Phase transitions in stochastic self-organizing maps

Thore Graepel, Matthias Burger, and Klaus Obermayer
Fachbereich Informatik, FR 2-1, Technische Universität Berlin, Franklinstraße 28/29, 10587 Berlin, Germany
(Received 25 April 1997)

We describe the development of neighborhood-preserving stochastic maps in terms of a probabilistic clustering problem. Starting from a cost function for central clustering that incorporates distortions from channel noise, we derive a soft topographic vector quantization algorithm (STVQ) which is based on the maximum entropy principle, and which maximizes the corresponding likelihood in an expectation-maximization fashion. Among other algorithms, a probabilistic version of Kohonen's self-organizing map (SOM) is derived from STVQ as a computationally efficient approximation of the E step. The foundation of STVQ in statistical physics motivates a deterministic annealing scheme in the temperature parameter β, and leads to a robust minimization algorithm of the clustering cost function. In particular, this scheme offers an alternative to the common stepwise shrinking of the neighborhood width in the SOM, and makes it possible to use its neighborhood function solely to encode the desired neighborhood relations between the clusters. The annealing in β, which corresponds to a stepwise refinement of the resolution of representation in data space, leads to the splitting of an existing cluster representation during the ''cooling'' process. We describe this phase transition in terms of the covariance matrix **C** of the data and the transition matrix **H** of the channel noise, and calculate the critical temperatures and modes as functions of the eigenvalues and eigenvectors of **C** and **H**. The analysis is extended to the phenomenon of the automatic selection of feature dimensions in dimension-reducing maps, thus leading to a ''batch'' alternative to the Fokker-Planck formalism for on-line learning. The results provide insights into the relation between the width of the neighborhood and the temperature parameter β: It is shown that the phase transition which leads to the representation of the excess dimensions can be triggered not only by a change in the statistics of the input data but also by an increase of β, which corresponds to a decrease in noise level. The theoretical results are validated by numerical methods. In particular, a quantity equivalent to the heat capacity in thermodynamics is introduced to visualize the properties of the annealing process.
[S1063-651X(97)01110-0]

PACS number(s): 64.60.−i, 07.05.Mh, 89.70.+c

I. INTRODUCTION

The tractability of pattern recognition and signal processing tasks depends strongly on the representation of the relevant input data. Usually, the input signals are high dimensional vectors which are hard to visualize and which—for reasons of complexity—cannot be processed directly. Therefore it is desirable to find some mapping of the high dimensional input space to some lower dimensional space in a way which captures the essential spatial relations of the data as faithfully as possible, and which at the same time performs a kind of lossy data compression. Algorithms of this kind are generally known as ''topology preserving vector quantizers'' [1,2].

The self-organizing map (SOM), first introduced by Kohonen [3,4], is an example of such an algorithm. The mapping is achieved by a heuristic on-line learning rule that leads to a correspondence between local regions in input space and neurons in a usually two-dimensional array, such that the spatial relations between data points are reflected by the spatial relations of the corresponding neurons in the array. The SOM has been applied to a wide range of technical tasks (see Refs. [5,6] for a review), and has become one of the standard modeling approaches for neural development in the computational neuroscience community (see Refs. [7,8] for a review), for which it was originally intended. Also, there exists a great amount of literature that deals with different theoretical aspects and applications of the SOM [9,5].

Another approach to the problem of lossy data compression is called clustering or vector quantization. The idea is to encode a set of data points by a reduced set of reference vectors in such a way as to minimize a given cost function based on a suitable distortion measure. The easiest and most well-known paradigm is k-means clustering [10], which uses an on-line learning rule and applies the squared Euclidean distance as a distortion measure to update its reference vectors. Recently, more elaborate schemes have been suggested, which take into account the complexity of the codebook or the robustness of the representation [1].

Rose, Gurewitz, and Fox [11] introduced deterministic annealing as a robust minimization procedure for the clustering cost function leading to a set of optimal reference vectors. Deterministic annealing was originally derived from statistical physics (cf. Refs. [12,13]) and is in this context based on fuzzy assignments of data points to clusters. The annealing process helps to avoid local minima in the possibly highly nonconvex cost function during the optimization procedure. After deriving a Gibbs distribution related to the cost function via the principle of maximum entropy, the unique maximum of the likelihood at high temperatures is determined and tracked through lower temperatures. Depending on the structure of the cost function, this procedure leads to good local minima or even to the global minimum of the cost function.

Luttrell [14–16] established a connection between the self-organizing map and noisy vector quantization. By

choosing a distortion measure for vector quantization that incorporates robustness with respect to noise-induced changes of assignments, he derived an algorithm which he named the topographic vector quantization (TVQ). He showed that the SOM can be viewed as an efficient approximation to a gradient descent on the TVQ cost function. Since the TVQ has a known cost function it is thus possible to find efficient optimization procedures (see, e.g., Ref. [2]). Those procedures can then—via the approximation—be applied to develop robust variants of the SOM. Additionally, the analysis of stationary states and convergence properties of the SOM [17,18] is facilitated by considering the link to the TVQ [19].

In this paper we apply the idea of deterministic annealing to the optimization of the TVQ cost function, and develop an algorithm for noisy vector quantization which we call soft topographic vector quantizer (STVQ). The STVQ can be used for the creation of topology-preserving maps by appropriately choosing the transition probabilities of the assumed channel noise, because the channel noise breaks the permutation symmetry of the clusters and thus provides a distance measure on the space of clusters similar to the neighborhood matrix in the SOM. The probabilistic formulation enables us to apply an annealing scheme in the temperature instead of in the range of the neighborhood function, which can thus be chosen freely to represent desired neighborhood relations of the clusters (e.g., random graphs in Ref. [20]). From an optimization point of view, the annealing process is viewed as a means to avoid local minima of the clustering cost function.

Our analysis also shows that the annealing leads to the splitting of existing cluster representations in data space. This process is identified as a phase transition, and is characterized in relation to the channel noise and the input data.

In Sec. II we derive a set of self-consistent equations for the cluster centers based on fuzzy assignments of data points to clusters using the principle of maximum entropy. These fixed-point equations are solved by an expectation-maximization (EM)-type algorithm [21] at a given temperature. In order to avoid local minima of the cost function we then employ an annealing procedure in the temperature parameter. Via an approximation in the E step of the STVQ, this leads to a deterministic annealing procedure for the SOM, as well. In Sec. III we analyze phase transitions that occur during the annealing process in the temperature. We calculate the critical temperatures and modes for the splitting of existing clusters in terms of eigenvalues and eigenvectors of the covariance matrix of the data and the transition matrix. The same technique is then applied in Sec. IV to the phenomenon of the automatic selection of feature dimensions, which was first analyzed for the on-line SOM by Ritter and Schulten [17] using a Fokker-Planck approach and later applied by Obermayer, Blasdel, and Schulten to pattern formation in neural systems [18]. The technique yields expressions for critical variance of the data and critical wavelength of the unstable mode in terms of temperature and transition matrix. Results are compared to the zero temperature case for the SOM, which had been obtained earlier [17]. In Sec. V numerical results are presented that demonstrate the behavior of the algorithm and confirm the theoretical results of the previous sections. We numerically explore the transitions which the cluster representation undergoes during the annealing, and introduce a quantity similar to the heat capacity in thermodynamics to better visualize the annealing process.

II. SOFT TOPOGRAPHIC VECTOR QUANTIZATION

A. Derivation of the STVQ algorithm

In clustering data points which are in some sense similar are grouped for the purpose of data interpretation as well as data compression. Given a set $\mathcal{X}$ of data points $\mathbf{x}_i \in \Re^d$, $i=1,\ldots,D$, and a set $\mathcal{C}$ of clusters $\mathcal{C}_\mathbf{r}$, $\mathbf{r}=1,\ldots,N$, the aim of any clustering algorithm is to assign each data point $\mathbf{x}_i$ to a cluster $\mathcal{C}_\mathbf{r}$ so as to minimize a given assignment cost function E. If we introduce binary assignment variables $m_{i\mathbf{r}}$, which take the value one if data point $\mathbf{x}_i$ is member of cluster $\mathcal{C}_\mathbf{r}$ and zero otherwise, the cost function can be written as

$$E(\{m_{i\mathbf{r}}\},\ \text{parameters}) = \sum_i \sum_\mathbf{r} m_{i\mathbf{r}} E_\mathbf{r}(i,\text{parameters}), \tag{1}$$

where $E_\mathbf{r}(i,$ parameters$)$ denotes the cost associated with assigning data point $\mathbf{x}_i$ to cluster $\mathcal{C}_\mathbf{r}$, and ''parameters'' parametrize the assignment costs $E_\mathbf{r}$. In central clustering $E_\mathbf{r}(i,$ parameters$)$ is taken to be the squared Euclidean distance $E_\mathbf{r}(i,\mathbf{w}_\mathbf{r}) = \|\mathbf{x}_i - \mathbf{w}_\mathbf{r}\|^2$ between a data point $\mathbf{x}_i$ and a parameter vector $\mathbf{w}_\mathbf{r} \in \Re^d$, which for central clustering is called cluster center, and which serves as the representative in data space for the data points assigned to the cluster $\mathcal{C}_\mathbf{r}$. The desired property of the assignment, that each data point is assigned to exactly one cluster, requires the constraints

$$\sum_\mathbf{r} m_{i\mathbf{r}} = 1, \quad \forall i. \tag{2}$$

The quantity $E(\{m_{i\mathbf{r}}\},$parameters$)$ takes its minimum with respect to the parameters when an optimal set of locations for the cluster centers, i.e., an optimal representation for each group of data points in data space, is achieved.

Following an idea by Luttrell [16] we consider the case that the cluster indices $\mathbf{r}$, which label the clusters, form a compressed encoding of the data for the purpose of transmission via a noisy channel (see Fig. 1). The distortion caused by the channel noise is modeled by a matrix $\mathbf{H}$ of transition probabilities $h_{\mathbf{rs}}$ for the noise induced change of the assignment of a data point $\mathbf{x}_i$ from cluster $\mathcal{C}_\mathbf{r}$ to cluster $\mathcal{C}_\mathbf{s}$. After transmission the received index $\mathbf{s}$ is decoded, i.e., mapped back to data space, using its cluster center $\mathbf{w}_\mathbf{s}$. Averaging the squared Euclidean distance $\|\mathbf{x}_i - \mathbf{w}_\mathbf{s}\|^2$ over all possible transitions thus yields what Luttrell calls the topographic Euclidean distortion

$$E_\mathbf{r}(i,\{\mathbf{w}_\mathbf{r}\}) = \tfrac{1}{2} \sum_\mathbf{s} h_{\mathbf{rs}} \|\mathbf{x}_i - \mathbf{w}_\mathbf{s}\|^2, \tag{3}$$

where the factor 1/2 is introduced for computational convenience. Since $h_{\mathbf{rs}}$ is the probability for the transition $\mathbf{r} \to \mathbf{s}$, the following constraint must hold:

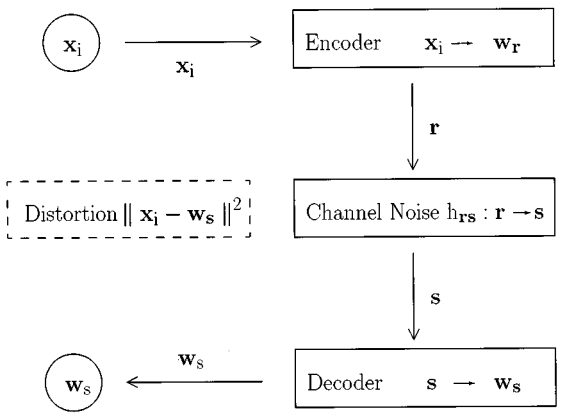

FIG. 1. Cartoon of a generic data communication problem. Input data $\mathbf{x}_i$ are grouped and the groups (clusters) are labeled with indices $\mathbf{r}$ (encoding stage). The indices are then transmitted via a noisy channel which is characterized by a set of transition probabilities $h_{\mathbf{rs}}$ for the noise process. As soon as an index $\mathbf{s}$ is received at the decoder the data is reconstructed via a vector $\mathbf{w}_\mathbf{s}$ (decoding stage) which represents all data points assigned to cluster $\mathbf{s}$ during encoding. In the following, we will measure the combined error due to clustering and channel noise via the squared Euclidean distance between the original data point $\mathbf{x}_i$ and the cluster center $\mathbf{w}_\mathbf{s}$. The final assignment cost is then given by an average over all transitions $\mathbf{r} \rightarrow \mathbf{s}$.

$$\sum_{\mathbf{s}} h_{\mathbf{rs}} = 1, \quad \forall \mathbf{r}. \tag{4}$$

The transition probabilities are closely related to the elements of the neighborhood matrix in the SOM [3,4]. The cost function (1) with distortion measure (3) takes its minimum, when a robust representation of the data with respect to the channel noise is achieved. Since the assignment of a data point $\mathbf{x}_i$ to cluster $\mathcal{C}_\mathbf{r}$ changes to cluster $\mathcal{C}_\mathbf{s}$ with probability $h_{\mathbf{rs}}$ the corresponding representatives or cluster centers $\mathbf{w}_\mathbf{r}$ and $\mathbf{w}_\mathbf{s}$ should be located close to each other in data space if $h_{\mathbf{rs}}$ is large in order to keep the assignment cost (1) low. In this way the noise-induced transitions lead—via Eq. (3)—to a coupling between different clusters. The transition probability can be interpreted as a measure for ''closeness'' between clusters: Clusters are ''close'' if the transition probabilities are high. In the special case that the transition probabilities are monotonically related to a metric they define a neighborhood in the sense of the SOM.

Now, given the cost function $E = E(\{m_i\hat{\mathbf{r}}\}, \{\mathbf{w}_\mathbf{r}\})$ as a quality criterion for the representation $\{\{m_{i\mathbf{r}}\}, \{\mathbf{w}_\mathbf{r}\}\}$ of the data, we determine a probability distribution $P = P(\{m_{i\mathbf{r}}\}, \{\mathbf{w}_\mathbf{r}\})$ over the space of all representations in the spirit of Bayesian model evaluation. In order to simplify notation here and in the following, bounds on sums and integrals are omitted if sums over i run over all D data points in $\mathcal{X}$, sums over $\mathbf{r}$ run over all N clusters in $\mathcal{C}$, and integrals are taken from $-\infty$ to ∞. Integrals over vectors are to be read as multiple integrals over the vectors' components. Since we do not make any assumptions about the distribution of data points we apply the principle of maximum entropy [22]. This amounts to choosing the probability distribution which maximizes the entropy,

$$S = \sum_{\{m_{i\mathbf{r}}\}} \int \cdots \int P \ln P \, d\mathbf{w}_1 \cdots d\mathbf{w}_N, \tag{5}$$

under the constraint of a given average cost

$$U = \sum_{\{m_{i\mathbf{r}}\}} \int \cdots \int EP \, d\mathbf{w}_1 \cdots d\mathbf{w}_N \tag{6}$$

and yields the Gibbs distribution as the probability distribution over the space of representations,

$$P(\{m_{i\mathbf{r}}\}, \{\mathbf{w}_\mathbf{r}\}) = \frac{1}{Z} \exp(-\beta \, E(\{m_{i\mathbf{r}}\}, \{\mathbf{w}_\mathbf{r}\})). \tag{7}$$

The Lagrange multiplier β is associated with the average cost U, and is interpreted as an inverse temperature. Z is the normalization constant or partition function and is given by

$$Z = \sum_{\{m_{i\mathbf{r}}\}} \int \cdots \int \exp(-\beta \, E(\{m_{i\mathbf{r}}\}, \{\mathbf{w}_\mathbf{r}\})) d\mathbf{w}_1 \cdots d\mathbf{w}_N. \tag{8}$$

Since we are primarily interested in determining the most probable set of cluster centers so as to generalize from a given set of training samples, the marginal probability

$$P(\{\mathbf{w}_\mathbf{r}\}) = \frac{1}{Z} \sum_{\{m_{i\mathbf{r}}\}} \exp(-\beta \, E(\{m_{i\mathbf{r}}\}, \{\mathbf{w}_\mathbf{r}\})) \tag{9}$$

is considered, where the summation runs over all sets $\{m_{i\mathbf{r}}\}$ of assignments which obey relation (2). Using the identity

$$\sum_{\{m_{i\mathbf{r}}\}} \exp(-\beta \, E(\{m_{i\mathbf{r}}\}, \{\mathbf{w}_\mathbf{r}\})) = \prod_i \sum_{\mathbf{r}} \exp(-\beta E_\mathbf{r}(i, \{\mathbf{w}_\mathbf{r}\})), \tag{10}$$

one obtains, for the log likelihood,

$$\ln P(\{\mathbf{w}_\mathbf{r}\}) = \sum_i \ln \sum_{\mathbf{r}} \exp(-\beta \, E_\mathbf{r}(i, \{\mathbf{w}_\mathbf{r}\})) - \ln Z. \tag{11}$$

Maximizing Eq. (11) with respect to $\{\mathbf{w}_\mathbf{r}\}$ at a given value of the temperature parameter β yields conditions

$$\mathbf{w}_\mathbf{r} = \frac{\sum_i \mathbf{x}_i \sum_{\mathbf{s}} h_{\mathbf{rs}} P(i \in \mathcal{C}_\mathbf{s})}{\sum_i \sum_{\mathbf{s}} h_{\mathbf{rs}} P(i \in \mathcal{C}_\mathbf{s})}, \quad \forall \mathbf{r}, \tag{12}$$

for the cluster centers $\{\mathbf{w}_\mathbf{r}\}$, where $P(i \in \mathcal{C}_\mathbf{s})$ is the assignment probability of data point $\mathbf{x}_i$ to cluster $\mathcal{C}_\mathbf{s}$ and is given by

$$P(i \in \mathcal{C}_\mathbf{s}) = \langle m_{i\mathbf{s}} \rangle = \frac{\exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}} h_{\mathbf{st}} \| \mathbf{x}_i - \mathbf{w}_\mathbf{t} \|^2 \right)}{\sum_{\mathbf{u}} \exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}} h_{\mathbf{ut}} \| \mathbf{x}_i - \mathbf{w}_\mathbf{t} \|^2 \right)}. \tag{13}$$

$\langle m_{i\mathbf{s}} \rangle$ is the expectation value of the binary assignment variable $m_{i\mathbf{s}}$ for a given set $\{\mathbf{w}_\mathbf{r}\}$ with respect to the probability

distribution (7). For a given β, Eqs. (12) and (13) can be solved by fixed-point iteration. We will call this optimization algorithm the STVQ.

Equations (12) and (13) can also be derived from a statistical physics framework (cf. Yuille *et al.* [12,13] for an expanded treatment). Starting from the Hamiltonian given in (1) and (3), we first consider the probability distribution over the $\{m_{i\mathbf{r}}\}$ for finite temperature and fixed $\{\mathbf{w_r}\}$. This yields

$$P_m(\{m_{i\mathbf{r}}\}|\{\mathbf{w_r}\}) = \frac{1}{Z_m}\exp(-\beta E(\{m_{i\mathbf{r}}\}|\{\mathbf{w_r}\})), \quad (14)$$

with the $\{\mathbf{w_r}\}$-dependent partition function

$$Z_m(\{\mathbf{w_r}\}) = \sum_{\{m_{i\mathbf{r}}\}} \exp(-\beta E(\{m_{i\mathbf{r}}\}|\{\mathbf{w_r}\})). \quad (15)$$

Using Eq. (10), the partition function Z_m can be evaluated exactly, which yields an expression for the free energy

$$\mathcal{F}_m = \frac{1}{\beta}\ln Z_m . \quad (16)$$

In statistical physics one is interested in expectation values rather than maximum likelihood estimates. The expectation value of $\mathbf{w_s}$ can be expressed as

$$\langle \mathbf{w_s}\rangle = \frac{1}{Z}\int \cdots \int \mathbf{w_s}\exp(-\beta\, \mathcal{F}_m(\{\mathbf{w_r}\}))d\mathbf{w}_1\cdots d\mathbf{w}_N , \quad (17)$$

where the degeneracy of solutions due to symmetries of the data space and the transition probabilities have to be taken into account by integrating only over one fundamental cell. Expression (17) can be evaluated using the saddle-point approximation, i.e., by expanding $\mathcal{F}_m(\{\mathbf{w_r}\})$ to second order around its minimum, which yields Eqs. (12) and (13) for the saddle-point conditions. The assignment probabilities $P(i\in\mathcal{C}_\mathbf{s})$ are equal to the mean fields $\langle m_{i\mathbf{s}}\rangle$ of the binary assignment variables for a given set of cluster centers $\{\mathbf{w_r}\}$.

From an algorithmic point of view, iteratively solving Eqs. (12) and (13) comprises an expectation-maximization algorithm [21]. The E step (13) consists of the calculation of the assignment probabilities $P(i\in\mathcal{C}_\mathbf{s})$ for all data points $\mathbf{x}_i$ and clusters $\mathcal{C}_\mathbf{s}$. Then in the M step (12) of the algorithm the positions of the cluster centers $\mathbf{w_r}$ are recalculated using the new assignment probabilities $P(i\in\mathcal{C}_\mathbf{s})$ from the E step. In Ref. [21] it is shown that the EM algorithm converges monotonically to a local maximum of the log likelihood (11) under mild conditions that are valid for our case. However, we are interested in finding the global minimum of E given by Eq. (1). Since the global minimum of E coincides with the global maximum of the log likelihood for $\beta\to\infty$, we can apply a deterministic annealing scheme in β. At low β, the local minima of E are washed out in the log likelihood, whose global maximum can then be found using the EM algorithm. The maximum is then tracked through higher values of β until it coincides at sufficiently high β with a minimum of E. Convergence to a (one-change optimal) local minimum was established by Puzicha, Hofmann, and Buhmann [23], who also pointed out that convergence to the global minimum should not be expected in the general case. Geman and Geman [24] gave an annealing schedule for simulated annealing according to which $\beta(t)\leqslant c\,\ln(1+t)$, where t is the number of the annealing step and c is a constant independent of t, and proved the convergence to a global minimum in distribution. This result hints at how the parameter β is to be handled in deterministic annealing. In practical applications, however, a linear or exponential annealing scheme for β could be allowed to save computation time, possibly at the cost of precision of the results. In analogy to Gaussian mixture models the parameter β can also be interpreted as an inverse variance in data space, thus determining the resolution of the clustering. Consequently, the annealing process corresponds to a stepwise refinement of the representation of the data, and it is possible to determine the resolution of the final representation by terminating the annealing schedule at an appropriate value of β. This is particularly appropriate to avoid an overfitting of the data in the presence of noise.

B. Derivatives of the STVQ algorithm

To put the above-derived algorithm (STVQ) into a familiar context we consider certain limits and approximations which lead to a family of topographic clustering algorithms. The limiting case $\beta\to\infty$ in the assignment probabilities (13) yields a batch version of the TVQ discussed by Luttrell [16] and Heskes and Kappen [19]. The TVQ is a winner-take-all algorithm for which Eqs. (12) and (13) become

$$\mathbf{w_r} = \frac{\sum_i \mathbf{x}_i \sum_\mathbf{s} h_{\mathbf{rs}} P_{\mathrm{TVQ}}(i\in\mathcal{C}_\mathbf{s})}{\sum_i \sum_\mathbf{s} h_{\mathbf{rs}} P_{\mathrm{TVQ}}(i\in\mathcal{C}_\mathbf{s})} \quad (18)$$

and

$$P_{\mathrm{TVQ}}(i\in\mathcal{C}_\mathbf{s}) = \delta_{\mathbf{st}},$$

$$\mathbf{t} = \arg\min_\mathbf{u} \sum_\mathbf{v} h_{\mathbf{uv}}\|\mathbf{x}_i - \mathbf{w_v}\|^2. \quad (19)$$

The approximation $h_{\mathbf{rs}}\to\delta_{\mathbf{rs}}$ in the assignment probabilities (13) leads to a fuzzy version of the SOM which we call soft-SOM (SSOM). This modification provides an important computational simplification because the omission of one convolution with $h_{\mathbf{rs}}$ saves a considerable amount of computation time. Equations (12) and (13) then become

$$\mathbf{w_r} = \frac{\sum_i \mathbf{x}_i \sum_\mathbf{s} h_{\mathbf{rs}} P_{\mathrm{SSOM}}(i\in\mathcal{C}_\mathbf{s})}{\sum_i \sum_\mathbf{s} h_{\mathbf{rs}} P_{\mathrm{SSOM}}(i\in\mathcal{C}_\mathbf{s})} \quad (20)$$

and

$$P_{\mathrm{SSOM}}(i\in\mathcal{C}_\mathbf{s}) = \frac{\exp\left(-\frac{\beta}{2}\|\mathbf{x}_i - \mathbf{w_s}\|^2\right)}{\sum_\mathbf{t}\exp\left(-\frac{\beta}{2}\|\mathbf{x}_i - \mathbf{w_t}\|^2\right)}. \quad (21)$$

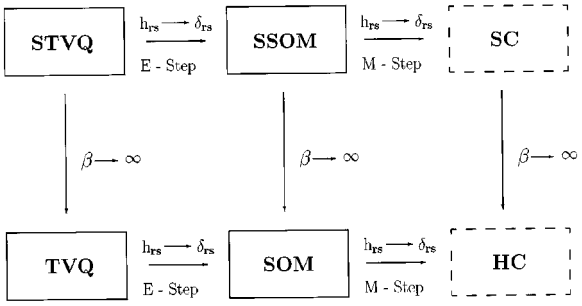

FIG. 2. The STVQ family of clustering algorithms.

It has been noted by Luttrell [16], however, that Eqs. (20) and (21), which correspond to a nearest-neighbor encoding, do not in general minimize the cost function (1) with Eq. (3) [14]. An exact minimization is only achieved, when the channel noise is taken into account not only in the update rule but also in the determination of the winner as in Eqs. (12) and (13) for the STVQ.

If one combines the limiting case $\beta\to\infty$ with the approximation $h_{\mathbf{rs}}\to\delta_{\mathbf{rs}}$ in Eq. (13), one obtains a batch version of the SOM [25], for which Kohonen's original algorithm [3,4] is a stochastic approximation [26]. Equations (12) and (13) then become

$$\mathbf{w_r}=\frac{\sum_i \mathbf{x}_i \sum_{\mathbf{s}} h_{\mathbf{rs}} P_{\mathrm{SOM}}(i\in\mathcal{C}_{\mathbf{s}})}{\sum_i \sum_{\mathbf{s}} h_{\mathbf{rs}} P_{\mathrm{SOM}}(i\in\mathcal{C}_{\mathbf{s}})} \tag{22}$$

and

$$P_{\mathrm{SOM}}(i\in\mathcal{C}_{\mathbf{s}})=\delta_{\mathbf{st}},$$

$$\mathbf{t}=\arg\min_{\mathbf{u}}\|\mathbf{x}_i-\mathbf{w_u}\|^2. \tag{23}$$

Finally, substituting $h_{\mathbf{rs}}\to\delta_{\mathbf{rs}}$ in both Eqs. (12) and (13) yields the soft clustering procedure proposed in Ref. [11], whose limit $\beta\to\infty$ recovers the well-known k-means clustering (HC) [10]. Figure 2 summarizes the family of topographic clustering algorithms.

III. ANALYSIS OF THE INITIAL PHASE TRANSITION

In order to understand the annealing process in the temperature parameter β it is instructive to look at how the representation of the data changes with β. From Refs. [11] and [1] it is known that the cluster centers split with increasing β, and that the number of relevant clusters for a resolution given by β is determined from the number of clusters that have split up to that point. In the STVQ, however, the permutation symmetry of the cluster centers is broken and couplings between clusters are introduced by the transition matrix **H**. This changes stationary states and the ''splitting'' behavior of the cluster centers.

For $\beta=0$, which corresponds to infinite temperature, every data point $\mathbf{x}_i$ is assigned to every cluster $\mathcal{C}_{\mathbf{r}}$ with equal probability $P^0(i\in\mathcal{C}_{\mathbf{r}})=1/N$, where N is the number of cluster centers. In this case the cluster centers are given by

$$\mathbf{w}_{\mathbf{r}}^0=\frac{1}{D}\sum_i \mathbf{x}_i, \quad \forall \mathbf{r}; \tag{24}$$

that is, all the cluster centers are located at the center of mass of the data. Without loss of generality we set $\mathbf{w}_{\mathbf{r}}^0=\mathbf{0}$, $\forall\ \mathbf{r}$. A Taylor expansion of the right-hand side of Eq. (12) around $\{\mathbf{w}_{\mathbf{r}}^0\}$, to first order in $\mathbf{w_t}$, yields

$$\mathbf{w_r}=\left[\frac{\sum_i \mathbf{x}_i \sum_{\mathbf{s}} h_{\mathbf{rs}} P(i\in\mathcal{C}_{\mathbf{s}})}{\sum_i \sum_{\mathbf{s}} h_{\mathbf{rs}} P(i\in\mathcal{C}_{\mathbf{s}})}\right]_{\{\mathbf{w}_{\mathbf{r}}^0\}} + \sum_{\mathbf{t}}\left[\frac{\partial}{\partial \mathbf{w_t}}\frac{\sum_i \mathbf{x}_i \sum_{\mathbf{s}} h_{\mathbf{rs}} P(i\in\mathcal{C}_{\mathbf{s}})}{\sum_i \sum_{\mathbf{s}} h_{\mathbf{rs}} P(i\in\mathcal{C}_{\mathbf{s}})}\right]_{\{\mathbf{w}_{\mathbf{r}}^0\}} \mathbf{w_t}+O(\mathbf{w_t}^2). \tag{25}$$

Under the assumption that **H** is symmetrical, i.e., $h_{\mathbf{rs}}=h_{\mathbf{sr}}$, $\forall \mathbf{r},\mathbf{s}$, this expression can be evaluated using the relation

$$\frac{\partial P(i\in\mathcal{C}_{\mathbf{s}})}{\partial \mathbf{w_t}}=\beta\,(\mathbf{x}_i-\mathbf{w_t})P(i\in\mathcal{C}_{\mathbf{s}})\left(h_{\mathbf{st}}-\sum_{\mathbf{u}} h_{\mathbf{tu}} P(i\in\mathcal{C}_{\mathbf{u}})\right), \tag{26}$$

and the linearized fixed-point equations become

$$\mathbf{w_r}=\beta\,\mathbf{C}\sum_{\mathbf{t}} g_{\mathbf{rt}}\mathbf{w_t}. \tag{27}$$

Here $\mathbf{C}=(1/D)\,\Sigma_i \mathbf{x}_i\mathbf{x}_i^T$ is the covariance matrix of the data, and

$$g_{\mathbf{rt}}=\sum_{\mathbf{s}} h_{\mathbf{rs}}\left(h_{\mathbf{st}}-\frac{1}{N}\right) \tag{28}$$

are the elements of a matrix **G** which acts on the cluster indices. The system of equations (27) decouples under transformation to the eigenbasis of the covariance matrix **C** in data space, and to the eigenbasis of the matrix **G** in cluster space. The former transformation is also known as principal component analysis (PCA) [27]. Denoting the transformed cluster centers by $\hat{\mathrm{w}}'_{\mu\mathbf{k}}$, where μ and **k** designate the components in the new bases of data space and cluster space and the prime and hat denote PCA and the transformation to the eigenbasis of **G**, Eq. (27) becomes

$$\hat{\mathrm{w}}'_{\mu\mathbf{k}}=(\beta\,\lambda_{\mu}^{\mathbf{C}}\,\lambda_{\mathbf{k}}^{\mathbf{G}})\hat{\mathrm{w}}'_{\mu\mathbf{k}}, \tag{29}$$

where $\lambda_{\mu}^{\mathbf{C}}$ and $\lambda_{\mathbf{k}}^{\mathbf{G}}$ are the eigenvalues for the eigenvectors $\mathbf{v}_{\mu}^{\mathbf{C}}$ and $\mathbf{v}_{\mathbf{k}}^{\mathbf{G}}$. Equation (29) can only have nonzero solutions for $\beta\,\lambda_{\mu}^{\mathbf{C}}\,\lambda_{\mathbf{k}}^{\mathbf{G}}=1$. Hence there is a critical β^*,

$$\beta^*=\frac{1}{\lambda_{\max}^{\mathbf{C}}\lambda_{\max}^{\mathbf{G}}}, \tag{30}$$

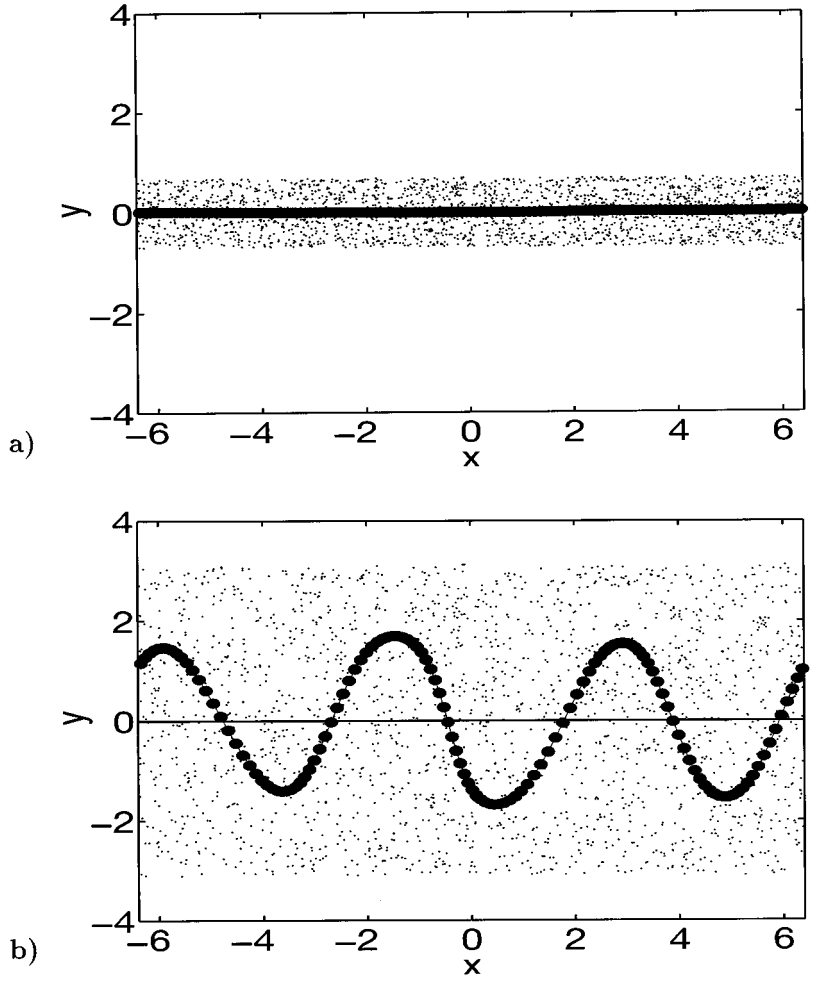

FIG. 3. The phenomenon of ''dimension reduction'' and the automatic selection of feature dimensions. States of minimal free energy are shown (a) before the phase transition ($\sigma_y = 0.4$ d.u.), and (b) after the transition ($\sigma_y = 1.8$ d.u.) for a one-dimensional array of $N = 128$ cluster centers and a two-dimensional data space. The chain of clusters as well as the x dimension in data space are subject to periodic boundary conditions. The x direction is referred to as the longitudinal dimension, the y direction is called the transversal dimension. The units of the axes are data space units (d.u.). The dots represent data points and the filled circles the locations $\mathbf{w_r}$ of the cluster centers. Those cluster centers whose labels differ by one are connected by lines. The transition probabilities $h_{\mathbf{rs}}$ correspond to a Gaussian neighborhood function of standard deviation $\sigma_h = 5.0$. Parameter values $\beta = 1.3$ d.u.$^{-2}$ and $\rho = 10.0$ d.u.$^{-1}$ lead to a critical standard deviation $\sigma_y^* = 1.25$ d.u. and a critical mode $k^* = 3$ for the transition.

at which the center of mass solution becomes unstable, clusters split, and a new representation of the data set emerges. β^* depends on the data via the largest eigenvalue $\lambda^{\mathbf{C}}_{\max}$ of the covariance matrix $\mathbf{C}$ whose eigenvector $\mathbf{v}^{\mathbf{C}}_{\max}$ denotes the direction of maximum variance $\sigma^2_{\max} = \lambda^{\mathbf{C}}_{\max}$ of the data. Consequently, the split of the clusters occurs along the principal axis in data space. β^* also depends on the transition matrix $\mathbf{H}$ via the largest eigenvalue $\lambda^{\mathbf{G}}_{\max}$ of the matrix $\mathbf{G}$. The largest eigenvalue $\lambda^{\mathbf{G}}_{\max}$ indicates which eigenvector $\mathbf{v}^{\mathbf{G}}_{\mathbf{k}} = \mathbf{v}^{\mathbf{G}}_{\max}$ is dominant, and therefore determines the direction in cluster space in which the split occurs. Any component $w'_{\mu\mathbf{r}}$ of vector $\mathbf{w}'_\mu = (w'_{\mu 1}, \ldots, w'_{\mu N})^T$ can be expressed as a linear combination $w'_{\mu\mathbf{r}} = \Sigma_{\mathbf{k}} \hat{w}'_{\mu\mathbf{k}} v^{\mathbf{G}}_{\mathbf{kr}}$ of components $v^{\mathbf{G}}_{\mathbf{kr}}$ of eigenvectors $\mathbf{v}^{\mathbf{G}}_{\mathbf{k}} = (v^{\mathbf{G}}_{\mathbf{k}1}, \ldots, v^{\mathbf{G}}_{\mathbf{k}N})^T$ of the matrix $\mathbf{G}$. Thus the development of cluster center component $w'_{\mu\mathbf{r}}$ under the linearized fixed-point equation (29) depends on the value of the $\mathbf{r}^{\text{th}}$ component of eigenvector $\mathbf{v}^{\mathbf{G}}_{\max}$. Given the principal axis in data space, the eigenvector $\mathbf{v}^{\mathbf{G}}_{\max}$ indicates in which direction along this axis as well as how far each cluster center moves relative to the other cluster centers in the linear approximation.

In order to express this result in terms of eigenvectors $\mathbf{v}^{\mathbf{H}}_{\mathbf{k}}$ and eigenvalues $\lambda^{\mathbf{H}}_{\mathbf{k}}$ of $\mathbf{H}$, it is observed that $\mathbf{G}$ and $\mathbf{H}$ have the same set of eigenvectors. It follows from Eq. (28) that $\mathbf{v}^{\mathbf{G}}_{\max}$ is identical to the eigenvector of $\mathbf{H}$ which corresponds to its second largest eigenvalue $\lambda^{\mathbf{H}}_{\mathbf{k}}$, with $(\lambda^{\mathbf{H}}_{\mathbf{k}})^2 = \lambda^{\mathbf{G}}_{\max}$.

The above results can be extended to the SSOM, which is based on the fixed-point equations (20) and (21). For the SSOM the matrix $\mathbf{G}$, whose elements are given by Eq. (28), must simply be replaced by $\mathbf{G}^{\text{SSOM}}$ with elements $g^{\text{SSOM}}_{\mathbf{rt}} = h_{\mathbf{rt}} - 1/N$.

IV. ANALYSIS OF THE AUTOMATIC SELECTION OF FEATURE DIMENSIONS

A similar analysis as above can be carried out with regard to the phenomenon of the automatic selection of feature dimensions, a term first used by Kohonen [9] in the context of dimension reduction [28,29]. Let us consider a d-dimensional data space and an n-dimensional array of clusters labeled by n-dimensional index vectors $\mathbf{r}$. The couplings $h_{\mathbf{rs}}$ of clusters are defined on this array, and are typically chosen to be a monotonically decreasing function of $\|\mathbf{r} - \mathbf{s}\|$. For $d > n$ a simple representation of the input data is achieved, if the data have significant variance only along n of the d dimensions. In this case, the vectors $\mathbf{w_r}$ lie in an n-dimensional subspace and the excess dimensions are effectively ignored [see Fig. 3(a)]. If, however, the variance of the data in the excess dimensions surpasses a critical value, the original representation becomes unstable, and the array of vectors $\mathbf{w_r}$ folds into the excess dimensions so as to represent them as well [see Fig. 3(b)]. This phenomenon was studied in a formal way by employing a Fokker-Planck approxima-

tion for the dynamics of the (zero temperature) SOM on-line learning algorithm [18,17]. In the following we provide an analysis for the full STVQ family by investigating the fixed-point equations (12) and (13), and compare the results to the limiting case of the SOM.

A. Phase transition in the discrete case

For this purpose, we examine the stability of Eqs. (12) and (13) around a known fixed point. Let us consider the case of an infinite number of data points generated by an underlying probability distribution $P(\mathbf{x})$. The fixed-point equations then read

$$\mathbf{w_r} = \frac{\int P(\mathbf{x}) \mathbf{x} \sum_{\mathbf{s}} h_{\mathbf{rs}} P(\mathbf{x} \in \mathcal{C}_{\mathbf{s}}) d\mathbf{x}}{\int P(\mathbf{x}) \sum_{\mathbf{s}} h_{\mathbf{rs}} P(\mathbf{x} \in \mathcal{C}_{\mathbf{s}}) d\mathbf{x}}, \quad \forall \mathbf{r}, \tag{31}$$

$$P(\mathbf{x} \in \mathcal{C}_{\mathbf{s}}) = \frac{\exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}} h_{\mathbf{st}} \|\mathbf{x} - \mathbf{w_t}\|^2 \right)}{\sum_{\mathbf{u}} \exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}} h_{\mathbf{ut}} \|\mathbf{x} - \mathbf{w_t}\|^2 \right)}, \tag{32}$$

where cluster indices $\mathbf{r}$ are now n-dimensional index vectors which lie on an n-dimensional cubic array, $\mathbf{r} \in \mathcal{N}^n$, $r_\nu \in \{1,2,\ldots,N\}$. For the following we assume that $h_{\mathbf{rs}} : \mathcal{N} \times \mathcal{N} \to [0,1]$ obey $h_{\mathbf{rs}} = h_{\|\mathbf{r}-\mathbf{s}\|}$. For notational convenience, the data space $\mathcal{X}$ is split into two subspaces, $\mathcal{X} = \mathcal{X}^{\|} \oplus \mathcal{X}^{\perp}$, one for the embedding or longitudinal dimensions $\mathcal{X}^{\|}$ with elements $\mathbf{x}^{\|}$ and one for the excess or transversal dimensions $\mathcal{X}^{\perp}$ with elements $\mathbf{x}^{\perp}$. We also assume the probability distribution $P(\mathbf{x})$ over data space $\mathcal{X}$ to factorize as $P(\mathbf{x}) = P(\mathbf{x}^{\|}) P(\mathbf{x}^{\perp})$, where the probability distribution $P(\mathbf{x}^{\perp})$ in the transversal dimensions has zero mean, i.e. $\int P(\mathbf{x}^{\perp}) \mathbf{x}^{\perp} \, d\mathbf{x}^{\perp} = 0$. In the longitudinal dimensions of data space we assume the factorization $P(\mathbf{x}^{\|}) = \Pi_\nu P(x_\nu^{\|})$, with $P(x_\nu^{\|}) = 1/l$ for $-l/2 \leqslant x_\nu^{\|} \leqslant l/2$ and $P(x_\nu^{\|}) = 0$ otherwise, and we consider the system in the approximation $N \to \infty$, $l \to \infty$ and $\rho := N/l$ finite. Since the variance in the longitudinal data space is effectively infinite, for the fixed point of Eqs. (31) and (32) (see Appendix A) we obtain

$$\mathbf{w_r}^{\|0} = \rho^{-1} \mathbf{r} \quad \text{and} \quad \mathbf{w_r}^{\perp\, 0} = \mathbf{0}, \quad \forall \mathbf{r}. \tag{33}$$

Equation (31) can again be expanded to first order in $\mathbf{w_t}$ around the fixed point $\{\mathbf{w_r^0}\}$, just as in Eq. (25). The assignment probability $P^0(\mathbf{x} \in \mathcal{C}_{\mathbf{s}})$ of a data point $\mathbf{x}$ to a cluster $\mathcal{C}_{\mathbf{s}}$ in the fixed-point state (33) depends on the longitudinal components of $\mathbf{x}$ only and—abusing notation—we can write $P^0(\mathbf{x} \in \mathcal{C}_{\mathbf{s}}) = P^0(\mathbf{x}^{\|} \in \mathcal{C}_{\mathbf{s}})$. Let us consider the stability of Eq. (33) along the transversal dimensions which determines the critical parameters for the phase transition depicted in Fig. 3. Using

$$\int P(\mathbf{x}) \, \mathbf{x}^{\perp} \sum_{\mathbf{s}} h_{\mathbf{rs}} \, P^0(\mathbf{x}^{\|} \in \mathcal{C}_{\mathbf{s}}) \, d\mathbf{x} = \mathbf{0} \tag{34}$$

[see Appendix A, Eq. (A4)], for the transversal components of the cluster centers $\mathbf{w_r^{\perp}}$ in the linear approximation we obtain

$$\mathbf{w_r^{\perp}} = \sum_{\mathbf{t}} \frac{\int P(\mathbf{x}) \mathbf{x}^{\perp} \sum_{\mathbf{s}} h_{\mathbf{rs}} \left[\frac{\partial P(\mathbf{x} \in \mathcal{C}_{\mathbf{s}})}{\partial \mathbf{w_t}} \right]_{\{\mathbf{w_r^0}\}} d\mathbf{x}}{\int P(\mathbf{x}) \sum_{\mathbf{s}} h_{\mathbf{rs}} \, P^0(\mathbf{x}^{\|} \in \mathcal{C}_{\mathbf{s}}) \, d\mathbf{x}} \mathbf{w_t}. \tag{35}$$

The denominator of Eq. (35) evaluates to N^{-n} [see Appendix A, Eq. (A7)] because on the average over data space for the fixed point no cluster is singled out. Inserting Eq. (26) into Eq. (35), we obtain

$$\mathbf{w_r^{\perp}} = \beta \, \mathbf{C} \sum_{\mathbf{t}} \sum_{\mathbf{s}} h_{\mathbf{rs}} \left(h_{\mathbf{st}} - \sum_{\mathbf{u}} h_{\mathbf{tu}} \, f_{\mathbf{us}} \right) \mathbf{w_t^{\perp}}, \tag{36}$$

in which $\mathbf{C} = \int P(\mathbf{x}^{\perp}) \, \mathbf{x}^{\perp} \, \mathbf{x}^{\perp T} \, d\mathbf{x}^{\perp}$ is the covariance matrix of the transversal dimensions of data space and

$$f_{\mathbf{us}} = \rho^n \int P^0(\mathbf{x}^{\|} \in \mathcal{C}_{\mathbf{u}}) \, P^0(\mathbf{x}^{\|} \in \mathcal{C}_{\mathbf{s}}) d\mathbf{x}^{\|} \tag{37}$$

is essentially the correlation function of the assignment probabilities of clusters $\mathcal{C}_{\mathbf{u}}$ and $\mathcal{C}_{\mathbf{s}}$ in the fixed-point state $\{\mathbf{w_r^0}\}$ taken over data space. $f_{\mathbf{us}}$ depends on β via the assignment probabilities $P^0(\mathbf{x}^{\|} \in \mathcal{C}_{\mathbf{u}})$. Note that Eq. (36) has the same form as Eq. (27) when $g_{\mathbf{rt}}$ is taken to be $g_{\mathbf{rt}} = \Sigma_{\mathbf{s}} h_{\mathbf{rs}} (h_{\mathbf{st}} - \Sigma_{\mathbf{u}} h_{\mathbf{tu}} \, f_{\mathbf{us}})$.

Equation (36) can again be decoupled in data space by a transformation to the eigenbasis of $\mathbf{C}$. Denoting the components of the transformed cluster centers by $w'^{\perp}_{\mu\mathbf{r}}$, where μ is the index with respect to the eigenvector $\mathbf{v}_\mu^{\mathbf{C}}$ with eigenvalue $\lambda_\mu^{\mathbf{C}}$, Eq. (37) reads

$$w'^{\perp}_{\mu\mathbf{r}} = \beta \, \lambda_\mu^{\mathbf{C}} \sum_{\mathbf{t}} \sum_{\mathbf{s}} h_{\mathbf{rs}} \left(h_{\mathbf{st}} - \sum_{\mathbf{u}} h_{\mathbf{tu}} \, f_{\mathbf{us}} \right) w'^{\perp}_{\mu\mathbf{t}}. \tag{38}$$

From $h_{\mathbf{rs}} = h_{\|\mathbf{r}-\mathbf{s}\|}$, it follows that $f_{\mathbf{rs}} = f_{\|\mathbf{r}-\mathbf{s}\|}$ (see Appendix B). Defining the discrete convolution for two lattice functions $a_{\mathbf{r}}$ and $b_{\mathbf{s}}$ to be $(a * b)_{\mathbf{r}} = \Sigma_{\mathbf{s}} a_{(\mathbf{r}-\mathbf{s})} b_{\mathbf{s}}$, Eq. (38) can be written as

$$w'^{\perp}_{\mu\mathbf{r}} = \beta \, \lambda_\mu^{\mathbf{C}} (h * (h - h * f) * w'^{\perp}_{\mu})_{\mathbf{r}}. \tag{39}$$

Application of the discrete Fourier transform, $\hat{a}_{\mathbf{k}} = \Sigma_{\mathbf{r}} a_{\mathbf{r}} \exp(\mathrm{i}(\mathbf{k} \cdot \mathbf{r}))$, to Eq. (39) leads to a decoupling of Eq. (39) in cluster space as well, and we obtain

$$\hat{w}'^{\perp}_{\mu k} = \beta \, \lambda_\mu^{\mathbf{C}} \, \hat{h}_k^2 (1 - \hat{f}_k) \hat{w}'^{\perp}_{\mu k}, \tag{40}$$

where we make use of the fact that the modes in $\mathbf{k}$ space depend only on the absolute value $k := \|\mathbf{k}\|$ due to the isotropy of the neighborhood function, of the data distribution, and of the fixed-point state. Equation (40) can only have nonzero solutions if $\beta \, \lambda_\mu^{\mathbf{C}} \, \hat{h}_k^2 (1 - \hat{f}_k(\beta)) = 1$. Since $\lambda_\mu^{\mathbf{C}} = \sigma_\mu^2$, where σ_μ^2 is the variance along the μ axis in data space, it is clear that the cluster centers will automatically select the direction in transversal data space with maximum variance $\sigma_{\max}^2$. Thus the eigenvector $\mathbf{v}_{\max}^{\mathbf{C}}$ gives the direction in data

space in which the array of cluster centers folds first. The critical value β^* of the temperature parameter at which this transition occurs is given implicitly by

$$\sigma_{\max}^2 \, \hat{h}_{k^*}^2 \, \beta^*(1-\hat{f}_{k^*}(\beta^*))-1=0, \tag{41}$$

where the critical mode k^* is the mode k for which Eq. (41) has a solution with minimal β. For a given β an explicit expression for the critical variance $(\sigma_{\max}^*)^2$ can be obtained:

$$(\sigma_{\max}^*)^2=\frac{1}{\beta \, \hat{h}_{k^*}^2(1-\hat{f}_{k^*}(\beta))}, \tag{42}$$

where

$$k^*=\arg\max_k \; \hat{h}_k^2(1-\hat{f}_k(\beta)). \tag{43}$$

Very similar results can be derived for the SSOM when the approximation to the E step (21) is applied. The resulting equations are identical to Eqs. (41), (42), and (43) except that $\hat{h}_k$ is not squared and $\hat{f}_k(\beta)$ has to be calculated using the approximation given in Eq. (21).

B. Continuous Gaussian case

To determine values for $(\sigma_{\max}^*)^2$ and k^* for a given β from Eqs. (42) and (43) analytically, we choose $h_{\mathbf{rs}}$ Gaussian with variance σ_h^2 on the distance $\|\mathbf{r}-\mathbf{s}\|$ between clusters $\mathbf{r}$ and $\mathbf{s}$ in the array. We also consider a continuum approximation, i.e., all index vectors $\mathbf{r}$ and their associated index vectors in $\mathbf{k}$ space are real and all functions that were previously defined on $\mathcal{N}^n$ are now defined on the corresponding continuum $\Re^n$. Under these conditions $h_{\mathbf{rs}}$ can be expressed as

$$h_{\mathbf{rs}} \rightarrow h(\|\mathbf{r}-\mathbf{s}\|)=\left(\frac{1}{\sqrt{2\pi}\sigma_h}\right)^n \exp\left(-\frac{\|\mathbf{r}-\mathbf{s}\|^2}{2\sigma_h^2}\right), \tag{44}$$

where n denotes the dimensionality of the cluster array. Inserting Eq. (44) into (37) and replacing sums by integrals yields (see Appendix C)

$$f_{\mathbf{rs}} \rightarrow f(\|\mathbf{r}-\mathbf{s}\|)=\left[\left(\frac{\beta}{4\pi\rho^2}\right)^{1/2}\right]^n \exp\left(-\frac{\beta}{4\,\rho^2}\|\mathbf{r}-\mathbf{s}\|^2\right). \tag{45}$$

Inserting the Fourier transformations of $h_{\|\mathbf{r}-\mathbf{s}\|}$ and $f_{\|\mathbf{r}-\mathbf{s}\|}$ into Eq. (43), we obtain

$$(k^*)^2=\frac{\beta}{\rho^2}\ln\left(1+\frac{\rho^2}{\beta\,\sigma_h^2}\right) \tag{46}$$

from $[(\partial/\partial k)\hat{h}_k^2(1-\hat{f}_k(\beta))]_{k^*}=0$. Inserting Eq. (46) into Eq. (42) finally provides the critical variance $(\sigma_{\max}^*)^2$,

$$(\sigma_{\max}^*)^2=\left(\frac{1}{\beta}+\frac{\sigma_h^2}{\rho^2}\right)\left(1+\frac{\rho^2}{\beta\sigma_h^2}\right)^{\beta\sigma_h^2/\rho^2} \tag{47}$$

for the mode k^*.

An interesting aspect of Eq. (47) is that $1/\beta$ and σ_h^2/ρ^2 appear to play a very similar role. If we interpret β as an inverse variance of the noise in data space, Eq. (47) is essentially the sum of the variance in data space given by $1/\beta$ and the variance σ_h^2 of the noise in cluster space scaled to data space by a factor ρ^{-2}.

The above results are also valid for the case $\beta\rightarrow\infty$ which corresponds to the TVQ given in Eqs. (18) and (19). From Eqs. (46) and (47), we obtain

$$\lim_{\beta\rightarrow\infty}(k^*)^2=\frac{1}{\sigma_h^2}, \tag{48}$$

$$\lim_{\beta\rightarrow\infty}(\sigma_{\max}^*)^2=\frac{\sigma_h^2\, e}{\rho^2}. \tag{49}$$

Equation (48) shows that high values of σ_h^2, i.e., long-ranged coupling between clusters, suppress high transversal modes. From (49) it can be seen, that the critical variance $(\sigma_{\max}^*)^2$ is proportional to the variance of the neighborhood function σ_h^2 scaled to data space by a factor ρ^{-2}. Thus the stability of the fixed-point state $\{\mathbf{w}_{\mathbf{r}}^{\perp 0}\}$ with respect to the variance of the data along the transversal direction in data space can be adjusted by changing σ_h^2.

All the above results carry over to the SOM versions of the algorithm, Eqs. (20)–(23), if σ_h^2 is replaced by ${\sigma'_h}^2/2$ in Eqs. (46)–(49), where ${\sigma'_h}^2$ denotes the variance of the SOM neighborhood function. For the wavelength λ^* of the critical mode we obtain $(\rho=1)$

$$\lambda^*=\frac{2\pi}{k^*}=\sigma'_h\pi\sqrt{2}\approx 4.44\;\sigma'_h. \tag{50}$$

If the critical variance $(\sigma_{\max}^*)^2$ is expressed in terms of the half-width s^* of a homogeneous data distribution we obtain

$$s^*=\sigma'_h\sqrt{3e/2}\approx 2.02\sigma'_h. \tag{51}$$

The last two results (50) and (51) are identical to those presented by Ritter and Schulten [17] for the on-line version of Kohonen's SOM algorithm with a Gaussian neighborhood function using the Fokker-Planck approach.

V. NUMERICAL RESULTS

In this section we present numerical results to validate the analytical calculations and to illustrate the deterministic annealing scheme. We first apply the STVQ to a toy problem with a sufficiently simple transition matrix $\mathbf{H}$ for which the eigenvectors and eigenvalues can be easily calculated. Then, in order to demonstrate the effects and advantages of the deterministic annealing scheme for the STVQ, we consider a two-dimensional array of clusters in a two-dimensional data space. Finally, we investigate the behavior of a one-dimensional ''chain'' of 128 clusters in a two-dimensional data space to validate the results of Sec. IV. Throughout this section components of data vectors will be measured in data space units, abbreviated ''d.u.'' The numerical simulations

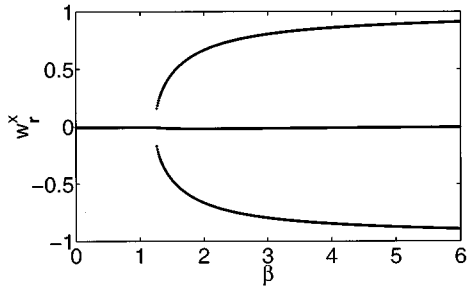

FIG. 4. Plot of the x positions $w_{\mathbf{r}}^x$ (in d.u.) of the cluster centers as functions of β (in d.u.$^{-2}$) for the toy problem with $N=3$ cluster centers and nearest-neighbor coupling. 2000 data points are chosen randomly and independently from the Gaussian probability distribution $P(\mathbf{x})=(2\pi)^{-1}|\mathbf{C}|^{-1/2}\exp(-\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}/2)$ with diagonal covariance matrix $\mathbf{C}$ = diag (1.0 d.u.^{2}, 0.04 d.u.^{2}). Cluster centers are initialized at the origin and STVQ is applied for different values of β. The STVQ iterations are stopped, when $\|\mathbf{w}_{\mathbf{r}}^{(t+1)}-\mathbf{w}_{\mathbf{r}}^{(t)}\| < 5\times10^{-10}$ d.u. for all $\mathbf{r}$. The analytically determined critical value of β is given by $\beta^*=1.21$ d.u.$^{-2}$ for a coupling strength of $s = 0.1$. It corresponds to the trifurcation point seen in the plot.

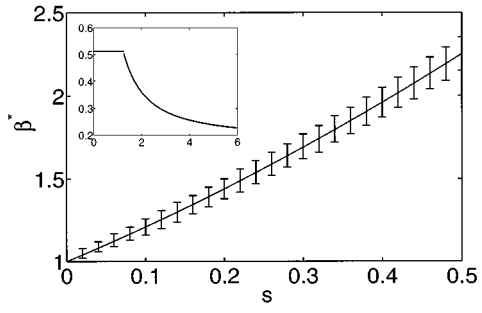

FIG. 5. Plot of the critical value β^* (in d.u.$^{-2}$) of the temperature parameter as a function of the coupling strength s for the STVQ toy problem of Fig. 4. Error bars denote the numerical results. For each value of s the cluster centers are initialized at the origin, and β is linearly annealed according to $\beta_{t+1}=\beta_t+0.02$ d.u.$^{-2}$, with $\beta_0=0.0$ d.u.$^{-2}$ and $\beta_{\text{final}}=6.0$ d.u.$^{-2}$, while monitoring $\langle E\rangle$. For low values of β, the average cost $\langle E\rangle$ is constant. The lower error margins denote the β values, for which the first change in $\langle E\rangle$ occurs and the upper error margins denote the β values, for which the large drop in $\langle E\rangle$ occurs. The line shows the theoretical prediction calculated from Eq. (30) for $\lambda_{\max}^{\mathbf{C}}=\sigma_x^2=1.0$ d.u.^{2} and $\lambda_{\max}^{\mathbf{G}}=1/(1+s)^2$. Inset: Plot of the average cost $\langle E\rangle$ (in d.u.^{2}) as a function of β (in d.u.$^{-2}$) for a typical example ($s = 0.1$). The visible drop in $\langle E\rangle$ occurs at $\beta=1.25$ d.u.$^{-2}$.

were implemented in ANSI C on Sun Sparc 20 and Sun Ultra Sparc workstations.

A. Toy problem

We consider a two-dimensional data space with 2000 data points which were generated by an elongated Gaussian probability distribution $P(\mathbf{x})=(2\pi)^{-1}|\mathbf{C}|^{-1/2}\exp(-\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}/2)$ with diagonal covariance matrix $\mathbf{C}$ = diag (1.0 d.u.^{2}, 0.04 d.u.^{2}). $N = 3$ cluster centers were coupled via a transition probability matrix $\mathbf{H}$,

$$\mathbf{H}=\frac{1}{1+s}\begin{pmatrix}1 & s & 0\\ s & 1-s & s\\ 0 & s & 1\end{pmatrix}. \tag{52}$$

This choice of $\mathbf{H}$ corresponds to a ''chain'' of clusters where each cluster is linked to its nearest neighbor via the transition probability $s/(1+s)$, while second-nearest neighbors are uncoupled because the transition probabilities $h_{13}=h_{31}$ vanish. The magnitude of s governs the coupling strength and the normalization factor is included to comply with condition (4).

Figure 4 shows the x coordinates of the positions $\mathbf{w}_{\mathbf{r}}$ of the cluster centers in data space as functions of the temperature parameter β for the configuration of minimal free energy. At a critical temperature $\beta^*=1.21$ d.u.$^{-2}$ the cluster centers split along the x axis, which is the principal axis of the distribution of data points. In accordance with the eigenvector $\mathbf{v}_{\max}^{\mathbf{G}}$,

$$\mathbf{v}_{\max}^{\mathbf{G}}=(-1 \text{ d.u.}, 0 \text{ d.u.}, 1 \text{ d.u.})^T, \tag{53}$$

for the largest eigenvalue $\lambda_{\max}^{\mathbf{G}}$ of the matrix $\mathbf{G}$ given in Eq. (28) two cluster centers move to opposite positions along the principal axis, while one remains at the center. Therefore, a topologically correct ordering is already established at the initial phase transition. Figure 5 shows the critical value β^* of the temperature parameter as a function of the nearest-neighbor coupling strength s. Error bars indicate the numerical results, which are in agreement with the theoretical prediction of (30) (solid line). The inset displays the average cost $\langle E\rangle$,

$$\langle E\rangle=\tfrac{1}{2}\sum_i \sum_{\mathbf{r}} P(i\in\mathcal{C}_{\mathbf{r}})\sum_{\mathbf{s}} h_{\mathbf{rs}}\|\mathbf{x}_i-\mathbf{w}_{\mathbf{s}}\|^2, \tag{54}$$

as a function of β for a coupling strength of $s = 0.1$. The visible drop of the average cost occurs at $\beta = 1.25$ d.u.$^{-2}$. Note that the transition zone is finite due to finite-size effects.

B. Annealing of a two-dimensional array of cluster centers

Let us now consider a two-dimensional data space and a set of 8×8 clusters labeled by two-dimensional index vectors $\mathbf{r}$, $r_\nu=\{1,2,\ldots,8\}$. The $D=8\times8$ data points lie equally spaced on a grid in the unit square. The transition probabilities $h_{\mathbf{rs}}$ are chosen from a Gaussian function of the distance between the index vectors $\mathbf{r}$ and $\mathbf{s}$,

$$h_{\mathbf{rs}}=\frac{1}{\Theta_{\mathbf{r}}}\exp\left(-\frac{\|\mathbf{r}-\mathbf{s}\|^2}{2\sigma_h^2}\right), \tag{55}$$

with

$$\Theta_{\mathbf{r}}=\sum_{\mathbf{u}}\exp\left(-\frac{\|\mathbf{r}-\mathbf{u}\|^2}{2\sigma_h^2}\right), \tag{56}$$

where the normalization constant $\Theta_{\mathbf{r}}$ is needed to satisfy Eq. (4). This set of transition probabilities corresponds to a ''square grid'' of clusters, and is commonly used in applications of the SOM. Figure 6 shows snapshots of a combined

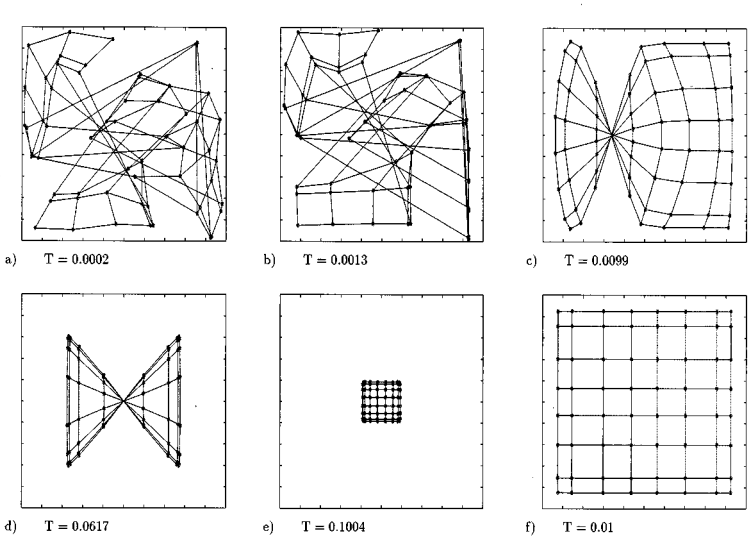

FIG. 6. ''Melting'' of topological defects. The plots show snapshots of cluster centers for a two-dimensional 8×8 cluster array and a two-dimensional data space using STVQ at different temperatures T (in d.u.^{2}). Dots indicate cluster centers with those centers connected by lines which correspond to pairs of clusters for which the transition probability $h_{\mathbf{rs}}$ is highest. Starting from a local minimum of the cost function introduced by random initialization and preserved at low temperature, as seen in (a), the temperature T is increased exponentially according to $T_{t+1}=1.01T_t$. (b)–(e) illustrate the corresponding ''melting'' of topological defects. (f) shows the positions of the cluster centers after ''recooling'' to $T=0.01$ d.u.^{2}. The Gaussian neighborhood function has standard deviation $\sigma_h=0.5$, and the input data consist of 64 data points on a square grid in the unit square.

''heating'' and ''cooling'' experiment which is best described in terms of the temperature $T:=1/\beta$.

For the ''heating'' process annealing starts at a low temperature $T=0.0002$ d.u.^{2} with randomly initialized cluster centers and then the temperature is increased according to an exponential scheme. Figures 6(a)–6(e) display a series of five snapshots of cluster centers during ''heating.'' Defects of the grid, which indicate a local minimum of E, are introduced by the random initialization of the cluster centers and are preserved at low temperatures. As T is gradually increased, shallow local minima vanish and the grid becomes more and more ordered. Finally, a topologically ordered state is reached, which corresponds to the global minimum of the free energy. Because T governs the resolution of the representation in data space, rather localized defects melt away at low temperature, which corresponds to a high resolution in data space, while global twists melt away last.

During ''cooling'' the temperature T is decreased starting from a very high value ($T=0.1$ d.u.^{2}), which corresponds to a state of the system where all cluster centers are merged at the center of mass of the data distribution. Annealing is performed according to the reverse ''heating'' schedule and terminates at $T=0.0002$ d.u.^{2}, which corresponds to the global minimum of the free energy and which is shown in Fig. 6(f). Note that an ordered two-dimensional grid of cluster centers is established at the initial phase transition, and remains in the ordered configuration throughout the ''cooling'' process.

Figure 7 shows the average cost $\langle E\rangle$, a measure for the quality of the data representation, as a function of the temperature T for both annealing experiments from Fig. 6, ''heating'' and ''cooling.''

Figure 8 displays $C:=d\langle E\rangle/dT$, the derivative of the average cost with respect to the temperature, as a function of T for ''heating.'' C is equivalent to the heat capacity in thermodynamics, and can be interpreted as a measure for the progress made in the quality of data representation per change in temperature during annealing. $C(T)$ exhibits pronounced peaks at temperatures which correspond to the ''steps'' in $\langle E\rangle$ during the annealing at which rearrangements of the cluster centers occur. This behavior is analogous to that of physical systems that undergo phase transitions, and reflects in our case a qualitative change in the assignment cost triggered by a small quantitative change in T. The ''heat capacity'' $C(T)$ may also serve to determine a reasonable annealing schedule in the temperature parameter because it indicates critical points during the annealing.

C. Automatic selection of feature dimensions for a chain of clusters

Finally, we consider a data set of 2000 data points drawn from a homogeneous probability distribution defined on a two-dimensional rectangular data space of length $l_x=12.8$

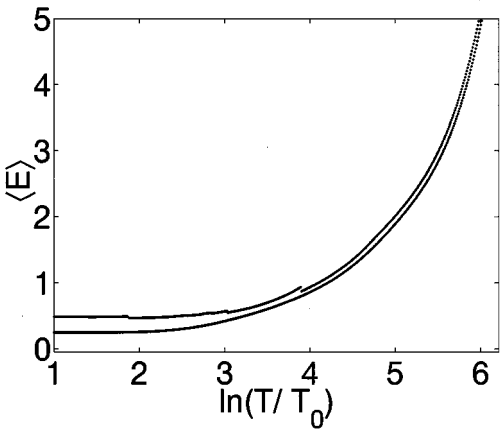

FIG. 7. Semilogarithmic plot of the average assignment cost $\langle E\rangle$ (in d.u.^{2}) as a function of temperature T (in d.u.^{2}) for the cluster array of Fig. 6. The upper curve shows the development of $\langle E\rangle$ for the exponential ''heating'' schedule from $T=0.0002$ to 0.1 d.u.^{2}, starting from the local minimum of the cost function shown in Fig. 6(a). The steps in the average cost occur at temperatures where ''twists'' in the spatial arrangements of cluster centers unfold. The lower curve shows the average cost $\langle E\rangle$ for the same exponential scheme now applied backwards, in the ''cooling'' direction from $T=0.1$ to 0.0002 d.u.^{2}. During ''cooling,'' the cluster centers remain in a ''topologically ordered'' grid-shaped arrangement [cf. Figs. 6(e) and 6(f)]. The normalization constant is $T_0=0.0002$ d.u.^{2}; other parameters are as given in Fig. 6.

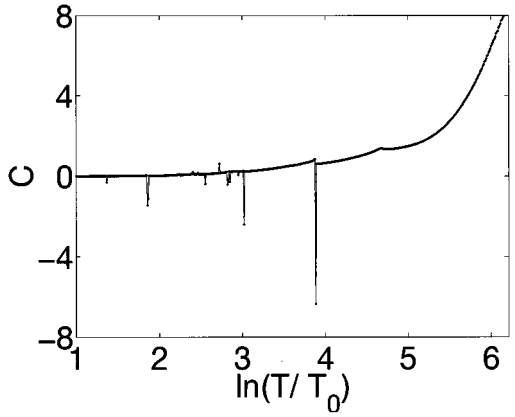

FIG. 8. Semilogarithmic plot of the heat capacity $C(T):=d\langle E\rangle/dT$ as a function of temperature T (in d.u.^{2}) for the ''heating'' as shown in Figs. 6(a)–6(e) and 7 (upper curve). The temperatures corresponding to the peaked minima of the heat capacity indicate transition points of the array of cluster centers as observed in Fig. 6. Parameters are as given in Figs. 6 and 7.

d.u. and a variable width $l_y=2\sqrt{3}\ \sigma_y$, where σ_y^2 is the variance of the probability distribution along the y axis in data space. A set of $N=128$ clusters is labeled by indices $\mathbf{r}=\{1,2,\ldots,N\}$. The transition probabilities $h_{\mathbf{rs}}$ are chosen from a Gaussian function of the distance between indices $\mathbf{r}$ and $\mathbf{s}$,

$$h_{\mathbf{rs}}=\frac{1}{\Theta_{\mathbf{r}}}\exp\left(-\frac{(\min(\|\mathbf{r}-\mathbf{s}\|\ ,\ N-\|\mathbf{r}-\mathbf{s}\|))^2}{2\ \sigma_h^2}\right), \quad (57)$$

with

$$\Theta_{\mathbf{r}}=\sum_{\mathbf{u}}\ \exp\left(-\frac{(\min(\|\mathbf{r}-\mathbf{u}\|\ ,\ N-\|\mathbf{r}-\mathbf{u}\|))^2}{2\sigma_h^2}\right). \quad (58)$$

This set of transition probabilities corresponds to a linear chain of clusters. A one-dimensional chain in a two-dimensional data space constitutes the simplest nontrivial case for which Eq. (47) has been derived.

Since Eq. (47) has been derived for a longitudinal space of infinite size and in the continuum limit, periodic boundary conditions were imposed in the longitudinal x dimension of data space and on the transition probabilities $h_{\mathbf{rs}}$. The cluster centers were initialized according to Eq. (33) [see Fig. 3(a)] with $\rho=10.0$ d.u.$^{-1}$. The size of the system to be examined was important in two aspects. The number of clusters was chosen as large as computationally feasible in order to reduce finite-size effects on the mode spectrum, as well as in order for the continuum approximation to be valid. The number of data points was chosen such that local inhomogeneities would not strongly bias the result, while keeping the computation time still tractable. Figure 3(b) shows the spatial distribution of cluster centers after the variance σ_y^2 has been gradually increased from $\sigma_y^2=0.0$ to 3.24 d.u.^{2} beyond the phase transition. The chain folds into the excess dimension y in a wavelike shape with a dominant wavelength λ^*. This is well illustrated in Fig. 9, which depicts the power in each of the first five Fourier modes as a function of σ_y. At the critical value $\sigma_y^*=1.27$ d.u. the critical mode $k^*=3$ increases in power and, finally, dominates the spatial arrangement of the cluster centers.

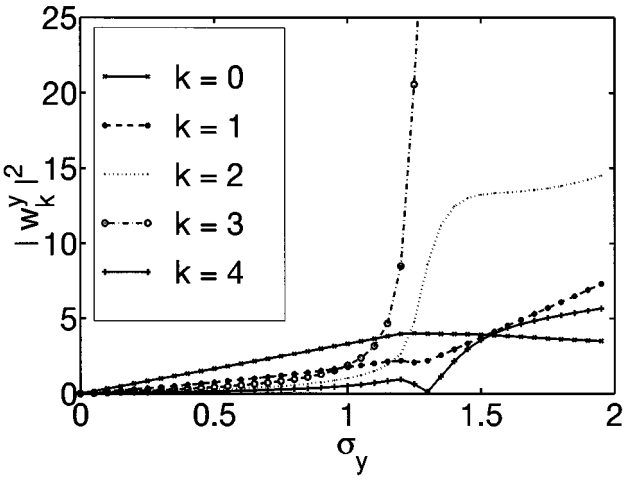

FIG. 9. Plot of the squared absolute amplitudes $\|w_k^y\|^2$ (in d.u.^{2}) of transversal Fourier modes k as functions of the standard deviation σ_y (in d.u.) of the data for the chain of $N=128$ cluster centers shown in Fig. 3. Only the five modes with the largest wavelength are shown. Beyond the phase transition at $\sigma_y^*=1.27$, d.u. the $k=3$ mode is selected, and the chain folds into a sine-wave-like curve. Parameters are given by $\beta=1.3$ d.u.$^{-2}$, $\rho=10.0$ d.u.$^{-1}$, and $\sigma_h=5.0$. The 2000 data points are distributed uniformly in the data plane given by $[-6.4$ d.u., 6.4 d.u.$]\times[-l_y/2,l_y/2]$, where $l_y=2\sqrt{3}\ \sigma_y$ is the width of the data distribution in the y direction.

Figure 10 shows the average cost $\langle E\rangle$ and its derivative with respect to σ_y as functions of σ_y for the numerical experiment shown in Fig. 9. At the critical standard deviation σ_y^* a kink occurs in $d\langle E\rangle/d\sigma_y$. The position of this kink was used to obtain the numerical results of Fig. 11, which

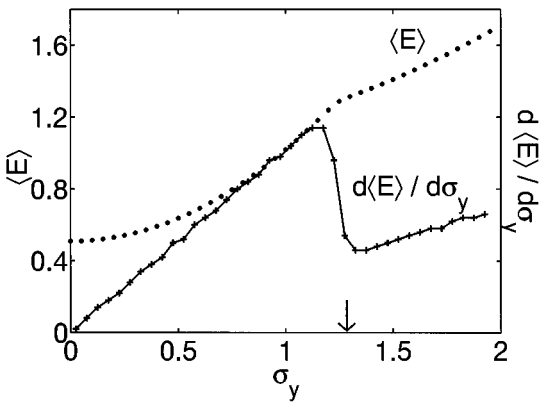

FIG. 10. Plot of the average cost $\langle E\rangle$ (in d.u.^{2}) and its derivative $d\langle E\rangle/d\sigma_y$ (in d.u. scaled by an arb. const.) as functions of the standard deviation σ_y (in d.u.) of the data set in the y dimension for the chain of $N=128$ cluster centers. The slope of the average cost shows a clear change at the critical value of σ_y. Interpolating between σ_y at the minimum and σ_y at the maximum of the derivative yields the critical value σ_y^*. The arrow indicates the theoretical prediction for the critical standard deviation $\sigma_y^*=1.27$ d.u. Parameters are as given in Fig. 9.

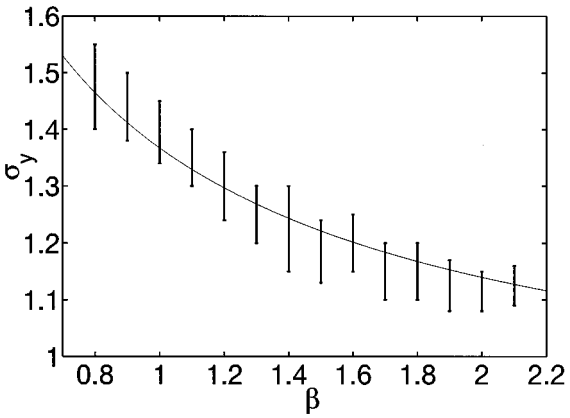

FIG. 11. Plot of the critical standard deviation σ_y^* (in d.u.) as a function of the temperature parameter β (in d.u.$^{-2}$) for the chain of $N=128$ cluster centers. The standard deviation σ_y of the data set in the transversal y dimension is linearly increased for fixed β and the critical value σ_y^* obtained from the derivative of the average cost, as shown in Fig. 10. The upper bound of the error bars is taken from the position of the minimum, and the lower bound from the position of the maximum of $d\langle E\rangle/d\sigma_y$. Parameters are as given in Fig. 9.

compares the theoretical values for σ_y^* (solid line) obtained from Eq. (47) with those that were obtained from the numerical simulations (error bars). The numerical results are in good agreement with the theoretical values obtained in Sec. IV, which justifies the approximations employed in the derivation of Eq. (47).

Similar transitions in the data representation occur during annealing in T for fixed σ_y and σ_h. It can be observed from Figure 12, which shows the heat capacity $C(T)$ for such a case, that a stepwise decrease in T leads to a smooth change of representation from the initial state (left inset) to a folded state (right inset) of the chain. This observation is of interest with regard to neural development in biological systems [18,30]. Interpreting $T=1/\beta$ as a noise parameter leads to the idea that the development of cortical maps may be triggered by a reduction of neuronal noise rather than—as is the widely accepted view [28]—by a change in the variance of the input data.

VI. CONCLUSION

Topographic vector quantizers are useful lossy data compression algorithms that produce encoding-decoding strategies which are robust against channel noise. In order to develop a robust optimization scheme for the TVQ cost function we employed the idea of deterministic annealing and we derived a fuzzy version of the TVQ algorithm in the form of an EM scheme. From this algorithm we then obtained a family of topographic clustering algorithms, among them the self-organizing map, as approximations. Since the annealing process is essential to the algorithm, we examined the behavior of the data representation as a function of temperature. Critical temperatures and modes of the resulting phase transitions were determined and were found to depend on the data distribution via its covariance matrix and on the channel noise, or cluster couplings, via its transition matrix. A similar analysis was performed with regard to the phenomenon of the automatic selection of feature dimensions, and analytical results with respect to the critical variance of the data and critical modes of the folding map were obtained.

Our numerical results confirmed the theoretical predictions and showed the essential features of the annealing process. Since the temperature can be considered as a resolution parameter in data space, the algorithms presented in this paper may prove particularly useful for applications for which optimal topographic vector quantization at different scales is desirable. Our results indicate that the first split of the clusters is in accordance with the desired structure of the data representation, as implicitly given by the transition matrix **H**. This demonstrates the usefulness of deterministic annealing in clustering and provides the STVQ (and the SSOM) with many possible applications. From the interpretation of the temperature as a noise parameter for cluster assignments it follows that phase transitions in topographic clustering can

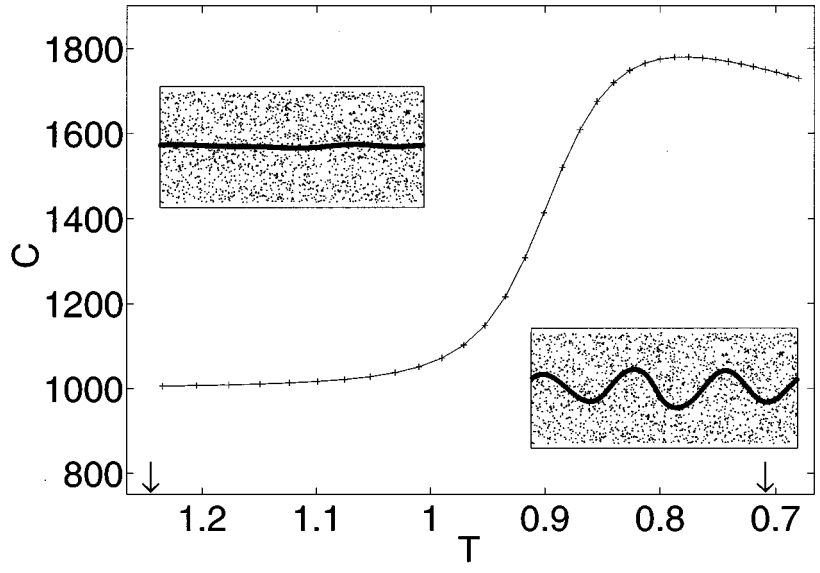

FIG. 12. Plot of the heat capacity $C(T):=d\langle E\rangle/dT$ as a function of the temperature T (in d.u.^{2}) for the chain of $N=128$ cluster centers. Starting from the initial state of the chain at high temperature (left inset), the temperature T is reduced in linear steps in $\beta=1/T$, for fixed σ_y. As T is lowered, the heat capacity C increases, the average cost $\langle E\rangle$ is reduced faster, and the chain is continuously transformed into a folded configuration of the cluster centers (right inset). The vertical arrows indicate the corresponding temperatures, $T=1.25$ and 0.714 d.u.^{2} for the left and right insets, respectively. Parameters are given by $\sigma_y=1.3$ d.u. and $\sigma_h=5.0$.

be induced by decreasing the noise level. This finding has implications for neural development in biological systems, and leads to the hypothesis that the development of cortical maps may be induced by a decrease of neuronal noise rather than by a change in the statistics of the input signals, as is currently believed.

ACKNOWLEDGMENT

This project was funded by the Technical University of Berlin via the Forschungsinitiativprojekt FIP 13/41.

APPENDIX A: PROOF OF THE FIXED-POINT PROPERTY

A set of cluster centers $\{\mathbf{w}_\mathbf{r}^0\}$ qualifies as a fixed point of Eq. (31) if it satisfies

$$\mathbf{w}_\mathbf{r}^0 = \frac{\int P(\mathbf{x})\mathbf{x}\sum_\mathbf{s} h_{\mathbf{rs}} P^0(\mathbf{x}\in\mathcal{C}_\mathbf{s})\, d\mathbf{x}}{\int P(\mathbf{x})\sum_\mathbf{s} h_{\mathbf{rs}} P^0(\mathbf{x}\in\mathcal{C}_\mathbf{s}) d\mathbf{x}}, \quad \forall \mathbf{r}, \tag{A1}$$

where

$$P^0(\mathbf{x}\in\mathcal{C}_\mathbf{s}) = \frac{\exp\left(-\frac{\beta}{2}\sum_\mathbf{t} h_{\mathbf{st}}\|\mathbf{x}-\mathbf{w}_\mathbf{t}^0\|^2\right)}{\sum_\mathbf{u}\exp\left(-\frac{\beta}{2}\sum_\mathbf{t} h_{\mathbf{ut}}\|\mathbf{x}-\mathbf{w}_\mathbf{t}^0\|^2\right)}. \tag{A2}$$

Let us first consider the transversal dimensions. Inserting Eq. (33) into Eq. (A1) yields conditions

$$\mathbf{w}_\mathbf{r}^{\perp 0} = \frac{\int P(\mathbf{x})\mathbf{x}^\perp \sum_\mathbf{s} h_{\mathbf{rs}} P^0(\mathbf{x}\in\mathcal{C}_\mathbf{s}) d\mathbf{x}}{\int P(\mathbf{x})\sum_\mathbf{s} h_{\mathbf{rs}} P^0(\mathbf{x}\in\mathcal{C}_\mathbf{s}) d\mathbf{x}} = \mathbf{0}^\perp, \quad \forall \mathbf{r}. \tag{A3}$$

Using $P^0(\mathbf{x}\in\mathcal{C}_\mathbf{s}) = P^0(\mathbf{x}^\parallel\in\mathcal{C}_\mathbf{s})$ we obtain for the numerator of Eq. (A3)

$$\int P(\mathbf{x})\mathbf{x}^\perp \sum_\mathbf{s} h_{\mathbf{rs}} P^0(\mathbf{x}^\parallel\in\mathcal{C}_\mathbf{s}) d\mathbf{x} = \int P(\mathbf{x}^\parallel)\sum_\mathbf{s} h_{\mathbf{rs}} P^0(\mathbf{x}^\parallel\in\mathcal{C}_\mathbf{s}) d\mathbf{x}^\parallel \int P(\mathbf{x}^\perp)\mathbf{x}^\perp d\mathbf{x}^\perp = \mathbf{0}^\perp \tag{A4}$$

because the mean of $P(\mathbf{x}^\perp)$ was assumed to be zero. Hence Eq. (A3) is satisfied.

For the evaluation of the longitudinal dimensions, we again insert Eq. (33) into Eq. (A1), and obtain conditions

$$\mathbf{w}_\mathbf{r}^{\parallel 0} = \frac{\int P(\mathbf{x})\mathbf{x}^\parallel \sum_\mathbf{s} h_{\mathbf{rs}} P^0(\mathbf{x}\in\mathcal{C}_\mathbf{s}) d\mathbf{x}}{\int P(\mathbf{x})\sum_\mathbf{s} h_{\mathbf{rs}} P^0(\mathbf{x}\in\mathcal{C}_\mathbf{s}) d\mathbf{x}} = \rho^{-1}\mathbf{r}, \quad \forall \mathbf{r}. \tag{A5}$$

Equation (A5) can be written as an average of $\int Q_\mathbf{r}(\mathbf{x}^\parallel)\mathbf{x}^\parallel d\mathbf{x}^\parallel$ over a probability distribution $Q_\mathbf{r}(\mathbf{x}^\parallel)$ given by

$$Q_\mathbf{r}(\mathbf{x}^\parallel) = \frac{P(\mathbf{x}^\parallel)\sum_\mathbf{s} h_{\mathbf{rs}} P^0(\mathbf{x}^\parallel\in\mathcal{C}_\mathbf{s})}{\int P(\mathbf{x}^\parallel)\sum_\mathbf{s} h_{\mathbf{rs}} P^0(\mathbf{x}^\parallel\in\mathcal{C}_\mathbf{s}) d\mathbf{x}^\parallel} = N^n P(\mathbf{x}^\parallel)\sum_\mathbf{s} h_{\mathbf{rs}} P^0(\mathbf{x}^\parallel\in\mathcal{C}_\mathbf{s}), \tag{A6}$$

where in the second step the identity

$$\int P(\mathbf{x}^\parallel)\sum_\mathbf{s} h_{\mathbf{rs}} P^0(\mathbf{x}^\parallel\in\mathcal{C}_\mathbf{s}) d\mathbf{x}^\parallel = \frac{1}{N^n} \tag{A7}$$

has been used. Equation (A7) can be shown by summing both sides over $\mathbf{r}$, yielding unity. To demonstrate the validity of Eq. (A5) we only need to show that $Q_\mathbf{r}(\mathbf{x}^\parallel)$ is symmetric with respect to $\mathbf{w}_\mathbf{r}^{\parallel 0} = \rho^{-1}\mathbf{r}$. Since $P(\mathbf{x}^\parallel)$ is homogeneous, this reduces to showing that $\Sigma_\mathbf{s} h_{\mathbf{rs}} P^0(\mathbf{x}^\parallel\in\mathcal{C}_\mathbf{s})$ is symmetric with respect to $\rho^{-1}\mathbf{r}$. This is equivalent to

$$\sum_\mathbf{s} h_{\mathbf{rs}} P^0(\mathbf{x}^\parallel\in\mathcal{C}_\mathbf{s}) = \sum_\mathbf{s} h_{\mathbf{rs}} P^0(2\rho^{-1}\mathbf{r}-\mathbf{x}^\parallel\in\mathcal{C}_\mathbf{s}) = \sum_\mathbf{s} h_{\mathbf{rs}} \frac{\exp\left(-\frac{\beta}{2}\sum_\mathbf{t} h_{\mathbf{st}}\|\mathbf{x}^\parallel-\rho^{-1}(2\mathbf{r}-\mathbf{t})\|^2\right)}{\sum_\mathbf{u}\exp\left(-\frac{\beta}{2}\sum_\mathbf{t} h_{\mathbf{ut}}\|\mathbf{x}^\parallel-\rho^{-1}(2\mathbf{r}-\mathbf{t})\|^2\right)}. \tag{A8}$$

From $h_{\mathbf{rs}} = h_{\|\mathbf{r}-\mathbf{s}\|}$, it follows that $h_{\mathbf{rs}} = h_{\mathbf{r}(2\mathbf{r}-\mathbf{s})}$. Substituting $\mathbf{s}\to\mathbf{s}' = 2\mathbf{r}-\mathbf{s}$ and $\mathbf{t}\to\mathbf{t}' = 2\mathbf{r}-\mathbf{t}$ we can write

$$\sum_\mathbf{s} h_{\mathbf{rs}} P^0(\mathbf{x}^\parallel\in\mathcal{C}_\mathbf{s}) = \sum_{\mathbf{s}'} h_{\mathbf{rs}'} \frac{\exp\left(-\frac{\beta}{2}\sum_\mathbf{t} h_{(2\mathbf{r}-\mathbf{s}')\mathbf{t}}\|\mathbf{x}^\parallel-\rho^{-1}(2\mathbf{r}-\mathbf{t})\|^2\right)}{\sum_\mathbf{u}\exp\left(-\frac{\beta}{2}\sum_\mathbf{t} h_{\mathbf{ut}}\|\mathbf{x}^\parallel-\rho^{-1}(2\mathbf{r}-\mathbf{t})\|^2\right)}$$

$$= \sum_{\mathbf{s}'} h_{\mathbf{r}\mathbf{s}'} \frac{\exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}'} h_{\mathbf{s}'\mathbf{t}'} \|\mathbf{x}^{\|} - \rho^{-1}\mathbf{t}'\|^2 \right)}{\sum_{\mathbf{u}} \exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}'} h_{\mathbf{u}\mathbf{t}'} \|\mathbf{x}^{\|} - \rho^{-1}\mathbf{t}'\|^2 \right)} = \sum_{\mathbf{s}'} h_{\mathbf{r}\mathbf{s}'} P^0(\mathbf{x}^{\|} \in \mathcal{C}_{\mathbf{s}'}). \tag{A9}$$

Thus the probability distribution $Q_{\mathbf{r}}(\mathbf{x}^{\|})$ is symmetric with respect to $\mathbf{w}_{\mathbf{r}}^{\|0} = \rho^{-1}\mathbf{r}$ and consequently $\int Q_{\mathbf{r}}(\mathbf{x}^{\|})\ \mathbf{x}^{\|} d\mathbf{x}^{\|} = \rho^{-1}\mathbf{r}$. Hence Eq. (A5) is correct, and Eq. (33) is a fixed point of Eq. (31).

APPENDIX B: DERIVATION OF THE SYMMETRY PROPERTIES OF THE ASSIGNMENT CORRELATIONS

Here we show that $f_{\mathbf{r}\mathbf{s}} = f_{\|\mathbf{r}-\mathbf{s}\|}$ follows from $h_{\mathbf{r}\mathbf{s}} = h_{\|\mathbf{r}-\mathbf{s}\|}$. Starting from Eq. (37), we can express f_{rs} as

$$f_{\mathbf{r}\mathbf{s}} = N^n \int P(\mathbf{x}^{\|}) \frac{\exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}} (h_{\|\mathbf{r}-\mathbf{t}\|} + h_{\|\mathbf{s}-\mathbf{t}\|}) \|\mathbf{x}^{\|} - \rho^{-1}\mathbf{t}\|^2 \right)}{\left[\sum_{\mathbf{u}} \exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}} h_{\|\mathbf{u}-\mathbf{t}\|} \|\mathbf{x}^{\|} - \rho^{-1}\mathbf{t}\|^2 \right) \right]^2} d\mathbf{x}^{\|}. \tag{B1}$$

Substituting $\mathbf{t} \to \mathbf{t}' = \mathbf{A}(\mathbf{t}-\mathbf{s})$, where $\mathbf{A}$ is any nonsingular, length-preserving transformation matrix, and using $\|\mathbf{A}\mathbf{r}\| = \|\mathbf{r}\|$, $\forall \mathbf{r}$, we obtain

$$f_{\mathbf{r}\mathbf{s}} = N^n \int P(\mathbf{x}^{\|}) \frac{\exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}'} (h_{\|\mathbf{A}(\mathbf{r}-\mathbf{s})-\mathbf{t}'\|} + h_{\|\mathbf{t}'\|}) \|\mathbf{x}^{\|} - \rho^{-1}(\mathbf{A}^{-1}\mathbf{t}' + \mathbf{s})\|^2 \right)}{\left[\sum_{\mathbf{u}} \exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}'} h_{\|\mathbf{u}-\mathbf{A}^{-1}\mathbf{t}'-\mathbf{s}\|} \|\mathbf{x}^{\|} - \rho^{-1}(\mathbf{A}^{-1}\mathbf{t}' + \mathbf{s})\|^2 \right) \right]^2} d\mathbf{x}^{\|}. \tag{B2}$$

Substituting $\mathbf{x}^{\|} \to \mathbf{x}^{\|\prime} = \mathbf{A}(\mathbf{x}^{\|} - \rho^{-1}\mathbf{s})$ and $\mathbf{u} \to \mathbf{u}' = \mathbf{A}(\mathbf{u}-\mathbf{s})$ leads to

$$f_{\mathbf{r}\mathbf{s}} = N^n \int P(\mathbf{A}^{-1}\mathbf{x}^{\|\prime} + \rho^{-1}\mathbf{s}) \frac{\exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}'} (h_{\|\mathbf{A}(\mathbf{r}-\mathbf{s})-\mathbf{t}'\|} + h_{\|\mathbf{t}'\|}) \|\mathbf{x}^{\|\prime} - \rho^{-1}\mathbf{t}'\|^2 \right)}{\left[\sum_{\mathbf{u}} \exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}'} h_{\|\mathbf{u}-\mathbf{A}^{-1}\mathbf{t}'-\mathbf{s}\|} \|\mathbf{x}^{\|\prime} - \rho^{-1}\mathbf{t}'\|^2 \right) \right]^2} d\mathbf{x}^{\|\prime}$$

$$= N^n \int P(\mathbf{A}^{-1}\mathbf{x}^{\|\prime} + \rho^{-1}\mathbf{s}) \frac{\exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}'} (h_{\|\mathbf{A}(\mathbf{r}-\mathbf{s})-\mathbf{t}'\|} + h_{\|\mathbf{t}'\|}) \|\mathbf{x}^{\|\prime} - \rho^{-1}\mathbf{t}'\|^2 \right)}{\left[\sum_{\mathbf{u}'} \exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}'} h_{\|\mathbf{u}'-\mathbf{t}'\|} \|\mathbf{x}^{\|\prime} - \rho^{-1}\mathbf{t}'\|^2 \right) \right]^2} d\mathbf{x}^{\|\prime}. \tag{B3}$$

Comparing Eqs. (B3) and (B1), it can be seen that $f_{\mathbf{r}\mathbf{s}}$ is a function of $\mathbf{A}(\mathbf{r}-\mathbf{s})$, if $P(\mathbf{A}^{-1}\mathbf{x}^{\|} + \rho^{-1}\mathbf{s}) = P(\mathbf{x}^{\|})$. This is the case for our particular choice $P(\mathbf{x}^{\|}) = l^{-n}$, and, since $\mathbf{A}$ can be any length-preserving linear transformation, it follows that $f_{\mathbf{r}\mathbf{s}} = f_{\|\mathbf{r}-\mathbf{s}\|}$.

APPENDIX C: EVALUATION OF THE ASSIGNMENT CORRELATION FOR GAUSSIAN NEIGHBORHOOD FUNCTIONS

Starting from Eq. (37), we calculate the approximation of $f_{\mathbf{r}\mathbf{s}}$ as given in Eq. (45) for the homogeneous isotropic Gaussian neighborhood function given in Eq. (44) in the continuum approximation. Inserting the assignment probabilities $P^0(\mathbf{x}^{\|} \in \mathcal{C}_{\mathbf{r}})$, Eq. (A2), for the fixed point (33) into Eq. (37) gives

$$f_{\mathbf{r}\mathbf{s}} = \rho^n \int \frac{\exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}} (h_{\mathbf{r}\mathbf{t}} + h_{\mathbf{s}\mathbf{t}}) \|\mathbf{x}^{\|} - \mathbf{w}_{\mathbf{t}}^{\|0}\|^2 \right)}{\left[\sum_{\mathbf{u}} \exp\left(-\frac{\beta}{2} \sum_{\mathbf{t}} h_{\mathbf{u}\mathbf{t}} \|\mathbf{x}^{\|} - \mathbf{w}_{\mathbf{t}}^{\|0}\|^2 \right) \right]^2} d\mathbf{x}^{\|}. \tag{C1}$$

First we evaluate the expression $\Sigma_{\mathbf{t}} h_{\mathbf{r}\mathbf{t}} \|\mathbf{x}^{\|} - \mathbf{w}_{\mathbf{t}}^{\|0}\|^2$ in the continuum approximation with sums replaced by integrals by using the property of the fixed point $\mathbf{w}_{\mathbf{t}}^{\|\ 0} = \rho^{-1}\mathbf{t}$ from (33). This gives

$$\sum_{\mathbf{t}} h_{\mathbf{r}\mathbf{t}} \|\mathbf{x}^{\|} - \mathbf{w}_{\mathbf{t}}^{\|0}\|^2$$

$$\approx\left(\frac{1}{\sqrt{2\pi}\sigma_h}\right)^n\int \exp\left(-\frac{\|\mathbf{r}-\mathbf{t}\|^2}{2\ \sigma_h^2}\right)\|\mathbf{x}^{\|}-\rho^{-1}\mathbf{t}\|^2\ d\mathbf{t}$$

$$=\left(\frac{1}{\sqrt{2\pi}\sigma_h}\right)^n\int \exp\left(-\frac{\|\mathbf{t}'\|^2}{2\ \sigma_h^2}\right)\|\mathbf{x}^{\|}-\rho^{-1}(\mathbf{t}'+\mathbf{r})\|^2\ d\mathbf{t}', \tag{C2}$$

for which $\mathbf{t}\to\mathbf{t}'=\mathbf{t}-\mathbf{r}$. Now the evaluation of the integral is straightforward, and we obtain

$$\sum_{\mathbf{t}} h_{\mathbf{rt}}\|\mathbf{x}^{\|}-\mathbf{w}_{\mathbf{t}}^{\|0}\|^2\approx\|\mathbf{x}^{\|}-\mathbf{w}_{\mathbf{r}}^{\|0}\|^2+n\rho^{-2}\sigma_h^2. \tag{C3}$$

Inserting this into Eq. (C1) and observing that the expression $\exp(-\beta n\rho^{-2}\sigma_h^2)$ appears as a factor in the numerator and denominator and thus cancels, we arrive at

$$f_{\mathbf{rs}}\approx\rho^n\int\frac{\exp\left(-\frac{\beta}{2}(\|\mathbf{x}^{\|}-\rho^{-1}\mathbf{r}\|^2+\|\mathbf{x}^{\|}-\rho^{-1}\mathbf{s}\|^2)\right)}{\left[\sum_{\mathbf{u}}\exp\left(-\frac{\beta}{2}\|\mathbf{x}^{\|}-\rho^{-1}\mathbf{u}\|^2\right)\right]^2}d\mathbf{x}^{\|}. \tag{C4}$$

The denominator of the integrand in Eq. (C4) is approximated by

$$\left[\sum_{\mathbf{u}}\exp\left(-\frac{\beta}{2}\|\mathbf{x}^{\|}-\rho^{-1}\mathbf{u}\|^2\right)\right]^2$$

$$\approx\left[\int\exp\left(-\frac{\beta}{2}\|\mathbf{x}^{\|}-\rho^{-1}\mathbf{u}\|^2\right)d\mathbf{u}\right]^2$$

$$=\left(\frac{2\pi\rho^2}{\beta}\right)^n, \tag{C5}$$

and the numerator of the integrand in Eq. (C4) can be rewritten as

$$\exp\left(-\frac{\beta}{2}(\|\mathbf{x}^{\|}-\rho^{-1}\mathbf{r}\|^2+\|\mathbf{x}^{\|}-\rho^{-1}\mathbf{s}\|^2)\right)$$

$$=\exp\left(-\frac{\beta}{4}(\|2\mathbf{x}^{\|}-\rho^{-1}(\mathbf{r}+\mathbf{s})\|^2+\|\rho^{-1}(\mathbf{r}-\mathbf{s})\|^2)\right). \tag{C6}$$

Inserting Eqs. (C5) and (C6) into Eq. (C4), and using

$$\int\exp\left(-\frac{\beta}{4}\|2\mathbf{x}^{\|}-\rho^{-1}(\mathbf{r}+\mathbf{s})\|^2\right)d\mathbf{x}^{\|}=\left(\frac{\pi}{\beta}\right)^{n/2},$$

we finally obtain the continuum approximation for $f_{\mathbf{rs}}$,

$$f_{\mathbf{rs}}\approx f(\|\mathbf{r}-\mathbf{s}\|)=\left[\left(\frac{\beta}{4\pi\rho^2}\right)^{1/2}\right]^n\exp\left(-\frac{\beta}{4\rho^2}\|\mathbf{r}-\mathbf{s}\|^2\right). \tag{C7}$$

[1] J. M. Buhmann and H. Kühnel, IEEE Trans. Inf. Theory **39**, 1133 (1993).
[2] J. M. Buhmann and T. Hofmann (unpublished).
[3] T. Kohonen, Biol. Cybern. **43**, 59 (1982).
[4] T. Kohonen, Biol. Cybern. **44**, 135 (1982).
[5] T. Kohonen, *Self-Organizing Maps* (Springer-Verlag Berlin, 1995).
[6] T. Kohonen, Bibliography on the SOM, http://nucleus.hut.fi/nnrc/refs/.
[7] E. Erwin, K. Obermayer, and K. Schulten, Neural Comput. **7**, 425 (1995).
[8] N. W. Swindale, Network **7**, 161 (1996).
[9] T. Kohonen, *Self-Organization and Associative Memory*, 3rd ed. (Springer-Verlag, Berlin, 1989).
[10] J. MacQueen, in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, edited by L. M. LeCam and J. Neyman (University of California Press, Berkeley, 1967), p. 281.
[11] K. Rose, E. Gurewitz, and G. C. Fox, Phys. Rev. Lett. **65**, 945 (1990).
[12] A. L. Yuille, P. Stolortz, and J. Utans, Neural Comput. **6**, 334 (1994).
[13] A. L. Yuille and J. J. Kosowsky, Neural Comput. **6**, 341 (1994).
[14] S. P. Luttrell, Proc. IJCNN **2**, 495 (1989).
[15] S. P. Luttrell, IEEE Trans. Neural Netw. **2**, 427 (1991).
[16] S. P. Luttrell, Neural Comput. **6**, 767 (1994).
[17] H. Ritter and K. Schulten, Biol. Cybern. **60**, 59 (1988).
[18] K. Obermayer, G. G. Blasdel, and K. Schulten, Phys. Rev. A **45**, 7568 (1992).
[19] T. M. Heskes and B. Kappen, Proc. IEEE-ICNN **3**, 1219 (1993).
[20] H. Ritter, *Artificial Neural Networks* (Elsevier/North-Holland, Amsterdam, 1991), p. 379.
[21] A. P. Dempster, N. M. Laird, and D. B. Rubin, J. R. Stat. Soc. B **39**, 1 (1977).
[22] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[23] J. Puzicha, T. Hofmann, and J. M. Buhmann (unpublished).
[24] S. Geman and D. Geman, IEEE Trans. Pattern. Anal. Mach. Intell. **6**, 721 (1984).
[25] F. Mulier and V. Cherkassky, Neural Comput. **7**, 1165 (1995).
[26] H. Robbins and S. Monroe, Ann. Math. Stat. **22**, 400 (1951).
[27] H. Karhunen, Ann. Acad. Sci. Fenn. Ser. A.I. **37**, 3 (1947).
[28] R. Durbin and G. Mitchison, Nature (London) **343**, 644 (1990).
[29] K. Obermayer, H. Ritter, and K. Schulten, Proc. Natl. Acad. Sci. USA **87**, 8345 (1990).
[30] H. Bauer, M. Riesenhuber, and T. Geisel, Phys. Rev. E **54**, 2807 (1996).