Biol. Cybern. 64, 95–105 (1990)

Biological Cybernetics

Self-organization of associative memory and pattern classification: recurrent signal processing on topological feature maps

P. Tavan, H. Grubmüller, and H. Kühnel

Physik-Department, Technische Universität München, James-Franck-Strasse, D-8046 Garching, Federal Republic of Germany

Received April 4, 1990/Accepted in revised form July 5, 1990

Abstract. We extend the neural concepts of topological feature maps towards self-organization of auto-associative memory and hierarchical pattern classification. As is well-known, topological maps for statistical data sets store information on the associated probability densities. To extract that information we introduce a recurrent dynamics of signal processing. We show that the dynamics converts a topological map into an auto-associative memory for real-valued feature vectors which is capable to perform a cluster analysis. The neural network scheme thus developed represents a generalization of non-linear matrix-type associative memories. The results naturally lead to the concept of a feature atlas and an associated scheme of self-organized, hierarchical pattern classification.

1 Introduction

Topological feature maps are ubiquitous in the brain (Knudsen et al. 1987). Such maps show up in a localization of cortical activity by sensory stimuli and are characterized by the fact that excitations on nearby positions of the cortical plane are caused by similar sensory signals. Examples are the tonotopic and retinotopic maps in the auditory and visual cortices, respectively. Detailed structures and contents of these maps cannot be genetically prespecified but evolve after birth and are structured by experiences.

The basic principles for the self-organization of topological feature maps from sensory input have been detected by v.d. Malsburg and Willshaw (1977); Willshaw and v.d. Malsburg (1976). They involve *(i)* competition of synaptic projections from a sensory part of the cortex onto the cortex area forming the map, *(ii)* competition among the neurons of the map for maximal response to a given signal and *(iii)* cooperation of neurons which are neighbours on the mapping cortex. To demonstrate the validity of these principles the authors quoted above have suggested a corresponding algorithm which, as an example, was capable to explain the self-organization of a retinotopic map.

A very simple algorithm for the adaptive formation of feature maps implementing these principles has been developed by Kohonen (Kohonen 1982a,b; 1984). Because of its simplicity, that algorithm not only allowed the derivation of valuable analytical results on the character of the evolving map but also has proven to be useful for a variety of applications as motor control in robotics, solution of complicated optimization problems or semantics (see e.g. Ritter and Schulten 1988a, b; Ritter and Kohonen 1989).

Topological feature maps are internal representations of the outside world as experienced by a sensory apparatus. The latter encodes events, objects or relations into feature combinations which are represented as spatio-temporal activity patterns of neurons on a sensory cortex (SC). Neural fibers connecting the SC to the mapping cortex (MC) evoke an initial activity on the MC which, due to competition of the MC neurons, localized itself in the region around the MC neuron of maximal initial response (cf. Fig. 1 for the network topology). Hence, an incoming feature combination is associated with the single MC neuron which exhibits maximal response and all feature combinations which evoke maximal response at that neuron define a class. Each class is characterized by a prototype feature combination. That is the one to which the classifying neuron is optimally tuned. Given some kind of metric, which allows to express similarity of feature combinations in terms of distances, the feature space is, thus, decomposed into as many small volumina centered at the prototype combinations as there are classifying MC neurons. Thus, a topological feature map provides a discretization of feature space.

Although one might be tempted to conclude from the above discussion, that topological feature maps could also be conceived as associative memories and statistical classifiers, that is actually not the case. First, since the number of MC neurons determines the size of the classes discretizing the feature space, both,

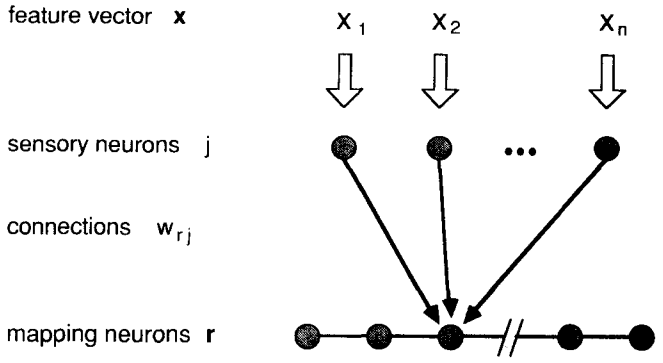

Fig. 1. Scheme of a network for self-organization of a topological feature map; the dimension n of the feature space is given by the number of sensory neurons

classification of feature combinations and their association to prototypes, strongly depend on discretization. Second, association and classification essentially proceed in terms of 'grandmother' neurons contrary to the well-known physiological evidence of distributed storage and of distributed neural activity in associative memories. Finally and most importantly, the type of association and classification characterized above does not exhibit any of the propeties generally required for unbiased statistical data analysis; these properties may be summarized as follows:

– For a set of n-dimensional statistical datá $\mathbf{x}$ characterized by a probability density $P(\mathbf{x})$ a classification without a priori bias has to be derived from the properties of $P(\mathbf{x})$. As is common in multi-variate analysis, classes have to be defined in terms of clusters of data, i.e., in terms of surroundings of local maxima of $P(\mathbf{x})$, and prototypes of classes in terms of averages over such clusters.

– Distance between clusters should enable a hierarchical classification. Thus, clusters of closely neighboring clusters which are well-separated from other clusters should combine to form superclasses and, only by closer inspection, decompose into subclasses.

In this article we show that feature maps become self-organizing statistical classifiers and auto-associative memories for n-dimensional feature vectors $\mathbf{x}$, which have all the desired properties sketched above, if an additional recurrent dynamics of signal processing between SC and MC is introduced.

After a short review of properties of topological feature maps in Sect. 2 we sketch in Sect. 3 the self-organization of neural connections which enable a regulated recurrent signal processing between SC and MC and introduce a corresponding algorithm. By methods of mathematical analysis we show in Sect. 4 that our recurrent dynamics converts topological feature maps into auto-associative memories and tools for cluster analysis. Results of simulations, which are presented in Sects. 5–8, illustrate the analytically derived properties and discuss effects of finite size and dimension of neural maps. Combining the results we develop in Sect. 9 the concept of a hierarchically organized feature atlas which consists of sequences of maps representing successively smaller portions of a feature space at a successively higher resolution. A short summary and discussion concludes the paper.

2 Topological feature maps

Topological feature maps resulting from Kohonen's algorithm form the basis of our construction. For a thorough understanding a short presentation and discussion of that algorithm is necessary.

In Kohonen's algorithm a feature combination is represented as an n-dimensional vector $\mathbf{x}$ composed of real numbers. A component x_j of that vector is interpreted as the activity of SC neuron j. Thus, as sketched in Fig. 1, the feature vectors $\mathbf{x}$ provide the input to a two-layered neural net consisting of the SC as input layer and of the MC as output layer. The MC neurons are labelled by a position vector $\mathbf{r}$ indicating their *physical position* within the cortical net. The two layers are fully interconnected by links of strengths w_{rj} which are combined to form n-dimensional weight vectors $\mathbf{w}_r$.

A weight vector $\mathbf{w}_r$ determines the response of MC neuron $\mathbf{r}$ to a signal $\mathbf{x}$ from the SC. Unlike in most neural network models the size of that response if not determined by the dot product $\mathbf{w}_r \cdot \mathbf{x}$ corresponding to a Hamming metric, but rather is given by the Euclidean distance $d_r(\mathbf{x})$ between weight vector $\mathbf{w}_r$ and feature vector $\mathbf{x}$

$$d_r(\mathbf{x}) = \|\mathbf{w}_r - \mathbf{x}\| . \tag{1}$$

Note, that Kohonen's algorithm is not confined to that simple Euclidean metric but is compatible also with more general metrics. Note, furthermore, that the use of such metrics instead of a Hamming metric is the first central point on which the new developments of this paper are based. An MC neuron is argued to exhibit strong response to a signal $\mathbf{x}$ if $d_r(\mathbf{x})$ is small. Hence, the weight vector $\mathbf{w}_r$ directly points to that position in the n-dimensional feature space to which MC neuron $\mathbf{r}$ is optimally tuned. Therefore, we call $\mathbf{w}_r$ the *virtual position* of MC neuron $\mathbf{r}$ in feature space.

To formally express the initial response $a_r^i(\mathbf{x})$ of an MC neuron $\mathbf{r}$ to a feature vector $\mathbf{x}$ one may choose any positive function of $d_r(\mathbf{x})$ peaked at $d_r(\mathbf{x}) = 0$ and decaying to zero over a characteristic distance ρ. For our simulations we have chosen a Gaussian

$$a_r^i(\mathbf{x}) = \exp(-d_r^2(\mathbf{x})/2\rho^2) , \tag{2}$$

although a linearly decaying or a step-like function, e.g.,

$$a_r^i(\mathbf{x}) = \begin{cases} 1 & \text{if } d_r(\mathbf{x}) < \rho , \\ 0 & \text{else} , \end{cases} \tag{3}$$

should serve the same purposes and should be computationally much more efficient in large scale calculations. We call the characteristic distance ρ the *selectivity parameter* as it determines the degree of fine tuning of MC neurons to incoming signals.

The initial response of MC neurons, which will be of central importance for our developments, plays a minor role in Kohonen's algorithm. Here, in order to achieve formation of a *topological* map, it is assumed to rapidly decay by a 'winner-takes-all' dynamics towards a final standard activity on the MC. Such dynamics can arise from long-range competition and short-range cooperation among MC neurons. The final activity $a_{\mathbf{r}}^{f}(\mathbf{r}')$ is centered around the MC neuron $\mathbf{r}'$ of largest initial response and decays on the MC over a characteristic distance σ. That distance measures the *range of cooperativity* among the MC neurons and is assumed to decrease during the formation of the feature map. Despite its computational inefficiency we have chosen a Gaussian for the final activity, too

$$a_{\mathbf{r}}^{f}(\mathbf{r}') = \exp[-(\mathbf{r}-\mathbf{r}')^2/2\sigma^2]\,. \tag{4}$$

It is important to note that the width σ determining the final activity refers to distances between *physical* positions $\mathbf{r}$ of MC neurons within the cortical net whereas the width ρ determining the initial activity refers to distances between *virtual* positions $\mathbf{w}_{\mathbf{r}}$ of MC neurons and feature vectors $\mathbf{x}$ within feature space.

The final activity enters a Hebbian learning rule for the update of weight vectors after presentation of a feature vector $\mathbf{x}$ chosen according to its corresponding a priori probability density $P(\mathbf{x})$

$$\mathbf{w}_{\mathbf{r}}^{\text{new}} = \mathbf{w}_{\mathbf{r}}^{\text{old}} + \epsilon a_{\mathbf{r}}^{f}(\mathbf{r}')[\mathbf{x} - \mathbf{w}_{\mathbf{r}}^{\text{old}}]\,. \tag{5}$$

Learning of weight vectors $\mathbf{w}_{\mathbf{r}}$ representing the map proceeds in discrete time steps $t = 0, 1, \ldots, t_{\max}$. In our simulations we have chosen $t_{\max} = 100 \cdot N^{\delta}$ where δ is the MC dimension and N is the number of MC neurons per cortex dimension. With increasing t the learning parameter ϵ and the range of cooperatively σ are decreased according to the formula $\alpha(t) = \alpha_{\max}(\alpha_{\min}/\alpha_{\max})^{t/t_{\max}}$ with $\alpha \in \{\epsilon, \sigma\}$. Useful values are $\epsilon_{\max} = .9$, $\epsilon_{\min} = .05$, $\sigma_{\max} = N/2$ and $\sigma_{\min} = 1$ (see Ritter and Schulten 1988a,b).

As a result of the self-organization process sketched above, the MC neurons span a topologically ordered, smooth, virtual net $\mathbf{W} \equiv \{\mathbf{w}_{\mathbf{r}}\}$ in feature space such that neighboring MC neurons $\mathbf{r}$ occupy also neighboring virtual positions $\mathbf{w}_{\mathbf{r}}$ within that virtual net. Thus, the mapping cortex has become a topological feature map. As shown by Ritter (1989) the point density $D(\mathbf{w}_{\mathbf{r}})$ of the virtual net $\mathbf{W}$ in feature space is a polynomial function of the probability density $P(\mathbf{x})$ of the feature vectors

$$D(\mathbf{w}_{\mathbf{r}}) \sim P(\mathbf{x})^{\gamma}\,, \tag{6}$$

with an exponent γ depending on cortex dimension δ and cooperativity range σ (for $\delta = 1$ the exponent γ is about 2/3). *Therefore, the point density $D(\mathbf{w}_{\mathbf{r}})$ of the virtual net W is a discretized, slightly deformed version of the probability density of feature combinations* $\mathbf{x}$. This is the second important property of topological feature maps on which our further arguments are based. No reference to the *topological* character of feature maps will be made. Kohonen's algorithm can render non-topological virtual nets obeying (6) if the requirement of cooperation among MC neurons is dropped, i.e., if the Gaussian in (4) is replaced by a δ-distribution (Ritter 1989). But a corresponding conventional scheme of vector quantization (Linde et al. 1980) exhibits much slower convergence, lacks biological relevance and those nice features of topological maps which enable simple graphical representations.

3 Recurrent signal processing

As exhibited by (6), the point density $D(\mathbf{w}_{\mathbf{r}})$ of the virtual net $\mathbf{W}$ attached to the feature map contains information on the structure of the probability density $P(\mathbf{x})$. We will now introduce a very simple algorithm which serves to extract that information from the map. The algorithm is based on the concept of regulated, recurrent signal processing between sensory and mapping cortex.

To enable recurrent signal processing we assume that after growth and self-organization of synaptic connections $w_{\mathbf{r}j}$ from SC neurons j towards MC neurons $\mathbf{r}$ the reverse process also occurs. We suggest that reverse synaptic connections $w_{j\mathbf{r}}$ from MC neurons towards SC neurons are formed according to the same Hebbian learning principle, cf. (5). As a result the connectivity between the two layers will become completely symmetrical, i.e., $w_{j\mathbf{r}} = w_{\mathbf{r}j}$. Due to the reverse connections, activities of MC neurons evoked by primary sensory signals will induce a secondary activity of SC neurons. We imagine that this recurrent signalling proceeds on a fast time scale. Correspondingly, we employ the initial MC activity expressed by (2) or (3) for a description of that process instead of the final activity Eq. (4) relevant for the formation of the map.

For regulation of the reverse signals we assume that a few additional neurons become linked to all MC neurons and to all reverse connections. The additional neurons are supposed to sum up the initial MC activity $a_{\mathbf{r}}^{i}$ caused by an incoming signal $\mathbf{x}(t)$. Depending on the size of the initial MC activity they are suggested to influence the weights of the reverse connections. A fast strengthening of the reverse connections is assumed if the activity evoked on the map is small and a fast weakening if it is large. Such regulation of synaptic weights could be achieved, e.g., by shunting inhibition. Figure 2 shows a scheme of a corresponding network.

Mathematically, the suggested regulation of the overall strength of the reverse signal corresponds to a normalization. Thus, the secondary activity $\mathbf{x}(t+1)$ of the SC neurons evoked by their primary activity $\mathbf{x}(t)$, mediated by the initial activity of the map and transmitted through the regulated reverse connections may be expressed

$$\mathbf{x}(t+1) = \frac{\Sigma_{\mathbf{r}}\, a_{\mathbf{r}}^{i}[\mathbf{x}(t)]\mathbf{w}_{\mathbf{r}}}{\Sigma_{\mathbf{r}}\, a_{\mathbf{r}}^{i}[\mathbf{x}(t)]}\,. \tag{7}$$

This equation defines a recurrent dynamics of signal processing between SC and MC. As we will demonstrate, that dynamics extends the range of applications of

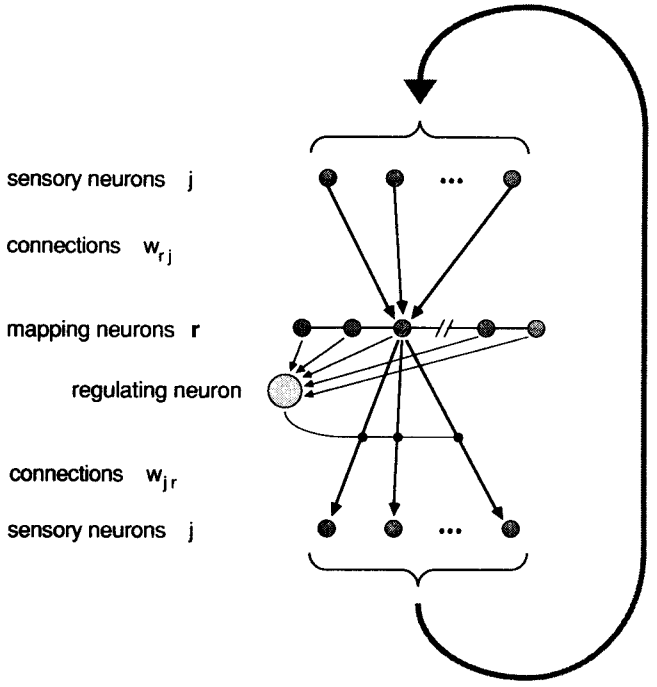

Fig. 2. Scheme of a network for recurrent signal processing on a topological feature map; for a most simple graphical representation of the network structure the sensory cortex has been duplicated; the large arrow indicates the identity of top and bottom layer

topological feature maps towards construction of self-organizing statistical classifiers and of auto-associative memories.

Note, that the network structure depicted in Fig. 2 resembles that of 'bidirectional associative memories' introduced by Kosko (1988). Upon closer analysis one may convince oneself, that our recurrent dynamics on topological feature maps represents a generalization of Kosko's concept since our similarity measure in feature space, which determines $a_r^i[\mathbf{x}(t)]$, is based on more general metrics than just the usual Hamming metric.

Note, furthermore, that the neural interpretation of the mechanism for regulation of the reverse signal is by no means unique. Instead of influencing the efficiency of the reverse connections, as formally suggested by Fig. 2, the regulating neurons also could be assumed to appropriately scale the neural activity on the mapping cortex. Both interpretations are compatible with (7). Our arguments are independent of a detailed neural interpretation; they solely rely on the validity of (7).

4 Properties of the recurrent dynamics

To explain the dynamics defined by (7) we first want to demonstrate that it is stable. Here, stability means that each initial activity pattern $\mathbf{x}(0)$ on the sensory cortex converges to a stable attractor $\mathbf{x}^p = \lim_{t\to\infty}\mathbf{x}(t)$. When regarded in terms of information processing such a process is an auto-association of a trial pattern $\mathbf{x}(0)$ to a prototype pattern $\mathbf{x}^p$.

The stability of our recurrent algorithm is most easily seen if one employs a step-like function as given by (3) to describe the initial activity of MC neurons and if one assumes the number of MC neurons to be large. In the corresponding continuum limit, the sums in (7) may be replaced by integrals and, using (6), the point density of the virtual net by the probability density of feature vectors. One obtains

$$\mathbf{x}(t+1) = \frac{\int_{S[\mathbf{x}(t)]} \mathrm{d}^n x' P^\gamma(\mathbf{x}')\mathbf{x}'}{\int_{S[\mathbf{x}(t)]} \mathrm{d}^n x' P^\gamma(\mathbf{x}')}, \tag{8}$$

where the volume $S[\mathbf{x}(t)]$ is an n-dimensional sphere of radius ρ centered at the primary feature vector $\mathbf{x}(t)$. According to (8) the updated SC activity $\mathbf{x}(t+1)$ is given in terms of a local average of the feature space weighted by the probability density. In comparison with the center $\mathbf{x}(t)$ of the sphere, the local average $\mathbf{x}(t+1)$ will be shifted towards regions of higher density. Such shifting will occur as long as there is a direction of larger probability density within the volume $S(\mathbf{x})$. Therefore, the recurrent dynamics entails a gradient ascent on a bounded, positive function $P_\rho(\mathbf{x})$ which is obtained by taking local averages of $P^\gamma(\mathbf{x})$ over volumes $S(\mathbf{x})$. In general $P_\rho(\mathbf{x})$ is a convolution of $P^\gamma(\mathbf{x})$ with the initial MC activity $a_r^i(\mathbf{x})$. The maxima $\tilde{\mathbf{x}}_i(\rho)$, $i = 1, \ldots, \nu(\rho)$, of $P_\rho(\mathbf{x})$ are the stable fixed points $\mathbf{x}_i^p(\rho)$ of the auto-associative dynamics $\mathbf{x}(0) \to \mathbf{x}_i^p(\rho)$. We call $P_\rho(\mathbf{x})$ the *effective potential* of the auto-associative process. Note, that the number $\nu(\rho)$ of different prototypes $\mathbf{x}_i^p(\rho)$ identified by the dynamics should monotonously increase with decreasing ρ.

To further elaborate these concepts, consider the extreme cases of very large and very small values of the selectivity parameter ρ. For very large values of ρ the spheres $S(\mathbf{x})$ will cover the complete feature space and, therefore, local averages will correspond to global averages. Then (8) renders the total average $\langle\mathbf{x}\rangle = \int \mathrm{d}^n x P^\gamma(\mathbf{x})\mathbf{x} / \int \mathrm{d}^n x P^\gamma(\mathbf{x})$ as a fixed point for any initial pattern $\mathbf{x}(0)$ after the first step. Hence, at very large values of ρ the auto-associative dynamics $\mathbf{x}(0) \to \mathbf{x}_1^p(\rho) = \langle\mathbf{x}\rangle$ identifies all feature combinations.

For very small values of ρ the effective potential $P_\rho(\mathbf{x})$ becomes identical with $P^\gamma(\mathbf{x})$, its maxima $\tilde{\mathbf{x}}_i(\rho)$ become identical with the maxima $\tilde{\mathbf{x}}_i$ of $P(\mathbf{x})$ and, hence, the dynamics corresponds to a gradient ascent on $P(\mathbf{x})$. For a most simple proof assume the feature space to be one-dimensional and $P(x)$ to be analytic. Expanding $P(x)$ into a Taylor series at $x(t)$ the integrals in (8) can be evaluated. Retaining terms up to first order one obtains

$$x(t+1) = x(t) + \frac{\gamma\rho^2}{3}\frac{P'[x(t)]}{P[x(t)]}, \tag{9}$$

where $P'(x)$ is the derivative of $P(x)$. Equation (9) proves that the dynamics actually is a gradient ascent on $P(x)$. Thus, for very small values of the selectivity parameter ρ the prototype $x_i^p(\rho)$ associated to an initial feature $x(0)$ is the local maximum $\tilde{x}_i$ of $P(x)$ which is closest to $x(0)$ in the direction selected by the gradient $P'[x(0)]$; at small values of ρ the auto-associative dynamics most selectively differentiates the various feature combinations.

As a consequence of the fact, that the recurrent dynamics represents gradient ascent on the effective potential $P_\rho(\mathbf{x})$, also minima of $P_\rho(\mathbf{x})$ are fixed points of the dynamics. However, these fixed points are unstable and, correspondingly, their basins of attraction are of measure zero. In our extended simulations we never happened to hit one of these spurious states as an initial point of the associative process. Therefore, these points are neglected in future discussions.

Summarizing we may state that (7) defines a sequence of auto-associative dynamics $\mathbf{x}(0) \to \mathbf{x}_i^\rho(\rho)$ for the various scales of distance in feature space which are given by the respective values of the selectivity parameter ρ. For small ρ the prototypes are given by the local maxima of $P(\mathbf{x})$, at intermediate ρ sets of local maxima clustering within a distance ρ define prototypes at a coarser scale of differentiation, whereas at large ρ eventually all patterns are identified. Hence, when considered as a function of ρ, (7) provides a scheme for hierarchical pattern classification.

The properties, which we have just derived, apply to the case of a very fine discretization of the feature space by the virtual net. In the remainder of the paper we will present the results of simple simulations which aim at an illustration of these properties. However, simulations have to rely on a limited number of MC neurons such that some of the properties will be modified by *discretization effects*. Furthermore, in numerical calculations convergence of a pattern $\mathbf{x}(t)$ towards its prototype $\mathbf{x}^\rho$ has to be judged employing a threshold criterion. In our simulations we have used a small number $\theta_1 = 10^{-4}$ to define prototypes according to

$$\mathbf{x}_i^\rho \equiv \mathbf{x}(t) \text{ if } \|\mathbf{x}(t) - \mathbf{x}(t+1)\| < \theta_1 . \tag{10}$$

In case of a shallow effective potential $P_\rho(\mathbf{x})$ the speed of convergence may become very small and a threshold criterion like the one given above may mimic stability and, therefore, may generate numerical artifacts. As a result meta-stable points or points close to a locally shallow maximum of the effective potential may become erroneously identified as fixed points of the dynamics. To exclude to some extent the latter type of artifacts we have identified closely spaced prototypes in our calculations

$$\mathbf{x}_1^\rho \equiv \mathbf{x}_2^\rho \quad \text{if} \quad \|\mathbf{x}_1^\rho - \mathbf{x}_2^\rho\| < \theta_2 . \tag{11}$$

As an estimate for the threshold θ_2 we have employed the optimal size of the discretization of the feature space provided by the virtual net, i.e., the minimum of the distances $\|\mathbf{w}_\mathbf{r} - \mathbf{w}_{\mathbf{r}'}\|$ between neighboring virtual positions of MC neurons $\mathbf{r}$ and $\mathbf{r}'$.

5 Hierarchical classification

Figure 3 provides a first example for the capability of our algorithm to perform a hierarchical classification of features. We have selected a one-dimensional feature space and a one-dimensional mapping cortex consisting of a chain of 100 neurons. For the self-organization of the map the features x were chosen at random from the interval [0, 1] according to the probability density $P(x)$ depicted at the top of Fig. 3. $P(x)$ is composed of four identical Gaussian distributions $g_i(x)$, $i = 1, \ldots, 4$, characterized by a standard deviation σ of 0.05. Gaussians g_1 and g_2 as well as g_3 and g_4 form two identical local clusters. The distances of maxima within the clusters measure 3σ and are smaller by a factor of two than the distance between the maxima of Gaussians g_2 and g_3 which belong to different clusters.

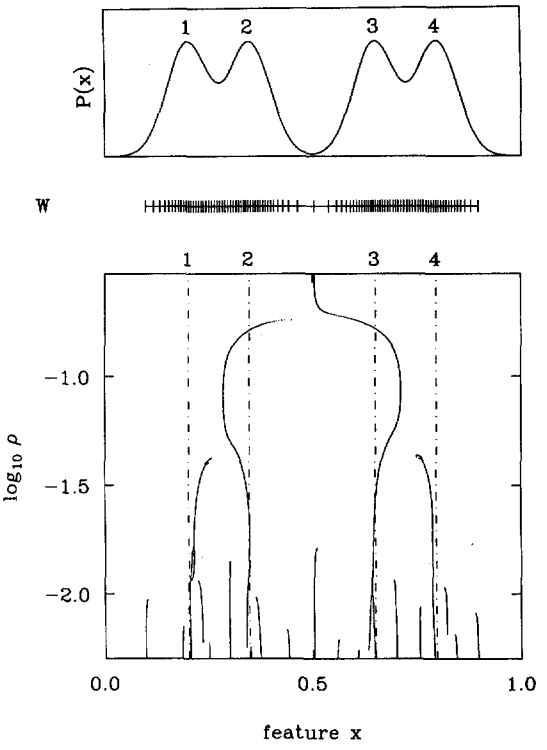

Fig. 3. Hierarchical pattern classification by recurrent signal processing for a one-dimensional feature space; *top:* probability density $P(x)$; *center:* virtual net **W** for a one-dimensional mapping cortex of 100 neurons; *bottom:* location of prototypes $x_i^\rho(\rho)$ in feature space as a function of the selectivity parameter ρ; for discussion see text

The virtual net **W** resulting after 10^4 learning steps is shown in Fig. 3 below the graph of $P(x)$. The virtual positions $w_\mathbf{r}$ of the MC neurons are marked by vertical bars at the respective positions within the feature space [0, 1]. By inspection one may convince oneself, that the point density $D(w_\mathbf{r})$ of virtual positions in feature space actually provides a mapping of $P(x)$.

The particular form of $P(x)$ has been chosen so as to encode four prototype features x_i^ρ by the locations $\tilde{x}_i$ of the maxima of $P(x)$ and a hierarchical class structure of features and prototypes by the selected distances between the maxima. When viewed on a coarse scale, the structure of $P(x)$ suggests a combination of the four feature classes $\mathscr{C}_i$ defined by the g_i into two superclasses $\mathscr{S}_1 = \mathscr{C}_1 \cup \mathscr{C}_2$ and $\mathscr{S}_2 = \mathscr{C}_3 \cup \mathscr{C}_4$. The prototypes of the superclasses are then the averages of the prototypes of the respective subclasses from which they are composed.

According to the analysis presented in Sect. 3, the auto-associative dynamics given by (7) should be perfectly capable to reveal the hierarchical structure of

feature classes encoded by $P(x)$. The bottom part of Fig. 3 proves that this is actually the case. That part of the figure depicts the locations $x_i^\rho(\rho)$, $i = 1, \ldots, \nu(\rho)$, of the prototypes within feature space as a function of the selectivity parameter ρ. The locations of the prototypes have been determined according to the prescriptions given in (10) and (11) using 21 different trial features $x_\alpha(0)$ regularly distributed onto the feature space for each value of ρ. For comparison the locations $\tilde{x}_i$ of the maxima of $P(x)$ are indicated by dashed-dotted lines. To the extent at which the continuum limit provides a valid approximation for the 100 neurons case considered here, the locations of the prototypes should approach the $\tilde{x}_i$ at small ρ.

At values $\rho > 0.184$ the auto-associative dynamics renders only one fixed point located at about the mean value $\langle x \rangle = 0.5$ of the whole distribution. In the range $0.044 < \rho < 0.184$ the dynamics identifies two prototypes labeling the two superclasses $\mathscr{S}_1$ and $\mathscr{S}_2$ introduced above. Four prototypes x_i^ρ are identified in the range $0.016 < \rho < 0.044$ and, as expected, these prototypes are located close to the respective maxima $\tilde{x}_i$ of $P(x)$. At the borders between the ranges the classification scheme exhibits bifurcations. Hence, as claimed further above, the recurrent dynamics given by (7) is able to provide a hierarchical classification of features if the selectivity ρ of the initial response of the MC neurons is taken as parameter for tuning classification.

The values ρ_b at which the classification exhibits *bifurcations* provide a measure for the proximity of the feature classes. At the chosen parameters the algorithm can resolve two Gaussian peaks of equal height and standard deviation σ only if the distance δ between the respective maxima approximately exceeds 3ρ. That estimate for the bifurcation value ρ_b of the selectivity parameter has been derived analytically evaluating the convolution of $P^\gamma(x)$ with the initial MC activity $a_r^i(x)$ given by (2). Differentiating $P_\rho(x)$ one finds the bifurcation value $\rho_b = \sqrt{(\delta/2)^2 - \sigma^2/\gamma}$. This analytical result for ρ_b is in perfect agreement with the values determined from the simulations.

For two Gaussians of different height or standard deviation one can derive implicit equations for ρ_b. As compared to ρ_b for Gaussians of equal height and standard deviation one finds that ρ_b becomes smaller if one of the heights is reduced or one of the widths is increased (see next section for examples).

Increase of the selectivity of the MC neurons beyond a critical lower bound ρ_c, which is about 0.016 in the case considered here, leads to a rapidly increasing number of predicted classes. These classes are spurious and due to the fact that at ρ_c the selectivity parameter becomes smaller than the discretization of the feature space by the virtual net. Under these conditions the initial activity of the mappig cortex essentially involves a single neuron [cf. (2)] and the virtual position of that neuron becomes an attractor of the auto-associative dynamics. An upper bound for the *discretization limit* ρ_c is provided by the maximum of the distances $\|\mathbf{w}_\mathbf{r} - \mathbf{w}_{\mathbf{r}'}\|$ between the virtual positions of MC neurons $\mathbf{r}$ and $\mathbf{r}'$ which are neighbours in the virtual net.

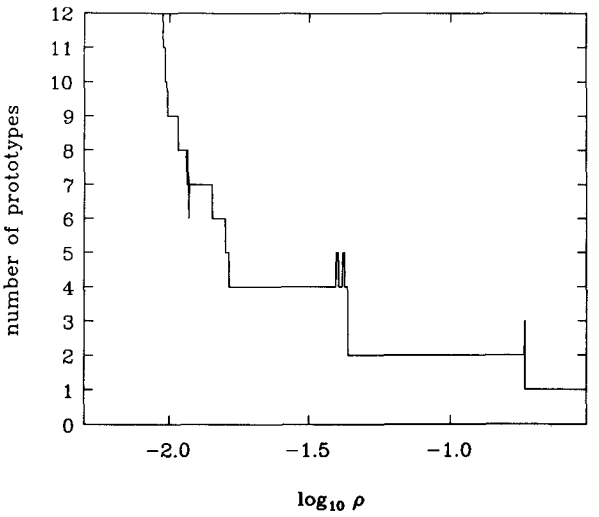

Fig. 4. Classification graph for the example presented in Fig. 3; the number of different prototypes identified by the auto-associative dynamics is plotted as a function of the selectivity parameter ρ

At selectivity parameters ρ close to the bifurcation values ρ_b, a *critical slowing down* of the speed of convergence towards the fixed points was observed in the simulations. This observation indicates that the maxima of the effective potential $P_\rho(x)$ become very flat at these values of ρ. Correspondingly, the prescriptions given by (10) and (11) for determination of fixed points render artifacts near ρ_b. That effect is illustrated in Fig. 4 which shows a *classification graph*. A classification graph represents the number $\nu(\rho)$ of prototypes as a function of ρ. The graph clearly identifies the two large ranges of ρ in which the algorithm, as shown in Fig. 3, identifies two and four prototypes, respectively. At the two bifurcation points the graph exhibits small spikes caused by misclassifications upon critical slowing down. The rapid increase of $\nu(\rho)$ for ρ smaller than 0.016 marks the discretization limit ρ_c.

For high-dimensional feature spaces a direct visualization of the bifurcation pattern of prototypes is impossible. In these cases classification graphs like the one shown in Fig. 4 provide a most important tool to judge classification and its hierarchical structure stored in a given feature map.

6 Discretization effects

To illustrate further properties of our classification scheme we have chosen a two-dimensional feature space as a second example. Feature vectors $\mathbf{x}$ are chosen from the rectangle $[0, 1] \times [0, 0.5]$ according to the probability distribution $P(\mathbf{x})$ shown in Fig. 5. $P(\mathbf{x})$ consists of four bivariate Gaussians of different shape and height. Therefore, that distribution encodes four prototype features associated with classes of features differing in frequency and variance. Two of the classes represent highly frequent features. These classes correspond to the Gaussian peaks depicted in the upper left and right

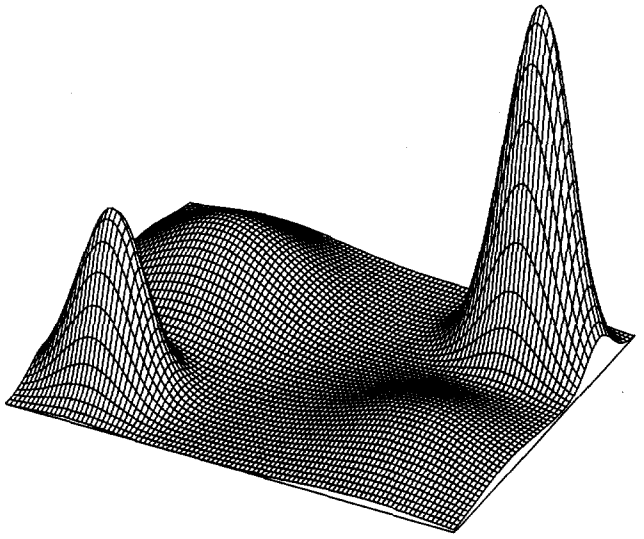

Fig. 5. Sample probability density $P(\mathbf{x})$ within the two-dimensional feature space $[0, 1] \times [0, 0.5]$

corners. Two classes of less frequent features are located in the lower corners. Like in the first example, the selected choice of distances between maxima groups the four classes into two superclasses with a corresponding partition perpendicular to the long axis of the feature space.

We present results for two different neural maps of $P(\mathbf{x})$ in order to discuss how the size of cortical maps influences their classification properties. Figure 6 shows the virtual nets of a very coarse map comprising 18 MC neurons (top) and of a finer map constructed from 200 MC neurons (bottom). For each of the two cases, we have considered a one-dimensional cortex topology in order to provide a simple example for a mapping of a higher-dimensional feature space onto a lower-dimensional cortex. Similar simulations with two-dimensional cortices have shown, that cortex topology actually is irrelevant for classification. In addition to the virtual nets, Fig. 6 exhibits as patterns of black pixels the sets of feature vectors employed for training of the respective maps. The probability density shown in Fig. 5 is reflected in the density of black pixels.

Figure 7 shows the classification graphs obtained for the two feature maps. To determine the number $v(\rho)$ of prototypes at a given value of the selectivity parameter ρ, convergence of the auto-associative dynamics given by (7) has been monitored for 231 different trial feature vectors $\mathbf{x}_\alpha(0)$. The trial vectors were chosen from a regular 21×11 grid covering the feature space.

At large values of the selectivity parameter ($\rho > 0.1$) the classification graphs of the two maps are very similar. Therefore, as far as the statistical analysis of the coarse structure of $P(\mathbf{x})$ is concerned, coarse and fine discretizations are equivalent. For both maps a first bifurcation of prototype identification occurs at $\rho_b \approx 0.25$ which is the value expected from our analytical estimate of ρ_b introduced in Sect. 5. That bifurcation corresponds to a partition of the feature space into

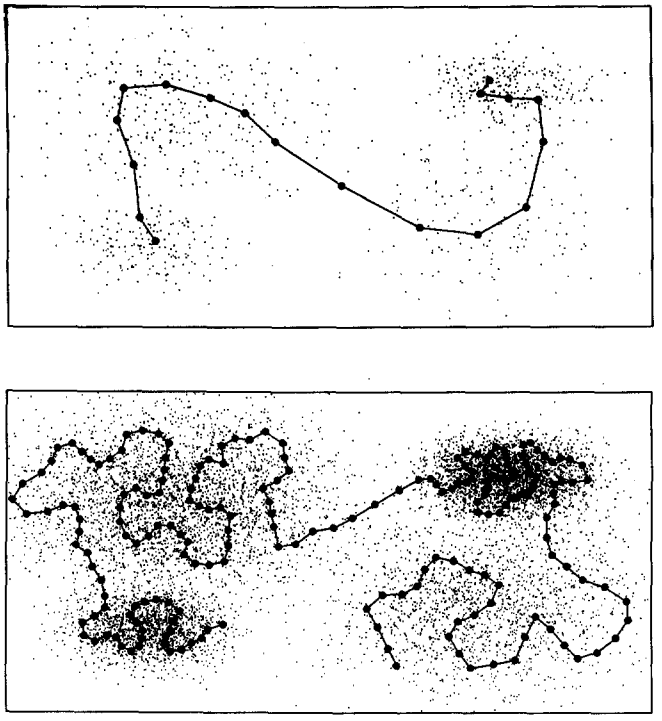

Fig. 6. Virtual nets **W** constructed in the two-dimensional feature space $[0,1] \times [0, 0.5]$ for one-dimensional mapping cortices; lines between virtual nodes (fat dots) indicate nearest neighbour relations on the mapping cortices; black pixels indicate the feature vectors used for training of the maps; *top:* 18 MC neurons; *bottom:* 200 MC neurons

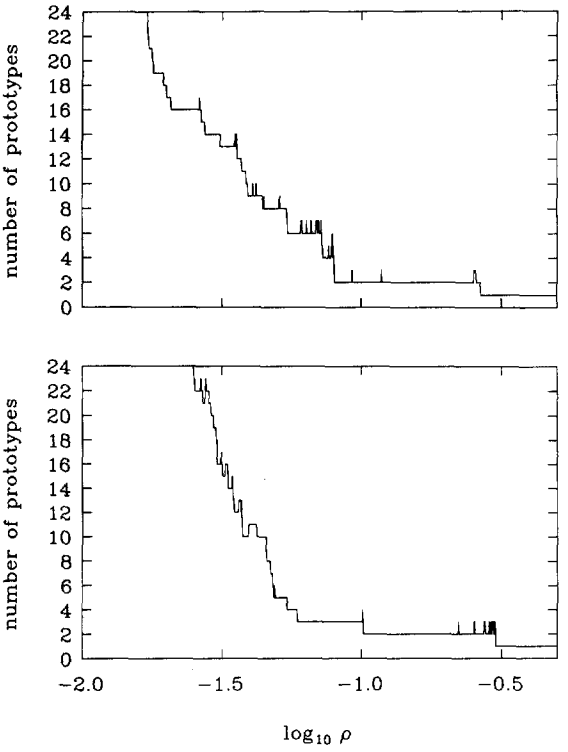

Fig. 7. Classification graphs for the maps shown in Fig. 6; *top:* coarse discretization based on 18 MC neurons; *bottom:* finer discretization with 200 MC neurons; see text and caption to Fig. 4 for further explanations

two parts perpendicular to its long axis. Thus, the two superclasses encoded by $P(\mathbf{x})$ are detected first.

However, upon approach to the discretization limit of the coarse map, classifications by the two maps become markedly different. As demonstrated by the graph at the top of Fig. 7, the small map immediately switches from identification of the two superclasses to identification of all four classes. The corresponding bifurcations occur at a value ρ_b of about 0.08 which is approximately a third of the distance between the maxima of the subclasses. Due to the coarse discretization that value somewhat deviates from the analytical estimate for ρ_b which applies to the continuum limit. According to the estimate the bifurcation pattern of the subclasses should proceed for the given $P(\mathbf{x})$ (Fig. 5) in two steps with a larger ρ_b at about 0.089 and a smaller ρ_b at 0.065. The larger ρ_b corresponds to the distinction of the subclasses in the left half of feature space and the smaller ρ_b to that in the right half.

Because of the proximity of the discretization limit ρ_c at 0.072, identification of four prototypes is highly unstable for the coarse map. Generally, classification becomes unreliable if the value of ρ is of the same order of magnitude as ρ_c. Therefore, inference on the number of prototypes actually encoded by $P(\mathbf{x})$ upon inspection of a classification graph is limited to a scale provided by ρ_c.

Correspondingly, the classification graphs of the finer map at the bottom of Fig. 7 renders an improved inference on the structure of $P(\mathbf{x})$ possible. That graph exhibits a wide range of the parameter ρ above the corresponding discretization ρ_c at 0.054 which indicates the existence of the three prototypes expected in that range and, hence, allows the safe conclusion that the corresonding classes actually do exist. However, also for the larger map the range at which all four existing classes are identified is too close to ρ_c as to allow their inference merely from consideration of the graph. As argued in Sect. 3, improvement of classification towards safe identification of all existing classes requires a reduction of the discretization limit by further increase of the size of the map.

7 Basins of attraction

According to Sect. 4 the associative dynamics entails a gradient ascent on the effective potential $P_\rho(\mathbf{x})$. Therefore, a graphical representation of many different trajectories of the dynamics can reveal the structure of $P_\rho(\mathbf{x})$, the location of its maxima and the basins of attraction.

To provide an example we consider the auto-associative dynamics on a moderately sized, two-dimensional map for the two-dimensional $P(\mathbf{x})$ shown in Fig. 5. The map comprises 20×10 MC-neurons and the corresponding virtual net is shown in Fig. 8.

Figure 9 shows 231 trajectories $\mathbf{x}_\alpha(t)$ of the auto-associative dynamics for a value $\rho = 0.08$ of the selectivity parameter. A regular grid of 21×11 trial feature vectors has been chosen for the starting points $\mathbf{x}_\alpha(0)$ of the dynamics.

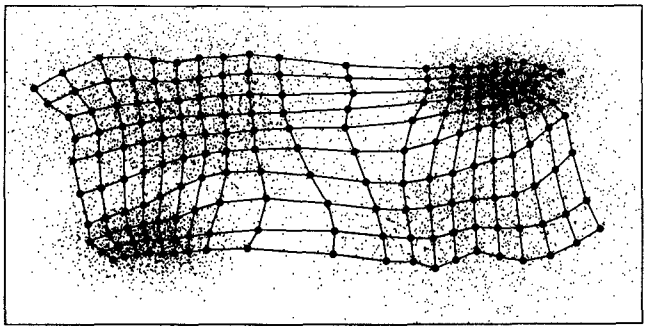

Fig. 8. Virtual net **W** constructed in the two-dimensional feature space $[0, 1] \times [0, 0.5]$ for a two-dimensional MC comprising 20×10 neurons; for further explanation see caption to Fig. 6

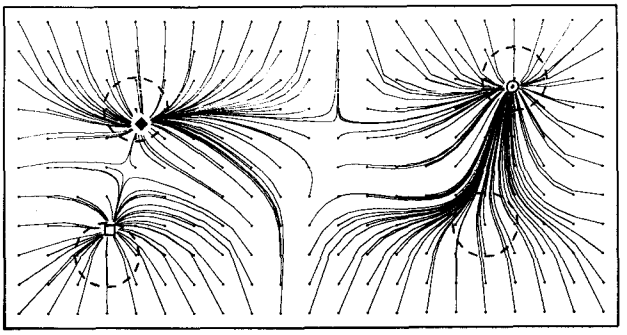

Fig. 9. Feature space trajectories $\mathbf{x}(t)$ of the auto-associative dynamics for a value $\rho = 0.08$ of the selectivity parameter employing the two-dimensional virtual net shown in Fig. 8; see text for further explanations

At the chosen value of ρ three different prototypes are identified by the dynamics. Thus, the two-dimensional map of 200 MC neurons considered here provides a similar classification as the one-dimensional map of the same size (cf. bottom of Fig. 7), which illustrates the independence of classification from cortex topology. Figure 9 reveals the three prototypes as end points of bundles of trajectories. All three prototypes are located near one of the maxima of $P(\mathbf{x})$ which in the figure are marked by large circles. Each prototype is surrounded by a basin of attraction forming a connected region in feature space. The size of the respective basin of attraction is determined by the frequency of features in the corresponding class or superclass. At the chosen value of ρ the shallow and broad maximum of $P(x)$, which is located in the lower right corner of the feature space, does not give rise to a separate maximum of the effective potential $P_\rho(\mathbf{x})$. Instead, as exhibited by the shape of the trajectories, that maximum of $P(\mathbf{x})$ generates a ridge in the effective potential $P_\rho(\mathbf{x})$ which funnels the trajectories towards the much higher and much more strongly peaked maximum of $P(\mathbf{x})$ in the upper right corner. In contrast, the two close maxima of $P(\mathbf{x})$ in the other part of the feature space, one being broad but intense, the other being weak but strongly peaked, both give rise to separate maxima of $P_\rho(\mathbf{x})$.

For high-dimensional feature spaces a direct visualization of the basins of attraction is impossible.

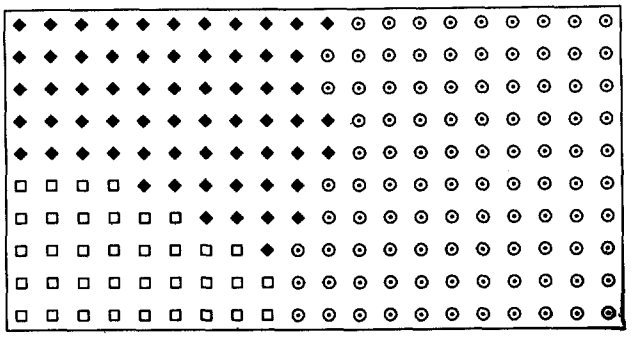

Fig. 10. View on the *physical* positions of the 20 × 10 neurons of the MC; the neurons are classified by the auto-associative dynamics at $\rho = 0.08$ using the *virtual* positions $\mathbf{w}_\mathbf{r}$ as starting points and are labeled by the classifying prototypes employing the symbols of Fig. 9

However, the topological character of the feature map entails the possibility to draw a map for the *probability weighted* sizes of these basins. For that purpose one may choose, instead of a regular grid, the virtual positions $\mathbf{w}_\mathbf{r}$ of the MC neurons $\mathbf{r}$ as starting points $\mathbf{x}_\alpha(0)$ of the auto-associative dynamics (cf. Fig. 8). The dynamics will classify each MC neuron $\mathbf{r}$ by that prototype whose basin of attraction contains $\mathbf{w}_\mathbf{r}$. Simple labeling of MC neurons by the associated prototypes will then reveal the relative frequencies of features in the corresponding classes by inspection of the labeled cortex since the point density of the virtual net is a measure for $P(\mathbf{x})$.

To provide an example, Fig. 10 shows a view on the regular 20 × 10 grid of the MC with its neurons labeled by their associated prototype features. Although parameter values and symbols coding prototypes are identical to those of Fig. 9, note that the two figures display different spaces. For understanding consider the difference between *virtual* positions of neurons in feature space (Fig. 8) and *physical* positions of neurons on the MC (Fig. 10). According to the figure 100 MC neurons are associated with each of the two superclasses encoded by $P(\mathbf{x})$. That partition of neural resources is the expected result since the total probabilities of the two superclasses had been chosen identical.

8 High-dimensional feature spaces

In the examples discussed above one- and two-dimensional feature spaces have been employed since they allow simple graphical representations. However, realistic problems of pattern recognition usually involve feature spaces of very high dimension. To demonstrate that our hierarchical scheme of pattern classification can also cope with somewhat more realistic problems we have selected a mapping of a five-dimensional feature space onto a one-dimensional cortex for our last example. As feature space we have chosen the five-dimensional hypercube $[0, 1]^5$ and as mapping cortex a chain of 120 neurons. In order to encode six different prototypes into the sample probability density $P(\mathbf{x})$ we have centered six Gaussian distributions g_i of standard deviation $\sigma = 0.05$ at the corners of a regular simplex. The distance between the corners of the simplex was 0.7. The choice of such a regular simplex ensures that the six clusters of probability density completely span the five-dimensional space.

Figure 11 shows the classification graph obtained by monitoring convergence of the auto-associative dynamics for 59,049 trajectories $\mathbf{x}_\alpha(t)$. For each value of ρ the trial feature vectors $\mathbf{x}(0)$ have been located on a regular grid of 9^5 starting points covering the hypercube. Figure 11 allows a clearcut inference on the existence of six prototypes. The six existing prototypes are safely identified over a wide range of the selectivity parameter ρ extending from $\rho_b \approx 0.25$ down to $\rho_c \approx 0.10$. Instead of stepwise bifurcations in prototype detection expected for non-symmetric probability densities, the high symmetry of the sample distribution entails a simultaneous onset of identification of all six prototypes at the "bifurcation" value ρ_b. In that parameter range the usual critical slowing down of the auto-associative dynamics and the corresponding numerical artifacts by misclassifications show up in the spikes of the classification graph. Since the 'bifurcation' actually is a 'hexfurcation', the critical slowing down is strongly enhanced. The discretization limit ρ_c of the auto-associative dynamics is clearly indicated by the rapid increase of calculated fixed points at values of ρ smaller than ρ_c.

These results indicate that our algorithm should be well-suited to analyze unknown distributions of high-dimensional patterns using low-dimensional topological maps. In the case considered here, each class of features has been described by about 20 MC neurons. We have checked that even a much poorer discretization involving only 5 MC neurons per class on a one-dimensional cortex still renders a satisfactory classification and auto-association of the five-dimensional patterns. Success of

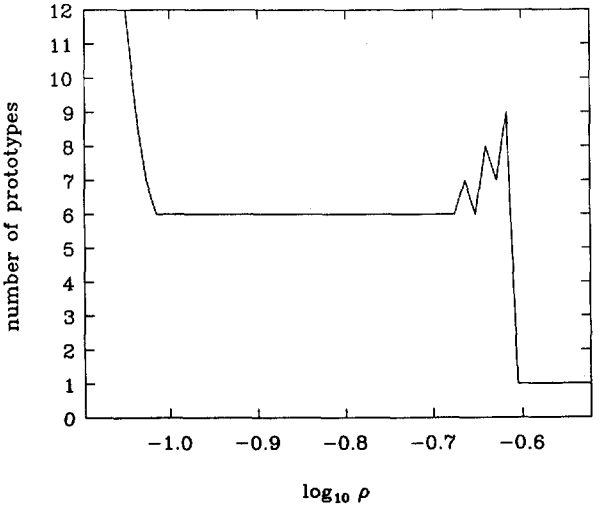

Fig. 11. Classification graph for a one-dimensional MC of 120 neurons which map a probability density $P(\mathbf{x})$ in a five-dimensional feature space; $P(\mathbf{x})$ encodes six prototypes located at the corners of a regular simplex

such poor discretization depends on the existence of sharp and well-separated maxima of $P(\mathbf{x})$.

9 Feature atlas

In a statistical data set prototypes may occur at strongly different distances. Our auto-associative dynamics can safely identify and discriminate classes only if the selectivity parameter ρ is smaller than the distance between corresponding prototypes and larger than the discretization limit ρ_c below which classification breaks down. ρ_c is determined by the size of the map. Therefore, if one wants to identify all existing classes, one must choose a very large number of MC neurons. Variation of ρ then renders the desired hierarchical classification.

However, the mapping strategy sketched above closely resembles the attempt to draw a single map of the complete surface of the earth. For a sufficiently high resolution such map would have to be of enormous size. Furthermore, to its largest part the map would be completely uninteresting since it covers oceans, deserts and so on. Therefore, geographic mapping commonly is provided by an *atlas* which contains a set of maps hierarchically ordered according to the employed scales. Large scale maps provide gross overviews whereas small scale maps reveal the details of 'interesting' regions.

We now want to show that our scheme for pattern recognition and classification naturally leads to the mapping strategy of an atlas. For that purpose we imagine that in a first step a primary feature map for the probability density $P(\mathbf{x})$ of a statistical data set is formed which comprises just enough MC neurons of low selectivity (large ρ) as to safely identify some of the coarse prototypes. As an example consider the virtual $\mathbf{W}$ spanned by a two-dimensional MC of 5×10 neurons which is depicted at the top of Fig. 12 and maps the probability density shown in Fig. 5. At a low selectivity of the MC neurons characterized by a value 0.15 of ρ the auto-associative dynamics discriminates the two superclasses encoded by $P(\mathbf{x})$.

In a second step we imagine that the coarse classification of feature vectors $\mathbf{x}$ by the primary map selectively can steer formation of additional maps. Each of these secondary maps will then be confined to the basin of attraction of the respective coarse prototype. Therefore, the basic resolution of these maps will exceed that of the primary map even without expansion of size. As a result, prototype recognition at a coarse scale can initialize classification at a finer scale if the MC neurons of the secondary maps respond with an increased selectivity to a presented pattern $\mathbf{x}$. Iteration of that procedure will render a hierarchical tree of maps, a feature atlas, which can discriminate patterns up to any desired resolution. It is not very difficult to devise neural circuits and self-organized learning schemes which implement such a hierarchical mapping strategy.

The two secondary virtual nets shown at the bottom of Fig. 12 illustrate the concepts developed above. For

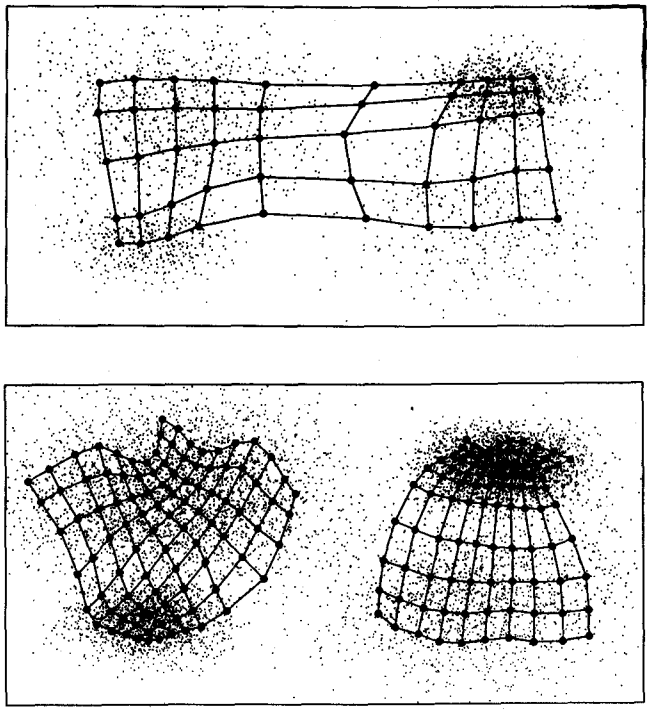

Fig. 12. Three virtual nets representing a simple example for a feature atlas; feature space and probability density are shown in Fig. 5; *top:* primary virtual net constructed from 5×10 MC neurons for identification of two superclasses at a value $\rho_s = 0.15$; *bottom:* secondary virtual nets mapping the basins of attraction of the two superclasses; each of the nets comprises 10×10 MC neurons

self-organization of these nets feature vectors $\mathbf{x}$ were first partitioned into two superclasses by auto-associative dynamics on the primary map shown at the top of the figure. Depending on the result of that primary classification each of the secondary nets has been trained only with members of its associated superclass. Therefore, the basins of attraction of the superclasses are selectively mapped by the two virtual nets. The secondary maps shown are of sufficiently high resolution as to safely identify all existing subclasses by the associative dynamics.

10 Summary and discussion

We have developed a general neural network scheme for self-organization of auto-associative memories and of classifiers for real valued patterns. The scheme employs topological feature maps as its building blocks. Such maps consist of a sensory cortex (SC) feeding its activity patterns into a mapping cortex (MC). For our scheme we have extended that concept by self-organizing feedback connections from MC to SC and by mechanisms for regulation of neural activity. As a result we have obtained a recurrent dynamics of signal processing which converts topological feature maps into auto-associative memories. We have shown that these networks become tools for hierarchical cluster analysis in feature space upon variation of the selectivity parameter ρ.

The selectivity parameter measures the response characteristics of MC neurons to an activity pattern on the SC. At large values of ρ many MC neurons respond to a given SC activity whereas at small values of ρ only the few MC neurons become active which are well-tuned to the particular signal. Therefore, other cortical areas by sending non-specific signals to the MC could change its background activity and, thereby, steer the value of ρ.

Upon adjustment of ρ an activity pattern repeatedly presented to an SC can become hierarchically classified. Increase of ρ corresponds to an inductive sequence of associations which starts at a highly specific classification and leads towards increasingly general notions, whereas decrease of ρ entails deductive associations. In contrast to these capabilities of large single maps, classification by the feature atlas introduced above is restricted to deduction.

The most prominent features of the algorithms employed in our scheme are computational simplicity and stability. Concerning their function as associative memories, our networks represent a generalization of Kosko's 'bidirectional associative memories' which in turn have been a generalization of Hopfield networks (see Kosko 1988, for discussion and further reference). Problems with spurious states of the kind occurring in such non-linear matrix-type associative memories are absent since our algorithms do not rely on the conventional Hamming metric but rather are based on Euclidean and even more general metrics in feature space. It has been the use of these metrics which has enabled our extension towards hierarchical clustering.

Finally we would like to note, that, upon presentation of time series on the SC and by Hebbian learning of non-symmetric connections between SC and MC, our scheme may easily be extended towards associative recall of limit cycles and towards autoregression (Kühnel 1990).

Acknowledgements. The authors would like to thank J. Rubner and J. Buhmann for helpful discussions. Valuable critiques of the referee enabled considerable improvements of the text. Support by the Deutsche Forschungsgemeinschaft (SFB143-C1) is gratefully acknowledged.

References

Knudsen EI, du Lac S, Esterly SD (1987) Computational maps in the brain. Ann Rev Neurosci 10:41–65

Kohonen T (1982a) Self-organized formation of topologically correct feature maps. Biol Cybern 43:59–69

Kohonen T (1982b) Analysis of a simple self-organizing process. Biol Cybern 44:135–140

Kohonen T (1984) Self-organization and associative memory. Springer, Berlin Heidelberg New York

Kosko B (1988) Bidirectional associative memories. IEEE Trans Syst Man Cybern 18:49–60

Kühnel H (1990) Diplomarbeit, Physik-Department, Technische Universität München

Linde Y, Buzo A, Gray RM (1980) An algorithm for vector quantizer design. IEEE Trans Comm 28:84–95

Malsburg C von der, Willshaw DJ (1977) How to label nerve cells so that they can interconnect in an ordered fashion. Proc Natl Acad Sci USA 74:5176–5178

Ritter H (1989) Asymptotic level density for a class of vector quantization processes. Technical Report, University of Helsinki

Ritter H, Kohonen T (1989) Self-organizing semantic maps. Biol Cybern 61:241–254

Ritter H, Schulten K (1988a) Extending Kohonen's self-organizing mapping algorithm to learn ballistic movements. In: Eckmiller R, Malsburg C von der (eds) Neural computers. Springer, Berlin Heidelberg New York, pp 393–406

Ritter H, Schulten K (1988b) Kohonen's self-organizing maps: exploring their computational capabilities. IEEE ICNN 88 Conference, San Diego, pp 109–116

Willshaw DJ, Malsburg C von der, (1976) How patterned neural connections can be set up by self-organization. Proc R Soc London B 194:431–445

Dr. Paul Tavan
Physik-Department
Theoretische Physik
Technische Universität München
James-Franck-Strasse
D-8046 Garching
Federal Republic of Germany