Capacity and information efficiency of the associative net

Bruce Graham and David Willshaw

Centre for Cognitive Science, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, Scotland, UK

Abstract.

Numerical calculations have been used to assess the performance of three different *winners-take-all* recall strategies for the associative net model of heteroassociative memory. Two strategies designed to improve recall when the net is partially connected or the input cues are noisy are shown to provide significantly greater capacity and information efficiency under these conditions. Estimates are made for the capacity of nets that are the size of the CA3 region of the rat hippocampus. These indicate that thousands of patterns can be stored, with the exact number highly dependent on pattern coding rates and the recall strategy. Analysis of nets with different types of structure shows that the capacity is much more sensitive to net size than to connectivity level. This has implications for neurobiology where there are examples of environmentally produced changes in both the size and connectivity of various brain regions.

Short title: Capacity and efficiency of the associative net

November 30, 1998

1. Introduction

The performance of a distributed associative memory can be judged on the basis of capacity and information efficiency. The capacity is the number of patterns that can be stored and recalled from the memory given an allowable error rate during recall. The information efficiency is the ratio of the amount of information that can be retrieved from the memory to the amount of storage available. These criteria provide different, but complementary, views of associative memory performance. A memory may be able to store a number of patterns of a given type, but be very inefficient in how the available storage is used. Only particular pattern types, such as very sparse patterns, can be stored if maximum efficiency is desired. Further, using the memory efficiently may involve an unacceptable degree of error during recall. In this paper we give results on the capacity and efficiency of the associative net model of heteroassociative memory (Willshaw *et al* 1969; Willshaw 1971) for varying pattern coding rates, connectivity and cue noise levels. This simple model consists of a set of input units connected to a set of output units by binary-valued synapses. Pairs of binary patterns are stored in the net using a 'clipped' Hebbian learning rule. Output patterns may be recalled from the net by using a previously stored input pattern as a cue. Theoretical analysis shows that a fully connected net can operate with 69% efficiency if the stored patterns are suitably sparse (Willshaw *et al* 1969). This drops to 24% if the connectivity is also sparse (Frolov and Murav'ev 1993).

In previous work we developed recall strategies based on a *winners-take-all* threshold that improve net performance when it is configured in a brain-like way, with partial connectivity and noisy input cues (Graham and Willshaw 1995b; Budinich *et al* 1996). Computer simulations were used to test these strategies when different numbers of patterns were stored and the recall cues were corrupted by different levels of noise. The strategies were designed to introduce minimal computational and informational overheads and we have argued for their possible biological implementation (Graham and Willshaw 1995b). Nonetheless their performance proved to be equivalent to that of the theoretically optimal strategy (Graham and Willshaw 1995b; Buckingham and Willshaw 1993; Buckingham 1991). However, capacity and information efficiency estimates are difficult to obtain with computer simulations due to the very large number of simulations that would be required. The size of net that can be simulated is also limited. Analytical results are only obtainable by making many approximations (Willshaw *et al* 1969; Palm 1980; Palm 1988; Nadal and Toulouse 1990; Frolov and Murav'ev 1993). Following the approach of Buckingham (Buckingham and Willshaw 1993; Buckingham and Willshaw 1992; Buckingham 1991), here we use more accurate numerical calculations to determine the capacity and efficiency over a wide range of net parameter values. This enables us to determine the relative merits of the different recall strategies and, in particular, the parameter regions in which the relative improvement due to the more complicated strategies is greatest. Preliminary results on the information efficiency of partially connected nets have been published elsewhere (Graham and Willshaw 1995a; Graham and Willshaw 1996).

The ability to calculate the capacity and information efficiency of this model of heteroassociative memory allows us to generate results that may be instructive from a neurobiological perspective. The hippocampal region of the mammalian cortex has long been implicated in short-term memory and learning. The architecture of different subregions in the hippocampus bears comparison with neural network models of associative memory (Marr 1971; Gardner-Medwin 1976; McNaughton and Morris 1987; Rolls 1989; Treves and Rolls 1994; Bennett *et al* 1994). Thus it is tempting to estimate the potential memory performance of the hippocampus on the basis of the performance of associative memory models. Here we give estimates of hippocampal capacity based on the associative net. The results are in broad agreement with data from other models (Treves and Rolls 1994; Bennett *et al* 1994) and indicate that in the order of thousands of patterns can be stored in the CA3 region of the rat hippocampus with high information efficiency. However, the exact capacity and efficiency are highly sensitive to net parameters such as the pattern coding rates, connectivity and the recall strategy.

Another issue of relevance to neurobiology is how best to construct an associative net to meet specific performance requirements. We examine the sensitivity of net capacity to changes in connectivity and net size. There are examples of environmentally-driven changes in the size and connectivity of various brain regions (Turner and Greenough 1985; Clayton and Krebs 1994). Our results show that the capacity of the associative net is much more sensitive to the number of units than to the specific connectivity.

2. The Associative Net

2.1. The model

The associative net is a simple two-layer feedforward neural network for performing heteroassociative memory. A set of N_{A} input units form connections with a set of N_{B} output units. The net may be either fully connected in which an output unit receives connections from all input units, or partially connected in which an output unit receives connections from a random fraction, Z, of the input units. The units have binary activity, with an active unit having an activation of one and an inactive unit having an activation of zero. Pairs of binary patterns are stored in the net using a 'clipped' Hebbian learning rule. The input patterns consist of M_{A} active input units and the output patterns consist of M_{B} active output units. Generally these patterns are sparse ($M_{\mathrm{A}}/N_{\mathrm{A}} = \alpha_{\mathrm{A}} \ll 1$; $M_{\mathrm{B}}/N_{\mathrm{B}} = \alpha_{\mathrm{B}} \ll 1$). The connection weights are unusual in that they are also binary. All weights are initially zero. A weight is changed to one during pattern storage if the input and output units are both active for the same pattern pair.

Once a set of pattern pairs have been stored in the net, an output pattern is recalled by using a previously stored input pattern as a retrieval cue. This cue may be a noisy version of the actual stored pattern, in which a fraction, s, of the M_{A} active units are spurious (active when they should be inactive). Recall consists of deciding which output units should be active on the basis of some measure of the input cue. We use particular *winners-take-all* recall strategies as explained in the next section. Output units that should be made active during recall will be referred to as *high* units. All other output units are *low* units.

2.2. *Winners-take-all recall*

During recall, every output unit measures the weighted sum of its inputs, or dendritic sum, d. Winners-take-all (*WTA*) recall involves applying an identical threshold to every output unit. The threshold is varied until the required number of output units (M_{B}) have a dendritic sum that is greater than the threshold. These units are then made active. We call this the *basic WTA* recall strategy. It is suitable for recall from a fully connected net with noise-free input cues. If the net is partially connected or the input cues are noisy then recall performance is degraded. Variations on this *basic* strategy are used to compensate for partial connectivity and noisy cues (Graham and Willshaw 1995b). During recall from a partially connected net, an output unit cannot distinguish between a missing connection and a synapse that was not modified during pattern storage. This is a source of noise that degrades recall. As was first suggested by Marr (1971), partial connectivity can be compensated for by setting the threshold to be the number of active inputs an output unit is actually connected to. We use this quantity, called the input activity, a, by applying the *WTA* threshold to a *normalized* dendritic sum which is the *basic* dendritic sum divided by the input activity, $d' = d/a$. The effect of noisy input cues can be lowered by minimizing the variance of the dendritic sums. We have done this by transforming the *normalized* dendritic sum by a function of the unit usage of each output unit which minimizes the variance of the *low* unit dendritic sums (Graham and Willshaw 1995b). The unit usage, r, is the number of stored patterns in which an output unit was active. The *transformed* dendritic sum is $d^* = 1 - (1 - d/a)^{1/r}$. This gives three recall strategies, depending on which form of dendritic sum the threshold is applied to: (a) *basic WTA* (b) *normalized WTA* and (c) *transformed WTA*. The performance of these strategies is examined and contrasted in this paper.

2.3. *Numerical calculations of net recall*

In order to determine the capacity and information efficiency of an associative net we first need to be able to determine the expected recall response of the net for a given input cue. The *WTA* recall response can be calculated using expressions for the probability distributions of the dendritic sums of *low* and *high* output units. These distributions differ for the *basic*, *normalized* and *transformed* sums. They also depend on the connectivity of the net and the noise level of the recall cue. Expressions for the distributions of each of these sums are given in Appendix A. For a net with a particular connectivity in which a given number of pattern pairs, R, have been stored, the *WTA* response to an input cue with a given noise level is calculated using the probability distributions by finding the threshold, T, that gives

$$(N_{\mathrm{B}} - M_{\mathrm{B}})P(d_{\mathrm{l}} \geq T) + M_{\mathrm{B}}P(d_{\mathrm{h}} \geq T) = M_{\mathrm{B}} \tag{1}$$

where $P(d_{\mathrm{l}} \geq T)$ ($P(d_{\mathrm{h}} \geq T)$) is the probability that a *low* (*high*) output unit has a dendritic sum greater than the threshold. The number of false positive and false negative errors in the response is given by

$$E = (N_{\mathrm{B}} - M_{\mathrm{B}})P(d_{\mathrm{l}} \geq T) + M_{\mathrm{B}}(1 - P(d_{\mathrm{h}} \geq T)) \tag{2}$$

The capacity of the net is defined to be the number of pattern pairs that can be stored before there is one bit in error in a recalled output pattern. We determine the capacity for the *WTA* recall strategy by calculating the recall response for different numbers of stored patterns, R, to find the minimum value of R for which a recall error occurs. The capacity, R_c, is one fewer than this value.

The information efficiency of pattern retrieval from the net is given by

$$\eta = \frac{R_c I}{Z N_A N_B} \quad \text{bits-per-synapse} \tag{3}$$

This is the ratio of the number of bits of information that can be retrieved from the net to the number of bits of storage available. The information in a single output pattern is

$$I = \log_2 \begin{pmatrix} N_B \\ M_B \end{pmatrix} \quad \text{bits} \tag{4}$$

As we only consider patterns recalled accurately, there is no loss in efficiency due to errors in an output pattern.

3. Results

The capacity and information efficiency of a very large associative net have been determined. The net has equal numbers of input and output units, with $N_A = N_B = 2^{18} = 262\,144$. This is of a similar order of magnitude to the recurrent collaterals of the CA3 region, or the projection from CA3 to CA1, in rat hippocampus (Amaral *et al* 1990; Boss *et al* 1987). The capacity and efficiency of this net vary with the pattern coding rates, connectivity, noise level in the input cue and the pattern recall strategy.

3.1. Fully connected net

Figure 1 shows the capacity (R_c) and information efficiency (η) of the fully connected net with noise-free input cues as functions of the input and output coding levels. All three *WTA* recall strategies are equivalent in this case. Both the capacity and efficiency are highly non-linear functions of the input coding rate, M_A. For a given output coding rate, M_B, both R_c and η reach maximum values when $M_A = \log_2(N_B) = 18$. For a given input coding rate, the capacity decreases exponentially with increasing $\log_2(M_B)$. This results in a linear decrease in information efficiency with increasing $\log_2(M_B)$, which can be approximated by

$$\eta = C\left(1 - \frac{\log_2(M_B)}{\log_2(N_B)}\right)\ln 2 \tag{5}$$

where $0 < C \leq 1$ is constant for a given M_A. Absolute maximum capacity and efficiency are achieved when $M_A = 18$ and $M_B = 8$, with $R_c = 313\,000\,000$ and $\eta = 0.58$ bits-per-synapse or 58%. At higher coding rates the capacity and efficiency drop to low levels. For example, $R_c = 340$ and $\eta = 0.0007$ when $M_A = M_B = 32\,768$ (1/8 of N_A, N_B).

3.2. Partially connected net

Figure 2 shows the effect of partial connectivity on the information efficiency when the input cues are noise-free and M_B is fixed at 1024. For noise-free cues, the *normalized*

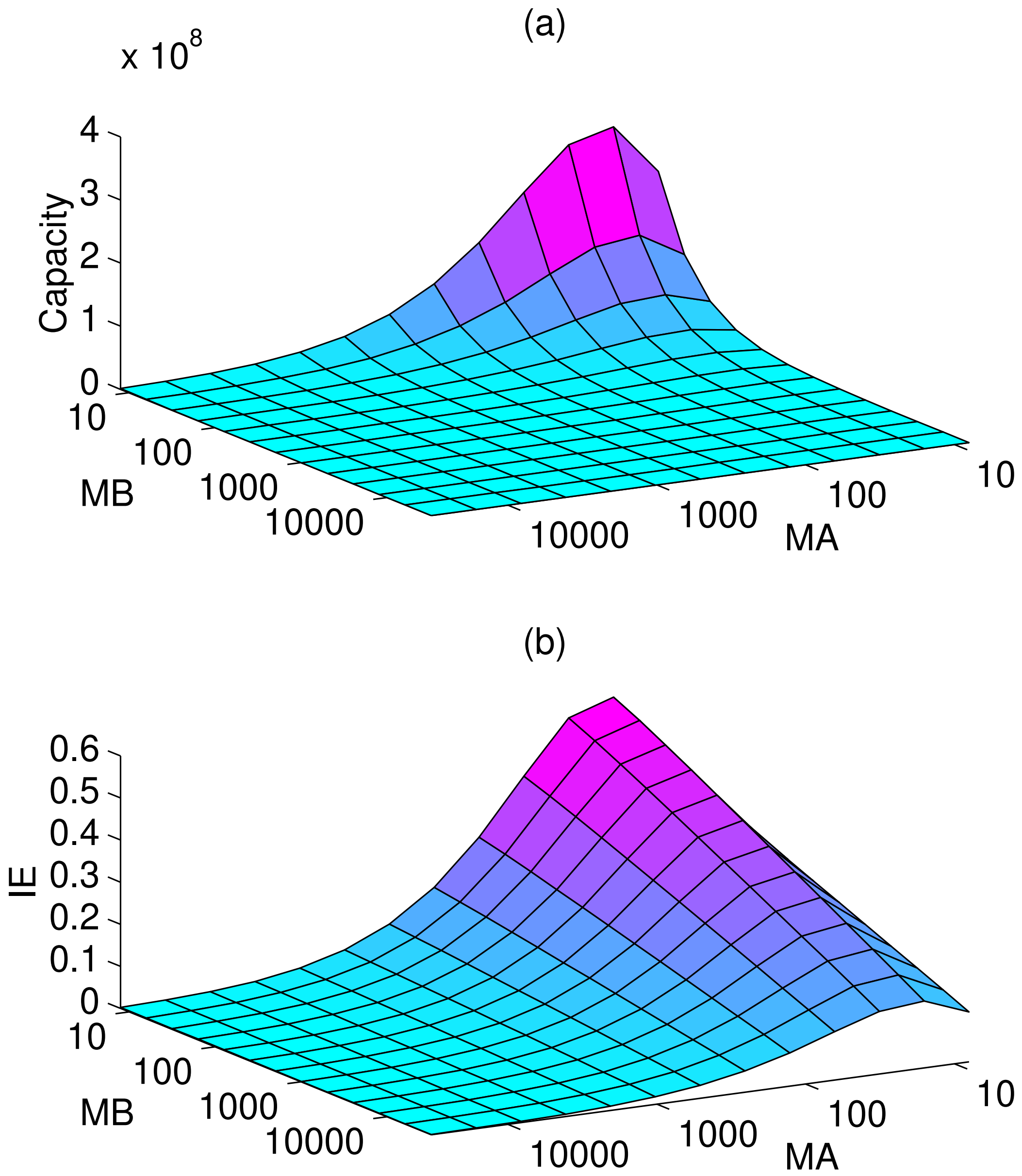

Figure 1. Noise-free input cues - (a) capacity and (b) information efficiency of the fully connected net as functions of input and output activity. In all figures $N_\mathrm{A} = N_\mathrm{B} = 2^{18}$.

and *transformed WTA* strategies are identical. When using either the *basic* or *normalized WTA* recall strategies for a given M_A, η reaches a maximum for a particular value of connectivity, Z, which may be less than one. The *normalized WTA* maintains a much higher efficiency over a range of values of M_A than does the *basic WTA*. Figure 3(a) shows the maximum value of η as a function of M_A for both strategies. Maximum efficiency ($\eta = 0.34$) is obtained for a fully connected net with $M_\mathrm{A} = 18$. The efficiency achieved by the *basic WTA* drops rapidly when M_A varies from this value. In contrast, the *normalized WTA* maintains near peak efficiency for a several orders of magnitude increase in M_A. The biggest difference between the maximum efficiencies of the two strategies is when $M_\mathrm{A} = 1024$ (Figure 3(b)). At this point the *normalized WTA* efficiency is still 95% of maximum ($\eta = 0.32$), while the *basic WTA* efficiency has dropped to 13% of maximum ($\eta = 0.04$). Figure 3(c) shows the connectivity level at which maximum efficiency is achieved for a given value of M_{A}. For the *normalized WTA*, maximum efficiency is achieved when $ZM_{\mathrm{A}} \approx \log_2(N_{\mathrm{B}}) = 18$. For low values of M_{A}, the *basic WTA* is most efficient with a fully connected net. At higher values of M_{A}, a partially connected net is more efficient, with the optimum connectivity being given by $ZM_{\mathrm{A}} \approx 100$.

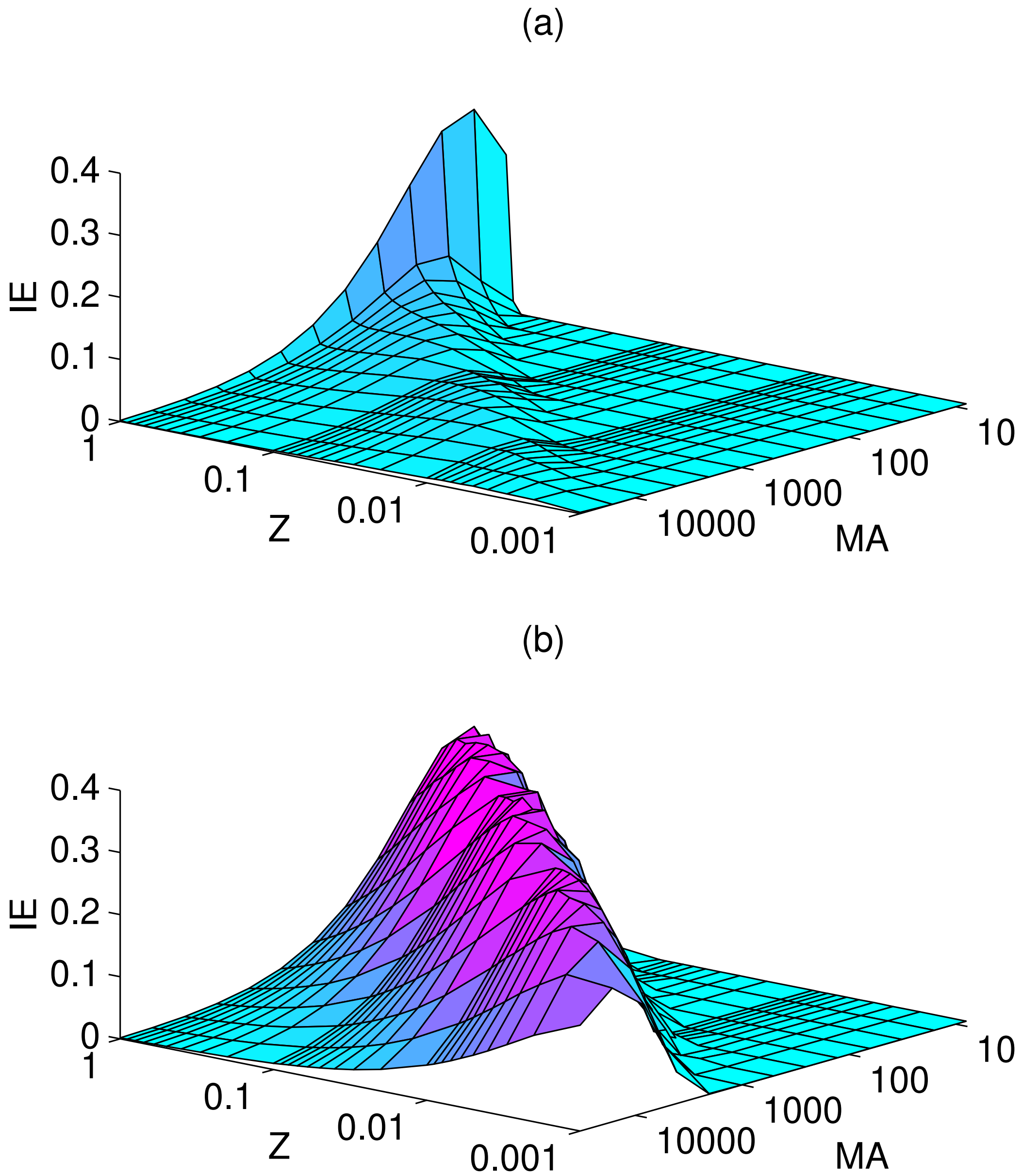

Figure 2. Noise-free input cues - information efficiency as a function of connectivity and input activity (a) *basic WTA* (b) *normalized WTA* ($M_{\mathrm{B}} = 1024$).

Similar results are obtained when the input cues are noisy, but now the performance of the *normalized* and *transformed WTA* strategies is different. Figure 4 shows the information efficiency of the net obtained by the three recall strategies when the input cues are corrupted by 40% noise ($s = 0.4$) and $M_{\mathrm{B}} = 1024$. Peak efficiency is now $\eta = 0.04$, obtained for the fully connected net when $M_{\mathrm{A}} = 128$. The efficiency of all three strategies drops rapidly to zero for $M_{\mathrm{A}} < 128$, as shown in Figure 5(a). At higher values of M_{A}, both the *normalized* and *transformed WTA* provide a much higher maximum efficiency

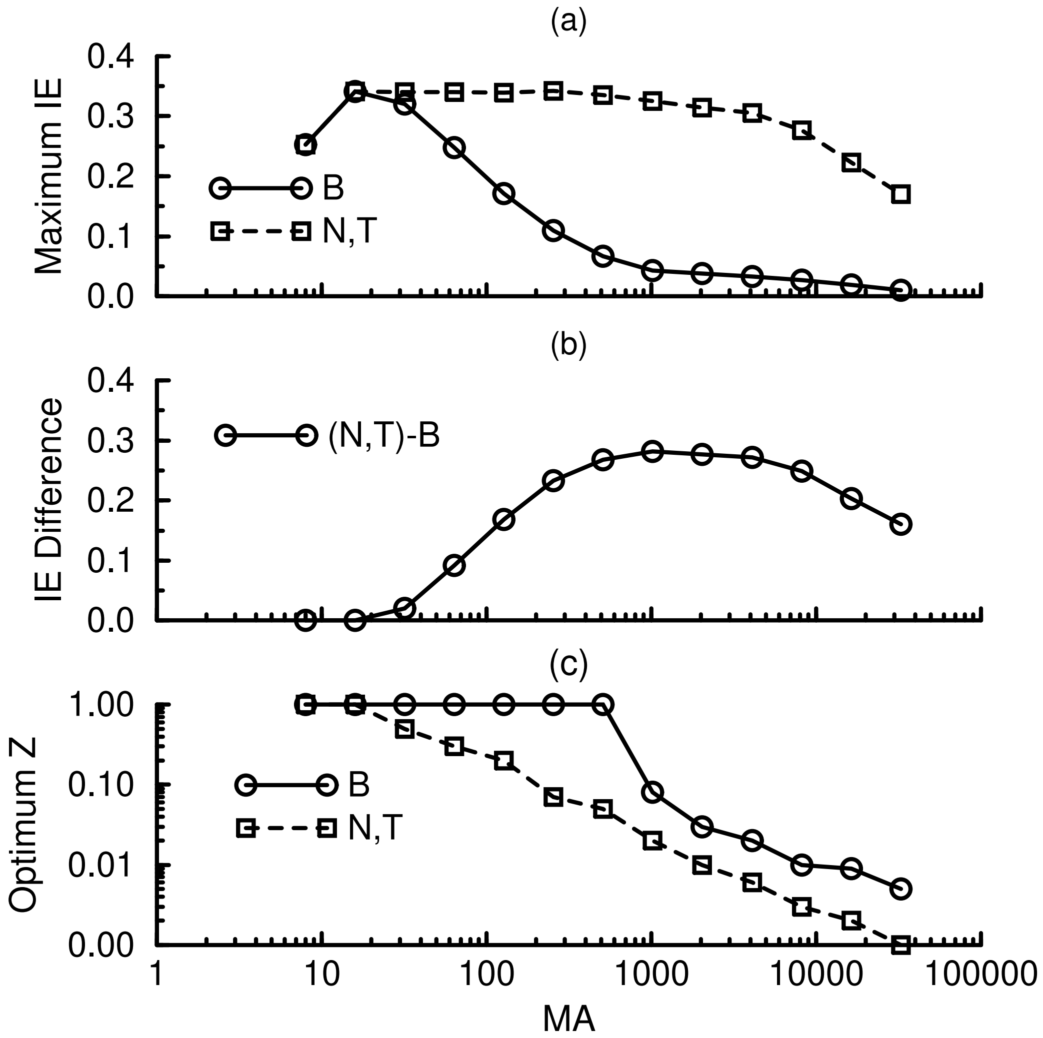

Figure 3. (a) Maximum information efficiency as a function of input activity. (b) Difference in efficiency between the *normalized (N)* or *transformed (T) WTA* and the *basic (B) WTA*. (c) Value of the connectivity (Z) at maximal efficiency.

than the *basic WTA*. The maximum difference in efficiencies between the *normalized* and *basic WTA* is again when $M_{\mathrm{A}} = 1024$ (Figure 5(b)). The maximum difference between the *transformed* and *basic WTA* occurs at $M_{\mathrm{A}} = 2048$, and between the *transformed* and *normalized WTA* at $M_{\mathrm{A}} = 8192$. Maximum efficiency is now obtained for the *normalized WTA* when $ZM_{\mathrm{A}} \approx 100$ (Figure 5(c)). The *basic WTA* is most efficient for a fully connected net, except at high values of M_{A} where the optimum connectivity is similar to that of the *normalized WTA*. For low values of M_{A} the *transformed WTA* has the same optimum connectivity as the *normalized WTA*. At high values of M_{A}, the *transformed WTA* is most efficient at slightly higher connectivity than the *normalized WTA*.

3.3. Variation in pattern coding rates

Figure 6 compares the efficiency of the *transformed* and *basic WTA* as a function of either input or output coding rate for a fully connected net with 40% noise in the cue. In this situation the behaviour of the *normalized WTA* is identical to the *basic WTA*. For $M_{\mathrm{B}} = 1024$, the *transformed WTA* outperforms the *basic WTA* only for moderate values

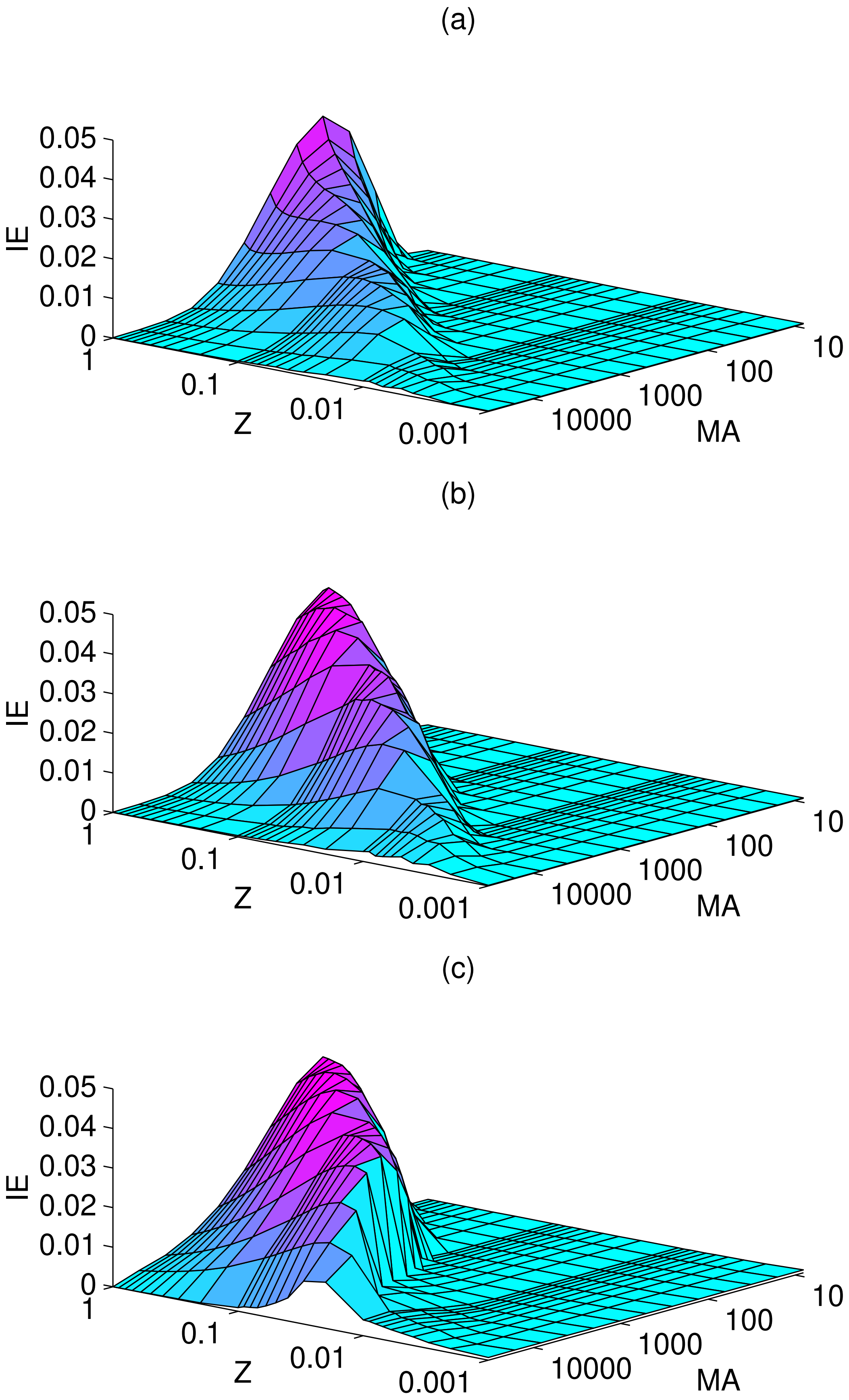

Figure 4. Noisy input cues - information efficiency as a function of connectivity and input activity for 40% noise (a) *basic WTA* (b) *normalized WTA* (c) *transformed WTA* ($M_{\mathrm{B}} = 1024$).

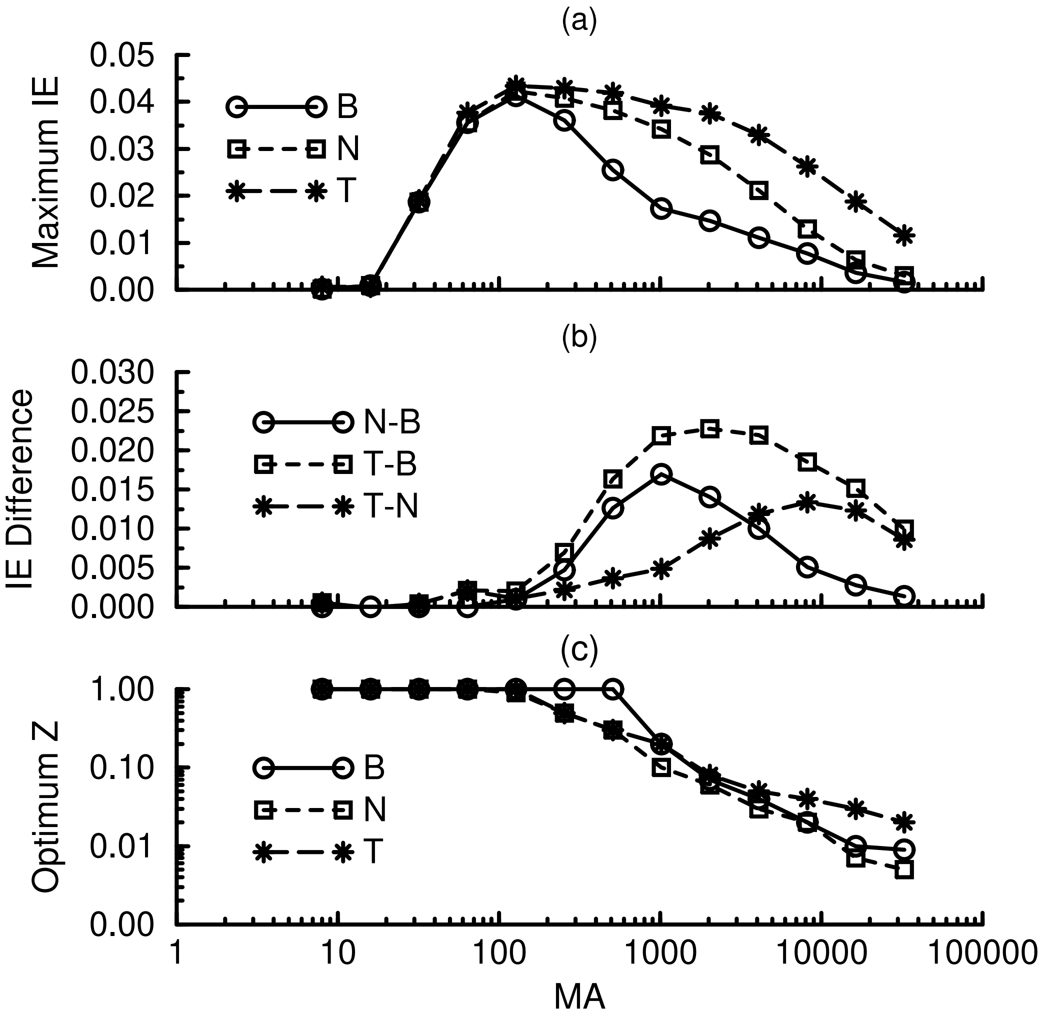

Figure 5. Maximum information efficiency as a function of input activity. As for Figure 3 but for 40% noise.

of M_A, with the maximum difference in efficiency occuring when $M_\mathrm{A} = 2048$ (Figure 6(c)). For $M_\mathrm{A} = 1024$, the efficiency of both strategies decreases with increasing M_B, and there is a corresponding decrease in the difference in efficiency. In relative terms, however, the *transformed WTA* is consistently 1.3 times more efficient than the *basic WTA* over the range of M_B values. The constant relative improvement depends on M_A, and rises to a maximum of about six times when $M_\mathrm{A} = 8192$ (data not shown).

3.4. Noisy input cues

In Figure 7 the performance of the three strategies is compared over a range of connectivities and noise levels when $M_\mathrm{A} = 8192$ and $M_\mathrm{B} = 1024$. For connectivities greater than about 0.03, the *transformed WTA* provides greater capacity and information efficiency over a range of noise levels than the other two strategies. At a given connectivity in this range, the capacity, and hence the information efficiency, of the *transformed WTA* decreases approximately linearly with increasing noise, whilst it decreases exponentially for the other strategies. At very low connectivities the capacity of the *transformed WTA* collapses to near zero, and the performance is actually worse than the *normalized WTA*.

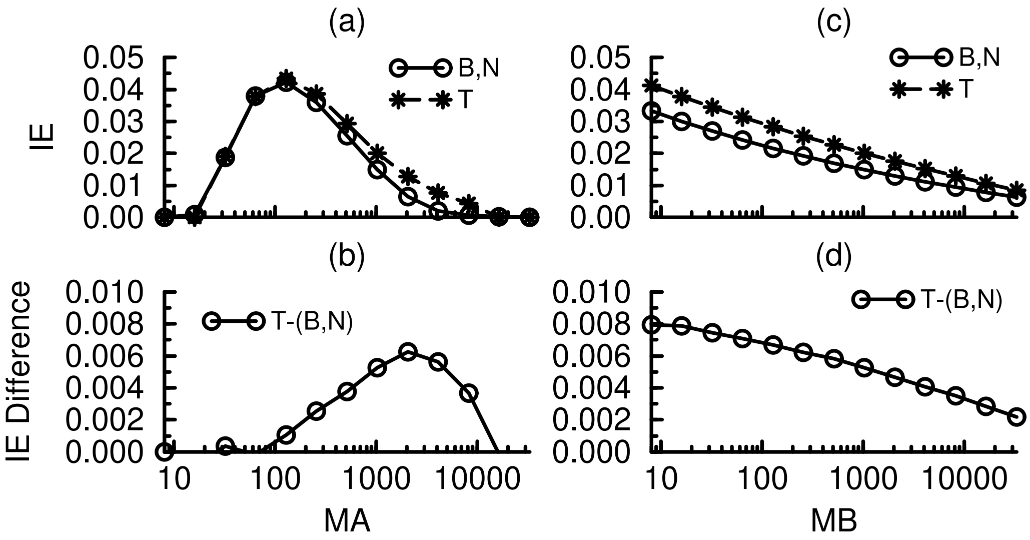

Figure 6. Noisy cues - information efficiency as a function of input and output activity levels for the fully connected net with 40% noise ((a),(b) $M_{\mathrm{B}} = 1024$ (c),(d) $M_{\mathrm{A}} = 1024$).

The maximum efficiency that can be achieved at any noise level is shown in Figure 8. At low noise levels, the *normalized* and *transformed WTA* are much more efficient than the *basic WTA*. At high noise levels the efficiency of the *normalized WTA* drops to that of the *basic WTA*, while the *transformed WTA* maintains slightly higher performance (Figure 8(a),(b)). In relative terms, the *transformed WTA* is four to five times more efficient than the *basic WTA*. At 40% noise the *transformed WTA* is twice as efficient as the *normalized WTA*, rising to nearly four times more efficient at 60% noise. Figure 8(c) shows the connectivity required to achieve the maximum efficiency at each noise level. For all three strategies this optimum connectivity increases as the noise level increases, with the *transformed WTA* requiring the highest connectivity at high noise. The *normalized WTA* is consistently most efficient at a lower connectivity than the other strategies.

4. Discussion

4.1. Performance of recall strategies

The main aim of this study has been to compare the performance of three *WTA* recall strategies. The *basic WTA* is the simplest implementation of a *WTA* recall strategy and provides a baseline against which to compare the more complicated *normalized* and *transformed WTA* strategies.

The *normalized WTA* is designed to improve the recall quality from partially connected nets. The results here show that it always provides greater capacity and information efficiency than the *basic WTA* when a net is partially connected. When the recall cue is noise-free, any reduction below full connectivity provides great advantage to the *normalized WTA*. For example, Figure 2 shows that though the two strategies are equivalent at full connectivity, with a maximum efficiency of 0.34, at 90% connectivity ($Z = 0.9$) the *normalized WTA* still has a maximum efficiency of 0.33, while the *basic WTA* has dropped to $\eta = 0.12$. The advantage increases with decreasing connectivity,

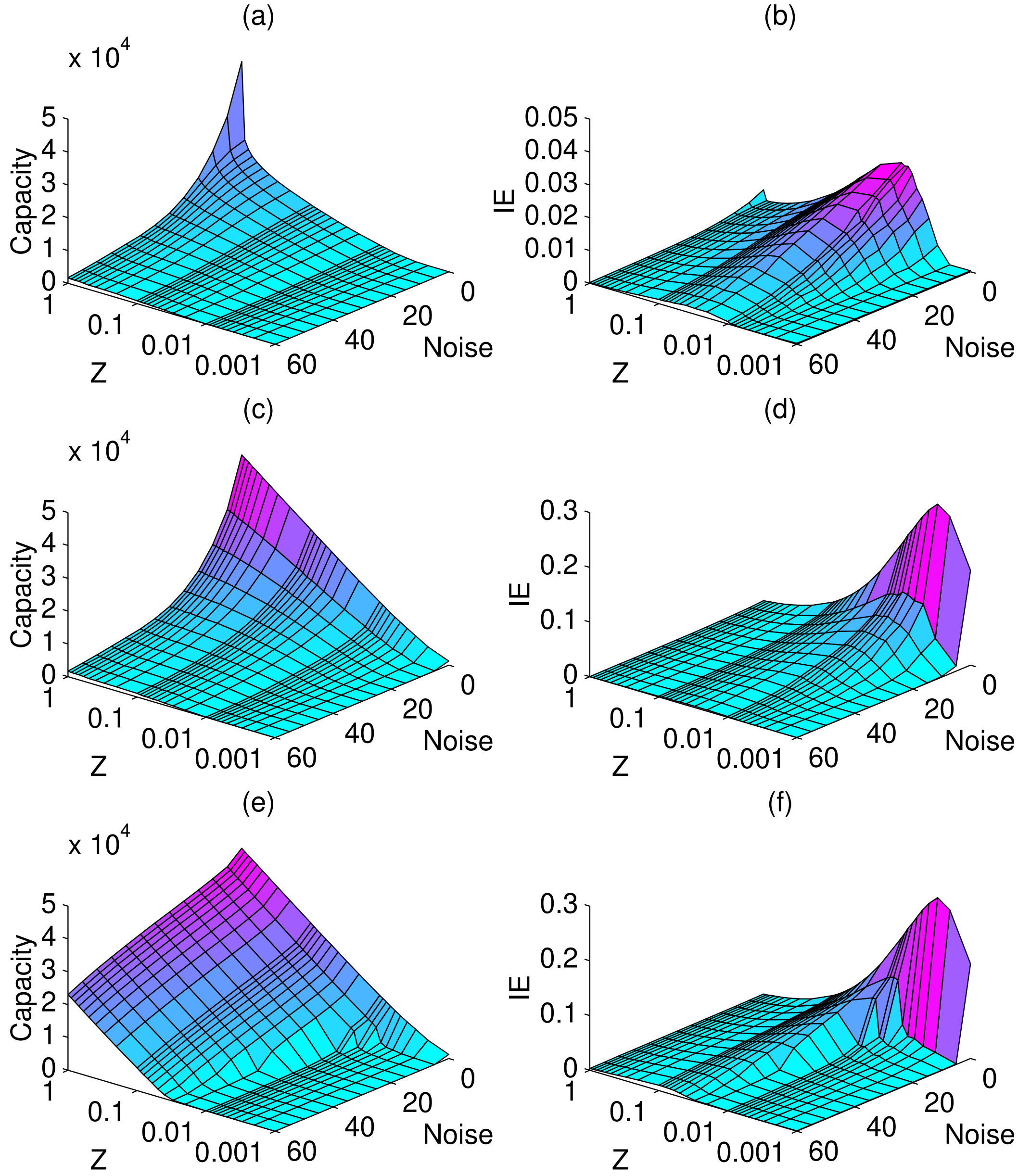

Figure 7. Capacity and information efficiency as a function of connectivity and input cue noise for specific pattern codings (a),(b) *basic WTA* (c),(d) *normalized WTA* (e),(f) *transformed WTA* ($M_{\rm A} = 8192$, $M_{\rm B} = 1024$).

so that at 10% connectivity the *normalized WTA* has a maximum efficiency of 0.337, greater than at 90% connectivity, compared to 0.043 for the *basic WTA*. Only at very low connectivity, when the performance of both strategies is poor, does the efficiency of the *normalized WTA* approach that of the *basic WTA*.

When the input cues are noisy, the relative advantage of the *normalized WTA* at near full connectivity is reduced. With 40% cue noise and 90% connectivity the *normalized WTA* has a maximum efficiency of 0.042, compared to 0.036 for the *basic WTA* (Figure 4). At 10% connectivity the maximum *normalized WTA* efficiency is 0.034, while the maximum *basic WTA* efficiency is 0.016. So the relative advantage again increases with decreasing connectivity. However, at high cue noise levels ($> 50\%$) the advantage

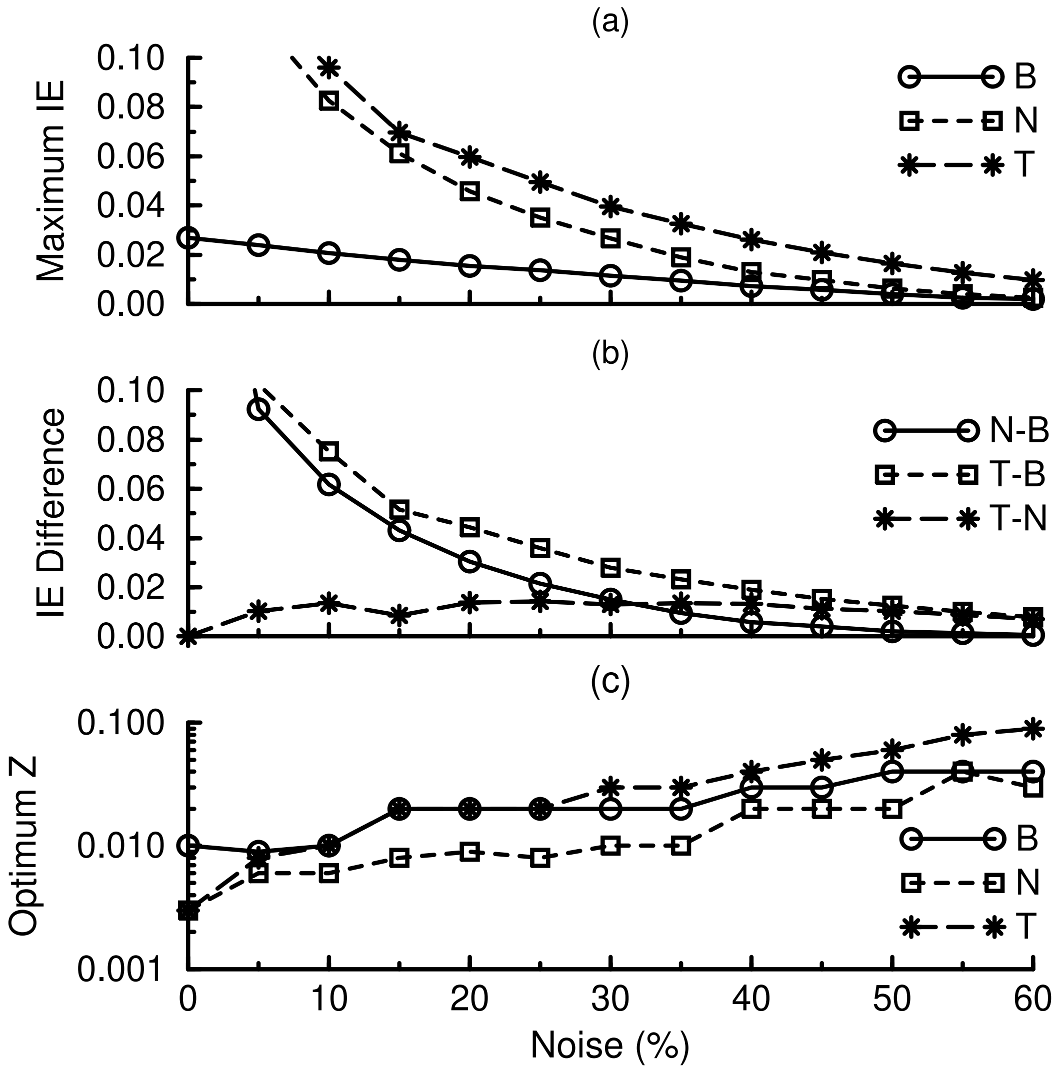

Figure 8. Maximum information efficiency as a function of input cue noise ($M_{\mathrm{A}} = 8192$, $M_{\mathrm{B}} = 1024$).

of the *normalized WTA* is virtually abolished and both strategies perform very poorly (Figure 8), regardless of connectivity.

The *transformed WTA* is a variation on the *normalized WTA* to improve recall when the input cues are noisy. However, its performance is dependent on the net configuration. For example, at full connectivity but with 40% cue noise, the *transformed WTA* outperforms the *basic WTA* only for $128 \leq M_{\mathrm{A}} \leq 8192$ (Figure 6). At $M_{\mathrm{A}} = 8192$ the *transformed WTA* has an efficiency of 0.0043 ($R_{\mathrm{c}} = 30\,600$), seven times the *basic WTA* efficiency of 0.0006 ($R_{\mathrm{c}} = 4600$). When connectivity is varied at this input coding rate, the *transformed WTA* has a maximum efficiency of 0.026, compared with 0.013 for the *normalized WTA* and 0.008 for the *basic WTA* (Figure 5). The relative advantage of the *transformed WTA* is small for low noise levels. For the previous situation, but with 10% noise, the maximum *transformed WTA* efficiency is 0.096, while for the *normalized WTA* it is 0.082 (Figure 8). These maxima are achieved at around 10% connectivity and both strategies significantly outperform the *basic WTA* which has a maximum efficiency of 0.021.

At connectivities less than about 5% the performance of the *transformed WTA* actually can be worse than that of the *normalized WTA*, due to the non-optimum nature of the transform employed (Figures 4 and 7). The transform is designed to minimize the variance of the dendritic sums of *low* output units by making the mean dendritic sum independent of unit usage. However, the mean dendritic sum of the *high* units remains a function of both unit usage and cue noise and the variance of these sums may be increased.

4.2. Maximum information efficiency

Theoretical analysis indicates that the associative net can be highly information efficient, but only with very sparse patterns. For a fully connected net with noise-free input cues, an efficiency of 0.69 bits-per-synapse theoretically can be achieved when the input coding rate is given by $M_{\mathrm{A}} = \log_2(N_{\mathrm{B}})$ (Willshaw *et al* 1969; Willshaw 1971; Palm 1988; Nadal and Toulouse 1990; Buckingham and Willshaw 1992). Frolov and Murav'ev (1993) have shown that if the information due to the recognition of an input cue as familiar is taken into account, maximum total efficiency can be maintained for values of M_{A} greater than $\log_2(N_{\mathrm{B}})$. However, the information extracted about output patterns decreases. We only consider this form of information here.

This level of efficiency is difficult to reach in practice. For the net studied here, an efficiency of 0.58 is achieved when $M_{\mathrm{A}} = \log_2(N_{\mathrm{B}}) = 18$ ($\alpha_{\mathrm{A}} = 7 \times 10^{-5}$) and $M_{\mathrm{B}} = 8$ ($\alpha_{\mathrm{B}} = 3 \times 10^{-5}$) (Figure 1). The output patterns must be even more sparse if the theoretical maximum is to be approached. Though the numerical calculations used here confirm this, strictly they are only valid for $M_{\mathrm{B}} \gg 1$. The efficiency drops rapidly with increasing coding rates and is, for example, only 0.038 for $M_{\mathrm{A}} = 2048$ ($\alpha_{\mathrm{A}} = 0.0078$) and $M_{\mathrm{B}} = 8$. Efficiency also drops for $M_{\mathrm{A}} < \log_2(N_{\mathrm{B}})$. It has been shown elsewhere (Nadal and Toulouse 1990; Buckingham and Willshaw 1992) that an efficiency of 0.69 can theoretically still be achieved in this coding region if errors are allowed during recall. We will restrict our discussion, however, to the efficiency of error-free recall, on the basis that high memory performance involves both accurate recall and efficient use of storage.

At low connectivity levels and for recall based purely on the *basic* dendritic sums, an efficiency of only 0.24 theoretically can be achieved (Frolov and Murav'ev 1993). For connectivity levels between 1% and 10%, the efficiency of the *basic WTA* is insensitive to the actual connectivity. By varying the input coding rate, a maximum efficiency of 0.1 is achieved for $Z = 0.01$, rising to 0.13 for $Z = 0.1$, when $M_{\mathrm{B}} = 8$. The maximum efficiency decreases for connectivities below 1%. The optimum input coding rate is given by $ZM_{\mathrm{A}} \approx 38$. The approximate theoretical analysis indicates that this optimum should be $ZM_{\mathrm{A}} \approx \log_2(N_{\mathrm{B}})$ (Frolov and Murav'ev 1993). Of the recall strategies used here, this is only true for the *normalized WTA*.

Using *normalized WTA* recall, and given noise-free input cues, maximum efficiency at any input coding rate greater than $\log_2(N_{\mathrm{B}})$ is achieved when the connectivity satisfies $ZM_{\mathrm{A}} \approx \log_2(N_{\mathrm{B}})$ (Figures 2 and 3). This relationship may be derived in a manner analogous to the fully connected case, as shown in Appendix B. In general, regardless of the recall strategy and the cue noise level, maximum efficiency is achieved when $ZM_{\mathrm{A}} \approx C \geq \log_2(N_{\mathrm{B}})$, where C is a constant. The value of C depends on the recall strategy and the cue noise level, and it increases with increasing noise. These results quantify to some extent the intuition expressed previously that an output unit needs to connect to a reasonable number of active input units, say 20, to guarantee accurate and efficient recall (Marr 1971; Gardner-Medwin 1976; Buckingham 1991). This number must increase when the cues are noisy so that an output unit still connects to a reasonable number of correctly active input units (Willshaw and Buckingham 1990). In our situation, the spuriously active input units act to decrease recall performance, so that at 50% noise, the *normalized WTA* achieves maximal efficiency with seven times the connectivity required for noise-free cues, compared with twice the connectivity needed to simply maintain the number of connections to correctly active inputs (Figure 8).

4.3. Implications for hippocampal capacity

The mammalian hippocampus is a possible site of short-term memory storage, and its architecture bears comparison with neural network models of associative memory (Marr 1971; Gardner-Medwin 1976; McNaughton and Morris 1987; Rolls 1989; Treves and Rolls 1994; Bennett *et al* 1994). The projections from entorhinal cortex to dentate gyrus and from CA3 to CA1 could store patterns in an heteroassociative fashion through modification of synaptic strengths (McNaughton and Morris 1987; Treves and Rolls 1994). Similarly, the recurrent collaterals of CA3 could form an autoassociative memory (Treves and Rolls 1994; Bennett *et al* 1994). It is interesting to speculate about the possible capacity and efficiency of the hippocampus on the basis of the performance of associative memory models. Marr (1971) designed a three-layered network that could store 100 000 patterns and accurately recall them with 10% partial cues. This model is based only loosely on hippocampal structure and parameters (Willshaw and Buckingham 1990; Buckingham 1991). A version that more closely matches data from the rat hippocampus results in a capacity of between 5000–14 000 patterns when using 10% partial cues (Buckingham 1991). Other estimates have been made for the autoassociative memory capacity of the CA3 region (Treves and Rolls 1994; Bennett *et al* 1994). In the Sprague-Dawley rat this region contains approximately 330 000 cells (Amaral *et al* 1990; Boss *et al* 1987). The 'clipped' Hebbian learning rule and a particular thresholding strategy results in 340 000 patterns being stable states of a net of this size when it has 5% connectivity with some spatial bias and the patterns contain 330 active units ($\alpha = 0.1\%$) (Bennett *et al* 1994). For a rather different model employing covariance Hebbian learning, with pattern coding $\alpha = 2\%$ and approximately 2% connectivity distributed uniformly across both halves of the hippocampus ($\approx 600\,000$ cells), this reduces to 36 000 retrievable patterns (Treves and Rolls 1991; Treves and Rolls 1994).

The capacity of a square heteroassociative net gives an estimate of the number of stable patterns that can be stored in an autoassociative net of the size of a single layer of the heteroassociative net. Table 1 compares the results of Bennett *et al* (1994) and Treves and Rolls (1994) with the capacities of associative nets using *basic* or *normalized WTA* recall when $N = N_{\mathrm{A}} = N_{\mathrm{B}}$ and $M = M_{\mathrm{A}} = M_{\mathrm{B}}$. Though the capacities are comparable they are clearly dependent on the net configuration and the recall strategy. The first configuration ($N = 330\,000$; $M = 330$; $Z = 0.05$) is nearly optimal for *normalized WTA* recall ($ZM_{\mathrm{A}} \approx \log_2(N_{\mathrm{B}})$) with an efficiency of 0.42. The size and connectivity of the second net requires much sparser pattern coding ($M = 950$) than that used by Treves and Rolls (1994) to achieve a maximum efficiency of 0.36 for *normalized WTA* recall.

Table 1. Comparison of capacities from different models.

N	M	Z	Bennett *et al*	Treves & Rolls	Basic WTA	Norm WTA
330 000	330	0.05	340 000	—	19 000	600 000
600 000	12 000	0.02	—	36 000	1 000	6 000
600 000	950	0.02	—	—	—	260 000

Given the sensitivity of the capacity to the exact configuration and operation of the associative memory model, all that can be said is that these models indicate that a brain area such as the CA3 region of the rat hippocampus can store thousands of patterns with high efficiency. It is not possible to be more specific than this without more exact data about the structure of the hippocampus, the biological plausibility of recall strategies (Graham and Willshaw 1995b), and, in particular, the activity levels in the hippocampus (Buckingham 1991; Treves and Rolls 1991; Bennett *et al* 1994).

4.4. Designing an associative net

In designing an associative memory, it is relevant to ask how much information is required to be stored, and how much does storage cost? Specifically, for patterns with a given information content, what size of net is necessary and what connectivity level should be employed to store the total amount of information required? The smaller the net and the lower the connectivity level, the lower the material cost in building the net. This has implications for brain function where evidence exists for structural modifications in response to a need to absorb and remember more information. For example, an increase of some 20–25% in the number of synapses per neuron in rat occipital cortex was measured for rats reared in a complex environment compared to rats reared in a very simple environment (Turner and Greenough 1985). This may indicate an increase in connectivity in this region of the rat brain due to the need to cope with greater informational input. In food-storing birds, an increase in the relative volume of the hippocampus, compared to overall brain volume, was measured for birds with food storing experience compared to birds denied such experience (Clayton and Krebs 1994). These experiments with marsh-tits showed a relative increase of some 28% in the volume of the hippocampus, largely due to an increase in the number of cells. There is no information about changes in the number of synapses per neuron.

The associative net has varying degrees of sensitivity to such structural changes. The effect of changing connectivity level depends very much on the operating regime and the recall strategy. Except at very low connectivity, the three recall strategies used here are quite insensitive to changes in connectivity, with the *normalized WTA* being specifically designed to ameliorate the effect of partial connectivity. Consider the square net with $N_{\mathrm{A}} = N_{\mathrm{B}} = 262\,144$, $M_{\mathrm{A}} = M_{\mathrm{B}} = 4096$ and noise-free input cues. At physiological levels of around 5% connectivity, a 20% change in connectivity only yields in the order of a 10% change in capacity. This rises to around 20% at 0.5% connectivity. Only at much lower connectivity does the sensitivity increase, but this is a region of very low capacity. So altering the connectivity is not an effective way of changing the capacity of an associative net. It is perhaps more likely that the increase in the number of synapses seen biologically reflects an increase in the number of connections between particular neurons, thus providing increased synaptic strength. Such changes have been associated with the induction of LTP in the mammalian hippocampus (Lee *et al* 1980; Chang and Greenough 1984) and correspond to an incremental Hebbian learning rule, as opposed to the 'clipped' Hebbian rule used here.

Generally the capacity is more sensitive to net size than to connectivity. Suppose patterns with the same information content are stored in nets of different size (the pattern coding rate is lowered as net size is increased to achieve this). When 5% connectivity is maintained, a 20% change in size of the above net yields a 120% change in capacity for the *basic* and *normalized WTA* strategies. If the number of synapses is kept constant by altering the connectivity, for a 20% change in net size the *basic WTA* still gives a 75% change in capacity, while the *normalized WTA* gives a 90% change. On this basis, a 28% increase in hippocampal volume could provide well over twice the information storage capacity, if the increase in volume equates with an increase in net size. Note that in both cases there is an optimum net size after which the capacity and information efficiency will actually drop as the net size is increased further. This is because ZM_{A} is decreasing while $\log_2(N_{\mathrm{B}})$ is increasing, so that eventually the optimum point of $ZM_{\mathrm{A}} = \log_2(N_{\mathrm{B}})$ is passed. If 5% connectivity is maintained this will happen for an extremely large net. If connectivity is lowered to maintain the number of synapses the capacity reaches a maximum for a net size of $N_{\mathrm{A}} = N_{\mathrm{B}} = 2^{22} = 4\,194\,304$.

5. Conclusions

Using numerical calculations of associative net recall we have quantified the performance of three different *WTA* recall strategies. For a partially connected net, the *normalized WTA* can provide many times the capacity and efficiency of the *basic WTA*. For moderate pattern coding rates and net connectivity, the *transformed WTA* is four to five times more efficient than the *basic WTA* and two to four times more efficient than the *normalized WTA*. The *transformed WTA* is computationally more expensive than the other strategies. However, even allowing one bit per synapse extra storage for the measurement of input activity and another bit for unit usage, the *transformed WTA* can provide much greater capacity at a similar efficiency.

Acknowledgments

To the Medical Research Council for financial support under Programme grant PG 9119632.

References

Amaral D G, Ishizuka N and Claiborne B 1990 Neurons, numbers and the hippocampal network. *Progress in Brain Research* vol 83 ed J Storm-Mathisen *et al* (Amsterdam: Elsevier) chapter 1 pp 1–11

Bennett M, Gibson W and Robinson, J 1994 Dynamics of the CA3 pyramidal neuron autoassociative memory network in the hippocampus *Phil. Trans. Roy. Soc. Lond. B* **343** 167–87

Boss B, Turlejski K, Stanfield B and Cowan W 1987 On the numbers of neurons in fields CA1 and CA3 of the hippocampus of Sprague-Dawley and Wistar rats *Brain Res.* **406** 280–87

Buckingham J 1991 *Delicate nets, faint recollections: a study of partially connected associative network memories* Ph.D. thesis, University of Edinburgh

Buckingham J and Willshaw D 1992 Performance characteristics of the associative net *Network* **3** 407–14

——1993 On setting unit thresholds in an incompletely connected associative net *Network* **4** 441–59

Budinich M, Graham B. and Willshaw D 1996 Multiple cueing of an associative net *Int. J. of Neural Networks*, to appear

Chang F-L and Greenough W 1984 Transient and enduring morphological correlates of synaptic activity and efficacy change in the rat hippocampal slice *Brain Res.* **309** 35–46

Clayton N and Krebs J 1994 Hippocampal growth and attrition in birds affected by experience *Proc. Nat. Acad. Sci.* **91** 7410–14

Frolov A and Murav'ev I 1993 Informational characteristics of neural networks capable of associative learning based on Hebbian plasticity *Network* **4** 495–536

Gardner-Medwin A 1976 The recall of events through the learning of associations between their parts *Proc. Roy. Soc. Lond. B* **194** 375–402

Graham B and Willshaw D 1995a Capacity and information efficiency of a brain-like associative net *Neural Information Processing Systems 7* ed G Tesauro *et al* (Cambridge: MIT Press) pp 513–20

——1995b Improving recall from an associative memory *Biol. Cybern.* **72** 337–46

——1996 Information efficiency of the associative net at arbitrary coding rates *Proc. of ICANN96 (Bochum)*, to appear

Lee K, Schottler F, Oliver M and Lynch G 1980 Brief bursts of high-frequency stimulation produce two types of structural change in rat hippocampus *J. Neurophys.* **44** 247–58

Marr D 1971 Simple memory: a theory for archicortex *Phil. Trans. Roy. Soc. Lond. B* **262** 23–81

McNaughton B and Morris R 1987 Hippocampal synaptic enhancement and information storage within a distributed memory system *TINS* **10** 408–15

Nadal J-P and Toulouse G 1990 Information storage in sparsely coded memory nets *Network* **1** 61–74

Palm G 1980 On associative memory *Biol. Cybern.* **36** 19–31

——1988 On the asymptotic information storage capacity of neural networks *Neural Computers* vol F41 of *NATO ASI*, ed R Eckmiller and Ch v d Malsburg (Berlin: Springer-Verlag) pp 271–80

Rolls E 1989 The representation and storage of information in neural networks in the primate cerebral cortex and hippocampus *The Computing Neuron* ed R Durbin *et al* (Wokingham: Addison-Wesley) pp 125–59

Treves A and Rolls E 1991 What determines the capacity of autoassociative memories in the brain? *Network* **2** 371–97

——1994 Computational analysis of the role of the hippocampus in memory *Hippocampus* **4** 374–91

Turner A and Greenough W 1985 Differential rearing effects on rat visual cortex synapses. I. Synaptic and neuronal density and synapses per neuron *Brain Res.* **329** 195–203

Willshaw D 1971 *Models of distributed associative memory* Ph.D. thesis, University of Edinburgh

Willshaw D and Buckingham J 1990 An assessment of Marr's theory of the hippocampus as a temporary memory store *Phil. Trans. Roy. Soc. Lond. B* **329** 205–15

Willshaw D, Buneman O and Longuet-Higgins H 1969 Non-holographic associative memory *Nature* **222** 960–62

Appendix A. Probability Distributions of Dendritic Sums

The *WTA* recall response can be calculated numerically using expressions for the distributions of the dendritic sums of *low* and *high* output units. This appendix gives details of the probability distributions of the *basic*, *normalized* and *transformed* dendritic sums. It is assumed that all patterns are sparse and the activity of individual units is approximately independent ($1 \ll M_\mathrm{A} \ll N_\mathrm{A}$ and $1 \ll M_\mathrm{B} \ll N_\mathrm{B}$). Also, the number of stored pattern pairs must be much smaller than the possible number of independent pattern pairs.

Appendix A.1. Basic WTA

The probability that the *basic* dendritic sum of a *low* or *high* output unit should have a particular value x is (Buckingham and Willshaw 1993; Buckingham 1991; Graham and Willshaw 1995a)

$$P(d_\mathrm{l} = x) = \sum_{r=1}^{R} \binom{R}{r} \alpha_\mathrm{B}^r (1-\alpha_\mathrm{B})^{R-r} \binom{M_\mathrm{A}}{x} (Z\rho[r])^x (1 - Z\rho[r])^{M_\mathrm{A}-x} \tag{A1}$$

$$P(d_\mathrm{h} = x) = \sum_{r=0}^{R-1} \binom{R}{r} \alpha_\mathrm{B}^r (1-\alpha_\mathrm{B})^{R-r} \binom{M_\mathrm{A}}{x} (Z\mu[r+1])^x (1 - Z\mu[r+1])^{M_\mathrm{A}-x} \tag{A2}$$

where $\rho[r]$ and $\mu[r]$ are the probabilities that an arbitrarily selected active input is on a connection with weight 1. For a *low* unit,

$$\rho[r] = 1 - (1-\alpha_\mathrm{A})^r \tag{A3}$$

For a *high* unit a good approximation for μ is

$$\mu[r+1] \simeq g + s\rho[r] = 1 - s(1-\alpha_\mathrm{A})^r \tag{A4}$$

where g and s are the probabilities that a particular active input in the cue pattern is *genuine* (belongs to the stored pattern) or *spurious*, respectively ($g + s = 1$) (Buckingham and Willshaw 1993).

Appendix A.2. Normalized WTA

A *normalized* dendritic sum is $d' = d/a$. The distributions of *normalized* sums can be approximated by the *basic* distributions for the situation where every unit has the mean input activity, $a_\mathrm{m} = M_\mathrm{A} Z$. In this case the *low* and *high* unit *normalized* distributions are given by

$$P(d'_\mathrm{l} = x') = \sum_{r=1}^{R} \binom{R}{r} \alpha_\mathrm{B}^r (1-\alpha_\mathrm{B})^{R-r} \binom{a_\mathrm{m}}{x''} (\rho[r])^{x''} (1 - \rho[r])^{a_\mathrm{m}-x''} \tag{A5}$$

$$P(d'_\mathrm{h} = x') = \sum_{r=0}^{R-1} \binom{R}{r} \alpha_\mathrm{B}^r (1-\alpha_\mathrm{B})^{R-r} \binom{a_\mathrm{m}}{x''} (\mu[r+1])^{x''} (1 - \mu[r+1])^{a_\mathrm{m}-x''} \tag{A6}$$

where $x'' = a_\mathrm{m} x'$. These equations are both estimates for the mean value of $P(d' = x'|a)$ over all possible a.

Appendix A.3. Transformed WTA

A *transformed* dendritic sum is $d^* = 1 - (1 - d/a)^{1/r}$, where r is the unit usage. The *transformed* distributions cannot be calculated as simple sums of binomials, so the following approach is used. For a given *transformed* sum, x^*, and for each possible value of unit usage, r, an equivalent *normalized* sum is calculated via

$$x'[r] = (1 - (1 - x^*)^r) \tag{A7}$$

The *transformed* cumulative probabilities can then be calculated from the *normalized* distributions:

$$P(d^*_{\mathrm{l}} \geq x^*) = \sum_{r=1}^{R} \left(\begin{array}{c} R \\ r \end{array} \right) \alpha_{\mathrm{B}}^r (1 - \alpha_{\mathrm{B}})^{R-r} P(d'_{\mathrm{l}} \geq x'[r]) \tag{A8}$$

$$P(d^*_{\mathrm{h}} \geq x^*) = \sum_{r=0}^{R-1} \left(\begin{array}{c} R \\ r \end{array} \right) \alpha_{\mathrm{B}}^r (1 - \alpha_{\mathrm{B}})^{R-r} P(d'_{\mathrm{h}} \geq x'[r+1]) \tag{A9}$$

The cumulative probabilities can be used directly in calculating the *transformed WTA* recall response.

An example of the correspondence between the theoretical distributions and the frequency of dendritic sums collected from simulations is shown in Figure A1. Note that the numerically calculated *transformed* cumulative probabilities were changed into actual probabilities for this plot.

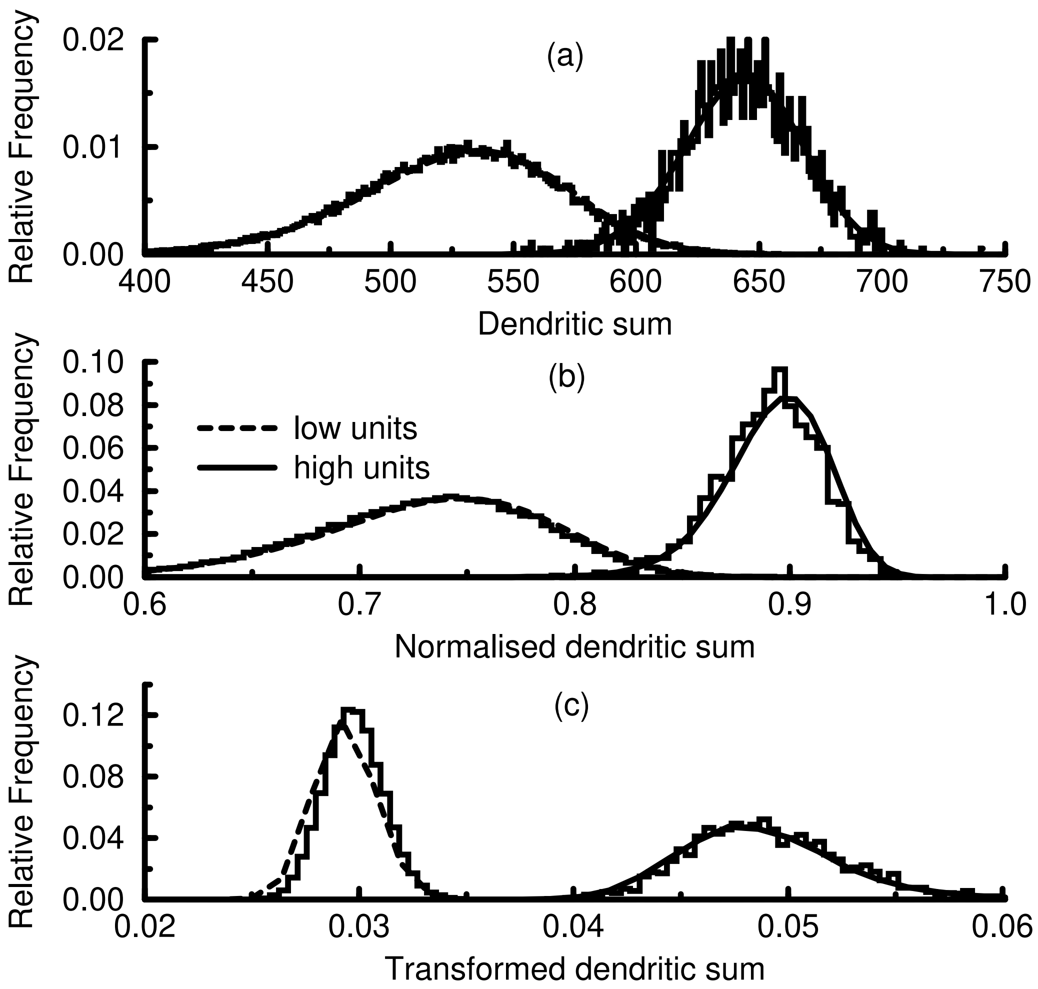

Figure A1. Comparison of the theoretical probability distributions of *low* and *high* unit dendritic sums with the frequency of the sums collected from simulations (staircase plots are simulation data) ($N_A = 48000$, $M_A = 1440$, $N_B = 6144$, $M_B = 180$, $R = 1500$, $Z = 0.5$, $s = 0.4$).

Appendix B. Optimum Information Efficiency

The information efficiency, η, of reliable recall from the fully connected net is easily derived in terms of the average probability, $\hat{\rho}$, that a synapse has been modified during pattern storage (Willshaw *et al* 1969; Willshaw 1971; Buckingham and Willshaw 1992). Optimum efficiency is achieved when $\hat{\rho} = 0.5$ and is approximately $\eta_o = \ln 2$. Under the condition that at most one bit in the output pattern is allowed to be in error, the optimum is achieved when the number of active units in an input pattern is given by $M_A = \log_2(N_B)$.

If the *normalized WTA* threshold strategy is used during recall, the information efficiency of a partially connected net can be derived in exactly the same manner. The average probability, $\hat{\rho}$, that a synapse has been modified during pattern storage is the same as for the fully connected net,

$$\hat{\rho} = 1 - \left(1 - \frac{M_A M_B}{N_A N_B}\right)^R \simeq 1 - \exp\left(-\frac{R M_A M_B}{N_A N_B}\right) \qquad \text{(B10)}$$

where R is the number of pattern pairs stored in the net. Rearranging terms we have,

$$R \simeq -\frac{N_{\mathrm{A}} N_{\mathrm{B}}}{M_{\mathrm{A}} M_{\mathrm{B}}} \ln(1-\hat{\rho}) \tag{B11}$$

The information efficiency, η, is the ratio of the number of bits of information that can be retrieved from the net to the number of bits of storage available,

$$\eta = \frac{R_{\mathrm{c}} I}{Z N_{\mathrm{A}} N_{\mathrm{B}}} \tag{B12}$$

where I is the number of bits of information in a single output pattern, R_{c} is the capacity of the net, and Z is the fraction of input units connected to any given output unit ($Z N_{\mathrm{A}} N_{\mathrm{B}}$ is the number of synapses in the net). The information in a pattern is

$$I = \log_2 \binom{N_{\mathrm{B}}}{M_{\mathrm{B}}} \tag{B13}$$

Substituting equations (B11) and (B13) into (B12) yields

$$\eta = -\log_2 \binom{N_{\mathrm{B}}}{M_{\mathrm{B}}} \ln(1-\hat{\rho})/(Z M_{\mathrm{A}} M_{\mathrm{B}}) \tag{B14}$$

The capacity of the net is defined to be the number of pattern pairs that can be stored in the net before 1 bit, or unit, is in error during recall. As for the fully connected net, except at extremely low connectivity, the only possible errors are false-positives, where output units that should not be active are made active because all their connections to active inputs have modified synapses. On average, an output unit is connected to $Z M_{\mathrm{A}}$ active inputs and the probability that all these connections have modified synapses is $\hat{\rho}^{Z M_{\mathrm{A}}}$. Thus the expected number of errors during recall is $(N_{\mathrm{B}} - M_{\mathrm{B}})\hat{\rho}^{Z M_{\mathrm{A}}}$. Following the analysis for the fully connected net, we take the criterion for good performance to be

$$N_{\mathrm{B}} \hat{\rho}^{Z M_{\mathrm{A}}} = 1 \tag{B15}$$

The net capacity that meets this criterion can be determined by rearranging this equation to find $\hat{\rho}$ and substituting this into (B11). We can also rearrange (B15) to give

$$Z M_{\mathrm{A}} = -\log_2(N_{\mathrm{B}})/\log_2(\hat{\rho}) \tag{B16}$$

Substituting this into (B12) and taking Stirling's approximation yields

$$\eta \simeq \log_2(\hat{\rho}) \ln(1-\hat{\rho}) \tag{B17}$$

This is the identical expression to that obtained for the fully connected net and gives an optimum value of $\eta_{\mathrm{o}} = \ln 2$ when $\hat{\rho} = 0.5$. The only difference is that this optimum is obtained when the criterion

$$Z M_{\mathrm{A}} = \log_2(N_{\mathrm{B}}) \tag{B18}$$

is satisfied. So as the input coding rate is raised above $M_{\mathrm{A}} = \log_2(N_{\mathrm{B}})$, it is simply a matter of lowering the connectivity level in proportion to retain optimum information efficiency.