Learning from Reward as an emergent property of Physics-like interactions between neurons in an artificial neural network.

Frédéric Davesne

Learning from Reward as an emergent property of Physics-like interactions between neurons in an artificial neural network.

Frédéric Davesne
LPPA CNRS UMR 7124, Collège de France

11 Place Marcelin-Berthelot, 75005 Paris, France

Abstract. We study a class of artificial neural networks in which a physics-like conservation law upon the activity of connected neurons is imposed at each time. We postulate that the modification of the network activities may be interpreted as a learning capability if a judicious conservation law is chosen. We illustrate our claim by modeling a rat behavior in a labyrinth: the exploration of the labyrinth permits to create connections between neurons (latent learning), whereas the discovery of food induces a one step backpropagation process over the activities (reinforcement learning). We give theoretical results about our learning algorithm CbL and show it is intrinsically faster than Q-Learning.

1 Introduction.

1.1 Why to utilize a conservation law ?

Living entities (from cells to humans) have the ability to adapt themselves to an environment they do not completely know *a priori*. Astonishingly, the *a priori* uncertainty is compatible with the *a posteriori* robustness of the entities behavior in their environment: it is *highly predictable* that a baby will learn successfully how to walk and speak his mother tongue.

This dilemma is the core issue in building adaptive and *robust* artifact behaviors. Learning methods cope with the adaptation of the artifact. But what "highly predictable" means in a modeling point of view ? It is often associated to convergence properties of a method to an *optimal or suboptimal behavior*. However, if we take the example of reinforcement learning (RL) which theoretically owns this property [7], experiments in the real world have stressed some practical limitations, mainly because the Markovian hypothesis (which ensures the convergence to an optimal behavior) is invalid in a lot of real applications [5]. This inadequacy between the theoretical hypothesis and the real world

may lead to chaotic or quasi-random behaviors after the learning stage [3],[2].

We think that algorithms like RL may fail in practice due to nature of the object the predictivity must be applied to: (statistical) goodness of the learnt behavior, *whatever the environment is*. If we apply a conversation law on the interaction between an artifact and its environment, we may be able to (a) deduce theoretically the set of possible evolutions of the artifact through its interaction to the set of all possible environments; (b) divide this set into two sub-sets: interesting evolutions in terms of the resulting behavior and uninteresting evolutions; (c) specify the set of environments that are likely to lead to interesting behaviors (we shall say learnt behaviors); (d) compare this theoretical set to the set of real environments. The artifact will learn the behavior in reality (i.e. follow an interesting evolution) if and only if the two sets of environments are "compatible".

1.2 The case study: facts, modeling and notations.

This paper is dedicated to the application of our strategy based on conservation law to the behavior of a simulated rat in a labyrinth. In this paragraph, we will specify the structure of our model, looking at the three following facts:
(a) It has been shown that the rat may use a *cognitive map* for its navigation [4]. (b) There are also evidences that the rat begins to learn the topology of an unexplored labyrinth before having any reward: this has been called *latent learning* [6]. (c) And, at the moment a rat has discovered food, it is able to navigate to it from any point of the labyrinth (reinforcement learning).

Our model is comparable to a Q-Learning (QL) model [7]. We suppose that the artifact has p actions $a_1, a_2, \ldots, a_p$. We consider a set of n states $s_1, s_2, \ldots, s_n$ associated to the firing of place-cells $c_1, c_2, \ldots, c_n$ for n locations covering the labyrinth (fact (a)); the activity of each cell c_i is q_i. For each cell c_i, we associate a set of p states $s_{i,1}, s_{i,2}, \ldots, s_{i,p}$. A $s_{i,j}$ is a transitory state meaning that cell c_i has fired and the execution of action a_j has been decided. The transitory state $s_{i,j}$ is supposed to be materialized by a cell $c_{i,j}$ which activity is $q_{i,j}$, representing the Q-value for the QL techniques. Two extra states, namely s_p and s_f, are supposed to exist: s_p is associated to the hit of an obstacle and s_f to the discovery of food. When the rat hits an obstacle, its state becomes s_p; when it discovers food, its states moves to s_f. s_p is linked to a cell c_p (activity is q_p) whereas s_f is associated to a cell c_f (activity is q_f).

Each cell c_i is *a priori* connected to the p cells $c_{i,1}, c_{i,2}, \ldots, c_{i,p}$. Connections $t_{i,j,k}$ from a cell $c_{i,j}$ to a cell c_k may be created during the artificial rat's exploration of the labyrinth to meet fact (b). The set of all connections $t_{i,j,k}$ starting from a cell $c_{i,j}$ is called $T_{i,j}$.

2 Constraint based Learning algorithm.

2.1 Chosen conservation law.

The conservation law has been chosen to permit a comparison between CbL and RL techniques. It is applied to the activities of connected cells and is very closed to the *minimax* algorithm. Its expression relies on the two following equations:

$$q_i = \alpha \left[\max_{j \in \{1,\ldots,p\}} \{q_{i,j}\} \right] \quad (1)$$

$$q_{i,j} = \alpha \left[\min_{k \in T(i,j)} \{q_k\} \right] \quad (2)$$

Where α is a scalar in $]0,1[$. If $T(i,j)$ is an empty set, we assume that q_i must be equal to 0. The conservation law implies that equations 1 and 2 must be fulfilled for *every* cell c_i and *every* cell $c_{i,j}$ of the network *at any time*.

2.2 Backpropagation process and decision making.

At the moment the rat is firstly introduced into the labyrinth, we suppose that no connections $t_{i,j,k}$ exist. Looking at equations 1 and 2, we deduce that initial activities q_i and $q_{i,j}$ must be equal to 0. A connection from $c_{i,j}$ to c_k (resp. c_p or c_f) is added if the artifacts moves for the first time from state $s_{i,j}$ to state s_k (resp. s_p or s_f).

The modification of the activities in the network may occur if: (a) a connection is created; (b) q_f or q_p are changed. If a connection is created, a $T_{i,j}$ grows and the equation associated to $c_{i,j}$ may not be satisfied. Whereas if q_f (resp. q_p) changes, the equations associated to all $c_{i,j}$ connected to c_f (resp. c_p) may not be fulfilled. If an equation is not fulfilled, the right term remains unchanged (as it is done in the temporal difference method) whereas the left term is set to the value of the right term: for equation 1 (resp. equation 2), q_i (resp. $q_{i,j}$) is modified. These modifications may lead to the unfulfillment of other equations, which involve the modification of other q_i or $q_{i,j}$, and so on: this is our *backpropagation process*.

The decision making process of the artifact is based on a greedy policy: when the artifact is in state s_i (c_i is firing), we determine a sub-set $S = \{a_{j_1^*}, \ldots, a_{j_u^*}, \ldots, a_{j_m^*}\}$ of $a_1, \ldots, a_p$: $q_{i,j_u^*} = \max_{j \in \{1,\ldots,p\}}\{q_{i,j}\}$. In the case S is not reduced to one element (which often happens), one action from S is chosen randomly. This permits the *exploration* of the labyrinth.

2.3 Theoretical results.

The use of a conservation law leads to a *one step backpropagation process*. If the conservation law is not chosen carefully, this process may turn into a loop

and never ends (oscillation of activities). It is fundamental to prove that this process is *stable*: this is the case for CbL.

The interesting configurations of the network activities (i.e. the artifact learns a good navigation behavior) are obtained if the interaction with the environment leads to the creation of one connection (at most) from any $c_{i,j}$. If this condition is not violated, the following facts may be proved (see [1] for details):

- As soon as the artifact has found food, it is able to return to it from every explored place (due to the one step backpropagation process).
- Moreover, if the labyrinth do not change and the q_f and q_p values are constant, the value of the activities converge in a *finite window time* (as soon as food is discovered).
- The navigation of the artificat through the *explored parts* of the labyrinth is then *optimal*; but it may not be optimal considering the whole labyrinth (with its unexplored parts): see fig. (b).
- CbL is *intrinsically* faster than QL (with eligibility trace)(see fig. (a)).
- The behavior of the artifact may be immediately changed by the way of q_f and q_p (see fig. (c) to (f)).

3 Experiments.

This section aims to show some experimental details that illustrate the theoretical results given in the former section.

3.1 Wandering and looking for food.

The model of the labyrinth is a 10×10 grid in which the artifact may use four actions (left,right,up,down). Each unit of the labyrinth is associated to a cell c_i. The "wandering" and "looking for food" behaviors are generated by considering a relation between the "energy" of the rat and q_f: $E + q_f = 0$. If we suppose that E diminishes at each move and is set to 1 if food is reached, the two behaviors alternate. Figure (f) shows that as soon as the artifacts finds food for the first time (step 120), q_f (hence E) oscillates regularly: the rat is able to go for food as soon as it is hungry ($E < 0$). Figure (d) shows the equipotentials of activities in the network in this situation: an attraction basin is built to guide the rat to the food place. The shape of this curve depends deeply on the exploration of the labyrinth (see fig. (c)): the second place of food has not been explored yet so that only one basin exists. The exploration of the labyrinth may be continuated, even if a first place of food has been discovered, because this place is avoided (as an obstacle) until $E < 0$ (fig. (e)).

3.2 Comparison between CbL and Q-Learning.

We evaluate the mean performance of CbL and Q-Learning (with eligibility trace) over 1000 learning phases. Each learning phase is composed of 1000 trials. Each trial ends if food ($q_f = 1$), obstacle ($q_p = -1$) is encountered or if 5000 learning steps have been performed. The artifact begins a trial in a unit chosen randomly.

The convergence is evaluated by looking at the cumulative absolute difference of Q-values for QL, and at the cumulative number of modified activities for CbL. The curves of graph (a) shows the mean convergence for QL and CbL over the 1000 learning phases, from learning step 1 to 5000. CbL curve starts to diminish long before QL, whereas the slope of the two diminishing curves are comparable. This difference is due to the one step backpropagation process of CbL. However, the navigation behavior for CbL may not be optimal: figure (b) compares the mean number of steps needed to go for food and we can see that QL is better than CbL. This disadvantage disappears in the case the whole labyrinth had been explored before finding food for the first time.

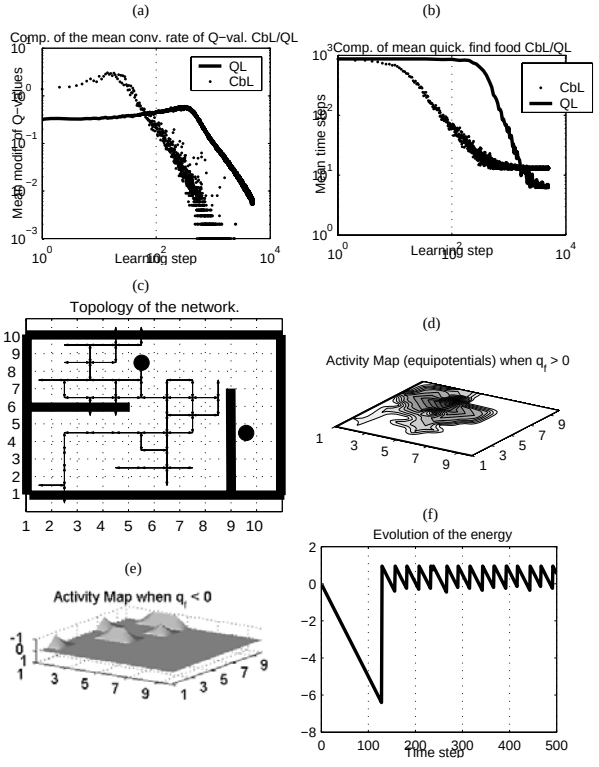

Figure 1: Experimental results.

4 Conclusion.

In this paper, we have focused on a class of artificial neural networks in which a constraint has been put on the activity of connected neurons; this constraint is comparable to a physics-like conservation law. This implies that the modification of the neural activities is caused by a change in the topology of the network (add or delete a connection between two neurons) or by a change in the activity of specific neurons associated with a reward: this induces a *one step backpropagation process*. Two main results may be stressed.
First, from a theoretical point of view, the convergence of a learning algorithm based on a conservation law relies mostly on the ability to prove that the backpropagation process is stable (i.e. do not oscillate). Second, our learning algorithm (CbL) has been compared favorably to a classical Q-Learning technique: (a) the learning performances are much faster; (b) CbL is truly incremental; (c) CbL has no internal parameter. The learning process is fast because the modification of the activity of a neuron is propagated to the other neurons in *one step*.
But there exists a strong limitation on CbL: noisy environments are likely to lead to uninteresting configurations (even if CbL has been proved to converge in all situations). It is due to the *deterministic* nature of the chosen conservation law. To overcome this issue, we are working on conservation laws that are not applied on real values but on *densities of probability*.

References

[1] F. Davesne. *Etude de l'émergence de facultés d'apprentissage fiables et prédictibles d'actions réflèxes, à partir de modèles paramétriques soumis à des contraintes internes.* PhD thesis, University of Evry, France, 2002.

[2] F. Davesne and C. Barret. Influence of the context of a reinforcement learning technique on learning performances a case study. In *AIA'03*, Benalmádena, Spain, 2003.

[3] U. Nehmzow and K. Walkery. The behaviour of a mobile robot is chaotic. *AISB*, 1(4), 2003.

[4] J.O. O'Keefe and L Nadel. *The Hyppocampus as a Cognitive Map.* Clarendon Press, Oxford, 1978.

[5] M.D. Pendrith. Reinforcement learning in situated agents: Some theoretical problems and practical solutions. In *EWLR'08*, 1999.

[6] J. Seward. An experimental analysis of latent learning. *Journal of Experimental Psychologie*, 39:177–186, 1949.

[7] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An introduction.* MIT Presss, Cambridge, MA, 1998.