Learning Continuous Attractors

Helge Malmgren

Dept. of Philosophy, Göteborg University

Abstract. Neural systems with continuous attractors have been invoked in explanations of spatial memory and the control of eye movements. Line attractors have also been implicated in cellular and synaptic memory.

In the present paper, formal conditions for the existence of continuous attractors with learning properties in two-dimensional dynamic systems are derived. Some actual and hypothetical biological examples of such learning continuous attractors are discussed, with special emphasis on subcellular mechanisms. It is argued that such systems may underly not only short-term and synaptic memory, but also the micro-genesis of perceptions. The theory of learning continuous attractors is placed in the context of a cybernetic theory of learning, in which learning is a necessary consequence of feedback in dynamic systems that can simulate input. Key words: continuous attractors, dynamic systems, memory, perception.

1. Introduction

Human beings do not only learn discrete pieces of information such as whom their father is, but also quasi-continuous values of variables such as the angle of vision to a nearby object and the present location of the own body in relation to the surroundings. Such learning would be easier to understand if one could assume that the state spaces of some systems of neurons contain connected sets of point attractors, or continuous attractors as these sets are also called [1]. Continuous attractor systems have indeed been postulated to explain the properties of the brain’s control of eye movements [2], [3] and our short-term memory for places [4]. In these theories, the existence of continuous attractors depends on tuned connections between many neurons. The possibility of using continuous attractor models for explaining intrinsic cellular memory [5] has recently received some attention; see for example [6], [7]. Synaptic memory has also been discussed in these terms [8].

In the present paper, the concept of a learning continuous attractor is given a strict definition, and the basic conditions for the existence of such attractors in two-dimensional autonomous systems are described. A number of actual and hypothetical biological realisations with possible relevance for perception and learning are then briefly discussed, with special emphasis on sub-cellular mechanisms and small neural systems.

The theory of learning continuous attractors is finally placed in the context of the author’s cybernetic theory of learning [9]. In this theory, which is inspired by W.R. Ashby’s ideas [10], learning is a necessary consequence of feedback in systems that can simulate input, i.e., systems for which the state space defined by the system’s output overlaps with the space of its inputs. The learning continuous attractor should be regarded as the final refinement, through evolutionary causes, of the basic learning principle which is valid for all such systems. An application of the integrated theory to the microgenesis of perception is also outlined.

2. Learning Continuous Attractors

2.1 The Learning Thermostat

Consider the learning thermostat. This device works as a textbook thermostat in that its internal state \( y \) determines, if held fixed, the final value of the room temperature \( x \). Hence there is a function \( B \) such that

\[
B: x_\infty = B(y)
\]

where, of course, \( x_\infty \) is the equilibrium to which \( x \) goes in the limit. With an ordinary thermostat, you set the desired value of \( x_\infty \) by directly manipulating \( y \) (presumably while reading off the corresponding \( x_\infty \) on some kind of display). Not so with the learning thermostat. If the desired value of \( x_\infty \) is \( x_D \), you take the thermostat to another room where the temperature is held fixed at \( x_D \). The environment influences the thermostat’s internal
state; indeed, it causes its internal state $y$ to be set, in the limit, at a certain value $y_\infty$, which is uniquely determined by $x_0$. When we return the thermostat to the first room and keep $y$ fixed at that value, we find that the temperature of the first room becomes identical in the limit with that of the second. I.e.,

$$x_\infty = x_0$$  \hspace{1cm} (2)

### 2.2 Learning Continuous Attractors Defined

We may summarize the behaviour of the learning thermostat in the following way: By being exposed for sufficiently long for a certain external stimulus, it acquires the ability to reproduce the same stimulus in other environments, given sufficient time. The formulation should make it clear why we use the term “learning”.

Now, the function according to which the environment, if held fixed, determines the learning thermostat’s final internal state, i.e.,

$$A: y_\infty = A(x)$$  \hspace{1cm} (3)

is obviously the inverse of $B$. Under the supposition that $B$ is a continuous function, this entails that $A$ (like $B$) is either monotonously increasing or monotonously decreasing. Figure 1 illustrates the increasing case. The curved line is the graph of the function $A$. The labels “presentation”, “representation”, “storage” and “retrieval” have been chosen in order to emphasize the possible application to learning.

![Figure 1](image)

**Fig. 1.** Storage and retrieval in a learning thermostat. The environmental ($x$) state is now labelled “presentation”, while the internal ($y$) state is labelled “representation”. The thick arrows labelled “storage” and “retrieval (re-presentation)” indicate trajectories of the system when the $x$ and $y$ states are kept constant at $x_0$ and $y_0$ ($= A(x_0)$), respectively.

We now suppose that the intrinsic dynamics of the system consisting of the environmental state $x$ and the internal state $y$ of the thermostat is independent of whether a variable is kept constant or not. In the “free-running” condition, the system must then go both $x$-wise and $y$-wise towards the graph of $A$. Indeed, it must go to a point on this graph between the two points determined by keeping $x$ and $y$ fixed, respectively (cf. Figure 1). Since every point on the graph of $A$ is an equilibrium, it follows that this graph is a continuous attractor for the system. This is why we are allowed to speak of learning continuous attractors. The characteristic dynamics of such a system is illustrated in Figure 2.
Fig. 2. Phase portrait of a two-dimensional system with a strictly increasing continuous attractor function A. Vertical and horizontal dotted arrows: trajectories when x and y are kept fixed, respectively. Curved dotted arrows: trajectories in the free-running condition. Open circles: equilibrium points. Solid line: Connected set of point attractors (continuous attractor).

2.3 Necessary Conditions

Suppose that the system portrayed in Figure 2 obeys the autonomous differential equations (4)–(5):

\[
\frac{dx}{dt} = f(x, y) \quad (4)
\]
\[
\frac{dy}{dt} = g(x, y) \quad (5)
\]

where f and g are continuous everywhere in the interval under consideration. From Figure 2, it is immediately clear that if the system shall go to \((x_0, y_0)\) in the limit, if started in a point to the left of it (here: \((x_1, y_0)\)) and y is kept constant, then \(\frac{dx}{dt}\) must be positive all the way on the trajectory to A. Else the system will be trapped on the way. If the system is started to the right of A, the derivative must be negative. Hence \(\frac{dx}{dt}\) must, for every value of x, have the same sign as \(x_0 - x\). An analogous argument gives at hand that \(\frac{dy}{dt}\) must have the same sign as \(y_0 - y\) for all values of y, i.e., goes from positive to negative at A. These conditions on the derivatives hold for every point \((x, y)\) in the state space and for the free-running as well as for the clamped case.

Figure 2 can also tell us the following about the dependence of \(\frac{dx}{dt}\) on y around \((x_0, y_0)\), and the converse dependence of \(\frac{dy}{dt}\) on x. When we keep x constant, increase y and pass through \((x_0, y_0)\), \(\frac{dx}{dt}\) switches sign from negative to positive. Likewise, when we keep y constant, increase x and pass through \((x_0, y_0)\), \(\frac{dy}{dt}\) switches sign from negative to positive.

We will hereafter designate the sign-switching behaviour of the derivatives which has just been described by adding a parenthesis with the proper pair of plus or minus signs after the function symbol. Let \(f(+ -)(x, y)\) stand for any function which switches signs in the positive direction (from negative to positive) when x grows, but in the negative direction (from positive to negative) when y grows. The symbol \(f(+ -)\) also entails that for each dimension of variation, there is one and only one switch of sign. Similarly for the symbols \(f(+ +)(x, y)\), \(f(- -)(x, y)\) and \(f(- +)(x, y)\), the meaning of which should be evident. In these terms, the found necessary conditions for a system with an increasing continuous attractor boil down to (6)–(7):

\[
\frac{dx}{dt} = f(- +)(x, y) \quad (6)
\]
\[ \frac{dy}{dt} = g(-)(x,y) \]  

Slightly different necessary conditions hold for the case of a decreasing attractor function \( A \). Figure 3 is a simplified phase portrait, containing only trajectories around a single equilibrium when one variable is kept constant, of a system with a decreasing continuous attractor.

From Figure 3, it is immediately clear that if the system shall move to the point \((x_0, y_0)\) in the limit when \( y \) is kept fixed, \(dx/dt\) must (as for the increasing attractor) have the same sign as \( x_0 - x \), i.e. it switches from positive to negative at \( A \). And if the system shall approach \((x_0, y_0)\) when \( x \) is kept constant and \( y \) grows, \(dy/dt\) must again have the same sign as \( y_0 - y \), i.e., go from positive to negative. However, when \( x \) is kept fixed and \( y \) grows, \(dx/dt\) switches from positive to negative at \( A \). This differs from the increasing case. Similarly, when \( y \) is kept fixed and \( x \) grows, \(dy/dt\) goes from positive to negative. In terms of the recently introduced symbolism, the found necessary conditions for a system with a decreasing continuous attractor are (8)–(9):

\[ \frac{dx}{dt} = f(-)(x,y) \]  
\[ \frac{dy}{dt} = g(-)(x,y) \]  

2.4 Necessary And Sufficient Conditions

So, that conditions (6)–(7), or (8)–(9), are necessary for the existence of continuous attractors follows immediately from what it is to be an increasing and a decreasing strictly monotonic function: the right side of its graph is down, and up, respectively. It may be just a little more surprising that (6)–(7), and (8)–(9), become sufficient conditions if it is added that the \( x \) and \( y \) time derivatives always switch signs together. But consider a
system fulfilling (6), choose any point where \( \frac{dx}{dt} \) is positive, and keep \( y \) constant. The system will now traverse the state space towards the right. Condition (6) entails that \( \frac{dx}{dt} \) will diminish until you have reached a point where it is zero, that it is not zero at any other point with either the same \( x \)-coordinate or the same \( y \)-coordinate, and that the \( x \)-wise and \( y \)-wise sign switches of \( \frac{dx}{dt} \) in this point are of the right kind for a point attractor. Assuming common sign switch, \( \frac{dy}{dt} \) will also be zero at this point. Furthermore, (7) entails that \( \frac{dy}{dt} \) is not zero at any other point with either the same \( x \)-coordinate or the same \( y \)-coordinate, and that the sign switches of \( \frac{dy}{dt} \) at the present point are of the right kind for a point attractor. Hence we have found a point on a strictly monotonous, continuous attractor. The same kind of argument is obviously valid for systems fulfilling (8)–(9).

Furthermore, the condition of common sign switch signs makes one of the conditions in (6)–(7), and one in (8)–(9), redundant. Hence

**Proposition 1.** An autonomous two-dimensional system \((x, y)\) has an increasing learning continuous attractor iff

\[
\frac{dx}{dt} = f(-+)(x, y) \tag{10}
\]

and

\[
\text{sgn}(\frac{dx}{dt}) = -\text{sgn}(\frac{dy}{dt}) \tag{11}
\]

It has a decreasing learning continuous attractor iff

\[
\frac{dx}{dt} = f(--)(x, y) \tag{12}
\]

and

\[
\text{sgn}(\frac{dx}{dt}) = \text{sgn}(\frac{dy}{dt}) \tag{13}
\]

Proposition 1 defines a very large class of systems through purely qualitative (topological) conditions. In other words, the combined task of mapping a variable \( x \) on another variable \( y \) in the limit, \( A: y \to \infty = A(x) \), and finding the inverse of this mapping, \( B: x \to \infty = B(y) = A^{-1}(y) \) can be performed through innumerably many different choices of dynamic laws. It may not be that difficult to build systems with learning continuous attractors, after all.

### 2.5 Note on the Concepts of ‘Error Function’, ‘Error Correction’ and ‘Neural Comparator’.

Control mechanisms, not least when they are postulated to exist in the context of neural systems, are not seldom conceptualised in terms of “error correcting” mechanisms. Feedback information in the form of the value of an “error function” is supposed to guide the controller’s actions. The error function, in turn, is defined in terms of some kind of comparison between the actual \( x \) and the desired \( x_D \) value of the controlled variable.

Consider a learning thermostat using the simple version of error correction where \( \frac{dy}{dt} \) is proportional to \( x_D - x \). Obviously, \( \frac{dy}{dt} = f(-+)(x, x_D) \). Also suppose that the relation between \( x \) and \( y \) is a strictly increasing one. Then \( x_D \) is the same as \( B(y_D) \) for a certain \( y_D \). Hence \( \frac{dy}{dt} = f(-+)(x, y) \). The same holds by implication for \( \frac{dx}{dt} \), and the signs of both derivatives switch in the proper way at \((x_D, y_D)\). Therefore (10)–(11) are fulfilled.

However, not only does (10)–(11) allow for non-linear error functions and non-linear learning algorithms; they allow for \( \frac{dy}{dt} \) being non-monotonically dependent on both \( x \) and \( x_D \). In other words: not all functions \( f(x, y) \) that fulfill (11) are decomposable as \( h(c(x, y)) \), where \( h \) (the correction mechanism) is an increasing function and \( c \) (the comparison underlying the error function) is decreasing in \( x \) and increasing in \( y \). And such decomposability seems to be a fair requirement on an “error-correcting” mechanism.

The biological significance of what has just been said is that neural and other control circuits might not always be analysable in terms of a concrete mechanism which transmits “error information”. In the author’s opinion, the opposite should always be the first guess. There is no a priori reason why the control mechanism should reduce the information contained in the two state variables \( x \) and \( y \) to a single value before it is put to use. And if this is so, the neural comparator should be regarded as a too simplistic solution for many neural problems (see also Section 4.2).
3. Special Attractor Functions and Ways to Realize Them

We will now look at some interesting subclasses of learning continuous attractors which may have biological relevance, starting with the simplest ones.

3.1 Creating the dynamics from the equation for the attractor

The direct way of creating a strictly increasing, learning continuous attractor \( y = A(x) \) (with \( B = A^{-1} \)) is to rewrite its equation as

\[
y - A(x) = 0
\]

and then to assign the following dynamic equations to a system:

\[
\begin{align*}
\frac{dx}{dt} &= y - A(x); \\
\frac{dy}{dt} &= A(x) - y &= y - B(x)
\end{align*}
\]

(11)

Since it is supposed that \( A \) is strictly increasing, it follows immediately that the system satisfies (6) - (7).

Starting instead with a strictly decreasing function \( A \), we may use the following dynamics:

\[
\begin{align*}
\frac{dx}{dt} &= \frac{dy}{dt} = A(x) - y &= y - B(x)
\end{align*}
\]

(12)

The time derivatives of a system created in either of these ways are strictly monotonous in both \( x \) and \( y \), so the systems are analysable in terms of (possibly non-linear) error correction mechanisms.

3.2 Linearly increasing attractors

If \( A \) in Eq. (14) is a linear function passing through the origin, i.e.

\[
y - cx = 0
\]

(13)

for some constant \( c \), (14) takes the form

\[
\begin{align*}
\frac{dx}{dt} &= y - cx; \\
\frac{dy}{dt} &= cx - y &= y - \frac{1}{c} x
\end{align*}
\]

(14)

This is equivalent to linear error correction (the delta rule) in both directions. From a mathematical viewpoint, (18) is indeed the easiest way of building a learning thermostat.

3.3 Self-inverse systems

Let us now consider the simplest linear case, i.e.

\[
\begin{align*}
\frac{dx}{dt} &= y - x; \\
\frac{dy}{dt} &= x - y
\end{align*}
\]

(15)

This dynamics gives rise to the attractor \( y = x \). The identity function is self-inverse, i.e. \( A(A(x)) \equiv x \). Indeed, since any self-inverse function is symmetric around the line \( y = x \), identity is the only increasing self-inverse function (for decreasing self-inverses, cf. below).

Note that (19) entails the same dynamics for \( dx/dt \) and \( dy/dt \) when these derivatives are considered in relation to the variables which they are derivatives of. Cf. (20) below for a formal characterization of this sense of “the same dynamics”. Self-inverse attractor functions therefore tend to emerge from systems where two identical elements are connected with each other in a symmetric way. This might be a good way for nature (e.g., the nervous system) to economize with resources.

Innumerable many linear and non-linear dynamic equations will lead up to the identity attractor. Take (19), and imagine clamping a strip of the phase diagram surrounding the line \( x = y \) while stretching another part in whatever direction you want (taking care not to violate the continuity of the derivatives). And, limiting ourselves to symmetrical systems of the kind.
reversible chemical reactions are often embedded in buffer systems, which tend to keep concentra-
approximat-
compartments. I.e., one may “store” concentrations of A as corresponding equilibrium concentrations of C, and
kinds of dynamics, the attractor may well be non-
linear. If, now, the system has its equilib-
umum far "to the right", the equilibrium lies. With other
order kinematics, it will have a
creasing linear continuous attractor passing through zero and with a slope indicating how far “to the right” or “to the left” the equilibrium lies. With other
kinds of dynamics, the attractor may well be non-linear. If, now, the system has its equilibrium far “to the right”,
it can be used in a way which is quite analogous to the way we used the diffusion system with differently sized
compartments. I.e., one may “store” concentrations of A as corresponding equilibrium concentrations of C, and
approximatively “retrieve” them through the free-running system.
In nature, diffusion and reversible reactions of the form (22) are of course often combined. Furthermore,
reversible chemical reactions are often embedded in buffer systems, which tend to keep concentrations constant.

\[
\frac{dx}{dt} = f(- +)(x, y); \quad \frac{dy}{dt} = f(- +)(y, x)
\]

(20)

any such system defines \( x = y \) as a learning continuous attractor if the function \( f \) is of the form \( h(x) + h(y) = 0 \),
for some strictly increasing \( h \). An illustrative choice may be

\[
\frac{dx}{dt} = x^2 - y^2; \quad \frac{dy}{dt} = y^2 - x^2
\]

(21)

However, in spite of all these possibilities a symmetrical system consisting of two interacting excitatory
neurons will not do. Real neurons, at least as conceived in standard textbooks, simply do not have the required
asymptotic behaviour. It is generally assumed that, if \( x \) is the net input to a neuron and \( y \) is the resulting output
frequency, the activation function \( y = A(x) \) is not linear. A fortiori, \( A \) is not the identity function. Therefore the
identity function cannot be a continuous attractor in the activation space of two symmetrically coupled,
excitatory neurons.

3.4 Diffusive Two-Compartment Systems

A biological system which does fulfil condition (20) is simple diffusion over a membrane separating two
compartments of equal size. Supposing that the system is closed to other influences, the net flow \( dx/dt \) into one of
the compartments is equal (with opposite sign) to the net flow \( dy/dt \) into the other compartment, and both are
proportional to the difference (with opposite signs) between the concentrations \( x \) and \( y \) on the two sides of the
membrane. Hence the concentrations on the two sides will, at equilibrium, be equal. The point is that this holds
from whatever point in state space you start and whether or not you hold the concentration on one of sides
constant. So, by keeping the concentration \( x \) in the one compartment (the input compartment) constant at any
value \( v \), you can store a “representation” of it in the other (the storage compartment) in the form of an identical
value \( v \) of the other concentration, \( y \). In principle, you should also be able to “retrieve” the original concentration
in the input compartment by keeping \( y \) constant at \( v \) and let the diffusion go the other way.

Now, simple diffusion may seem not to be very well suited as a biological learning mechanism since it is not
quite easy to imagine how the “representation” should be kept constant while the original concentration is
“retrieved”, except by postulating another learning system which ensures the stability of the representation (by
somehow continuously refilling the storage compartment). However, let us assume that the storage compartment
is very much bigger than the in-

If we change the relevant state variables in this models from concentrations to absolute amounts of the
involved substance on either side, the resulting attractor will not be identity but another line passing through
zero. There is not place here to go into the details of this obvious generalisation of the simple diffusion model.
Neither will I discuss the equally obvious extensions in terms of non-linear time derivatives and non-linear
increasing attractors.

3.5 Diffusion, Learning Buffers and Cellular Memory

The great merits of all diffusive systems is that the nature of the process as an exchange of substances guarantees
that \( dx/dt = -dy/dt \), which in turn ensures common points of sign switch for the derivatives and thereby an
increasing continuous attractor. Now, diffusion is only one example of reversible physico-chemical processes
involving (something like) an exchange of substances. All stoichiometric equations of the form

\[
aA + bB + ... \rightleftharpoons cC + dD + ...
\]

(22)

where \( A, B \) etc. are now names of chemical compounds, define increasing continuous attractors in two or
more dimensions (cf. [8]). In the case where \( a = c = 1 \), all other constants are zero and the remaining
two-dimensional system obeys first-order kinematics, it will have an increasing linear continuous attractor passing
through zero and with a slope indicating how far “to the right” or “to the left” the equilibrium lies. With other
kinds of dynamics, the attractor may well be non-linear. If, now, the system has its equilibrium far “to the right”,
it can be used in a way which is quite analogous to the way we used the diffusion system with differently sized
compartments. I.e., one may “store” concentrations of \( A \) as corresponding equilibrium concentrations of \( C \), and
approximatively “retrieve” them through the free-running system.
In nature, diffusion and reversible reactions of the form (22) are of course often combined. Furthermore,
reversible chemical reactions are often embedded in buffer systems, which tend to keep concentrations constant.
Hence information about the extracellular concentration of a substance could be stored as the concentration of another substance inside the cell, and retrieval from a “chemical memory” could be supported by activation of a buffer system pertinent to the “storage” side of the relevant reaction.

One possible application (out of many) of this conceptual model to cellular memory has been outlined in [6], while [7] contains a critical discussion of the usefulness of the idea of continuous attractors for the understanding of cellular memory.

Another near-lying application is to network memory of (graded) stimuli as (graded) synaptic weight change; cf. [8]. Neural system are supposed to store associative connections between stimuli $S_1$ and $S_2$ as sets of weights between neurons. Here $S_2$ is allowed to vary along a continuum. The biological function of the storage is certainly to reproduce (at least something like) $S_2$, when $S_1$ occurs. Suppose that the weights and the neural activations are just two ends of a chain of reversible physico-chemical exchange processes with the overall equilibrium curve displaced very much towards the substance the concentration of which codes the weights. Explanations in terms of learning continuous attractors are immediately forthcoming for both the slow learning and the fast retrieval phase of the process. The only glitch is that the weights are not kept absolutely constant during retrieval. But how could they be kept constant, on any biologically plausible theory of graded learning? It is all too easy to just move the problem to a supposed controlling instance.

### 3.6 Decreasing Attractors Open More Possibilities

We will now go on to the decreasing learning continuous attractors, and first discuss the symmetric systems defined by self-inverse functions. Remember that the graph of a self-inverse is symmetrical around $x = y$. There are innumerable many such functions of the decreasing kind, e.g. $y = 1 - x, y = 1/x$ and $y^2 + x^2 = 1$. The latter two are non-linear functions. We will not dwell on linear decreasing attractors here, whether self-inverse or not, but it may be worthwhile to note the possibility of finding such systems, too, in the context of biochemical equilibria.

So, a symmetrical system does not have to have elements which are asymptotically linear in order to behave according to a decreasing continuous attractor. Why, except for this mathematical fact, should decreasing self-inverse continuous attractors be of any interest for brain/cognitive theory? First of all, the decreasing property reminds us of the possibility that representations may be rather like photographic negatives of the original. Maybe our tendency to think of mental representations as plain copies is due to a tendency to confuse representing, which certainly must reproduce the original, with representing, which need not. Second, the self-inverse property tells us that negative representations, too, can be retrieved by using the same operation on them as that which created them – as one does with photographic negatives and (some) moulds. So, again, we can build these systems using pairs of identical units. And to repeat, the extra benefit of the decreasing systems is that these units need not be linear.

The direct way (cf Section 3.2) of deriving fitting dynamic equations from the decreasing attractor $y = 1/x$ starts with putting the function on the form $xy - 1 = 0$. This leads to the following equations

$$\frac{dx}{dt} = \frac{dy}{dt} = 1 - xy$$

(23)

It is not wildly implausible to suppose that there are real neurons, the activation of which follow the non-linear rule $dy/dt = 1 - xy$, where $x$ is the (inhibitory) input. After all, that equation entails that the effect of an inhibitory input is higher the greater the input, and higher the higher the initial activation, and that the effects of input and state are multiplicative. None of these properties seems unfamiliar from neural theory. However, the hypothesis is implausible since no known neuron shows an asymptotic input dependence of exactly the required form. The interaction between the input and the state is not quite multiplicative.

Indeed, I think that the prospects are not altogether bright for finding a symmetric two-neuron system, the activation space of which contains any self-inverse decreasing attractor. This is because there is no plausible mechanism which could materially guarantee that the attractor function is symmetric, in the way the exchange of substances guarantees symmetry in the case of diffusive systems. Instead this property would have to be upheld by a delicate coordination of the different mechanisms by means of which the activation and input levels, respectively, affect the time derivative of the activation. I do not say that this is impossible, and it could well be that some such coordinative mechanism has been invented by evolution. There are also, of course, other kinds of neural dynamics than the standard textbook one [11], and it could be worthwhile to look for neurons with a decreasing, self-inverse activation function.
4 Some Stability Considerations

In this last section, I will first shortly comment on the issue of the stability of continuous attractor equilibria. Then, the ideas presented above will be put in the context of a cybernetic theory of learning, where stability considerations play a central role.

4.1 The Stability of Continuous Attractors

A point on a continuous attractor is not asymptotically stable since it is not true that the system will return to it, given any small displacement. How, then, could continuous attractors possibly form the basis of memory? First, note that the displaced system will always return to a nearby point, and stay there. “Nearby” is here synonymous with “between the two asymptotic points determined by keeping x and y fixed, respectively” (cf Section 2.2). So the points on a continuous attractor are nearly asymptotically stable. Second and even more important, “not asymptotically stable” is not equivalent to “easy to disturb”. The position of a heavy piece of furniture standing on a floor, with substantial friction between them, defines a state space where no point is asymptotically stable. But we usually trust that the sofa stays where we put it. The main reason why we can trust our continuous memories is the simple fact that it is hard to displace their states, except by using learning. In this sense they could have almost any degree of stability. The limits for their stability are instead set by the nature of thermodynamic equilibria (cf. Section 3.5).

4.2 On-line learning As a Necessary Consequence of Feedback

The argument in this section is inspired by a proposal by W.R. Ashby [10].

Consider a finite deterministic system S with partially shared input and output space. I.e., the system can produce some outputs which lie in the same state space as its inputs. From now on, we only consider this shared space. For simplicity we think of it as one-dimensional, and label its state x. The system also has memory states y, and there is a switch which determines whether the system is input-driven or memory-driven. In the input-driven phase, no output is produced and the memory state is updated according to

\[ y_{t+1} = a(y_t, x_t) \]  

In the memory-dependent phase, the system’s output is some function of the current memory state. The system also reads this output as an input at the next moment. This it can do because of the shared state space.

It is now easy to show that the system will tend to learn to reproduce inputs. Suppose that the system is given input x0 for some time, and that the memory state goes to a point attractor y0. Now the system switches to memory-driven mode. If it then produces output x0, the memory state will receive the same input as before, and so the system must remains stable in A. If it produces an output which differs from x, it may go to another state y1, etcetera. With repeated presentations of x0, the probability therefore rises that the system is in a memory state which can produce x0. More specifically, this will occur in the vast majority of all possible finite automata. It is not difficult to extend the argument to associative learning [12].

The similarities between the discrete automaton just presented and the continuous attractor model should be obvious. Both assume a shared output-input space x, and in both the output tends to be equal to the earlier input. Neither model relies on explicit comparisons and traditional error correction; x and y may control the movements of the system in many more manners. The crucial difference between the models is, of course, that the continuous version offers guaranteed, gradual learning while the finite automaton uses a quasi-random trial-and-error process.

My hypothesis is that the central nervous system uses a mix of these two mechanisms for both perception and learning. Let me talk about perception for a while, without loss of generality. The micro-genesis of perceptions, or percept genesis for short, consists in the progressive stabilisation of the percept over a time scale of less than a half second. Usually, this process is smooth, but sometimes – especially when we had what the psychologists call the wrong “perceptual set” – it is not. We do not see anything until we change the set. I would describe what happens as follows. When percept genesis is smooth, it is because the perceptual system started in a region of its state space where there is a learning continuous attractor which fits the input. When it is not smooth, quasi-random jumps in state space occur until such a region is found. This is in my view the proper way of looking at the role of top-down mechanisms in perception. They come into play not only when the bottom-up processes give an unexpected result, but – more importantly – when they do not give any result.
References


