Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008)

Perpetual Learning for Non-Cooperative Multiple Agents

Luke Dickens
Email: luke.dickens@imperial.ac.uk , Tel.: 020 7594 8351
Address: Imperial College London,
South Kensington Campus,
SW7 2AZ, UK.

Abstract

This paper examines, by argument, the dynamics of sequences of behavioural choices made, when non-cooperative restricted-memory agents learn in partially observable *stochastic games*. These sequences of combined agent strategies (*joint-policies*) can be thought of as a walk through the space of all possible joint-policies. We argue that this walk, while containing random elements, is also driven by each agent's drive to improve their current situation at each point, and posit a *learning pressure* field across policy space to represent this drive. Different learning choices may skew this *learning pressure*, and affect the simultaneous joint learning of multiple agents.

Motivation

Multi-Agent Stochastic Processes are becoming increasingly popular as a modelling paradigm. Game theoretic approaches commonly rely on the participating agents having full access to the process dynamics in advance, and then solve to find the best solution analytically, but with large problems this approach is impractical. For single agent stochastic processes, on-line learning is often proposed as an alternative, where control strategies are often restricted by internal memory or computational constraints. If we apply this approach to multiple agents, we lose the traditional game theoretic axiom of intelligent, informed agents; instead agents independently explore choices, exploiting successful strategies when they stumble across them. If these are ongoing games or endlessly repeated tasks, individual agents may continue to benefit by regularly updating their policy indefinitely. This leads to a continually changing game dynamic, which demands regular adaption by all other agents. We call such a situation *perpetual learning*, and it requires strict guarantees on policy evaluation for agents to perform well, both in terms of accuracy and timeliness.

For an example where perpetual learning may be appropriate consider an electronic trading environment, staffed by autonomous trading agents. The agents have a finite memory and trade goods with one another to achieve targets (for buying and/or selling), preferring to do so at the best rates. An individual agent may adopt a trading strategy, which is successful at turning a profit over the short term. However, this will cause other agents to be less successful, and, if they have appropriate learning techniques, they will adapt quickly to improve matters for themselves. Agents that anticipate changes, or adapt very quickly, can outperform others. Mistakes may be costly though, and predictable behaviour could leave agents open to exploitation.

Poster Description

Our work examines the properties of finite state dependent *stochastic games*[1], under independent perpetual learning of memory-restricted agents. In previous work, we proposed a family of frameworks for modelling finite analytic stochastic processes (FASPs), which can model multiple agents with independent reward functions, and showed how these related to a *state encapsulated* formulation of the POMDP (Dickens, Broda, & Russo 2008). In this abstract we discuss how each agent's control strategy can be composed with the others to give an expectation of average-reward per time-step for each agent, and this equates to a set of *preference relations* over the set of combined control strategies (joint-policy space). Using these preference relations, we posit the existence of a *learning pressure field*, which, we argue, predicts the combined drive to improve experienced by on-line learning agents satisfying certain constraints. This, in turn, provides certain guarantees for such learning agents, and should inform the development of algorithms appropriate for on-line learning in non-cooperative multi-agent stochastic domains.

Our argument is structured as follows:

1. The FASP lends itself to natural representations of stochastic games supporting perpetual learning. These multi-agent FASPs can be rewritten as a family of Markov-chains (MCs), each chain indexed by a combined control strategy (or *joint policy*).
2. Given certain assumptions (e.g. ergodicity), these MCs can be used to predict the probability distribution over states at some distant future time, and we use this distribution to predict each agent's average reward per time-step, giving us a set of real valued evaluation measures for that particular chain.

[1] Non-cooperative multi-agent stochastic decision processes (Bowling & Veloso 2004)

3. Combining results from steps 1 and 2, we can define a preference relation over joint policy space for each agent.
4. Learning methods which approximate gradients on this preference relation, allow agents to *climb* to local policies with better values, by making small policy changes.
5. Each agent's direction of improvement is a vector within its policy space. These improvement vectors are orthogonal, and can be combined to give a general vector of perceived improvement, at every point in joint-policy space. This creates a *learning pressure field*, an expectation of the local direction of change to joint-policies, induced by simultaneous multi-agent learning.
6. Knowledge of the existence and implications of a learning pressure field will inform the design of learning algorithms for use in simultaneous multi-agent on-line learning, and with work might be used to define the meta-level stochastic processes of perpetual learning.

In slightly more detail: any (single-agent) POMDP can, for any fixed *policy*, be remodelled as a Markov-chain (MC) (Baxter, Bartlett, & Weaver 2001). If we parametrise our set of available policies, we can define a parametrised MC which represents the entire family of MCs in the process. In (Dickens, Broda, & Russo 2008), we defined a family of modelling frameworks, (FASP), which allow us to model finite-state multi-agent stochastic processes (*stochastic games*). For each framework, we showed how any multi-agent FASP model can be transformed into the simple FASP, which has the same state-to-state transition probability representation as our *state-encapsulated* form of the POMDP, and hence the translation to fixed-policy MCs can be applied to these FASPs. Multi-Agent FASPs are different from the POMDP, in that each policy is made up of agent specific parts, and there are multiple measure signals. These differences allow us to model various multi-agent scenarios; here we concentrate on the general-sum case.

The use of the term general-sum is not accidental: our problem domain is similar to the game theoretic extensive-form game, defined in (Myerson 1997). However, our stochastic games differ in three crucial ways: agents are restricted by how much historical knowledge they can retain; rewards are evaluated at each state, rather than once at the end of the game; and agents are not assumed to have full knowledge of the game's properties in advance of playing. These differences mean we can model *infinite-horizon* (ongoing) games, suited to on-line learning techniques such as gradient ascent (Baxter, Bartlett, & Weaver 2001).

Since these FASPs with fixed policies can be rewritten as Markov-chains, and assuming that the system is ergodic for all policies, we can exploit the Perron-Frobenius theorem for non-negative primitive stochastic matrices (Horn & Johnson 1986), providing guarantees on the probability distribution over states after a large number of time-steps. This also provides guarantees on the average reward per time-step for each agent, again after a sufficient time, so given a fixed joint policy, we can prove that there is some well defined average reward for each agent.

We assume that each agent specific policy is parametrised independently, and that the space of all associated policy values equates to that agent's policy space. Further, combining all the agent specific parameters into a vector specifies a point in joint-policy space. This allows us to define a reward function for each agent over the entire joint-policy space, and it is the gradient of this function that the associated agent seeks to approximate with gradient ascent learning techniques, such as those in (Baxter, Bartlett, & Weaver 2001; Buffet & Aberdeen 2006). However, each agent only has control over its own policy, so can only climb this gradient by varying its own policy parameters. All agents will learn along vectors orthogonal to all others, and hence the gradients for each reward function are projected onto these sub-spaces of *free control*, each sub-space orthogonal to all others, and the span of the sub-spaces taken together is joint-policy space. Subsequent combination of agent gradients forms a general gradient vector field, representing the combined direction of perceived improvement. Rather than having a deterministic interpretation, this represents the unbiased expectation of joint adaption at any joint-policy, termed the *learning pressure field*.

For illustration purposes, we reformulate Zinkevich et al.'s NoSDE example (Zinkevich, Greenwald, & Littman 2005), of a non-cooperative general-sum stochastic process, as a FASP (allowing stochastic policies), then compute the learning pressure field. We discuss the potential of the learning pressure field to describe on-line learning as a meta-level stochastic process, and what is needed to predict the average reward per time-step, for each agent, within a perpetual learning environment.

Acknowledgements

My thanks go to Krysia Broda and Alessandra Russo, my PhD supervisors, whose criticisms and insights helped progress this work to where it is.

References

Baxter, J.; Bartlett, P. L.; and Weaver, L. 2001. Experiments with Infinite-Horizon, Policy-Gradient Estimation.

Bowling, M., and Veloso, M. 2004. Existence of Multiagent Equilibria with Limited Agents. *Journal of Artificial Intelligence Research 22*. Submitted in October.

Buffet, O., and Aberdeen, D. 2006. The Factored Policy Gradient Planner(ipc-06 version). In *Proceedings of the Fifth International Planning Competition*.

Dickens, L.; Broda, K.; and Russo, A. 2008. Transparent Modelling of Finite Stochastic Processes for Multiple Agents. Technical Report 2008/2, Imperial College London.

Horn, R. A., and Johnson, C. R. 1986. *Matrix analysis*. New York, NY, USA: Cambridge University Press.

Myerson, R. B. 1997. *Game Theory: Analysis of Conflict*. Harvard University Press.

Zinkevich, M.; Greenwald, A.; and Littman, M. 2005. Cyclic Equilibria in Markov Games. In *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press. 1641–1648.