Stigmergy in multi-agent reinforcement learning

Raghav Aras, Alain Dutech, François Charpillet

Stigmergy in Multi Agent Reinforcement Learning

Raghav Aras, Alain Dutech and François Charpillet
Loria \ INRIA-Lorraine
Campus Scientific, B.P. 239,
Cedex 54506, Vandœuvre-lès-Nancy, France
{aras, dutech, charpillet}@loria.fr

Abstract

In this paper, we describe how certain aspects of the biological phenomena of stigmergy can be imported into multi-agent reinforcement learning (MARL), with the purpose of better enabling coordination of agent actions and speeding up learning. In particular, we detail how these stigmergic aspects can be used to define an inter-agent communication framework.

1 Stigmergy

The term stigmergy describes a kind of cooperative phenomena that emerges in a group of simple animals, like ants. For instance, when in a colony of ants, a mass of dead ants is strewn all over it, the live ants are able to make progressively lesser and larger heaps of the carcasses until finally they pile them onto just one big heap, which is the ideal solution to the problem of ridding the colony of obstacles [2]. The stigmergic pull here is directly proportional to the amassed dead ants in any given spot. An ant is more likely to place a dead ant in a spot where there is already a mass of dead-ants rather than in a spot where there are no dead-ants. Similarly, it is more likely to pick up an isolated carcass than disturb a heap of carcasses. When each ant follows this simple, local rule, the desired global behavior "emerges". Ants do not communicate directly (through sounds or signals), but rather let their past actions message themselves as well as other ants in the future. To cite another instance of stigmergic cooperation, ants are able to determine the shorter of two given paths, between two points, using ephemeral pheromone traces [3]. No single ant can establish the necessary differential of pheromone (since it evaporates) to induce a reinforcement loop. The key properties of stigmergy can be recapitulated as follows, labeling ants as agents: $1)$ The past actions of each agent influence the current actions of other agents as well as the agent itself. $2)$ Observing local rules leads to desired global behavior. $3)$ Agents work only with a localised view of the environment.

2 Multi-agent Reinforcement learning

In reinforcement learning [6], an agent tries to learn the optimal way of doing a task. The model of the task is a Markov decision process (MDP) [6], and it is unknown to the agent. The agent does not know how the system will evolve on taking an action. The agent collects experiences of the form $< state, action, reward >$, and progressively finds the optimal action for each state. Such tasks have the Markov property which implies that the current optimal action can be decided solely on the basis of the current system state. The past is redundant. The game of chess has this property. In effect, the agent is trying to maximize the "reward" that the environment gives it on taking actions. A particular action executed in a state may be reinforced or not by the reward that the environment gives the agent for taking the action. Thus the agent does not learn the model of the system, but learns *policies* of optimal behavior directly through experience. The key features of conventional reinforcement learning are:

- Agents are memoryless. They have no direct "memory" of past states visited and actions taken. In this, they resemble stigmergic ants. Agent policies are of the form $\pi : state \rightarrow action$.
- Agents see the global state of the system perfectly .
- The state space of the system is static, and is not changed by the agents. In order to have the Markov property, the definition of state must be parameterized by all agent-influenceable variables.
- The reward function of the task is static. Agents must adapt their behavior according to the reward function.

The Q-learning algorithm is a standard single-agent reinforcement learning algorithm [6], and it calculates Q-values (expected accumulated reward on taking action a

in state s, using the following *update rule*: $Q(s,a) = (1-\alpha)Q(s,a) + \alpha(R(s,a) + \gamma max_{a'}Q(s',a'))$, where α is the learning rate and γ is the discount factor $\in (0,1)$. The result of the algorithm is an optimal reactive policy $\pi^* : s \rightarrow a$, giving best action to take in state s. When more than one controls a MDP, the process is defined as a stochastic game (SG) [4]. Adaptations of single agent RL algorithms to multiagent reinforcement learning(MARL) make restrictive assumptions: $1)$ Agents see the complete state of the system. $2)$ Each agent sees every other agent's actions and rewards. $3)$ Agents use the Q-learning rule over the *joint action space* and the *global state space*. Clearly this is not the way ants operate. They work with much less information. However, they have at their disposal, *shared, external memory*. We are interested in formulating an independent, multi-agent reinforcement learning approach, where each agent works only with its own action space. The two properties of stigmergy that make it different from an MDP (and a SG for MARL) can be summarized as: $1)$ Non-static state space (e.g., ants begin with a certain, pheromone-empty state space, and they transform it). $2)$ Non-static reward function (for e.g., ants have no particular spot in mind to gather all the dead ants; the reward function is thus undefined at the beginning).

3 Communication as Stigmergy

We can point to three kinds of external memory for agents: $1)$ individual memory $2)$ common or shared memory $3)$ Individual or private memory that is accessible to others. Individual memory has been explored to a limited extent in reinforcement learning, for partially observable environments where the optimal policy of an agent may depend on the history of the past states visited and actions taken, rather than just on the current state. In other words, the environment may be non-Markov. Common or shared memory has not received much attention for multi-agent reinforcement learning. We are more interested in the third kind of external memory, that points to direct inter-agent communication. So how is communication related to stigmergy ? We observe that ants leaving pheromones in the environment can be compared to them leaving messages in letterboxes. The messages must be *read* before they expire just as the pheromones are "readable" before they evaporate. We note that the pheromones themselves are semantic-less. It is rather their concentration (a direct function of depositors) that is of significance. All ants have access to a given pheromone concentration. In this sense, we can say that stigmergy is a sort of *persistent, broadcast communication* mechanism.

3.1 Sitgmergic MARL framework

In order to import the two properties of non-staticity of stigmergy stated previously, into MARL, we propose a framework of stigmergy-like communication. It has two components $1)$ Stigmergic messages $2)$ Reward interpretation rule. Consider the simple idea of an agent A sending agent B a message. Three related properties are associated with this communication act: $a)$ the semantic of the message $b)$ time taken for delivery $c)$ duration for which the message can be retained by the recipient. Thus we define stigmergic messages to be such that: $1)$ Messages are contentless. This is in keeping with the non-semanticity of stigmergic "communication" (pheromone traces, pile density etc) and $2)$ Messages are "beep-like" - they disappear immediately after the receiver hears them. This is in order to retain the reactive, memory-less feature of reinforcement learning agents.
Since messages last for just one time unit, they must be used/assimilated into whatever Q-learning like rule the agent uses. The agent has to "put" the received message somewhere; the state space is static and exclusionary of the communication framework (since that is purely prescriptive). So, we propose that the agents assimilate the message into their rewards, thus re-interpreting a given reward into a *virtual reward*. The idea is that if an agent finds itself in a "good" state, it can learn to send messages to other agents to attract them to that state. The receivers would be compelled to visit states where messages are to be obtained, in order to obtain higher virtual rewards. The sender has an incentive to send because the rule works two-ways, in a manner similar to that employed in encryption protocols: the sender uses a public key and a private key. The reader is asked to consult [1] for a fuller description of this approach, and wherein we present results of the efficiency of such an approach.

References

[1] R. Aras, A. Dutech, and F. Charpillet. Cooperation through communication in decentralized markov games. *Advances in Intelligent Systems - Theory and Applications*, 2004.

[2] J.-L. Deneubourg, S. Goss, N. Franks, A. Sendova-Franks, D. C., and L. Chretien. The dynamics of collective sorting robot-like ants and ant-like robots. *From Animals to Animats*, 1990.

[3] M. Dorigo and L. Gambardella. Ant colony system: A cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation*, 1997.

[4] M. Littman. Memoryless policies: Theoretical limitations and practical results. *From Animals to Animats 3*, 1994.

[5] A. McCallum. Learning to use selective attention and short-term memory in sequential tasks. *Fouth International Conference on Simulating Adaptive Behaviour*, 1996.

[6] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.