
Effects of Spatial and Temporal Contiguity on the Acquisition of Spatial Information

Thea B. Ghiselli-Crippa and Paul W. Munro
Department of Information Science and Telecommunications
University of Pittsburgh
Pittsburgh, PA 15260
tbgst@sis.pitt.edu, munro@sis.pitt.edu

Abstract

Spatial information comes in two forms: direct spatial information (for example, retinal position) and indirect temporal contiguity information, since objects encountered sequentially are in general spatially close. The acquisition of spatial information by a neural network is investigated here. Given a spatial layout of several objects, networks are trained on a prediction task. Networks using temporal sequences with no direct spatial information are found to develop internal representations that show distances correlated with distances in the external layout. The influence of spatial information is analyzed by providing direct spatial information to the system during training that is either consistent with the layout or inconsistent with it. This approach allows examination of the relative contributions of spatial and temporal contiguity.

1 Introduction

Spatial information is acquired by a process of exploration that is fundamentally temporal, whether it be on a small scale, such as scanning a picture, or on a larger one, such as physically navigating through a building, a neighborhood, or a city. Continuous scanning of an environment causes locations that are spatially close to have a tendency to occur in temporal proximity to one another. Thus, a temporal associative mechanism (such as a Hebb rule) can be used in conjunction with continuous exploration to capture the spatial structure of the environment [1]. However, the actual process of building a cognitive map need not rely solely on temporal associations, since some spatial information is encoded in the sensory array (position on the retina and proprioceptive feedback). Laboratory studies show different types of interaction between the relative contributions of temporal and spatial contiguities to the formation of an internal representation of space. While Clayton and Habibi's [2] series of recognition priming experiments indicates that priming is controlled only by temporal associations, in the work of McNamara et al. [3] priming in recognition is observed only when space and time are both contiguous. In addition, Curiel and Radvansky's [4] work shows that the effects of spatial and temporal contiguity depend on whether location or identity information is emphasized during learning. Moreover, other experiments ([3]) also show how the effects clearly depend on the task and can be quite different if an explicitly spatial task is used (e.g., additive effects in location judgments).

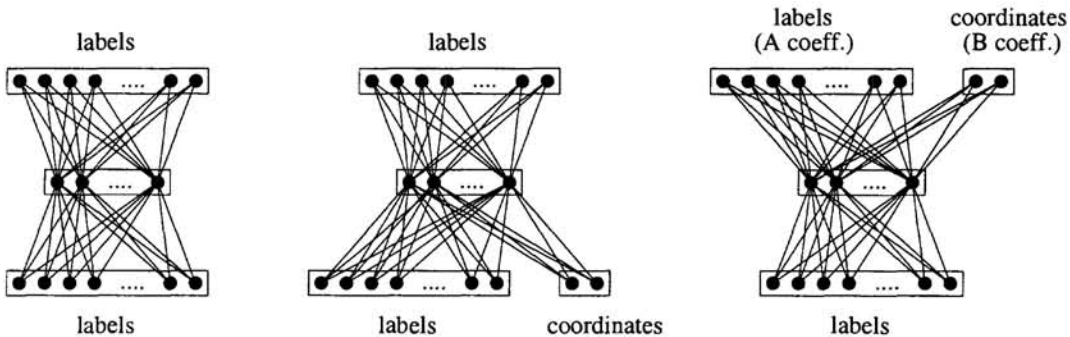


Figure 1: Network architectures: temporal-only network (left); spatio-temporal network with spatial units part of the input representation (center); spatio-temporal network with spatial units part of the output representation (right).

2 Network architectures

The goal of the work presented in this paper is to study the structure of the internal representations that emerge from the integration of temporal and spatial associations. An encoder-like network architecture is used (see Figure 1), with a set of N input units and a set of N output units representing N nodes on a 2-dimensional graph. A set of H units is used for the hidden layer. To include space in the learning process, additional spatial units are included in the network architecture. These units provide a representation of the spatial information directly available during the learning/scanning process. In the simulations described in this paper, two units are used and are chosen to represent the (x, y) coordinates of the nodes in the graph. The spatial units can be included as part of the input representation or as part of the output representation (see Figure 1, center and right panels): both choices are used in the experiments, to investigate whether the spatial information could better benefit training as an input or as an output [5]. In the second case, the relative contribution of the spatial information can be directly manipulated by introducing weighting factors in the cost function being minimized. A two-term cost function is used, with a cross-entropy term for the N label units and a squared error term for the 2 coordinate units,

$$E = A \left[- \sum_{i=1}^N t_i \log_2(r_i) + (1 - t_i) \log_2(1 - r_i) \right] + B \left[\frac{1}{2} \sum_{i=1}^2 (t_i - r_i)^2 \right] \quad (1)$$

r_i indicates the actual output of unit i and t_i its desired output. The relative influence of the spatial information is controlled by the coefficients A and B .

3 Learning tasks

The left panel of Figure 2 shows an example of the type of layout used; the effective layout used in the study consists of $N = 28$ nodes. For each node, a set of neighboring nodes is defined, chosen on the basis of how an observer might scan the layout to learn the node labels and their (spatial) relationships; in Figure 2, the neighborhood relationships are represented by lines connecting neighboring nodes. From any node in the layout, the only allowed transitions are those to a neighbor, thus defining the set of node pairs used to train the network (66 pairs out of $C(28, 2) = 378$ possible pairs). In addition, the probability of occurrence of a particular transition is computed as a function of the distance to the corresponding neighbor. It is then possible to generate a sequence of visits to the network nodes, aimed at replicating the scanning process of a human observer studying the layout.

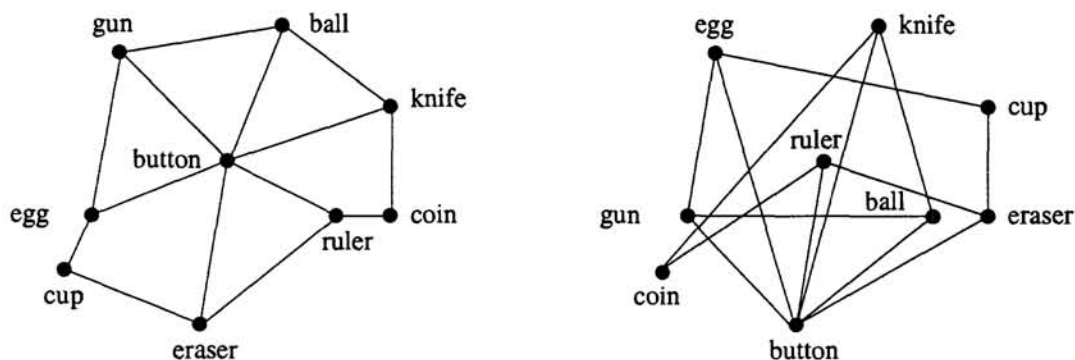


Figure 2: Example of a layout (left) and its permuted version (right). Links represent allowed transitions. A larger layout of 28 units was used in the simulations.

The basic learning task is similar to the grammar learning task of Servan-Schreiber et al. [6] and to the neighborhood mapping task described in [1] and is used to associate each of the N nodes on the graph and its (x, y) coordinates with the probability distribution of the transitions to its neighboring nodes. The mapping can be learned directly, by associating each node with the probability distribution of the transitions to all its neighbors: in this case, batch learning is used as the method of choice for learning the mapping. On the other hand, the mapping can be learned indirectly, by associating each node with itself and one of its neighbors, with online learning being the method of choice in this case; the neighbor chosen at each iteration is defined by the sequence of visits generated on the basis of the transition probabilities. Batch learning was chosen because it generally converges more smoothly and more quickly than online learning and gives qualitatively similar results. While the task and network architecture described in [1] allowed only for temporal association learning, in this study both temporal and spatial associations are learned simultaneously, thanks to the presence of the spatial units. However, the temporal-only (T-only) case, which has no spatial units, is included in the simulations performed for this study, to provide a benchmark for the evaluation of the results obtained with the spatio-temporal (S-T) networks.

The task described above allows the network to learn neighborhood relationships for which spatial and temporal associations provide consistent information, that is, nodes experienced contiguously in time (as defined by the sequence) are also contiguous in space (being spatial neighbors). To tease apart the relative contributions of space and time, the task is kept the same, but the data employed for training the network is modified: the same layout is used to generate the temporal sequence, but the x, y coordinates of the nodes are randomly permuted (see right panel of Figure 2). If the permuted layout is then scanned following the same sequence of node visits used in the original version, the net effect is that the temporal associations remain the same, but the spatial associations change so that temporally neighboring nodes can now be spatially close or distant: the spatial associations are no longer consistent with the temporal associations. As Figure 4 illustrates, the training pairs (filled circles) all correspond to short distances in the original layout, but can have a distance anywhere in the allowable range in the permuted layout. Since the temporal and spatial distances were consistent in the original layout, the original spatial distance can be used as an indicator of temporal distance and Figure 4 can be interpreted as a plot of temporal distance vs. spatial distance for the permuted layout.

The simulations described in the following include three experimental conditions: temporal only (no direct spatial information available); space and time consistent (the spatial coordinates and the temporal sequence are from the same layout); space and time inconsistent (the spatial coordinates and the temporal sequence are from different layouts).

Hidden unit representations are compared using Euclidean distance (cosine and inner product measures give consistent results); the internal representation distances are also used to compute their correlation with Euclidean distances between nodes in the layout (original and permuted). The correlations increase with the number of hidden units for values of H between 5 and 10 and then gradually taper off for values greater than 10. The results presented in the remainder of the paper all pertain to networks trained with $H = 20$ and with hidden units using a \tanh transfer function; all the results pertaining to S-T networks refer to networks with 2 spatial output units and cost function coefficients $A = 0.625$ and $B = 6.25$.

4 Results

Figure 3 provides a combined view of the results from all three experiments. The left panel illustrates the evolution of the correlation between internal representation distances and layout (original and permuted) distances. The right panel shows the distributions of the correlations at the end of training (1000 epochs). The first general result is that, when spatial information is available and consistent with the temporal information (original layout), the correlation between hidden unit distances and layout distances is consistently better than the correlation obtained in the case of temporal associations alone. The second general result is that, when spatial information is available but not consistent with the temporal information (permuted layout), the correlation between hidden unit distances and original layout distances (which represent temporal distances) is similar to that obtained in the case of temporal associations alone, except for the initial transient. When the correlation is computed with respect to the permuted layout distances, its value peaks early during training and then decreases rapidly, to reach an asymptotic value well below the other three cases. This behavior is illustrated in the boxplots in the right panel of Figure 3, which report the distribution of correlation values at the end of training.

4.1 Temporal-only vs. spatio-temporal

As a first step in this study, the effects of adding spatial information to the basic temporal associations used to train the network can be examined. Since the learning task is the same for both the T-only and the S-T networks except for the absence or presence of spatial information during training, the differences observed can be attributed to the additional spatial information available to the S-T networks. The higher correlation between internal representation distances and original layout distances obtained when spatial information is

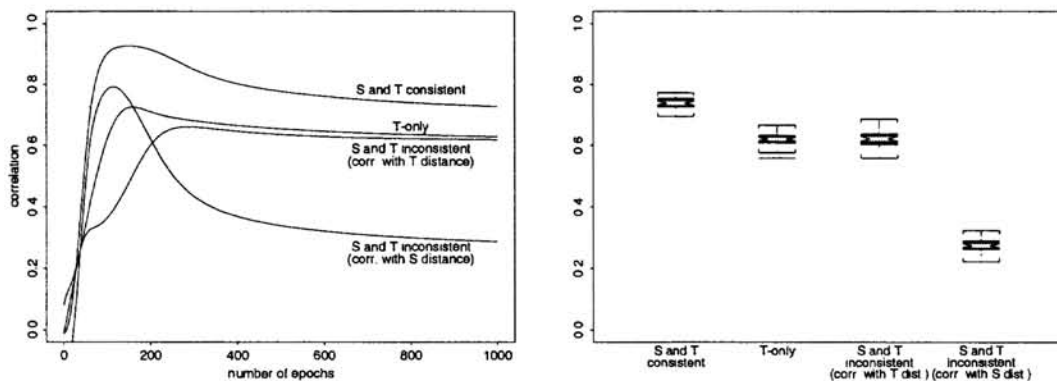


Figure 3: Evolution of correlation during training (0 - 1000 epochs) (left). Distributions of correlations at the end of training (1000 epochs) (right).

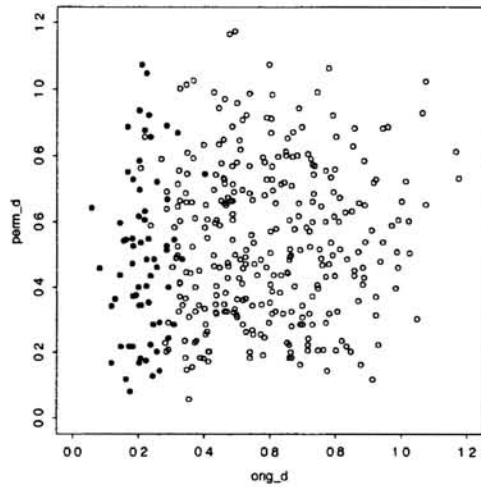


Figure 4: Distances in the original layout (x) vs. distances in the permuted layout (y). The 66 training pairs are identified by filled circles.

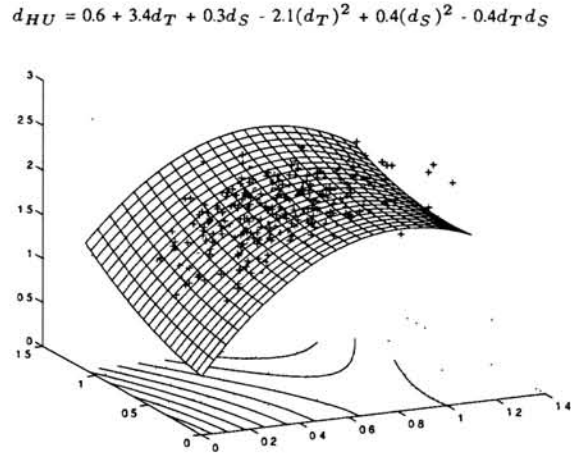


Figure 5: Similarities (Euclidean distances) between internal representations developed by a S-T network (after 300 epochs). Figure 4 projects the data points onto the x, y plane.

available (see Figure 3) is apparent also when the evolution of the internal representations is examined. As Figure 6 illustrates, the presence of spatial information results in better generalization for the pattern pairs outside the training set. While the distances between training pairs are mapped to similar distances in hidden unit space for both the T-only and the S-T networks, the T-only network tends to cluster the non-training pairs into a narrow band of distances in hidden unit space. In the case of the S-T network instead, the hidden unit distances between non-training pairs are spread out over a wider range and tend to reflect the original layout distances.

4.2 Permuted layout

As described above, with the permuted layout it is possible to decouple the spatial and temporal contributions and therefore study the effects of each. A comprehensive view of the results at a particular point during training (300 epochs) is presented in Figure 5, where the x, y plane represents temporal distance vs. spatial distance (see also Figure 4) and the z axis represents the similarity between hidden unit representations. The figure also includes a quadratic regression surface fitted to the data points. The coefficients in the equation of the surface provide a quantitative measure of the relative contributions of spatial (d_S) and temporal distances (d_T) to the similarity between hidden unit representations (d_{HU}):

$$d_{HU} = k_0 + k_1 d_T + k_2 d_S + k_3 (d_T)^2 + k_4 (d_S)^2 + k_5 d_T d_S \quad (2)$$

In general, after the transient observed in early training (see Figure 3), the largest and most significant coefficients are found for d_T and $(d_T)^2$, indicating a stronger dependence of d_{HU} on temporal distance than on spatial distance.

The results illustrated in Figure 5 represent the situation at a particular point during training (300 epochs). Similar plots can be generated for different points during training, to study the evolution of the internal representations. A different view of the evolution process is provided by Figure 7, in which the data points are projected onto the x, z plane (top panel) and the y, z plane (bottom panel) at four different times during training. In the top panel,

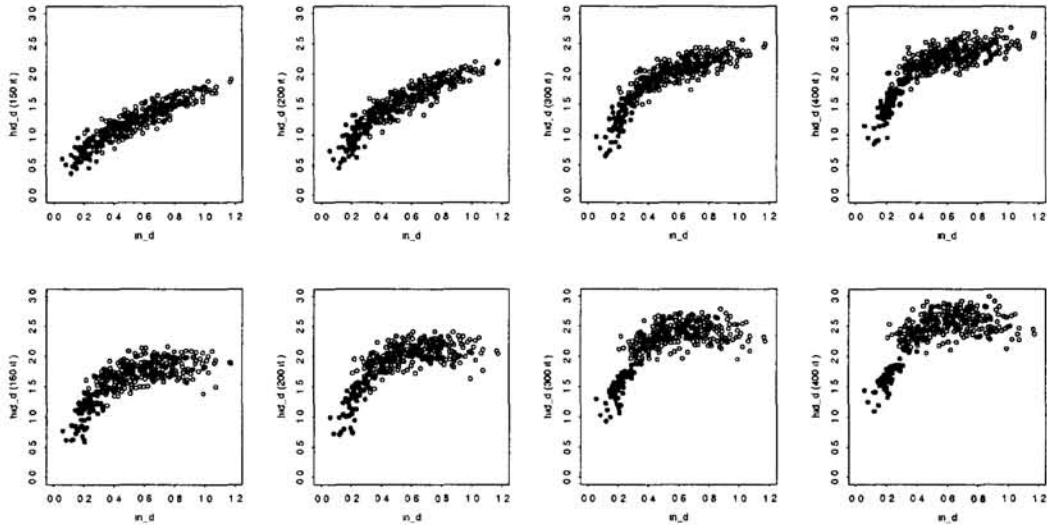


Figure 6: Internal representation distances vs. original layout distances: S-T network (top) vs. T-only network (bottom). The training pairs are identified by filled circles. The presence of spatial information results in better generalization for the pairs outside the training set.

the internal representation distances are plotted as a function of temporal distance (i.e., the spatial distance from the original layout), while in the bottom panel they are plotted as a function of spatial distance (from the permuted layout). The higher asymptotic correlation between internal representation distances and temporal distances, as opposed to spatial distances (see Figure 3), is apparent also from the examination of the evolutionary plots, which show an asymptotic behavior with respect to temporal distances (see Figure 7, top panel) very similar to the T-only case (see Figure 6, bottom panel).

5 Discussion

The first general conclusion that can be drawn from the examination of the results described in the previous section is that, when the spatial information is available and consistent with the temporal information (original layout), the similarity structure of the hidden unit representations is closer to the structure of the original layout than that obtained by using temporal associations alone. The second general conclusion is that, when the spatial information is available but not consistent with the temporal information (permuted layout), the similarity structure of the hidden unit representations seems to correspond to temporal more than spatial proximity. Figures 5 and 7 both indicate that temporal associations take precedence over spatial associations. This result is in agreement with the results described in [1], showing how temporal associations (plus some high-level constraints) significantly contribute to the internal representation of global spatial information. However, spatial information certainly is very beneficial to the (temporal) acquisition of a layout, as proven by the results obtained with the S-T network vs. the T-only network.

In terms of the model presented in this paper, the results illustrated in Figures 5 and 7 can be compared with the experimental data reported for recognition priming ([2], [3], [4]), with distance between internal representations corresponding to reaction time. The results of our model indicate that distances in both the spatially far and spatially close condition appear to be consistently shorter for the training pairs (temporally close) than for the non-training pairs (temporally distant), highlighting a strong temporal effect consistent with the data reported in [2] and [4] (for spatially far pairs) and in [3] (only for the spatially close

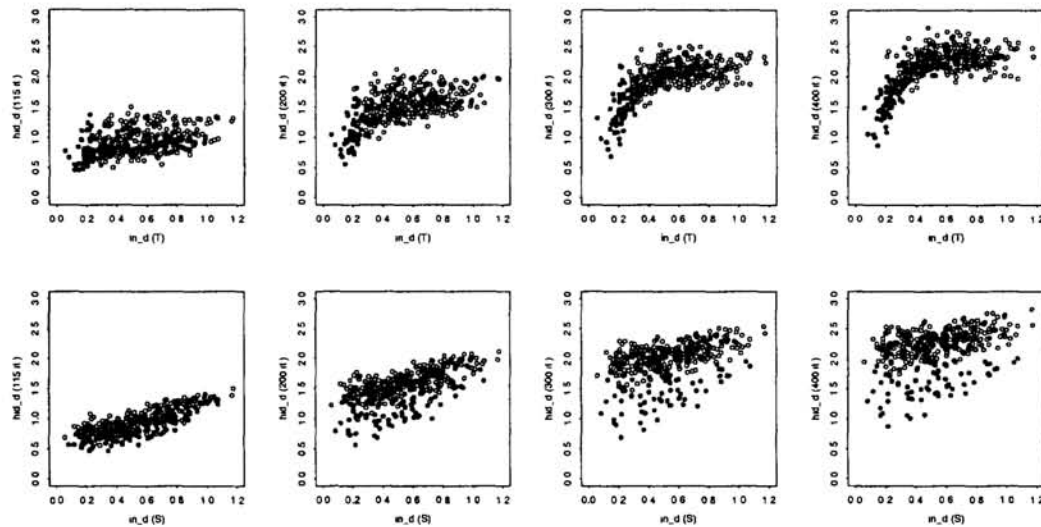


Figure 7: Internal representation distances vs. temporal distances (top) and vs. spatial distances (bottom) for a S-T network (permuted layout). The training pairs are identified by filled circles. The asymptotic behavior with respect to temporal distances (top panel) is similar to the T-only condition. The bottom panel indicates a weak dependence on spatial distances.

case). For the training pairs (temporally close), slightly shorter distances are obtained for spatially close pairs vs. spatially far pairs; this result does not provide support for the experimental data reported in either [3] (strong spatial effect) or [2] (no spatial effect). For the non-training pairs (temporally distant), long distances are found throughout, with no strong dependence on spatial distance; this effect is consistent with all the reported experimental data. Further simulations and statistical analyses are necessary for a more conclusive comparison with the experimental data.

References

- [1] Ghiselli-Crippa, T.B. & Munro, P.W. (1994). Emergence of global structure from local associations. In J.D. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in Neural Information Processing Systems 6*, pp. 1101-1108. San Francisco, CA: Morgan Kaufmann.
- [2] Clayton, K.N. & Habibi, A. (1991). The contribution of temporal contiguity to the spatial priming effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **17**:263-271.
- [3] McNamara, T.P., Halpin, J.A. & Hardy, J.K. (1992). Spatial and temporal contributions to the structure of spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **18**:555-564.
- [4] Curiel, J.M. & Radvansky, G.A. (1998). Mental organization of maps. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24**:202-214.
- [5] Caruana, R. & de Sa, V.R. (1997). Promoting poor features to supervisors: Some inputs work better as outputs. In M.C. Mozer, M.I. Jordan, & T. Petsche (Eds.), *Advances in Neural Information Processing Systems 9*, pp. 389-395. Cambridge, MA: MIT Press.
- [6] Servan-Schreiber, D., Cleeremans, A. & McClelland, J.L. (1989). Learning sequential structure in simple recurrent networks. In D.S. Touretzky (Ed.), *Advances in Neural Information Processing Systems 1*, pp. 643-652. San Mateo, CA: Morgan Kaufmann.

Neural Representation of Multi-Dimensional Stimuli

Christian W. Eurich, Stefan D. Wilke and Helmut Schwegler
Institut für Theoretische Physik
Universität Bremen, Germany
(*eurich,swilke,schwegler*)@*physik.uni-bremen.de*

Abstract

The encoding accuracy of a population of stochastically spiking neurons is studied for different distributions of their tuning widths. The situation of identical radially symmetric receptive fields for all neurons, which is usually considered in the literature, turns out to be disadvantageous from an information-theoretic point of view. Both a variability of tuning widths and a fragmentation of the neural population into specialized subpopulations improve the encoding accuracy.

1 Introduction

The topic of neuronal tuning properties and their functional significance has focused much attention in the last decades. However, neither empirical findings nor theoretical considerations have yielded a unified picture of optimal neural encoding strategies given a sensory or motor task. More specifically, the question as to whether narrow tuning or broad tuning is advantageous for the representation of a set of stimulus features is still being discussed. Empirically, both situations are encountered: small receptive fields whose diameter is less than one degree can, for example, be found in the human retina [7], and large receptive fields up to 180° in diameter occur in the visual system of tongue-projecting salamanders [10]. On the theoretical side, arguments have been put forward for small [8] as well as for large [5, 1, 9, 3, 13] receptive fields.

In the last years, several approaches have been made to calculate the encoding accuracy of a neural population as a function of receptive field size [5, 1, 9, 3, 13]. It has turned out that for a firing rate coding, large receptive fields are advantageous provided that $D \geq 3$ stimulus features are encoded [9, 13]. For binary neurons, large receptive fields are advantageous also for $D = 2$ [5, 3].

However, so far only radially symmetric tuning curves have been considered. For neural populations which lack this symmetry, the situation may be very different. Here we study the encoding accuracy of a population of stochastically spiking neurons. A Fisher information analysis performed on different distributions of tunings widths will indeed reveal a much more detailed picture of neural encoding strategies.

2 Model

Consider a D -dimensional stimulus space, X . A stimulus is characterized by a position $\mathbf{x} = (x_1, \dots, x_D) \in X$, where the value of feature i , x_i ($i = 1, \dots, D$), is measured relative to the total range of values in the i -th dimension such that it is dimensionless. Information about the stimulus is encoded by a population of N stochastically spiking neurons. They are assumed to have independent spike generation mechanisms such that the joint probability distribution for observing $\mathbf{n} = (n^{(1)}, \dots, n^{(k)}, \dots, n^{(N)})$ spikes within a time interval τ , $P_s(\mathbf{n}; \mathbf{x})$, can be written in the form

$$P_s(\mathbf{n}; \mathbf{x}) = \prod_{k=1}^N P_s^{(k)}(n^{(k)}; \mathbf{x}), \quad (1)$$

where $P_s^{(k)}(n^{(k)}; \mathbf{x})$ is the single-neuron probability distribution of the number of observed spikes given the stimulus at position \mathbf{x} . Note that (1) does not exclude a correlation of the neural firing rates, i.e., the neurons may have common input or even share the same tuning function.

The firing rates depend on the stimulus via the local values of the tuning functions, such that $P_s^{(k)}(n^{(k)}; \mathbf{x})$ can be written in the form $P_s^{(k)}(n^{(k)}; \mathbf{x}) = S(n^{(k)}, f^{(k)}(\mathbf{x}), \tau)$, where the tuning function of neuron k , $f^{(k)}(\mathbf{x})$, gives its mean firing rate in response to the stimulus at position \mathbf{x} . We assume here a form of the tuning function that is not necessarily radially symmetric,

$$f^{(k)}(\mathbf{x}) = F \phi \left(\sum_{i=1}^D \frac{(x_i - c_i^{(k)})^2}{\sigma_i^{(k)2}} \right) =: F \phi \left(\xi^{(k)2} \right), \quad (2)$$

where $\mathbf{c}^{(k)} = (c_1^{(k)}, \dots, c_D^{(k)})$ is the center of the tuning curve of neuron k , $\sigma_i^{(k)}$ is its tuning width in the i -th dimension, $\xi_i^{(k)2} := (x_i - c_i^{(k)})^2 / \sigma_i^{(k)2}$ for $i = 1, \dots, D$, and $\xi^{(k)2} := \xi_1^{(k)2} + \dots + \xi_D^{(k)2}$. $F > 0$ denotes the maximal firing rate of the neurons, which requires that $\max_{z \geq 0} \phi(z) = 1$.

We assume that the tuning widths $\sigma_1^{(k)}, \dots, \sigma_D^{(k)}$ of each neuron k are drawn from a distribution $P_\sigma(\sigma_1, \dots, \sigma_D)$. For a population of tuning functions with centers $\mathbf{c}^{(1)}, \dots, \mathbf{c}^{(N)}$, a density $\eta(\mathbf{x})$ is introduced according to $\eta(\mathbf{x}) := \sum_{k=1}^N \delta(\mathbf{x} - \mathbf{c}^{(k)})$.

The encoding accuracy can be quantified by the Fisher information matrix, \mathbf{J} , which is defined as

$$J_{ij}(\mathbf{x}) := E \left[\left(\frac{\partial}{\partial x_i} \ln P(\mathbf{n}; \mathbf{x}) \right) \left(\frac{\partial}{\partial x_j} \ln P(\mathbf{n}; \mathbf{x}) \right) \right], \quad (3)$$

where $E[\dots]$ denotes the expectation value over the probability distribution $P(\mathbf{n}; \mathbf{x})$ [2]. The Fisher information yields a lower bound on the expected error of an unbiased estimator that retrieves the stimulus \mathbf{x} from the noisy neural activity (Cramér-Rao inequality) [2]. The minimal estimation error for the i -th feature x_i , $\epsilon_{i,\min}$, is given by $\epsilon_{i,\min}^2 = (\mathbf{J}^{-1})_{ii}$ which reduces to $\epsilon_{i,\min}^2 = 1/J_{ii}(\mathbf{x})$ if \mathbf{J} is diagonal.

We shall now derive a general expression for the population Fisher information. In the next chapter, several cases and their consequences for neural encoding strategies will be discussed.

For model neuron (k), the Fisher information (3) reduces to

$$J_{ij}^{(k)}(\mathbf{x}; \sigma_1^{(k)}, \dots, \sigma_D^{(k)}) = \frac{1}{\sigma_i^{(k)} \sigma_j^{(k)}} A_\phi \left(\xi^{(k)2}, F, \tau \right) \xi_i^{(k)} \xi_j^{(k)}, \quad (4)$$

where the dependence on the tuning widths is indicated by the list of arguments. The function A_ϕ depends on the shape of the tuning function and is given in [13]. The independence assumption (1) implies that the population Fisher information is the sum of the contributions of the individual neurons, $\sum_{k=1}^N J_{ij}^{(k)}(\mathbf{x}; \sigma_1^{(k)}, \dots, \sigma_D^{(k)})$. We now define a population Fisher information which is averaged over the distribution of tuning widths $P_\sigma(\sigma_1, \dots, \sigma_D)$:

$$\langle J_{ij}(\mathbf{x}) \rangle_\sigma = \sum_{k=1}^N \int d\sigma_1 \dots d\sigma_D P_\sigma(\sigma_1, \dots, \sigma_D) J_{ij}^{(k)}(\mathbf{x}; \sigma_1, \dots, \sigma_D). \quad (5)$$

Introducing the density of tuning curves, $\eta(\mathbf{x})$, into (5) and assuming a constant distribution, $\eta(\mathbf{x}) \equiv \eta \equiv \text{const.}$, one obtains the result that the population Fisher information becomes independent of \mathbf{x} and that the off-diagonal elements of \mathbf{J} vanish [13]. The average population Fisher information then becomes

$$\langle J_{ij} \rangle_\sigma = \eta D K_\phi(F, \tau, D) \left\langle \frac{\prod_{l=1}^D \sigma_l}{\sigma_i^2} \right\rangle_\sigma \delta_{ij}, \quad (6)$$

where K_ϕ depends on the geometry of the tuning curves and is defined in [13].

3 Results

In this section, we consider different distributions of tuning widths in (6) and discuss advantageous and disadvantageous strategies for obtaining a high representational accuracy in the neural population.

Radially symmetric tuning curves. For radially symmetric tuning curves of width $\bar{\sigma}$, the tuning-width distribution reads

$$P_\sigma(\sigma_1, \dots, \sigma_D) = \prod_{i=1}^D \delta(\sigma_i - \bar{\sigma});$$

see Fig. 1a for a schematic visualization of the arrangement of the tuning widths for the case $D = 2$. The average population Fisher information (6) for $i = j$ becomes

$$\langle J_{ii} \rangle_\sigma = \eta D K_\phi(F, \tau, D) \bar{\sigma}^{D-2}, \quad (7)$$

a result already obtained by Zhang and Sejnowski [13]. Equation (7) basically shows that the minimal estimation error increases with $\bar{\sigma}$ for $D = 1$, that it does not depend on $\bar{\sigma}$ for $D = 2$, and that it decreases as $\bar{\sigma}$ increases for $D \geq 3$. We shall discuss the relevance of this case below.

Identical tuning curves without radial symmetry. Next we discuss tuning curves which are identical but not radially symmetric; the tuning-width distribution for this case is

$$P_\sigma(\sigma_1, \dots, \sigma_D) = \prod_{i=1}^D \delta(\sigma_i - \bar{\sigma}_i),$$

where $\bar{\sigma}_i$ denotes the fixed width in dimension i . For $i = j$, the average population Fisher information (6) reduces to [11, 4]

$$\langle J_{ii} \rangle_\sigma = \eta D K_\phi(F, \tau, D) \frac{\prod_{l=1}^D \bar{\sigma}_l}{\bar{\sigma}_i^2}. \quad (8)$$

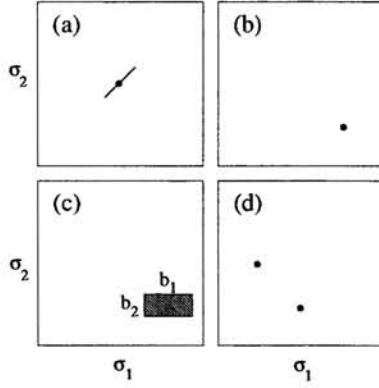


Figure 1: Visualization of different distributions of tuning widths for $D = 2$. (a) Radially symmetric tuning curves. The dot indicates a fixed $\bar{\sigma}$, while the diagonal line symbolizes a variation in $\bar{\sigma}$ discussed in [13]. (b) Identical tuning curves which are not radially symmetric. (c) Tuning widths uniformly distributed within a small rectangle. (d) Two subpopulations each of which is narrowly tuned in one dimension and broadly tuned in the other direction.

Equation (8) contains (7) as a special case. From (8) it becomes immediately clear that the expected minimal square encoding error for the i -th stimulus feature, $\epsilon_{i,\min}^2 = 1/\langle J_{ii}(\mathbf{x}) \rangle_\sigma$, depends on i , i. e., *the population specializes in certain features*. The error obtained in dimension i thereby depends on the tuning widths in all dimensions.

Which encoding strategy is optimal for a population whose task it is to encode a single feature, say feature i , with high accuracy while not caring about the other dimensions? In order to answer this question, we re-write (8) in terms of receptive field overlap.

For the tuning functions $f^{(k)}(\mathbf{x})$ encountered empirically, large values of the single-neuron Fisher information (4) are typically restricted to a region around the center of the tuning function, $\mathbf{c}^{(k)}$. The fraction $p(\beta)$ of the Fisher information that falls into a region $E_D : \sqrt{\xi^{(k)2}} \leq \beta$ around $\mathbf{c}^{(k)}$ is given by

$$p(\beta) := \frac{\int_{E_D} d^D x \sum_{i=1}^D J_{ii}^{(k)}(\mathbf{x})}{\int_X d^D x \sum_{i=1}^D J_{ii}^{(k)}(\mathbf{x})} = \frac{\int_0^\beta d\xi \xi^{D+1} A_\phi(\xi^2, F, \tau)}{\int_0^\infty d\xi \xi^{D+1} A_\phi(\xi^2, F, \tau)}, \quad (9)$$

where the index (k) was dropped because the tuning curves are assumed to have identical shapes. Equation (9) allows the definition of an effective receptive field, $\text{RF}_{\text{eff}}^{(k)}$, inside of which neuron k conveys a major fraction p_0 of Fisher information, $\text{RF}_{\text{eff}}^{(k)} := \{\mathbf{x} | \sqrt{\xi^{(k)2}} \leq \beta_0\}$, where β_0 is chosen such that $p(\beta_0) = p_0$. The Fisher information a neuron k carries is small unless $\mathbf{x} \in \text{RF}_{\text{eff}}^{(k)}$. This has the consequence that a fixed stimulus \mathbf{x} is actually encoded only by a subpopulation of neurons. The point \mathbf{x} in stimulus space is covered by

$$N_{\text{code}} := \eta \frac{2\pi^{D/2}(\beta_0)^D}{D\Gamma(D/2)} \prod_{j=1}^D \bar{\sigma}_j \quad (10)$$

receptive fields. With the help of (10), the average population Fisher information (8) can be re-written as

$$\langle J_{ii} \rangle_\sigma = \frac{D^2 \Gamma(D/2)}{2\pi^{D/2}(\beta_0)^D} K_\phi(F, \tau, D) \frac{N_{\text{code}}}{\bar{\sigma}_i^2}. \quad (11)$$

Equation (11) can be interpreted as follows: We assume that the population of neurons encodes stimulus dimension i accurately, while all other dimensions are of secondary importance. The average population Fisher information for dimension i , $\langle J_{ii} \rangle_\sigma$, is determined by the tuning width in dimension i , $\bar{\sigma}_i$, and by the size of the active subpopulation, N_{code} . There is a tradeoff between these quantities. On the one hand, the encoding error can be decreased by decreasing $\bar{\sigma}_i$, which enhances the Fisher information carried by each single

neuron. Decreasing $\bar{\sigma}_i$, on the other hand, will also shrink the active subpopulation via (10). This impairs the encoding accuracy, because the stimulus position is evaluated from the activity of fewer neurons. If (11) is valid due to a sufficient receptive field overlap, N_{code} can be increased by increasing the tuning widths, $\bar{\sigma}_j$, in all other dimensions $j \neq i$. This effect is illustrated in Fig. 2 for $D = 2$.

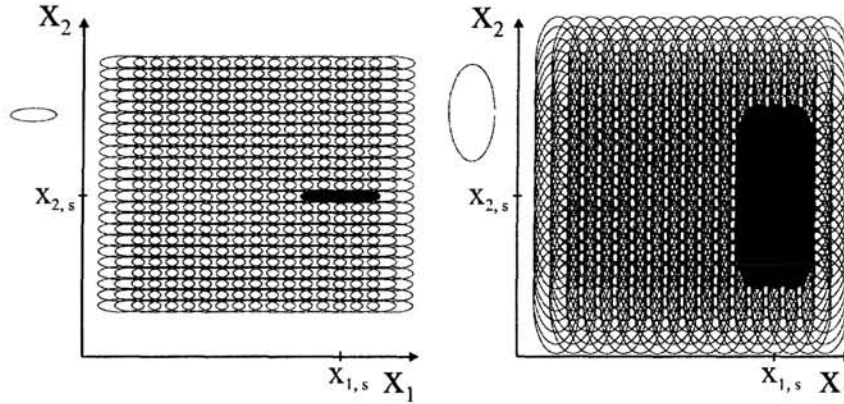


Figure 2: Encoding strategy for a stimulus characterized by parameters $x_{1,s}$ and $x_{2,s}$. Feature x_1 is to be encoded accurately. Effective receptive field shapes are indicated for both populations. If neurons are narrowly tuned in x_2 (left), the active population (solid) is small (here: $N_{\text{code}} = 3$). Broadly tuned receptive fields for x_2 (right) yield a much larger population (here: $N_{\text{code}} = 27$) thus increasing the encoding accuracy.

It shall be noted that although a narrow tuning width $\bar{\sigma}_i$ is advantageous, the limit $\bar{\sigma}_i \rightarrow 0$ yields a bad representation. For narrowly tuned cells, gaps appear between the receptive fields: The condition $\eta(\mathbf{x}) \equiv \text{const.}$ breaks down, and (6) is no longer valid. A more detailed calculation shows that the encoding error diverges as $\bar{\sigma}_i \rightarrow 0$ [4]. The fact that the encoding error decreases for both narrow tuning and broad tuning – due to (11) – proves the existence of an *optimal tuning width*. An example is given in Fig. 3a.

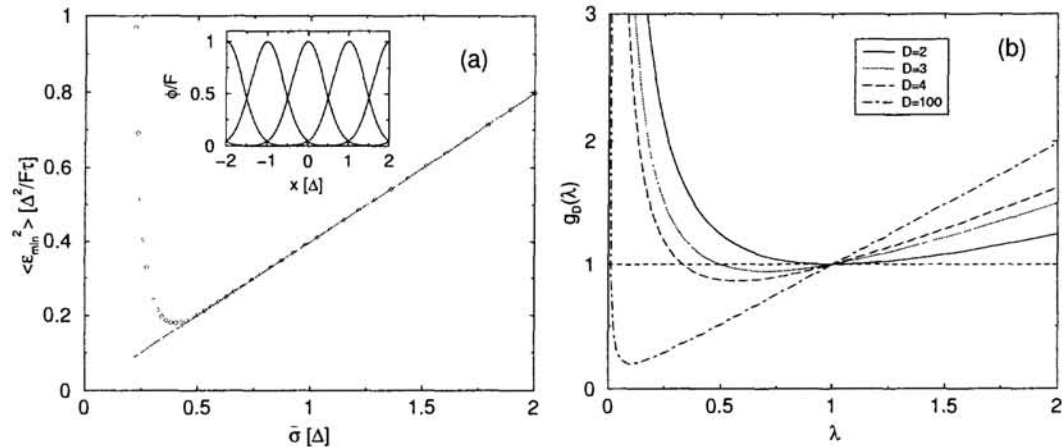


Figure 3: (a) Example for the encoding behavior with narrow tuning curves arranged on a regular lattice of dimension $D = 1$ (grid spacing Δ). Tuning curves are Gaussian, and neural firing is modeled as a Poisson process. Dots indicate the minimal square encoding error averaged over a uniform distribution of stimuli, $\langle \epsilon_{\min}^2 \rangle$, as a function of $\bar{\sigma}$. The minimum is clearly visible. The dotted line shows the corresponding approximation according to (8). The inset shows Gaussian tuning curves of optimal width, $\bar{\sigma}^{\text{opt}} \approx 0.4\Delta$. (b) $g_D(\lambda)$ as a function of λ for different values of D .

Narrow distribution of tuning curves. In order to study the effects of encoding the stimulus with distributed tuning widths instead of identical tuning widths as in the previous cases, we now consider the distribution

$$P_\sigma(\sigma_1, \dots, \sigma_D) = \prod_{i=1}^D \frac{1}{b_i} \Theta \left[\sigma_i - \left(\bar{\sigma}_i - \frac{b_i}{2} \right) \right] \Theta \left[\left(\bar{\sigma}_i + \frac{b_i}{2} \right) - \sigma_i \right], \quad (12)$$

where Θ denotes the Heaviside step function. Equation (12) describes a uniform distribution in a D -dimensional cuboid of size b_1, \dots, b_D around $(\bar{\sigma}_1, \dots, \bar{\sigma}_D)$; cf. Fig. 1c. A straightforward calculation shows that in this case, the average population Fisher information (6) for $i = j$ becomes

$$\langle J_{ii} \rangle_\sigma = \eta DK_\phi(F, \tau, D) \frac{\prod_{l=1}^D \bar{\sigma}_l}{\bar{\sigma}_i^2} \left\{ 1 + \frac{1}{12} \left(\frac{b_i}{\bar{\sigma}_i} \right)^2 + \mathcal{O} \left[\left(\frac{b_i}{\bar{\sigma}_i} \right)^4 \right] \right\}. \quad (13)$$

A comparison with (8) yields the astonishing result that an increase in b_i results in an increase in the i -th diagonal element of the average population Fisher information matrix and thus in an improvement in the encoding of the i -th stimulus feature, while the encoding in dimensions $j \neq i$ is not affected. Correspondingly, the total encoding error can be decreased by increasing an arbitrary number of edge lengths of the cube. *The encoding by a population with a variability in the tuning curve geometries as described is more precise than that by a uniform population. This is true for arbitrary D .* Zhang and Sejnowski [13] consider the more artificial situation of a correlated variability of the tuning widths: tuning curves are always assumed to be radially symmetric. This is indicated by the diagonal line in Fig. 1a. A distribution of tuning widths restricted to this subset yields an average population Fisher information $\propto \langle \bar{\sigma}^{D-2} \rangle$ and does not improve the encoding for $D = 2$ or $D = 3$.

Fragmentation into D subpopulations. Finally, we study a family of distributions of tuning widths which also yields a lower minimal encoding error than the uniform population. Let the density of tuning curves be given by

$$P_\sigma(\sigma_1, \dots, \sigma_D) = \frac{1}{D} \sum_{i=1}^D \delta(\sigma_i - \lambda \bar{\sigma}) \prod_{j \neq i} \delta(\sigma_j - \bar{\sigma}), \quad (14)$$

where $\lambda > 0$. For $\lambda = 1$, the population is uniform as in (7). For $\lambda \neq 1$, the population is split up into D subpopulations; in subpopulation i , σ_i is modified while $\sigma_j \equiv \bar{\sigma}$ for $j \neq i$. See Fig. 1d for an example. The diagonal elements of the average population Fisher information are

$$\langle J_{ii} \rangle_\sigma = \eta DK_\phi(F, \tau, D) \bar{\sigma}^{D-2} \left\{ \frac{1 + (D-1)\lambda^2}{D\lambda} \right\}, \quad (15)$$

where the term in brackets will be abbreviated as $g_D(\lambda)$. $\langle J_{ii} \rangle_\sigma$ does not depend on i in this case because of the symmetry in the subpopulations. Equation (15) and the uniform case (7) differ by $g_D(\lambda)$ which will now be discussed. Figure 3b shows $g_D(\lambda)$ for different values of D . For $\lambda = 1$, $g_D(\lambda) = 1$ and (7) is recovered as expected. $g_D(\lambda) = 1$ also holds for $\lambda = 1/(D-1) < 1$: narrowing one tuning width in each subpopulation will at first decrease the resolution provided $D \geq 3$; this is due to the fact that N_{code} is decreased. For $\lambda < 1/(D-1)$, however, $g_D(\lambda) > 1$, and the resolution exceeds $\langle J_{ii} \rangle_\sigma$ in (7) because each neuron in the i -th subpopulation carries a high Fisher information in the i -th dimension. $D = 2$ is a special case where no impairment of encoding occurs because the effect of a decrease of N_{code} is less pronounced. Interestingly, an increase in λ also yields an improvement in the encoding accuracy. This is a combined effect resulting from an increase in N_{code} on the one hand and the existence of D subpopulations, $D - 1$ of

which maintain their tuning widths in each dimension on the other hand. The discussion of $g_D(\lambda)$ leads to the following encoding strategy. For small λ , $\langle J_{ii} \rangle_\sigma$ increases rapidly, which suggests a fragmentation of the population into D subpopulations each of which encodes one feature with high accuracy, i.e., one tuning width in each subpopulation is small whereas the remaining tuning widths are broad. Like in the case discussed above, the theoretical limit of this method is a breakdown of the approximation of $\eta \equiv \text{const.}$ and the validity of (6) due to insufficient receptive field overlap.

4 Discussion and Outlook

We have discussed the effects of a variation of the tuning widths on the encoding accuracy obtained by a population of stochastically spiking neurons. The question of an optimal tuning strategy has turned out to be more complicated than previously assumed. More specifically, the case which focused most attention in the literature – radially symmetric receptive fields [5, 1, 9, 3, 13] – yields a worse encoding accuracy than most other cases we have studied: uniform populations with tuning curves which are not radially symmetric; distributions of tuning curves around some symmetric or non-symmetric tuning curve; and the fragmentation of the population into D subpopulations each of which is specialized in one stimulus feature.

In a next step, the theoretical results will be compared to empirical data on encoding properties of neural populations. One aspect is the existence of sensory maps which consist of neural subpopulations with characteristic tuning properties for the features which are represented. For example, receptive fields of auditory neurons in the midbrain of the barn owl have elongated shapes [6]. A second aspect concerns the short-term dynamics of receptive fields. Using single-unit recordings in anaesthetized cats, Wörgötter et al. [12] observed changes in receptive field size taking place in 50–100 ms. Our findings suggest that these dynamics alter the resolution obtained for the corresponding stimulus features. The observed effect may therefore realize a mechanism of an adaptable selective signal processing.

References

- [1] Baldi, P. & Heiligenberg, W. (1988) *Biol. Cybern.* **59**:313–318.
- [2] Deco, G. & Obradovic, D. (1997) *An Information-Theoretic Approach to Neural Computing*. New York: Springer.
- [3] Eurich, C. W. & Schwegler, H. (1997) *Biol. Cybern.* **76**: 357–363.
- [4] Eurich, C. W. & Wilke, S. D. (2000) *Neural Comp.* (in press).
- [5] Hinton, G. E., McClelland, J. L. & Rumelhart, D. E. (1986) In Rumelhart, D. E. & McClelland, J. L. (eds.), *Parallel Distributed Processing, Vol. 1*, pp. 77–109. Cambridge MA: MIT Press.
- [6] Knudsen, E. I. & Konishi, M. (1978) *Science* **200**:795–797.
- [7] Kuffler, S. W. (1953) *J. Neurophysiol.* **16**:37–68.
- [8] Lettvin, J. Y., Maturana, H. R., McCulloch, W. S. & Pitts, W. H. (1959) *Proc. Inst. Radio Eng. NY* **47**:1940–1951.
- [9] Snippe, H. P. & Koenderink, J. J. (1992) *Biol. Cybern.* **66**:543–551.
- [10] Wiggers, W., Roth, G., Eurich, C. W. & Straub, A. (1995) *J. Comp. Physiol. A* **176**:365–377.
- [11] Wilke, S. D. & Eurich, C. W. (1999) In Verleysen, M. (ed.), *ESANN 99, European Symposium on Artificial Neural Networks*, pp. 435–440. Brussels: D-Facto.
- [12] Wörgötter, F., Suder, K., Zhao, Y., Kerscher, N., Eysel, U. T. & Funke, K. (1998) *Nature* **396**:165–168.
- [13] Zhang, K. & Sejnowski, T. J. (1999) *Neural Comp.* **11**:75–84.