8. Self-organizing Maps and Adaptive Filters

Helge Ritter, Klaus Obermayer, Klaus Schulten and Jeanne Rubner

With 13 Figures

Synopsis. Topographically organized maps and adaptive filters fulfill important roles for information processing in the brain and are also promising to facilitate tasks in digital information processing. In this contribution, we report results on two important network models. A first network model comprises the "self-organizing feature maps" of Kohonen. We discuss their relation to optimal representation of data, present results of a mathematical analysis of their behavior near a stationary state, demonstrate the formation of "striped projections", if higher-dimensional feature spaces are to be mapped onto a two-dimensional cortical surface, and present recent simulation results for the somatosensory map of the skin surface and the retinal map in the visual cortex. The second network model is a hierarchical network for principal component analysis. Such a network, when trained with correlated random patterns, develops cells the receptive fields of which correspond to Gabor filters and resemble the receptive fields of "simple cells" in the visual cortex.

8.1 Introduction

One essential task of neural-network algorithms is optimal storage of data. Different criteria for optimal storage are conceivable, and correspondingly different neural-network algorithms have been derived. Many of them fall into one of two major, and to some extent complementary, categories.

The first category is that of so-called *attractor networks* [8.1]. Such networks are fully connected: information is stored in a distributed way and retrieved by a dynamical relaxation process. The distributed storage mechanism makes these systems very tolerant to partial damage or degradation in their connectivity but also introduces a tendency for "crosstalk" between similar patterns [8.2, 3]. This type of storage does not reduce the information content of patterns stored. In fact, it stores prototype patterns completely, e.g. as pixel images, and allows classification of presented patterns according to the stored prototypes.

The second category is formed by so-called *competitive learning networks*, in which a set of "grandmother cells" is used for storage of the presented patterns [8.4, 5]. Such networks involve an input and output layer of "grandmother cells" and storage is achieved through the development of receptive fields which resemble stored patterns. The receptive fields act as filters: when a pattern similar

to one of the patterns in a training set is offered, the output cell the receptive field of which best matches the input becomes activated. Since a single cell provides the network response, such systems lack any tolerance against hardware failure, but they avoid crosstalk between patterns of even very high overlap. Although each "grandmother cell" might appear as a fully localized storage device, part of the information is actually distributed: the "grandmother cell" selected by the input pattern is only determined by competition among many candidate cells and, therefore, depends crucially on information from many different cells in the network.

The usable storage capacity of both types of network is similar and can be brought close to the information inherent in the required weight values (see e.g. [8.6, 7]). Generalization or "associative completion" of partial inputs is also very similar: in the absence of any special preprocessing the stored pattern of maximal overlap with the presented input is usually retrieved.

While attractor networks have been investigated very much in recent years, competitive learning networks have received less attention. There are many non-trivial and interesting properties of competitive networks that deserve more study. This is particularly true for a generalization of these networks where weight adjustments of "grandmother cells" lose their independence and are mutually coupled in some prespecified way. These networks, introduced by Kohonen under the name *"self-organizing feature maps"* [8.8–10], possess properties which make them particularly interesting for both understanding and modeling the biological brain [8.11, 56–58] and for practical applications such as robotics [8.12, 13].

In the following, we will present several mathematical results pertaining to Kohonen networks and review some work concerning the application of Kohonen networks to modeling of neural tissue in the cortex. A more comprehensive account can be found in [8.14].

Another important issue for optimal storage, relevant to both types of model discussed above, is efficient preprocessing of information. It is, of course, most desirable to achieve dense information storage through filters which rapidly discern the important features of input data and restrict storage to these features. In the visual system of biological species such filters operate on the lowest levels of the system in the optical cortex and extract important visual features such as edges and bars (see e.g. [8.15, 16]). An answer to the question how the brain achieves the neural connectivity which establishes optimal filters for preprocessing is extremely desirable for the development of neural-network algorithms for computer vision and other information-processing tasks characterized by huge amounts of data. Only a small part of the architecture of the brain is genetically specified; most of the brain's synaptic connections are achieved through self-organization. Postnatal visual input plays an essential role in the organization of synaptic patterns of the optical cortex of mature animals (see e.g [8.17]). These observations raise the question of how a sensory system, in response to input information, can organize itself so as to form feature detectors which encode mutually independent aspects of the information contained in patterns presented to it. In Sect. 8.6 we will present a two-layered network as a model for such

system. It will be demonstrated that simple local rules for synaptic connectivities allow the model to learn in an unsupervised mode. Presented with a set of input patterns the network learns to discern the most important features defined as the principal components of the correlation matrix of a set of training patterns. The network described generalizes a model of Linsker [8.18] which established the possibility of self-organized formation of feature detectors.

8.2 Self-organizing Maps and Optimal Representation of Data

The basic aim of "competitive networks" as well as of "self-organizing maps" is to store some, usually large, set V of patterns, encoded as "vectors" $\boldsymbol{v} \in V$, by finding a smaller set W of "prototypes" $\boldsymbol{w}_r$ such that the set $W := \{\boldsymbol{w}_{r_1}, \boldsymbol{w}_{r_2}, \ldots \boldsymbol{w}_{r_N}\}$ of prototypes provides a good approximation of the original set V. Intuitively, this should mean that for each $\boldsymbol{v} \in V$ the distance $||\boldsymbol{v} - \boldsymbol{w}_{s(v)}||$ between $\boldsymbol{v}$ and the closest prototype $\boldsymbol{w}_{s(v)}$ in the set W shall be small. Here, the "mapping function" $s(.)$ has been introduced to denote for each $\boldsymbol{v} \in V$ the (index of the) closest prototype in W. (Therefore, the function $s(.)$ depends on all prototypes $\boldsymbol{w}_r$ and is equivalent to a full specification of their values: given $s(.)$, one can reconstruct each $\boldsymbol{w}_r$ as the centroid of the subset of all $\boldsymbol{v} \in V$ for which $s(\boldsymbol{v}) = r$.)

For a more precise formulation of the notion "good approximation of V" we assume that the pattern vectors $\boldsymbol{v}$ are subject to a probability density $P(\boldsymbol{v})$ on V, and then require that the set of prototypes should be determined such that the *expectation value* E of the square error,

$$E[\boldsymbol{w}] = \int ||\boldsymbol{v} - \boldsymbol{w}_{s(v)}||^2 P(\boldsymbol{v}) \mathrm{d}^d \boldsymbol{v} \tag{8.1}$$

is minimized. Here $\boldsymbol{w} := (\boldsymbol{w}_{r_1}, \boldsymbol{w}_{r_2}, \ldots \boldsymbol{w}_{r_N})$ represents the vector of prototypes in W.

Minimization of the functional $E[\boldsymbol{w}]$ is the well-known problem of *optimal vector quantization* [8.19,20] and is related to *data compression* for efficient transmission of the pattern set V: if an average error $E[\boldsymbol{w}]$ can be tolerated, sender and receiver can agree to use the mapping function $s(.)$ to transmit only the (usually much shorter) "labels" $s(\boldsymbol{v})$ of the approximating prototypes $\boldsymbol{w}_s$ instead of the complete pattern vectors $\boldsymbol{v}$ themselves.

A straightforward approach to find a local minimum of (8.1) is gradient descent for the functional $E[.]$, i.e. the prototypes are changed according to

$$\dot{\boldsymbol{w}}_r = \int_{s(v)=r} (\boldsymbol{v} - \boldsymbol{w}_{s(v)}) P(\boldsymbol{v}) \mathrm{d}^d \boldsymbol{v} \ . \tag{8.2}$$

Equation (8.2) is equivalent to the discrete "learning rule"

$$\Delta \boldsymbol{w}_r = \varepsilon \delta_{r,s(v)} (\boldsymbol{v} - \boldsymbol{w}_r) \ , \tag{8.3}$$

applied for a sequence of random "samples" $\boldsymbol{v} \in V$ that are distributed according to the probability density $P(\boldsymbol{v})$ in the limit of vanishing "step size" ε. Equation (8.3) is a well-known "learning rule" found in many competitive networks: for each "presentation" of an input $\boldsymbol{v}$, $s(\boldsymbol{v})$ "selects" the best-matching prototype vector $\boldsymbol{w}_{s(\boldsymbol{v})}$, for an adjustment towards $\boldsymbol{v}$. For rapid convergence, one usually starts with a larger initial learning step size $\varepsilon_i < 1$, which is gradually lowered to a final, small value $\varepsilon_f \geq 0$.

The *self-organizing feature maps* [8.8–10] generalize this scheme for optimal storage (in the sense of minimal average error $E[\boldsymbol{w}]$, (8.1)) by considering the prototypes $\boldsymbol{w}_r$ to be associated with points $\boldsymbol{r}$ in some "image domain" A and requiring that *a "structured representation" or "map" of the data V is created on A during the storage process.* The "map" arises through the selection function $s(\boldsymbol{v})$, which maps each pattern vector to a point $\boldsymbol{s} \in A$. The discrete set A is endowed with some *topology*, e.g. by arranging the points as a (often two-dimensional) lattice. The aim of the algorithm of self-organizing feature maps then is, besides approximating V by the prototypes $\boldsymbol{w}_r$, also to arrange the $\boldsymbol{w}_r$ in such a way that the associated mapping $s(.)$ from V to A maps the topology of the set V, defined by the metric relationships of its vectors $\boldsymbol{v} \in V$, onto the topology of A in a least distorting way. This requires that (metrically) *similar patterns v are mapped onto neighboring points in A.* The desired result is a (low-dimensional) image of V in which *the most important similarity relationships among patterns from V are preserved and transformed into spatial neighborhood relationships in the chosen "image domain".*

To achieve this, the adjustments of the prototypes $\boldsymbol{w}_r$ must be coupled by replacing the Kronecker δ_{rs} in (8.3) by a "neighborhood function" h_{rs},

$$\Delta \boldsymbol{w}_r = \varepsilon h_{rs(\boldsymbol{v})}(\boldsymbol{v} - \boldsymbol{w}_r) \,. \tag{8.4}$$

The function h_{rs} is (usually) a unimodal function of the lattice distance $d = ||\boldsymbol{r} - \boldsymbol{s}||$ in A, decaying to zero for $d \to \infty$ and with maximum at $d = 0$. A suitable choice, for example, is a Gaussian $\exp(-d^2/2\sigma^2)$. Therefore, vectors $\boldsymbol{w}_r$, $\boldsymbol{w}_s$ associated with *neighboring* points $\boldsymbol{r}$, $\boldsymbol{s} \in A$ are coupled more strongly than vectors associated with more distant points and tend to converge to more similar patterns during learning. This mechanism enforces a good "match" between the topologies of V and A on a local scale. Consequently, it is no longer $E[\boldsymbol{w}]$ but instead the functional

$$F[\boldsymbol{w}] = \sum_{rr'} h_{rr'} \int\limits_{s(\boldsymbol{v})=r'} ||\boldsymbol{v} - \boldsymbol{w}_r||^2 P(\boldsymbol{v}) \mathrm{d}^d \boldsymbol{v} \tag{8.5}$$

that is (approximately) minimized by the new process. Equation (8.5) for a discrete set V has been stated in [8.21] and there it was shown that its minimization is related to the solution of the "traveling-salesman problem". A more general interpretation was subsequently given by Luttrell [8.22]. He considers again the case when the prototypes in W are used to obtain "labels" $s(\boldsymbol{v})$ to compress the pattern set V for the purpose of transmission, but in addition assumes that this

transmission is "noisy", i.e. there is a probability of h_{rs} that label s is confused with label r as a result of the transmission process. Then the reconstruction of a transmitted pattern v will not always be the closest prototype $w_{s(v)}$ but will be w_r with a probability of $h_{rs(v)}$. Hence the expected mean square error on transmitting a pattern v will be given by $\sum_r h_{rs(v)}(v - w_r)^2$ and $F[w]$ can thus be seen to represent *the expected mean square transmission error over the noisy channel*, averaged over the whole pattern set V.

In Sect. 8.4 we will return to this interpretation and discuss the relationship between some aspects of brain function and optimization of $F[w]$.

8.3 Learning Dynamics in the Vicinity of a Stationary State

After clarifying the optimization task underlying the formation of self-organizing maps, the next task is to characterize the convergence properties of the map formation process, based on (8.4). For a more detailed account of the mathematical analysis the reader is referred to [8.23].

First, one needs to address the question how far convergence to a global minimum can be achieved. Even for simple distributions $P(v)$, the functional $F[w]$ can exhibit many different local minima [8.59, 60]. As the adaptation equation (8.4) is based on a gradient descent procedure, one generally cannot hope to find the global optimum but must be content with some more or less optimal local minimum. The function h_{rs} plays an important role in finding a good minimum [8.59, 60]. Inspection of (8.4) shows that in the long-range limit (i.e. $h_{rs} = \text{const}$) the functional $F[w]$ approaches a simple quadratic function with a single minimum. Therefore, by starting an adaptation process with a long-ranged h_{rs} and then reducing the range of h_{rs} slowly to the intended, smaller final value, one gradually deforms the "optimality landscape" from a simple, convex shape to a final, multi-minimum surface. Such a strategy facilitates convergence to a good minimum of $F[.]$. The choice of a good starting configuration is also helpful in this respect. As will be pointed out in Sect. 8.4, the optimization process can be interpreted as a model for the formation of topographically ordered neural projections in the brain. In this case, a good starting configuration is provided by an initial coarse ordering of the neural connectivity.

We focus in the following on the behavior of the optimization process in the vicinity of a good minimum of F. For the mathematical analysis we consider an ensemble of systems, each characterized by a set $w = (w_{r_1}, w_{r_2}, \ldots w_{r_N})$, $r_j \in A$, of prototype vectors. $S(w, t)$ shall denote the distribution of the ensemble, after t adaptation steps, in the state space Ω spanned by the prototype sets w. We make the assumption that the pattern vectors v are statistically independent samples from V, subject to the probability density $P(v)$. In this case (8.4) defines a Markov process in the space Ω with transition probability

$$Q(w, w') = \sum_r \int_{F_r(w')} \mathrm{d}v \, \delta\Big(w - w' - \varepsilon h_{r,s'(v)}(v - w')\Big) P(v) \tag{8.6}$$

for the transition from a state $\boldsymbol{w}'$ to a state $\boldsymbol{w}$. $F_r(\boldsymbol{w}')$ denotes the set of all $\boldsymbol{v}$ which are closer to $\boldsymbol{w}'_r$ than to any other $\boldsymbol{w}'_s$, $s \neq r$, and $s'(\boldsymbol{v})$ is defined in analogy to $s(\boldsymbol{v})$, but using the primed reference vectors $\boldsymbol{w}'_r$ instead. $F_r(\boldsymbol{w}')$ and $s'(.)$ are related: the former is the set of patterns $\boldsymbol{v} \in V$ mapped onto the same point $r \in A$ by the latter.

Assuming a small learning step size ε and a distribution function $S(\boldsymbol{w}, t)$ that is concentrated in the vicinity of the selected local optimum $\bar{\boldsymbol{w}}$ of $F[.]$, one can derive the following Fokker–Planck equation for the time development of S:

$$\frac{1}{\varepsilon}\partial_t S(\boldsymbol{u}, t) = \sum_{rmr'n} \frac{\partial}{\partial \mathrm{u}_{rm}} B_{rmr'n} \mathrm{u}_{r'n} S(\boldsymbol{u}, t) + \frac{\varepsilon}{2} \sum_{rmr'n} D_{rm,r'n} \frac{\partial^2 S(\boldsymbol{u}, t)}{\partial \mathrm{u}_{rm} \partial \mathrm{u}_{r'n}}, \tag{8.7}$$

where we have tacitly shifted the origin of $S(., t)$ to the selected state $\bar{\boldsymbol{w}}$, using now the new argument variable $\boldsymbol{u} = \boldsymbol{w} - \bar{\boldsymbol{w}}$ instead of $\boldsymbol{w}$. $\boldsymbol{B}$ is a matrix given by

$$B_{rmr'n} := \left(\frac{\partial V_{rm}(\boldsymbol{w})}{\partial \mathrm{w}_{r'n}} \right)_{\boldsymbol{w}=\bar{\boldsymbol{w}}}, \tag{8.8}$$

and the quantities V_{rm} and $D_{rmr'n}$ are the expectation values $\langle -\delta \boldsymbol{w}_{rm} \rangle$ and $\langle \delta \boldsymbol{w}_{rm} \delta \boldsymbol{w}_{r'n} \rangle$ under one adaptation step, where $\delta \boldsymbol{w}_{rm}$ is the change of the mth component of prototype $\boldsymbol{w}_r$, but scaled to $\varepsilon = 1$. Their explicit expressions are are given in [8.23].

With the help of (8.7) one can answer an important question about the convergence of the process: Which control of the learning step size ε guarantees that each ensemble member converges (with probability 1) to the optimum $\bar{\boldsymbol{w}}$? It turns out that, if $\bar{\boldsymbol{w}}$ is a stable stationary state (a necessary and sufficient condition for this is that $\boldsymbol{B} + \boldsymbol{B}^{\mathrm{T}}$ be a positive definite matrix), the two conditions

$$\lim_{t \to \infty} \int_0^t \varepsilon(t') \mathrm{d}t' = \infty, \tag{8.9}$$

$$\lim_{t \to \infty} \varepsilon(t) = 0 \tag{8.10}$$

provide the desired answer [8.23]. Equation (8.10) is rather obvious, but (8.9) sets a limit on the rate of reduction of the learning step size. If ε decreases faster than allowed according to condition (8.9) there is some nonzero probability that the adaptation process "freezes" before the local optimum is reached.

In reality, of course, (8.9) can never be met exactly; however, for $\varepsilon_f = 0$, (8.7) can be used to show that the remaining deviations from the optimum are exponentially small in the quantity $\int \varepsilon(t) \mathrm{d}t$ [8.23].

A second important question concerns the *statistical fluctuations* that are present for a non-vanishing learning step size and that are due to the randomness of the sequence of input patterns $\boldsymbol{v}$. If ε = const, (8.7) admits as a stationary solution a Gaussian with correlation matrix

$$\langle u_{rm} u_{sn} \rangle = \langle (w_{rm} - \bar{w}_{rm})(w_{sn} - \bar{w}_{sn}) \rangle_S = \varepsilon \left[(B + B^{\mathrm{T}})^{-1} D \right]_{rm,sn} \quad (8.11)$$

where $\langle \ldots \rangle$ denotes averaging over the ensemble. For an explicit evaluation of the matrices B and D one needs to know the configuration $\bar{w}$ of the local optimum chosen for the discussion. Even for simple distributions $P(v)$ there can be numerous and complex configurations $\bar{w}$ leading to a local minimum of $F[.]$. To study a simple, but still interesting, case we assume that A is a two-dimensional lattice of $N \times N$ points and that the set V is continuous and of higher dimensionality. We choose V as the three-dimensional volume $0 \leq x, y \leq N$, $-s \leq z \leq s$ and assume a constant probability density $P(v) = [2sN^2]^{-1}$. Clearly, for a topology-preserving mapping between V and A there exists some "conflict" because of the different dimensions of A and V. The parameter s can be considered a measure of this conflict: if s is small (i.e. $2s \ll N$), a "perpendicular" projection obviously provides a very good match between the topologies of V and of A (Fig. 8.1a). However, this is no longer the case if s becomes larger. Then minimization of $F[.]$ is expected to require a more complicated mapping from V to A (Fig. 8.1b).

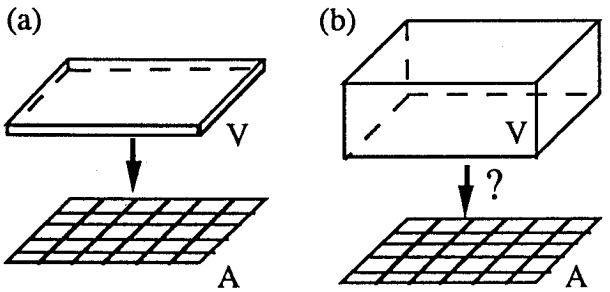


Fig. 8.1. (a): For small height $2s$ of the volume V, a "perpendicular" projection of V onto the lattice A provides a good match between the topologies of V and A. (b) For larger values of s the mapping $V \mapsto A$ required to minimize the functional $F[.]$ (8.5) is no longer obvious

We want to find the limiting value of s for which the "perpendicular" projection loses its optimality and to discuss in which way the mapping can be "improved" then.

To avoid any edge effects, we assume periodic boundary conditions for the lattice A and for V along the x- and y-axes. One can then evaluate (8.11) for $\bar{w}_r = r$, $r = me_x + ne_y$, which, by symmetry, must be a stationary solution for F (as well as any other configuration obtained by translations or rotations of $\bar{w}$). This choice for $\bar{w}$ corresponds to the mapping $\bar{s}(v) = \mathrm{nint}(v_x)e_x + \mathrm{nint}(v_y)e_y$ ($\mathrm{nint}(x)$ = nearest integer to x), i.e. to a perpendicular projection suppressing the v_z-coordinate. Besides $P(v)$ and $\bar{w}$, the remaining important determinant of the behavior of the system is the function h_{rs} that defines the coupling between different lattice points. A simple choice is the Gaussian

$$h_{rr'} = \sum_s \delta_{r+s,r'} \exp\left(-\frac{s^2}{2\sigma^2}\right), \quad (8.12)$$

with lateral width σ, for which we will require $1 \ll \sigma \ll N$. Owing to the translational invariance, both $D_{rmr'n}$ and $B_{rmr'n}$ depend only on the difference

$\boldsymbol{r} - \boldsymbol{r}'$ and on m, n. Therefore, we can decouple (8.7) if we represent $S(\boldsymbol{u}, t)$ in terms of the Fourier mode amplitudes

$$\hat{\boldsymbol{u}}(\boldsymbol{k}) = \frac{1}{N} \sum_{\boldsymbol{r}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} \boldsymbol{u}_{\boldsymbol{r}} \tag{8.13}$$

of $\boldsymbol{u}$, where $\boldsymbol{k} = (l/2\pi N, m/2\pi N)$ is a two-dimensional wave vector of the lattice A. Each mode amplitude turns out to be distributed independently, and its fluctuations can be calculated explicitly by separating (8.7) into a set of independent equations for each mode. The exact result is fairly complicated (for details, see [8.23]), but if one neglects "discretization effects" due to the finite lattice spacing and uses axes parallel ($||$) and perpendicular ($\perp$) to the wave vector instead of the fixed x- and y-directions, one can bring the mean square value of the equilibrium fluctuations of the different Fourier modes into a simpler form:

$$\langle \hat{u}_{\perp}(\boldsymbol{k})^2 \rangle = \varepsilon\pi\sigma^2 \frac{\exp(-k^2\sigma^2)}{12(1 - \exp(-k^2\sigma^2/2))} , \tag{8.14}$$

$$\langle \hat{u}_{\parallel}(\boldsymbol{k})^2 \rangle = \varepsilon\pi\sigma^2 \frac{(12k^2\sigma^4 + 1)\exp(-k^2\sigma^2)}{12 - 12(1 - k^2\sigma^2)\exp(-k^2\sigma^2/2)} , \tag{8.15}$$

$$\langle \hat{u}_3(\boldsymbol{k})^2 \rangle = \varepsilon\pi\sigma^2 \frac{s^2 \exp(-k^2\sigma^2)}{3 - s^2k^2 \exp(-k^2\sigma^2/2)} . \tag{8.16}$$

Figures 8.2–4 compare the theoretical prediction (curves) and data points from a Monte Carlo simulation of the process on a 32×32-lattice for the square roots $f_{1,2,3} = \langle \hat{u}^2_{\perp,||,3} \rangle^{1/2}$ of the mode fluctuations for $\varepsilon = 0.01$. To make the Monte Carlo simulation computationally more feasible, $h_{\boldsymbol{rs}}$ was not chosen according to (8.12), but instead as unity for $\boldsymbol{r} = \boldsymbol{s}$ and all nearest-neighbor pairs $\boldsymbol{r}, \boldsymbol{s}$ in the lattice and zero otherwise. This corresponds roughly to $\sigma = 1$ in (8.12), but the corresponding theoretical predictions are somewhat different from (8.14–16) (they are given in [8.23]); however, all essential features discussed below remain. Each mode can be interpreted as a periodic distortion of the equilibrium mapping. The first set of modes ($\hat{\boldsymbol{u}}_{\perp}$, Fig. 8.2) represents distortions that are "transverse" to their direction of periodicity, while the second set of modes ($\hat{\boldsymbol{u}}_{||}$, Fig. 8.3) represents distortions that are "longitudinal". For both kinds of mode, the fluctuations increase with increasing wavelength. This is to be expected, since the "restoring force" for modes with very long wavelengths is determined by the boundary conditions, which are assumed periodic, and, therefore, allow an arbitrary translational shift ($k = 0$-mode).

The most interesting set of modes are those perpendicular to the xy-directions ($\hat{u}_3$). These modes describe fluctuations of the prototypes $\boldsymbol{w}_{\boldsymbol{r}}$ along the additional dimension, which is "lost" in the "perpendicular" projection $\bar{s}(.)$ associated with the equilibrium configuration $\bar{\boldsymbol{w}}$. For values $s \ll \sigma$, inspection of (8.16) shows that the amplidude of these modes is of the order of s and, therefore, is small for any $\boldsymbol{k}$. This indicates that, although some information is lost, for $s \ll \sigma$ the mapping defined by $\bar{\boldsymbol{w}}$ cannot be improved by small distortions. However,

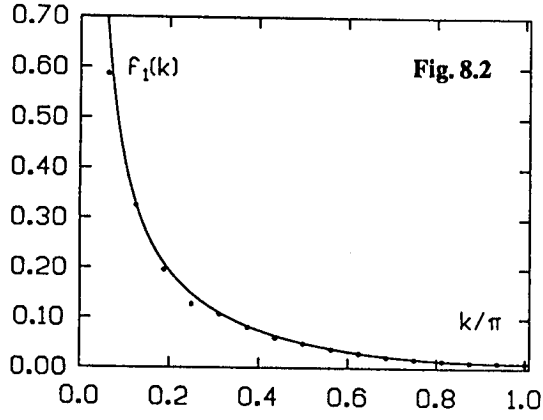


Fig. 8.2

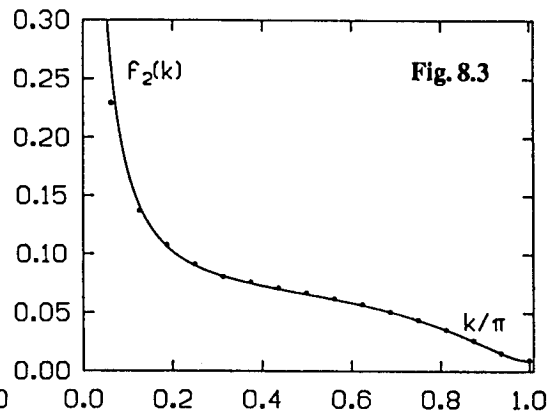


Fig. 8.3

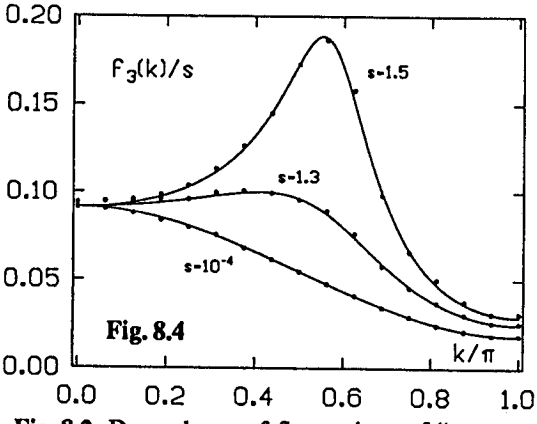


Fig. 8.4

Fig. 8.2. Dependence of fluctuations of "transverse mode" $u_\perp$ on the wave number k. The "neighborhood function" was chosen to be $h_{rs} = 1$ for $r = s$ and all nearest-neighbor pairs r, s and zero otherwise. The data points were obtained from a Monte Carlo simulation with 10000 samples of the Markov process (8.4) for fixed $\varepsilon = 0.01$ and $s = 10^{-4}$. The curve represents the analytical result

Fig. 8.3. Fluctuations of the "longitudinal mode" $u_{||}$ of the same simulation as in Fig. 8.2 above. For small wave numbers the fluctuations are smaller than for u_1

Fig. 8.4. Fluctuations of the "perpendicular mode" u_3 for three different values of the thickness parameter s: for $s = 10^{-4}$, i.e. essentially a two-dimensional input distribution, only small fluctuations arise. For $s = 1.3$ fluctuations begin to exhibit a broad maximum at about $k^* = 0.58\pi$, which becomes very pronounced for $s = 1.5$, i.e. a value closely below the critical value $s^* = \sqrt{12/5}$

as s increases, (8.16) shows that for s close to a "threshold value" of $s^* = \sigma\sqrt{3e/2} \approx 2.02\sigma$ the denominator can become very small for $||\boldsymbol{k}||$-values in the vicinity of $||\boldsymbol{k}|| = k^* = \sqrt{2}/\sigma$, and correspondingly large fluctuations are exhibited by these modes. Finally, at $s = s^*$, all modes with $||\boldsymbol{k}|| = k^*$ become unstable: s has become so large that the mapping $\bar{\boldsymbol{s}}(.)$ has lost its optimality and can be further optimized if the prototypes $\boldsymbol{w}_r$ assume a wavelike "modulation" along their $\boldsymbol{w}_{r3}$-direction. The characteristic wavelength of this modulation is $\lambda^* = \sigma\pi\sqrt{2} \approx 4.44\sigma$ [8.21]. For $s > s^*$, a whole "window" of unstable modes appears. This is also discernible in Fig. 8.4, where the different choice of the function h_{rs}, however, leads to changed values of $s^* = \sqrt{12/5} \approx 1.54$ and $k^* \approx 0.58\pi$ (for $\boldsymbol{k}$ directed along the x-direction).

We can summarize now the following answer to our initial question: the simple, "perpendicular" projection $\bar{\boldsymbol{s}}(.)$ is the optimal mapping as long as s, its

maximal "error" in the vertical direction, is below a value $s^* = \sigma\sqrt{3e/2}$. In this case, apart from fluctuations, all prototypes $\boldsymbol{w}_r$ have the same value of w_{r3}. The threshold value s^* *can be interpreted as being the distance in the space* V *that corresponds to the range of the "neighborhood function"* h_{rs} *in the lattice* A. For $s > s^*$, the "perpendicular" projection can be optimized further by distortions. These distortions arise from the components w_{r3}, which now must vary with $\boldsymbol{r}$. Their variation, and, therefore, the pattern of distortions, is dominated by a wavelength of $\lambda^* = \sigma\pi\sqrt{2}$, i.e. λ^* *is also proportional to the range of the "neighborhood function"* h_{rs}.

In the previous context, V being a set of patterns, the x- and y-coordinates would correspond to two "primary" features characterized by a large variation, whereas the z-coordinate would correspond to a "secondary" feature, characterized by a smaller variation that is measured by s. Then, as long as $s < s^*$, the system converges to a topographic map of the two "primary" features only. However, when the variation of the "secondary" feature, compared to the two "primary" ones, becomes large enough, the "secondary" feature begins to be reflected in the values of the prototypes and, therefore, in the topographic map. The variation of the prototypes along the axis of the "secondary" feature is dominated by the wavelength λ^* and gives rise to an irregular pattern of "stripes" if each lattice point $\boldsymbol{r}$ is assigned a gray value that indicates the value w_{r3} of its prototype $\boldsymbol{w}_r$ along the axis of the "secondary" feature (Fig. 8.5).

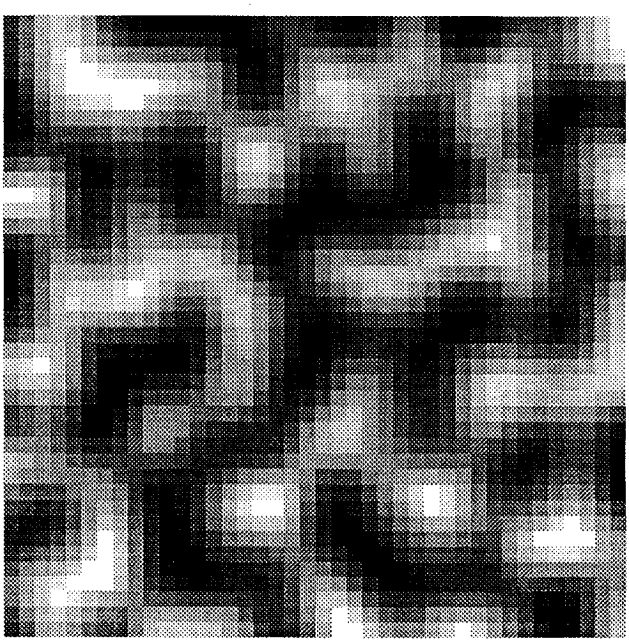


Fig. 8.5. "Striped projection". The displayed 40×40-lattice was used to obtain a "topographic map" of a 3-dimensional "feature space" given by $0 \leq x, y \leq 40$, $-4 \leq z \leq 4$ with Kohonen's algorithm ((8.4), $\sigma = 1.4$, 10^4 steps). The height (z-) dimension plays the role of the "secondary" feature, and gray values indicate its distribution over the lattice. The resulting pattern closely resembles the "ocularity stripes" found in the visual cortex. These are alternating bands of cells with stronger preference to input from either the right or the left eye (see, e.g. [8.24])

Interestingly, in the brain there are many two-dimensional arrangements of cells on which such "striped projections" seem to be realized. Prominent examples are the "ocular dominance stripes", where in addition to the "primary" two-dimensional retinal location, the additional feature "ocularity" (i.e. the degree to which a cell receives input from each eye) is mapped [8.24, 25], and the "orientation stripes", where the additional feature is line orientation [8.26, 27].

Models of the development of such "striped projections" have been previously suggested (see, e.g. [8.25, 28–31]). Here we want to emphasize that the previous analysis demonstrates that also *the particularly simple model by Kohonen can account for the phenomenon of striped projections*, a fact that has been observed already in simulations presented in [8.10] but seems to have received only little attention subsequently. This brings us to the issue of the "neural interpretation" of the model and its properties, a topic taken up in the following section.

8.4 Relation to Brain Modeling

One of the major architectural features within the brains of higher animals are *topographically organized "maps"* of various "feature spaces". They can be found in nearly all sensory and motor areas within the brain, e.g. the visual, auditive, and somatosensory fields as well as in the motor cortex, and there is both theoretical and some experimental evidence that maps of more abstract features might turn out also to play a role on higher processing levels [8.32].

In the somatosensory system the "feature space" is particularly simple. It is mapped onto a certain part of the cortex called the "somatosensory cortex". Experiments on the cortical representation of the hand surface in owl monkeys have revealed a very precise correspondence between hand locations and neurons in the cortical field [8.33]: each neuron can be excited only from receptors in some small "receptive field" in the hand surface, and the arrangement of the neurons in the cortex is a distorted, but still topographic "image" of the arrangement of their receptive fields on the skin. There is evidence that the required, very precise connectivity is not genetically prespecified but instead evolves gradually under the influence of sensory experience. Maps in different individuals show considerable variations, and they are not rigidly fixed even in adult animals. The somatotopic map can undergo adaptive changes, which have been found to be strongly driven by afferent input [8.34, 35].

The "self-organizing maps" are perhaps the simplest model that can account for the adaptive formation of such topographic representations (for other modeling approaches, see e.g. [8.30, 31, 36, 37]). In this case, the lattice A of prototypes $\boldsymbol{w}_r$ corresponds to a sheet of laterally interacting adaptive neurons, one for each lattice site $\boldsymbol{r}$, that are connected to a common bundle of n input fibers from the receptors in the receptor sheet. The ith component of vector $\boldsymbol{w}_r$ is interpreted as the connection strength between input fiber i and neuron $\boldsymbol{r}$.

The formation of the map is assumed to be driven by random sensory stimulation. Tactile stimuli on the receptor sheet excite clusters of receptors and thereby

cause activity patterns on the input lines that are described by n-dimensional real vectors and that take the role of the input patterns $\boldsymbol{v} \in V$. The total synaptic input to each neuron $\boldsymbol{r}$ is measured by the dot product $\boldsymbol{x} \cdot \boldsymbol{w}_r$. Each tactile stimulus is considered as a discrete event that leads to excitation of a localized group of neurons in the lattice A. The function h_{rs} (with $\boldsymbol{s}$ fixed and $\boldsymbol{r}$ taken as argument) is interpreted as the spatial variation of this neural excitation in A, and (8.4) can then be interpreted as a Hebbian rule together with an "activity-gated" memory loss term for the change in the synaptic strengths $\boldsymbol{w}_r$ following the stimulus.

The spatial shape h_{rs} of the neural response is modeled by a Gaussian and is assumed to arise from lateral competitive interactions within the cortical sheet (for details, see e.g. [8.11]). Its center location $\boldsymbol{s}$ is assumed to coincide with the neuron receiving the largest synaptic input $\boldsymbol{w}_r \cdot \boldsymbol{v}$. Strictly speaking, this is only equivalent to (8.4), where $\boldsymbol{s}$ was defined to be minimizing the Euclidean difference $||\boldsymbol{v} - \boldsymbol{w}_s||$, if all vectors $\boldsymbol{w}_r$ and $\boldsymbol{v}$ are assumed to be normalized. If, however, all $\boldsymbol{w}_r$ are kept normalized, one can even drop the "memory loss term" $-h_{rs(v)}\boldsymbol{v}$, as its main purpose is only to keep the vectors $\boldsymbol{w}_r$ bounded. Therefore, the simulations presented below are based on the modified adaptation equation

$$w_{kli}(t+1) = (w_{kli}(t) + \varepsilon(t) h_{rs;kl}(t) v_i) / \sqrt{\Sigma_i (w_{kli}(t) + \varepsilon(t) h_{rs;kl(t)} v_i)^2}, \quad (8.17)$$

where we have also altered the notation, replacing $\boldsymbol{r}$ by (k, l) when referring to the synaptic weights of a neuron at a lattice site $\boldsymbol{r} = (k, l)$.

We still need to specify the input patterns $\boldsymbol{v}$. To this end, each input line i is taken to belong to one tactile receptor, located at a position $\boldsymbol{x}_i$ in the receptor sheet. The tactile stimuli are assumed to excite spatial clusters of receptors. As a convenient mathematical representation for a stimulus centered at $\boldsymbol{x}_{\text{stim}}$, we choose a Gaussian "excitation profile"

$$v_i = N \exp\left(-\frac{(\boldsymbol{x}_i - \boldsymbol{x}_{\text{stim}})^2}{\sigma_r^2}\right) \quad (8.18)$$

where σ_r is a measure of the "radius" of each stimulus.

With this interpretation, the algorithm, discussed in Sect. 8.2 as a means to generate a representation of a data set V that is optimal for transmission over some noisy channel, is seen as an adaptive process shaping the connectivity between two sheets of nerve cells. We can now return to the significance of the minimization of the functional $F[\boldsymbol{w}]$ in the present context.

We will assume that the primary task of a cortical map is to prepare a suitable encoding of the afferent sensory information for use in subsequent processing stages. Part of this encoded information is represented by the location of the excited neurons. Therefore, the correct transmission of this information to subsequent processing stages is equivalent to the target neurons being able to assess the origin of their afferent excitation correctly. However, such neurons typically integrate information from several different brain areas. Therefore, their "receptive fields" in these areas tend to be the larger, the higher their level in the

processing hierarchy is, and their reaction to input from one source may be influenced by the current inputs from other sources. This may make it impossible to tie their response precisely to excitation in a precise location of a cortical predecessor map. However, if such neurons cannot "read" the precise source location of their input excitation, they are in a position very similar to that of a receiver at the other end of a noisy transmission channel. Therefore, *a cortical map based on minimization of the functional $F[\boldsymbol{w}]$ (8.5) might help to minimize the average transmission error between neural layers arising from fluctuations of their "functional connectivity"*.

8.5 Formation of a "Somatotopic Map"

In this section, we shall present a computer simulation for the adaptive ordering of an initially random projection between tactile receptors on the skin and neurons in a "model cortex" [8.11, 38]. The "model cortex" consists of 16384 neurons that are arranged as a 128×128 square lattice and connected to 800 tactile receptors randomly scattered over a "receptor surface" (Fig. 8.6). The initial values w_{kli} were given independent random values chosen from the unit interval, thereby "connecting" each neuron in a random fashion to the 800 receptors.

In experiments, neurons are frequently characterized by their "receptive field properties". A "receptive field" of a neuron is the set of stimulus locations $\boldsymbol{x}_{\text{stim}}$ that lead to a noticeable excitation of the neuron. The center of the receptive field of neuron (k, l) is in the model defined as the average

$$\boldsymbol{s}_{kl} = \sum_i \boldsymbol{x}_i w_{kli} \Big/ \sum_i w_{kli} \,. \tag{8.19}$$

The mean square radius of the receptive field is a measure of the stimulus selectivity of neuron (k, l) and is defined as

$$G_{kl} = \sum_i (\boldsymbol{x}_i - \boldsymbol{s}_{kl})^2 w_{kli} \Big/ \sum_i w_{kli} \tag{8.20}$$

Figure 8.7 shows the hand-shaped receptor surface of the model used for the simulations. Each dot represents one of the 800 tactile receptors. Different regions of the hand are coded by different colors. Figure 8.8a shows the initial state of the network. Each pixel in the image corresponds to one neuron (k, l) of the neural sheet, and the pixel color simultaneously encodes two different properties of its receptive field in the hand surface: the hue indicates the location $\boldsymbol{s}_{kl}$ of the field center on the receptor surface shown in Fig. 8.7, and the saturation of the color indicates the spatial spread ($\propto G_{kl}^{1/2}$) of the receptive field. Neurons with large, diffuse receptive fields, i.e. those neurons that are connected to many, widely scattered receptors, are represented with low color saturation, while neurons with small, spatially highly specific receptive fields, i.e. connected only to receptors within a small spatial domain, are represented with bright, highly saturated colors.

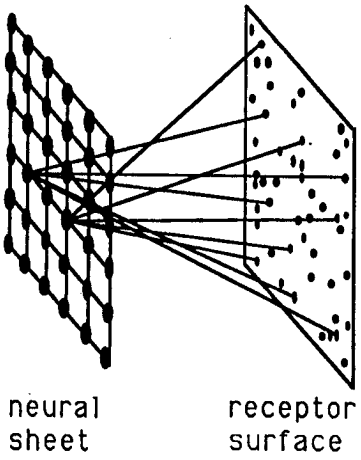


Fig. 8.6. Schematic drawing of the model system. The neuron sheet ("cortex") is represented by a square lattice of model neurons, each connected by modifiable links to all of 800 randomly located "receptors" in the receptor surface

As each neuron has its connections randomly initialized, all receptive fields, apart from statistical fluctuations, are initially very similar: they diffusely extend over the whole hand surface and, therefore, are "centered" in the middle of the receptor surface. This is reflected in Fig. 8.8a by the fairly uniform and pale color within the entire cortical sheet. Figure 8.8b shows the map after about 200 adaption steps. The brighter colors indicate the onset of a specialization of the neurons for smaller receptive fields. Finally, Fig. 8.8c shows the completed and refined map obtained after 10000 stimuli. It shows all parts of the hand surface in their correct topographic order, and the cortical map exhibits only small fluctuations during further "stimulation". Note that each neuron is still connected to every receptor, although the connection strengths outside the neuron's receptive field are very small. The connections can, however, be "revived" if the distribution of the input pattern changes, leading to an input-driven reorganization of the cortical map [8.34].

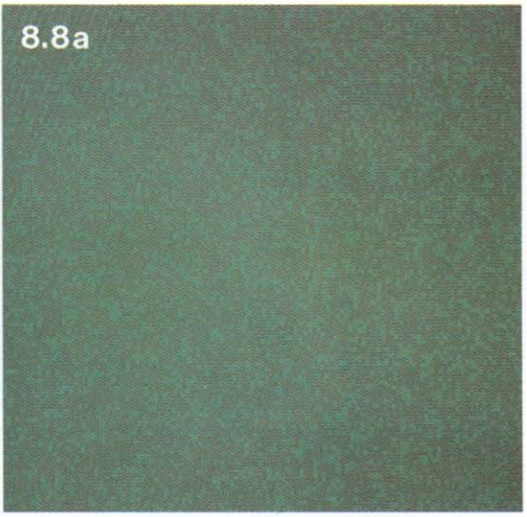
The emergence of selectively tuned neurons with spatially restricted receptive fields in the receptor surface can also be analytically demonstrated [8.39]. For sufficiently small adaptation steps (i.e. $\varepsilon \ll 1$), one can derive the following equation for the change of the receptive-field sizes G_{kl} under one adaptation step:

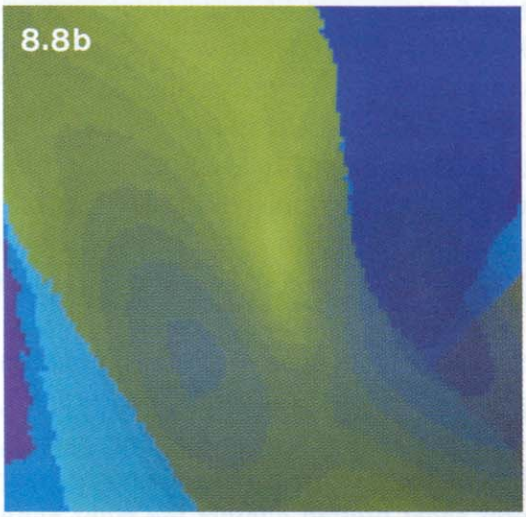
$$G_{kl}(t+1) = G_{kl}(t) + \frac{\varepsilon(t) h_{rs;kl}(t)}{\sum_i w_{kli}(t)} \sum_i (\boldsymbol{x}_i - \boldsymbol{s}_{kl})^2 \left(v_i - w_{kli}(t) \frac{\sum_j v_j}{\sum_j w_{klj}(t)} \right). \tag{8.21}$$

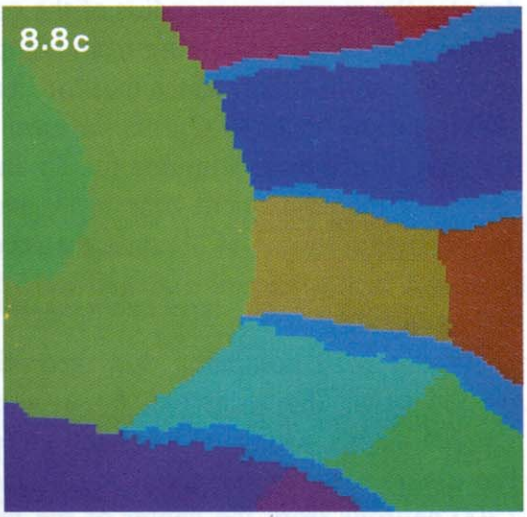
From this relation, one can derive an equation for G_{kl} when the system has reached a stationary state,

$$G_{kl} = \frac{\int h_{rs;kl} \langle [\Gamma(\boldsymbol{x}_s) + (\boldsymbol{x}_s - \boldsymbol{s}_{kl})^2] \sum_i v_i \rangle P(\boldsymbol{x}_s) \mathrm{d}^2 \boldsymbol{x}_s}{\int h_{rs;kl} \langle \sum_i v_i \rangle P(\boldsymbol{x}_s) \mathrm{d}^2 \boldsymbol{x}_s} \,. \tag{8.22}$$

Here $P(.)$ denotes the probability density of the stimuli centers $\boldsymbol{x}_{\mathrm{stim}}$ in the receptor surface, and $\langle .. \rangle$ denotes the average over all stimulus shapes (in case the stimuli are more general than the Gaussians of (8.18)). $\Gamma(.)$ represents the







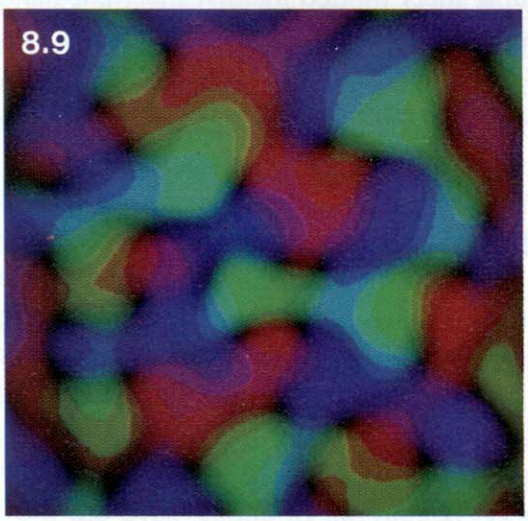


Fig. 8.7. (*top left*) Hand surface with receptor locations. Colors are used to identify different subregions of the hand

Fig. 8.8a–c. (*from top right to center right*) Development of "somatotopic map". Each pixel denotes a neuron, its color encodes the origin of its input in the hand surface shown in Fig. 8.7: (**a**) initial, (**b**) after about 200 stimuli, and (**c**) after 10 000 stimuli. At this stage, a clear topographic order has emerged

Fig. 8.9. Spatial organization of neuronal orientation preference and selectivity, formed in a simulated patch of "visual cortex" by the self-organizing process (8.17). A "rainbow" palette of colors indicates orientation preferences from 0° to 180°. Dark regions correspond to low, bright regions to high, directional selectivity

mean square radius of each stimulus, defined by

$$\Gamma(\boldsymbol{x}_s) = \frac{1}{\sum_i v_i} \sum_i (\boldsymbol{x}_i - \boldsymbol{x}_s)^2 v_i \,. \tag{8.23}$$

Equation (8.22) can be approximated well by the much simpler relation (for details see [8.40])

$$G_{kl} \approx \Gamma + M^{-1} \sigma_h^2 \,, \tag{8.24}$$

where

$$\sigma_h^2 = \frac{2 \sum_{m,n} h_{rs;mn} \left[(r-m)^2 + (s-n)^2 \right]}{\sum_{m,n} h_{rs;mn}} \tag{8.25}$$

denotes the mean square radius of the output function and M is the *local magnification factor* of the mapping from stimulus centroids $\boldsymbol{x}_{\text{stim}}$ to neuron coordinates (r, s). Equation (8.24) states that the neurons develop receptive fields the area of which (proportional to G_{kl}) is the sum of two terms: the first term is essentially the area of a typical stimulus ($\propto \Gamma$) and the second term is essentially the area ($\propto \sigma^2$) of the adjustment zone in the neuron layer, but "projected back" (inverse magnification factor M^{-1}) onto the receptor sheet. Therefore, for predominantly localized tactile stimuli and narrow $h_{rs;mn}$ the neurons will develop *localized receptive fields*.

Figure 8.10 compares this theoretical result with data from a simulation (16 384 cells, 784 receptors, $h_{rs,kl} = \exp(-[r-k]^2 + [s-l]^2/\sigma_h^2), \sigma_r = 0.15$, 6×10^4 steps), where σ_h has been varied slowly between $\sigma_h = 100$ and $\sigma_h = 5$. The diagram shows the mean square radius of the receptive field averaged over 2 300 neurons from the center of the neural sheet plotted against the right-hand side of (8.24). The dots represent the results of the simulation and the solid line corresponds to (8.24). The agreement is very satisfactory, except for parameter values leading to large receptive fields, for which edge effects become noticeable.

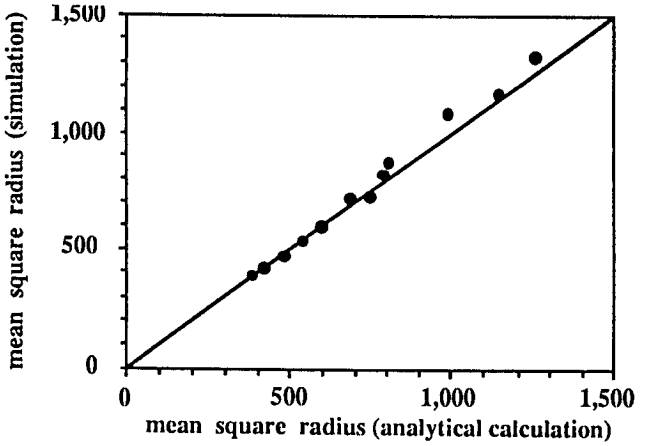


Fig. 8.10. Development of receptive field radii. The analytical result (8.24) is compared with results of a computer simulation (mean square radii are given in arbitrary, relative units)

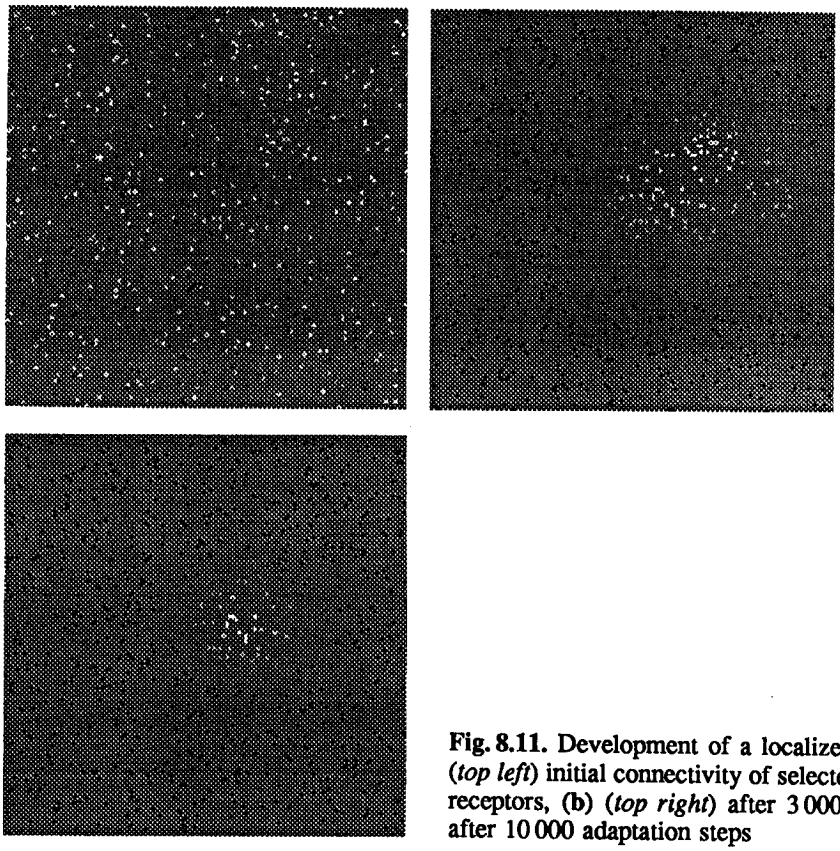


Fig. 8.11. Development of a localized receptive field: **(a)** (*top left*) initial connectivity of selected neuron with tactile receptors, **(b)** (*top right*) after 3000 and **(c)** (*lower left*) after 10000 adaptation steps

Figure 8.11 illustrates the state of a typical receptive field at the beginning of another simulation (Fig. 8.11a), after 3000 iterations (Fig. 11b) and after 10000 iterations (Fig. 8.11c) (for this run $\sigma_h = 50 \ldots 5$, $\sigma_p = 0.12$ and $t_{max} = 10^4$). The dot locations mark the positions of the tactile receptors on the receptor surface, while their brightness encodes the strength of their connection to the neuron under consideration. Initially, the field is very diffuse (Fig. 8.11a), but contracts rapidly (Fig. 8.11b), until finally it is localized well (Fig. 8.11c).

An important aspect of these simulations is the demonstration that the formation of the correct topographic map of the skin surface succeeds also in the case when the two-dimensional position information about the stimuli is provided by the very high dimensional ($d = 800$) input vectors that represent the excitation patterns of the tactile receptors. To accomplish this, the algorithm has to "detect" the relevant, two-dimensional submanifold formed by the stimuli, and has to "ignore" all remaining "orthogonal" coordinates.

The optical cortex faces a self-organization task similar to that of the somatosensory cortex. However, in contrast to the somatosensory map, the map from the eyes' retina to the optical cortex is known to represent, in addition to the "primary" features "retinal location" (pair of coordinates!), further features, such as *orientation* and *ocularity*. The spatial organization of the visual map has been studied in great detail (see e.g. [8.24, 26, 27]) and cells with similar ocularity or tuned to similar orientations were found to be arranged in *irregular*

"bands" or "stripes". Adding "ocularity" or "orientation" to the two "primary" features "retinal location" results in the need to map a three- or higher-dimensional feature space onto a two-dimensional surface [8.58, 61]. Figure 8.5 showed a self-organizing map which gave rise to a pattern of the "secondary" feature that resembles observed patterns in the visual map. A closer agreement with natural maps can be obtained if the three-dimensional input representation is replaced by high-dimensional excitation patterns on a "model retina" [8.56, 57]. Figure 8.9 shows an "orientation map" on a "model cortex" of 256×256 cells obtained in this way. "Stimuli" were of elliptic shape with randomly chosen orientations. The "stimuli" produced excitations on a "model retina" covered with 900 randomly distributed light-sensitive "receptors". Each pixel in Fig. 8.9 represents one neuron. The pixel color encodes the orientation of the stimulus to which the neuron is maximally responsive ("orientation preference"), and the brightness the degree of specificity of the neuron. Various features, such as "slabs" along which orientation selectivity changes continuously, dark "foci" of unspecific cells around which orientation changes in a clockwise or anti-clockwise fashion, and "fractures", across which orientation changes discontinuously, can be discerned and correspond well to observations from actual mapping experiments (see, e.g. [8.26, 27]).

8.6 Adaptive Orientation and Spatial Frequency Filters

In this section we consider the issue of data compression through preprocessing by local filters which select geometrically significant features from their input. A major part of the material in this section is taken from [8.40]. The input will be spatially varying patterns, for example, correlated random dot patterns or textures. The architecture of the network is very similar to that of the networks described above. It consists of an input layer with N_i neurons and an output layer with N_o neurons. Input and output units exhibit real, continuous-valued activities $\boldsymbol{i} = (i_1, .., i_{N_\mathrm{i}})$ and $\boldsymbol{o} = (o_1, .., o_{N_\mathrm{o}})$. The output layer should develop synaptic connections with the input layer to establish suitable receptive fields. The development of connections is driven by training with a set of N_π representative patterns $\{\boldsymbol{p}^\pi = (p_1^\pi, .., p_{N_\mathrm{i}}^\pi), \pi = 1, \ldots, N_\pi\}$.

The formation of the receptive fields will depend on adaptive lateral interactions between units in the output layer. These lateral interactions serve functions similar to lateral connections in Kohonen-type networks. In the latter such connections are needed to determine the output unit of maximal activity as well as to induce neighborhoodwide activities described by the functions h_{rs}, which leads to a topology-conserving neural projection between input and output layers. In the present network, lateral interactions serve the role of molding the projection such that characteristic correlations between input activities $i_1, \ldots, i_{N_\mathrm{i}}$ are detected. Such correlations are described by the the covariance matrix C of a set of characteristic patterns. This matrix has elements $C_{jk} = \langle p_j^\pi p_k^\pi \rangle$ where p_j^π denotes a sample pattern and $\langle \ldots \rangle$ denotes the average over the training set $\{\boldsymbol{p}^\pi, \pi = 1, \ldots, N_\pi\}$.

The desired filters are achieved by synaptic weights w_{jm} between input unit j and output unit m. The set of weights connecting an output unit m with all input units forms the weight vector $\boldsymbol{w}_m$, the transpose of which is the mth row of the weight matrix $\boldsymbol{W}$. Activities of the input units correspond to the presented patterns, i.e., $\boldsymbol{i} = \boldsymbol{p}^\pi$. Activities of the output units, in response to a pattern $\boldsymbol{p}^\pi$, are linear sums of the inputs weighted by the synaptic strengths, i.e., $\boldsymbol{o}^\pi = \boldsymbol{W}\,\boldsymbol{p}^\pi$. The desired filters, i.e. synaptic weights, are the eigenvectors of the covariance matrix C_{jk} defined through

$$\sum_j C_{jk} w_{km} = \lambda_m w_{jm} \,. \tag{8.26}$$

Application of such filters corresponds to the statistical technique of *principal component analysis* (see, e.g., [8.41]). The network should yield the first eigenvectors of the covariance matrix corresponding to the largest eigenvalues λ_m of C_{jk}. To render such network robust against failure of single units one may represent the mth eigenvalue by a large number of output units rather than a single unit. However, in the following we will assume for the sake of simplicity that single output units represent an eigenvector of C_{jk}.

The output units should also discern the spatial location of these characteristics in a spatially extended pattern $\boldsymbol{p}$. This latter capacity can be achieved through a translation-invariant duplication of network characteristics. We will neglect the position dependence and focus in the following only on preprocessing in a small neighborhood of input cells. We choose, therefore, two completely interconnected layers, i.e. w_{jm} initially is nonzero for all j and m.

We will describe now how the network through adjustment of its synaptic weights can adopt appropriate receptive fields when exposed to the set of training patterns. Weights between layers are adjusted upon presentation of an input pattern $\boldsymbol{p}^\pi$ according to a Hebbian rule, leading to an increase in synaptic strength if the corresponding pre- and postsynaptic potentials are of the same sign. If weight changes are small enough, the update can be performed after presentation of all patterns, i.e.

$$\Delta \boldsymbol{w}_m = \eta \langle (\boldsymbol{p}^\pi - \langle \boldsymbol{p}^\pi \rangle)(o_m^\pi - \langle o_m^\pi \rangle) \rangle \,, \tag{8.27}$$

where η is a positive parameter and where the brackets $\langle \ldots \rangle$ denote again the average over the set of patterns. The subtraction of averages in (8.27) can be interpreted as the existence of thresholds of the units. On the other hand, subtracting averages is convenient from a mathematical point of view [8.42] and allows one to assume that $\langle \boldsymbol{p}^\pi \rangle = 0$ and $\langle \boldsymbol{o}^\pi \rangle = 0$.

Let us first consider the case of a single output unit. Linsker showed that the weights of a unit that is subject to the Hebbian rule (8.27) evolve to maximize the variance of the output for a set of presented patterns [8.18]. If the weights are normalized after every update such that $\sum_i w_{i1}^2 = 1$, the Hebbian rule renders weights which characterize the direction of maximal variance of the pattern set [8.42]. Equivalently, the weight vector $\boldsymbol{w}_1$ converges to the eigenvector with

the largest eigenvalue λ_1 of the covariance matrix C of the pattern set. Thus, a Hebbian learning rule for Euclidian normalized weights yields the first principal component of the input data set. The nonvanishing weights w_{i1} of the output unit define its receptive field. The output unit then acts as a feature detector which analyzes the principal feature of a presented pattern and corresponds to a so-called "matched linear filter" [8.42].

However, a single principal component usually describes only a fraction of the total information contained in a pattern. In order to transmit the complete information between the two layers, as many output cells as the rank of the covariance matrix C are required. Furthermore, in order to develop into filters of mutually orthogonal features, the output cells need to become uncorrelated. For this purpose we assume the existence of lateral, hierarchically organized connections with weights u_{lm} between output units l and m, where $l < m$. The activity of the mth output cell is then given by $o_m^{\pi} = \boldsymbol{w}_m \cdot \boldsymbol{p}^{\pi} + \sum_{l<m} u_{lm} \boldsymbol{w}_l \cdot \boldsymbol{p}^{\pi}$. According to (8.27), changes of synaptic connection strengths between input units and output unit m are given by

$$\Delta \boldsymbol{w}_m = \eta \left(C\boldsymbol{w}_m + \sum_{k<m} u_{km} C\boldsymbol{w}_k \right) . \tag{8.28}$$

Figure 8.12 presents the architecture of the network, in particular, the hierarchical arrangement of lateral connections. This arrangement has been chosen to guide the network to a final state in which the output units assume receptive fields corresponding to the different eigenvectors of the covariance matrix C. The cell in the output layer at the top of the hierarchy will adopt a receptive field corresponding to the eigenvector with the largest eigenvalue, the cell next in the hierarchy will represent the eigenvector with the second largest eigenvalue, and so forth.

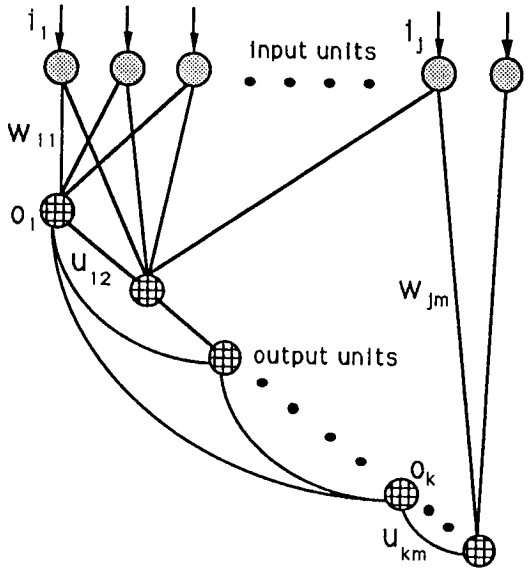


Fig. 8.12. Schematic drawing of the hierarchical network used for feature extraction

To ascertain that the cells adopt different receptive fields the lateral weights u_{lm} adapt themselves according to an *anti-Hebbian* rule: the strength of a lateral synapse is lowered if the corresponding pre- and postsynaptic activities are of the same sign. Again we assume that changes of the synaptic weights are small. The anti-Hebbian rule leads to a decrease in synaptic strength if the corresponding output units have correlated activities and is described by

$$\Delta u_{lm} = -\mu \langle o_l^\pi \, o_m^\pi \rangle \, . \tag{8.29}$$

Here μ is a positive learning parameter. The anti-Hebbian rule is similar to the learning rule of Kohonen's novelty filter [8.10] and to the "unlearning" rule proposed by Hopfield [8.43].

Because of the hierarchical arrangement, the cell at the top of the hierarchy will force all other output units to become uncorrelated to it; the top cell develops its receptive field in accordance with the first eigenvector and suppresses any attempt of cells lower in the hierarchy to develop the same receptive field. The second cell in the hierarchy prevents all cells below it developing a receptive field similar to its own, the latter being shaped to agree with the second eigenvector of the covariance matrix C. This chain is continued down to the cell last in the hierarchy. The selection of receptive fields in the order of decreasing eigenvalues λ_m originates from the fact that the weights w_{jm} grow fastest in the direction of the distribution of the first eigenvector, second fastest in the direction of the second eigenvector, and so on.

As a result of the proposed learning scheme, the weight vector w_m converges to the mth eigenvector of C. Convergence requires that the learning parameters η and μ governing the weights w_{jm} and u_{mn}, respectively, need to obey the inequality (we assume the ordering of eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_{N_o}$)

$$\mu > \frac{\eta(\lambda_1 - \lambda_n)}{\lambda_1 (1 + \eta \lambda_n)} \tag{8.30}$$

for $n = 1, 2, \ldots, N_o$ [8.44]. Since C is a real symmetric matrix, the weight vectors w_{jm} become orthogonal and, consequently, the output units with different receptive fields in the mature network are uncorrelated. This implies that in the mature network the lateral connections vanish after they have completed their important function of yielding orthogonal receptive fields.

Several authors have proposed inhibitory connections between output units in order to render their activities uncorrelated [8.5, 10, 46, 47]. In our scheme, lateral connections are both excitatory and inhibitory before they vanish. This results in a purely feed-forward network, which represents an important computational advantage for a parallel system. Principal component analysis has also been associated with linear feed-forward networks using optimization methods with respect to a quadratic error function, i.e., back-propagation [8.48]. The advantage of our model consists in optimal feature extraction without supervision and in the existence of biologically plausible, local adaptation rules for the weights, namely Hebbian and anti-Hebbian rules.

We will illustrate now the performance of the network for patterns of spatially varying intensity and show that the network develops feature cells with receptive fields which are similar to those of "simple cells" found in the striate cortex and which select features of different orientation and different spatial frequencies.

For this purpose, we consider a rectangular lattice of $N_i \times N_i'$ sensory input units representing the receptive field of N_o output units, with $N_o \leq N_i N_i'$. We generate two-dimensional input patterns of varying intensity by first selecting random numbers s_{ij}^{π}, $\pi = 1, \ldots, N_\pi$ from the interval $[-1, +1]$. Then, in order to introduce information about the topological structure of the receptive field, the random input intensities are correlated, e.g., with their nearest neighbors in both directions. As a result, the component p_{ij}^{π} of a pattern $\boldsymbol{p}^{\pi}$ at the coordinate (i, j) of the receptive field is given by $p_{ij}^{\pi} = s_{ij}^{\pi} + s_{i-1\,j}^{\pi} + s_{i+1\,j}^{\pi} + s_{i\,j-1}^{\pi} + s_{i\,j+1}^{\pi}$. We assume vanishing boundary conditions, i.e., $s_{0j} = s_{i0} = s_{N_i+1\,j} = s_{i\,N_i'+1} = 0$. Note that this averaging of neighboring signals corresponds to introducing an additional layer with random activities and with fixed and restricted connections to the input layer.

Receptive fields of simple cells in cat striate cortex as recorded by Jones et al. [8.49, 50] can be described by Gabor functions which consist of an oscillatory part, namely a sinusoidal plane wave, modulated by a Gaussian, exponentially decaying part. To localize the receptive fields in our model system correspondingly, we scale the weights between layers, i.e. the weight $w(ij, m)$ between the input unit at lattice location (i, j) and the mth output unit, according to $w'(ij, m) = D(i, j) w(ij, m)$, where $D(i, j)$ is a Gaussian distribution with $D(i, j) \sim \exp[-(i - i_0)^2/\sigma_1 - (j - j_0)^2/\sigma_2]$. Here, σ_1 and σ_2 control the width of the distribution and (i_0, j_0) is the coordinate of the lattice center, i.e., $(i_0, j_0) = (N_i/2, N_i'/2)$.

Imposing a Gaussian distribution of synaptic weights will change the eigenvalue spectrum of the covariance matrix of the input pattern. Therefore, such a network, in a strict sense, cannot develop receptive fields according to a principal component analysis. However, if the restriction to neighborhoods described by a Gaussian were not applied, the weights w_{jm} develop towards the exact eigenvectors of C. The localization of w_{jm} can be exploited to prevent degeneracies between eigenvalues, which can lead to a mixing of receptive fields, resulting in asymmetrical fields. If the Gaussian distribution is chosen not to be rotationally symmetric, i.e., if $\sigma_1/\sigma_2 \neq 1$, the orientation of receptive fields is predetermined owing to imposed symmetry axes.

Figure 8.13 displays contour plots of the receptive fields of the first eight output cells after 10 000 learning cycles (from left to right and top to bottom). Solid lines correspond to positive, dashed lines to negative synaptic weights. The input lattice was a square of 20×20 units. We imposed a Gaussian distribution of synaptic weights with parameters $\sigma_1 = 12$ and $\sigma_2 = 15$. Learning parameters η and μ were equal to 0.05 and 0.1, respectively. Owing to the nonsymmetric Gaussian distribution of weights, all units have slightly elongated receptive fields. The first unit corresponds to a simple cell with all-inhibitory synaptic weights.

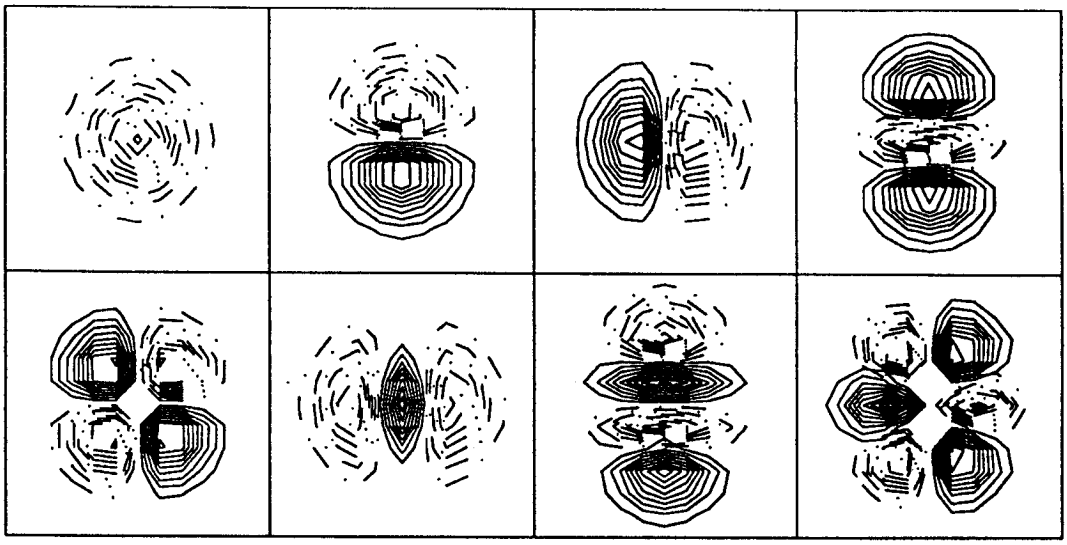


Fig. 8.13. From left to right and top to bottom: contour plots of receptive fields of output units 1–8 in the case of a square lattice of 20×20 input units. The synaptic distribution $D(i, j)$ between layers was Gaussian with $\sigma_1 = 12$ and $\sigma_2 = 15$. Solid lines indicate positive, dashed lines negative, weights. The number of learning cycles was 10000; the learning parameters η and μ were equal to 0.05 and 0.1, respectively

The receptive fields of the second and third units display an excitatory and an inhibitory region and resemble simple cells, selective to edges of a fixed orientation. The fourth and sixth units have receptive fields with two zero crossings, corresponding to simple cells, selective to bars of a fixed orientation. In addition, the seventh unit is orientation selective, with four alternating excitatory and inhibitory regions. This unit would respond maximally to two parallel lines or bars with fixed distance and orientation.

The described units have receptive fields that resemble recorded receptive fields of simple cells in the primary visual cortex [8.16, 49, 50]. Up to now, there has not been any experimental evidence for receptive fields of the type of the fifth and eighth units, displaying four and six lobes.

8.7 Conclusion

Our previous analysis of self-organizing maps and of a hierarchical network for learning feature extraction has focused on two complementary aspects: the first aspect concerned the information-processing task carried out by each system, while the second concerned the capability of each system to account for observed phenomena of brain organization.

Regarding the first aspect, we demonstrated that both systems have in common the ability to *compress data*. This ability is realized in different ways for the two systems: self-organizing maps achieve data compression by a nonlinear mapping of their input patterns onto a lower-dimensional manifold. The points of this

manifold can be considered "code labels" requiring less storage space than, and allowing an approximate reconstruction of, the original data. The self-organizing maps lead to an encoding that is a compromise between minimization of the reconstruction error for the original data and preservation of their similarity relationships under the encoding transformation.

The hierarchical network for feature extraction achieves data compression by performing a principal component analysis of its input data. The available cells organize their connectivity such that they automatically extract the principal components with the largest eigenvalues of the signal correlation matrix. Their output values represent the amplitudes of these principal components and, therefore, provide a lower-dimensional signal from which the original signal can be reconstructed with minimal expected square error.

With regard to the second aspect, we showed that self-organizing maps can explain several properties of the organization of cortical areas, such as the ubiquitous "striped projections" and the hierarchical feature maps found in the visual cortex, as a consequence of a *single* principle. This principle is related to the minimization of the functional $F[w]$ and interpretable as a process of adaptive synaptic modification. The hierarchical feature-extraction network complements this ability by explaining the formation of the small-scale structure observed in the various receptive fields encountered in cells of the visual cortex. The receptive-field properties of these cells resemble Gabor filters, and very similar receptive fields are developed by the cells of the artificial network.

A better understanding of the operation of the brain involves the investigation of several levels of organization. The research presented in this contribution was meant to be a small step towards this goal.

Acknowledgement. We would like to thank R. Kufrin and G. Quinn for their help and support in all technical matters concerning the use of the Connection Machine system. The authors are grateful to the Boehringer-Ingelheim Fonds for providing a fellowship to K. Obermayer. This research has been supported by the University of Illinois at Urbana-Champaign. Computer time on the Connection Machine CM-2 has been made available by the National Center for Supercomputer Applications.

References

8.1 D.J. Amit: *Modeling Brain Function* (Cambridge University Press 1989)

8.2 D.J. Amit, H. Gutfreund, H. Sompolinsky: Storing infinite number of patterns in a spin-glass model of neural networks, Phys. Rev. Lett. **55**, 1530–1533 (1985)

8.3 D.J. Amit, H. Gutfreund, H. Sompolinsky: Information storage in neural networks with low level of activity, Phys. Rev. A **35**, 2293–2303 (1987)

8.4 S. Grossberg: Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors, Biol. Cybern. **23**, 121–134 (1976)

8.5 D.E. Rumelhart, D. Zipser: Feature discovery by competitive learning, Cognitive Science **9**, 75–112 (1985)

8.6 J. Buhmann, R. Divko, K. Schulten: Associative memory with high information content, Phys. Rev. A **39**, 2689–2692 (1989)

8.7 E. Gardner, B. Derrida: Optimal storage properties of neural network models, J. Phys. A **21**, 271–284 (1988)

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

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

8.10 T. Kohonen: *Self-Organization and Associative Memory*, Springer Series in Information Sciences 8 (Springer, Berlin, Heidelberg 1984)

8.11 K. Obermayer, H. Ritter, K. Schulten: Large-scale simulation of self-organizing neural network on parallel computers: Application to biological modelling, Parallel Computing **14**, 381–404 (1990)

8.12 H. Ritter, T. Martinetz, K. Schulten: Topology conserving maps for learning visuomotor-coordination, Neural Networks **2**, 159–168 (1989)

8.13 T. Martinetz, H. Ritter, K. Schulten: Three-dimensional neural net for learning visuomotor-coordination of a robot arm, IEEE Transactions on Neural Networks **1**, 131–136 (1990)

8.14 H. Ritter, T. Martinetz, K. Schulten: *Neuronale Netze – Eine Einführung in die Neuroinformatik selbstorganisierender Netzwerke* (Addison-Wesley, Bonn 1990) (in German); *Neural Computation and Self-organizing Maps: An Introduction* (Addison-Wesley, New York 1992)

8.15 D.H. Hubel, T.N. Wiesel: Receptive fields of single neurones in cat's striate cortex, J. Physiol. **148**, 574–591 (1959)

8.16 J.P. Jones, L.A. Palmer: An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex, J. Neurophysiol. **58**, 1133–1258 (1987)

8.17 W.M. Cowan: The development of the brain, Sci. Am. **241**, 107–117 (1979)

8.18 R. Linsker: Self-organization in a perceptual network, IEEE Computer **21**, 105–117 (1988)

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

8.20 J. Makhoul, S. Roucos, H. Gish: Vector quantization in speech coding, Proc. IEEE (1985), 73-1551-1558

8.21 H. Ritter, K. Schulten: Kohonen's self-organizing maps: Exploring their computational capabilities, *Proc. IEEE ICNN 88*, San Diego (IEEE Computer Society press 1988), Vol. I, pp. 109–116

8.22 S.P. Luttrell: "Self-organisation: A derivation from first principles of a class of learning algorithms", in: *Proc. IJCNN 89*, Washington DC (IEEE Computer Society Press, 1989), Vol. II pp. 495–498

8.23 H. Ritter, K. Schulten: Convergence properties of Kohonen's topology conserving maps: Fluctuations, stability and dimension selection, Biol. Cybern. **60**, 59–71 (1989)

8.24 S. LeVay, T.N. Wiesel, D.H. Hubel: The development of ocular dominance columns in normal and visually deprived monkeys, J. Comp. Neurol. **191**, 1–51 (1974)

8.25 K.D. Miller, J.B. Keller, M.P. Stryker: Ocular dominance column development: analysis and simulation, Science **245**, 605–615 (1989)

8.26 G.G. Blasdel, G. Salama: Voltage-sensitive dyes reveal a modular organization in monkey striate cortex, Nature **321**, 579–585 (1986)

8.27 D.H. Hubel, T.N. Wiesel: Sequence regularity and geometry of orientation columns in the monkey striate cortex, J. Comp. Neurol. **158**, 267–294 (1974)

8.28 C. von der Malsburg: Development of ocularity domains and growth behavior of axon terminals, Biol. Cybern. **32**, 49–62 (1979)

8.29 C. von der Malsburg, J. Cowan: Outline of a theory for the ontogenesis of iso-orientation domains in visual cortex, Biol. Cybern. **45**, 49–56 (1982)
8.30 A. Takeuchi, S. Amari: Formation of topographic maps and columnar microstructures, Biol. Cybern. **35**, 63–72 (1979)
8.31 D.J. Willshaw, C. von der Malsburg: How patterned neural connections can be set up by self-organization, Proc. R. Soc. London B **194**, 431–445 (1976)
8.32 H. Ritter, T. Kohonen: Self-organizing semantic maps, Biol. Cybern. **61**, 241–254 (1989)
8.33 J.H. Kaas, R.J. Nelson, M. Sur, C.S. Lin, M.M. Merzenich: Multiple representations of the body within the primary somatosensory cortex of primates, Science **204**, 521–523 (1979)
8.34 J.H. Kaas, M.M. Merzenich, H.P. Killackey: The reorganization of somatosensory cortex following peripheral nerve damage in adult and developing mammals, Annual Rev. Neurosci. **6**, 325–256 (1983)
8.35 M.M. Merzenich et al.: Somatosensory cortical map changes following digit amputation in adult monkeys, J. Comp. Neurol. **224**, 591 (1984)
8.36 C. von der Malsburg, D.J. Willshaw: How to label nerve cells so that they can interconnect in an ordered fashion, Proc. Natl. Acad. Sci. USA **74**, 5176–5178 (1977)
8.37 J.C. Pearson, L.H. Finkel, G.M. Edelman: Plasticity in the organization of adult cerebral maps: A computer simulation based on neuronal group selection, J. Neurosci. **12**, 4209–4223 (1987)
8.38 K. Obermayer, H. Ritter, K. Schulten: Large-scale simulation of a self-organizing neural network: Formation of a somatotopic Map, *Parallel Processing in Neural Systems and Computers*, edited by R. Eckmiller, G. Hartmann, and G. Hauske (North-Holland, Amsterdam 1990) 71–74
8.39 K. Obermayer, H. Ritter, K. Schulten: A neural network model for the formation of topographic maps in the CNS: Development of receptive fields, IJCNN-90, Conf. Proceedings II, 423–429, San Diego (1990)
8.40 J. Rubner, K. Schulten: A self-organizing network for complete feature extraction, Biol. Cybern. **62**, 193–199 (1990)
8.41 D.N. Lawlwy, A.E. Maxwell: *Factor Analysis as a Statistical Method* (Butterworths, London 1963)
8.42 E. Oja: A simplified neuron model as a principal component analyzer, J. Math. Biology **15**, 267–272 (1982)
8.43 J.J. Hopfield, D.I. Feinstein, R.G. Palmer: "Unlearning" has a stabilizing effect in collective memories, Nature **304**, 158–159 (1983)
8.44 J. Rubner, P. Tavan: A self-organizing network for principal component analysis, Europhys. Lett. **10**, 693–698 (1989)
8.45 H. Ritter: Asymptotic level density for a class of vector quantization processes, Internal Report A9, Helsinki Univ. of Technology (1989)
8.46 C. von der Malsburg: Self-organization of orientation sensitive cells in the striate cortex, Kybernetik **14**, 85–100 (1973)
8.47 A.L. Yuille, D.M. Kammen, D.S. Cohen: Quadrature and the development of orientation selective cortical cells by Hebb rules, Biol. Cybern. **61**, 183–194 (1989)
8.48 P. Baldi, K. Hornik: Neural networks and principal component analysis: leraning from examples without local minima, Neural Networks **2**, 53–58 (1989)
8.49 J.P. Jones, L.A. Palmer: The two-dimensional spatial structure of simple receptive fields in cat striate cortex, J. Neurophysiol. **58**, 1187–1211 (1987)
8.50 J.P. Jones, A. Stepnoski, L.A. Palmer: The two-dimensional spectral structure of simple receptive fields in cat striate cortex, J. Neurophysiol. **58**, 1112–1232 (1987)
8.51 M. Cottrell, J.C. Fort: A stochastic model of retinotopy: A self-organizing process, Biol. Cybern. **53**, 405–411 (1986)
8.52 N.G. van Kampen: *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam 1981)
8.53 E.I. Knudsen, S. du Lac, S.D. Esterly: Computational maps in the brain, Ann. Rev. Neurosci. **10**, 41–65 (1987)
8.54 T. Kohonen: *Clustering, Taxonomy and Topological Maps of Patterns*, Proc 6th Int. Conf. on Pattern Recognition, Munich pp. 114–128 (1982)
8.55 H. Ritter, K. Schulten: On the stationary state of Kohonen's self-organizing sensory mapping, Biol. Cybern. **54**, 99–106 (1986)

8.56 K. Obermayer, H. Ritter, K. Schulten: A Principle for the Formation of the Spatial Structure of Cortical Feature Maps, Proc. Natl. Acad. Sci. USA **87**, 8345–8349 (1990)
8.57 K. Obermayer, H. Ritter, K. Schulten: A Model for the Development of the Spatial Structure of Retinotopic Maps and Orientation Columns, IEICE Trans. Fund. Electr. Comm. Comp. Sci., in press (1992)
8.58 K. Obermayer, G.G. Blasdel, K. Schulten: A Statistical Mechanical Analysis of Self-Organization and Pattern Formation during the Development of Visual Maps, Phys. Rev. A, in press (1992)
8.59 E. Erwin, K. Obermayer, K. Schulten: Self-Organizing Maps: Ordering, Convergence Properties and Energy Functions, Biol. Cybern., in press (1992)
8.60 E. Erwin, K. Obermayer, K. Schulten: Self-Organizing Maps: Stationary States, Metastability and Convergence Rate, Biol. Cybern., in press (1992)
8.61 K. Obermayer, K. Schulten, G.G. Blasdel: A Comparison of a Neural Network Model for the Formation of Brain Maps with Experimental Data, in: Advances in Neural Information Processing Systems 4, Eds. D.S. Touretzky, R. Lippman, Morgan Kaufmann Publishers, in press (1992)