Learning Options with Interest Functions

Khimya Khetarpal, Doina Precup
Mila - Reasoning and Learning Lab
McGill University, 3480 University St. Montreal, Quebec H3A 0E9
Email: khimya.khetarpal@mail.mcgill.ca, dprecup@cs.mcgill.ca

Introduction

Learning temporal abstractions which are partial solutions to a task and could be reused for solving other tasks is an ingredient that can help agents to plan and learn efficiently. In this work, we tackle this problem in the *options* framework (Sutton, Precup, and Singh 1999; Precup 2000). We aim to learn options which are specialized in different state space regions by proposing a notion of *interest functions*. We build on the option-critic framework (Bacon, Harb, and Precup 2017) to derive policy gradient theorems for interest functions leading to a new *interest-option-critic* architecture.

Preliminaries

A finite, discrete-time Markov Decision Processes (MDP) (Sutton and Barto 1998) is a tuple $\langle S, A, r, P, \gamma \rangle$, where S is the set of states, A is the set of actions, $r : S \times A \to \mathbb{R}$ is the reward function, P is the state-transition probability, and $\gamma \in [0, 1)$ is the discount factor. At each time step, the learning agent perceives a state $S_t \in S$, takes an action $A_t \in A$ drawn from a policy, $\pi : S \times A \to [0, 1]$, and with probability $P(S_{t+1}|S_t, A_t)$, enters into next state S_{t+1}, receiving a numerical reward R_{t+1} from the environment. The value function of policy π is defined as: $V_\pi(s) = E_\pi[\sum_{t=0}^{\infty} \gamma^t R_{t+1}|S_0 = s]$ and its action-value function as $Q_\pi(s, a) = E_\pi[\sum_{t=0}^{\infty} \gamma^t R_{t+1}|S_0 = s, A_0 = a]$.

A Markovian option (Sutton, Precup, and Singh 1999) $\omega \in \Omega$ is composed of an *intra-option policy* π_ω, a termination condition $\beta_\omega : S \to [0, 1]$, and an initiation set $I_\omega \subseteq S$. In the *call-and-return* option execution model; the agent chooses an option ω according to the policy over options π_Ω, follows the option policy π_ω, until option termination governed by β_ω, at which point this process is repeated. The option-value function is defined as:

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

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:

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

where U is the option-value function upon arrival in a state:

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

Learning Options with Interest Functions

Recent research has demonstrated that options can be learned automatically and end-to-end for a given task (Bacon, Harb, and Precup 2017; Bacon 2018). Unfortunately, this can result in degenerate solutions, with either one option being used for the entire task, or option duration collapsing to single time steps. This type of degenerate solution is potentially due to a simplifying assumption used in the option-critic (Bacon, Harb, and Precup 2017): that all options are available in all states. This assumption is not present in the original options paper, where an option is limited to act in a subset of states. However, sets are inconvenient for learning, as they do not lend themselves to gradient-based adjustments. In order to learn options that represent specialized and meaningful skills for lifelong learning, we revisit the idea of an initiation set, used in the options framework, but through a formulation that is more amenable to learning.

We introduce the notion of *interest functions* $I_\omega : S \to \mathbb{R}$. The idea is inspired by human visual attention: 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.

Note that we will interpret $I_\omega(s)$ as an indicator of the extent to which an option is applicable in a state. Initiation set can then be implemented through their characteristic function, which is a special type of interest function with binary output. However, in general it is more convenient to consider differentiable interest functions, $I_{\omega,z}$ parameterized by a parameter vector z, in order to be able to adjust them with gradients.

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

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

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:

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

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 can derive the interest function gradient, obtaining the following result:

Theorem 1. *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:*

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

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 $I_{\omega,z}$.

We can then derive the policy gradients for intra-option policies and termination functions which are assumed to be stochastic and differentiable in θ and ν respectively. The proofs are in the appendix[1]. This gives us the following two results in Theorem 2 and 3.

Theorem 2. *Given a set of Markov options with stochastic, differentiable intra-option policies $\pi_{\omega,\theta}$, the gradient of the expected discounted return with respect to θ and initial condition (s_0,ω_0) is:*

$$\sum_{s,\omega} \hat{\mu}_\Omega(s,\omega|s_0,\omega_0)\sum_a \frac{\partial \pi_{\omega,\theta}(a|s)}{\partial \theta} Q_U(s,\omega,a)$$

where $\hat{\mu}_\Omega(s,\omega|s_0,\omega_0)$ is the discounted weighting of the state-option pairs along trajectories starting from (s_0,ω_0) sampled from the new option sampling distribution determined by $I_{\omega,z}(s)$.

Theorem 3. *Given a set of Markov options with stochastic, differentiable termination functions $\beta_{\omega,\nu}$, the gradient of the expected discounted return with respect to ν and initial condition (s_0,ω_0) is:*

$$-\sum_{s',\omega} \hat{\mu}_\Omega(s',\omega|s_0,\omega_0)\sum_a \frac{\partial \beta_{\omega,\nu}(s')}{\partial \nu} A_\Omega(s',\omega)$$

where $\hat{\mu}_\Omega(s,\omega|s_0,\omega_0)$ is the discounted weighting of the state-option pairs along trajectories starting from (s_0,ω_0) sampled from the new option sampling distribution determined by $I_{\omega,z}(s)$.

Here $A_\Omega(s',\omega)$ is the advantage function over options. Note that these two results remain similar to the ones in (Bacon, Harb, and Precup 2017) with the key difference in the discounted weighting of state-option pairs now sampled from the new option sampling distribution determined by $I_{\omega,z}(s)$. This is natural as the introduction of interest-function should only impact the choice of options in each state.

An implementation of the interest-option-critic in the tabular setting using intra-option Q-learning is shown in Algorithm 1. The algorithm is also applicable to function approximation. Experiments are in progress. After empirical evidence in simulated environments, we aim to extend the work to the robotics domain to demonstrate its efficacy in a real world scenario.

Algorithm 1 Interest-Option-Critic with tabular intra-option Q-learning

$s \longleftarrow s_0$
Initialize policy over options π_Ω
Initialize $I_{\omega,z}$ parameterized by z such that all options are everywhere at the start
$\pi_{I_{\omega,z}}(\omega|s) = I_{\omega,z}(s)\pi_\Omega(\omega|s) \Big/ \sum_\omega I_{\omega,z}(s)\pi_\Omega(\omega|s)$
Choose ω according to $\pi_{I_{\omega,z}}$
repeat
  Choose a according to option-policy $\pi_{\omega,\theta}(a|s)$
  Take action a in s, observe s', r

  1. Options evaluation:
  $\delta \leftarrow r - Q_U(s,\omega,a)$
  if s' is non terminal **then**
    $\delta \leftarrow \delta + \gamma(1-\beta_{\omega,\nu}(s'))Q_\Omega(s',\omega) + \gamma\beta_{\omega,\nu}(s')\max_{\omega'} Q_\Omega(s',\omega')$
  end if
  $Q_U(s,\omega,a) \leftarrow Q_U(s,\omega,a) + \alpha\delta$

  2. Options improvement:
  $\theta \leftarrow \theta + \alpha_\theta \frac{\partial \log \pi_{\omega,\theta}(a|s)}{\partial \theta} Q_U(s,\omega,a)$
  $\nu \leftarrow \nu - \alpha_\nu \frac{\partial \beta_{\omega,\nu}(s')}{\partial \nu}(Q_\Omega(s',\omega) - V_\Omega(s'))$
  $z \leftarrow z + \alpha_z \beta_{\omega,\nu}(s')\frac{\partial \pi_{I_{\omega,z}}(\omega'|s')}{\partial z} Q_\Omega(s',\omega')$

  if $\beta_{\omega,\nu}$ terminates in s' **then**
    Choose new ω according to $\pi_{I_{\omega,z}}$
  end if
until s' is a terminal state

Interest functions enable end-to-end autonomous construction of options that are specialized in different regions. Emergence of such options would enable generalization over multiple tasks requiring similar options, and facilitate lifelong and hierarchical learning.

References

Bacon, P.-L.; Harb, J.; and Precup, D. 2017. The option-critic architecture. In *AAAI*, 1726–1734.

Bacon, P.-L. 2018. *Temporal Representation Learning*. Ph.D. Dissertation, McGill University, Montreal.

Precup, D. 2000. Temporal abstraction in reinforcement learning. *Ph. D. thesis, University of Massachusetts*.

Sutton, R. S., and Barto, A. G. 1998. *Introduction to Reinforcement Learning*. Cambridge, MA, USA: MIT Press, 1st edition.

Sutton, R. S.; Precup, D.; and Singh, S. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112.

[1] https://sites.google.com/view/learninterest