A Remark on Alan H. Kawamoto: Nonlinear Dynamics in the Resolution of Lexical Ambiguity: A Parallel Distributed Processing Account

REGINALD FERBER

Fachbereich 2, Universität Paderborn, D-33095 Paderborn, Germany

ferber@psycho2.uni-paderborn.de

Kawamoto (1993) reports a simulation of word recognition processes with a recurrent network of 216 units. 48 patterns of activity were used to train the network and yield data which are compared with data from word recognition experiments. Beside the presentation of his interesting data, Kawamoto describes his network as a dynamical system, sharing properties with a Hopfield net such as nonincreasing energy and convergence toward stable states.

This note gives counter examples to some of the statements made in this description: Simple nets with increasing energy and not converging to stable states. It shows why the notion of "local energy minima" is not of much use for networks updated in parallel and points out that there are important differences between the two types of networks. A simulation shows the strong influence of the representation of words Kawamoto uses.

In his article "Nonlinear Dynamics in the Resolution of Lexical Ambiguity: A Parallel Distributed Processing Account" Alan H. Kawamoto (1993) describes a simulation of word recognition processes with a recurrent network of 216 units. 48 patterns of activity were used to train the network and yielded data which are compared to data from word recognition experiments. Beside the presentation of his interesting data, Kawamoto describes his network as a dynamical system, sharing properties with a Hopfield net. This note addresses this description. It shows that there are many differences between the type of net used by Kawamoto and a Hopfield net and that the two networks can show very different dynamical behavior. It further shows, that some of the statements made by Kawamoto are not true in the general form he states them and that some arguments from the theory of Hopfield nets cannot be applied to the net used by Kawamoto in the way he does. Further some results of simulations of the net defined by Kawamoto are reported. Runs with different training material show, that the results given by Kawamoto depend heavily on the representation of the words he uses.

The Kawamoto Net

Kawamoto (1993) defines a network with units $\{1..N\}$ where "each unit receives input from the environment (...) as well as from every other unit in the network (...)." (Kawamoto, 1993 p. 481). "The activity of a unit in the network is represented as a real value ranging between -1.0 and $+1.0$" (p. 482). The activity of unit i " ... at time $t+1$ is

$$a_i(t+1) = LIMIT\left[\delta a_i(t) + \left[\sum_j W_{ij}(t)a_j(t)\right] + s_i(t)\right] \tag{1}$$

where δ is a decay constant, $s_i(t)$ is the influence of the input stimulus on unit i and $LIMIT$ bounds the activity to the range from -1.0 to $+1.0$." (Kawamoto (1993) p. 482 formulas [3] and [4]). W_{ij} denotes the (symmetric) connection strength between the units i and j. The meaning of the parameter (t) after W_{ij} is neither defined nor clear.

There is no further definition of the function $LIMIT$, so I will assume that it is defined as

$$LIMIT(x) = \begin{cases} 1.0 & if\ x > 1.0 \\ -1.0 & if\ x < -1.0 \\ x & otherwise \end{cases} \tag{2}$$

(See also Golden, 1986). If the input is removed $s_i(t)$ is set to 0.

To calculate the strengths of the connections the activities of the units are repeatedly set to different target patterns and in each such learning step the connections are modified according to the following rule: "This change in the connection strength from a given unit j to a unit i, ΔW_{ij}, can be expressed in terms of

$$\Delta W_{ij} = \eta(t_i - i_i)t_j \tag{3}$$

$$i_i = \sum_j W_{ij}t_j, \tag{4}$$

where η is a scalar learning constant, t_i and t_j are the target activation levels of units i and j, and i_i is the net input to unit i." (Kawamoto, 1993, p. 482).

Kawamoto further defines a Hopfield like "energy" function (Hopfield, 1982)

$$E = -(1/2)\sum_i \sum_j W_{ij}a_i a_j \tag{5}$$

(formula [5] in Kawamoto, 1993) and states that "each subsequent state of the network has an equal or lower energy relative to the previous state" (p. 483).

This statement is not true in general as can be seen from the following very simple example.

An Example

Consider a net with only two units and a target pattern $\{1, 1\}$. Under the assumption that the connection strengths at the beginning of the learning process are set to 0, one learning step yields symmetric connection strengths $W_{1,2} = W_{2,1} = \eta$ (Kawamoto, 1993, p. 485). In what follows this connection strength will be denoted by $W_{1,2} = W_{2,1} =: x$. Assume further that the two units have the starting activations a_1, a_2 and that $\delta = s_1 = s_2 = 0$. The sequence of activations and energy values is given in Table 1

For many values of a_1, a_2 and x the sequence of energy values is obviously increasing. For example for $a_1 = a_2 = x = 0.1$ it is -10^{-3}, -10^{-5}, -10^{-7}, ...

Table 1. Activities and energy of the two unit network

time	activity unit 1	activity unit 2	energy
1	a_1	a_2	$-\frac{1}{2}(a_1a_2x + a_2a_1x) = -a_1a_2x$
2	a_2x	a_1x	$-\frac{1}{2}(a_1xa_2xx + a_2xa_1xx) = -a_1a_2x^3$
3	a_1x^2	a_2x^2	$-\frac{1}{2}\left(a_1x^2a_2x^2x + a_2x^2a_1x^2x\right) = -a_1a_2x^5$
n	$a_{p(n)}x^{n-1}$	$a_{q(n)}x^{n-1}$	$-a_1a_2x^{2n-1}$

with $p(n) = ((n+1)\ mod\ 2) + 1$ *and* $q(n) = (n\ mod\ 2) + 1$.

The Difference

The net defined by Kawamoto differs from a Hopfield net (Hopfield, 1982) in several ways. These differences lead to different dynamic behavior:

– in a Hopfield net there are only two possible activity values 0 and 1
– Hopfield uses a threshold function to calculate the new activation value of a unit
– in a Hopfield net only one unit is updated per time step (sequential updating)

To show the effect of the differences I will show how changes of the definitions of Kawamoto towards the definition of Hopfield change the behavior of the system.

With a Hopfield like threshold function

$$a_i(t+1) = \begin{cases} 1 & if \sum_{k=1}^{N} W_{ik} a_k(t) > 0 \\ 0 & otherwise \end{cases} \tag{6}$$

for the updating of the units and positive values of a_1, a_2 and x the example would look like given in Table 2. The sequence of energy values for $a_1 = a_2 = x = 0.1$ would be -10^{-3}, -10^{-1}, -10^{-1}, -10^{-1}, ... a nonincreasing, finally constant sequence! The existence of activity values different from 0 and 1 in the Kawamoto net allows the increasing energy sequences.

Table 2. Activities and energy of the two unit network with a threshold function

time	activity unit 1	activity unit 2	energy
1	a_1	a_2	$-\frac{1}{2}(a_1 a_2 x + a_2 a_1 x) = -a_1 a_2 x$
2	1	1	$-x$
3	1	1	$-x$
n	1	1	$-x$

There are also other ways to guarantee nonincreasing energy in the net used by Kawamoto (Golden, 1986, Hui & Zak, 1992). For this purpose the inner part of formula (1) can be rewritten as a matrix vector multiplication, i.e. a linear function in a N–dimensional space:

$$a(t+1) = L[(\delta I + W)a(t) + s(t)] \tag{7}$$

with $a(t)$ being a vector containing the activities at time t and I denoting the unit matrix with 1 in the main diagonal and 0 elsewhere; $W = \{W_{i,j}\}_{i=1..N, j=1..N}$ is the matrix containing the weights, and $s(t)$ is a vector containing the external input at time t. L denotes the application of the function $LIMIT$ to the single components of a vector.

Golden (1986) introduces this formula in the form

$$b(t) = a(t) + \gamma W a(t) \tag{8}$$

and

$$a(t+1) = L(b(t)) \tag{9}$$

He shows that the energy is decreasing, if the matrix W is positive semidefinite or if $\gamma < 2/|\lambda_{min}|$ with λ_{min} denoting the minimum eigenvalue of W. Hui and Zak (1992) give a related condition for the extreme points of the hypercube to be fixed points.

These assumptions are not mentioned by Kawamoto. In fact, the only way to manipulate the properties of the eigenvalues of the matrix in his setting is to choose the learning parameter η appropriately, or to construct the training patterns representing the words appropriately.

Table 3. Activities of the two unit network with constant energy

time	activity unit 1	activity unit 2	energy
1	$a_1 = 1$	$a_2 = 0$	$-a_1 a_2 x = 0$
2	$a_2 x = 0$	$a_1 x = 1$	$-a_1 a_2 x^3 = 0$
3	$a_1 x^2 = 1$	$a_2 x^2 = 0$	$-a_1 a_2 x^5 = 0$
n	$a_{p(n)} x^{n-1} = n \bmod 2$	$a_{q(n)} x^{n-1} = (n+1) \bmod 2$	$-a_1 a_2 x^{2n-1} = 0$

with $p(n) = ((n+1) \bmod 2) + 1$ *and* $q(n) = (n \bmod 2) + 1$.

Nonincreasing Energy, Local Minima and Stable States

Kawamoto (1993) argues the following way:

> The activity in the network then changes in such a way as to consistently move down the energy gradient until a local minimum in the energy is reached (Hopfield, 1982; Golden, 1986). These states are called *stable states* because the state of the network generally does not change once the network reaches one of these states. (p. 483)

Here again the differences between the Hopfield net and the net used by Kawamoto may look small but they are important for the dynamic of the system. In a Hopfield net a pattern of activities may have many possible successors in the development of the system due to the fact, that the next unit for processing is selected at random. All these possible successors differ at most in the activity of one unit since the activity of only one unit is changed. If we assume a Hamming distance in the space of activity patterns (i. e. the distance between two patterns is the number of units in which the activities are different) all successors of an activity pattern lie in the immediate neighborhood since they have at most distance 1. If all patterns with distance 1 from a pattern a have a higher energy, this pattern a is a local minimum of the energy. In this case it *is* a stable state because the next state would have to be identical or have distance 1, which it cannot because the energy cannot increase.

In the case of the Kawamoto net things are quite different. On the one hand a pattern has a unique successor since there is no random selection of a unit, but all units are updated in parallel. On the other hand this means that many units can change their activity. This in turn means, that the Hamming distance to a successor is not bounded by 1 but the system can "jump around" over large distances. Hence "local minima" on the basis of the Hamming distance (as they are defined by Hopfield) are rather uninteresting for the description of the dynamics of the system.

To use the "local energy minima" for the Kawamoto net in a sense analog to the use in the Hopfield net it would be necessary to define a "distance" (i. e. a metric) for the space of activity patterns that guaranties that the distance between a pattern and its successor is bounded. (Golden, 1986 did not mention "local minima" at all.) If the energy of a system were nonincreasing a pattern would then be a stable state if all patterns within that distance had higher energy.

To give an example that even with a nonincreasing sequence of energy values there need not be convergence in the sequence of activity patterns for the net defined by Kawamoto (1993), we return to the example given above and set $a_1 = x = 1$ and $a_2 = 0$. (See Table 3)

The energy values are constantly 0 but the pattern of activity switches between the two patterns $\{1, 0\}$ and $\{0, 1\}$ in a cycle of length two for infinity and hence the system never converges. The Hamming distance of the two patterns is maximal (The activity of both units changes).

The difference to the Hopfield net is that Kawamoto uses parallel updating. If only one unit (chosen by chance) would be updated at a time, the system of the example would either converge

to the activity pattern $\{1, 1\}$ (if unit 2 were chosen first) or the activity pattern $\{0, 0\}$ (if unit 1 were chosen first). Both patterns are stable states of the system.

In the discussion of the "energy landscape" Kawamoto uses distances "in Cartesian coordinates" (Kawamoto 1993, p. 490); I will assume, that this means euclidean norm and the respective distance.

Note that the euclidean norm $|a| = \sqrt{\sum_{i=1}^{N} a_i^2}$ is small compared to the hamming distance or to $|a|_1 = \sum_{i=1}^{N} abs(a_i)$, since the range of a_i is restricted to $[-1, 1]$. The length of the unit vector is $\left|(1, ..., 1)^t\right| = \sqrt{216} = 14.70$ and the maximum distance between two vectors is $\sqrt{216 \cdot (1+1)^2} = 29.39$.

An upper bound for the maximum distance between two consecutive states of the system can be given. The growth of a pattern vector a when multiplied with the matrix W is bounded by the spectral radius ρ: $|Wa| \leq \rho|a|$. Hence the distance between two consecutive states of the net is bounded by

$$\begin{aligned} |L(\delta a + Wa) - a| &\leq |L(\delta a + Wa)| + |a| \leq |\delta a + Wa| + |a| \\ &\leq |\delta a| + |Wa| + |a| \leq \delta|a| + \rho|a| + |a| \leq (\delta + \rho + 1)|a| \end{aligned} \tag{10}$$

The growth of the length of a pattern vector $a(t)$ can be given as the *growth quotient* of the length of two consecutive states. It can be bounded from above in the following way:

$$\frac{|a(t+1)|}{|a(t)|} \leq \frac{(\delta + \rho)|a(t)|}{|a(t)|} \leq (\delta + \rho) \tag{11}$$

This upper bound is of interest only for very small pattern vectors $a(t)$ or matrices with small spectral radius because otherwise the influence of the function $LIMIT$ is too strong.

Simulations

The system described by Kawamoto (1993) has been implemented with the parameters given in his article, except for the learning material. This was replaced by 48 random patterns of values 1 and -1, which were presented 33 times in random sequence; altogether 1584 learning trials. The simulations were made to check whether the properties shown in the counter examples can also be found in networks of the size used by Kawamoto. Random patterns were used to check the influence of the representation constructed for the words.

The net was tested with the 48 patterns it was trained with and 48 different random patterns. For this purpose the activities of the units were set to $+0.25$ if the respective activity in the input pattern was 1, and to -0.25 if it was -1 (compare Kawamoto, 1993 p. 485). Then iterations were performed without any further external input. The iteration was terminated if either the (euclidean) distance between two consecutive Patterns was less than 0.01 or a total of 50 iterations were performed.

The net did not converge within 50 iterations; neither for any pattern of the training set, nor for one of the new patterns. The energy increased within all iteration sequences and the distance between two consecutive patterns in a sequence varied between 12.84 and 20.54. (Note, that the length of the input pattern vectors is only $\left|(0.25, ..., 0.25)^t\right| = 3.67$ and the maximum distance is 29.39.) The spectral radius of the matrix W can be estimated as $\rho \approx 213.5$. With the value $\delta = 0.95$ this yields an upper bound of $\delta + \rho = 214.45$ for the growth quotient. This upper bound can be reached with very small pattern vectors ($abs(a_i) = 10^{-12}$). With the parameters given by Kawamoto it varied between 0.99 and 4.00 because most of the activities after the first iteration are close to 1 or -1 due to the large spectral radius and the function $LIMIT$ (4.00 is the maximum possible value for the first iteration) .

After 50 iterations, the distance to the input pattern was on average 21.10 $(sd = 0.56)$ for the training patterns and 20.89 $(sd = 0.42)$ for the new patterns. The distance to the closest pattern of the training set was on average 17.10 $(sd = 0.66)$ for the patterns that started with a training pattern and 17.07 $(sd = 0.63)$ for those patterns, that originated from a new pattern.

If saturation of all units was taken as criterium (i. e. the iteration was terminated, if all activities in the net were either > 0.99 or < -0.99, compare Kawamoto, 1993, pp. 485 and 486) the average number of steps needed until the net was saturated was 3.77 $(sd = 2.35)$ for the training patterns and 3.65 $(sd = 1.86)$ for the new patterns. However, the distance to the input pattern was on average 22.19 $(sd = 2.24)$ for the training patterns and 21.17 $(sd = 0.80)$ for the new patterns. The minimum distance between the saturated patterns and the set of training patterns was on average 16.81 $(sd = 0.72)$ for the patterns that started with a training pattern and 16.94 $(sd = 0.70)$ for those saturated patterns, that originated from a new pattern. In short: nothing was learned.

These results show, that some of the assumptions made by Golden (1986) are not fulfilled by the matrix generated with the 48 random patterns. Probably the minimum eigenvalue of the matrix is much too small and hence the spectral radius much too big, compared to the value $\delta = 0.95$ used.

The spectral radius of the matrix can be reduced by reducing the learning parameter. With a value of $\eta = 0.0003$ the system was in general able to distinguish the training patterns from random patterns. The spectral radius was about 1.08. For 44 of the 48 training patterns the system converges within 50 iterations. For these cases the average number of iterations until convergence is 13.18 $(sd = 10.00)$. In 47 of 48 new random patterns the system did not converge within 50 iterations. For patterns from the training set the average distance between input patterns and the patterns to which they converged was 0.75 $(sd = 2.70)$. For new patterns the distance after 50 iterations or convergence was 19.01 $(sd = 0.84)$. For the patterns that started with a training pattern the distance between the patterns after iteration and the closest training pattern was again on average 0.75 $(sd = 2.70)$; for those patterns, that originated from a new pattern it was 15.65 $(sd = 2.59)$.

The maximum distances between two consecutive patterns in the iteration sequences is still 5.78 (the maximum growth quotient is 1.91), hence compared to the maximum possible distance between any two patterns (29.39) the “jumps” of the system are still quite big.

Better results were obtained with more training cycles: For 100 presentations of each training pattern and $\eta = 0.00008$ the average number of iterations was 8.52 $(sd = 4.70)$ for the training patterns. None of the new patterns converged within 50 steps. All inputs from the training set converged to the correct training pattern. For the new patterns the average distance between the input pattern and the pattern to which it converged was 17.77 $(sd = 0.66)$; the distance to the closest training pattern was 16.49 $(sd = 1.00)$. The largest distance between two consecutive iterations was 5.31, the maximum growth quotient 1.80

If we assume, that the net of Kawamoto really converged (as definition of the “number of iterations through the network” Kawamoto (1993, p. 486) mentions only saturation) there have to be reasons for this difference. One possible reason is the strong structure of the representation patterns of the words. The learning algorithm (equation (3)) changes the connection strength only as long as the incoming activation is different from the target activation. If the training patterns share many identical subpatterns, like the representations of ambiguous words used by Kawamoto, the sum of all changes made during learning is reduced. This leads to a smaller spectral radius of the matrix. This effect is further strengthened, if patterns that share identical subpatterns, like those representing ambiguous words, are presented more often than other patterns.

In this light the fact, that the net converges for (most) of the patterns constructed by Kawamoto but does not converge for any of the random patterns used in the present simulations raises the question in how far the results presented by Kawamoto are due to his specific (hand-made) representations and how much they are due to the learning of the net.

Discussion

The counter examples given above demonstrate, that two properties assumed by Kawamoto (1993) for his system do not hold in the general form he states them: convergence and decreasing energy. For the convergence this does not change much as long as the simulations he ran converged toward stable states and the validity of the simulation is restricted to the 48 examples he simulated. For the decreasing energy the situation is different. All the arguments using the "energy landscape" are based on this property. They need a new proof that the energy of the system is really nonincreasing under iteration (if that is in fact the case) and they need the definition of a distance in the space of patterns that makes the notion of local minima meaningful to the system, i. e. that guaranties that the movement of the system in the space of patterns is substantially bounded. Simulations with 48 random patterns as training material show, that the problems given in the counter example are not restricted to these small nets. The iteration sequences did not converge, and, if the saturation of the activities was taken as criterium to terminate iteration, the system showed no learning at all. Even if the learning parameter was reduced in such a way, that the system converged for the training patterns, the motion in the space of patterns was not substantially bounded. Altogether there seems to be a very strong influence of the special (hand-made) representation Kawamoto (1993) constructed for the words. The question can be raised, how far the results are due to these representations and how far they are due to the learning of the net.

References

Golden, R. M. (1986). The "Brain-State-in-a-Box" neural model is a gradient descent algorithm. *Journal of Mathematical Psychology 30*(1), 73-80.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA.*

Hui, S., & Zak, S. H. (1992). Dynamical analysis of the brain-state-in-a-box (bsb) neural models. *IEEE Transactions on Neural Networks 3*(1).

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language 32*, 474-516.