Coordinated Learning in Infinite Markov Games

Francisco S. Melo
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA
fmelo@cs.cmu.edu

Extended Abstract

Recent years have witnessed increasing interest in extending reinforcement learning (RL) to multiagent problems. Markov games have thoroughly been used to model multiagent RL problems [4] and several researchers have applied single-agent RL methods (with adequate adaptations) to this multiagent framework.

However, most multiagent RL research focuses on Markov games with finite state-spaces and only a few works on multiagent learning address problem with infinite state-spaces. Some examples include the work by Bowling and Veloso [1], Guestrin et al. [2] and Kok et al. [3]. Singh et al. [7] also refer the interest of applying gradient ascent techniques to games with infinite state-spaces.

In this work, we focus on infinite fully-cooperative Markov games, also known as *team Markov games* or *multiagent MDPs* (MMDPs). In MMDPs, all agents must commit upon a common joint behavior. We assume that no explicit communication takes place, *i.e.*, consensus emerges from the mutual interaction among the different agents and with the environment. We thus feature cooperation as coordination: the multiple decision-makers must *coordinate* their individual decisions to yield an optimal joint behavior.

In our approach we consider two distinct "subproblems" that must be solved simultaneously:

Learning the game where each agent must learn a compact representation of the game from which an optimal policy can be determined (namely the optimal Q-function for the MMDP);

Learning to coordinate where, in the presence of multiple optimal policies, all decision-makers agree upon one such policy *without any explicit communication*;

To address the first of the two problems above, we propose the use of Q-learning with soft-state aggregation (Q-SSA). The combination of Q-learning and soft-state aggregation has been widely studied in the RL literature [6]. One appealing property of Q-SSA for our purposes is that function approximation can be used to compute in an off-policy manner the optimal Q-function, Q^*, associated with the MMDP. Our results heavily rely on this fundamental property that Q-SSA converges w.p.1 to an approximation of Q^* when some fixed learning policy is used.

In addressing the problem of coordination, we propose the use of *approximate biased adaptive play* (ABAP). This method was introduced and shown to converge in [5], when Q^* is known. The ABAP method differs from other coordination methods in several aspects. First of all, ABAP assumes that no communication takes place. Furthermore, no assumption is made regarding previous knowledge on the coordination algorithm of the other decision-makers. In particular, we do not assume that all decision-makers follow the same decision-making or coordination algorithm. This is an important advantage of ABAP: in the presence of a heterogeneous group of decision-makers, ABAP is still able coordinate to the best decision-rule possible if, for some reason, the other decision-makers act sub-optimally.

To combine ABAP with Q-SSA, two main difficulties must be addressed. One first difficulty lies on the fact that Q-SSA is implemented with a fixed learning policy. However, it is expected that, as the agents in the game learn how to coordinate, their policy will *slowly change*. This means that the policy learned by the agents running ABAP cannot be the one used to generate the sample transitions for the learning algorithm. To tackle this difficulty, we consider an auxiliary process from which the sample transitions are generated.

The second difficulty lies on the fact that, at each time instant t, the agents do not know the optimal function Q^* but, instead, have access only to an estimate, $\hat{Q}_t$, that they must use to choose their actions and coordinate. The intuitive idea to overcome this difficulty is to use the estimate $\hat{Q}_t$ to build at each state a *virtual matrix game* that the agents can use to learn how to coordinate in that state. This virtual game will include not only the optimal actions according to the current estimate $\hat{Q}_t$, but also actions that "appear" to be close to optimal. Then, as $t \to \infty$, sub-optimal actions are gradually removed.

We dub the algorithm obtained by combining ABAP with Q-SSA as *coordinated approximate Q-learning* (CAQL). Its basic procedure is as follows. Given an MMDP and a fixed learning policy, we apply standard Q-SSA to build successive estimates $\hat{Q}_t$ of Q^*. At each state visited during learning, all agents engage in the matrix virtual game defined from $\hat{Q}_t$. The sole purpose of the agents in this game is to coordinate in an optimal policy, disregarding any knowledge otherwise on the underlying learning process. As such, the actions determined by the learning policy are irrelevant in the coordination process and vice-versa.

In the setting considered herein, we focus on games in which all agents have full access to the state of the game and to the actions played by the other agents. However, there are many important problems in which one or both such assumptions do not hold. There is a great body of work on problems with partial observability, and many models of different complexity have been proposed to address such problems (*e.g.*, Dec-MDPs, Dec-POMDPs, I-POMDPs). An important and interesting topic for future research is the extension of some of the ideas in this paper to problems with partial observability.

References

[1] Michael Bowling and Manuela Veloso. Scalable learning in stochastic games. In *Proceedings of the AAAI Workshop on Game Theoretic and Decision Theoretic Agents (GTDT'02)*, pages 11–18. The AAAI Press, 2000. Published as AAAI Technical Report WS-02-06.

[2] Carlos Guestrin, Michail G. Lagoudakis, and Ronald Parr. Coordinated reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning (ICML'02)*, pages 227–234, 2002.

[3] Jelle R. Kok, Matthijs T. J. Spaan, and Nikos Vlassis. An approach to noncommunicative multiagent coordination in continuous domains. In Marco Wiering, editor, *Benelearn 2002: Proceedings of the Twelfth Belgian-Dutch Conference on Machine Learning*, pages 46–52, Utrecht, The Netherlands, 2002.

[4] Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In Ramon López de Mántaras and David Poole, editors, *Proceedings of the 11th International Conference on Machine Learning (ICML'94)*, pages 157–163, San Francisco, CA, 1994. Morgan Kaufmann Publishers.

[5] F. Melo and M.I. Ribeiro. Emerging coordination in infinite team Markov games. In *Proc. 7th Int. Conf. Autonomous Agents and Multiagent Systems*, pages 355–362, 2008.

[6] Satinder P. Singh, Tommi Jaakkola, and Michael I. Jordan. Reinforcement learning with soft state aggregation. In *Advances in Neural Information Processing Systems*, volume 7, pages 361–368, 1994.

[7] Satinder P. Singh, Michael Kearns, and Yishay Mansour. Nash convergence of gradient dynamics in general-sum games. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI'00)*, pages 541–548, 2000.