14. Mathematical Analysis of Kohonen's Model

The preceding chapters have shown the versatility of self-organizing maps for sensory processing and control by means of a series of examples. The properties of self-organizing maps that became evident in these examples will be characterized in this chapter from a more general mathematical point of view, and their relationship to other signal processing algorithms will be pointed out.

14.1. Overview

First, there is an important connection between self-organizing maps and algorithms for *adaptive data compression*. The latter algorithms are dealing with the coding of given data in a more compact form, so that later the original data can be recovered with as little error as possible. Obviously, in the interest of obtaining the highest possible "compression factor," a certain reconstruction error must be permitted. The method of *vector quantization* is a class of compression procedures leading to minimization of a prescribed measure of the reconstruction error. We will show that self-organizing maps can be regarded as a generalization of this approach. The neighborhood function modifies the error quantity to be minimized as compared to that minimized in conventional procedures.

Maps have a second important connection to the various procedures of *principal component analysis* of data. In these procedures, one seeks to describe as faithfully as possible the dis-

tribution of data points embedded in a high-dimensional space, using only a space of lower dimension. In principal component analysis, this occurs by linear projection onto a space spanned by those eigenvectors of the data distribution that belong to the largest eigenvalues of the two-point correlation matrix. Topology preserving maps offer a generalization of this linear procedure by providing a projection onto nonlinear, so-called *principal manifolds*. Projections onto principal manifolds can yield a low-dimensional image of the original data with smaller projection errors, *i.e.*, more faithful representations of the original data compared to linear procedures that use the same projection dimension.

The problems of data compression and of obtaining "good" projections onto lower-dimensional spaces are related and play an important role for numerous information processing tasks. A large part of the applicability of topology preserving maps is due to their relevance to both kinds of problems. Hence, it may not be too surprising that topology preserving maps are found in various areas of the brain.

The map formation process is adaptive and is driven by a random sequence of input signals. Mathematically, the process corresponds to an adaptively changing map that gradually evolves toward a stationary state. This leads to the question of the *convergence properties* of the process. We will investigate this question in Sections 14.6–14.9 more closely and, among other things, we will derive convergence conditions as well as expressions for the magnitude of fluctuations that occur due to the random distribution of input signals.

For this purpose we derive a *Fokker-Planck equation* describing the adaptation process and allowing a more precise discussion of the dependence of the stationary map on the input signal distribution. We can then show that, under certain conditions, the map takes on a structure which is *spatially periodic* with respect to a subset of the components of the input signal. This result is especially interesting in view of experimentally established spatial periodicities in the response behavior of many neurons belonging to cortical and noncortical areas of the brain. A well-known example of such periodicity is provided by the *ocular dominance stripes* observed in the visual cortex, along which neurons segregate into groups with preference for one eye or the other. A similar structure on a smaller scale than occular

dominance stripes in the striate cortex are orientation columns, a segregation of neurons with receptive fields favouring different orientations in the visual field of an animal (Blasdel and Salama 1986).

14.2. Vector Quantization and Data Compression

An important prerequisite for any kind of information processing is the establishment of an appropriate encoding for the data under consideration. In the case of the brain, this encoding has to a large extent been determined by nature, and indeed one of the principal research questions is to decipher the coding schemes that underly brain function. In the case of artificial information processing systems, the decision for an appropriate encoding of the data is left to the designer, and a determination of which features are to play an important role and must be coded well may be very task specific. However, there are also important aspects of more general relevance. One such aspect is the average code length required for transmission of a specified amount of information. For example, one current scheme for text encoding uses 8-bit words, the so-called ASCII-characters, for individual letters. Theoretically, $2^8 = 256$ distinct characters can be encoded in this way. However, in many cases 128 characters suffice, which can be coded with only 7 bits per character. Taking into account the different frequency with which different characters occur, one can find still more efficient codes. If the characters in the sequence are statistically independent of one another, and if p_i denotes the probability for the occurrence of the ith character, then the lower limit for the most efficient code is

$$S = -\sum_i p_i \operatorname{ld}(p_i) \tag{187}$$

bits per character ("ld" denotes the logarithm to the base two). The quantity S is known as the so-called *Shannon information* (see for example Khinchin 1957) transmitted on the average by a single character. However, for most character sequences, the assumption of statistically independent characters does not hold. By exploiting correlations between several characters, one can find even more efficient codes. For example, in the case

of language, one can encode whole words in place of individual letters, thus achieving a further compactification. Written Chinese provides an example of this strategy.

This sort of code optimization is of particular importance when large quantities of data are to be stored or transmitted. This happens particularly in image processing. The bitwise transmission of a raster image with a resolution of 1,000×1,000 pixels and 256 gray levels per pixel requires the transfer of about 1 Mbyte of data. However, in most images, adjacent pixels are strongly correlated, and significantly more efficient coding schemes than simple bitwise transmission can be found. Interestingly enough, the brain also seems to make use of such possibilities. The optic nerve contains only about 10^6 nerve fibres, whereas the retina is covered by about 10^8 light sensitive receptors (Kandel und Schwartz 1985). Hence, the optic nerve constitutes a kind of "bottleneck" for the transmission of visual information from the retina to the brain. However, before transmission occurs, the signal is subjected to an extensive preprocessing stage in the retina, involving nearly 100 different cell types and enabling a subsequent transmission of all necessary information through the optic nerve.

Hence, data compression is an equally important task for both artificial and natural information processing systems. A general approach developed for the solution of this task is the method of *vector quantization* (see, *e.g.*, Makhoul et al. 1985). This method supposes that the data are given in the form of a set of data vectors $\mathbf{v}(t)$, $t = 1, 2, 3, \ldots$ (possibly of rather high dimension). The index t numbers the individual vectors. The components of a vector $\mathbf{v}(t)$ may take binary, integer, or analogue values, corresponding for example to bits of a binary sequence, gray level values of image pixels, or amplitudes of a speech signal. "Compression" of the data occurs by approximating every data vector $\mathbf{v}(t)$ by a *reference vector* $\mathbf{w}_s$ of equal dimension. This presupposes that a fixed, finite set W of reference vectors $\mathbf{w}_s$ has been established, determined such that a "good" approximate vector $\mathbf{w}_s \in W$ can be found for every data vector that may arise. The set W of reference vectors plays the role of a *code book* assigning to each data vector $\mathbf{v}$ that reference vector $\mathbf{w}_s \in W$ for which the norm of the difference $\delta = \|\mathbf{v} - \mathbf{w}_s\|$ assumes its minimum over all code book vectors. As the new code for the data vector $\mathbf{v}$, it then suffices to specify the index s of the reference vector $\mathbf{w}_s$ that yielded the most

accurate approximation. In the case of a code book with N reference vectors, this requires specification of at most ld N bits. Therefore, the smaller the code book can be chosen, the better the resulting data compression factor.† However, this gain has its price: the original data can no longer be exactly recovered from the codes s. For reconstruction of the original data vector $\mathbf{v}$ from its code s, only the reference vector $\mathbf{w}_s$ is available. This gives rise to a "reconstruction error" that is equal to the approximation error $\delta = \|\mathbf{v} - \mathbf{w}_s\|$.

Crucial for the whole procedure is the construction of a good code book W. It should contain sufficiently many appropriately distributed reference vectors to enable a good approximation to any data vector $\mathbf{v}$ by a reference vector $\mathbf{w}_s$. For a mathematical formulation of this requirement, one often considers the expectation value of the squared reconstruction error, *i.e.*, the quantity

$$E[W] = \int \|\mathbf{v} - \mathbf{w}_{s(\mathbf{v})}\|^2 P(\mathbf{v})\, d\mathbf{v}, \tag{188}$$

where $P(\mathbf{v})$ is the probability density describing the distribution of data vectors $\mathbf{v}$. $E[W]$ depends on the ensemble W of all code book vectors $\mathbf{w}_s$. A frequently appropriate requirement demands the minimization of E subject to the constraint of a fixed, prescribed number of code book vectors $\mathbf{w}_s$ (without such a constraint, E could be reduced to arbitrarily small positive values simply by increasing the number of code vectors. However, this would also entail an arbitrary reduction of the compression effect, since the effort required to specify a single value of s increases with the number N of reference vectors).

The minimization of E with respect to reference vectors $\mathbf{w}_s$ is a complicated, nonlinear optimization problem, for which in most cases no closed solutions are known. Hence, one must resort to iterative approximation methods. In Chapter 15 we will see that these approximation methods are closely related to Kohonen's map-formation algorithm. The maps provided by Kohonen's procedure can be regarded in this context as code books of a vector quantization procedure in which the topology preserving property of the maps leads to a modification of the original error quantity (188).

†The astute reader will notice that the probability distribution of the discrete codes s may be nonuniform. Exploiting this circumstance in the assignment of code words (shorter code words for more frequent codes), one can improve the code efficiency still further.

14.3. Self-Organizing Maps and Vector Quantization

The construction of a good code book requires the minimization of the average reconstruction error $E[\mathbf{w}]$ with respect to the reference vectors $\mathbf{w}_r$. The simplest procedure for this is gradient descent. Starting with initial values $\mathbf{w}_r(0)$, all reference vectors are changed according to

$$\mathbf{w}_r(t+1) = \mathbf{w}_r(t) - \frac{\epsilon}{2} \cdot \frac{\partial E}{\partial \mathbf{w}_r} \tag{189}$$

$$= \mathbf{w}_r(t) + \epsilon \cdot \int_{s(\mathbf{v})=r} (\mathbf{v} - \mathbf{w}_r(t)) P(\mathbf{v})\, d\mathbf{v}, \tag{190}$$

where we employed (188).

The integration condition $s(\mathbf{v}) = r$ restricts the region of integration to those v-values for which $\mathbf{w}_r$ is the most suitable reference vector ($s(\mathbf{v})$ defined through $\|\mathbf{w}_{s(\mathbf{v})} - \mathbf{v}\| = \min_{r'} \|\mathbf{w}_{r'} - \mathbf{v}\|$). For a sufficiently small step size parameter ϵ, repeated application of (190) leads to a decrease of $E[W]$ until a *local* minimum is reached. Equation (190) was first suggested by Linde, Buzo, and Gray (1980) and is known as the "LBG"-procedure. Although this procedure does not guarantee that a *global* minimum is achieved, in many important cases the *local* minimum reached provides a sufficiently good solution. If required, better *local* minima can be found by repeating the procedure with different initial values or with the help of "annealing techniques" (see, for example, Kirkpatrick et al. 1983).

However, carrying out the procedure in this form requires a knowledge of the probability distribution $P(\mathbf{v})$ of the data vectors. Usually, $P(\mathbf{v})$ is not known explicitly. This difficulty can be avoided by replacing (190) with the simpler prescription

$$\mathbf{w}_{s(\mathbf{v})}(t+1) = \mathbf{w}_{s(\mathbf{v})}(t) + \epsilon \cdot \left(\mathbf{v} - \mathbf{w}_{s(\mathbf{v})}(t)\right), \tag{191}$$

where for each step (191) a new data vector v selected at random from the (unknown) distribution is used. For sufficiently small step size ϵ, the accumulation of many individual steps (191) will lead to an approximate realization of the integration in (190)

(the "step counting parameters" t of (190) and (191) of course no longer agree).

Comparison of equation (191) with the adaptation rule (70) in Kohonen's model of self-organizing maps shows that (191) represents a special case of Kohonen's algorithm which results in the limit of vanishing neighborhood cooperation (*i.e.*, $h_{rs} = \delta_{rs}$). Kohonen's algorithm can thus be understood as a generalization of a vector quantization procedure for data compression. The "synaptic strengths" $\mathbf{w_r}$ correspond to the reference vectors, the map provides the code book, and the choice of the excitation center s for an input signal $\mathbf{v}$ defines the mapping $\mathbf{v} \mapsto s(\mathbf{v})$, *i.e.*, corresponds to the *coding step* of the vector quantization procedure. The "receptive fields" F_s introduced earlier (Eq.(99)) comprise just those input signals for which the coding step leads to the same excitation center s.

The shift of a reference vector $\mathbf{w}_s$ in the LBG-procedure (190) always occurs in the direction of the center of gravity $\int_{F_s} \mathbf{v} P d\mathbf{v}$ of the density distribution of the input data, but restricted to the field F_s. This leads to a distribution of reference vectors, in which each reference vector coincides with the center of gravity of the data in "its" field F_s.

The introduction of the neighborhood functions h_{rs} leads to a modification of the distribution of reference vectors compared to standard vector quantization. The average shift of a reference vector then becomes

$$\langle \Delta \mathbf{w}_r \rangle = \sum_s h_{rs} \int_{F_s} (\mathbf{v} - \mathbf{w}_s) P(\mathbf{v})\, d\mathbf{v}, \tag{192}$$

i.e., the shift of $\mathbf{w}_r$ now occurs in the direction of the mean center of gravity of all fields F_s, the contribution of each field being weighted by the neighborhood function h_{rs}. In a stationary state, every reference vector $\mathbf{w}_r$ therefore coincides with a weighted average density, where the weighting is taken over a neighborhood and includes contributions with relative weight h_{rs} from all neighboring fields s for which $h_{rs} \neq 0$.

This no longer leads to minimization of the reconstruction error (188), but to a minimization of a modified expression. For the case of a one-dimensional "chain" of reference vectors, each with n neighbors on both sides (*i.e.*, $h_{rs} = 1$ for $\|r - s\| \leq n$, and $h_{rs} = 0$ otherwise), one finds

$$E[W] = \int \|\mathbf{v} - \mathbf{w}_{s(\mathbf{v})}\|^r P(\mathbf{v})\, d\mathbf{v}, \tag{193}$$

where the exponent r now differs from the value $r = 2$ in (188) taking instead the smaller value

$$r = \frac{1}{2} + \frac{3}{2(2n+1)^2} \tag{194}$$

(Ritter 1989). This can be interpreted as implying that the inclusion of a neighborhood region in each adaptation step leads to a vector quantizer which, relative to a vector quantizer minimizing the quadratic error quantity (188), suppresses small quantization errors.

14.4. Relationship to Principal Component Analysis

Gaining deeper insights into an observed phenomenon often depends crucially on the discovery of a more effective description, involving a smaller number of variables than needed before. This has motivated the search for algorithms that could, at least to some extent, automate the generation of more effective data descriptions.

A rather general and frequent case is the availability of a number of measurements $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots$ of the parameters $\mathbf{v} = (v_1, v_2, \ldots, v_L)^T$ of an experiment. As a rule, the individual parameters v_i will not vary completely independently of one another, but rather will be correlated to a greater or lesser extent, the type of correlation being often unknown. This entails the following question: to what extent can one attribute the observed variation of the measurements to a dependence of the v_i on a smaller number of "hidden" variables $r_1, r_2, \ldots, r_D$, $D < L$? If such dependency exists, one can find L functions $f_1, \ldots, f_L$ of the hidden variables for which

$$v_i = f_i(r_1, r_2, \ldots, r_D), \quad i = 1, \ldots, L, \tag{195}$$

holds. The variables r_i enable then a more economical description of the observed phenomenon compared to the directly available measurements v_i. In particular, they are more likely to correspond to the true "degrees of freedom" that are involved and the number for which, in many cases, is smaller than the number of observed parameters v_i.

Here, one should keep in mind that the new parameters — if such a simplification is possible — are not uniquely determined. Any invertible one-to-one mapping of the r_i onto an equal number of new variables r_i' provides, a priori, an equally "good" set of parameters for a description of the variation of the original variables v_i. Mathematically, each of these different, but equivalent parametrizations can be regarded as a "coordinate system" on an abstract manifold (indeed, this manifold characterizes the system independently of any special choice of coordinates).

However, the non-uniqueness of the parameters r_i makes their general determination difficult. The procedure most frequently applied, *principal component analysis*, makes the simplifying assumption of a *linear relationship between the variables* r_i *and* v_i. This assumption can be viewed geometrically as the introduction of a D-dimensional "hyperplane" lying in the L-dimensional data space, the location and orientation of which are chosen such that every data point can be approximated well by a point of the hyperplane (Fig. 14.1).

This corresponds to a representation of each data point in the form

$$\mathbf{v} = \mathbf{w}^0 + \sum_{i=1}^{D} \mathbf{w}^i r_i(\mathbf{v}) + d_{\mathbf{w}}(\mathbf{v}), \tag{196}$$

where $\mathbf{w}^0, \ldots, \mathbf{w}^D \in R^L$ are $D+1$ vectors specifying the hyperplane and $r_1(\mathbf{v}), \ldots, r_D(\mathbf{v})$ are the new parameters belonging to data point $\mathbf{v}$. Since, as a rule, not all data points will be located within the hyperplane, for most data points a nonvanishing distance $d_{\mathbf{w}}(\mathbf{v})$ perpendicular to the hyperplane results (the index $\mathbf{w}$ is a reminder of the fact that this distance depends on the choice of hyperplane). The choice of hyperplane is optimal if the vectors $\mathbf{w}_i$ are determined such that the weighted mean square residual error $\langle d_{\mathbf{w}}(\mathbf{v})^2 \rangle$, where the weighting factor is the probability density $P(\mathbf{v})$ of the data, takes its smallest possible value, *i.e.*,

$$\int \|\mathbf{v} - \mathbf{w}^0 - \sum_{i=1}^{D} \mathbf{w}^i r_i(\mathbf{v})\|^2 P(\mathbf{v})\, d^L\mathbf{v} = \text{Minimum!} \tag{197}$$

One can show that the solution of this minimization problem yields

$$\mathbf{w}^0 = \int \mathbf{v} P(\mathbf{v})\, d^L\mathbf{v}, \tag{198}$$

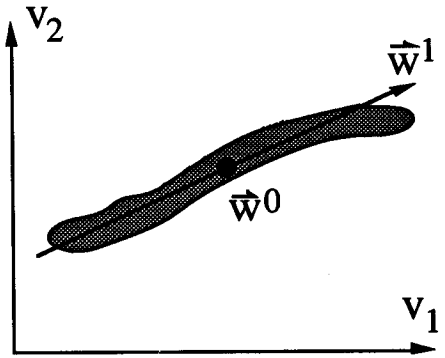

Figure 14.1 Description of a two-dimensional data distribution (shaded region) by a straight line (one-dimensional "hyperplane"). The best description of the distribution results if the line passes through the center of gravity $\mathbf{w}^0$ and is directed parallel to the "principal eigenvector" $\mathbf{w}^1$ (*i.e.*, the eigenvector with largest eigenvalue) of the correlation matrix C.

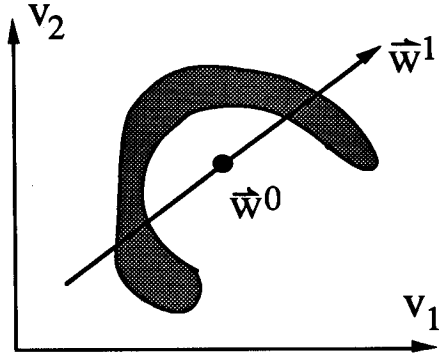

Figure 14.2 If the form of the data distribution is too "nonlinear," no straight line (lowerdimensional hyperplane) leading to a good description of the data can be found.

i.e., $\mathbf{w}^0$ coincides with the center of gravity of the data distribution, whereas the remaining vectors $\mathbf{w}^i$, $i = 1, 2 \ldots, D$, must form a basis of the eigenspace spanned by those D eigenvectors of the correlation matrix that have the largest eigenvalues

$$\mathbf{C} = \int (\mathbf{v} - \mathbf{w}^0) \otimes (\mathbf{v} - \mathbf{w}^0)^T P(\mathbf{v})\, d^L\mathbf{v} \tag{199}$$

(see for example Lawley and Maxwell 1963) when $\otimes$ denotes the tensor product of two vectors, *i.e.*, $(\mathbf{u} \otimes \mathbf{v})_{jk} = u_j v_k$. One possible special choice for the $\mathbf{w}^i$ $(i > 0)$ are the D normalized eigenvectors of C corresponding to the largest eigenvalues. In this case, the new parameters r_i turn out to be the projections of the data vectors along D "principal axes" of their distribution and are called "principal components" of the distribution:

$$r_i = \mathbf{w}^i \cdot \mathbf{v}, \quad i = 1, 2, \ldots, D. \tag{200}$$

Geometrically, this implies that the hyperplane passes through the center of gravity $\mathbf{w}^0$ of the data distribution and is spanned by the D eigenvectors or "principal axes" of the correlation matrix that have the largest eigenvalues. One can show that the orientation of the hyperplane determined in this way maximizes the variance of the perpendicular projection of the data points. The D variables r_i can thus be characterized by the property

to account for (with a linear ansatz) the total data variation as much as possible . However, for the quality of such a description the adequacy of the underlying linearity assumption is crucial: the more the actual distribution of data points deviates from a hyperplane, the worse the description resulting from a projection onto the principal axes of the distribution (Fig. 14.2).

Topology-preserving maps overcome this problem by replacing the linear principal axes or hyperplanes with curved surfaces, which enable a better description of nonlinear data distributions. Here, the maps approximate so-called *principal curves* or *principal surfaces*, which represent a generalization of linear principal axes or eigenspaces. In the following section, we discuss this generalization and its relation to topology-preserving maps.

14.5. Principal Curves, Principal Surfaces and Topology Preserving Maps

Principal component analysis yields a linear description of a prescribed data distribution by a hyperplane that is characterized by the property (198), (197). This can be interpreted geometrically as a minimization of the "mean squared perpendicular distance" $\langle d_{\mathbf{w}}(\mathbf{v})^2 \rangle$ between the data points and the hyperplane. This property motivates a generalization from a hyperplane to nonlinear manifolds (Hastie and Stuetzle 1989). Let us first consider the one-dimensional case. Let $f(s)$ be a "smooth" curve in the space V parametrized by arc length. To every point $\mathbf{v} \in \mathbf{V}$ one can define then a distance $d_f(\mathbf{v})$ to the curve f. Thus, for any such curve and for any density distribution $P(\mathbf{v})$ of points in V we can define a mean squared distance D_f, given by

$$D_f = \int d_f^2(\mathbf{v}) P(\mathbf{v}) \, d^L\mathbf{v}. \tag{201}$$

We call the curve f a *principal curve* of the density distribution $P(\mathbf{v})$, if D_f is extremal, *i.e.*, if the curve is stationary with respect to small, "sufficiently smooth" deformations of the curve.†

†A precise mathematical discussion requires consideration of the special situation at the curve endpoints. We will not go into this problem here. The reader interested in a more thorough discussion is refered to Hastie and Stuetzle (1989).

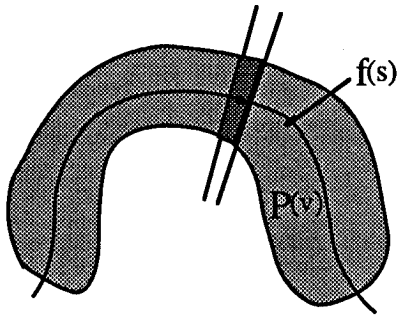

Figure 14.3 Principal curve as nonlinear generalization of the concept of principal axes of a density distribution (shaded). We consider the center of gravity of the density distribution in the region between two "infinitesimally seperated" curve normals. For a principal curve, this center of gravity must always lie on the curve itself.

Intuitively, this requirement demands that a principal curve pass "right through the middle" of its defining density distribution. For better illustration of this situation, we consider a principal curve for the two-dimensional distribution presented in Fig. 14.3. The figure demonstrates that for the principal curve the center of gravity of the density distribution enclosed by two "infinitesimally" distant normals lies on the principal curve. This property must hold, in fact, for every such pair of normals since, otherwise, the mean squared distance D_f could be decreased by a local deformation of the curve in the direction of the deviation, which would contradict the extremal property of D_f. Conversely, the extremality of D_f follows from the fact that the center of gravity of every such "normal strip" coincides with a point on the curve f.

Principal axes arise as a special case of principal curves. One has the following *theorem:* If $P(\mathbf{v})$ has zero mean and a straight line as principal curve, then this principal curve coincides with one of the principal axes of the distribution P (Hastie and Stuetzle 1989).

The generalization to *principal surfaces* and higher-dimensional *"principal manifolds"* proceeds analogously to the one-dimensional case:

> *Definition of a principal surface:* Let $f(\mathbf{s})$ be a surface in the vector space V, *i.e.*, $\dim(\mathbf{f}) = \dim(\mathbf{V}) - 1$, and let $d_f(\mathbf{v})$ be the shortest distance of a point $\mathbf{v} \in V$ to the surface f. f is a principal surface corresponding to a density distribution $P(\mathbf{v})$ in V, if the "mean squared distance"
>
> $$D_f = \int d_f^2(\mathbf{v}) P(\mathbf{v})\, d^L\mathbf{v} \tag{202}$$
>
> is extremal with respect to local variations of the surface.

Thus, Kohonen's algorithm can be interpreted as an approximation procedure for the computation of principal curves, surfaces,

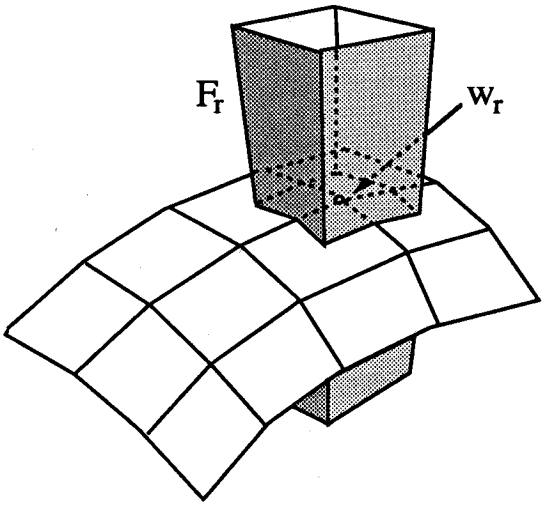

Figure 14.4 2d-Kohonen lattice as discrete approximation to a "principal surface". To each lattice point $\mathbf{w_r}$, a volume $F_\mathbf{r}$ is assigned which is bounded by planes perpendicularly bisecting the distances to the lattice neighbors. The lattice possesses the property of a principal surface, if each lattice point $\mathbf{w_r}$ coincides with the center of gravity of the part of the density distribution enclosed within the volume $F_\mathbf{r}$. This state is approximately achieved as a result of the adaptation procedure of Kohonen's algorithm.

or higher-dimensional principal manifolds. The approximation consists in the discretization of the function f defining the manifold. The discretization is implemented by means of a lattice A of corresponding dimension, where each weight vector $\mathbf{w_r}$ indicates the position of a surface point in the embedding space V. In Kohonen's algorithm, a volume region $F_\mathbf{r}$ was assigned to each point $\mathbf{r}$ of the surface, containing all those points $\mathbf{v}$ for which $\mathbf{w_r}$ is the surface point with the shortest distance (Eq.(99)). $F_\mathbf{r}$ is thus the realization of a volume region which in the continuous limit would be bounded by a "bundle of normals" of infinitesimal cross section penetrating the surface perpendicularly at the point $\mathbf{w_r}$ (Fig. 14.4). The crucial property of Kohonen's algorithm now consists in iteratively deforming the discretized surface in such a way that the center of gravity of the density distribution $P(\mathbf{v})$ contained within the volume $F_\mathbf{r}$ coincides with the surface point $\mathbf{w_r}$ for every $\mathbf{r}$. But this is just the (discretized form) of the condition leading to extremality of the mean squared distance D_f and thus to the "principal surface property" of the stationary state.

As we saw in Section 14.3, however, this property results if and only if $h_{\mathbf{rs}} = \delta_{\mathbf{rs}}$ holds for the neighborhood function. Otherwise, in addition to $F_\mathbf{r}$, other volumes $F_\mathbf{s}$ contribute to the calculation of the equilibrium location of $\mathbf{w_r}$. These volumes lie in a neighborhood about $F_\mathbf{r}$ whose extension is determined by the size of the region within which $h_{\mathbf{rs}}$ differs significantly from zero. This has the effect of "broadening" the volume

region over which the averaging of the probability density is performed in order to obtain the equilibrium location of the center of gravity for the determination of $\mathbf{w_r}$. This is a desirable property for the practical application of the procedure, because most of the data are not given as continuous distributions, but rather as discrete distributions of a finite number of "trials." Strictly speaking, continuous principal manifolds can no longer be defined for such discrete data. A way out of this predicament consists in "smearing out" the data to obtain a better approximation of their underlying probability distribution. The neighborhood function $h_{\mathbf{rs}}$ has just such a "smearing effect," where the amount of "smearing " can be adjusted through the range σ of the neighborhood function. The optimal choice of σ depends on the density of the available data: the principal surfaces thus obtained yield a good description of the data if the neighborhood determined by σ contains sufficiently many data points. For values of the "smearing " that are too small, the surface attempts to touch every single data point, and the desired "smooth" interpolation of the data by a principal surface is lost. In the case of a one-dimensional lattice A, we encountered this behavior (which in that context was desired) in the "traveling salesman problem" of Chapter 6: the curve obtained at the end of the simulation touched every one of the prescribed "cities." The "Peano curve" which, as discussed in Section 4.3, results for a one-dimensional lattice of an infinite number of nodes embedded in a two-dimensional space is another example. For a further discussion of this problem, see also Hastie and Stuetzle (1989).

This section can thus be summarized as follows. Kohonen's algorithm for topology-preserving maps leads to a generalization of standard principal component analysis. The mathematical background of this generalization consists of a nonlinear extension of the concept of principal axes and eigenspaces to so-called *principal curves* and *principal manifolds*. These nonlinear concepts allow one to find dimensionally reduced descriptions even for very nonlinear data distributions, and Kohonen's model can be regarded as an implementation of the required calculations in a neural network.

14.6. Learning as a Stochastic Process

Many learning systems, including Kohonen's self-organizing maps, achieve their goal by means of a sequence of finite adaptation steps. Every single adaptation step results from an "interaction" with the environment. Through these "interactions" information about the environment is obtained. To ensure that the whole sensory space V is explored a random process is employed to generate the sequence of adaptation steps. For example, in the case of sensory maps each of the sensory stimuli $\mathbf{v}$ are chosen at random.

Nevertheless, the assumption of some probability distribution $P(\mathbf{v})$ (usually unknown to the system) for the sensory stimuli frequently provides a reasonable idealization (at least in a stationary environment). The random sequence of input stimuli $\mathbf{v}$ leads to a corresponding random sequence of adaptation steps. Let us denote by $\mathbf{w}$ the ensemble of system parameters which are subject to the learning process (in our case $\mathbf{w} = (\mathbf{w}_{\mathbf{r}_1}, \mathbf{w}_{\mathbf{r}_2}, \ldots, \mathbf{w}_{\mathbf{r}_N})$ is again, as in Section 5.4, the ensemble of all synaptic strengths of a network). Each adaptation step then induces a transformation

$$\mathbf{w}^{new} = \mathbf{T}(\mathbf{w}^{old}, \mathbf{v}). \tag{203}$$

Here, $\mathbf{v}$ is a random variable with probability distribution $P(\mathbf{v})$. Equation (203) does not describe a fixed, deterministic sequence, but rather a "stochastic process." The simulation of such a process provides in each case only one of its infinitely many realizations, a so-called *"sample,"* of the process. To what extent a specific realization represents a "typical" case can only be judged by sufficiently frequent repetition of the simulation. In this way, an "ensemble" of realizations is created, by means of which typical realizations can be identified through their particularly frequent occurrence. Thus, ideally one would like to know the distribution function $\tilde{S}(\mathbf{w}, t)$ of the realizations of an ensemble of infinitely many simulation runs after t time steps, $t = 1, 2, \ldots$. An intuitive picture of $\tilde{S}(\mathbf{w}, t)$ can be given as follows: We consider the space spanned by the synaptic strengths of a network and regard each network of the ensemble as a

point with position vector $\mathbf{w}$ in this space (for a network with N neurons and D synaptic strengths per neuron, this space is a $N \cdot D$-dimensional space). The ensemble can thus be regarded as a cloud of points in this space, and $\tilde{S}(\mathbf{w},t)$ is the density distribution of the points in the cloud. Thus, after t adaptation steps, an "infinitesimal" volume element $d^N\mathbf{w}$ centered at $\mathbf{w}$ contains a fraction $\tilde{S}(\mathbf{w},t)\ d^N\mathbf{w}$ of all ensemble members.

If $\tilde{S}(\mathbf{w},t)$ is known, then all of the statistical properties of the stochastic process can be calculated from it. A typical question can be posed as follows: one has some function $F(\mathbf{w})$ of the synaptic strengths $\mathbf{w}$ and is interested in the average value $\langle F\rangle_t$ to be expected after t adaptation steps. This "expectation value" is then given by

$$\langle F\rangle_t = \int F(\mathbf{w})\tilde{S}(\mathbf{w},t)\ d^N\mathbf{w}. \tag{204}$$

Hence, $\tilde{S}(\mathbf{w},t)$ contains all information to calculate the expectation values of arbitrary functions of the system parameters $\mathbf{w}$. If, for example, one wishes to know the average value $\bar{\mathbf{w}}$ of the synaptic strengths, one chooses $F(\mathbf{w}) = \mathbf{w}$, whereas the choice $F(\mathbf{w}) = (\mathbf{w} - \bar{\mathbf{w}})^2$ yields their mean squared deviation due to the statistical sequence of adaptation steps.

By sufficiently many simulations, one can in principle generate a large ensemble and with it the approximate distribution function $\tilde{S}(\mathbf{w},t)$. However, the required computational effort rapidly rises to an unfeasible level as the desired accuracy and the complexity of the stochastic process increase. In that case, the derivation of analytic results becomes indispensable. This will be the aim of the following sections. The technical point of departure is the derivation of a so-called *Fokker-Planck equation*, which describes the evolution of the distribution function $\tilde{S}(\mathbf{w},t)$ in the vicinity of an equilibrium state and which is valid in the limit of small learning step size ϵ. From this we obtain a necessary and sufficient condition for convergence of the learning procedure to an asymptotic equilibrium state during the final phase of the algorithm. The condition involves an appropriate decrease of the learning step size $\epsilon(t)$. Provided the distribution $P(\mathbf{v})$ is restricted to a multidimensional box volume and is constant there, the statistical fluctuations about the asymptotic equilibrium state can be computed explicitly. From this result, one can conclude that the learning step size must be chosen inversely proportional to the number

of lattice points in order that the remaining fluctuations not exceed a fixed tolerance threshold. We also investigate the ability of the algorithm to automatically use the directions of maximal signal variation as the primary map dimensions. We show that this property derives from an instability which arises when the variance of the sensory events v along a direction which is "poorly" represented by the map exceeds a critical value. The occurrence of this instability manifests itself by strong fluctuations of a characteristic wavelength. Both the critical variance and the characteristic wavelength are computed for the case of a multidimensional box volume.

14.7. Fokker-Planck Equation for the Learning Process

For the derivation of a Fokker-Planck equation that governs the stochastically driven learning process, we consider an ensemble of systems whose states w after t learning steps are distributed according to a distribution function $\tilde{S}(\mathbf{w}, t)$. As in Chapter 5, we assume that all systems are close to the same asymptotic equilibrium state $\bar{\mathbf{w}}$ and that the learning step size ϵ is sufficiently small so that transitions into the neighborhood of different equilibrium states can be neglected. We thus restrict our attention to the asymptotic phase of the convergence behavior which, actually, takes up the largest part of the total computing time in simulations. We obtain the new distribution $\tilde{S}(\mathbf{w}, t+1)$ after an additional learning step from the previous distribution $\tilde{S}(\mathbf{w}, t)$ by integrating over all transitions from states $\mathbf{w}'$ to states w. Each transition contributes with a weight given by the product of the transition probability $Q(\mathbf{w}, \mathbf{w}')$ from $\mathbf{w}'$ to w to the probability $\tilde{S}(\mathbf{w}', t)$ of the occurrence of the state $\mathbf{w}'$ in the ensemble. Both factors were first introduced in Section 5.4. This yields

$$\begin{aligned}\tilde{S}(\mathbf{w}, t+1) &= \int d^N\mathbf{w}'\, Q(\mathbf{w}, \mathbf{w}')\tilde{S}(\mathbf{w}', t) \\ &= \sum_{\mathbf{r}} \int d^N\mathbf{w}' \int\limits_{F_{\mathbf{r}}(\mathbf{w}')} d\mathbf{v}\; P(\mathbf{v})\delta(\mathbf{w} - \mathbf{T}(\mathbf{w}', \mathbf{v}, \epsilon))\tilde{S}(\mathbf{w}', t)\end{aligned} \tag{205}$$

where $P(\mathbf{v})$ and $\mathbf{T}(\mathbf{w}', \mathbf{v}, \epsilon)$ are defined in Section 5.4. In order to carry out the $\mathbf{w}'$-integration, which is taken over all N vector

variables $\mathbf{w}'_{\mathbf{r}}$, $\mathbf{r} \in A$, we require the inverse Jacobian

$$J(\epsilon) = \left[\det \frac{\partial \mathbf{T}}{\partial \mathbf{w}}\right]^{-1}. \tag{206}$$

By assuming for the moment $\mathbf{v} \in F_{\mathbf{s}}(\mathbf{w}')$, we obtain

$$J(\epsilon) = \left[\prod_{\mathbf{r}} (1 - \epsilon h^0_{\mathbf{rs}})\right]^{-d}. \tag{207}$$

Here, d is the dimension of the input vectors $\mathbf{v}$ and we have denoted the excitatory response by $h^0_{\mathbf{rs}}$. Since $h^0_{\mathbf{rs}}$ should only depend on the difference $\mathbf{r}-\mathbf{s}$, J is independent of $\mathbf{s}$ and depends only on ϵ.
The $\mathbf{w}'$-integration yields

$$\begin{aligned}\tilde{S}(\mathbf{w}, t+1) =& J(\epsilon) \sum_{\mathbf{r}} \int \chi_{\mathbf{r}}\left(\mathbf{T}^{-1}(\mathbf{w}, \mathbf{v}, \epsilon), \mathbf{v}\right) \\ & \times P(\mathbf{v}) \tilde{S}\left(\mathbf{T}^{-1}(\mathbf{w}, \mathbf{v}, \epsilon), t\right) d\mathbf{v}.\end{aligned} \tag{208}$$

Here, $\chi_{\mathbf{r}}(\mathbf{w}, \mathbf{v})$ is the characteristic function of the region $F_{\mathbf{r}}(\mathbf{w})$, *i.e.*,

$$\chi_{\mathbf{r}}(\mathbf{w}, \mathbf{v}) = \begin{cases} 1, & \text{if } \mathbf{v} \in F_{\mathbf{r}}(\mathbf{w}); \\ 0, & \text{otherwise.} \end{cases} \tag{209}$$

$\mathbf{T}^{-1}$ denotes the inverse of the transformation $\mathbf{T}(\ ., \mathbf{v}, \epsilon)$. For $\mathbf{v} \in F_{\mathbf{s}}(\mathbf{w})$, $\mathbf{T}^{-1}(\mathbf{w}, \mathbf{v}, \epsilon)$ is given by

$$\left[\mathbf{T}^{-1}(\mathbf{w}, \mathbf{v}, \epsilon)\right]_{\mathbf{r}} = \mathbf{w}_{\mathbf{r}} + \epsilon h_{\mathbf{rs}}(\mathbf{w}_{\mathbf{r}} - \mathbf{v}), \tag{210}$$

where we have introduced the new function $h_{\mathbf{rs}} := h^0_{\mathbf{rs}}/(1 - \epsilon h^0_{\mathbf{rs}})$. In this section $h_{\mathbf{rs}}$ stands for a rescaled excitatory response differing from the original excitatory response $h^0_{\mathbf{rs}}$ only in order ϵ.

For $\epsilon \ll 1$ and $\mathbf{v} \in F_{\mathbf{s}}(\mathbf{w})$, we can expand $\tilde{S}(\mathbf{T}^{-1}(\mathbf{w}, \mathbf{v}, \epsilon), t)$ as

$$\begin{aligned}\tilde{S}(T^{-1}(\mathbf{w}, \mathbf{v}, \epsilon), t) =\ & \tilde{S}(\mathbf{w}, t) + \epsilon \sum_{\mathbf{rm}} h_{\mathbf{rs}}(\mathbf{w}_{\mathbf{r}m} - \mathbf{v}_m) \frac{\partial \tilde{S}}{\partial \mathbf{w}_{\mathbf{r}m}} + \\ & + \frac{1}{2}\epsilon^2 \sum_{\mathbf{rm}} \sum_{\mathbf{rn}} h_{\mathbf{rs}} h_{\mathbf{r}'\mathbf{s}} (\mathbf{w}_{\mathbf{r}m} - \mathbf{v}_m)(\mathbf{w}_{\mathbf{r}'n} - \mathbf{v}_n) \frac{\partial^2 \tilde{S}}{\partial \mathbf{w}_{\mathbf{r}m} \partial \mathbf{w}_{\mathbf{r}'n}} \\ & + O(\epsilon^3).\end{aligned} \tag{211}$$

Correspondingly, $J(\epsilon)$ can be expanded as

$$J(\epsilon) = 1 + \epsilon J_1 + \frac{1}{2}\epsilon^2 J_2 + \ldots, \tag{212}$$

where

$$J_1 = d \cdot \sum_{\mathbf{r}} h_{\mathbf{rs}} = d \cdot \sum_{\mathbf{r}} h_{\mathbf{r0}} \tag{213}$$

is independent of s. Substituting Eq.(211) and (212) into (208) while keeping derivatives up to second order and of these only the leading order in ϵ, we obtain

$$\begin{aligned}\frac{1}{\epsilon}\left[\tilde{S}(\mathbf{w}, t+1) - \tilde{S}(\mathbf{w}, t)\right] = \; & J_1 \tilde{S}(\mathbf{w}, t) \\ & + \sum_{\mathbf{s}} \int_{F_{\mathbf{s}}(\mathbf{w})} d\mathbf{v}\; P(\mathbf{v}) \sum_{\mathbf{rm}} h_{\mathbf{rs}} (\mathrm{w}_{\mathbf{r}m} - \mathrm{v}_m) \frac{\partial \tilde{S}}{\partial \mathrm{w}_{\mathbf{r}m}} \\ & + \frac{\epsilon}{2} \sum_{\mathbf{s}} \int_{F_{\mathbf{s}}(\mathbf{w})} d\mathbf{v}\; P(\mathbf{v}) \\ & \times \sum_{\mathbf{rm}} \sum_{\mathbf{rn}} h_{\mathbf{rs}} h_{\mathbf{r's}} (\mathrm{w}_{\mathbf{r}m} - \mathrm{v}_m)(\mathrm{w}_{\mathbf{r'}n} - \mathrm{v}_n) \frac{\partial^2 \tilde{S}}{\partial \mathrm{w}_{\mathbf{r}m} \partial \mathrm{w}_{\mathbf{r'}n}}.\end{aligned} \tag{214}$$

In the vicinity of the stationary state we expect $\tilde{S}(\mathrm{w}, t)$ to be peaked around the asymptotic equilibrium value $\bar{\mathbf{w}}$. Therefore, we shift variables and define

$$S(\mathbf{u}, t) := \tilde{S}(\bar{\mathbf{w}} + \mathbf{u}, t), \tag{215}$$

i.e., $S(\mathrm{u}, t)$ is the distribution function of the deviations u from the asymptotic equilibrium value $\bar{\mathbf{w}}$. In what follows it is useful to introduce the quantities

$$\hat{P}_{\mathbf{r}}(\mathbf{w}) := \int_{F_{\mathbf{r}}(\mathbf{w})} d\mathbf{v} P(\mathbf{v}), \tag{216}$$

$$\bar{\mathbf{v}}_{\mathbf{r}} := \frac{1}{\hat{P}_{\mathbf{r}}(\mathbf{w})} \int_{F_{\mathbf{r}}(\mathbf{w})} d\mathbf{v} P(\mathbf{v}) \mathbf{v}, \tag{217}$$

$$\mathrm{V}_{\mathbf{r}m}(\mathbf{w}) := \sum_{\mathbf{s}} (\mathrm{w}_{\mathbf{r}m} - \bar{\mathrm{v}}_{\mathbf{s}m}) h_{\mathbf{rs}} \hat{P}_{\mathbf{s}}(\mathbf{w}), \tag{218}$$

$$\begin{aligned}\mathrm{D}_{\mathbf{r}m\mathbf{r'}n}(\mathbf{w}) := & \sum_{\mathbf{s}} h_{\mathbf{rs}} h_{\mathbf{r's}} \Big[(\mathrm{w}_{\mathbf{r}m} - \bar{\mathrm{v}}_{\mathbf{s}m})(\mathrm{w}_{\mathbf{r'}n} - \bar{\mathrm{v}}_{\mathbf{s}n}) \hat{P}_{\mathbf{s}}(\mathrm{w}) \\ & + \int_{F_{\mathbf{s}}(\mathbf{w})} (\mathrm{v}_m \mathrm{v}_n - \bar{\mathrm{v}}_{\mathbf{s}m} \bar{\mathrm{v}}_{\mathbf{s}n}) P(\mathbf{v}) d\mathbf{v} \Big].\end{aligned} \tag{219}$$

$\hat{P}_{\mathbf{r}}(\mathbf{w})$ is the probability for neuron $\mathbf{r}$ to be selected as excitation center, and $\bar{\mathbf{v}}_{\mathbf{r}}$ is the expectation value of all input signals giving rise to this case. $-\mathrm{V}_{\mathbf{r}m}(\mathbf{w})$ can be interpreted as the expectation value for the change $\delta \mathrm{w}_{\mathbf{r}m}$ (change of the synapse between incoming axon m and neuron $\mathbf{r}$) under an infinitesimal learning step, but normalized to $\epsilon = 1$. Correspondingly, $\mathrm{D}_{\mathbf{r}m\mathbf{r}'n}(\mathbf{w})$ is the expectation value of the product $\delta \mathrm{w}_{\mathbf{r}m} \delta \mathrm{w}_{\mathbf{r}'n}$, also normalized to $\epsilon = 1$.

For sufficiently small ϵ we can evaluate the $O(\epsilon)$-term in (214) directly at $\mathbf{w} = \bar{\mathbf{w}}$ and replace $S(\mathbf{u}, t+1) - S(\mathbf{u}, t)$ by $\partial_t S(\mathbf{u}, t)$. This yields the *Fokker-Planck equation*

$$\frac{1}{\epsilon}\partial_t S(\mathbf{u},t) = J_1 S(\mathbf{u},t) + \sum_{\mathbf{r}m} \mathrm{V}_{\mathbf{r}m}(\bar{\mathbf{w}}+\mathbf{u}) \frac{\partial S(\mathbf{u},t)}{\partial \mathrm{u}_{\mathbf{r}m}} + \frac{\epsilon}{2} \sum_{\mathbf{r}m\mathbf{r}'n} \mathrm{D}_{\mathbf{r}m\mathbf{r}'n}(\bar{\mathbf{w}}) \frac{\partial^2 S(\mathbf{u},t)}{\partial \mathrm{u}_{\mathbf{r}m} \partial \mathrm{u}_{\mathbf{r}'n}}. \tag{220}$$

The term with the first derivative represents a "back driving force." It vanishes for $\mathbf{u} = 0$ and must therefore be kept up to linear order in $\mathbf{u}$. This gives

$$\sum_{\mathbf{r}m} \mathrm{V}_{\mathbf{r}m}(\bar{\mathbf{w}}+\mathbf{u}) \frac{\partial S(\mathbf{u},t)}{\partial \mathrm{u}_{\mathbf{r}m}} = -\sum_{\mathbf{r}m} \frac{\partial \mathrm{V}_{\mathbf{r}m}}{\partial \mathrm{w}_{\mathbf{r}m}} S + \sum_{\mathbf{r}m\mathbf{r}'n} \frac{\partial}{\partial \mathrm{u}_{\mathbf{r}m}} \left(\frac{\partial \mathrm{V}_{\mathbf{r}m}}{\partial \mathrm{w}_{\mathbf{r}'n}}(\bar{\mathbf{w}}) \mathrm{u}_{\mathbf{r}'n} S \right). \tag{221}$$

In order to obtain a more convenient form of $\sum_{\mathbf{r}m} \partial \mathrm{V}_{\mathbf{r}m} / \partial \mathrm{w}_{\mathbf{r}m}$, we make use of

$$\begin{aligned} \mathbf{V}_{\mathbf{r}}(\mathbf{w}) &= \sum_{\mathbf{s}} h_{\mathbf{r}\mathbf{s}} \int\limits_{F_{\mathbf{s}}(\mathbf{w})} d\mathbf{v} P(\mathbf{v})(\mathbf{w}_{\mathbf{r}} - \mathbf{v}) \\ &= \frac{1}{\epsilon} \int d\mathbf{v}\ P(\mathbf{v}) \left(\mathbf{w}_{\mathbf{r}} - \mathbf{T}(\mathbf{w}, \mathbf{v}, \epsilon)_{\mathbf{r}} \right) \end{aligned} \tag{222}$$

and obtain

$$\sum_{\mathbf{r}m} \frac{\partial \mathrm{V}_{\mathbf{r}m}}{\partial \mathrm{w}_{\mathbf{r}m}} = \frac{1}{\epsilon} \int d\mathbf{v}\ P(\mathbf{v})\ \mathrm{Tr} \left(1 - \frac{\partial \mathbf{T}}{\partial \mathbf{w}} \right) \tag{223}$$

where Tr denotes the trace operation. The deviation of the Jacobi matrix $\partial \mathbf{T} / \partial \mathbf{w}$ from the unit matrix is of order ϵ. Hence, $\frac{\partial \mathbf{T}}{\partial \mathbf{w}} = \mathbf{1} + \epsilon \mathbf{A}$, and together with (206), one has

$$J(\epsilon) = \det(\mathbf{1} - \epsilon \mathbf{A}) + O(\epsilon^2) = 1 - \epsilon \cdot\ \mathrm{Tr}\ \mathbf{A} + O(\epsilon^2). \tag{224}$$

Comparison with (212) yields

$$J_1 = -\operatorname{Tr} \mathbf{A} = \frac{1}{\epsilon} \operatorname{Tr} \left(\mathbf{1} - \frac{\partial \mathbf{T}}{\partial \mathbf{w}} \right). \tag{225}$$

Substituting this into Eq. (223), we obtain the relation

$$\sum_{\mathbf{rm}} \frac{\partial \mathrm{V}_{\mathbf{rm}}}{\partial \mathrm{w}_{\mathbf{rm}}} = J_1. \tag{226}$$

This leads us to the final form of our equation for the distribution density $S(\mathbf{u}, t)$

$$\begin{aligned} \frac{1}{\epsilon} \partial_t S(\mathbf{u}, t) = & \sum_{\mathbf{rmr'n}} \frac{\partial}{\partial \mathrm{u}_{\mathbf{rm}}} \mathrm{B}_{\mathbf{rmr'n}} \mathrm{u}_{\mathbf{r'n}} S(\mathbf{u}, t) \\ & + \frac{\epsilon}{2} \sum_{\mathbf{rmr'n}} \mathrm{D}_{\mathbf{rmr'n}} \frac{\partial^2 S(\mathbf{u}, t)}{\partial \mathrm{u}_{\mathbf{rm}} \partial \mathrm{u}_{\mathbf{r'n}}}, \end{aligned} \tag{227}$$

where the constant matrix $\mathbf{B}$ is given by

$$\mathrm{B}_{\mathbf{rmr'n}} := \left(\frac{\partial \mathrm{V}_{\mathbf{rm}}(\mathbf{w})}{\partial \mathrm{w}_{\mathbf{r'n}}} \right)_{\mathbf{w}=\bar{\mathbf{w}}}. \tag{228}$$

(227) is the desired Fokker-Planck equation for the asymptotic phase of the map formation process.

One can derive explicit expressions for the expectation value $\bar{\mathrm{u}}_{\mathbf{rm}}(t) = \langle \mathrm{u}_{\mathbf{rm}} \rangle_S$ and the correlation matrix $\mathrm{C}_{\mathbf{rmsn}}(t) = \langle (\mathrm{u}_{\mathbf{rm}} - \bar{\mathrm{u}}_{\mathbf{rm}})(\mathrm{u}_{\mathbf{sn}} - \bar{\mathrm{u}}_{\mathbf{sn}}) \rangle_S$ of the distribution S (see, for example, van Kampen 1981; Gardiner 1985). Defining

$$\mathbf{Y}(t) = \exp\Big(-\mathbf{B} \int_0^t \epsilon(\tau)\, d\tau \Big), \tag{229}$$

one obtains for $\bar{\mathbf{u}}(t)$, The vector with components $\bar{\mathrm{u}}_{\mathbf{rm}}$,

$$\bar{\mathbf{u}}(t) = \mathbf{Y}(t) \bar{\mathbf{u}}(0). \tag{230}$$

Here, $\bar{\mathbf{u}}(0)$ is the expectation value at $t = 0$. The quantity $\bar{\mathbf{u}}(t)$ gives the trajectory of the expectation value of the synaptic strengths and provides a good approximation for the evolution of the system in the limit of sufficiently small learning step size ϵ. For the correlation matrix $\mathbf{C}(t)$, one has (van Kampen, 1981)

$$\mathbf{C}(t) = \mathbf{Y}(t) \Big[\mathbf{C}(0) + \int_0^t \epsilon(\tau)^2 \mathbf{Y}(\tau)^{-1} \mathbf{D} (\mathbf{Y}(\tau)^{-1})^T \, d\tau \Big] \mathbf{Y}(t)^T. \tag{231}$$

If the initial distribution is δ-like, *i.e.*, if $S(\mathbf{u},0) = \prod_{rm} \delta(\mathrm{u}_{rm} - \mathrm{u}(0)_{rm})$ and $\mathbf{C}(t)$ is positive definite, then $S(\mathbf{u},t)$, the solution of Eq. (227), is a Gaussian distribution

$$S(\mathbf{u},t) = \det(2\pi\mathbf{C})^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{u}-\bar{\mathbf{u}})^T\mathbf{C}^{-1}(\mathbf{u}-\bar{\mathbf{u}})\right). \quad (232)$$

If $\epsilon(t)$ is chosen such that the initial conditions become irrelevant in the limit $t \to \infty$, for example if ϵ = constant, the stationary solution can by obtained by substituting the asymptotic values for $\mathbf{C}$ and $\bar{\mathbf{u}}$. If $\mathbf{B}$ and $\mathbf{D}$ commute and ϵ is constant, a further simplification occurs. In this case, one can carry out the integration of (231) explicitly and obtains for the stationary distribution the Gaussian (232) with

$$\mathbf{C} = \epsilon(\mathbf{B} + \mathbf{B}^T)^{-1}\mathbf{D}. \quad (233)$$

14.8. Convergence Condition on Sequences of Learning Step Sizes

The goal of the algorithm is convergence to an asymptotic equilibrium state $\bar{\mathbf{w}}$. In order for this to occur with probability one for every member of the ensemble, the sequence of learning step sizes $\epsilon(t)$ must decrease sufficiently slowly with the number t of learning steps, so that both the variance of the distribution function and the average $\bar{\mathbf{u}}(t)$ of its deviation $\bar{\mathbf{w}}$ vanish in the limit $t \to \infty$. In the following, we derive a necessary and sufficient condition for this.

From (231), one has (van Kampen 1981)

$$\dot{\mathbf{C}} = -\epsilon(t)\left(\mathbf{BC} + \mathbf{CB}^T\right) + \epsilon(t)^2\mathbf{D}. \quad (234)$$

Hence, one obtains for the time derivative of the Euclidean matrix norm $\|\mathbf{C}\|^2 := \sum_{rmr'n} \mathrm{C}^2_{rmr'n}$

$$\frac{1}{2}\partial_t\|\mathbf{C}\|^2 = -\epsilon(t)\ \mathrm{Tr}\ \mathbf{C}(\mathbf{B}+\mathbf{B}^T)\mathbf{C} + \epsilon(t)^2\ \mathrm{Tr}\ \mathbf{DC}. \quad (235)$$

In the following, we require that $\mathbf{C}$ remains bounded if $\epsilon(t)$ is constant and the initial correlation matrix $\mathbf{C}(0)$ is sufficiently small, but otherwise arbitrarily chosen. This is a stability requirement on the equilibrium state $\bar{\mathbf{w}}$. Since $\mathbf{C}$ and $\mathbf{D}$ are both

symmetric and nonnegative, one has $\mathrm{Tr}\,\mathbf{DC} \geq 0$. Hence, by the stability requirement, $(\mathbf{B} + \mathbf{B}^T)$ must be positive. Thus, there exist constants $\beta > 0$ and $\gamma > 0$ such that

$$\mathrm{Tr}\,\mathbf{C}\left[\mathbf{B}(\bar{\mathbf{w}}) + \mathbf{B}(\bar{\mathbf{w}})^T\right]\mathbf{C} > \beta\|\mathbf{C}\|^2/2, \tag{236}$$

and, hence,

$$\partial_t\|\mathbf{C}\|^2 \leq -\epsilon(t)\beta\|\mathbf{C}\|^2 + \epsilon(t)^2\gamma. \tag{237}$$

Integration yields the inequality

$$\|\mathbf{C}(t)\|^2 \leq \gamma \int_0^t \epsilon(t')^2 \exp\left(-\beta \int_{t'}^t \epsilon(t'')\, dt''\right) dt'. \tag{238}$$

Every positive function $\epsilon(t)$ for which the RHS of (238) vanishes asymptotically guarantees the desired convergence of C to zero. In the appendix at the end of this chapter, it is shown that this condition is equivalent to the requirement $\lim_{t\to\infty} \epsilon(t) = 0$, together with

$$\lim_{t\to\infty} \int_0^t \epsilon(t')\, dt' = \infty. \tag{239}$$

With $\lim_{t\to\infty} \mathbf{C}(t) = 0$, this also guarantees $\lim_{t\to\infty} \bar{\mathbf{u}}(t) = 0$ and, hence, convergence to the equilibrium average $\bar{\mathbf{w}}$ with probability one. This criterion cannot be weakened: because of Eq.(234), $\lim_{t\to\infty} \epsilon(t) = 0$ is necessary for the asymptotic vanishing of the variance, and according to (229) and (230), condition (239) is required for $\lim_{t\to\infty} \bar{\mathbf{u}}(t) = 0$. Hence, for convergence to an asymptotic equilibrium state $\bar{\mathbf{w}}$ satisfying the stability requirement, we have shown the following:

> Let $\epsilon(t) > 0$ for all sufficiently small t so that the Markov process (70) can be described by the Fokker-Planck equation (227) in the neighborhood of an equilibrium state. Then the two conditions
>
> $$\lim_{t\to\infty} \int_0^t \epsilon(t')\, dt' = \infty, \tag{240}$$
>
> $$\lim_{t\to\infty} \epsilon(t) = 0 \tag{241}$$
>
> together are necessary and sufficient for the convergence to $\bar{\mathbf{w}}$ of any initial state lying sufficiently close to $\bar{\mathbf{w}}$.

The demand (240) is identical to the first convergence condition of Cottrell and Fort (1986) for a closely related process. Their

second condition, the requirement $\int_0^\infty \epsilon(t)^2\ dt < \infty$, is overly strict in the present case and has been replaced by the weaker condition (241). In particular, (240) and (241) are satisfied for all functions $\epsilon(t) \propto t^{-\alpha}$ with $0 < \alpha \leq 1$. In contrast, the conditions of Cottrell and Forts require $1/2 < \alpha \leq 1$. For $\alpha > 1$ or exponential vanishing of $\epsilon(t)$, (240) is no longer satisfied, and a nonvanishing residual deviation remains even in the limit $t \to \infty$. Nevertheless, (230) and (231) show that the residual error $\bar{\mathbf{u}}$ of the average becomes exponentially small with increasing $\int_0^\infty \epsilon(t)\ dt$. For $\int_0^t \epsilon(t')\ dt' \gg 1$, the main contributions to the residual error come from the equilibrium fluctuations of the correlation matrix C. Hence this error is of order ϵ. Thus, in practical applications, aside from a small residual $\epsilon(t)$, the condition, $\int \epsilon(t)dt \gg 1$ is sufficient, and the precise behavior of $\epsilon(t)$ is of little importance as long as the decrease is monotonic.

14.9. Uniform Signal Density Restricted to a Rectangular Box

In the following sections we consider a Kohonen net which is a two-dimensional lattice A with a three-dimensional input space V. The probability density $P(\mathbf{v})$, $\mathbf{v} \in V$ is assumed to be uniform and restricted to the region of a rectangular box. We also assume that the learning step size ϵ varies sufficiently slowly with the number of learning steps such that at any time t the density $S(\mathbf{u}, t)$ may be replaced by its stationary value for fixed ϵ. Since the input vectors $\mathbf{v}$ are drawn from a volume of dimension three, *i.e.*, larger than the dimension two of the Kohonen net, the Markov process will attempt to project onto the Kohonen net those two directions along which the distribution has its largest variance. In this way, the resulting map is a two-dimensional projection reproducing the higher-dimensional region V as faithfully as possible. Figure 14.5 illustrates this for a three-dimensional rectangular box V of size $40\times40\times10$ and a 40×40-lattice A. Figure 14.5a shows the resulting map again as an "imbedding" in the box V. Since the box is relatively flat, the map is basically a simple projection onto the subspace that is aligned with the two longest sides of the rectangular box.

For nonvanishing ϵ, the learning steps cause continual fluc-

tuations about an average "equilibrium map." These fluctuations appear in Fig. 14.5a as shallow "bumps" and as weak tangential distortions of the lattice. These "bumps"are destortions which will be described quantitatively in this section.

If inputs in case of a d-dimensional input space scatter too much along some or all of the additional $d-2$ dimensions not represented by a two-dimensional Kohonen net, then for many vectors v the restriction of the projection to a reproduction of the two principal directions of V would be unsatisfactory. In this case, the simple projection just described loses its stability and changes into a more complicated equilibrium map. Usually, this new map possesses a lower symmetry and corresponds to an imbedding of the lattice A in V that is strongly folded in the direction of the additional dimensions. This property, known as "automatic choice of feature dimensions," (Kohonen 1984a) is apparent in Fig. 14.5b. In comparison to Fig. 14.5a, the height of the box was increased from 10 to 14 units. The symmetric projection is now no longer stable, and the corresponding imbedding seeks a new configuration. This new configuration breaks the symmetry of the probability distribution $P(\mathbf{v})$ in order to enable a better reproduction of the vertical variation of v by means of an appropriate folding. In the following, we will show that this change to a new equilibrium state arises at a critical value $2s^*$ of the height of the box and that, approaching that value from below, the maps exhibit increasing equilibrium fluctuations of a typical wavelength λ^*. Both values s^* and λ^* will be calculated in the following.

In the mapping of a multidimensional box volume (dimension d) onto a two-dimensional neural net A, each of the $d-2$ "height dimensions" contributes in the same manner and independently of the other dimensions to the instability and to the equilibrium fluctuations. Hence, there is no loss of generality if we consider a three-dimensional box V. We choose for A a square lattice of N×N points[†] and for V the volume $0 \leq x,y \leq N,\ -s \leq z \leq s$. This yields $P(\mathbf{v}) = [2sN^2]^{-1}$ as a homogeneous distribution. In order to avoid boundary effects, we assume periodic boundary conditions along the x- and y-directions. From symmetry considerations, we expect that for sufficiently small s the assignment $\bar{\mathbf{w}}_\mathbf{r} = \mathbf{r},\ \mathbf{r} = me_x + ne_y$ represents the average for $\tilde{S}(\mathbf{w}, t \to \infty)$. In this case, the state

[†]Note that the number of lattice points is now N^2 instead of N.

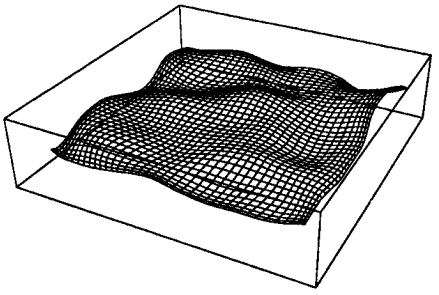

Figure 14.5a "Snapshot" of a Monte Carlo simulation for a 40 x 40-Kohonen net A and a 40 x 40 x 10 units large rectangular box representing the input space V. Due to the sufficiently small box height (10 units), the resulting mapping is essentially a projection perpendicular to the two principal (long) directions of the box. Fluctuations about the equilibrium value due to the statistical sequence of learning steps are evident as shallow "bumps."

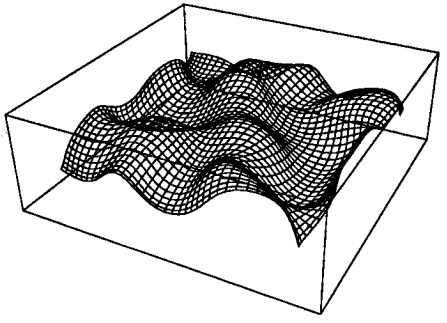

Figure 14.5b The same simulation as in Fig. 14.5a, but for a box height of 14 units. In this case, the state of the net in Fig. 14.5a is no longer stable and a less symmetric configuration emerges. The resulting imbedding achieves a better reproduction of the vertical direction of the map by means of folds extending along this direction.

$\bar{\mathbf{w}}$ is stable up to equilibrium fluctuations. The equilibrium fluctuations can be computed from Eq. (232).

In the following, let $S(\mathbf{u}) = \lim_{t\to\infty} S(\mathbf{u}, t)$ be the stationary distribution of the deviations $\mathbf{u} = \mathbf{w} - \bar{\mathbf{w}}$ from the average value (let ϵ be constant). Due to translational invariance, both $\mathrm{D}_{\mathbf{r}m\mathbf{r}'n}$ and $\mathrm{B}_{\mathbf{r}m\mathbf{r}'n}$ depend only on the difference $\mathbf{r} - \mathbf{r}'$ and on n and m. Hence, we can decouple Eq. (227) if we express $S(\mathbf{u})$ in terms of Fourier amplitudes

$$\hat{\mathbf{u}}_{\mathbf{k}} = \frac{1}{N} \sum_{\mathbf{r}} e^{i\mathbf{k}\cdot\mathbf{r}} \mathbf{u}_{\mathbf{r}}. \tag{242}$$

In fact, the individual amplitudes are distributed independently of one another, *i.e.*, one can express

$$S(\mathbf{u}) = \prod_{\mathbf{k}} \hat{S}_{\mathbf{k}}(\hat{\mathbf{u}}_{\mathbf{k}}), \tag{243}$$

and obtains a set of mutually independent, stationary Fokker-Planck equations for the distributions $\hat{S}_{\mathbf{k}}$ of the individual modes

$$\sum_{mn} \hat{\mathrm{B}}(\mathbf{k})_{mn} \frac{\partial}{\partial \mathbf{u}_m} \mathbf{u}_n \hat{S}_{\mathbf{k}}(\mathbf{u}) + \frac{\epsilon}{2} \sum_{mn} \hat{\mathrm{D}}(\mathbf{k})_{mn} \frac{\partial^2}{\partial \mathbf{u}_m \partial \mathbf{u}_n} \hat{S}_{\mathbf{k}}(\mathbf{u}) = 0. \tag{244}$$

Here, $\hat{\mathbf{D}}(\mathbf{k})$ and $\hat{\mathbf{B}}(\mathbf{k})$ are the d×d matrices

$$\begin{aligned}\hat{\mathbf{D}}(\mathbf{k}) &= \sum_{\mathbf{r}} e^{i\mathbf{k}(\mathbf{r}-\mathbf{r}')}\mathbf{D}_{\mathbf{r}\mathbf{r}'} \\ &= \frac{1}{N^2}\left[(\nabla_k \hat{h}(\mathbf{k}))(\nabla_k \hat{h}(\mathbf{k}))^T + \mathbf{M}\,\hat{h}(\mathbf{k})^2\right]\end{aligned} \tag{245}$$

and

$$\hat{\mathbf{B}}(\mathbf{k}) = \frac{\hat{h}(\mathbf{0})}{N^2}\left[\mathbf{1} - \frac{\hat{h}(\mathbf{k})}{\hat{h}(\mathbf{0})}\hat{\mathbf{a}}(\mathbf{k})\right] - \frac{1}{N^2}\left(i\nabla_k \hat{h}(\mathbf{k})\right)\hat{\mathbf{b}}(\mathbf{k})^T. \tag{246}$$

For a more compact notation, we defined $\mathbf{k} := (k_x, k_y, 0)^T$. M is given by

$$\mathbf{M} = \frac{1}{2s}\int_{F_{\mathbf{r}}(\bar{\mathbf{w}})} d\mathbf{v}\,(\mathbf{v}\mathbf{v}^T - \bar{\mathbf{v}}_{\mathbf{r}}\bar{\mathbf{v}}_{\mathbf{r}}{}^T) = \begin{pmatrix} 1/12 & 0 & 0 \\ 0 & 1/12 & 0 \\ 0 & 0 & s^2/3 \end{pmatrix}, \tag{247}$$

i.e., M is the correlation matrix of the distribution $\hat{P}(\mathbf{v})$ restricted to one of the regions $F_{\mathbf{r}}(\bar{\mathbf{w}})$. Since all of the $F_{\mathbf{r}}(\bar{\mathbf{w}})$ are equal and since $\hat{P}(\mathbf{v})$ is constant, M is independent of the choice of r. The function $\hat{h}(\mathbf{k})$ is the discrete Fourier transform of the neighborhood function $h_{\mathbf{r}\mathbf{s}}$, *i.e.*,

$$\hat{h}(\mathbf{k}) = \sum_{\mathbf{r}} e^{i\mathbf{k}\cdot\mathbf{r}} h_{\mathbf{r}\mathbf{0}}. \tag{248}$$

The matrix $\hat{\mathbf{a}}(\mathbf{k})$ and the vector $\hat{\mathbf{b}}(\mathbf{k})$ are the Fourier transforms of the functions

$$\mathbf{a}_{\mathbf{r}\mathbf{r}'} := \left.\frac{\partial \bar{\mathbf{v}}_{\mathbf{r}}(\mathbf{w})}{\partial \mathbf{w}_{\mathbf{r}'}}\right|_{\bar{\mathbf{w}}}, \tag{249}$$

$$\mathbf{b}_{\mathbf{r}\mathbf{r}'} := \frac{1}{\hat{P}_{\mathbf{r}}}\left.\frac{\partial \hat{P}_{\mathbf{r}}(\mathbf{w})}{\partial \mathbf{w}_{\mathbf{r}'}}\right|_{\bar{\mathbf{w}}}. \tag{250}$$

respectively. The quantities $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ depend only on the geometry of the vectors $\mathbf{w}_{\mathbf{r}}$ in the equilibrium state, but not on the excitatory response h. The matrix $\hat{\mathbf{a}}$ describes the shift of the center of gravity of a region $F_{\mathbf{r}}$ under an infinitesimal change of the equilibrium state, and $\hat{\mathbf{b}}$ describes essentially the corresponding volume change of $F_{\mathbf{r}}$. In the present case, $F_{\mathbf{r}}(\mathbf{w})$ is the volume that is enclosed by the four planes perpendicularly bisecting the distances $\mathbf{w}_{\mathbf{r}} - \mathbf{w}_{\mathbf{r}'}$ ($\mathbf{r}'$ are the nearest lattice neighbors of r)

together with the two planes $z = \pm s$. For this geometry and after some calculation, one obtains

$$\begin{aligned}\mathbf{a}_{\mathbf{r}\mathbf{r}'} = \delta_{\mathbf{r}\mathbf{r}'} &\begin{pmatrix} 2/3 & 0 & 0 \\ 0 & 2/3 & 0 \\ 0 & 0 & 4s^2/3 \end{pmatrix} \\ &- \begin{pmatrix} -1/4 & 0 & 0 \\ 0 & 1/12 & 0 \\ 0 & 0 & s^2/3 \end{pmatrix} \cdot (\delta_{\mathbf{r}+\mathbf{e}_x,\mathbf{r}'} + \delta_{\mathbf{r}-\mathbf{e}_x,\mathbf{r}'}) \\ &- \begin{pmatrix} 1/12 & 0 & 0 \\ 0 & -1/4 & 0 \\ 0 & 0 & s^2/3 \end{pmatrix} \cdot (\delta_{\mathbf{r}+\mathbf{e}_y,\mathbf{r}'} + \delta_{\mathbf{r}-\mathbf{e}_y,\mathbf{r}'})\end{aligned} \tag{251}$$

and

$$\mathbf{b}_{\mathbf{r}\mathbf{r}'} = \frac{1}{2} \sum_{\mathbf{n}=\pm\mathbf{e}_x,\mathbf{e}_y} \mathbf{n}\, (\delta_{\mathbf{r}+\mathbf{n},\mathbf{r}'} - \delta_{\mathbf{r}\mathbf{r}'}). \tag{252}$$

The corresponding Fourier transforms are then

$$\begin{aligned}\hat{\mathbf{a}}(\mathbf{k}) = &\frac{1}{6}(4 + 3\cos k_x - \cos k_y)\mathbf{e}_x\mathbf{e}_x^T \\ &+ \frac{1}{6}(4 - \cos k_x + 3\cos k_y)\mathbf{e}_y\mathbf{e}_y^T \\ &+ \frac{2s^2}{3}(2 - \cos k_x - \cos k_y)\mathbf{e}_z\mathbf{e}_z^T, \end{aligned} \tag{253}$$

$$\hat{\mathbf{b}}(\mathbf{k}) = -i \cdot (\mathbf{e}_x \sin k_x + \mathbf{e}_y \sin k_y). \tag{254}$$

With this, we can discuss the behavior of the system in the vicinity of the state $\bar{\mathbf{w}}$. We can see from $\lim_{k\to 0} \hat{\mathbf{b}}(\mathbf{k}) = 0$ and $\lim_{k\to 0} \hat{\mathbf{a}}_{mn}(\mathbf{k}) = \delta_{mn}(1 - \delta_{m,3})$ that, in the limit of small wavenumbers, for deviations of $\bar{\mathbf{w}}$ along the x- and y-directions the restoring force vanishes, which is consistent with the two vanishing eigenvalues of $\hat{\mathbf{B}}(\mathbf{k})$ in this limit. Hence, long-wavelength fluctuations of these modes can become very large. In contrast, the restoring force to displacements along the z-direction is always nonvanishing even at $\mathbf{k} = 0$.

However, displacements in the z-direction are subject to a different instability. Since $\hat{\mathbf{a}}_{33}(\mathbf{k}) \propto s^2$, $\hat{\mathbf{B}}(\mathbf{k})$ according to (246) can develop a negative eigenvalue for these modes, if s becomes too large. Hence, some or all of these modes can become unstable if s exceeds a critical value s^*. If the variance of $P(\mathbf{v})$ along the "transverse" dimensions is too large, this causes

the system to assume a new equilibrium state which as a rule breaks the symmetry of the distribution $P(\mathbf{v})$. A precursor to this symmetry breaking is an increase of fluctuations of a characteristic wavelength λ^*.

For a more detailed analysis and calculation of λ^* and s^* we now turn to the two cases of long- and short-range interactions (neighborhood functions) $h_{\mathbf{rs}}$.

14.9.1. Long-Range Interaction

We consider as the interaction a Gaussian

$$h_{\mathbf{rr}'} = \sum_{\mathbf{s}} \delta_{\mathbf{r}+\mathbf{s},\mathbf{r}'} \exp\left(-\frac{s^2}{2\sigma^2}\right) \tag{255}$$

with range σ, where we require $1 \ll \sigma \ll N$. In this case, we can replace the discrete Fourier series to a good approximation by the continuous transform and obtain

$$\hat{h}(\mathbf{k}) = 2\pi\sigma^2 \exp(-\sigma^2 k^2/2). \tag{256}$$

Substitution of (256) into (245) yields

$$\hat{\mathbf{D}}(\mathbf{k}) = \frac{4\pi^2\sigma^4}{N^2}\left[\mathbf{k}\mathbf{k}^T\sigma^4 + \mathbf{M}\right] \exp(-k^2\sigma^2). \tag{257}$$

The nonvanishing elements of $\hat{\mathbf{B}}(\mathbf{k})$ are

$$\hat{\mathbf{B}}_{11} = \frac{2\pi\sigma^2}{N^2}\left[1 - \frac{1}{6}(4 + 3\cos k_x - 6k_x\sigma^2 \sin k_x - \cos k_y) \cdot e^{-\frac{1}{2}k^2\sigma^2}\right], \tag{258}$$

$$\hat{\mathbf{B}}_{22} = \frac{2\pi\sigma^2}{N^2}\left[1 - \frac{1}{6}(4 - \cos k_x - 6k_y\sigma^2 \sin k_y + 3\cos k_y) \cdot e^{-\frac{1}{2}k^2\sigma^2}\right], \tag{259}$$

$$\hat{\mathbf{B}}_{33} = \frac{2\pi\sigma^2}{N^2}\left[1 - \frac{2s^2}{3}(2 - \cos k_x - \cos k_y) \exp(-k^2\sigma^2/2)\right], \tag{260}$$

$$\hat{\mathbf{B}}_{12} = \frac{2\pi\sigma^4}{N^2} \cdot k_x \sin k_y \cdot \exp(-k^2\sigma^2/2), \tag{261}$$

$$\hat{\mathbf{B}}_{21} = \frac{2\pi\sigma^4}{N^2} \cdot k_y \sin k_x \cdot \exp(-k^2\sigma^2/2). \tag{262}$$

In order to simplify these expressions, we use the fact that for $\sigma \gg 1$ either $e^{-\sigma^2 k^2}$ is very small or k_x and k_y admit an expansion of the angular functions. Neglecting as well the k^2-terms compared to $k^2\sigma^2$-terms, we obtain for $\hat{\mathbf{B}}$ the simpler form

$$\hat{\mathbf{B}}(\mathbf{k}) \approx \frac{2\pi\sigma^2}{N^2}\left[1 - \left(1 - \sigma^2\mathbf{k}\mathbf{k}^T + \frac{s^2k^2}{3}\mathbf{e}_z\mathbf{e}_z^T\right) \exp\left(-k^2\sigma^2/2\right)\right]. \tag{263}$$

In this approximation, $\hat{\mathbf{B}}(\mathbf{k})$ and $\hat{\mathbf{D}}(\mathbf{k})$ commute with each other and, in fact, possess the same eigenvectors, *i.e.*, $\vec{\xi}_3 = \mathbf{e}_z$, $\vec{\xi}_2 = \mathbf{k}$ and the vector $\vec{\xi}_1 = \mathbf{k}^\perp$ perpendicular to both of these. The corresponding eigenvalues λ_n^B and λ_n^D for $\hat{\mathbf{B}}(\mathbf{k})$ and $\hat{\mathbf{D}}(\mathbf{k})$ are

$$\begin{aligned} \lambda_1^B(\mathbf{k}) &= \frac{2\pi\sigma^2}{N^2}\left(1 - e^{-k^2\sigma^2/2}\right); \\ \lambda_1^D(\mathbf{k}) &= \frac{\pi^2\sigma^4}{3N^2} e^{-k^2\sigma^2}; \end{aligned} \tag{264}$$

$$\begin{aligned} \lambda_2^B(\mathbf{k}) &= \frac{2\pi\sigma^2}{N^2}\left(1 - (1 - k^2\sigma^2)e^{-k^2\sigma^2/2}\right); \\ \lambda_2^D(\mathbf{k}) &= \frac{\pi^2\sigma^4}{3N^2}(12k^2\sigma^4 + 1)e^{-k^2\sigma^2}; \end{aligned} \tag{265}$$

$$\begin{aligned} \lambda_3^B(\mathbf{k}) &= \frac{2\pi\sigma^2}{N^2}\left(1 - \frac{s^2k^2}{3} e^{-k^2\sigma^2/2}\right); \\ \lambda_3^D(\mathbf{k}) &= \frac{4\pi^2\sigma^4}{3N^2}\, s^2 e^{-k^2\sigma^2}. \end{aligned} \tag{266}$$

$\hat{\mathbf{B}}$ gives the strength of the "drift term" driving the expectation value of the distribution toward the equilibrium average. Hence, by (264) and (265), the system exhibits more "stiffness" against displacements along the $\vec{\xi}_2$-mode and, thus, parallel to k than against displacements along the $\vec{\xi}_1$-mode and, thus, perpendicular to k. For wavelengths large compared to the range σ of h_{rs}, we have asymptotically $\lambda_2^B(\mathbf{k}) = 3\lambda_1^B(\mathbf{k}) = O(k^2)$, *i.e.*, the $\vec{\xi}_2$-mode is three times stiffer as the $\vec{\xi}_1$-mode, and both "stiffnesses" vanish in the limit $\mathbf{k} \to 0$. However, this does not hold for the $\vec{\xi}_3$-mode, which owes its stability to sufficiently small values of s. If s becomes too large, then $\lambda_3^B(\mathbf{k})$ can become negative for a whole band of k-values. The corresponding modes $\vec{\xi}_3(\mathbf{k})$ then become unstable, the chosen state $\bar{\mathbf{w}}$ no longer represents the average equilibrium value, and the system seeks a new equilibrium. This can be seen even more clearly from the fluctuations of the corresponding mode amplitudes u_n. From (233) follows

$$\langle u_n(\mathbf{k})^2 \rangle = \frac{\epsilon\lambda_n^D(\mathbf{k})}{2\lambda_n^B(\mathbf{k})}, \quad n = 1, 2, 3. \tag{267}$$

All other correlations vanish. We thus obtain

$$\langle u_1(\mathbf{k})^2 \rangle = \epsilon\pi\sigma^2 \frac{\exp(-k^2\sigma^2)}{12(1 - \exp(-k^2\sigma^2/2))}, \tag{268}$$

$$\langle u_2(\mathbf{k})^2\rangle = \epsilon\pi\sigma^2 \frac{(12k^2\sigma^4+1)\exp(-k^2\sigma^2)}{12-12(1-k^2\sigma^2)\exp(-k^2\sigma^2/2)}, \quad (269)$$

$$\langle u_3(\mathbf{k})^2\rangle = \epsilon\pi\sigma^2 \frac{s^2\exp(-k^2\sigma^2)}{3-s^2k^2\exp(-k^2\sigma^2/2)}. \quad (270)$$

For the fluctuations of u_1 and u_2, the deviation of $\mathbf{w_r}$ from the equilibrium $\bar{\mathbf{w}}_\mathbf{r}$ lies along one of the two principal directions of the map. In the map, these fluctuations affect the image locations $\mathbf{r}$ of the region $F_\mathbf{r}$ and, therefore, are called "longitudinal" in what follows. From (268) and (269), we see that these fluctuations for wavelengths shorter than σ are practically absent. Hence, the main contribution to statistical distortions of the map comes from fluctuations of long wavelength, whose amplitudes are subject to a $1/k^2$-singularity. For an estimate of the influence of these fluctuations, we expand (268) for the lowest possible wavenumber $k = 2\pi/N$, where we assume $k\sigma = 2\pi\sigma/N << 1$. This yields

$$\langle u_1^2\rangle^{1/2} \approx N\sqrt{\epsilon/24\pi} \approx 0.12N\epsilon^{1/2}. \quad (271)$$

In order for this not to exceed a fixed, prescribed number of lattice constants, ϵ must be chosen inversely proportional to the number N^2 of lattice points of A. For practical applications, these distortions, which are smooth and distributed over large distances, are not disturbing, since one is often mainly interested in the correct, two-dimensional reproduction of the neighborhood relationships in the original higher-dimensional space V. Therefore, for many applications, a significantly larger learning step size ϵ is allowable even in the final phase of the algorithm.

The u_3-mode describes the deviation of each $\mathbf{w_r}$ along the direction perpendicular to the local imbedding plane of A in V. According to (270), its amplitude remains bounded, in contrast to u_1 and u_2, even at $\mathbf{k} = 0$, but, as mentioned previously, its stability depends crucially on s. Instability occurs for s-values for which the denominator of (270) no longer is positive for all k-values. This is the case for $s > s^* = \sigma\sqrt{3e/2} \approx 2.02\ \sigma$. For $s = s^*$, the wavelength of the marginally unstable mode is $\lambda^* = \sigma\pi\sqrt{2} \approx 4.44\ \sigma$. A mapping is hence stable if and only if the variance of $P(\mathbf{v})$ transverse to the imbedding plane does not exceed a maximal value that is proportional to the range σ of $h_{\mathbf{rs}}$. If necessary, the algorithm enforces this condition by an appropriate folding of the imbedding. By the choice of σ, one can control what variance will be tolerated before

folds occur. If s approaches the limiting value s^* from below, the system exhibits fluctuations which grow as the difference to s^* becomes smaller, and which are particularly evident in the vicinity of the wavelength λ^*. The fluctuations, in case of further increasing s, lead to a destabilization of the symmetric equilibrium distribution above s^*.

14.9.2. Short-Range Interaction

We consider now the short-range limit, in which $h_{\mathbf{rs}}$ extends only as far as the nearest neighbors, *i.e.*,

$$h_{\mathbf{rs}} = \delta_{\mathbf{rs}} + \sum_{\mathbf{n}=\pm\mathbf{e}_x,\mathbf{e}_y} \delta_{\mathbf{r}+\mathbf{n},\mathbf{s}}. \tag{272}$$

In this case holds

$$\hat{h}(\mathbf{k}) = 1 + 2\cos k_x + 2\cos k_y. \tag{273}$$

For the representative case $k_y = 0$, $k := k_x$, one has

$$\langle u_1(\mathbf{k})^2\rangle = \frac{\epsilon\cdot(3+2\cos k)^2}{4(1-\cos k)(9-2\cos k)}, \tag{274}$$

$$\langle u_2(\mathbf{k})^2\rangle = \frac{\epsilon\cdot(44\sin^2 k + 12\cos k + 13)}{12(1-\cos k)(11+6\cos k)}, \tag{275}$$

$$\langle u_3(\mathbf{k})^2\rangle = \frac{\epsilon s^2\cdot(1+2\kappa)^2}{2(4s^2\kappa^2 - 6s^2\kappa + 15 - 4s^2)}, \tag{276}$$

with $\kappa := \cos k_x + \cos k_y$. Expression (276) also holds for $k_y \neq 0$. There is again a $1/k^2$-singularity of the longitudinal fluctuations. As before, $\hat{\mathbf{B}}_{11}(\mathbf{k}) > \hat{\mathbf{B}}_{22}(\mathbf{k})$, *i.e.*, the restoring force for displacements in the direction of k is again higher than for displacements perpendicular to it. Due to $\hat{\mathbf{D}}_{11}(\mathbf{k}) = \hat{\mathbf{D}}_{22}(\mathbf{k})$, this behavior arises also for the smaller fluctuations of the "stiffer" mode. By considerations similar to those of section 14.9.1, one has $\langle u_1(\mathbf{k})^2\rangle_{max}^{1/2} \approx 0.2\epsilon^{1/2}N$. Hence, the limitation of the fluctuations to a fixed number of lattice constants requires $\epsilon \propto 1/N^2$. The critical limit for the occurrence of the transverse instability becomes $s^* = \sqrt{12/5} = 1.549$, and the corresponding first unstable modes belong to $\kappa^* = 3/4$. For $k_y = 0$, this corresponds to the relatively small wavelength of 3.45 lattice constants, *i.e.*, again as in the long wavelength case, a wavelength of the order of the range of $h_{\mathbf{rs}}$.

14.9.3. Comparison with Monte-Carlo Simulations

In this section, we compare the analytical results obtained in Sections 14.9.1 and 14.9.2 with data from Monte-Carlo simulations of the Markov process (70) for the cases of long-range (Eq.(255)) and short-range (Eq.(272)) excitatory response $h_{\mathbf{rs}}$.

In the first simulation, we use a square 32×32-lattice (*i.e.*, $N = 32$) with the short-range excitatory response (272) and constant learning step size $\epsilon = 0.01$. Beginning with the equilibrium state $\bar{\mathbf{w}}_{\mathbf{r}} = m\mathbf{e}_x + n\mathbf{e}_y$, $m, n = 1, 2, \ldots, 32$, 20,000 "snapshots" of the Markov process described by (70) were generated in intervals of 2,000 Markov steps for the evaluation of the correlation function $\langle u_n(\mathrm{k})^2 \rangle$. For the ensemble of states obtained in this manner, the correlation function $f_n(\mathrm{k}) := \langle u_n(\mathrm{k})^2 \rangle^{1/2}$, $n = 1, 2, 3$ was evaluated at the discrete wave vectors $\mathrm{k} = \mathbf{e}_x \cdot 2\pi l/N$, $l = 1, \ldots, 32$. The data points, thus obtained for the "hard" mode u_1 and the "soft" mode u_2, are presented for $s = 10^{-4}$ in Fig. 14.6 and Fig. 14.7. Also shown are the predictions on the basis of (274) and (275). Obviously, the analytical description agrees very well with the simulation data. Figure 14.8 shows the dependence of the transverse fluctuations (in units of s) on the height $2s$ of the box for parameter values $s = 10^{-4}$, $s = 1.3$, and $s = 1.5$. The transverse fluctuations are described by the correlation function $f_3(\mathrm{k})$ and were obtained through simulations and from Eq. (276). For $s = 10^{-4}$, *i.e.*, essentially a very flat, two-dimensional box, the fluctuations decrease monotonically with wavelength. As s approaches the critical value s^*, the fluctuations of the modes near $k^* = 0.58\pi$ increase markedly. At $s = 1.5$, *i.e.*, just below $s^* \approx 1.54$, the fluctuations already take up a significant fraction of the box volume height and, thus, indicate the incipient instability. For all three parameter values, the agreement between the theoretical graphs and simulation data is very good.

A similar Monte-Carlo simulation for the long-range excitatory response is difficult to perform because of the considerably higher computational effort. Therefore, for this case we have carried out a simulation for a one-dimensional lattice consisting of $N = 128$ points. The box volume is replaced by a rectangular strip of length N and vertical extension $2s$. The learning step size was again $\epsilon = 0.01$. In this case, we generated an ensemble of states consisting of $10,000$ "snapshots" at intervals of 1000 Markov steps. The derivations of the preceding Section are eas-

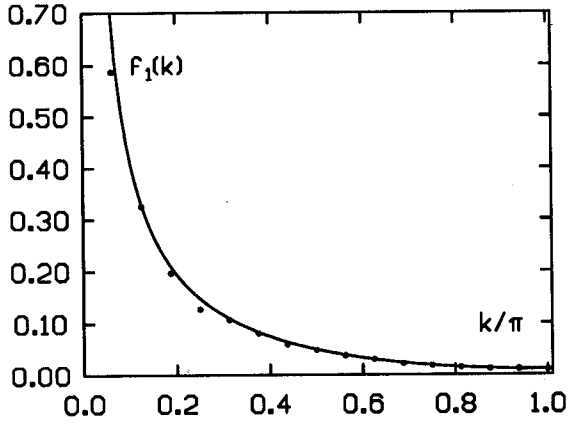

Figure 14.6 Dependence of the fluctuations of the "soft" mode u_1 for a short-range excitatory response of Eq. (272) on the wavenumber k. The data points are from a Monte-Carlo simulation of the Markov process (70) with fixed $\epsilon = 0.01$ and $s = 10^{-4}$. Superimposed is the dependence according to Eq.(274).

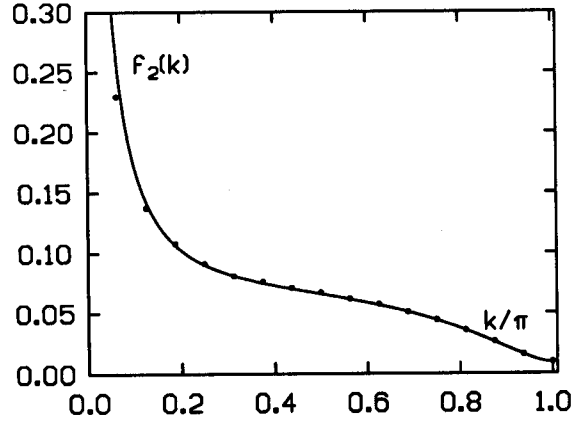

Figure 14.7 Fluctuations of the "hard" mode u_2, obtained from the same simulation as in Fig.14.6 (analytic result according to Eq.(275)). For small wavenumbers, the fluctuations are smaller than for u_1.

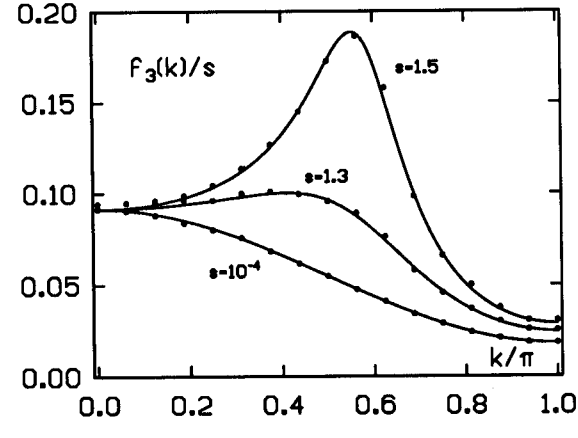

Figure 14.8 Fluctuations of the "transverse" mode u_3 (analytic results according to Eq. (276)) for three different values of the height parameter s: for $s = 10^{-4}$, *i.e.*, an essentially two-dimensional probability distribution, there are only small transverse fluctuations. For $s = 1.3$, the fluctuations begin to show a broad maximum near $k = 0.58\pi$. This is quite evident for $s = 1.5$, *i.e.*, just below s^*.

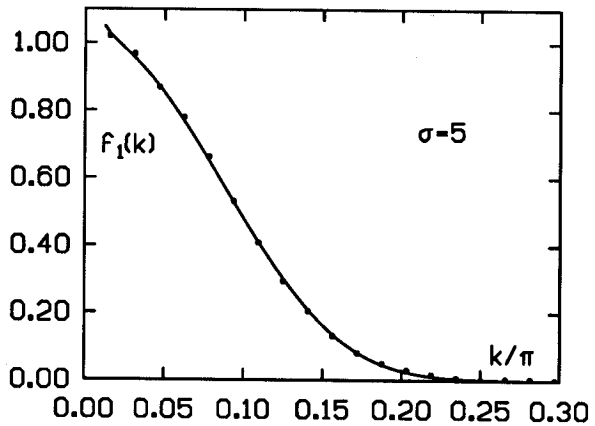

Figure 14.9 Dependence of the longitudinal fluctuations on the wavenumber k for a Gaussian excitatory response (255) with $\sigma = 5$. The data points pertain to a Monte-Carlo simulation of a chain with $N = 128$ points. Superimposed is the theoretical graph according to Eq.(277). The exponential fall-off at large wavenumbers is correctly reproduced by the data.

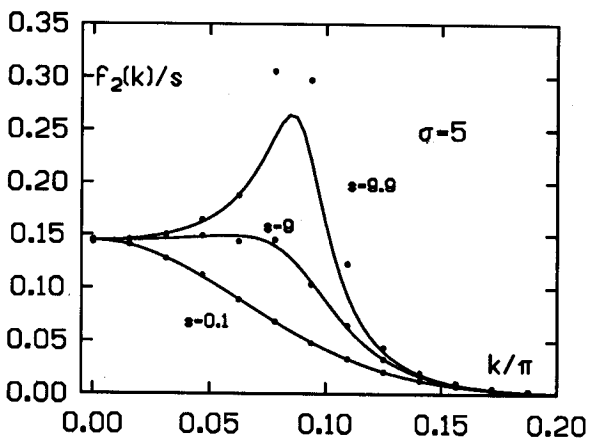

Figure 14.10 The corresponding transversale fluctuations for three different values of s (analytical results according to Eq.(278)). In comparison to Fig. 14.8, the critical value is now $s^* \approx 10.1$, and the fluctuations show an exponential fall-off for larger k-values. The maximum, related to the transverse instability, is shifted in comparison to Fig. 14.8 toward lower k-values.

ily adapted to the present situation and yield for the equilibrium fluctuations of the longitudinal (u_1) and transverse (u_2) modes (here the only ones):

$$\langle u_1(\mathbf{k})^2\rangle = \frac{\epsilon\sigma\sqrt{2\pi}(12k^2\sigma^4+1)\exp(-k^2\sigma^2)}{12(2-[1+\cos k-2\sigma^2 k\sin k]\exp(-k^2\sigma^2/2))}, \quad (277)$$

$$\langle u_2(\mathbf{k})^2\rangle = \frac{\epsilon\sigma\sqrt{2\pi}s^2\exp(-k^2\sigma^2)}{6-4s^2(1-\cos k)\exp(-k^2\sigma^2/2)}. \quad (278)$$

These expressions are, up to an additional factor of $(\sigma\sqrt{2\pi})^{-1}$, identical to the results (268) and (270) for the two-dimensional lattice in the limit $k \to 0$. In particular, for s^* and λ^* we obtain the same values as before. Figure 14.9 and Fig. 14.10 show a comparison of the shape of the theoretical correlation functions according to (277) and (278) with the data from a Monte-Carlo simulation at $\sigma = 5$. Figure 14.9 shows the data points of the

simulation for the longitudinal fluctuations $f_1(k)$ and $s = 0.1$. The expected exponential fall off for $k^2\sigma^2 > 1$ is reproduced well. On the other hand, the expected $1/k$-singularity for $f_1(k)$ is not visible, since the very small k-values required are possible only for considerably longer chains. Figure 14.10 shows the transverse fluctuations $f_2(k)$ for the three cases $s = 0.1$, *i.e.*, essentially a one-dimensional input vector distribution, $s = 9.0$ (still significantly below the critical value $s^* \approx 2.02\sigma \approx 10.1$), and $s = 9.9$ which is just below s^*. The main differences between the present case and the short-range case presented in Fig. 14.8 turn out to be the shift of the instability (maximum of $f_2(k)$) to shorter wavenumbers and the exponential fall-off of the fluctuations for $k\sigma >> 1$.

14.10. Interpretation of Results

In this section, we summarize the results of the preceding sections 14.8–14.9 and interpret them in terms of biological maps.

The situation analysed in Section 14.8 can be regarded as the simplest possible "scenario" in which a "dimensionality conflict" arises between the manifold of input signals (3-dimensional rectangular box) and the topology of the map (two-dimensional surface). The quantity determining the "strength" of the "conflict" is the height dimension $2s$ of the box volume. For small values of s, the variation of the input signal along the vertical dimension is hardly noticeable, and the structure of the resulting map is not affected by this part of the input signal variation. In this case the map corresponds geometrically to a vertical projection of the box onto a horizontal plane.

However, as shown by our analysis, this map only remains stable as long as $s \leq s^* = \sigma\sqrt{3e/2}$ is satisfied. In this stability region, the components w_{r3} of all weight vectors fluctuate about their common average value zero, and the size of the fluctuations decreases with the square root of the adaptation step size. The "stability threshold" s^* *can be interpreted essentially as that distance in the space of input signals which corresponds to the range of the neighborhood function* h_{rs} *in the lattice.* For $s > s^*$, a map with periodic "distortions" develops. Mathematically, these "distortions" stem from those components w_{r3} the average values of which are no longer spatially constant above

the stability threshold, but rather vary with position $\mathbf{r}$ in the map. This variation exhibits a periodic pattern and begins to makes itself felt even below the threshold s^* by an increase of wavelike fluctuations about the equilibrium value $\mathbf{w}_{\mathbf{r}3} = 0$. Here, contributions from fluctuations of wavelength $\lambda^* = \sigma\pi\sqrt{2}$ dominate, that is, *the scale of the dominant wavelengths is also determined by the range of the neighborhood function.*

In the context of a pattern processing task, the x- and y-coordinate would have the interpretation of two "primary" features, characterized by large variations. In contrast, the z-coordinate would correspond to a "secondary" feature with less strongly evident variation. As long as $s < s^*$, the system converges to a topographic map of the two "primary" features alone, and the "secondary" feature remains completely invisible in the map. As soon as the variation of the "secondary" feature exceeds the threshold value given by s^*, a map is created in which the "secondary" feature is also visible. This happens in such a way that the components of the weight vector $\mathbf{w}_\mathbf{r}$ become position dependent in the direction of the "secondary" feature. If one represents the values $\mathbf{w}_{\mathbf{r}3}$ of these components by gray levels, one finds an irregular pattern consisting of black and white stripes, as shown in Fig. 14.11.

Interestingly enough, in the brain one finds a whole series of two-dimensional arrangements of neurons whose response properties are distributed in qualitatively similar spatial patterns. The best-known examples of this are the "ocular dominance stripes," an irregular pattern of stripes containing neurons that prefer either the left or the right eye as their input, as well as the "orientation columns," along which neurons reacting to stimulation of the retina by brightness edges of the same orientation are grouped. In both cases, the response behavior of the neurons is described (to a first approximation) by three "stimulus variables," and there is a "dimensionality conflict" for the distribution of these parameters on the two-dimensional visual cortex: in addition to the two "primary" stimulus variables "retinal position" (x- and y-coordinates), the relative weight of the input of both eyes is a "secondary" feature in the case of the ocular dominance stripes. On the other hand, in the orientation stripes, the "secondary" feature is the orientation of the brightness edge, and each neuron — in addition to its specialization to a particular retinal position — will respond well to a small range of edge orientations only. Several models

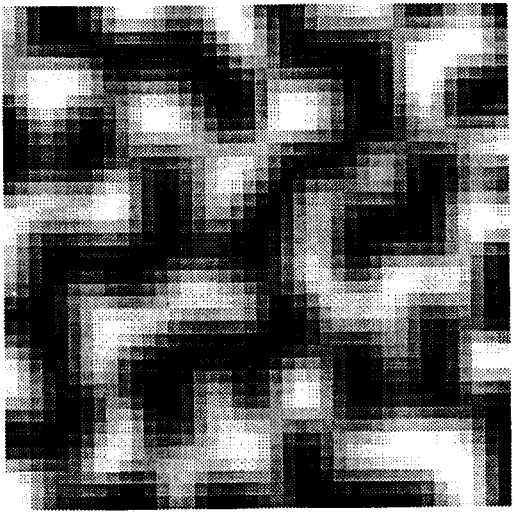

Figure 14.11 Topographic map with periodic structure of stripes. The input signals came from a three-dimensional feature space $0 \leq x, y \leq 40$, $-4 \leq z \leq 4$. The map was generated by Kohonen's algorithm on a 40×40-lattice ($\sigma = 1.4$, 10^4 steps). The height (z-dimension) plays the role of the "secondary" feature, whose distribution in the map is represented by gray levels. The resulting pattern qualitatively resembles the pattern of *ocular dominance stripes* observed in the visual cortex, into which cells with a preference for the same eye become segregated, or of *orientation columns* in the striate cortex separating cells with receptive fields of different orientation.

for the description of such spatial patterns of neural stimulus variables have been suggested in the past. The papers of von der Malsburg (1979, 1982), Willshaw and von der Malsburg (1976), Takeuchi and Amari (1979), as well as Miller et al. (1989) represent some selected contributions to this area. In particular, the ability of Kohonen's model to generate such striped patterns was noticed very early by Kohonen himself in computer simulations (Kohonen 1982a). However, until recently this important property of the model received only little attention by other researchers. The derivation given here augments the earlier simulation results by means of a mathematical analysis that can serve as a point of departure for the mathematical treatment of more realistic versions of Kohonen's model. It shows that stripe formation can be regarded as an instability against wavelike "distortions" resulting from a 'dimensionality conflict" between input signals and the neuron layer.

14.11. Appendix

In this appendix, we show that for every positive function $\epsilon(t)$ the conditions

$$\lim_{t\to\infty}\int_0^t \epsilon(\tau)\,d\tau = \infty \qquad \lim_{t\to\infty}\epsilon(t) = 0 \tag{i}$$

and

$$\lim_{t\to\infty}\int_0^t \epsilon(t')^2 \exp\Big(-\beta\int_{t'}^t \epsilon(t'')\,dt''\Big)\,dt' = 0. \tag{ii}$$

are equivalent for arbitrary $\beta > 0$.

Proof: $(ii) \to (i)$ is obvious for $\epsilon > 0$; $(i) \to (ii)$:

Choose $\delta > 0$ arbitrarily small and $a > 0$ such that $\epsilon(t) < \beta\delta$ holds for all $t > a$. Let $\epsilon_{max} := \max_t \epsilon(t)$. Then a $b > a$ can be chosen such that $\exp(-\beta\int_a^t \epsilon(\tau)\,d\tau) < \beta\delta/\epsilon_{max}$ holds for all $t > b$. It then follows for all $t > b$ that:

$$\begin{aligned}
&\int_0^t \epsilon(t')^2 \exp\Big(-\beta\int_{t'}^t \epsilon(t'')\,dt''\Big)\,dt' = \\
&= \frac{1}{\beta}\Big(\int_0^a + \int_a^t\Big)\Big[\epsilon(t')\frac{\partial}{\partial t'}\exp\Big(-\beta\int_{t'}^t \epsilon(t'')\,dt''\Big)\Big]\,dt' \\
&\le \frac{\epsilon_{max}}{\beta}\Big[\exp\Big(-\beta\int_{t'}^t \epsilon(t'')\,dt''\Big)\Big]_{t'=0}^{t'=a} + \delta\cdot\Big[\exp\Big(-\beta\int_{t'}^t \epsilon(t'')\,dt''\Big)\Big]_{t'=a}^{t'=t} \\
&\le \frac{\epsilon_{max}}{\beta}\cdot\frac{2\beta\delta}{\epsilon_{max}} + \delta = 3\delta.
\end{aligned}$$

Since δ may be chosen arbitrarily small, (ii) must hold.