**Project no.:** **027657**

**Project full title:** **Perception, Action & Cognition through Learning of Object-Action Complexes**

**Project Acronym:** **PACO-PLUS**

# Deliverable no.: D6.3

# Title of the deliverable: Strategies for modification of Actor-Critic loops

| | | |
|---|---|---|
| **Contractual Date of Delivery to the CEC:** | **31/07-07** | |
| **Actual Date of Delivery to the CEC:** | **08/08-07** | |
| **Organisation name of lead contractor for this deliverable:** | **BCCN** | |
| **Author(s): BCCN** | | |
| **Participants(s): BCCN** | | |
| **Work package contributing to the deliverable:** | **WP6** | |
| **Nature:** | **R/D** | |
| **Version:** | **1.1** | |
| **Total number of pages:** | **6** | |
| Start date of project: | 1st Feb. 2006 | **Duration:** 48 month |

**Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)**

**Dissemination Level**

| | | |
|---|---|---|
| **PU** | Public | **X** |
| **PP** | Restricted to other programme participants (including the Commission Services) | |
| **RE** | Restricted to a group specified by the consortium (including the Commission Services) | |
| **CO** | Confidential, only for members of the consortium (including the Commission Services) | |

**Abstract:**

This deliverable discusses Actor-Critic Architectures for reinforcement learning in PACO-PLUS. Given the recent developments achieved by the consortium, we conclude that Actor-Critic Architectures should by replaced by a neurally compatible version of SARSA learning.

# Strategies for modification of Actor-Critic Loops

Note: Parts of this deliverable have been presented already in the first progress report. The goal of this deliverble is to assess Actor-Critic architectures in conjunction with PACO-PLUS.

The idea behind using an Actor-Critic algorithm lies in its biological realism. For instance there exist many models of the basal-ganglia (see [JNR02] for a review) which try to implement the Actor-Critic in a biologically realistic way.

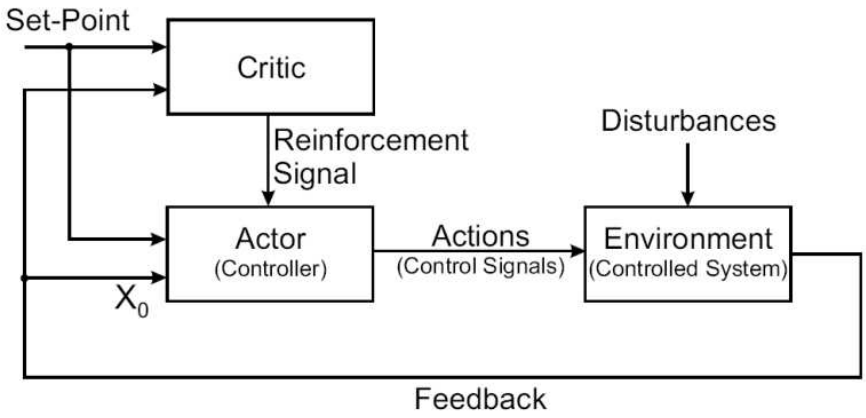


Figure 1: Actor-Critic control system, where a Critic influences action selection by means of a reinforcement signal.

The algorithm (see Fig 1) originated from the field of Reinforcement Learning (RL, [SB98]). In RL an agent maximizes the rewards $r$ it will receive when following a policy traveling along states $s$. There exist many algorithms where almost all of them rely on the temporal difference (TD) learning (Eq. 1, [Sut88]) rule similar to the *critic* in the Actor-Critic algorithm.

$$V(s_i) \rightarrow (1-\alpha)V(s_i) + \alpha(r(s_{i+1}) + \gamma V(s_{i+1})), \tag{1}$$

where $V$ is the value of a state $s_i$, $r$ the reward, $\gamma$ the discount factor and $\alpha$ is the learning rate.

Additionally the Actor-Critic has a separate memory structure to explicitly represent the policy which is named *actor*

$$p(s_i, a_i) \rightarrow p(s_i, a_i) + \beta\delta_{s_i}, \tag{2}$$

where $p$ is the probability for a certain action $a_i$ to be taken from state $s_i$ and $\beta$ a rate factor. It chooses actions that will lead to states with higher reward expectations according to the TD-error:

$$\delta_{s_i} = r(s_{i+1}) + \gamma(V(s_{i+1}) - V(s_i)) \tag{3}$$

However, in the general case an Actor-Critic converges badly and the convergence as such cannot be guaranteed.

An alternative description that also uses a modified version of the temporal difference learning rule but lacks biological realism is Q- or SARSA learning (the difference between both algorithms will be discussed later, $b(s_{i+1})$ in Eq. 4). For these algorithms no explicit *critic* is necessary:

$$Q(s_i, a_i) \rightarrow (1 - \alpha)Q(s_i, a_i) + \alpha(r(s_{i+1}) + \gamma Q(s_{i+1}, b(s_{i+1}))). \tag{4}$$

This gives us the opportunity to handle a much more compact framework for both: evaluating conducted actions and selecting new actions. Furthermore it is known that these computations are superior when dealing only with a restricted and well-defined set of actions which is the case for e.g. the ARMAR robot.

The only difference between Q- and SARSA learning is the dependence of the update rule on the policy. In SARSA learning [SJLS00] the actually conducted action is applied on the update rule [on-policy, $b(s_{i+1}) = a(s_{i+1})$], Q-learning [WD92] uses always the optimal action independent of the last choice [off-policy, $b(s_{i+1}) = argmax_\eta(\eta(s_{i+1}))$]. Both strategies have advantages and disadvantages but it is important to mention that the Actor-Critic algorithm is only on-policy however Q- and SARSA learning incorporate both on- and off-policy. To switch between these both policy learning strategies only one simple module must be changed. This offers a flexible mechanism for future tasks.

Additionally, a recent publication [MNA+06] showed that SARSA learning and not an Actor-Critic is used by primates. However, until recently it could not be ascertained that SARSA learning can be implemented in a biologically realistic way.

But now it has been finally been shown by us that SARSA learning can indeed be emulated with biological realistic neurons that use correlation based learning [PW03, WP04]. The architecture is depicted in Fig. 2 and the results in Fig. 3.

Additionally the equivalence between RL and Correlation based learning could be proven [KPW07].

As RL depends on discrete states and actions, an extension to a continuous space is required. Within the PACO-PLUS project a function approximation method was developed for the framework of Q- and SARSA learning [TAK+ed]. The achieved results improve on former approaches and, under general conditions, the algorithm guarantees convergence.

# 1 Summary Argumentation and Conclusion

**Status at the time of writing the TA:**

- Reinforcement learning (RL) in the context of PACO-PLUS must be efficient and should be biologically motivated.

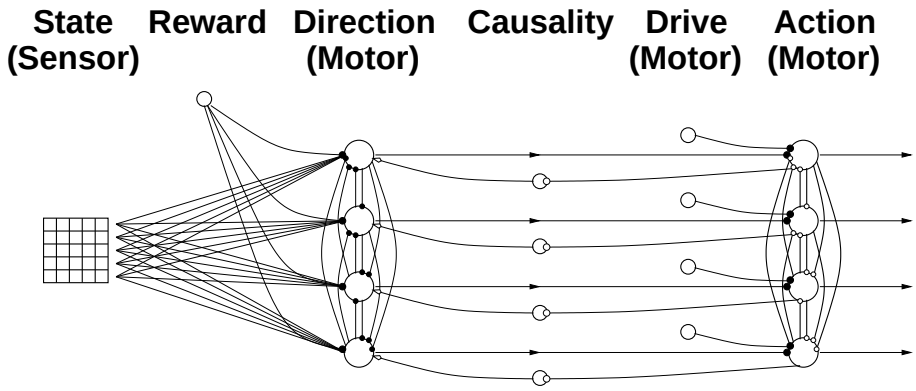


Figure 2: This diagram shows the architecture of the neuronal equivalent of SARSA Learning. Filled circles depict excitatory connections, empty circles inhibitory connections and empty arrows the third input $r(t)$ (see equ 4). The Q-values are located between the states-layer and the directions-layer. The lateral excitation within the directions-layer guarantees an excitation of all layer neurons when an action is conducted. In contrast the lateral inhibition of the motor-layer ensures an excitation only of the neuron that fires first. The driving neurons perform a random movement when the weights of the current state are zero and the causality neurons feed back the information which motor neuron actually fired. Additionally a reward is provided to the direction-layer.

- Actor-Critic Architetures where the only known RL-Method that supported these two notions at least to some degree.
- Convergence control for Actor-Critics, however, was known to be problematic.
- As a consequence better alternatives had to be found and Actor-Critics were only to be considered if the search for better alternatives failed.

**Status now:**

- Q-learning and SARSA are more reliable but not really compatible to neuronal functions[1].
- Q- and SARSA-learning have beneficial convegence properties.
- A neuronal implementation of SARSA has been achieved by us (see also first PACO-PLUS report). SARSA relies on the TD-rule.

[1] A detailed argument exists here about the function of the Dopaminergic system, which supports the biological realism of TD-learning [Sch98] but this does not immediatelly also hold for Q- and SARSA.

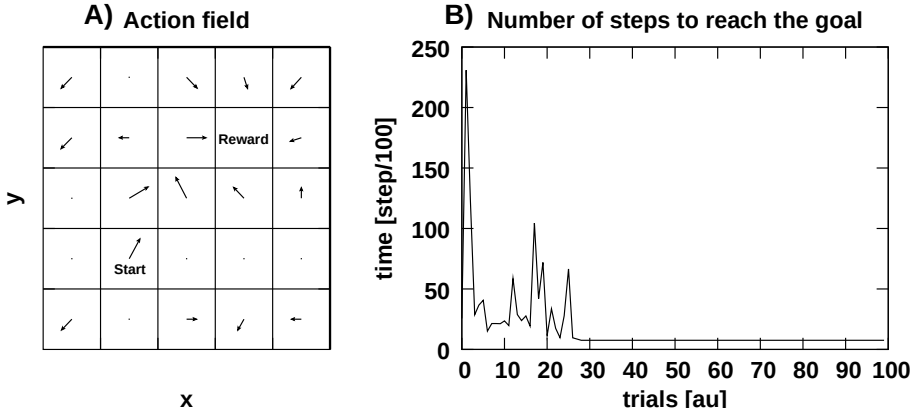


Figure 3: Results for Neuronal SARSA. Panel A shows an action field of each state. The arrows point in the direction of the most probable action. In panel B the time until the reward is found is plotted.

- A proof has been achieved by us that TD-learning can be made fully equivalent to differential Hebbian learning
- Physiological evidence exists for the use of SARSA in primates [MNA+06]

Conclusion: This offers the option to implement the much more efficient SARSA-algorithm in a biologically realistic way. Hence, Actor-Critic architectures will not be pursued any longer in the context of PACO-PLUS.

# References

[JNR02] D. Joel, Y. Niv, and E. Ruppin. Actorcritic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*, 15:535–547, 2002.

[KPW07] C. Kolodziejski, B. Porr, and F. Wörgötter. On the equivalence between hebbian and reinforcement learning. *Physical Review Letters (submitted)*, 2007.

[MNA+06] G. Morris, A. Nevet, D. Arkadir, E. Vaadia, and H. Bergman. Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience*, 9 (8):1057–1063, 2006.

[PW03] B. Porr and F. Wörgötter. Isotropic Sequence Order Learning. *Neural Computation*, 15:831864, 2003.

[SB98] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[Sch98] W. Schultz. Predictive reward signal of dopamine neurons. *J. Neurophysiol.*, 80:1–27, 1998.

[SJLS00] S. P. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38(3):287–308, 2000.

[Sut88] R. S. Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, 3:9–44, 1988.

[TAK+ed] M. Tamosiunaite, J. Ainge, T. Kulvicius, B. Porr, P. Dudchenko, and F. Wörgötter. Path-finding in real and simulated rats: On the usefulness of forgetting and frustration for navigation learning. *J. Comp. Neurosci.*, submitted.

[WD92] C. Watkins and P. Dayan. Technical note:Q-Learning. *Mach. Learn.*, 8:279–292, 1992.

[WP04] F. Wörgötter and B. Porr. Temporal Sequence Learning, Prediction, and Control - A Review of different models and their relation to biological mechanisms. *Neural Computation*, 17:245–319, 2004.