Randomized Coordination Search for Scalable Multiagent Planning

(Extended Abstract)

N. Kemal Ure [*]
Laboratory of Information and Decision Systems
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA, USA
ure@mit.edu

Jonathan P. How [†]
Laboratory of Information and Decision Systems
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA, USA
jhow@mit.edu

John Vian [‡]
Boeing Research & Technology
Seattle, WA
john.vian@boeing.com

ABSTRACT

Multiagent Markov Decision Processes (MMDPs) are difficult problems to solve due to the exponential increase in the size of the planning space in the number of agents. One of the most successful approaches for solving MMDPs utilizes coordination graphs (CGs), which encode the decouplings between the agents to reduce the dimension of the value function, which in turn reduces the computational complexity. However, it is typically assumed that the structure of the CG is available a priori, which is a limiting assumption for many practical scenarios. This work presents a randomized planning scheme based on the Bayesian optimization algorithm to probabilistically search over the space of CGs to discover CG structures that yield high return policies. The results demonstrate that the proposed method is superior in terms of convergence speed and accumulated reward.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed AI

Keywords

Multi Agent Systems, Planning Under Uncertainty

1. INTRODUCTION

Many robotic missions involve teams of mobile agents operating in an uncertain environment with the objective of achieving a common goal or maximizing a joint reward. Many of these missions can be formulated as stochastic sequential decision making problems and written as Markov Decision Processes (MDPs). However, the algorithms that solve MDPs exactly scale poorly with the number of agents. The alternative of using computationally feasible suboptimal algorithms usually requires extensive domain knowledge and manual parameter tuning to obtain a good approximation to the original problem. The contribution of this work is that it presents a randomized search algorithm for discovering coordination graphs that obtain a good trade-off between computational efficiency and the quality of the resulting policy.

Dynamic Programming (DP) [1] and its variants are often used to solve MDPs. Although these methods are guaranteed to converge to the optimal solution, the computational complexity increases exponentially in the number of agents, which renders these exact approaches infeasible for multiagent domains. Approximate Dynamic Programming (ADP) [2] methods partially address the issue of scalability by approximating the value/policy function by a finite set of basis functions, so that the approximate problem has fewer unknown parameters compared to the original problem. However ADP methods typically require the designer to hand-code the basis of the approximation, which usually relies on domain expertise. Recently, significant effort has been applied into automating the basis function selection process based on the observed/simulated data [3]. Overall, these approximate techniques were shown to improve the scalability and relaxed the constraints on specifying a fixed set of basis functions a priori. However, in practice these methods do not scale well to multiagent missions because the basis function automation slows down significantly in large-scale planning spaces.

A highly scalable multiagent planning algorithm with factored MDPs was proposed by Guestrin and Parr [4] in which agents solve their individual MDPs and the joint return optimization is achieved using a coordination graph (CG). The main drawback of the approach in [4] is the assumption that the structure of the the coordination graph is known to the designer. Kok and Valassis [5] proposed an algorithm for learning the structure of the coordination graphs greedily based on statistical tests, however no theoretical analysis was done and the algorithm was shown to be effective for only a small number of moderately sized problems.

Although this work focuses on centralized planning, the problem of discovering coordination structures is also a relevant topic in decentralized multiagent planning for partially observable domains[6]. Most of the existing work in this field performs incremental search, which can fail to converge.

Fig. 1 provides an outline of the Randomized Coordination Discovery (RCD) algorithm developed in this paper, which builds on the CG framework [4, 5]. In a nutshell, the algorithm performs a randomized search over the space of coordination graphs in order to discover coordination structures that yield high return policies.

[*]PhD Candidate at Department of Aeronautics And Astronautics

[†]Richard C. Maclaurin Professor of Aeronautics and Astronautics

[‡]Technical Fellow

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May 4–8, 2015, Istanbul, Turkey.*

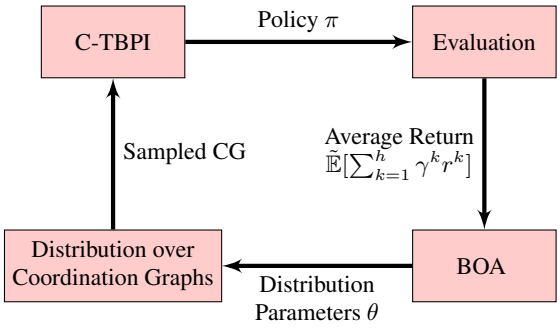

Figure 1: Representation of Randomized Coordination Discovery (RCD) algorithm.

The standard decomposition-based MMDPs solvers assume a fixed coordination graph structure a priori. This structure is usually obtained from domain knowledge or parameter tuning. The planning algorithm uses this coordination graph to compute a multiagent policy. In contrast, RCD does not assume a fixed coordination graph structure, and keeps generating new coordination graphs based on the planning performance. RCD (Fig. 1) defines a parametric probability distribution over the space of Coordination Graphs. At the start of each iteration, a number of CGs are sampled from this distribution. These sampled CGs are used by the planning algorithm [1] to compute a policy corresponding to each CG. These policies are evaluated by computing the corresponding discounted returns, and the return-CG pairs are used in the Bayesian Optimization Algorithm (BOA)[8] algorithm to update the sampling distribution by using the return-CG pairs. The updated distribution puts more probability weight on the CGs with higher returns. As a result, CGs with higher returns are sampled with higher probability in the next round. The main contribution of this work is the randomized coordination graph search architecture displayed in the Fig. 1 and the use of BOA to efficiently search over the space of coordination graphs. By automating the CG search with RCD, we obtain good suboptimal solutions to large-scale multiagent planning problems without the need for domain expert inputs. This is an improvement over the majority of the previously mentioned works, which assumed a fixed structure of Coordination Graphs. It is also shown in simulations that the developed algorithm is superior to existing adaptive approximation techniques in terms of convergence speed and solution quality. For more information on the development of the algorithm and theoretical analysis, the reader is referred to [7].

2. RESULTS AND CONCLUSIONS

This mission involves a group of unmanned aerial vehicles managing a forest fire [7]. The stochastic fire spread dynamics are affected by the wind direction, fuel left in the location and the vegetation. A total of 16 UAVs are present in the mission. Actions are traveling within a fixed distance in the forest or dropping water to kill the fire at the current location. The team receives a negative reward every time the fire spreads into a new location. The size of the planning space for this problem is approximately 10^{42} state-action pairs. The following approaches were compared with RCD. iFDD+ [3], which applies linear function approximation directly to the value function. The method starts with a fixed number of binary basis functions and grows the representation by adding new basis functions formed by taking conjunctions of the basis functions from the initial set. In order to emphasize the value of automating the coordination graph search, an approach that involves a fixed CG is also included in the results. We use intuition/domain knowledge to fix a CG beforehand, and use the same structure at every planning step. We also attempted to implement the Sparse Greedy Discovery approach [5], however the algorithm failed to converge for this domain. The algorithms were evaluated 30 times, and the average cumulative cost (negative reward) were computed per iteration. Each algorithm was allocated the same CPU time per iteration to ensure a fair comparison. The results are displayed in Fig. 2. iFDD+ resulted in a poorly performing policy. Although the fixed CG approach yielded a passable performance, on average RCD outperformed the competing approaches, in terms of both the convergence rate and the average cost.

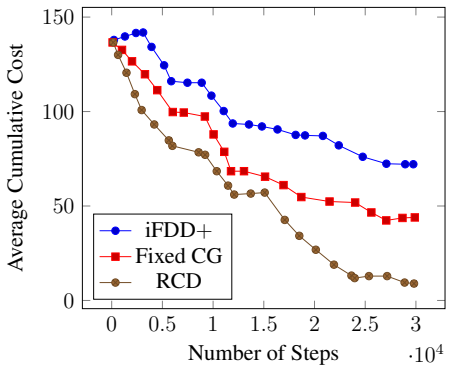

Figure 2: Performance comparison of three algorithms on firefighting domain with 40×40 grid and 16 agents.

This work developed a novel algorithmic framework to discover coordination structures in Multiagent Markov Decision Processes. Simulation results showed that the algorithm yields policies with significantly higher returns than the compared approaches. The future work consists of analysis of the performance guarantees.

[1] C-TBPI stands for Coordinated Trajectory Based Value Iteration algorithm, which is a variant of real time dynamic programming that works with coordination graphs[7].

REFERENCES

[1] R. Bellman, "Dynamic programming and stochastic control processes," *Information and Control*, vol. 1, no. 3, pp. 228–239, 1958.

[2] L. Busoniu, R. Babuska, B. D. Schutter, and D. Ernst, *Reinforcement Learning and Dynamic Programming Using Function Approximators*. CRC Press, 2010.

[3] A. Geramifard, T. J. Walsh, N. Roy, and J. How, "Batch iFDD: A Scalable Matching Pursuit Algorithm for Solving MDPs," in *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, (Bellevue, Washington, USA), AUAI Press, 2013.

[4] C. Guestrin, M. Lagoudakis, and R. Parr, "Coordinated reinforcement learning," in *ICML*, 2002.

[5] J. R. Kok and N. Vlassis, "Collaborative multiagent reinforcement learning by payoff propagation," *The Journal of Machine Learning Research*, vol. 7, pp. 1789–1828, 2006.

[6] F. A. Oliehoek, M. T. Spaan, S. Whiteson, and N. Vlassis, "Exploiting locality of interaction in factored dec-pomdps," in *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, pp. 517–524, International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[7] N. K. Ure, *Multiagent Planning and Learning Using Random Decompositions and Adaptive Representations*. PhD thesis, Massachusetts Institute of Technology, Department of Aeronautics and Astronautics, Cambridge, MA, February 2015.

[8] M. Pelikan, "Bayesian optimization algorithm," in *Hierarchical Bayesian Optimization Algorithm*, pp. 31–48, Springer, 2005.