Learning Generalized Temporal Abstractions Across Both Action and Perception

Khimya Khetarpal

Ph.D. Advisor: Doina Precup

Reasoning and Learning Lab
Mila - McGill University

AAAI Doctoral Consortium 2019
Mentor: Michael Littman

Overview

- **Research goals**
- **Temporal abstraction**
- **Theme I: Learning options with interest functions**
- **Theme II: Learning temporal abstractions across action and perception**
- **Timeline**

Overview

- Research goals
- Temporal abstraction
- Theme I: Learning options with interest functions
- Theme II: Learning temporal abstractions across action and perception
- Timeline

Research Goals

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

Overview

- Research goals
- Temporal abstraction
- Theme I: Learning options with interest functions
- Theme II: Learning temporal abstractions across action and perception
- Timeline

Temporal Abstraction

Consider a simple morning routine of preparing breakfast.

Higher level steps

Choosing the kind of eggs, the type of toast

Medium level steps

Chop vegetables, get butter, put ingredients in a skillet

Low level steps

Wrist and arm movements in chopping vegetables, etc.

Temporal Abstraction

Consider a simple morning routine of preparing breakfast.

Higher level steps

Choosing the choice of eggs, the type of toast

Medium level steps

Chop vegetables, Get butter, Put ingredients in a skillet, toast bread

Low level steps

Wrist and arm movements in chopping vegetables, making eggs, etc.

The ability to abstract knowledge temporally over many different time scales is seamlessly integrated in human decision making!

Why Temporal Abstraction

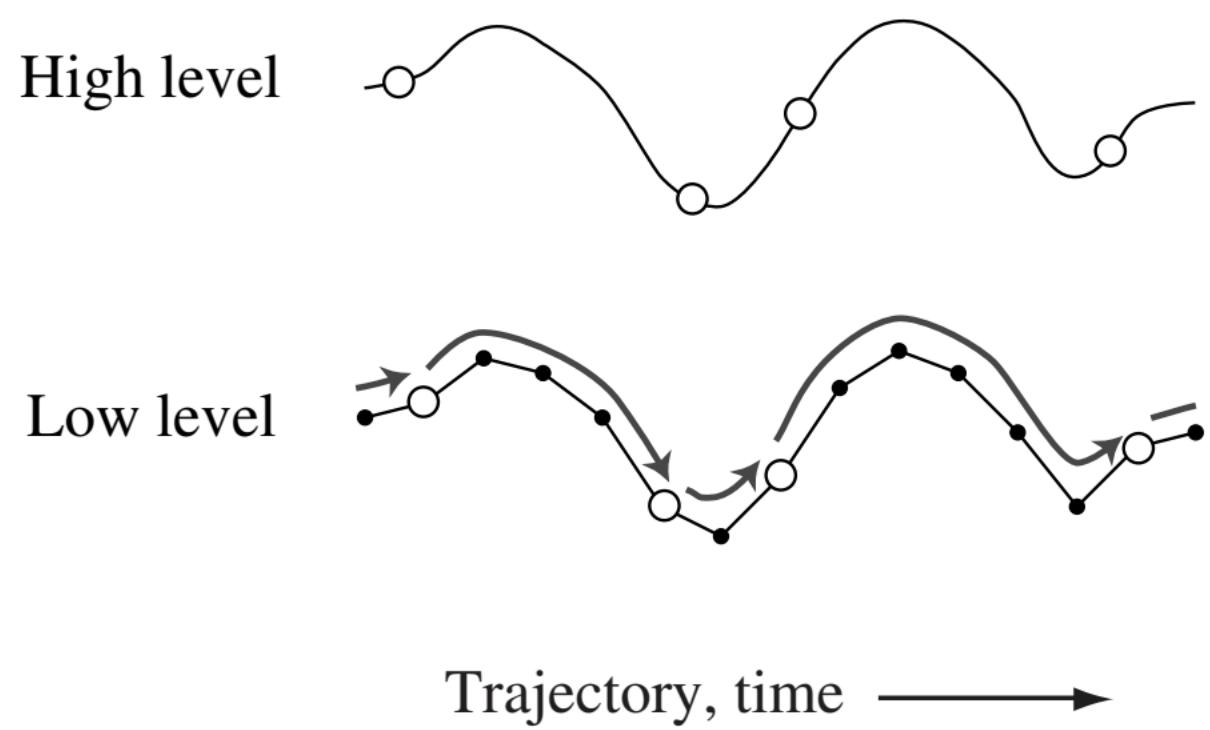
It has been shown to:

- Reduce the complexity of choosing actions
- Generate shorter plans
- Improve exploration by taking shortcuts in the environment
- Help in transfer learning

Slide Source: Pierre-Luc Bacon

Temporal abstraction in RL

Options (Sutton, Precup, and Singh, 1999) formalize the idea of temporally extended actions also known as **skills.**



Slide Source: Pierre-Luc Bacon

Options Framework

- Definition

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

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

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 $\pi_\Omega : S \times \Omega \rightarrow [0,1]$

- 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

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

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

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

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

Assumption: All options are available in all states

This is counterintuitive, leads to degeneracies and options learned lack in meaning.

Overview

- **Research goals**
- **Temporal abstraction**
- **Theme I: Learning options with interest functions**
- **Theme II: Learning temporal abstractions across action and perception**
- **Timeline**

Theme I: Learning options with interest functions

Hypothesis:

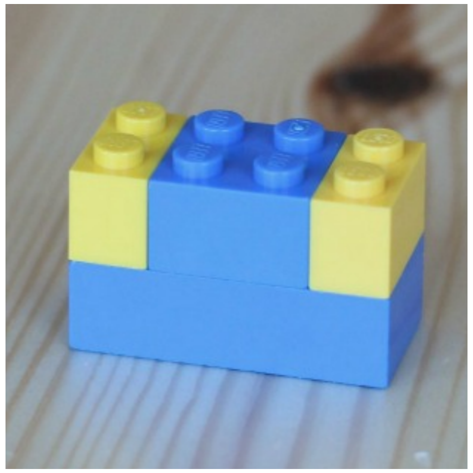
Learning options which are specialized in situations of specific interest can be leveraged to learn <u>meaningful</u>, <u>interpretable</u> and <u>reusable</u> temporal abstractions.

Theme I: Learning options with interest functions

Motivation:

- Just like humans acquire skills, reuse and build on top of already existing skills to solve more complicated tasks

- AI agents should be able to learn and develop skills ***continually, hierarchically*** *and* ***incrementally*** over time [referred as continual / lifelong learning]




Theme I: Learning options with interest functions

Motivation:

Theme I: Learning options with interest functions

Motivation:

Open the door

Go to the kitchen

Exit Hall

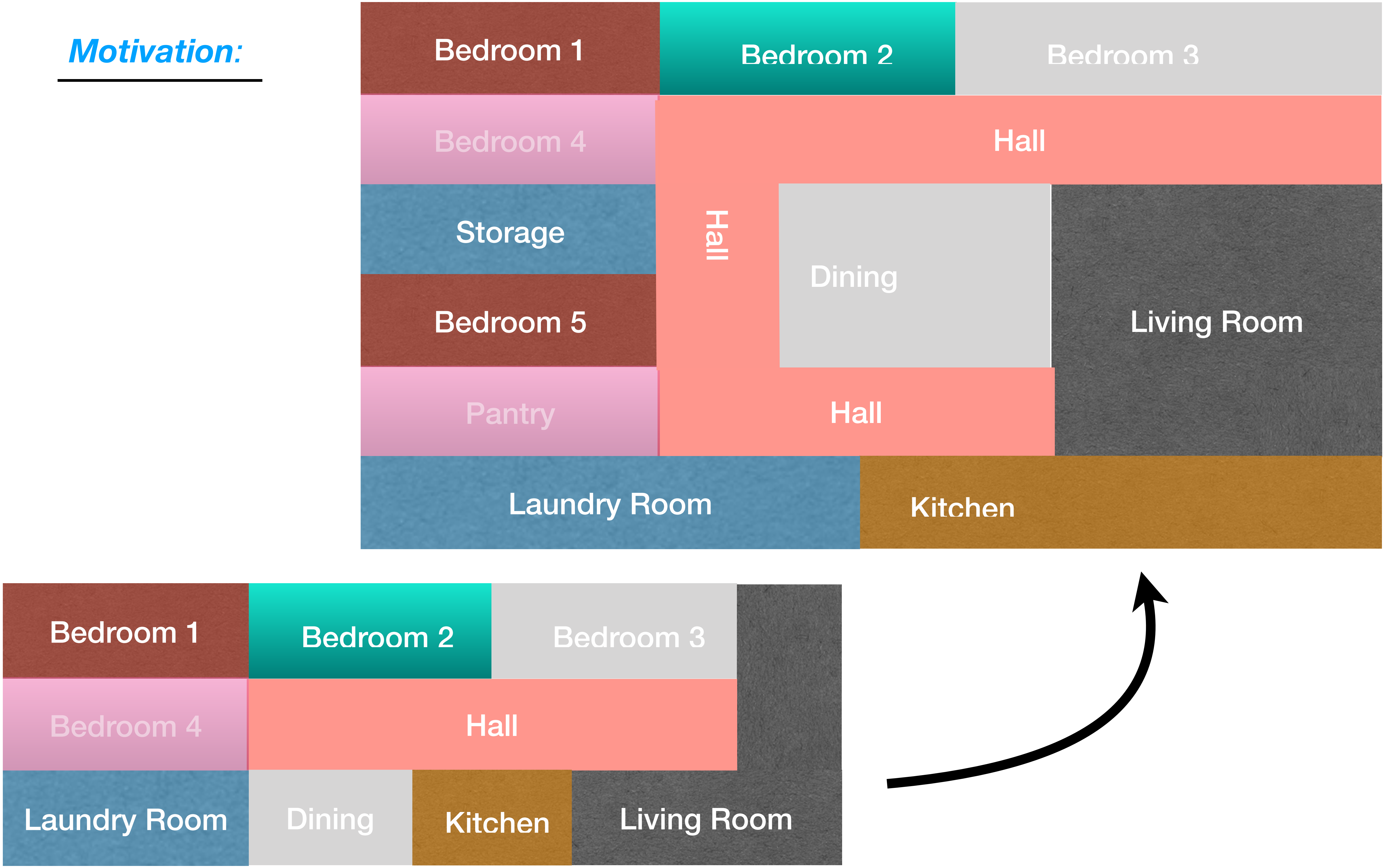
Go to living room

Find laundry room

Fetch phone from Bedroom 2

Theme I: Learning options with interest functions

Motivation:



Theme I: Learning options with interest functions

- Break the assumption that all options are present in all states.

- **Definition**: Interest Function $\mathbf{I}_{\omega,\mathbf{z}} : \mathbf{S} \times \mathbf{O} \longrightarrow \mathbb{R}^{+}$ is an indication of the extent to which an option ω is interested in a state s.

- Here we consider differentiable interest functions parameterized with z.

Theme I: Learning options with interest functions

Formulation

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

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

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

Theme I: Learning options with interest functions

Formulation

- 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.

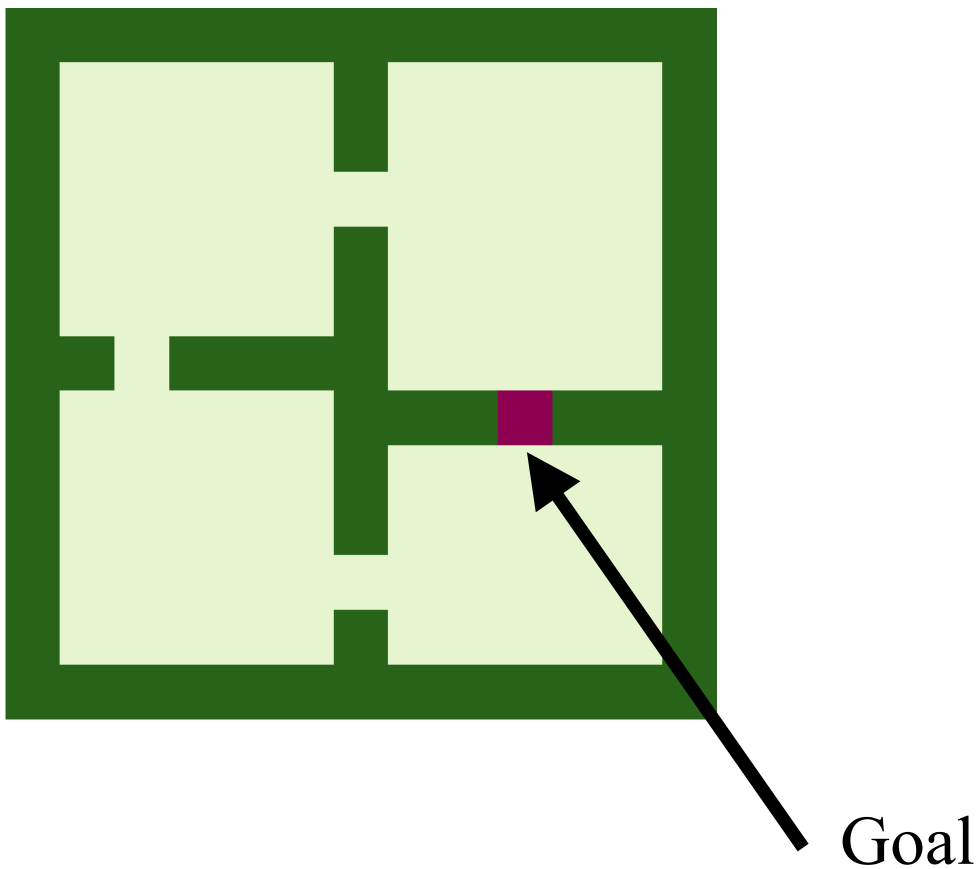
- **TL;DR** all three components of options are parameterized and learned

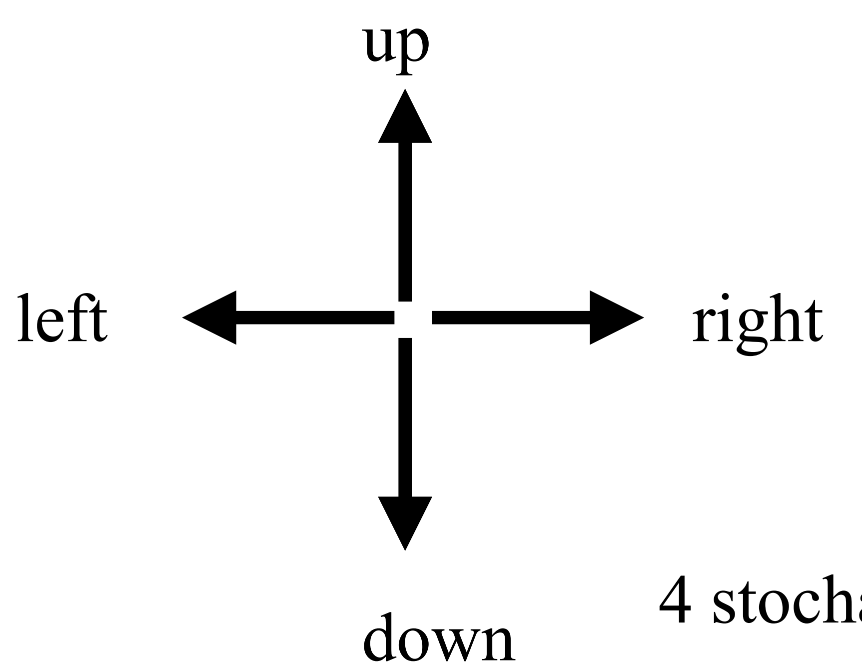
$$\left(\mathbf{I}_\omega \subseteq \mathbf{S} \ , \ \pi_\omega : \mathbf{S} \times \mathbf{A} \to [\mathbf{0}, \mathbf{1}] \ , \ \beta_\omega : \mathbf{S} \to [\mathbf{0}, \mathbf{1}]\right)$$

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

Theme I: Learning options with interest functions

Four Rooms Domain

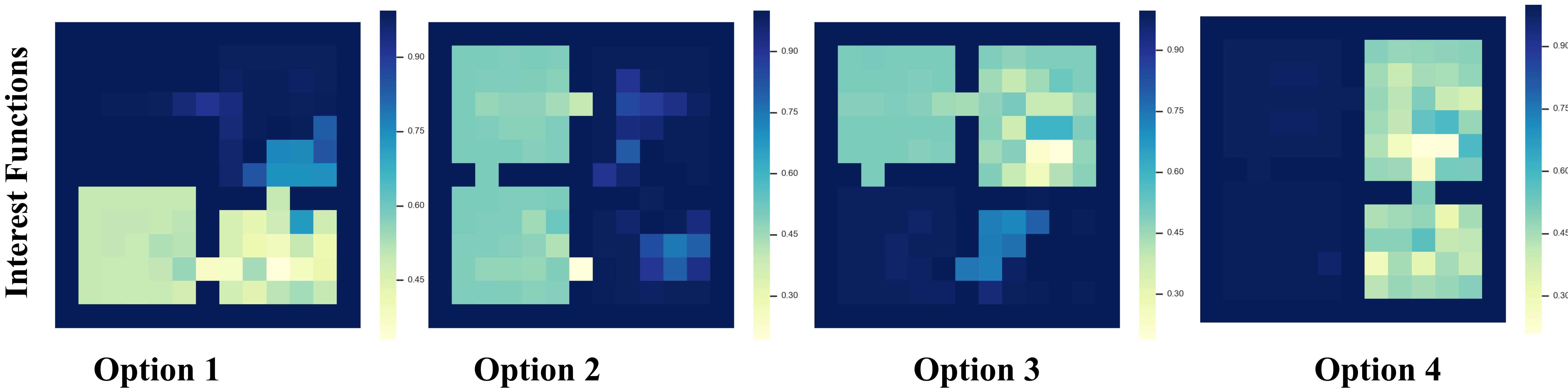




4 stochastic primitive actions

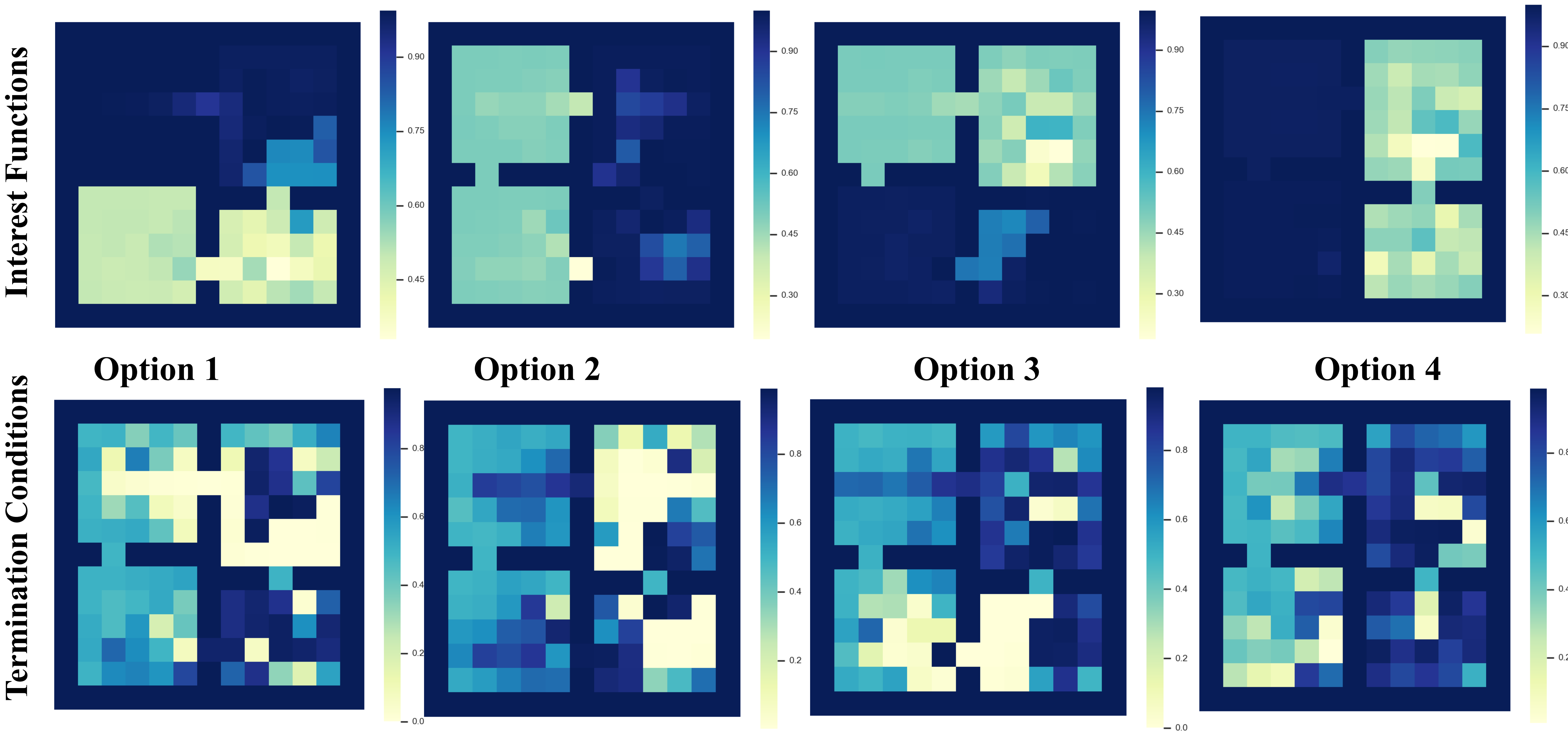
Theme I: Learning options with interest functions

Four Rooms Domain



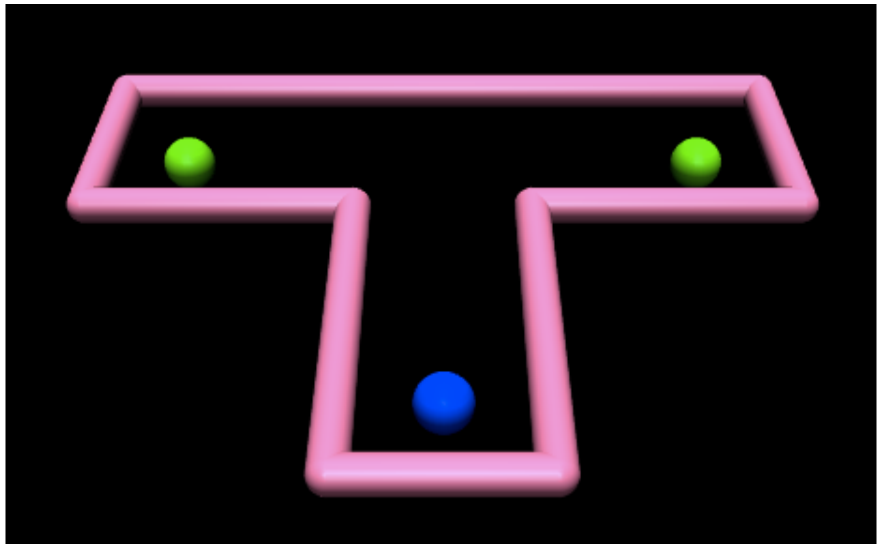
Theme I: Learning options with interest functions

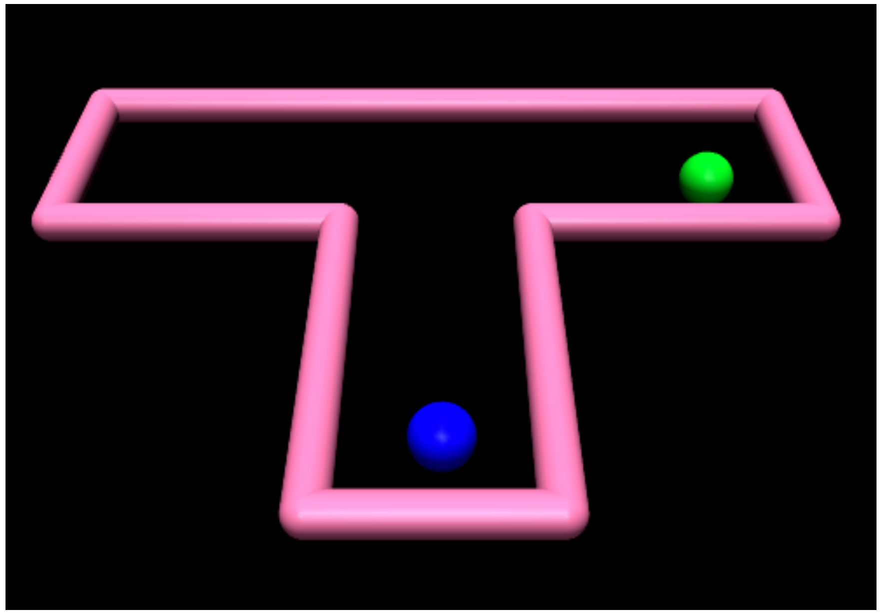
Four Rooms Domain

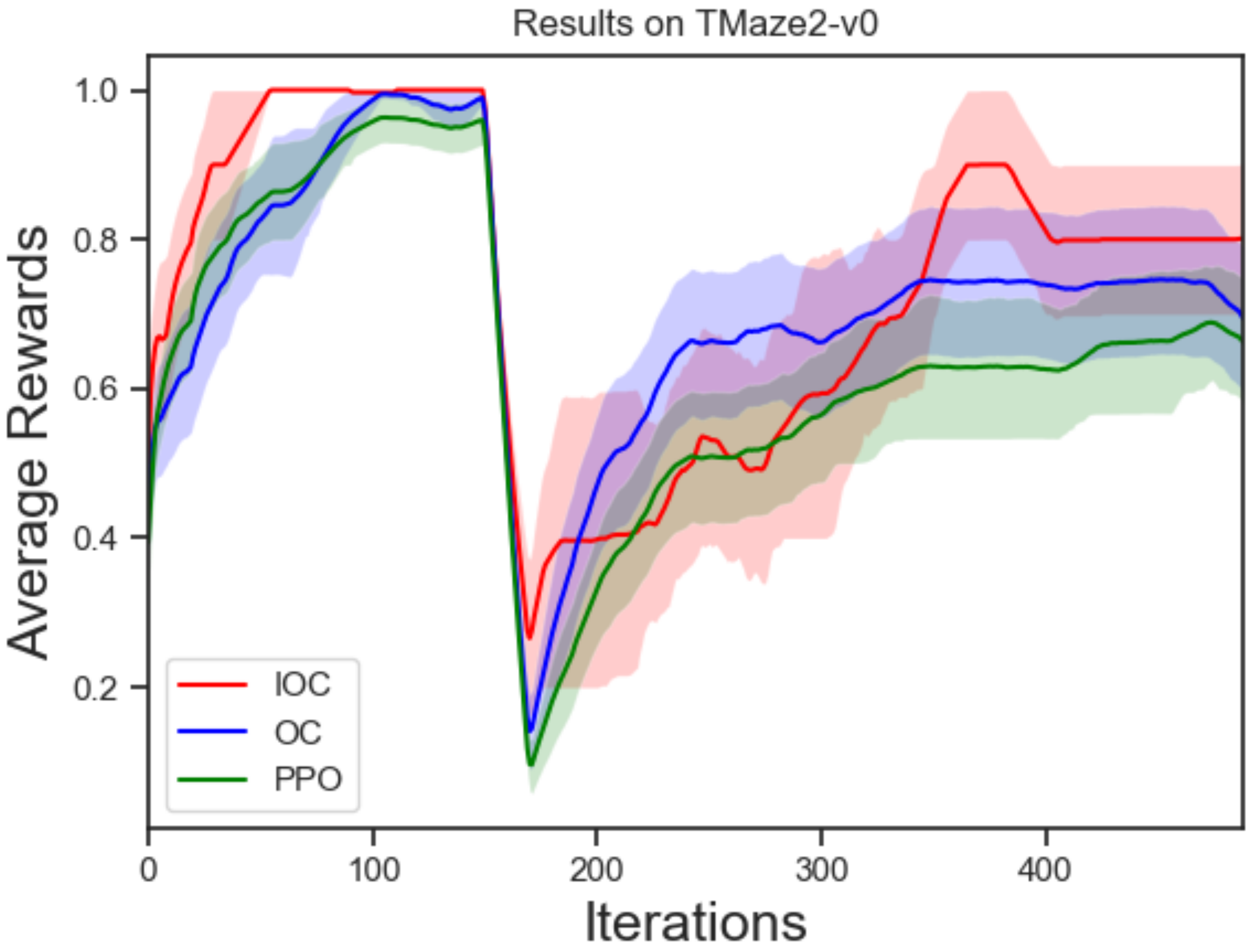


Theme I: Learning options with interest functions

Continuous Control

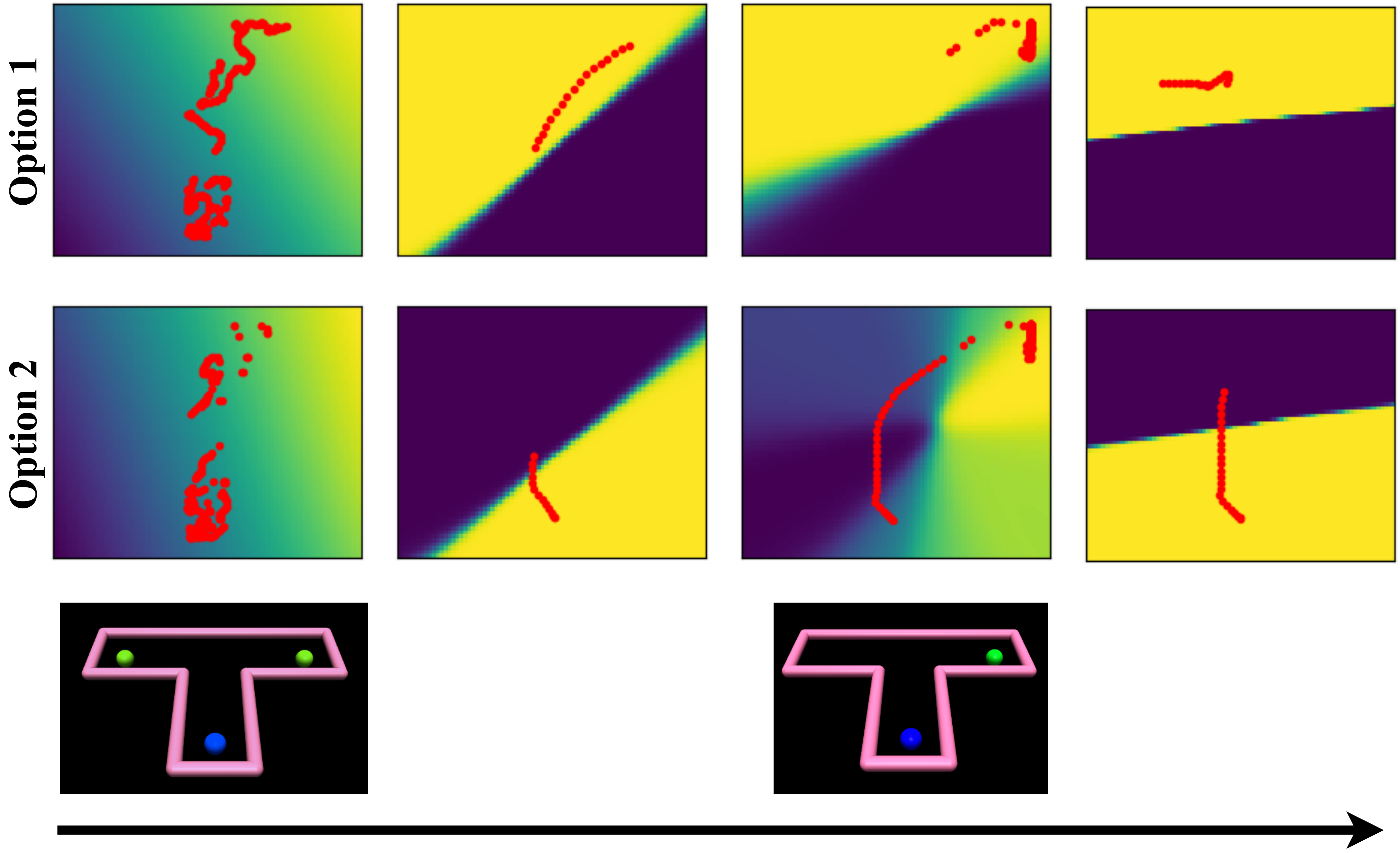






Learning Options with Interest Functions

Continuous Control



Theme I: Summary

- Introduced a generalization of initiation sets for options: ***interest functions***

- Proposed an approach to learn interest functions leading to options that are ***specialized*** to different regions of the space

- Experiments demonstrate the utility in continual learning tasks

- Options learned are ***interpretable***, ***reusable***, and ***meaningful***

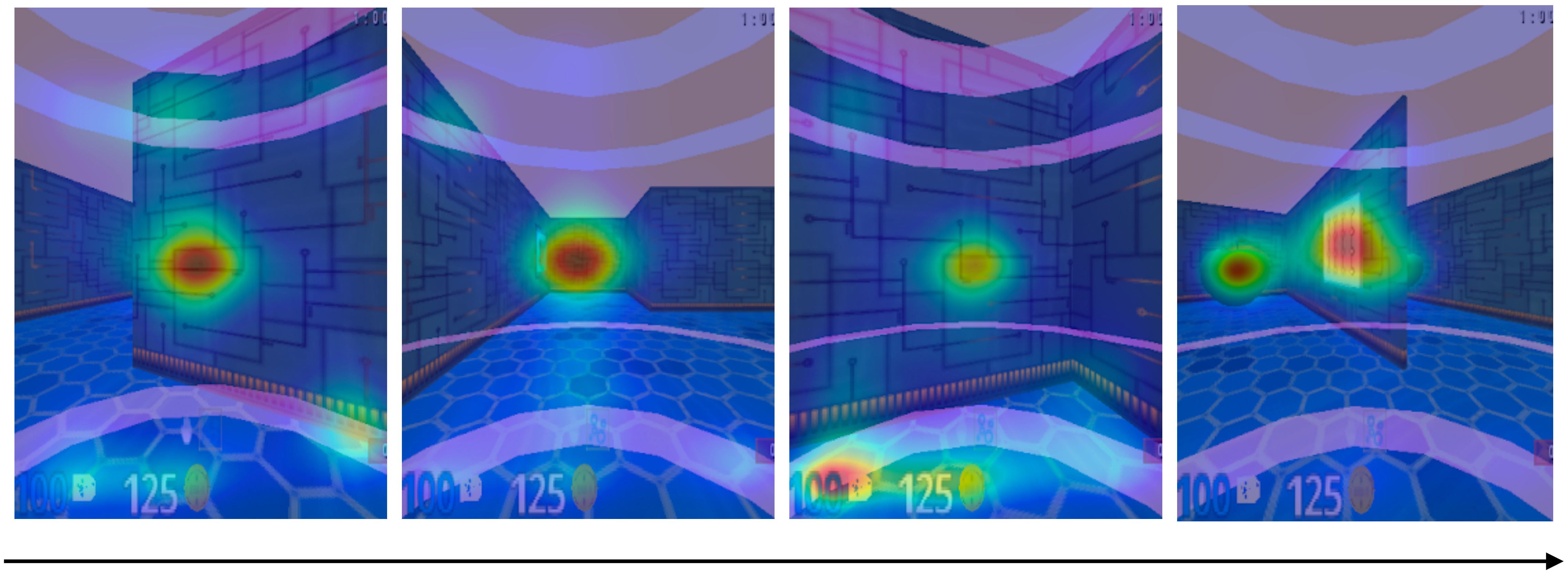
- Work is submitted and under review

Overview

- **Research goals**
- **Temporal abstraction**
- **Theme I: Learning options with interest functions**
- **Theme II: Learning temporal abstractions across action and perception**
- **Timeline**

Theme II: Learning temporal abstractions across action and perception

- Never-ending stream of rich sensorimotor data
- What we ***see*** forms an important source of information



Time

Attend Before you Act: Leveraging human visual attention for continual learning [The 2nd Lifelong Learning: A Reinforcement Learning Approach (LLARLA) Workshop, ICML 2018]

Theme II: Learning temporal abstractions across action and perception

- Never ending stream of sensorimotor data
- What we *see* forms an important source of information

Idea: ***Learn temporally extended perception + action***

- Embodied interaction allows the agent to understand objects and associated affordances
- Allow perceptual features to represent multiple time steps in synchrony with the agent's option

Theme II: Learning temporal abstractions across action and perception

Current Challenges:

- How can the agent automatically learn features which are meaningful pseudo rewards ?
- Where do task descriptions come from ?
- How can we achieve most generalized temporal abstractions without hand designing tasks and rewards associated with each task ?
- Evaluation in a lifelong learning task
  - Need of benchmarks ?

Overview

- **Research goals**
- **Temporal abstraction**
- **Theme I: Learning options with interest functions**
- **Theme II: Learning temporal abstractions across action and perception**
- **Timeline**

Timeline

- **Theme I: Learning options with interest functions**
  - Formulation
  - Derivation
  - Algorithm, design of experiments
  - Experiments
    - Tabular
    - Function approximation
  - What are the theoretical guarantees for this work, Can we do better ?
- **Theme II: Learning temporal abstractions across action and perception**
  - Attend Before you Act: Leveraging human visual attention for continual learning [Lifelong Learning: A Reinforcement Learning Approach Workshop, ICML 2018]
  - Formulation (in progress)
- **Misc:**
  - Environments for Lifelong Reinforcement Learning [Continual Learning Workshop, NeurIPS 2018]

Discussion

Extra Slides

Option-Critic

Formulation

All options are available in all states

The option value function is defined as

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

Option-Critic

Formulation

All options are available in all states

The option value function is defined as

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

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:

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

Option-Critic

Formulation

All options are available in all states

The option value function is defined as

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

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:

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

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

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

Learning Options with Interest Functions

Formulation

- 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.

- Starting with the option value function, we derive the policy gradient theorem for interest functions, intra-option policy and the termination function.

Learning Options with Interest Functions

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 $I_{\omega,z}(s)$

Learning Options with Interest Functions

Formulation

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:

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

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

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

Learning Options with Interest Functions

Formulation

- 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.

- Starting with the option value function, we derive the policy gradient theorem for interest functions, intra-option policy and the termination function.

Learning Options with Interest Functions

Main Result : Interest Function Gradient Updates

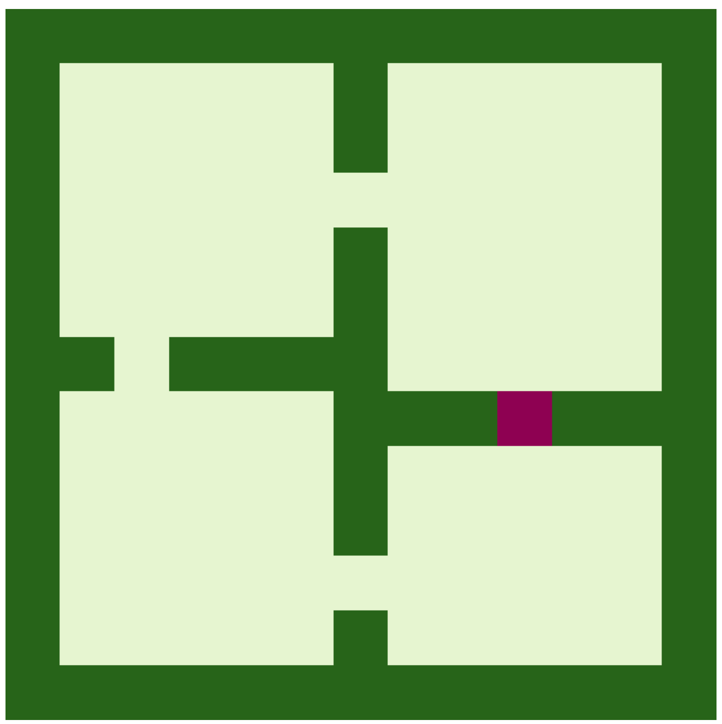
Given a set of Markov options with stochastic, differentiable interest functions, the gradient of the expected discounted return with respect to z at (s, ω) is:

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

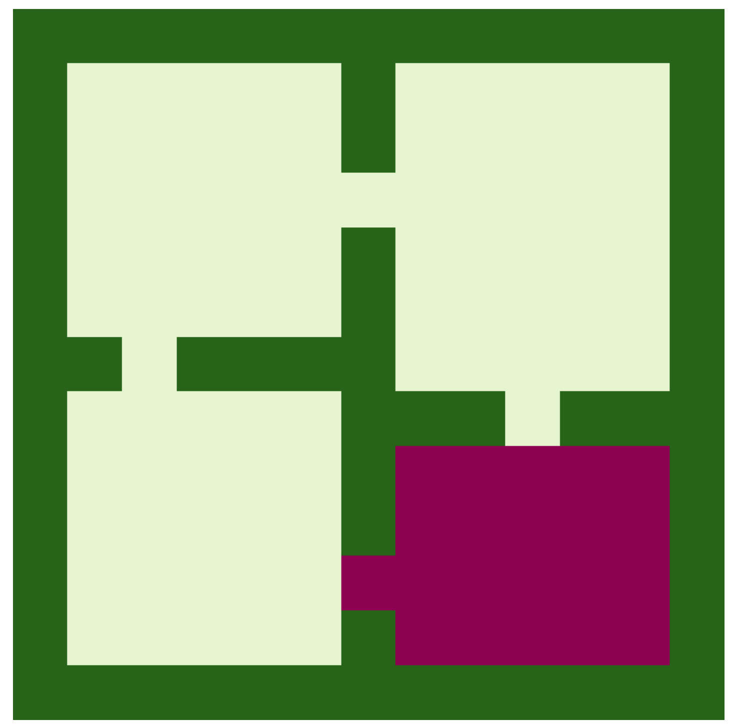
where $\hat{\mu}_\Omega(s', \omega' \,|\, s, \omega)$ is the discounted weighting of the state-option pairs along trajectories starting from (s, ω) sampled from the sampling distribution determined by $I_{\omega,z}(s)$

Learning Options with Interest Functions

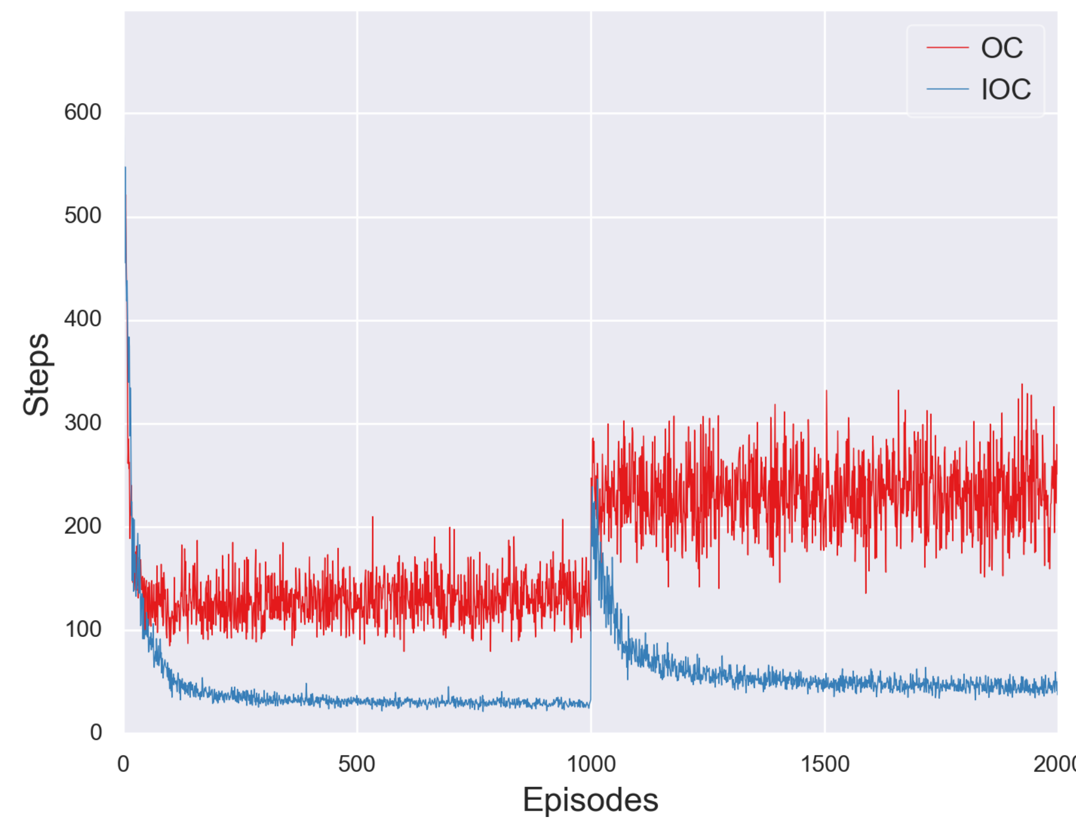
Four Rooms Domain



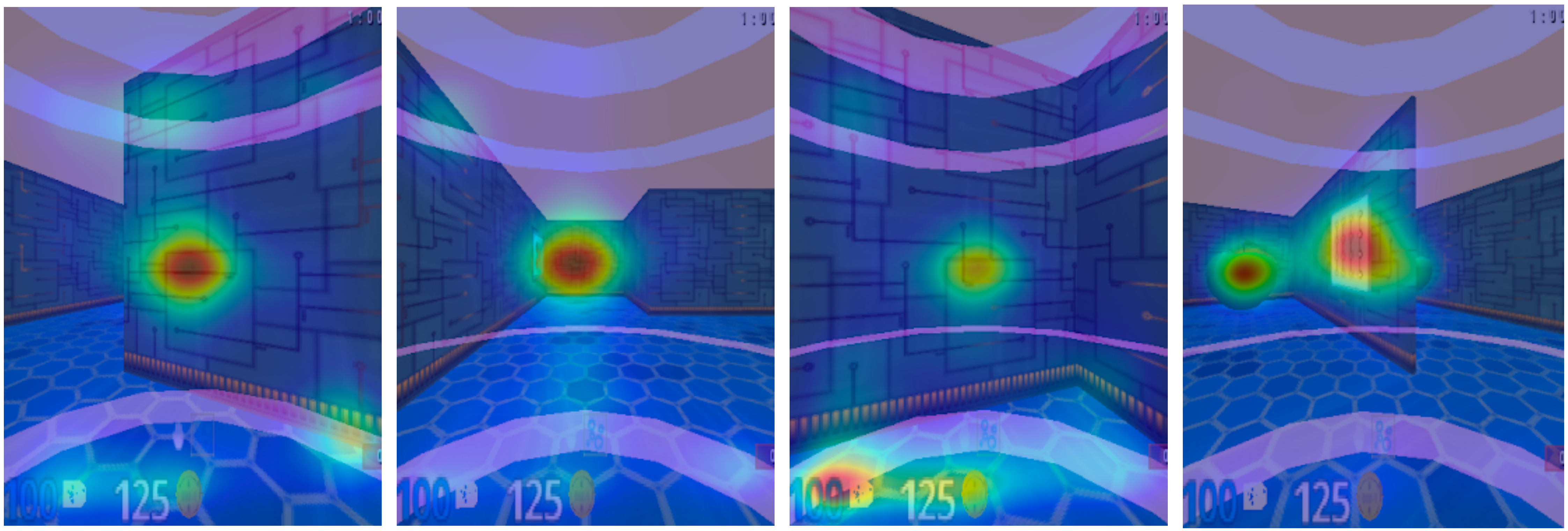
Initial Goal



Random Goal



Temporally extended perception + action



Related Work

- State abstractions
    -

- Related Work