Deterministic Annealing for Topographic Vector Quantization and Self-Organizing Maps

Thore Graepel
Matthias Burger
Klaus Obermayer

Technical University of Berlin
Computer Science Department, FR 2-1, KON
Franklinstraße 28/29, 10587 Berlin, Germany

Abstract

We have developed a robust optimization scheme for self-organizing maps in the framework of noisy vector quantization. Based on a cost function that takes distortions from channel noise into account we derive a fuzzy algorithm of EM-type for topographic vector quantization (STVQ) which employes deterministic annealing. This annealing process leads to phase transitions in the cluster representation for which we are able to calculate critical modes and temperatures as a function of the neighbourhood function and the covariance matrix of the data. Similar results are obtained for the automatic selection of feature dimensions. Deterministic annealing also offers an alternative to the heuristic stepwise shrinking of the neighbourhood width in the SOM and makes it possible to use the neighbourhood solely to encode desired neighbourhood relations between the clusters. A soft version of the SOM (SSOM) is derived as a computationally efficient approximation to the E-step of STVQ. Both methods are numerically tested on a two-dimensional map of the plane and we conclude that the temperature annealing can be precisely controlled and could for many applications be the method of choice.

1 Introduction

The SOM algorithm was originally proposed by Kohonen [4, 5] on the grounds of biological plausibility and was formulated as an iterative "learning" procedure. Since then it has been successfully applied to a vast variety of problems in data analysis [6]. Its transparency, however, suffered from the fact, that SOM had not originally been formulated as an optimization procedure. Luttrell [8] established a connection between the SOM and vector quantization. By choosing a cost function for vector quantization based on a distortion measure that incorporates robustness w.r.t. noise-induced change of representation, he derived an algorithm which he named topographic vector quantization (TVQ), and showed that the SOM can be considered as a computationally efficient approximation to TVQ. The relation of TVQ and SOM to cost functions makes it possible to apply elaborate learning procedures also to SOM. In this paper we investigate one of these procedures, which is based on deterministic annealing [1, 11].

Deterministic annealing is an optimization method derived from statistical physics which has been successfully applied to data clustering [1, 2, 11]. By defining fuzzy assignments of data points to clusters via a temperature parameter β, the solution of the optimization problem can be tracked from high to low temperature thereby avoiding local minima of the cost function. In this contribution, we apply the idea of deterministic annealing to the TVQ cost function and derive an EM-type algorithm which we call soft-TVQ (STVQ). Using the fact that SOM is an approximation to TVQ we derive a deterministic annealing scheme for the SOM and we mathematically analyse phase transitions that occur during the annealing. The annealing in the temperature parameter

β is then compared to the stepwise shrinking of the neighbourhood width in the SOM, which has previously been used to ensure correct topological ordering of maps. We conclude that deterministic annealing is a robust optimization procedure for self-organizing maps that allows precise control over the optimization process and leads to favourable results.

2 Derivation of the Algorithm

Self-organizing maps can be viewed as topology preserving clustering algorithms. 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 a clustering algorithm is to assign each data point $\mathbf{x}_i$ to a cluster $\mathcal{C}_\mathbf{r}$, such that a given energy function E is minimized. The membership of data point $\mathbf{x}_i$ in cluster $\mathcal{C}_\mathbf{r}$ is indicated by a binary assignment variable $m_{i\mathbf{r}}$ which takes the value one if data point $\mathbf{x}_i$ is member of cluster $\mathcal{C}_\mathbf{r}$ and zero otherwise. In central clustering there exists a cluster center $\mathbf{w}_\mathbf{r}$ for each cluster $\mathcal{C}_\mathbf{r}$, which is the representative in data space of those data points that are members of cluster $\mathcal{C}_\mathbf{r}$. Commonly, the squared Euclidean distance between data point $\mathbf{x}_i$ and cluster center $\mathbf{w}_\mathbf{r}$ is taken as the cost of assigning a data point $\mathbf{x}_i$ to a cluster $\mathcal{C}_\mathbf{r}$.

Following an idea by Luttrell [8], however, we consider the case that the cluster indices $\mathbf{r}$ are sent via a noisy channel. The channel noise induces a further distortion, which is described by transition probabilities $h_{\mathbf{rs}}$ for a transition from cluster $\mathcal{C}_\mathbf{r}$ to cluster $\mathcal{C}_\mathbf{s}$. Averaging the squared Euclidean distance over all possible noise induced transitions and summing over all assignments yields the cost or energy function

$$E\left(\{m_{i\mathbf{r}}\}, \{\mathbf{w}_\mathbf{r}\}\right) = \frac{1}{2} \sum_i \sum_\mathbf{r} m_{i\mathbf{r}} \sum_\mathbf{s} h_{\mathbf{rs}} \left\|\mathbf{x}_i - \mathbf{w}_\mathbf{s}\right\|^2 , \tag{1}$$

whose variables $m_{i\mathbf{r}}$ and $h_{\mathbf{rs}}$ are subject to the constraints

$$\sum_\mathbf{r} m_{i\mathbf{r}} = 1 \quad \forall i \qquad \text{and} \qquad \sum_\mathbf{s} h_{\mathbf{rs}} = 1 \quad \forall \mathbf{r} . \tag{2}$$

The transition probabilities have been termed $h_{\mathbf{rs}}$ because they are closely related to the elements of the neighbourhood matrix $\mathbf{H}$ in the SOM. The energy function (1) takes its minimum when a noise robust representation of the data is achieved. If the assignment of a data point $\mathbf{x}_i$ changes from cluster $\mathcal{C}_\mathbf{r}$ to cluster $\mathcal{C}_\mathbf{s}$ with a high probability $h_{\mathbf{rs}}$ the corresponding representatives or cluster centers $\mathbf{w}_\mathbf{r}$ and $\mathbf{w}_\mathbf{s}$ must be located close to each other in data space in order to prevent the noise from considerably deteriorating the representation. This leads to an effective coupling between the cluster centers. The symmetry of the clustering cost function w.r.t. permutations of indicies is broken and the transition probability can be interpreted as a neighbourhood function in the space of clusters.

From (1) the Gibbs-distribution $P\left(\{m_{i\mathbf{r}}\}, \{\mathbf{w}_\mathbf{r}\}\right) = \frac{1}{Z} \exp\left(-\beta\, E\left(\{m_{i\mathbf{r}}\}, \{\mathbf{w}_\mathbf{r}\}\right)\right)$ can be obtained via the principle of maximum entropy under the constraint of a given mean energy $\langle E \rangle$. Here β is the Lagrange multiplier associated with $\langle E \rangle$ and is interpreted as an inverse temperature which determines the fuzzyness of the assignments. Since we are primarily interested in finding the most probable set of cluster centers $\{\mathbf{w}_\mathbf{r}\}$ so as to generalize from a given set of training samples, this Gibbs-distribution is marginalized over all legal sets of associations $\{m_{i\mathbf{r}}\}$. Considering the log-likelihood we obtain

$$\log P\left(\{\mathbf{w}_\mathbf{r}\}\right) = \sum_i \log \sum_\mathbf{r} \exp\Big(-\frac{\beta}{2} \sum_\mathbf{s} h_{\mathbf{rs}} \left\|\mathbf{x}_i - \mathbf{w}_\mathbf{s}\right\|^2\Big) - \log Z . \tag{3}$$

Maximizing (3) w.r.t. $\{\mathbf{w}_\mathbf{r}\}$ yields conditions for the cluster centers $\{\mathbf{w}_\mathbf{r}\}$ at a given β,

$$\mathbf{w}_\mathbf{r} = \frac{\sum_i \mathbf{x}_i \sum_\mathbf{s} h_{\mathbf{rs}} P(i \in \mathcal{C}_\mathbf{s})}{\sum_i \sum_\mathbf{s} h_{\mathbf{rs}} P(i \in \mathcal{C}_\mathbf{s})} \qquad \forall \mathbf{r} , \tag{4}$$

where $P(i \in \mathcal{C}_{\mathbf{r}}) = \langle m_{i\mathbf{r}} \rangle$ is the assignment probability of data point $\mathbf{x}_i$ to cluster $\mathcal{C}_{\mathbf{r}}$ and is given by

$$P(i \in \mathcal{C}_{\mathbf{r}}) = \frac{\exp\left(-\frac{\beta}{2}\sum_{\mathbf{s}} h_{\mathbf{rs}} \|\mathbf{x}_i - \mathbf{w}_{\mathbf{s}}\|^2\right)}{\sum_{\mathbf{t}} \exp\left(-\frac{\beta}{2}\sum_{\mathbf{s}} h_{\mathbf{ts}} \|\mathbf{x}_i - \mathbf{w}_{\mathbf{s}}\|^2\right)}. \tag{5}$$

For a given β, (4) and (5) can be iteratively solved in an EM fashion. Starting from $\beta = 0$ (infinite temperature) β can be subjected to an annealing schedule which, when carefully chosen, leads to the global minimum of the cost function (1) corresponding to a topologically ordered state representing the data.

The above derived algorithm (STVQ) can be put into a familiar context by considering certain limiting cases and approximations which are summarized in Figure 1. In particular the approximation $h_{\mathbf{rs}} \to \delta_{\mathbf{rs}}$ in the assignment probabilities (5) leads to a fuzzy version of the SOM algorithm which we call soft-SOM (SSOM).

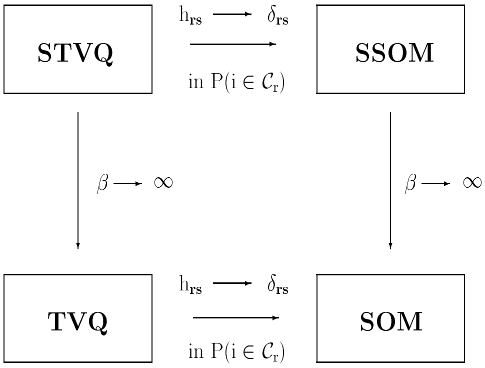

Figure 1: The family of topological clustering algorithms. The limiting case $\beta \to \infty$ in the assignment probabilities (5) leads to the batch version of the TVQ. The approximation $h_{\mathbf{rs}} \to \delta_{\mathbf{rs}}$ in (5) yields a fuzzy version of the SOM algorithm which we have called soft-SOM (SSOM). Employing $h_{\mathbf{rs}} \to \delta_{\mathbf{rs}}$ and $\beta \to \infty$ in (5), leads to the batch version of the SOM.

3 Analysis of the Algorithm

Since we want to better understand the annealing process in the temperature parameter β it is instructive to look at how the representation of the data changes with β. It is known [1, 11] from soft clustering that the cluster centers split as β is increased. Thus β determines the resolution via the number of relevant clusters that have split up to that point. In STVQ, however, the permutation symmetry of the cluster centers is broken and couplings between clusters are introduced by the transition matrix $\mathbf{H}$. This changes stationary states and the "splitting" behaviour of the cluster centers. From (4) it can be seen that at $\beta = 0$, 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 clusters. This results in a fixed-point $\mathbf{w}_{\mathbf{r}}^0 = \frac{1}{D}\sum_i \mathbf{x}_i \; \forall \mathbf{r}$. To determine the stability of this fixed-point w.r.t. changes in β we make a Taylor-expansion of the r.h.s. of (4) around the fixed-point $\{\mathbf{w}_{\mathbf{r}}^0\}$. For $h_{\mathbf{rs}} = h_{\mathbf{sr}} \; \forall \, \mathbf{r}, \mathbf{s}$, we arrive at a critical value β^*,

$$\beta^* = \frac{1}{\lambda_{\max}^{\mathbf{C}} \lambda_{\max}^{\mathbf{G}}}, \tag{6}$$

at which the first split of the cluster centers occurs. Here $\lambda_{\max}^{\mathbf{C}}$ is the largest eigenvalue of the covariance matrix of the data $\mathbf{C} = \frac{1}{D}\sum_i \mathbf{x}_i \mathbf{x}_i^T$ and is identified with the variance $\lambda_{\max}^{\mathbf{C}} = \sigma_{\max}^2$ of the data along the principal axis. The associated eigenvector $\mathbf{v}_{\max}^{\mathbf{C}}$ corresponds to the principal axis in data space along which the splitting of the cluster centers takes place. $\lambda_{\max}^{\mathbf{G}}$ is the largest eigenvalue of a matrix $\mathbf{G}$ whose elements are given by $g_{\mathbf{rt}} = \sum_{\mathbf{s}} h_{\mathbf{rs}} \left(h_{\mathbf{st}} - \frac{1}{N}\right)$. The $\mathbf{r}^{\text{th}}$ component of the corresponding eigenvector $\mathbf{v}_{\max}^{\mathbf{G}}$ determines for each cluster center $\mathbf{w}_{\mathbf{r}}$ in which direction along the principal axis it departs from $\mathbf{w}_{\mathbf{r}}^0$ and how it moves relative to the other cluster centers. These results can be extended to the SSOM algorithm introduced in the previous section. In the above result the matrix $\mathbf{G}$ is simply replaced by $\mathbf{G}^{\mathrm{SSOM}}$ with elements $g_{\mathbf{rt}}^{\mathrm{SSOM}} = h_{\mathbf{rt}} - \frac{1}{N}$. Along similar lines as above we also carried out an analysis of the automatic selection of feature dimensions - so termed by Kohonen [6] - which was later applied by [10, 9] to the formation of cortical maps. This phenomenon occurs, when the map has to perform a dimension reduction and represents the "excess" dimensions by folding. For STVQ, Gaussian neighbourhood function of variance σ_{h}^2 and a continuum approximation with cluster density ρ we obtain for the critical mode $(\mathrm{k}^*)^2$ and for the critical variance in data space $(\sigma^*_{\max})^2$, respectively,

$$(\mathrm{k}^*)^2 = \frac{\beta}{\rho^2} \log\left(1 + \frac{\rho^2}{\beta\,\sigma_{\mathrm{h}}^2}\right), \quad (\sigma^*_{\max})^2 = \left(\frac{1}{\beta} + \frac{\sigma_{\mathrm{h}}^2}{\rho^2}\right)\left(1 + \frac{\rho^2}{\beta\sigma_{\mathrm{h}}^2}\right)^{\frac{\beta\sigma_{\mathrm{h}}^2}{\rho^2}}. \tag{7}$$

In the limit $\beta \to \infty$ and $h_{\mathbf{rs}} \to \delta_{\mathbf{rs}}$ in (5), which corresponds to the traditional SOM, we recover exactly the results of Ritter and Schulten [10], who obtained $\lambda^* = \frac{2\pi}{\mathrm{k}^*} = \sigma_{\mathrm{h}}\pi\sqrt{2} \approx 4.44\,\sigma_{\mathrm{h}}$ for the wavelength of the critical mode and $\mathrm{s}^* = \sigma_{\mathrm{h}}\sqrt{3e/2} \approx 2.02\sigma_{\mathrm{h}}$ for the critical half width of the homogenious data distribution by describing the on-line learning of the SOM by a Fokker-Planck-equation.

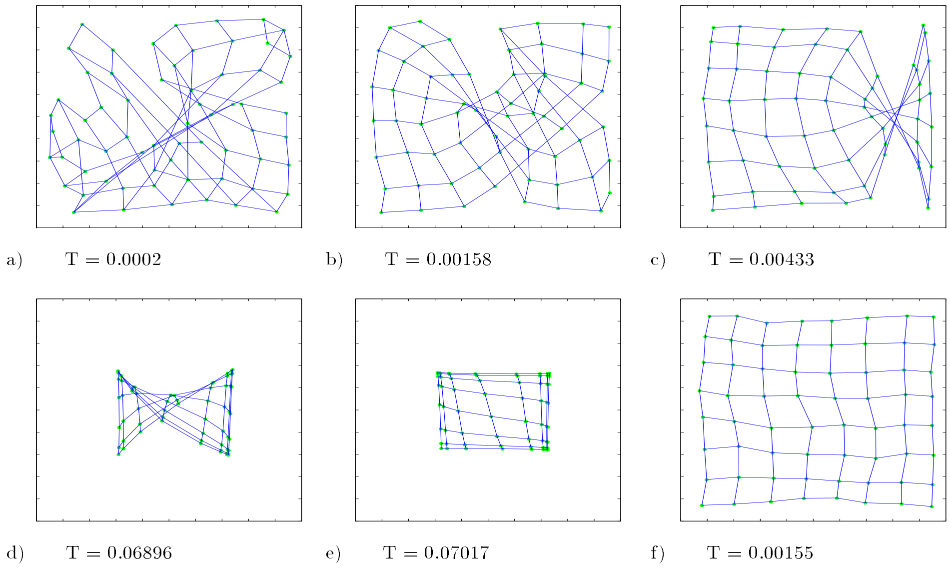

Figure 2: Melting of topological defects for SSOM. The snapshots are taken from an exponential annealing process in T for a net of 8×8 cluster centers. Starting from a random initialization of the cluster centers at $\mathrm{T} = 0.0002$ a local minimum of the cost function is reached (Fig. 2a). The snapshots in Fig. 2b-e are taken at successively higher values of T. Subsequent cooling yielded the ordered configuration in Fig. 2f, where $\mathrm{T} = 0.00155$. σ_{h} was set to 0.5 and the data set consisted of 1024 points equally spaced on a square grid.

4 Numerical Results for a 2-dimensional Net of Clusters

In this section we compare the optimization procedure of STVQ to that of SSOM and TVQ in the simple case of a two-dimensional net of clusters, a Gaussian neighbourhood function and a two-dimensional data space. In order to illustrate the properties of the annealing processes we run it forward ("cooling") and backwards ("heating") using the temperature $\mathrm{T} = 1/\beta$ as the natural parameter for describing the annealing procedure.

Fig. 2 shows how SSOM with annealing overcomes local minima of the cost function (1) at sufficiently high temperatures. Random initialization and a short range neighbourhood function,

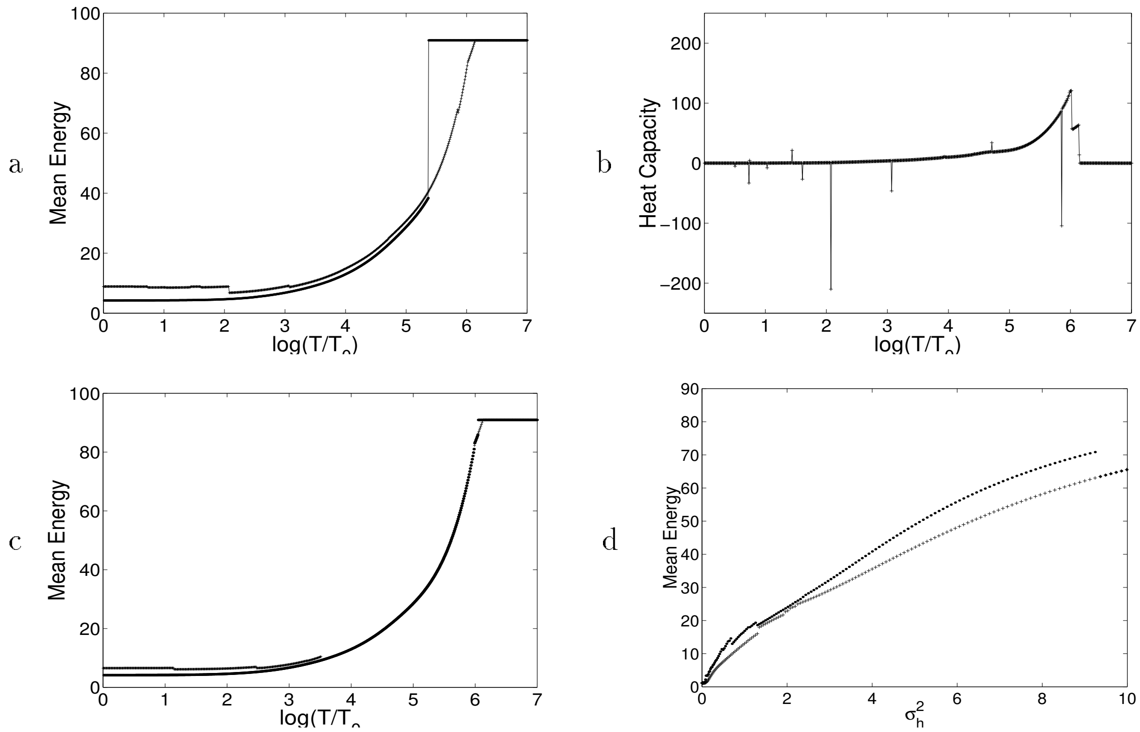

Figure 3: Mean energy and heat capacity for the SSOM (Fig. 3 a,b) are plotted semi-logarithmically vs. temperature T for an exponential annealing scheme. The upper curve on the right in Fig. 3 a shows the cooling phase, where the discontinuity at 5.32 (T = 0.041) arises from a delayed split of the ground state. The upper curve on the left depicts the heating phase, where the observed discontinuities correspond to changes in the topology of the map. In Fig. 3 b, the corresponding heat capacity is shown for the heating phase, where the discontinuity at 5.857 (T = 0.07) is identified as melting of the global twist (cf. Fig. 2 d). In Fig. 3 c $\langle \mathrm{E} \rangle$ is plotted for STVQ, applying the same exponential annealing scheme. In Fig. 3 d for the TVQ case, $\langle \mathrm{E} \rangle$ is plotted vs. the variance σ_h^2 of the neighbourhood function, where the upper curve corresponds to a linear increase in σ_h which shows, how the topological defects induced by the random initialization melt during the annealing. In the lower curve the shrinking of the neighbourhood is depicted, which leads to a perfectly ordered state at $\sigma_\mathrm{h} = 0.02$. The normalization constant was $\mathrm{T}_0 = 0.0002$. Input data were the same as for Fig. 2.

$\sigma_\mathrm{h} = 0.5$, lead the SSOM convergence to a local minimum at low temperature (Fig. 2 a), which corresponds to a so-called topological defect. During heating (Figs. 2 b-d), however, the "twists" gradually unfold until a topologically ordered state (Fig. 2 e) is reached. Subsequent cooling lets this state evolve into the global minimum of the assignment cost (Fig. 2 f). The interpretation of β as a resolution parameter in data space explains why the more localized defects melt first: As the temperature rises data points are soft-assigned to more and more clusters with distant cluster centers.

Figures 3 a,b show the corresponding plots of the average assignment cost $\langle \mathrm{E} \rangle$ and the heat capacity $\mathrm{C}(\mathrm{T}) := \mathrm{d}\langle \mathrm{E} \rangle/\mathrm{dT}$, respectively. The upper curve on the left in Fig. 3 a corresponds to the heating process, and its discontinuities indicate the temperatures at which certain topological defects become unstable and the map folds into a somewhat more ordered configuration. The discontinuities appear as pronounced peaks in the heat capacity (Fig. 3 b) as a result of the sudden decrease of the cost per unit temperature when a defect "unfolds". The upper right curve in Fig. 3 a shows the average assignment cost under re-cooling starting from the ordered configuration. The costs stays constant until the split from the ground state to the ordered configuration occurs indicated by the large step in the energy after which no transitions into local minima occur and the map converges to the global optimum as $T \to 0$. Figure 3 c is analogous to Fig. 3 a for the STVQ which exhibits essentially the same behaviour under deterministic annealing, although the rapid changes in the specific heat occur at lower values of T thus arriving earlier at the ordered state than SSOM. Note, that the global minimum was reached despite σ_{h} remained constant at a small value ($\sigma_{\mathrm{h}} = 0.5$).

Figure 3 d shows the assignment costs as a function of σ_{h} for the commonly used annealing in the width of the neighbourhood function. Increasing and subsequent shrinking of σ_{h} also leads to the unfolding of defects during the increasing phase, and to the global optimum during subsequent shrinking. Note, however, that this σ_{h}-annealing is easy to implement only for certain classes of neighbourhood matrices, like the often used Gaussian case, and does not necessarily carry over to realistic source-channel coding problems as given by (1). Interestingly, in all the numerical simulations so far we have observed that the initial phase transition at high T immediately leads to topologically ordered states, which are then tracked down to the global optimum during annealing for STVQ and SSOM. In summary, our results indicate that deterministic annealing is the method of choice for topographic vector quantization and self-organizing maps. Benchmarks for real-world applications are currently under way.

Acknowledgments

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

References

[1] J. M. Buhmann, and H. Kühnel. *Vector Quantization with Complexity Costs.* IEEE Transactions on Information Theory, vol. 39, p. 1133-1145 (1993).

[2] J. M. Buhmann, and T. Hofmann. *Robust Vector Quantization by Competitive Learning.* Proceedings of the International Conference on Accoustics, Speech and Signal Processing ICASSP'97, Munich, (1997).

[3] E. Erwin, K. Obermayer, and K. Schulten. *Self-Organizing Maps: Stationary States, Metastability and Convergence Rate.* Biol. Cybern., vol. 67, p. 35-45 (1992).

[4] T. Kohonen. *Self-Organized Formation of Topologically Correct Feature Maps.* Biol. Cybern., vol. 43, p. 59-69 (1982).

[5] T. Kohonen. *Analysis of a Simple Self-Organizing Process.* Biol. Cybern., vol. 44, p. 135-140 (1982).

[6] T. Kohonen. *Self-Organizing Maps.* Springer-Verlag, 1995.

[7] S. P. Luttrell. *Self-Organisation: A Derivation from first Principles of a Class of Learning Algorithms.* IEEE International Joint Conference on Neural Networks: IJCNN, Washington DC, vol. 2, p. 495-498 (1989).

[8] S. P. Luttrell. *A Baysian Analysis of Self-Organizing Maps.* Neural Computation, vol. 6, p. 767-794 (1994).

[9] K. Obermayer, G. G. Blasdel, and K. Schulten. *Statistical-Mechanical Analysis of Self-Organization and Pattern Formation during the Development of Visual Maps.* Physical Review A, vol. 45, No. 10, p. 7568-7589 (1992).

[10] H. Ritter, and K. Schulten. *Convergence Properties of Kohonen's Topology Conserving Maps: Fluctuations, Stability, and Dimension Selection.* Biol. Cybern., vol. 60, p. 59-71 (1988).

[11] K. Rose, E. Gurewitz, and G. C. Fox. *Statistical Mechanics and Phase Transitions in Clustering.* Physical Review Letters, vol. 65, No. 8, p. 945-948 (1990).

[12] N. W. Swindale. *The Development of Topography in the Visual Cortex: A Review of Models.* Computation in Neural Systems, vol. 7, p. 161-247 (1996).