On the Performance of On-line Concurrent Reinforcement Learners

Bikramjit Banerjee
Department of Electrical Engineering & Computer Science
Tulane University
New Orleans, LA 70118. USA
banerjee@eecs.tulane.edu

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent Systems; I.2.6 [**Artificial Intelligence**]: LearningConcept Learning

General Terms

Algorithms, Theory, Performance

Keywords

Multiagent Learning, Game Theory

Multiagent Learning (MAL) is significantly complicated relative to Machine Learning (ML) by the fact that multiple learners render each other's environments non-stationary. While ML focuses on learning a fixed target function, MAL deals with learning a "moving target function". In contrast to classical Reinforcement Learning, MAL deals with an extra level of uncertainty in the form of the behaviors of the other learners in the domain. Existing learning methods provide guarantees about the performance of the learners only in the limit since a learner approaches its desired behavior asymptotically. There is little insight into how well or how poorly an on-line learner can perform *while* it is learning. This is the core problem studied in this thesis, resulting in the following contributions.

First, it sets up a novel mix of goals for a new MAL algorithm that will achieve some basic learning objectives *without knowing the type of the other agents*, such as (1) learn the best response behavior when other agents in the domain are (eventually) stationary and (2) jointly converge on a mutual equilibrium behavior in case the other agents are using the same learning algorithm, and will also ensure that in case the other agents are neither of the above types, it will achieve some minimum average payoff that is "good" in some sense. We use the concept of "no external regret" from Game and Decision Theory to define "good" payoffs. This thesis explores the difficulty of achieving these goals simultaneously and with the help of the insights gained, introduces ReDVaLeR [1], a novel MAL algorithm and its variant, $RV_{\sigma(t)}$, that achieves all of these goals.

Second, the thesis explores the cost of learning when the opponents are also adaptive. This is the main challenge of MAL but has not been addressed explicitly until now. We focus on arbitrary opponents that use bounded memories to compute their strategies, a.k.a 'recent history adversaries" (RHA). An example of such strategies is the well-known Tit-for-tat strategy in the Prisoner's Dilemma game. The thesis contributes a uniform modeling framework for such agents and introduces "μ-PSAIM" [2], an efficient learning technique against RHA, that is shown to achieve non-stationary best response payoffs against RHA in polynomial time. This effectively ensures that the payoff (short term) cannot be poor for long.

Third, the thesis extends [3] the class of no external regret algorithms to yield a class of algorithms that are shown to achieve (1) close to best response payoffs against (eventually) stationary opponents, (2) close to the best possible asymptotic payoffs against converging opponents, and (3) close to at least the minimax payoffs against any other opponents, in polynomial time with high likelihood. This new approach assumes only that a learner can observe its payoffs and does not need to observe the opponents' payoffs or actions. It also produces polynomial bounds that are significantly improved over previous work. It may be possible to further characterize the class of "other" opponents to include stronger payoff guarantees against such opponents than the minimax payoff. This approach makes it possible to build practical algorithms with reasonable payoff guarantees, for MAL applications in uncertain environments with limitations on observability.

Lastly, the thesis validates all its novel techniques and algorithms empirically comparing them with existing techniques. A network scheduling testbed is developed to study the relative efficacy of these techniques in a multi-user scenario where all users are allowed to employ intelligent strategic reasoning and learning to maximize their individual utilities. The thesis demonstrates that exploratory learning does not necessarily associate high cost with exploration and that the payoffs of reinforcement learners in concurrent learning scenarios can be "good enough" even if the learners do not incorporate any knowledge about the learning strategies of the other agents.

AAMAS'05, July 25-29, 2005, Utrecht, Netherlands.

1. REFERENCES

[1] B. Banerjee and J. Peng. Performance bounded reinforcement learning in strategic interactions. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI-04)*, pages 2–7, San Jose, CA, 2004. AAAI Press.

[2] B. Banerjee and J. Peng. Efficient learning of multi-step best response. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, (to appear) 2005.

[3] B. Banerjee and J. Peng. Efficient No-regret Multiagent Learning. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05)*, (to appear) 2005.