# Cost minimisation and Reward maximisation. A neuromodulating minimal disturbance system using anti-hebbian spike timing-dependent plasticity.

Karla Parussel
*Department of Computer Science and Maths,
University of Stirling,
Stirling, FK9 4LA, Scotland
kmp@cs.stir.ac.uk

Leslie S. Smith
†Department of Computer Science and Maths,
University of Stirling,
Stirling, FK9 4LA, Scotland
lss@cs.stir.ac.uk

**Abstract**

In the brain, different levels of neuro-active substances modulate the behaviour of neurons that have receptors for them, such as sensitivity-to-input, Koch (1999). An artificial neural network is described that learns which actions have the immediate effect of minimising cost and maximising reward for an agent. Two versions of the network are compared, one that uses neuromodulation and one that does not. It is shown that although neuromodulation can decrease performance it agitates the agent and stops it from over-fitting the environment.

## 1 Introduction

Fellous (1999) proposes that emotion can be seen as continuous patterns of neuromodulation of certain brain structures. It is argued that theories considering emotions to emanate from certain brain structures and from non-localised diffuse chemical processes should be integrated. Three brain structures are considered in this way; the hypothalamus, amygdala and prefrontal cortex.

Fellous (2004) further suggests that the focus of study should be on the *function* of emotions rather than on what they are. Seen in this way, animals can be seen functionally as having emotions, whether or not we empathise with them. Given this, robots can functionally have emotions as well. One function of emotions mentioned that has a robotic counterpart is to achieve a multi-level communication of simplified but high impact information.

One way of studying the functionality of emotions, is to identify the extra functionality provided by neuromodulation compared to a non-modulating solution. Modulation is used here to signal agent needs in a neural network that is used for the purpose of action selection. The structures of both solutions are inherently the same but the modulating version has the added interaction between neuromodulation and neural network.

Although neuromodulators and hormones have been emulated for the purpose of action selection before, Avila-Garcia and Canamero (2004) Shen and Chuong (2002), they have not been applied to a purely neural network solution and have not been compared to non-modulating versions. Husbands (1998) evolves controllers inspired by the modulatory effects of diffusing NO. This speeds up evolutionary production of successful controllers.

The difficulty is that what can be done with a modulating network, can also be done with a non-modulating network if it has been evolved for that purpose. Therefore the comparison needs to be made in an environment that the agent has not been evolved for.

## 2 Application of the model

An adaptive agent needs a reason to adapt in order to do so. A common reason is to maximise and retain resources. In this context a resource is a single continuous scalar value that correlates with a characteristic of the state of the agent or environment. A resource can correlate with a single quantifiable level such as a battery charge level for a physical robot, or be an estimation of a virtual non-measurable level such as utility or safety. An adaptive agent is faced with two tasks when maximising these resources, that of learning to perform actions which result in an increase in a resource level, and that of learning not to perform actions which result in a decrease of resource.

Here, the Artificial Life animat concept is abstracted to provide the simplest possible context for testing the effect of neuromodulation applied to an artificial neural network. The agents have no external senses to adapt to and can only sense their internal state. The choice of output directly and immediately alters the internal state of the agent, which therefore alters the strength of the input signal to the network.

The agent has a body that requires two resources, energy and water. It keeps track of the largest increase and decrease of each. The current change in resource level is then scaled to these maxima to be within the range [0,1] before being passed to the network.

The agent is given a set of actions that increase or decrease by one or two resource points[1], or are neutral to, either the energy or water level in the body. There is one action for each permutation making 10 in total.

# 3 Implementation

## 3.1 Topology

The network consists of three layers of adaptive leaky integrate and fire neurons learning via spike timing-dependent plasticity, G. Q. Bi (2002). The model learns which outputs should be most frequently and strongly fired to minimise the level of input signal. There is one output neuron per action. The action has an effect on the internal state of the agent, which determines the strength of the input signal to the network. For the modulating network, the input layer neurons increase modulator strengths when fired, while the middle layer neurons have receptors for those modulators.

There are situations in which an effective behaviour for an agent may decrease a need but not satisfy it. For example, if it is in an environment which is temporarily bereft of resources then waiting and conserving its current levels may be the optimal behaviour. Alternatively there may be situations in which an agent needs to store more resources than it normally does. In this case the need for the resource will be signalled despite that need being signalled as satisfied. An example would be an agent expecting to find itself in an environment bereft of resources.

For each resource the input layer has two neurons that output to the middle layer. One neuron signals the need for the resource and the other neuron signals the satisfaction of that need. If a previous action performed by the agent results in a decrease in hunger or an increase in resource satiation, then the corresponding input signal is momentarily decreased.

The model uses a feed forward network that can be iterated over a number of times within a single turn, after which the winning output neuron is chosen. Which neuron wins is determined by summing up the total charge of each neuron over all the iterations and choosing the neuron with the greatest sum. This stops a neuron with strong inputs from losing because it just has spiked and thus has low activity or is in a refractory period.

## 3.2 Modulators

Two variants of the network were created; modulating and non-modulating. They were the same except that the modulating network had in addition two modulators, one used to signify hunger and the other thirst.

As used here, a modulator is a global signal that can influence the behaviour of a neuron if that neuron has receptors for it. The signal decays over time, specified by the re-uptake rate, and can be increased by firing neurons that have secretors for it.

Neurons within the middle layer are given a random number of receptors. These can be modulated by neurons in the input layer that have secretors. These neurons were given a random number of secretors. The receptors modulate either the neuron's sensitivity to input or probability of firing. The extent of this is determined by the level of the associated modulator and whether the receptor is inhibitory or excitatory. The secretors increase an associated modulator. The modulation rate of the receptors and the increment rate of the secretors is determined by evolution.

# 4 Parameter Optimisation

The network has a number of parameters which must be set correctly for it to adapt successfully. These are parameters that have no obvious value, such as the number of neurons in the middle layer, secretion, modulation rates etc. Automated parameter optimisation was performed for the modulating and non-modulating agents. Afterwards the parameter sets were hard-coded and tests were performed upon a population of agents using them.

The fitness of an agent was determined by

$$Energy + Water + Age - |Energy - Water|$$

The difference between the energy and water resource was subtracted from the fitness as both resources were essential for the agent to stay alive.

[1] Points are used as it is an arbitrary level that has no correlation with any real physical quantity.

# 5 Results

After optimisation, a modulating and a non-modulating agent were picked for further testing. The fitness of the genotypes were equivalent and both were typical of the solutions that were evolved. Because there was a stochastic mapping from genotype to phenotype and to provide multiple evaluations, the agents were hard-coded so that they could be tested as a population.

Parameter optimisation converged upon a fully hebbian network for the non-modulating network and a hybrid anti-hebbian / hebbian network for the modulating network.

## 5.1 Initial tests

When viewed over the course of the agent lifetime it can be seen that a typical agent learns which actions provide minimal disturbance to its inputs. It initially chooses a neutral action before settling on the most rewarding water action. The agent then alternates between this and the most rewarding energy action; see figure 1. Figure 2 shows the initial learning process before one output neuron wins over all the others.

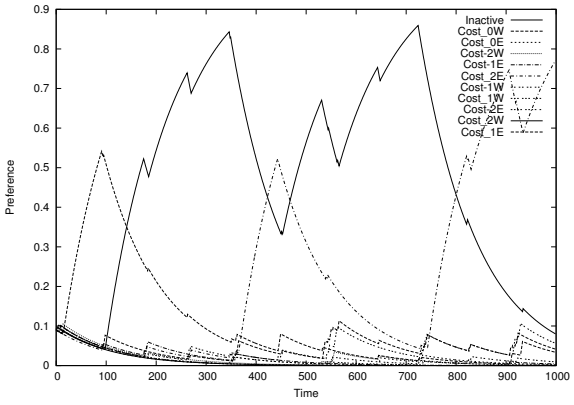


Figure 1: Actions chosen over lifetime of a single modulating agent.

The performance of the non-modulating and the modulating agents were similar although on average the non-modulating network would reach higher levels of fitness and would be optimised by the parameter search more quickly.

## 5.2 Extended tests

During parameter optimisation, each genotype was tested for 1,000 turns before being evaluated. The evaluation was cut short if the agent died prematurely because a resource had decreased to nothing.

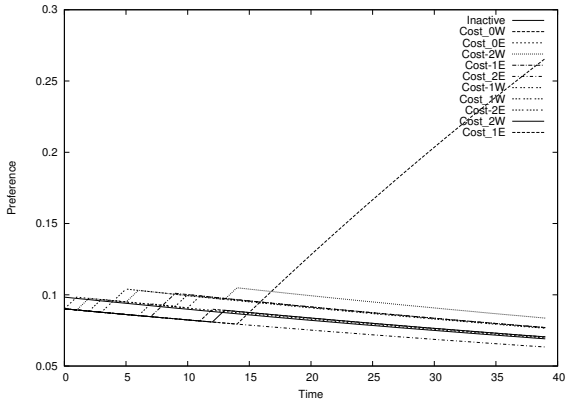


Figure 2: The first 40 cycles of the run in figure 1 showing the initial learning process.

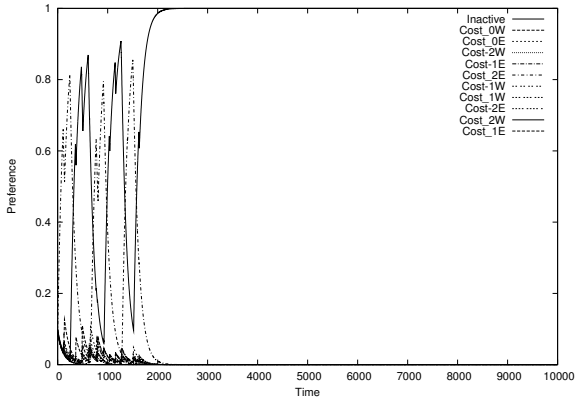


Figure 3: Non-Modulating agent run over an extended period of time (10,000 turns).

After parameter optimisation, when testing a population of hard-coded non-modulating agents for longer than 1,000 turns, the activity in the network ceased over time. The charge of the output neurons would slowly decay over time with the winning action remaining the same each time; see figure 3.

The limited use of artificial evolution for parameter optimisation had settled upon a brittle strategy which depended on how long each agent was evaluated for.

A population of hard-coded modulating agents were then tested for the same extended period of time. They were shown to continue transitioning between the same two winning output neurons that caused a maximum increase in energy and water, with other neurons very occasionally being chosen; see figure 4. Modulation had prevented the artificial evolution

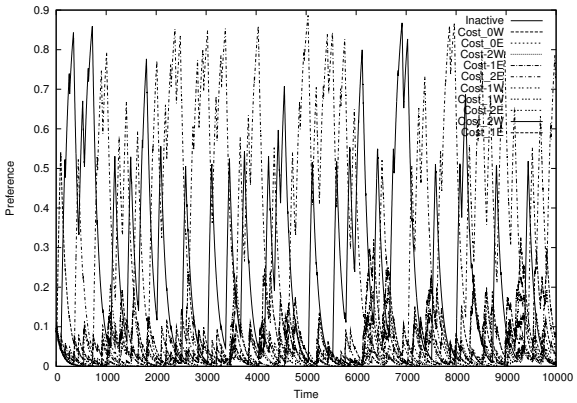


Figure 4: Modulating agent run over same extended period of time.

used for the parameter optimisation from over-fitting the test environment.

# 6 Discussion

It was discovered that the network performed most effectively when the actions it chose could minimise input activity. Wörgötter and Porr (2004) provide an overview of the field of temporal sequence learning. They discuss how the learning paradigm of disturbance minimisation, as opposed to reward maximisation, removes the problem of credit structuring and assignment. The two paradigms are not equivalent. Whereas maximal return is associated with a few points on a decision surface, minimal disturbance uses all of the points. Every input into the system drives the learning and when there are no inputs then the system is in a desirable, stable state.

Modulation agitates the network, stopping it from settling into a stable state for too long or letting activity decline to a point whereby the network stops alternating between actions. When tested using an extended run, the modulating network, unlike the non-modulating version, continues to alternate between the actions causing the least input disturbance throughout its lifetime. Figure 4 shows that other actions always have a chance of being selected.

When comparing the modulating and non-modulating agents in environments that they were not evolved for, in this case evaluated for an extended length of time, then it is shown that modulation makes the agent more robust. This robustness carries with it a performance cost.

This suggests that one functional use of emotions is to provide agitation to the agent in order to not let it settle into a stable state. Even though the environment may allow for it or make this the optimal behaviour. An explanation for this could be that natural agents have not evolved for such environments because they rarely exist and cannot be relied upon to last.

# 7 Acknowledgements

Karla Parussel acknowledges the financial support of the EPSRC and the University of Stirling. The authors wish to thank the anomynous referees for their useful comments.

# References

O. Avila-Garcia and L. Canamero. Using hormonal feedback to modulate action selection in a competitive scenario. In *From Animals to Animats 8: Proceedings of the eighth international conference on the simulation of adaptive behavior*, pages 243–252. MIT Press, 2004. ISBN 0262693410.

Jean-Marc Fellous. The neuromodulatory basis of emotion. *The neuroscientist*, 5(5):283–294, 1999.

Jean-Marc Fellous. From human emotions to robot emotions. In *Architectures for Modeling Emotions: Cross-Disciplinary Foundations. Papers from the 2004 AAAI Spring Symposium*, pages 37–47. AAAI Press, 2004.

H. X. Wang G. Q. Bi. Temporal asymmetry in spike timing-dependent synaptic plasticity. *Physiol Behav.*, 77(4-5):551–555, 2002.

P. Husbands. Evolving robot behaviours with diffusing gas networks. In P. Husbands and J.-A. Meyer, editors, *Evolutionary Robotics: First European Workshop, EvoRobot98*, pages 71–86. Springer-Verlag, April 1998.

Christof Koch. *Biophysics of Computation*. Oxford University Press., 1999. ISBN 0-19-510491-9.

Wei-Min Shen and Cheng-Ming Chuong. The digital hormone model for self-organization. In *From animals to animats 7: Proceedings of the seventh international conference on simulation of adaptive behavior*, pages 242–243. MIT Press, 2002. ISBN 0-262-58217-1.

F. Wörgötter and B. Porr. Temporal sequence learning, prediction and control - a review of different models and their relation to biological mechanisms, 2004.