Laplacian Framework for Eigen-option Discovery

Manan Tomar : ED14B023

Abstract

This work (Machado et al., 2017) presents an option (Sutton et al., 1999) discovery method which uses Proto-value functions (PVFs) (Mahadevan & Maggioni, 2007) to learn task independent options that allow for efficient exploration. The proposed method is tested on simple grid worlds and on the Atari domain.

1. Approach

Proto-value functions are the eigenvectors produced by the eigendecomposition of the graph Laplacian L which is defined as $D - A$, where A is the adjacency matrix of the graph and D is a diagonal matrix with the elements being the row sums of A. (Mahadevan & Maggioni, 2007) showed that PVFs capture the inherent geometry of the environment and thus are a good representation of task agnostic value functions.

The authors define *eigenpurpose* as an intrinsic reward for learning the option policy for a given proto-value function e as

$$r^e(s, s^{'}) = e^T(\phi(s^{'}) - \phi(s)) \tag{1}$$

where ϕ is the state feature representation. Such a reward is dense, as it is available at all states and directed towards reaching the state where the PVF value is the biggest. Such a policy is termed as an *eigenbehavior* and the corresponding option as an *eigenoption*.

Learning eigenoptions requires an available state graph and is shown to enhance exploration in reward based settings and provide options that do not necessarily relate to bottleneck options and thus are more general.

2. Experiments

2.1. Grid Worlds

Learning with and without eigenoptions is shown for 3 grid world tasks, a simple 10 x 10 grid, I-maze and the 4-room domain. It is observed that using a sufficient amount of eigenoptions allow learning faster than when only learning using primitive actions. Note that the performance is hurt when the number of eigenoptions is too high or too low, which is regarded to either visiting specific parts (termination states of eigenoptions) of the state space very frequently or rarely respectively.

2.2. Atari

The authors consider learning eigenoptions for continuous state domains such as Atari by constructing a transition matrix (called as *Incidence Matrix*) using a random policy of primitive actions. The proto-value functions are then constructed by doing SVD decomposition over such a matrix and using the right vectors of the V matrix. To learn the options corresponding to the generated PVFs, the one step probability of the emulator is used.

3. Critique and Future Work

- Although the authors mention that eigenoptions are not exactly capturing the bottleneck idea, they do not clearly explain the intuition behind why eigenoptions are better just discovering bottleneck otpions. They also do not learn for a constantly changing goal (it is fixed at the top right corner for the 4-room domain).
- The authors also do not provide a scheme for constructing the state graph automatically, especially for cases where primitive actions are insufficient in exploring the whole state space in finite time.
- The options discovered in atari are termed useful in that some of them are able to produce good reward accumulating policies, however the authors do not explain how useful options need to be selected out of all possible eigenoptions generated.

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.

Mahadevan, Sridhar and Maggioni, Mauro. Proto-value functions: A laplacian framework for learning representation and control in markov decision processes. *Journal of Machine Learning Research*, 8(Oct):2169–2231, 2007.

Sutton, Richard S, Precup, Doina, and Singh, Satinder. Between mdps and semi-mdps: A framework for temporal abstraction

in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.