Conditions for forced learning of graded responses

Helge Malmgren

Department of Philosophy, Göteborg University, Sweden

e-mail: helge.malmgren@phil.gu.se

Poster presentation at the *Sixth International Conference on Cognitive and Neural Systems*, Boston, May 30-June 1, 2002

1. Introduction

In this poster, I will address a set of related issues pertaining to learning systems at different levels of organization. The basic question is, *how can a learning system store graded responses,* i.e., *any chosen* response within a certain range? An obvious case is, *Under which conditions can a neuron uphold any chosen graded rate of firing through its interconnection with another neuron?* But the question should also be asked at the single-cell level, since it is possible that the ability to uphold any chosen rate of firing could be backed by a biochemical mechanism in the cell or its immediate environment. Finally, looking at *forced unsupervised learning* where exposure of a neural net to a stimulus contingency CS → UCS causes the net to react to CS as it originally reacted only to the UCS,

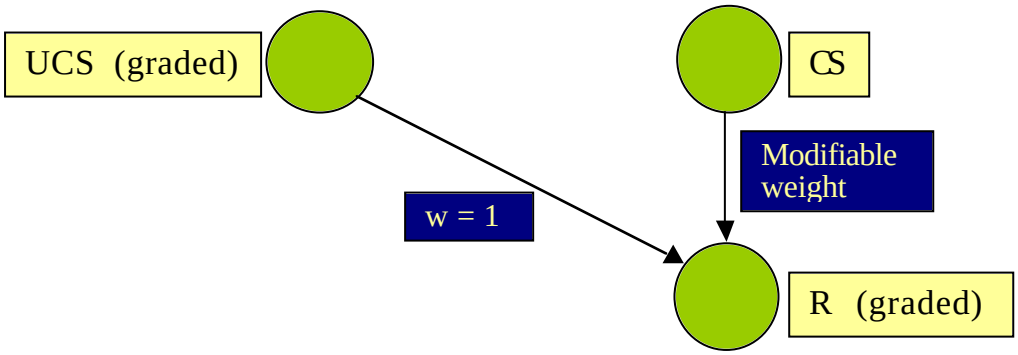

Figure 1. A neural net candidating for forced graded learning

one can ask what the conditions are under which a network can learn any chosen degree of a continuous UCS. What is required of a system which can learn (as we can) a contingency between a certain light and a tone of *any* chosen frequency? Is it possible to design a three-neuron system which, given suitable activation functions and appropriate weights, accomplishes this? If so, is there a biologically plausible weight update rule which assigns these weights under forced learning?

As will soon be clear, these questions require an analysis of systems with *continuous attractors* (i.e., connected sets of point attractors).[1] Such systems have recently been discussed in the context of the control of eye movements.[2,3] These discussions are however limited to networks of many interacting neurons upholding each other's activity, the models used are as a rule quasi-linear, and the resulting attractors are not very robust.[4,5] Another many-cell system which has received an analysis in terms of continuous attractors are the hippocampal place cells.[6,7] No plausible synaptic or single-cell system for upholding graded activity has yet been proposed. Discrete attractor mechanisms ("biochemical switches") have been proposed[8,9,10] but continuous attractors have not been implicated. And none of the modifications of Hebb's rule for synaptic plasticity which have been suggested can account for graded learning in a three-neuron forced learning system.

In this presentation, a large class of biologically not wholly implausible *non-linear* continuous attractor systems is described, which may have the potential to clarify several of these remaining problems.

2. Storage and retrieval of graded responses: definitions

In the case of a system with two variables x and y, learning of graded responses means the following: for a certain region of state space (below sometimes referred to as "everywhere"), clamping the system at *any* value x_0 of x causes y to go to a value y_0, such that clamping the system at y_0 causes x to go to x_0. (When we say that the system goes to an equilibrium point, we always mean that it goes to this point in the limit.) Clamping the system x-wise will be referred to as *presentation*, the process which it initiates in the system as *storage*, and its final y-wise result as a *representation*. Clamping the system y-wise, i.e., letting the representation control the system state, will be called *retrieval*. Retrieval may occur from "everywhere": the system goes to (x_0, y_0) from (x, y_0) for any x in the region. If the system is clamped y-wise when it is in an equilibrium point, the x value is simply *upheld*.

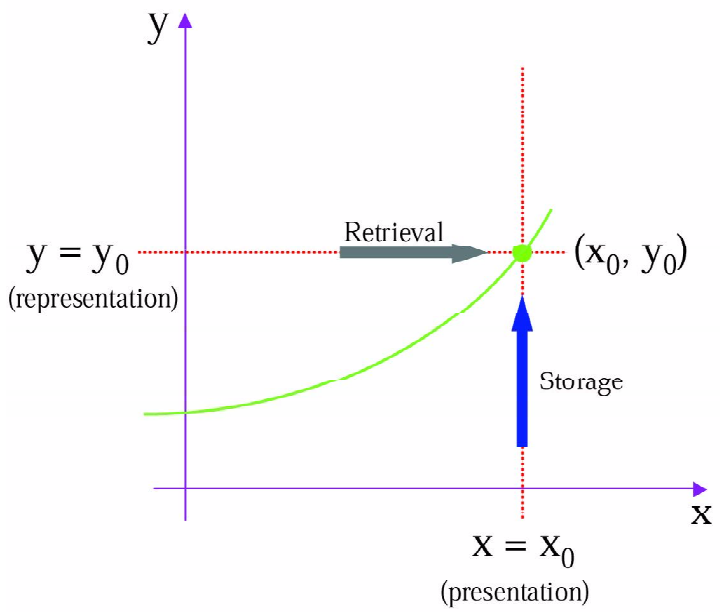

Figure 2. Storage and retrieval

Let us refer to the function which delivers an equilibrium value for y from a clamped value of x as the *storage function* **a**(x). The inverse of a is the *retrieval function* **b**(y) = $\mathbf{a}^{-1}$(y).

In the following we will presuppose that the intrinsic dynamics of the system is not changed by the clamping. Hence even when the system is free-running, all points on the graph **A** of **a** are equilibriums and attract points in their surrounding. (The surrounding points go towards **A** both x-wise and y-wise.) In other words, **a** defines a continuous attractor.

Now we have defined what me mean by a *learning continuous attractor system*. We have not yet addressed the question how to construct such systems. Below we will show how they can be built from biologically plausible components. But let us first illustrate the principle by means of two fictive examples.

3. The learning thermostat and the hydraulic memory system

Think of the system in Figure 2 as a *learning thermostat.* Such a device not only regulates the temperature x of a small room to a value which depends on the internal state y of the thermostat. It also regulates its own internal state according to the external temperature. Any constant external temperature x_0 (for example, outdoors) drives it towards an internal state y_0 such that the thermostat, when later kept in internal state y_0 in the small room, gives the room the temperature x_0.

It is not a wholly trivial task to construct a learning thermostat. First, it is of course essential that the functions involved in the two control loops are each other's inverses. How to achieve this will be a main theme in the following. Second, the internal state must be kept constant during regulation of room temperature. We might refer to this as *consolidation of the representation*. Third, a learning thermostat must be able to switch between keeping its internal state constant and letting the internal state be regulated by the external temperature. We will refer to this switch as the *memory gate.* However, if the time scale of learning is very long compared to the time it takes to regulate the room, and if there is no time delay between the storage and retrieval phases, the consolidation and memory gate mechanisms could both be dispensed with for practical purposes. In this case the drift of the internal state during regulation of room temperature will be negligible.

Another physical model, the design of which already contains the mechanisms required for learning, is offered by the following hydraulic memory system. The large container to the left in the figure is connected to a thin vertical tube which can be filled to any desired height from an external source. The connection tube can be opened and closed by a valve.

Figure 3. The hydraulic memory module

Starting with an empty container and an open valve, one should add water up to a certain level in the tube ("the desired level") and replace it continually until the water in the container has also reached this level. The water reserve will now uphold the desired level in the tube. One may also close the valve. Then at any later occasion, and starting with any level in the tube, opening the valve will again result in the water level being the same in the container and the tube – which means that the system has retrieved *almost* exactly the desired level of water in the tube. Here, of course, the large size of the container compared to the tube, which ensures that the level in the container does not change notably during retrieval, corresponds to widely differing time constants in the case of the learning thermostat.

4. Conditions for learning continuous attractors

By representing any level in the tube by means of the same level in the container, the hydraulic model exemplifies the simplest system which learns graded responses. It is given by the identity function **a**(x) = x. This function has itself as the inverse, and systems which use it can be said to *represent by means of similarity.* We will soon show that there are many other monotonic functions, including a set of non-linear functions having themselves as inverses, which can learn graded responses.

But let us first note that not all monotonic continuous attractors will do. Whether a line in state-space is a continuous attractor or not is decided by the system's behavior in the *free-running* state, while a learning system must also behave in a special way when clamped in either variable. If the system state S shall move to the graph **A** of a strictly increasing function **a**, if clamped in a point (x, y) to the left of **A**, dx/dt(x, y) must be positive all the way to **A**; else it will be trapped on the way:

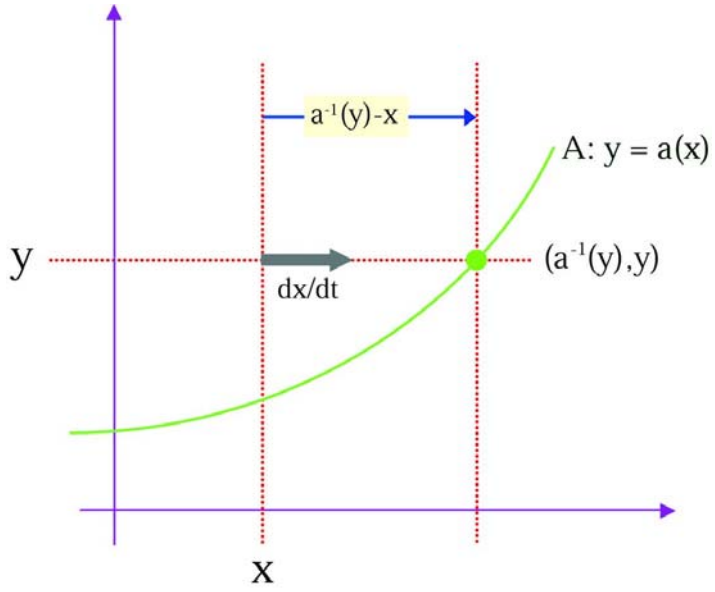

Figure 4. Conditions on dx/dt under y-wise clamping

Similarly, if (x, y) lies to the right of the graph, dx/dt(x, y) must be negative, and in the case of x-clamping the corresponding conditions must hold for dy/dt.

The mentioned conditions are not only necessary. *If* the signs of the two derivatives are the right ones *everywhere* (i.e., within the range considered), the system *will* behave as desired when clamped in either direction. Noting that the location of (x, y) in relation to the graph **A** is indicated by the signs of ($\mathbf{a}^{-1}$(y) - x) and (**a**(x) - y) respectively, we can therefore formulate the following general conditions:

(1) The strictly increasing graph **A**: y = **a**(x) always attracts S when y-clamped in M *iff* for all (x, y) in M, dx/dt(x, y) everywhere has the same sign as ($\mathbf{a}^{-1}$(y) - x)).

(2) The strictly increasing graph **A**: y = **a**(x) always attracts S when x-clamped in M *iff* for all (x, y) in M, dy/dt(x, y) everywhere has the same sign as (**a**(x) - y)).

5. Conditions for learning continuous attractors (continued)

Now consider *any* dynamic system S which fulfils the right side of (1) and (2) for *the same* strictly increasing function **a**. Obviously, the system goes to the *same* line (namely, **A**) when clamped in either direction, and stays there. Since clamping does not change the system's intrinsic dynamics, it will go to **A** from everywhere even when in the free-running state. Hence **A** is a continuous attractor. So,

(3) A system has a strictly increasing learning continuous attractor *iff* it fulfils the right side of (1) and (2) for the *same* strictly increasing function **a**.

The following is an equivalent but more intuitive way of formulating the relevant conditions on the time derivatives dx/dt and dy/dt. Suppose first that **a** is strictly increasing, as in our argument so far. Given a fixed y and with x growing from somewhere to the left of **A**, dx/dt should first be positive, then zero (at **A**) and then always negative. Let us express this by saying that dx/dt is *descending in x*. When x is fixed and y grows from somewhere below **A**, dx/dt instead goes from negative via zero to always positive. Hence dx/dt should be *ascending in y*. Note that an ascending function need not be monotonically ascending; likewise, an ascending function need not be monotonically ascending.

Now let the formula, dx/dt = f-+(x, y) express that dx/dt is descending in x and ascending in y. The meaning of similar locutions, such as dx/dt = f--(x, y), should be immediately clear without definitions. The above argument establishes that if an increasing function **a** shall be a learning continuous attractor for a system (x, y), it must be the case that dx/dt = f-+(x, y). (This corresponds to the right side of Equation 1.) Similarly, it must be the case that dy/dt = f+-(x, y) (cf. Equation 2). Finally, the two derivatives must switch sign *together*. But these three conditions also guarantee the existence of a strictly increasing function where both dx/dt and dy/dt switch signs. Hence we get the following equivalent of (3):

(4) The system S(x, y) has a strictly increasing learning continuous attractor *iff:* dx/dt = f-+(x, y) *and* dy/dt has the opposite sign of dx/dt everywhere.

So far we have not used the graphs of *decreasing* functions as attractors. However, it is easy to verify that in these cases (for examples see next page), the counterpart of (4) becomes:

(5) The system S(x, y) has a strictly decreasing learning continuous attractor *iff*: dx/dt = f--(x, y) *and* dy/dt has the same sign as dx/dt everywhere.

The concept of a time derivative dx/dt which is descending in x can be said to be a more precise version of the concept of *negative feedback.* So, conditions (4) and (5) formulate two different ways in which negative feedback can underlie a continuous learning attractor for a system. It is often said that negative feedback subserves homeostasis. We are here arguing that if the set of all possible homeostatic equilibriums in a region of state space is exploited, it may also subserve learning.

6. Self-inverses, decreasing attractors, and symmetric systems

A function is *self-inverse* if and only if $f(f(x)) \equiv x$. Geometrically, the graph of a self-inverse function is symmetric around the line $x = y$. The function $x = y$ is itself a self-inverse. There are innumerable *decreasing* self-inverse functions; examples include $y = - x$, $y = k/x$ (for any constant k) and $x^2 + y^2 = 1$ restricted to the first quadrant. Analytically, they can all be written as $f(x, y) = 0$, where f is a positive symmetric function (for example, $x + y = 0$). Now,

> (6) Any system with $dx/dt = dy/dt = - f(x, y)$, where f is positive and symmetric, has the decreasing, self-inverse function $f(x, y) = 0$ as a learning continuous attractor.

This follows immediately from Equation (5) and simple considerations of symmetry. For example, a system obeying the dynamic equations $dx/dt = dy/dt = 1 - xy$ will have the graph of the function $xy = 1$ as a learning continuous attractor:

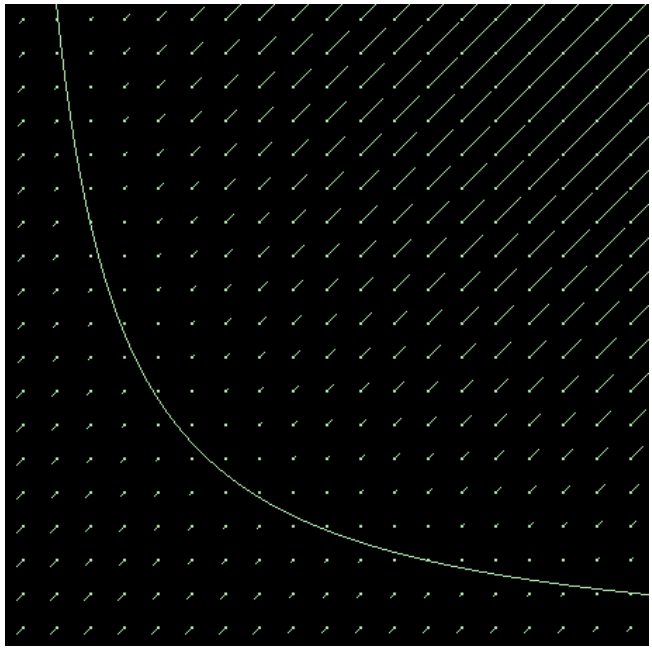

Figure 5. The flow field for dx/dt = dy/dt = 1-xy

Other simple self-inverse systems include $dx/dt = dy/dt = x + y$ (which has the line $y = - x$ as its attractor) and $dx/dt = dy/dt = x + y + 1$ (with $y = 1 - x$ as the attractor). Depending on the coding of photographic colors, one of these systems may be taken to formalize the properties of photographic negatives. Note that in traditional photography, the original signal is restored by taking the negative of the negative. This means that we are using a self-inverse representational system.

Of course, a main formal reason for using self-inverse functions when building continuous learning attractor systems is that one does not have to search independently for the inverse of the storing function. It is there already because of the symmetric negative feedback. But there is also a material reason for preferring these systems.

7. Biological self-inverse systems?

The material reason is that actual self-inverse systems can, in principle, be built by *using any two identical components in any symmetrical inhibitory arrangement obeying Equation (6)*. The essential point is that as long as Equation (6) is fulfilled, it does not matter *which* symmetric function one uses to control the two derivatives. And it seems that it is a much easier task for ontogeny and phylogeny to produce two identical units and to couple them symmetrically, than to try to fine-tune the properties of non-identical units in order to find the proper inverse of a storing function. No fine tuning is required once we know that the system is symmetric and, in the right way, inhibitory.

To be sure, "in the right way" may not be easily attainable. In theory, a symmetrical neural network system with two identical neurons obeying an inhibitory dynamics of the kind dx/dt = dy/dt = f--(x, y) = f--(y, x) would be able to store information in a self-invertible form, and hence to uphold and re-create graded activity. One way of realizing this would seem to be a system where each neuron sends activity through identical inhibitory connections to *themselves* as well as to the other neuron:

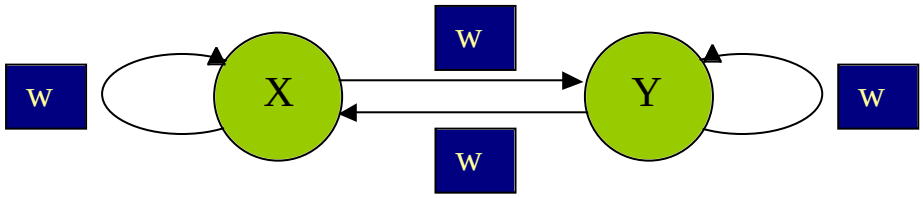

Figure 6. A possibly self-inverse neuronal net (w < 0)

However, we also have to suppose that the activation function is such that the time derivative of the activity is *only* (descendingly) *dependent on the input.*

I do not know of any real neural system which obeys such an "input-driven" dynamics, but since there are infinitely many solutions of this kind and since they all automatically provide the decoder together with the coder, it would be surprising if the CNS did not use any of them.

Here it is also of great interest that among all the self-inverse functions, only one is increasing: identity. Hence if you want to build a learning continuous attractor system by means of symmetrical couplings which provide both positive and negative feedback (i.e., a symmetric system obeying Equation 4 above), you will have to use representation by similarity. Theoretically this is of course not more difficult than building the symmetric decreasing systems: for example, let the neurons obey dx/dt = y - x, dy/dt = x - y, and they will have the continuous attractor x = y. But I think the consensus is that for a biological two-neuron system the attractor x = y is not very plausible, while I guess that it is too early to say the same of all the possible decreasing systems.

In other words, there may be other reasons than energy considerations behind the ubiquity of inhibitory connections in our central nervous systems.

8. Diffusion and chemical equilibrium attractors

Just as the self-inverse functions constitute an especially interesting class of decreasing continuous attractors, increasing attractors with dy/dt = -dx/dt may have a special biological relevance. Consider, abstractly, diffusion over a membrane. Let x and y stand for the concentrations of a chemical C on the two sides of the membrane. If no other reaction than diffusion occurs, dy/dt = -dx/dt. If dx/dt and dy/dt depend on x and y in the way required by Equation (4) – which we may suppose – then we are dealing with a system with an increasing learning continuous attractor. Note that the attractor need not be a straight line. This is because the exchange process may obey different dynamics at different total concentrations.

Here is the formal backbone of a very simple, single-neuron memory system which is based on a diffusion attractor. The external input to the cell sets the concentration of a chemical C in a compartment X to a value x which, in turn, determines the frequency of spiking. C can diffuse between X and another, much larger compartment Y, from which it cannot escape in any way except than to X. Finally, diffusion is only possible when membrane channel D is open:

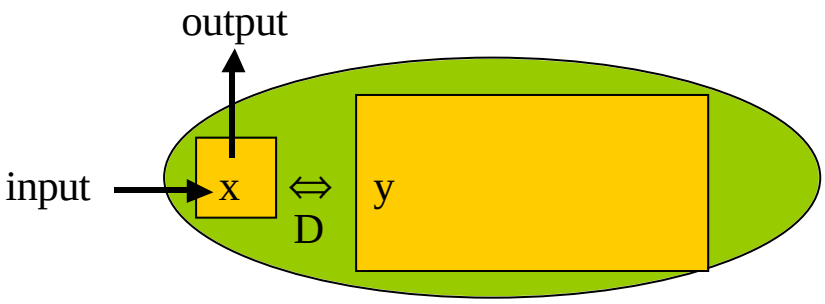

Figure 7. A diffusion-based single-neuron memory

Depending on the state of the membrane channel (the memory gate), this system will uphold, store or approximately recreate any firing frequency which has been forced upon it for a sufficiently long time.

Not only diffusion systems, but also a host of other biochemical equilibrium reaction systems, determine learning continuous attractors. If x and y denote the concentrations of chemicals A and AB in an isolated simple system catalyzed by C,

$$A + B \underset{C}{\leftrightarrows} AB$$

then dx/dt = -dy/dt = f-+(x, y), and Equation (4) is fulfilled. Again, the increasing attractor which is involved need not be linear since the reaction kinetics may change qualitatively at certain concentrations of A.

Also note that everything else being equal, clamping A at a higher level will cause the equilibrium concentration of B to go down. Hence the system defined by the concentrations of A and B, respectively, will have a *decreasing* learning attractor.

9. Forced associative learning, and simplicity in complexity

Our argument leads naturally to the following model of forced learning. The UCS produces a temporary concentration x of a chemical A, which controls the output. The CS activates enzyme D, which alone enables a reaction between A and A*. There is a large functional store of A* through its participation in further reactions.

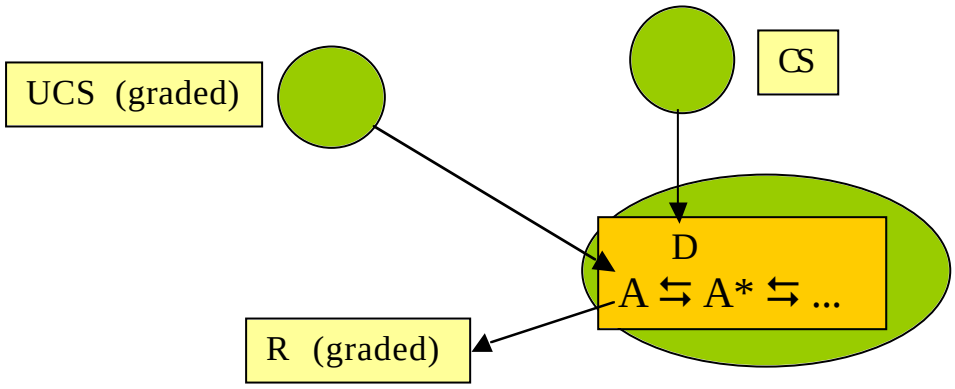

Figure 8. A continuous attractor model of forced, graded learning

The Pavlovian learning condition will lead to a build-up of A* (and its further products) until there is an equilibrium with the input concentration x of A. After that, CS will by itself make the "chemical store" release just enough A to achieve the concentration x, and the system's output will be very like the desired signal.

Of course, our simplistic model is a caricature of cell biochemistry. But note that the conditions for learning attractors are so permissive that many complex systems fulfill them at least in circumscribed regions of their state-space. Suppose, for example, that A is involved in two equilibrium reactions with different kinetics:

$$A + B \leftrightarrows AB;\ A + D \leftrightarrows AD$$

Here it is no longer the case, with x and y again being the concentrations of A and AB, that dx/dt = -dy/dt. It is not even true, considering a large state-space, that dx/dt and dy/dt always have opposite signs, since with a well chosen starting point – a lot of B and a lot of AD – the reactions can run in opposite dominant directions for a while. However, if we consider only the part of the state-space which results from starting in the overall equilibriums of the system and manipulating the concentrations of A or AB, dx/dt will have the opposite sign of dy/dt in this region. If, for example, we add some A to the system in equilibrium, both reactions will run with a dominant direction towards the right, so $dx/dt < 0$ and $dy/dt > 0$.

I dare a guess: even some very complex biochemical reaction systems, the detailed dynamics of which are intractable with analytical and simulation methods, may well be amenable to simple descriptions in terms of learning continuous attractors.[11] And it may be a fruitful venture to search, among these attractors, for the ones which explain forced associative learning of graded responses.

Acknowledgements

I am much indebted to Bo Berntsson, Lars-Johan Erkell, Björn Haglund, Daniel Ruhe and Holger Wigström for kind assistance and comments.

Selected references

1. Amari S, Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics 27* (1977), 77-87.

2. Seung HS, How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences USA 93* (1996), 13339-44.

3. Seung HS, Lee DD, Reis BY, Tank DW, Stability of the memory of eye position in a recurrent network of conductance-based model neurons. *Neuron 26* (2000), 259-71.

4. Durstewitz D, Seamans JK, Sejnowski TJ, Neurocomputational models of working memory. *Nature Neuroscience, Suppl. 3* (2000), 1184-91.

5. Wang XJ, Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neuroscience 24* (2001), 455-63.

6. Samsonovich A, McNaughton BL, Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience 17* (1997), 5900-20.

7. Stringer SM, Trappenberg TP, Rolls ET, de Aranjo IET, Self-organising continuous attractor networks and path integration: one-dimensional models of head direction cells. *Network: Computation in Neural Systems 13* (2002), 217-42.

8. Bhalla US, Iyengar R, Emergent properties of networks of biological signalling pathways. *Science 283* (1999), 381-7.

9. Katz PS, Clemens S, Biochemical networks in nervous systems: expanding neuronal information capacity beyond voltage signals. *Trends in Neuroscience 24* (2001), 18-25.

10. Lisman JE, Zhabotinsky AM, A model of synaptic memory: a CaMKII/PP1 switch that potentiates transmission by organizing an AMPA receptor anchoring assemply. *Neuron 31* (2001), 191-201.

11. Weng G, Bhalla US, Iyengar R, Complexity in biological signaling systems. *Science 284* (1999), 92-6.