Theoretical Property of Topological Efficiency Measurements for Markov Decision Problems

Seung-Joon Yi
and Byoung-Tak Zhang
Department of Computer Engineering
Seoul National University
Seoul, Korea
Email: {sjlee,btzhang}@bi.snu.ac.kr

Abstract— **Many real world problems can be formalized as Markov decision problems (MDPs). When the model of MDP is known, it can be solved using dynamic programming algorithms such as value iteration and policy iteration. When the model is not known, reinforcement learning (RL) algorithms such as Q-learning algorithm can be used.**

Due to high computational complexity of MDP solving algorithms, A number of temporal abstraction approaches have been suggested so far to increase the efficiency of solving MDPs. As most of these approaches require a priori design of temporal abstraction structure, there have been attempts to automatically learn temporal abstractions recently.

However, although the structure of temporal abstraction can significantly affect the efficiency of solving the MDP, to our knowledge none of current temporal abstraction approaches explicitly give performance guarantees as they lack the measurement of efficiency.

To tackle this problem, an explicit measurement of efficiency is suggested based on the topology of state transition graph of MDP. We also give theoretical proof that the running times of value iteration and Q-learning algorithm are linearly bounded by this measurement under some simplifying assumptions.

I. Introduction

A number of problems can be modeled as Markov decision processes (MDPs). To solve MDPs, various algorithms have been suggested such as value iteration or policy iteration. For more general setting where there is no knowledge of the model of MDP, reinforcement learning (RL) approaches are being used. But as these algorithms typically assume a discrete state space and do not scale well with the size of MDP, their practical applications with real problems with large or continuous state space are still limited. A common solution is using a function approximator such as a neural network. However, it is also known that the number of parameters to be estimated grows exponentially with the size of any compact encoding of a state, which is called the curse of dimensionality [1].

Attempts to combat this curse of dimensionality lead to temporal abstraction where decisions are not required to perform every single action. This naturally leads to hierarchical control architectures and thus the associated learning algorithms are called hierarchical reinforcement learning (HRL) algorithms [1,2,3]. But major shortcoming of temporal abstraction approaches is that most approaches require a priori design of hierarchy. Furthermore, the learning efficiency of resulting MDP varies significantly with the design of hierarchy [4]. There have been approaches to automatically learn hierarchy [5,6,7], but to our knowledge, none of these approaches have a performance guarantee for solving resulting MDP.

To have a performance guarantee, we need a measurement of efficiency first. With the measurement, we can design the hierarchy so that the resulting MDP is efficient to solve, and further we can get the performance guarantee for the resulting MDP as well. But, to our knowledge, none of former temporal abstraction approaches have such an explicit measurement of efficiency. In this work, we introduce a topological measurement of efficiency and analyze the relationship between the performance of MDP solving algorithms and the suggest measurement of efficiency. Further, we show that the running time of two MDP solving algorithms are linearly bounded with the measurement under some simplifying assumptions.

II. Background

A. Markov decision problems

A Markov decision process M is a four-turple $< S, A, T, R >$ [8]. S is a finite set of states s, A is a finite set of actions a, R is the reward function that defines the immediate reward $r(s, a, s')$. T is the state transition function that describes the probability $p(s'|s, a)$ that the environment will move from state s to s' after action a is performed. And the policy is defined as a mapping $\pi: S \rightarrow A$. A Markov decision problem (MDP) is a Markov decision process together with a performance criterion. In this paper we consider the performance criterion of expected discounted cumulative cost defined as

$$V^{\pi}(s) = E[\sum_{k=0}^{\infty} \gamma^k r(s_k, \pi(s_k), s_k + 1)|\pi, s_0 = s] \quad (1)$$

where γ denotes the discount factor.

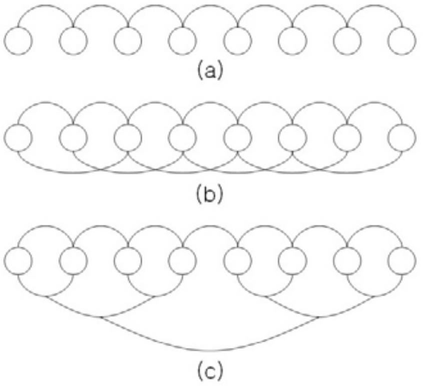

Fig. 1. Three different MDP structures

Solving an MDP is finding the optimal policy that maximizes the cumulative cost for every state. When the model of MDP is known, which consists of T and R , the MDP can be solved by dynamic programming based algorithms such as value iteration or policy iteration algorithm. Even when the model of the environment is not known, MDP can still be solved by reinforcement learning (RL) algorithms such as Q-learning.

B. *Temporal abstraction approaches*

A common way to solve large MDP efficiently is the temporal abstraction approach, which try to reduce the number of decision steps between two states using hierarchical decomposition of MDP[2] or multi-step actions (options) [3]. Figure 1 is an example of MDP with temporal abstraction. Figure 1 (a) shows the flat MDP where each state is only connected to its neighboring states. Figure 1 (b) shows the MDP augmented with options, and figure 1 (c) shows the MDP with a hierarchical structure [2].

C. *State trajectory graph*

MDP can be represented as a weighted directed graph where each node i corresponds to the state of the MDP s_i and each edge (i, j) corresponds to the possible state change $(s_i a_{ik}, s_j)$, and the weight of each edge corresponds to the transition probability between two states [7]. For the simple case when the state transition is deterministic, each edge directly corresponds to the state-action pair (s, a). .

III. Topological complexity measurements

A. *Relation between topology and efficiency*

It is known that temporal abstraction can generally accelerate the process of solving MDP [1,2]. This means that by temporal abstraction, we can effectively convert the original MDP into more efficient one. However, computational complexity of solving an MDP can vary much according to its topology [4].

Figure 2 shows the scaling behavior of time needed to solve three types of MDPs shown in figure 1 using value iteration algorithm. We can see that with options the running time is halved, but its scaling behavior is still linear with the number of states, which is the same as flat MDP. However, hierarchical MDP has wholly different logarithmic scaling behavior.

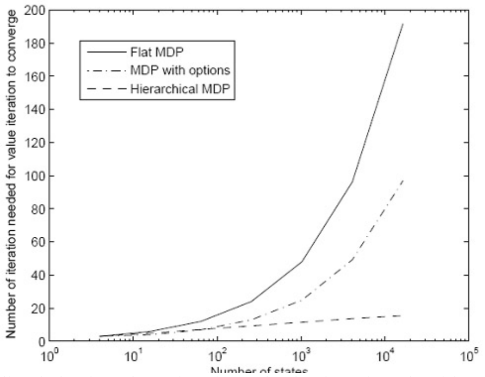

Fig. 2. Scaling behavior of running time of value iteration algorithm over the number of states for different types of MDPs

B. *Topological measurements of efficiency*

Each complex network presents specific topological features which characterize its connectivity and influence the dynamics and function of processes executed on the network. Network measurements that can express these topological features are used for the analysis, discrimination, and synthesis of complex networks. Depending on the network and analysis task one has in mind, a specific set of suitable measurements may be chosen. A complete survey on network measurements can be found in [11].

To find suitable measurements that express the efficiency of solving corresponding MDP, we empirically test a number of network measurements over different network models with varying parameters in our former work [12]. MDPs are generated according to 4 different network models, Regular lattice, Erdős-renyl, Watts-Strogatz[9] and Barabasi-Albert[10], with various sets of parameters. Generated MDPs are solved using value iteration and Q-learning algorithm and the network measurements of corresponding state trajectory graph, average distance, maximum distance, clustering coefficient, mean geodesic distance, subgraph centrality, fractal dimension, are evaluated. From the figure 3, we can see that the network measurement named mean geodesic distance is highly correlated with the running time of both MDP solving algorithms, so it can be used as a general efficiency measurement of MDP.

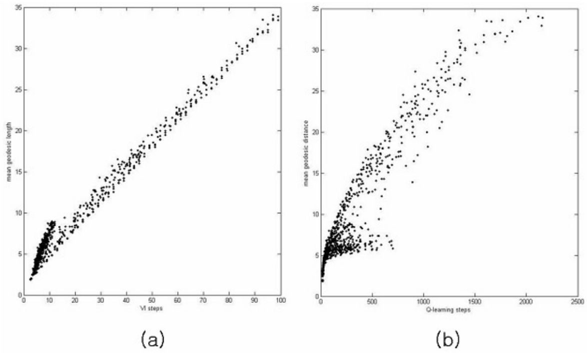

Fig. 3. Relationship between mean geodesic distance of state trajectory graph and time needed to solve corresponding MDP.
(a): Value iteration (b):Q-Learning

IV. Theoretical analysis

A. Deterministic single-reward MDP

First we will consider the simple single-reward case where positive reward is given at only one state. Among the single-reward case we will handle the deterministic MDP first. At each iteration of the value iteration, the value function V of each state *s* is updated using value functions of successor states $s', V(s')$. When the goal state which gives a positive reward is s_{goal} and all value functions are initialized to zero, $V(s)$ is updated after exactly $GD(s, s_{goal})$ iterations from initialization, where $GD(s_1, s_2)$ is the geodesic distance between two states s_1 and s_2. This value function is the final value function as the triangle inequality holds. So, for the deterministic single-reward MDP, value iteration has averaged run time of

$$\frac{1}{|s|^2}\sum_{s_1}\sum_{s_2} GD(s_1, s_2) = \frac{l}{|s|^2} \tag{2}$$

which is proportional to the mean geodesic distance l of the corresponding state trajectory graph.

For Q-learning case where update is local and action is selected off-policy[8], we need another simplifying assumption. We assume that by some exploration mechanism, all state-action pair is repeatedly selected every v_{int} iteration. This assumption somewhat artificial but not too restrictive as at the early stage of learning, explorative policy is preferred (High ϵ value for ϵ-greedy action selection rule and high temperature for Boltzmann action selection rule), and at the later stage of learning, state selection is being concentrated to 'important' region of the state space. By this assumption, Q-learning becomes a asynchronous version of value iteration and the expected running time of Q-learning is bounded by

$$\frac{v_{int}}{|s|^2}\sum_{s_1}\sum_{s_2} GD(s_1, s_2) = \frac{l\, v_{int}}{|s|^2} \tag{3}$$

B. Stochastic single-reward MDP

Note that if the every entry of the transition matrix $T(s, a, s')$ is nonzero, the resulting state trajectory graph will be a complete graph which renders the graph-based approach meaningless. So we make another assumption that the state trajectory graph is sparse, which means that each state, action pair (s, a) has a small number of resulting states, rs_{max} , and all the states are expected to occur with at most p_{sel} repeated selection of (s, a). This assumption is not too restrictive as in many types of RL problems, states are only connected to small number of its neighbor states and not connected to remote states. Due to the added randomness, the expected running time of Q-learning for the stochastic single-reward MDP is bounded by

$$\frac{v_{int}p_{sel}}{|s|^2}\sum_{s_1}\sum_{s_2} GD(s_1, s_2) = \frac{l\, p_{sel}v_{int}}{|s|^2} \tag{4}$$

C. Multi-reward MDP

Finally we consider the multi reward case that a number of states $s_{goal_1}, s_{goal_2} \cdots s_{goal_k}$ have positive reward values. Unfortunately, even for the simpler value iteration case, the value function of a state $V(s)$ is finalized only after $\max_i GD(s, s_{goal_i})$ iterations. With the same analysis as above, we can conclude that the running time of value iteration and Q-learning for general multiple-reward MDP is, under a number of simplifying assumptions, linearly bounded by maximum geodesic distance. However, as the network with small maximum geodesic distance tend to have small mean geodesic distance and vice versa, we can still expect high correlation between mean geodesic distance and the efficiency of solving MDP.

V. Conclusion and future work

In this work, we adopt a network measurement from complex network literature, the mean geodesic distance of state trajectory graph, as a topological complexity measurement of MDP. And we analyze the relation between the topological complexity measurement and the running time of two MDP solving algorithms, value iteration and Q-learning.

As a result, we show that for single reward MDP, the running time of two MDP solving algorithms are linearly bounded with mean geodesic distance of corresponding state trajectory graph. Although this result does not extend to the multi reward MDP case, we can still expect high correlation between the complexity measurement and the running time of MDP solving algorithms for the multiple reward case.

Future work include deriving more tighter bounds, deriving a tighter bound for multi-reward MDP considering the distribution of geodesic distances, and designing an efficient MDP structure using this analysis.

References

[1] Barto, A.G., Mahadevan, S. Recent advances in hierarchical reinforcement learning. Discrete Event Systems Journal, 13, 41-77, 2003.

[2] Dietterich, T. G. (1998). Hierarchical reinforcement learning with the MAXQ value function decomposition. In Proceedings of the 15th International Conference on Machine Learning ICML'98.

[3] Sutton, R.S., Precup, D., Singh, S.P. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artificial Intelligence, 112, 181-211, 1999.

[4] Rust, J., 1997. A Comparison of Policy Iteration Methods for Solving Continuous-State, Infinite-Horizon Markovian Decision Problems Using Random, Quasi-random, and Deterministic Dircretizations, Computational Economics 9704001, EconWPA.

[5] McGovern, A. and Barto, A.G., Automatic Discovery of Subgoals in Reinforcement Learning using Diverse Density, Proceedings of the Eighteenth International Conference on Machine Learning, p.361-368, June 28-July 01, 2001

[6] Digney, B., Learning Hierarchical Control Structure for Multiple Tasks and Changing Environments, Proceedings of the Fifth Conference on the Simulation of Adaptive Behavior: SAB 98, 1998

[7] Mannor, S., Menache, I., Hoze, A., & Klein, U. (2004) Dynamic abstraction in reinforcement learning via clustering. ICML, 21: 560--567. 13

[8] Sutton, R.S., Barto, A.G. Reinforcement learning: an introduction. MIT press, 1998.

[9] Watts, D.J., Strogatz, S.H. Collective dynamics of 'small-world' networks. Nature, 393, 404-407, 1998

[10] Barabasi, A.-L., Albert, R. Emergence of scaling in random networks. Science, 286, pp. 509-512,. 1999.

[11] L. da F. Costa a; F. A. Rodrigues a; G. Travieso a; P. R. Villas Boas a. Characterization of complex networks: A survey of measurements, Advances in Physics, Volume 56, Issue 1 January 2007 , pages 167 - 242

[12] Yi, S.J. and Zhang, B.T., Using topological properties of complex networks for analysis of the efficiency of MDP-based learning, in proceedings of the 33rd Korean information science society spring conference, 2006.