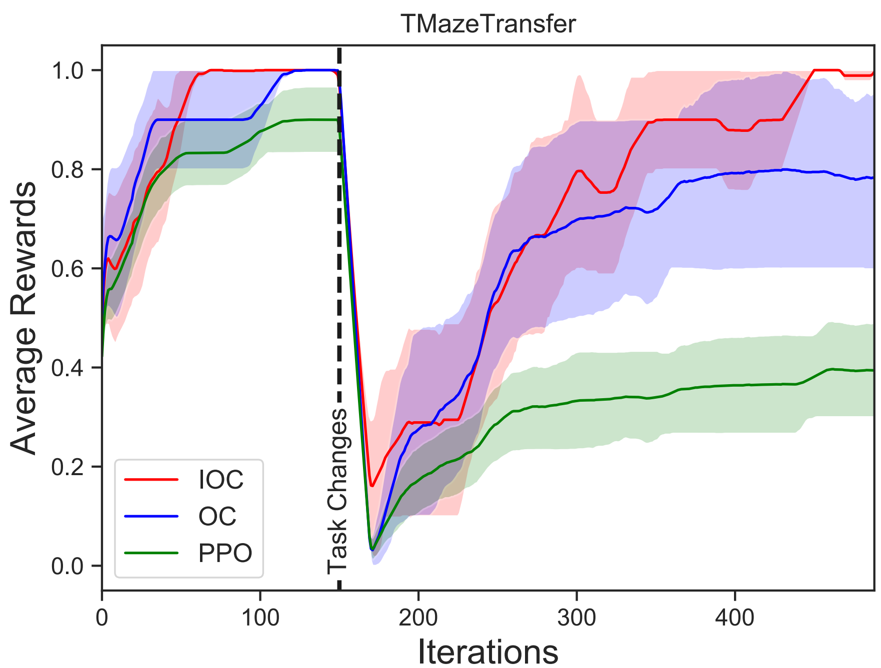


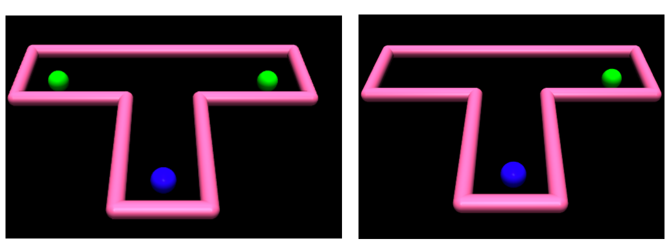
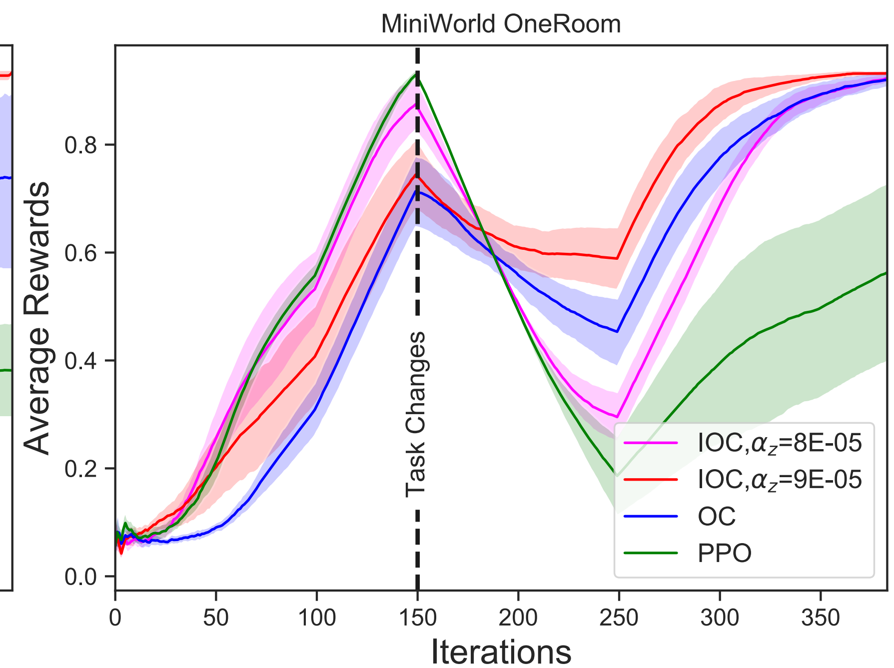
Reward: 1.0 - 0.2 * (step_count / max_episode_steps), agent needs to generalize to unseen blue box after 150 iterations

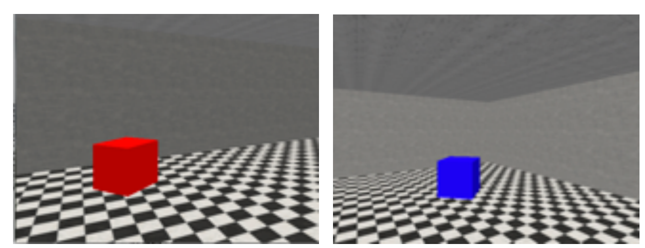
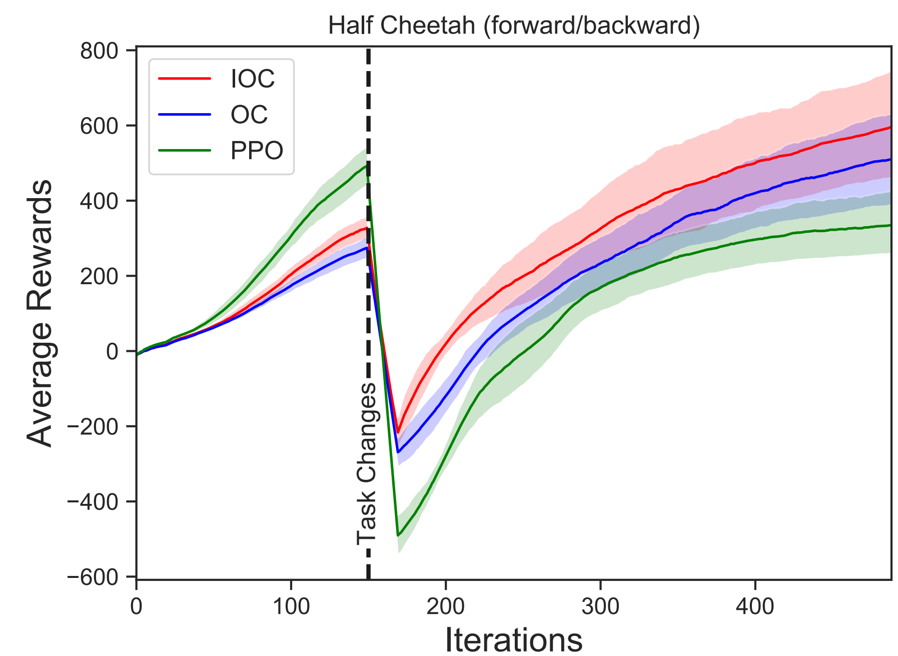


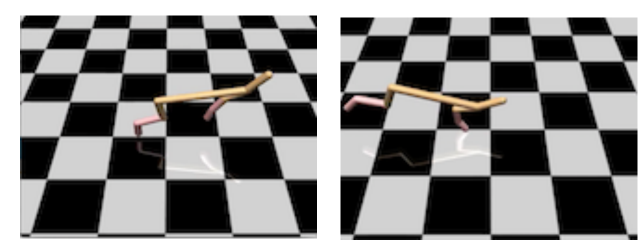
Reward: magnitude of the velocity in forward direction, after 150 iterations agent is rewarded to move backward as fast as possible with |v|

Do interest functions facilitate learning reusable options?





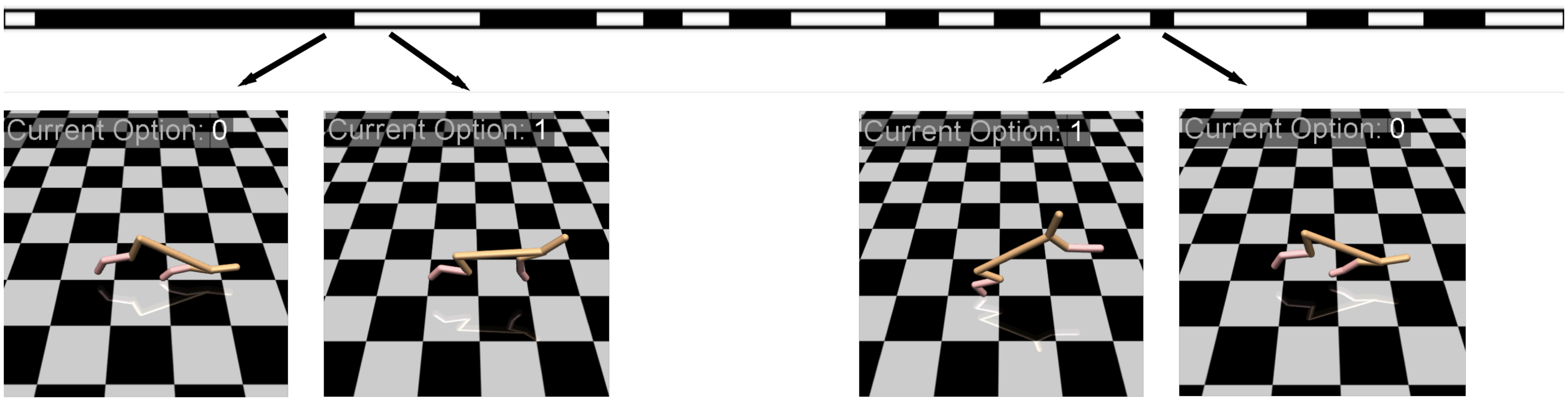
Reward: +1 upon successful navigation to goal, 0 otherwise, equi-rewarding goals, goal changes after 150 iterations

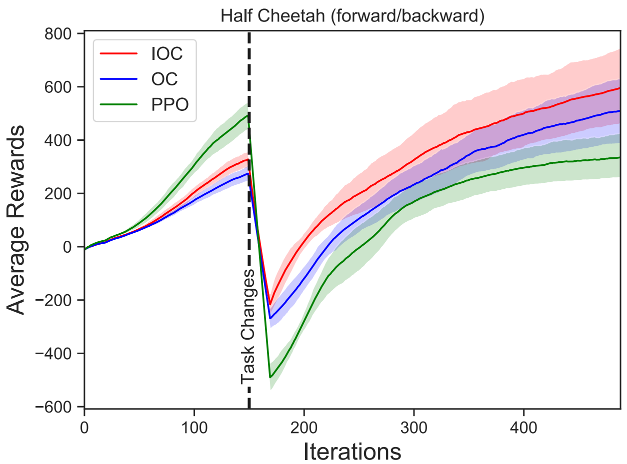
Reward: 1.0 - 0.2 * (step_count / max_episode_steps), agent needs to generalize to unseen blue box after 150 iterations

Reward: magnitude of the velocity in forward direction, after 150 iterations agent is rewarded to move backward as fast as possible with |v|

Do interest functions lead to better interpretability of learned options?

HalfCheetah



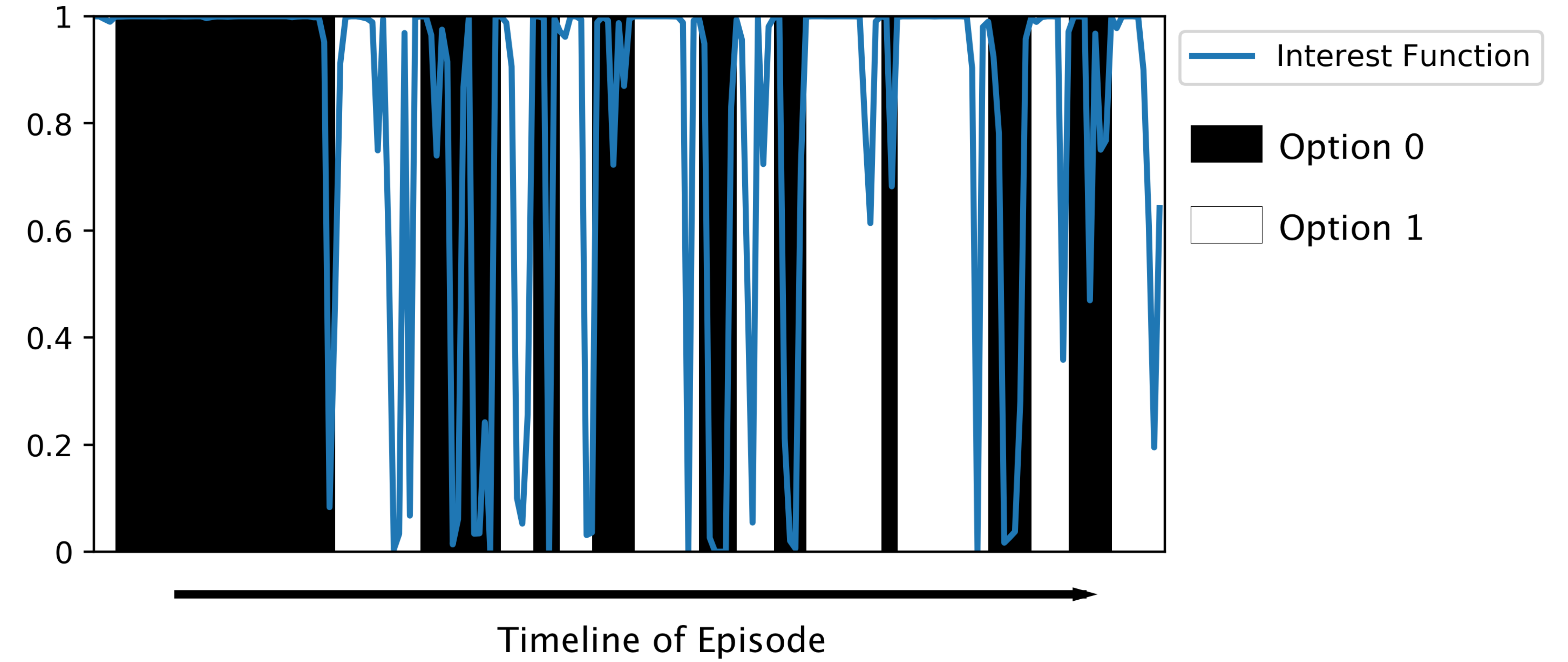


Do interest functions lead to better interpretability of learned options?

Link to videos

Interest as an Attention Mechanism

HalfCheetah



Discussion & Future Directions

- Introduced the notion of *interest functions for options*, which generalize initiation sets in a way which allows graceful learning

- Our approach is able to learn *options which are specialized*, and therefore are able to both learn faster in a single task as well as quickly adapt to changes in the task.

- To some extent, the interest functions learnt are able to override termination degeneracies as well

- Limitation: The agent optimizes a task-based external reward. Interest functions could similarly be learned driven by intrinsic task-agnostic rewards

Thank you

Extra Slides

Interest-Option-Critic

Algorithm 1: IOC with tabular intra-option Q-learning

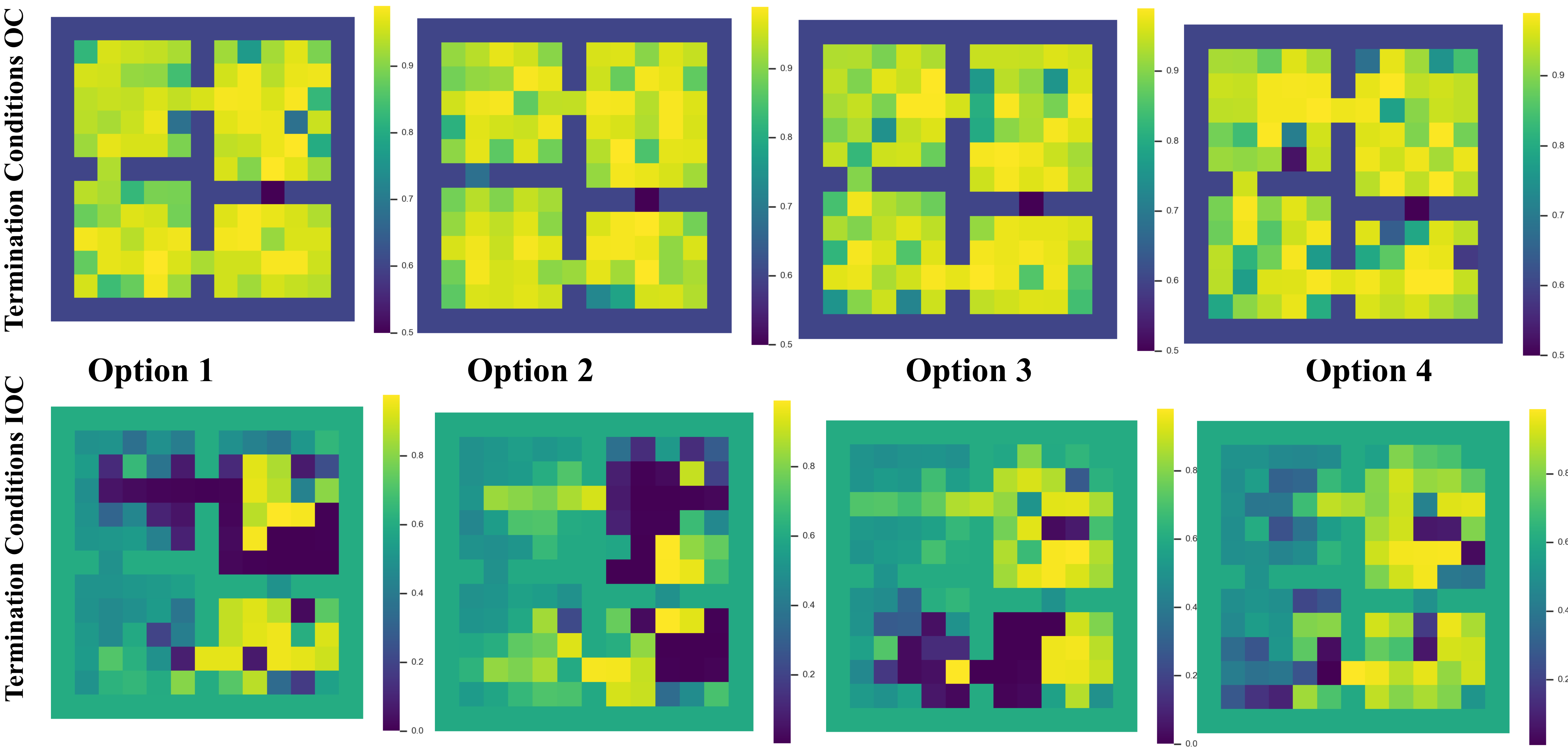
Initialize policy over options π_Ω
Initialize $I_{\omega,z}$ parameterized by z such that all options are available everywhere to some extent
Initialize $\pi_{I_{\omega,z}}(\omega|s)$ as in Eq.(1)
Set $s \longleftarrow s_0$ and ω at s according to $\pi_{I_{\omega,z}}$
repeat
  Choose a according to $\pi_{\omega,\theta}(a|s)$
  Take action a in s, observe s', r
  Sample termination from $\beta_\omega, \nu(s')$
  if ω terminates in s' **then**
    Sample ω' according to $\pi_{I_{\omega,z}}(\cdot|s')$
  else
    $\omega' = \omega$
  end if
  1. Evaluation step:
  $\delta \leftarrow r - Q_U(s,\omega,a)$
  $\delta \leftarrow r + \gamma(1-\beta_{\omega,\nu}(s'))Q_\Omega(s',\omega) + \gamma\beta_{\omega,\nu}(s')\max_{\omega'} Q_\Omega(s',\omega')$
  $Q_U(s,\omega,a) \leftarrow Q_U(s,\omega,a) + \alpha\delta$
  2. Improvement step
  $\theta \leftarrow \theta + \alpha_\theta \frac{\partial \log \pi_{\omega,\theta}(a|s)}{\partial\theta} Q_U(s,\omega,a)$
  $\nu \leftarrow \nu - \alpha_\nu \frac{\partial\beta_{\omega,\nu}(s')}{\partial\nu}(Q_\Omega(s',\omega) - V_\Omega(s'))$ where$V_\Omega(s') = \sum'_\omega \pi_{I_{\omega,z}}(\omega'|s')Q_\Omega(s',\omega')$
  $z \leftarrow z + \alpha_z \beta_{\omega,\nu}(s') \frac{\partial\pi_{I_{\omega,z}}(\omega'|s')}{\partial z} Q_\Omega(s',\omega')$
  $s \leftarrow s'$
until s' is a terminal state

$$\left(\mathbf{I}_{\omega,\mathbf{z}} : \mathbf{S}\times\boldsymbol{\Omega} \to \mathbb{R}^+,\ \pi_{\omega,\theta} : \mathbf{S}\times\mathbf{A} \to [\mathbf{0},\mathbf{1}],\ \beta_{\omega,\nu} : \mathbf{S} \to [\mathbf{0},\mathbf{1}]\right)$$

Interest Functions **Intra option policy** **Termination condition**

Are interest functions useful in a single task?

Four Rooms Domain



Options learned in OC terminate almost everywhere as all options are applicable in all states.

Interest-Option-Critic

The option value function is defined as

$$Q_\Omega(s,\omega) = \sum_a \pi_{\omega,\theta}(a\,|\,s) Q_U(s,\omega,a)$$

Taking the derivation w.r.t. z

$$\frac{\partial Q_\Omega(s,\omega)}{\partial z} = \frac{\partial}{\partial z}\left\{ \sum_a \pi_{\omega,\theta}(a\,|\,s) Q_U(s,\omega,a) \right\}$$

$$= \sum_a \pi_{\omega,\theta}(a\,|\,s) \sum_{s'} \gamma P(s'\,|\,s,a) \left\{ (1-\beta_{\omega,\nu}(s'))\frac{\partial Q_\Omega(s',\omega)}{\partial z} + \beta_{\omega,\nu}(s')\frac{\partial V_\Omega(s')}{\partial z} \right\}$$

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

$$\frac{\partial V_\Omega(s')}{\partial z} = \sum_\omega \left(\frac{\partial \pi_{I_{\omega,z}}(\omega\,|\,s')}{\partial z} Q_\Omega(s',\omega) + \pi_{I_{\omega,z}}(\omega\,|\,s')\frac{\partial Q_\Omega(s',\omega)}{\partial z} \right)$$

Interest-Option-Critic

$$\frac{\partial Q_\Omega(s,\omega)}{\partial z} = \frac{\partial}{\partial z}\left\{ \sum_a \pi_{\omega,\theta}(a\,|\,s) Q_U(s,\omega,a) \right\}$$

$$= \sum_a \pi_{\omega,\theta}(a\,|\,s) \sum_{s'} \gamma P(s'\,|\,s,a) \left\{ (1-\beta_{\omega,\nu}(s')) \frac{\partial Q_\Omega(s',\omega)}{\partial z} + \beta_{\omega,\nu}(s') \frac{\partial V_\Omega(s')}{\partial z} \right\}$$

……

$$= \sum_a \pi_{\omega,\theta}(a\,|\,s) \sum_{s'} \gamma P(s'\,|\,s,a) \sum_{\omega'} \beta_{\omega,\nu}(s') \frac{\partial \pi_{I_{\omega,z}}(\omega'\,|\,s')}{\partial z} Q_\Omega(s',\omega') + \sum_{s'} \sum_{\omega'} \left(\sum_a \pi_{\omega,\theta}(a\,|\,s) \gamma P(s'\,|\,s,a) \Big((1-\beta_{\omega,\nu}(s')) + \beta_{\omega,\nu}(s') \pi_{I_{\omega,z}}(\omega'\,|\,s') \Big) \right) \frac{\partial Q_\Omega(s',\omega')}{\partial z}$$

In the above equation, one-step discounted transition probability in the augmented space is given as

$$P_\gamma^{(1)}(s',\omega'\,|\,s,\omega) = \sum_a \pi_{\omega,\theta}(a\,|\,s) \gamma P(s'\,|\,s,a) \Big((1-\beta_{\omega,\nu}(s')) 1_{\omega=\omega'} + \beta_{\omega,\nu}(s') \pi_{I_{\omega,z}}(\omega'\,|\,s') \Big)$$

……

……

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