

The relationship of local context codes to sequence length memory capacity

William B Levy† and Xiangbao Wu

Department of Neurological Surgery, Box 420, University of Virginia Health Sciences Center, Charlottesville, VA 22908, USA

Received 10 October 1995

Abstract. This paper pursues part of a theory to understand sequence prediction networks that use local context neuron encodings. In particular, a previously described neural network model of hippocampal CA3 is studied. General expressions relating CA3 interpattern distances (which reflect local context codes) to sequence length memory capacity are created and verified by computer simulations. As a result, we confirm a very simple relationship between the sequence length memory capacity and a combination of average activity level and the average local context lifetimes. Sequence length memory capacity is also bounded for networks using such local context neuronal codes. Thus, this simple theory quantifies an important limitation on a fully adaptive (i.e. self-supervising) neural model capable of creating context-dependent learning and memory.

1. Introduction

Two goals of neuroscience are to understand how information is encoded in the brain and how this information is stored and retrieved. One way of reaching these goals is to study biologically consistent neural network models. Recurrently connected networks of neuron-like elements, for instance, have been widely used for associative storage of pattern sequences (Amari 1972, Fukushima 1973, Sompolinsky and Kanter 1986, Kleinfeld 1986, Buhmann and Schulten 1987, Coolen and Gielen 1988, Bauer and Krey 1990, Jordan 1986, Mozer 1989, Elman 1990, Reiss and Taylor 1991, Heskes and Gielen 1992, Bartholomeus and Coolen 1992, Sutton and Hobson 1994).

In such research, a common question about such networks concerns memory capacity. Moreover, because network simulations use many fewer neurons and synapses than actually exist, a theory of scalable capacity is needed to understand the fundamental limits of neural networks. Such values are available for static pattern recognition systems (e.g. Amit *et al* 1985, Amari 1989, Treves and Rolls 1991). In contrast to stable, point attractor networks that learn individual patterns, however, here we are concerned with neural network models that use dynamics to encode sequences. For sequence learning, we must define sequence length memory capacity. In particular, sequence length capacity is the maximal length of a sequence that can be learned so that approximate sequence completion exists. Sequence completion is an analogue of the pattern recognition problem of pattern completion. In the sequence completion problem defined here, only the first pattern of a long sequence is given as an external input. The network is then allowed to run on its own without further inputs.

† E-mail wbl@virginia.edu

In addition to the generic issue of memory capacity, there is another issue that concerns us, and this arises from the larger problem we are studying. Specifically, we are studying hippocampal function using a highly simplified, biologically consistent neural network model of hippocampal region CA3. At the cognitive level, our theory of the hippocampus is a modification of Hirsh's (1974) context theory that, in addition to the learning and storage of context in the older theory, emphasizes sequence prediction as using the encoded context. Thus, from this perspective, the most interesting aspect of our CA3 model is its ability to create, in an unsupervised fashion, its own code for context (Levy 1989, 1994). By computer simulations, we have shown that the network can learn and use context codes to solve the problems of simple sequence completion in a single trial learning (Minai and Levy 1993c), jump-ahead prediction after learning a sequence (Prepscius and Levy 1994) and sequence completion under ambiguous conditions (Minai *et al* 1994, Levy *et al* 1995). Local context neurons, the cells that combine to make up the overall context code, are hypothesized to be analogous to place cells (O'Keefe and Nadel 1978) in the hippocampus. With such a context code, the network also solves two other sequence prediction problems: finding shortcuts and seeking goals without search (Levy *et al* 1995).

It turns out that the embodiment of the network-wide encoding of context is easily evident at the level of single neurons. We call a neuron with such a prototypical firing pattern a 'local context neuron'. A local context neuron identifies a subsequence of a sequence. That is, such a neuron fires in response to temporally contiguous patterns while a non-context neuron fires to multiple unrelated portions of the sequence. Thus, a local context neuron, compared with a non-context neuron, recognizes the existence of a local subsequence or its parts. The accumulation of enough local context neurons may be very related to the correlated firing patterns in other (e.g. Griniasty *et al* 1993) models. Thus, one question of general interest here concerns the correlation of coded sequences of patterns (Amit *et al* 1994, Griniasty *et al* 1993, Levy and Wu 1995, Wu and Levy 1995). Such correlation values describe the similarity of codewords as a function of temporal separation. The similarity of codewords as a function of time (or sequence position) is interesting because it indicates how distinguishable different patterns of a sequence are and the relationship between subsequences. Such distinguishability and relatedness will, in turn, determine performance on sequence disambiguation problems and control the extent of jump-ahead predictions. In addition, when the codes change very slowly (see e.g. Prepscius and Levy, 1994), they provide a possible mechanism for solving the problem of temporal chunking (e.g. Hulse and Dorsky 1977, Fountain and Annau 1984, Dallal and Meck 1990, Macuda and Roberts 1995).

Two interrelated problems are addressed in this paper. First, we resolve the sequence length memory capacity for networks that learn sequences using local context codes. Second, a measurement of correlation of local context codes for such networks is investigated and related to capacity and activity. Although we doubt that any approach, including the one here, will be totally general, the approach here will be relevant to all sequence prediction networks that use local context neuron encodings based on adaptive recodings.

2. The model and the methods

2.1. The computational architecture

A generally accepted, gross hippocampal computational architecture (e.g. Levy 1989, Eichenbaum and Buckingham 1991, O'Reilly and McClelland 1994, Hasselmo and Schnell 1994) has three parts:

- (i) an input layer;
- (ii) a recoder inspired by the hippocampal CA3 region;
- (iii) a decoder of the recoded signals inspired by the hippocampal CA1 region.

However, this paper focuses on computations performed by a CA3-like structure.

Our specification of the CA3-like portion of the network has four essential aspects:

- (i) Sparse recurrent excitatory connectivity that produces more overall excitation than the external input.
- (ii) A neuronal delay of at least one time step in converting an input to an output (i-o).
- (iii) An associative modification rule that spans at least the i-o time step.
- (iv) A certain generic feedback inhibition that bounds total activity narrowly, albeit imperfectly in comparison with competitive networks.

The networks consist of an input layer (whose effect is loosely analogous to the combined entorhinal cortex and dentate gyrus inputs to CA3) and a sparsely connected feedback layer (CA3-like). For simplicity, there is only a single input line, x_j , from the input layer to each CA3 neuron. This input will always fire a CA3 neuron if active and, as indicated by the subscript j , the connectivity for this external input is 1:1 with CA3 neurons. Note that an active external input always produces a firing, but that neurons with no active external input are *not* forced to the zero state and can be fired through feedback connections. The recurrent CA3-like layer consists of 1024 binary (0/1) primary neurons with identical firing thresholds, θ . External activity is always kept very sparse, typically less than 0.1% of the neurons are externally driven at any one time step. The neurons in the CA3-like layer are interconnected via a Bernoulli process: each neuron j has a probability p of receiving a modifiable excitatory connection from each neuron i in this recurrent layer (in simulations p ranged from 0.05 to 0.15, here 0.1 is used). The presence or absence of such a connection is indicated by the binary variable c_{ij} . Feedback inhibition is mediated by a single interneuron that receives input from all primary neurons in the CA3 layer; it then provides an identical shunting conductance to all j , proportional to its input. There is a fast feedforward inhibitory effect, mediated through the constant K_I , that accompanies external activation and whose effect is proportional to the summed external input. At time t , taking $w_{ij}(t)$ as the excitatory weight from neuron i to j , K_I as the equivalent fixed inhibitory weight from the input layer and K_R as the equivalent fixed weight for feedback inhibition, the excitation y_j of CA3 neuron j is given by:

$$y_j(t) = \frac{\sum_i w_{ij} c_{ij} z_i(t-1)}{\sum_i w_{ij} c_{ij} z_i(t-1) + K_I \sum_i x_i(t) + K_R \sum_i z_i(t-1)}$$

and its output by,

$$z_j(t) = \begin{cases} 1 & \text{if } y_j(t) \geq \theta \text{ or } x_j(t) = 1 \\ 0 & \text{otherwise.} \end{cases}$$

2.2. The learning rule

A Hebbian-type postsynaptic associative modification rule is used in our simulations here. We have examined several learning rules and found little difference; we choose one of these rules here (Levy and Steward 1979, Levy 1982). For input i and output j , this postsynaptic rule is given by: $w_{ij}(t) = w_{ij}(t-1) + \epsilon z_j(t)[z_i(t-1) - w_{ij}(t-1)]$.

Because of the NMDA receptor, associative modification is asymmetric in time (Levy and Steward 1983, Holmes and Levy 1990) and the rule used here allows the network to

learn an association that spans only one time step. Note, however, that this time span is chosen as the minimal one of interest not that of greatest biological relevance.

2.3. The input sequences

For the quantitative simulations in this paper, each external input pattern contains eight on-bits out of the 1024 possible. In a single sequence, the successive patterns are constantly moving away from all the previous input patterns. From one pattern to the next there is a shift of k bit(s) per unit of time ($k = 1, 2, 3, 4, 5, 6, 7$ or 8). As a result, the Hamming distance between a pair of successive prototype input patterns is $2k$. In other words, the overlap length of a pair of successive input patterns is $8 - k$. Thus, given the set of k , the overlap length can be 7, 6, 5, 4, 3, 2, 1 or 0. If $k = 8$, all the input patterns are orthogonal to one another and the overlap is 0. At the other extreme, if $k = 1$ then the overlap length is 7. This amount of overlap gives the slowest possible shifting input sequence. We also study noisy sequences described in Wu *et al* (1996).

There are at least as many possible definitions of memory capacity in sequence learning networks as there are sequences; in other words, any definition must be a function of the inputs. The definitions used here represent an attempt to span a reasonable range of input types. At the same time, these inputs seem both simple and sufficiently general that they, or their near cousins, might be applied elsewhere. That is, actual sequences will largely be a mixture of the sequence types studied here. Finally, it should also be pointed out that our networks and their outputs are random variables. Thus, single maximal observations of capacity are not very useful. Therefore, to avoid unrepresentative capacity values, we present only robust capacity values, where robust is defined as replicable in four out of five randomly constructed networks. For example, if a capacity value of 165 is stated as the sequence length memory capacity, 166 was found to be too long for consistently satisfactory recall.

In our simulations here, a selected input sequence is presented to the network for 300 trials with a learning rate constant, ε , set at 0.01. With such a value and so many trials, additional learning trials do not affect the results, so long as activity is kept approximately constant.

Before each learning trial (i.e. fully driven sequence), CA3 neurons are randomly excited. During recall testing, the network is randomized, given the first pattern of the sequence as a prompt and then allowed to run without further input.

2.4. The average normalized Hamming distance and its relationship to a correlation measure

To evaluate network performance, the normalized Hamming distance of each state generated in response to the single pattern probe test is compared with every state produced by the fully driven CA3 codings. The normalized value is the Hamming distance divided by the maximal possible Hamming distance. If a pattern vector always consists of M neurons on, for example, the normalized Hamming distance between two such patterns is defined as the Hamming distance divided by $2M$.

Here we define a summary statistical characterization of the firing patterns of local context neurons, the average normalized Hamming distance between two vectors, $\mathbf{Z}(t)$ and $\mathbf{Z}(t + \tau)$. The Hamming distance between binary $\{0, 1\}$ vectors can be calculated as

$$d_H(\tau) \equiv d_H(\mathbf{Z}(t), \mathbf{Z}(t + \tau)) = \sum_j Z_j(t) + \sum_j Z_j(t + \tau) - 2\mathbf{Z}(t)^T \mathbf{Z}(t + \tau).$$

On average then,

$$E[d_H(\tau)] = 2E \left[\sum Z_j(t) \right] - 2E \left[Z(t)^T Z(t + \tau) \right].$$

Note that this quantity ranges from 0 to $2E[\sum Z_j(t)]$. Therefore, we define the average normalized Hamming distance as

$$E[d_{nH}(\tau)] = \frac{2E \left[\sum Z_j(t) \right] - 2E \left[Z(t)^T Z(t + \tau) \right]}{2E \left[\sum Z_j(t) \right]} = 1 - \frac{E \left[Z(t)^T Z(t + \tau) \right]}{an}$$

where in the last step, we denote $E \left[\sum Z_j(t) \right]$ as the average neuronal activity, a , times the number of neurons, n .

The interpretation of this measure is straightforward. If codewords are very similar, this measure is near zero and if codewords are very different, this measure is near one. As is obvious by inspection, this normalized distance measure is the first cousin to something approximating a complement of a correlation or a slightly misnormalized calculation of the sine between two vectors. The use of this measure is justified here by its straightforward meaning, the natural use of a Hamming measure for comparing vertices on the n -cube and finally, by the simple form our theory takes when using this measure.

The idea of comparing codewords in a sequence with something like a correlation function is not new (Griniasty *et al* 1993, Amit *et al* 1994) and, as we shall see, their results are not dissimilar. Indeed, the important underlying similarity between our model and theirs (e.g. the use of a time-spanning associative modification rule) predicts such similar results.

3. Results

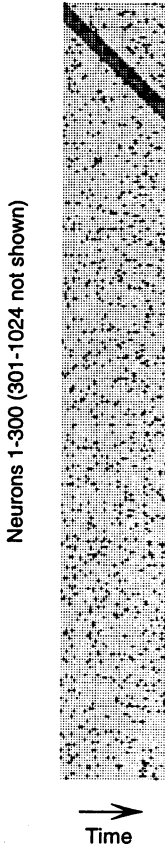
Using the normalized Hamming distances between output pattern codes as a function of time span between codewords, the results show certain simple qualitative relationships between the overlap length of the input sequence patterns, activity levels and local context lifetimes. Here context codes are quantified by measuring pattern similarities over time because such similarities are at the heart of what a context neuron is and how a neural network makes use of them. Finally, we give a general expression for interpattern distances of the local context codes.

But first, we describe some simulation results and analyse the learned patterns of neuronal firing.

3.1. Viewing and evaluating local context codes

3.1.1. Local context neuron firings. Figure 1 shows what it means for a sequence learning network to create its own compressed code (Levy 1989). After training, the firing activity of neurons (figure 1(b)) becomes more regular in comparison to before training (figure 1(a)). Note how the CA3 encodings of successive patterns are similar to each other but still different enough to be distinguished. Note also the distinctive activity patterns of the many local context neurons. Such neurons usually turn on for one short sequence of firings just once. That is once on, they tend to stay on for a few time steps in succession. In this example they stay on for about 4–10 successive time points. Such local context neurons greatly outnumber the externally driven neurons and it is the interleaved patterning of active local context neurons that supplies the information for context-dependent prediction in a flexible fashion.

(a) CA3 activities before training



(b) CA3 activities after training

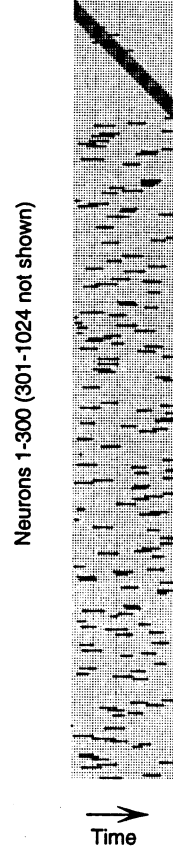


Figure 1. The development of context cell firing. (a) Neuronal firings in response to an input sequence of 40 patterns before learning. The externally driven neurons (1–47) are visible as the diagonal stripe of activity that begins in the upper left corner. (b) Neuronal firings in response to the input sequence after 300 presentations of the input. Note the firing pattern for individual neurons goes from a somewhat random pattern in (a) to a pattern of continuous on–continuous off in (b). Time goes from left to right and neurons go from top to bottom. A large dot stands for a firing (1) and a small dot stands for a non-firing (0).

3.1.2. Average normalized Hamming distance of the local context codes. Figures 2 and 3 quantify some of the self-created CA3 codes once the synaptic weights have stabilized. Figures 2 and 3 are plots of the average normalized Hamming distance between patterns as a function of the time between patterns in a sequence. Note that the curve in figure 2 with the (●) is noticeably different from the others. This is from a network with no synaptic modification but the same activity level ($\sim 5\%$) as the other networks. For this case the average distance immediately rises from zero to approximately 0.95. This curve shows that neighbouring patterns, before learning, are essentially uncorrelated and that uncorrelated patterns have the appropriate random overlap with other patterns. This randomness arises from the network connectivity (see Minai and Levy 1993a). On the other hand, the simulations with functioning synaptic modification produce a different result. Synaptic modification produces increased sequence similarity for temporal neighbours. How

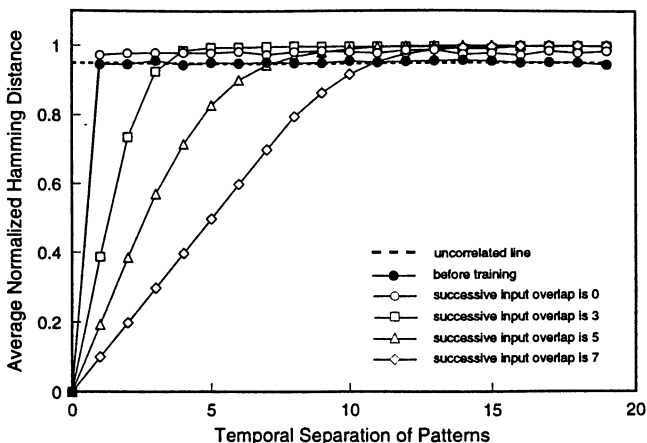


Figure 2. Comparison of the normalized Hamming distance between successive learned patterns against an unlearned sequence. The average normalized Hamming distance ($E[d_{nH}(\tau)]$) between CA3 states is plotted as a function of the time between states. As the overlap length between successive input patterns increases, the relatedness of the self-generated pattern codes within a sequence increases. The curve with the (●) is from a network with no synaptic modification. The other four curves are different from each other because the overlap lengths of the successive slowly shifting input patterns of sequences are different (as described in the legend). The dashed line is the average Hamming distance between uncorrelated patterns. Note how all sequences except that without learning go from correlated to anticorrelated. In the exceptional curve (the one before learning), the asymptotic value of $E[d_{nH}(\tau)]$ is the uncorrelated value. (If two patterns are uncorrelated $E[d_{nH}(t)] = 1 - E[d_{nH}(t)]^T E[d_{nH}(t)]/an = 1 - (a^2n/an) = 1 - a$, here 0.95). All networks for these plots were run with approximately the same activity level (5%) by adjusting feedback inhibition. Specifically, K_R for different overlap lengths of the successive input patterns was 0.0162 for 7, 0.0138 for 5, 0.0118 for 3 and 0.01 for 0. The other parameters were $K_I = 0.018$, $\theta = 0.8$.

far away this similarity extends is a function of the similarity of the input patterns that make up each sequence. For example, the similarity extends for temporal neighbours up to about 10 steps away for overlap length seven of the successive input patterns, but about seven steps away for overlap length five of the successive input patterns. Moreover, more distant neighbours are not just uncorrelated, but also anticorrelated in that the codewords far apart tend towards orthogonality. These qualitative results are true regardless of input overlap.

3.2. Memory capacity as a function of average activity level and the context life of the network

3.2.1. Sequence length memory capacity. As mentioned in the introduction, we define the sequence length memory capacity of the network as the maximal length of a sequence that can be learned so that approximate sequential pattern completion exists. Specifically, at least 75% of the different patterns in the learned sequence must be sequentially recalled during test for a test trial to be successful. For instance, suppose a sequence of length 10, say ABCDEFGHIJ, is learned by the network. When tested, the network only receives pattern A as an input. In response to this one input, it might produce ABBDEEGHIJ. This would be a successful recall because 80% of the different patterns were recalled in the proper sequence. Although the definition seems a little bit arbitrary, we believe some amount of noise is always tolerable in a neural network and we have to choose some value. In fact,

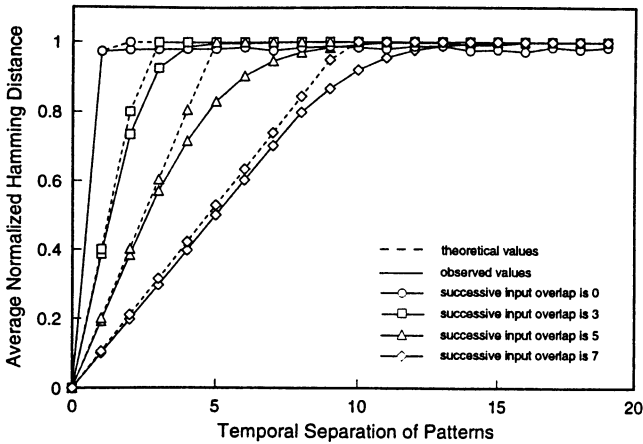


Figure 3. Interpattern distance of the local context codes. Theoretical predictions (equation (3)) compare well with the simulation results for the early time points and for the asymptotic anticorrelation that is predicted. These simulations are identical to those in figure 2.

based on pilot research not presented here, it does not matter much exactly how such a definition is parametrized in terms of the qualitative results we see.

3.2.2. An estimate of sequence length memory capacity. Before actually measuring sequence length memory capacity, let us estimate it by assuming all neurons that fire are local context neurons. Denote C as the maximum sequence length capacity and a as the average activity over all time and neurons. Let $N(l)$ be the number of context runs (i.e. sequential firings) of specified length l , where l can be 0, 1, 2, ..., up to the longest possible length which is, in fact, C .

It is now possible to write down a relationship between capacity (C), average activity (a) and the average length of a local context neuron firing sequence ($E[l]$). If we total up all neuronal firings and divide by the number of neurons, n and the full time span of activity, C , we obtain average activity

$$a = \frac{\sum N(l)l}{Cn}.$$

Rewriting this in terms of C gives

$$C = \frac{\sum N(l)l}{an} = \frac{\sum P(l)l}{a}$$

because $n = \sum N(l)$ and, by virtue of the counts $N(l)$ and their normalization, $N(l)/\sum N(l)$, there is a probability distribution, $P(l)$.

The relationship between average local context lifetime, average activity and sequence length capacity follows immediately from this form:

$$C = \frac{E[l]}{a}. \quad (1)$$

3.2.3. Verification of the theory. Figure 4 plots the average local context life, $E[l]$, against the capacity, C , for a series of simulations as well as for the theory. The predicted results compared quite favourably with the computational results. Note that in producing these

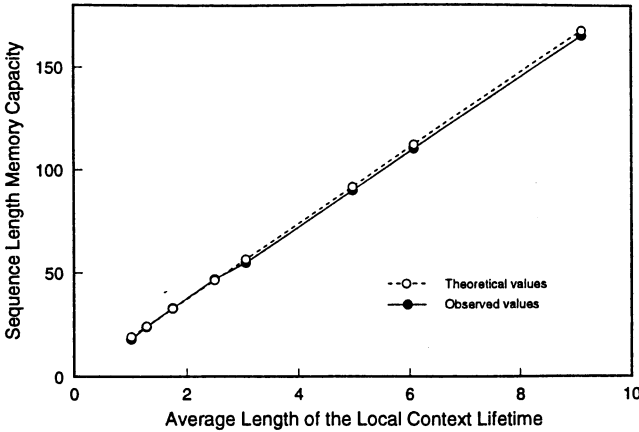


Figure 4. Memory capacity as a function of average local context lifetime. Here the average activity level is kept at a nearly constant value of 5%. Comparing the theoretical plot (equation (1), dotted line) to the empirically defined regression curve fit (solid line) shows that sequence length memory capacity C is, for all practical purposes, a simple linear function of the average context lifetime $E[l]$. The eight different C values correspond to the different sequences. The different sequences are generated by input sequences with different overlap lengths of successive input patterns, i.e. 0, 1, 2, 3, 4, 5, 6 and 7 respectively. To produce equal activity values, K_R was set differently, i.e. 0.01, 0.0104, 0.0114, 0.0118, 0.0123, 0.014, 0.0153 and 0.0165 for the respective overlaps. The threshold θ was 0.8 and K_I was 0.018.

results, we have maintained relatively constant levels of average activity (5%) by adjusting K_R as needed. From figure 4, it can be seen that our estimation works very well for slowly shifting inputs and errs by less than 10% for orthogonal inputs.

3.3. A general expression for interpattern distances of the local context codes

3.3.1. A theoretical value of the average normalized Hamming distance. Now, let us create a general expression to estimate the average normalized Hamming distance. First, we assume that the number of active neurons at each time step is approximately constant (this is essentially what the shunting inhibition tries to produce). Then, for each time step, on average R_{off} local context neurons turn off and, perforce, approximately R_{on} local context neurons must turn on. Thus, we are assuming $R_{\text{on}} \approx R_{\text{off}} \approx R$. Then, $E[Z(t)^T Z(t + \tau)] \approx an - \tau R$ for $0 < \tau R \leq an$ and 0 otherwise. Now, we recall that $E[d_{\text{NH}}(\tau)] = 1 - E[Z(t)^T Z(t + \tau)]/an$. Thus, substitution yields

$$E[d_{\text{NH}}(\tau)] \approx \tau R/an. \quad (2)$$

The idealized (and typical) local context neuron goes from non-firing to firing just once, so let this be our definition. Therefore, we can think of these neurons as a limited resource that is used up. Since R is the average number of neurons that pass from firing to non-firing in one time step, then in n/R time steps all neurons are used up. Therefore, if we ignore one end of the sequence, we can write $C \approx n/R$. But recall that $C = E[l]/a$. Therefore, (2) can be written as

$$E[d_{\text{NH}}(\tau)] \approx \begin{cases} \tau/E[l] & \text{for } \tau \leq E[l] \\ = 1 & \text{otherwise.} \end{cases} \quad (3)$$

Note that as $E[l]$ gets larger, that is, as the average local context length gets longer, $E[d_{\text{NH}}(\tau)]$ gets smaller for a fixed τ . This makes sense. The longer lived the local context neurons are, the more slowly the system moves away from its present position.

3.3.2. Predicting interpattern distances by controlling local context lifetimes. If we fix connectivity and activity, there are two ways of controlling $E[l]$: (i) by changing the time span of the associative modification rule or (ii) by changing the input. We postpone examination of (i) because it leads to complexities that require independent analysis, but we shall examine (ii) as a test of our second prediction, equation (3).

In figure 3, such predictions are compared with the simulation results. Based on this comparison, it seems we have at least qualitatively, an acceptable first-order theory of the model. Importantly, the theory captures three characteristics of the simulation data:

- (i) small temporal separation Hamming distances are well approximated;
- (ii) the theory correctly predicts the shift right of the average normalized Hamming distance curves with increasing overlap length of successive input patterns (figure 3);
- (iii) it predicts anticorrelations, that is orthogonality, at longer time points. (Anticorrelations are not predicted by other simple models with independence. For example, suppose that rather than R_{off} and R_{on} , the fixed rates of neurons turning off and on, we instead assumed a fixed probability for each neuron turning off or on. Then this example gives values that converge more closely to uncorrelatedness with increasing time.)

However, as clear as this qualitative agreement is, it is also clear that the statistical model requires quantitative improvement. For example, the convergence to maximal anticorrelations ($d_{\text{NH}} = 1$) is too fast.

3.3.3. Another way of producing covarying activity levels and context lifetimes. In addition to varying the average normalized Hamming distance of the input to change activity levels, and thereby change context cell lifetime as is essentially illustrated in figure 2, we can vary the statistical characteristics of the input. In a previous paper (Wu *et al* 1996), we examined the effect of noise on network performance when the input has a slowly shifting input ($k = 1$). Here, we present some novel results from that study by plotting average activity level against average length of local context lifetime. As can be seen in figure 5, there is a clear linear relationship ($r = 0.99$) between average activity level and average context lifetime as predicted by equation (1), i.e. $E[l] = Ca$.

3.4. Bounding sequence length memory capacity for local context neuron based codes

Since simulations sustain our estimates so far, we shall point out the bounds placed on memory capacity by this analysis. Suppose we only restrict firing by the definition of a local context neuron and we require this feature of all neurons in the network. Suppose also the uniqueness of all codewords. If activity values are allowed to fluctuate from 0 to n then memory capacity could, in theory, reach $2n$. That is, there are $2n$ unique codewords that can be constructed if each local context unit is on for n steps, with $R_{\text{on}} = 1$ and $R_{\text{off}} = 0$ for the first n steps and then $R_{\text{on}} = 0$ and $R_{\text{off}} = 1$ for the last n steps.

Interestingly, this code of maximum sequence length capacity has an average activity level of $1/2$, and activity levels fluctuate as widely as possible. However, this code would not occur because feedback inhibition constrains activity levels to remain around a particular value (Minai and Levy 1993b, 1994).

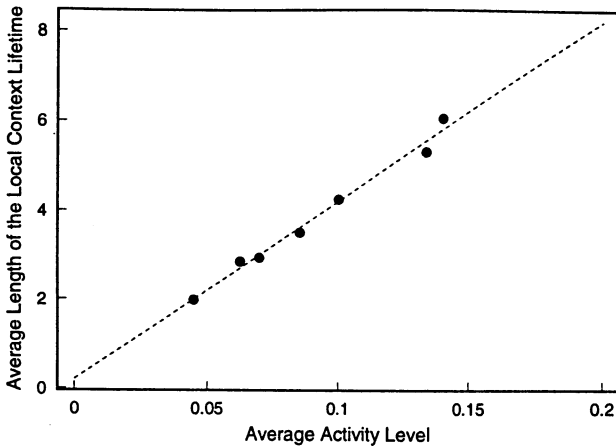


Figure 5. Average length of the local context lifetime is a linear function of average activity of the network. Different activity levels are achieved by different noise levels while keeping the network parameters constant. The correlation coefficient of the linear fit here is 0.99. The noise-to-signal level varied from 0.11 to 0.95. All networks learned a sequence of 40 patterns. For more details of this simulation, see Wu *et al* (1996).

Thus, we now move to the other, more sensible extreme we used earlier and strictly enforce $R_{\text{on}} = R_{\text{off}}$. Then a maximum capacity of just n can be constructed. This occurs at the lowest possible activity level, $a = n^{-1}$ when $R = 1$ (implying $E[I] = 1$). Such a capacity can also be created using longer context life times if we relax the strict enforcement of $R_{\text{on}} = R_{\text{off}}$ throughout the whole sequence. However, such codes use but a single value of local context lifetimes and are not really desirable because, in all likelihood, they fail to reflect the fluctuating nature of most input sequences.

In contrast to this theoretical upper bound of length n , we have, elsewhere (Levy and Wu 1995), examined capacity as an empirical problem and found substantially lower values than n . For simple deterministic sequences, we find sequence length capacity to vary profoundly as a function of input shifting rates (k). For a network of 1024 neurons running at 5% activity, we can regularly achieve a capacity of 165 for slowly shifting ($k = 1$) input sequences, or just 20 for orthogonal ($k = 8$) input sequences. As one might guess, lower activity levels give greater sequence length memory capacity, which is a result not just limited to local context neuron codes (Treves and Rolls 1991). In fact (albeit with some difficulty and probably lacking the robustness required here) an activity level of 3.5% was achieved to produce the sequence length capacity of 200 in a network of 1024 neurons.

4. Discussion

We have presented a first analysis of a very simplified neural model inspired by our knowledge of the physiology and anatomy of hippocampal region CA3. As mentioned in the introduction, this model displays many properties that are fundamental to hippocampal theories. This simple model can learn and use context. The network can perform several types of sequence prediction including: simple sequence completion, sequence completion under ambiguous conditions, finding a shortcut through a sequence, or following an appropriate path to a particular place, that is weakly specified in a sequence. Of these five types of predictions, the last four all require that the information within the problem

be coded by local context neurons to produce good predictions.

The fact that the model solves these various problems by creating and using a local context code (much like hippocampal place cells (O'Keefe and Nadel 1978)), motivated us to study the network at a slightly more fundamental level. In particular, the theory provided by this paper helps us understand the fundamental limitations that constrain the capabilities of our CA3 model and, if our model is valid, perhaps region CA3 itself. In particular, the theory here focuses on the constraints of sequence length memory capacity (figure 4) and local context lifetimes (figure 5) and on the interplay of activity with both of these.

The limitations on sequence length memory capacity arise from the characteristic of a local context neuron as a specific pattern recognition device that recognizes a limited subsequence of the whole sequence that the network is learning. Because, in our idealized definition of sequence length capacity, the local context neurons are used for only one subsequence, we can think of these neurons as an exhaustible resource. Thus, the higher the activity level in the network, the greater the rate at which these neurons are being used up. In theory, there can be some tendency to increase memory capacity by having longer runs of context cell firing. Unfortunately, it is only by lowering activity levels, i.e. creating a sparser coding, that we have reliably increased capacity.

In tuning a network to solve any particular problem we must pay special attention to the relationship (derived from equation (1))

$$a = \frac{E[l]}{C}. \quad (4)$$

Unfortunately, if, as is usually the case, we can only control the variables that control activity (K_R , K_I and θ), then we cannot increase both C and $E[l]$ simultaneously. That is, if we can only alter activity, we are forced to choose between capacity and successive similarities. From equation (4) it is easy to see that the network will fail at any problem that requires a combination of long local context lifetime and large sequence length memory capacity. The basic disambiguation problem (Minai *et al* 1994, Levy *et al* 1995) provides a simple example. If a long sequence must be learned and a long shared subsequence must be spanned and if both requirements are long enough, no amount of tuning of the network parameters will succeed in producing a network that solves the problem. Adjustment of a non-standard parameter or more circuitry is a definite requirement for a successful solution.

Finally, we might reflect on the possibility that the brain itself can adjust parameters to alter activity in the region CA3 of the hippocampus. Most of the time in the awake animal the hippocampus works at very low activity levels, or at least so one would infer because the cognitive mapping problem is characterized by such low activity levels (Thompson and Best 1989). On the other hand, in the classical conditioning paradigm (and the hippocampus does participate in the learning of classical conditioning (Kim *et al* 1995)), relatively high levels of activity are seen throughout the hippocampus (Berger *et al* 1976). Thus, neuroscientists may have already observed a pair of behaviours in which capacity and context lifetime are differentially optimized.

Acknowledgments

This work was supported by NIH MH48161, MH00622, EPRI RP8030-08 and Pittsburgh Supercomputing Center Grant No BNS950001P to WBL and by the Department of Neurosurgery, Dr John A Jane, Chairman. We thank Dr Robert A Baxter for his feedback during the course of the research and Mr A Amarasingham for his help on the programming. We also appreciate the observations and comments of Dr Joanna M Tyrcha that led to the definition of robust capacity.

References

- Amari S 1972 Learning patterns and pattern sequences by self-organizing nets of threshold elements *IEEE Trans. Comput.* **C 21** 1197–206
- 1989 Characteristics of sparsely encoded associative memory *Neural Networks* **2** 451–7
- Amit D J, Brunel N and Tsodyks M V 1994 Correlations of cortical Hebbian reverberations: theory versus experiment *J. Neurosci.* **14** 6435–45
- Amit D J, Gutfreud H and Sompolinsky H 1985 Spin-glass models of neural networks *Phys. Rev. A* **2** 1007–18
- Bartholomeus M and Coolen A C C 1992 Sequences of smoothly correlated patterns in neural networks with random transmission delays *Biol. Cybern.* **67** 285–90
- Bauer K and Krey U 1990 On learning and recognition of temporal sequences of correlated patterns *Z. Phys. B* **79** 461–75
- Berger T W, Alger B and Thompson R F 1976 Neuronal substrate of classical conditioning in the hippocampus *Science* **192** 483–5
- Buhmann J and Schulten K 1987 Noise-driven temporal association in neural networks *Europhys. Lett.* **4** 1205–9
- Coolen A C C and Gielen C C A M 1988 Delays in neural networks *Europhys. Lett.* **7** 281–5
- Dallal N L and Meck W H 1990 Hierarchical Structures: chunking by food type facilitates spatial memory *J. Exp. Psychol.: Animal Behavior Processes* **16** 69–84
- Eichenbaum H and Buckingham J 1991 Studies on hippocampal processing: experiment, theory and model *Neurocomputation and Learning: Foundations of Adaptive Networks* ed M Gabriel and J Moore (Cambridge, MA: MIT Press) pp 171–231
- Elman J L 1990 Finding structure in time *Cog. Sci.* **14** 179–211
- Fountain S and Annau Z 1984 Chunking, sorting and rule-learning from serial patterns of brain-stimulation reward by rats *Animal Learning Behav.* **12** 265–74
- Fukushima K 1973 A model of associative memory in the brain *Kybernetik* **12** 58–63
- Griniasty M, Tsodyks M V and Amit D J 1993 Conversion of temporal correlations between stimuli to spatial correlations between attractors *Neural Comput.* **5** 1–17
- Hasselmo M E and Schnell E 1994 Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: computational modeling and brain slice physiology *J. Neurosci.* **14** 3898–914
- Heskes T M and Gielen S 1992 Retrieval of pattern sequences at variable speeds in a neural network with delays *Neural Networks* **5** 145–52
- Hirsh R 1974 The hippocampus and contextual retrieval of information from memory *Behav. Biol.* **12** 421–44
- Holmes W R and Levy W B 1990 Insights into long-term potentiation from computational models on NMDA receptor-mediated calcium influx and intracellular calcium concentration changes *J. Neurophysiol.* **63** 1148–68
- Hulse S H and Dorsky N P 1977 Structural complexity as a determinant of serial pattern learning *Learning and Motivation* **8** 488–506
- Jordan M I 1986 Attractor dynamics and parallelism in a connectionist sequential machine *Proc. 8th Conf. Cog. Sci. Soc.* (Hillsdale, NJ: Erlbaum) pp 531–46
- Kim J J, Clark R E and Thompson R F 1995 Hippocampectomy impairs the memory of recently, but not remotely, acquired trace eyeblink conditioned-responses *Behav. Neurosci.* **109** 195–203
- Kleinfeld D 1986 Sequential state generation by model neural networks *Proc. Natl Acad. Sci. USA* **83** 9469–73
- Levy W B 1982 Associative encoding at synapses *Proc. 4th Ann. Conf. Cog. Sci. Soc.* pp 135–6
- 1989 A computational approach to hippocampal function *Computational Models of Learning in Simple Neural Systems* ed R D Hawkins and G H Bower (New York: Academic) pp 243–305
- 1994 Unification of hippocampal function via computational considerations *INNS World Congress on Neural Networks IV* (Int. Neural Network Soc.) pp 661–6
- Levy W B and Steward O 1979 Synapses as associative memory elements in the hippocampal formation *Brain Res.* **175** 233–45
- 1983 Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus *Neurosci.* **8** 791–7
- Levy W B and Wu X B 1995 Controlling performance by controlling activity levels in a model of hippocampal region CA3. II: memory capacity comes at the expense of context cell firing and compressed coding *INNS World Congress on Neural Networks I* (Int. Neural Network Soc.) pp 582–6
- Levy W B, Wu X B and Baxter R A 1995 Unification of hippocampal function via computational/encoding considerations (*Proc. 3rd Workshop on Neural Networks: from Biology to High Energy Physics*) *Int. J. Neural Sys.* **6** suppl. 71–80
- Macuda T and Roberts W A 1995 Further evidence for hierarchical chunking in rat spatial memory *J. Exp. Psychol.: Animal Behavior Processes* **21** 20–32

- Minai A A and Levy W B 1993a The dynamics of sparse random networks *Biol. Cybern.* **70** 177–87
- 1993b Predicting complex behavior in sparse asymmetric networks *Advances in Neural Information Processing Systems* ed C L Giles *et al* (San Mateo, CA: Morgan Kaufmann) p 556–63
- 1993c Sequence learning in a single trial *INNS World Congress on Neural Networks II* (Int. Neural Network Soc.) pp 505–8
- 1994 Setting the activity level in sparse random networks *Neural Comput.* **6** 85–99
- Minai A, Barrows G and Levy W B 1994 Disambiguation of pattern sequences with recurrent networks *INNS World Congress on Neural Networks IV* (Int. Neural Network Soc.) pp 176–81
- Mozier M C 1989 A focused back propagation algorithm for temporal pattern recognition *Complex Sys.* **3** 349–81
- O'Keefe J and Nadel L 1978 *The Hippocampus as a Cognitive Map* (Oxford: OUP)
- O'Reilly R C and McClelland J L 1994 Hippocampal conjunctive encoding, storage and recall: avoiding a tradeoff *Hippocampus* **4** 661–82
- Prepscius C and Levy W B 1994 Sequence prediction and cognitive mapping by a biologically plausible neural network *INNS World Congress on Neural Networks IV* (Int. Neural Network Soc.) pp 164–9
- Reiss M and Taylor J G 1991 Storing temporal sequences *Neural Networks* **4** 773–87
- Sompolinsky H and Kanter I 1986 Temporal association in asymmetric neural networks *Phys. Rev. Lett.* **57** 2861–4
- Sutton J P and Hobson J A 1994 State-dependent sequencing and learning *Computation in Neurons and Neural Systems* ed F H Eeckman (Norwell, MA: Kluwer) p 275–80
- Thompson L T and Best P J 1989 Place cells and silent cells in the hippocampus of freely-behaving rats *J. Neurosci.* **9** 2382–90
- Treves A and Rolls E T 1991 What determines the capacity of autoassociative memories in the brain? *Network* **2** 371–97
- Wu X B and Levy W B 1995 Controlling performance by controlling activity levels in a model of hippocampal region CA3. I: overcoming the effect of noise by adjusting network excitability parameters *INNS World Congress on Neural Networks I* (Int. Neural Network Soc.) pp 577–81
- Wu X B, Baxter R A and Levy W B 1996 Context codes and the effect of noisy learning on a simplified hippocampal CA3 model *Biol. Cybern.* **74** 159–65