CS7011 : Topics in Reinforcement Learning
Summary of A Laplacian Framework for Option Discovery in Reinforcement Learning : Marlos C. Machado, Marc G. Bellemare and Michael Bowling

Sahana Ramnath : EE15B109

Abstract

'Laplacian Framework for Option Discovery in Reinforcement Learning' (Machado et al., 2017) proposes discovering options by using Proto Value Functions **PVFs** to learn a representation of the underlying MDP. The aim is to learn options which can be used for multiple tasks. This is made achievable by discovering options without using the environment's rewards. To learn options, they introduce *eigenpurposes*, which are intrinsic reward functions derived from the PVFs. They encourage the agent to traverse the state space in the direction of the principal components of the PVFs. Each *eigenpurpose* leads to a different option, the *eigenbehaviour*, which is the optimal policy for that reward function. These act at different time scales (and so help in exploration) and are defined over the whole state space. Experiments were conducted in tabular domains and in Atari games.

1. Proto Value Functions

Proto Value Functions PVFs are representations that capture large-scale geometries of an environment, such as bottlenecks and symmetries. They are obtained by diagonalizing a diffusion model constructed from the MDP's transition matrix. Different diffusion models can be used, such as the *combinatorial graph Laplacian* matrix $L = D - A$ or the *normalized graph Laplacian* matrix $L = D^{-1/2}(D-A)D^{-1/2}$, where A is the MDP graph's adjacency matrix and D is a diagonal matrix whose diagonal elements are the rowsums of A. (A can be generalized to a weight matrix W). PVFs are the eigenvectors corresponding to the K smallest eigenvalues of L; they are used as basis functions to define options.

2. Option Discovery using PVFs

An *eigenpurpose* defines an intrinsic reward function $r_i^e(s, s')$ of a PVF $e \in \Re^{|S|} : e^T(\phi(s') - \phi(s))$, where $\phi(x)$ is the feature representation of state x. Using PVFs to learn options can also be viewed as trying to reach the highest point of the graph of the *eigenpurpose* versus states(basically maximize the intrinsic rewards). Since the sign of an eigenvector is arbitrary, a PVF can also be viewed as a desire to reach the lowest point of the graph.

Now, a new MDP can be defined to learn options associated with the *eigenpurpose* : $M_i^e =< S, A \cup \{terminate\}, r_i^e, p, \gamma >$. The discount rate γ affects the timescale the option encodes. The state space and transition kernel remain unchanged from the original MDP. r_i^e is a dense reward function and helps avoid exploration challenges.Corresponding to this MDP, a new state-value and action-value function can be computed. The options obtained from this are available in every state where it's possible to achieve its purpose and terminate when it is acieved.

3. Empirical Evaluation of Eigenbehaviours

- They present specific purposes, not limited to finding bottlenecks.
- They improve exploration. This is shown using a new metric *diffusion time*. The first options added hurt exploration since they were at a high temporal scale and intermediate states were never explored, but as options were introduced at different timescales, exploration greatly improved compared to random walks.
- They accumulate rewards faster and speed up learning.

4. Approximate Option Discovery

With large state spaces, obtaining the adjacency matrix is infeasible. So, sample based methods are used. Each sample transition when encountered for the first time is added to a matrix T as $\phi(s') - \phi(s)$. Once enough transitions have been sampled, the SVD of T is calculated : $T = U \sum V^T$. The columns of V which are the right eigenvectors of T are used for the *eigenpurpose*. T is called the *incidence matrix*.

- Proved : $T^T T = 2L$, where $L = D - W$
- Function approximation can be used for continuous state spaces

5. Experiments

The model was tested on tabular and Atari domains. The options when visualized showed meaningful intentions such as picking up a key or opening a door in Montezuma's Revenge. One of the drawbacks observed was that this method predicts too many options for the task at hand.

References

Machado, Marlos C, Bellemare, Marc G, and Bowling, Michael. A laplacian framework for option discovery in reinforcement learning. *arXiv preprint arXiv:1703.00956*, 2017.