A non supervised multi-reinforcement agents architecture
to model the development of behavior of living organisms

Philippe Preux   PREUX@LIL.UNIV-LITTORAL.FR
Christophe Cassagnabère   CASSAGNA@LIL.UNIV-LITTORAL.FR
Samuel Delepoulle   DELEPOULLE@LIL.UNIV-LITTORAL.FR
Laboratoire d'Informatique du Littoral, Université du Littoral Côte d'Opale, UPRES-JE 2335, B.P. 719, 62228
Calais Cedex, France

Jean-Claude Darcheville   DARCHEVILLE@UNIV-LILLE3.FR
Unité de Recherche sur l'Évolution des Comportements et des Apprentissages, Université de Lille 3, UPRES-EA
1059, B.P. 149, 59653 Villeneuve d'Ascq Cedex, France

Our project is based on computer science and psychology. Its aim is to assess to which extent reinforcement learning (RL) can account for animal dynamics of behavior, that is, how behaviors are acquired, learnt, and evolve during lifespan. To this end, we are incrementally building an artificial creature that we call MAABAC which behavior is entirely controlled by RL algorithm. As such, MAABAC is subject to its enviroment to which it continuously adapt itself. We want MAABAC to be able to learn to reach and grasp objects, to walk, to run, ... in its "physical" environment. We have interest in the ability to reach an object as it is one of the first complex motor behavior that is exhibited by human babies. In the same time, we want that some features of learning are shared by animals and MAABAC (Multi-Agent Animat for Behavioral Arm Control). We put a strong emphasis on the development of new behaviors, that is, behaviors have to be acquired, or learnt, rather than merely being added as new components to MAABAC. This idea is shared with the BabyBot project (Metta et al., 2000). At first sight, it might seem that we follow a project which is close to the COG project at MIT (Brooks et al., 1998). However, as argued below, we want to restrict ourselves to the use of RL algorithms because that is of interest for the study of animal behavior, while COG has a more ingineer, pragmatic approach to build a robot. In the sequel, we present motivations and methodological choices, the basic architecture of MAABAC, the first stages of development of MAABAC, the first experiments and current conclusions. Along the text, we outline some points directly related to the research being performed in the RL community.

Reinforcement algorithms (Sutton & Barto, 1998) provide an interesting implementation of a fundamental law of animal behavior, namely the law of effect (Thorndike, 1911). Then, assuming that the behavior of living organisms is fundamentally based on the law of effect, we can try to build artefacts based on TD that learn their behavior through interacting with their environment.

Basically, the architecture of MAABAC is a non supervised multi-agent system (MAS). There are two reasons for this. The first reason regards modeling: it is natural that each "organ" maps one agent. The second is that of reducing the size of search space; assuming that each organ can be in one of any N states, a system made of k agents would yield a search space of size N^k, whereas a MAS approach leads to k problems of size N, thus $k \times N$, which is far smaller than N^k; interactions between the agents guarantee that constraints are respected; in summary, a MAS approach simplifies the search of space by each individual agent. Each agent is controlled by a TD algorithm.

We have designed an arm, originally introduced by S. Delepoulle in his PhD thesis (Delepoulle, 2000). This arm is made of two joint segments (bones); one extremity is the shoulder, the other is the fist (no hand for the moment), while the joint is the elbow. The position of each segment is controlled by two muscles, one flexor and one tensor. Each muscle can either contract, or relax itself a little bit more, or remains in its current state of contraction: these are the 3 possible actions of the agent, and the number of states of contraction is $N = 50$. The position of each segment of the arm is simply given by the difference of contraction of the two muscles controlling it. Each muscle is

an agent which behavior is controlled by a Q-learner so that to follow the law of effect. When the arm is activated, each state of contraction leads to a certain energetic cost, the more contracted, the more costly; the energetic cost is handled as a negative reward in each Q-learner; when the arm puts its fist in a certain target zone of the space, it gets a positive reward, while otherwise, it only receives a negative reward (energetic cost). Among the various tasks that could have been used, we consider the one aiming at putting the fist into the target zone. As far as each agent has only access to its current state of contraction, the whole task is non markovian. The whole architecture is thus a hybrid of non supervised and reinforcement learning.

This overall presentation being made, let us present and discuss the simulations that have been performed. When activated in an initially random position, the arm first shows erratic behaviors until its fist reaches the target zone for the first time. Then, the arm receives a positive reward for the first time and at that time, it "learns" that it can receive a reward. More precisely, the 4 muscles receive the same positive reward, either they have contributed to receive it, or not. After resetting its position and after a certain amount of reachings of the target zone, the movement of the arm becomes smooth and straightforward towards the zone. We call a "movement" the whole sequence of behaviors from the initial position to the target zone. Then, we have shown that the arm exhibits various abilities and perform multi-task reinforcement: learning to reach different positions of the target zone, learning to reach them from different initial positions, acquiring movements with an optimal trajectory (with regards to their energetic cost, considering this precise representation of space), extinction, generalization, shaping, tracking a moving target, escaping an uncomfortable zone, and avoiding such a zone. Clearly, these behaviors and movements are the outcome of the use of TD algorithms. The coordination and self-organization of muscles come from the fact that the reward is shared by the muscles while, in the same time, the negative reward due to energetic cost leads to the optimization of movements, thus their straightforwardness and their smoothness. The precision of the movements is bounded by the number of states of contraction of the muscles: the more states, the more precise the movement. To obtain finer movements while avoiding too long learning time, we have added a mechanism of state splitting (Moore & Atkeson, 1995).

Then, we have added a second arm to the artefact, itself driven by 4 muscles (exactly the same architecture as the first arm). We have done the same experiments as with the first arm. To reduce the time to learn correct movements for the two arms, we have also used the following idea. Once one arm had learnt several movements, we have used its Q-values to initialize the second arm. Then, the second arm is readily able to perform the same movements as the first, whereas both may still adapt freely their own behaviors to their own contingencies.

The third step has been to add a body on which the two arms are attached by their shoulders. The body can turn on itself under the action of a muscle, again controlled by a Q-learner. The whole artefact is then controlled by 9 Q-learners, without any central supervisor. It has shown its ability to learn optimal movements. We have noted that the learning time for this third step artefact scales very favorably with regards to the simple arm: actually, a movement is acquired faster by this artefact than by the version made of only two arms (without a body).

The next steps will be to add legs to the artefact as well as senses to let it perceive its environment. We are also considering using TD(λ) instead of Q-learning, foreseeing that it will speed-up learning, while remaining realistic with regards to the law of effect. We are also paying some attention to training to speed-up the initial exploratory phase, and study the interaction between supervised, reinforcement, and non supervised learnings.

References

Brooks, R., (Ferrell), C. B., Irie, R., Kemp, C., Marjanovic, M., Scassellati, B., & Williamson, M. (1998). Alternate essences of intelligence. *AAAI*.

Delepoulle, S. (2000). *Coopération entre agents adaptatifs ; étude de la sélection des comportements sociaux, expérimentations et simulations*. Doctoral dissertation, Université de Lille 3, URECA, Villeneuve d'Ascq. Thèse de doctorat de Psychologie.

Metta, G., Manzotti, R., Panerai, F., & Sandini, G. (2000). Development: is it the right way towards humanoid robotics? *IAS-6*. Venice, Italy.

Moore, A., & Atkeson, C. (1995). The parti-game algorithm for variable resolution reinforcement learning in multidimensional state-spaces. *Machine Learning, 21*.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: an introduction*. MIT Press.

Thorndike, E. (1911). *Animal intelligence: Experimental studies*. Mac Millan.