6. APPLICATION TO THE "TRAVELING SALESMAN PROBLEM"

The properties that have the most significant influence on the maps constructed by Kohonen's algorithm are the dimensionality of the neural network and the dimensionality and distribution of the input signals. The simplest case arises for a one-dimensional net, *i.e.*, a chain of neurons, and one-dimensional input signals. As shown in Chapter 5, one encounters this apparently quite unbiological case in the auditory cortex of mammals, where approximately linearly arranged neurons are assigned to a frequency interval. This situation is especially interesting also from a theoretical point of view, because it admits a closed solution, yielding the dependence of the resultant mapping on the probability density of the input signals. In this chapter, we extend our discussion to the case of multidimensional input signals, but we continue to assume a one-dimensional chain for the arrangement of the neurons. An analytical solution for the stationary maps which are possible under these circumstances can no longer be given. Instead, we will see that under appropriate conditions the resulting maps can be interpreted as approximate solutions of an interesting but analytically not tractable optimization problem, the "traveling salesman problem."

6.1 Paths as One-Dimensional Maps

In the case of a chain of neurons there exists a fixed order among the neurons given by their arrangement along the chain. Each neuron r carries a vector $\mathbf{w}_r$ marking a point in the space V of input signals. Hence, the corresponding "map" of V is one-dimensional. Whenever V is of higher dimension than the space of the lattice A, a substantial loss of information is inevitable, and the topology of V can only be reproduced to a very limited degree by a map $V \mapsto A$. Nevertheless, such maps may contain important and highly nontrivial information. We demonstrate this for the example of one-

dimensional maps.
If one runs through the neurons of the chain A, the points $\mathbf{w}_r$ run through a corresponding sequence of stations in the space V, which can be thought of as a *path*. This path is the image of the neuron chain under the mapping $r \mapsto \mathbf{w}_r$. From this point of view, Kohonen's algorithm for the formation of a one-dimensional map appears as a procedure for the stepwise optimization of a path in the space V (Angeniol et al. 1988). Initially, the path visits N randomly distributed stations. Each input signal $\mathbf{v} \in V$ chooses that station $\mathbf{w}_s$ of the path which is closest to $\mathbf{v}$ and deforms the path a bit toward $\mathbf{v}$ by shifting all stations corresponding to neurons in the neighborhood of s towards $\mathbf{v}$ as well. Thus, a path gradually develops, whose course favors regions from which input signals $\mathbf{v}$ are frequently chosen. Hence, by specification of an appropriate probability density $P(\mathbf{v})$, one can influence how important the presence of the path is in the individual regions of V. Since neurons neighboring on the chain become assigned to points $\mathbf{w}_r$ adjacent in the space V, the resulting path tends to be as short as possible. Hence, one-dimensional topology-conserving maps (approximately) solve an interesting optimization problem, that is, to find the shortest possible path, where $P(\mathbf{v})$ plays the role of a position-dependent "utility function" (cf. Angeniol et al. 1988).

6.2 The Model for a Discrete Stimulus Distribution

In this section we supplement the very qualitative remarks of the preceding section by a more precise, mathematical formulation. This is simplified if we assume a discrete probability distribution $P(\mathbf{v})$ instead of a continuous one. In this case, the input signal $\mathbf{v}$ can only take values from a discrete set $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_L\}$. Denoting by p_i the probability that $\mathbf{v}$ takes the value $\mathbf{q}_i$ ($\sum_i p_i = 1$), we see that $P(\mathbf{v})$ is of the form

$$P(\mathbf{v}) = \sum_{i=1}^{L} p_i \delta(\mathbf{v} - \mathbf{q}_i), \quad \mathbf{q}_i \in V, \tag{6.1}$$

where $\delta(.)$ denotes the "Dirac delta function" or "unit point measure" and represents a probability density concentrated entirely at the origin. In the context of the path optimization problem described above, the $\mathbf{q}_i$, $i = 1, 2, \ldots, L$, designate the location L of specified positions where the probability function is entirely concentrated, and through which the path is supposed

to pass. The p_i enable one to vary the relative importance of the positions. Taking the discrete probability density (6.1), we can drop the assumption of a one-dimensional neuron chain for the following derivation and temporarily admit an arbitrary topology of the neural network, without introducing additional complications. We now ask for the expectation value $E(\Delta \mathbf{w_r}|\mathbf{w}')$ for the change $\Delta \mathbf{w_r} := \mathbf{w_r} - \mathbf{w}'_\mathbf{r}$ of the synaptic strengths of neuron $\mathbf{r}$ under a single learning step. The notation $E(\Delta \mathbf{w_r}|\mathbf{w}')$ indicates that the expectation value is conditional, *i.e.*, it depends on the state $\mathbf{w}'$ of the neural network before the learning step. In analogy to Eq. (70), $E(\Delta \mathbf{w_r}|\mathbf{w}')$ is given by

$$\begin{aligned} E(\Delta \mathbf{w_r}|\mathbf{w}') &= \epsilon \int h_{\mathbf{r}\phi_{\mathbf{w}'}(\mathbf{v})}(\mathbf{v} - \mathbf{w}'_\mathbf{r}) P(\mathbf{v})\, d\mathbf{v} \\ &= \epsilon \sum_\mathbf{s} h_\mathbf{rs} \int\limits_{F_\mathbf{s}(\mathbf{w}')} (\mathbf{v} - \mathbf{w}'_\mathbf{r}) P(\mathbf{v})\, d\mathbf{v}. \end{aligned} \tag{6.2}$$

Here, $F_\mathbf{s}(\mathbf{w})$ is the set of all $\mathbf{v} \in V$ leading to the selection of "neuron" $\mathbf{s}$, *i.e.*,

$$F_\mathbf{s}(\mathbf{w}) = \Big\{ \mathbf{v} \in V \;\Big|\; \|\mathbf{v} - \mathbf{w_s}\| \leq \|\mathbf{v} - \mathbf{w_r}\| \;\forall \mathbf{r} \in A \Big\}. \tag{6.3}$$

Since we will encounter the set $F_\mathbf{s}(\mathbf{w})$ (called "indicator function" in probability theory) very often throughout the rest of this book, let us give a further explanation of (6.3): $F_\mathbf{s}(\mathbf{w})$ entails the sub-volume of the space V, whose center of gravity is given by $\mathbf{w_s}$, enclosing all points of V lying closer to $\mathbf{w_s}$ than to any other $\mathbf{w_r}$, $\mathbf{r} \neq \mathbf{s}$. With regard to the biological interpretation of Kohonen's model, $F_\mathbf{s}(\mathbf{w})$ thus plays the role of the set of all input patterns exciting the "neuron" $\mathbf{s}$ most strongly and, hence, can be interpreted as the "receptive field" of this neuron.
For the discrete probability distribution (6.1), expression (6.2) simplifies to

$$E(\Delta \mathbf{w_r}|\mathbf{w}) = \epsilon \sum_\mathbf{s} h_\mathbf{rs} \sum_{\mathbf{q}_i \in F_\mathbf{s}(\mathbf{w})} p_i(\mathbf{q}_i - \mathbf{w_r}). \tag{6.4}$$

The right-hand side (RHS) can be expressed as the gradient of a "potential function"

$$E(\Delta \mathbf{w_r}|\mathbf{w}) = -\epsilon \nabla_{\mathbf{w_r}} V(\mathbf{w})$$

where $V(\mathbf{w})$ is given by [1]

$$V(\mathbf{w}) = \frac{1}{2} \sum_\mathbf{rs} h_\mathbf{rs} \sum_{\mathbf{q}_i \in F_\mathbf{s}(\mathbf{w})} p_i(\mathbf{q}_i - \mathbf{w_r})^2. \tag{6.5}$$

[1] For a continuous probability density, a potential cannot be derived in this manner because of the dependence of (6.2) on the regions of integration, $F_\mathbf{r}(\mathbf{w})$.

According to (6.4), a single learning step *on the average* leads to a decrease

$$E(\Delta V|\mathbf{w}) = -\epsilon \sum_{\mathbf{r}} \|\nabla_{\mathbf{w_r}} V\|^2 \tag{6.6}$$

of $V(\mathbf{w})$. However, an *individual* learning step can also lead to an increase in $V(\mathbf{w})$. Hence, as in Monte-Carlo annealing (Kirkpatrick et al. 1983, Kirkpatrick 1984), for $\epsilon > 0$ there is some possibility of escaping from local minima. However, for this to happen, the RHS of (6.6) must be comparable to the depth of the minimum. Otherwise, escaping the minimum requires the joint action of several steps. But the change in the potential for k steps tends approximately to $k \cdot E(\Delta V|\mathbf{w})$, *i.e.*, to a strongly negative value. Therefore, the chance of leaving the minimum by the joint action of several steps is small. This indicates that the learning step size ϵ is qualitatively analogous to the temperature in Monte-Carlo annealing. In particular, in the limit $\epsilon \to 0$, a deterministic trajectory in the potential $V(\mathbf{w})$ results.

For small ϵ, the stationary states correspond to the stationary points of $V(\mathbf{w})$. If $N \geq L$, then $V(\mathbf{w})$ assumes particulary small values if one sets $\mathbf{w_r} \approx \mathbf{q}_{i(\mathbf{r})}$, where $i(\mathbf{r})$ is an assignment of lattice sites $\mathbf{r}$ to positions $\mathbf{q}_i$ with the property that lattice sites $\mathbf{r}$, $\mathbf{s}$ for which $h_{\mathbf{rs}}$ has large values, are assigned to positions $\mathbf{q}_{i(\mathbf{r})}$, $\mathbf{q}_{i(\mathbf{s})}$ that are as close as possible in V. The minimization of $V(\mathbf{w})$ can thus be viewed as the mathematical formalization of seeking a mapping from the positions $\mathbf{q}_i$ to the lattice A such that the neighborhood relations in the image on A (being defined by the function $h_{\mathbf{rs}}$: the larger $h_{\mathbf{rs}}$, the closer $\mathbf{r}, \mathbf{s}$) reproduce the corresponding neighborhood relations of the $\mathbf{q}_i \in V$ as faithfully as possible. The success of this minimization, and hence the "quality" of the obtained mapping, depends to a considerable degree on the form of the potential surface $V(\mathbf{w})$ and on the possible presence of local minima corresponding to "more poorly arranged" maps.

Now, $V(\mathbf{w})$ is differentiable for all configurations $\mathbf{w}$ in which none of the $\mathbf{q}_i$ happens to be on the boundary $\partial F_{\mathbf{s}}$ of one of the regions $F_{\mathbf{s}}(\mathbf{w})$, and in this case one has

$$\frac{\partial^2 V}{\partial w_{\mathbf{r}m} \partial w_{\mathbf{s}n}} = \delta_{\mathbf{rs}} \delta_{mn} h_{\mathbf{rs}} \sum_{\mathbf{q}_i \in F_{\mathbf{s}}(\mathbf{w})} p_i \geq 0. \tag{6.7}$$

At those values $\mathbf{w}$ for which one of the $\mathbf{q}_i$ lies on the border between two regions $F_{\mathbf{r}}$ and $F_{\mathbf{s}}$, one has $\|\mathbf{q}_i - \mathbf{w_r}\| = \|\mathbf{q}_i - \mathbf{w_s}\|$ and hence V is still continuous. However, the first derivative has a discontinuity at these positions, and the potential surface above the state space has a "cusp." Thus, in spite

of (6.7), V as a rule possesses numerous local minima of finite width. This situation is shown in Fig. 6.1, where the state space is represented schematically as a one-dimensional abscissa. For sufficiently small ϵ, the system can become trapped in any one of the "valleys" and converges in the limit $\epsilon \to 0$ to that state $\bar{\mathbf{w}}$ which corresponds to the local minimum of the "valley" that the system has chosen.

Abb. 6.1: Behavior of the potential $V(\mathbf{w})$ above the state space. This space is actually $N \cdot d$-dimensional, and its representation in the figure as a one-dimensional abscissa is only schematic.

The number of minima depends on the range of $h_{\mathbf{rs}}$. For an infinite range, *i.e.*, $h_{\mathbf{rs}} = h =$const., V becomes

$$V(\mathbf{w}) = \frac{h}{2} \sum_{i,\mathbf{r}} p_i (\mathbf{q}_i - \mathbf{w}_\mathbf{r})^2 \tag{6.8}$$

with a single minimum at $\mathbf{w}_\mathbf{r} = \sum_i p_i \mathbf{q}_i$. Decreasing the range, the cusps in V emerge, and with decreasing range of $h_{\mathbf{rs}}$ they become more prominent. In this way, additional local minima enter the picture. Finally, in the limit $h_{\mathbf{rs}} = \delta_{\mathbf{rs}}$, one has

$$V(\mathbf{w}) = \frac{1}{2} \sum_{\mathbf{q}_i \in F_\mathbf{r}(\mathbf{w})} p_i (\mathbf{q}_i - \mathbf{w}_\mathbf{r})^2. \tag{6.9}$$

For $N \geq L$, every configuration $\mathbf{w}_\mathbf{r} = \mathbf{q}_{i(\mathbf{r})}$ for which $i(\mathbf{r})$ is surjective is a local minimum of V. For instance, for $N = L$ this leads to $N!$ minima. For $N >> L$, one has about L^N such minima (aside from these, there are further minima in which some of the $\mathbf{w}_\mathbf{r}$ are averages of several of the $\mathbf{q}_i$). Hence, for short-range $h_{\mathbf{rs}}$, V possesses very many local minima, and the minimization of V generally represents an extremely difficult problem.

Nevertheless, one can obtain a close to minimal path in this case by beginning with a very long-range $h_{\mathbf{rs}}$, for which V has only a single minimum. If the $h_{\mathbf{rs}}$ are slowly adjusted toward their desired final values, additional local minima successively emerge. For a sufficiently slow change, the system will fall into those new minima which are created in the current valley. But these are just the most promising candidates for an especially low final value. We can thus appreciate the importance of a slow decrease of the range of $h_{\mathbf{rs}}$ for the construction of a good map.

6.3 Application to the "Traveling Salesman Problem"

The occurrence of numerous local minima is a frequent characteristic of difficult optimization problems that belong to the class of so-called NP-complete problems and is one of the causes for the difficulty of finding their solution (although there are also NP-complete problems without local minima; see for example Baum 1986). For a problem to be efficiently tractable, there must exist a deterministic algorithm that generates a solution with a computational effort that rises no faster asymptotically than polynomially with the size of the problem. The set of all problems with this property forms the class P of so-called deterministic Polynomial problems. The class NP of Non-deterministic Polynomial problems arises if one weakens this requirement and just demands that the *correctness* of a solution is verifiable with a computational effort growing at most as some polynomial with the size of the problem. Evidently $P \subset NP$, but it is to be expected that NP contains in addition problems that are considerably more "difficult" than those in P, since every problem in NP not contained in P must require a computational effort for finding a solution which by definition grows faster asymptotically than any power of the problem size (Garey and Johnson 1979). A subclass of NP which is not contained in P is the class of so-called NP-*complete* problems. NP-complete problems can be characterized as being at least as hard as any other NP problem and not being solvable deterministically in polynomial time. Today, many NP-complete problems are known, however, it is not possible in any case to decide whether a deterministic solution procedure may be discovered someday that would reduce the computational effort to within polynomial bounds. (It has not been proven that $NP \neq P$, *i.e.*, *every* NP-complete problem might be reducible to a "merely" P problem,

although at present hardly anyone believes this).

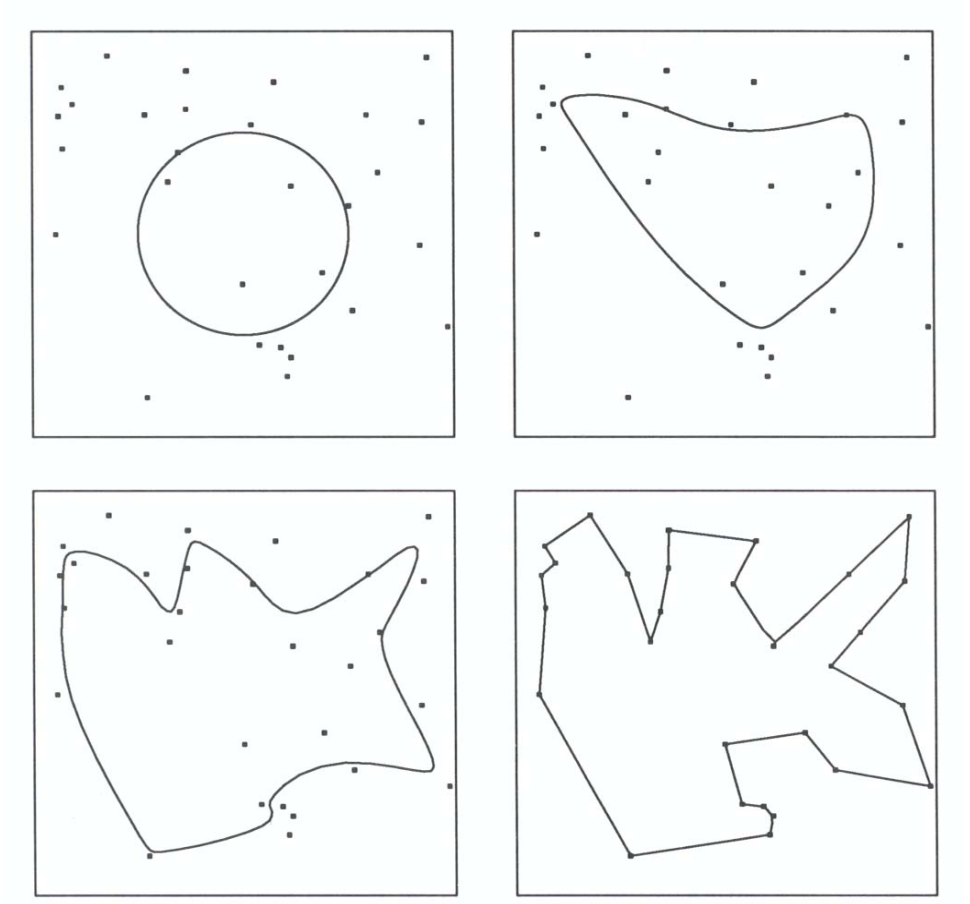

Abb. 6.2: Simulation of the Markov-Process (70) for the TSP problem with $L = 30$ cities chosen at random in the unit square. Top left to bottom right: Initially chosen polygon tour, polygon tour obtained after 5,000, 7,000 and 10,000 learning steps, respectively. Simulation parameters: $N = 100$, $\epsilon = 0.8$, $\sigma(0) = 50$, $\sigma(10,000) = 1$.

The best-known example of an NP-complete problem, for which the computational effort rises exponentially with the problem size for every algorithm known up to now, is the "Traveling Salesman Problem" (TSP). In this problem, one seeks the shortest possible tour passing through N given cities. By

testing all $\frac{1}{2}(N-1)!$ possible tours, one can always find the shortest tour, but the computational effort for this "direct" strategy, called "exhaustive search," rises exponentially with N and rapidly becomes unmanageable (for $N = 30$ the required processing time, even using a Cray–XMP supercomputer, would exceed the age of the universe.) The exponential character of this growth behavior persists for all improved algorithms discovered so far, although one has been successful at postponing the increase to considerably large values N. The root of this difficulty lies in the extremely irregular structure of the function "path length" over the state space of the problem. In particular, this function possesses numerous local minima, very many of which lie only very little above the global minimum. In order to find at least good approximations to the global minimum for such functions, several methods have been developed (Lin and Kerninghan 1973; Kirkpatrick et al. 1983). They are mostly based on a stochastic sampling of the state space in the direction of decreasing path lengths, together with some provision to escape from unfavorable local minima.

The usefulness of models of the formation of neural projections for treating the traveling salesman problem was first recognized by Durbin and Willshaw (1987). In the following, we demonstrate in a computer simulation how an approximate solution can be obtained by means of Kohonen's model (see also Angeniol et al. 1988). To this end, we choose a closed chain of "neurons" in the form of a ring. The vectors $\mathbf{w}_r$ of the neurons are changed iteratively according to equation (70), where in each step an element of the set $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_L\}$ of position vectors $\mathbf{q}_i$ of the L cities is selected as the input vector $\mathbf{v}$. For each $\mathbf{q}_i$, the same selection probability $p_i = 1/L$ is chosen. The Gaussian (68) was chosen for $h_{\mathbf{rs}}$, and the remaining simulation data were $N = 800$, $\epsilon = 0.8$, $\sigma(t) = 50 \cdot 0.02^{t/t_{max}}$ and $t_{max} = 10,000$ Markov steps. For the simulation example, $L = 30$ cities, randomly located in a unit square, were given. The initial values of the N vectors $\mathbf{w}_r$ were assigned to the corners of a regular 30-sided polygon. This results in the initial configuration shown in the upper left part of Fig. 6.2. Each iteration causes a local deformation of this path. Initially, as long as $h_{\mathbf{rs}}$ is still long-range, each deformation affects rather large path segments. In this way, first the rough outline of the eventual path is formed (Fig. 6.2, upper right, 5000 iterations). As the range of $h_{\mathbf{rs}}$ gradually decreases, the deformations along the chain become more localized and finer details of the path emerge (Fig. 6.2, lower left, 7000 iterations). Towards the end of the simulation $h_{\mathbf{rs}}$ differs significantly from zero only for immediate chain neighbors r, s. In this phase, the path takes

on its final shape, passing through all of the given cities (Fig. 6.2, lower right, 10000 iterations). The path found after 10,000 steps has length 4.5888 and, in this example, happened to be the optimal solution.[2] However, this is not guaranteed for every case. Depending on the initial conditions, a slightly longer path may result, especially if the number of cities becomes larger.
We have seen in the previous chapters how even one-dimensional maps make possible interesting applications. In the following chapters, we will extend the discussion to two-dimensional maps. In Chapter 7, we will use them to model the formation of a "somatotopic map" of the palm of the hand. An extension of the algorithm to the task of *learning of output values* will then open up applications to *control problems* and thus introduce the subject of the Part III of this book.

[2] Only a "naive" comparison of all possible paths would require a computational time which exceeds the age of the universe. In fact, there are clever search techiques which reduce the computational effort significantly. With those sophisticated search techniques it has even been possible to find the shortest path through a nontrivial distribution of 2430 points, the current "world record" (1990).